SỔ TAY THỰC HÀNH

LLM SETTINGS

Thiết lập tham số để kiểm soát đầu ra của mô hình ngôn ngữ lớn

PROMPT	LLM SETTINGS	RESPONSE
Yêu cầu đầu vào	Cách mô hình lựa chọn đầu ra	Kết quả được sinh ra

Biên soạn tiếng Việt từ Prompt Engineering Guide

Giữ nguyên thuật ngữ chuyên ngành bằng English và giải thích tiếng Việt trong ngoặc

Phiên bản biên tập: 09/2026

CHƯƠNG MỞ ĐẦU

Cách đọc và sử dụng cuốn sổ tay

Cuốn sổ tay này trình bày các thiết lập phổ biến khi làm việc với Large Language Model - LLM (mô hình ngôn ngữ lớn) thông qua API (giao diện lập trình ứng dụng) hoặc các công cụ cho phép điều chỉnh tham số. Trọng tâm là hiểu đúng vai trò của từng setting (tham số thiết lập), biết khi nào nên tăng hoặc giảm, và biết cách thử nghiệm để tìm cấu hình phù hợp cho từng loại tác vụ.

Nội dung cốt lõi được biên soạn lại từ trang "LLM Settings" của Prompt Engineering Guide. Phần diễn giải, ví dụ, bảng so sánh và quy trình thực hành trong tài liệu này được viết lại theo hướng học tập, không sao chép nguyên văn nguồn.

Nguyên tắc đọc thuật ngữ

Mỗi thuật ngữ chuyên ngành được giữ bằng English; ngay lần xuất hiện quan trọng sẽ có phần giải thích tiếng Việt trong ngoặc. Ví dụ: token (đơn vị mã hóa văn bản), deterministic (có tính xác định), sampling (lấy mẫu).

Mục tiêu sau khi đọc

- Hiểu 6 settings chính và phân biệt các cặp dễ nhầm: Temperature/Top P; Frequency/Presence Penalty.
- Biết chọn cấu hình khởi điểm theo mục tiêu: chính xác, cô đọng, đa dạng, sáng tạo hoặc đúng cấu trúc.
- Biết chạy experiment (thử nghiệm) có kiểm soát và lưu ý sự khác biệt giữa model, version và provider.

Mục lục nội dung

Phần 1-6	Phần 7-12
1. LLM Settings là gì?	7. Presence Penalty
2. Temperature	8. Tương tác giữa các Settings
3. Top P / nucleus sampling	9. Cấu hình theo Use Case
4. Max Length	10. Quy trình thử nghiệm và tối ưu
5-6. Stop Sequences + Frequency Penalty	11-12. Cheat Sheet + Glossary + nguồn

CHƯƠNG 1

LLM Settings là gì?

Các nút điều khiển giúp định hình cách mô hình sinh câu trả lời



Hình 1.1. Vị trí của LLM Settings trong quá trình tạo response.

Khi gửi một prompt (chỉ dẫn đầu vào) đến LLM, mô hình không "nghĩ" theo cách con người mà tính toán xác suất cho các token (đơn vị mã hóa văn bản) có thể xuất hiện tiếp theo. Quá trình sinh câu trả lời là chuỗi các bước lựa chọn token nối tiếp nhau. LLM Settings là tập tham số giúp người dùng tác động lên cách lựa chọn đó.

VÍ DỤ DỄ HIỂU • Token & sampling (lấy mẫu token)

Tình huống: Bạn gõ: "Viết một câu chúc sinh nhật ngắn." Sau cụm "Chúc bạn", mô hình phải chọn token tiếp theo.

Thiết lập: Các token như "luôn", "một", "thật" có xác suất khác nhau. Sampling quyết định token nào được chọn từ các ứng viên đó.

Kết quả: Nếu ưu tiên token xác suất cao, câu trả lời thường quen thuộc và ổn định; nếu cho phép nhiều ứng viên hơn, câu chữ có thể đa dạng hơn.

Ghi nhớ: LLM tạo câu trả lời từng token một; settings tác động trực tiếp lên quá trình chọn token kế tiếp.

Trong thực tế, cùng một prompt có thể cho ra kết quả khác nhau nếu thay đổi setting. Một cấu hình phù hợp cho factual QA (hỏi đáp dựa trên sự kiện) có thể lại quá cứng nhắc cho creative writing (viết sáng tạo). Ngược lại, cấu hình phù hợp cho sáng tạo có thể làm câu trả lời kém ổn định khi cần dữ kiện chính xác.

1.1. Ba mục tiêu lớn của việc điều chỉnh settings

Mục tiêu	Ý nghĩa	Ví dụ
Reliability (độ ổn định)	Giảm biến động không cần thiết giữa các lần chạy.	Phân loại văn bản, hỏi đáp theo dữ liệu, trích xuất thông tin.
Desirability (mức độ phù hợp mong muốn)	Định hình độ sáng tạo, độ đa dạng và phong cách đầu ra.	Viết ý tưởng quảng cáo, tên sản phẩm, kịch bản.
Control (khả năng kiểm soát)	Khống chế độ dài, điểm dừng và mức lặp lại.	Giới hạn số mục, tạo JSON, tạo câu trả lời ngắn.

1.2. Mô hình khái niệm đơn giản

Prompt -> phân phối xác suất -> sampling -> response

Prompt xác định ngữ cảnh và nhiệm vụ. Model tạo probability distribution (phân phối xác suất) cho token tiếp theo. Các tham số như Temperature và Top P tác động đến quá trình sampling (lấy mẫu). Các tham số khác như Max Length và Stop Sequences tác động đến việc khi nào quá trình sinh dừng lại.

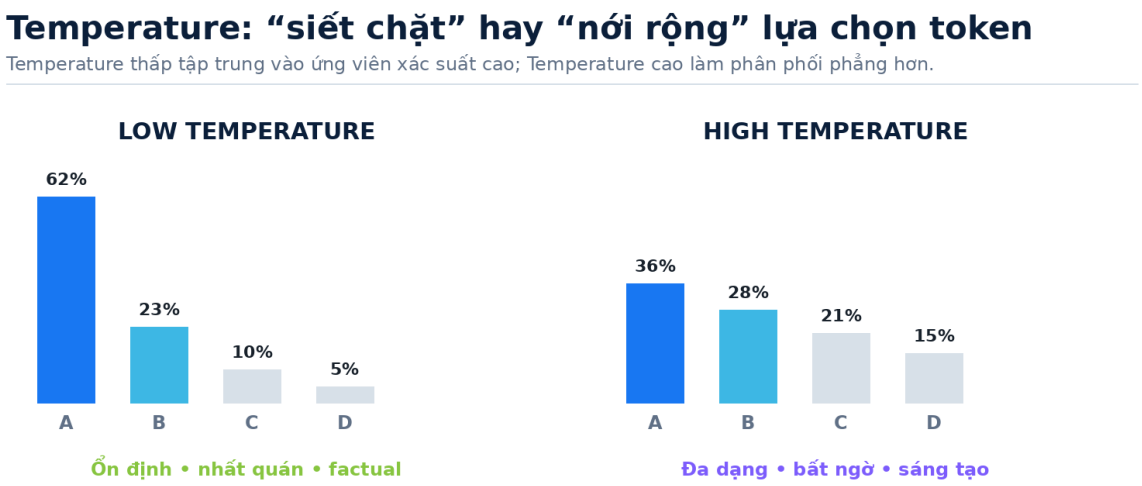
1.3. Điều quan trọng nhất: không có một cấu hình "tốt nhất" cho mọi trường hợp

Một setting chỉ có ý nghĩa khi gắn với use case (trường hợp sử dụng), yêu cầu chất lượng và loại model đang dùng. Do đó, thay vì tìm một bộ tham số "chuẩn", hãy tìm một cấu hình khởi điểm hợp lý rồi đánh giá bằng test set (tập kiểm thử) đại diện cho công việc thực tế.

CHƯƠNG 2

Temperature

Điều chỉnh mức độ ngẫu nhiên và đa dạng của lựa chọn token



Hình 2.1. Trực giác về tác động của Temperature lên phân phối lựa chọn token.

Temperature là tham số phổ biến nhất khi điều khiển cách LLM chọn token. Có thể hình dung Temperature như mức "nới rộng" hoặc "siết chặt" sự chênh lệch xác suất giữa các phương án tiếp theo.

VÍ DỤ DỄ HIỂU • Temperature

Tình huống: Prompt: “Viết một slogan cho quán cà phê dành cho người làm việc sáng tạo.”

Thiết lập: Temperature thấp: mô hình thiên về cách diễn đạt an toàn. Temperature cao: mô hình chấp nhận nhiều lựa chọn ít hiển nhiên hơn.

Kết quả: Thấp → “Cà phê ngon, cảm hứng mỗi ngày.” Cao → có thể xuất hiện câu độc đáo hơn như “Nạp một ngụm, bật cả ý tưởng.”

Ghi nhớ: Temperature chủ yếu điều khiển mức độ ngẫu nhiên/đa dạng, không phải nút “tăng độ thông minh”.

2.1. Khi Temperature thấp

- Mô hình có xu hướng ưu tiên token có xác suất cao nhất hoặc nhóm token rất có khả năng.
- Kết quả thường ổn định hơn giữa các lần chạy.
- Phù hợp khi cần factuality (tính đúng sự kiện), consistency (tính nhất quán) và câu trả lời ngắn gọn.
- Ví dụ: trích xuất dữ liệu, trả lời câu hỏi dựa trên tài liệu, phân loại, viết câu trả lời theo mẫu.

2.2. Khi Temperature cao

- Mô hình có khả năng chọn các token ít hiển nhiên hơn.
- Đầu ra thường đa dạng, bất ngờ và sáng tạo hơn.
- Phù hợp với brainstorming (động não), creative writing (viết sáng tạo), đề xuất nhiều phương án.
- Đổi lại, mức biến động cao hơn có thể làm giảm tính nhất quán hoặc làm câu trả lời đi xa khỏi trọng tâm nếu prompt chưa đủ chặt.

Mức Temperature	Đặc điểm đầu ra	Nhóm tác vụ phù hợp
Thấp	Ổn định, ít biến động, thiên về phương án xác suất cao.	QA theo dữ kiện, extraction, classification, chuẩn hóa văn bản.
Trung bình	Cân bằng giữa ổn định và đa dạng.	Tóm tắt, giải thích, viết nội dung nghiệp vụ.
Cao	Đa dạng hơn, dễ xuất hiện ý tưởng hoặc cách diễn đạt mới.	Brainstorming, slogan, ý tưởng chiến dịch, viết sáng tạo.

Gợi ý thực hành

Nếu nhiệm vụ có tiêu chí đúng/sai rõ ràng, hãy bắt đầu với Temperature thấp. Nếu nhiệm vụ đánh giá cao sự đa dạng hơn là tính lặp lại, có thể tăng dần Temperature và so sánh nhiều lần chạy.

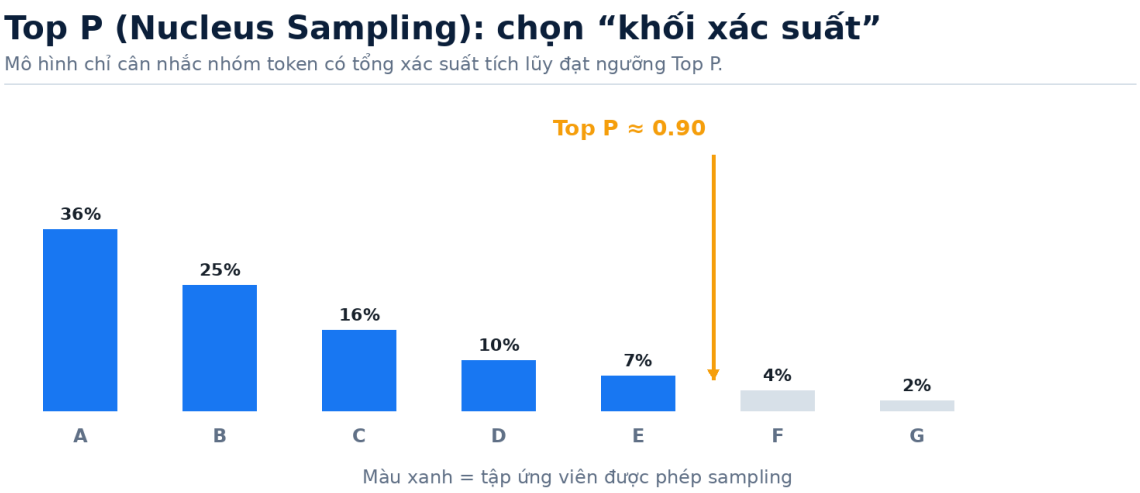
2.3. Ví dụ tư duy

Prompt: “Đề xuất 10 tên cho một ứng dụng giúp nhân viên ghi chú cuộc họp.” Với Temperature thấp, các tên có thể tập trung quanh các từ quen thuộc như Note, Meet, Memo. Với Temperature cao hơn, danh sách thường xuất hiện nhiều biến thể bất ngờ hơn. Đây không phải là đảm bảo tuyệt đối; nó là xu hướng trong cơ chế sampling.

CHƯƠNG 3

Top P

Nucleus sampling - lấy mẫu trong một “khối xác suất” được chọn



Hình 3.1. Top P giới hạn tập token theo xác suất tích lũy.

Top P là một kỹ thuật sampling thường được gọi là nucleus sampling (lấy mẫu theo hạt nhân). Thay vì xem xét toàn bộ token có thể xuất hiện, hệ thống chỉ giữ lại một tập token có tổng probability mass (khối xác suất) đạt ngưỡng Top P, sau đó lấy mẫu trong tập đó.

VÍ DỤ DỄ HIỂU • Top P / Nucleus Sampling

Tình huống: Giả sử token tiếp theo có xác suất: A=50%, B=25%, C=15%, D=7%, E=3%.

Thiết lập: Nếu Top P = 0,80, hệ thống chỉ cần giữ A+B+C vì tổng xác suất đã đạt 90% và bỏ các token còn lại khỏi tập lấy mẫu.

Kết quả: Mô hình chỉ chọn trong nhóm ứng viên xác suất cao đã được giữ lại, thay vì cân nhắc toàn bộ từ có thể xuất hiện.

Ghi nhớ: Top P điều khiển “phạm vi ứng viên”, còn Temperature điều chỉnh “độ chênh xác suất” giữa các ứng viên.

3.1. Top P thấp và Top P cao

Thiết lập	Mô hình được phép cân nhắc	Xu hướng
Top P thấp	Một tập nhỏ gồm các token có xác suất cao nhất.	Tập trung hơn, ít đa dạng hơn.
Top P cao	Một tập rộng hơn, bao gồm cả token có xác suất thấp hơn.	Đa dạng hơn, nhiều hướng diễn đạt hơn.

3.2. Temperature và Top P khác nhau như thế nào?

Cả hai đều ảnh hưởng đến mức độ đa dạng, nhưng theo hai cơ chế khác nhau. Temperature điều chỉnh hình dạng phân phối xác suất; Top P giới hạn tập token được phép đi vào quá trình lấy mẫu dựa trên tổng xác suất tích lũy.

> **Nguyên tắc đơn giản**
>
> Trong nhiều tình huống, nên ưu tiên điều chỉnh Temperature hoặc Top P, thay vì thay đổi đồng thời cả hai. Cách này giúp bạn biết rõ biến nào thực sự tạo ra thay đổi trong kết quả.

3.3. Khi nào ưu tiên Top P?

- Khi hệ thống hoặc provider cung cấp Top P như một tham số sampling chính.
- Khi muốn kiểm soát "phạm vi ứng viên" thay vì chỉ thay đổi độ phẳng của phân phối.
- Khi đang xây dựng experiment có mục tiêu so sánh rõ nucleus sampling với cấu hình mặc định.

CHƯƠNG 4

Max Length

Giới hạn số lượng token mà mô hình được phép sinh



Hình 4.1. Max Length như một "trần" kỹ thuật cho output.

Max Length là cách đặt giới hạn trên cho lượng output (đầu ra) mà mô hình có thể tạo. Tên tham số cụ thể có thể khác giữa các API, nhưng ý tưởng chung là kiểm soát độ dài sinh văn bản.