

Tối ưu hóa bộ nhớ ngắn hạn bằng cách tóm tắt hội thoại

Trong bài học trước, bạn đã học cách **giảm số lượng tin nhắn trong bộ nhớ ngắn hạn** để tiết kiệm token khi làm việc với mô hình ngôn ngữ.

Trong bài học này, chúng ta sẽ tìm hiểu một **phương pháp thay thế: Tạo bản tóm tắt hội thoại** thay vì lưu toàn bộ danh sách tin nhắn.

Tóm tắt hội thoại: Ý tưởng chính

- Thay vì lưu trữ toàn bộ tin nhắn, chúng ta sẽ **tóm tắt nội dung cuộc trò chuyện**.
- Bộ nhớ ngắn hạn ("ba lô" mà chatbot mang theo) sẽ chỉ chứa **bản tóm tắt**, thay vì toàn bộ các tin nhắn.

Ưu điểm:

- Bản tóm tắt nhỏ gọn hơn nhiều so với danh sách tất cả các tin nhắn → **tiết kiệm token**.

Nhược điểm:

- Mức độ **chính xác** của phản hồi sẽ **giảm nhẹ**, vì thông tin chi tiết có thể bị mất trong quá trình tóm tắt.

Khi nào nên dùng tóm tắt hội thoại?

- **Giai đoạn phát triển / demo / beta**: Ưu tiên tiết kiệm chi phí và token → Dùng tóm tắt.
- **Giai đoạn sản phẩm chính thức (production)**: Ưu tiên hiệu suất và độ chính xác → Có thể dùng toàn bộ tin nhắn.

Ghi nhớ:

Hiệu suất và chi phí luôn cần **cân bằng** tùy vào mục tiêu dự án của bạn.

Cách triển khai

a. Cấu trúc State mới

Trong `state` của ứng dụng, ngoài khóa mặc định `messages`, bạn sẽ thêm một khóa mới:

```
state = {  
  "messages": [...],  
  "summary": "..."  
}
```

- `messages`: Lưu danh sách tin nhắn.
- `summary`: Lưu bản tóm tắt nội dung cuộc trò chuyện.

b. Logic hoạt động

1. **Bắt đầu cuộc trò chuyện** như bình thường.
2. Khi số lượng tin nhắn trong bộ nhớ **vượt quá ngưỡng** (ví dụ: 6 tin nhắn), **tạo hoặc cập nhật bản tóm tắt**.
3. Nếu chưa vượt ngưỡng, tiếp tục hội thoại bình thường.

Điều kiện:

Nếu `len(messages) > 6` → tạo / cập nhật bản tóm tắt.

c. Các thành phần cần lập trình

- **Conditional Edge**: Kiểm tra số lượng tin nhắn để quyết định có cần tóm tắt hay không.
- **Function để tóm tắt hội thoại**: Lấy danh sách tin nhắn, tạo ra một bản tóm tắt.
- **Function gửi yêu cầu tới ChatGPT**:
 - Nếu đã có `summary` → dùng tóm tắt làm context.
 - Nếu chưa có `summary` → dùng danh sách tin nhắn.

Lưu ý khi thực thi

- **Mỗi hội thoại / user / session** phải gắn với **một thread ID** riêng.
- Khi gửi request, bạn cần truyền `thread_id` để phân biệt từng cuộc trò chuyện khác nhau.

Tổng kết

- Đây là **cách thứ hai** để giảm lượng token tiêu thụ.
- Khác với cách giảm số lượng tin nhắn, ở đây ta **chuyển đổi toàn bộ nội dung thành một bản tóm tắt**.
- Bạn có thể quay lại tài liệu chi tiết và ví dụ khi cần áp dụng kỹ thuật này vào những ứng dụng thực tế.

Ghi nhớ

“Hiệu suất vs Chi phí luôn cần cân bằng. Hiểu kỹ các kỹ thuật ngay từ giai đoạn phát triển sẽ giúp bạn xây dựng các ứng dụng AI tối ưu và chuyên nghiệp hơn.”

Phiên bản #1

Được tạo 29 tháng 4 2025 03:42:03 bởi Đỗ Ngọc Tú

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