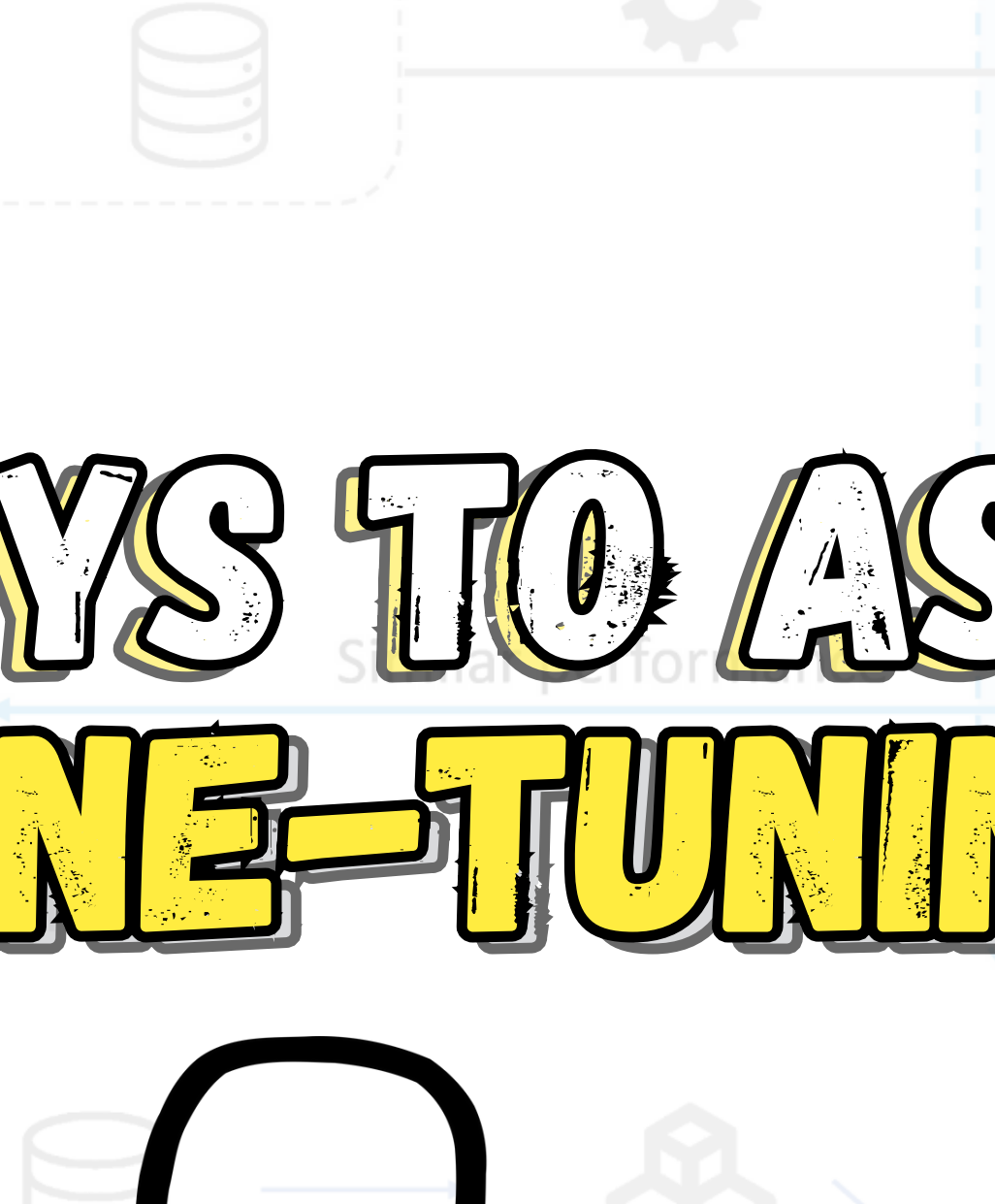
4 WAYS TO ASSESS FINE-TUNING



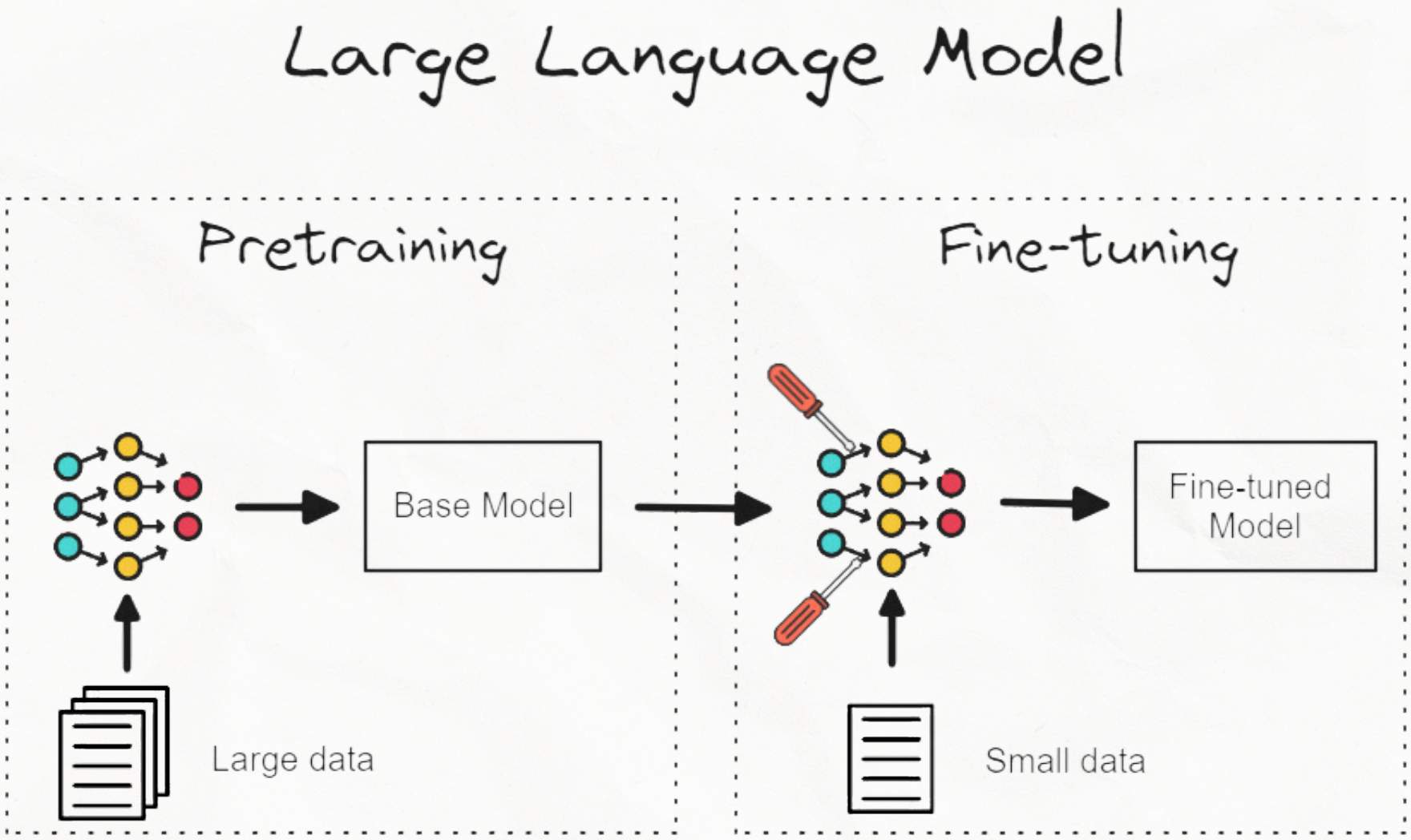
Bhavishya Pandit

INTRODUCTION

What does it mean to fine tune a model?

-Fine-tuning is the process of training a pre-trained model on a smaller, specific dataset to improve its performance for a particular task.

A model is first **pre-trained** on a **large dataset** to learn general knowledge. Then, it is **fine-tuned** on a **small dataset** focused on a specific domain, like medical or legal texts. The result is a **fine-tuned model** that gives more accurate answers in that field while still retaining its general knowledge.



Finetuning an LLM

Whenever we query a fine-tuned model, how do we know if the response is based on the new knowledge or the foundation knowledge? Let's discuss -

source: https://medium.com/@prasadmahamulkar/fine-tuning-phi-2-a-step-by-step-guide-e672e7f1d009

1. RESPONSE CONSISTENCY

One of the simplest ways to check if a fine-tuned LLM is using its newly trained data is by **analyzing response consistency**.

If the model starts producing answers that introduce new explanations, terminology, or details it previously lacked, it's a **strong indicator** that it has **learned** from the fine-tuning process.

e.g. What is RAG?

Before finetuning

A method combining retrieval and generation for LLMs.

VS

After finetuning

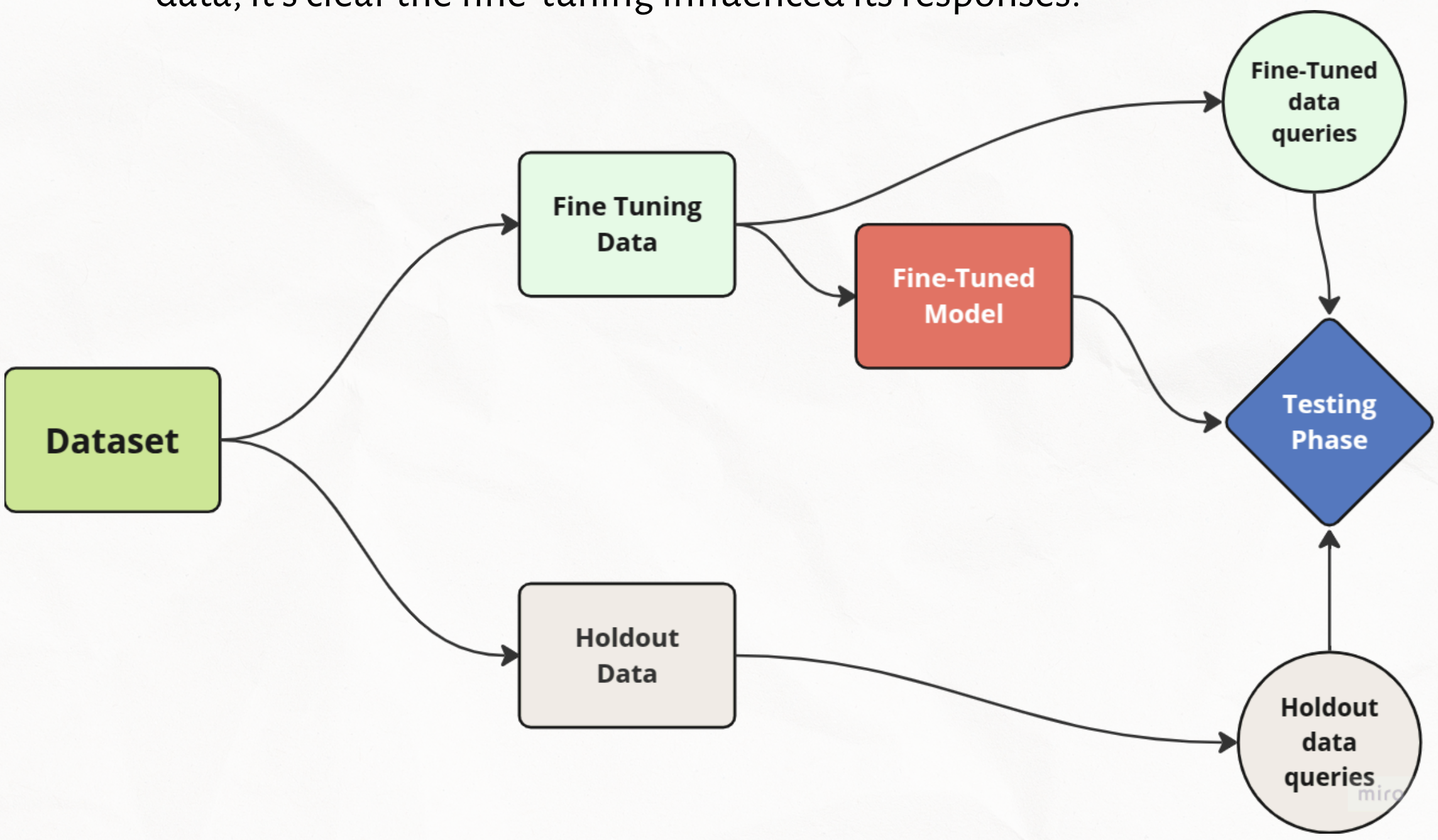
RAG integrates vector-based retrieval to fetch relevant documents before generating a response, improving factual accuracy.

2. USE HOLDOUT DATA

A simple way to check if an LLM is using newly fine-tuned data is by using holdout data—a portion of the dataset excluded from fine-tuning.

How It Works

1. Fine-tune the model on a subset of new data.
2. Keep a separate holdout set that the model never sees.
3. Test the model on both:
    - If it answers well using the fine-tuned subset but struggles with the holdout data, it's clear the fine-tuning influenced its responses.

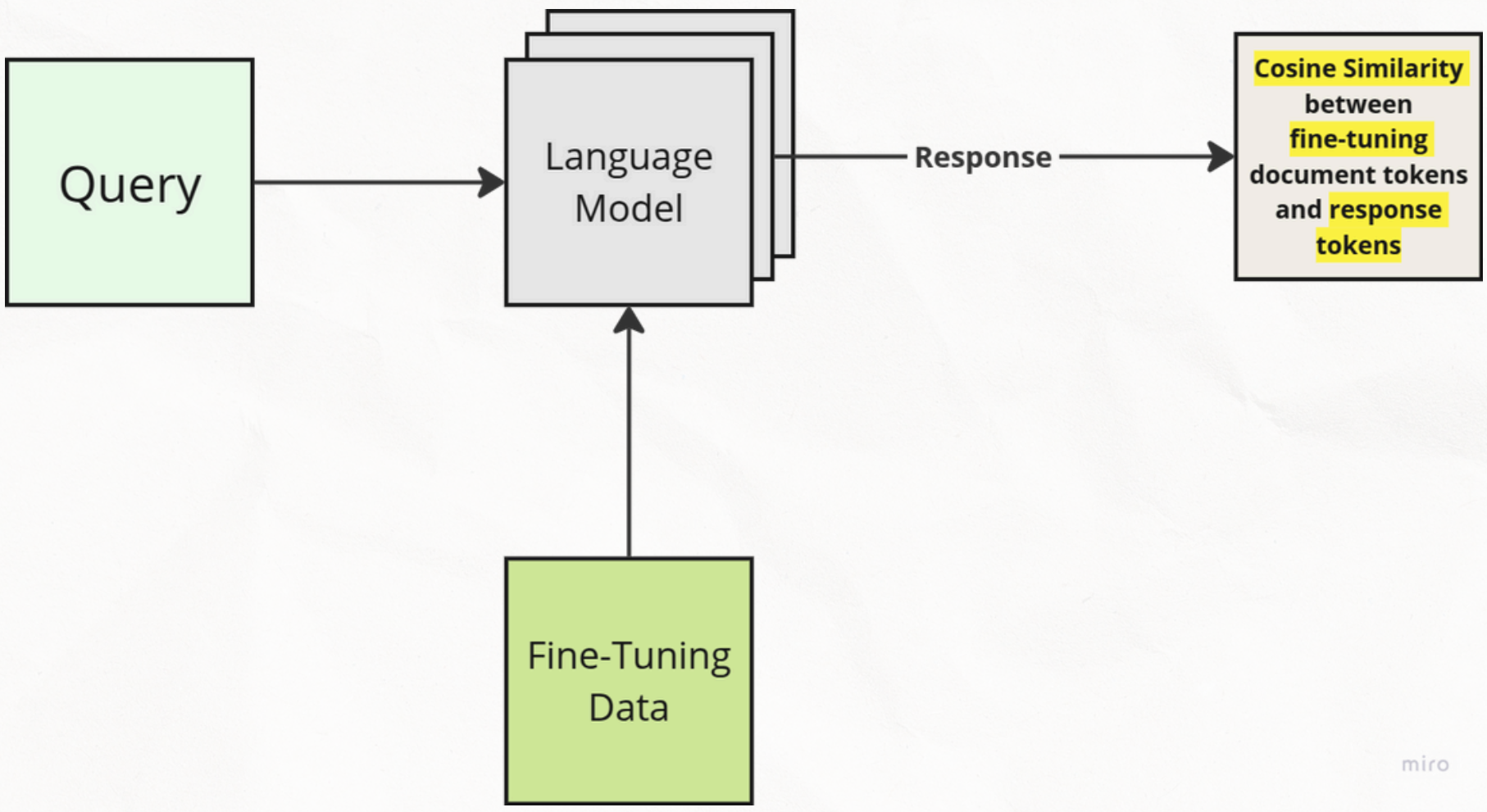


3. TOKEN ATTRIBUTION

Token Attribution analyzes which parts of the input the model focuses on when generating a response.

How Token Attribution Works -

- Fine-tuning **updates model weights** based on new data.
- Attribution tools (like attention mapping or gradient-based methods) can highlight whether the model prioritizes tokens from this fine-tuned data when forming answers.
- If a model assigns **higher attention** scores to words or phrases **from** the **fine-tuned** dataset, it's clear that it's relying on that knowledge.



The approach maps answer tokens to document sections using cosine similarity to identify key influences.

source: https://arxiv.org/html/2405.17980v1

4. BENCHMARKING

A fairly straight forward way is to **benchmark** your fine-tuned model against a custom dataset that measures the accuracy of responses.

FineTuneBench is one such evaluation framework for understanding how well fine-tuned models can successfully learn **new and updated knowledge**.

source: https://arxiv.org/html/2411.05059v2

Follow to stay updated on Generative AI

LIKE

COMMENT

REPOST