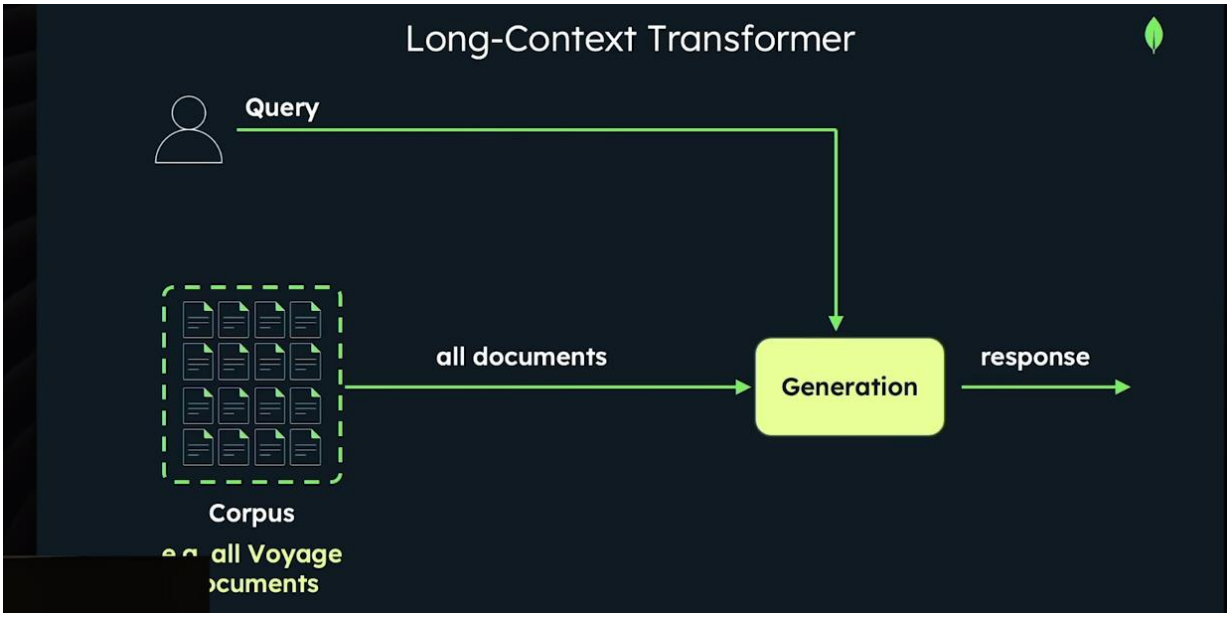
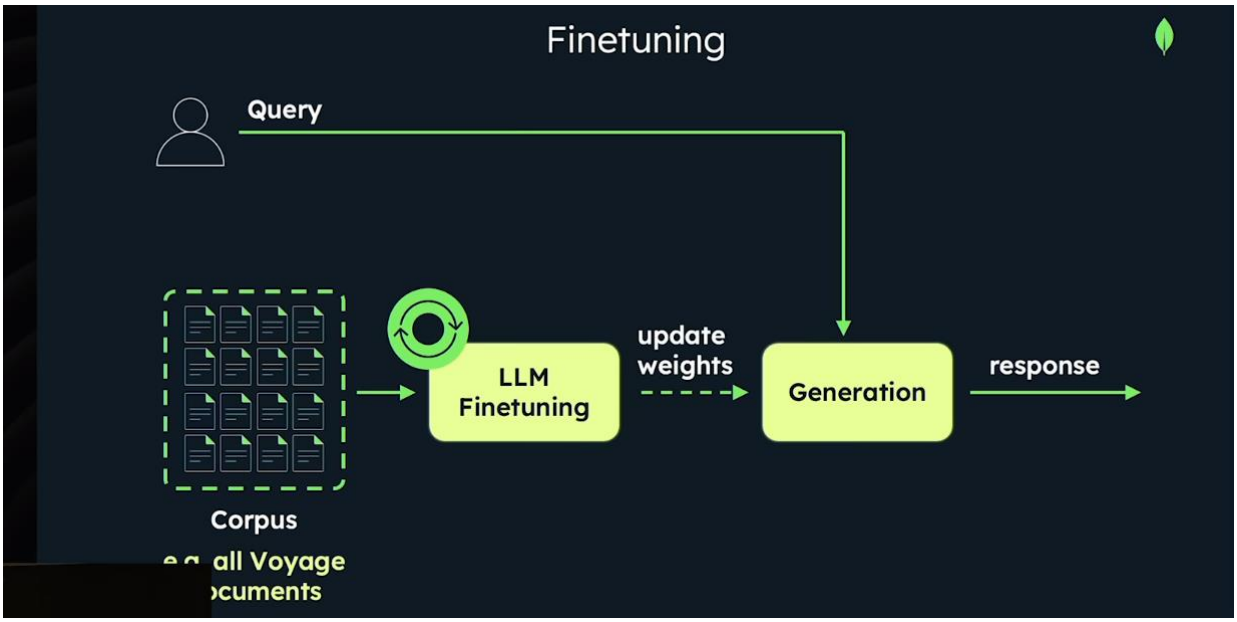
RAG in 2025: State of the Art and the Road Forward — Tengyu Ma, MongoDB (acq. Voyage AI)

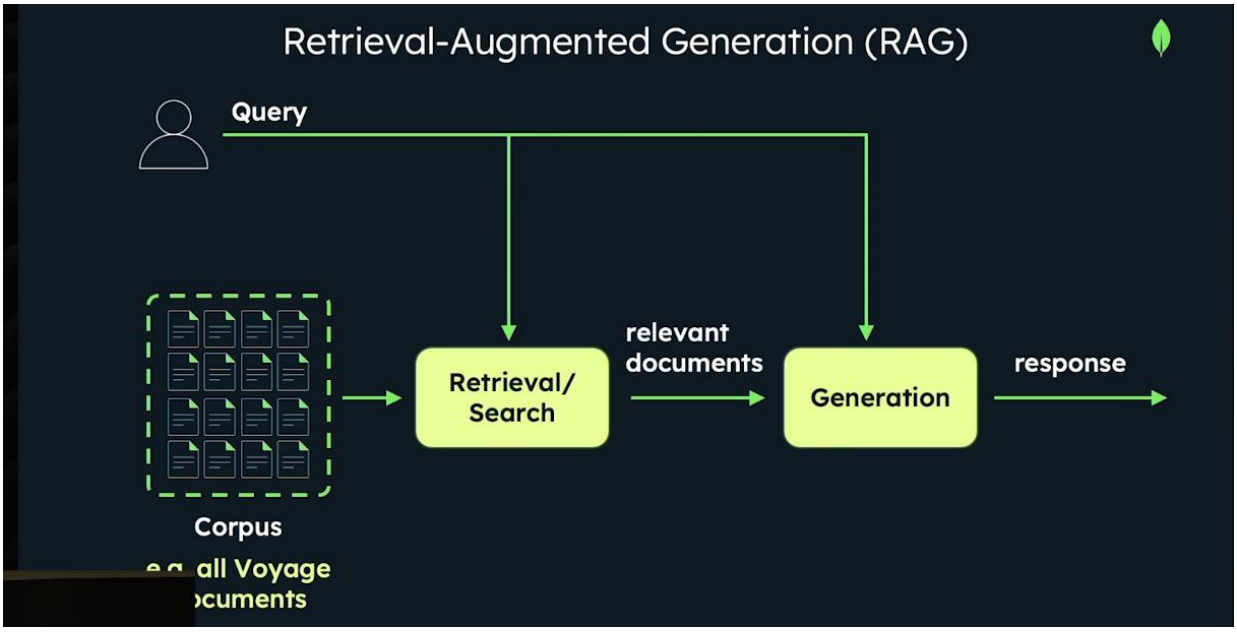
3,264 views Jun 27, 2025 AIEWF 2025 Complete Playlist

The talk will have three parts
1.Roadmap debate: RAG vs. finetuning vs. long-context
2.RAG today: benefits, challenges, and current solutions
3.RAG tomorrow: AI models do more work

About Tengyu Ma
Tengyu Ma is the Chief AI Scientist @ MongoDB and an Assistant Professor @ Stanford. He was the co-founder and CEO of Voyage AI before the acquisition by MongoDB.







Long-context vs fine-tuning vs RAG

Long-context

- Enormous cost
- Quality loss due to lots of irrelevant context

Skim an entire library to answer one question

Fine-tuning

- Needs lots of high-quality data (and paraphrasing)
- Acquiring and forgetting knowledge is difficult

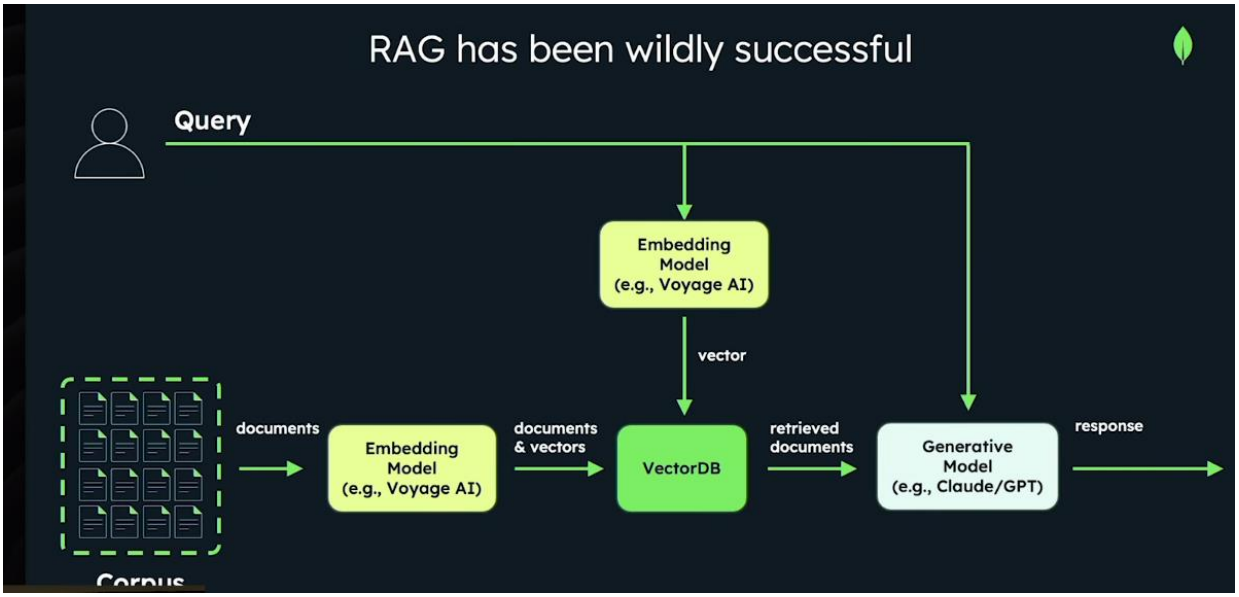
Muscle-memorize the library, which is both difficult and unnecessary

RAG

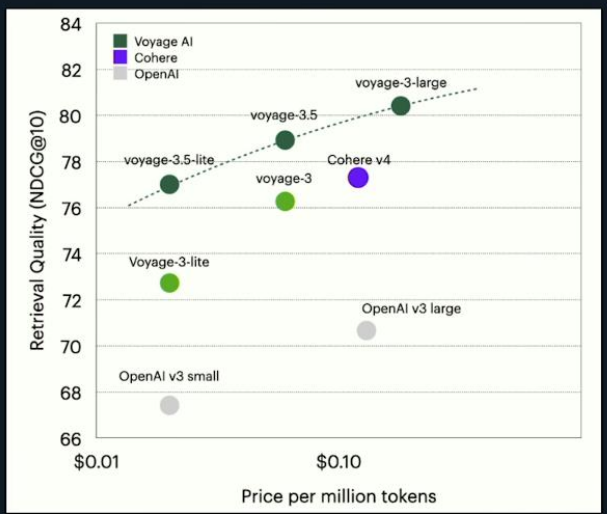
- Reliable, modular, fast, and cheap
- quality responses and no hallucination

Retrieve the most relevant book chapters from the library to answer the question

RAG has been wildly successful



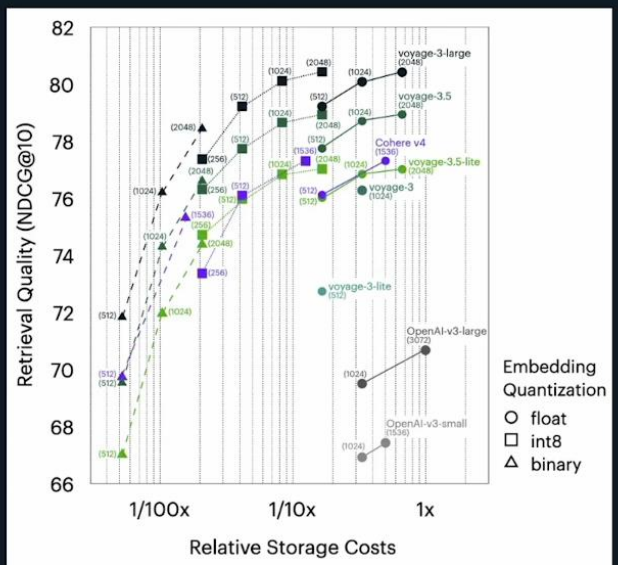
Significant Improvements in Retrieval Accuracy Over Last 2 Years



Matryoshka Learning & Quantization-Aware Training

Lower vectorDB cost

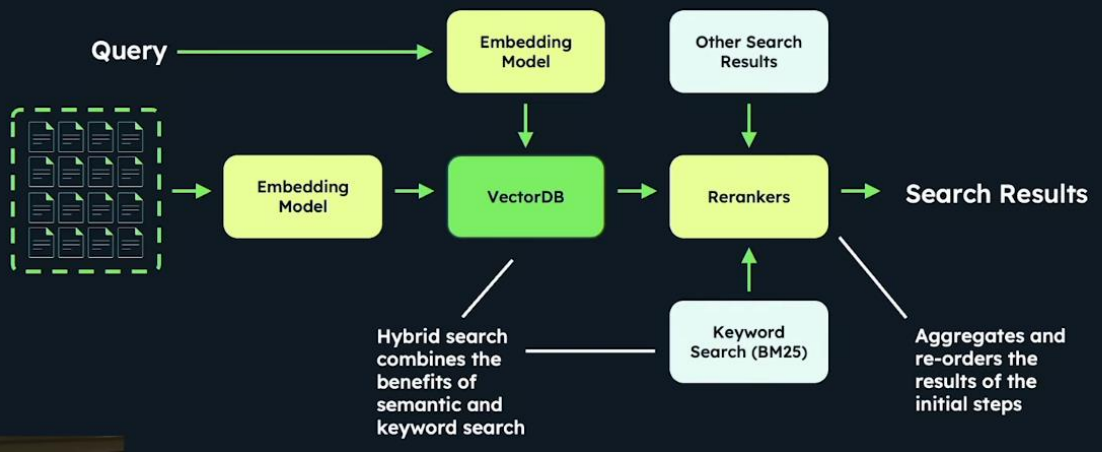
- **Matryoshka: given a 2048-dimensional embedding, its first 256 dimension is also a very good embedding, but is much cheaper to store.**
- **Quantization: the lower-precision version of the embedding still works**



How to Improve RAG Performance?

(Besides Using Better Embedding Models)

0. Hybrid Search and Rerankers



1. Enhancing Queries and Documents

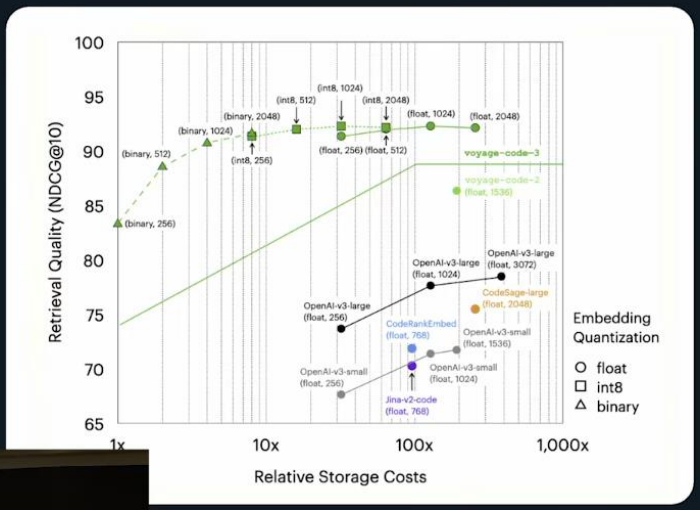
Query decomposition: rephrase or add context to the queries

- Original query: “RAG”
- Improved query: “Explain Retrieval-Augmented Generation (RAG) to me”
- Decomposed query: “Explain retrieval in RAG” + “Explain generation in RAG”

Document enrichment: add extra global context to document chunks

- E.g., document title & headers, categories, authors, dates, etc.
- Add LLM-generated contexts to the chunks

2. Use Domain-specific Embeddings (optimized for the particular domains)



voyage-code-3

- Note: the loss due to lower-dimensional embedding and quantization is much smaller for code-3 than general purposes.
- Voyage also has finance and legal models; finance-3 and law-3 are coming in a few months.

3. Finetune Embedding Models with Your Own Data

Find "positive" pairs with semantic relationships

- (title/header, document)
- (question, supporting evidence)
- (caption, image)
- (generated query, document)

Finetune embedding models with contrastive loss.

4. Flavor-of-the-month-RAG

Non-comprehensive list of different types of RAG

- Self-RAG
- Golden-Retriever
- Corrective RAG
- Speculative RAG
- GraphRAG
- Iterative/recursive retrieval
- ...

Agenda

1. **Roadmap Debate**: RAG vs. Finetuning vs. Long-context
2. **RAG Today:** Benefits, Challenges, and Current Solutions
3. **RAG Tomorrow:** AI Models Do More Work

Prior to GenAI/Foundation Models Era

CS229: Machine Learning

Instructors


Tengyu Ma


Chris Ré

Course Description This course provides a broad introduction to learning (generative/discriminative learning, parametric/non-param (clustering, dimensionality reduction, kernel methods); learning the

The 7 steps of ML Systems

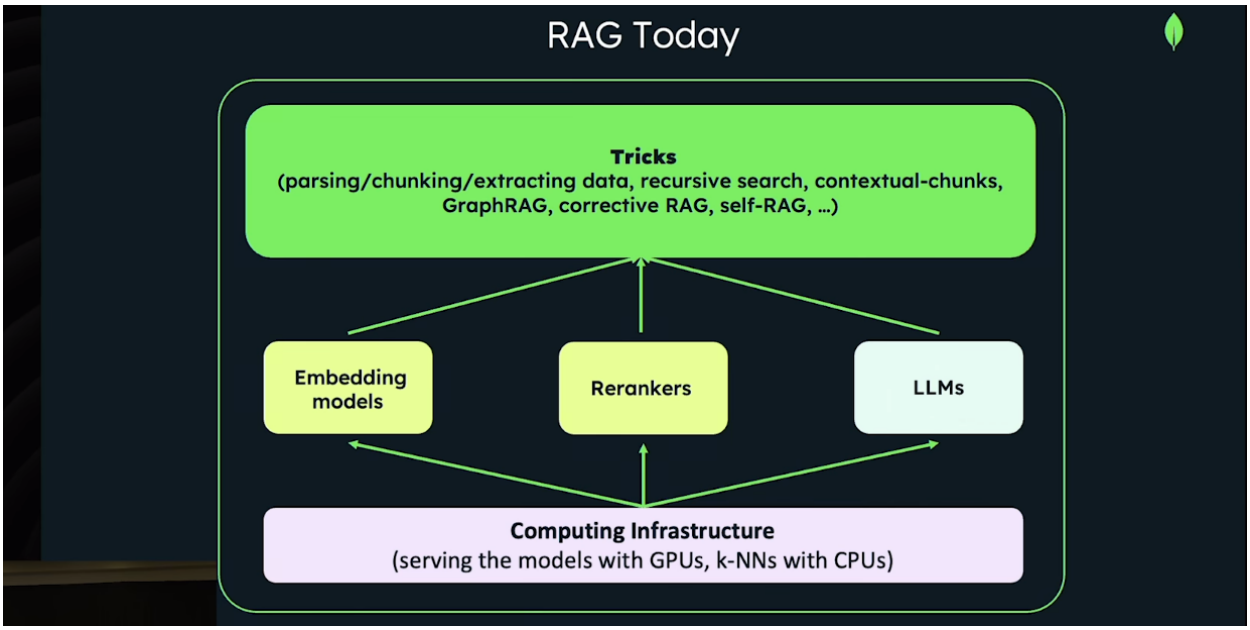
Step 1: Acquire Data
Step 2: Look at your data* - after every step.
Step 3: Create train/dev/test splits
Step 4: Create/refine a specification
Step 5: Build model (simplest that works!)
Step 6: Measurement
Step 7: Repeat.

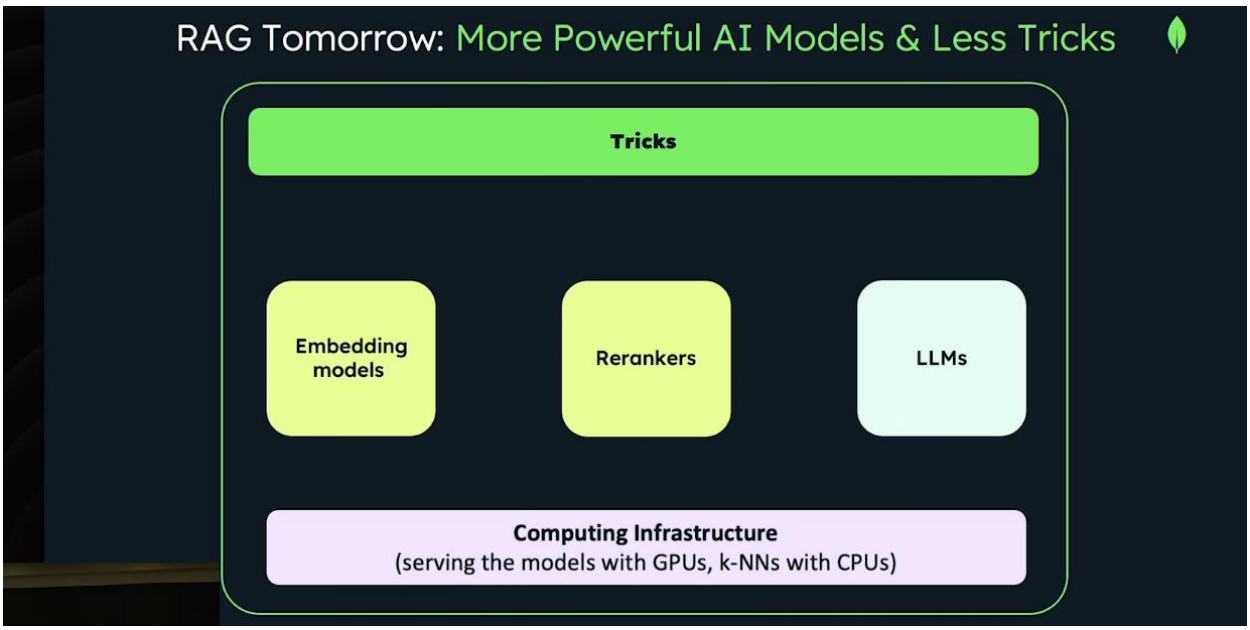
2024

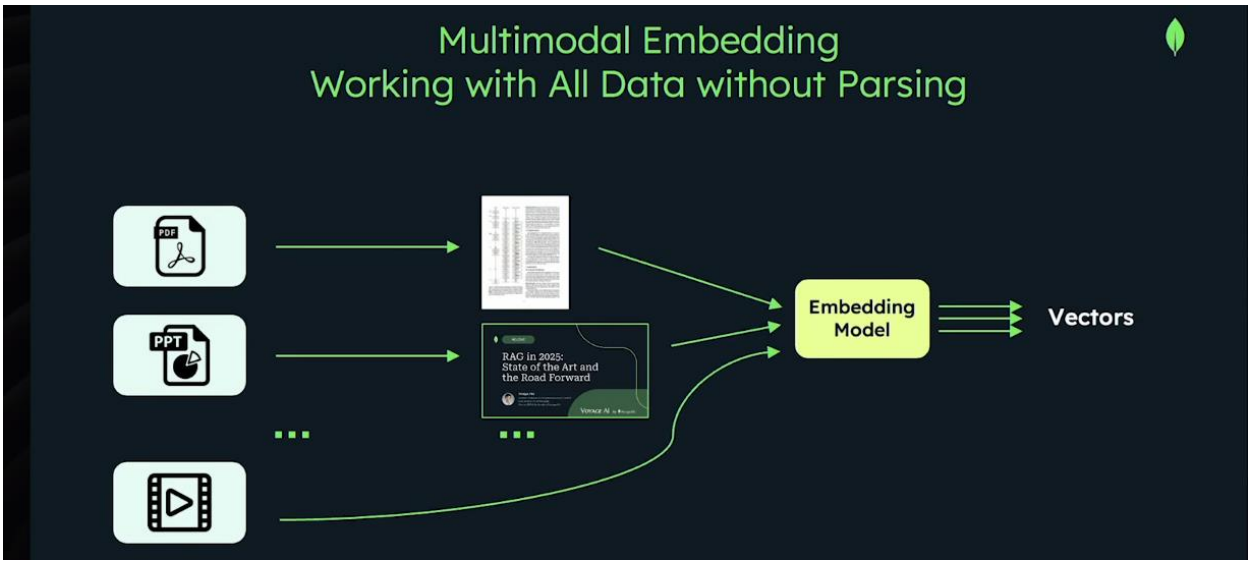
LLM
(Still need RAG for proprietary data)

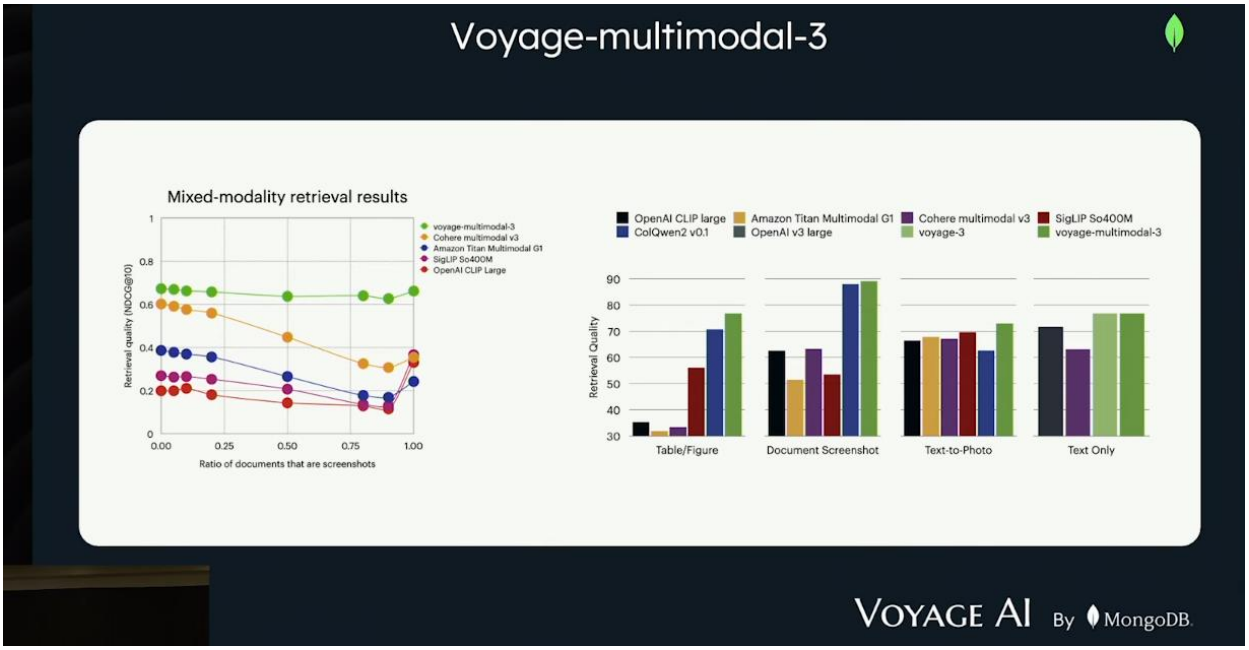
The 7 steps of ML Systems

Step 1: Acquire Data
Step 2: Look at your data* - after every step.
Step 3: Create train/dev/test splits
Step 4: Create/refine a specification
Step 5: Build model (simplest that works!)
Step 6: Measurement
Step 7: Repeat.

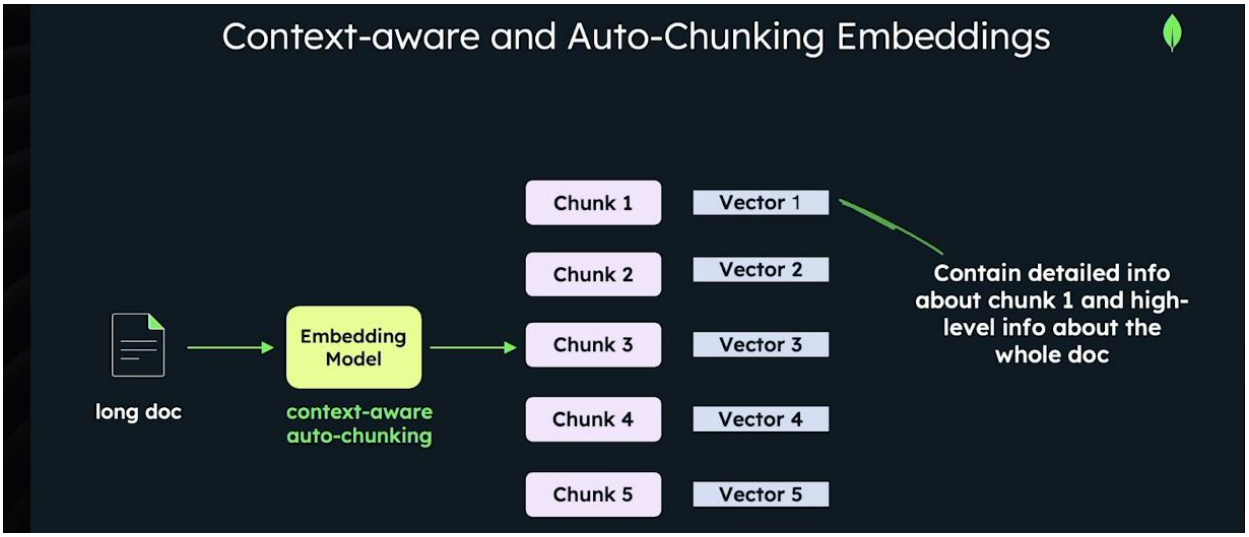

RAG Today
Tricks
(parsing/chunking/extracting data, recursive search, contextual-chunks, GraphRAG, corrective RAG, self-RAG, ...)
Embedding models
Rerankers
LLMs
Computing Infrastructure
(serving the models with GPUs, k-NNs with CPUs)

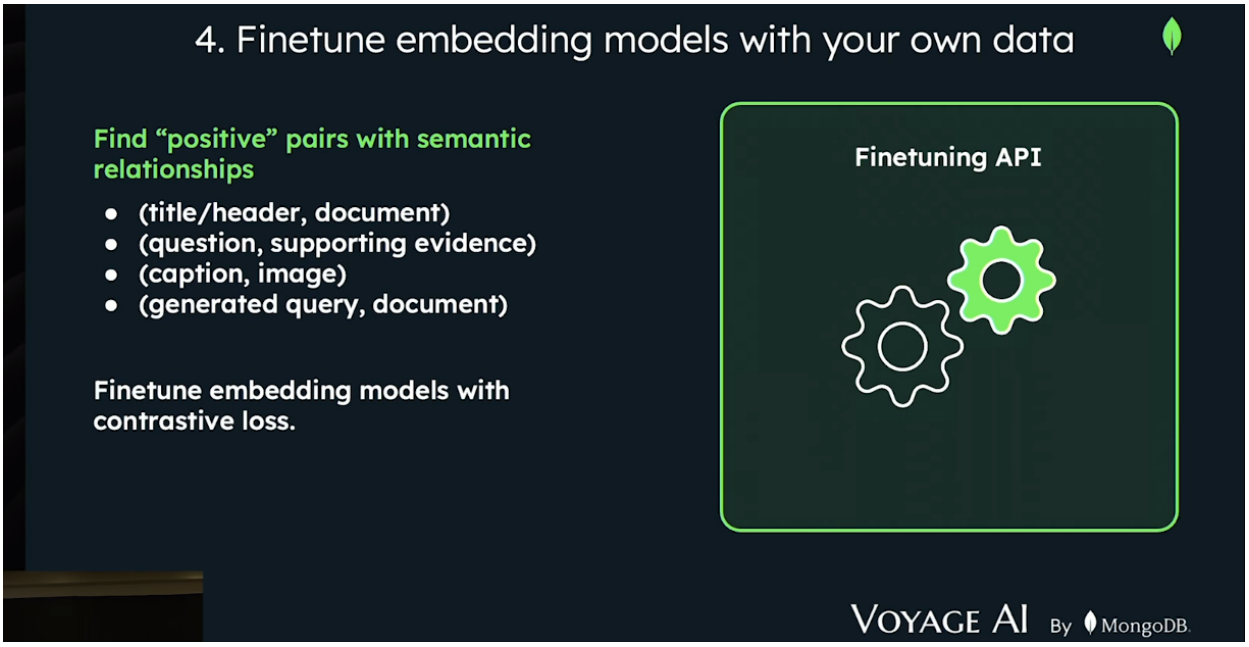

RAG Tomorrow: More Powerful AI Models & Less Tricks
Tricks
Embedding models
Rerankers
LLMs
Computing Infrastructure
(serving the models with GPUs, k-NNs with CPUs)

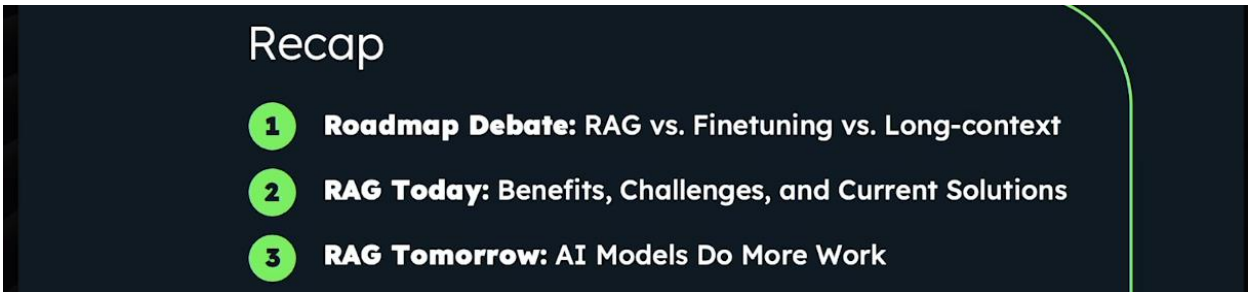

Multimodal Embedding
Working with All Data without Parsing
PDF
PPT
RAG in 2025: State of the Art and the Road Forward
Embedding Model
Vectors



Now you can just take screenshots of the data and give it to the multimodal embedding model






Recap
1
Roadmap Debate: RAG vs. Finetuning vs. Long-context
2
RAG Today: Benefits, Challenges, and Current Solutions
3
RAG Tomorrow: AI Models Do More Work