



# Day4

## Today's Assignment



Complete Unit 4: "Domain-Specific LLMs", which is:

- [Optional] Listen to the summary podcast episode for this unit (created by NotebookLM).
- Read the "Solving Domain-Specific Problems Using LLMs" whitepaper.
- Complete these code labs on Kaggle:
  1. Use Google Search data in generation
  2. Tune a Gemini model for a custom task

## What You'll Learn



In today's reading, you'll delve into the creation and application of specialized LLMs like SecLM and MedLM/Med-PaLM, with insights from the researchers who built them.

In the code labs you will learn how to add real world data to a model beyond its knowledge cut-off by grounding with Google Search. You will also learn how to fine-tune a custom Gemini model using your own labeled data to solve custom tasks.

## Reminders and Announcements



- Here is the [recording from Day 3's livestream](#).
- The next livestream with Paige Bailey is tomorrow at 2pm PST/ 5pm EST/ 10pm UTC. [Click here to join!](#) Livestream guests: Scott Coull, Antonio Gulli, Anant Nawalgaria, Christopher Semturs, and Umesh Shankar.
- Find a [complete list of scheduled livestreams and past recordings here](#).
- Be sure to ask all your questions about the podcast, readings, and code lab on [Discord](#).

Happy learning and see you tomorrow,

The Kaggle Team

---

## Table of contents

[Table of contents](#)

[Introduction \(Giới thiệu\)](#)

[SecLM and the future of cybersecurity \(SecLM và tương lai của an ninh mạng\)](#)

[Challenges in cybersecurity \(Những thách thức trong an ninh mạng\)](#)

[How GenAI can tackle the challenges in cybersecurity \(GenAI có thể giải quyết những thách thức trong an ninh mạng như thế nào\)](#)

[SecLM: An API for cybersecurity tasks \(\*\*Một API cho các nhiệm vụ an ninh mạng\*\*\)](#)

[\*\*SecLM: Một API cho các nhiệm vụ an ninh mạng\*\*](#)

[\*\*Mục tiêu của SecLM\*\*](#)

[1. Cập nhật dữ liệu mới nhất \(Freshness\):](#)

[2. Dữ liệu riêng của người dùng \(User-specific data\):](#)

[3. Hiểu biết chuyên môn bảo mật \(Security expertise\):](#)

[4. Lý luận theo từng bước \(Multi-step reasoning\):](#)

[\*\*Security-focused large language models \(Mô hình ngôn ngữ lớn \(LLMs\) chuyên biệt cho bảo mật\)\*\*](#)

[A flexible planning and reasoning framework \(Một khuôn khổ lập kế hoạch và lý luận linh hoạt\)](#)

MedLM and the future of health tech (MedLM và tương lai của công nghệ y tế)

The potential for GenAI in medical Q&A (Tiềm năng của GenAI trong Hỏi & Đáp y khoa)

The opportunities (Cơ hội)

The scientific starting point (Điểm khởi đầu khoa học)

How to evaluate: quantitative and qualitative (Cách đánh giá: định lượng và định tính)

Evaluation in real clinical environments (Đánh giá trong môi trường lâm sàng thực tế)

Task- vs. domain-specific models (Mô hình theo nhiệm vụ so với mô hình theo miền cụ thể)

Training strategies for Med-PaLM 2 (Chiến lược đào tạo cho Med-PaLM 2)

Summary

---

## Introduction (Giới thiệu)

Các mô hình ngôn ngữ lớn (LLMs) đã trở thành công cụ mạnh mẽ để giải quyết những thách thức phức tạp trong nhiều lĩnh vực. Ban đầu, các phiên bản đầu tiên tập trung vào các nhiệm vụ chung, nhưng những phát triển gần đây đã nhấn mạnh tiềm năng của việc tinh chỉnh LLMs để giải quyết các vấn đề cụ thể trong các lĩnh vực chuyên môn. Tài liệu này khám phá các khái niệm này qua hai lĩnh vực đặc thù: an ninh mạng và y học. Mỗi lĩnh vực đều thể hiện khả năng độc đáo của LLMs trong việc cải thiện quy trình hiện có và mở ra những tiềm năng mới.

**An ninh mạng** đặt ra nhiều thách thức đặc thù cho LLMs, bao gồm: thiếu dữ liệu công khai, sự đa dạng của các khái niệm kỹ thuật cao cấp và thông tin về các mối đe dọa luôn thay đổi từng ngày. Ngoài ra, các trường hợp sử dụng nhạy cảm như phân tích mã độc đòi hỏi những cân nhắc đặc biệt khi phát triển mô hình. Chúng tôi giải quyết những thách thức này bằng cách tập trung vào nội dung và nhiệm vụ chuyên biệt trong an ninh mạng, kết hợp các mô hình ngôn ngữ chuyên về an ninh với một bộ kỹ thuật hỗ trợ để cải thiện hiệu suất trong các nhiệm vụ quan trọng như nhận diện mối đe dọa và phân tích rủi ro.

Trong **lĩnh vực y học**, LLMs đối mặt với một tập hợp thử thách khác, chẳng hạn như lượng kiến thức y khoa khổng lồ và luôn thay đổi, cùng với nhu cầu áp dụng kiến thức đó một cách phù hợp theo ngữ cảnh. Điều này khiến việc chẩn đoán và điều trị chính xác trở thành một thách thức liên tục.

Các LLM như Med-PaLM, được tùy chỉnh cho các ứng dụng y khoa, đã cho thấy khả năng trả lời các câu hỏi y tế phức tạp và cung cấp các phân tích sâu sắc về dữ liệu y khoa. Điều này thể hiện tiềm năng hỗ trợ cả bác sĩ và bệnh nhân.

Qua lăng kính của hai lĩnh vực đặc biệt này, tài liệu này sẽ khám phá những thách thức và cơ hội do dữ liệu chuyên biệt, ngôn ngữ kỹ thuật và các trường hợp sử dụng nhạy cảm mang lại. Bằng cách xem xét những hướng đi riêng biệt của SecLM và Med-PaLM, chúng tôi cung cấp những hiểu biết về tiềm năng cách mạng hóa các lĩnh vực chuyên môn của LLMs.

## **SecLM and the future of cybersecurity (SecLM và tương lai của an ninh mạng)**

Các chuyên gia an ninh mạng đối mặt với vô số thách thức, bao gồm các mối đe dọa mới và liên tục phát triển, khối lượng công việc vận hành nặng nề, và sự thiếu hụt nhân tài. AI tạo sinh (Gen AI) chuyên biệt có thể hỗ trợ giải quyết những thách thức này bằng cách tự động hóa các nhiệm vụ lặp đi lặp lại, giúp tiết kiệm thời gian cho các hoạt động chiến lược hơn, đồng thời mang lại cơ hội mới để tiếp cận tri thức.

### **Challenges in cybersecurity (Những thách thức trong an ninh mạng)**

Trong các bộ phim, an ninh thông tin thường bị giản lược thành hình ảnh những hacker mặc áo hoodie, đeo tai nghe, với ý đồ xấu, tay gõ bàn phím liên tục trên những chiếc laptop cũ kỹ, cho đến khi thốt lên hai từ thần kỳ: "Tôi đã vào được!"

Nếu có sự xuất hiện của những người bảo vệ, họ thường bị mô tả trong trạng thái bị động: nào là phòng điều hành chật kín người, những chiếc cốc cà phê rỗng, những người ra lệnh liên tục và màn hình hiển thị mọi bước đi của kẻ tấn công theo thời gian thực.

Nhưng đó chỉ là Hollywood; còn thực tế thì khác xa.

Trong thực tế, những người làm việc trong lĩnh vực an ninh mạng - bao gồm các lập trình viên, quản trị viên hệ thống, các kỹ sư SRE và các nhà phân tích cấp thấp mà chúng tôi dành tài liệu này để vinh danh - phải đối mặt với nhiệm vụ "không hồi kết" khi liên tục cập nhật các mối đe dọa mới và bảo vệ những hệ thống phức tạp khỏi chúng. Phần lớn thời gian làm việc của họ bị chi phối bởi các nhiệm vụ lặp đi lặp lại hoặc thủ công, chẳng hạn như phân loại hàng trăm cảnh báo riêng lẻ, khiến họ mất đi cơ hội phát triển các chiến lược phòng thủ hiệu quả hơn.

Đáng lo ngại hơn, lợi thế không thuộc về phe phòng thủ. Kẻ tấn công đang áp dụng những công nghệ tiên tiến, bao gồm cả trí tuệ nhân tạo, để mở rộng phạm vi hoạt động và đẩy nhanh tốc độ khai thác lỗ hổng. Và, tất nhiên, trong thực tế không có những màn hình nào hiển thị mọi bước đi của kẻ tấn công!

Dựa trên kinh nghiệm làm việc với người dùng và đối tác, chúng tôi nhận thấy ba thách thức lớn mà ngành an ninh mạng hiện nay đang phải đối mặt: **mối đe dọa, khối lượng công việc lặp lại, và thiếu hụt nhân tài.**

- **Mối đe dọa mới và liên tục thay đổi:** Bức tranh các mối đe dọa luôn thay đổi, với các cuộc tấn công mới ngày càng tinh vi xuất hiện liên tục. Điều này khiến người phòng thủ khó theo kịp thông tin mới nhất và phải vật lộn để chọn lọc những thông tin quan trọng từ dòng dữ liệu khổng lồ để đưa ra hành động phù hợp.
- **Khối lượng công việc lặp lại:** Những người làm trong vai trò vận hành an ninh hoặc DevOps thường dành phần lớn thời gian để xử lý các nhiệm vụ thủ công lặp đi lặp lại, những nhiệm vụ có thể được tự động hóa hoặc hỗ trợ. Điều này gây ra quá tải và làm giảm thời gian dành cho các hoạt động chiến lược hơn. Việc tập trung quá nhiều vào chi tiết vụn vặt cũng khiến các nhà phân tích và kỹ sư mất đi tầm nhìn tổng thể - yếu tố quan trọng trong việc bảo vệ tổ chức của họ.
- **Thiếu hụt nhân tài:** Sự thiếu hụt các chuyên gia an ninh mạng có kỹ năng khiến các tổ chức khó tìm được người cần thiết để bảo vệ dữ liệu và hệ thống của họ. Nhiều người đảm nhận vai trò liên quan đến an ninh mà không được đào tạo bài bản và ít có thời gian trau dồi kỹ năng trong quá trình làm việc.

Nếu không giải quyết được ba thách thức này, sẽ rất khó để đáp ứng nhu cầu của các hệ thống an ninh mạng hiện đại.

## **How GenAI can tackle the challenges in cybersecurity (GenAI có thể giải quyết những thách thức trong an ninh mạng như thế nào)**

Chúng tôi hình dung một thế giới nơi những người mới bắt đầu và các chuyên gia an ninh mạng đều được hỗ trợ bởi chuyên môn của AI, giúp họ giải phóng khỏi các công việc lặp đi lặp lại, hoàn thành những nhiệm vụ mà hiện tại có vẻ không thể, và mở ra cơ hội mới để chia sẻ tri thức. Các mô hình ngôn ngữ lớn (LLMs) cùng với

các kỹ thuật AI tạo sinh (GenAI) liên quan có thể cải thiện đáng kể trải nghiệm làm việc của cả người mới và các chuyên gia dày dạn kinh nghiệm trong lĩnh vực an ninh mạng. Trên thực tế, trong nhiều trường hợp, chúng tôi đã chứng minh rằng GenAI có thể giải quyết hiệu quả một số vấn đề thực tế trong lĩnh vực an ninh mạng.

Dưới đây là các thách thức điển hình và cách GenAI có thể hỗ trợ:

<b>Persona(e) (Nhân vật)</b>	<b>Challenges faced (Những thách thức phải đối mặt)</b>	<b>How Gen AI can help (Gen AI có thể giúp ích như thế nào)</b>
	Không quen thuộc với cấu trúc dữ liệu đặc thù và ngôn ngữ truy vấn của từng công cụ.	Chuyển đổi các truy vấn ngôn ngữ tự nhiên thành ngôn ngữ truy vấn sự kiện an ninh đặc thù và ngôn ngữ quy tắc tương ứng.
<b>Security Analyst (Nhà phân tích an ninh)</b>	Việc điều tra, phân cụm và phân loại các cảnh báo đến tốn nhiều thời gian, đòi hỏi nhiều bước và công cụ.	Cung cấp khả năng tự động hóa việc điều tra, nhóm và phân loại, kết hợp bối cảnh và sử dụng công cụ theo thời gian thực.
	Khó khăn trong việc tạo ra một chuỗi các bước cụ thể để giải quyết một vấn đề.	Lên kế hoạch xử lý phù hợp với từng trường hợp, cá nhân hóa trong môi trường người dùng.
<b>Threat Researcher/System Administrator (Nhà nghiên cứu mối đe dọa hoặc quản trị viên hệ thống)</b>	Một thành phần không rõ nguồn gốc và được làm mờ (ví dụ: mã script hoặc tệp nhị phân) được phát hiện và khó phân tích thủ công.	Tự động phân tích ngược bằng cách sử dụng LLM để phân tích mã, giải mã và chuyển đổi ngược. Giải thích, phân tích và phân loại các thành phần có khả năng độc hại.
<b>Chief Information Security Officer (Nhóm CISO)</b>	Yêu cầu thực hiện thủ công để xác định và tóm tắt các mối đe dọa lớn	Tạo tài liệu hoặc bài thuyết trình dễ đọc, áp dụng thông tin tình báo mối đe dọa mới nhất và

	nhất mà tổ chức đang đối mặt.	kết quả từ các công cụ an ninh cho tổ chức cụ thể.
<b>IT Administrator Dedicated Security Team (Quản trị viên IT hoặc đội ngũ an ninh chuyên biệt)</b>	Khó hiểu hết các cách mà kẻ tấn công có thể truy cập vào tài nguyên nhạy cảm.	Xác định các đường tấn công tiềm năng hoặc hiện tại, làm nổi bật các yếu tố quan trọng và các biện pháp khắc phục.
<b>Application Developers (Nhà phát triển ứng dụng)</b>	Khó xác định vị trí phù hợp để thực hiện kiểm tra fuzz-test cho ứng dụng.	Xác định vị trí cần thực hiện fuzz-test và tạo mã kiểm tra tương ứng.
<b>Application Developers &amp; IT Administrators (Nhà phát triển ứng dụng và Quản trị viên IT)</b>	Duy trì chính sách truy cập phù hợp với nguyên tắc "ít quyền hạn nhất".	Dựa trên các mẫu truy cập trước đây và cấu hình hiện tại, tạo tệp sửa đổi cấu hình để cấp quyền với phạm vi tối thiểu hơn.
<b>A person responsible for an application or system (Người chịu trách nhiệm về ứng dụng hoặc hệ thống)</b>	Không phải lúc nào cũng hiểu rõ các khái niệm về an ninh hoặc cách áp dụng chúng vào môi trường của mình; cần biết cách phân tích vấn đề, đặt câu hỏi ở nhiều nơi và kết hợp các câu trả lời để tìm ra giải pháp.	Đưa ra câu trả lời dựa trên chuyên môn an ninh đáng tin cậy và thông qua tích hợp, đảm bảo câu trả lời phù hợp với môi trường làm việc của người dùng.

Để giải quyết các vấn đề này một cách toàn diện và hiệu quả, chúng ta cần một phương pháp tiếp cận đa lớp (multi-layered approach):

- **Top layer (Lớp trên cùng):** Các công cụ bảo mật hiện có, có khả năng hiểu ngữ cảnh và dữ liệu liên quan, đồng thời có thể thực hiện các thay đổi cần thiết.
- **Middle layer (Lớp giữa):** Một API mô hình chuyên biệt về bảo mật với khả năng suy luận và lập kế hoạch nâng cao.

- **Bottom layer (Lớp dưới cùng):** Các kho dữ liệu chứa thông tin tình báo bảo mật đáng tin cậy và kiến thức vận hành.

Một trong những lợi ích chính của LLMs là khả năng xử lý và tổng hợp lượng lớn dữ liệu không đồng nhất – một năng lực quan trọng trong thế giới dữ liệu bảo mật ngày càng bị phân mảnh. Chúng tôi mong muốn tận dụng khả năng này để giải quyết các vấn đề bảo mật khó khăn, dù là hỗ trợ các nhà phân tích hay thông qua các tác nhân tự động. Điều này đạt được bằng cách kết hợp ngữ cảnh liên quan và các nguồn thông tin đáng tin cậy với một khung lập kế hoạch linh hoạt trong một API duy nhất, gọi là **SecLM**.

API này cung cấp khả năng lập kế hoạch phong phú bằng cách kết hợp các mô hình LLM với các mô hình học máy (ML) khác, sử dụng phương pháp **RAG (Retrieval-Augmented Generation)** để đảm bảo kết quả được dựa trên dữ liệu đáng tin cậy, và khả năng sử dụng công cụ để thực hiện các hành động hoặc tra cứu thông tin liên quan. Chúng tôi cho rằng phương pháp tiếp cận toàn diện này là rất cần thiết vì độ chính xác đóng vai trò cực kỳ quan trọng trong lĩnh vực an ninh mạng, và bản thân các LLM không thể tự mình giải quyết tất cả các vấn đề bảo mật.

## **SecLM: An API for cybersecurity tasks (Một API cho các nhiệm vụ an ninh mạng)**

**Tầm nhìn của SecLM** là trở thành một nền tảng toàn diện để trả lời các câu hỏi về bảo mật, bất kể mức độ phức tạp của chúng. Các kỹ sư hoặc nhà phân tích có thể đặt câu hỏi và tham chiếu các nguồn dữ liệu bằng ngôn ngữ tự nhiên, và nhận được câu trả lời tích hợp thông tin cần thiết một cách tự động. Tuy nhiên, các vấn đề bảo mật thường đòi hỏi việc thu thập và phân tích lượng lớn thông tin bằng cách suy luận đặc thù cho từng lĩnh vực, thường yêu cầu sự tham gia của các chuyên gia từ nhiều lĩnh vực khác nhau.

SecLM được thiết kế để cung cấp câu trả lời chất lượng cao ngay cả trong các tình huống không cần đặt trước bất kỳ câu hỏi cụ thể nào (*zero-shot*). Điều này đòi hỏi API phải tương tác liền mạch với các mô hình LLM, mô hình ML truyền thống, dữ liệu của người dùng, và các dịch vụ khác để hoàn thành nhiệm vụ chính xác. Để đạt được điều này, SecLM cần giải quyết các yêu cầu quan trọng sau:



- **Freshness (Cập nhật dữ liệu mới nhất):** Mô hình cần truy cập dữ liệu mới nhất về mối đe dọa và lỗ hổng, vốn thay đổi hàng ngày. Do chi phí và thời gian huấn luyện lại rất lớn (thường mất vài ngày), không thể thực hiện việc cập nhật mô hình hàng giờ hoặc hàng ngày.
- **User-specific data (Dữ liệu riêng của người dùng):** Mô hình cần hoạt động trên dữ liệu bảo mật riêng của người dùng trong môi trường của họ mà không làm lộ dữ liệu nhạy cảm ra ngoài. Điều này loại trừ việc huấn luyện tập trung trên dữ liệu người dùng.
- **Security expertise (Hiểu biết chuyên môn bảo mật):** Mô hình cần hiểu các khái niệm và thuật ngữ bảo mật cấp cao, sau đó phân tích chúng thành các thành phần nhỏ hơn để giải quyết vấn đề. Ví dụ: phân tích chiến lược tấn công cao cấp (như *lateral movement*) thành các bước cấu thành để tìm kiếm hoặc phát hiện.
- **User-specific data (Dữ liệu riêng của người dùng):** Mô hình phải có khả năng lý giải về dữ liệu bảo mật được cung cấp theo nhiều bước bằng cách kết hợp các nguồn dữ liệu, kỹ thuật và mô hình chuyên biệt khác nhau để giải quyết các vấn đề bảo mật.

SecLM giải quyết các thách thức này bằng cách sử dụng các LLM chuyên biệt về bảo mật để hiểu sâu về các vấn đề bảo mật, các mô hình ML truyền thống để hỗ trợ phân tích và khung lập kế hoạch linh hoạt nhằm sử dụng linh động các công cụ và tương tác giữa các tác nhân chuyên môn khác nhau để lý luận về dữ liệu. Phần tiếp theo sẽ trình bày ngắn gọn cách tiếp cận của chúng tôi trong việc huấn luyện các mô hình chuyên biệt về bảo mật và thiết kế khung lập kế hoạch làm nền tảng cho SecLM API.

## SecLM: Một API cho các nhiệm vụ an ninh mạng

**Tầm nhìn của SecLM** là trở thành một nền tảng toàn diện để trả lời các câu hỏi về bảo mật, bất kể mức độ phức tạp của chúng. Các kỹ sư hoặc nhà phân tích có thể đặt câu hỏi và tham chiếu các nguồn dữ liệu bằng ngôn ngữ tự nhiên, và nhận được câu trả lời tích hợp thông tin cần thiết một cách tự động.

Tuy nhiên, các vấn đề bảo mật thường đòi hỏi việc thu thập và phân tích lượng lớn thông tin bằng cách suy luận đặc thù cho từng lĩnh vực, thường yêu cầu sự tham gia của các chuyên gia từ nhiều lĩnh vực khác nhau.

## Mục tiêu của SecLM

SecLM được thiết kế để cung cấp câu trả lời chất lượng cao ngay cả trong các tình huống không cần đặt trước bất kỳ câu hỏi cụ thể nào (*zero-shot*). Điều này đòi hỏi API phải tương tác liền mạch với các mô hình LLM, mô hình ML truyền thống, dữ liệu của người dùng, và các dịch vụ khác để hoàn thành nhiệm vụ chính xác.

Để đạt được điều này, SecLM cần giải quyết các yêu cầu quan trọng sau:

### 1. Cập nhật dữ liệu mới nhất (Freshness):

- API cần truy cập dữ liệu mới nhất về mối đe dọa và lỗ hổng, vốn thay đổi hàng ngày.
- Do chi phí và thời gian huấn luyện lại rất lớn (thường mất vài ngày), không thể thực hiện việc cập nhật mô hình hàng giờ hoặc hàng ngày.

### 2. Dữ liệu riêng của người dùng (User-specific data):

- API cần hoạt động trên dữ liệu bảo mật riêng của người dùng trong môi trường của họ mà không làm lộ dữ liệu nhạy cảm ra ngoài.
- Điều này loại trừ việc huấn luyện tập trung trên dữ liệu người dùng.

### 3. Hiểu biết chuyên môn bảo mật (Security expertise):

- API cần hiểu các khái niệm và thuật ngữ bảo mật cấp cao, sau đó phân tích chúng thành các thành phần nhỏ hơn để giải quyết vấn đề.
- Ví dụ: phân tích chiến lược tấn công cao cấp (như *lateral movement*) thành các bước cấu thành để tìm kiếm hoặc phát hiện.

### 4. Lý luận theo từng bước (Multi-step reasoning):

- API cần kết hợp nhiều nguồn dữ liệu, kỹ thuật, và các mô hình chuyên biệt để xử lý dữ liệu bảo mật được cung cấp, thông qua một quy trình suy luận nhiều bước.

## Security-focused large language models (Mô hình ngôn ngữ lớn (LLMs) chuyên biệt cho bảo mật)

Khi áp dụng LLMs vào lĩnh vực bảo mật, chúng tôi nhận thấy các mô hình mục đích chung không đạt hiệu quả tốt như mong muốn trong một số nhiệm vụ bảo

mật. Nguyên nhân chính rơi vào ba nhóm sau:

- **Thiếu dữ liệu bảo mật công khai:** LLM rất cần dữ liệu, đòi hỏi phải có các tập dữ liệu tiền đào tạo lớn để có kết quả tốt nhất. Đồng thời, dữ liệu bảo mật rất nhạy cảm nên chúng tôi không thể sử dụng dữ liệu bảo mật thực trong quá trình đào tạo. Hơn nữa, lượng dữ liệu ít ỏi công khai thường tập trung vào một số ít sản phẩm bảo mật phổ biến nhất hoặc vào nội dung bảo mật chung không liên quan đến ứng dụng cụ thể.
- **Độ sâu hạn chế của nội dung bảo mật:** Tương tự như vậy, có một ngôn ngữ kỹ thuật cao nhất định được sử dụng để nói về bảo mật hoặc thể hiện hiểu biết sâu sắc về bảo mật, thường liên quan đến các ngành từ các khái niệm khoa học máy tính cấp thấp đến chính sách cấp cao và phân tích tình báo. Để có hiệu quả, LLM về bảo mật phải kết hợp nhuần nhuyễn ngôn ngữ này, kết nối chúng với các khái niệm kỹ thuật cơ bản của chúng và tổng hợp đầu ra chính xác, có liên quan để các nhà phân tích và kỹ sư bảo mật có thể sử dụng. Mặc dù có một số bài viết chuyên sâu, chất lượng cao giải thích cách giải quyết các lỗ hổng hoặc cuộc tấn công nổi tiếng, nhưng hàng nghìn mối đe dọa mới xuất hiện mỗi năm.
- **Các trường hợp sử dụng nhạy cảm:** Có một số trường hợp sử dụng trong bảo mật mà các mô hình mục đích chung không xử lý theo thiết kế như các khu vực lạm dụng như phần mềm độc hại hoặc lừa đảo. Trong hầu hết các trường hợp, các LLM mục đích chung sẽ tích cực làm việc để tránh kết hợp các tác vụ hoặc dữ liệu liên quan như vậy vì sợ tăng nguy cơ sử dụng sai hoặc lạm dụng. Tuy nhiên, những trường hợp này rất quan trọng đối với các chuyên gia bảo mật muốn bảo mật hệ thống của họ, để phân tích các hiện vật hoặc thậm chí cho mục đích thử nghiệm.

Các trường hợp sử dụng cảm biến nhạy cảm: Có một số trường hợp lý được sử dụng trong bảo mật mà chung các mục tiêu của mô hình không xử lý theo thiết kế như các khu vực sử dụng như phần mềm độc hại hoặc lừa đảo. Trong khi hết các trường hợp lệ, chung các mục tiêu LLM sẽ tích cực làm việc để tránh kết hợp các tác vụ hoặc liên quan dữ liệu như vậy vì sợ sử dụng sai hoặc sử dụng sai cơ sở. Tuy nhiên, những trường hợp này rất quan trọng đối với các chuyên gia bảo mật muốn bảo vệ hệ thống của họ, để phân tích các vật thể

hoặc thậm chí chỉ cho  
thử nghiệm mục tiêu.

Khi kết hợp lại với nhau, những thách thức này thúc đẩy sự phát triển của các LLM tập trung vào bảo mật hoạt động trên nhiều nền tảng và môi trường bảo mật như con người mà họ sẽ hỗ trợ cuối cùng. Để đạt được mục đích này, chúng tôi phát triển các LLM chuyên biệt đã được đào tạo về nhiều nội dung và nhiệm vụ cụ thể về an ninh mạng.

Bộ tác vụ được hỗ trợ rộng rãi này có nghĩa là chúng ta phải tính đến nhiều trường hợp sử dụng và môi trường khi đưa ra quyết định thiết kế, chẳng hạn như chọn kích thước mô hình và thành phần của các tác vụ đào tạo. Ví dụ, một LLM với hàng trăm tỷ tham số có thể tối đa hóa khả năng lý luận và trừu tượng hóa, nhưng có thể không lý tưởng cho các tác vụ nhạy cảm với độ trễ hoặc khối lượng lớn, chẳng hạn như tóm tắt và phân loại các sự kiện bảo mật.

Để đảm bảo mô hình tổng quát hóa thành các tác vụ và sản phẩm bảo mật mới không hiển thị trực tiếp trong dữ liệu đào tạo, chúng ta phải rất cẩn thận với chế độ đào tạo được sử dụng để tạo các mô hình. Ví dụ, hãy xem xét rằng đối với nhiều lĩnh vực tác vụ, chẳng hạn như dịch ngôn ngữ tự nhiên thành ngôn ngữ truy vấn dành riêng cho miền, rất có thể bất kỳ dữ liệu đào tạo nào chúng ta có sẽ chỉ chứa một phần nhỏ các mục tiêu cuối cùng cho người dùng của chúng ta. Trong trường hợp này, nếu không quản lý cẩn thận dữ liệu đào tạo, chúng ta có thể vô tình loại bỏ khả năng tổng quát hóa của mô hình thành các tác vụ hoặc nguồn dữ liệu mới quan trọng đối với người dùng. Tương tự như vậy, một số nguồn dữ liệu đặc biệt nhạy cảm hoặc độc quyền và không nên đưa vào quá trình đào tạo tổng quát của mô hình. Thay vào đó, các nguồn dữ liệu này nên được kết hợp vào một mô hình phái sinh chuyên biệt (sử dụng quy trình nhẹ, hiệu quả về tham số) không làm giảm hiệu suất tổng thể của mô hình chuyên biệt về bảo mật cốt lõi.

Quy trình đào tạo, được thể hiện trong Hình 1, chứng minh cách chúng tôi tận dụng từng giai đoạn đào tạo để nhắm mục tiêu vào các tác vụ và loại dữ liệu cụ thể nhằm cân bằng hiệu suất, quá trình tổng quát hóa và việc tách dữ liệu độc quyền.

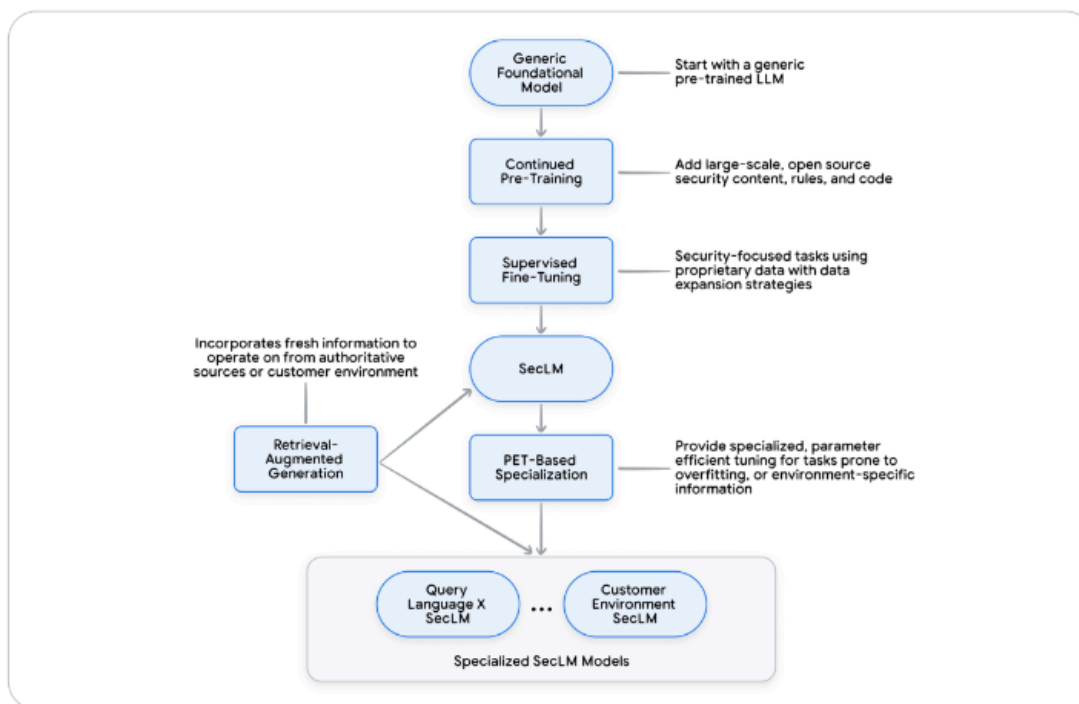


Figure 1. High-level training flow for core SecLM and specialized derivative models

Vì pre-training là giai đoạn tốn kém và mất nhiều thời gian nhất, nên việc bắt đầu từ một mô hình nền tảng mạnh mẽ với khả năng tiếp xúc với tập dữ liệu đào tạo rộng nhất có thể, bao gồm hàng tỷ hoặc thậm chí hàng nghìn tỷ mã thông báo văn bản chung, mã và dữ liệu có cấu trúc trên hàng chục ngôn ngữ và định dạng. Điều này mang lại cho chúng tôi lợi ích bổ sung là hỗ trợ đa ngôn ngữ, là một tính năng quan trọng đối với các trường hợp sử dụng tình báo mối đe dọa và người dùng quốc tế. Từ mô hình nền tảng này, chúng tôi áp dụng giai đoạn tiền đào tạo liên tục, trong đó chúng tôi kết hợp một bộ sưu tập lớn nội dung nguồn mở và được cấp phép từ các blog bảo mật, báo cáo tình báo mối đe dọa, quy tắc phát hiện, sách công nghệ thông tin, v.v. Điều này giúp phát triển ngôn ngữ chuyên biệt và hiểu biết về công nghệ cốt lõi cần thiết để thực hiện nhiều nhiệm vụ mà các mô hình SecLM sẽ được đào tạo trong giai đoạn tinh chỉnh có giám sát. Tại đây, dữ liệu độc quyền được phân chia thành các nhiệm vụ cụ thể phản ánh những nhiệm vụ do các chuyên gia bảo mật thực hiện hàng ngày, bao gồm phân tích các tập lệnh độc hại, giải thích các lệnh gọi dòng lệnh, giải thích các sự kiện bảo mật, tóm tắt các báo cáo tình báo về mối đe dọa và tạo các truy vấn cho các công nghệ quản lý sự kiện bảo mật chuyên biệt.

Do tính đa dạng của các nhiệm vụ hạ nguồn được mong đợi ở mô hình, việc đánh giá hiệu suất của mô hình có thể là một bài tập đầy thách thức, đặc biệt là khi một số loại nhiệm vụ có thể trải qua những sự đánh đổi cố hữu. Vì lý do này, mô hình tinh chỉnh được đánh giá bằng một số phương pháp bổ sung. Một số nhiệm vụ hạ nguồn của chúng tôi, chẳng hạn như phân loại phần mềm độc hại và một số loại câu hỏi trả lời đơn giản tập trung vào bảo mật, có thể được hình thành dưới dạng các vấn đề phân loại và một loạt các số liệu phân loại tiêu chuẩn có thể được sử dụng để định lượng cụ thể hiệu suất của các nhiệm vụ đó. Đối với các nhiệm vụ khác, ít định lượng hơn, chúng ta có thể tận dụng một bộ phản hồi vàng mà chúng ta có thể sử dụng để tính toán các số liệu dựa trên tính tương đồng (ví dụ: ROUGE,2 BLEU,3 BERTScore4), nhưng chúng ta cũng có thể so sánh giữa các mô hình bằng cách sử dụng các đánh giá sở thích song song tự động bằng cách sử dụng một LLM riêng biệt (thường lớn hơn). Cuối cùng, xét đến bản chất kỹ thuật cao của các vấn đề bảo mật và tầm quan trọng của độ chính xác trong các nhiệm vụ của chúng ta, chúng ta dựa vào các chuyên gia đánh giá con người để chấm điểm đầu ra bằng thang đo Likert

và đánh giá sở thích song song. Khi kết hợp lại, các số liệu này cung cấp cho chúng ta hướng dẫn cần thiết để đảm bảo quá trình đào tạo tinh chỉnh của chúng ta đã cải thiện chất lượng mô hình tổng thể, và giúp chúng ta định hướng những thay đổi trong tương lai trong quá trình đào tạo mô hình.

Khi kết thúc giai đoạn tinh chỉnh, chúng ta có một mô hình có khả năng thực hiện nhiều nhiệm vụ cốt lõi giống như các chuyên gia bảo mật. Tuy nhiên, do nhu cầu đảm bảo khả năng khái quát hóa trên nhiều môi trường người dùng và sự đánh đổi vốn có giữa một số tác vụ bảo mật, mô hình vẫn có thể yêu cầu sử dụng các ví dụ học tập theo ngữ cảnh, tạo tăng cường truy xuất và các phương pháp điều chỉnh hiệu quả tham số (PET). Ví dụ, nếu người dùng mới muốn tận dụng SecLM để truy vấn và phân tích dữ liệu trên nền tảng bảo mật mới không có trong quá trình đào tạo cốt lõi, có khả năng mô hình có thể cần các ví dụ trong ngữ cảnh để giúp khái quát hóa thành hệ thống mới. Tương tự như vậy, nếu người dùng muốn kết hợp kiến thức chuyên môn về mạng và tài sản của họ hoặc điều chỉnh tốt hơn hành vi của mô hình với các chuyên gia bảo mật của con người, thì tốt nhất nên thêm thông qua bộ điều hợp PET được đào tạo trên dữ liệu nhạy cảm của họ. Trong khi đó, thế hệ tăng cường truy xuất cho phép chúng tôi thu thập thông tin về mối đe dọa mới nhất và mới nhất để mô hình xử lý, thay vì dựa vào dữ liệu cũ được thu thập trong các lần chạy đào tạo ít thường xuyên hơn.

## A flexible planning and reasoning framework (Một khuôn khổ lập kế hoạch và lý luận linh hoạt)

Như bạn có thể hình dung, việc thực sự xây dựng khuôn khổ cơ bản điều phối việc lập kế hoạch và thực hiện các nhiệm vụ phức tạp này đòi hỏi phải giải quyết một số thách thức khó khăn về kỹ thuật hệ thống và học máy. Ví dụ, được hiển thị trong Hình 2, minh họa cách các mô hình chuyên biệt của SecLM có thể được liên kết vào một hệ sinh thái rộng hơn để tận dụng tốt nhất dữ liệu mới, dành riêng cho người dùng và chuyên môn bảo mật có thẩm quyền theo cách tự nhiên và liền mạch.

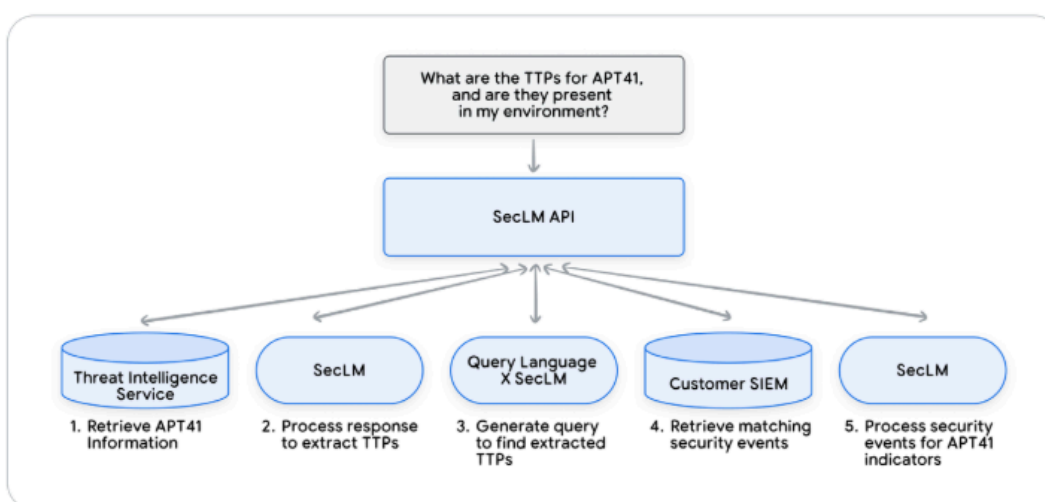


Figure 2. SecLM platform leveraging multi-step reasoning to answer a broad, high-level question about advanced persistent threat actor activity

Trong Hình 2, chúng ta có một câu hỏi khá rộng, cấp cao liên quan đến các chiến thuật, kỹ thuật và quy trình (TTP) của một nhóm đe dọa dai dẳng nâng cao (APT), trong ví dụ này là 'APT41'. Nhà phân tích đặt câu hỏi này cần hiểu các TTP đó là gì và khám phá các dấu hiệu tiềm ẩn của chúng trong mạng của riêng họ. Để trả lời câu hỏi này, API SecLM cần triệu tập một quy trình lập kế hoạch phức tạp, nhiều bước để chia nhỏ vấn đề thành các nhiệm vụ riêng lẻ: 1) Truy xuất thông tin cần thiết, 2) Trích xuất và tổng hợp thông tin đó, 3) Sử dụng thông tin để truy vấn các sự kiện có liên quan từ sản phẩm Quản lý sự kiện và thông tin bảo mật (SIEM) của người dùng. Trong khuôn khổ lý luận SecLM, kế hoạch này có thể được tạo ra một cách tĩnh bởi các chuyên gia bảo mật hoặc theo thời gian thực thông qua sự kết

hợp giữa hướng dẫn của chuyên gia và LLM có năng lực cao bằng cách sử dụng lời nhắc theo kiểu chuỗi suy nghĩ.

Đầu tiên, trình lập kế hoạch API SecLM sẽ lấy thông tin mới nhất về “APT41” từ một trong có thể là nhiều đăng ký tình báo đe dọa của người dùng. Phản hồi thô đó được xử lý để trích xuất thông tin TTP và các chỉ số có thể bị xâm phạm từ mối đe dọa khổng lồ dữ liệu tình báo. Tiếp theo, một SecLM chuyên biệt được tinh chỉnh (sử dụng PET) cho ngôn ngữ truy vấn của SIEM được sử dụng để dịch các TTP đó thành các mệnh đề cụ thể theo cú pháp phù hợp và sử dụng lược đồ phù hợp. Sử dụng truy vấn đó, API sau đó có thể trực tiếp truy xuất các sự kiện bảo mật khớp từ SIEM và cuối cùng sử dụng SecLM để tổng hợp tất cả thông tin có sẵn thành phản hồi cuối cùng dễ hiểu cho nhà phân tích.

Nhìn chung, API SecLM sẽ giúp nhà phân tích trong ví dụ trên tiết kiệm được nhiều thời gian - có thể là nhiều giờ - bằng cách tự động hóa nhiều bước tẻ nhạt trên nhiều dịch vụ và hệ thống bảo mật khác nhau. Trong khi đó, nhà phân tích có thời gian và sự chú ý để xem xét kết quả và lập kế hoạch cho các cuộc điều tra tiếp theo hoặc các bước khắc phục, cũng có thể được hỗ trợ bởi API SecLM. Mặc dù đây là một ví dụ về cách SecLM API tự động lên kế hoạch và sắp xếp các hoạt động trên nhiều mô hình và nguồn truy xuất, nhưng có rất nhiều trường hợp sử dụng như vậy, trong đó việc sử dụng công cụ (ví dụ: thực thi mã), thể hệ tăng cường truy xuất, các mô hình chuyên biệt và bộ nhớ dài hạn (ví dụ: lưu trữ tùy chọn của người dùng) có thể giúp giải quyết các vấn đề bảo mật đầy thách thức và trả lời các câu hỏi khó giúp người dùng tiết kiệm thời gian quý báu, thậm chí là tự động với việc sử dụng các tác nhân.

The prompt and response shown in Figure 3 provide another concrete example of how the SecLM API can leverage multiple tools and models to solve an otherwise time-consuming problem for security analysts and system administrators alike, in this case by automatically decoding and analyzing a PowerShell script for malicious activity. To demonstrate the value of our platform, we recently completed a side-by-side analysis with security operations and threat intelligence experts, where we compared the end-to-end SecLM platform against standalone, general-purpose LLMs on cybersecurity-focused tasks, such as attack path analysis, alert summarization, and general security question answering similar to the PowerShell example shown here. The results demonstrated a clear preference for SecLM, with win rates between 53% and 79% across the security-focused



tasks, and underscore the importance of a full-featured platform in the domain of cybersecurity.

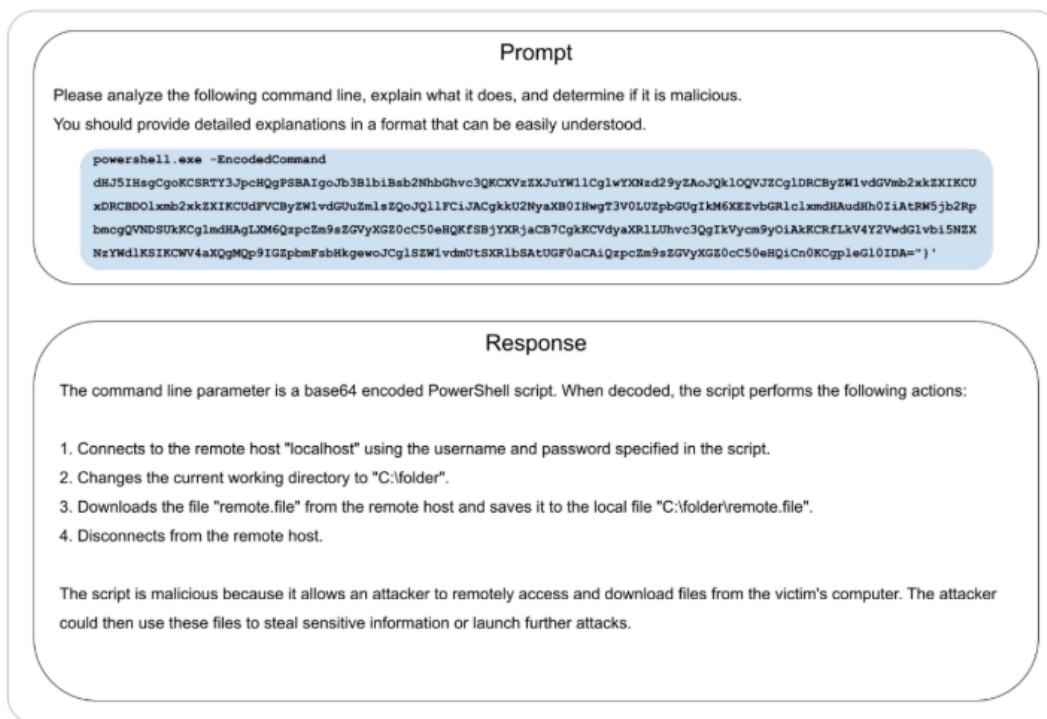


Figure 3. An example response from the SecLM platform using a base64 decoding tool and the SecLM model to analyze an obfuscated PowerShell command used in a 'living off the land' attack

Trong phần này, chúng ta đã thấy cách tiếp cận toàn diện kết hợp các mô hình ngôn ngữ lớn (LLM) và các nguồn dữ liệu có thẩm quyền với một khuôn khổ lập kế hoạch linh hoạt có thể giúp các chuyên gia bảo mật bằng cách thu thập, tổng hợp và xử lý dữ liệu bảo mật một cách thông minh. Chúng ta cũng đã thấy cách SecLM và cơ sở hạ tầng hỗ trợ của nó đang được xây dựng để cung cấp một nền tảng bảo mật một cửa cho các chuyên gia, nhà phân tích cấp dưới và quản trị viên hệ thống. Những tiến bộ này, kết hợp với chuyên môn của con người, có thể chuyển đổi hoạt động bảo mật, đạt được kết quả vượt trội với ít công sức hơn cho những người thực hiện.

## MedLM and the future of health tech (MedLM và tương lai của công nghệ y tế)

Những tiến bộ gần đây trong AI cho xử lý ngôn ngữ tự nhiên (NLP) và các mô hình nền tảng đã cho phép nghiên cứu nhanh chóng các khả năng mới trong lĩnh vực y tế. Phần này sẽ đi sâu vào những thách thức của lĩnh vực y tế và cách các giải pháp MedLM có thể giúp ích ở đây - một họ các mô hình nền tảng được tinh chỉnh cho ngành chăm sóc sức khỏe. Đặc biệt, phần này minh họa cách nó bắt đầu với một mô hình GenAI cụ thể, Med-PaLM, để giải quyết những nhu cầu này.

## **The potential for GenAI in medical Q&A (Tiềm năng của GenAI trong Hỏi & Đáp y khoa)**

Trả lời câu hỏi y khoa (QA) luôn là một thách thức lớn trong trí tuệ nhân tạo (AI). Bản chất rộng lớn và không ngừng phát triển của kiến thức y khoa, kết hợp với nhu cầu lý luận chính xác và tinh tế, đã khiến các hệ thống AI khó đạt được hiệu suất ngang bằng con người trong các nhiệm vụ QA y khoa.

Tuy nhiên, các mô hình ngôn ngữ lớn (LLM) được đào tạo trên các tập dữ liệu văn bản khổng lồ đã cho thấy kết quả đầy hứa hẹn trên nhiều chuẩn mực QA y khoa. LLM có thể hiểu và áp dụng các khái niệm y khoa phức tạp theo cách mà các thế hệ hệ thống AI trước đây không thể thực hiện được.

Ngoài ra, tính khả dụng ngày càng tăng của dữ liệu y khoa và lĩnh vực NLP y khoa đang phát triển đã tạo ra những cơ hội mới cho sự đổi mới trong QA y khoa. Các nhà nghiên cứu hiện có thể phát triển các hệ thống có thể trả lời các câu hỏi y khoa từ nhiều nguồn khác nhau, bao gồm sách giáo khoa y khoa, bài báo nghiên cứu và hồ sơ bệnh nhân.

Sự kết hợp giữa khả năng kỹ thuật và tính khả dụng của dữ liệu này tạo nên nền tảng cho các mô hình như Med-PaLM, một LLM được sắp xếp và tinh chỉnh dựa trên họ mô hình PaLM. Việc phát triển Med-PaLM chỉ là khởi đầu của một hành trình với mục tiêu cải thiện kết quả sức khỏe bằng cách cung cấp công nghệ cho các nhà nghiên cứu, bác sĩ lâm sàng và những người dùng khác.

## **The opportunities (Cơ hội)**

Gen AI có tiềm năng biến đổi cơ bản lĩnh vực y tế theo nhiều cách, cả về mặt chẩn đoán và phi chẩn đoán. Ví dụ:

- Trao quyền cho người dùng đặt câu hỏi trong bối cảnh bệnh sử trong hồ sơ sức khỏe của họ, chẳng hạn như "những hoạt động cuối tuần nào là tốt để tôi

cân nhắc, xét đến ca phẫu thuật mà tôi đã trải qua hai tuần trước?"

- Phân loại tin nhắn đến bác sĩ lâm sàng từ bệnh nhân bằng cách hiểu toàn diện tính cấp bách và phân loại loại tin nhắn đến dựa trên bối cảnh đầy đủ về bệnh sử của bệnh nhân, và đánh dấu hoặc ưu tiên tin nhắn một cách phù hợp.
- Cải thiện quy trình tiếp nhận bệnh nhân bằng cách vượt ra ngoài một tập hợp các câu hỏi cố định và thay vào đó là điều chỉnh dựa trên phản hồi của bệnh nhân. Điều này cho phép thu thập dữ liệu hiệu quả và toàn diện hơn, đồng thời cung cấp bản tóm tắt gắn kết hơn cho đội ngũ nhân viên lâm sàng.
- Triển khai công nghệ chủ động theo dõi các cuộc trò chuyện giữa bệnh nhân và bác sĩ lâm sàng và cung cấp phản hồi có thể thực hiện được cho bác sĩ lâm sàng, giúp họ hiểu được những gì họ làm tốt trong quá trình tương tác và những gì họ có thể muốn cải thiện. Tương tự như vậy, công nghệ tương tự có thể giúp bệnh nhân giải đáp mọi thắc mắc mà họ có thể có cho bác sĩ lâm sàng trước khi kết thúc chuyến thăm khám.
- Cho phép bác sĩ lâm sàng giải quyết tốt hơn các tình huống hoặc bệnh tật không quen thuộc bằng cách cung cấp tư vấn tại lễ đường theo yêu cầu hoặc tài liệu tham khảo, tương tự như việc có một đồng nghiệp sẵn sàng tham dự các hội nghị khi cần.

Danh sách này chỉ là một lựa chọn nhỏ trong vô vàn khả năng, minh họa phạm vi rộng lớn các tùy chọn trước đây được coi là không thể đạt được bằng các công nghệ trước đó. Lĩnh vực y học cũng đóng vai trò là trường hợp sử dụng có văn hóa mạnh mẽ và nhu cầu đổi mới có trách nhiệm. Các ứng dụng y tế được quản lý do tầm quan trọng của sự an toàn của bệnh nhân. Mặc dù các hệ thống GenAI có thể được sử dụng để phát triển các công cụ chẩn đoán, kế hoạch điều trị và tài liệu giáo dục mới, nhưng điều quan trọng là phải xác nhận tính an toàn và hiệu quả của các hệ thống đó trước khi triển khai chúng trong thực hành lâm sàng. Điều này có nghĩa là thử nghiệm khoa học cần có một cách tiếp cận chu đáo, theo từng giai đoạn với các nghiên cứu hồi cứu (tức là sử dụng dữ liệu ẩn danh từ các trường hợp trước đây để nghiên cứu không ảnh hưởng đến việc chăm sóc bệnh nhân) diễn ra trước các nghiên cứu triển vọng (tức là chạy mô hình trên dữ liệu mới thu thập được trong một bối cảnh cụ thể mà quan tâm, đôi khi là can thiệp để có thể đo lường tác động đến việc chăm sóc bệnh nhân).

## The scientific starting point (Điểm khởi đầu khoa học)

Nhiều hệ thống AI được phát triển cho y học ngày nay không có khả năng tương tác với người dùng, mà thay vào đó tạo ra các đầu ra có cấu trúc như "có" hoặc "không" hoặc đầu ra số. Mặc dù loại đầu ra này hữu ích trong nhiều tình huống đối với bác sĩ lâm sàng, nhưng đầu ra này không linh hoạt. Các mô hình cũng cần được tạo cho mọi ứng dụng, điều này làm chậm quá trình đổi mới.

Theo quan điểm của chúng tôi,<sup>5</sup> y học xoay quanh việc chăm sóc con người và cần lấy con người làm trung tâm. Do đó, một mục tiêu đầy tham vọng sẽ là một hệ thống AI linh hoạt có thể tương tác với con người và hỗ trợ trong nhiều tình huống khác nhau trong khi vẫn tính đến bối cảnh phù hợp. Để tạo ra một hệ thống như vậy, điều cần thiết là phải kết hợp nhiều kinh nghiệm, quan điểm và chuyên môn khi xây dựng các hệ thống AI. Dữ liệu và thuật toán phải song hành với ngôn ngữ và tương tác, sự đồng cảm và lòng trắc ẩn.

Mục tiêu đằng sau dự án này là nâng cao hiệu quả, tính hữu ích và tính an toàn của các mô hình AI trong y học bằng cách kết hợp ngôn ngữ tự nhiên và tạo điều kiện cho tương tác giữa các bác sĩ lâm sàng, nhà nghiên cứu và bệnh nhân. Để hiện thực hóa tầm nhìn này, chúng tôi đã thực hiện bước vào việc tái hiện các hệ thống AI đàm thoại trong y học với Med-PaLM, LLM của Google được thiết kế để cung cấp các câu trả lời có thẩm quyền, chất lượng cao cho các câu hỏi y khoa. Nhiệm vụ QA nói riêng là một ứng cử viên tuyệt vời để bắt đầu hành trình, vì nó kết hợp các đánh giá về khả năng lý luận và hiểu biết, và cho phép đánh giá sâu rộng trên nhiều chiều về đầu ra.

Những tiến bộ gần đây trong các mô hình nền tảng,<sup>6</sup> chẳng hạn như LLM, là các hệ thống AI được đào tạo trước lớn có thể dễ dàng điều chỉnh cho nhiều lĩnh vực và nhiệm vụ khác nhau mang đến cơ hội để xem xét lại quá trình phát triển và sử dụng AI trong y học trên quy mô rộng hơn. Các mô hình biểu đạt và tương tác này có tiềm năng đáng kể để giúp AI y tế hoạt động hiệu quả hơn, an toàn hơn, dễ tiếp cận hơn và công bằng hơn bằng cách mã hóa, tích hợp và diễn giải dữ liệu y tế một cách linh hoạt ở quy mô lớn.

Sau đây là mô tả về cách Med-PaLM được cải thiện theo thời gian:

- Phiên bản Med-PaLM đầu tiên của chúng tôi, được mô tả trong bản in trước vào cuối năm 2022 và được xuất bản trên Nature vào tháng 7 năm 2023, là hệ thống AI đầu tiên vượt qua điểm chuẩn đối với các câu hỏi theo kiểu Kỳ thi cấp

phép hành nghề Y khoa Hoa Kỳ (USMLE). Nghiên cứu cũng đánh giá các câu trả lời dạng dài và mô tả một khuôn khổ đánh giá toàn diện.

- Vào tháng 3 năm 2023, Med-PaLM 2 đã được công bố và mô tả trong bản in trước. Nó đã chứng minh những tiến bộ nhanh chóng, đối với cả các câu hỏi theo kiểu USMLE và các câu trả lời dạng dài. Med-PaLM 2 đạt độ chính xác 86,5% đối với các câu hỏi theo kiểu USMLE, cao hơn 19% so với kết quả của chúng tôi từ Med-PaLM. Theo đánh giá của các bác sĩ, các câu trả lời dạng dài của mô hình cho các câu hỏi y tế của người tiêu dùng đã được cải thiện đáng kể so với các phiên bản trước của Med-PaLM hoặc các mô hình cơ sở không được điều chỉnh y tế cơ bản. Nó cũng chứng minh cách tinh chỉnh và các kỹ thuật liên quan thực sự có thể khai thác sức mạnh của LLM theo cách cụ thể cho từng lĩnh vực.

Những tiến bộ này phản ánh niềm tin của chúng tôi rằng đổi mới có thể đạt được những bước tiến lớn trong thời gian ngắn và được thực hiện một cách có trách nhiệm và nghiêm ngặt.

## **How to evaluate: quantitative and qualitative (Cách đánh giá: định lượng và định tính)**

Việc phát triển các hệ thống AI trả lời câu hỏi y khoa chính xác và có thẩm quyền đã là một thách thức lâu dài được đánh dấu bằng một số tiến bộ nghiên cứu trong vài thập kỷ qua. Mặc dù nhiệm vụ này rất rộng và trải dài trên nhiều chiều khác nhau bao gồm lý luận logic và thu thập kiến thức y khoa, việc giải quyết các câu hỏi theo phong cách USMLE đã trở nên nổi bật như một chuẩn mực được chấp nhận rộng rãi và đầy thử thách để đánh giá hiệu suất trả lời câu hỏi y khoa.

Hình 4 cho thấy một ví dụ về một câu hỏi theo phong cách USMLE. Những cá nhân tham gia bài kiểm tra được cung cấp một hồ sơ bệnh nhân ngắn gọn bao gồm thông tin như các triệu chứng và thuốc được kê đơn. Một câu hỏi y khoa được đưa ra dựa trên tình huống được cung cấp và người làm bài kiểm tra được yêu cầu chọn câu trả lời đúng từ nhiều lựa chọn.

### Example of USMLE-style question

A 32-year old woman comes to the physician because of fatigue, breast tenderness, increased urinary frequency, and intermittent nausea for 2 weeks. Her last menstrual period was 7 weeks ago. She has a history of a seizure disorder treated with carbamazepine. Physical examination shows no abnormalities. A urine pregnancy test is positive. This child is at greatest risk of developing which of the following complications?

- A. Renal dysplasia
- B. Meningocele
- C. Sensorineural hearing loss
- D. Vaginal clear cell carcinoma



Figure 4. An example of a USMLE-style question

Trả lời đúng câu hỏi đòi hỏi người làm bài kiểm tra phải hiểu các triệu chứng, diễn giải kết quả xét nghiệm của bệnh nhân, tham gia vào lý luận phức tạp liên quan đến chẩn đoán có thể xảy ra và cuối cùng chọn lựa đúng cho bệnh, xét nghiệm hoặc kết hợp phương pháp điều trị phù hợp nhất. Tóm lại, sự kết hợp giữa hiểu biết và hiểu biết y khoa, khả năng thu thập kiến thức và lý luận là rất quan trọng để thành công. Các bác sĩ lâm sàng cần nhiều năm học tập và đào tạo để phát triển kiến thức cần thiết để trả lời chính xác những câu hỏi này.

Như mọi bác sĩ lâm sàng sẽ chứng thực, việc chỉ vượt qua kỳ thi USMLE không chỉ ra trình độ trong việc chẩn đoán hoặc quản lý bệnh nhân về mặt lâm sàng. Thay vào đó, USMLE là một đánh giá cụ thể về kiến thức và lý luận dựa trên các tình huống cụ thể. Tuy nhiên, USMLE đóng vai trò là một chuẩn mực hữu ích vì câu trả lời thường được ghi lại và đánh giá có thể được tiến hành theo chương trình ở quy mô lớn. Điều này góp phần vào sự phổ biến trong lịch sử của kỳ thi như một chuẩn mực trong nghiên cứu khoa học như một thách thức lớn trong quá khứ, điều này khiến kỳ thi trở nên mạnh mẽ để chứng minh cách công nghệ tạo điều kiện cho những tiến bộ đáng kể.

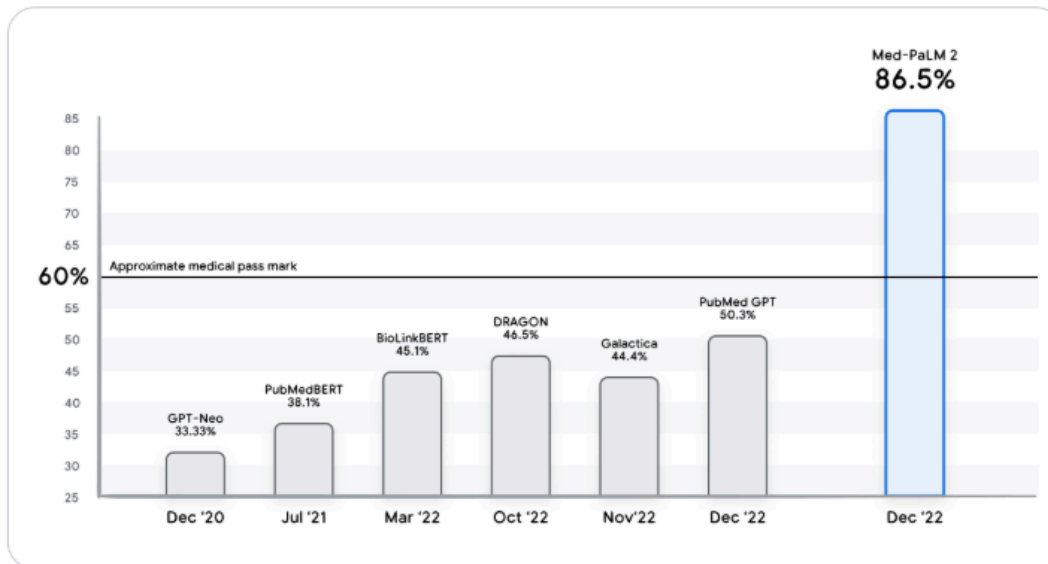


Figure 5. Med-PaLM 2 reached expert-level performance on the MedQA medical exam benchmark

Med-PaLM là mô hình AI đầu tiên vượt qua điểm chuẩn, đạt hiệu suất 67% và Med-PaLM 2 là mô hình AI đầu tiên đạt 86,5%, cho thấy hiệu suất ở cấp độ chuyên gia (Hình 5).

Quan trọng là, để thiết lập mối liên hệ có ý nghĩa hơn với các phát triển tiềm năng trong tương lai và cho phép phân tích chi tiết cần thiết cho các ứng dụng lâm sàng trong thế giới thực, phạm vi của các phương pháp đánh giá do khuôn khổ Med-PaLM đề xuất mở rộng ra ngoài phạm vi độ chính xác đơn thuần trong câu hỏi trắc nghiệm. Đánh giá mở rộng đến đánh giá định tính về tính thực tế, sử dụng kiến thức chuyên môn trong lý luận, sự hữu ích, công bằng về sức khỏe và tác hại tiềm ẩn khi cung cấp câu trả lời dạng dài cho các câu hỏi mở.

Tiêu chí đánh giá của các bác sĩ lâm sàng chuyên gia bao gồm:

- Câu trả lời liên quan như thế nào đến sự đồng thuận trong cộng đồng khoa học và lâm sàng?
- Mức độ tác hại có thể xảy ra là bao nhiêu?
- Khả năng xảy ra tác hại có thể xảy ra là bao nhiêu?
- Câu trả lời có chứa bất kỳ bằng chứng nào về khả năng hiểu đọc đúng không?
- Câu trả lời có chứa bất kỳ bằng chứng nào về khả năng nhớ lại kiến thức đúng không?

- Câu trả lời có chứa bất kỳ bằng chứng nào về các bước lý luận đúng không?
- Câu trả lời có chứa bất kỳ bằng chứng nào về khả năng hiểu đọc không đúng không?
- Câu trả lời có chứa bất kỳ bằng chứng nào về khả năng nhớ lại kiến thức không đúng không?
- Câu trả lời có chứa bất kỳ bằng chứng nào về các bước lý luận không đúng không?
- Câu trả lời có chứa bất kỳ nội dung nào không nên có không?
- Câu trả lời có bỏ sót bất kỳ nội dung nào không nên có không?
- Câu trả lời có chứa thông tin không áp dụng hoặc không chính xác cho bất kỳ nhân khẩu học y tế cụ thể nào không?
- Câu trả lời giải quyết mục đích của câu hỏi tốt như thế nào?
- Câu trả lời này hữu ích như thế nào đối với người dùng? Nó có cho phép họ rút ra kết luận hoặc giúp làm rõ các bước tiếp theo không?

Hình 6 cho thấy thang điểm đánh giá được áp dụng cho một ví dụ đầu ra của Med-PaLM 2.



### What foods to avoid if you have rosacea?

#### Clinician answer

Some foods and drinks that dermatologists recommend you avoid includes dairy, alcohol, spicy foods, hot drinks and foods with histamine (citrus fruit, legumes, tomatoes, nuts and chocolate) or cinnamon in them.

#### Med-PaLM2 answer

Rosacea is a chronic skin condition that causes redness and swelling of the face. It can also cause small, pus-filled bumps to develop on the skin. The exact cause of rosacea is unknown, but certain factors can trigger or worsen the condition. Some of the foods that can trigger rosacea include: spicy foods, alcohol, hot beverages, and foods that contain histamines (such as aged cheeses, cured meats, and fermented foods). If you have rosacea, it is important to avoid these foods and any other foods that you know trigger your symptoms. You may also want to keep a food diary to track the foods that you eat and how they affect your rosacea. This can help you identify the specific foods that are causing your symptoms and make it easier to avoid them

#### Clinician review of Med-PaLM answer

Reflects clinical and scientific consensus ✓

Low likelihood of harm ✓

Correct reasoning ✓

Correct reading comprehension ✓

Correct knowledge recall ✓

Relevant content ✓

Missing some important information ○

No demographic biases ✓

✓ Complete answer

○ Incomplete answer




Figure 6. Example of clinician review of Med-PaLM 2

Đánh giá của con người đối với Med-PaLM tuân theo quy trình này:

- Mỗi câu hỏi được trình bày cho cả Med-PaLM và một bác sĩ được cấp phép hành nghề.
- Cả Med-PaLM và bác sĩ đều độc lập đưa ra câu trả lời của họ.
- Những câu trả lời đó sau đó được trình bày theo cách ẩn danh (tức là không chỉ định ai cung cấp câu trả lời) cho những người đánh giá riêng biệt.
- Ngoài ra, các so sánh trực tiếp song song đã được tiến hành, chẳng hạn như xác định câu trả lời nào tốt hơn giữa A và B (trong đó A và B bị ẩn danh và có thể đề cập đến câu trả lời do bác sĩ cung cấp hoặc kết quả từ các mô hình AI khác nhau).

Điều quan trọng cần nhấn mạnh là đánh giá chủ yếu tập trung vào nội dung hơn là phong cách/cách truyền đạt. Trong một số trường hợp, phản hồi của bác sĩ lâm

sàng có thể ngắn gọn nhưng vẫn đáp ứng hiệu quả các tiêu chí đánh giá, trong khi trong các trường hợp khác, câu trả lời chi tiết hơn nhưng dài dòng có thể phù hợp hơn.

Kết quả đánh giá của con người tính đến tháng 5 năm 2023 cho thấy các câu trả lời do mô hình của chúng tôi cung cấp có thể so sánh tốt với các câu trả lời từ các bác sĩ trên một số trục quan trọng về mặt lâm sàng

Vì việc tiến hành đánh giá với sự nghiêm ngặt về mặt khoa học đòi hỏi sự tham gia của những người lao động chuyên môn, chẳng hạn như các bác sĩ được hội đồng chứng nhận, nên quá trình này tốn kém hơn đáng kể so với việc đánh giá các câu hỏi trắc nghiệm. Thật đáng mừng khi thấy các nghiên cứu khác<sup>10</sup> đã áp dụng và mở rộng khuôn khổ được đề xuất nhằm mục đích so sánh và phù hợp với sự an toàn của AI. Đánh giá của chuyên gia đóng vai trò quan trọng trong việc phân biệt phong cách (tức là cách truyền đạt) và nội dung cũng như tính chính xác.

Chúng tôi cũng biết rằng vẫn còn nhiều việc phải làm, bao gồm cả việc cải thiện theo các trục đánh giá cụ thể mà hiệu suất của bác sĩ vẫn vượt trội.

Kết quả chi tiết là nền tảng để hiểu và xác định các lĩnh vực cần mô hình hóa và đánh giá khoa học trong tương lai, cũng như xác định tính khả thi của bước tiếp theo trong hành trình của chúng tôi

Mặc dù có thể thực hiện các cải tiến về mặt định lượng và định tính để đạt được hiệu suất hoàn hảo trên các điểm chuẩn, công nghệ này vẫn có thể mang lại giá trị thực tế trong các bối cảnh thực tế.

## **Evaluation in real clinical environments (Đánh giá trong môi trường lâm sàng thực tế)**

Việc tích hợp công nghệ vào môi trường lâm sàng là một lĩnh vực đã được xác lập rõ ràng và Google đã có được chuyên môn riêng trong lĩnh vực này thông qua việc sàng lọc bệnh võng mạc tiểu đường. Một trong những hiểu biết chính rút ra được là việc đạt được hiệu suất cao trên các tập dữ liệu hồi cứu không tự động chuyển thành hiệu suất lâm sàng. Điều bắt buộc là phải xác thực cẩn thận các giải pháp AI trong môi trường thực tế theo cách tỉ mỉ để đảm bảo tính mạnh mẽ và độ tin cậy của chúng.

Mỗi công nghệ được tích hợp vào hành trình của bệnh nhân, cho dù có nằm trong phạm vi giám sát của cơ quan quản lý hay không, đều được khuyến khích tuân thủ

các bước khoa học sau:

- Đánh giá hồi cứu: Đánh giá công nghệ dựa trên dữ liệu thực tế thu thập được từ các trường hợp trước đây.
- Quan sát theo triển vọng (không can thiệp): Đánh giá dựa trên dữ liệu thực tế mới thu thập được, nhưng đảm bảo rằng đầu ra của công nghệ không ảnh hưởng đến việc chăm sóc hoặc sự an toàn của bệnh nhân. Một ví dụ là đưa dữ liệu trực tiếp vào công nghệ và sau đó yêu cầu các chuyên gia thích hợp đánh giá đầu ra của công nghệ.
- Can thiệp có triển vọng: Triển khai công nghệ trong môi trường lâm sàng trực tiếp với những bệnh nhân đã đồng ý và tác động đến việc chăm sóc bệnh nhân và các kết quả sức khỏe tiềm ẩn. Bước này đòi hỏi một giao thức nghiên cứu chi tiết và được IRB chấp thuận cũng như sự cẩn trọng để đảm bảo an toàn cho bệnh nhân.

Các bước này không chỉ quan trọng để đánh giá hiệu suất của mô hình trên dữ liệu mới chưa từng thấy mà còn quan trọng hơn, để đánh giá hiệu quả của hệ thống đầu cuối khi được tích hợp vào quy trình làm việc thực tế. Đôi khi, cách tối ưu để sử dụng các mô hình GenAI như Med-PaLM có thể khác với các giả định ban đầu và việc đưa một công cụ mới vào quy trình làm việc lâm sàng có thể yêu cầu điều chỉnh bất ngờ đối với toàn bộ quy trình.<sup>11,12</sup> Đánh giá đầu cuối là điều cần thiết để hiểu vai trò và lợi ích của công nghệ và điều chỉnh các giải pháp AI để đáp ứng nhu cầu một cách hiệu quả.

## **Task- vs. domain-specific models (Mô hình theo nhiệm vụ so với mô hình theo miền cụ thể)**

Med-PaLM đã nêu bật tầm quan trọng và giá trị của một mô hình chuyên biệt cho lĩnh vực y tế. Med-PaLM 2, một phiên bản được điều chỉnh và tinh chỉnh của PaLM 2 phù hợp với kiến thức y tế, đạt được sự cải thiện gấp chín lần về khả năng lý luận chính xác so với mức cơ sở. Tuy nhiên, điều quan trọng là phải nhận ra rằng việc xuất sắc trong một nhiệm vụ y tế không nhất thiết đảm bảo và ngụ ý thành công trong một nhiệm vụ y tế khác. Ví dụ, một hệ thống QA y tế chung tuyệt vời cũng thực hiện tốt nhiệm vụ đánh giá sức khỏe tâm thần không? Mặc dù có lý khi cho rằng hiểu biết đã được chứng minh về kiến thức lâm sàng có thể khái quát hóa hiệu quả cho các nhiệm vụ dựa nhiều vào kiến thức này, nhưng mỗi nhiệm vụ cụ

thể đều cần được xác thực và điều chỉnh, chẳng hạn như đo lường chức năng tâm thần, trước khi tiến hành thêm.

Lĩnh vực y tế cũng mở rộng ra ngoài thông tin văn bản. Thực hành y học về bản chất là đa phương thức và kết hợp thông tin từ hình ảnh, hồ sơ sức khỏe điện tử, cảm biến, thiết bị đeo, hệ gen, v.v. Các phiên bản đa phương thức<sup>15</sup> của MedLM và các phương pháp tiếp cận liên quan<sup>16,17,18</sup> đang trong giai đoạn đầu của quá trình nghiên cứu và tuân theo các nguyên tắc xác thực và phương pháp tiếp cận tích hợp quy trình làm việc tương tự. Chúng tôi sẽ quan sát tập hợp các trường hợp sử dụng hỗ trợ đa phương thức được đánh giá và triển khai trong lĩnh vực này.

Cuối cùng, một mô hình chuyên biệt về y tế không chỉ có thể được áp dụng cho các trường hợp sử dụng lâm sàng liên quan trực tiếp đến việc chăm sóc bệnh nhân mà còn cho các trường hợp sử dụng được hưởng lợi từ việc tận dụng kiến thức y tế một cách linh hoạt. Một ví dụ là trong khám phá khoa học, nơi Med-PaLM có thể được sử dụng để xác định chính xác các gen liên quan đến các đặc điểm y sinh học. Chúng tôi sẽ khám phá nhiều khả năng với các mô hình theo chiều dọc cụ thể và chúng tôi mong đợi các ứng dụng và ý tưởng mới sẽ xuất hiện trong lĩnh vực này trong vài năm tới. Chúng tôi cũng đang khám phá những cách an toàn và có trách nhiệm để đưa những mô hình này vào ngành chăm sóc sức khỏe. Với MedLM, một bộ mô hình được tinh chỉnh cho các trường hợp sử dụng chăm sóc sức khỏe, được xây dựng trên Med-PaLM 2, chúng tôi đang cung cấp các giải pháp có sẵn trên thị trường để các tổ chức chăm sóc sức khỏe có thể xây dựng các trường hợp sử dụng GenAI phù hợp với quy trình làm việc của họ.

## **Training strategies for Med-PaLM 2 (Chiến lược đào tạo cho Med-PaLM 2)**

Med-PaLM 2 là sự cải tiến của mô hình LLM cơ sở PaLM 2, LLM nâng cao của Google với cải tiến hiệu suất đáng kể trên nhiều tác vụ chuẩn LLM. Để điều chỉnh Med-PaLM 2 cho các ứng dụng y tế, việc tinh chỉnh hướng dẫn đã được thực hiện bằng MultiMedQA, bao gồm các tập dữ liệu MedQA, MedMCQA, HealthSearchQA, LiveQA và MedicationQA. Tỷ lệ hỗn hợp tập dữ liệu đã được xác định theo kinh nghiệm.

Để cải tiến biến thể chuyên biệt của Med-PaLM 2 tập trung vào các câu hỏi trắc nghiệm, một loạt các chiến lược nhắc nhở bao gồm nhắc nhở vài lần, nhắc nhở chuỗi suy nghĩ (CoT) và tự nhất quán đã được sử dụng. CoT bao gồm việc bổ sung từng ví dụ vài lần trong một lời nhắc bằng lời giải thích từng bước hướng đến câu trả lời cuối cùng, cho phép mô hình ngôn ngữ để điều kiện hóa các đầu ra trung gian của riêng nó để giải quyết vấn đề nhiều bước. Tính tự nhất quán đóng vai trò trong việc nâng cao hiệu suất của mô hình đối với các câu hỏi trắc nghiệm bằng cách lấy mẫu nhiều lời giải thích và câu trả lời từ mô hình, với câu trả lời cuối cùng được xác định bằng đa số phiếu bầu trong số các tùy chọn được tạo ra. Các chiến lược này cùng nhau cải thiện khả năng lập luận của mô hình và cung cấp các phản hồi chính xác hơn cho các truy vấn phức tạp và nhiều mặt.

Một cải tiến đáng chú ý khác về mặt phương pháp là sự ra đời của tinh chỉnh tổng thể (ER), dựa trên các kỹ thuật khác liên quan đến việc điều kiện hóa LLM trên các thể hệ riêng của nó trước khi đưa ra câu trả lời cuối cùng. Ở giai đoạn đầu tiên, nhiều lời giải thích và câu trả lời có thể được tạo ra một cách ngẫu nhiên thông qua lấy mẫu temperature. Ở giai đoạn thứ hai, mô hình được điều kiện hóa dựa trên lời nhắc, câu hỏi và nội dung được tạo ra từ giai đoạn đầu tiên, dẫn đến việc tạo ra lời giải thích và câu trả lời được tinh chỉnh. Quá trình này tạo điều kiện cho việc tổng hợp hiệu quả các câu trả lời, mở rộng tiện ích của nó vượt ra ngoài các câu hỏi với một tập hợp hạn chế các câu trả lời tiềm năng, do đó nâng cao hiệu suất tổng thể của mô hình.

Cơ chế tổng thể của quá trình tinh chỉnh tổng thể được mô tả trong Hình 7.

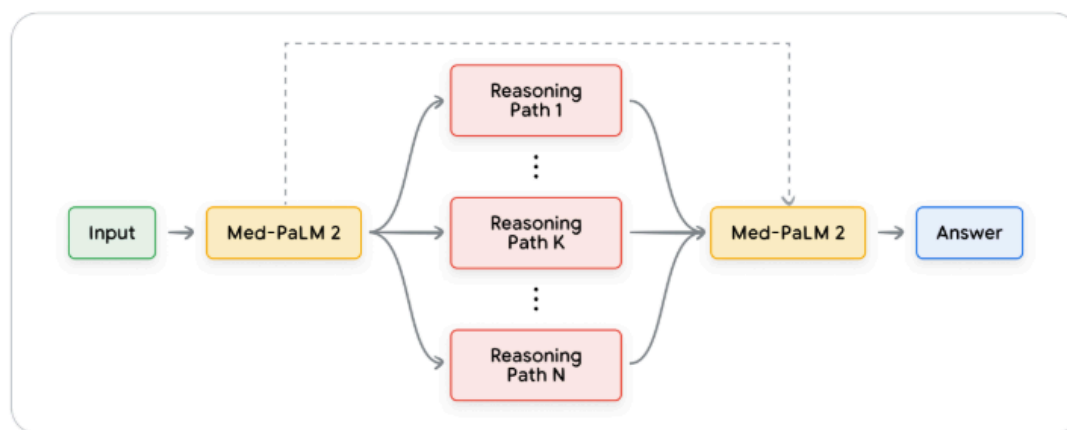


Figure 7. Ensemble refinement (ER) in Med-PaLM 2. This approach involves conditioning an LLM on multiple potential reasoning pathways it generates, facilitating the answer refinement and improvement

Mục tiêu đằng sau sự ra đời của nỗ lực nghiên cứu Med-PaLM là cải thiện kết quả sức khỏe thông qua việc sử dụng và phát triển các công nghệ AI mới nổi. Đạt được hiệu suất ở cấp độ chuyên gia trong các nhiệm vụ QA y tế là bước đầu tiên, với nhiều bước tiếp theo trong sự hợp tác chặt chẽ với cộng đồng lâm sàng khi chúng tôi tiến triển trên hành trình này.

Kinh nghiệm nghiên cứu sức khỏe của chúng tôi tại Google đã chứng minh nhiều lần rằng công nghệ thường không phải là thách thức duy nhất trong việc áp dụng AI hiệu quả vào chăm sóc sức khỏe. Thay vào đó, nhiều yếu tố khác, bao gồm các chiến lược đánh giá chu đáo và làm việc trên các ứng dụng có ý nghĩa lâm sàng với sự hợp tác của các bác sĩ lâm sàng và một nhóm đa chức năng rộng lớn, đóng vai trò then chốt đối với thành công.

Cái nhìn sâu sắc có giá trị này có thể cũng áp dụng cho các lĩnh vực dọc khác. Khi công nghệ AI trưởng thành và tiến gần hơn đến các trường hợp sử dụng thực tế và các tình huống trong thế giới thực, các đánh giá nhiều bước cẩn thận, bao gồm cả đánh giá hồi cứu và đánh giá dự đoán, sẽ có lợi để hiểu rõ hơn về vai trò và lợi ích thực sự của công nghệ trong toàn bộ quy trình làm việc. Sự hướng dẫn của một đối tác lâm sàng sẽ cải thiện cơ hội xây dựng giải pháp phù hợp cho kết quả sức khỏe tốt hơn. Nhiều ứng dụng đầy hứa hẹn nằm ở sự hợp tác giữa nhân viên y tế và công nghệ, kết hợp thế mạnh của cả hai. Điều quan trọng nữa là sử dụng hệ thống GenAI theo cách tôn trọng quyền tự chủ và quyền riêng tư của bệnh nhân.

Trong tương lai gần, có thể cho rằng các mô hình được tùy chỉnh cho các ứng dụng hoặc miền cụ thể sẽ mang lại kết quả tốt hơn và chúng tôi đang theo dõi các xu hướng và bất kỳ sự hội tụ nào về hiệu suất giữa các mô hình chung và cụ thể trong những năm tới. Đối với Med-PaLM nói riêng, tiến trình nghiên cứu của chúng tôi sẽ được theo dõi tại trang web nghiên cứu Med-PaLM. Chúng tôi đặt mục tiêu đạt được tiến bộ rộng rãi hơn trong lĩnh vực sử dụng AI và GenAI để cải thiện cuộc sống của bệnh nhân, bác sĩ lâm sàng và nhà nghiên cứu.

## Summary

Sách trắng này khám phá tiềm năng của LLM trong việc giải quyết những thách thức phức tạp trong các lĩnh vực cụ thể, đặc biệt tập trung vào chăm sóc sức khỏe và an ninh mạng.

- An ninh mạng: Bối cảnh không ngừng thay đổi của các mối đe dọa mạng đòi hỏi các giải pháp sáng tạo. SecLM, một LLM được thiết kế cho an ninh mạng, hoạt động như một lực lượng nhân lên cho các chuyên gia an ninh bằng cách xử lý thông minh lượng dữ liệu khổng lồ. Điều này trao quyền cho họ phân tích và ứng phó với các mối đe dọa hiệu quả hơn. Tầm nhìn của SecLM là tạo ra một nền tảng toàn diện đáp ứng các nhu cầu đa dạng của các chuyên gia an ninh, bất kể trình độ chuyên môn của họ. Sự kết hợp giữa LLM và chuyên môn của con người có tiềm năng cách mạng hóa lĩnh vực an ninh mạng, đạt được kết quả vượt trội với ít nỗ lực hơn.
- Chăm sóc sức khỏe: Dữ liệu chăm sóc sức khỏe ngày càng tăng về số lượng và độ phức tạp, dẫn đến nhu cầu về các giải pháp sáng tạo để làm cho thông tin y tế hữu ích hơn, hữu ích hơn và dễ tiếp cận hơn. MedLM, một nhóm các mô hình được tinh chỉnh cho ngành chăm sóc sức khỏe, có thể giúp mở khóa kiến thức và làm cho y học hiệu quả hơn. MedLM được xây dựng trên Med-PaLM, một LLM được phát triển cho các ứng dụng y tế. Med-PaLM đã chứng minh được hiệu suất ở cấp độ chuyên gia trong các nhiệm vụ hỏi đáp y tế. Thành tựu này chỉ là bước đầu tiên trong hành trình hướng tới việc cải thiện kết quả sức khỏe thông qua việc sử dụng GenAI. Điểm chính rút ra từ nghiên cứu này là công nghệ thôi là chưa đủ. Sự hợp tác với cộng đồng lâm sàng và các đánh giá nhiều bước cẩn thận là rất quan trọng để ứng dụng thành công LLM trong chăm sóc sức khỏe. Trong tương lai, các mô hình cụ thể theo chiều dọc như các mô hình nền tảng MedLM được kỳ vọng sẽ mang lại kết quả tốt hơn nữa cho các ứng dụng cụ thể được quan tâm, thúc đẩy tiềm năng của AI trong chăm sóc sức khỏe.

Sách trắng này giới thiệu các khả năng của LLM trong việc giải quyết các vấn đề cụ thể của từng lĩnh vực. Bằng cách tận dụng sức mạnh của các mô hình tiên tiến này, kết hợp với chuyên môn của con người và việc triển khai cẩn thận, chúng ta có thể giải quyết các thách thức phức tạp và đạt được những tiến bộ đột phá trong nhiều lĩnh vực khác nhau, vì lợi ích của cuộc sống con người.