AGENT ENGINEERING

Context Management & RAG

Hai

AGENDA

What’s on the menu today?

1. RAG
2. Context management
3. Different retrieval types
4. Semantic search
5. Improve RAG

CONTEXT MANAGEMENT

Context Mgnt. is the most important job of agent engineering

This sets your agent apart.

> “Prompt engineering” was coined as a term for the effort needing to write your task in the ideal format for a LLM chatbot. “Context engineering” is the next level of this.
>
> — WALDEN YAN @ COGNITION

WHAT IS RAG

retrieval augmented generation

LLMs have limited context window but we can selectively insert information into it for the best possible generation.

Retrieve

Fetch information related to user’s query

Augment

Manage the context so relevant information are prioritized

Generate

Generate output / answer with LLM

SEMANTIC SEARCH

Semantic search turns original knowledge into vectors and fetches closest ones to user’s query

1. Turn document into vectors with an embedding model
2. Store embedding in a vector database
3. Turn user’s query into vectors using the same embedding model
4. Find the closest vectors to user’s query, then fetch the documents based on that

RETRIEVAL TYPES

RAG / Retrieval is more than just vector search

Due to VC-backed vector DB startups’ marketing

- ✓ Keyword / metadata search
- ✓ Semantic search
- ✓ Agentic search
- ✓ Deep research

IMPROVE RAG

First-attempt RAG is never perfect.

Try to identify the root cause of retrieval errors and use the right solution (re-ranking, hybrid search, agentic search, etc..)

> “
>
> You can either actually try different ways to make your search results good, or you can just have seven meetings with VectorDB vendors who tell you their way is the best.
>
> — BRYAN BISCHOF, HEAD OF AI @ THEORY VENTURES

RECOMMENDED FOR EVERYONE

- Use Vectorize for all in one source ingestion, embedding, storing in vector DB, and retrieval.

FOR MOST DEVELOPERS

Beginner

- Pinecone for vector database

Advanced

- Hybrid search in Pinecone
- Cohere Rerank for reranker / Jina’s Reranker

RESOURCES

Additional resources to read

This expands on the topics we talked about today

RAG is more than just embedding search (Jason Liu)

What is Agentic RAG (Weaviate)

How we built our multi-agent research system (Anthropic) - this one is about deep research!

Hybrid Search (Pinecone)

Naive RAG vs Hybrid Search (Video) - Hai

Reranking (Video) - Hai

HOMEWORKS

Homework this week

1. Embed some documents of choice
2. Upsert those embeddings to a vector database
3. Build a question-answer chat interface that pulls in knowledge before answering

Feel free to do this manually with Pinecone + Python/TS, or with Vectorize (beginner-friendly)

FINAL

Q&A