CẨM NANG THUẬT NGỮ AI: TỪ NỀN TẢNG ĐẾN HỆ THỐNG AI AGENTS TỰ CHỦ

CHUYÊN ĐỀ 1: TỔNG QUAN VỀ TRÍ TUỆ NHÂN TẠO VÀ MÔ HÌNH NGÔN NGỮ LỚN

Trong kỷ nguyên chuyển đổi số toàn diện, việc nắm bắt chính xác bản chất kiến trúc của Trí tuệ Nhân tạo (AI) không còn dừng lại ở mức độ cập nhật thông tin kỹ thuật đơn thuần, mà đã trở thành năng lực định hình chiến lược cạnh tranh cốt lõi cho mọi doanh nghiệp. Sự bùng nổ của các mô hình trí tuệ nhân tạo hiện đại đang tái định hình cách thức vận hành thông tin, xử lý quy trình và ra quyết định kinh doanh. Nếu tiếp cận AI một cách cảm tính hoặc coi AI là một "chiếc hộp đen" vạn năng, doanh nghiệp rất dễ rơi vào bẫy kỳ vọng phi thực tế, triển khai sai giải pháp cho bài toán kinh doanh, dẫn đến lãng phí ngân sách và suy giảm hiệu năng hệ thống. Việc thấu hiểu tường tận bản chất toán học, các cột mốc tiến hóa kỹ thuật, cùng những hạn chế nội tại của Mô hình Ngôn ngữ Lớn (LLM) sẽ giúp các nhà quản trị và kiến trúc sư hệ thống xây dựng một tư duy nền tảng vững chắc, từ đó lựa chọn đúng công nghệ, tối ưu hóa chi phí vận hành và làm chủ các hệ thống AI tự chủ trong thực tiễn.

Bài 1: AI thực chất là gì? Hiểu AI trong 10 phút

(1) Giải thích khái niệm

Trí tuệ Nhân tạo (Artificial Intelligence - AI) dưới góc nhìn kiến trúc hệ thống là một tập hợp các mô hình toán học và thuật toán máy tính được thiết kế để tiếp nhận dữ liệu đầu vào (Inputs/Context), xử lý không gian trạng thái thông qua các cơ chế suy luận xác suất, và sinh ra dữ liệu đầu ra (Outputs) nhằm đạt được các mục tiêu định trước. Bản chất của AI hiện đại không phải là một bộ lưu trữ tri thức tĩnh hay một thực thể "hiểu" ngữ nghĩa như con người, mà là một bộ máy xử lý thông tin động (Information Processing Engine). Hệ thống AI tiếp nhận các tín hiệu đầu vào dưới dạng văn bản, hình ảnh, hoặc chỉ số định lượng, phân tích mối quan hệ phụ thuộc giữa các biến số thông qua các mạng thần kinh nhân tạo (Neural Networks), và tính toán phân phối xác suất để tạo ra phản hồi tối ưu nhất dựa trên tập dữ liệu đã được học.

(2) Ứng dụng khi nào

AI được triển khai hiệu quả nhất trong các bài toán doanh nghiệp cần tự động hóa quy trình xử lý thông tin phi cấu trúc quy mô lớn hoặc hỗ trợ gia tăng tốc độ ra quyết định. Bối cảnh áp dụng tối ưu bao gồm: phân loại văn bản tự động, tổng hợp dữ liệu đa nguồn, phân tích cảm xúc phản hồi của khách hàng, và trích xuất tri thức từ tài liệu nghiệp vụ.

- **Đánh đổi chiến lược (Trade-offs)** : Triển khai AI mang lại khả năng mở rộng quy mô xử lý không giới hạn (Scalability) với chi phí biên thấp, nhưng doanh nghiệp phải chấp nhận đánh đổi về độ trễ xử lý (Latency) ban đầu và sự xuất hiện của độ rủi ro tính ngẫu nhiên (Non-deterministic behavior).
- **Chiến lược xử lý lỗi (Fallback)** : Khi độ tin cậy của mô hình nằm dưới ngưỡng an toàn (ví dụ: confidence score < 85%), hệ thống bắt buộc phải kích hoạt cơ chế chuyển giao tác vụ cho con người xử lý (Human-in-the-loop).

(3) Ví dụ minh họa

Trong một quy trình xử lý hồ sơ bồi thường bảo hiểm, thay vì nhân viên pháp chế phải đọc thủ công tài liệu dài 30 trang để xác minh các điều khoản loại trừ, hệ thống AI tiếp nhận toàn bộ file PDF làm dữ liệu đầu vào. Mô hình phân tích các mối quan hệ ngữ nghĩa, đối chiếu các điều khoản bồi thường với dữ liệu lịch sử vụ việc, và trích xuất chính xác 3 căn cứ pháp lý cốt lõi kèm theo chỉ số tin cậy 92% dưới dạng cấu trúc JSON chuẩn hóa chỉ trong 2 giây, giúp nhân viên rút ngắn 90% thời gian phê duyệt hồ sơ.

Bài 2: Các nấc thang phát triển của AI

(1) Giải thích khái niệm

Sự phát triển của AI thể hiện sự chuyển dịch kiến trúc qua ba nấc thang năng lực chính:

1. **Rule-based Systems (Hệ thống dựa trên quy tắc cố định)** : Vận hành dựa trên các tập lệnh If-Else logic được con người lập trình cứng. Hệ thống hoàn toàn thiếu tính linh hoạt và gãy đổ ngay khi gặp dữ liệu nằm ngoài kịch bản.
2. **Prediction AI / Machine Learning (AI dự báo và phân loại)** : Vận hành dựa trên các thuật toán học máy thống kê (mạng CNN, RNN, Gradient Boosting). Hệ thống có khả năng phân loại dữ liệu, tính toán xác suất và dự báo xu hướng dựa trên dữ liệu lịch sử, nhưng không có khả năng sinh ra nội dung mới.
3. **Generative & Agentic AI (AI tạo sinh và AI tự chủ)** : Vận hành trên kiến trúc Transformer tiên tiến, có khả năng hiểu ngữ cảnh sâu sắc, sinh ra văn bản/hình ảnh mới từ ngữ cảnh thô, và tự chủ lập kế hoạch, sử dụng công cụ để thực thi các mục tiêu phức tạp đa bước.

(2) Ứng dụng khi nào

Doanh nghiệp cần định vị chính xác vị trí của bài toán nghiệp vụ để chọn nấc thang công nghệ tối ưu nhằm cân bằng giữa chi phí và hiệu năng:

- *Hệ thống Rule-based* : Áp dụng cho các bài toán kiểm soát giao dịch tài chính cốt lõi, nơi tính chính xác phải đạt 100% và logic xử lý hoàn toàn cố định.
- *Prediction AI* : Áp dụng cho các bài toán dự báo nhu cầu kho vận, chấm điểm tín dụng hoặc phát hiện gian lận thẻ.
- *Generative & Agentic AI* : Bắt buộc áp dụng cho các bài toán mở, xử lý văn bản/hình ảnh phi cấu trúc, tư vấn khách hàng linh hoạt và nghiên cứu dữ liệu tự động.

(3) Ví dụ minh họa

- **Kịch bản xử lý Tra cứu số dư ngân hàng** :
- *Rule-based* : Người dùng bắt buộc phải gõ đúng cú pháp TC SD Số_Tài_Khoản. Nếu gõ Kiểm tra tiền giúp tôi, hệ thống báo lỗi không nhận diện được cú pháp.
- *Generative/Agentic AI* : Người dùng nhắn Tháng này tôi còn đủ tiền trả tiền nhà 10 triệu không?. Hệ thống AI tự động phân tích ngữ cảnh, hiểu câu hỏi liên quan đến số dư hiện tại, gọi API truy vấn số dư thực tế là 15 triệu, và tự động tính toán phản hồi: Số dư hiện tại của bạn là 15.000.000 VNĐ, hoàn toàn đủ để thanh toán khoản tiền nhà 10.000.000 VNĐ.

Bài 3: Generative AI là gì?

(1) Giải thích khái niệm

Trí tuệ Nhân tạo Tạo sinh (Generative AI - GenAI) là lớp mô hình AI có khả năng tạo ra các thực thể dữ liệu mới—bao gồm văn bản, mã nguồn, hình ảnh, âm thanh—dựa trên việc thụ cảm ngữ cảnh từ dữ liệu đầu vào (Prompt Context). Cơ chế cốt lõi của GenAI dựa trên kiến trúc Transformer với cơ chế Tự chú ý (Self-Attention Mechanism). Mô hình phân tích các mối tương quan phụ thuộc không gian và thời gian giữa các token trong tập dữ liệu huấn luyện, từ đó tính toán phân phối xác suất để dự đoán và sinh ra chuỗi token tiếp theo ($Token_{t+1}$) sao cho văn bản mới tạo ra duy trì được độ mượt mà, đúng ngữ pháp và phù hợp nhất với yêu cầu của lời nhắc.

(2) Ứng dụng khi nào

Generative AI được áp dụng tối ưu trong các bài toán sáng tạo nội dung, tự động hóa bản thảo email truyền thông, tóm tắt tài liệu hợp đồng, dịch thuật đa ngôn ngữ và hỗ trợ lập trình viên viết mã nguồn.

- **Chỉ số đánh giá (KPIs)** : Đánh giá dựa trên Tỷ lệ tiết kiệm thời gian (Time-saved ratio), Độ bao phủ ngữ nghĩa (Semantic coverage) và Tỷ lệ cần con người hiệu chỉnh lại (Human-edit rate).
- **Rủi ro & Đánh đổi** : GenAI ưu tiên độ linh hoạt và tính sáng tạo của văn bản nhưng đối mặt với rủi ro biến đổi kết quả giữa các lần chạy (Non-determinism). Đối với các bài toán yêu cầu dữ liệu chính xác tuyệt đối, bắt buộc phải kết hợp kỹ thuật RAG hoặc kiểm soát tham số Temperature thấp.

(3) Ví dụ minh họa

Một chuyên viên marketing cần soạn kịch bản chiến dịch quảng cáo. Thay vì mất 4 giờ viết bản thảo, người dùng cung cấp ngữ cảnh đầu vào: "Sản phẩm: Phần mềm kế toán SaaS. Đối tượng: Doanh nghiệp SME. Mục tiêu: Tăng đăng ký dùng thử. Văn phong: Chuyên nghiệp, súc tích". Generative AI phân tích tập xác suất ngữ nghĩa và sinh ra một kịch bản hoàn chỉnh gồm tiêu đề, 3 luận điểm bán hàng cốt lõi và câu kêu gọi hành động (CTA) chuẩn hóa chỉ trong 3 giây.

Bài 4: Prediction AI là gì

(1) Giải thích khái niệm

AI Dự báo (Prediction AI) là nhánh AI chuyên biệt tập trung vào việc phân tích các mô hình dữ liệu lịch sử để tính toán xác suất xảy ra của các sự kiện trong tương lai hoặc phân loại các thực thể dữ liệu vào các nhóm chỉ định. Khác với Generative AI (vốn tập trung vào việc tạo ra dữ liệu mới), Prediction AI vận hành dựa trên các mô hình học máy thống kê và mạng thần kinh phân loại (Classification & Regression Models). Mô hình nhận diện các đặc trưng (Features) từ dữ liệu đầu vào, tối ưu hóa hàm mất mát (Loss function), và đưa ra kết quả đầu ra dưới dạng các điểm số xác suất (Probability Scores), nhãn định danh (Labels), hoặc các chỉ số dự báo định lượng.

(2) Ứng dụng khi nào

Prediction AI là công cụ nòng cốt trong các bài toán quản trị rủi ro doanh nghiệp, bao gồm: phân loại cảm xúc phản hồi khách hàng (Sentiment Analysis), dự báo nhu cầu chuỗi cung

ứng, phân khúc khách hàng mục tiêu, tính toán tỷ lệ rời bỏ dịch vụ (Churn Rate) và phát hiện giao dịch tài chính bất thường.

- **Chỉ số đo lường** : Hiệu năng của Prediction AI được đo lường khắt khe bằng các chỉ số thống kê chuẩn mực như Precision, Recall, F1-Score, và AUC-ROC.
- **Đánh đổi** : Mô hình yêu cầu tập dữ liệu gán nhãn lịch sử lớn và chất lượng cao để huấn luyện, nhưng mang lại thời gian phản hồi (Latency) cực thấp và tính toán quán tuyệt đối (Deterministic).

(3) Ví dụ minh họa

Trong hệ thống quản trị trải nghiệm khách hàng (CX), một mô hình Prediction AI được tích hợp để tự động quét 10.000 đánh giá hàng ngày từ ứng dụng. Mô hình không viết ra văn bản mới mà phân tích cú pháp, trích xuất điểm đặc trưng và gán nhãn chính xác từng phản hồi: Query: "Giao hàng quá chậm, đồ ăn bị nguội" $\rightarrow$ Nhãn: Tiêu cực | Độ tin cậy: 98.4% | Bộ phận: Chăm sóc Khách hàng Logistics.

Bài 5: Agentic AI là gì

(1) Giải thích khái niệm

AI Tự chủ (Agentic AI) thể hiện bước tiến kiến trúc quan trọng nhất của Trí tuệ Nhân tạo, chuyển đổi từ mô hình xử lý thụ động (Passive Generation) sang mô hình chủ động theo đuổi mục tiêu (Goal-driven Autonomy). Một hệ thống Agentic AI lấy LLM làm bộ não suy luận trung tâm, được trang bị khả năng tự phân rã mục tiêu phức tạp thành các tác vụ nhỏ (Task Decomposition), lập kế hoạch thực thi (Planning), tự tương tác với môi trường thông qua các công cụ bên ngoài (Tool Utilization), lưu trữ trạng thái vào bộ nhớ (Memory), và liên tục đánh giá, tự sửa lỗi (Self-Reflection) để đạt được mục tiêu cuối cùng mà không cần con người can thiệp từng bước.

(2) Ứng dụng khi nào

Agentic AI bắt buộc áp dụng cho các bài toán nghiệp vụ phức tạp, mang tính chất mở, nhiều bước xử lý và môi trường dữ liệu biến động cao. Bối cảnh điển hình bao gồm: tự động nghiên cứu thị trường chuyên sâu, vận hành quy trình kinh doanh đa hệ thống, kiểm thử mã nguồn tự động và xử lý sự cố công nghệ thông tin.

- **Thách thức kiến trúc** : Chi phí Token lớn, độ trễ xử lý cao (Latency từ vài chục giây đến vài phút), và khả năng bùng nổ vòng lặp vô tận (Infinite Loops) nếu không thiết kế ranh giới kiểm soát (Guardrails) và chỉ số dừng (Termination Conditions) chặt chẽ.

(3) Ví dụ minh họa

Doanh nghiệp giao mục tiêu cao cấp: "Hãy phân tích giá bán của 3 đối thủ cạnh tranh chính và lập bản thảo điều chỉnh giá sản phẩm X". Agentic AI sẽ tự động kích hoạt vòng lặp:

1. *Suy luận* : Cần tìm danh sách 3 đối thủ $\rightarrow$ Gọi công cụ Web Search.
2. *Hành động* : Trích xuất bảng giá từ 3 website đối thủ $\rightarrow$ Lưu kết quả vào Bộ nhớ làm việc.
3. *Phản ánh* : Phát hiện đối thủ B đang có chương trình khuyến mãi $\rightarrow$ Gọi tiếp API tính toán biên lợi nhuận để đề xuất mức giá tối ưu.
4. *Hoàn thành* : Biên soạn báo cáo tổng hợp gửi cho Giám đốc Kinh doanh.

Bài 6: LLM là gì?

(1) Giải thích khái niệm

Mô hình Ngôn ngữ Lớn (Large Language Model - LLM) là một kiến trúc trí tuệ nhân tạo chuyên sâu dựa trên mạng Transformer, sở hữu hàng tỷ đến hàng nghìn tỷ tham số (Parameters), được huấn luyện trên tập dữ liệu văn bản đa ngữ cảnh khổng lồ. Về mặt bản chất, LLM đóng vai trò là một bộ máy suy luận ngôn ngữ (Language Reasoning Engine). LLM vận hành bằng cách mã hóa văn bản đầu vào thành các không gian vectơ cao chiều (Embeddings), áp dụng các lớp Attention để thấu hiểu mối quan hệ ngữ cảnh phức tạp giữa các từ, và tính toán xác suất để đưa ra phản hồi logic, chính xác về mặt cú pháp và ngữ nghĩa.

(2) Ứng dụng khi nào

LLM đóng vai trò là hạ tầng trí tuệ trung tâm (Core Intelligence Layer) trong hầu hết các ứng dụng AI hiện đại: các hệ thống Chatbot doanh nghiệp, trợ lý ảo đa năng, công cụ trích xuất và tổng hợp tri thức, hệ thống dịch thuật ngữ cảnh sâu, và hạt nhân điều phối cho các Agentic AI.

- **Đánh giá kiến trúc** : Triển khai LLM đòi hỏi phải cân nhắc giữa việc sử dụng các mô hình thương mại đóng (như OpenAI GPT-4o, Claude 3.5 Sonnet với năng lực suy luận vượt trội) hoặc tự triển khai các mô hình mở (như Llama 3, Qwen với khả năng bảo mật dữ liệu nội bộ cao nhưng cần hạ tầng GPU đắt đỏ).

(3) Ví dụ minh họa

Khi nhân viên pháp chế gửi một câu hỏi phức tạp: "So sánh trách nhiệm pháp lý của Bên B trong hợp đồng kinh tế theo Luật Thương mại 2005 và Luật Doanh nghiệp 2020". Mô hình LLM tiếp nhận câu hỏi, phân tích cấu trúc ngôn ngữ, truy xuất tri thức được mã hóa trong không gian tham số, suy luận các điểm tương đồng và khác biệt, từ đó tổng hợp thành một văn bản phản hồi được cấu trúc mạch lạc thành các mục so sánh rõ ràng chỉ trong vài giây.

Bài 7: AI hallucination là gì?

(1) Giải thích khái niệm

Hiện tượng AI Ảo giác (AI Hallucination) là trạng thái mô hình ngôn ngữ sinh ra các thông tin hoàn toàn bịa đặt, sai lệch sự thật, không có trong dữ liệu huấn luyện hoặc phản trực giác, nhưng lại được trình bày với văn phong cực kỳ tự tin, logic và thuyết phục. Nguyên nhân kỹ thuật cốt lõi xuất phát từ bản chất vận hành của LLM: mô hình sinh từ dựa trên xác suất thống kê chuỗi Token ($P(Word_{t} | Word_{1}, ..., Word_{t-1})$) chứ không sở hữu một cơ sở dữ liệu sự thật khách quan (Ground Truth Database). Khi gặp khoảng trống dữ liệu hoặc lời nhắc gợi mở sai lệch, mô hình vẫn buộc phải hoàn thành chuỗi xác suất, dẫn đến việc "tự tạo ra" thông tin giả lập.

(2) Ứng dụng khi nào

Việc kiểm soát AI Hallucination là yêu cầu sinh tử khi triển khai AI trong các ngành nghiệp vụ có rủi ro cao như: pháp lý, y tế, phân tích tài chính, và quy trình vận hành kỹ thuật cốt lõi của doanh nghiệp.

- **Giải pháp giảm thiểu (Mitigation)** : Bắt buộc áp dụng kiến trúc Truy xuất dữ liệu bổ sung (RAG), yêu cầu trích xuất nguồn trích dẫn rõ ràng, tinh chỉnh tham số

Temperature = 0.0, xây dựng lớp kiểm định độc lập (Verification Layer / LLM-as-a-Judge) và duy trì cơ chế giám sát con người (Human-in-the-loop).

(3) Ví dụ minh họa

Một nhân viên tư vấn pháp lý sử dụng LLM thuần túy và hỏi: "Hãy trích dẫn Điều 105 Luật Bảo hiểm Xã hội về quyền lợi thai sản năm 2024". Do không truy cập dữ liệu thực tế, LLM tự tin trả lời: "Theo Điều 105 Luật Bảo hiểm Xã hội 2024, lao động nữ sinh con được hưởng 8 tháng lương...". Khi đối soát văn bản pháp luật thực tế, Điều 105 hoàn toàn quy định về nội dung khác và thời gian hưởng chế độ thai sản chỉ là 6 tháng. Đây là minh họa điển hình cho ảo giác AI gây rủi ro pháp lý nghiêm trọng nếu không có lớp kiểm định RAG.

Bài 11: Các cấp độ của AI

(1) Giải thích khái niệm

Các cấp độ của AI phản ánh sự tiến hóa về năng lực tự chủ, khả năng xử lý bài toán phức tạp và kiến trúc vận hành trong thực tiễn ứng dụng doanh nghiệp:

1. **Level 1 (Chatbot đơn bước)** : Phản hồi câu hỏi theo cơ chế Hỏi-Đáp đơn lẻ (Single-turn QA), không có bộ nhớ dài hạn và không dùng công cụ.
2. **Level 2 (AI Workflow)** : Thực thi chuỗi tác vụ có luồng điều khiển (Control Flow) được lập trình cố định bằng mã nguồn, AI chỉ xử lý tại các mắt xích chỉ định.
3. **Level 3 (AI Agent - Single Agent)** : Có năng lực tự chủ ra quyết định động, tự phân rã mục tiêu, tự chọn công cụ và quản lý bộ nhớ.
4. **Level 4 (Multi-Agent Systems)** : Hệ thống đa tác tử chuyên môn hóa, phối hợp và giám sát lẫn nhau để giải quyết các hệ thống bài toán quy mô lớn.| Cấp độ AI | Cơ chế vận hành chính | Độ tự chủ | Độ tin cậy (Predictability) | Thích hợp cho bài toán || ------ | ------ | ------ | ------ | ------ || **Level 1: Chatbot** | Hỏi - Đáp trực tiếp | Rất thấp | Cao (trong phạm vi ngắn) | Tra cứu nhanh, tóm tắt đơn giản || **Level 2: Workflow** | Luồng code cố định + LLM | Thấp | Tuyệt đối (Deterministic) | Quy trình chuẩn mực, ít biến số || **Level 3: Agent** | Lập kế hoạch + Gọi công cụ | Cao | Trung bình (Probabilistic) | Bài toán mở, cần linh hoạt || **Level 4: Multi-Agent** | Phân rã đa tác tử + Giám sát | Rất cao | Phức tạp (Cần Guardrails) | Hệ thống quy mô lớn, đa vai trò |

(2) Ứng dụng khi nào

Định vị đúng cấp độ AI giúp doanh nghiệp tránh việc "dùng dao mổ trâu để cắt lông gà" (triển khai Multi-Agent cho bài toán nhập liệu đơn giản) hoặc dùng Chatbot thụ động cho quy trình tự động hóa phức tạp.

(3) Ví dụ minh họa

- *Level 1* : Người dùng nhập yêu cầu viết bản thảo bài đăng Facebook quảng cáo sản phẩm.
- *Level 2* : Quy trình nhập liệu cố định: File PDF gửi đến $\rightarrow$ Code chuyển thành ảnh $\rightarrow$ LLM trích xuất số tiền $\rightarrow$ Code lưu vào cơ sở dữ liệu.
- *Level 3* : AI Agent tự nghiên cứu thị trường, tự viết nội dung, gọi API gửi email, đo lường tỷ lệ tương tác và tự điều chỉnh văn phong cho chiến dịch tiếp theo.

Bài 21: Multimodal AI: Khi AI hiểu cả chữ và hình ảnh

(1) Giải thích khái niệm

AI Đa phương thức (Multimodal AI) là kiến trúc trí tuệ nhân tạo có khả năng tiếp nhận, xử lý và đồng bộ hóa ngữ cảnh từ nhiều dạng dữ liệu đầu vào khác nhau—bao gồm văn bản (Text), hình ảnh (Images), âm thanh (Audio), và tài liệu cấu trúc phức tạp—vào cùng một không gian biểu diễn vectơ chung (Shared Embedding Space). Cơ chế này giúp mô hình không chỉ "đọc" văn bản mà còn "nhìn" và hiểu được mối tương quan không gian, biểu đồ, hình ảnh thực tế và cấu trúc thị giác của tài liệu, loại bỏ rào cản phân tách giữa các dạng dữ liệu khác nhau.

(2) Ứng dụng khi nào

Multimodal AI bắt buộc triển khai trong các bài toán kinh doanh thực tế nơi thông tin tồn tại dưới dạng hỗn hợp phi cấu trúc: phân tích báo cáo tài chính chứa sơ đồ/biểu đồ cột, đọc hiểu hóa đơn thanh toán dạng ảnh quét (OCR + Semantic Understanding), kiểm tra tổn thất bảo hiểm qua ảnh chụp hiện trường, và tự động hóa kiểm thử giao diện phần mềm (UI/UX Automation).

- **Thách thức** : Tốn kém tài nguyên Token (hình ảnh được quy đổi thành hàng trăm Tokens thị giác - Visual Tokens) và tăng độ trễ phản hồi của hệ thống.

(3) Ví dụ minh họa

Trong quy trình thẩm định tài chính, chuyên viên tải lên bức ảnh chụp biểu đồ doanh thu dạng cột phức tạp của doanh nghiệp kèm lời nhắc: "Trích xuất số liệu doanh thu từng quý năm 2023, tính tốc độ tăng trưởng Q4 so với Q1 và nhận xét xu hướng". Multimodal AI phân tích các Visual Tokens, nhận diện trục tọa độ, trích xuất chính xác các con số hiển thị trên biểu đồ, thực hiện phép tính và xuất ra báo cáo phân tích tài chính hoàn chỉnh theo cấu trúc Markdown.Thấu hiểu các khái niệm nền tảng về AI và LLM là bước đệm bắt buộc. Tuy nhiên, để điều khiển mô hình tạo ra đầu ra chính xác và ổn định, chúng ta cần chuyển đổi tư duy từ việc "ra lệnh" cảm tính sang việc làm chủ "Kỹ thuật Prompting và Quản lý Context".

CHUYÊN ĐỀ 2: NỀN TẢNG PROMPT ENGINEERING VÀ QUẢN LÝ CONTEXT

Trong kiến trúc ứng dụng mô hình ngôn ngữ lớn, chất lượng đầu ra của AI không phụ thuộc vào "sự thông minh" ngẫu nhiên của mô hình mà được quyết định hoàn toàn bởi chất lượng của môi trường thông tin đầu vào. Để khai thác AI hiệu quả trong môi trường doanh nghiệp, kỹ sư và chuyên viên cần thực hiện một sự chuyển dịch tư duy quan trọng: từ việc đưa ra các câu lệnh (Prompts) ngắn ngơ đơ cảm tính sang việc thiết kế toàn bộ môi trường thông tin (Context Engineering). Khi ngữ cảnh đầu vào được cấu trúc hóa chặt chẽ—bao gồm định danh vai trò, ranh giới nhiệm vụ, cấu trúc dữ liệu, tiêu chí hoàn thành và định dạng đầu ra—mô hình ngôn ngữ sẽ vận hành như một bộ máy suy luận chính xác, loại bỏ tối đa ảo giác và đáp ứng hoàn hảo các tiêu chuẩn nghiệp vụ.

Bài 8: Context Engineering là gì

(1) Giải thích khái niệm

Context Engineering (Kỹ thuật Ngữ cảnh) là quá trình thiết kế, cấu trúc hóa, thử nghiệm, giám sát và tối ưu hóa toàn bộ môi trường thông tin đưa vào mô hình AI để định hình hành vi và nâng cao độ tin cậy của hệ thống. Trong kiến trúc sản xuất (Production), Context không

chỉ đơn thuần là một lời nhắc (Prompt), mà là một hệ thống thông tin đa tầng bao gồm: Chỉ dẫn hệ thống (System Instructions), Tiêu chuẩn tác vụ (Task Constraints), Mô tả công cụ (Tool Schemas), Lịch sử trạng thái (Memory) và Lớp giám sát vận hành (Observability Layer).Khái niệm cốt lõi của Context Engineering bao gồm việc duy trì tính minh bạch vận hành (Observability) thông qua bộ theo dõi tác vụ (Task Tracker với các trường bắt buộc: Task ID, Query, Status, Result Summary, Timestamp) và tuân thủ **4 tiêu chí đánh giá Context chuẩn mực** :

1. **Completeness (Tính đầy đủ)** : Bao phủ mọi kịch bản biên và ngoại lệ.
2. **Clarity (Tính rõ ràng)** : Loại bỏ tuyệt đối sự mơ hồ khiến AI phải tự đoán.
3. **Consistency (Tính nhất quán)** : Các quy tắc không mâu thuẫn lẫn nhau.
4. **Testability (Tính kiểm thử)** : Đầu ra có thể đo lường và xác minh tự động.

(2) Ứng dụng khi nào

Context Engineering là kỹ thuật bắt buộc khi xây dựng các ứng dụng AI chuyên nghiệp, hệ thống AI Agent tự chủ hoặc các quy trình tự động hóa phức tạp trong doanh nghiệp. Đây là lớp bảo vệ giúp AI vận hành ổn định, tuân thủ chính sách bảo mật và tránh gãy đổ quy trình.

(3) Ví dụ minh họa

Cấu trúc kiến trúc Context đa tầng cho một Agent hỗ trợ pháp lý:

```
[SYSTEM LAYER]: Bạn là Trợ lý Pháp lý Cao cấp. Bắt buộc tuân thủ Luật Doanh nghiệp 2020.
[TASK LAYER]: Hãy phân tích khoản bồi thường trong Hợp đồng X. Chỉ xuất ra các khoản > 100 triệu.
[TOOL LAYER]: Schema tool `search_law_database(keyword: string, year: int)`.
[MEMORY LAYER]: Lịch sử phân tích 2 phụ lục hợp đồng trước đó.
[OBSERVABILITY]: Task_ID: #4092 | Status: Processing | Timestamp: 2026-10-01 10:15:00
```

Bài 9: Prompt Engineering là gì

(1) Giải thích khái niệm

Prompt Engineering (Kỹ thuật Thiết kế Lời nhắc) là quá trình cấu trúc, thiết kế và tinh chỉnh ngôn ngữ đầu vào nhằm hướng dẫn mô hình ngôn ngữ lớn thực hiện một tác vụ cụ thể một cách chính xác, ổn định và có thể dự đoán được. Prompt Engineering đóng vai trò là giao diện giao tiếp (Interface) giữa ngôn ngữ tự nhiên của con người và bộ máy suy luận xác suất của mô hình. Một prompt được tối ưu hóa chuẩn mực không nhất thiết phải kéo dài, mà cần tập trung cung cấp đầy đủ các chỉ dẫn cốt lõi để mô hình hiểu rõ: nhiệm vụ cần làm, phạm vi dữ liệu được sử dụng, các ràng buộc cấm và định dạng phản hồi mong muốn.

(2) Ứng dụng khi nào

Prompt Engineering được áp dụng trong công việc hàng ngày của cán bộ nhân viên để tối ưu hóa hiệu suất giao tiếp với AI, chuẩn hóa các văn bản báo cáo, tóm tắt tài liệu nghiệp vụ hoặc trích xuất dữ liệu phi cấu trúc.

- **Đánh đổi** : Tốn thời gian thử nghiệm và tinh chỉnh ban đầu (Prompt Iteration), nhưng giúp giảm thiểu 80% thời gian chỉnh sửa văn bản thủ công về sau và tiết kiệm lượng lớn token không cần thiết.