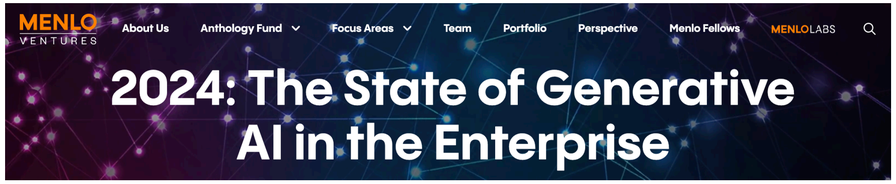
Vector search, embeddings, retrieval augmented generation

David Lee | TIM 175

Today

- Vector embeddings and search
- Retrieval augmented generation
- Project 2 + HW 7 Prelab
- Upcoming deliverables
  - Homework 7 (prelab) due TODAY at 11:59pm
  - Homework 7 (individual) due SATURDAY at 11:59pm
  - Homework 7 (team) due NEXT MONDAY at 11:59pm

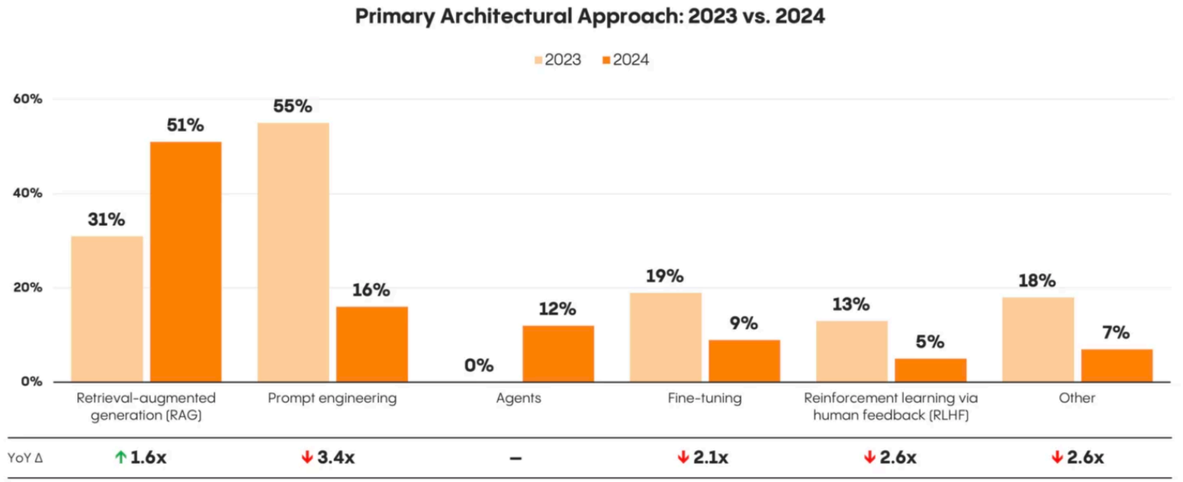


https://menlovc.com/2024-the-state-of-generative-ai-in-the-enterprise/

Design Pattern Trends: RAG Gains, Fine Tuning Is Rare, and Agents Break Out

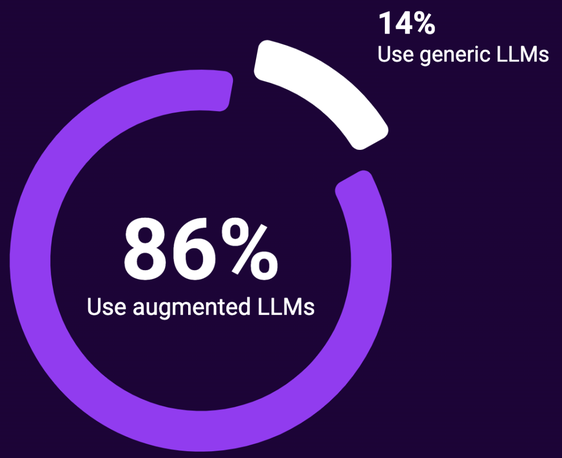
Enterprise AI design patterns—standardized architectures for building efficient, scalable AI systems—are evolving rapidly. RAG (retrieval-augmented generation) now dominates at 51% adoption, a dramatic rise from 31% last year. Meanwhile, fine-tuning—often touted, especially among leading application providers—remains surprisingly rare, with only 9% of production models being fine-tuned.

The year's biggest breakthrough? Agentic architectures made their debut and already power 12% of implementations.



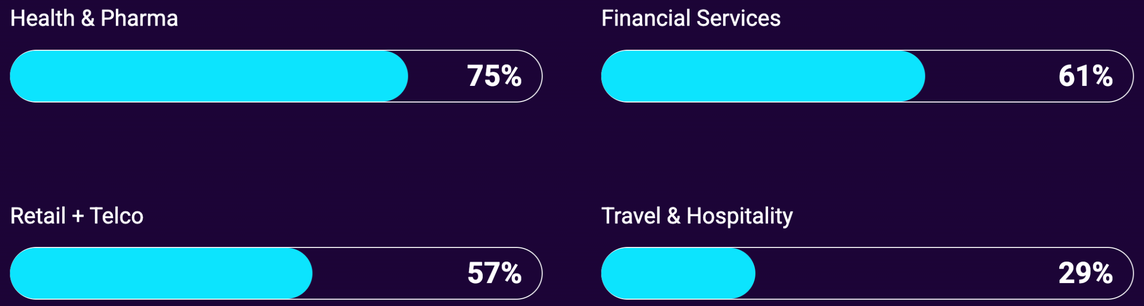
LLM augmentation: A must-have for 86% of organizations

When looking at GenAI adoption, the overwhelming majority—86%—are opting to augment their LLMs, using frameworks like Retrieval Augmented Generation (RAG), recognizing that out-of-the-box models often lack the customization needed to meet specific business needs.



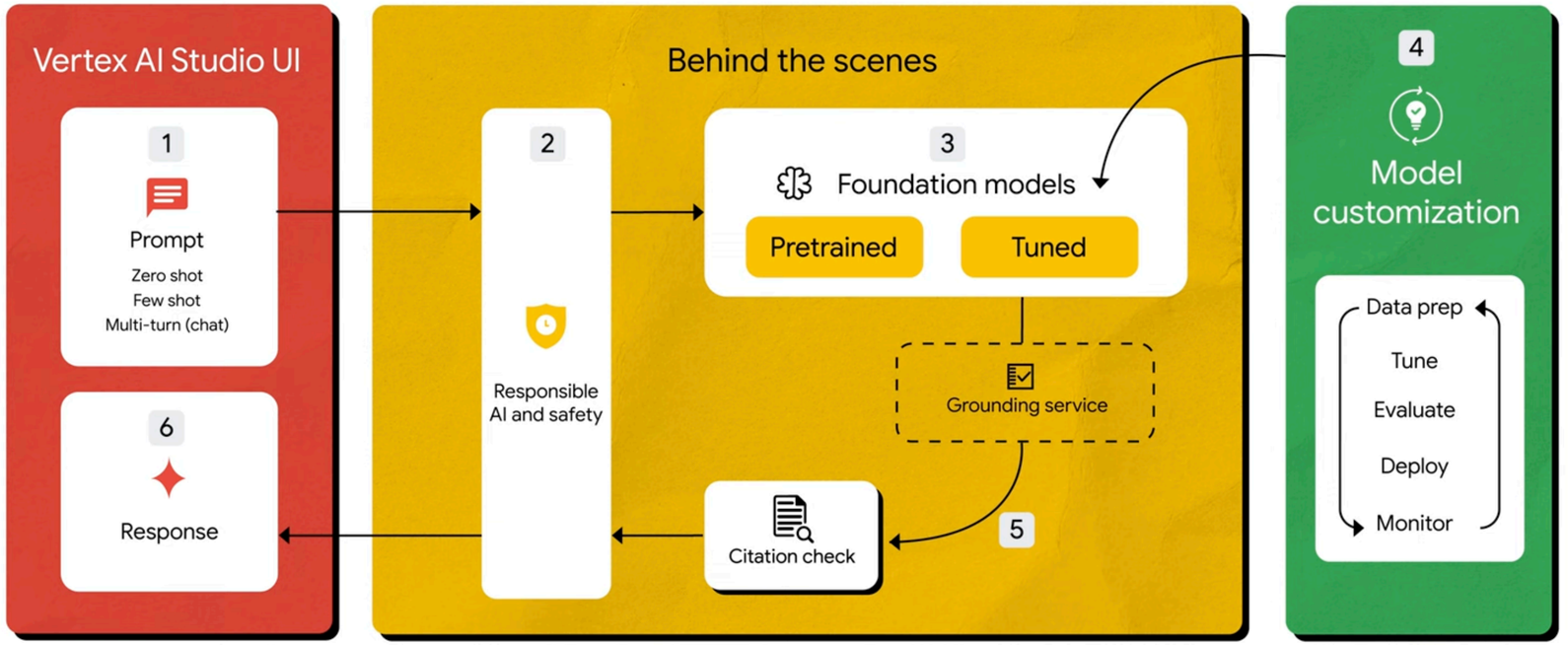
Status of RAG adoption

We see a strong momentum in adopting RAG for pilot projects, particularly in industries where data privacy and response accuracy are critical. The transition to full production is still limited, reflecting the complexities of scaling these technologies.



Vector embeddings, search, and RAG

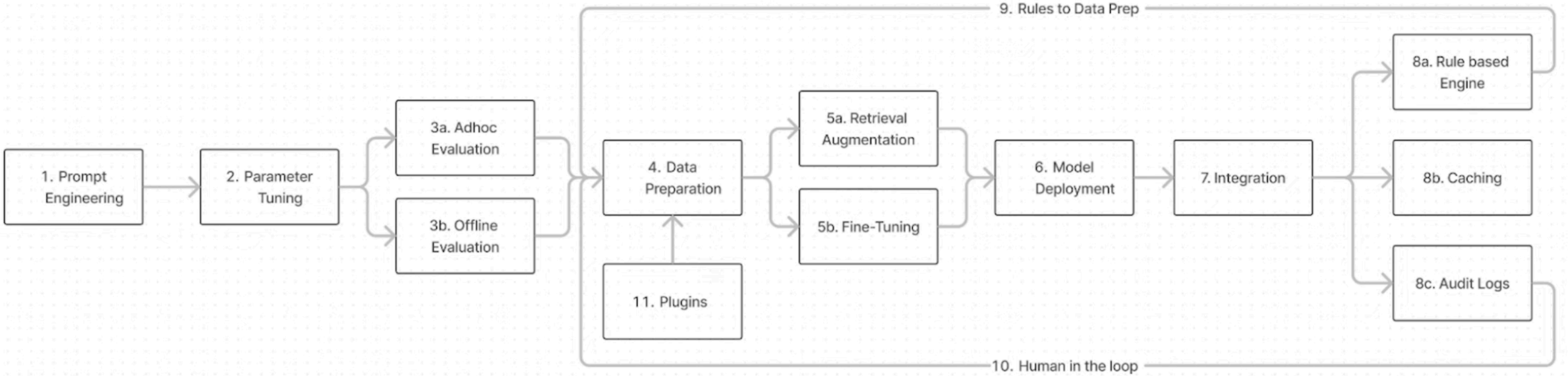
Generative AI workflow on Vertex AI



Introduction to Vertex AI Studio
https://www.cloudskillsboost.google/paths/183/course_templates/552/video/469801

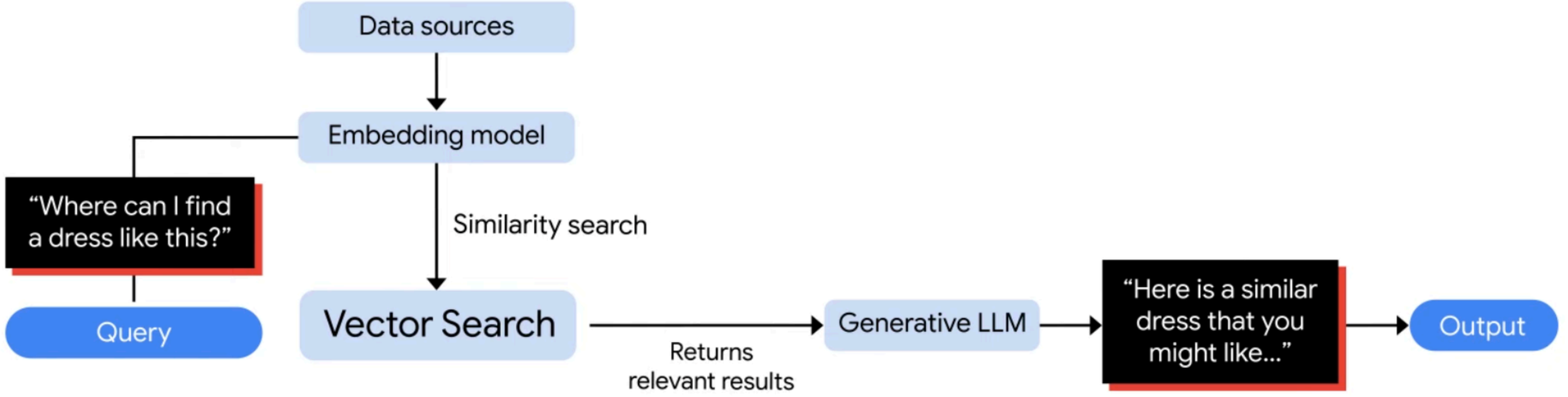
the ***last mile*** problem

It's monumental effort to deploy reliable generative AI to production

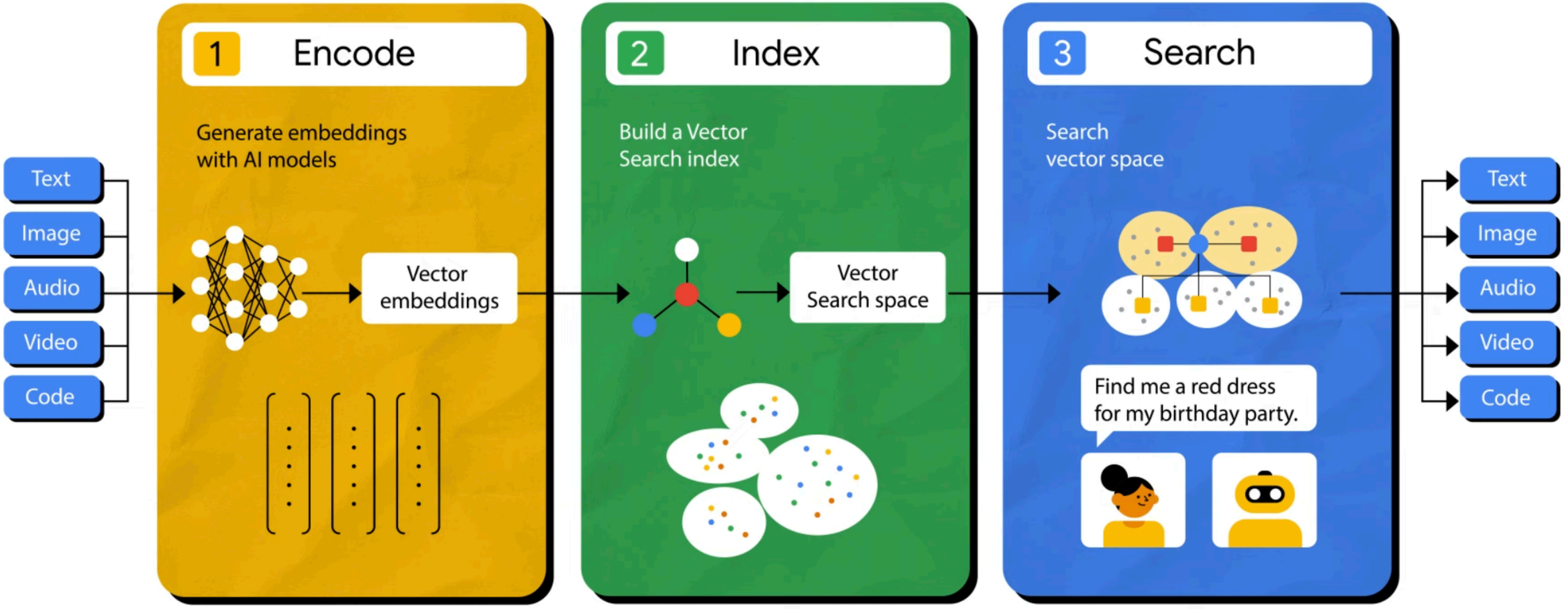


Deploying an application with Generative AI best practices
https://www.youtube.com/watch?v=dRf4DdA1o5c

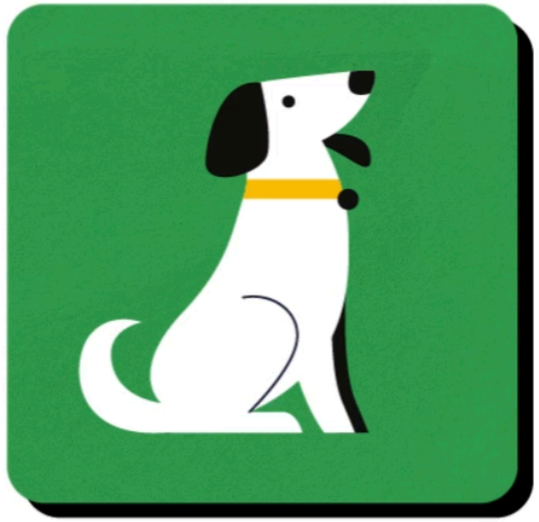
Vector Search is an important component in generative AI applications

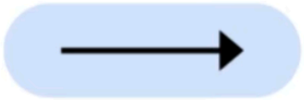


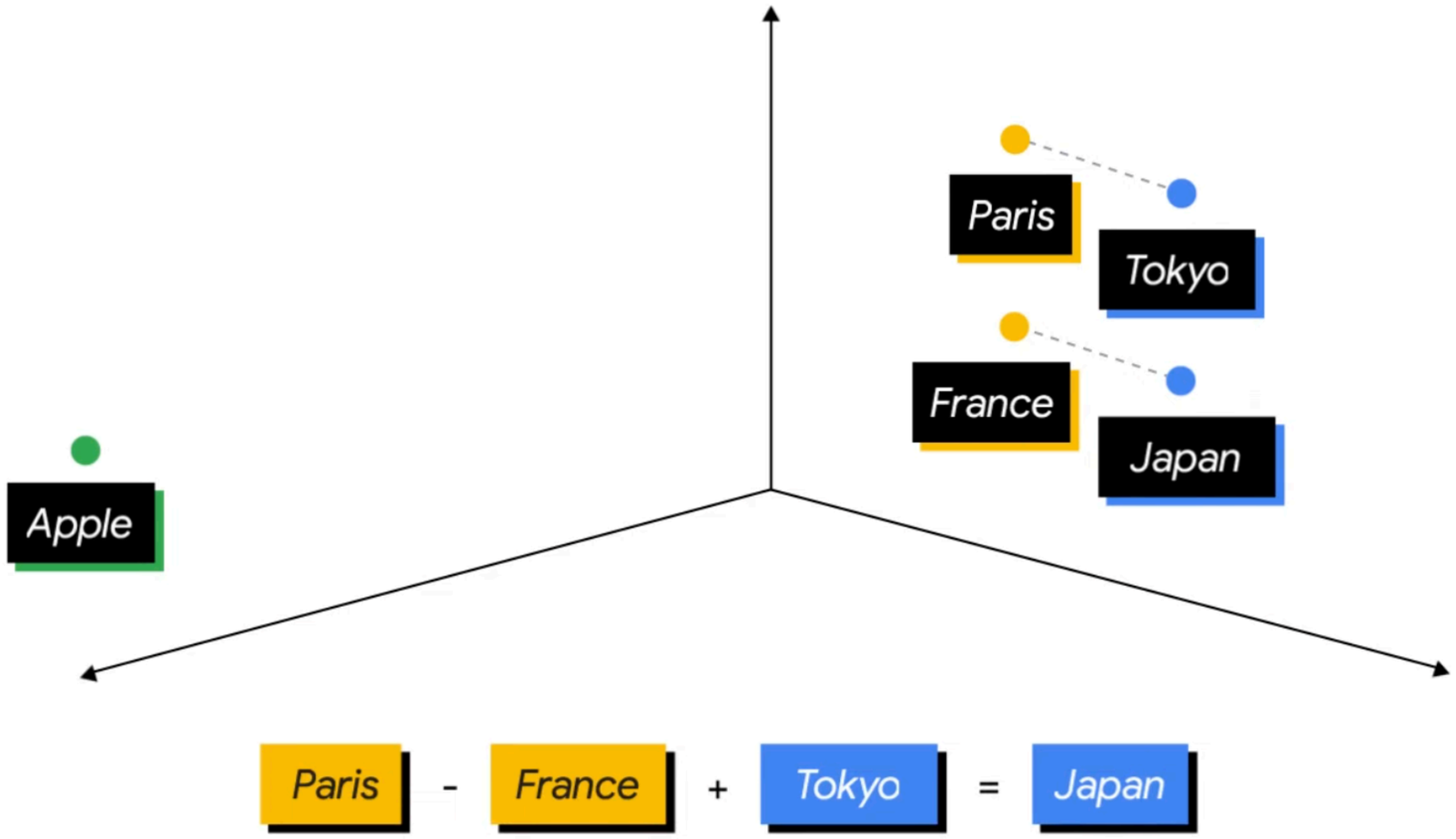
Vector Search process



How do you describe a dog?

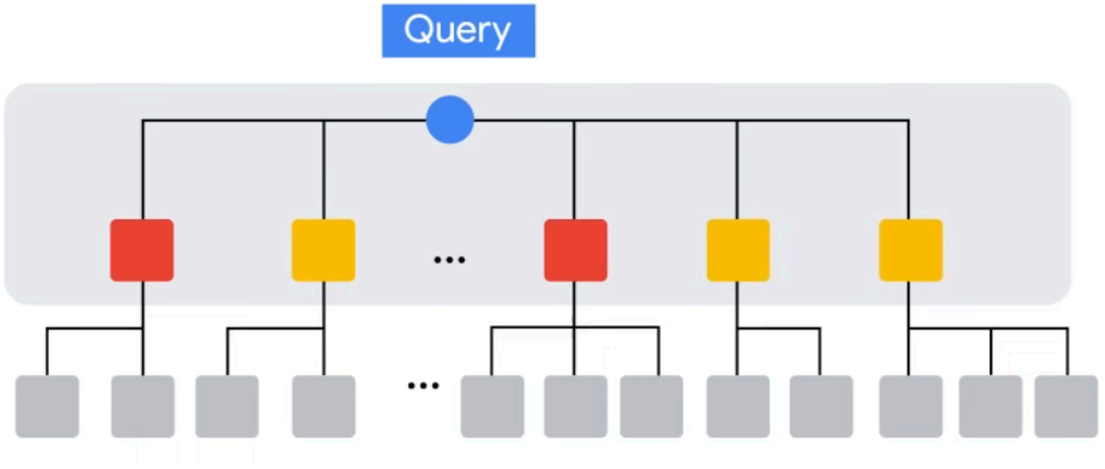





Paris
Tokyo
France
Japan
Apple
Paris - France + Tokyo = Japan

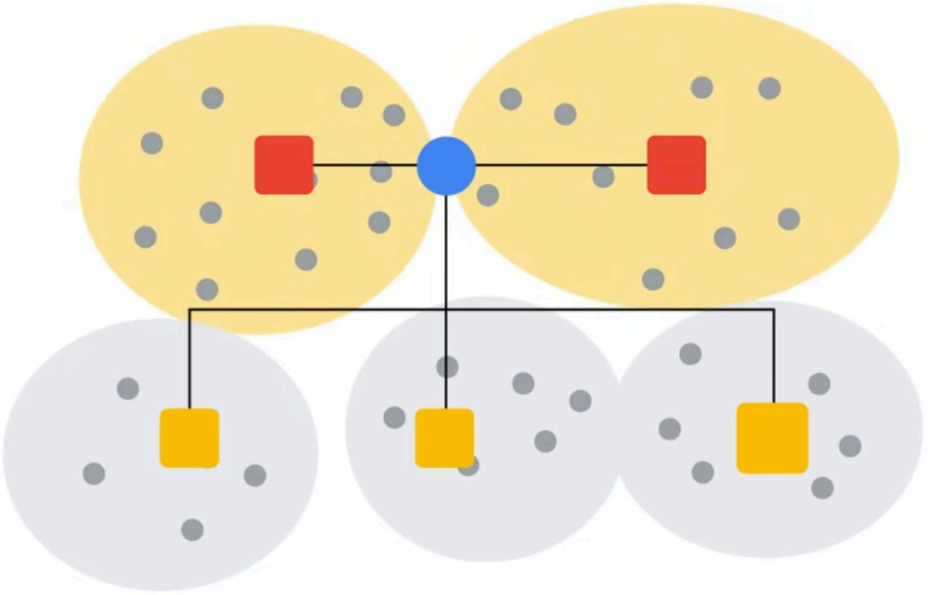
1 Space pruning: Multilevel tree search

Search tree view



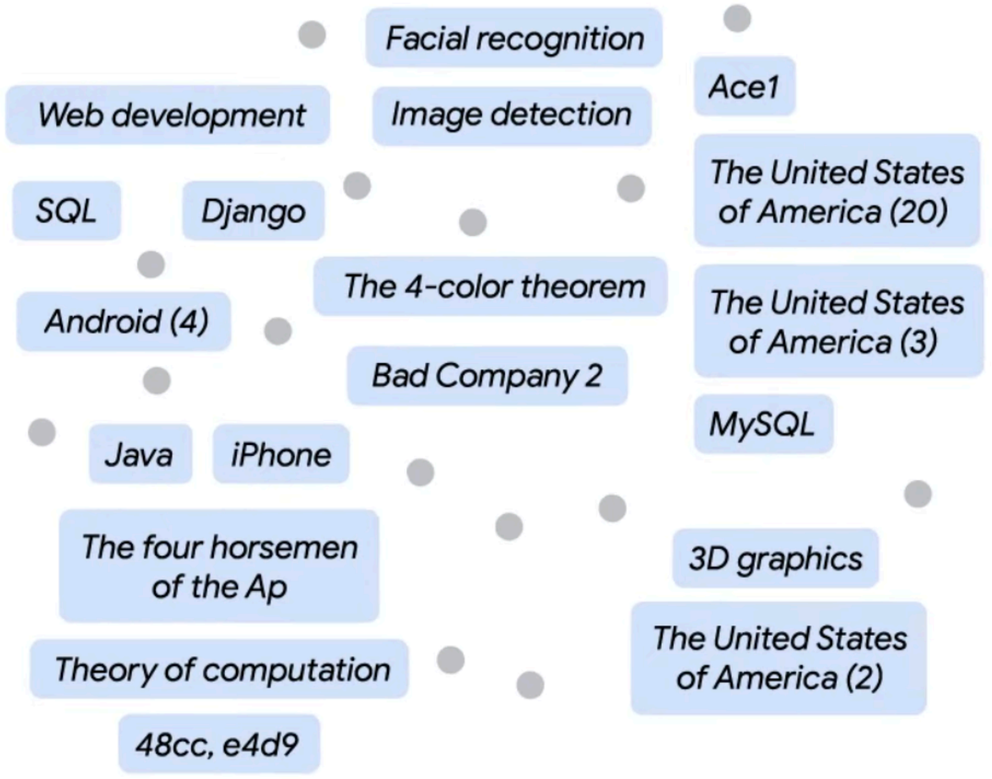
2 Select the closest partition

Vector space view

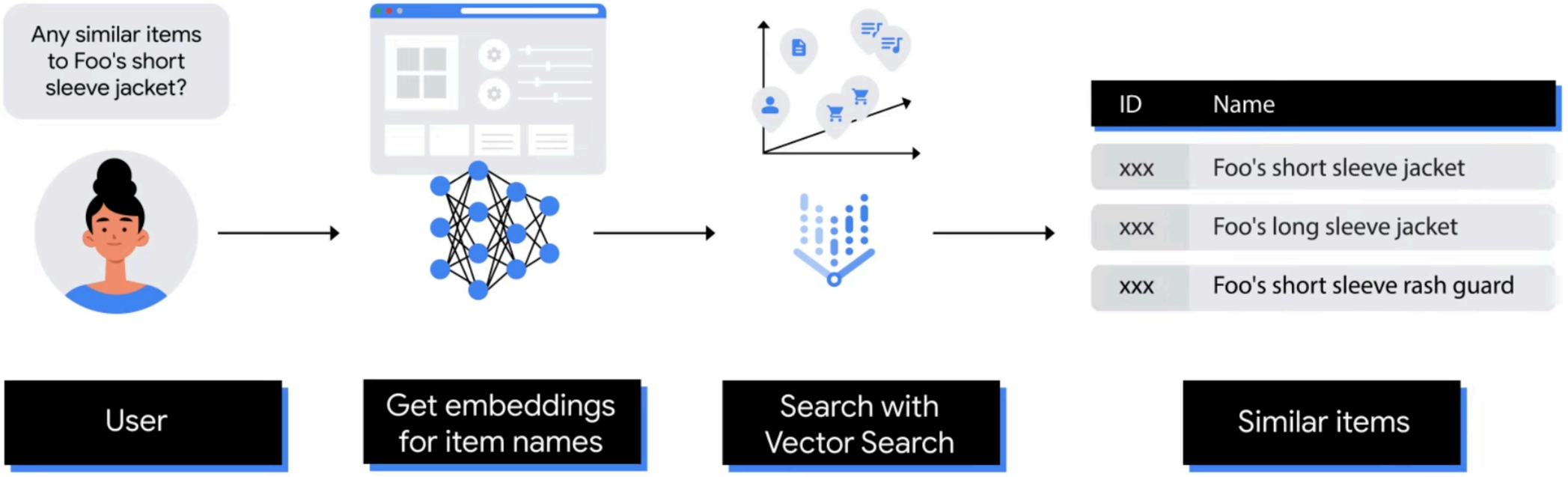


TreeAh algorithm:

Approximate Nearest Neighbor (ANN)



Example: Finding similar items by their names



Typical use cases of Vector Search

Search

Semantic search
Document retrieval, ranking

Multimedia search
Video, music, image search

Personalization

Ads targeting
Keywords targeting, similar user targeting

Recommendation
Candidate generation, filter and diversity

Trust and safety

Trust and safety
Spam, abuse content, label propagation

Content re-use
Copyright detection, video reupload

LLMs

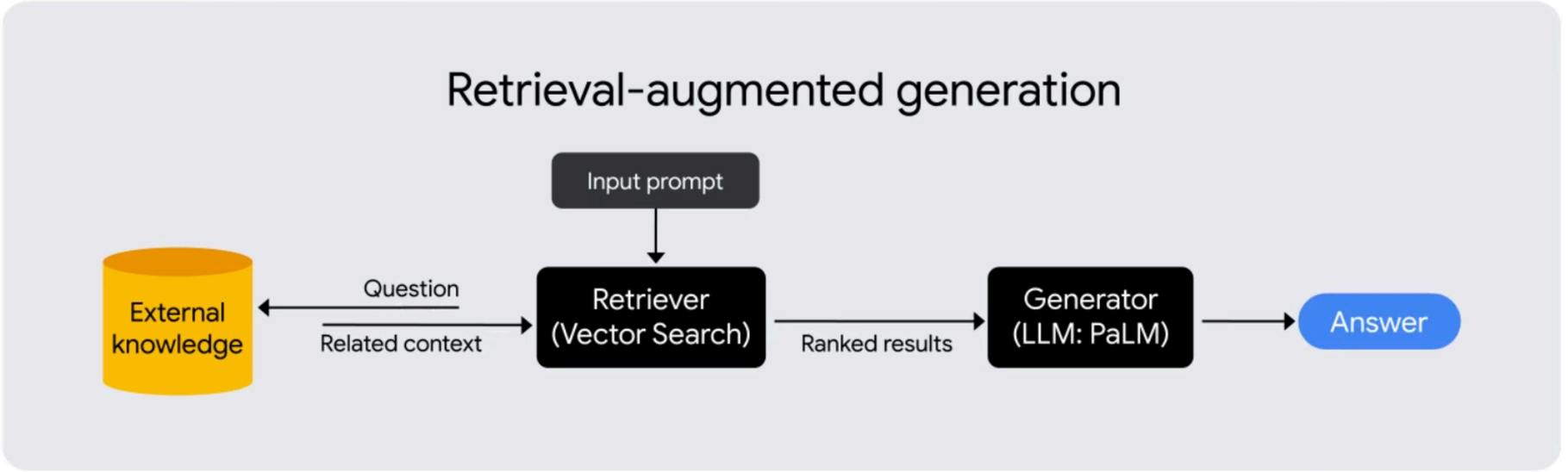
Retrieval augmented generation
RetrievalQA, agents personal knowledge base

Example mining
Training private embedding models

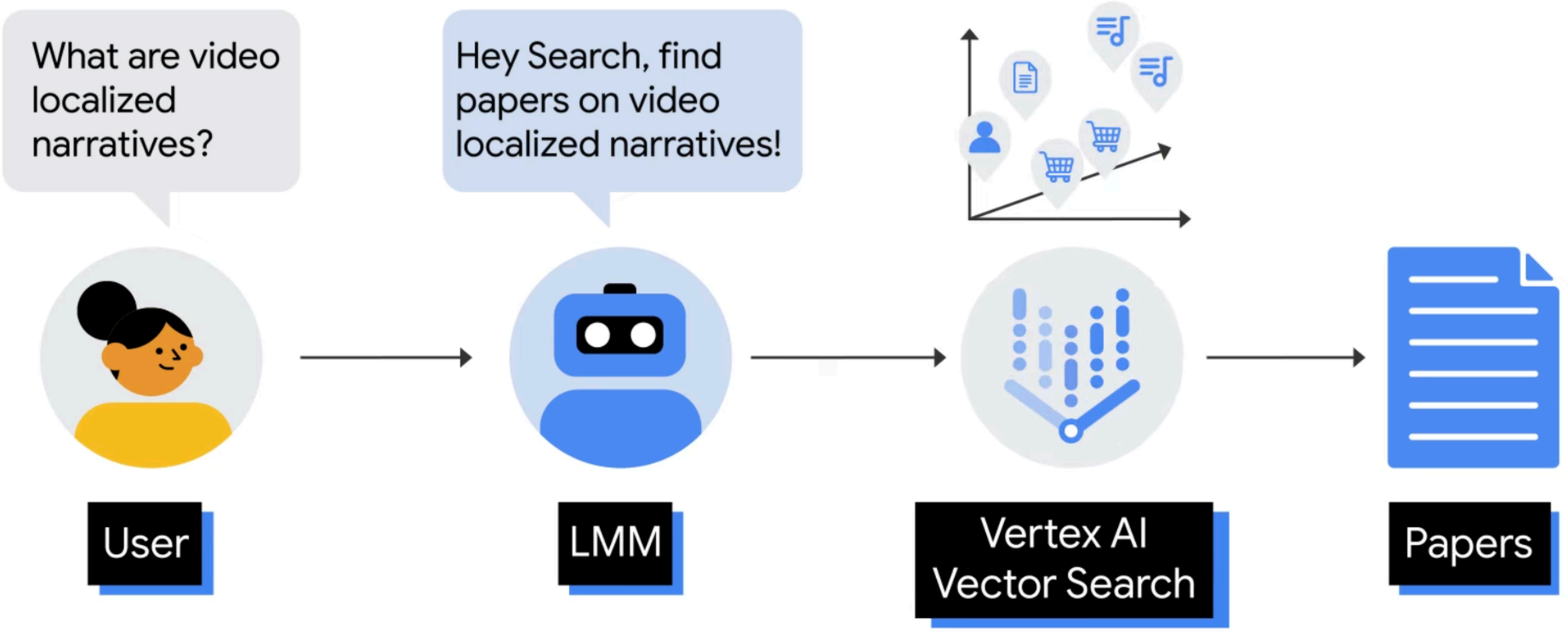
Retrieval-augmented generation (RAG)

"Grounding" on user data

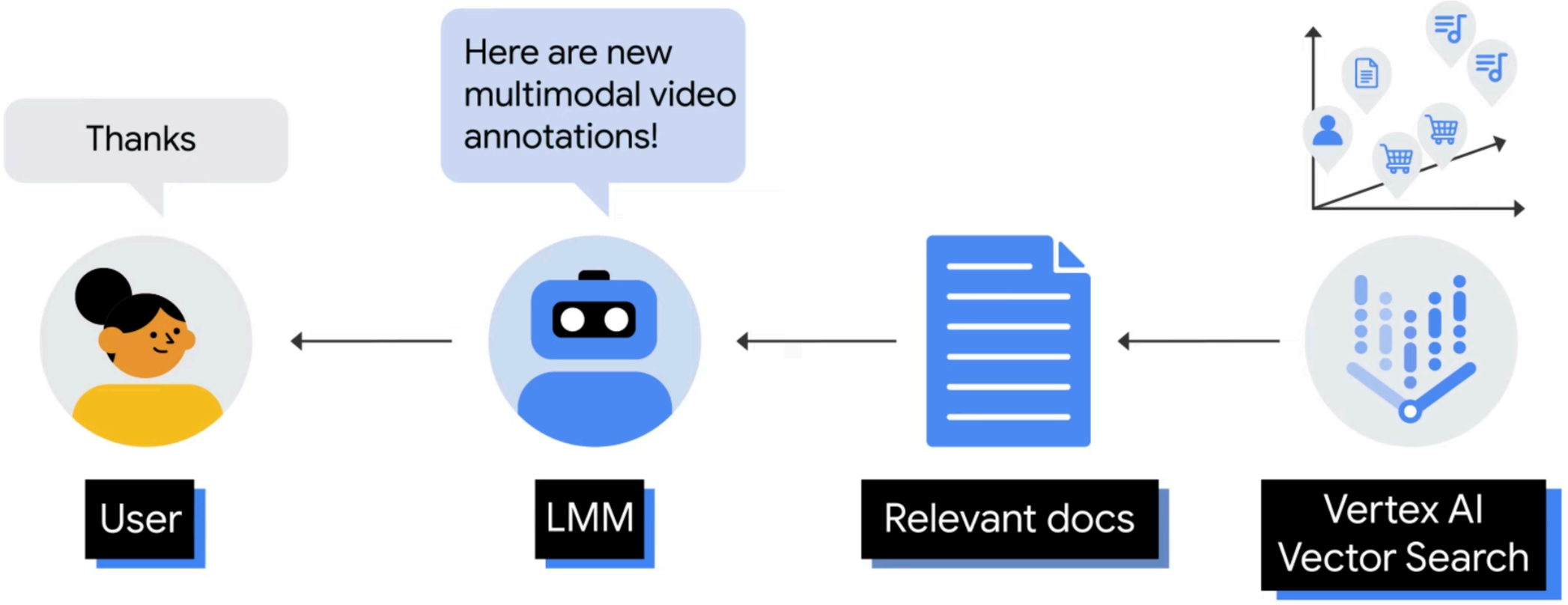
Feed the LLM relevant context in real time, by using an information retrieval system like Vector Search.



Solution: Use Vector Search with RAG



Solution: Use Vector Search with RAG



Prompt for the LLM in RAG

SYSTEM: You are an intelligent assistant helping the users with their questions on research papers.

Question: What are video localized narratives?

Strictly Use *only* the following pieces of context to answer the question at the end. Think step-by-step.

Do not try to make up an answer:
- If the answer to the question cannot be determined from the context alone, say "I cannot determine the answer.
- If the context is empty, just say "I do not know the answer to that."

We propose Video Localized Narratives, a new form of multimodal video annotations connecting vision and langal Localized Narratives [40], annotators speak and move their mouse simultaneously on an image, thus groun a mouse trace segment.
However, this is challenging on a video. Our new protocol empowers annotators to tell t with Localized Narratives, capturing even complex events involving multiple actors interacting with each oth al passive objects. We annotated 20k videos of the OVIS, UVO, and Oops datasets,totalling 1.7M words. Based oo construct new benchmarks for the video narrative grounding and video question answering tasks, and provide r om strong baseline models. Our annotations are available at https://google. github.io/videolocalizednarrativ

The **question** from the user

The **instructions** for the LLM

The paper retrieved by **Vector Search**

Project 2 +
HW 7 Prelab

Two mini-projects

- **Weeks 2-5**: extract and synthesize podcast themes to share career insights and support exploration
- **Weeks 7-9**: develop a career Q&A experience with responses grounded in podcast insights and stories

TIM 175 Your Future Is Our Business Project Survey

As you know, our course assignments are organized around a community-engaged project with Your Future Is Our Business, where they have allowed us to use their What-To-Be podcast data to explore prototyping generative AI experiences to augment youth career exploration.

A student has reached out anonymously with strong complaints of their work being given to the client we are partnering with (and they were not satisfied by my Canvas announcement explanation). As I stated before, this is a common practice for capstone projects (and in the community-engaged setting, it is generally perceived as unethical to take community partner time and resources to benefit student learning without providing them with some benefit too), so I am reaching out to those more knowledgeable than myself on how this is typically handled and the detailed policies around it for the future.

However, regardless of the policy details, I think it is also reasonable to allow students to decide themselves whether they would like to contribute to the final client deliverable for this year. Since we have a large number of students working with a single client partner, we should still be able to do the community project in an ethical way where they we are not just taking from them without giving back.

I will also be changing the teams for Project 2 based on these preferences so that teams can be formed out of people who are willing to contribute their work to the non-profit partner (for Project 1 work, we will only share team work in which all team members agree to share)

Fill out preference survey now

(will be used for Project 2 teams)

Vector Embeddings and RAG introduction

TIM 175 WEEK 7 PRELAB

Brief Task Overview

Complete the following activities to understand vector embeddings, vector databases and their role in similarity-based retrieval tasks.

1. Activity 1: Introduction to word embeddings
2. Activity 2: Introduction to sentence embeddings
3. Activity 3: Retrieval as a similarity task
4. Activity 4: Introduction to creating and querying Vector databases
5. Activity 5: Using vector databases for RAG