DầuKhí
KHOA HỌC, CÔNG NGHỆ VÀ ĐỔI MỚI SÁNG TẠO PETROVIETNAM

KHAI THÁC DỮ LIỆU CHUYÊN NGÀNH VỚI SỰ HỖ TRỢ CỦA AI: KINH NGHIỆM TỪ PHẦN MỀM QUẢN LÝ CƠ SỞ DỮ LIỆU ĐƯỜNG CONG ĐỊA VẬT LÝ GIẾNG KHOAN BỂ CỬU LONG

Vũ Tuyết Vy, Nguyễn Trung Sơn
Viện Dầu khí Việt Nam (VPI)
Email: vyvt@vpi.pvn.vn
https://doi.org/10.47800/PVSI.2025.03-04

Tóm tắt

Sự phát triển của trí tuệ nhân tạo (AI), đặc biệt là các mô hình ngôn ngữ lớn (LLMs) đã tạo ra sự thay đổi đáng kể trong quản lý và xử lý dữ liệu chuyên ngành, bao gồm dữ liệu có cấu trúc và phi cấu trúc. Bài viết giới thiệu kinh nghiệm ứng dụng AI trong khai thác dữ liệu tại Viện Dầu khí Việt Nam (VPI), đặc biệt trong hệ thống cơ sở dữ liệu đường cong địa vật lý giếng khoan bể Cửu Long. Đây là sản phẩm được phát triển dựa trên việc kết hợp AI để tối ưu quá trình quản lý, khai thác và phân tích dữ liệu địa vật lý giếng khoan, đảm bảo các yêu cầu về bảo mật.

Kết quả nghiên cứu cho thấy sản phẩm có khả năng tích hợp với các công cụ hỗ trợ khác, góp phần nâng cao trải nghiệm của người dùng; từ đó đề xuất các hướng phát triển trong tương lai nhằm tối ưu hóa hiệu quả và tính bảo mật của hệ thống.

Từ khóa: Trí tuệ nhân tạo, mô hình ngôn ngữ lớn, cơ sở dữ liệu, địa vật lý giếng khoan, bể Cửu Long.

1. Giới thiệu

Trong ngành công nghiệp - năng lượng, AI và học máy không chỉ giúp cải thiện quy trình thăm dò và khai thác mà còn tối ưu hóa việc bảo trì thiết bị, quản lý rủi ro và giảm thiểu chi phí sản xuất thông qua các công nghệ như phân tích dữ liệu thời gian thực và học sâu [1]. Các nghiên cứu cũng cho thấy AI hỗ trợ cải thiện an toàn lao động thông qua các hệ thống dự đoán và đánh giá rủi ro bằng phân tích dữ liệu quy mô lớn [2].

Việc khai thác hiệu quả dữ liệu là "chìa khóa" để đưa ra các quyết định chính xác và tối ưu hóa quy trình. Trong lĩnh vực khí tự nhiên, AI đã chứng minh khả năng phân tích dữ liệu từ các quy trình vận hành để hỗ trợ ra quyết định kịp thời và chính xác, giảm thiểu rủi ro và tối ưu hóa hoạt động [3]. Tuy nhiên, việc áp dụng AI không chỉ yêu cầu nguồn dữ liệu chất lượng cao mà còn đối mặt với các thách thức như an ninh mạng, quản trị dữ liệu và đảm bảo tính minh bạch của các thuật toán [4]. Các hệ thống AI tiên tiến trong ngành công nghiệp năng lượng đang tập trung vào bảo trì dự đoán và giảm thời gian ngừng hoạt động, giúp tiết kiệm chi phí và tăng hiệu quả sản xuất [2].

Trí tuệ nhân tạo mở ra tiềm năng to lớn cho việc phát triển các sản phẩm số thông minh. Việc tích hợp AI vào các sản phẩm này có thể mang lại nhiều lợi ích như cá nhân hóa trải nghiệm người dùng, tự động hóa các tác vụ và dự đoán chính xác. Đặc biệt, AI giúp giảm chi phí sản xuất, tăng năng suất và nâng cao chất lượng sản phẩm. AI không chỉ cải thiện hiệu quả hoạt động mà còn hỗ trợ quản lý năng lượng bền vững và tối ưu hóa sản xuất hydrocarbon thông qua các chiến lược dự đoán và tối ưu hóa dựa trên AI (Gowe AI [1]). Tuy nhiên, các thách thức lớn vẫn tồn tại, bao gồm đào tạo nhân sự chuyên môn cao, đầu tư ban đầu lớn và giải quyết các vấn đề đạo đức liên quan đến dữ liệu và tính minh bạch [3]. Ngoài ra, việc xây dựng các khung quản trị phù hợp cũng là yếu tố then chốt để triển khai AI hiệu quả và bền vững.

2. Ứng dụng AI trong sản phẩm số đặc thù cho lĩnh vực công nghiệp - năng lượng tại Viện Dầu khí Việt Nam (VPI)

2.1. Ứng dụng AI trong khai thác dữ liệu có cấu trúc

Dữ liệu có cấu trúc là dữ liệu được tổ chức thành các bảng với các trường thông tin rõ ràng, dễ dàng phân tích

Ngày nhận bài: 30/12/2024.
Ngày phản biện đánh giá và sửa chữa: 30/12/2024 - 23/1/2025.
Ngày bài báo được duyệt đăng: 23/1/2025.

và truy vấn bằng các công cụ cơ sở dữ liệu truyền thống. Việc ứng dụng trí tuệ nhân tạo và học máy truyền thống đối với dữ liệu có cấu trúc trong lĩnh vực dầu khí diễn ra rất phổ biến. Tại VPI, các ứng dụng này đã sớm được triển khai, đặc biệt phát triển trong vài năm trở lại đây với hàng loạt các sản phẩm số trong tất cả các khâu từ thượng nguồn đến hạ nguồn, chia thành các nhóm chính: phân tích xu hướng tương quan và dự báo, phát hiện bất thường, tối ưu hóa.

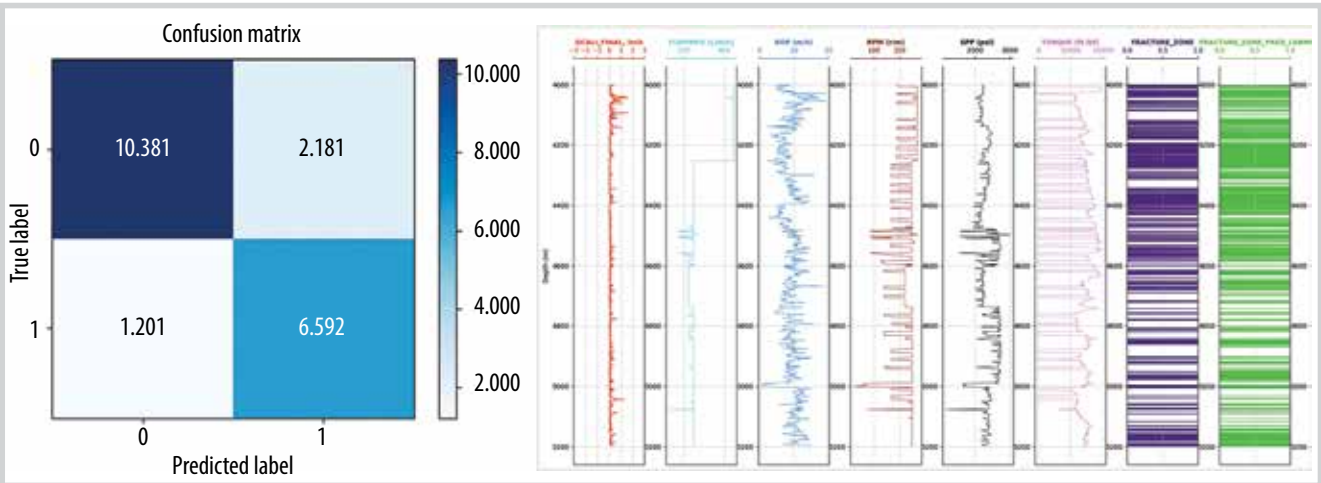
Đối với nghiệp vụ phân tích xu hướng tương quan và dự báo, AI hỗ trợ nhận diện mẫu (pattern recognition), phân tích đa chiều, khai phá, phân loại dữ liệu, dự báo… giúp các nhà nghiên cứu hiểu rõ hơn về các mối quan hệ phức tạp trong dữ liệu. Các sản phẩm tiêu biểu trong nhóm bao gồm: các mô hình kỹ thuật như mô hình dự báo độ thấm của vỉa chứa, mô hình dự báo đường cong địa vật lý giếng khoan, mô hình phân tích mối quan hệ giữa các thông số vỉa chứa…; các mô hình kinh tế như dự báo giá dầu thô Dated Brent, mô hình dự báo trích chi quỹ bình ổn xăng dầu…

Phát hiện bất thường trong dữ liệu có cấu trúc là một ứng dụng quan trọng của AI, giúp nhận diện các giá trị không bình thường hoặc các xu hướng bất thường trong quá trình khai thác dầu khí, được áp dụng trong ứng dụng quản lý khai thác của VPI. AI có thể phát hiện các sự thay đổi đột ngột trong các thông số đo đạc trong quá trình khai thác dầu khí, ví dụ thay đổi đột ngột trong nhiệt độ hoặc áp suất của miệng giếng hoặc đáy giếng, từ đó giúp các kỹ sư nhanh chóng điều chỉnh quá trình khai thác. Bên cạnh đó, các mô hình AI có thể nhận diện các dấu hiệu sớm của các vấn đề tiềm ẩn trong quá trình khai thác, như nguy cơ nổ hoặc rò rỉ dầu, giúp giảm thiểu rủi ro và đảm bảo an toàn cho quá trình vận hành.

Các bài toán tối ưu là một trong những bài toán điển hình của ngành (tối ưu vận hành nhà máy, tối ưu chi phí, tối ưu lựa chọn địa điểm…). Bên cạnh những thuật toán tối ưu truyền thống, việc ứng dụng đồng thời AI sẽ hỗ trợ tìm ra phương án tối ưu nhanh chóng, hiệu quả và tăng tính linh hoạt cho mô hình. Ngoài ra, khả năng học tăng cường, học theo thời gian cũng sẽ giúp mô hình thích nghi với các dữ liệu mới. Một số ứng dụng điển hình phải kể đến tối ưu hóa vị trí khoan, tối ưu hóa kỹ thuật khoan, tối ưu điều phối hệ thống vận chuyển nhiên liệu bằng xe bồn…

Khả năng về ngôn ngữ tự nhiên mạnh mẽ của các mô hình ngôn ngữ lớn đã mở ra hướng tiếp cận mới giúp khai thác dữ liệu có cấu trúc hiệu quả hơn. LLM rút ngắn khoảng cách giữa người dùng và cơ sở dữ liệu có cấu trúc với khả năng chuyển câu lệnh của người dùng thành các câu lệnh truy vấn hệ thống hiểu được (programming language) và diễn giải kết quả truy vấn bằng ngôn ngữ của người dùng. Kỹ thuật này được ứng dụng nhiều trong các sản phẩm số của VPI, chủ yếu trong lĩnh vực hạ nguồn, báo cáo kinh tế thị trường do các dữ liệu quen thuộc, phổ biến và yêu cầu bảo mật thấp. Việc ứng dụng kỹ thuật với dữ liệu tìm kiếm thăm dò khai thác gặp phải nhiều thách thức hơn do đặc thù của dữ liệu và yêu cầu bảo mật dữ liệu cao, cần phải thiết lập nền tảng và công nghệ chặt chẽ hơn, tiêu biểu là thử nghiệm với cơ sở dữ liệu địa vật lý giếng khoan bể Cửu Long (chi tiết trong phần tiếp theo).

Ngoài ra, sự kết hợp của LLM, các thuật toán AI truyền thống và hệ thống tự động hóa tác vụ (robotic process automation - RPA) giúp nhân rộng các ứng dụng đối với dữ liệu có cấu trúc thông qua các tác nhân (agents). Agents có khả năng tự lập kế hoạch, tự hành động và đưa ra quyết định dựa trên bối cảnh và yêu cầu mà người thiết



Hình 1. *Ứng dụng AI/ML trong việc xác định các đới nứt nẻ trong tìm kiếm thăm dò mỏ dầu khí. Nguồn: VPI, 2022.*

lập và người dùng cung cấp. Một dạng agent tiêu biểu đối với dữ liệu có cấu trúc đó là agent xây dựng báo cáo tự động. Sau khi nhận yêu cầu của người dùng (ví dụ như yêu cầu tạo báo cáo liên quan đến một chủ đề nhất định), agent tự động lập kế hoạch từng bước để xây dựng báo cáo, dựa vào nguồn dữ liệu đã được thiết lập sẵn hoặc có quyền truy cập để tự động trích xuất các dữ liệu liên quan, tự động phân tích, tính toán, vẽ biểu đồ và viết báo cáo theo chủ đề được yêu cầu.

2.2. Ứng dụng của AI trong khai thác dữ liệu phi cấu trúc

Sự phát triển của mô hình ngôn ngữ lớn cùng với khả năng kết nối với các thuật toán AI/ML mở ra tiềm năng khai thác dữ liệu phi cấu trúc (dữ liệu không có cấu trúc cố định, khó để sắp xếp vào các bảng cơ sở dữ liệu truyền thống nhưng lại chiếm tỷ trọng lớn). Các dạng dữ liệu phi cấu trúc phổ biến gồm: báo cáo thông tin địa chất, báo cáo mỏ, nhật ký khoan, hình ảnh mẫu lõi, mẫu vụn, mẫu chất lưu… và rất nhiều các loại báo cáo ở định dạng chữ, hình ảnh.

Một trong những ứng dụng phổ biến nhất ứng dụng AI đối với dữ liệu phi cấu trúc là xử lý ngôn ngữ tự nhiên, với đầu vào là tài liệu dạng văn bản. Xử lý ngôn ngữ tự nhiên (natural language processing - NLP) cho phép máy tính hiểu và xử lý ngôn ngữ tự nhiên của con người để thực hiện các tác vụ nâng cao khác như: tóm tắt, trích xuất thông tin, phân loại tài liệu.... Trong lĩnh vực thăm dò khai thác dầu khí, NLP có thể được ứng dụng để phân tích và trích xuất thông tin từ các báo cáo địa chất và nhật ký khoan. Trước khi có sự phát triển vượt bậc của các mô hình ngôn ngữ lớn, các ứng dụng NLP cũng đã xuất hiện tuy nhiên khá khiêm tốn, tập trung vào việc phân tích báo cáo địa chất, tóm tắt nhật ký khoan, dịch thuật tự động… Với sự phát triển của các mô hình NLP hiện nay, có thể tự động phân tích nội dung của các báo cáo địa chất, trích xuất các thông tin quan trọng như đặc điểm cấu trúc địa chất, thành phần khoáng vật và tự đánh giá tiềm năng chứa dầu khí theo những yêu cầu đã được thiết lập sẵn hoặc gợi ý của người sử dụng. Điều này giúp giảm thời gian xử lý dữ liệu và tăng độ chính xác trong việc nhận diện các yếu tố quan trọng. NLP có thể được sử dụng để tự động tóm tắt các ghi chép hàng ngày từ nhật ký khoan, giúp nhanh chóng nắm bắt được các thông tin quan trọng mà không cần đọc toàn bộ tài liệu.

Cùng với sự xuất hiện của ChatGPT và các mô hình ngôn ngữ lớn, việc xử lý ngôn ngữ tự nhiên nói chung và trong lĩnh vực dầu khí nói riêng cũng phổ biến và dễ dàng tiếp cận hơn, với độ chính xác cao hơn và khối lượng dữ liệu được xử lý cũng tăng mạnh. VPI đã tiếp cận hướng đi mới để phát triển các sản phẩm số thử nghiệm cho việc khai thác kho dữ liệu phi cấu trúc của VPI. Các ứng dụng chủ yếu tập trung vào sử dụng khả năng thấu hiểu ngôn ngữ tự nhiên của mô hình ngôn ngữ lớn kết hợp với kỹ thuật xử lý tài liệu theo hình thức vector hóa hoặc ứng dụng bản đồ tri thức (knowledge graph) và các kỹ thuật tìm kiếm nâng cao (retrieval-augmented generation - RAG) để xây dựng cơ sở dữ liệu tài liệu phi cấu trúc cùng với hệ thống chatbot để tìm kiếm, hỏi đáp các thông tin liên quan. Ngoài ra, LLM cũng hỗ trợ đắc lực trong việc tự động hóa tìm kiếm thông tin trên internet theo chủ đề xác định trước, tổng hợp và phân loại thông tin.

Sản phẩm thử nghiệm tiêu biểu phải kể đến việc ứng dụng LLM để phân tích các báo cáo địa chất và nhận diện các yếu tố quan trọng ảnh hưởng đến khả năng chứa dầu khí của bể Sông Hồng. Kết quả cho thấy, hệ thống có tích hợp LLM đã giảm được 20 - 30% thời gian xử lý dữ liệu và tăng độ chính xác trong việc phân loại các đặc điểm địa chất so với phương pháp thủ công, từ đó giúp việc tổng hợp và đánh giá thông tin trở nên dễ dàng và hiệu quả hơn.

Bên cạnh xử lý ngôn ngữ tự nhiên, việc ứng dụng thị giác máy tính (computer vision) cũng phổ biến trong lĩnh vực dầu khí nhằm phân tích hình ảnh mẫu lõi, tự động nhận dạng các đặc điểm địa chất như các tầng trầm tích khác nhau, vết nứt hoặc dấu hiệu của chất lưu trong mẫu lõi. Trong phân loại hình ảnh mẫu lõi, bằng việc sử dụng các neural network tích chập (CNN), hệ thống có thể tự động phân loại các loại đá và xác định các đặc điểm địa chất quan trọng từ hình ảnh mẫu lõi. Đồng thời có thể trợ giúp xác định những loại cổ sinh trong lát mỏng. Điều này không chỉ giúp tiết kiệm thời gian mà còn nâng cao độ chính xác trong việc phân tích mẫu lõi [3]. Đối với nhận diện đặc điểm các yếu tố địa chất, thị giác máy tính có thể phát hiện các dấu hiệu bất thường hoặc các đặc điểm địa chất quan trọng mà con người có thể bỏ sót như các hệ thống đứt gãy trên mặt cắt địa chấn, từ đó hỗ trợ quá trình thăm dò khai thác hiệu quả hơn.

3. Kinh nghiệm ứng dụng AI trong phần mềm quản lý cơ sở dữ liệu địa vật lý giếng khoan bể Cửu Long

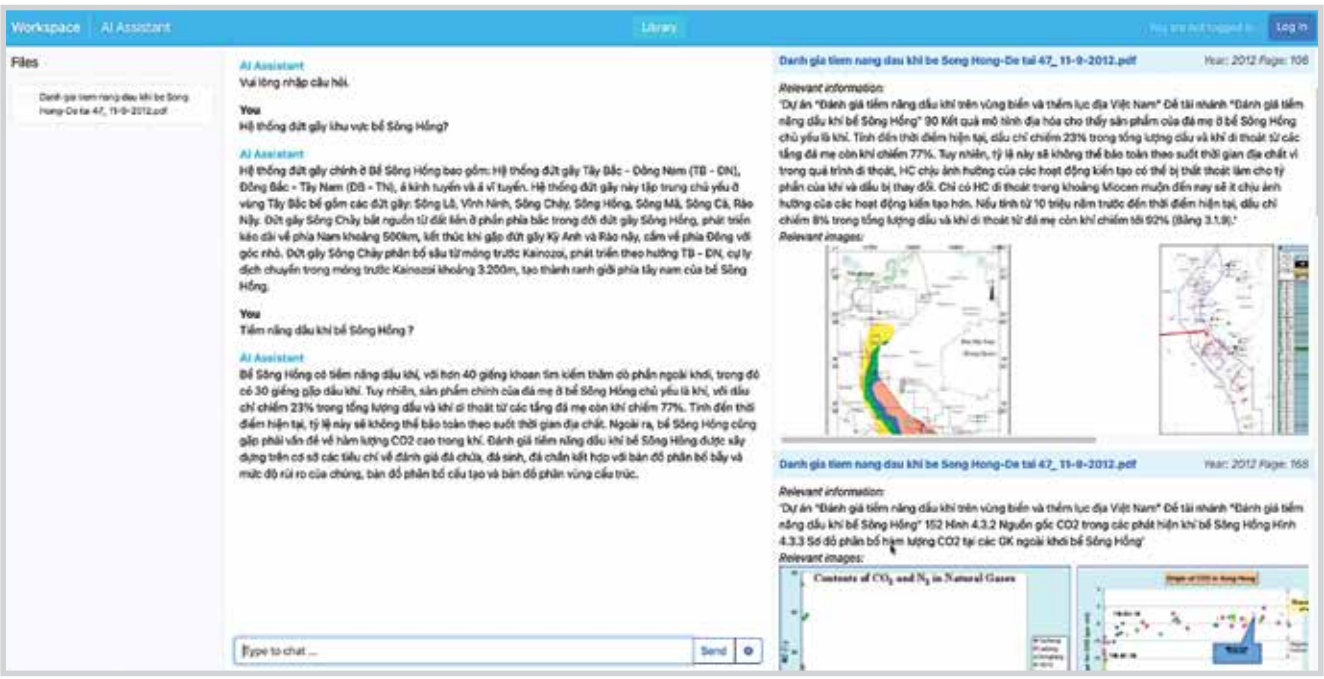
3.1. Phần mềm quản lý cơ sở dữ liệu địa vật lý giếng khoan bể Cửu Long

Cơ sở dữ liệu địa vật lý giếng khoan bể Cửu Long chứa lượng thông tin rất lớn về đặc điểm địa chất của các giếng khoan, bao gồm dữ liệu đo ghi địa vật lý giếng khoan, mẫu lõi và các kết quả phân tích (Hình 3). Việc ứng dụng AI vào quản lý và khai thác cơ sở dữ liệu này mang lại những lợi

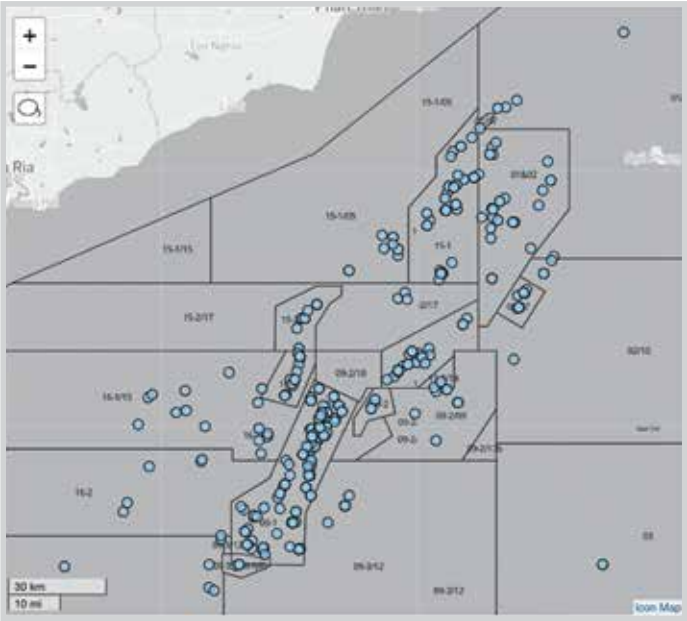
ích to lớn, như tăng cường độ chính xác trong phân tích và dự báo tiềm năng dầu khí [4]. Phần mềm quản lý cơ sở dữ liệu địa vật lý giếng khoan được xây dựng nhằm cung cấp một giải pháp toàn diện cho việc quản lý, khai thác và phân tích dữ liệu địa vật lý giếng khoan. Phần mềm được thiết kế với giao diện Power BI thân thiện, dễ sử dụng, đáp ứng các yêu cầu về an toàn bảo mật. Điểm nổi bật của sản phẩm này so với các sản phẩm truyền thống là khả năng tích hợp AI, cho phép người dùng tương tác với dữ liệu bằng ngôn ngữ tự nhiên và tự động hóa nhiều tác vụ, chạy trên máy tính nội bộ.

Giao diện người dùng đáp ứng nhu cầu trực quan và các yêu cầu kỹ thuật chuyên môn, giúp người dùng dễ truy cập, phân tích và khai thác dữ liệu hiệu quả. Việc xây dựng giao diện trên nền tảng Power BI với khả năng tương tác động mang lại nhiều lợi thế vượt trội. Power BI cho phép người dùng tương tác trực tiếp với dữ liệu thông qua các biểu đồ, bảng và bản đồ. Người dùng có thể dễ dàng sàng lọc dữ liệu, "đào sâu" vào chi tiết và thay đổi cách hiển thị dữ liệu phù hợp với nhu cầu phân tích. Power BI có thể kết nối với nhiều nguồn dữ liệu khác nhau, bao gồm cơ sở dữ liệu, tệp Excel và các dịch vụ đám mây. Điều này cho phép phần mềm dễ tích hợp với cơ sở dữ liệu địa vật lý giếng khoan và các hệ thống khác (Hình 4).

Địa vật lý giếng khoan đóng vai trò quan trọng trong việc thu thập, quản lý và phân tích dữ liệu liên quan đến giếng khoan (ví dụ: logs, khảo sát địa chấn xung quanh giếng...). Để đảm bảo dữ liệu nhất quán, dễ truy cập, dễ tích hợp và chia sẻ, VPI đã áp dụng mô hình dữ liệu PPDM (professional petroleum data management) để chuẩn hóa và quản lý dữ liệu địa vật lý giếng khoan. Mô hình dữ liệu (data model) được xây dựng chặt chẽ, liên kết vùng thông tin chung của giếng với các vùng thông tin khác: Tầng đánh dấu (marker), quỹ đạo giếng khoan (deviation), đường cong địa vật lý giếng khoan (logs), đồng thời duy trì tính nhất quán về thông tin giếng khoan xuyên suốt các vùng dữ liệu. Các khái niệm và đơn vị đo đã được chuẩn hóa và thống nhất trong từ điển dữ liệu chung (data dictionary) để tránh nhầm lẫn và dễ dàng chuyển giao, kế thừa trong các sản phẩm số khác. Bên cạnh đó, sản phẩm cơ sở dữ liệu địa vật lý giếng khoan bể Cửu Long được kế thừa dữ liệu đã qua kiểm soát chất lượng (QC) và chuẩn hóa bởi nhóm chuyên gia lĩnh vực địa vật lý giếng khoan của VPI. Vấn đề bảo mật và phân quyền dữ liệu cũng được đặt lên hàng đầu nhằm đảm bảo việc quản trị thông tin minh bạch, hiệu quả và đáp ứng tốt nhu cầu khác nhau của người dùng trong lĩnh vực dầu khí. Do đó cấu trúc cơ sở dữ liệu địa vật lý giếng khoan bể Cửu Long có khả năng mở rộng dữ liệu áp dụng cho các bể trầm tích khác, như bể Nam Côn Sơn, hay bể Sông Hồng, nơi cũng có nhiều giếng khoan với lượng dữ liệu đồ sộ, với thiết kế tổng quát và linh hoạt. Đồng thời, hệ thống đồng bộ sẽ hỗ trợ so sánh và phân tích dữ liệu hiệu quả giữa các trường thông tin khác nhau, ví dụ như thông tin khai thác và thông tin thăm dò trong cùng một khu vực hay các khu vực khác nhau, góp phần nâng cao khả năng quản lý và ra quyết



Hình 2. *Chatbot hỏi đáp báo cáo chuyên ngành. Nguồn: VPI, 2024.*



Hình 3. *Vị trí thông tin giếng khoan trong bể Cửu Long. Nguồn: VPI, 2024.*

Hình 4. *Giao diện ứng dụng quản lý cơ sở dữ liệu giếng khoan. Nguồn: VPI, 2024.*

định. Trên thực tế, nhóm tác giả đã và đang ứng dụng công nghệ và kỹ thuật xây dựng cơ sở dữ liệu của bể Cửu Long cho bể Nam Côn Sơn mà không cần thay đổi về cấu trúc dữ liệu.

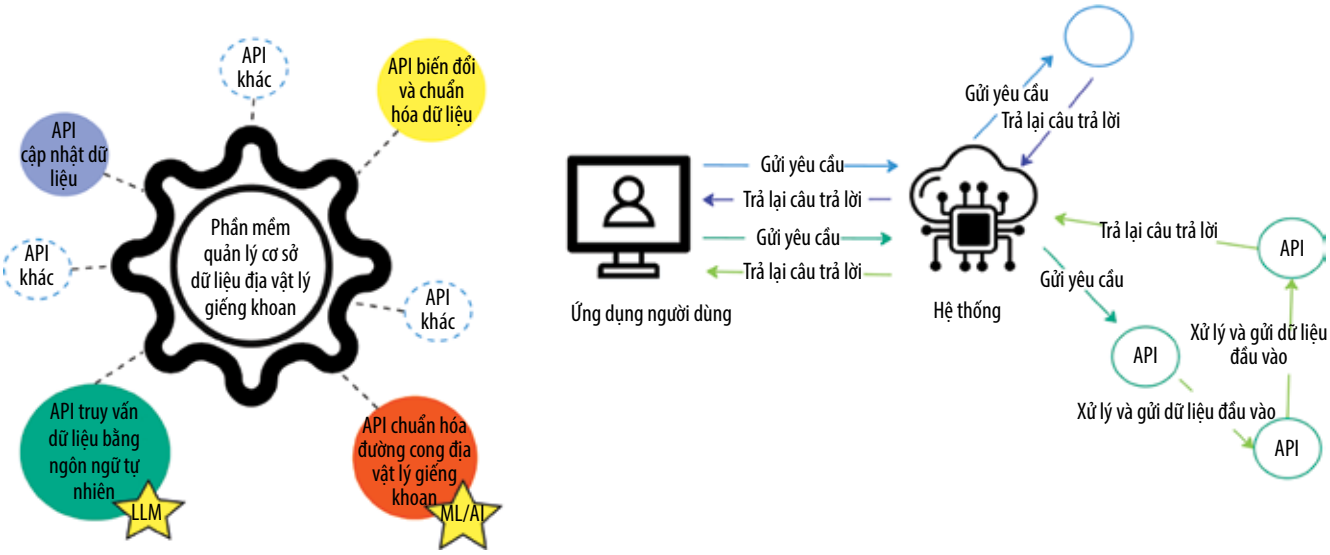
Với đặc thù của dữ liệu lĩnh vực tìm kiếm thăm dò, nền tảng yêu cầu truy cập an toàn, bảo mật và phân quyền chặt chẽ. Cơ sở dữ liệu tại chỗ (on-premise) kết hợp với hệ thống phân quyền chi tiết là giải pháp toàn diện mà nhóm tác giả đã triển khai để đảm bảo an toàn và kiểm soát dữ liệu. Dữ liệu được lưu trữ trên máy chủ nội bộ thay vì trên đám mây, từ đó giảm thiểu nguy cơ bị tấn công mạng hoặc mất mát dữ liệu do các lỗ hổng trên môi trường trực tuyến. Người dùng được giới hạn truy cập trực tiếp từ hạ tầng nội bộ theo phân quyền truy cập dựa trên các vai trò cụ thể, và có thể giới hạn phạm vi dữ liệu được truy cập (như theo giếng, mỏ, lô, bể…).

Một trong những điểm nổi bật trong thiết kế hệ thống của nền tảng quản lý cơ sở dữ liệu giếng khoan là khả năng tích hợp với các công cụ chuyên môn khác thông qua mảnh ghép số (API). Với định hướng của VPI trong việc phát triển hình thức mảnh ghép số và tích hợp, tái sử dụng trong các sản phẩm số khác nhau, nhóm tác giả lựa chọn nền tảng triển khai sản phẩm dễ dàng mở rộng trong tương lai, với định hướng nền tảng là nơi tập trung công cụ duy nhất hỗ trợ nhóm chuyên môn trong việc khai thác dữ liệu địa vật lý giếng khoan. Tính đến nay sản phẩm đã tích hợp thêm 2 mảnh ghép số khác từ các tác giả bộ phận chuyên môn và 1 mảnh ghép số hỗ trợ khai thác dữ liệu bằng ngôn ngữ tự nhiên phát triển bởi bộ phận Phân tích dữ liệu: API dự báo đường cong địa vật lý giếng khoan bị thiếu, API hiệu chỉnh đường cong địa vật lý giếng khoan, API chatbot hỏi đáp về dữ liệu.

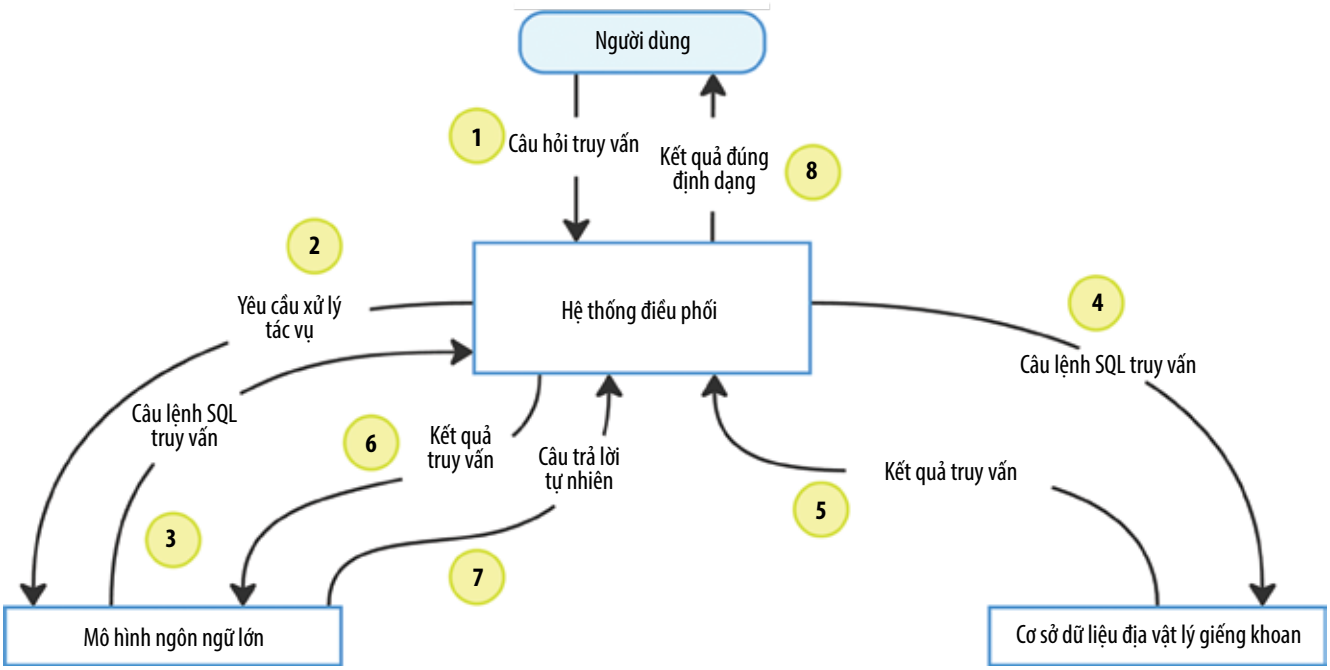
3.2. Kinh nghiệm ứng dụng AI trong phần mềm quản lý cơ sở dữ liệu địa vật lý giếng khoan bể Cửu Long

Khả năng hỏi đáp bằng ngôn ngữ tự nhiên là tính năng đột phá của ứng dụng, giúp người dùng tương tác với dữ liệu dễ dàng và tự nhiên hơn. Thay vì phải sử dụng các câu lệnh phức tạp hoặc truy vấn SQL, người dùng có thể đặt câu hỏi bằng ngôn ngữ hàng ngày và nhận được câu trả lời chính xác từ cơ sở dữ liệu. Nguyên lý hoạt động của hệ thống hỏi đáp như sau: Câu hỏi của người dùng sẽ được hệ thống gửi đến mô hình LLM để phân tích ngữ nghĩa và cú pháp của câu hỏi để hiểu rõ nội dung yêu cầu. Sau đó, mô hình sẽ chuyển đổi những yêu cầu được nhận biết, cùng với những hiểu biết về cơ sở dữ liệu, thành một câu lệnh truy vấn SQL mà hệ thống có thể thực hiện trên cơ sở dữ liệu. Câu lệnh truy vấn này sẽ được hệ thống gửi tới cơ sở dữ liệu để thực thi câu lệnh trên cơ sở dữ liệu và trả lại dữ liệu liên quan. Kết quả truy vấn (dạng dữ liệu, bảng biểu) một lần nữa được gửi qua LLM để chuyển thành ngôn ngữ tự nhiên, được trình bày dưới dạng văn bản, bảng biểu hoặc biểu đồ, giúp người dùng dễ dàng tiếp nhận thông tin (Hình 6).

Thách thức lớn nhất trong việc áp dụng mô hình ngôn ngữ lớn với cơ sở dữ liệu địa vật lý giếng khoan là yêu cầu bảo mật. Các mô hình có chất lượng tốt, có khả năng thấu hiểu người dùng và tạo ra câu lệnh truy vấn sát nhất đều sử dụng công nghệ đám mây, ngược lại, các mô hình có khả năng triển khai trên máy nội bộ, đảm bảo bảo mật lại cho chất lượng truy vấn thấp hơn, thậm chí trong một số trường hợp không hiểu được nhu cầu người dùng, hoặc diễn đạt không chính xác ý nghĩa của kết quả truy vấn. Để giải quyết vấn đề này, nhóm tác giả đã thiết kế và triển



Hình 5. *Ứng dụng cơ sở dữ liệu kết nối với các API khác. Nguồn: VPI, 2024.*

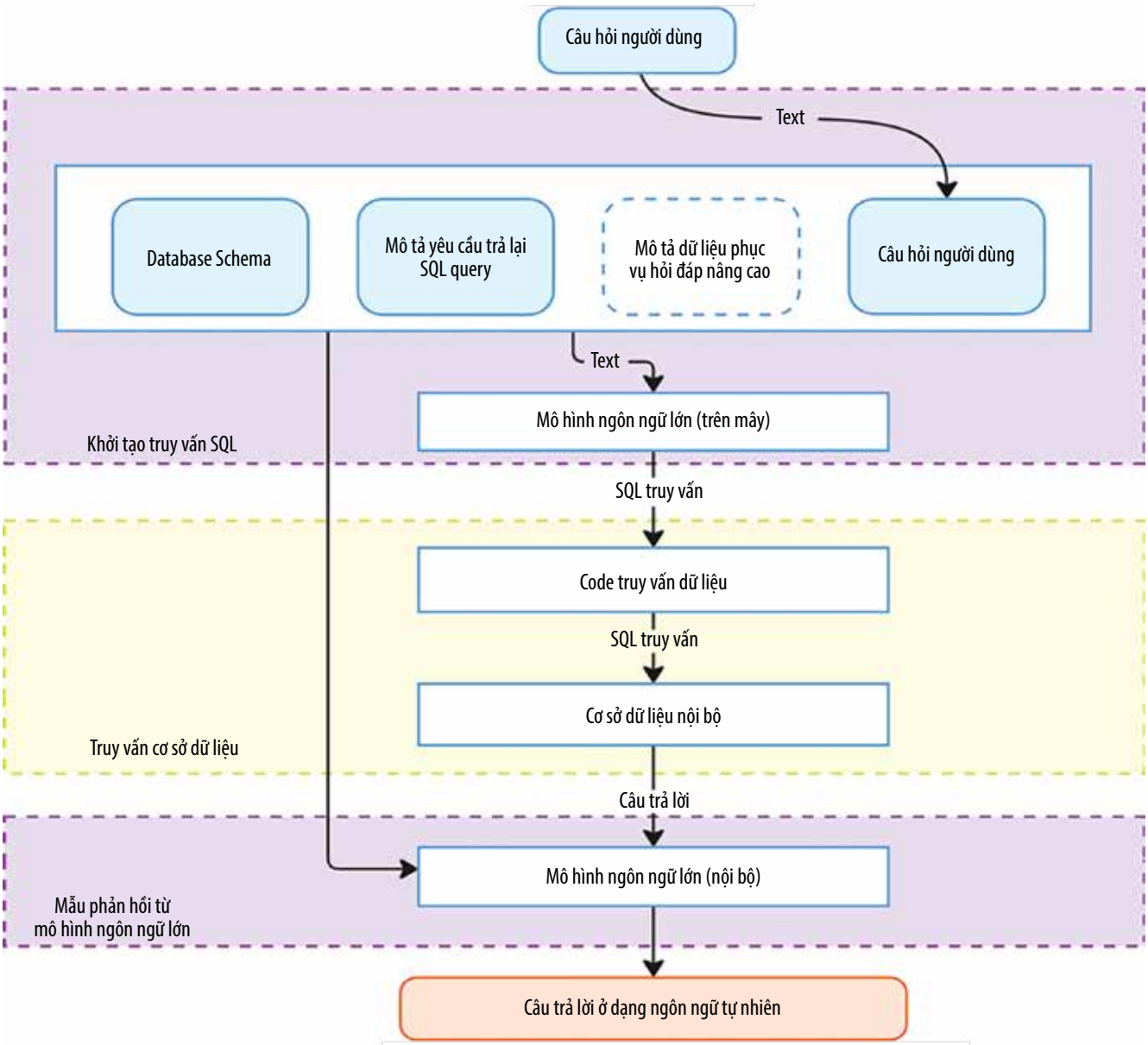


Hình 6. *Quy trình hoạt động sử dụng mô hình ngôn ngữ lớn hỏi đáp với cơ sở dữ liệu. Nguồn: VPI, 2024.*

khai phương án kết hợp giữa mô hình AI sử dụng cloud (mây) và mô hình trên máy nội bộ để đảm bảo chất lượng tốt và AI local để bảo mật thông tin.

Chiến lược kết hợp giữa mô hình AI sử dụng cloud và mô hình trên máy nội bộ được triển khai như sau: Đối với tác vụ tạo ra câu hỏi truy vấn, do đây là tác vụ cần chất lượng mô hình tốt, nhưng không yêu cầu đưa dữ liệu trực tiếp vào mô hình, nhóm tác giả sử dụng AI trên nền tảng đám mây của OpenAI, kết hợp với kỹ thuật xây dựng câu lệnh (prompt engineering). Kết quả cho thấy câu lệnh truy vấn rất tốt và sát với nhu cầu người dùng. Đối với tác vụ minh giải kết quả truy vấn, do tác vụ yêu cầu đưa dữ liệu trực tiếp vào mô hình để diễn giải, nhóm tác giả sử dụng mô hình LLM trên máy nội bộ với kích thước và số lượng tham số vừa đủ để đảm bảo hiệu suất hoạt động và bảo mật dữ liệu. Vấn đề về chất lượng diễn giải câu trả lời cũng được khắc phục thông qua kỹ thuật xử lý câu lệnh (Hình 7).

Câu lệnh (prompt) là đoạn văn bản hoặc thông tin đầu vào cung cấp cho mô hình. Vai trò của nó là "gợi ý" hoặc định hướng cho mô hình biết nội dung cần triển khai, các yêu cầu cụ thể, hoặc ngữ cảnh liên quan đến vấn đề. Do đó, việc đưa ra câu lệnh cụ thể và phản ánh rõ được nhu cầu của người dùng rất quan trọng và ảnh hưởng trực tiếp đến chất lượng câu trả lời của mô hình. Prompt engineering (kỹ thuật xây dựng câu lệnh) là quá trình thiết kế, tinh chỉnh và tối ưu prompt để có được



Hình 7. *Quy trình xử lý câu hỏi truy vấn dữ liệu kết hợp giữa AI cloud (trên mây) và AI local (nội bộ). Nguồn: VPI, 2024.*

câu trả lời mong muốn từ mô hình ngôn ngữ. Các tác vụ prompt engineering thường sẽ bao gồm lựa chọn ngôn từ, cú pháp, cung cấp ngữ cảnh chi tiết, quy ước về các tác vụ, quy ước về định dạng đầu ra, hoặc các tác vụ nâng cao khác. Việc lựa chọn câu lệnh phù hợp đòi hỏi sự tinh chỉnh, thử nghiệm lặp lại và sự "thấu hiểu" từng mô hình ngôn ngữ sử dụng của người xây dựng câu lệnh. Trong trường hợp này, do đặc thù dữ liệu chuyên môn nên việc để mô hình tự hiểu về cấu trúc cũng như ý nghĩa của dữ liệu trong cơ sở dữ liệu là không khả thi. Do đó, đối với tác vụ tạo câu lệnh truy vấn SQL, nhóm tác giả đã cung cấp cho mô hình chi tiết về cấu trúc của cơ sở dữ liệu (database schema không bao gồm dữ liệu) để đảm bảo mô hình biết được phạm vi của cơ sở dữ liệu và mối quan hệ giữa các bảng dữ liệu trong cơ sở dữ liệu. Bên cạnh đó, nhóm tác giả đã tận dụng từ điển dữ liệu đã được khai báo trước đó trong cơ sở dữ liệu (theo quy định về dữ liệu số của VPI) để mô hình ngôn ngữ hiểu về ý nghĩa của các bảng và các trường dữ liệu, đồng thời có thông tin về loại dữ liệu, tần suất, nguồn dữ liệu… từ đó đưa ra câu lệnh truy vấn chính xác hơn. Đối với tác vụ diễn giải câu trả lời, bên cạnh dữ liệu kết quả, trong câu lệnh gửi tới mô hình trên máy nội bộ, nhóm tác giả đã đưa kèm câu hỏi của người dùng và toàn bộ thông tin của cơ sở dữ liệu, cùng với chỉ dẫn cụ thể cho các trường hợp thường xuyên được hỏi liên quan đến chuyên môn để mô hình đưa ra câu trả lời sát nhất với nhu cầu và tuân thủ định dạng yêu cầu.

Ví dụ, khi thống kê giếng khoan theo lô (Hình 8), câu hỏi của người dùng được đưa vào hệ thống và gửi cùng với câu lệnh để đưa vào LLM trên đám mây để phân tích

câu hỏi và tạo ra câu lệnh truy vấn kèm theo định dạng hiển thị mong muốn. Trong trường hợp này, nội dung của LLM tạo ra như Hình 8.

Bên cạnh câu lệnh SQL phù hợp với cấu trúc dữ liệu trong cơ sở dữ liệu, LLM tự định dạng loại đồ thị phù hợp với phương thức thống kê này là đồ thị hình cột, đồng thời tự động đặt tên và mô tả cho hiển thị. Các thông tin này sau đó được bóc tách câu lệnh để gửi vào hệ thống cơ sở dữ liệu và nhận lại dữ liệu tương ứng, sau đó dữ liệu này dùng phần định nghĩa về hình vẽ và các mô tả khác sẽ được gửi tiếp vào hệ LLM trên máy nội bộ để đưa ra kết quả hiển thị cuối cùng ở dạng câu trả lời và hình vẽ như Hình 9.

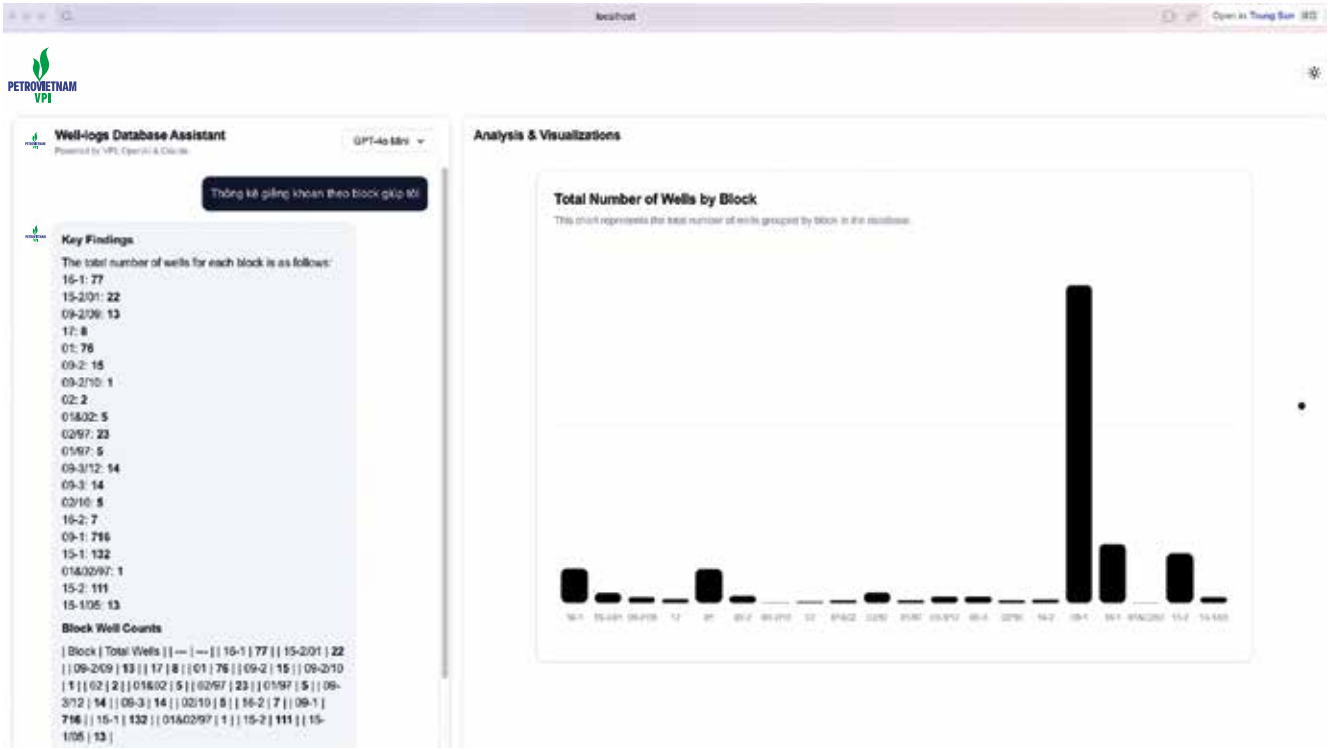
Kinh nghiệm ứng dụng kết hợp giữa AI trên máy nội bộ và trên mây có thể ứng dụng được trong rất nhiều trường hợp. Việc phân tách dựa trên yêu cầu bảo mật dữ liệu là nguyên tắc có thể áp dụng cho các dữ liệu chuyên ngành cần yêu cầu bảo mật cao nhưng vẫn đòi hỏi chất lượng truy vấn chính xác. Trong một số trường hợp, thay vì sử dụng mô hình trên máy nội bộ, người dùng có thể cân nhắc sử dụng AI trên mây để tạo ra mẫu câu trả lời và ghép dữ liệu một cách cơ học vào câu trả lời đưa tới người dùng. Việc này cũng đảm bảo dữ liệu không bị rò rỉ trên mây thông qua mô hình. Ngoài ra, việc phân tách cũng có thể dựa trên nguyên tắc yêu cầu về chất lượng và chi phí. Ví dụ những câu hỏi đơn giản hoặc yêu cầu độ chính xác không quá cao có thể xử lý trên máy nội bộ để tiết kiệm chi phí, những câu hỏi phức tạp hơn, yêu cầu chất lượng cao hơn sẽ được xử lý với các mô hình trên đám mây.

```
{

Chart Type: bar

Table Name: wellinfo

Title: Total Number of Wells by Block

Description: This chart represents the total number of wells grouped by block in
the database.

SQL Query:

SELECT COUNT("WELL_UWI") AS "count", "BLOCK"

FROM "wellinfo"

GROUP BY "BLOCK";

}
```

Hình 8. *Mẫu câu lệnh truy vấn cơ sở dữ liệu tạo ra từ mô hình ngôn ngữ lớn.*

4. Kết luận

Phần mềm quản lý cơ sở dữ liệu địa vật lý giếng khoan bể Cửu Long có tính ứng dụng cao và phù hợp với nhiều sản phẩm số tương tự, đặc biệt là các sản phẩm yêu cầu bảo mật dữ liệu. Với tốc độ phát triển công nghệ, chất lượng phản hồi của chatbot đối với dữ liệu chuyên môn sẽ ngày càng được cải thiện, ngay cả khi sử dụng các mô hình trên máy nội bộ. Điều này có thể đạt được thông qua việc đào tạo và tinh chỉnh các mô hình chuyên biệt dựa trên dữ liệu đầu vào từ các lĩnh vực chuyên ngành như: mô hình embedding



Hình 9. *Giao diện hỏi đáp với cơ sở dữ liệu địa vật lý giếng khoan bằng ngôn ngữ tự nhiên. Nguồn: VPI, 2024.*

(xử lý tài liệu đầu vào và xử lý câu hỏi người dùng), hay mô hình dịch thuật tài liệu chuyên ngành…

Đối với các mô hình AI/ML dựa trên dữ liệu có cấu trúc, việc thiết lập hệ thống chỉ số đo lường tương đối đơn giản và có tính nhất quán cao. Tuy nhiên, đối với các ứng dụng sử dụng ngôn ngữ tự nhiên như chatbot, việc đo lường chính xác câu trả lời trở nên khó khăn hơn, đặc biệt là do mô hình có thể thiếu hiểu biết về vấn đề, hoặc gặp phải hiện tượng ảo giác (hallucination). Để cải thiện, cần có sự can thiệp chuyên môn và hệ thống bộ câu hỏi thử nghiệm tiêu chuẩn để đánh giá hiệu quả. Ngoài ra, việc cập nhật tri thức và nâng cao chất lượng câu trả lời cần phải được cải tiến liên tục theo thời gian. Ưu điểm rõ rệt nhất là việc giảm thiểu thời gian người dùng tiếp cận và khai thác dữ liệu, nâng cao trải nghiệm người dùng, xóa bỏ rào cản công nghệ, giúp người dùng dễ dàng hơn trong việc sử dụng và tương tác với hệ thống.

Tài liệu tham khảo

[1] Ganesh Shank Gowekar, "How oil and gas industry are transforming with AI and ML", *World Journal of Advanced Research and Reviews*, Volume 23, Issue 3, pp. 1234 - 1238, 2024. DOI: 10.30574/wjarr.2024.23.3.2722.

[2] Chuka Anthony Arinze, Vincent Onuegbu Izionworu, Daniel Isong, Dominic Daudu, and Adedayo Adefemi "Integrating artificial intelligence into engineering processes for improved efficiency and safety in oil and gas operations", *Open Access Research Journal of Engineering and Technology*, Volume 6, Issue 1, pp. 39 - 51, 2024. DOI: 10.53022/oarjet.2024.6.1.0012.

[3] Siddhartha Nuthakki, Chinmay Shripad Kulkarni, Suraj Kumar, Satish Kathiriya, and Yudhisthir Nuthakki, "Artificial intelligence applications in natural gas industry: A literature review", *International Journal of Engineering and Advanced Technology*, Volume 13, Issue 3, 2024. DOI: 10.35940/ijeat.C4383.13030224.

[4] Muoi Duy Nguyen, Hoa Minh Nguyen, and Ngan Thi Bui, "Application of artificial neural network and seismic attributes to predict the distribution of Late Oligocene sandstones in the Cuu Long basin", *Journal of Mining and Earth Sciences*, Volume 64, Issue 3, pp: 24 - 31, 2023. DOI: 10.46326/jmes.2023.64(3).03.

[5] I.K. Turgazinov and T.M. Nassigazin, "Application of artificial intelligence for construction of well IPR", *Neft'i Gaz*, Volume 6, Issue 13, pp. 136 - 150, 2023. DOI: 10.37878/2708-0080/2023-6.13.

[6] Vivian Ofure Eghaghe, Olajide Soji Osundare, Chikezie Paul-Mikki Ewim, and Ifeanyi Chukwunonso Okeke, "Navigating the ethical and governance challenges of AI deployment in AML practices within the financial industry", *International Journal of Scholarly Research and Reviews*, Volume 5, Issue 2, pp. 30 - 51, 2024. DOI: 10.56781/ijsrr.2024.5.2.0047.

[7] PPDM Association, *"PPDM data model version 3.2. PPDM Standard, 3.2.1"*, 2020.

AI-INTEGRATED DOMAIN-SPECIFIC DATA MANAGEMENT: EXPERIENCE FROM DEVELOPING WELL-LOG DATABASE MANAGEMENT SOFTWARE IN THE CUU LONG BASIN

Vu Tuyet Vy, Nguyen Trung Son
Vietnam Petroleum Institute (VPI)
Email: vyvt@vpi.pvn.vn

Summary

The rapid advancement of artificial intelligence (AI), particularly Large Language Models, has significantly transformed data management and processing practices for both structured and unstructured data. This paper presents the experience of designing, implementing, and operating an AI-integrated geophysical well log database software for the Cuu Long basin at the Vietnam Petroleum Institute (VPI). The software was developed by leveraging AI to optimize the management, retrieval, and analysis of well logging data while ensuring strict security requirements.

The research results show that the product can be integrated with other supporting tools, contributing to an enhanced user experience. Based on these findings, future development directions are proposed to optimize the system's efficiency and security.

Key words: Artificial intelligence, large language models, database, well log geophysics, Cuu Long basin.