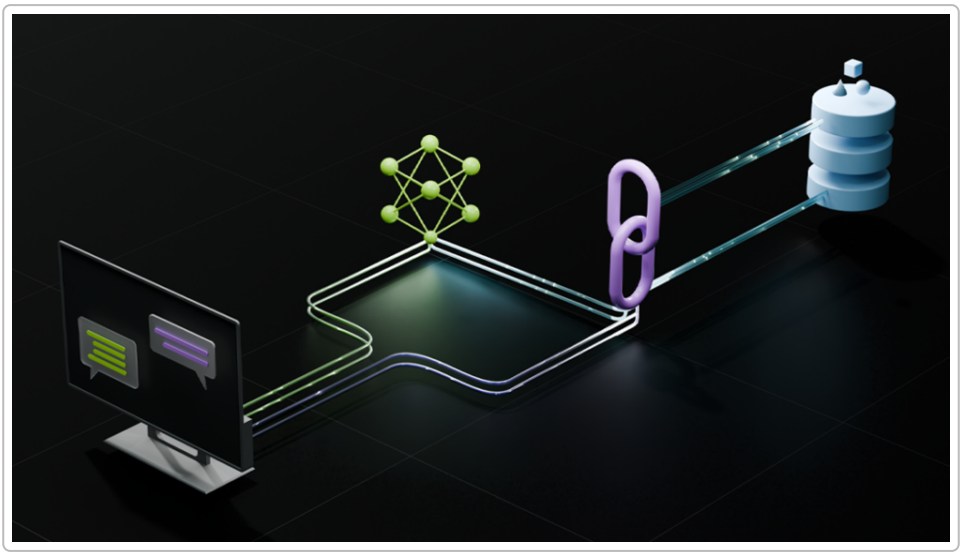
ChatGPT

Kiến trúc và Mô hình AI Agent hiện đại

Các hệ thống gợi ý ý tưởng nội dung video thường kết hợp mô hình ngôn ngữ lớn (LLM) với cơ chế RAG. Một kiến trúc điển hình là: **(1)** thu thập dữ liệu liên quan (ví dụ: transcript video, tiêu đề, mô tả, xu hướng hiện tại); **(2)** tạo embedding và lưu vào cơ sở dữ liệu vector (Pinecone, Weaviate...); **(3)** khi nhận truy vấn hoặc ngữ cảnh mới, truy vấn vector để lấy ngữ cảnh liên quan rồi **(4)** đưa vào LLM để sinh đề xuất. Ví dụ, qua LangChain ta có thể xây dựng pipeline RAG: chuyển truy vấn thành vector, tìm kiếm trong vector DB và đưa ngữ cảnh vào LLM [1] [2]. Ngoài ra, khung đa-tác nhân ngày càng phổ biến: Google ADK (Agent Development Kit) cho phép phân tách nhiệm vụ thành nhiều agent (ví dụ Agent LLM, Agent hướng dẫn luồng công việc, Agent tùy chỉnh) phối hợp với nhau [3] [4]. ADK định nghĩa rõ **LLM Agent** dùng các mô hình mạnh (Gemini 1.5, GPT-4o, ...) cho các tác vụ suy luận và lập kế hoạch [3]; **Workflow Agent** điều phối thứ tự thực thi giữa các bước; **Custom Agent** gọi API bên ngoài (ví dụ YouTube API). Quy trình chung có thể mô tả như sau (minh họa sơ đồ bên dưới): hệ thống RAG đầu tiên ingest dữ liệu (video, tin tức, v.v.) và tách nhỏ, chuyển thành embedding [2] [5]. Khi người dùng tương tác, hệ thống truy vấn những vector phù hợp và cung cấp ngữ cảnh đó cho LLM sinh phản hồi. Các thư viện như LangChain và LlamaIndex (GPT Index) đặc biệt hữu ích để kết nối LLM với nguồn dữ liệu. LlamaIndex có các bộ nạp (loader) chuyên biệt, ví dụ `YoutubeTranscriptReader`, giúp tự động lấy transcript từ YouTube và tạo index vector [6] [7].



Hình ảnh: Sơ đồ luồng RAG tiêu biểu – dữ liệu được ingest, chia nhỏ và embedding, lưu trong vector DB, rồi khi có truy vấn hệ thống truy xuất ngữ cảnh (vector) đưa vào LLM để sinh kết quả [2] [5].

- **Hành trình truy vấn-đáp:** Ví dụ, người dùng nhập chủ đề (“animals”), hệ thống chuyển sang vector rồi truy vấn cơ sở dữ liệu (các video liên quan, tin tức, trending topics). Kết quả tìm được (mô tả, transcript hay tiêu đề) được đưa cùng truy vấn vào LLM (Gemini-1.5) để sinh đề xuất cụ thể (như chuỗi ý tưởng video).
- **Đa-hạng mục kết nối:** LangChain/LlamaIndex cho phép tạo và quản lý index từ nhiều nguồn (file văn bản, API, YouTube, v.v.) [1] [7]. Pinecone hay Weaviate dùng để lưu trữ vector và tìm kiếm tương tự (ví dụ bài viết hướng dẫn dùng Pinecone cho chatbot RAG [8]).
- **Agent đa-tác vụ:** Với Google ADK (Agent Development Kit), ta cấu trúc hệ thống thành các agent chuyên trách; ví dụ một dự án tạo kịch bản video ngắn đã dùng agent LLM (Gemini-1.5) làm tổng

chỉ huy và các agent con (Script Writer, Visualizer, Formatter) chuyên phân tích nội dung, thêm hiệu ứng hình ảnh, định dạng kịch bản [3] [9]. ADK hỗ trợ giao thức Agent2Agent để các agent tương tác qua chuẩn chung [4], cho phép kết hợp cả công cụ nguồn mở (LangChain, Crew.ai) lẫn hạ tầng Google Cloud.

Vai trò của RAG trong đề xuất nội dung video

Kỹ thuật RAG giúp chatbot **cập nhật và sử dụng kiến thức ngoài mô hình** để gợi ý sáng tạo. Thay vì chỉ dựa trên dữ liệu đã học, RAG cho phép tra cứu tức thời các nguồn dữ liệu mới hoặc chuyên biệt. Ví dụ:
- **Xu hướng và dữ liệu thời gian thực:** Chatbot có thể tích hợp các API như Google Trends, YouTube Data API để lấy chủ đề đang thịnh hành, tiêu đề video nổi bật, hoặc thông tin phản hồi người xem. Những dữ liệu này được chuyển thành embedding và lưu vào kho dữ liệu để trả lời truy vấn sáng tạo.
- **Dữ liệu sẵn có:** Nội dung transcript của các video YouTube hiện có (của kênh hoặc lĩnh vực liên quan) giúp hiểu phong cách và chủ đề đã làm. LlamaIndex có bộ reader cho YouTube để tự động lấy transcript và xây dựng chỉ mục nội dung [6] [7]. Khi hỏi chatbot, hệ thống RAG sẽ truy vấn các đoạn transcript liên quan làm ngữ cảnh, nhờ đó đề xuất mới gắn kết với nội dung đã có.
- **Tệp tài liệu khác:** Ngoài video, có thể ingest wiki, bài báo, diễn đàn,... liên quan đến chủ đề video (vd. tin tức công nghệ mới, xu hướng game, v.v.). Vertex AI RAG Engine của Google hỗ trợ ingest dữ liệu từ Cloud Storage, Drive, Slack, Jira, thậm chí Google Search để làm nguồn tri thức [2] [10]. Sau đó, mỗi khi người dùng hỏi, hệ thống tìm văn bản/đoạn thích hợp đưa vào Gemini sinh ý tưởng video.

Bằng cách này, RAG **mở rộng kiến thức** của chatbot: nó không chỉ "bắt chước" bản thân LLM mà còn "tra cứu" thông tin thời sự, đảm bảo đề xuất thiết thực, cập nhật. Đồng thời RAG giúp giảm sai sót (hallucination) khi LLM có ngữ cảnh xác thực làm nền tảng.

Fine-tune nhẹ Gemini-1.5-Flash

Gemini-1.5-Flash (phiên bản miễn phí) có khả năng rất cao và hỗ trợ ngữ cảnh dài (tới 1 triệu tokens [11]), nhưng có thể cần tinh chỉnh cho sát nhu cầu cụ thể. **Fine-tuning giám sát (supervised fine-tuning – SFT)** của Gemini trên Vertex AI/AI Studio cho phép "dạy" mô hình các phản hồi mẫu theo định dạng mong muốn [12] [13]. Ví dụ ta có thể chuẩn bị bộ ví dụ gồm: *Input:* một mô tả nội dung hoặc câu hỏi (ví dụ "Hãy gợi ý ý tưởng video về 'du lịch Đà Lạt'"), *Output:* là một dàn ý hoặc tiêu đề sáng tạo phù hợp. Đưa bộ này vào tính năng fine-tune sẽ tạo ra phiên bản Gemini tùy chỉnh. Sau fine-tune, mô hình có xu hướng tuân thủ định dạng hoặc phong cách đã học (ví dụ luôn trả lời dưới dạng tiêu đề, gạch đầu dòng ý tưởng, v.v.) [13] [12].

Google hỗ trợ nhiều mức độ tinh chỉnh: các mô hình Gemini-2.0 (tương tự Gemini-1.5) dùng cơ chế adapter có kích cỡ 1, 2, 4, 8 (kích thước adapter càng nhỏ, mô hình càng giữ nguyên nhiều thông tin gốc) [14]. Với "fine-tune nhẹ", ta có thể chọn adapter nhỏ (1–4) và ít ví dụ (vài nghìn câu) để tinh chỉnh hành vi sáng tạo của Gemini-1.5-Flash. Ngoài ra, Gemini còn có khái niệm "Gem" (cho Gemini Advanced) – cách tạo chatbot chuyên gia bằng mô tả mô phỏng đầu vào [15] [16]. Dù "Gem" là tính năng trong ứng dụng Gemini, tương tự ý tưởng fine-tune, giúp điều chỉnh tính cách và phản hồi của AI mà không cần re-prompt liên tục.

Công cụ áp dụng: Việc fine-tune Gemini-1.5 thường được thực hiện qua **Vertex AI** hoặc **Google AI Studio**. Google AI Studio cung cấp giao diện để upload dataset, khởi chạy job huấn luyện và quản lý API key cho Gemini [17] [12]. Sau khi fine-tune, ta có một mô hình con dựa trên Gemini gốc, có thể triển khai qua Vertex hoặc gọi API giống như gọi LLM bình thường. Một lưu ý: Gemini 1.5-Flash mới đang được thay bằng Gemini 2.0, nhưng nguyên tắc huấn luyện (dataset JSONL, adapter) vẫn tương tự [18].

Nền tảng và Thư viện chuyên dụng

Hiện nay có nhiều thư viện và nền tảng hỗ trợ phát triển AI Agent RAG/LLM:

- **LangChain (Python)** – khung làm việc phổ biến để xây dựng ứng dụng LLM. LangChain giúp xâu chuỗi bước (lấy dữ liệu, tạo embedding, tìm kiếm, gọi LLM) một cách thuận tiện [1]. Nó hỗ trợ nhiều loại connectors (openAI, Google Gemini qua `langchain-google-genai`, v.v.) và tích hợp dễ với Pinecone hay Vertex RAG.
- **LlamaIndex (GPT Index)** – thư viện tập trung vào xây dựng chỉ mục ngữ liệu. Cung cấp hàng loạt *reader/loader* cho các nguồn khác nhau (văn bản, PDF, YouTube, website...). Ví dụ `YoutubeTranscriptReader` của LlamaIndex giúp thu thập transcript và tạo index vector cho RAG YouTube [6] [7].
- **Cơ sở dữ liệu vector** – như *Pinecone*, *Weaviate*, *ChromaDB*, *Qdrant*... dùng để lưu embeddings và hỗ trợ tìm kiếm nhanh. Ví dụ, trong hướng dẫn xây chatbot bán hàng, Pinecone lưu trữ embedding của mô tả sản phẩm để tìm sản phẩm tương tự [8].
- **Google Vertex AI RAG Engine** – nền tảng quản lý RAG trên đám mây GCP. Tự động ingest dữ liệu từ file/Cloud Storage/Drive, tạo chỉ mục ("corpus"), và hỗ trợ tìm kiếm ngữ cảnh [2]. Đặc biệt hữu ích khi dữ liệu doanh nghiệp đã có sẵn trên GCP (BigQuery, Cloud Storage...), vì chỉ cần cấu hình vài bước là có pipeline RAG đầy đủ.
- **Google Vertex AI Agent Builder (ADK)** – khung xây dựng hệ thống agent đa-tác vụ. ADK cho phép định nghĩa agent (LLM Agent, Workflow Agent, Custom Agent) bằng cấu hình đơn giản. Hỗ trợ biên dịch luồng đa bước, kết nối audio/video đa chiều, và tích hợp với **A2A Protocol** (Agent2Agent) để giao tiếp giữa các agent trên các framework khác nhau [4] [19]. ADK tương thích cả với Gemini và các mô hình khác.
- **Google AI Studio / Generative AI Studio** – giao diện web để quản lý API key Gemini, tạo hoặc fine-tune mô hình. Ngoài ra còn có **Vertex AI Search** (tìm kiếm kết hợp từ khóa và vector) cho RAG và hàng trăm connector tích hợp sẵn (Google Search, Maps, v.v.) để "ground" kết quả [19] [10].
- **Crew.ai, LangGraph, Agent Playground...** – các nền tảng/thư viện khác cho phép thiết kế workflow agent (mã nguồn mở). Google cũng liệt kê chúng tương thích với hệ thống của mình [4].
- **Nền tảng no-code (như Momen.ai)** – một số công ty cung cấp giải pháp không cần code để xây chatbot đa agent, tích hợp API YouTube/DALL-E. Ví dụ, Momen đã xây chatbot tạo kịch bản video với workflow tùy chỉnh, gọi API YouTube và OpenAI DALL-E tự động [20] [21].

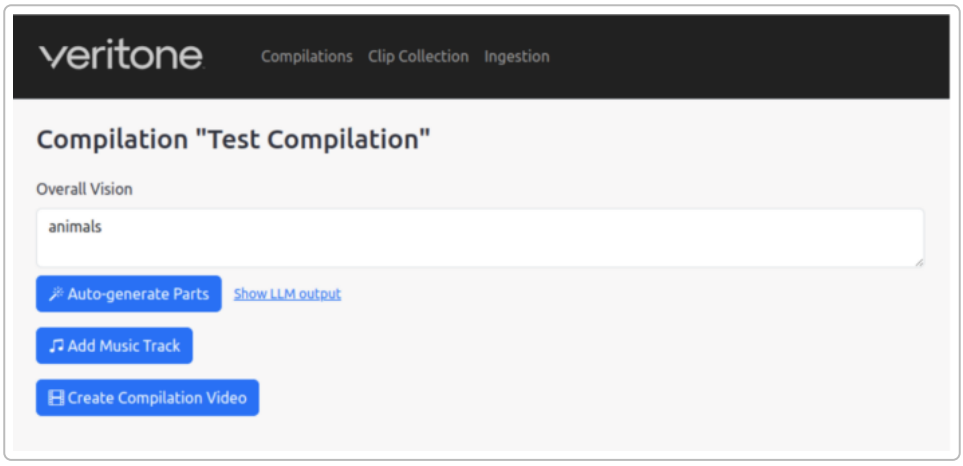
Bảng dưới đây so sánh một số công cụ và nền tảng tiêu biểu:

Công cụ/Nền tảng	Loại	Chức năng chính / Ứng dụng tiêu biểu
LangChain	Thư viện Python	Xây dựng pipeline LLM/RAG; quản lý luồng giữa các thành phần (embedding, RAG, LLM) [1]. Dễ tích hợp với Gemini qua plugin `langchain-google-genai`.
LlamaIndex (GPT-Index)	Thư viện Python	Tạo chỉ mục và vector từ nhiều nguồn (YouTube, tài liệu, web,...). Có `YoutubeTranscriptReader` tự động lấy transcript cho RAG YouTube [6] [7].
Pinecone / Weaviate	Cơ sở vector (DB)	Lưu embeddings và tìm kiếm tương tự. Ví dụ dùng Pinecone lưu mô tả sản phẩm cho chatbot RAG [8].

Công cụ/Nền tảng	Loại	Chức năng chính / Ứng dụng tiêu biểu
Vertex AI RAG Engine	Nền tảng đám mây GCP	Quản lý quy trình RAG toàn bộ: ingest dữ liệu, tách nhỏ, embed, index, truy vấn. Kết nối tốt với GCP (Cloud Storage, BigQuery) [2] .
Vertex AI Agent (ADK)	Nền tảng đa-agent	Phát triển hệ thống đa-agent (LLM Agents, Workflow Agents) với Gemini/GPT; hỗ trợ viết config Python <100 dòng [3] [4] .
Google AI Studio	Nền tảng GCP	Quản lý API key và fine-tune cho Gemini; cung cấp Gemini API (text, code, image, video) và Gemini Gems để tạo chatbot theo yêu cầu.
Khác (Crew.ai, LangGraph,...)	Framework agent	Các thư viện cộng đồng để xây agent; có thể phối hợp với Vertex AI Agent. Google hỗ trợ giao tiếp Agent2Agent giữa các hệ sinh thái [4] .

Ví dụ điển hình và Sản phẩm thực tế

- **Veritone Labs – Video Compilation Agent:** Ví dụ minh họa (hình dưới) của Veritone Lab cho thấy chatbot lấy một "vision" ban đầu ("animals"), rồi LLM đề xuất 5 phần của video biên tập và các truy vấn tìm đoạn phim tương ứng [22] [23] . Đầu tiên, người dùng nhập ý tưởng tổng quát ("animals"); hệ thống gọi LLM để sinh từng ý tưởng chính ("Majestic Wild Animals", "Cute Baby Animals",...) và gợi ý các truy vấn tìm clip (như "Lion roaring in savanna" hoặc "Kitten playing with yarn"). Người dùng sau đó có thể click chọn video từ kết quả tìm kiếm. Cách tiếp cận này kết hợp **RAG (tìm video/ảnh nhờ embedding)** với **tạo ý tưởng ngôn ngữ** qua LLM, giúp tạo nhanh ý tưởng cho video biên tập [22] [23] .



Hình ảnh: Ví dụ prototype Veritone – chatbot AI nhận "overall vision" (ở đây là "animals"), tự động sinh ra 5 phần ý tưởng cho video biên tập và cung cấp truy vấn tìm clip tương ứng (theo ngữ cảnh) [22] [23] .

- **Momen AI – YouTube Script Assistant:** Momen.ai cung cấp nền tảng no-code để tạo chatbot nội dung YouTube. Ứng dụng mẫu của họ tự động **phân tích kênh YouTube** (lấy 30 video phổ biến), rồi tìm kiếm 50 video nổi bật khác dựa trên chủ đề do người dùng đưa. Sau đó LLM kết hợp thông tin về xu hướng và phong cách kênh để đề xuất 5 tiêu đề video sáng tạo riêng biệt [20] . Người dùng chọn tiêu đề, sau đó chatbot tiếp tục tạo kịch bản chi tiết (title, mô tả, tags, nội dung) và cả gợi ý thumbnail với DALL-E [20] [21] . Ứng dụng này tích hợp API YouTube (lấy thông

tin kênh, tìm kiếm, chi tiết video) và AI (GPT hoặc Gemini) để tổng hợp ý tưởng, minh họa việc dùng RAG tra cứu dữ liệu từ YouTube kết hợp LLM cho sáng tác nội dung.

- **Google ADK – Short Video Script Writer:** Tại sự kiện Google Cloud Next 2025, một bài viết kỹ thuật giới thiệu dùng Google ADK để xây "nhóm agent" tự động viết kịch bản video ngắn [3] [24]. Hệ thống gồm một agent tổng chỉ huy (chatbot Gemini-1.5), một "Script Writer Agent" (tìm kiếm thông tin trên Google), một "Visualizer Agent" (thêm gợi ý hình ảnh), và một "Formatter Agent" (định dạng kịch bản). Các agent giao tiếp với nhau (qua khoá chung output_key) để lần lượt xây dựng nội dung từ thô đến tinh chỉnh. Ví dụ, Script Writer Agent dùng LLM tìm dữ liệu liên quan, Visualizer Agent thêm chi tiết minh họa, Formatter Agent gom lại thành kịch bản hoàn chỉnh [24]. Thiết kế nhiều agent như trên cho thấy cách tổ chức mô-đun với Gemini và RAG trong một hệ thống tự động sáng tạo nội dung.

- **Ví dụ khác:** Google cũng công bố hướng dẫn tạo AI Agent theo dõi xu hướng thực tế, như **Trip Planner Agent** dùng Gemini 1.5 Pro kết hợp API sự kiện/vé/phòng khách sạn [25] [26]. Mặc dù mục đích khác (du lịch), nó cho thấy Gemini 1.5 có thể gọi API bên ngoài (function calling) để lấy dữ liệu thời gian thực, tương tự ta có thể gọi YouTube API, Google Trends... để gợi ý nội dung video. Chẳng hạn, tài liệu này nhấn mạnh chức năng *function calling* giúp Gemini truy cập hệ thống bên ngoài và *grounding* giúp nó xử lý dữ liệu trực tuyến chính xác hơn [25].

Nhìn chung, hàng loạt công cụ và sản phẩm mới (từ nghiên cứu đến thương mại) đều áp dụng kiến trúc RAG + LLM để hỗ trợ ý tưởng sáng tạo. Các ví dụ trên cho thấy quy trình chung: **Kết hợp dữ liệu định hướng (đã có và mới) với khả năng sinh ngôn ngữ của Gemini-1.5 hoặc GPT,** đồng thời tận dụng framework như LangChain/ADK để tổ chức hệ thống sao cho linh hoạt và mở rộng được.

Nguồn tham khảo: Các thông tin trên được tổng hợp từ tài liệu kỹ thuật và hướng dẫn của Google Cloud [3] [25], bài viết/đồ án mã nguồn mở về RAG và chatbot trên Youtube [7] [20], cùng các ví dụ thực tế được công bố công khai [22] [20].

---

[1] [8] [17] Building a RAG Chatbot: LangChain, Pinecone, and Gemini | by Wasay Ali | Medium
https://medium.com/@wasay.abbs/building-a-rag-chatbot-langchain-pinecone-and-gemini-d6f6b4be1015

[2] Vertex AI RAG Engine overview | Generative AI on Vertex AI | Google Cloud
https://cloud.google.com/vertex-ai/generative-ai/docs/rag-engine/rag-overview

[3] [9] [24] Building a Short Video Script Writer with Google's Agent Development Kit (ADK) | by Malaya Khuntia | Google Cloud - Community | Apr, 2025 | Medium
https://medium.com/google-cloud/building-a-short-video-script-writer-with-googles-agent-development-kit-adk-7b0e55d132cc

[4] [10] [19] Vertex AI Agent Builder | Google Cloud
https://cloud.google.com/products/agent-builder

[5] RAG 101: Demystifying Retrieval-Augmented Generation Pipelines | NVIDIA Technical Blog
https://developer.nvidia.com/blog/rag-101-demystifying-retrieval-augmented-generation-pipelines/

[6] [7] Retrieval Augmented Generation(RAG) — Chatbot for Youtube with LlamaIndex | by A B Vijay Kumar | Medium
https://abvijaykumar.medium.com/retrieval-augmented-generation-rag-chatbot-for-youtube-with-llamaindex-f17e92d8886a

[11] Introducing Gemini 1.5, Google's next-generation AI model
https://blog.google/technology/ai/google-gemini-next-generation-model-february-2024/

12 Fine-tuning with the Gemini API | Google AI for Developers
https://ai.google.dev/gemini-api/docs/model-tuning

13 14 18 About supervised fine-tuning for Gemini models | Generative AI on Vertex AI | Google Cloud
https://cloud.google.com/vertex-ai/generative-ai/docs/models/gemini-supervised-tuning

15 16 How to Create a Personalized Chatbot With Google Gemini
https://promevo.com/blog/google-gemini-chatbots

20 21 We Built an AI Agent for Youtube Content Creation - DEV Community
https://dev.to/momen_hq/we-built-an-ai-agent-for-youtube-content-creation-42ih

22 23 An AI Agent Prototype for Assistance in Creating Video Compilations - Veritone
https://www.veritone.com/blog/an-ai-agent-prototype-for-assistance-in-creating-video-compilations/

25 26 Learn how to create an AI agent for trip planning with Gemini 1.5 Pro | Google Cloud Blog
https://cloud.google.com/blog/topics/developers-practitioners/learn-how-to-create-an-ai-agent-for-trip-planning-with-gemini-1-5-pro