

# Why Context Matters?

Retrieval-Augmented Generation

# Context Window: The Model's Memory



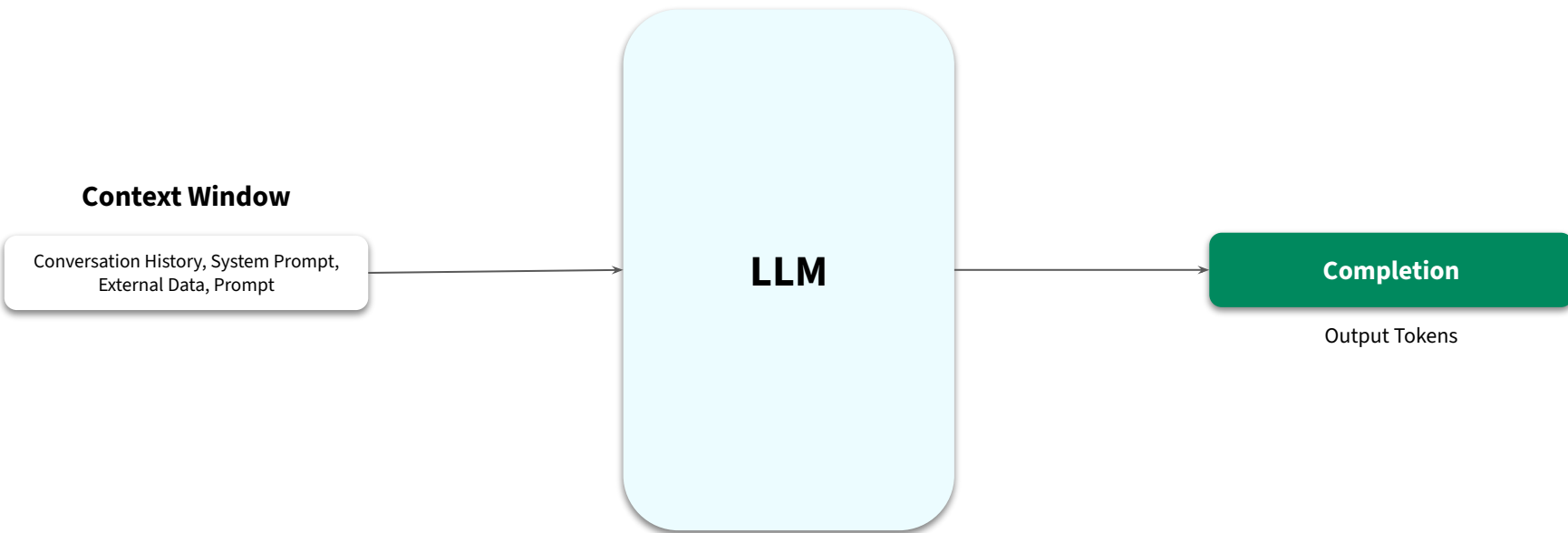
## Context Window

Conversation History, System Prompt,  
External Data, Prompt

**LLM**

**Completion**

Output Tokens





**Good Context = Good Results**



Does the new 'Aether-5' drone have  
4K video recording?

User

The Aether drone series is known for its excellent  
camera quality. Most modern drones in its class  
support 4K video, so it is highly likely that the Aether-5  
does as well.

Chatbot



Does the new 'Aether-5' drone have  
4K video recording?

User

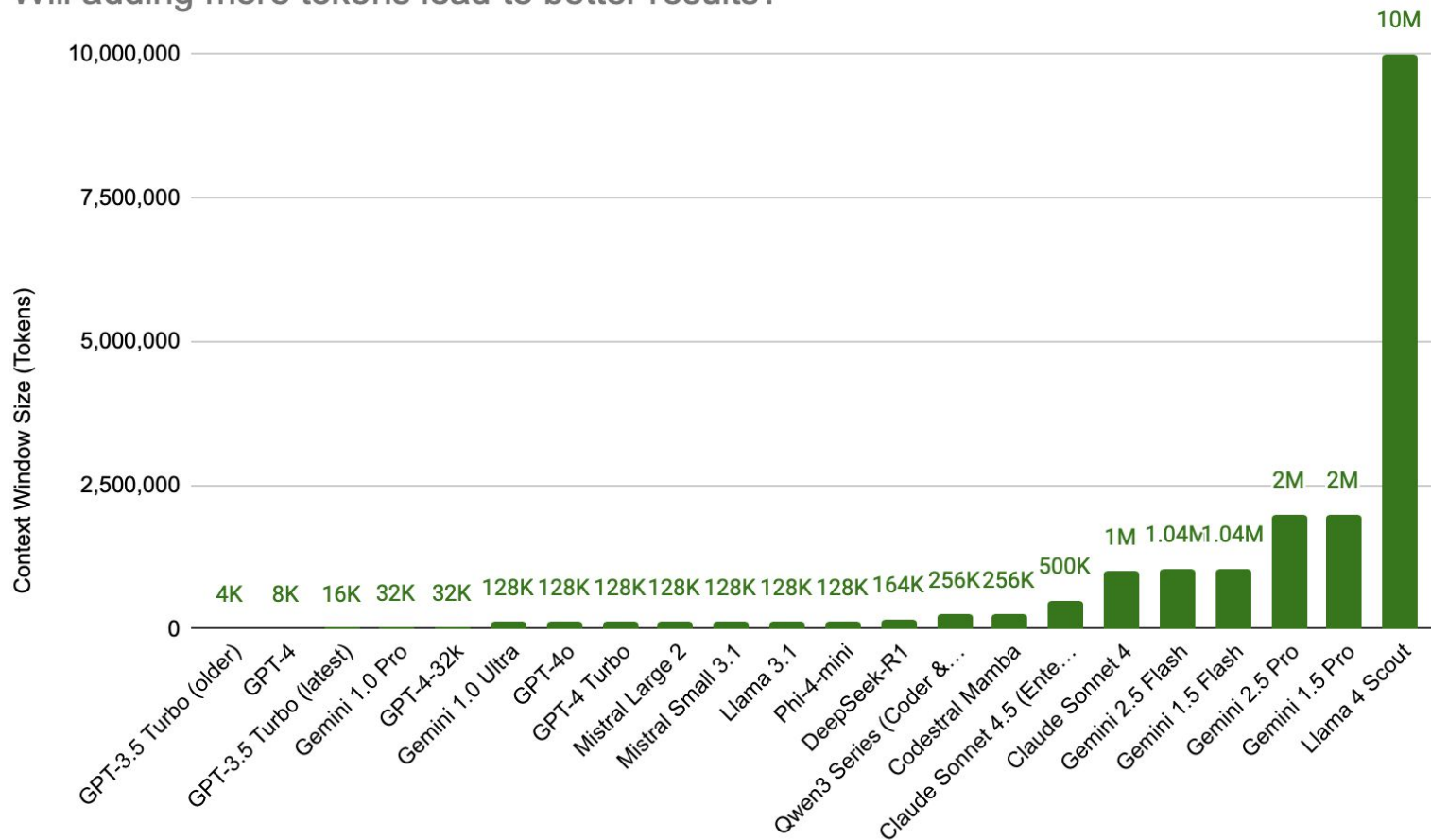
**Context:** The Aether-5 drone, launched in September 2025, features a new 1-inch sensor. This enables video capture in 6K at 30fps and 4K at up to 120fps, a significant upgrade from the Aether-4's 1080p capability

Yes, the Aether-5 drone supports 4K video recording at  
up to 120fps. It can also record in 6K at 30fps.

Chatbot



## Will adding more tokens lead to better results?



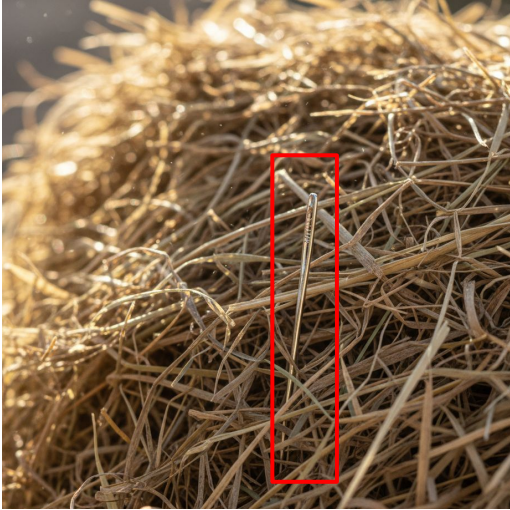


**No.**



## Needle in the Haystack









**Good, Small Context = Good Results**

# When to use RAG

Retrieval-Augmented Generation



## **Business Factors**

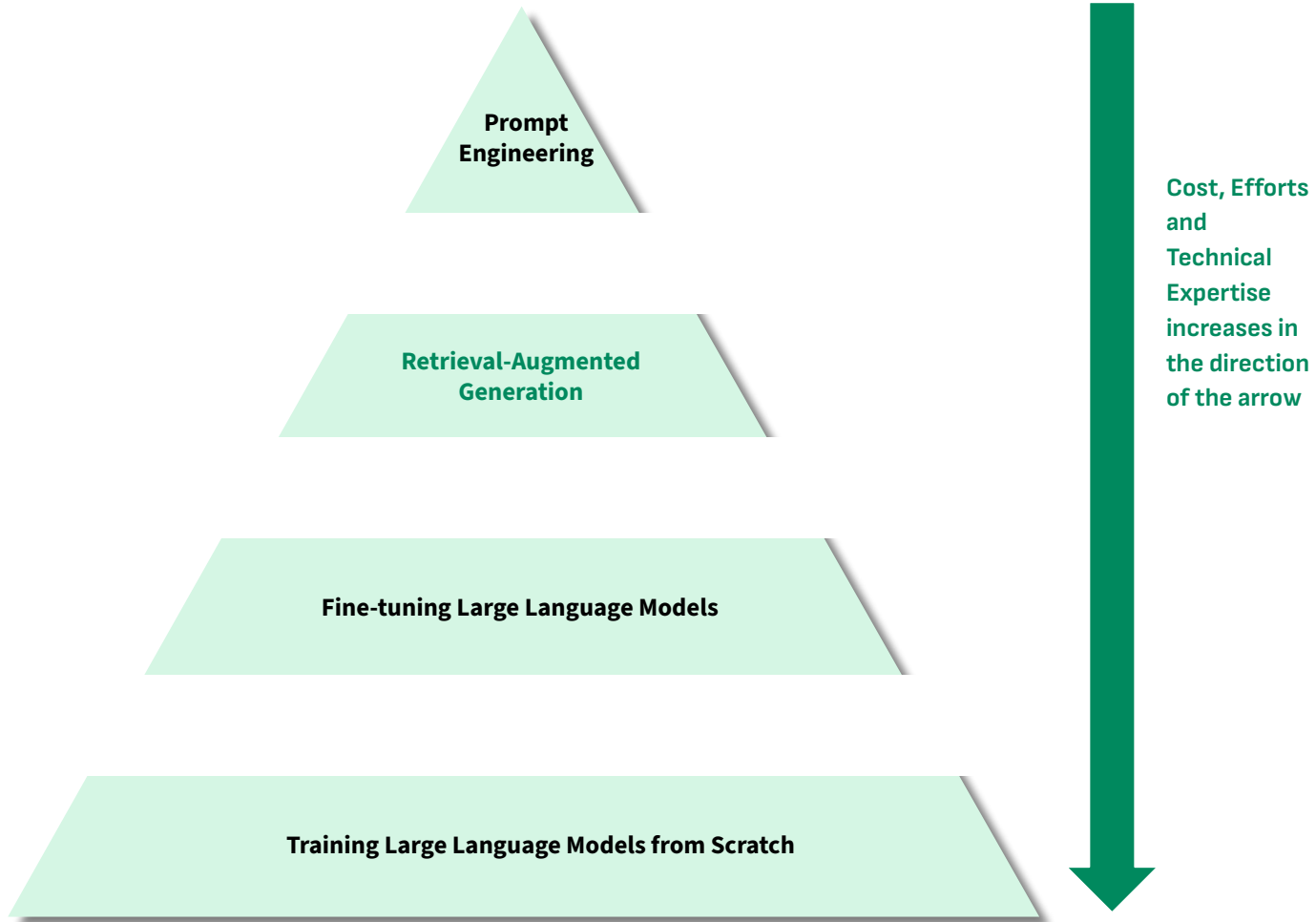


## Business Factors

**Cost**

**Efforts**

**Technical Expertise**



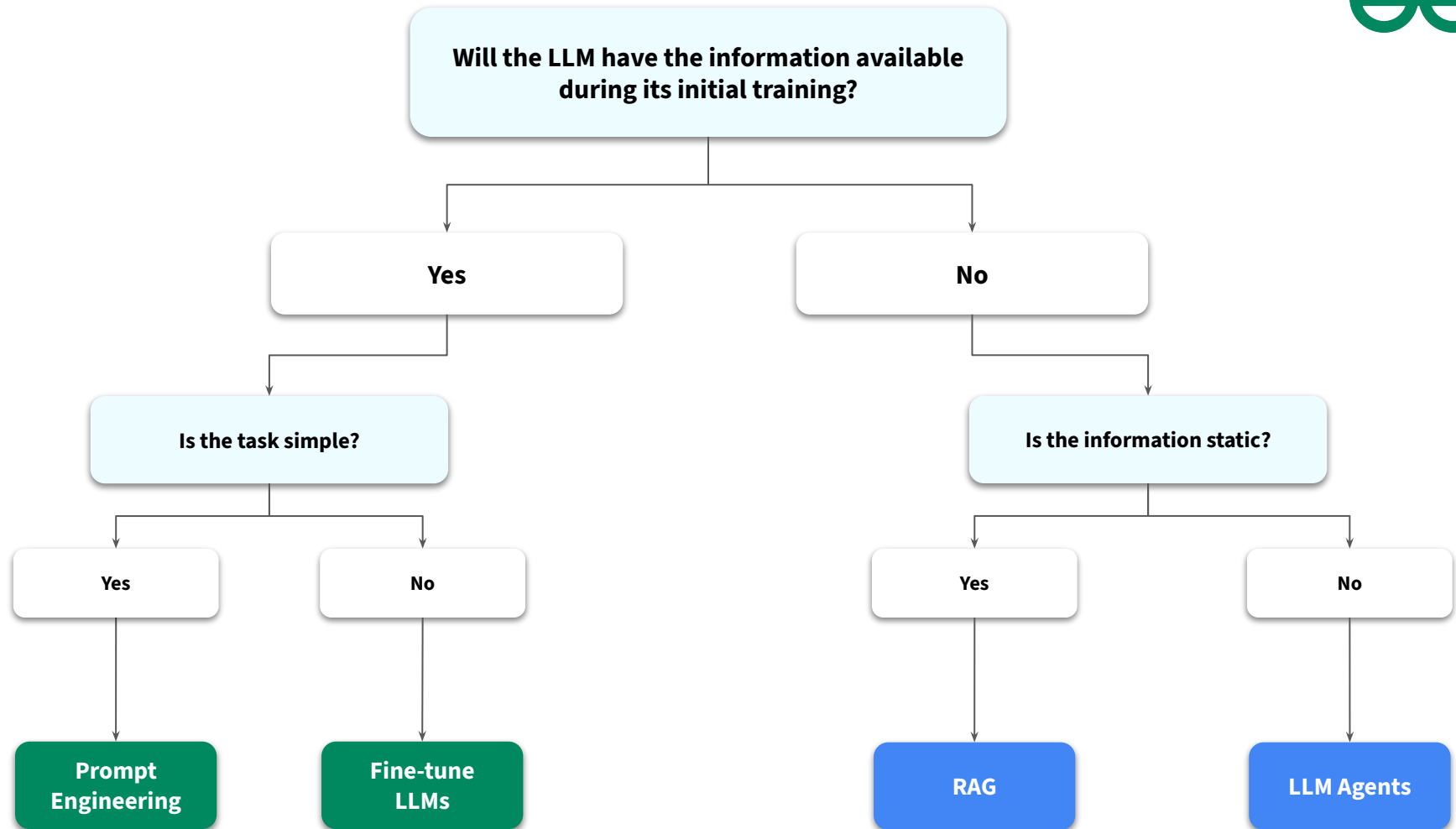


## Technical Perspective





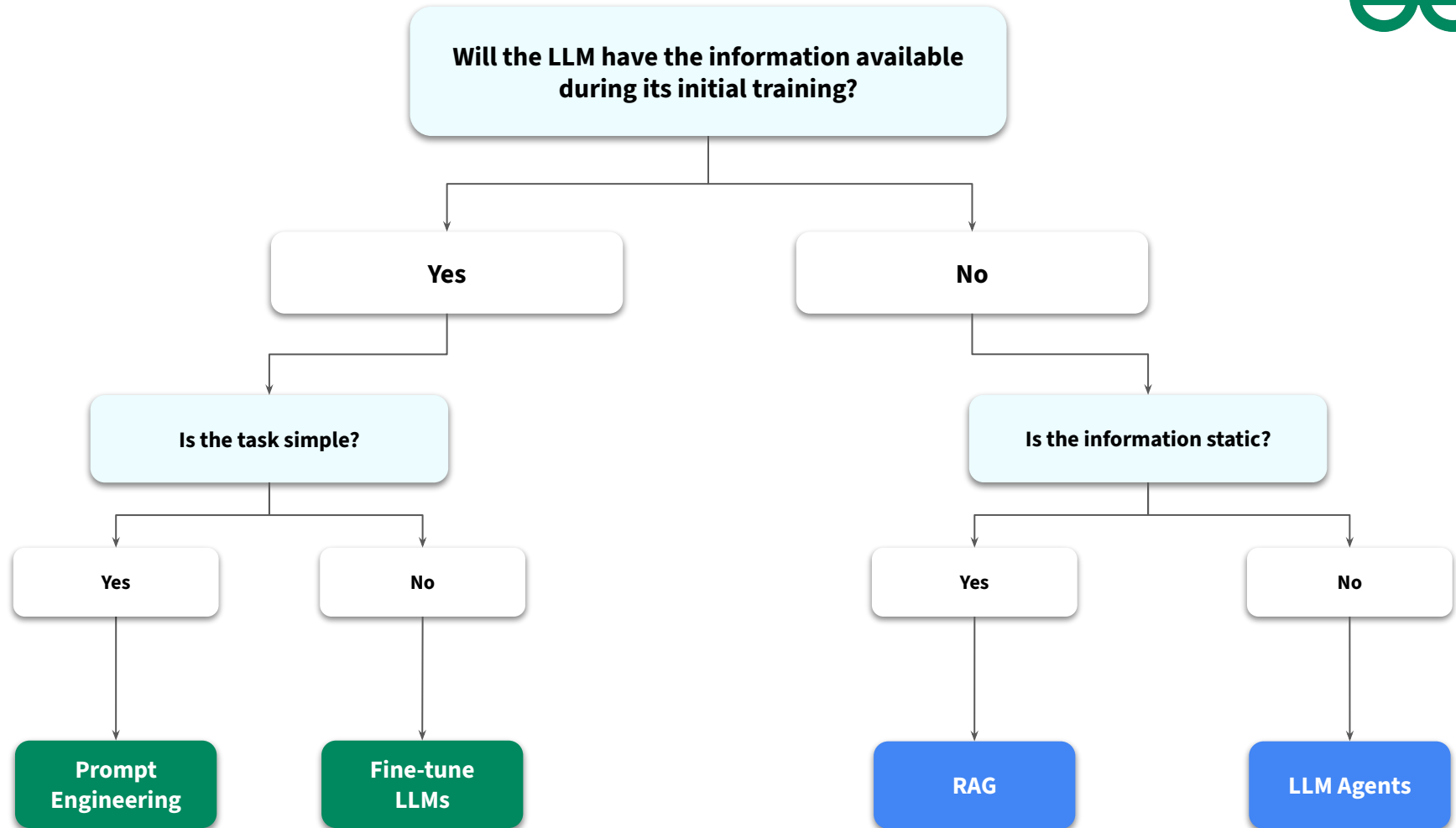
## **Information Required for the Task**

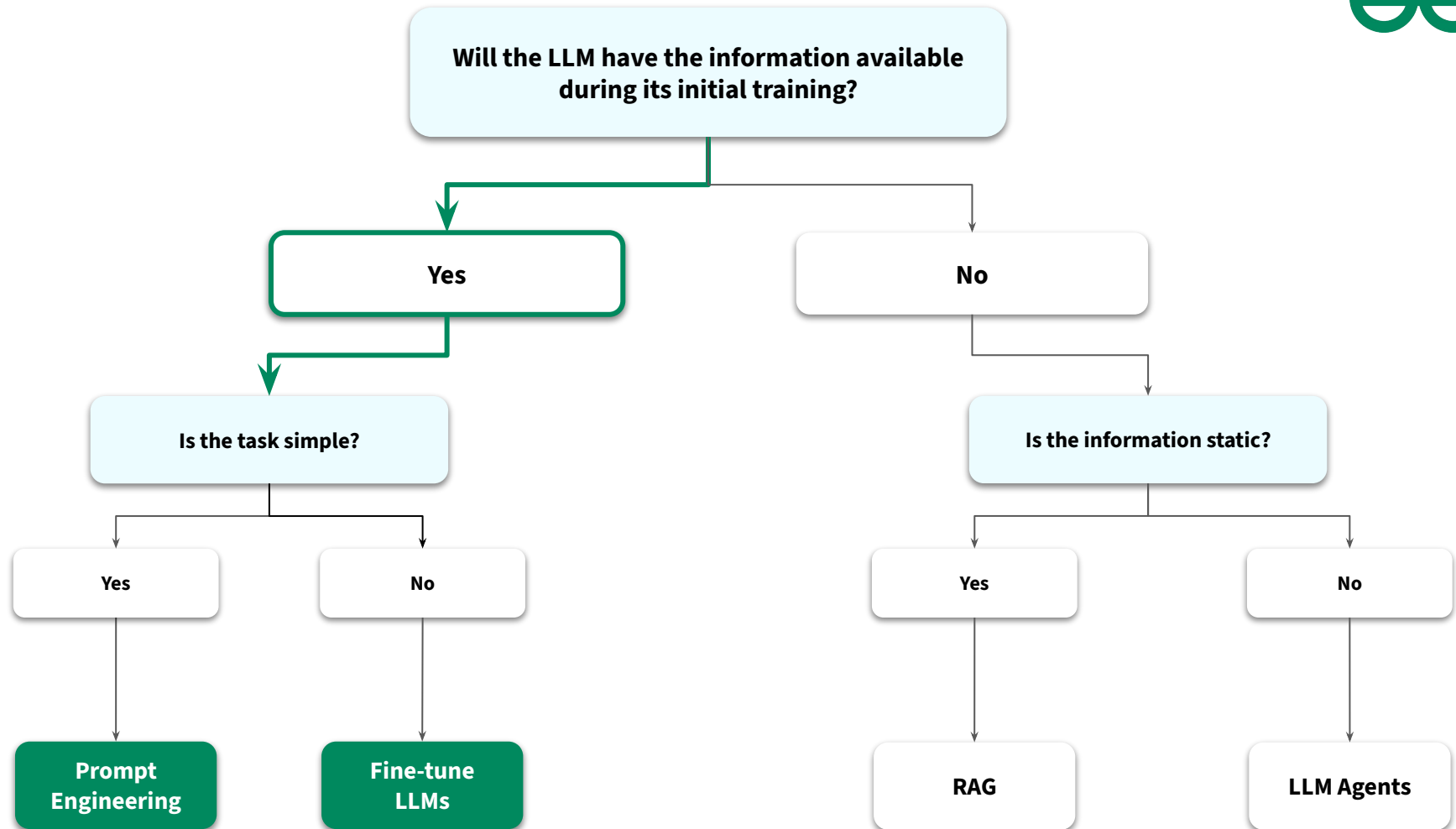


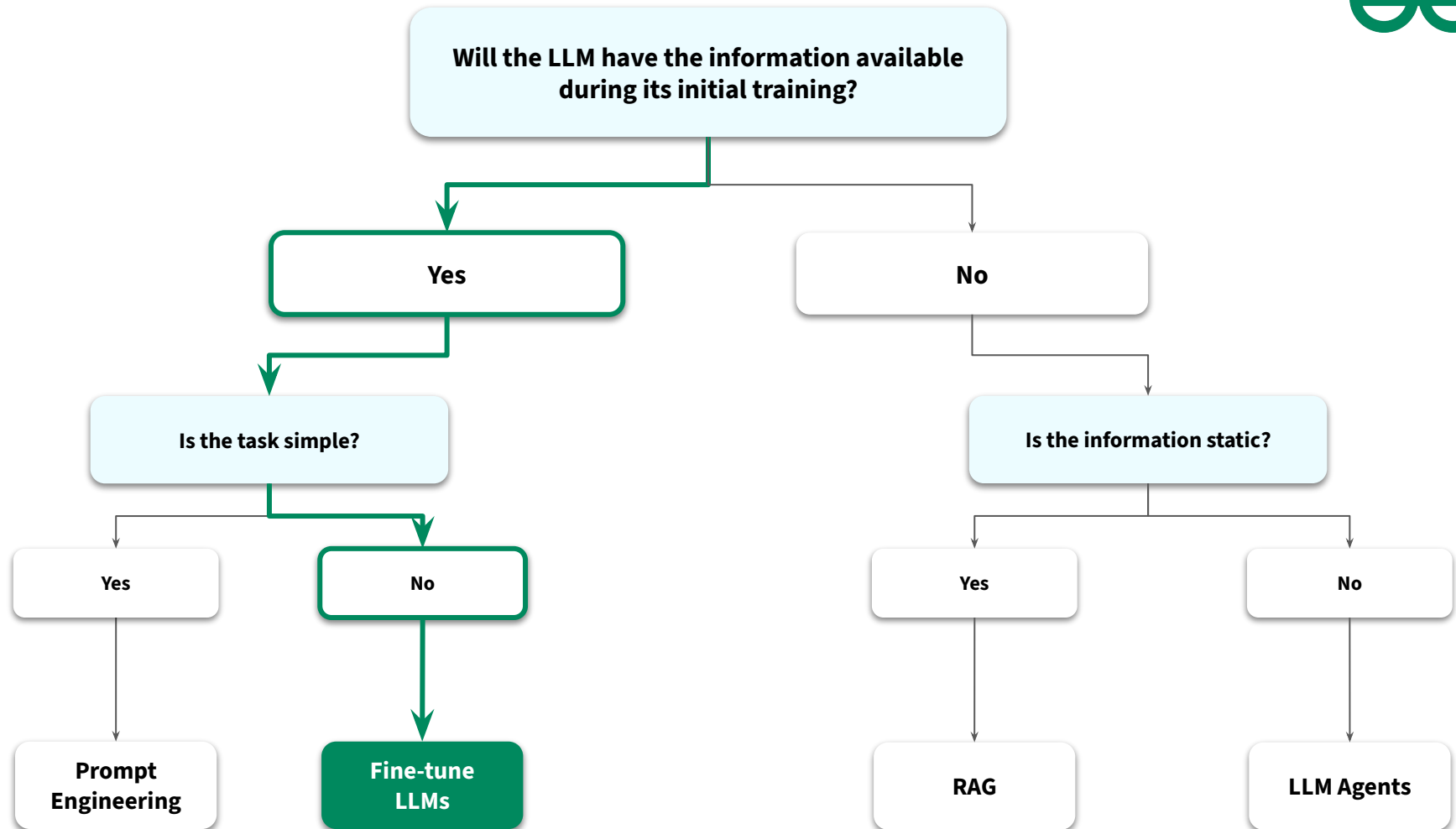


## Scenario One

You want to use LLMs to generate medical diagnosis using medical records



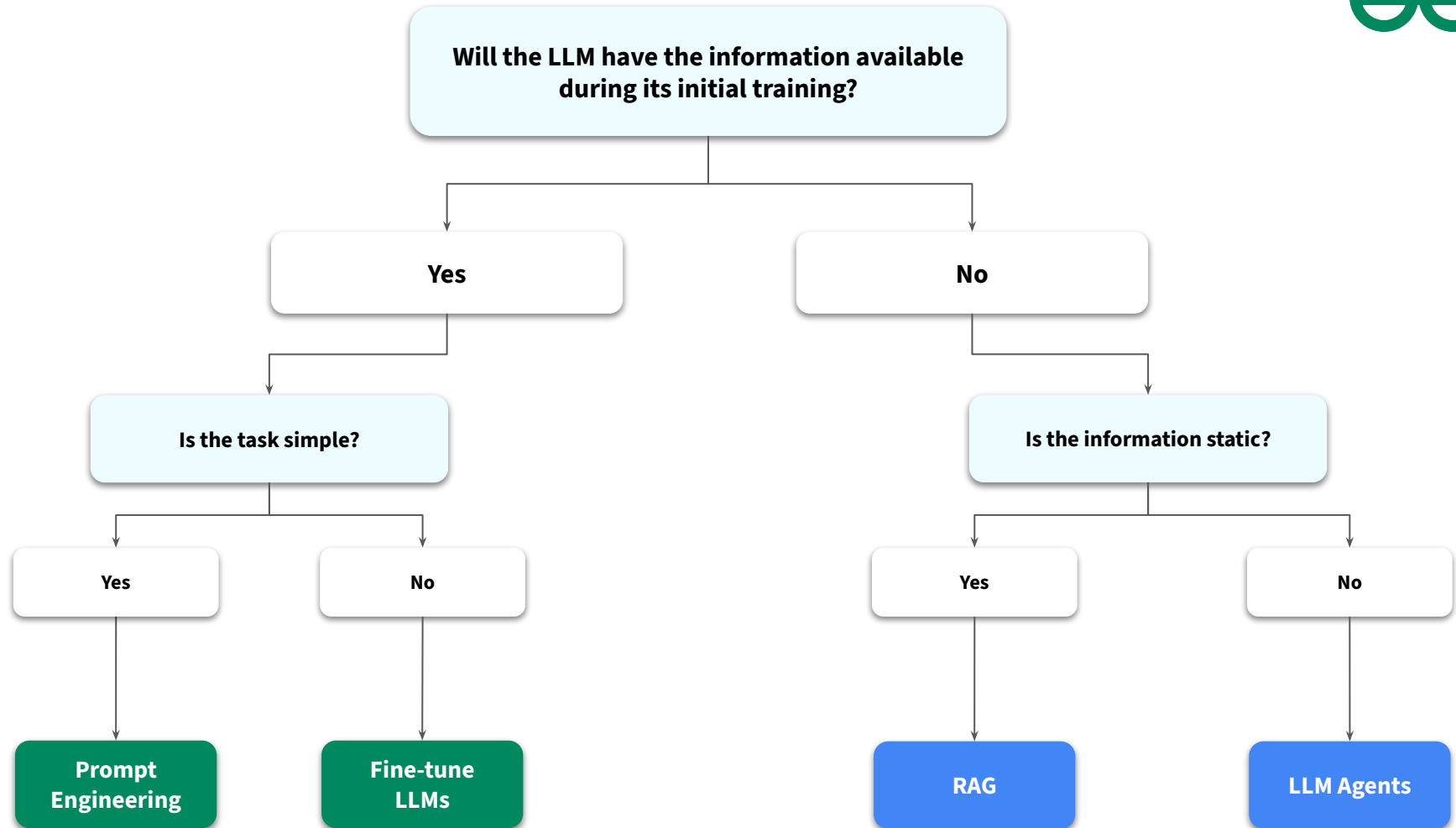




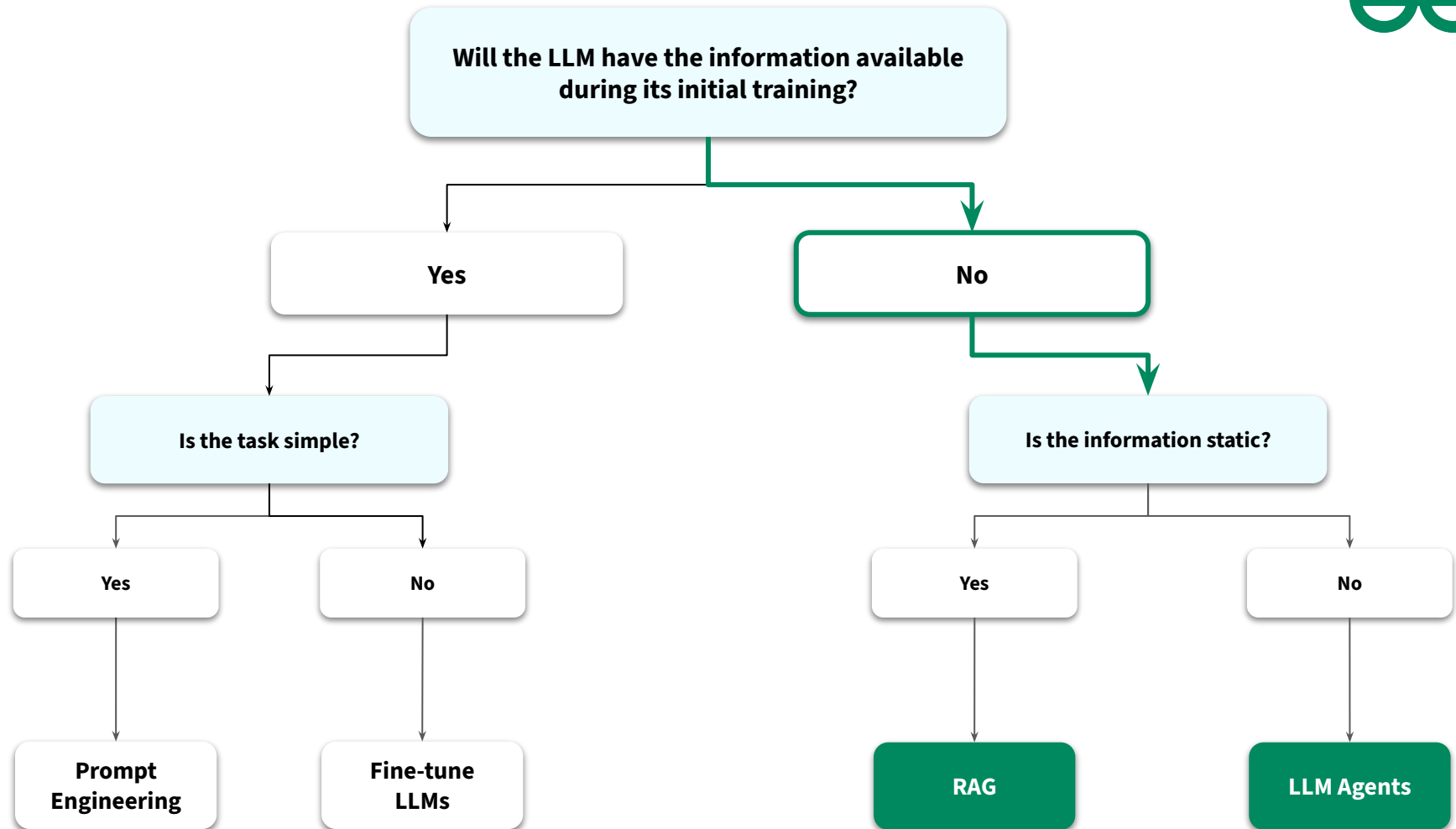


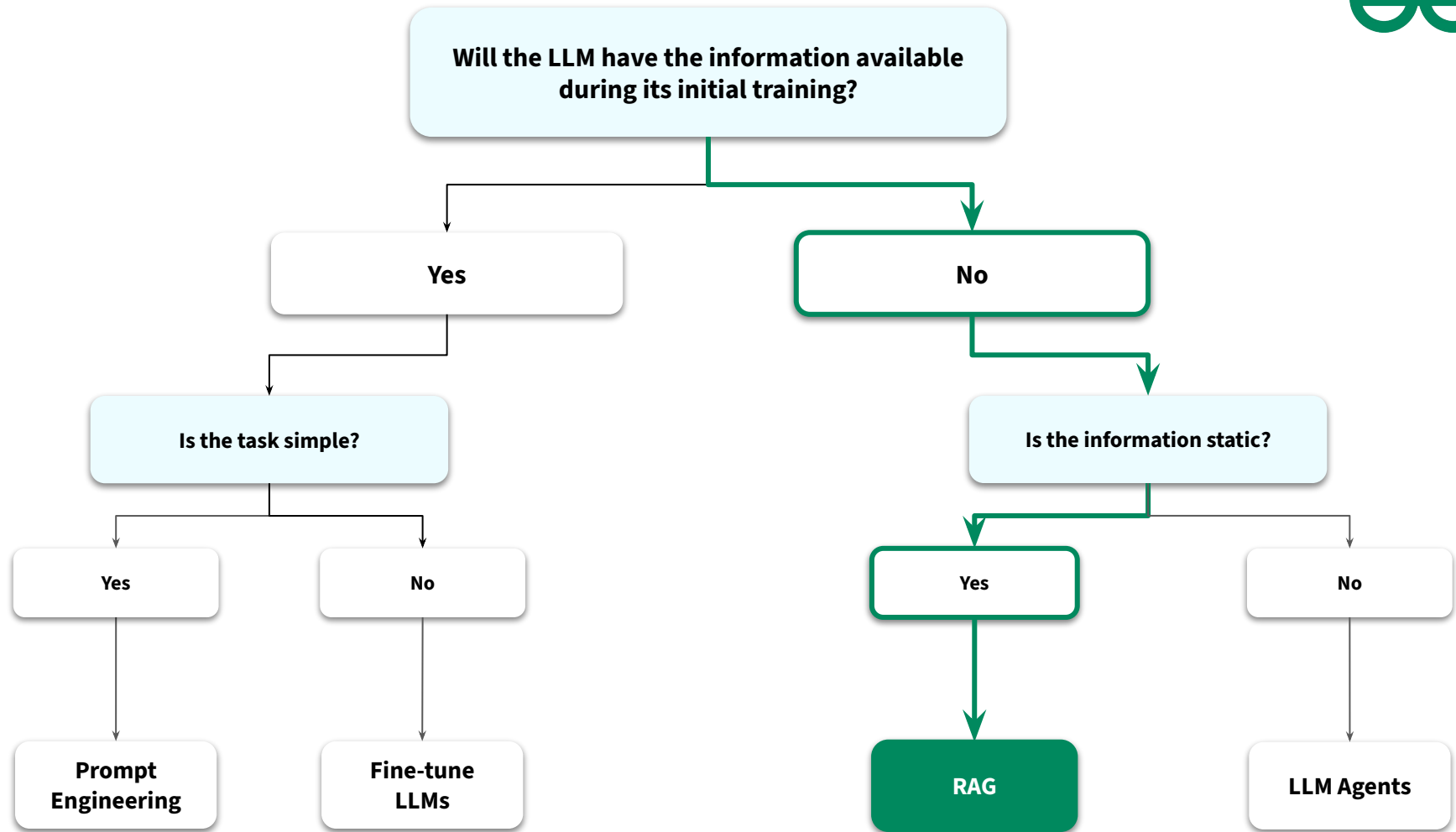
## Scenario Two

You want to develop a chatbot that helps your team access all company policy documents and answers any related questions.







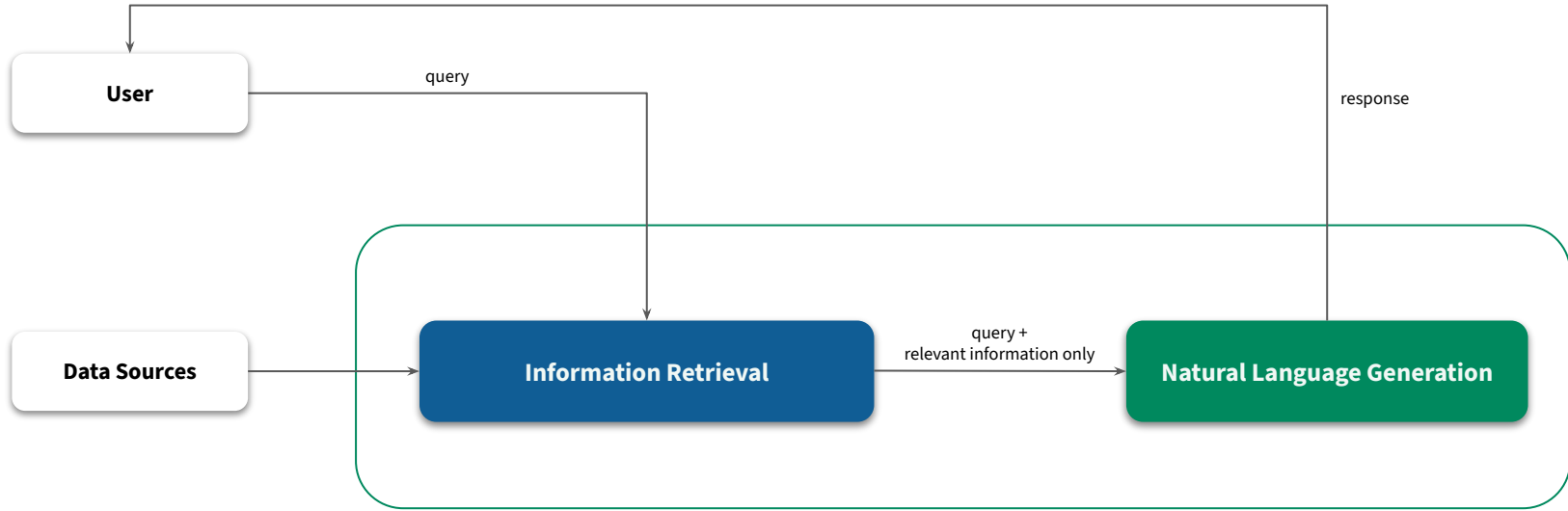


# RAG Deep Dive

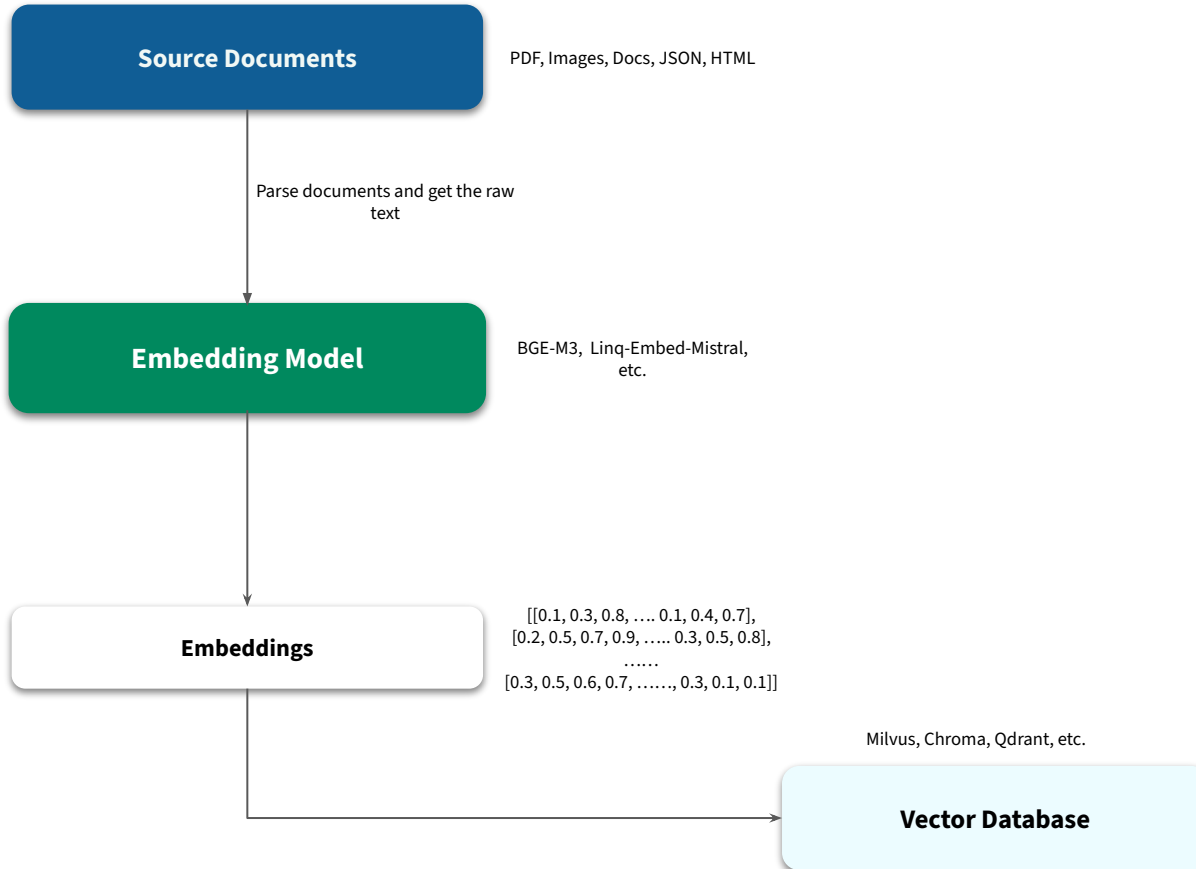
Retrieval-Augmented Generation



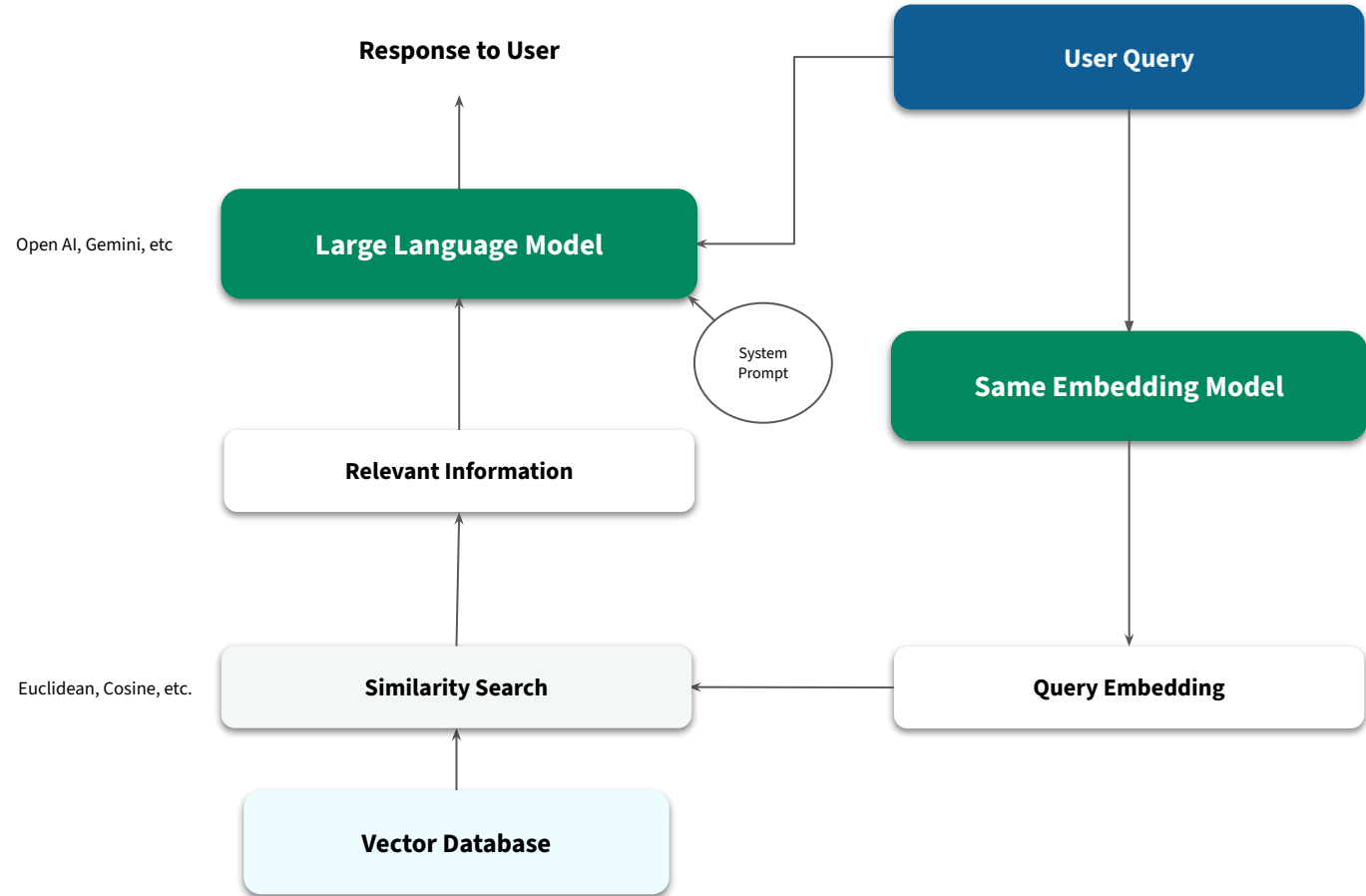
# RAG



# Information Retrieval: The Quest for Relevance



# Natural Language Generation: The Art of Articulation



# RAG Flow

