Nghiên cứu paper: LightRAG: Simple and Fast Retrieval-Augmented Generation

Link:<https://arxiv.org/abs/2410.05779>

# Kiến trúc

## 1. Tổng quan

LightRAG là một hệ thống Retrieval-Augmented Generation (RAG) được thiết kế để khắc phục hạn chế của các hệ thống RAG truyền thống. Thay vì chỉ dựa vào các phân đoạn văn bản flat text chunks để lập chỉ mục và truy xuất thông tin, LightRAG tích hợp cấu trúc đồ thị (knowledge graph) nhằm khai thác mối quan hệ giữa các thực thể trong dữ liệu. Điều này giúp hệ thống:

* Hiểu rõ hơn ngữ cảnh toàn cục (comprehensive context) qua các liên kết nhiều bước.
* Cải thiện hiệu suất truy xuất và giảm thời gian xử lý nhờ việc kết hợp vector hóa và biểu diễn đồ thị.

## 2. Graph-based Text Indexing (Lập chỉ mục văn bản dựa trên đồ thị)

Quá trình lập chỉ mục trong LightRAG gồm ba bước chính:

* **Trích xuất Thực thể và Mối quan hệ:** Dữ liệu văn bản được chia thành các đoạn nhỏ (chunks) để dễ xử lý. Sử dụng LLM (mô hình ngôn ngữ lớn) để nhận diện và trích xuất các thực thể (như tên, ngày tháng, địa điểm, sự kiện) và mối quan hệ giữa chúng.
* **LLM Profiling cho Sinh cặp Key-Value**: Sau khi trích xuất, hệ thống sử dụng một hàm profiling dựa trên LLM để gán cho mỗi nút (thực thể) và cạnh (mối quan hệ) một cặp key-value.

Key: Thường là từ khóa hoặc cụm từ ngắn (ví dụ: tên của thực thể) được dùng để truy xuất nhanh.

Value: Là đoạn văn bản tóm tắt các thông tin quan trọng liên quan, hỗ trợ cho việc sinh văn bản phản hồi.

* **Deduplication (Hợp nhất dữ liệu trùng lặp)**: Vì các thực thể và mối quan hệ có thể được trích xuất từ nhiều đoạn văn bản khác nhau, nên hệ thống thực hiện bước hợp nhất các thông tin trùng lặp.

Việc này giúp giảm kích thước của đồ thị, từ đó cải thiện hiệu quả xử lý khi truy xuất. Nhờ vào quá trình này, LightRAG xây dựng được một kiến trúc đồ thị toàn diện, giúp hệ thống thấy được mối liên hệ giữa các thông tin ở nhiều cấp độ khác nhau, từ đó nâng cao khả năng hiểu và truy xuất ngữ cảnh phức tạp.

## 3. Dual-level Retrieval Paradigm (Mô hình truy xuất hai cấp độ)

Để giải quyết đa dạng các truy vấn từ người dùng, LightRAG áp dụng một mô hình truy xuất hai cấp:

* **Truy xuất cấp độ thấp (Low-Level Retrieval):**Tập trung vào truy xuất các thực thể cụ thể cùng với các thuộc tính, mối quan hệ trực tiếp.

Ví dụ: Truy vấn “Ai là tác giả của ‘Pride and Prejudice’?” sẽ dẫn đến việc lấy ra thực thể “Jane Austen” cùng với các mối liên hệ liên quan.

* **Truy xuất cấp độ cao (High-Level Retrieval):** Được thiết kế để khai thác các chủ đề chung, các ý tưởng trừu tượng, hay mối quan hệ gián tiếp giữa các thực thể.

Ví dụ: Truy vấn “Trí tuệ nhân tạo ảnh hưởng như thế nào đến giáo dục hiện đại?” đòi hỏi hệ thống phải tổng hợp thông tin từ nhiều thực thể và mối quan hệ khác nhau để tạo ra một cái nhìn tổng quan.

**Kết hợp với Vector hóa:** Khi có truy vấn, hệ thống thực hiện:

* Trích xuất từ khóa: Phân chia truy vấn thành các từ khóa cục bộ (local) và toàn cục (global).
* So khớp từ khóa: Sử dụng cơ sở dữ liệu vector để so khớp các từ khóa với các thực thể và mối quan hệ đã được lập chỉ mục.
* Khai thác mối liên hệ bậc cao: Hệ thống còn mở rộng truy vấn bằng cách thu thập các nút lân cận (one-hop neighbors) để nắm bắt thông tin liên quan thêm.

Nhờ vậy, LightRAG có khả năng cung cấp kết quả truy xuất chính xác và toàn diện, vừa đáp ứng được các yêu cầu chi tiết lẫn tổng quát của người dùng.

## 4. Retrieval-Augmented Answer Generation (Sinh phản hồi kết hợp truy xuất)

Sau khi đã truy xuất được các thực thể và mối quan hệ liên quan, LightRAG thực hiện các bước sau để sinh phản hồi:

**Tích hợp thông tin:** Các đoạn văn được thu thập từ các thực thể và mối quan hệ được kết hợp lại thành một ngữ cảnh tổng hợp.

**Sinh phản hồi**: LLM sử dụng ngữ cảnh tổng hợp cùng với truy vấn gốc để sinh ra câu trả lời hoàn chỉnh, đảm bảo tính chính xác và phù hợp với yêu cầu người dùng.

Cách tiếp cận này giúp hệ thống không chỉ dựa vào kiến thức đã được huấn luyện mà còn bổ sung thông tin mới từ các nguồn bên ngoài, cho ra những câu trả lời phong phú và có căn cứ hơn.

# Độ đo để đánh giá

## So sánh cặp câu trả lời:

Với mỗi truy vấn, hệ thống so sánh câu trả lời từ LightRAG với câu trả lời của một mô hình RAG khác (ví dụ NaiveRAG, GraphRAG, …). Sự so sánh này được thực hiện cho từng khía cạnh:

* Comprehensiveness: Đánh giá mức độ bao quát, chi tiết của câu trả lời.
* Diversity: Đánh giá sự đa dạng, phong phú của các góc nhìn được cung cấp.
* Empowerment: Đánh giá khả năng giúp người dùng hiểu rõ và đưa ra quyết định dựa trên thông tin được cung cấp.

## Đánh giá theo cặp sử dụng LLM:

Một mô hình ngôn ngữ GPT-4o-mini được sử dụng làm đánh giá để so sánh trực tiếp hai câu trả lời. Mô hình này được hướng dẫn qua một prompt đặc biệt, đưa ra lựa chọn cho mỗi tiêu chí (ví dụ: “Winner”: "Answer 2 (LightRAG)" nếu câu trả lời của LightRAG tốt hơn).

# Chạy thử nghiệm

1. **Yêu cầu**

* Python phiên bản 3.11.0
* Các thư viện: aiohttp, future, gensim, pipmaster, pydantic, python-dotenv, setuptools, tenacity, tiktoken:
  + aiohttp: Thư viện HTTP bất đồng bộ (async) cho Python.
  + future: Hỗ trợ tương thích giữa Python 2 và Python 3.
  + gensim: Xử lý ngôn ngữ tự nhiên (NLP), chủ yếu dùng cho mô hình chủ đề (topic modeling).
  + pipmaster: Không phổ biến, có thể là một công cụ hỗ trợ quản lý package.
  + pydantic: Xác thực và xử lý dữ liệu với kiểu dữ liệu Python.
  + python-dotenv: Quản lý biến môi trường từ .env.
  + setuptools: Hỗ trợ tạo và cài đặt package Python.
  + tenacity: Cung cấp cơ chế retry (thử lại) khi gặp lỗi.
  + tiktoken: Mã hóa token, dùng cho OpenAI LLM (GPT, ChatGPT).

1. **Các tính năng đã thử nghiệm**

* Truy vấn bằng các mô hình Ollama.
* Trực quan đồ thị bằng html.

**III. Các bước thử nghiệm**

* Tạo 1 bản sao github directory <https://github.com/HKUDS/LightRAG.git>:

git clone <https://github.com/HKUDS/LightRAG.git>

* Direct sang thư mục LightRAG, thiết lập môi trường

ảo và tải thư viện cần thiết.

cd lightrag

conda create -n lightrag\_3 python=3.11

activate lightrag\_3

pip install lightrag-hku

pip install -e .

* Tải ollama và sử dụng ollama để tải các model: nomic-embed-text

và qwen2

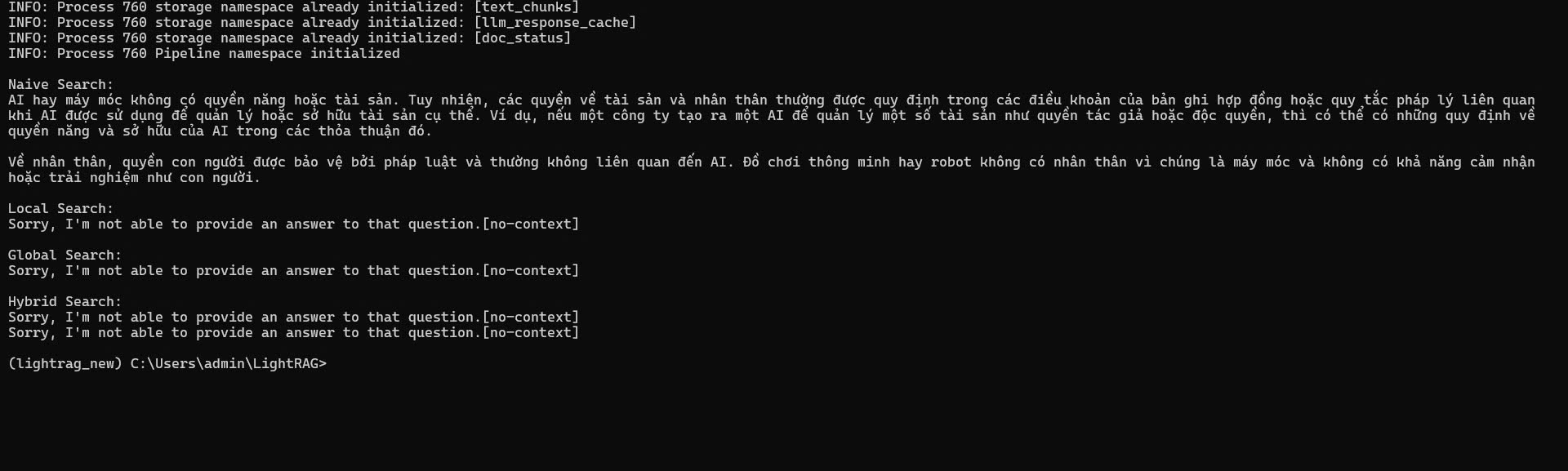
ollama pull nomic-embed-text

ollama pull qwen2

* Tải data dưới định dạng txt và tạo modelfile.
* Đặt các tham số: ctx\_num= llm max token size =2048, embedding dim = 768, embedding max token size = 512.
* Chạy file examples\lightrag\_ollama\_demo.py và examples\graph\_visual\_with\_html.py để thu kết quả.

**IV. Kết quả**

* Kết quả truy vấn:



* Kết quả đồ thị: Không có.

### **So sánh câu trả lời model LightRAG trên và RAG (BAAI/bge-small-en + Mistral-7B-Instruct-v0.1):** 1. Cá nhân có quy ền kết hôn, ly hôn, quy ền bình đẳng của vợ chồng, quyền xác định cha, mẹ, con, quyền được nhận làm con nuôi, quyền nuôi con nuôi và các quyền nhân thân khác trong quan h ệ hôn nhân, quan h ệ cha mẹ và con và quan h ệ giữa các thành viên gia đình. Con sinh ra không phụ thuộc vào tình trạng hôn nhân của cha, mẹ đều có quyền và nghĩa vụ như nhau đối với cha, mẹ của mình.

### 2. Cá nhân có quy ền kết hôn, ly hôn, quy ền bình đẳng của vợ chồng, quyền xác định cha, mẹ, con, quyền được nhận làm con nuôi, quyền nuôi con nuôi và các quyền nhân thân khác trong quan h ệ hôn nhân, quan h ệ cha mẹ và con và quan h ệ giữa các thành viên gia đình. Con sinh ra không phụ thuộc vào tình trạng hôn nhân của cha, mẹ đều có quyền và nghĩa vụ như nhau đối với cha, mẹ của mình. **dựa trên các tiêu chí:**

| **Tiêu chí** | **LightRAG** | **RAG** |
| --- | --- | --- |
| **Comprehensiveness (Mức độ bao quát, chi tiết)** | ⚠️ Chỉ tập trung vào AI và máy móc, chưa đề cập rõ ràng đến quyền nhân thân và tài sản của con người theo pháp luật. | ✅ Trả lời chi tiết về quyền nhân thân trong quan hệ hôn nhân, gia đình. Tuy nhiên, chưa đề cập đến quyền tài sản. |
| **Diversity (Sự đa dạng, phong phú của góc nhìn)** | ✅ Cung cấp góc nhìn mới về quyền nhân thân và tài sản liên quan đến AI, nhưng thiếu thông tin về pháp luật chung. | ⚠️ Chỉ tập trung vào quan hệ gia đình, chưa đề cập đến quyền nhân thân trong các lĩnh vực khác (danh dự, uy tín, hình ảnh…) và quyền tài sản. |
| **Empowerment (Khả năng giúp người dùng hiểu rõ và đưa ra quyết định)** | ⚠️ Mở rộng suy nghĩ về AI nhưng không hữu ích trong việc hiểu quyền nhân thân và tài sản theo pháp luật. | ✅ Giúp người dùng hiểu rõ về quyền nhân thân trong hôn nhân và gia đình, có thể áp dụng ngay trong thực tế. |

### **Tổng kết:**

* **Câu trả lời RAG tốt hơn** vì cung cấp thông tin đúng trọng tâm về quyền nhân thân trong quan hệ hôn nhân và gia đình. Tuy nhiên, nó **chưa đề cập đến quyền tài sản**.
* **Câu trả lời LightRAG có góc nhìn thú vị về AI nhưng không trả lời đúng câu hỏi theo pháp luật**, nên không hữu ích nếu người dùng muốn tìm hiểu quy định thực tế.