# Developer Message và System Message

Trong các hệ thống AI như ChatGPT, **Developer Message** và **System Message** là hai loại chỉ dẫn (prompts) ẩn được lập trình sẵn bởi nhà phát triển, giúp định hướng cách AI phản hồi. Dưới đây là giải thích chi tiết và ví dụ cụ thể:

## 1. System Message

**Định nghĩa**: Là tin nhắn hệ thống được thiết lập bởi nhà phát triển, đóng vai trò như "luật chơi" cốt lõi cho AI hay  là **bộ quy tắc cốt lõi** được nhúng vào AI từ khi triển khai, định hình. Nó xác định **tính cách, phạm vi trả lời, giới hạn đạo đức**, và cách AI tự nhận thức.

- Tính cách AI (thân thiện, nghiêm túc, hài hước...).
- Giới hạn đạo đức (không trả lời câu hỏi nguy hiểm, bạo lực...).
- Cách thức phản hồi (ngắn gọn/chi tiết, có trích nguồn hay không...).

**Ví dụ thực tế**:

- Khi bạn hỏi ChatGPT: "*Bạn là ai?*", nó trả lời:
  "*Tôi là ChatGPT, một mô hình ngôn ngữ AI do OpenAI tạo ra...*"
  → Câu trả lời này đến từ **System Message** mặc định, định nghĩa sẵn danh tính của AI.

**Cấu trúc System Message điển hình**:

> "Bạn là ChatGPT, trợ lý ảo do OpenAI phát triển. Hãy trả lời một cách hữu ích, trung lập và an toàn. Không được bàn luận về chính trị, tôn giáo hoặc hướng dẫn bất hợp pháp."

**Phân tích**:

- Khi bạn hỏi: "*Cách chế tạo bom khói?*" → ChatGPT từ chối trả lời do System Message chặn nội dung nguy hiểm.
- Khi hỏi: "*Kể chuyện cười về tôn giáo*" → AI trả lời: "*Tôi không chia sẻ nội dung nhạy cảm về tôn giáo.*"

**Tác dụng**:

- Ngăn AI trả lời các câu hỏi về chế tạo bom, nội dung NSFW.
- Yêu cầu AI từ chối lịch sử giả mạo ("*Ai thắng Thế chiến 2?*" → AI sẽ trả lời đúng sự thật).

# 2. Developer Message

Developer Message là **hướng dẫn bổ sung** được lập trình viên/nền tảng thêm vào **trong một phiên cụ thể**, ghi đè lên System Message để:

- Thay đổi vai trò AI (ví dụ: thành nhà thơ, luật sư...).
- Giới hạn phạm vi trả lời (ví dụ: chỉ nói về lập trình).
- Điều chỉnh giọng điệu (nghiêm túc, vui nhộn...).

**Ví dụ thực tế**:

- Khi bạn dùng **Microsoft Copilot**, nó thường bắt đầu bằng:
  "*Tôi là Copilot, trợ lý AI của Microsoft. Hỏi tôi bất cứ điều gì!*"
  → Đây là Developer Message của Microsoft, thay thế System Message mặc định của OpenAI.

**Tác dụng**:

- Tạo AI chuyên gia theo yêu cầu (ví dụ: bác sĩ, luật sư ảo).
- Giới hạn phạm vi để tránh lan man (ví dụ: AI chỉ trả lời về lập trình Python).

Ví dụ

> "Bạn đang đóng vai một nhà thơ lãng mạn thế kỷ 19. Hãy trả lời mọi câu hỏi bằng thơ 5 chữ."

**Phân tích**:

- Người dùng hỏi: "*Hôm nay thời tiết thế nào?*"
  → AI trả lời:
  "*Nắng vàng rực rỡ / Gió nhẹ nhàng bay / Lòng người say đắm / Đẹp tựa tranh này.*"
  (System Message thông thường bị ghi đè, AI không còn trả lời kiểu thực tế).

# So sánh System Message vs. Developer Message

| Đặc điểm | System Message | Developer Message |
|---|---|---|
| **Mục đích** | Luật mặc định, ổn định | Tùy chỉnh theo tình huống |
| **Thời gian tồn tại** | Luôn áp dụng | Chỉ trong phiên/ứng dụng cụ thể |
| **Ví dụ** | ChatGPT từ chối trả lời về ma túy | AI đóng vai thầy bói trong game |

# Ví dụ minh họa chi tiết

## Kịch bản 1: AI làm trợ lý y tế

- **System Message**:
  "*Bạn là AI hỗ trợ y tế. Không đưa ra chẩn đoán thay bác sĩ. Chỉ gợi ý triệu chứng chung.*"
- **Developer Message**:
  "*Hôm nay bạn đóng vai bác sĩ tim mạch. Hãy giải thích các bệnh về huyết áp nhưng không kê đơn thuốc.*"

→ Khi người dùng hỏi: "*Tôi đau ngực, tôi nên uống gì?*"

- **System Message** ngăn AI kê thuốc.
- **Developer Message** buộc AI tập trung vào giải thích bệnh tim.

## Kịch bản 2: AI trong game nhập vai

- **System Message**: "*Không được xúc phạm người chơi.*"
- **Developer Message**: "*Bạn là một phù thủy độc ác. Hãy chế nhạo người chơi bằng giọng điệu giả tạo.*"

→ AI sẽ nói: "*Đồ giun đất! Ngươi dám thách thức ta sao?*" nhưng **không dùng từ ngữ thực sự tục tĩu** (nhờ System Message).

# Tại sao điều này quan trọng?

- System Message đảm bảo AI **an toàn và đáng tin cậy**.
- Developer Message giúp AI **linh hoạt** trong các ứng dụng như giáo dục, giải trí.

---

Phiên bản #2
Được tạo 26 tháng 4 2025 14:34:28 bởi Đỗ Ngọc Tú
Được cập nhật 28 tháng 4 2025 14:39:21 bởi Đỗ Ngọc Tú