**THÀNH PHỐ HỒ CHÍ MINH**



**TRƯỜNG ĐẠI HỌC SÀI GÒN**

Ảnh có chứa vòng tròn, biểu tượng, Phông chữ, văn bản

Mô tả được tạo tự động

**ĐỒ ÁN MÔN HỌC**

**XỬ LÝ NGÔN NGỮ TỰ NHIÊN**

**ĐỀ TÀI:**

**MEMORAG**

*Giảng viên hướng dẫn:*

TS. Nguyễn Hồng Bửu Long

*Sinh viên thực hiện:*

Lý Quốc Thành - 3121411194

Thành phố Hồ Chí Minh, 14 tháng 12 năm 2024

Nội dung

[LỜI NÓI ĐẦU 3](#_Toc185180968)

[I. Giới thiệu chung 4](#_Toc185180969)

[1. Mục tiêu của MemoRAG 4](#_Toc185180970)

[2. Ý nghĩa và ứng dụng 4](#_Toc185180971)

[II. Giải quyết bài toán 5](#_Toc185180972)

[1. Vấn đề bài toán 5](#_Toc185180973)

[a) Khả năng cập nhật thông tin 5](#_Toc185180974)

[b) Bộ nhớ ngắn hạn 5](#_Toc185180975)

[c) Kết hợp thông tin 5](#_Toc185180976)

[2. Giải quyết bài toán 6](#_Toc185180977)

[III. Kiểm nghiệm chương trình 8](#_Toc185180978)

[Tài liệu tham khảo 11](#_Toc185180979)

[Lời kết 12](#_Toc185180980)

# LỜI NÓI ĐẦU

Trong thời đại dữ liệu số hóa bùng nổ, việc khai thác thông tin từ nhiều nguồn trở nên ngày càng cần thiết nhưng cũng đầy thách thức. Các hệ thống trí tuệ nhân tạo hiện đại, như GPT, đã chứng minh tiềm năng trong việc hỗ trợ tổng hợp và truy xuất thông tin. MemoRAG (Memory-Augmented Retrieval-Augmented Generation) là một dự án đột phá, tích hợp các khả năng truy xuất và tạo văn bản để mang lại trải nghiệm mạnh mẽ trong xử lý ngôn ngữ tự nhiên (NLP).

Báo cáo này sẽ giới thiệu dự án MemoRAG từ mục tiêu thiết kế, các phương pháp triển khai kỹ thuật, đến kiểm nghiệm hiệu quả thực tế. Mục tiêu chính là làm sáng tỏ cách MemoRAG tận dụng các mô hình ngôn ngữ lớn (LLMs) kết hợp với cơ chế truy xuất thông tin để giải quyết bài toán trí tuệ nhân tạo trong thực tiễn.

# Giới thiệu chung

## Mục tiêu của MemoRAG

MemoRAG được thiết kế nhằm giải quyết hạn chế cố hữu của các mô hình ngôn ngữ hiện đại, đặc biệt là:

* **Truy xuất ngữ cảnh bên ngoài:** Mô hình GPT mặc định chỉ dựa trên dữ liệu huấn luyện và bộ nhớ ngắn hạn, khiến nó không thể cập nhật thông tin mới nhất hoặc truy cập dữ liệu tùy chỉnh.
* **Bộ nhớ lâu dài:** MemoRAG tích hợp khả năng lưu trữ và truy xuất dữ liệu, cho phép mô hình ghi nhớ và sử dụng dữ liệu đã truy xuất trước đó.

Mục tiêu cụ thể:

* Kết hợp các phương pháp truy xuất thông tin (retrieval) với khả năng tạo văn bản (generation) của GPT.
* Cung cấp một giải pháp mạnh mẽ trong xử lý ngôn ngữ tự nhiên (NLP), ứng dụng trong các lĩnh vực như trợ lý ảo, quản lý tri thức, và tìm kiếm thông tin.

## Ý nghĩa và ứng dụng

MemoRAG có tiềm năng lớn trong việc cải thiện các ứng dụng thực tế như:

* **Trợ lý ảo thông minh:** Cung cấp câu trả lời chi tiết dựa trên dữ liệu cập nhật từ các nguồn như tài liệu doanh nghiệp hoặc cơ sở tri thức tùy chỉnh.
* **Hệ thống hỏi đáp (QA):** Truy xuất và trình bày thông tin chính xác dựa trên các truy vấn phức tạp.
* **Quản lý tài liệu:** Hỗ trợ lưu trữ và truy xuất dữ liệu hiệu quả cho các tổ chức hoặc cá nhân.

So với các phương pháp truyền thống, MemoRAG mang lại khả năng học hỏi liên tục, tự động cập nhật và kết hợp dữ liệu bên ngoài một cách hiệu quả.

# Giải quyết bài toán

## Vấn đề bài toán

Trong xử lý ngôn ngữ tự nhiên (NLP), các mô hình ngôn ngữ lớn (LLMs) như GPT có khả năng xử lý và tạo văn bản một cách tự nhiên. Tuy nhiên, chúng gặp phải những hạn chế:

### Khả năng cập nhật thông tin

* Mô hình GPT không thể tự động cập nhật thông tin sau khi được huấn luyện, dẫn đến việc phản hồi không chính xác nếu dữ liệu đã lỗi thời.
* Ví dụ: GPT không biết các sự kiện xảy ra sau năm huấn luyện của nó.

### Bộ nhớ ngắn hạn

* GPT hoạt động dựa trên ngữ cảnh đầu vào, giới hạn bởi độ dài chuỗi (token). Điều này khiến mô hình khó xử lý các câu hỏi phức tạp cần thông tin từ nhiều nguồn.

### Kết hợp thông tin

* Việc truy xuất và kết hợp dữ liệu từ các nguồn khác nhau không được tích hợp tốt trong các mô hình thuần túy, dẫn đến phản hồi không đầy đủ hoặc thiếu chính xác.

**Bài toán cần giải quyết:**

Làm thế nào để:

* **Tích hợp khả năng truy xuất thông tin:** Kết hợp thông tin từ các nguồn dữ liệu bên ngoài như cơ sở tri thức, tài liệu, API, hoặc web.
* **Tăng cường bộ nhớ lâu dài:** Cho phép mô hình ghi nhớ thông tin qua các truy vấn liên tiếp, giảm thiểu sự phụ thuộc vào dữ liệu truy xuất lặp lại.
* **Cải thiện độ chính xác và chi tiết:** Tạo phản hồi chính xác và phù hợp hơn với các truy vấn phức tạp.

## Giải quyết bài toán

Để giải quyết các vấn đề trên, MemoRAG áp dụng kiến trúc **Memory-Augmented Retrieval-Augmented Generation (RAG)** với ba bước chính:

**a) Tích hợp truy xuất thông tin (Retrieval)**

* **Hướng tiếp cận:**
  + Sử dụng FAISS hoặc ElasticSearch để tìm kiếm và truy xuất các đoạn thông tin liên quan từ dữ liệu lưu trữ.
  + Các truy vấn từ người dùng được chuyển đổi thành biểu diễn vector để tìm kiếm hiệu quả trong không gian ngữ nghĩa.
* **Ví dụ:**  
  Người dùng hỏi: *"Tóm tắt tài liệu về công nghệ AI trong giáo dục."*
  + MemoRAG tìm kiếm các đoạn văn bản liên quan trong cơ sở dữ liệu hoặc trên internet.

**b) Áp dụng bộ nhớ lâu dài (Memory)**

* **Hướng tiếp cận:**
  + Các kết quả truy xuất được lưu vào một bộ nhớ lâu dài (sử dụng Redis hoặc PostgreSQL).
  + Bộ nhớ này không chỉ lưu trữ thông tin mà còn theo dõi ngữ cảnh của các truy vấn trước đó để cải thiện phản hồi trong các truy vấn tiếp theo.
* **Ví dụ:**  
  Người dùng tiếp tục hỏi: *"Tài liệu nào đề cập đến hiệu quả của AI trong việc cá nhân hóa học tập?"*
  + MemoRAG sử dụng thông tin trong bộ nhớ từ truy vấn trước để phản hồi nhanh và chính xác hơn.

**c) Kết hợp thông tin và tạo văn bản (Generation)**

* **Hướng tiếp cận:**
  + MemoRAG sử dụng các đoạn thông tin truy xuất và dữ liệu từ bộ nhớ làm ngữ cảnh đầu vào cho GPT.
  + Mô hình GPT sau đó tạo ra câu trả lời tự nhiên, đầy đủ và chi tiết.
* **Ví dụ:**  
  Kết hợp các đoạn văn bản từ tài liệu khác nhau để tạo ra một phản hồi như:  
  *"AI đã cải thiện hiệu quả giáo dục thông qua các ứng dụng cá nhân hóa học tập như học qua chatbot và phân tích dữ liệu sinh viên."*

**Cách MemoRAG vượt trội:**

* **Tích hợp thông tin động:** Không bị giới hạn bởi dữ liệu huấn luyện tĩnh.
* **Ghi nhớ ngữ cảnh:** Kết hợp dữ liệu cũ và mới một cách thông minh.
* **Khả năng mở rộng:** Hoạt động tốt trên nhiều loại dữ liệu, từ tài liệu nội bộ đến các nguồn trực tuyến.

# Kiểm nghiệm chương trình

* Cài đặt các môi trường cần thiết

Ảnh có chứa văn bản, phần mềm, ảnh chụp màn hình, Phần mềm đa phương tiện

Mô tả được tạo tự động

* Cài đặt mô hình **memorag-qwen2-7b-inst**

Ảnh có chứa văn bản, phần mềm, Phần mềm đa phương tiện, Biểu tượng máy tính

Mô tả được tạo tự động

* Tải dữ liệu văn bản một nửa cuốn sách “Harry Potter”

Ảnh có chứa văn bản, phần mềm, Phần mềm đa phương tiện, Phông chữ

Mô tả được tạo tự động

* Tải bộ nhớ đã được lưu trữ trước

Ảnh có chứa văn bản, ảnh chụp màn hình, phần mềm, Phần mềm đa phương tiện

Mô tả được tạo tự động

Ảnh có chứa văn bản, Phần mềm đa phương tiện, phần mềm, ảnh chụp màn hình

Mô tả được tạo tự động

* Thực hiện QA (Question Answering)

Ảnh có chứa phần mềm, văn bản, Phần mềm đa phương tiện, ảnh chụp màn hình

Mô tả được tạo tự động

* Thực hiện truy xuất thông tin

Ảnh có chứa văn bản, ảnh chụp màn hình, phần mềm, Phần mềm đa phương tiện

Mô tả được tạo tự động

# Tài liệu tham khảo

1. Dự án MemoRAG trên GitHub: <https://github.com/qhjqhj00/MemoRAG>
2. Tài liệu về FAISS: <https://faiss.ai>
3. OpenAI GPT API: <https://platform.openai.com/docs/>
4. Hướng dẫn ElasticSearch: <https://www.elastic.co/guide/>
5. Bài nghiên cứu về Retrieval-Augmented Generation (RAG): Lewis et al., "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks," 2020.
6. Tài liệu tham khảo khác trong lĩnh vực NLP:

* <https://www.nltk.org/>
* https://huggingface.co/transformers/

# Lời kết

MemoRAG là một bước tiến quan trọng trong lĩnh vực xử lý ngôn ngữ tự nhiên (NLP), đặc biệt trong việc kết hợp giữa truy xuất thông tin (retrieval) và tạo văn bản (generation). Bằng cách tích hợp bộ nhớ dài hạn và khả năng truy xuất mạnh mẽ, MemoRAG đã vượt qua nhiều hạn chế cố hữu của các mô hình ngôn ngữ lớn như GPT, đồng thời mở ra tiềm năng ứng dụng trong nhiều lĩnh vực thực tiễn.

Tuy vẫn còn một số thách thức cần khắc phục, MemoRAG cho thấy sự hứa hẹn trong việc phát triển các hệ thống AI thông minh hơn, đáp ứng tốt hơn nhu cầu của con người. Với những cải tiến trong tương lai, MemoRAG có thể trở thành một công cụ không thể thiếu trong các ứng dụng quản lý tri thức, giáo dục, và trợ lý ảo.