

# Productized & Enterprise RAG

October 2025





## Lucian Gruia

Data Science Principal Lead

## Speaker

- Head of Engineering, Romania
- 13+ years of experience in Software Engineering
- Led cross-functional teams across 8 time zones in Telecom, Fintech, Aerospace, and MedTech
- Mentor @Google Developer Groups Romania
- Research Assistant, AI Multimedia Lab (National University of Science & Technology, Bucharest)
- Supervised 12 BSc and MSc diploma projects

# Agenda

---

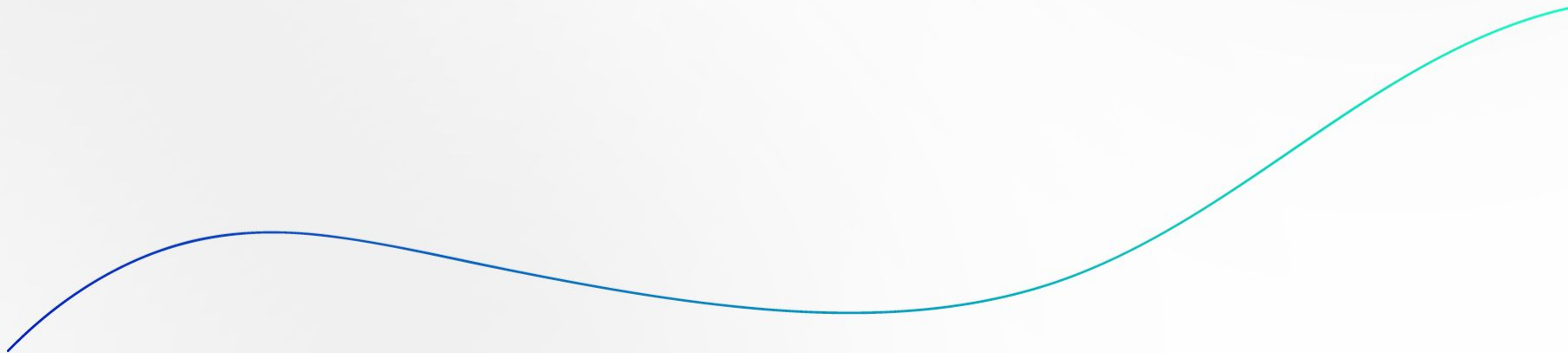
01 Existing repos & products

---

02 Pros, Cons, Limitations and  
Considerations

---

# Quick recap

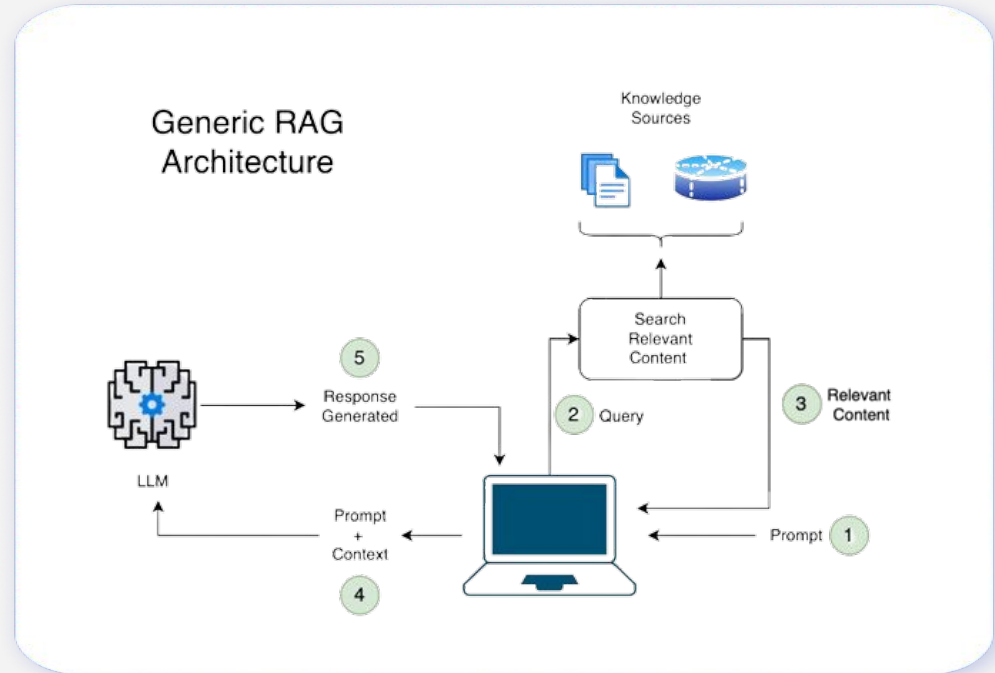


# What is RAG

A RAG system essentially correlates a **user's prompt** with a relevant **data chunk**. It does this by identifying **the most semantically similar** chunk from the database.

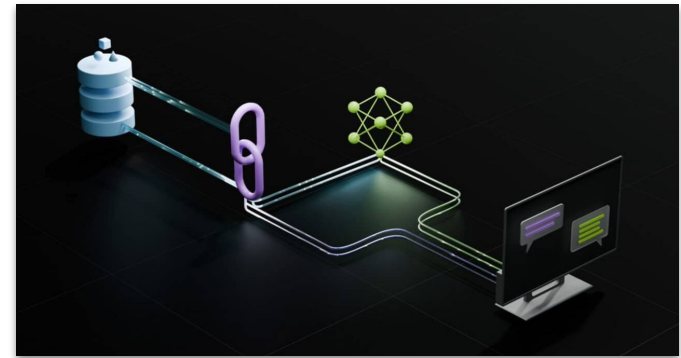
This chunk then becomes **the context** for the prompt.

When **passed to the Large Language Model (LLM)**, it enables the system to provide a relevant answer within the **given context**.



# Why do we need RAG?

- **Expands Knowledge Base**  
RAG accesses a vast external database, enriching its knowledge beyond initial training data
- **Improves Accuracy**  
Enhances response precision by integrating relevant, real-time information
- **Adaptable**  
Effectively handles novel and niche queries
- **Increases Efficiency**  
Streamlines information retrieval and generation process
- **Versatile Applications**  
Useful across various fields, from customer support to research



Source: [What Is Retrieval-Augmented Generation, aka RAG?](#)

**LLMs are limited to the knowledge they were originally trained on, and retraining or fine-tuning them is slow and expensive.**

# Embeddings. Similarity

- **Embeddings**

Numerical representations of concepts, in a high-dimensional space, capturing semantic meaning.

- **Similarity:**

- Lexical: entities are alike in appearance
- **Semantic:** entities are alike in meaning

- In RAG we represent entities **by describing** them.

This is a form of **knowledge representation**.

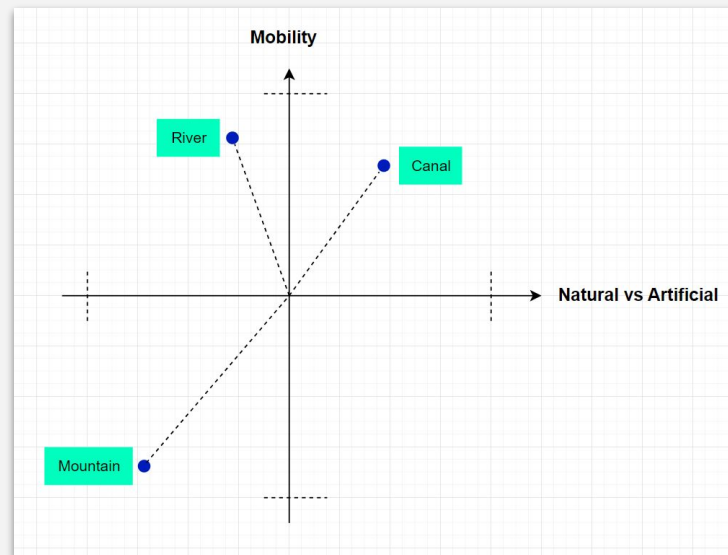
## Example: Mountain, River, Canal

### One hot encoding

Mountain: 1  
River: 2  
Canal: 3

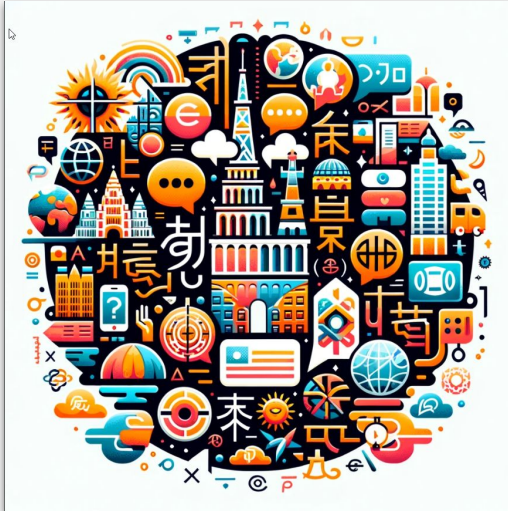
### 2-Dimensional Space

[Natural vs Artificial, Mobility]  
Mountain: [-0.7, -0.8]  
River: [-0.3, 0.7]  
Canal: [ 0.4, 0.5]



Read more: [Wikipedia - Cosine Similarity](https://en.wikipedia.org/wiki/Cosine_similarity)

# Multilinguality



Generated with DALL·E 3

- **Multilingual Capabilities**  
RAG and chatbots can process and respond in multiple languages, enhancing global reach and accessibility.
- **Cross-lingual Understanding**  
These systems understand queries in one language and retrieve information in another, facilitating cross-lingual interactions.
- **Language-Agnostic Embeddings**  
Use of embeddings that capture semantic meanings across languages.
- **Challenges**  
Handling idiomatic expressions, cultural nuances, and varying sentence structures in different languages.



# Existing repos & products

A decorative wavy line spanning the width of the slide, starting with a blue segment on the left and transitioning into a green segment on the right.



## Understand **Anything**

Your research and thinking partner, grounded in the information you trust, built with the latest Gemini models.

Try NotebookLM



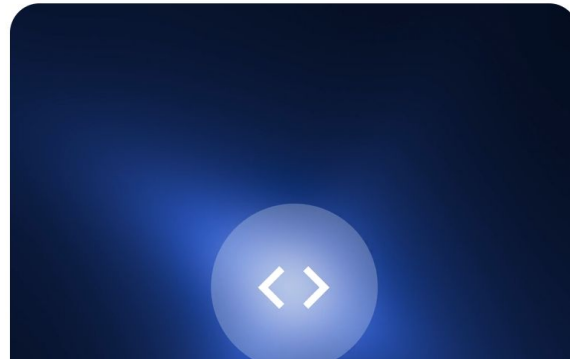
Your AI-Powered Research Partner

# GEMs

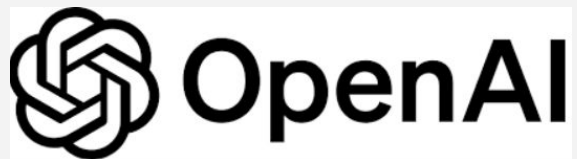


## Build custom experts with Gems

Gems are your custom AI experts for help on any topic. Gems can be anything from a career coach or brainstorm partner to a coding helper. Get started with our suite of premade Gems, or build your own custom Gems tailored to your unique needs.



# Custom GPTs



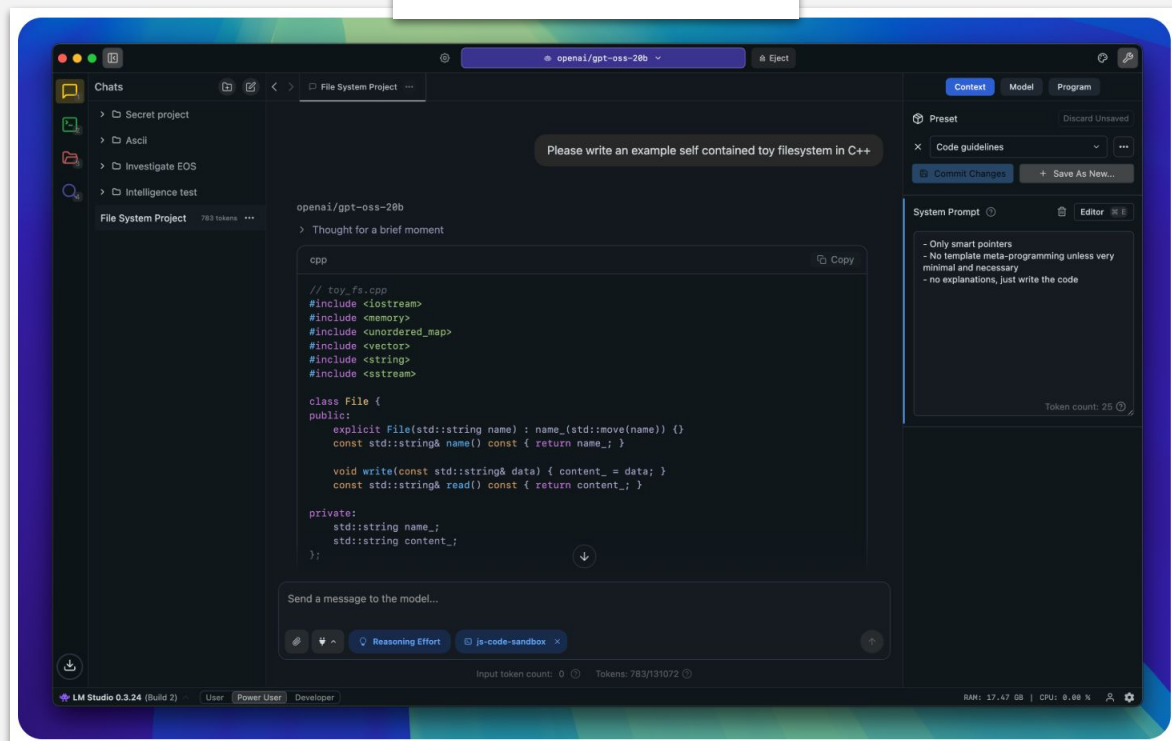
November 6, 2023 Product

## Introducing GPTs

You can now create custom versions of ChatGPT that combine instructions, extra knowledge, and any combination of skills.

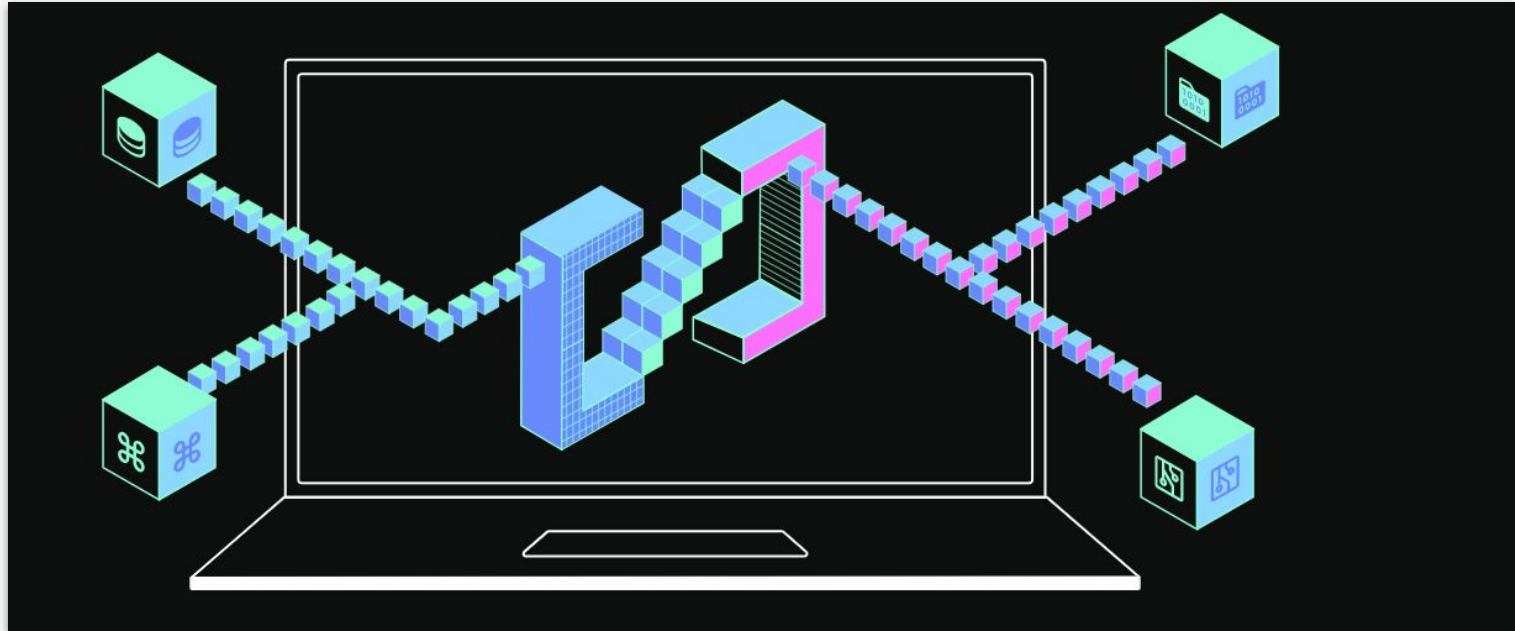


# LM Studio



# Anything LLM

 Anything LLM



## OI Open WebUI

Prompts

Search Prompts

+

Sponsored by Join the Open WebUI Team

OI

We are hiring! Join us to build the future of AI applications.

New Functions

See All

#1 FILTER v1.0.0

**Memory System**

Memory system with intelligent skip detection, hybrid retrieval, and background consolidation.

@tayfur

#2 PIPE v0.6.1

**Nano Banana**

This code implements Google Flash 2.5 Image with API usage for image generation and editing. You can...

@anaumer

#3 FILTER v0.2.0

**Qwen Thinking Filter**

filter qwens thinking into a closable element

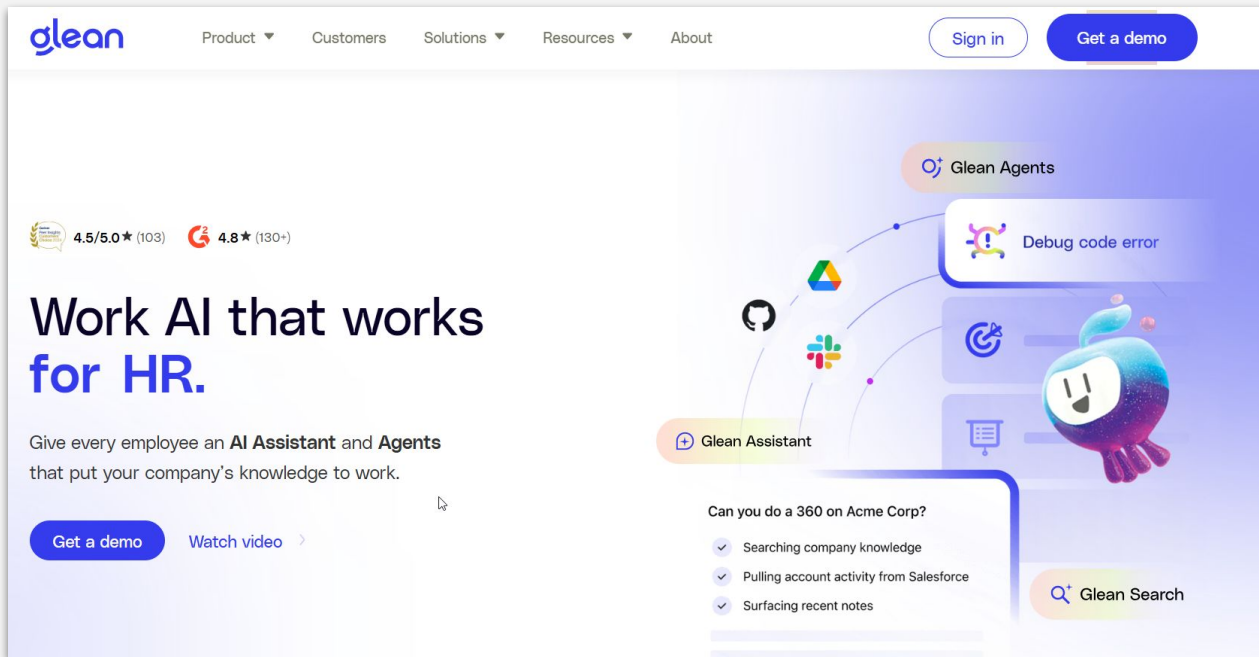
@ekisf

#4 FILTER v1.0

**Apriel-1.5-15B-Thinker - Final+Think**

Makes Apriel-1.5-15B-Thinker integrate cleanly with Open WebUI by rendering a native <think> panel an...

@supczinskib



The screenshot shows the Glean website's homepage. At the top is a navigation bar with the Glean logo, links for Product, Customers, Solutions, Resources, and About, and buttons for Sign in and Get a demo. The main content area features a large heading 'Work AI that works for HR.' with subtext 'Give every employee an AI Assistant and Agents that put your company's knowledge to work.' and buttons for Get a demo and Watch video. To the right, a diagram illustrates the Glean ecosystem, showing a central 'Glean Assistant' connected to various integrations like Google Drive, Slack, and GitHub. A 'Glean Agents' section lists tasks like 'Debug code error'. A 'Glean Search' section shows a search query 'Can you do a 360 on Acme Corp?' with results for searching company knowledge, pulling account activity from Salesforce, and surfacing recent notes. A cartoon robot character is also present in the diagram.

glean

Product ▾ Customers Solutions ▾ Resources ▾ About

Sign in Get a demo

4.5/5.0★ (103) 4.8★ (130+)

## Work AI that works for HR.

Give every employee an **AI Assistant** and **Agents** that put your company's knowledge to work.

Get a demo Watch video >

**Glean Agents**

- Debug code error

**Glean Assistant**


Can you do a 360 on Acme Corp?



- ✓ Searching company knowledge
- ✓ Pulling account activity from Salesforce
- ✓ Surfacing recent notes




**Glean Search**



# Sourcegraph

 **Sourcegraph**

 `context:global` Search for code or files... Aa 

 **Get Sourcegraph for your favorite editor**  
Download the extension to get the power of Sourcegraph right where you code.   [Download](#) ×

Search in files, paths, and repository-names

`test server` (both terms anywhere)  
`"Error 1001"` (specific string)  
`"\"Error 1001\""` (specific string containing quotes)  
`foo OR bar`  
`/open(File|Dir)/` (regular expression)

Search in commit diffs

`type:diff after:1week fix`  
`type:diff author:alice add`

Filter by...

`file:file.go foo`  
`repo:org/repo`  
`repo:org/` (all repositories in org)  
`lang:javascript`

Advanced

`repo:has.description(foo)`  
`repo:has.commit.after(1week)`  
`file:~some_path file:has.owner(alice)`  
`file:~some_path select:file.owners`

# Pros, Cons, Limitations and Considerations

A decorative wavy line in blue and green colors spans the bottom of the slide, starting from the left and curving upwards towards the right.

# Pros & Cons

## Pros

- Enhances factual accuracy via external knowledge
- Reduces hallucinations vs. pure LLMs
- Enables dynamic updates without retraining
- Modular, domain-adaptable architecture
- Scalable across multimodal or multi-source data

## Cons

- Retrieval quality depends on data prep & embeddings
- Latency increases with retrieval pipeline
- Requires infra (vector DBs, indexing, chunking, etc.)
- Context window still limits answer depth

# Limitations & Considerations

## Limitations

- Not reasoning-first, still retrieval-driven
- Poor handling of implicit or fuzzy queries
- No built-in memory or learning from feedback
- Struggles with conflicting or outdated sources

## Considerations

- Balance retrieval vs. generation cost & latency
- Continuous evaluation of relevance & drift
- Governance for data freshness and accuracy
- Potential upgrade path: **Agentic / Tool-use RAG**

# Enterprise Data Protection and Privacy

- **Safeguarding sensitive information** and maintaining customer trust.
- **Regulatory Compliance:** Adherence to laws like GDPR, which mandate data security and privacy.
- **Encryption & Anonymization:** Secure data and protect individual identities.
- **Access Control:** Implementing strict permissions and access policies to sensitive data.
- **Regular Audits:** Monitoring and auditing to identify and rectify any vulnerabilities.
- Importance of using **External Data** & RAG models.
- **Augment context** in order to get accurate completions.



# Quiz Time



## **What is the main goal of RAG?**

- A. To train larger models
- B. To reduce hallucinations using external knowledge
- C. To replace fine-tuning
- D. To speed up token generation

**What is a key limitation of RAG systems?**

- A. They cannot retrieve multimodal data
- B. They require GPU inference for retrieval
- C. They depend heavily on data quality and embeddings
- D. They cannot use APIs



## **What's the next evolution beyond basic RAG?**

- A. Fine-tuned GPT models
- B. Static FAQ chatbots
- C. Agentic or tool-using RAG architectures
- D. Prompt templates with zero retrieval

# Key Takeaways from the session

- Boosts **factual accuracy** by grounding LLMs in **real data**
- **Reduces hallucinations**, but depends on retrieval quality
- **No retraining needed**, knowledge updated via data refresh
- Latency-complexity **trade-off**: retrieval adds overhead
- Still **limited** reasoning and context awareness
- Next **evolution** → Agentic, monitored, tool-using RAGs



Thank you!

