

# Fine-Tuning là gì

**Fine-tuning** là quá trình **đào tạo lại (huấn luyện tiếp)** một mô hình ngôn ngữ đã được huấn luyện trước (pre-trained) trên **dữ liệu cụ thể của bạn**, để mô hình:

- Hiểu tốt hơn về lĩnh vực bạn quan tâm (ví dụ: y tế, pháp lý, kỹ thuật...),
- Trả lời chính xác và phù hợp hơn với yêu cầu của ứng dụng thực tế,
- Tuân theo phong cách viết, giọng điệu hoặc cấu trúc riêng.

## Ví dụ dễ hiểu

Giả sử bạn có một mô hình GPT đã học hàng tỷ câu văn từ Internet. Nếu bạn muốn nó:

- Trả lời theo cách **ngghiêm túc, ngắn gọn, kỹ thuật**,
  - Hoặc chuyên trả lời **câu hỏi về luật Việt Nam**,
- thì bạn sẽ **fine-tune** nó bằng cách huấn luyện thêm trên tập dữ liệu nhỏ gồm các ví dụ bạn mong muốn.

## Fine-tuning khác với Prompt Engineering thế nào?

Prompt Engineering	Fine-Tuning
Thay đổi prompt để điều khiển đầu ra	Huấn luyện lại mô hình để thay đổi hành vi
Không tốn chi phí đào tạo lại	Tốn tài nguyên (GPU, RAM, thời gian)
Dễ làm, nhưng hiệu quả có giới hạn	Mạnh hơn, dùng cho yêu cầu phức tạp
Không cần dữ liệu huấn luyện	Cần dữ liệu huấn luyện có nhãn

## Có mấy loại Fine-Tuning?

### 1. Full Fine-Tuning (ít dùng với LLM lớn)

- Cập nhật **toàn bộ trọng số (weights)** của mô hình.
- Yêu cầu **nhiều tài nguyên** → không hiệu quả với mô hình lớn như LLaMA-2-13B, GPT-J...

### 2. PEFT (Parameter-Efficient Fine-Tuning) – [rất phổ biến hiện nay]

Gồm các kỹ thuật như:

- **LoRA** (Low-Rank Adaptation)
- **QLoRA** (LoRA kết hợp nén và huấn luyện trên GPU yếu hơn)

- **Prefix Tuning, Adapter Tuning...**

Ưu điểm:

- Chỉ fine-tune **rất ít tham số** (~0.1%-1%) → tiết kiệm chi phí và tài nguyên.
- Có thể lưu nhiều “bản cập nhật nhỏ” cho các mục tiêu khác nhau.

## QLoRA là gì?

**QLoRA = LoRA + Quantization (4-bit)**

Mục tiêu là fine-tune mô hình lớn (ví dụ: LLaMA-2-13B) **trên laptop hoặc GPU yếu**.

- **Quantization**: giảm độ chính xác xuống 4-bit để tiết kiệm RAM/GPU.
- **LoRA**: chèn các lớp nhỏ để học thêm nhưng không thay đổi mô hình gốc.

Rất phổ biến để fine-tune mô hình lớn với ngân sách thấp!

Khi nào nên fine-tune?

Trường hợp	Có nên fine-tune?
Muốn mô hình hiểu văn bản nội bộ, tài liệu chuyên ngành	Có
Muốn mô hình trả lời theo phong cách riêng	Có
Chỉ cần thay đổi nhẹ câu trả lời	<input type="checkbox"/> Dùng prompt thôi là đủ
Không có dữ liệu huấn luyện	<input type="checkbox"/> Không fine-tune được

## Các công cụ hỗ trợ Fine-tuning

- **Hugging Face Transformers + PEFT + TRL** (Python, rất phổ biến)
- **LoRA, QLoRA** với **bitsandbytes**
- **OpenAI Fine-tuning API** (với GPT-3.5-Turbo)
- **Axolotl, SFTTrainer, AutoTrain...**

---

Phiên bản #1

Được tạo 7 tháng 5 2025 05:56:17 bởi Đỗ Ngọc Tú

Được cập nhật 7 tháng 5 2025 05:59:37 bởi Đỗ Ngọc Tú