

Thực hành điều chỉnh tham số với LM Studio

LM Studio là một phần mềm giao diện GUI giúp bạn **chạy mô hình ngôn ngữ LLM** (như Mistral, LLaMA, Phi-2, v.v.) **ngay trên máy tính cá nhân**, thông qua GGUF và backend như llama.cpp hoặc Ollama.

Trong **LM Studio**, bạn có thể điều chỉnh các tham số này ở phần **Advanced Settings**:

- Temperature
- Top-k
- Top-p
- Repetition penalty
- Max tokens
- Sampling mode là mặc định luôn bật (`do_sample = true`) nếu bạn có `temperature > 0`.

Bài thực hành 1: Temperature - Điều chỉnh độ ngẫu nhiên

Mục tiêu:

Hiểu cách **temperature ảnh hưởng đến mức độ sáng tạo** và ổn định của mô hình.

Cách thực hiện:

1. Mở **LM Studio** và chọn một mô hình như **Mistral-7B Instruct GGUF** hoặc bất kỳ mô hình nào bạn đã cài.
2. Đặt prompt:

Viết một đoạn văn giới thiệu về Việt Nam như thể bạn là một hướng dẫn viên du lịch chuyên nghiệp.

3. Thử 3 lần với các mức `temperature` khác nhau:
 - `0.1` → Siêu chính xác, ít sáng tạo
 - `0.7` → Trung bình, cân bằng giữa sáng tạo và logic
 - `1.2` → Rất sáng tạo, nhưng có thể "nói bậy" (hallucinate)
4. So sánh kết quả.

Kết luận mong đợi:

- 0.1: Câu trả lời giống sách giáo khoa, ít biến thể.
- 0.7: Có chút cảm xúc, dùng từ phong phú hơn.
- 1.2: Có thể thêm chi tiết không đúng sự thật hoặc nói lan man.

Bài thực hành 2: Top-k Sampling - Giới hạn số lượng từ khả thi

Mục tiêu:

Hiểu cách giới hạn lựa chọn từ tiếp theo bằng số lượng cố định.

Cách làm:

1. Prompt giống như trên.
2. Giữ temperature ở 0.7.
3. Thay đổi top_k:
 - top_k = 5: Chọn từ trong top 5
 - top_k = 50: Từ trong top 50
 - top_k = 100: Rộng hơn

Kết luận:

- top_k thấp: Câu trả lời dễ đoán, lặp lại nhiều.
- top_k cao: Câu trả lời phong phú hơn, đôi khi bất ngờ.

Bài thực hành 3: Top-p Sampling (Nucleus Sampling)

Mục tiêu:

Thay vì số lượng từ, bạn giới hạn theo xác suất cộng dồn.

Cách làm:

1. Prompt giữ nguyên.
2. temperature = 0.7, top_k = 0 (tắt top_k).
3. Thử các giá trị top_p:
 - top_p = 0.3 → Chọn từ rất chắc chắn
 - top_p = 0.9 → Cho phép đa dạng hơn

Kết luận:

- `top_p` thấp: Trả lời ngắn gọn, an toàn
- `top_p` cao: Phong cách viết đa dạng hơn

Bài thực hành 4: Repetition Penalty - Tránh lặp lại

Mục tiêu:

Ngăn mô hình nói đi nói lại một ý.

Cách làm:

1. Prompt:

Hãy viết một đoạn giới thiệu ngắn về lợi ích của việc đọc sách.

2. Chạy với:

- `repetition_penalty = 1.0` (mặc định)
- `repetition_penalty = 1.2` (tránh lặp nhiều hơn)
- `repetition_penalty = 1.5` (rất ghét lặp)

Kết luận:

- Không penalty: Có thể lặp cụm như "Đọc sách giúp bạn..." nhiều lần.
- Có penalty: Câu trau chuốt hơn, tránh lặp từ.

Bài thực hành 5: Sampling Mode (`do_sample = True`)

Mục tiêu:

Bật/tắt chế độ lấy mẫu (sampling) – chọn từ ngẫu nhiên hoặc chọn từ xác suất cao nhất.

Cách làm:

1. Prompt:

Viết một lời chào sáng tạo cho một ứng dụng học tiếng Anh.

2. So sánh khi:

- `do_sample = False` (greedy decoding – luôn chọn từ xác suất cao nhất)
- `do_sample = True` + `temperature = 0.7`

Kết luận:

- `do_sample = False`: Câu trả lời giống nhau mỗi lần chạy.
- `do_sample = True`: Mỗi lần chạy cho ra câu khác nhau.

Phiên bản #1

Được tạo 24 tháng 4 2025 15:09:47 bởi Đỗ Ngọc Tú

Được cập nhật 26 tháng 4 2025 02:30:05 bởi Đỗ Ngọc Tú