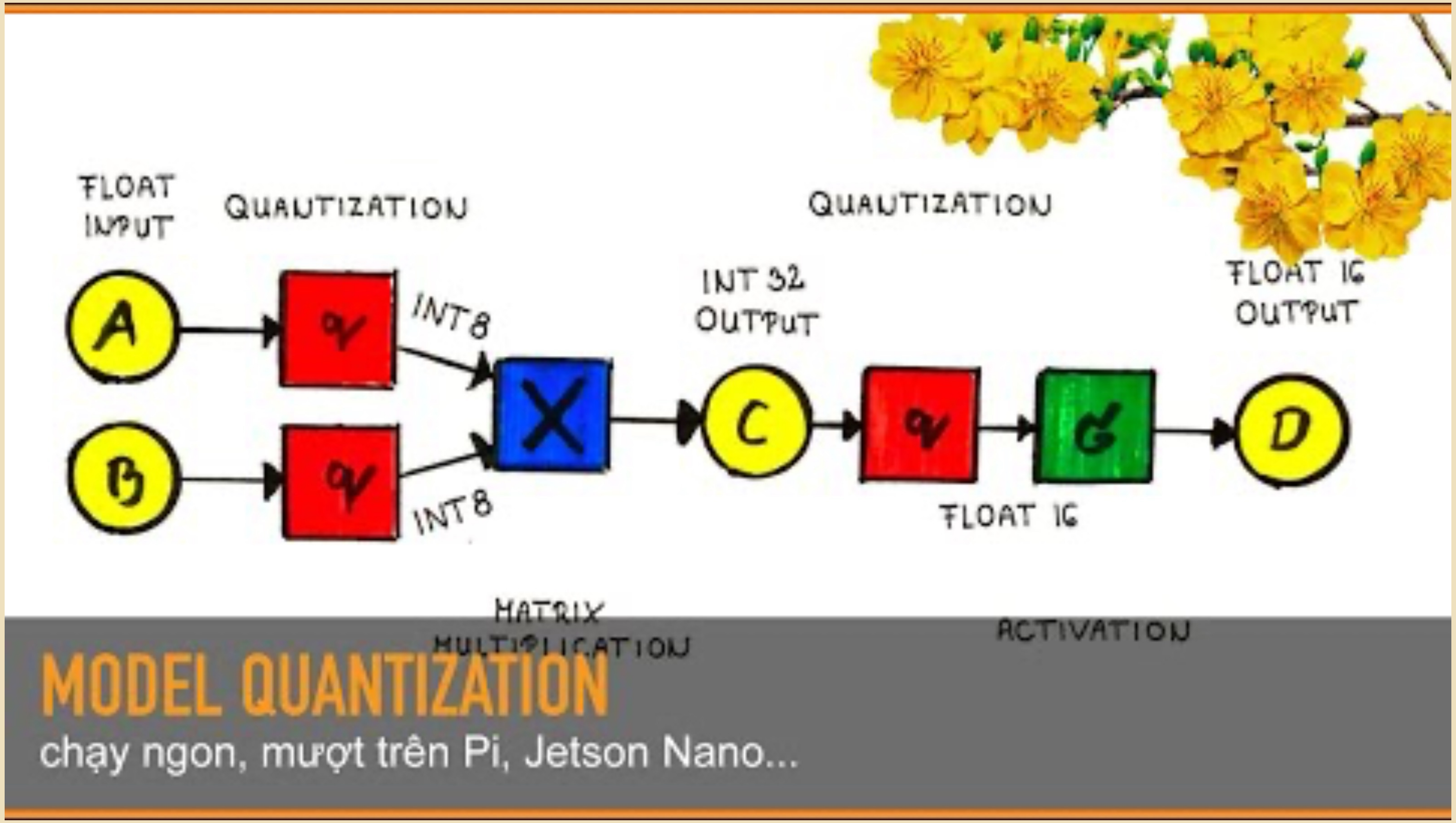
LLM

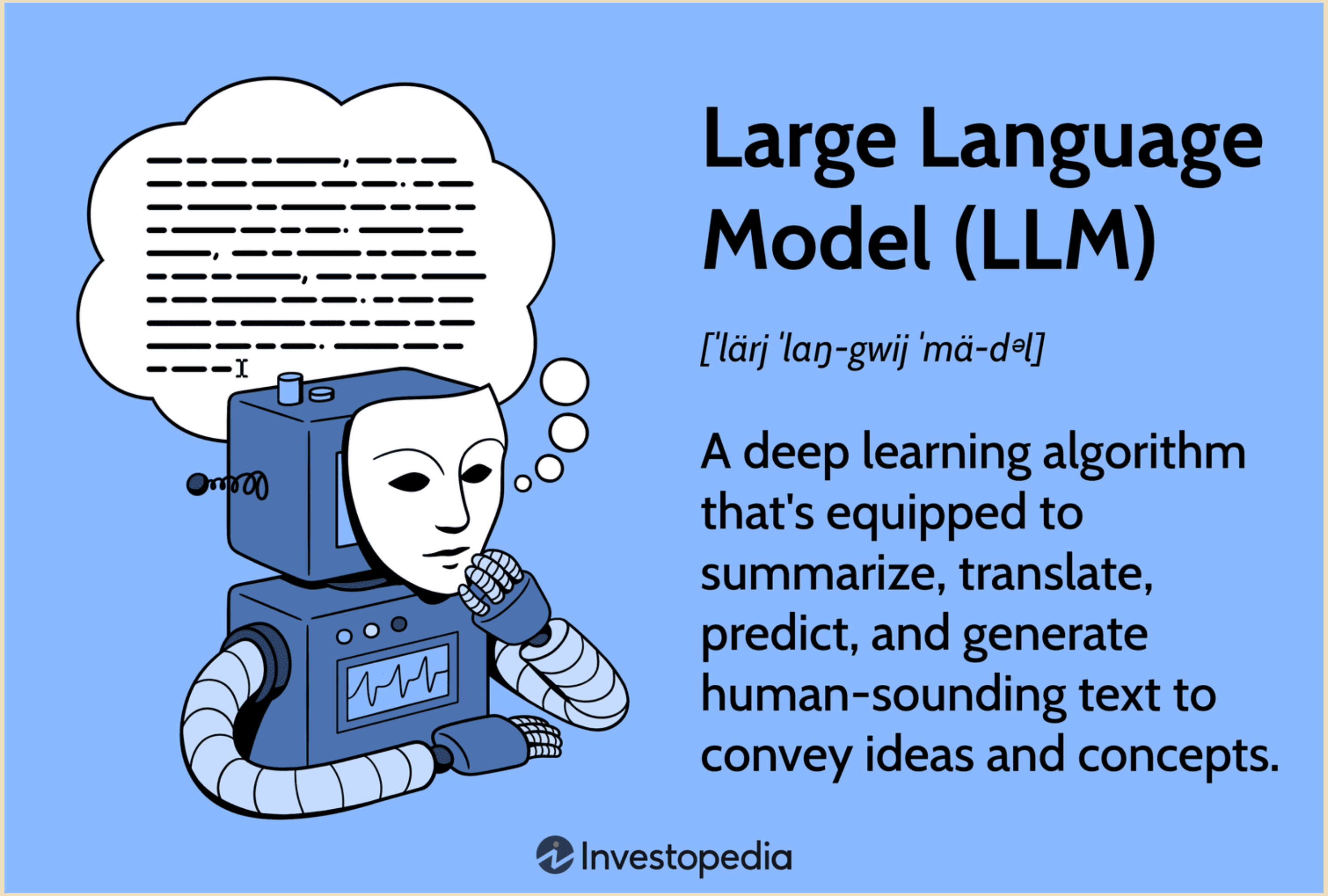
Finetuning with PEFT, LoRA

What is Quantization?



Search “miai quantization” là ra video :D

What is LLM?



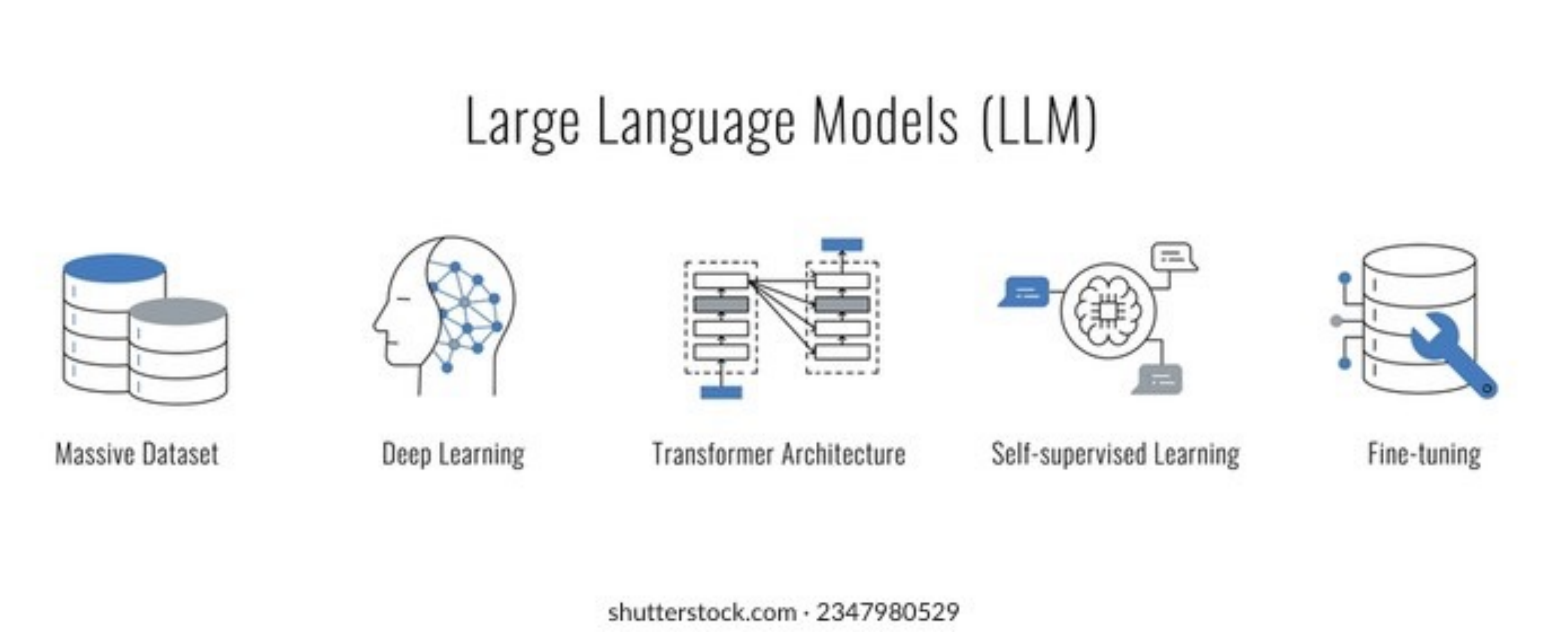
Large Language Model (LLM)

['lärj 'laŋ-gwij 'mä-dᵊl]

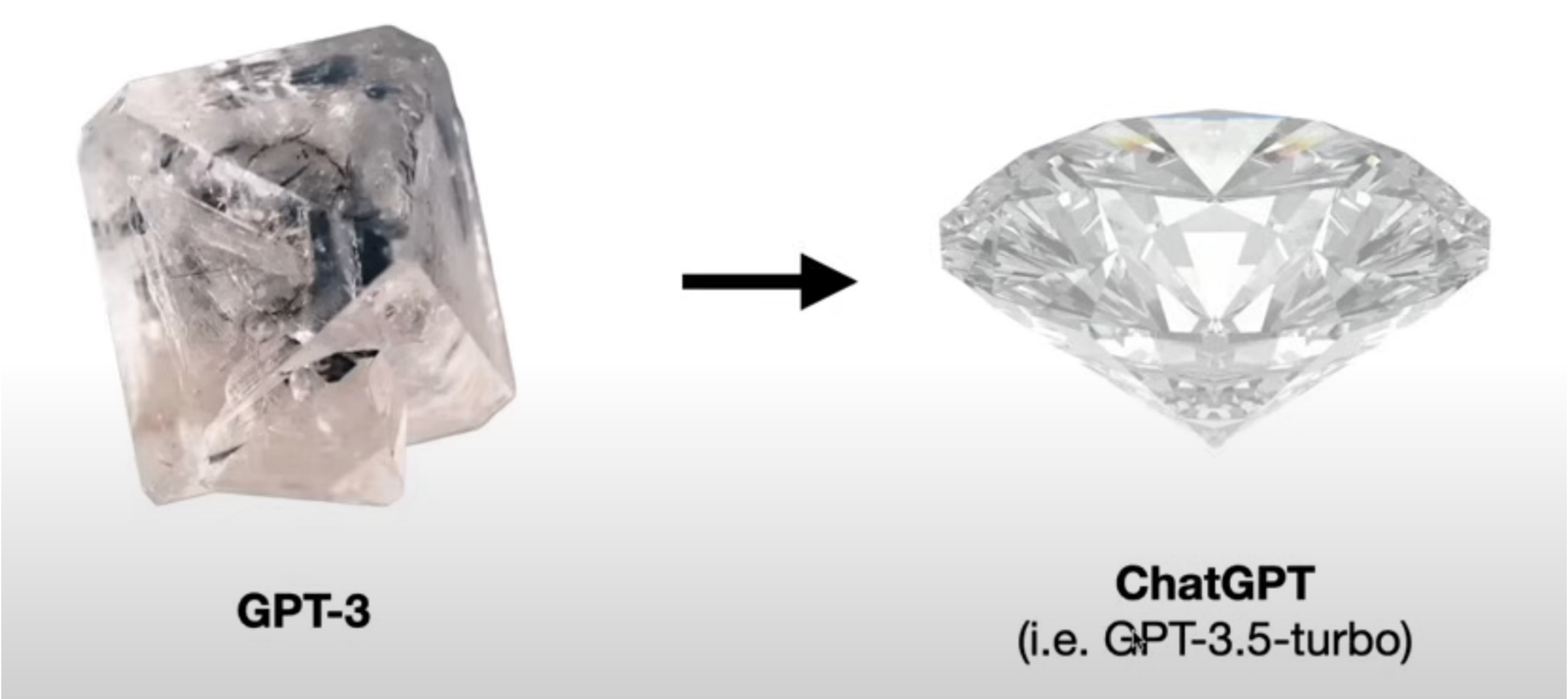
A deep learning algorithm that's equipped to summarize, translate, predict, and generate human-sounding text to convey ideas and concepts.

Investopedia

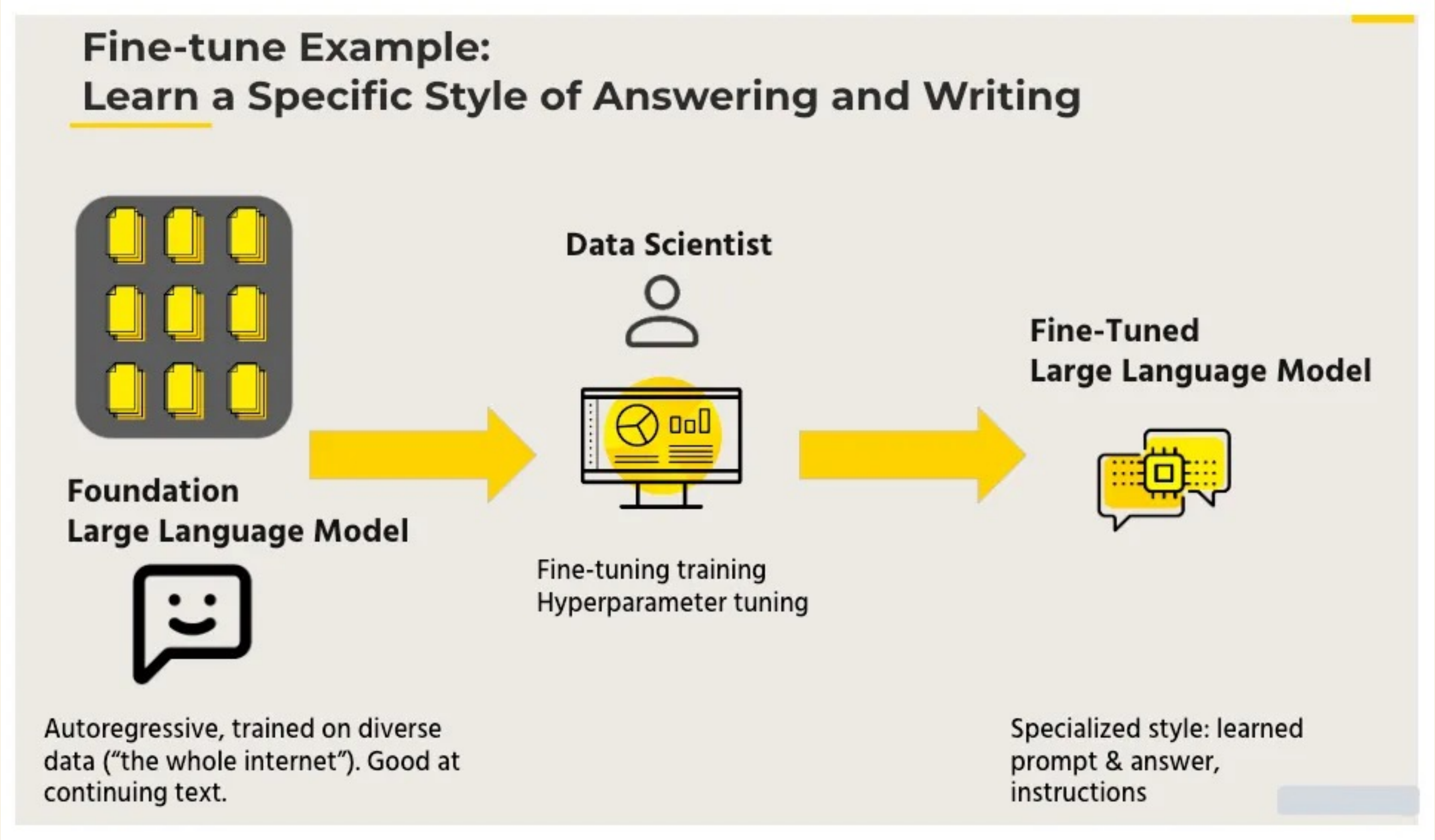
What is LLM?


Large Language Models (LLM)
Massive Dataset
Deep Learning
Transformer Architecture
Self-supervised Learning
Fine-tuning
shutterstock.com · 2347980529

Fine-tuning is taking a pre-trained model and training at least one internal model parameter



Why finetuning?



Fine-tune Example:
Learn a Specific Style of Answering and Writing

Foundation
Large Language Model

Autoregressive, trained on diverse data ("the whole internet"). Good at continuing text.

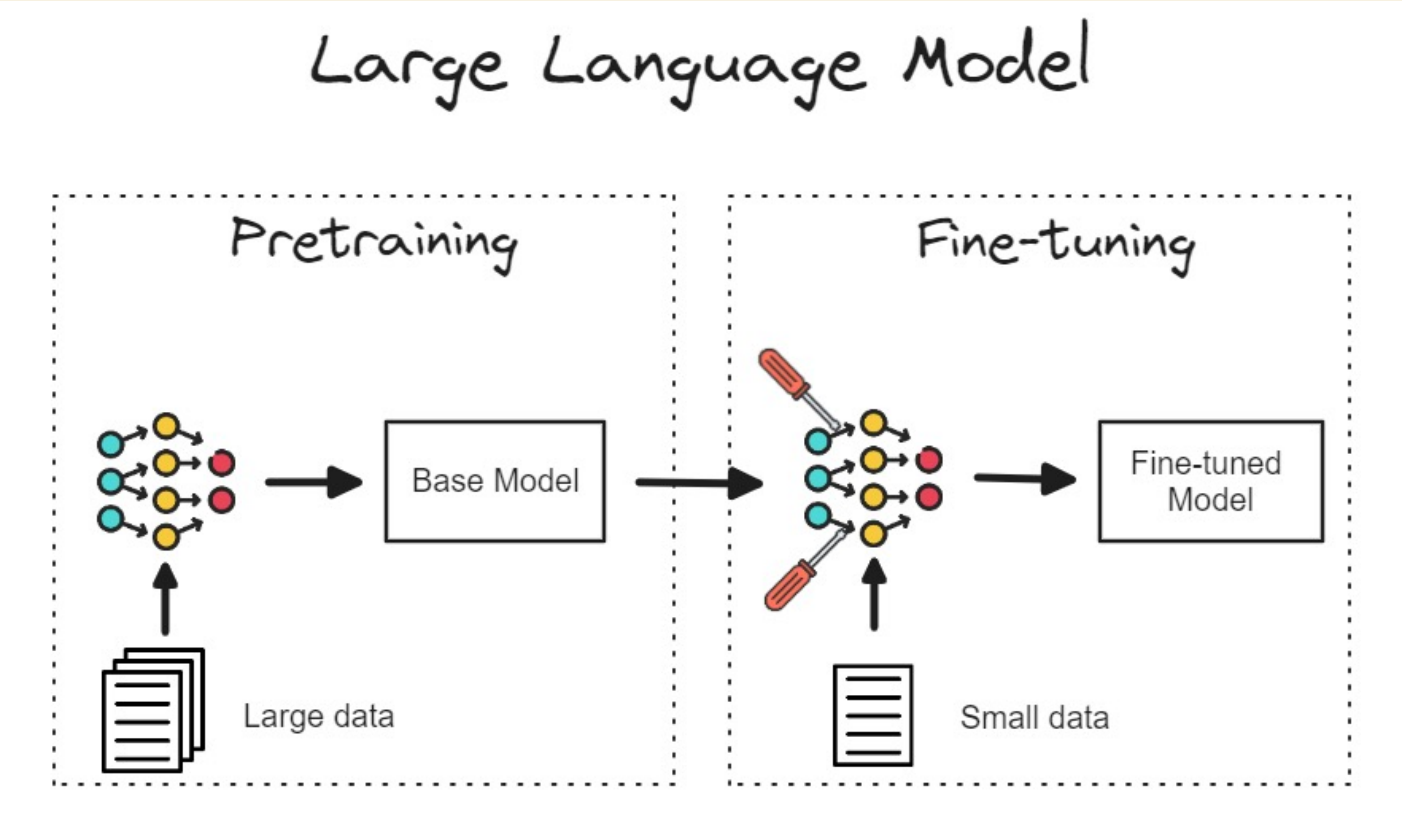
Data Scientist

Fine-tuning training
Hyperparameter tuning

Fine-Tuned
Large Language Model

Specialized style: learned prompt & answer, instructions

Why finetuning?



Base Model

(davinci (GPT-3))

Prompt: Tell me how to fine-tune a model

Completion:
How can I control the complexity of a model?
How do I know when my model is done?
How do I test a model?
How do I define a task for my model?
How do I scale a model?
How do I make a probabilistic model?
How do I start using a model once it's complete?
How do I cope with missing data?
How do I make my modeling efforts sustainable?

Fine-tuned Model

(text-davinci-003)

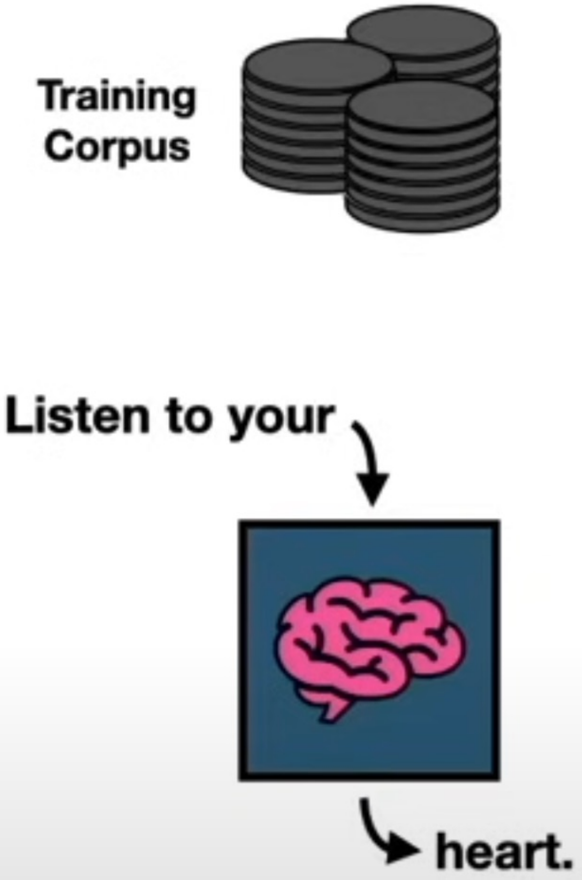
Prompt: Tell me how to fine-tune a model

Completion:
Fine-tuning a model involves adjusting the parameters of a pre-trained model in order to make it better suited for a given task. There are generally three steps to fine-tuning a model:

1. Select a base model - Select a pre-trained deep learning model that has been trained on a large dataset.
2. Adjust parameters - Adjust parameters of the pre-trained model to better suit the desired task. This may include changing the number of layers, adjusting learning rate, adding regularization, or tweaking the optimizer.
3. Train the model - Train the new model on the desired dataset. The amount of data and the amount of training required will depend on the task and the model.

How to finetuning?

03 ways of finetuning



1) Self-supervised

Input	Output

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

Output: *John F. Kennedy*

```
"""Please answer the following question.

Q: {Question}

A: {Answer}"""
```

2) Supervised

i. Supervised FT

ii. Train Reward Model

3) Reinforcement Learning

Flow of Supervised finetuning

Chọn một task fine-tuning: hỏi đáp về nội quy lao động, về quy định sinh viên...

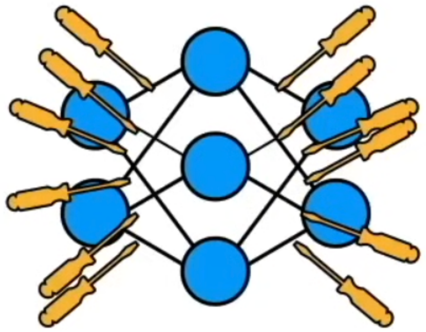
Chuẩn bị cơ sở dữ liệu: nội quy lao động, bộ câu hỏi - trả lời...

Chọn một model gốc: GPT, LLAMA, PhoGPT...

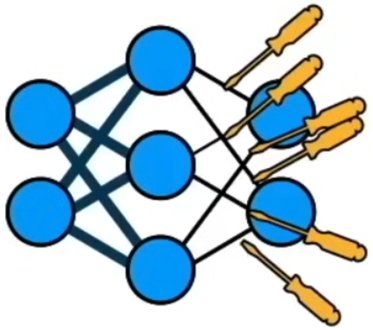
Finetune model & evaluate model

03 options for model Parameter training

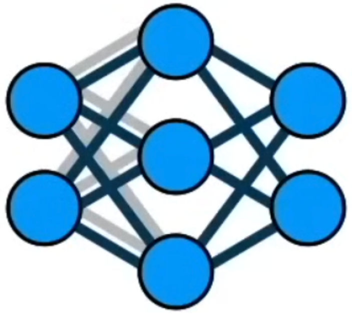
1) Retrain all parameters



2) Transfer Learning

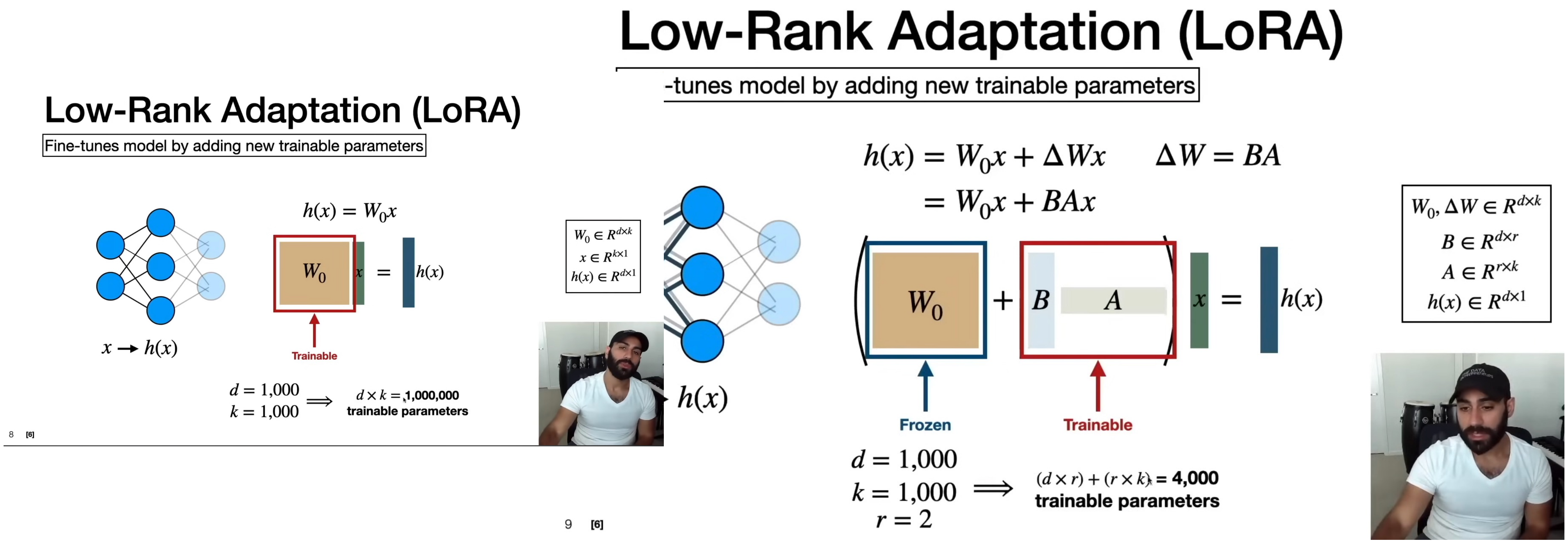


3) Parameter Efficient Fine-tuning (PEFT)



LoRA

PEFT involves **augmenting a base model with a relatively small number of trainable parameters.**
PEFT encapsulates a family of techniques, one of which is the popular **LoRA (Low-Rank Adaptation)** method [6]



Low-Rank Adaptation (LoRA)

Fine-tunes model by adding new trainable parameters

$$h(x) = W_0 x$$

$$x \rightarrow h(x)$$

$W_0 \in R^{d\times k}$
$x \in R^{k\times 1}$
$h(x) \in R^{d\times 1}$

$$\begin{matrix} d = 1{,}000 \\ k = 1{,}000 \end{matrix} \implies d \times k = \mathbf{1{,}000{,}000}$$
trainable parameters

8 [6]

Low-Rank Adaptation (LoRA)

-tunes model by adding new trainable parameters

$$h(x) = W_0 x + \Delta W x \qquad \Delta W = BA$$
$$= W_0 x + BAx$$

$W_0, \Delta W \in R^{d\times k}$
$B \in R^{d\times r}$
$A \in R^{r\times k}$
$h(x) \in R^{d\times 1}$

$$\begin{matrix} d = 1{,}000 \\ k = 1{,}000 \\ r = 2 \end{matrix} \implies (d \times r) + (r \times k) = \mathbf{4{,}000}$$
trainable parameters

9 [6]

https://www.youtube.com/watch?v=eC6Hd1hFvos&t=676s

Explain a bit about LORA

Hand-on on Colab and VPS