Tham số của mô hình Generative AI trong kiến trúc RAG

RAG (Retrieval-Augmented Generation) là một kiến trúc kết hợp giữa mô hình sinh (generative model) và mô hình tìm kiếm (retrieval model).
Mục tiêu là giúp mô hình tạo ra các câu trả lời chính xác hơn bằng cách **truy xuất thông tin từ dữ liệu bên ngoài**, rồi **dùng dữ liệu đó làm input** cho mô hình sinh văn bản (ví dụ GPT hay BERT).

Đây là kiến trúc dùng trong các chatbot trả lời tài liệu nội bộ, trợ lý ảo doanh nghiệp, v.v.

Tham số sinh (generation parameters) là gì

Tham số sinh (generation parameters) là các thiết lập giúp **kiểm soát cách mô hình AI tạo ra văn bản**.

> **Nói cách khác:** Bạn có thể xem chúng như các "nút điều chỉnh" giúp quyết định liệu mô hình nên sáng tạo hay nghiêm túc, nên ngắn gọn hay chi tiết, nên logic hay phong phú.

1. TEMPERATURE – Điều khiển độ ngẫu nhiên

Định nghĩa:
Temperature điều chỉnh độ ngẫu nhiên trong câu trả lời bằng cách **co giãn xác suất** (logits) trước khi chọn từ tiếp theo.

Giá trị	Ý nghĩa	Kết quả
Gần 0	Cực kỳ chắc chắn	Câu trả lời chính xác, ít sáng tạo
~1.0	Trung bình	Cân bằng sáng tạo và logic
>1.0	Rất ngẫu nhiên	Câu trả lời có thể lệch lạc, "ảo tưởng" (hallucination)

Ví dụ:

- `temperature = 0.2`: Thích hợp cho chatbot chăm sóc khách hàng
- `temperature = 0.9`: Thích hợp viết thơ hoặc nội dung sáng tạo

> *Lưu ý cá nhân:* Mình thường đặt **temperature rất thấp** (gần 0) cho các ứng dụng nghiêm túc như tư vấn pháp lý hoặc kỹ thuật. sử dụng LM studio

2. TOP-K SAMPLING – Giới hạn theo số lượng từ có xác suất cao nhất

Cơ chế hoạt động:
Chỉ chọn từ trong **k từ có xác suất cao nhất** tại mỗi bước.

Giá trị K	Ý nghĩa
10	Rất hạn chế – gần như luôn chọn từ phổ biến nhất
50	Cân bằng – vẫn sáng tạo nhưng tránh "nói bậy"
100+	Rất đa dạng – dễ lệch ngữ nghĩa

Ví dụ:
`top_k = 50` → Mô hình chỉ chọn từ tiếp theo từ 50 từ khả thi nhất.

3. TOP-P (Nucleus Sampling) – Giới hạn theo tổng xác suất

Cơ chế hoạt động:
Thay vì chọn số lượng cố định như top-k, **top-p chọn số từ sao cho tổng xác suất ≥ p**.

Giá trị P	Ý nghĩa
0.9	Cân bằng – dùng nhiều trong thực tế
0.8	Hạn chế hơn – ít rủi ro hơn
1.0	Không giới hạn – gần như không lọc

Ví dụ:

- `top_p = 0.9` → Chọn những từ sao cho tổng xác suất đạt 90% → tránh các từ "hiếm gặp" gây lệch ngữ cảnh.

4. REPETITION PENALTY – Tránh lặp lại

Cơ chế hoạt động:
Thêm "hình phạt" cho việc lặp từ, giúp đầu ra đa dạng và giống người hơn.

Giá trị	Ý nghĩa

1.0	Không phạt – có thể lặp lại nhiều
1.1	Hơi phạt – khuyến khích sự đa dạng
>1.2	Phạt nặng – tránh lặp từ gần như tuyệt đối

Ứng dụng:

- Tạo nội dung marketing hoặc viết truyện → dùng `1.2`
- Trả lời khoa học hoặc kỹ thuật → dùng `1.05–1.1` để vẫn giữ từ khóa

5. SAMPLING MODE – Có chọn ngẫu nhiên hay không

Tham số: `do_sample = True/False`

Chế độ	Kết quả
`True`	Có chọn ngẫu nhiên – đầu ra đa dạng hơn
`False`	Luôn chọn từ có xác suất cao nhất – đầu ra chắc chắn, nhưng đơn điệu

Tác giả: Đỗ Ngọc Tú

Công Ty Phần Mềm VHTSoft

Phiên bản #1
Được tạo 24 tháng 4 2025 14:58:00 bởi Đỗ Ngọc Tú
Được cập nhật 26 tháng 4 2025 02:30:05 bởi Đỗ Ngọc Tú