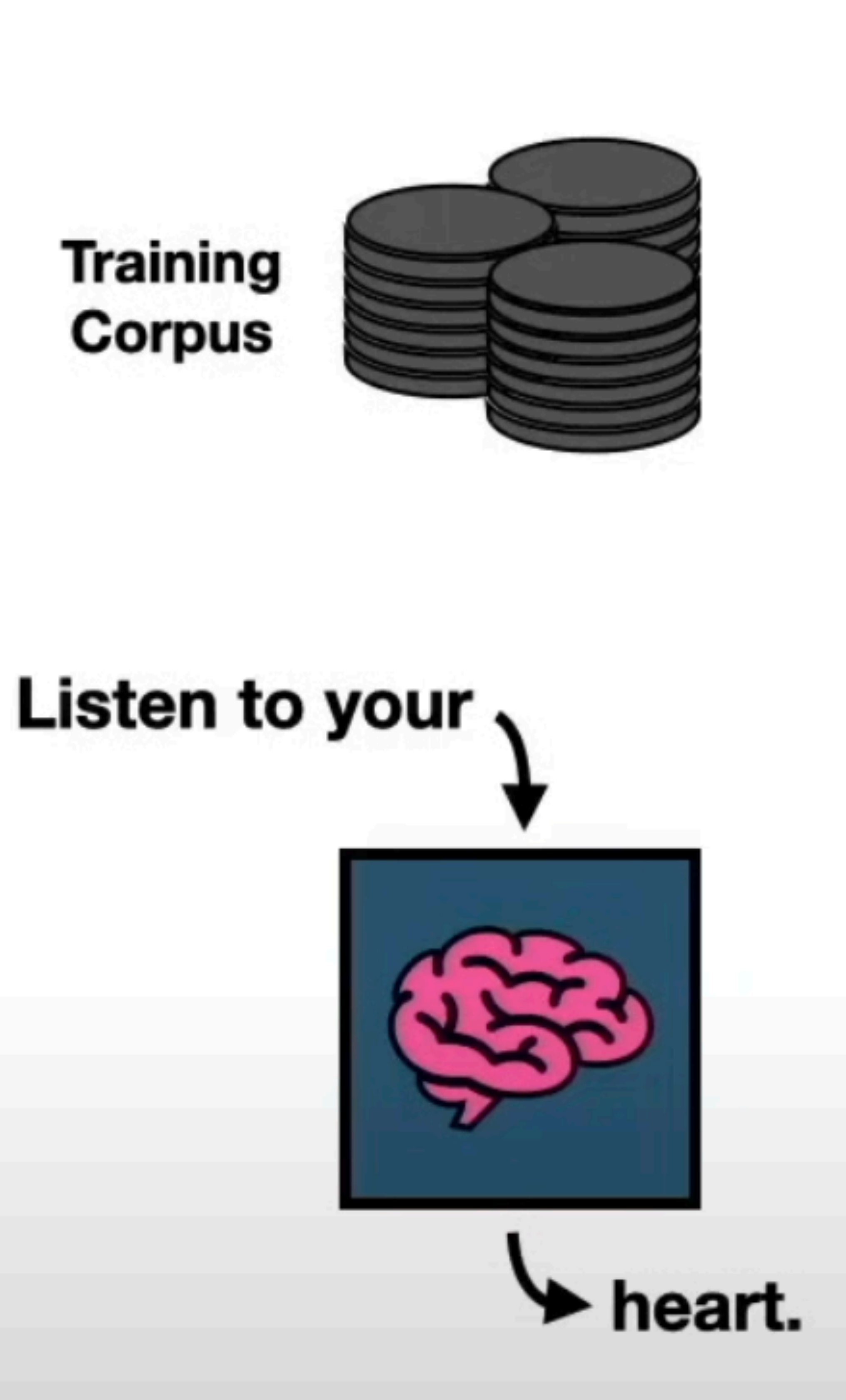
Module 3 - Topic 1

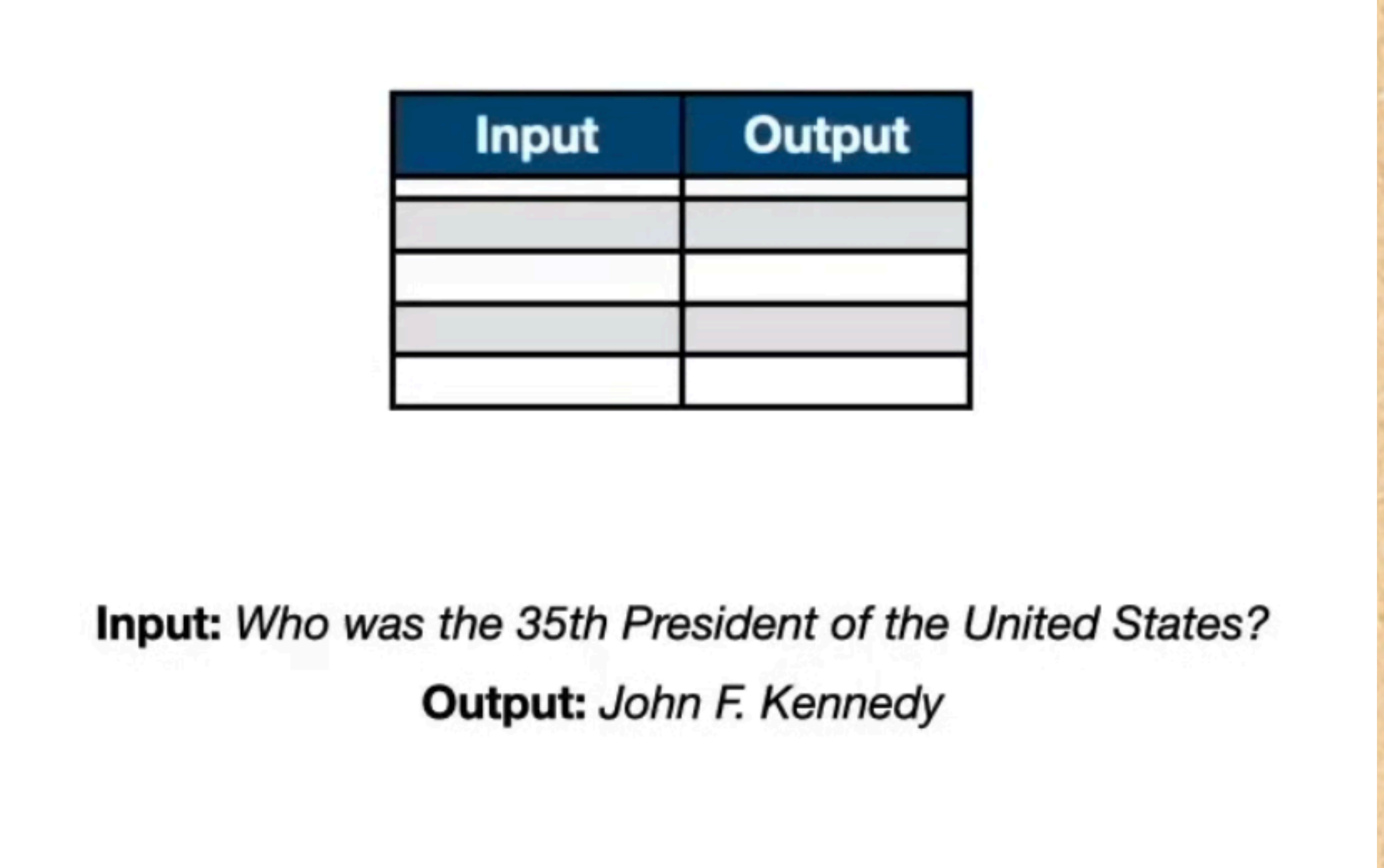
3.1.2 Instruction Tuning

Instruction
Tuning

Types of fine-tuning



Pre-training
(Unsupervised)



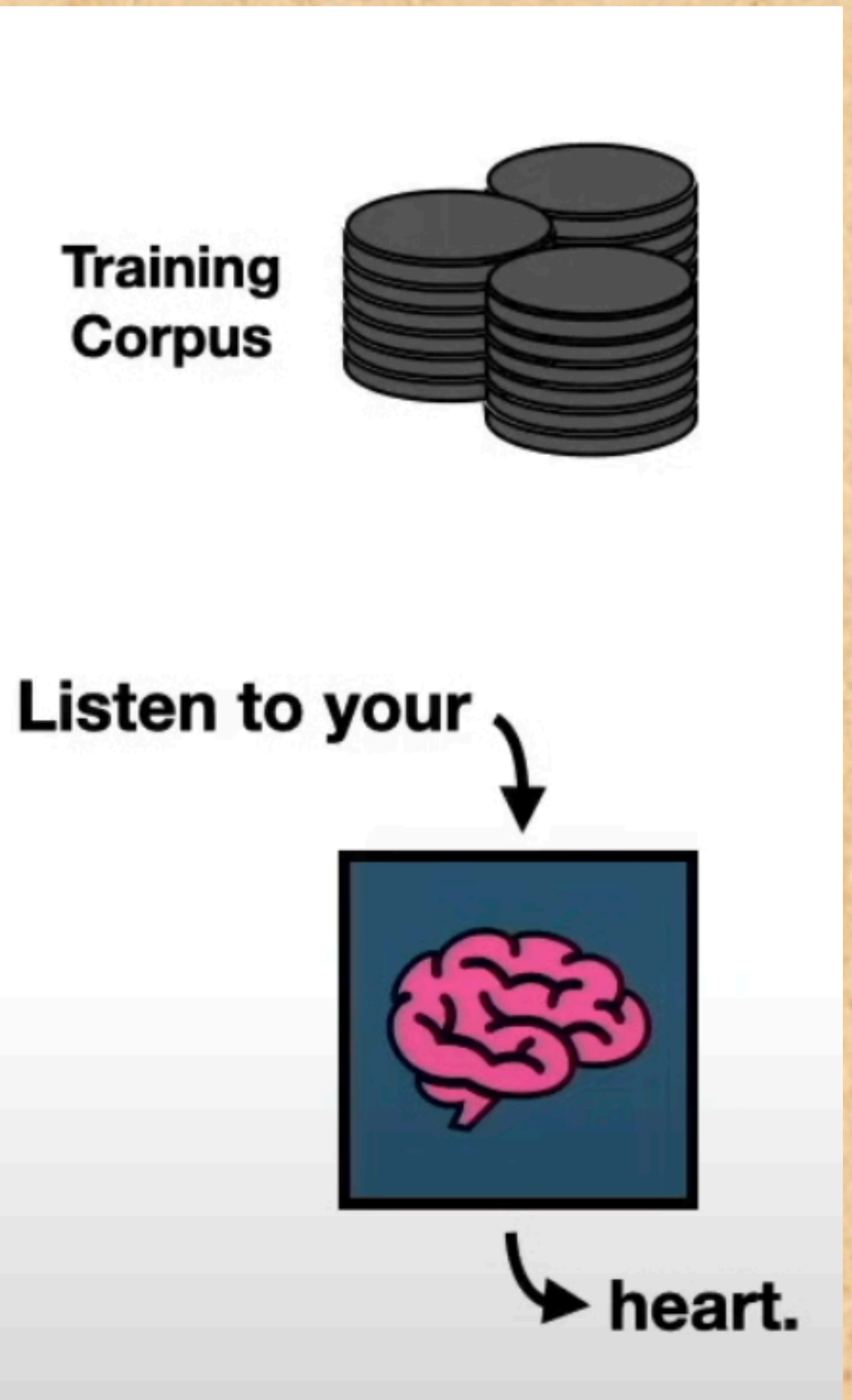
Input	Output

Input: *Who was the 35th President of the United States?*

Output: *John F. Kennedy*

Instruct-tuning
(Supervised)

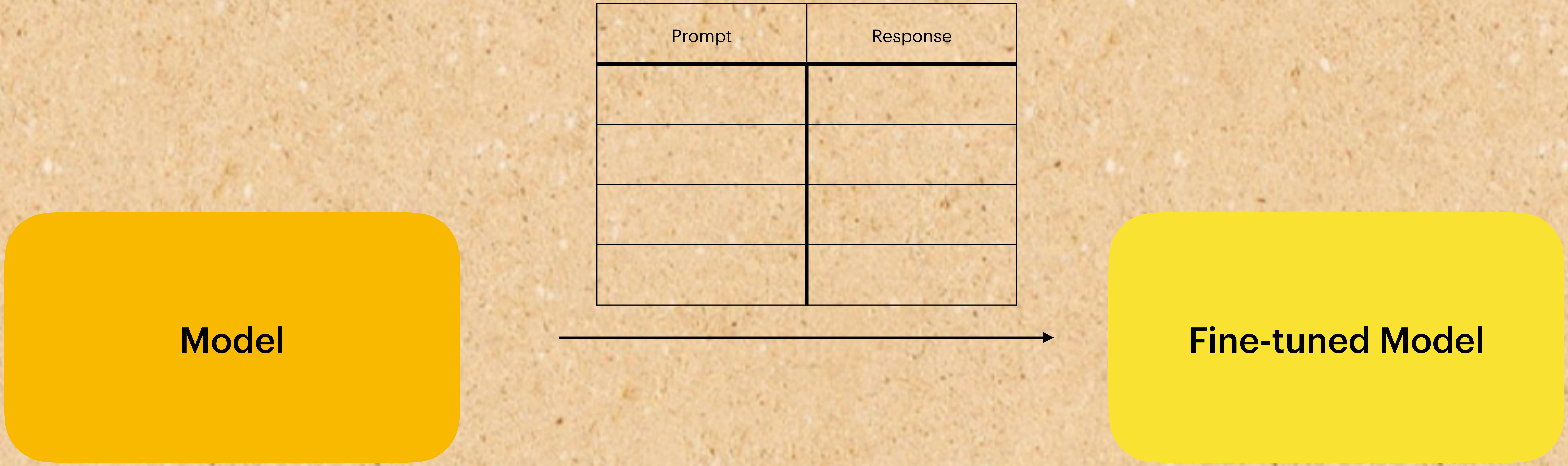
Pre-training



Pre-training
(Unsupervised)

- Initial phase of training where the model learns from a large corpus of text data.
- Model learns patterns and structures from the data without specific labeled outputs.

What is Instruction Tuning?



Prompt	Response

- More focused, task-specific type of fine-tuning.
- It uses paired input-output data to teach the model how to respond to specific types of prompts or questions.

What is Instruction Tuning?

Instruction fine-tuning is a specialized technique for adapting pre-trained LLMs using a labeled dataset that consists of prompts and corresponding outputs.

- **Instructional Prompts:** Input samples resemble user requests, guiding the model on how to responds
- **Output Responses:** The model learns to generate responses that align with these instructions

How does Instruction tuning work?

1. **Dataset Preparation:**
    - Create a labeled dataset with input-output pairs that include specific instructions.
    - Datasets can be manually curated or generated using other LLMs.
2. **Training Process:**
    - The model is fine-tuned on this dataset, adjusting its weights to minimize the difference between its predictions and the actual outputs.

Demo: Instruction Tuning

Example 1: Guardrails

Training the bot not to respond to certain topics

Google AI Studio

Example 2: Structured Output

Generate MCQs in predefined format

Prompt	Response

aistudio.google.com

Why is Instruction tuning useful?

- Enhanced Instruction Following

- Reduced Need for Prompt Engineering

- Generalisation Across Tasks

- Improved User Experience