

“Hàng Rào An Toàn” Cho AI - Guardrails

Guardrails là gì?

Guardrails (tạm dịch: *hàng rào an toàn*) là các cơ chế, quy tắc, hoặc giới hạn **được thiết kế để kiểm soát hành vi của AI**, đảm bảo AI hoạt động **đúng mục đích, an toàn, và đáng tin cậy**.

Chúng giống như "lan can" trên đường cao tốc, ngăn AI vượt khỏi phạm vi mong muốn – dù AI có khả năng suy diễn linh hoạt hoặc tự động xử lý nhiều bước.

Tại sao cần Guardrails?

Khi sử dụng **AI agent** hoặc **LLM (mô hình ngôn ngữ lớn)** trong các hệ thống tự động như:

- Gửi email tự động
- Trò chuyện với khách hàng
- Gọi API bên ngoài
- Quyết định lựa chọn công cụ

Thì **việc thiếu kiểm soát** có thể gây ra:

Tình huống	Rủi ro
Agent tự tạo nội dung độc hại	Vi phạm đạo đức/hành vi
Agent gửi email không phù hợp	Ảnh hưởng danh tiếng doanh nghiệp
Agent gọi API sai	Gây lỗi hệ thống hoặc tốn chi phí
Agent trả lời sai lệch	Gây hiểu nhầm hoặc sai thông tin

Vì vậy, **Guardrails giúp định hướng và giới hạn hành vi AI một cách có kiểm soát**.

Guardrails hoạt động như thế nào?

Một số kỹ thuật hoặc chiến lược để thiết lập Guardrails gồm:

1. Instructions rõ ràng (hướng dẫn vai trò cụ thể)

Ví dụ:

"Bạn là Giám đốc bán hàng. Không bao giờ tự tạo email. Luôn sử dụng các công cụ được cung cấp."

→ Hạn chế AI không tự viết email, buộc phải sử dụng tool hợp lệ.

2. Tách biệt giữa "Tool" và "Handoff"

Loại	Vai trò
Tool (công cụ)	Gọi API, thực hiện hành động cụ thể
Handoff (bàn giao)	Chuyển quyền điều khiển sang agent khác

Giúp đảm bảo rõ ràng từng phần trong quy trình được kiểm soát và có thể audit.

Trace & quan sát (theo dõi agent)

Sử dụng tính năng **trace** để theo dõi:

- Agent nào đã làm gì?
- Dùng công cụ nào?
- Thời điểm bàn giao ra sao?

Điều này **giúp kiểm tra và khôi phục lại hành vi nếu có sai sót**.

Các kiểu Guardrails phổ biến

Loại Guardrails	Mô tả
Ràng buộc logic	Nếu không hài lòng với kết quả, thử lại bằng công cụ khác
Luật đạo đức / an toàn	Không thảo luận về chủ đề nhạy cảm hoặc cá nhân hóa quá mức
Giới hạn hành động	Chỉ được phép gửi email sau khi kiểm tra toàn bộ tools
Theo dõi chi tiết (trace)	Kiểm tra lại toàn bộ hành trình trước khi gửi

Bài học thực tế từ Sales Automation

Trong ví dụ "From Function Calls to Agent Autonomy":

- AI sales manager **được hướng dẫn cụ thể** không tự tạo email.
- AI **phải dùng đủ 3 công cụ tạo email trước khi chọn cái tốt nhất**.
- Sau đó **handoff sang Email Manager** để gửi.
- Mọi hành động đều được **trace và kiểm soát**.

→ Đây là mô hình agent có Guardrails rõ ràng.

Thực hành đề xuất

Bài tập 1: Xác định Guardrails trong đoạn sau

““Bạn là Giám đốc bán hàng. Không bao giờ tự tạo email. Luôn sử dụng các công cụ được cung cấp.””

Câu hỏi:

- Guardrails nào đang được áp dụng?
- Mục tiêu của guardrails đó là gì?

Bài tập 2: Viết Guardrails cho Agent Tuyển dụng

Hãy viết các hướng dẫn guardrails cho một AI agent tuyển dụng, với các quy tắc như:

- Luôn hỏi ứng viên 3 câu hỏi trước khi gửi thông tin cho quản lý
- Không tự quyết định nhận ứng viên

Bài tập 3: Quan sát hành vi agent qua Trace

Dựa trên dữ liệu trace của ví dụ trên:

- Agent nào thực hiện những bước nào?
- Handoff xảy ra ở đâu?
- Guardrails đã giúp kiểm soát hành vi nào?

Guardrails là **linh hồn của hệ thống AI có trách nhiệm**. Chúng:

Giữ AI hoạt động đúng vai trò

Tránh sai lệch, vi phạm đạo đức hoặc kỹ thuật

Tạo cơ sở để mở rộng agent một cách an toàn

Tác giả: **Đỗ Ngọc Tú**
Công Ty Phần Mềm **VHTSoft**

Phiên bản #2

Được tạo 30 tháng 4 2025 03:11:48 bởi Đỗ Ngọc Tú

Được cập nhật 2 tháng 5 2025 07:45:21 bởi Đỗ Ngọc Tú