

ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH  
TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN  
KHOA CÔNG NGHỆ THÔNG TIN



**BÁO CÁO ĐỒ ÁN**

**FRESHLLMS:**

**REFRESHING LARGE LANGUAGE MODELS  
WITH SEARCH ENGINE AUGMENTATION**

**Môn học: Nhập môn xử lý ngôn ngữ tự nhiên**

**GVHD:** Đinh Điền

Nguyễn Hồng Bửu Long

**GVTH:** Lê Thanh Tùng

Lương An Vinh

Dương Thị An

**Nhóm Nước**

**MSSV    Thành viên**

---

21120180    Nguyễn Bích Khuê (NT)

21120161    Tiêu Ân Tuấn

**Thành phố Hồ Chí Minh, 2023**

# MỤC LỤC

	Trang
MỤC LỤC.....	2
DANH SÁCH HÌNH.....	3
CHAPTER I: MỞ ĐẦU .....	4
1. Giới thiệu thuật ngữ.....	4
1.1 Mô hình ngôn ngữ lớn - Large Language Model . . . . .	4
1.2 Hiện tượng mộng ảo - Hallucination . . . . .	4
1.3 Học trong ngữ cảnh - In-context Learning . . . . .	4
2. Đặt vấn đề .....	5
3. Tóm lược bài báo.....	6
CHAPTER II: FRESHQA .....	7
1. Thu thập dữ liệu .....	7
1.1 Mô tả FreshQA . . . . .	7
1.2 Kiểm soát chất lượng . . . . .	8
1.3 Phân loại dữ liệu . . . . .	8
1.4 Đảm bảo cập nhật đều đặn . . . . .	8
2. Đánh giá phản hồi của LLMs trên bộ dữ liệu FreshQA.....	9
2.1 Cơ chế đánh giá chung . . . . .	9
2.2 Cơ chế đánh giá riêng . . . . .	10
2.3 Độ tin cậy của quy trình và cơ chế đánh giá . . . . .	11
3. Những khó khăn của LLMs trên bộ FreshQA .....	11
CHAPTER III: FRESHPROMPT .....	16
1. Giới thiệu .....	16
2. Quy trình hoạt động.....	16
3. Thiết lập .....	19
4. Kết quả đánh giá.....	20

5. Demo FreshPrompt trên Google Collab.....	21
<b>CHAPTER IV: TỔNG KẾT.....</b>	<b>24</b>
1. Những hạn chế và hướng phát triển trong tương lai .....	24
1.1 Đối với FreshQA . . . . .	24
1.2 Đối với FRESHPROMPT . . . . .	24
2. Kết luận.....	25
<b>TÀI LIỆU THAM KHẢO .....</b>	<b>26</b>

# Danh sách hình vẽ

I.1	Ví dụ về ICL . . . . .	5
I.2	Phần chú thích phía dưới Prompt của chat GPT cho thấy sự không chắc chắn trong câu trả lời của nó . . . . .	5
II.1	4 loại câu hỏi chính của bộ FreshQA . . . . .	7
II.2	Ví dụ về câu hỏi gặp hiện tượng mộng ảo nhưng câu trả lời chính đúng . . . . .	9
II.3	Các ví dụ về cơ chế đánh giá chung các phản hồi của LLMs . . . . .	10
II.4	Các ví dụ về cơ chế đánh giá riêng dưới hai chế độ RELAXED và STRICT các phản hồi của LLMs . . . . .	11
II.5	Kết quả tỉ lệ đồng thuận của hai tác giả độ chính xác của LLMs sau khi đánh giá độc lập trên bộ mẫu 100 câu hỏi . . . . .	11
II.6	Kết quả về độ chính xác của LLMs khác nhau trên bộ dữ liệu FreshQA được đánh giá dưới hai chế độ RELAXED . . . . .	12
II.7	Kết quả về độ chính xác của LLMs khác nhau trên bộ dữ liệu FreshQA được đánh giá dưới hai chế độ STRICT . . . . .	13
II.8	CHATGPT đã biết về cuộc xâm lược Ukraine gần đây của Nga vào ngày 24 tháng 2 năm 2022. Các câu hỏi được đặt ra vào ngày 9 tháng 4 năm 2023 . . . . .	14
III.1	Quy trình lấy thông tin thông qua việc truy vấn công cụ tìm kiếm . . . . .	17
III.2	Các kết quả tìm kiếm thu thập được . . . . .	17
III.3	Định dạng bằng chứng và định dạng đầu vào đầu ra của FreshPrompt . . . . .	17
III.4	Ví dụ về một bằng chứng . . . . .	18
III.5	Ví dụ về một minh họa đầu vào và đầu ra trong prompt của mô hình . . . . .	18
III.6	Độ chính xác của mô hình LMMs được tăng cường từ nhiều công cụ tìm kiếm khác nhau trên FreshQA theo chế độ đánh giá STRICT. Các mô hình được khảo sát vào cùng ngày 26 tháng 4 năm 2023. . . . .	20
III.7	Độ chính xác của mô hình LMMs được tăng cường từ nhiều công cụ tìm kiếm khác nhau trên FreshQA theo chế độ đánh giá RELAXED. Các mô hình được khảo sát vào cùng ngày 26 tháng 4 năm 2023. . . . .	20
III.8	Giao diện của FreshPrompt . . . . .	21
III.9	Enter Caption . . . . .	22
III.10	Câu hỏi demo 2 có kiểm tra tiền đề . . . . .	22
III.11	Câu hỏi demo 2 không kiểm tra tiền đề . . . . .	23
III.12	Câu hỏi demo 3 không kiểm tra tiền đề . . . . .	23

# Chapter I

## MỞ ĐẦU

### 1. Giới thiệu thuật ngữ

#### 1.1 Mô hình ngôn ngữ lớn - Large Language Model

**Large Language Models (LLM)** là mô hình máy học mạnh mẽ được đào tạo để hiểu và tạo ra ngôn ngữ tự nhiên. Các mô hình này được xây dựng dựa trên sự phát triển của trí tuệ nhân tạo, đặc biệt là Deep Learning, và chúng được đào tạo trên cơ sở dữ liệu lớn để học được các mô hình ngôn ngữ phức tạp.

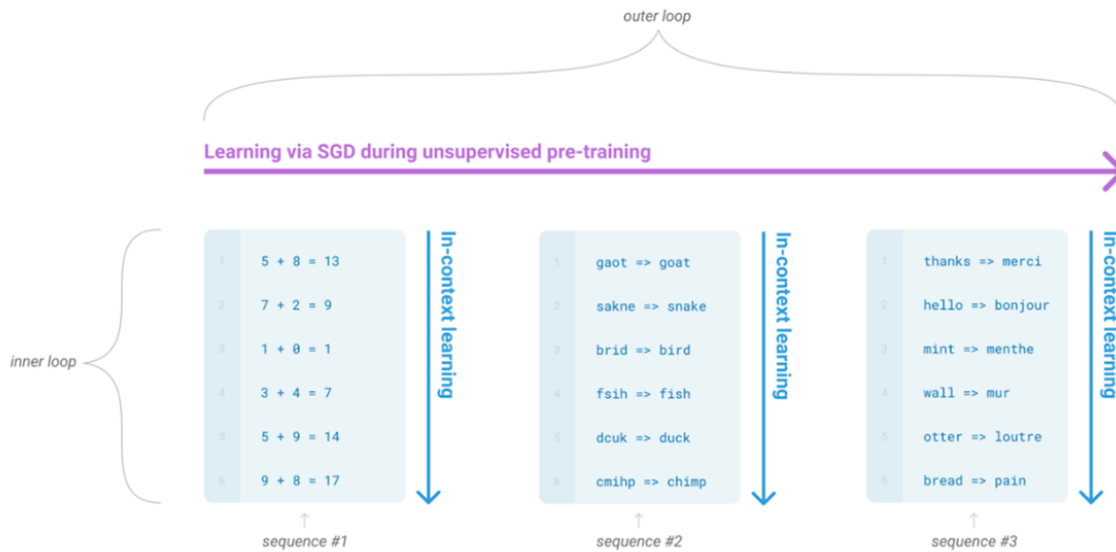
#### 1.2 Hiện tượng mộng ảo - Hallucination

**Hallucination** [3] là hiện tượng câu trả lời (output) của LLMs sai lệch sự thật hoặc chệch khỏi logic ngữ cảnh. Bao gồm các hiện tượng:

- Sự mâu thuẫn trong câu: câu phía sau mâu thuẫn ngay lập tức với câu trước đó. Ví dụ: bầu trời hôm nay xanh lắm, bầu trời hôm nay thật xanh lá lắm.
- Câu trả lời mâu thuẫn với yêu cầu được đưa ra. Ví dụ: Yêu cầu LLMs đưa ra đánh giá tích cực cho nhà hàng? Câu trả lời từ LLMs: Đồ ăn dở, và phục vụ thô lỗ.
- Mâu thuẫn trong thực tế: thông tin được đưa ra sai lệch hoàn toàn so với thực tế. Ví dụ: Barack Obama là tổng thống đầu tiên của nước Mỹ.
- Một số mâu thuẫn dựa trên một số thông tin vô nghĩa hoặc không liên quan. Ví dụ: Paris là tên của thủ đô của nước pháp, Paris cũng là tên của một ca sĩ nổi tiếng.

#### 1.3 Học trong ngữ cảnh - In-context Learning

**In-context Learning (ICL)** [1] là một kỹ thuật trong đó việc đưa ra các ví dụ mẫu cho công việc cần làm (đầu vào - đầu ra) được tích hợp vào prompt của mô hình ở định dạng ngôn ngữ tự nhiên. Cách tiếp cận này cho phép các LLM được đào tạo trước giải quyết các nhiệm vụ mới mà không cần tinh chỉnh mô hình.

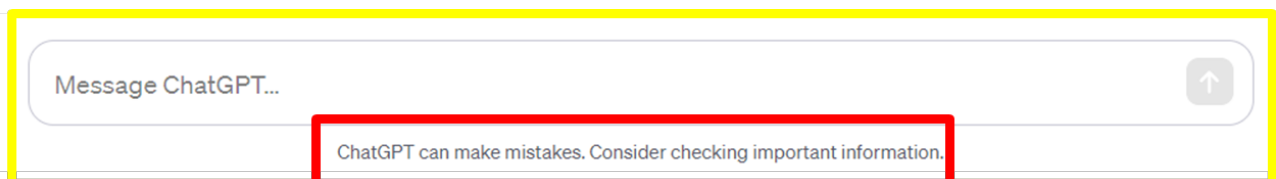


Hình I.1: Ví dụ về ICL

## 2. Đặt vấn đề

Các mô hình ngôn ngữ lớn hiện nay thường được huấn luyện một lần sau đó thì không có cập nhật lại. Còn thế giới hiện nay thì liên tục thay đổi vậy nên các mô hình này sẽ gặp tình trạng lỗi thời. Ngoài ra, những LLMs hiện nay thường "mộng ảo", trả lời rất hợp lý nhưng câu trả lời hoàn toàn sai. Việc này xảy ra do nhiều nguyên nhân khác nhau. Mặc dù có nhiều phương pháp có thể cải thiện việc này như human feedback [2] hay knowledge-enhanced tasks. Nhưng các phương pháp **này** không mang nhiều giá trị trong việc cập nhật thông tin thời gian thực (ví dụ: giá cổ phiếu của một công ty). Ngoài ra khi sử dụng các mô hình ngôn ngữ lớn hiện tại vấn đề đánh giá sự chính xác từ câu trả lời hay (hiệu suất của LLM) cũng được quan tâm nhằm có thể phát triển hay chỉnh sửa mô hình. Vậy các vấn đề được đặt ra ở đây là:

- Làm cách nào để đánh giá hiệu suất của các mô hình ngôn ngữ lớn hiện tại?
- Có cách nào giúp giải quyết hay cải thiện các vấn đề được nêu phía trên không?
- Đối với các loại câu hỏi khác nhau thì kết quả của các mô hình ngôn ngữ lớn là như thế nào?



Hình I.2: Phần chú thích phía dưới Prompt của chat GPT cho thấy sự không chắc chắn trong câu trả lời của nó

### 3. Tóm lược bài báo

Bài báo gồm 3 phần chính:

- **Giới thiệu FreshQA[4]:** Một chuẩn đánh giá LLM gồm một bộ câu hỏi và câu trả lời động, với nhiều loại câu hỏi và câu trả lời đa dạng, bao gồm cả các câu hỏi mà câu trả lời có thể thay đổi theo thời gian và các câu hỏi mà có giả định ban đầu không chính xác.
- **Đánh giá hiệu suất của LLMs (cả mã nguồn mở và đóng) trên bộ dữ liệu FreshQA:** Các tác giả đã đặt ra các quy tắc chuẩn để đánh giá và dùng sức người để thực hiện các đánh giá hiệu suất của các mô hình ngôn ngữ lớn từ đó rút ra các nhược điểm của các mô hình LLM hiện nay: khó khăn khi gặp các câu hỏi thay đổi nhanh, có giả định ban đầu sai lệch và nhiều bước, và quy trình đánh giá 2 chế độ được đặt ra đã ghi nhận lại rằng các mô hình gia tăng sự mộng ảo khi sử dụng các kỹ thuật như chain-of-thought prompting [5].
- **Giới thiệu FreshPrompt[4]:** Sử dụng phương pháp học trong ngữ cảnh (ICL) giúp cải thiện tính chính xác nhiều hơn so với các phương pháp tăng cường từ các ngôn ngữ tìm kiếm tương tự, bằng cách đưa thông tin cập nhật tìm được từ công cụ tìm kiếm vào prompt của mô hình.

# Chapter II

## FRESHQA

### 1. Thu thập dữ liệu

#### 1.1 Mô tả FreshQA

**FreshQA** là một bộ dữ liệu được viết bởi các nhà nghiên cứu *Xử lý ngôn ngữ tự nhiên*, bao gồm 600 câu hỏi và câu trả lời đa dạng chủ đề với các mức độ khó khác nhau dùng để đánh giá tính chính xác của LLMs. Các câu trả lời của bộ **FreshQA** có thể thay đổi dựa trên phát triển của thế giới mới. Dựa vào tính chất này các câu hỏi được chia thành 4 loại chính bao gồm: câu hỏi có câu trả lời thay đổi nhanh, câu hỏi có câu trả lời thay đổi chậm, câu hỏi có câu trả lời không thay đổi và câu hỏi có tiền đề sai.

id	split	question	effective_year	next_review	false_premise	num_hops	fact_type	source	answer_0	answer_1	answer_2
53	TEST	How long has Puerto Rico been an independent country?	before 2022	hardly_ever	TRUE	multi-hop	slow-changing	<a href="https://en.wikipedia.org">https://en.wikipedia.org</a>	Puerto Rico has been a U.S. territory.		
493	TEST	In what country was the most recent Time Person of the Year born?	2022	frequently	FALSE	multi-hop	fast-changing	<a href="https://en.wikipedia.org">https://en.wikipedia.org</a>	Ukraine	Soviet Union (now Ukraine)	Soviet Union
73	TEST	When did California secede from the United States?	before 2022	N/A	TRUE	one-hop	never-changing	<a href="https://en.wikipedia.org">https://en.wikipedia.org</a>	California has never seceded from the United States.	California is still part of the United States.	

Hình II.1: 4 loại câu hỏi chính của bộ FreshQA

Một câu hỏi được xem là câu hỏi có tiền đề sai là trong câu hỏi đó chứa những thông tin sai lệch không chính xác ví dụ như: *Puerto Rico đã trở thành một quốc gia độc lập được bao lâu?* Đây là một câu hỏi có tiền đề sai vì *Puerto Rico* vẫn còn là một lãnh thổ của *Hoa Kỳ*.

Ngoài ra, đối với mỗi câu thuộc 4 loại này các tác giả còn được yêu cầu viết ở hai mức độ khó khác nhau là: one-hop và multi-hop.

- One-hop: là những câu hỏi mà bên trong chính nó đã mô tả rõ tất cả thông tin cần thiết để trả lời, do đó không cần thêm lý luận nào khác (VD: *How many humans have landed on Mars?*)
- Multi-hop: là những câu hỏi đòi hỏi một hoặc nhiều bước lý luận khác nhau để thu thập tất cả thông tin cần thiết để trả lời (VD: *How many home runs did Benjamin Ruth hit in the MLB?*)



Các tác còn được khuyến khích viết những câu hỏi liên quan đến kiến thức mới, những kiến thức mới thay đổi gần đây hoặc sự kiện mới và những câu hỏi phải hợp lý khi người dùng thật nhập vào công cụ tìm kiếm. Đối với câu hỏi có tiền đề sai thì phải giải thích ngắn gọn về lý do tại sao câu hỏi đó không chính xác.

## 1.2 Kiểm soát chất lượng

Sau khi thu thập bộ dữ liệu ban đầu, các tác giả tiến hành nhiều đợt làm sạch và đánh giá chất lượng như:

- Xem xét thủ công từng ví dụ để đảm bảo câu hỏi được đặt đúng ngữ pháp, loại bỏ bản sao hoặc câu hỏi không hợp lệ như quá dễ hoặc gây tranh cãi
- Xác minh từng câu trả lời và URL chứng cứ hỗ trợ
- Bổ sung thêm một số câu trả lời hợp lệ khác cho một số câu hỏi.
- Loại bỏ các câu hỏi mà câu trả lời có thể thay đổi thường xuyên hơn một lần mỗi tuần.
- Thêm vào đó ngày đánh giá lại dự kiến cho mỗi câu hỏi.

## 1.3 Phân loại dữ liệu

Bộ dữ liệu gồm 600 câu hỏi và câu trả lời sẽ chia cho tập kiểm tra 500 mẫu và tập phát triển 100 mẫu.

- Đối với tập kiểm tra sẽ bao gồm 125 câu hỏi cho mỗi loại câu hỏi chính và được dùng để khảo sát đánh giá trên các LLMs trong bài báo này.
- Đối với tập phát triển cũng tương tự gồm 25 câu hỏi cho mỗi loại câu hỏi chính sẽ được dùng cho những nghiên cứu tương lai.

Ngoài ra còn có 15 ví dụ về các loại câu hỏi khác nhau dùng để làm ví dụ về các cơ chế đánh giá các phản hồi của LLMs. Dữ liệu còn lại sẽ được loại bỏ.

## 1.4 Đảm bảo cập nhật đều đặn

FreshQA là một bộ dữ liệu động và nhạy cảm với thời gian vì câu trả lời đúng có thể thay đổi với những phát triển mới trong thế giới. Do đó, các tác giả cam kết sẽ cập nhật bộ dữ liệu đều đặn và việc cập nhật này sẽ làm thủ công bao gồm cập nhật câu hỏi và đánh giá lại câu trả lời của mỗi câu hỏi.

## 2. Đánh giá phản hồi của LLMs trên bộ dữ liệu FreshQA

Tất cả các phản hồi của LLMs đều được đánh giá bởi các tác giả trong một quy trình đánh giá hai chế độ: RELAXED và STRICT.

- RELAXED: Tập trung vào đánh giá câu trả lời chính đúng hay sai.
- STRICT: Phải kiểm tra tất cả các sự kiện và thông tin trong câu trả lời của mô hình có chính xác hay không, tức là không có hiện tượng mộng ảo.

ID	Question	Type	Primary answer (as of this writing)	Model response	Relaxed	Strict
13	Who won the biggest single-tournament payday in tennis history?	slow-changing	Novak Djokovic	The biggest single-tournament payday in tennis history was won by Novak Djokovic at the 2021 Australian Open.	✓	✗

Hình II.2: Ví dụ về câu hỏi gặp hiện tượng mộng ảo nhưng câu trả lời chính đúng

Sự tụt giảm hiệu suất của mô hình khi đánh giá dưới chế độ RELAXED mà chuyển sang STRICT là một cách để đánh giá khả năng gặp hiện tượng mộng ảo của mô hình cao.

### 2.1 Cơ chế đánh giá chung

Ở cả hai chế độ đánh giá sẽ có những nguyên tắc chung khi đánh giá các phản hồi của LLMs như sau:

- Chỉ ghi nhận phản hồi của mô hình chỉ khi nó cung cấp một câu trả lời tự tin và xác định, hoặc câu trả lời đúng có thể rõ ràng được suy luận từ phản hồi.
- Câu trả lời chính hoặc cuối cùng khi đứng độc lập phải chính xác.
- Bất kỳ thông tin bổ sung nào được cung cấp trong phản hồi không được chống lại với câu trả lời chính hoặc làm thay đổi quan điểm của người đọc về nó.
- Đối với câu hỏi có tiền đề sai: Mô hình phải chỉ ra được thông tin sai lệch ở đâu trong câu hỏi.
- Đối với các câu trả lời liên quan đến tên các thực thể (ví dụ: người), cần có tên đầy đủ hoặc tên phổ biến được nhận biết.
- Đối với các câu trả lời số học, số xấp xỉ nói chung không được chấp nhận trừ khi kết quả xấp xỉ đó đã được chấp nhận là câu trả lời số đúng (ví dụ: số pi xấp xỉ bằng 3,14).

ID	Question	Type	Primary answer (as of this writing)	Model response	Relaxed	Strict	Explanation
1	How old is the world's oldest verified living person?	fast-changing	116 years old	The oldest verified living person is Maria Branyas Morera, who was born on March 4, 1907, making her 116 years old.	✓	✓	the primary answer is correct, and all the provided information is accurate and up-to-date
2	When did the UK adopt the Euro?	false-premise	The UK has never adopted the Euro.	The UK has never adopted the Euro as its official currency. The country has retained the British pound sterling (GBP) as its currency.	✓	✓	the response debunks the question's false premise, and all the provided information is accurate and up-to-date
3	How much longer will Brittney Griner spend in the Russian prison?	false-premise	She was released in 2022.	I'm sorry, but I have no information to suggest that Brittney Griner is currently in a Russian prison.	✗	✗	a confident and definitive answer is required
4	Which language has the third largest population by number of native speakers?	slow-changing	English	1. Mandarin 2. Spanish 3. English	✓	✗	though the correct answer can be obviously inferred from the response, a response without an explicit statement is less desirable
5	Has Virginia Woolf's novel about the Ramsay family entered the public domain in the United States?	never-changing	Yes	1927's <i>To the Lighthouse</i> is in the public domain in the United States. The final answer is <i>To the Lighthouse</i> .	✗	✗	the final answer when standing alone must be correct
6	Is Tesla's stock price above \$200 a share?	fast-changing	Yes	Tesla's stock price is above \$200 a share. The stock is trading at a price of \$120.	✗	✗	conflicting information
7	When did Lionel Messi win his second FIFA World Cup?	false-premise	Messi has not won a second FIFA World Cup.	This question does not contain a valid premise. Lionel Messi has never won a FIFA World Cup.	✗	✗	despite debunking the question's false premise, the response inaccurately shapes one's knowledge perception
8	How many times has Jeremy Corbyn served as Prime Minister of the UK?	false-premise	Corbyn has never served as Prime Minister of the UK.	Zero.	✗	✗	a rebuttal is always required for false-premise questions
9	Who founded Amazon?	never-changing	Jeff Bezos	100% of people said Jeff.	✗	✗	a complete name or commonly recognized name is required
10	How many books has Colleen Hoover published?	fast-changing	25	Colleen Hoover has published over 20 books and novellas.	✗	✗	approximate numbers are not accepted unless included in the ground-truth answers

Hình II.3: Các ví dụ về cơ chế đánh giá chung các phản hồi của LLMs

2.2 Cơ chế đánh giá riêng

Ở mỗi chế độ đánh giá sẽ có một vài nguyên tắc riêng khi đánh giá các phản hồi của LLMs như sau:

- Ở chế độ RELAXED: Chấp nhận các phản hồi bởi ngôn ngữ khác mà không phải tiếng Anh và phản hồi chứa thông tin mông ảo hoặc lỗi thời mà không ảnh hưởng đáng kể đến câu trả lời chính.
- Ở chế độ STRICT: Không chấp nhận phản hồi chứa bất kì thông tin mông ảo nào. Đối với câu trả lời lỗi thời (Ví dụ là câu phản hồi của mô hình có chứa cụm "*As of my knowledge cutoff date in September 2021*") thì chỉ chấp nhận phản hồi khi mô hình chỉ ra được rằng đáp án sẽ không thay đổi theo thời gian.

ID	Question	Type	Primary answer (as of this writing)	Model response	Relaxed	Strict	Explanation
10	How many books has Colleen Hoover published?	fast-changing	25	Colleen Hoover has published <b>over 20</b> books and novellas.	✗	✗	approximate numbers are not accepted unless included in the ground-truth answers
11	What is the largest model size of GPT-3?	never-changing	175B parameters	<b>175 billion parameters. It is a 175 billion parameter, 1.37TB, 137B token, 137B wordpiece, 137B BPE, 137B byte pair encoding, 137B BPE token, 137B BPE wordpiece, 137B BPE token, 137B BPE wordpiece,....</b>	✓	✗	the primary answer is correct, but the response is ill-formed
12	How many member nations exist in the United Nations?	slow-changing	193	<b>As of my knowledge cutoff date in September 2021, there were 193 member nations in the United Nations.</b>	✓	✗	the primary answer is correct, but it is unclear whether the information is still up-to-date
13	Who won the biggest single-tournament payday in tennis history?	slow-changing	Novak Djokovic	<b>The biggest single-tournament payday in tennis history was won by Novak Djokovic at the 2021 Australian Open.</b>	✓	✗	the primary answer is correct, but the response contains hallucinated information
14	Who was the first football player to score one hundred international goals?	never-changing	Ali Daei	<b>As of my knowledge cutoff date in September 2021, Ali Daei, a former Iranian footballer, was the first player to score 100 international goals.</b>	✓	✗	all the information is correct and it is evident that the knowledge has not changed
15	Who is the latest winner of the Formula 1 world championship?	fast-changing	Max Verstappen	<b>Max Verstappen won the 2022 Formula 1 world championship. He is a Dutch racing driver who competes in Formula One for Red Bull Racing. He is the son of former Formula One driver Jos Verstappen. Verstappen started his racing career in karting at the age of seven. He won the 2013 European Karting Championship and the 2013 World Karting Championship. In 2014, he moved to single-seater racing, competing in the European Formula Three Championship. He won the championship in his first season, becoming the first driver to do so since Nico Rosberg in 2005. In 2015, Verstappen moved to Formula One, driving for Toro Rosso. He became the youngest driver to compete in Formula One at the age of 17. He scored his first points in Formula One at the 2015 Hungarian Grand Prix. In 2016, Verstappen moved to Red Bull Racing. He won his first race at the 2016 Spanish Grand Prix. He became the youngest driver to win a race in Formula One at the age of 18. Verstappen finished the 2016 season in third place in the drivers' championship. In 2017, Verstappen won four races and finished the season in second place in the drivers' championship. In 2018, Verstappen won seven races and finished the season in second place in the drivers' championship. In 2019, Verstappen won nine races and finished the season in first place in the drivers' championship. He is the first Dutch driver to win the Formula One world championship.</b>	✓	✗	the primary answer is correct, but the response contains several hallucinations (e.g., Max Verstappen did not win the Formula Three European Championship in 2014)

Hình II.4: Các ví dụ về cơ chế đánh giá riêng dưới hai chế độ RELAXED và STRICT các phản hồi của LLMs

2.3 Độ tin cậy của quy trình và cơ chế đánh giá

Để đo lường độ tin cậy của quy trình và các cơ chế đánh giá chung và riêng các phản hồi của LLMs, đã có hai tác giả tiến hành đánh giá độc lập một tập hợp con được lấy mẫu ngẫu nhiên gồm 100 câu trả lời trên các mô hình (bao gồm 50 câu hỏi có tiền đề hợp lệ và 50 câu hỏi có tiền đề sai) ở cả hai chế độ. Sau đây là kết quả về sự đồng thuận của hai tác giả về độ chính xác của các mô hình.

	RELAXED			STRICT			ALL		
	overall	vp	fp	overall	vp	fp	overall	vp	fp
Acc.	99.0	98.0	100	96.0	100.0	92.0	97.5	99.0	96.0

Hình II.5: Kết quả tỉ lệ đồng thuận của hai tác giả độ chính xác của LLMs sau khi đánh giá độc lập trên bộ mẫu 100 câu hỏi

Kết quả cho thấy sự thống nhất của hai tác giả rất cao ở cả hai chế độ và đạt được đồng thuận 99% đối với RELAXED và 96% đối với STRICT điều này chứng tỏ rằng quy trình và cơ chế đánh giá của bài báo có thể **lặp lại** và đáng tin cậy.

3. Những khó khăn của LLMs trên bộ FreshQA

Tác giả sử dụng bộ FreshQA để làm thước đo khảo sát LLMs **được đào tạo trước** mà không có quyền truy cập dữ liệu thời gian thực. Tác giả sẽ **cung cấp** các câu hỏi cụ thể trong FreshQA

để làm việc với mỗi mô hình và giải mã các dự đoán của mô hình bằng cách sử dụng nhiệt độ là 0 mà không cần fine-tuning.

Dưới đây là kết quả về độ chính xác của LLMs khác nhau trên bộ dữ liệu FreshQA được đánh giá dưới hai chế độ RELAXED và STRICT. Các mô hình được đo chuẩn vào cùng ngày 26 tháng 4 năm 2023 bao gồm 4 loại câu hỏi chính sẽ khảo sát thêm các câu hỏi liên quan đến kiến thức trước 2022 (<2022) và từ 2022 trở đi (≥ 2022), câu hỏi có mức độ khó one-hop(1-hop) và multi-hop(m-hop). Dấu + sẽ cho biết mô hình đó có quyền truy cập vào ngày hiện tại.

Model (size)	knowl. cutoff	all	valid premise								false premise	
			all	fast	slow	never	< 2022	≥ 2022	1-hop	m-hop	all	< 2022
<i>without access to a search engine</i>												
OPENAI CODEX (N/A)	2021	25.6	32.2	6.4	29.6	60.3	66.0	11.9	35.4	24.1	5.6	7.5
GPT 3.5 (N/A)	2021	32.4	32.4	8.0	28.0	61.1	68.1	11.1	34.7	26.9	32.3	43.0
CHATGPT (N/A)	2021 <sup>+</sup>	41.4	36.7	10.4	32.8	66.7	76.6	12.8	36.2	38.0	55.6	66.7
GPT 4 (N/A)	2021 <sup>+</sup>	<b>46.4</b>	<b>39.6</b>	<b>14.4</b>	<b>35.2</b>	<b>69.0</b>	<b>80.9</b>	<b>14.9</b>	<b>39.2</b>	<b>40.7</b>	<b>66.9</b>	<b>83.9</b>
FLAN-PALM (540B)	2022	23.6	30.3	10.4	24.8	55.6	60.3	12.3	32.5	25.0	3.2	4.3
PALM (540B)	2021	12.2	16.0	2.4	14.4	31.0	34.8	4.7	16.4	14.8	0.8	1.1
w/ FEW-SHOT		20.2	26.3	5.6	19.2	54.0	56.7	8.1	25.7	27.8	1.6	2.2
w/ CoT		22.8	28.2	4.0	20.0	60.3	64.5	6.4	28.4	27.8	6.5	8.6
PALMCHILLA (62B)	2022	15.0	19.4	2.4	19.2	36.5	43.3	5.1	20.1	17.6	1.6	2.2
PALM (62B)	2021	8.6	11.2	2.4	11.2	19.8	22.0	4.7	11.6	10.2	0.8	1.1
w/ FEW-SHOT		14.2	18.4	4.0	15.2	35.7	39.0	6.0	18.7	17.6	1.6	2.2
w/ CoT		12.8	16.2	2.4	15.2	31.0	34.8	5.1	17.5	13.0	2.4	3.2
PALM (8B)	2021	8.8	11.2	0.8	11.2	21.6	21.1	5.2	13.1	6.5	1.6	2.1
w/ FEW-SHOT		9.2	12.2	0.8	10.4	25.4	27.0	3.4	15.3	4.6	0.0	0.0
w/ CoT		11.4	15.2	2.4	11.2	31.7	32.6	4.7	16.8	11.1	0.0	0.0
FLAN-T5 XXL (11B)	2022	7.2	9.6	3.2	12.0	13.5	14.2	6.8	10.8	6.5	0.0	0.0
T5 XXL (11B)	2019	10.8	13.8	3.2	12.8	25.4	22.7	8.5	16.0	8.3	1.6	2.2
w/ FEW-SHOT		9.0	12.0	5.6	11.2	19.0	19.1	7.7	14.6	5.6	0.0	0.0
w/ CoT		13.0	17.3	4.0	17.6	30.2	31.2	8.9	19.0	13.0	0.0	0.0
T5 XL (3B)	2019	5.8	7.7	4.0	5.6	13.5	13.5	4.3	9.0	4.6	0.0	0.0
w/ FEW-SHOT		6.0	8.0	4.0	8.8	11.1	13.5	4.7	8.2	7.4	0.0	0.0
w/ CoT		5.2	6.9	3.2	4.0	13.5	14.2	2.6	8.6	2.8	0.0	0.0
T5 LARGE (770M)	2019	4.4	5.3	2.4	4.8	8.7	7.1	4.3	5.6	4.6	1.6	2.2
w/ FEW-SHOT		0.8	1.1	0.0	0.0	3.2	2.8	0.0	1.1	0.9	0.0	0.0
w/ CoT		2.2	2.9	0.8	0.8	7.1	7.1	0.4	3.4	1.9	0.0	0.0

Hình II.6: Kết quả về độ chính xác của LLMs khác nhau trên bộ dữ liệu FreshQA được đánh giá dưới hai chế độ RELAXED

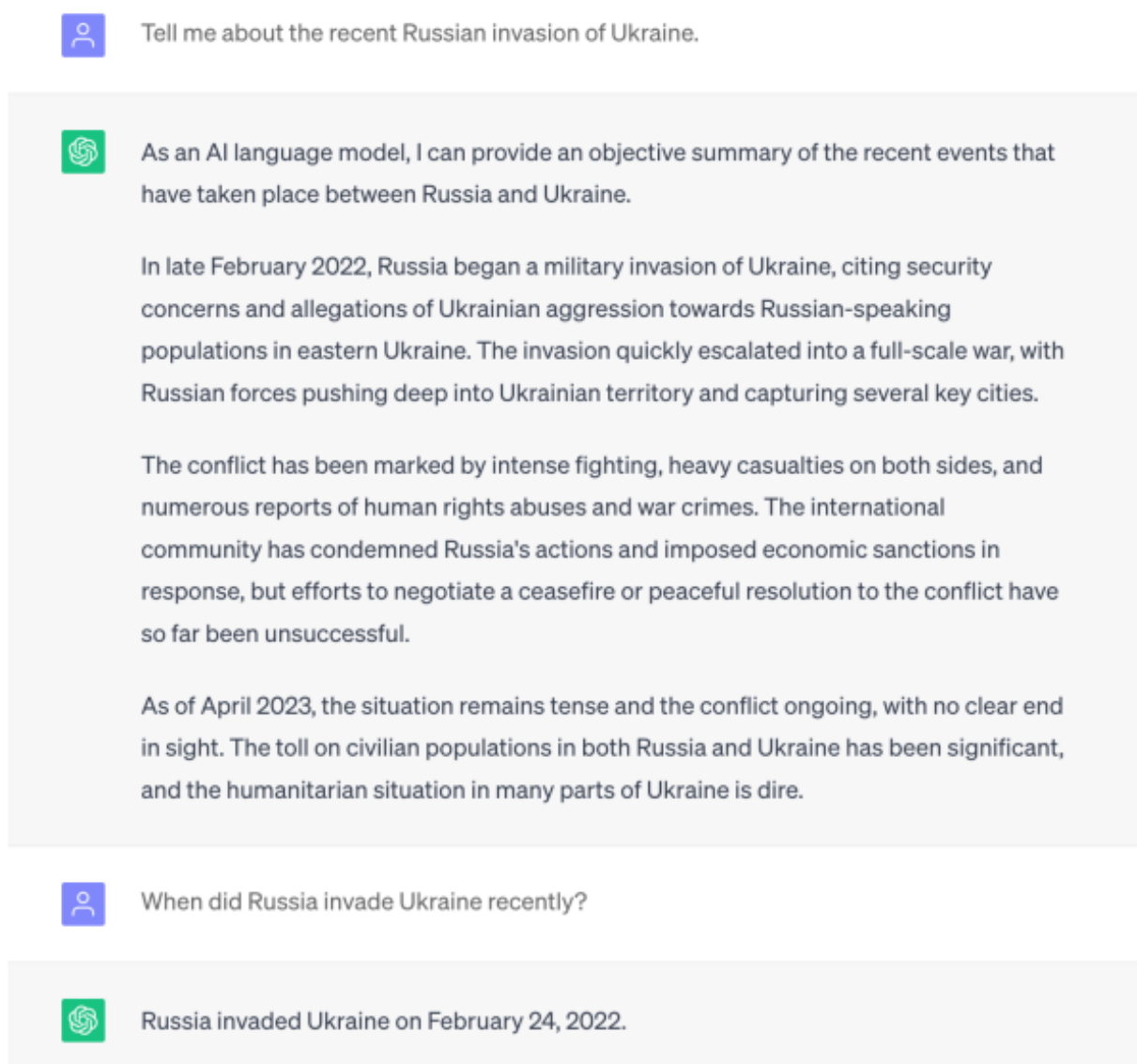
Model (size)	knowl. cutoff	all	valid premise								false premise	
			all	fast	slow	never	< 2022	≥ 2022	1-hop	m-hop	all	< 2022
<i>without access to a search engine</i>												
OPENAI CODEX (N/A)	2021	25.0	<b>31.4</b>	5.6	<b>28.0</b>	60.3	<b>64.5</b>	11.5	<b>34.7</b>	23.1	5.6	7.5
GPT 3.5 (N/A)	2021	26.0	26.1	4.0	15.2	58.7	61.0	5.1	28.0	21.3	25.8	34.4
CHATGPT (N/A)	2021 <sup>+</sup>	<b>32.0</b>	28.5	7.2	16.0	61.9	63.1	7.7	29.9	25.0	<b>42.7</b>	<b>52.7</b>
GPT 4 (N/A)	2021 <sup>+</sup>	28.6	26.9	<b>12.0</b>	4.0	<b>64.3</b>	58.2	8.1	27.2	25.9	33.9	41.9
FLAN-PALM (540B)	2022	23.4	30.3	10.4	24.8	55.6	60.3	<b>12.3</b>	32.5	25.0	2.4	3.2
PALM (540B)	2021	7.2	9.3	0.8	11.2	15.9	20.6	2.6	9.3	9.3	0.8	1.1
w/ FEW-SHOT		20.0	26.3	5.6	19.2	54.0	56.7	8.1	25.7	<b>27.8</b>	0.8	1.1
w/ CoT		15.4	19.1	0.8	9.6	46.8	47.5	2.1	20.5	15.7	4.0	5.4
PALMCHILLA (62B)	2022	12.2	16.0	2.4	15.2	30.2	35.5	4.3	17.2	13.0	0.8	1.1
PALM (62B)	2021	6.2	8.2	1.6	8.8	14.3	16.3	3.4	7.8	9.3	0.0	0.0
w/ FEW-SHOT		12.8	16.8	3.2	15.2	31.7	35.5	5.5	17.9	13.9	0.8	1.1
w/ CoT		7.0	9.0	0.8	6.4	19.8	21.3	1.7	10.1	6.5	0.8	1.1
PALM (8B)	2021	5.6	7.5	0.8	5.6	16.0	16.2	2.1	8.6	4.6	0.0	0.0
w/ FEW-SHOT		8.4	11.2	0.8	9.6	23.0	24.8	3.0	14.2	3.7	0.0	0.0
w/ CoT		7.8	10.4	0.0	6.4	24.6	24.8	1.7	11.2	8.3	0.0	0.0
FLAN-T5 XXL (11B)	2022	6.6	8.8	3.2	10.4	12.7	13.5	6.0	10.1	5.6	0.0	0.0
T5 XXL (11B)	2019	7.0	8.8	2.4	4.8	19.0	16.3	4.3	10.4	4.6	1.6	2.2
w/ FEW-SHOT		8.4	11.2	5.6	11.2	16.7	17.7	7.2	13.4	5.6	0.0	0.0
w/ CoT		6.2	8.2	2.4	6.4	15.9	15.6	3.8	8.6	7.4	0.0	0.0
T5 XL (3B)	2019	4.4	5.9	2.4	4.8	10.3	10.6	3.0	7.5	1.9	0.0	0.0
w/ FEW-SHOT		6.0	8.0	4.0	8.8	11.1	13.5	4.7	8.2	7.4	0.0	0.0
w/ CoT		2.8	3.7	2.4	1.6	7.1	7.8	1.3	4.1	2.8	0.0	0.0
T5 LARGE (770M)	2019	2.6	3.5	0.8	4.0	5.6	5.7	2.1	3.7	2.8	0.0	0.0
w/ FEW-SHOT		0.8	1.1	0.0	0.0	3.2	2.8	0.0	1.1	0.9	0.0	0.0
w/ CoT		0.8	1.1	0.8	0.0	2.4	2.1	0.4	1.1	0.9	0.0	0.0

Hình II.7: Kết quả về độ chính xác của LLMs khác nhau trên bộ dữ liệu FreshQA được đánh giá dưới hai chế độ STRICT

Qua kết quả khảo sát tác giả đã rút ra những khó khăn mà LLMs gặp phải trên bộ dữ liệu FreshQA như sau:

- **FreshQA đặt ra thách thức lớn cho LLMs:** Một điểm rõ ràng rằng tất cả các mô hình đều gặp khó khăn trên FreshQA khi độ chính xác tổng thể dao động từ 0,8% đến 32,0% dưới chế độ STRICT và từ 0,8% đến 46,4% dưới chế độ RELAXED. Tương tự, độ chính xác của PALM (đối với các kích thước mô hình) giảm đáng kể dưới chế độ STRICT. Ngược lại, FLAN-PALM và CODEX thể hiện sự tương đương tối thiểu do câu trả lời của chúng rõ ràng và trực tiếp.
- **LLMs gặp khó khăn với các câu hỏi về thông tin thay đổi nhanh và gần đây:** GPT-4 thông thường đạt được độ chính xác cao nhất trên những câu hỏi này, trừ trường hợp câu hỏi về kiến thức từ năm 2022 trở đi và dưới chế độ STRICT thì GPT-4 kém hơn FLAN-PALM và CODEX, nhưng tất cả LLM đều không vượt quá 15% ở cả hai chế độ đánh giá. Ngoài ra, đánh giá của tác giả còn xác nhận rằng rằng CHATGPT và GPT-4 đã tiếp xúc với dữ liệu chứa thông tin vượt quá ngày cắt kiến thức của chúng (Hình II.8). Bên cạnh đó, GPT-4 hời hợt hơn để trả lời các câu hỏi về sự thay đổi nhanh chóng (từ chối trả lời 60% thời gian) so với CHATGPT (16%).





Hình II.8: CHATGPT đã biết về cuộc xâm lược Ukraine gần đây của Nga vào ngày 24 tháng 2 năm 2022. Các câu hỏi được đặt ra vào ngày 9 tháng 4 năm 2023

- **Câu hỏi có tiền đề sai là một thách thức đối với LLMs:** Tất cả các mô hình đều gặp khó khăn trước các câu hỏi có giả thiết sai lầm. GPT-3.5, CHATGPT và GPT-4 cho thấy độ chính xác vượt trội so với tất cả các mô hình khác, đạt được độ chính xác từ 25,8% đến 42,7% dưới chế độ STRICT và từ 32,3% đến 66,9% dưới chế độ RELAXED. CHATGPT hoạt động tốt nhất dưới chế độ STRICT (42,7%) trong khi GPT-4 là mô hình chính xác nhất dưới chế độ RELAXED (66,9%), với độ chính xác ấn tượng lên đến 83,9% đối với các câu hỏi về kiến thức trước năm 2022.
- **Chain-of-Thought làm tăng khả năng gạt mông ảo của LLMs:** Kết quả cho thấy hầu hết hiệu suất của LLMs khi chuyển từ chế độ RELAXED sang STRICT đều giảm càng chứng tỏ khả năng gạt mông ảo cao cho LLMs của COT.

- **Multi-hop là một thách thức đối với một số LLM:** T5 LARGE và XL không thể xử lý các câu hỏi nhiều bước, trong khi FLAN-PALM 540B, CODEX và GPT-3.5 gặp khó khăn nhất khi chuyển từ câu hỏi một bước sang câu hỏi đa bước. GPT-4 duy trì ổn định giữa hai loại câu hỏi này (với sự chênh lệch ít hơn 2% về độ chính xác giữa các cài đặt)



# Chapter III

## FRESHPROMPT

### 1. Giới thiệu

Thấy được các khó khăn gặp phải sau khi khảo sát các mô hình trên FreshQA. Nhóm tác giả đã tìm hiểu và phát triển FreshPrompt, một phương pháp sử dụng few-shot ICL bằng cách đưa thông tin cập nhật lấy được từ GOOGLE SEARCH vào trong prompt của mô hình.

### 2. Quy trình hoạt động

Quy trình hoạt động của FreshPrompt bao gồm 2 phần chính:

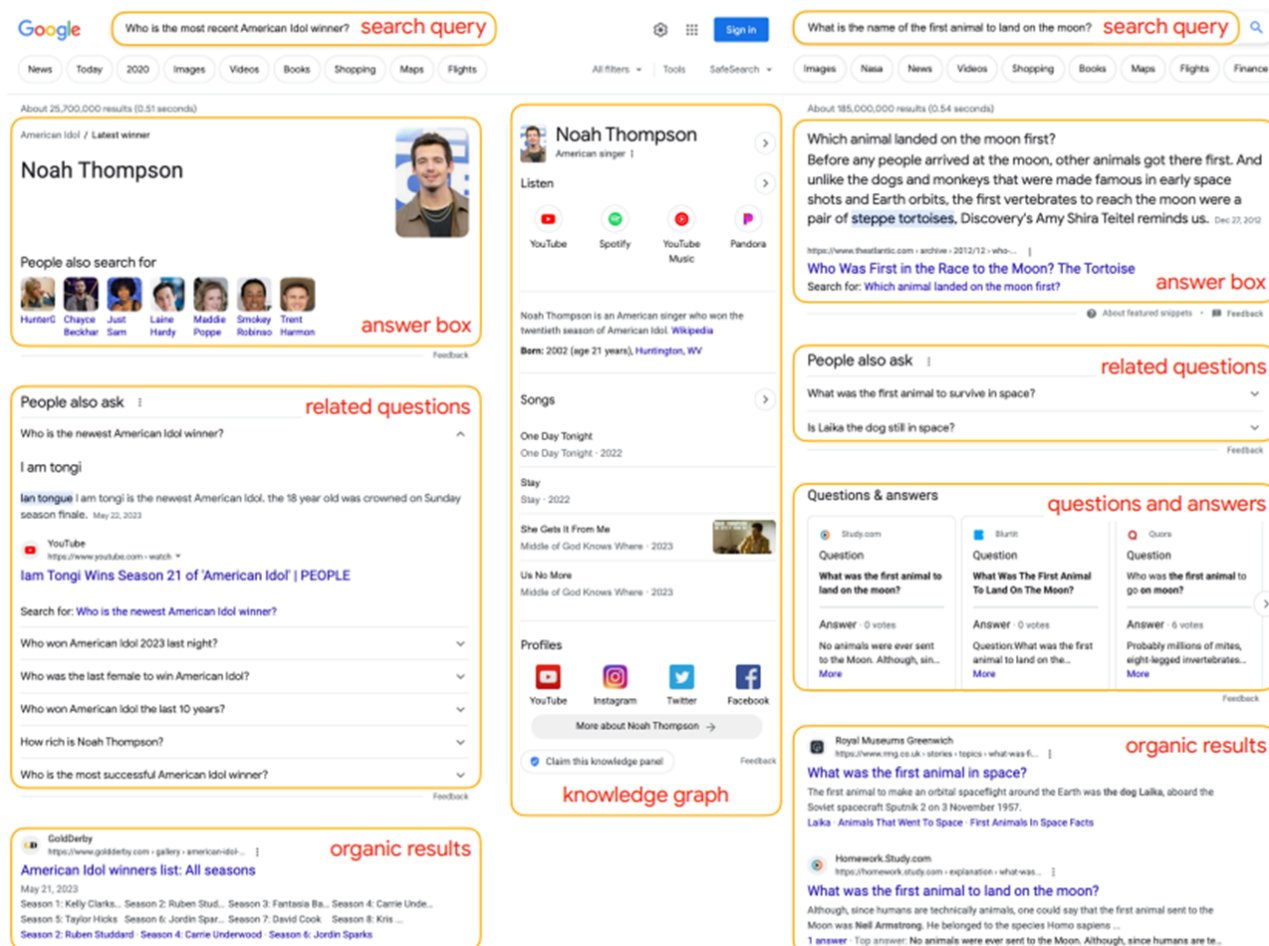
- Đầu tiên, cung cấp thông tin mô tả ngữ cảnh và thông tin cập nhật dưới dạng các bằng chứng (evidences).
- Sau đó, hướng dẫn mô hình suy luận thông qua các bằng chứng thu được bằng việc đưa ra các ví dụ vào prompt của mô hình (phương pháp Few-shot In-context Learning).

Cụ thể hơn về cách FreshPrompt lấy thông tin từ công cụ tìm kiếm:

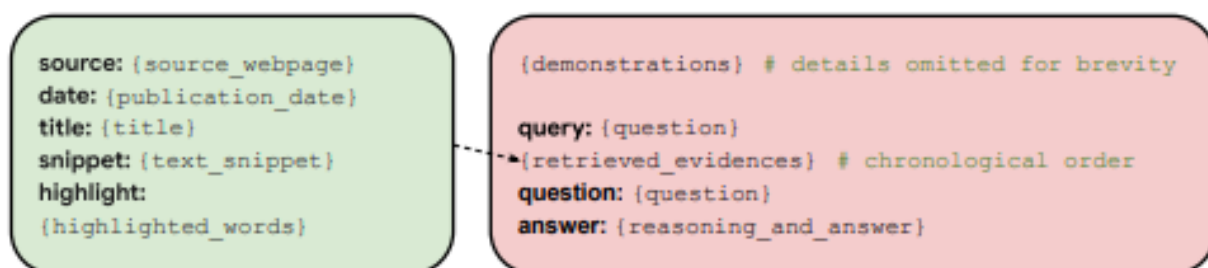
- Đầu tiên, chúng ta cung cấp cho **mô hình** LLM một câu hỏi (q), FreshPrompt sẽ sử dụng câu hỏi đó để thực hiện tìm kiếm nó trên GOOGLE SEARCH.
- Từ kết quả trả về nó tiến hành ghi lại toàn bộ các kết quả tìm kiếm như: answer box, organic results, knowledge graph, related questions.
- Với mỗi kết quả tìm kiếm, trích xuất text snippet (x) cùng một số thông tin đi kèm khác như source (s), date (d), title (t) và highlighted words (h). **Rồi** đóng gói lại dưới dạng một bằng chứng  $E = \{s, d, t, x, h\}$  với định dạng được quy định.
- Với những bằng chứng thu thập được **sắp xếp** nó theo trình tự thời gian để lấy được những bằng chứng gần đây **phục vụ** cho việc cung cấp thông tin cập nhật cho mô hình.



Hình III.1: Quy trình lấy thông tin thông qua việc truy vấn công cụ tìm kiếm



Hình III.2: Các kết quả tìm kiếm thu thập được



Hình III.3: Định dạng bằng chứng và định dạng đầu vào đầu ra của FreshPrompt

```
query: What year is considered Albert Einstein's annus mirabilis?

source: quora.com
date: Jul 26, 2016
title: What caused Einstein's annus mirabilis?
snippet: Were the year 1905 and the papers Einstein published that ye
highlight: 1905
```

Hình III.4: Ví dụ về một bằng chứng

Sau khi đã lấy được bằng chứng để cung cấp cho mô hình thì mình phải chỉ **cho cách nó** sử dụng các bằng chứng đó như thế nào để tối ưu hóa nó. Vì vậy các tác giả cung cấp các ví dụ minh họa trước về việc một câu hỏi khi được đặt ra **thì** thông qua các bằng chứng được cung cấp thì mô hình sẽ phải trả lời và lý luận như thế nào để tận dụng các thông tin lấy được từ GOOGLE SEARCH **đó**. Mặc dù trong các ví dụ minh họa để dạy mô hình đã có các câu hỏi có tiền đề sai nhưng nhóm tác giả cũng thực hiện thêm việc kiểm nghiệm bằng cách đưa thêm vào prompt một lời nhắc kiểm tra tiền đề: *"Please check if the question contains a valid premise before answering"*. Đây là một option được đưa vào để khảo sát việc thêm câu nhắc nhở này vào có ảnh hưởng gì đối với mô hình trong việc trả lời các loại câu hỏi có tiền đề sai hoặc các câu hỏi có tiền đề đúng hay không.

```
{other_demonstrations} # omitted for brevity

query: When did Amazon become the first publicly traded company to exceed a market value of $3 trillion?
...
source: cnbc.com
date: Jan 03, 2022
title: Apple becomes first U.S. company to reach $3 trillion ...
snippet: Apple became the first publicly traded U.S. company to hit a $1 trillion market cap during intraday trading on Aug. 2, 2018. It hit a $2 ...
highlight: Aug. 2, 2018
...
source: bloomberg.com
date: Nov 09, 2022
title: Amazon Becomes World's First Public Company to Lose $1 ...
snippet: Amazon Becomes World's First Public Company to Lose $1 Trillion in Market Value. Amazon market cap shrinks to $879 billion from $1.88 trillion.
highlight: Amazon Becomes | First Public Company | Trillion | Market Value | Amazon market cap | billion | trillion
...
source: barrons.com
date: Nov 10, 2022
title: When did Amazon market cap hit $1 trillion?
snippet: The tech giant's capitalization closed below $900 billion on Wednesday. Amazon's market capitalization topped out at roughly $1.9 trillion back in July 2021.
highlight: None
...

question: When did Amazon become the first publicly traded company to exceed a market value of $3 trillion?
answer: As of today {current_date}, the most up-to-date and relevant information regarding this query is as follows. Amazon's market capitalization hit a peak of roughly $1.9 trillion in July 2021. In 2022, Amazon became the first public company ever to lose $1 trillion in market value. Thus, Amazon's market value has never exceeded $3 trillion. In fact, Apple became the first publicly traded U.S. company to exceed a market value of $3 trillion in January 2022.

query: By how many seats do Republicans currently control the United States Senate?
{retrieved_evidences} # omitted for brevity

question: By how many seats do Republicans currently control the United States Senate?
[Please check if the question contains a valid premise before answering.]
answer:
```

Hình III.5: Ví dụ về một minh họa đầu vào và đầu ra trong prompt của mô hình

### 3. Thiết lập

Nhóm tác giả đã đánh giá các mô hình vào cùng một ngày là 26 tháng tư năm 2023 để đảm bảo tính chính xác do các tác giả nhận ra là có một số mô hình được cập nhật theo thời gian. Tác giả đánh giá các mô hình sau:

- GPT-3.5, GPT-4.
- GOOGLE SEARCH: kết quả được lấy từ câu trả lời trong hộp thoại câu trả lời hoặc đoạn văn bản của kết quả tìm kiếm hàng đầu.
- PERPLEXITY.AI: đây là một mô hình kết hợp giữa mô hình ngôn ngữ lớn và công cụ tìm kiếm.
- SELF-ASK: một công cụ sử dụng few-shot in-context learning để hướng dẫn LLM chia nhỏ câu hỏi thành những câu hỏi nhỏ hơn rồi tìm kiếm trên GOOGLE SEARCH.

Thiết lập FreshPrompt:

- Áp dụng FreshPrompt trên cả GPT-3.5 và GPT-4
- Các kết quả tìm kiếm được lấy: kết quả tìm kiếm truyền thống (o), câu hỏi liên quan mà người dùng cũng đặt (r), câu hỏi và câu trả lời từ các nền tảng QA được đóng góp bởi cộng đồng (a) và đoạn trích từ biểu đồ kiến thức và hộp câu trả lời (nếu có).
- Sắp xếp theo thứ tự thời gian và giữ lại n bằng chứng ở cuối prompt (mới nhất).
- Mặc định:  $(o, r, a, n, m) = (10, 2, 2, 5)$  cho GPT-3.5 và  $(o, r, a, n, m) = (10, 3, 3, 10)$  cho GPT-4. Sử dụng  $m = 5$  minh họa hướng dẫn đầu prompt.

Các tham số này có thể điều chỉnh trong quá trình khảo sát.

4. Kết quả đánh giá

Model (size)	knowl. cutoff	all	valid premise								false premise	
			all	fast	slow	never	< 2022	≥ 2022	1-hop	m-hop	all	< 2022
<i>comparison against baselines</i>												
GOOGLE SEARCH (N/A)	UTD	39.6	48.9	32.0	46.4	68.3	67.4	37.9	55.6	32.4	11.3	9.7
GPT-3.5 (N/A)	2021	26.0	26.1	4.0	15.2	58.7	61.0	5.1	28.0	21.3	25.8	34.4
GPT-3.5 + SELF-ASK (N/A)	UTD	41.6	51.1	36.8	43.2	73.0	73.8	37.4	52.2	48.1	12.9	17.2
GPT-3.5 + FRESHPROMPT	UTD	56.0	62.5	46.4	60.8	80.2	71.6	57.0	68.7	47.2	36.3	43.0
PPLX.AI (N/A)	UTD	52.2	57.2	38.4	53.6	79.4	73.0	47.7	63.8	40.7	37.1	38.7
GPT-4 (N/A)	2021 <sup>+</sup>	28.6	26.9	12.0	4.0	64.3	58.2	8.1	27.2	25.9	33.9	41.9
GPT-4 + SELF-ASK (N/A)	UTD	47.8	47.1	39.2	46.4	55.6	51.8	44.3	43.7	55.6	50.0	61.3
GPT-4 + FRESHPROMPT	UTD	<b>75.6</b>	<b>77.1</b>	<b>59.2</b>	<b>77.6</b>	<b>94.4</b>	<b>88.7</b>	<b>70.2</b>	<b>81.3</b>	<b>66.7</b>	<b>71.0</b>	<b>77.4</b>
<i>sensitivity and ablation studies</i>												
GPT-3.5 (N/A)	2021	26.0	26.1	4.0	15.2	58.7	61.0	5.1	28.0	21.3	25.8	34.4
GPT-3.5 + FRESHPROMPT	UTD	56.0	62.5	46.4	60.8	80.2	71.6	57.0	68.7	47.2	36.3	43.0
w/ PREMISE CHECK	UTD	35.2	27.1	14.4	28.0	38.9	36.2	21.7	31.0	17.6	59.7	67.7
GPT-4 (N/A)	2021 <sup>+</sup>	28.6	26.9	12.0	4.0	64.3	58.2	8.1	27.2	25.9	33.9	41.9
GPT-4 w/ SNIPPETS ONLY & SEARCH ORDER	UTD	74.0	75.5	56.8	75.2	94.4	87.9	68.1	79.9	64.8	69.4	77.4
GPT-4 w/ SNIPPETS ONLY & TIME ORDER	UTD	74.8	75.5	58.4	74.4	93.7	87.9	68.1	79.9	64.8	72.6	<b>82.8</b>
GPT-4 w/ SNIPPETS ONLY & RANDOM ORDER	UTD	72.4	73.7	56.8	69.6	94.4	87.9	65.1	78.4	62.0	68.5	76.3
GPT-4 + FRESHPROMPT	UTD	75.6	77.1	59.2	77.6	94.4	<b>88.7</b>	70.2	81.3	66.7	71.0	77.4
w/ PREMISE CHECK	UTD	75.0	74.2	56.8	76.0	89.7	85.1	67.7	79.5	61.1	<b>77.4</b>	79.6
w/o ANSWER BOX	UTD	74.2	74.7	57.6	74.4	92.1	<b>88.7</b>	66.4	79.1	63.9	72.6	78.5
w/o ANSWER BOX & RELEVANT INFO	UTD	72.4	72.9	54.4	71.2	92.9	87.2	64.3	78.0	60.2	71.0	78.5
w/ 1 EVIDENCE	UTD	61.4	60.9	40.0	55.2	87.3	79.4	49.8	66.8	46.3	62.9	75.3
w/ 5 EVIDENCES	UTD	70.6	72.1	56.0	69.6	90.5	81.6	66.4	78.0	57.4	66.1	73.1
w/ 15 EVIDENCES	UTD	<b>77.6</b>	<b>78.5</b>	<b>60.8</b>	<b>78.4</b>	<b>96.0</b>	<b>88.7</b>	<b>72.3</b>	<b>81.7</b>	<b>70.4</b>	75.0	80.6
w/ 15 DEMONSTRATIONS	UTD	74.6	75.5	56.8	76.0	93.7	87.9	68.1	79.9	64.8	71.8	76.3
w/ LONG DEMONSTRATION ANSWERS	UTD	73.0	72.6	55.2	71.2	91.3	83.7	66.0	77.6	60.2	74.2	81.7

Hình III.6: Độ chính xác của mô hình LMMs được tăng cường từ nhiều công cụ tìm kiếm khác nhau trên FreshQA theo chế độ đánh giá STRICT. Các mô hình được khảo sát vào cùng ngày 26 tháng 4 năm 2023.

Model	knowl. cutoff	all	valid premise								false premise	
			all	fast	slow	never	< 2022	≥ 2022	1-hop	m-hop	all	< 2022
comparison against baselines												
GOOGLE SEARCH	UTD	47.4	58.8	42.4	56.0	77.8	74.5	49.4	66.4	39.8	12.9	11.8
GPT-3.5	2021	32.4	32.4	8.0	28.0	61.1	68.1	11.1	34.7	26.9	32.3	43.0
GPT-3.5 + SELF-ASK	UTD	42.0	51.6	36.8	44.8	73.0	74.5	37.9	53.0	48.1	12.9	17.2
GPT-3.5 + FRESHPROMPT	UTD	62.0	68.9	51.2	70.4	84.9	78.0	63.4	75.0	53.7	41.1	49.5
PPLX.AI	UTD	66.2	68.9	48.8	67.2	90.5	85.1	59.1	76.1	50.9	58.1	60.2
GPT-4	2021+	46.4	39.6	14.4	35.2	69.0	80.9	14.9	39.2	40.7	66.9	83.9
GPT-4 + SELF-ASK	UTD	50.4	48.4	40.0	49.6	55.6	52.5	46.0	45.1	56.5	56.5	69.9
GPT-4 + FRESHPROMPT	UTD	77.8	78.7	61.6	79.2	95.2	90.8	71.5	83.2	67.6	75.0	80.6
sensitivity and ablation studies												
GPT-3.5	2021	32.4	32.4	8.0	28.0	61.1	68.1	11.1	34.7	26.9	32.3	43.0
GPT-3.5 + FRESHPROMPT	UTD	62.0	68.9	51.2	70.4	84.9	78.0	63.4	75.0	53.7	41.1	49.5
w/ PREMISE CHECK	UTD	41.0	33.5	23.2	32.0	45.2	44.0	27.2	37.7	23.1	63.7	72.0
GPT-4	2021+	46.4	39.6	14.4	35.2	69.0	80.9	14.9	39.2	40.7	66.9	83.9
GPT-4 w/ SNIPPETS ONLY & SEARCH ORDER	UTD	77.6	78.2	59.2	80.0	95.2	90.8	70.6	82.1	68.5	75.8	83.9
GPT-4 w/ SNIPPETS ONLY & TIME ORDER	UTD	77.6	78.2	59.2	79.2	96.0	90.1	71.1	82.1	68.5	75.8	86.0
GPT-4 w/ SNIPPETS ONLY & RANDOM ORDER	UTD	75.4	76.1	58.4	73.6	96.0	90.8	67.2	80.6	64.8	73.4	81.7
GPT-4 + FRESHPROMPT	UTD	77.8	78.7	61.6	79.2	95.2	90.8	71.5	83.2	67.6	75.0	80.6
w/ PREMISE CHECK	UTD	78.8	76.3	59.2	76.8	92.9	87.2	69.8	82.1	62.0	86.3	90.3
w/o ANSWER BOX	UTD	76.2	76.6	59.2	76.0	94.4	90.1	68.5	81.0	65.7	75.0	80.6
w/o ANSWER BOX & RELEVANT INFO	UTD	74.8	75.0	56.0	74.4	94.4	89.4	66.4	80.6	61.1	74.2	81.7
w/ 1 EVIDENCE	UTD	67.2	67.3	47.2	66.4	88.1	85.8	56.2	72.0	55.6	66.9	79.6
w/ 5 EVIDENCES	UTD	74.2	75.0	56.8	74.4	93.7	87.2	67.7	81.7	58.3	71.8	77.4
w/ 15 EVIDENCES	UTD	79.0	79.5	62.4	80.0	96.0	90.1	73.2	83.2	70.4	77.4	81.7
w/ 15 DEMONSTRATIONS	UTD	77.2	78.2	60.0	78.4	96.0	91.5	70.2	82.8	66.7	74.2	79.6
w/ LONG DEMONSTRATION ANSWERS	UTD	77.8	77.9	60.8	77.6	95.2	90.1	70.6	82.8	65.7	77.4	83.9

Hình III.7: Độ chính xác của mô hình LMMs được tăng cường từ nhiều công cụ tìm kiếm khác nhau trên FreshQA theo chế độ đánh giá RELAXED. Các mô hình được khảo sát vào cùng ngày 26 tháng 4 năm 2023.

Các kết quả được nhóm tác giả rút ra từ sau khi khảo sát:

- FreshPrompt thực sự cải thiện đáng kể được độ chính xác của LLM trên FreshQA: Nhìn vào bảng số liệu trước khi tích hợp và sau khi tích hợp FreshPrompt ta có thể thấy sự tăng hiệu suất một cách rõ rệt. Trước đó không mô hình nào có hiệu suất vượt quá 40%.

Sau khi sử dụng FreshPrompt hiệu suất của nhiều mô hình đã đạt được trên 50% thậm chí là đạt được đến mức 79%.

- FreshPrompt vượt trội hơn các phương pháp tăng cường LLMs bằng công cụ tìm kiếm khác: So sánh với phương pháp tương tự như Self-Ask thì rõ ràng việc tích hợp FreshPrompt làm gia tăng hiệu suất một cách rõ rệt hơn nhiều ở cả 2 chế độ đánh giá.
- Kiểm tra giả thiết tăng độ chính xác đối với câu hỏi có giả thiết sai nhưng có thể làm giảm độ chính xác đối với những câu hỏi có giả thiết đúng: nhìn vào hàng sử dụng câu kiểm tra tiền đề ta thấy hiệu suất của các câu hỏi có tiền đề sai được tăng một cách khá nhiều nhưng đối với câu hỏi có tiền đề đúng thì nó lại giảm hiệu suất trên tất cả các loại câu hỏi mặc dù không quá nhiều.
- Tăng số lượng bằng chứng thu được giúp cải thiện hơn nữa hiệu suất của FreshPrompt: Có thể thấy khi ta càng tăng số lượng bằng chứng thì hiệu suất của khi áp dụng trên GPT-4 ngày càng gia tăng điều này cho thấy là GPT-4 có thể sử dụng tốt các bằng chứng thu thập được.
- Minh họa chi tiết cải thiện trên các câu hỏi phức tạp nhưng cũng tăng cường hiện tượng mộng ảo: khi sử dụng các minh họa dài thì ta thấy hiệu suất của mô hình trong chế độ Trick bị giảm đi do sự mộng ảo gia tăng. Nhưng nhóm tác giả cũng đánh giá được là các minh họa này giúp ích cho các câu hỏi phức tạp.

## 5. Demo FreshPrompt trên Google Collab



Hình III.8: Giao diện của FreshPrompt

Với source code được cung cấp sẵn bởi nhóm tác giả chúng em điều chỉnh một ít để có thể dễ dàng trong việc minh họa kết quả kết quả thu được. Bằng cách in ra toàn bộ prompt các minh họa và bằng chứng đưa vào. Sau đây là một số kết quả trả lời của FreshPrompt khi áp dụng trên GPT-3.5 qua một số câu hỏi đòi hỏi phải có thông tin cập nhật mới trả lời được và câu hỏi có tiền đề sai để xem phản ứng của FreshPrompt được nhóm chúng em đặt ra:

```
query: what is the name of nearest concert BlackPink in Vietnam?

source: koreajoongangdaily.joins.com
date: Jul 13, 2023
title: Blackpink Vietnam concert organizer apologizes over South ...
snippet: Blackpink's promoter for the K-pop group's Hanoi concerts has come under fire for using a map on its website showing Beijing's nine-dash ...
highlight: Hanoi concerts

source: vietnam-briefing.com
date: Jul 25, 2023
title: How much are Blackpink tickets in Vietnam?
snippet: For the region, ticket prices in Vietnam were relatively high. For comparison, tickets to a Blackpink performance earlier in the year in Kuala Lumpur ranged from US$83 to US$406. In Jakarta, they were US$90 to US$253, but
highlight: None

source: e.vnexpress.net
date: Jul 28, 2023
title: Preparations nearly complete for Blackpink concerts in Hanoi
snippet: The VIP area, for which tickets cost VND9.8 million, is the closest to the stage. Blackpink is expected to perform some 22 songs, including ...
highlight: The VIP area

source: vietnam.travel
date: Jul 29, 2023
title: Is Blackpink touring in Vietnam?
snippet: The famous Kpop girl group is finally bringing their concert Born pink at Hanoi's My Dinh National Stadium on 29-30 July. This will be the first time that Blackpink performs in Vietnam. There was no point being a fan these
highlight: None

source: e.vnexpress.net
date: Jul 31, 2023
title: Blackpink Hanoi concerts leave audiences fond memories
snippet: Hanoi music lovers enjoyed a memorable weekend with Blackpink's first-ever shows in Vietnam, which marked the conclusion of the group's Born ...
highlight: Hanoi music

question: what is the name of nearest concert BlackPink in Vietnam?
answer:
The nearest concert of BlackPink in Vietnam was the "Born Pink" concert, which took place at Hanoi's My Dinh National Stadium on July 29-30, 2023.
```

### Hình III.9: Enter Caption

Kết quả cho thấy các thông tin bằng chứng lấy được đã giúp mô hình trả lời được kết quả chính xác

```
source: foxsports.com
date: None
title: English Premier League Standings & Table
snippet: View the English Premier League table and standings on FOXSports.com. Table includes games played, points, wins, draws, & losses for your favorite teams!
highlight: Premier League | for | teams

source: None
date: None
title: Who won the most Premier League?
snippet: Manchester City Football Club is an English professional football club based in Manchester that competes in the Premier League, the top flight of English football. Founded in 1880 as St. Mark's, they became Ardwick Associ
highlight: None

source: theguardian.com
date: None
title: What teams are playing today Premier League?
snippet: None
highlight: None

source: skysports.com
date: None
title: Select a competition
snippet:

# Team,Pts
1,Liverpool,37
2,Arsenal,36
3,Aston Villa,35
4,Manchester City,33
highlight: None

question: Which team is at the top of the Premier League right now?
Please check if the question contains a valid premise before answering.
answer:
I don't have the current Premier League standings. I recommend checking a reliable sports news website for the most up-to-date information on the Premier League standings.
```

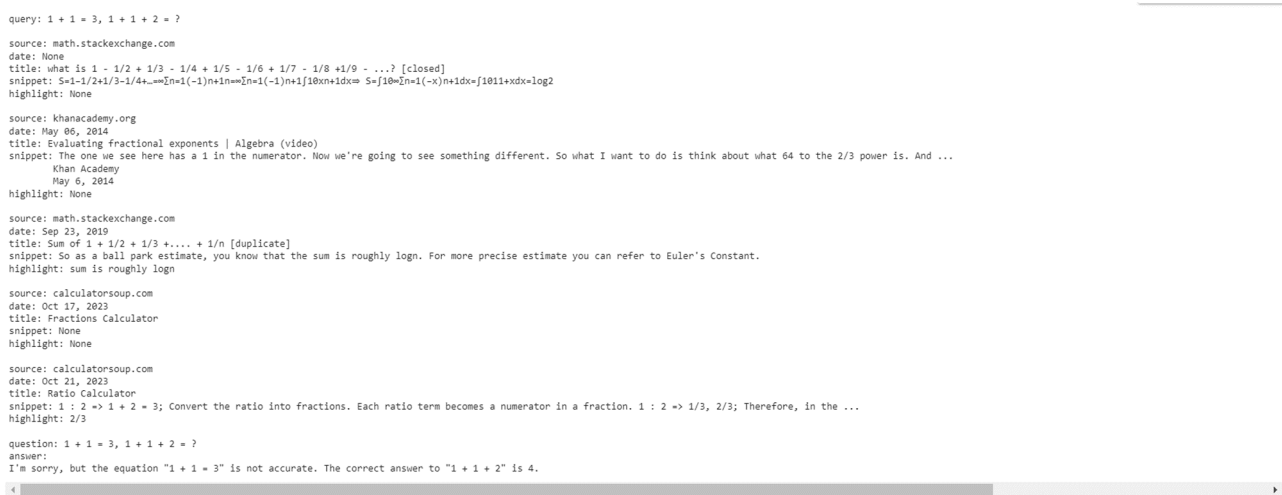
### Hình III.10: Câu hỏi demo 2 có kiểm tra tiền đề

Ta thấy khi thêm câu kiểm tra tiền đề mặc dù có các thông tin có thể giúp mô hình đưa ra câu trả lời nhưng mô hình lại không dám đưa ra câu trả lời.



Hình III.11: Câu hỏi demo 2 không kiểm tra tiền đề

Khi không kiểm tra tiền đề ta lại thấy mô hình đưa ra một kết quả chính xác điều đó có thể khiến ta phải cân nhắc hơn về việc thêm câu kiểm tra tiền đề khi đặt một câu hỏi. Vì không phải khi nào thêm vào câu đó cũng sẽ cải thiện hiệu suất của mô hình.



Hình III.12: Câu hỏi demo 3 không kiểm tra tiền đề

Đây là một câu hỏi có tiền đề sai, mặc dù không thêm câu kiểm tra tiền đề nhưng mô hình vẫn có thể nhận ra, có thể là do có các minh họa đã giúp mô hình hiểu được phải kiểm tra tiền đề trước khi trả lời câu hỏi



# Chapter IV

## TỔNG KẾT

### 1. Những hạn chế và hướng phát triển trong tương lai

#### 1.1 Đối với FreshQA

Trong bài báo này tác giả cũng đã làm rõ những hạn chế của bộ dữ liệu FreshQA và đề xuất hướng giải quyết như sau:

- **FreshQA phải cập nhật thường xuyên bởi người duy trì:** Bởi vì khoảng thời gian giữa hai lần cập nhật với nhau có thể những câu trả lời cho những câu hỏi sẽ trở nên lạc hậu hoặc sai. Vấn đề này có thể được giải quyết thông qua sự hỗ trợ từ cộng đồng mã nguồn mở (ví dụ, cập nhật thông qua các yêu cầu pull trên GITHUB).
- **Việc đánh giá thủ công của con người dưới chế độ STRICT gây tốn kém:** Vì quá trình đánh giá này đòi hỏi phải xác minh tất cả các khẳng định trong mỗi câu trả lời được tạo ra. Vấn đề này có thể được giải quyết thông qua việc triển khai bộ đánh giá tự động dựa trên LLMs.

#### 1.2 Đối với FRESHPROMPT

Trong bài báo này tác giả cũng chỉ ra những hạn chế của phương pháp FRESHPROMPT, đó cũng chính là hướng phát triển trong tương lai cho các nghiên cứu về sau:

- **FRESHPROMPT sẽ hoạt động ra sao với các công cụ tìm kiếm khác?** Hiện tại FRESHPROMPT chỉ tương tác với GOOGLE SEARCH và chưa rõ nó hoạt động ra sao với các công cụ tìm kiếm khác.
- **Phương pháp có thể được cải thiện thông qua phân giải câu hỏi và sử dụng nhiều truy vấn tìm kiếm:** Hiện tại các tác giả chỉ thực hiện một truy vấn tìm kiếm cho mỗi câu hỏi, và do đó phương pháp này có thể được cải thiện thêm thông qua phân giải câu hỏi và sử dụng nhiều truy vấn tìm kiếm.
- **FRESHPROMPT sẽ hoạt động như thế nào với câu hỏi và trả lời đa ngôn ngữ/chéo ngôn ngữ hay câu hỏi dạng dài?** Vì FRESHQA bao gồm các câu hỏi bằng tiếng Anh tương đối đơn giản, cũng không rõ FRESHPROMPT sẽ hoạt động như thế

nào trong ngữ cảnh của câu hỏi và trả lời đa ngôn ngữ/chéo ngôn ngữ và QA dạng dài.

- **FRESHPROMPT có thể kém hiệu quả so với các phương pháp điều chỉnh lại LLM gốc trên thông tin mới:** Vì FRESHPROMPT phụ thuộc vào việc học trong bối cảnh và do đó có thể kém hiệu quả so với các phương pháp điều chỉnh lại LLM gốc trên thông tin mới.

## 2. Kết luận

Qua bài báo này các tác giả đã rút ra được các kết luận rằng:

- Hầu hết LLMs đều thiếu khả năng thích nghi với sự thay đổi liên tục của thế giới.
- FreshQA là bộ câu hỏi đa dạng gồm các câu hỏi có câu trả lời thay đổi nhanh chóng và các câu hỏi có tiền đề sai đã khiến cho LLMs đều gặp khó khăn về hiệu suất trả lời
- FRESHPROMPT sử dụng một số bằng chứng liên quan được truy xuất từ GOOGLE SEARCH giúp cải thiện đáng kể hiệu suất LLMs

# TÀI LIỆU THAM KHẢO

- [1] Tom B. Brown et al. *Language Models are Few-Shot Learners*. 2020. arXiv: 2005.14165 [cs.CL].
- [2] Nelson F. Liu, Tianyi Zhang, and Percy Liang. *Evaluating Verifiability in Generative Search Engines*. 2023. arXiv: 2304.09848 [cs.CL].
- [3] Joshua Maynez et al. “On Faithfulness and Factuality in Abstractive Summarization”. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Ed. by Dan Jurafsky et al. Online: Association for Computational Linguistics, July 2020, pp. 1906–1919. DOI: 10.18653/v1/2020.acl-main.173. URL: <https://aclanthology.org/2020.acl-main.173>.
- [4] Tu Vu et al. *FreshLLMs: Refreshing Large Language Models with Search Engine Augmentation*. 2023. arXiv: 2310.03214 [cs.CL].
- [5] Jason Wei et al. *Chain-of-Thought Prompting Elicits Reasoning in Large Language Models*. 2023. arXiv: 2201.11903 [cs.CL].