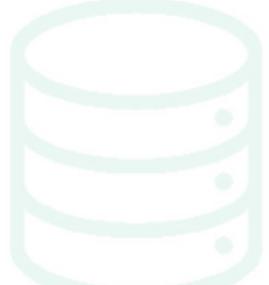


Original training dataset

Remaining data



Removed data



Original training



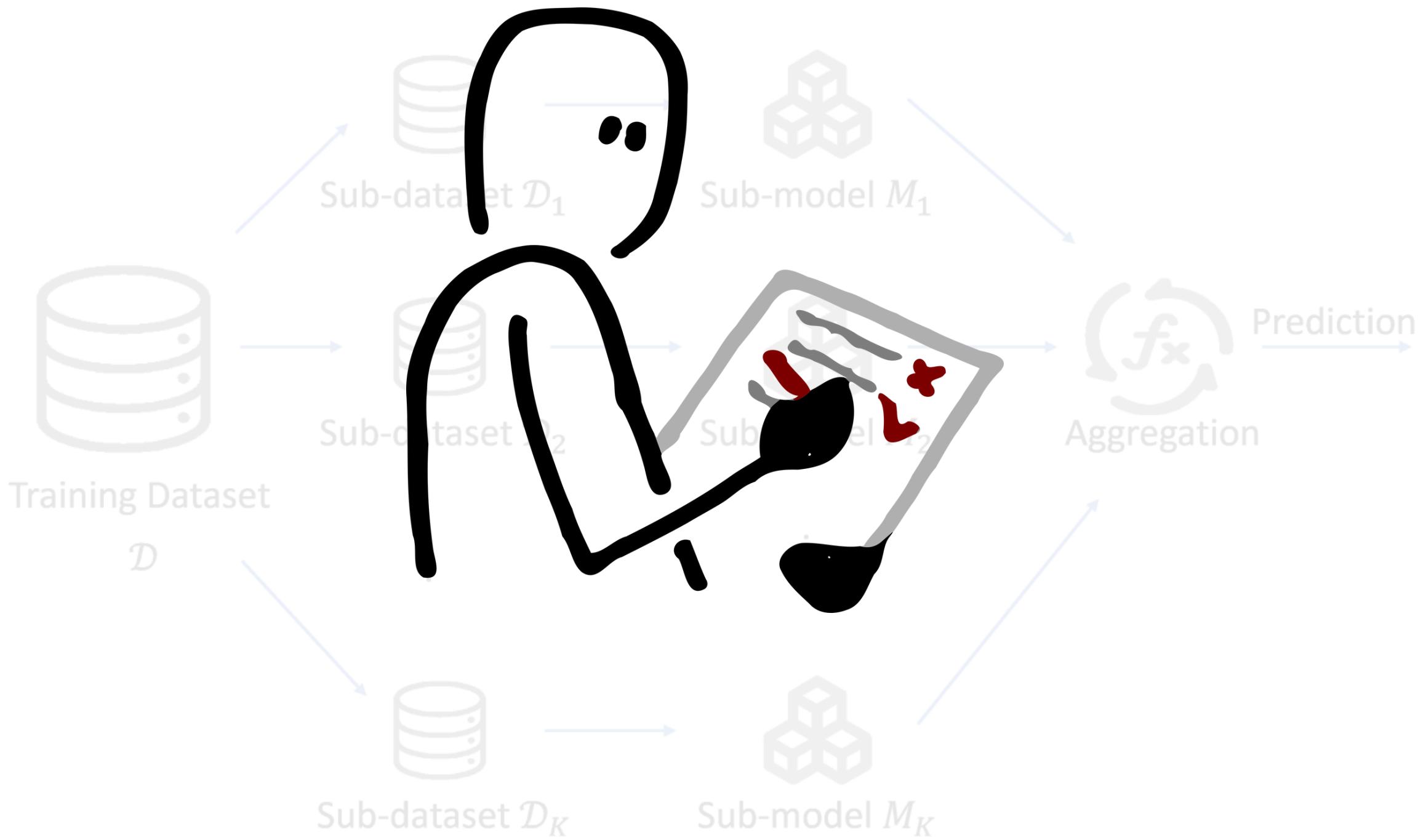
Original Model



Naive  
train



# 4 WAYS TO ASSESS FINE-TUNING

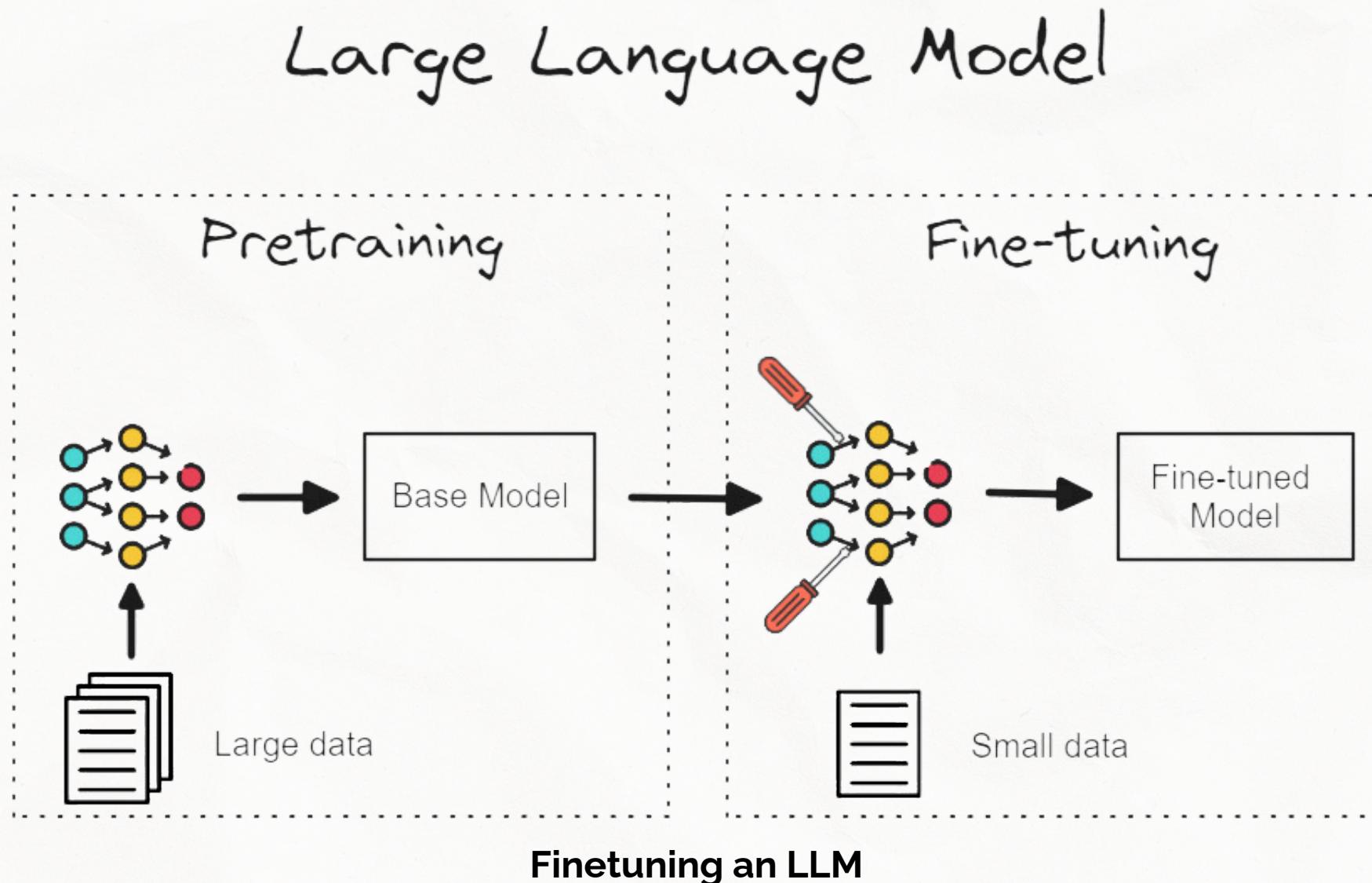


# INTRODUCTION

## 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.



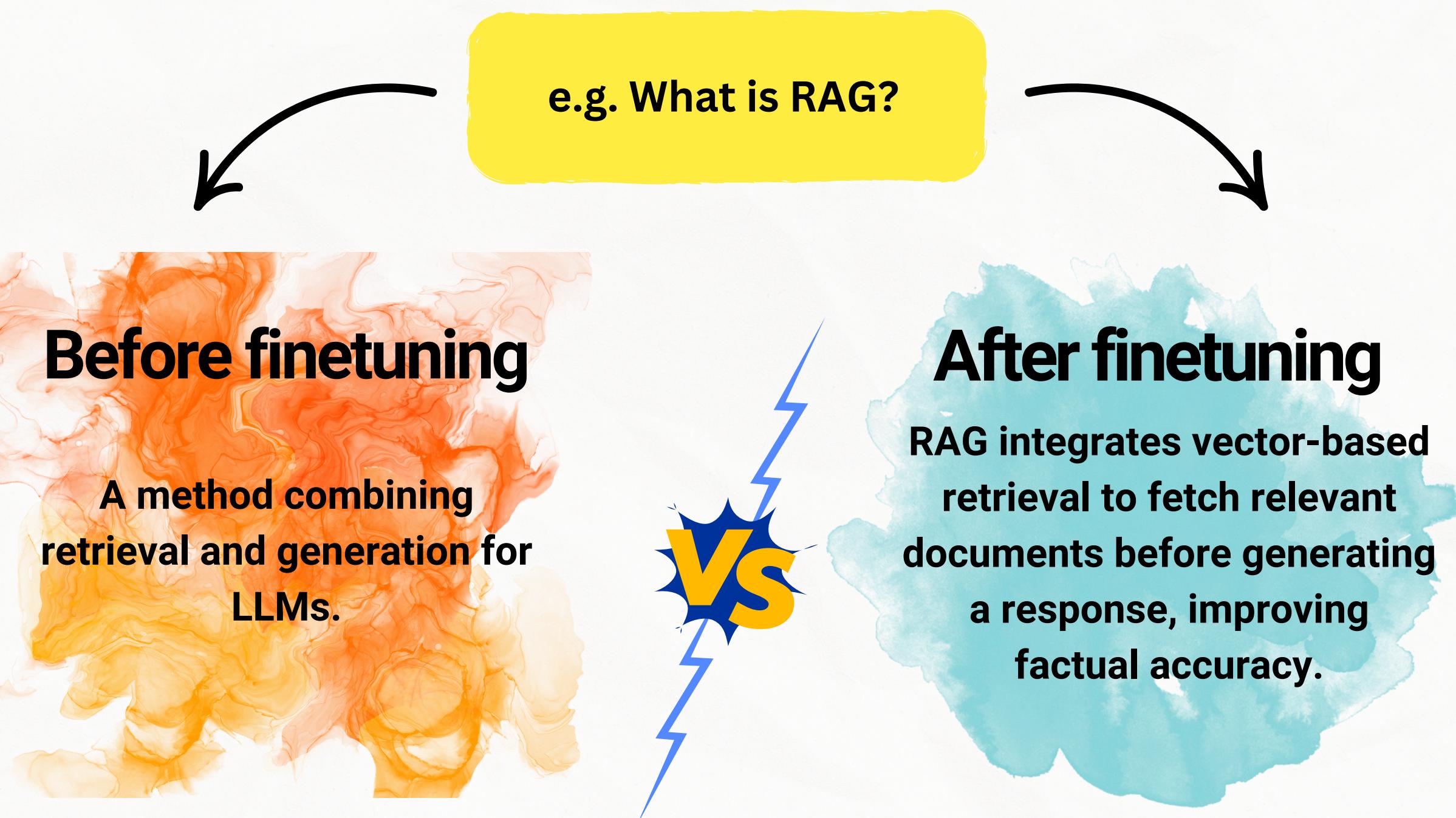
Whenever we query a fine-tuned model, how do we know if the response is based on the new knowledge or the foundation knowledge? Let's discuss -

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.

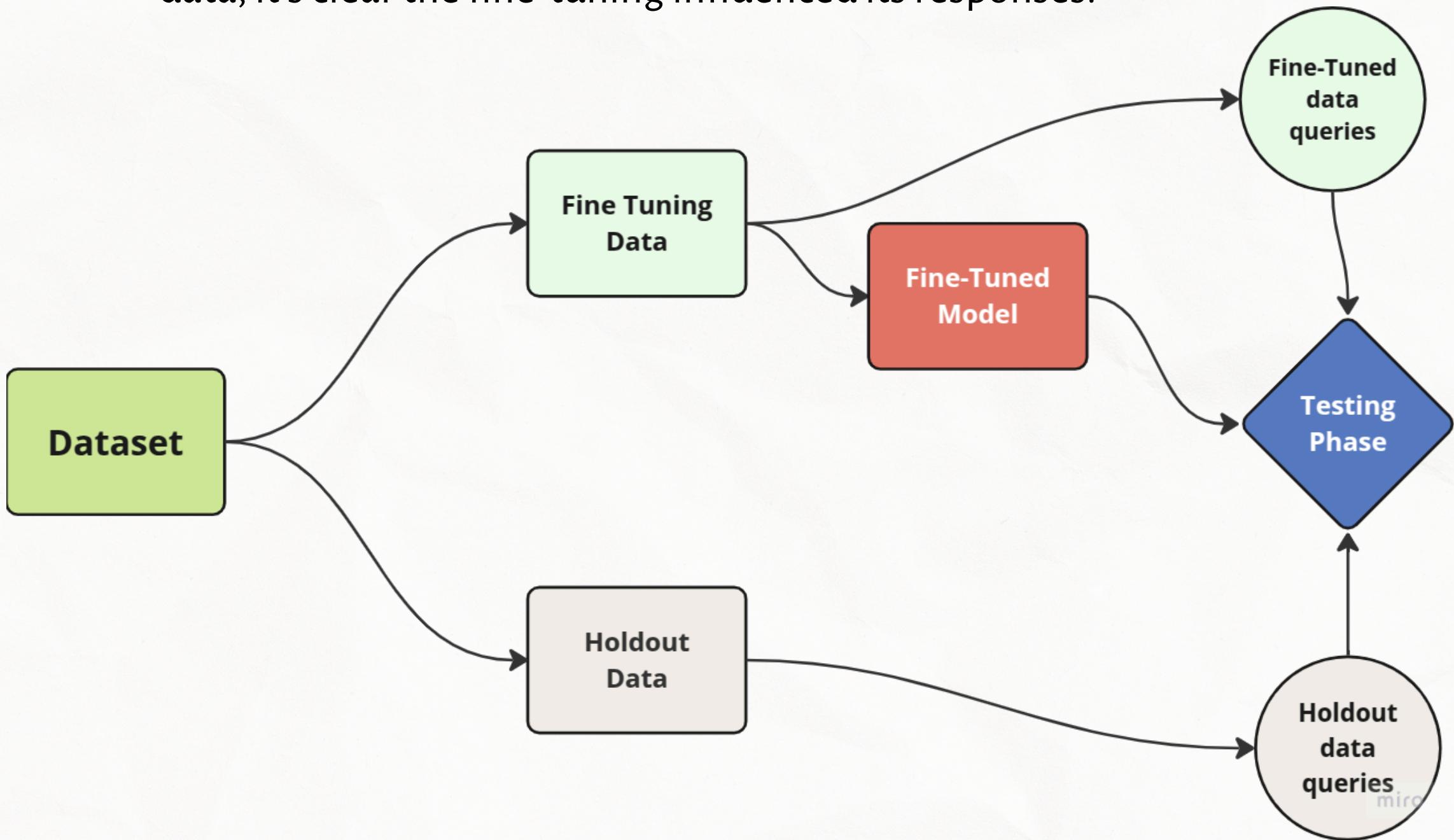


# 2. USE HOLDOUT DATA

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.

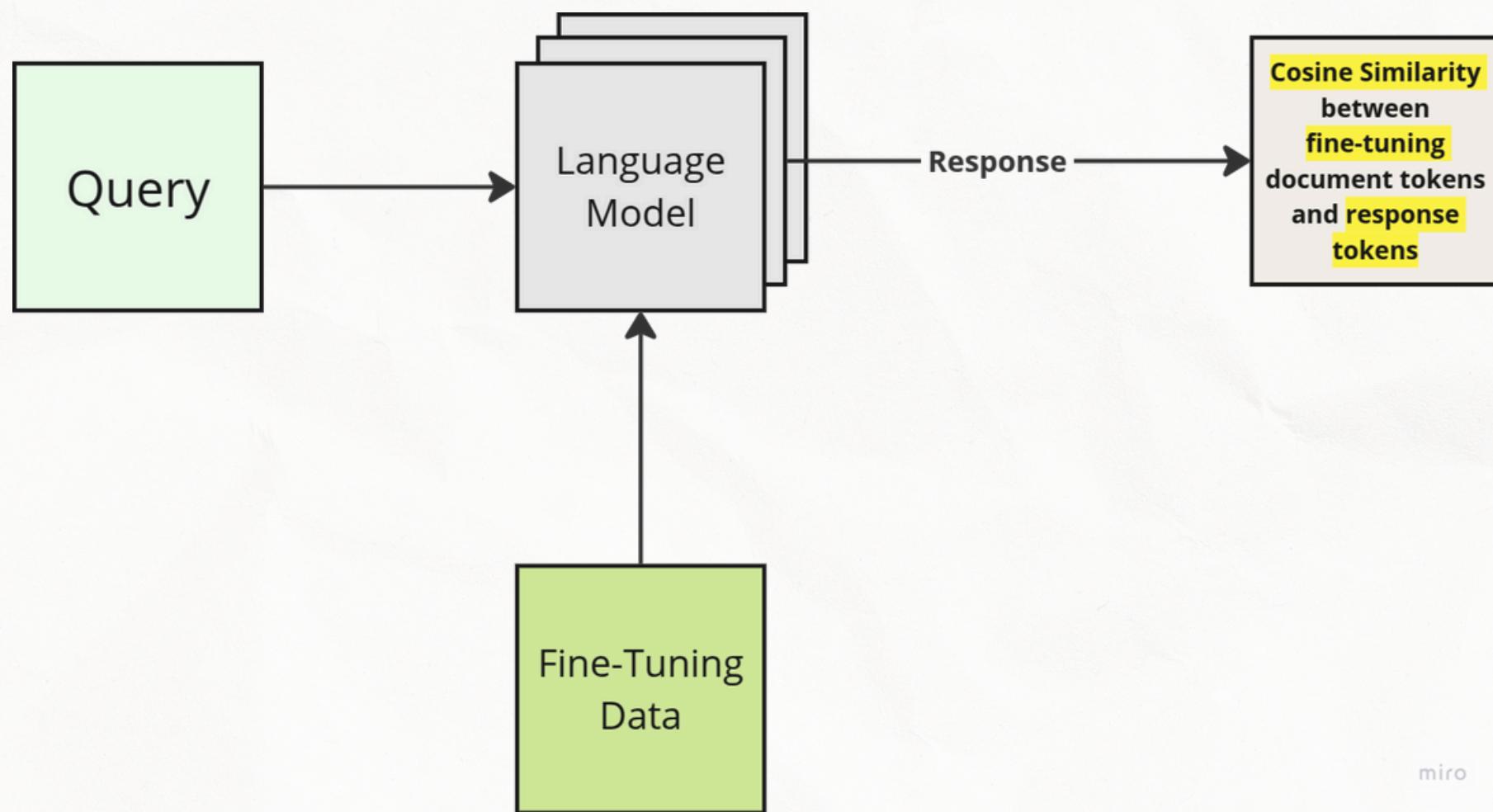


# 3. TOKEN ATTRIBUTION

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**.

A

Commercial fine-tuning services



Create a fine-tuned model

Base Model: Select...

Training data: Upload new or drag and drop here (20MB)

Validation data: Add a file to use for validation metrics. Upload new or Select existing.

Seed: The seed controls the reproducibility of the job. Placing it the same seed and all parameters should produce the same results, but may differ on new seeds. If a seed is not specified, one will be generated for you.

Configure hyperparameters: Batch size, Learning rate multiplier, Number of epochs.

Learn about fine-tuning > Cancel Create



Tuning can improve the performance of a model for specific tasks. Model tuning can also help it adhere to specific output requirements when instructions aren't sufficient. Learn more >

Tuning method:

- Supervised tuning: Uses labeled examples to teach a model to exhibit a desired behavior or task. Suitable for well-defined tasks like classification, sentiment analysis, extraction, and generation.
- Reinforcement learning from human feedback (RLHF): Uses human feedback to align a model with human preferences and reduce undesired outputs by increasing where people have complex intuitions about a task.

Model details: Tuned model name:  Base model:  Region:  Tuning setting: Number of epochs:  Learning rate multiplier:  Adapter size:  Show less Continue

Fine-tune and evaluate on **FineTuneBench** datasets



Latest News



Fictional People



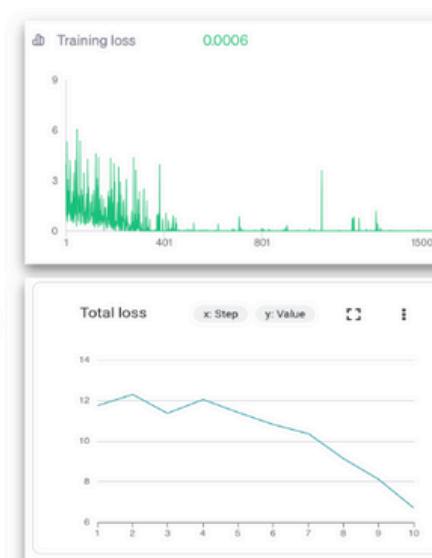
Medical Guidelines



Code

} New Knowledge

} Updating Knowledge



B

Training data

Q: Who hit a two-run homer and drove in three runs for the Cleveland Guardians on September 24, 2024?  
A: Lane Thomas

Baseline model

Modify

Rephrased

Q: Which player for the Cleveland Guardians hit a two-run home run and had three RBIs on September 24, 2024?

Modify

Date Changed

Q: Who hit a two-run homer and drove in three runs for the Cleveland Guardians on September 24, 2025?

Train

Fine-tuned model

A: Josh Naylor

A: Lane Thomas

(Model overfits to the training data)

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



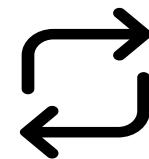
**Follow to stay updated on  
Generative AI**



**LIKE**



**COMMENT**



**REPOST**