Prompt Engineering: Leveraging LLMs Day 3: RAG & Multimodal LLMs

Max Moundas

July 17, 2025

Agenda

- Retrieval-Augmented Generation (RAG)
- Assistants
- Deep Research
- Multimodal Capabilities
- Voice, Image and Video Generation

Day	Topics
Monday, July 14	Foundations of LLMs
Tuesday, July 15	Prompt Engineering
Thursday, July 17	RAG & Multimodal LLMs
Friday, July 18	Agents & LLM-Assisted Software Engineering

Recent Research

- December 18, 2024
  - Alignment faking in large language models & AI Sandbagging
- June 21, 2025
  - Agentic Misalignment: How LLMs could be insider threats
- July 15, 2025
  - Chain of Thought Monitorability: A New and Fragile Opportunity for AI Safety

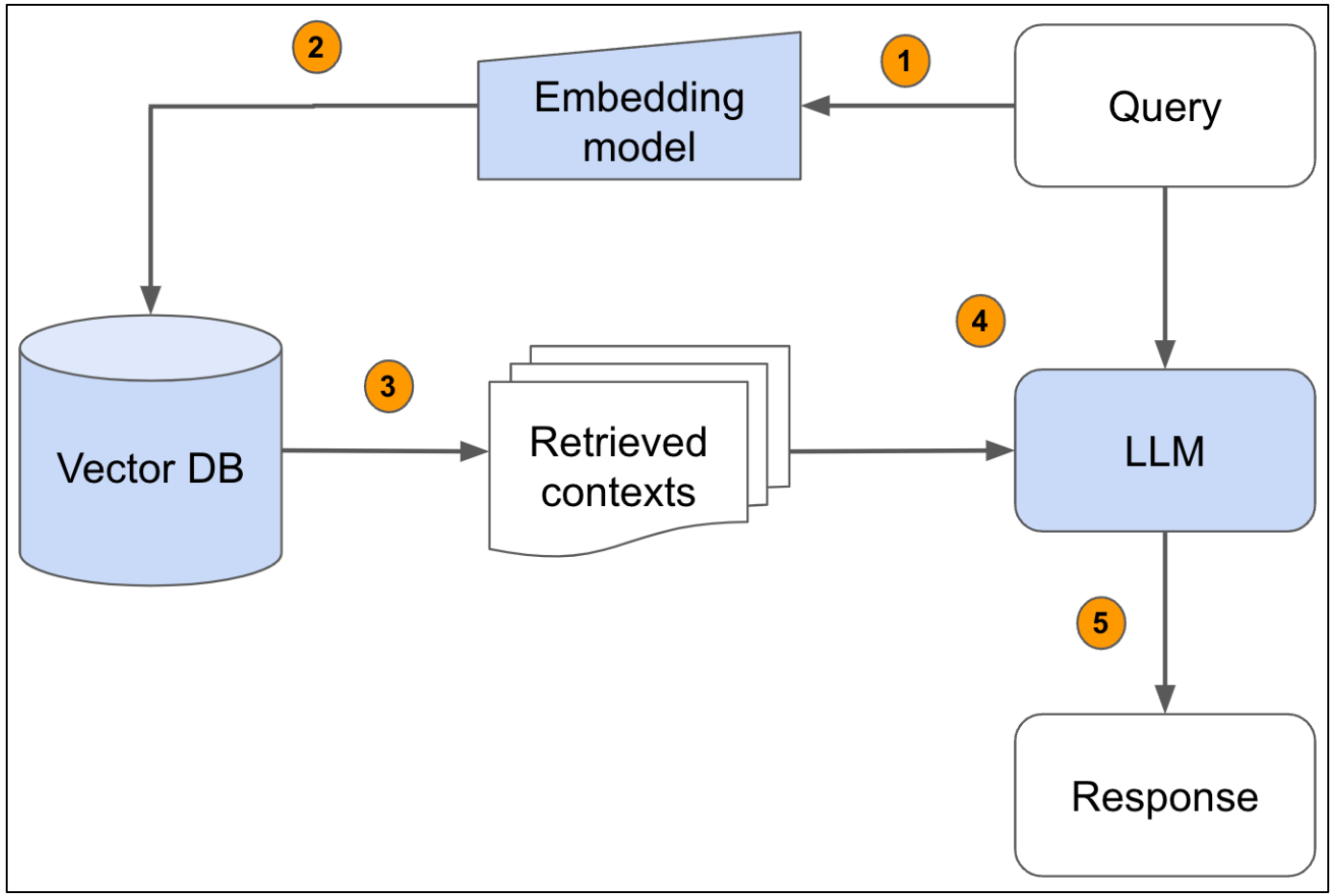
https://www.anthropic.com/research/alignment-faking
https://arxiv.org/abs/2406.07358
https://www.anthropic.com/research/agentic-misalignment
https://arxiv.org/abs/2507.11473

Limitations Of LLMs

- Can’t access real-time or private data without retrieval or APIs
- Struggle with long context, memory, or document traversal
- Tend to hallucinate facts, citation and math
- No built-in source grounding
- Outputs vary based on phrasing, temperature and recency of training

Retrieval-Augmented Generation (RAG)

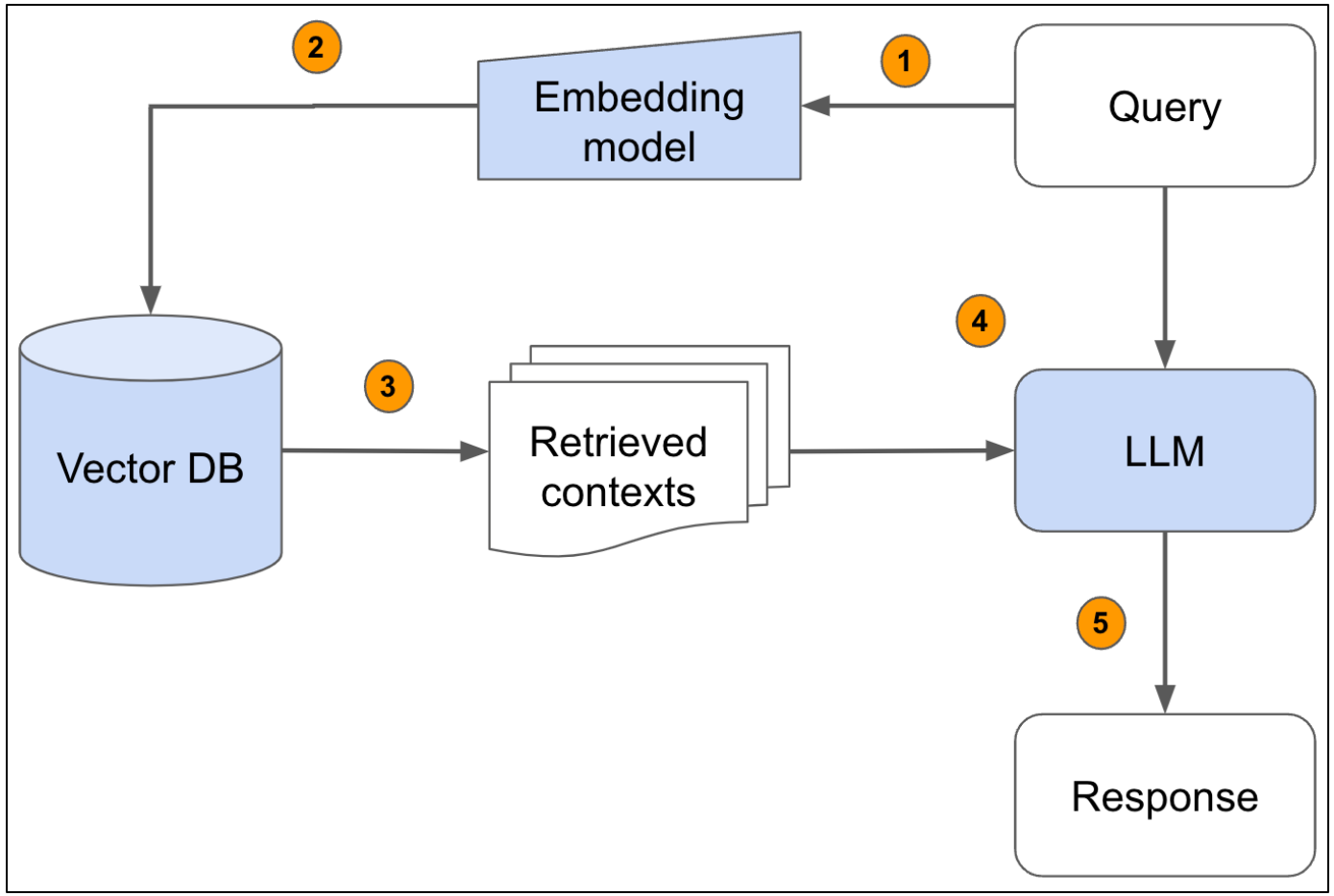
- Augments LLMs with external information retrieval
- Retrieves relevant text chunks from documents before generation
- Keeps outputs grounded in actual source material
- Useful for research questions, internal corpora, and long documents



https://medium.com/@sahin.samia/what-is-retrieval-augmented-generation-rag-in-llm-and-how-it-works-a8c79e35a172

RAG Architecture

- User Query → Embedding → Vector Search → Contextual Prompt → LLM Output
- Key Components:
  - Embedder: Converts query into vector space
  - Retriever: Finds relevant chunks from vector DB
  - LLM: Generates based on retrieved context



https://medium.com/@sahin.samia/what-is-retrieval-augmented-generation-rag-in-llm-and-how-it-works-a8c79e35a172

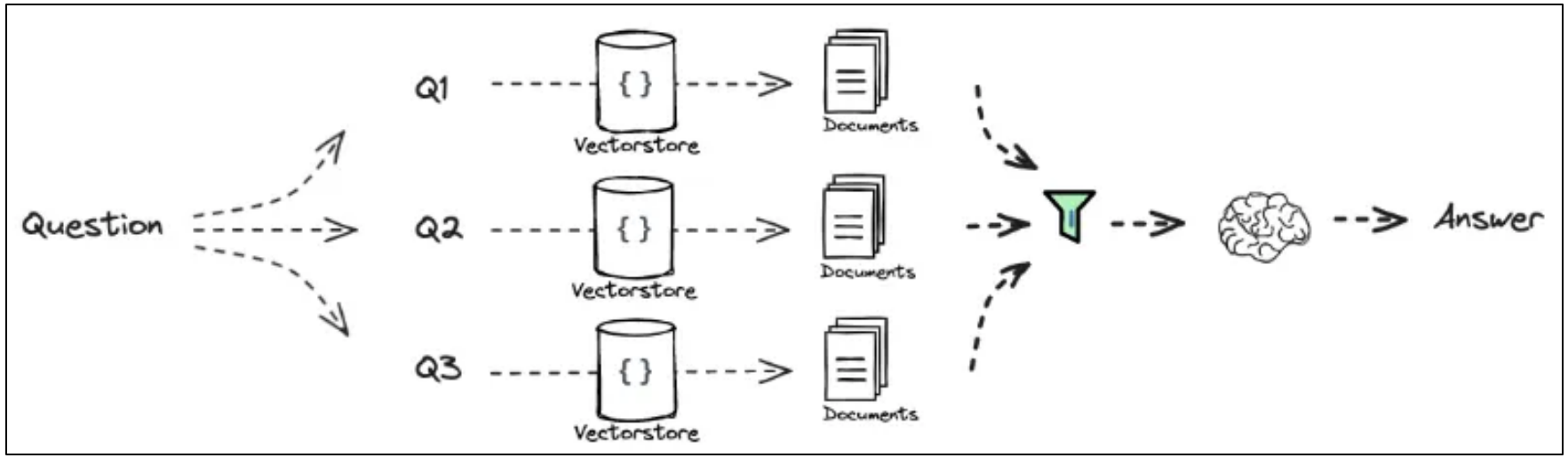
Basic RAG Implementation

1. Convert document to embeddings
2. Store embeddings in vector DB
3. Convert user query into embedding
4. Retrieve top relevant chunks
5. Inject relevant chunks into prompt for LLM

https://medium.com/@sahin.samia/what-is-retrieval-augmented-generation-rag-in-llm-and-how-it-works-a8c79e35a172

RAG Fusion

- Improves RAG results using multiple queries similar to the original
- Fuses results from different perspectives: original query, rephrased queries, related follow ups
- Aggregates top results before generation
- RAG fusion output tends to be more comprehensive than traditional RAG



https://gautam75.medium.com/unlocking-the-power-of-rag-fusion-53437e77d9c2

RAG Citations

- RAG enables LLMs to cite retrieved sources for transparency
- Citations help verify outputs, reduce hallucinations
- Supports trust in legal, academic, and scientific domains
- Format varies by platform (e.g., Claude vs. OpenAI)

https://docs.anthropic.com/en/docs/build-with-claude/citations

Large Document Processing

- RAG is especially useful for large corpora (manuals, research papers)
- Enables chunked retrieval: splits long docs into indexed segments
- LLM only sees relevant slices of content
  - Keeps context window lean
  - Reduces unnecessary token usage (cost)
- Use cases:
  - Summarization
  - Semantic Search

https://medium.com/@ricrivero3/implement-large-document-search-using-rag-semantic-search-with-a-vector-database-and-llm-b798675dca08

Assistants

- Go by many names: assistants, GPTs, copilots
- Combine LLM chat with:
  - Custom system instructions
  - Domain-specific knowledge bases
- Abstracts prompting logic from end user
- Streamlines prompting process

https://openai.com/index/introducing-gpts/

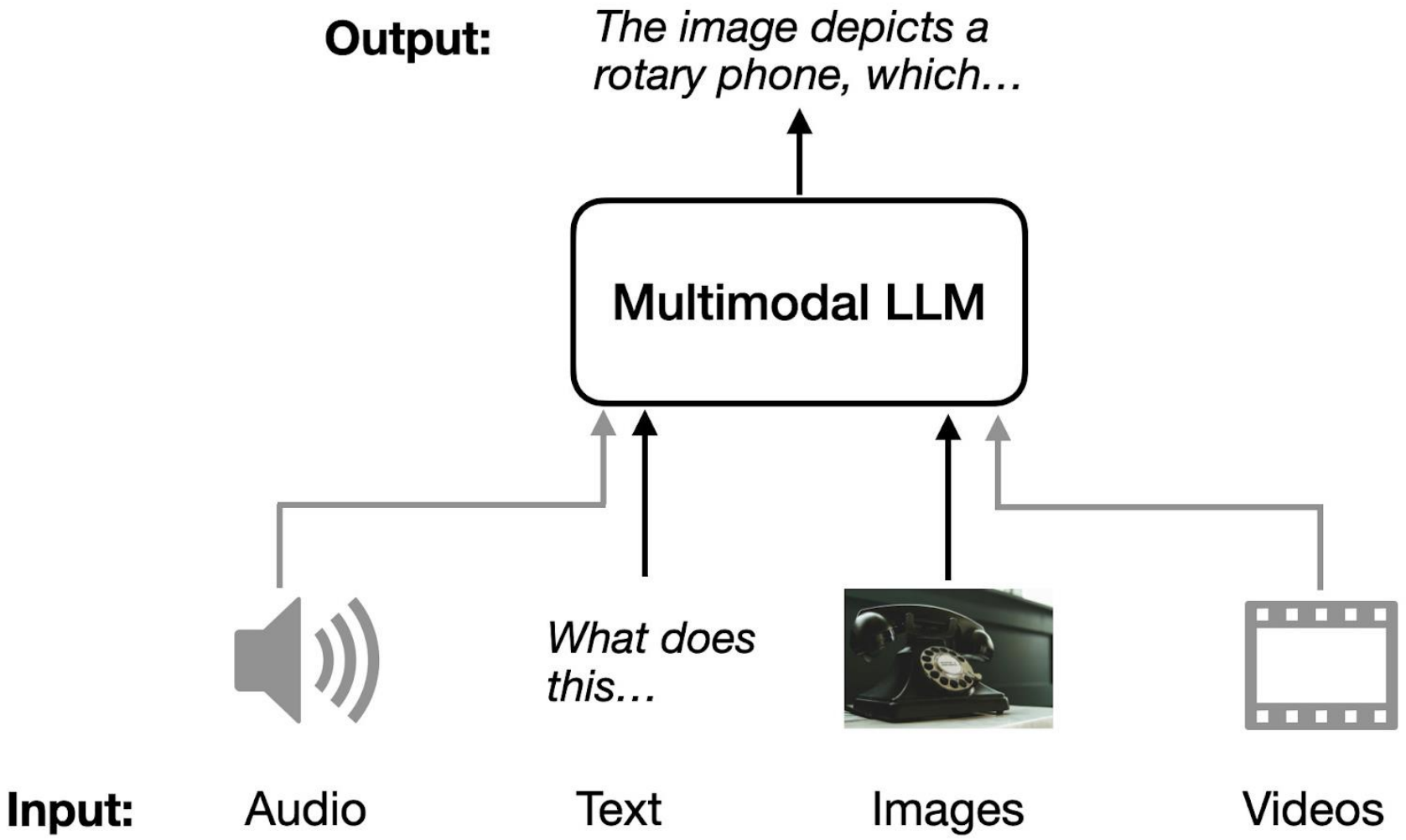
Deep Research

- For multi-faceted, domain-specific inquiries where depth and detail are critical, deep research's ability to conduct extensive exploration and cite each claim is the difference between a quick summary and a well-documented, verified answer that can be usable as a work product
- Prompt examples:
    - Conduct a comprehensive literature review on retrieval-augmented generation (RAG), including recent papers (2023–2025), taxonomy of methods, and open research challenges. Cite all sources.
    - Summarize current legal and ethical frameworks for regulating LLMs in the US and EU. Highlight open policy debates and cite relevant regulations, working groups, or whitepapers.
    - Survey modern LLM-based software engineering tools (e.g., Cursor, Claude Code, Gemini Code Assist). For each, summarize core features, architecture, and developer feedback based on published benchmarks or testimonials.

https://openai.com/index/introducing-deep-research/

Multimodal Capabilities

- Some LLMs support text, image, audio, and video inputs and outputs
- Combines vision, hearing and language for versatile usage



https://medium.com/@tenyks_blogger/multimodal-large-language-models-mllms-transforming-computer-vision-76d3c5dd267f

Visual Understanding

- LLMs can interpret and reason over almost any image
- Enables structured data extraction from unstructured visuals
- Use Cases:
  - Classify, describe or understand the content of images
  - Copy text in a screenshot
  - Turn app design on a whiteboard into documentation or code
  - Transcribe and summarize handwritten notes
  - Extract info from images of complex tables or forms
  - Translate menu from foreign language to English

https://medium.com/@cout.shubham/exploring-multimodal-large-language-models-a-step-forward-in-ai-626918c6a3ec
https://medium.com/@tenyks_blogger/multimodal-large-language-models-mllms-transforming-computer-vision-76d3c5dd267f

Voice & Video Interfaces

- Voice input enables real-time, hands-free interaction
- Useful for accessibility, field work, and mobile agents
- Hypothetical use cases:
  - Real-time navigation assistance for visually impaired users, combining audio questions with visual scene understanding
  - Communication aids for individuals with speech impairments, interpreting partial vocalizations alongside lip movements and gestures
  - Language tutoring
  - Medical training simulations where the AI observes surgical techniques and provides immediate feedback on hand positioning and procedural accuracy

https://openai.com/index/hello-gpt-4o/
https://www.instagram.com/reel/DGeV18UygCo/?igsh=MWNsNnRtMXR5djJ5dg==
https://www.instagram.com/reel/DAq9oO_P1qg/?igsh=dW4yd3V2ajd3dzBx

Image Generation

- GPT-4o includes native image generation, tightly integrated with chat context
- Generates photorealistic, diagrammatic, or surreal imagery with high fidelity
- Supports multi-turn refinement—images evolve through natural conversation
- Use Cases:
  - Edit images
  - Generate professional headshot from one image

https://openai.com/index/introducing-4o-image-generation/
https://www.youtube.com/watch?v=Sp6K3qpVFO0
https://www.artnews.com/art-news/news/signatures-lensa-ai-portraits-1234649633/

Voice Generation

- Text-to-Speech (TTS) synthesis using neural vocoders and diffusion models
- Voice cloning from minimal audio samples (few-shot learning)
- Real-time voice conversion with preservation of prosody and emotional tone
- Multilingual synthesis with cross-lingual voice transfer capabilities
- Use Cases:
  - Personalized voice assistants and accessibility tools
  - Content localization and dubbing
  - Interactive storytelling and gaming
  - Therapeutic applications

https://www.youtube.com/shorts/4p-YbyHJsi4
https://www.youtube.com/watch?v=uJ7_uetMtdo
https://www.tiktok.com/@ifonlytimeline_youtube/video/7519629564855864590?_r=1&_t=ZT-8y6Dktwgh82

Video Generation

- Veo 3 (Google DeepMind) is the current state-of-the-art in AI video generation
- Generates high-resolution video with synchronized audio from text or images
- Trained on large, annotated multimodal corpora (e.g., YouTube-scale)
- Capable of:
  - Cinematic CGI scenes with in-world sound
  - Photorealistic and artistic styles
  - Realistic HUDs, camera work, and ambient audio
- Far exceeds open-source alternatives in visual quality, prompt adherence, and audio alignment

https://www.digitalocean.com/community/conceptual-articles/veo3-next-generation-video
https://www.instagram.com/reel/DKPu4MppDnj/
https://www.tiktok.com/@declassifriedtv/video/7512332686158023979?_r=1&_t=ZT-8y6DnT3yBq3
https://www.tiktok.com/@stormtrooper.vlogz/video/7513091886882622751?_r=1&_t=ZT-8y6DzVplDTx
https://www.instagram.com/reel/DMBbYPrBCE6/?igsh=MXNjY2h4bWc5b2NtYQ%3D%3D

Feedback & Project Requests