LangSmith, Promptfoo, và TruLens

1. **LangSmith** – Giám sát và kiểm thử pipelines trong LangChain

LangSmith là một nền tảng **được phát triển bởi LangChain** giúp bạn:

- Ghi lại và giám sát các pipeline tương tác với LLM.
- Kiểm tra và đánh giá chất lượng các lời gọi đến LLM.
- Phát hiện lỗi, theo dõi hiệu suất và so sánh prompt/agent chains.

Tính năng chính:

- **Trace** toàn bộ luồng hoạt động trong LangChain (gồm các agent, tool, retriever…).
- **Compare** giữa các phiên bản prompt hoặc mô hình.
- **Test Suites**: tạo và chạy bộ test trên các prompt.
- **Feedback System**: thêm đánh giá thủ công hoặc tự động.

Dùng khi:

- Bạn đang dùng LangChain để xây dựng app dùng LLM.
- Muốn kiểm tra, debug hoặc theo dõi các phiên bản của mô hình/prompt.

2. **Promptfoo** – Kiểm thử và benchmark các prompt

Promptfoo là một **công cụ dòng lệnh và dashboard** giúp bạn **kiểm thử (test), so sánh (benchmark)** và đánh giá hiệu suất của **prompt**.

Tính năng chính:

- Viết **test cases** giống như unit tests cho prompt.
- So sánh nhiều mô hình (GPT-4, Claude, Mistral…) với cùng một prompt.
- Đo hiệu suất (latency, độ dài, token usage, v.v).
- Hỗ trợ tích hợp CI/CD – kiểm thử prompt tự động mỗi lần đẩy mã.

Ví dụ:

Bạn có thể viết một test YAML:

```
prompts:
  - "Summarize: {{input}}"
tests:
  - input: "This is a very long article about..."
    expected_output: "A short summary"
```

```
promptfoo test
```

Dùng khi:

- Muốn so sánh đầu ra từ nhiều mô hình hoặc nhiều phiên bản prompt.
- Muốn đảm bảo chất lượng prompt trước khi đưa vào production.

3. TruLens – Giám sát và đánh giá đạo đức, độ tin cậy, tính đúng đắn của LLM

TruLens là một framework mã nguồn mở giúp bạn:

- **Đánh giá chất lượng đầu ra LLM** (như factuality, relevance, toxicity…).
- **Tích hợp feedback tự động** (qua các đánh giá rule-based hoặc LLM-based).
- Ghi lại lịch sử lời gọi API và visual hóa qua dashboard.

Tính năng chính:

- **Instrumenting**: thêm ghi chú (instrumentation) vào app sử dụng LLM (OpenAI, LangChain...).
- **Evaluation**: cung cấp thước đo sẵn như:
  - Groundedness (tính gắn với dữ liệu truy xuất)
  - Harmfulness
  - Answer relevance
- **TruLens App**: Giao diện trực quan để duyệt và phân tích.

Dùng khi:

- Muốn **theo dõi độ đúng đắn** và **đạo đức** của LLM app.
- Cần đo lường LLM có sinh ra phản hồi sai, lệch, gây hiểu nhầm không.

So sánh nhanh:

Công cụ	Mục tiêu chính	Điểm mạnh	Khi nào dùng?

LangSmith	Giám sát & kiểm thử pipeline LLM (LangChain)	Giao diện mạnh, có trace	Khi dùng LangChain
Promptfoo	Benchmark & test prompt	CLI, CI/CD, so sánh nhiều mô hình	Khi muốn kiểm thử prompt
TruLens	Đánh giá đầu ra LLM (relevance, safety)	Tích hợp đánh giá đạo đức, factual	Khi cần đo lường chất lượng LLM

Phiên bản #1
Được tạo 7 tháng 5 2025 10:10:25 bởi Đỗ Ngọc Tú
Được cập nhật 7 tháng 5 2025 10:13:53 bởi Đỗ Ngọc Tú