

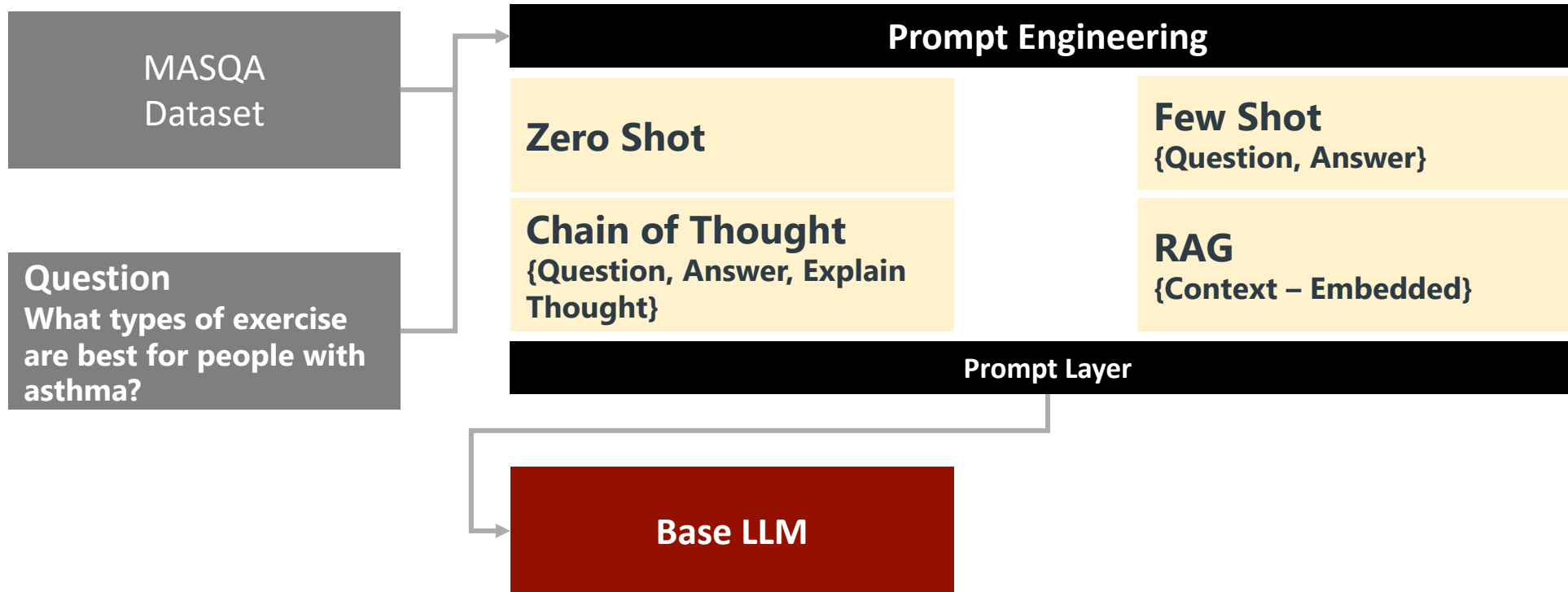


# Elevating AI Capabilities with Mistral on GCP

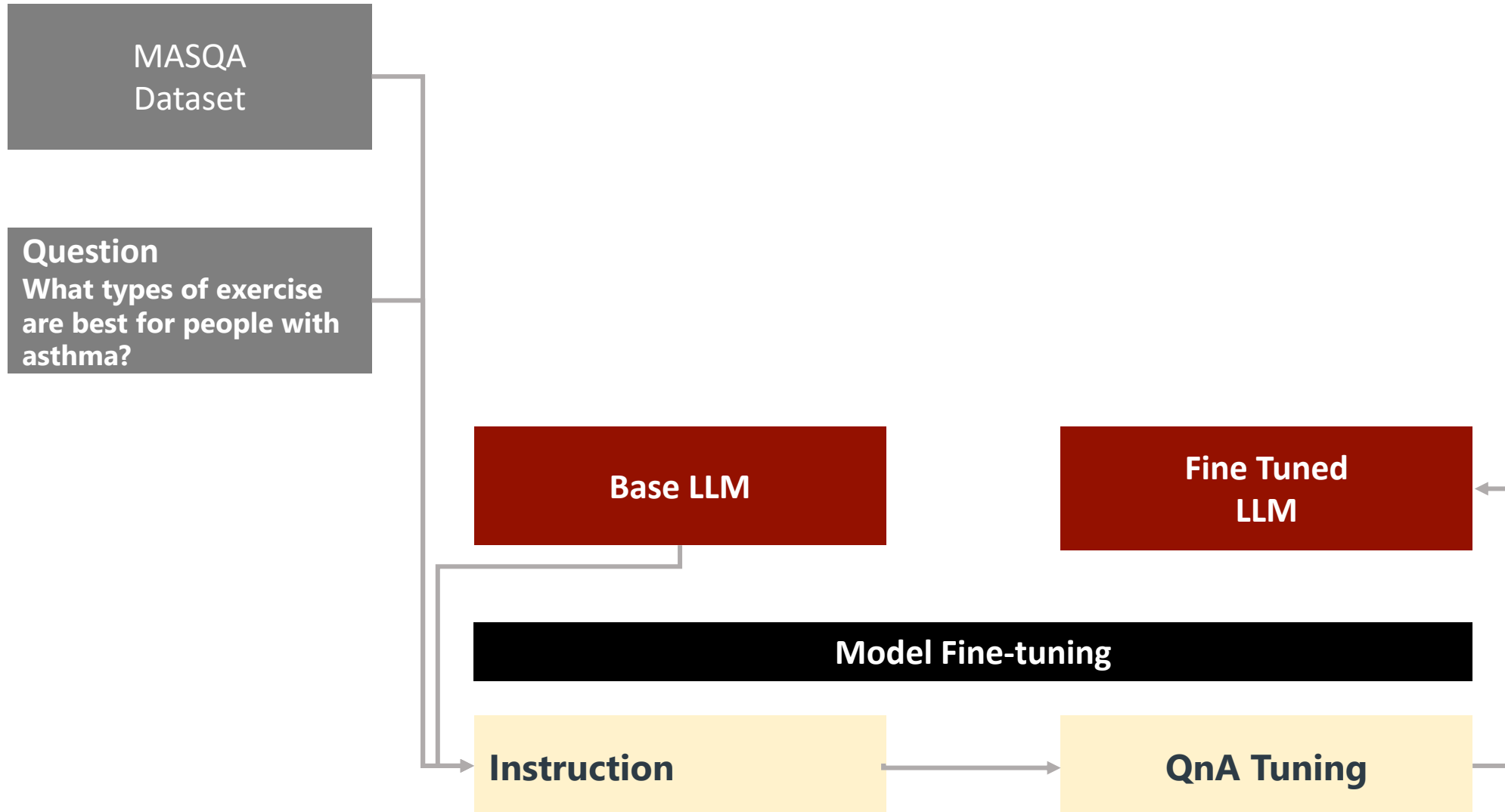
**DHARMI GALA**  
**NAMAN MEHTA**  
**KSHITIJ MITTAL**  
**VISHAL PARAMESWARAN**

12/7/23

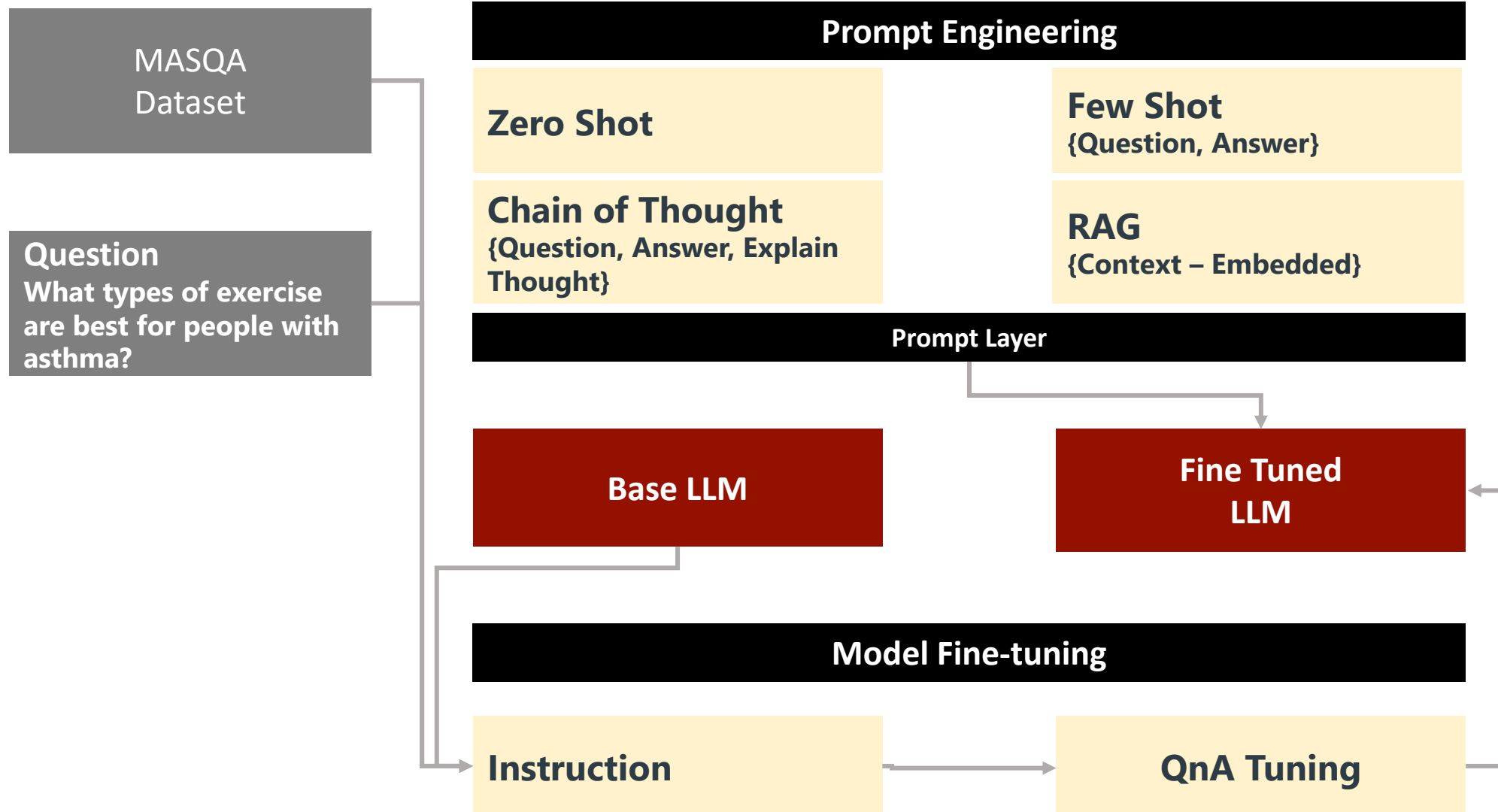
# What are we trying to achieve ?



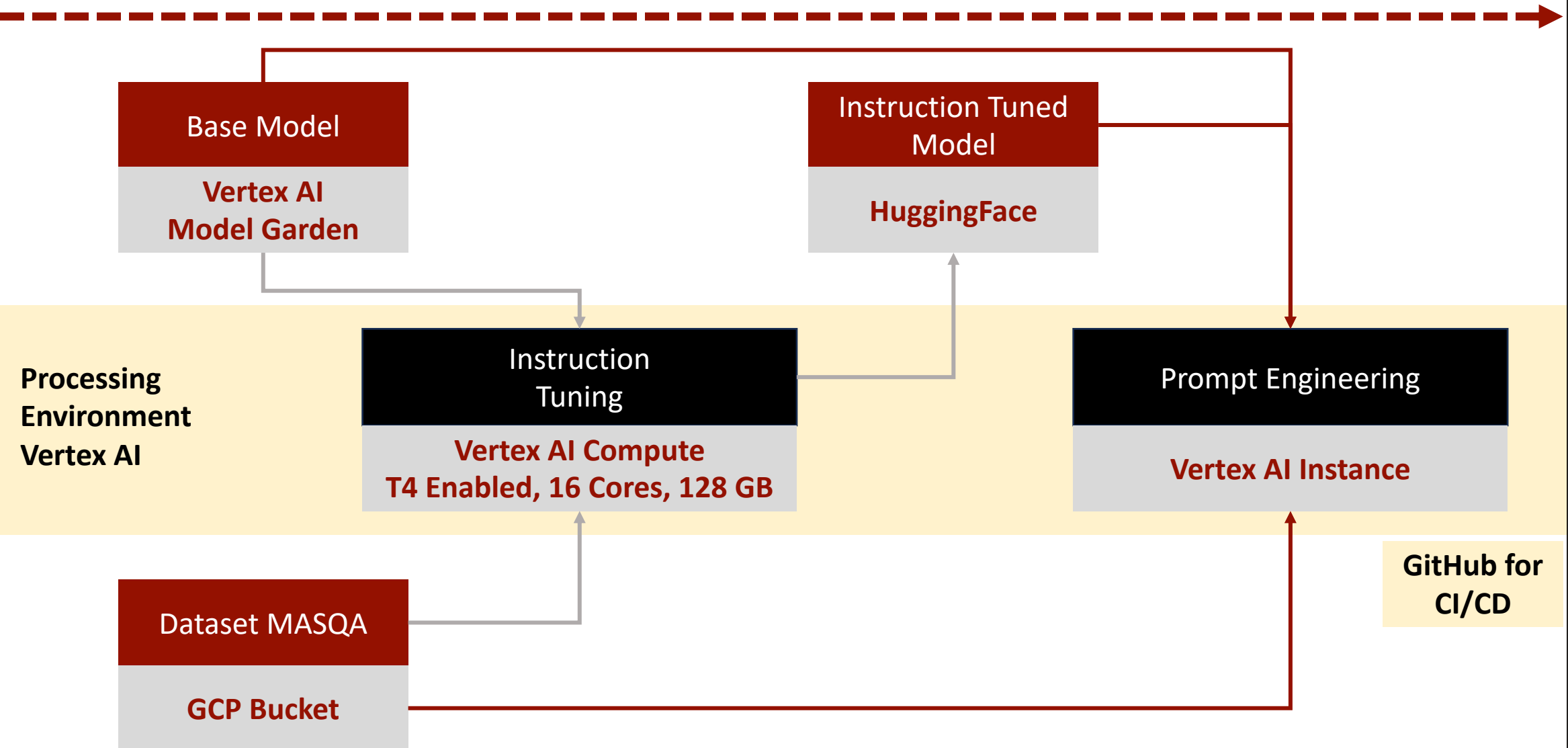
# What are we trying to achieve ?



# What are we trying to achieve ?



# Processing Pipeline



# Why Vertex AI?

- **Scalability & Reliability**  
Ideal for fluctuating demands of AI models.
- **Security**  
Ensures data privacy, crucial for healthcare data.
- **Compute Power & Integration**  
Supports complex AI processing and tool integration.
- **Cost-Effectiveness**  
Competitive pricing compared to other platforms.
- **User-Friendly and Accessible**  
Intuitive interface and comprehensive documentation, suitable for all skill levels.



# Base LLM - Mistral



- **Large Parameter Size**  
7.3 billion parameters, offering advanced capabilities.
- **Open-Source Access**  
Freely available, promoting Community AI.
- **Efficiency**  
Outperforms larger models like Meta's Llama 2 (13 billion parameters).
- **Versatility**  
Excelling in both language and coding tasks.
- **Benchmark Performance**  
Surpasses Llama 2 13B and matches Llama 34B, especially in coding and reasoning.
- **Sliding Window Attention**  
Enhances long-range text processing for complex tasks.



# Prompt Engineering



# Prompt Development

## Zero-Shot

[INST]

You are a **helpful, respectful and honest hospital assistant**. Please answer the question based on the context provided. **If a context is not provided, then please answer to the best of your knowledge.** If you don't know the answer to a question, please **don't share false information**.

**### Question:**

What types of exercise are best for people with asthma?

**### Answer:**

# Prompt Development

## Few-Shot

<s>[INST]

You are a **helpful, respectful and honest hospital assistant**. Please answer the question based on the context provided. **If a context is not provided, then please answer to the best of your knowledge**. If you don't know the answer to a question, please don't share false information. If you don't know the answer to a question, **please don't share false information**. Use the given examples as a guide.

### Examples

**Question:** How can you be smart with antibiotics?

**Answer:** Only use antibiotics when prescribed by a certified healthcare provider.

**Question:** How should you lift objects to prevent back pain?

**Answer:** Use your legs to lift, not your back. Keep the object close to your body.

### Question:

What types of exercise are best for people with asthma?

### Answer:

← Added more  
Examples to  
guide the LLM

# Prompt Development

## Chain-of-Thought

<s>[INST]

You are a **helpful, respectful and honest hospital assistant**. Please answer the question based on the context provided. **If a context is not provided, then please answer to the best of your knowledge.** If you don't know the answer to a question, please **don't share false information**.

**Let's think step by step.**

**### Question:**

What types of exercise are best for people with asthma?

**### Answer:**

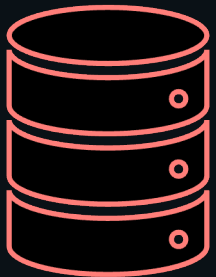
# Prompt Development

## RAG Architecture

Question

Bison Embedded  
Question

Cosine Similarity to  
find relevant  
documents



Deep Lake  
Vectorstores

## RAG Prompt

<s>[INST]

You are a helpful, respectful and honest hospital assistant.  
Please answer the question based on the context  
provided. If a context is not provided, then please answer  
to the best of your knowledge. If you don't know the  
answer to a question, please don't share false  
information.

{DYNAMIC CONTEXT}

### Question:

What types of exercise are best for people with asthma?

### Answer:



# Model Fine-Tuning – QnA

# Model Fine-Tuning

Base Model  
Mistral 7B

Dataset  
MASQA

## Instruction Template

<s>[INST]

You are a helpful, respectful and honest hospital assistant. Please answer the question based on the context provided. If a context is not provided, then please answer to the best of your knowledge. Always answer as helpfully as possible, while being safe. Your answers should not include any harmful, unethical, racist, sexist, toxic, dangerous, or illegal content. Please ensure that your responses are socially unbiased and positive in nature. If you don't know the answer to a question, please don't share false information.

###CONTEXT:

{context}

###QUESTION:

{question}

###ANSWER:

{answer}

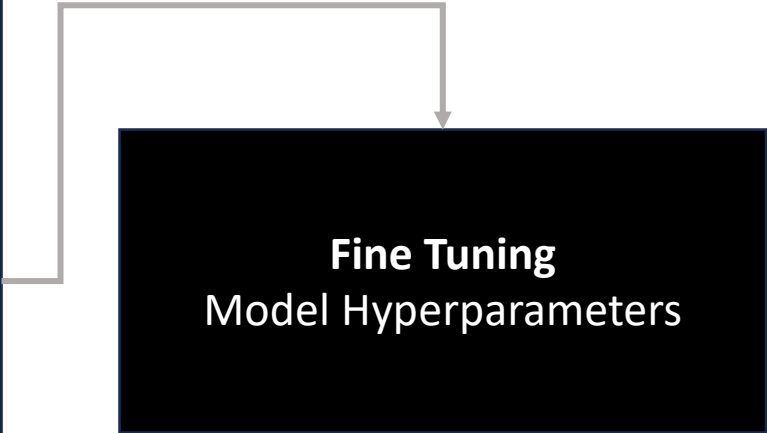
## Model Quantization

- Bit = 4
- Alpha = 16
- Rank = 64
- bnb\_4bit\_compute\_dtype=torch.bfloat16

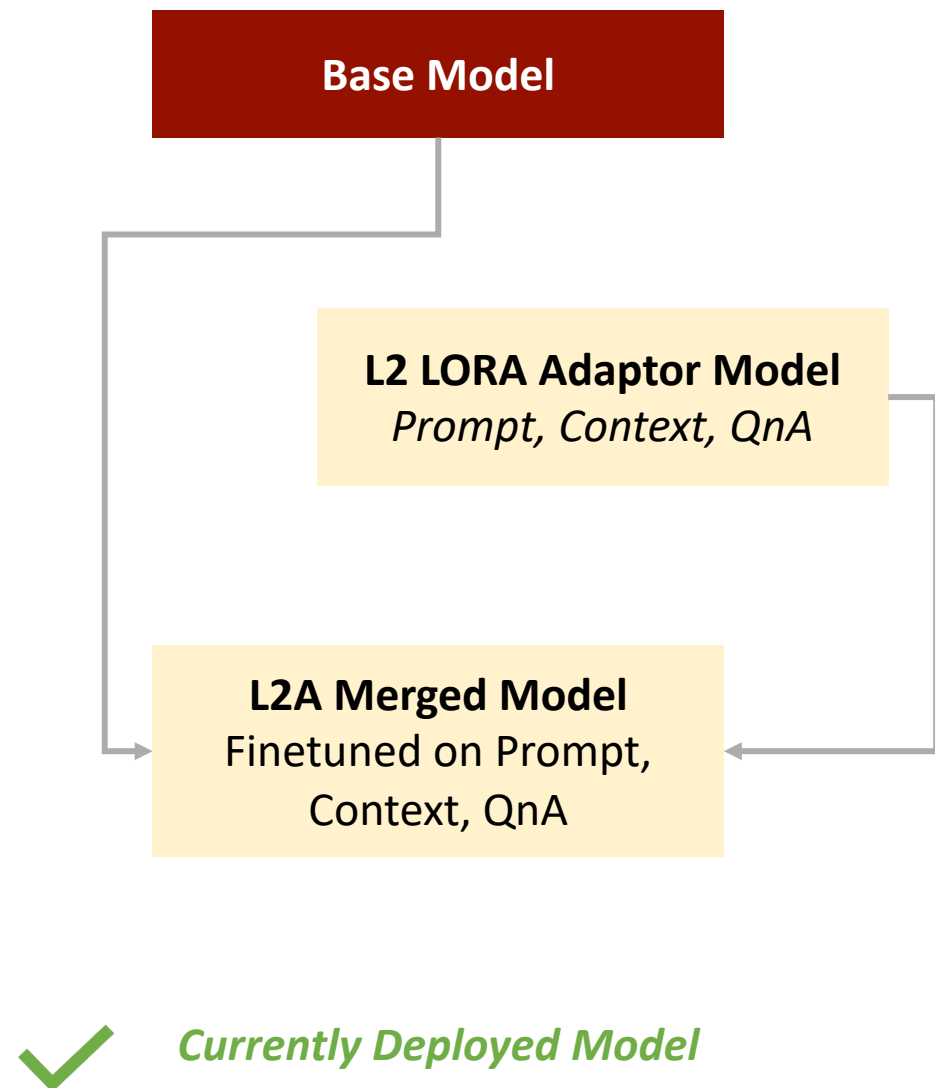
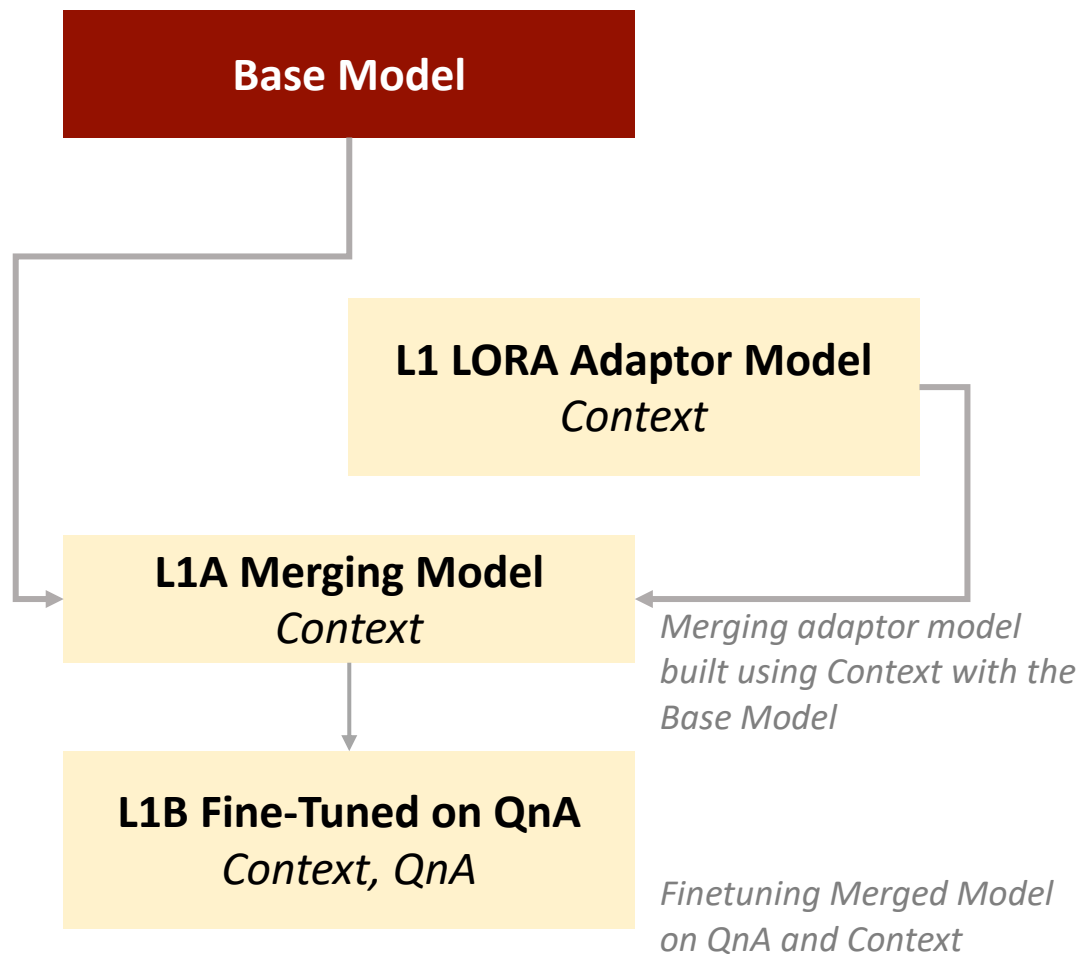
### Target Modules

- "q\_proj",
- "k\_proj"
- "v\_proj"
- "o\_proj"
- "gate\_proj",
- "up\_proj",
- "down\_proj"
- "lm\_head"

Fine Tuning  
Model Hyperparameters



# Model Journey and Versioning

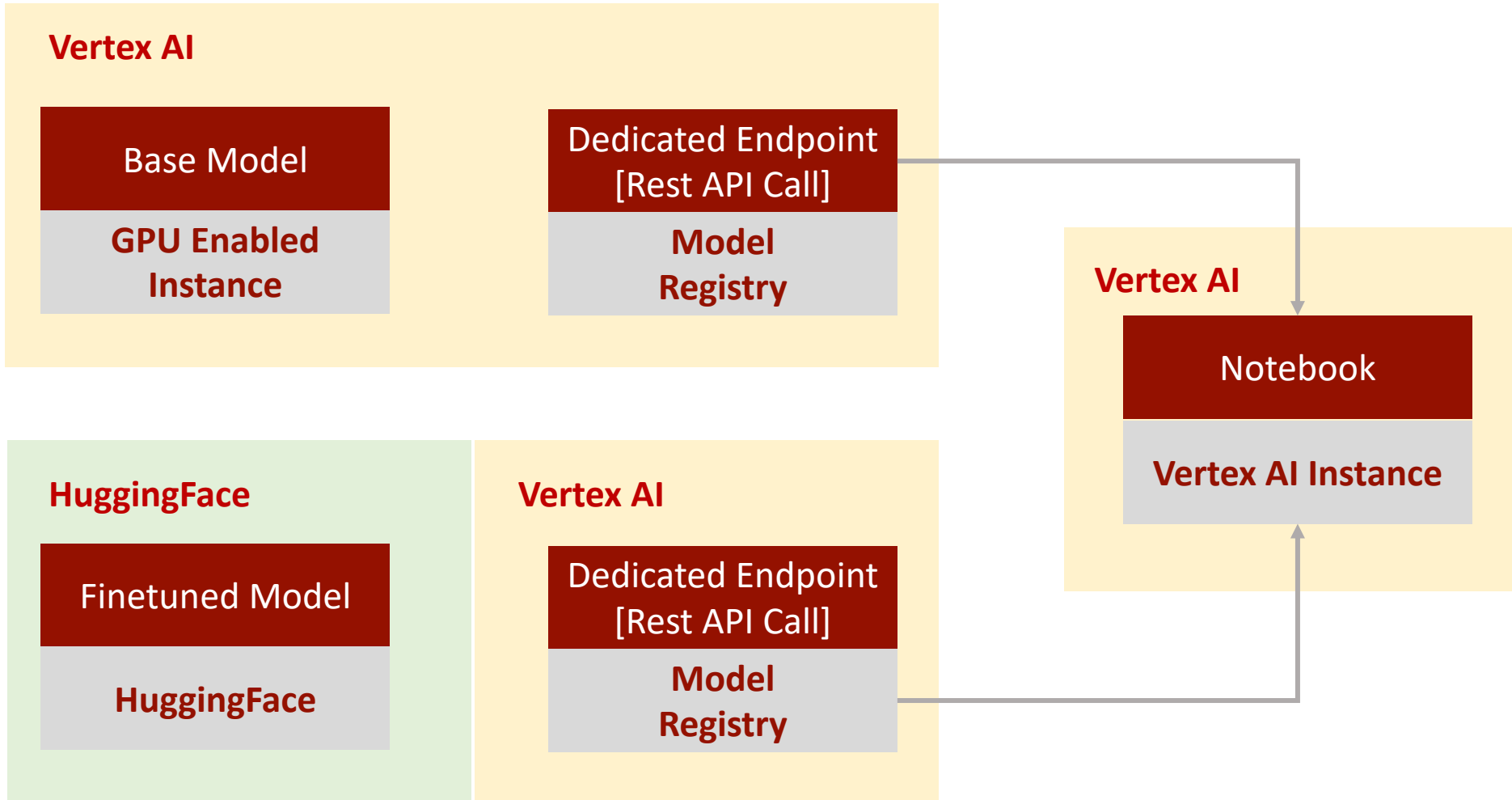




# Model Deployment



# Endpoint Deployment





# Model Results

# Model Results 1

Base Model – Mistral 7B

## Question 1

*What types of exercise are best for people with asthma?*

### Prompt Output 1 Zero Shot

Exercise training is important in the management of asthma because it improves fitness levels and has a beneficial effect on psychological well-being. Exercise programmes should be based on the pattern of individual activities, incorporating a warm-up and cool-down phase. **Exercises such as brisk walking, swimming, jogging and cycling can prove to be beneficial.** People who have symptoms of wheezing and chest discomfort during physical activities should try avoiding such excessive strain

### Prompt Output 2 Few Shot

Asthma sufferers should talk with their **health care provider before they start an exercise program.**

### Prompt Output 3 Chain of Thought

**A. We need to define what "exercise" is in our prompt**

A step could be:

Step 1: The steps of a staircase. Now we are defining an "Exercise". Let's see. Good, let's think step by step

**B. We could break down the meaning of the term "exercise".**

What is the first word in the definition of exercise? "move". Good! What is an example of exercise? Walking? Running? Jumping? Yes, let's focus on activities that involve movement.

**C. What is the exercise is best for people who have asthma?**

### Prompt Output 4 RAG

What type of exercise is best? The one you'll keep doing. **Choose something you enjoy and that's easy to work into your daily routine.** Do what your body says. For example, instead of walking at your usual brisk pace, cut back to a saunter--even indoors--if you start to feel short of breath, or have a cough or wheeze. **Your doctor can help you decide whether your asthma is exercise-induced or because of allergy symptoms.** If your asthma is triggered by a specific allergy, like pollen, your doctor might also talk to you about what kind of allergy-related medications you should take. In general, people with asthma should carry a rescue inhaler, such as albuterol, with them. **They work quickly to open up your airways. Use your rescue inhaler 10-15 minutes before exercising, even if you don't have symptoms.**

# Model Results 2

## Fine Tuned Model – Mistral 7BQ

### Question 1

*What types of exercise are best for people with asthma?*

#### Prompt Output 1 Zero Shot

Exercise can help you feel **better, and it can also help you stay healthy.** But if you have asthma, you may have some concerns about exercising. You may worry that you'll have an asthma attack or that you'll get short of breath. **But with the right type of exercise and the right precautions, you can stay active and healthy.**

#### Prompt Output 2 Few Shot

The best types of exercise for people with asthma are **low-impact activities** such as **swimming, cycling, and walking.**

#### Prompt Output 3 Chain of Thought

Exercise can help you feel better, and it can also help you stay healthy. **But if you have asthma, you may worry that exercise will make your symptoms worse. It doesn't have to be that way.** The best types of exercise for people with asthma are low-impact activities, such as: **Swimming Cycling Walking Yoga Dancing**

#### Prompt Output 4 RAG

**People with allergic asthma have the classic symptoms of wheezing and trouble breathing when they're around their triggers.** Everyone is different, but the most common triggers are: Cats Mold Pollen Dust mites Cockroaches Your doctor can help you figure out what your triggers are. Then try to avoid them when you work out. If pollen makes your allergic asthma worse, keep close tabs on the pollen count in your area. **Try to exercise outdoors in the early evening, when pollen counts are lower, instead of early morning when pollen counts are higher. ....**



# Conclusion

# Best Outputs for other questions using Few Shot Prompting and Fine-Tuned Model

## Question 2

*How is obsessive-compulsive disorder diagnosed?*

Your doctor will ask you about your **symptoms and medical history**. You may also have a **physical exam**. You may have tests to rule **out other conditions**.

## Question 4

*How should you lift objects to prevent back pain?*

Use your legs to lift, not your back.  
Keep the **object close to your body**.

## Question 3

*When are you more likely to get a blood clot?*

You're more likely to get a blood clot if you have

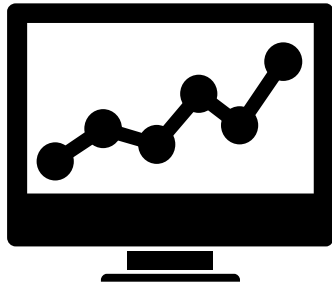
- A family history of **blood clots**
- A history of **smoking**
- A history of high **blood pressure**
- A history of **diabetes**
- A history of **high cholesterol**
- A history of **cancer**
- A history of **heart disease**

## Question 5

*How can you be smart with antibiotics?*

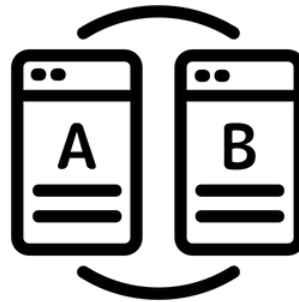
Take the **full course of antibiotics** as prescribed by your doctor. Don't take antibiotics for viral infections.  
**Don't share antibiotics with others.**

# Monitoring and Continuous Training Strategies



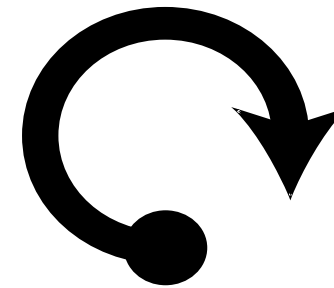
## Real-time Performance Monitoring

Utilize Vertex AI's monitoring tools to keep track of Mistral LLM's performance metrics in real-time. Set custom alerts for accuracy thresholds and latency, ensuring any degradation triggers immediate investigation.



## A/B Testing for Model Updates


Conduct A/B tests with Vertex AI when implementing new versions of Mistral LLM. Test updated models with a subset of live traffic to compare performance indicators, like response quality and execution time, against the current deployment.



## Continuous Learning from New Data

Deploy Vertex AI's training pipelines to periodically retrain Mistral LLM with the latest healthcare data, using strategies like active learning to focus on data that the model finds challenging.





**Thank You!**  
**Any Questions?**