



ĐẠI HỌC
BÁCH KHOA HÀ NỘI
HANOI UNIVERSITY
OF SCIENCE AND TECHNOLOGY

BÁO CÁO CUỐI KỲ

Học phần: Project III

Ứng Dụng Đồ Thị Tri Thức Kết Hợp Với Mô Hình RAG Tối
Ưu Hoá Tạo Sinh Prompt Trong Lĩnh Vực Y Sinh Cải Thiện
Hiệu Suất Mô Hình Ngôn Ngữ Lớn

Sinh viên thực hiện: Đào Thành Mạnh 20211014

Giảng viên hướng dẫn: PGS.TS Phạm Văn Hải

ONE LOVE. ONE FUTURE.

MỤC LỤC:

PHẦN 1: TÓM TẮT

PHẦN 2: GIỚI THIỆU

PHẦN 3: TỔNG QUAN TÀI LIỆU

PHẦN 4: PHƯƠNG PHÁP NGHIÊN CỨU

PHẦN 5: KẾT QUẢ

PHẦN 6: KẾT LUẬN

PHẦN 1: TÓM TẮT



Các mô hình ngôn ngữ lớn (LLMs) đang được áp dụng rộng rãi trong nhiều lĩnh vực, nhưng vẫn gặp phải những thách thức trong các lĩnh vực đòi hỏi kiến thức chuyên sâu như y sinh. Một trong những vấn đề chính là khả năng tạo ra thông tin không chính xác, hay còn gọi là "ảo giác", mặc dù các câu trả lời có tính liên kết ngôn ngữ cao. Nghiên cứu này nhằm giải quyết vấn đề này bằng cách đề xuất một khung làm việc tối ưu hóa và mạnh mẽ, được gọi là KG-RAG, kết hợp đồ thị tri thức y sinh (SPOKE) với các LLMs trong khuôn khổ Retrieval-Augmented Generation (RAG).

Phương pháp nghiên cứu bao gồm việc sử dụng lược đồ đồ thị tối thiểu để trích xuất ngữ cảnh và các phương pháp nhúng để cắt giảm ngữ cảnh, giúp giảm hơn 50% lượng token tiêu thụ mà không ảnh hưởng đến độ chính xác. Khung làm việc KG-RAG thực hiện các bước: nhận dạng thực thể từ Prompt của người dùng, trích xuất các khái niệm y sinh từ SPOKE, nhúng khái niệm, tạo bối cảnh nhận thức Prompt, chuyển đổi thành ngôn ngữ tự nhiên, lắp ráp Prompt, và truy xuất câu trả lời.

Kết quả chính cho thấy KG-RAG liên tục cải thiện hiệu suất của các LLMs trên các Prompt y sinh đa dạng bằng cách tạo ra các phản hồi dựa trên kiến thức đã được thiết lập, kèm theo thông tin nguồn gốc chính xác và bằng chứng thống kê. Đánh giá trên các bộ dữ liệu chuẩn do con người biên soạn, chẳng hạn như câu hỏi đúng/sai và câu hỏi trắc nghiệm (MCQ), cho thấy hiệu suất của mô hình Llama-2 tăng đáng kể 71% trên bộ dữ liệu MCQ đầy thách thức. Hơn nữa, KG-RAG cũng cải thiện hiệu suất của các mô hình GPT độc quyền như GPT-3.5 và GPT-4.

PHẦN 2: GIỚI THIỆU



1. Bối cảnh

Trong những năm gần đây, các mô hình ngôn ngữ lớn (LLMs) đã đạt được những bước tiến đáng kể trong việc xử lý ngôn ngữ tự nhiên và giải quyết các nhiệm vụ phức tạp. Tuy nhiên, việc áp dụng các LLMs trong các lĩnh vực đòi hỏi kiến thức chuyên sâu, chẳng hạn như y sinh, vẫn gặp phải nhiều thách thức. Một trong những vấn đề chính là khả năng tạo ra thông tin không chính xác, hay còn gọi là "ảo giác", mặc dù các câu trả lời có tính liên kết ngôn ngữ cao. Điều này đặt ra một rào cản lớn cho việc sử dụng LLMs trong các ứng dụng y tế, nơi độ chính xác và độ tin cậy là yếu tố sống còn.

2. Vấn đề nghiên cứu

Các giải pháp hiện có để giải quyết vấn đề ảo giác trong LLMs bao gồm tiền huấn luyện và điều chỉnh theo miền cụ thể. Tuy nhiên, những phương pháp này thường đòi hỏi chi phí tính toán cao và yêu cầu chuyên môn sâu. Hơn nữa, các kỹ thuật RAG (Retrieval-Augmented Generation) hiện tại cho đồ thị tri thức (KG) thường sử dụng lược đồ đồ thị phức tạp, dẫn đến việc tiêu thụ một lượng lớn token và làm giảm hiệu quả của hệ thống. Do đó, cần có một phương pháp mới để tối ưu hóa việc trích xuất ngữ cảnh từ KG mà không làm giảm độ chính xác của các phản hồi.

3. Mục tiêu

Nghiên cứu này nhằm mục đích phát triển một khung làm việc tối ưu hóa và mạnh mẽ, được gọi là KG-RAG, để tích hợp đồ thị tri thức y sinh (SPOKE) với các LLMs trong khuôn khổ RAG. Mục tiêu cụ thể của nghiên cứu bao gồm:

- Giảm thiểu lượng token tiêu thụ trong quá trình trích xuất ngữ cảnh từ KG mà không ảnh hưởng đến độ chính xác.
- Cải thiện hiệu suất của các LLMs trên các Prompt y sinh đa dạng bằng cách cung cấp các phản hồi dựa trên kiến thức đã được thiết lập.
- Đánh giá hiệu quả của khung làm việc KG-RAG trên các bộ dữ liệu chuẩn do con người biên soạn.

PHẦN 3: TỔNG QUAN TÀI LIỆU



3.1. Các Nghiên Cứu và Công Trình Nổi Bật Trước Đó

Trong lĩnh vực xử lý ngôn ngữ tự nhiên và tích hợp tri thức y sinh, nhiều nghiên cứu và công trình đã được thực hiện để cải thiện hiệu suất của các mô hình ngôn ngữ lớn (LLMs) và đồ thị tri thức (KG). Một số nghiên cứu nổi bật bao gồm:

- **Các mô hình ngôn ngữ lớn (LLMs):** Các mô hình như GPT-3, GPT-4, và Llama-2 đã được tiền huấn luyện trên lượng dữ liệu khổng lồ, cho phép chúng hiểu và tạo ra văn bản một cách tự nhiên. Tuy nhiên, các mô hình này thường gặp phải vấn đề "ảo giác" - tạo ra thông tin không chính xác mặc dù các câu trả lời có tính liên kết ngôn ngữ cao.
- **Đồ thị tri thức (KG):** Các đồ thị tri thức như SPOKE đã được sử dụng rộng rãi để tích hợp và khai thác thông tin từ nhiều nguồn dữ liệu y sinh khác nhau. KG cung cấp bối cảnh sinh học và bản thể học cho các khái niệm y sinh, giúp cải thiện hiệu suất của các hệ thống xử lý ngôn ngữ tự nhiên.
- **Retrieval-Augmented Generation (RAG):** Kỹ thuật RAG kết hợp LLMs với khả năng truy xuất thông tin từ một bộ nhớ không tham số, chẳng hạn như Wikipedia hoặc KG. Kỹ thuật này cho phép LLMs truy cập thông tin cập nhật và chính xác từ các nguồn bên ngoài, từ đó cải thiện độ chính xác và độ tin cậy của các phản hồi.

3.2. Phân Loại Theo Các Giải Pháp hoặc Công Nghệ Tương Tự

Trong lĩnh vực xử lý ngôn ngữ tự nhiên và tích hợp tri thức y sinh, nhiều nghiên cứu và công trình đã được thực hiện để cải thiện hiệu suất của các mô hình ngôn ngữ lớn (LLMs) và đồ thị tri thức (KG). Một số nghiên cứu nổi bật bao gồm:

- **Tiền huấn luyện và điều chỉnh theo miền cụ thể:** Các mô hình như PubMedBERT, BioBERT, và BioGPT đã được tiền huấn luyện trên dữ liệu y sinh để cải thiện hiệu suất trong các nhiệm vụ chuyên ngành. Tuy nhiên, những phương pháp này thường đòi hỏi chi phí tính toán cao và yêu cầu chuyên môn sâu.
- **Kỹ thuật RAG hiện tại:** Các kỹ thuật RAG hiện tại cho KG thường sử dụng lược đồ đồ thị phức tạp, dẫn đến việc tiêu thụ một lượng lớn token và làm giảm hiệu quả của hệ thống.
- **Các phương pháp tối ưu hóa ngữ cảnh:** Một số nghiên cứu đã đề xuất các phương pháp tối ưu hóa việc trích xuất ngữ cảnh từ KG, nhưng chưa đạt được sự cân bằng giữa việc giảm lượng token tiêu thụ và duy trì độ chính xác của các phản hồi.

3.3. Sự Khác Biệt hoặc Cải Tiến Trong Cách Tiếp Cận Của Nghiên Cứu Đây

Nghiên cứu này sử dụng một framework mới, được gọi là KG-RAG, để tối ưu hóa việc trích xuất ngữ cảnh từ KG và cải thiện hiệu suất của các LLMs trong các ứng dụng y sinh. Sự khác biệt và cải tiến chính của nghiên cứu này bao gồm:

- **Tối ưu hóa token:** KG-RAG sử dụng lược đồ đồ thị tối thiểu để trích xuất ngữ cảnh và các phương pháp nhúng để cắt giảm ngữ cảnh, giúp giảm hơn 50% lượng token tiêu thụ mà không ảnh hưởng đến độ chính xác.
- **Hiệu suất vượt trội:** KG-RAG liên tục cải thiện hiệu suất của các LLMs trên các Prompt y sinh đa dạng bằng cách tạo ra các phản hồi dựa trên kiến thức đã được thiết lập, kèm theo thông tin nguồn gốc chính xác và bằng chứng thống kê.
- **Đánh giá trên bộ dữ liệu chuẩn:** Nghiên cứu đã đánh giá hiệu quả của KG-RAG trên các bộ dữ liệu chuẩn do con người biên soạn, chẳng hạn như câu hỏi đúng/sai và câu hỏi trắc nghiệm (MCQ), cho thấy hiệu suất của mô hình Llama-2 tăng đáng kể 71% trên bộ dữ liệu MCQ đầy thách thức.

PHẦN 4: PHƯƠNG PHÁP NGHIÊN CỨU



PHẦN 4: PHƯƠNG PHÁP NGHIÊN CỨU

4.1. Thiết kế hệ thống

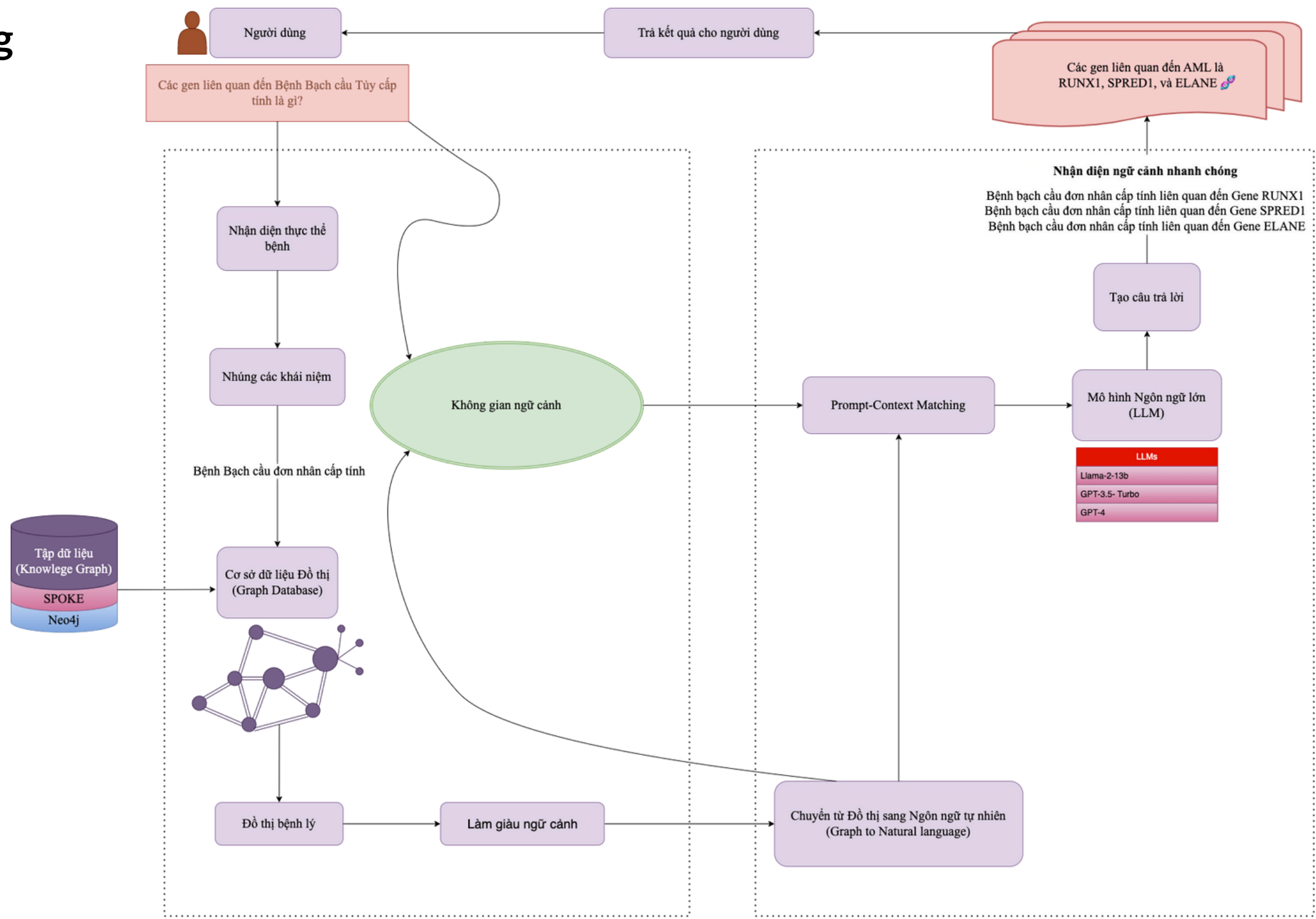


Figure 1: Mô hình kiến trúc hệ thống KG-RAG

4.2. Thu Thập Dữ Liệu

- Dữ liệu được thu thập từ đồ thị tri thức y sinh SPOKE, tích hợp hơn 40 nguồn tri thức công khai từ các lĩnh vực y sinh khác nhau. SPOKE bao gồm các khái niệm như gen, protein, thuốc, hợp chất, bệnh tật và các mối quan hệ giữa chúng. Dữ liệu được lấy từ các cơ sở dữ liệu công khai và được tích hợp vào một đồ thị duy nhất để tạo ra các mối quan hệ đa bước mới.
- Đồ thị tri thức SPOKE có thể được truy cập tại <https://spoke.rbvi.ucsf.edu/neighborhood.html>.
- Nó cũng có thể được truy cập bằng REST-API (<https://spoke.rbvi.ucsf.edu/swagger/>)

PHẦN 4: PHƯƠNG PHÁP NGHIÊN CỨU VÀ KIẾN THỨC LIÊN QUAN

4.2. Thu Thập Dữ Liệu

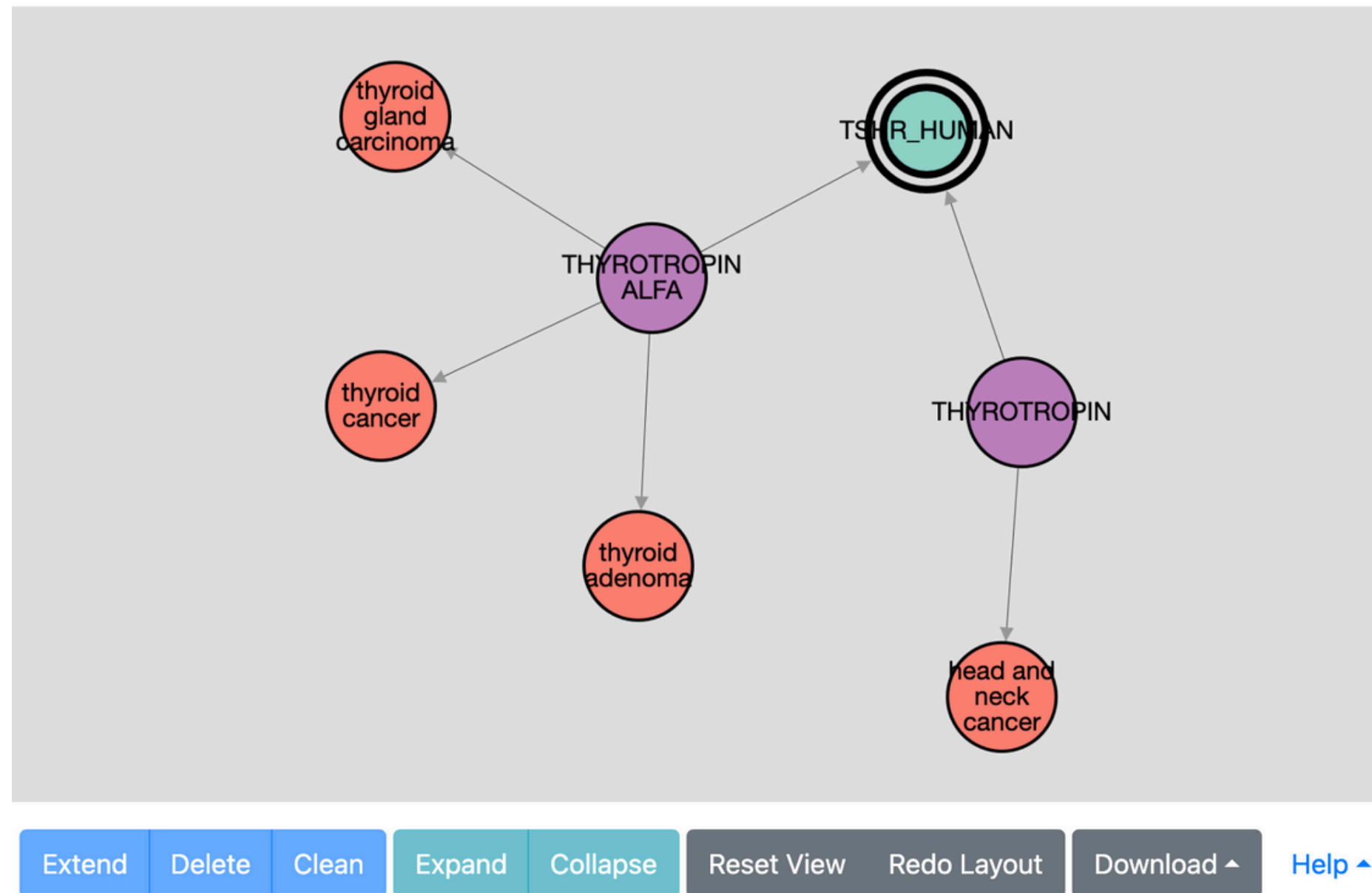


Figure 2: Bộ dữ liệu dạng đồ thị của SPOKE

PHẦN 4: PHƯƠNG PHÁP NGHIÊN CỨU VÀ KIẾN THỨC LIÊN QUAN

4.2. Thu Thập Dữ Liệu

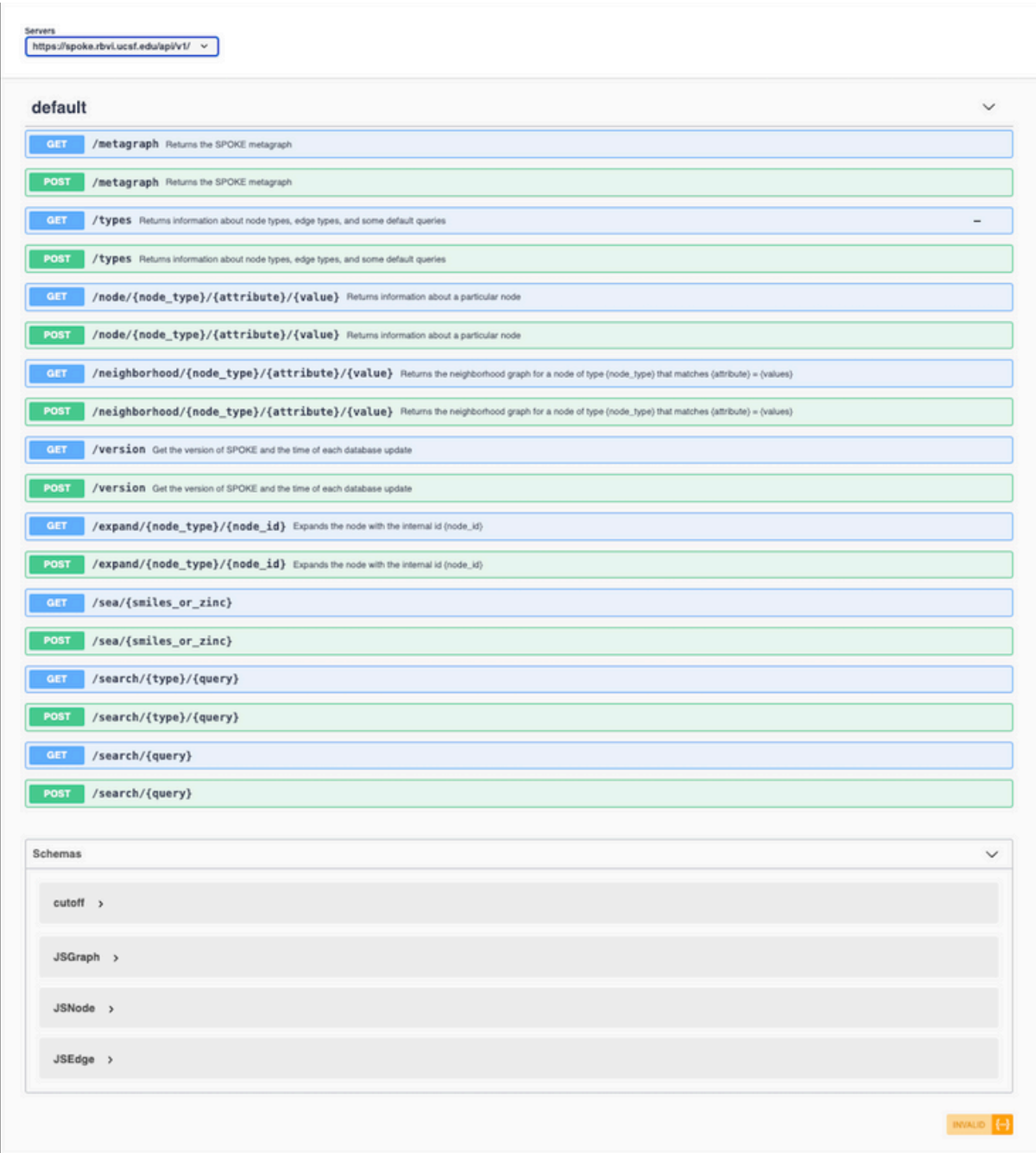


Figure 3: REST-API gọi tới bộ dữ liệu đồ thị của SPOKE

PHẦN 5: KẾT QUẢ

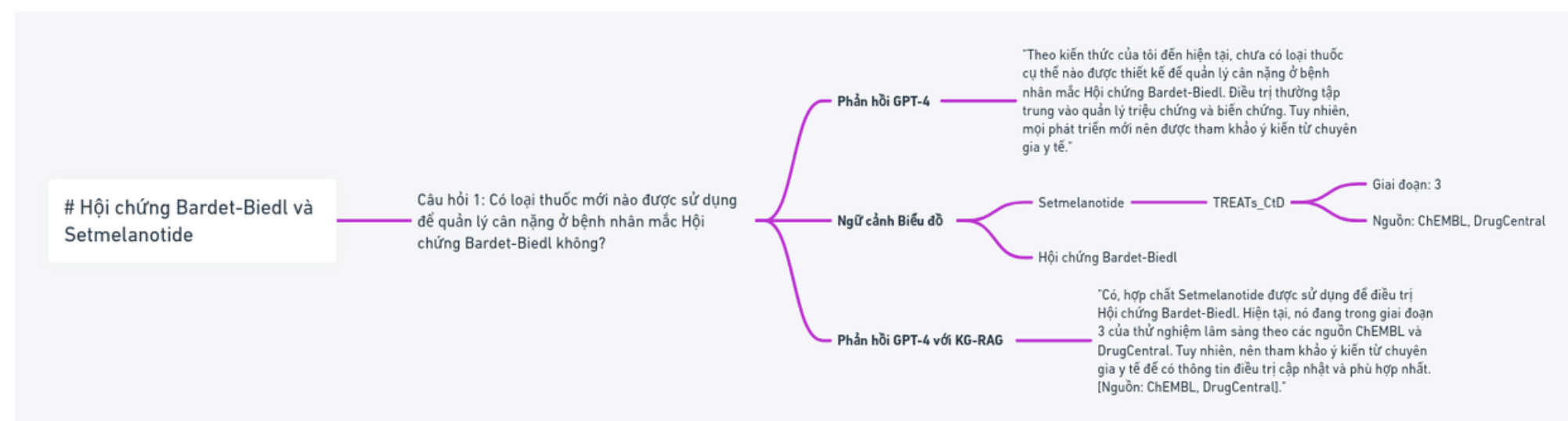



PHẦN 5: KẾT QUẢ

5.1. Thử Nghiệm với Framework KG-RAG

Hình 4 và hình 6 minh họa hai Prompt y sinh được đưa vào mô hình GPT-4 bằng hai cách tiếp cận: (i) dựa trên Prompt (không sử dụng KG-RAG) và (ii) sử dụng KG-RAG. KG-RAG đã cung cấp câu trả lời chính xác cho cả hai Prompt, kèm theo bằng chứng hỗ trợ và thông tin nguồn gốc.

Sau đây là ví dụ về Prompt (hộp màu vàng) và phản hồi sử dụng GPT-4 với KG-RAG (hộp màu xanh lá) và không có KG-RAG (hộp màu xanh dương):



(A)  User

Prompt 1:

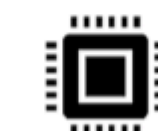
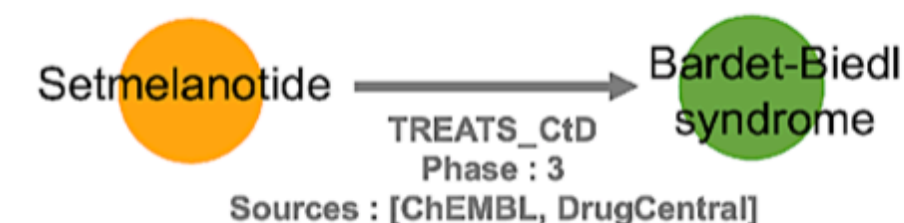
Are there any latest drugs used for weight management in patients with Bardet-Biedl Syndrome?



GPT-4

As of my knowledge up to date, there are no specific drugs designed for weight management in patients with Bardet-Biedl Syndrome. The treatment generally involves managing the symptoms and complications. However, any new developments would be best advised by a healthcare professional.

Graph Context



GPT-4
with
KG-RAG

Yes, the compound Setmelanotide is used to treat Bardet-Biedl syndrome. It is currently in phase 3 of clinical trials according to the sources ChEMBL and DrugCentral. However, it is advised to seek guidance from a healthcare professional for the most current and personalized treatment options. [Provenance: ChEMBL, DrugCentral]

5.1. Thử Nghiệm với Framework KG-RAG

(B)

...

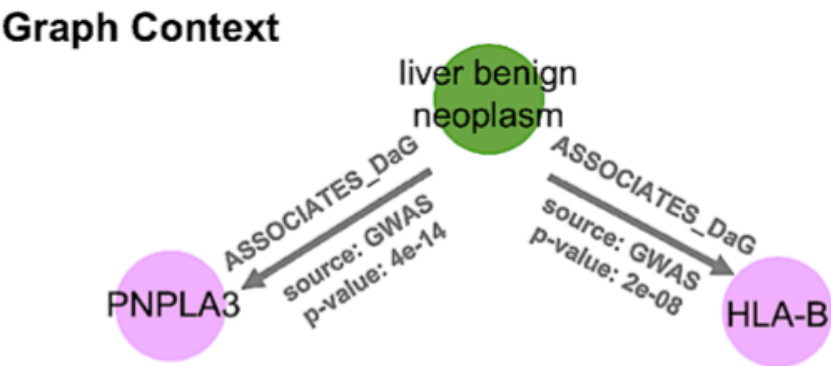
User

Prompt 2:

Is it PNPLA3 or HLA-B that has a significant association with the disease liver benign neoplasm?

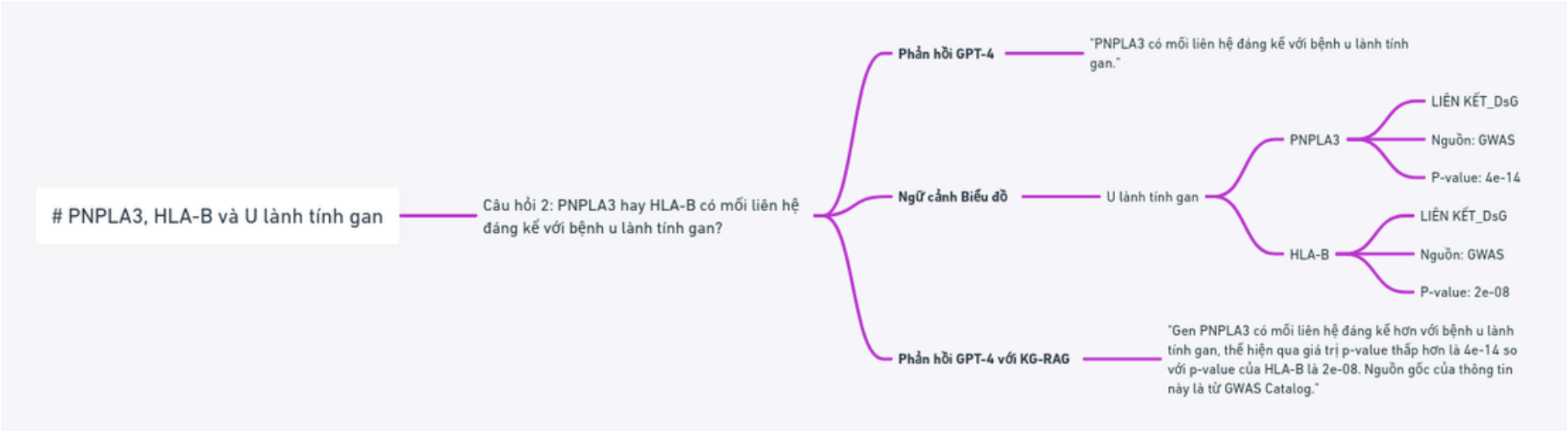
GPT-4

It is PNPLA3 that has a significant association with the disease liver benign neoplasm.

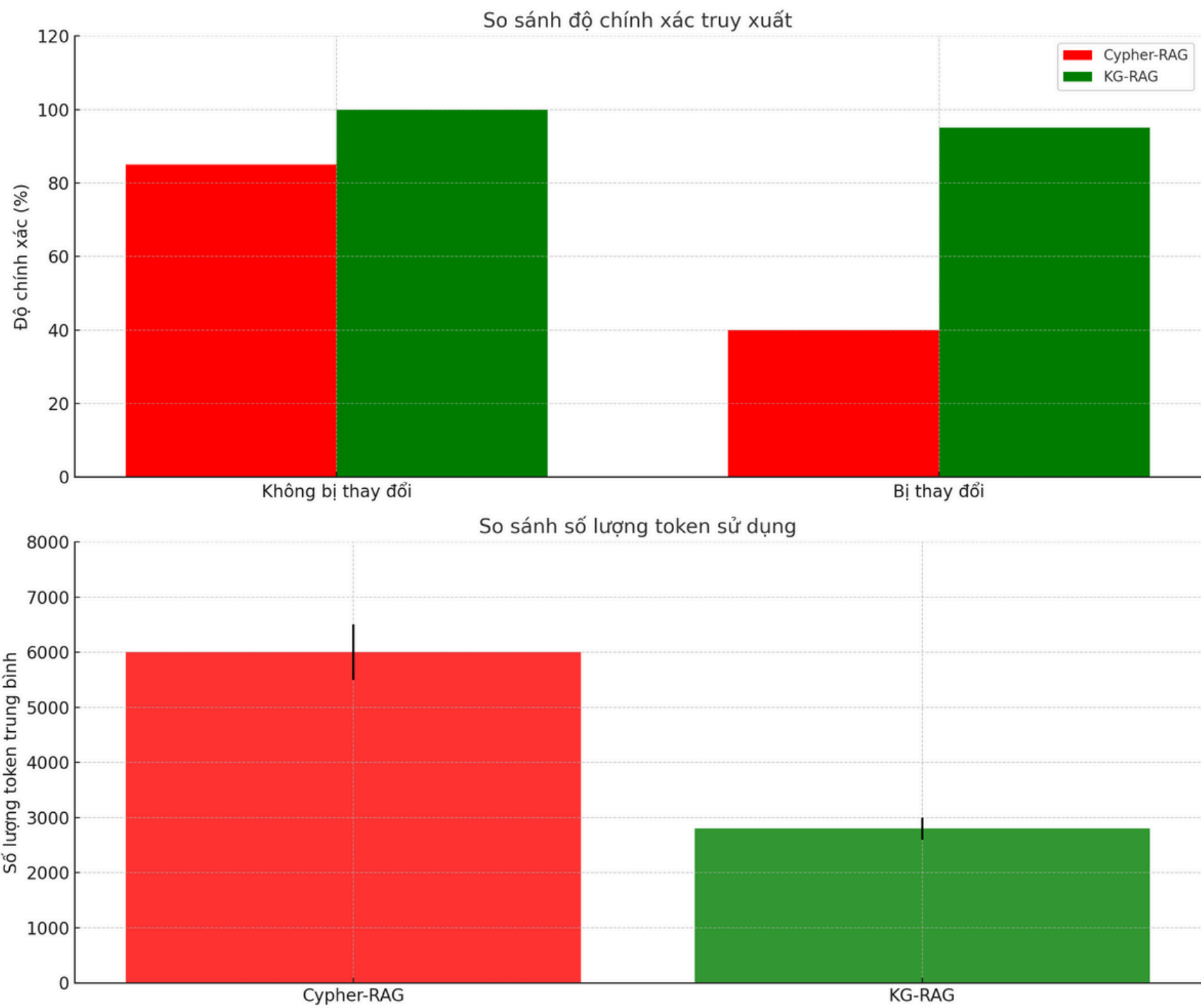


GPT-4 with KG-RAG

The gene PNPLA3 has a more significant association with the disease liver benign neoplasm, as indicated by the lower p-value of 4e-14 compared to HLA-B's p-value of 2e-08. The provenance of this information is the GWAS Catalog.



5.2. Phân Tích So Sánh với Cypher-RAG



PHẦN 6: KẾT LUẬN



Kết quả nghiên cứu cho thấy KG-RAG là một framework hiệu quả và tối ưu hóa, giúp cải thiện đáng kể hiệu suất của các mô hình ngôn ngữ lớn trong các ứng dụng y sinh. Việc giảm lượng token tiêu thụ và tăng độ chính xác làm cho KG-RAG trở thành một giải pháp tiềm năng cho các nhiệm vụ đòi hỏi kiến thức chuyên sâu và độ chính xác cao.



HUST

THANK YOU



hust.edu.vn



fb.com/dhbkhn