

Exploring Advanced Patterns in RAG

Confoo 2025

Hi! I'm Carl Lapierre

- Software developer from Montreal, Quebec
- 10+ years of experience in software
- Rich background in the education and online learning industry
- Developed search solutions
- Created digitization workflows
- Built content management platforms
- And more!



carllapierre





/OSEDEA

Osedea is a highly regarded **innovation firm** in Montreal that combines creativity, technical expertise, and passion to bring cutting-edge solutions to life.

- Product design
- Software development
- Artificial intelligence
- Robotics



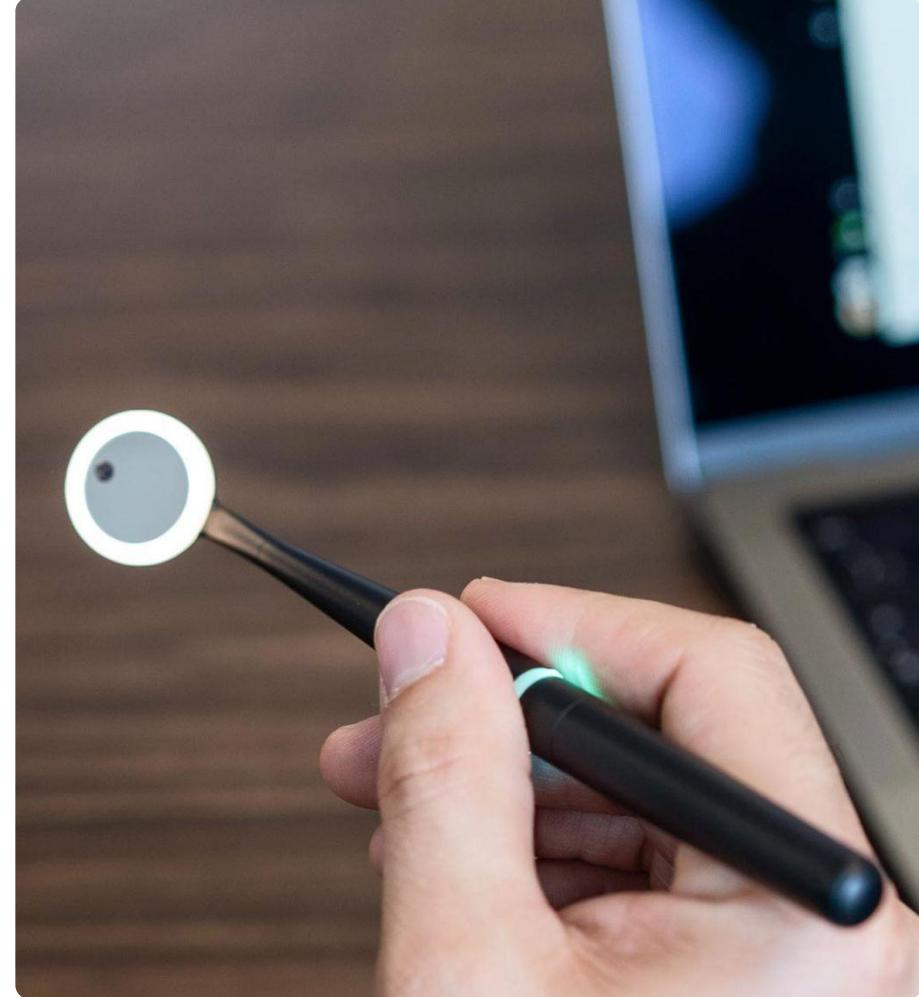
Mila



aws







Story Time





(581)-222-5152

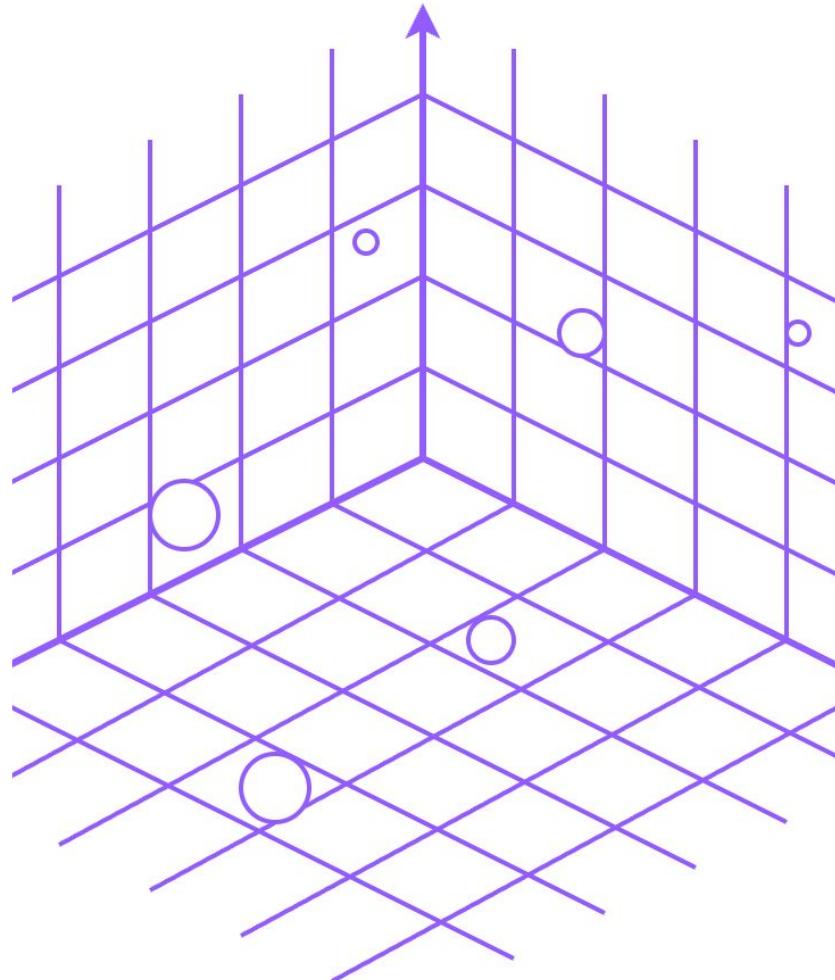
Margherita Pizza	9.99\$
Pepperoni Pizza	11.99\$
Meat Lovers Pizza	14.99\$
Veggie Pizza	12.99\$
Hawaiian Pizza	10.99\$
Garlic Bread	3.99\$
Caesar Salad	7.99\$
Soft Drink	1.99\$
Bottled Water	0.99\$
Pizza Combo	16.99\$
Salad Combo	12.99\$

Start Ordering Pizza

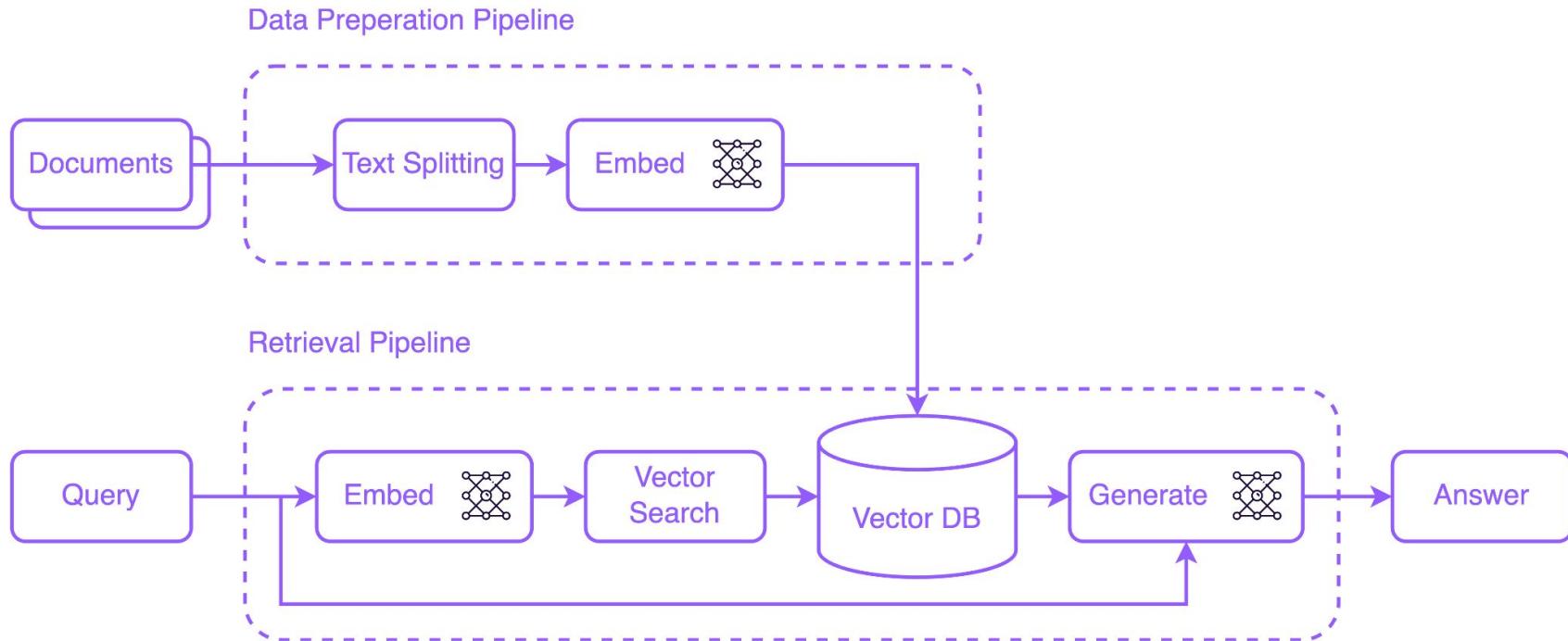
The need for RAG

One of the most common problems our clients face today involves **navigating and searching** over their internal data.

Natural language queries and RAG solves these problem.



Naive RAG



Some past RAG Projects

RFP Assistance

Request for proposal specialists were manually looking through case studies to answer thousands of request for proposals each year.

Engineering Assistance

Chemical engineers were spending hundreds of hours sifting through specification documents to find tolerances and material properties.

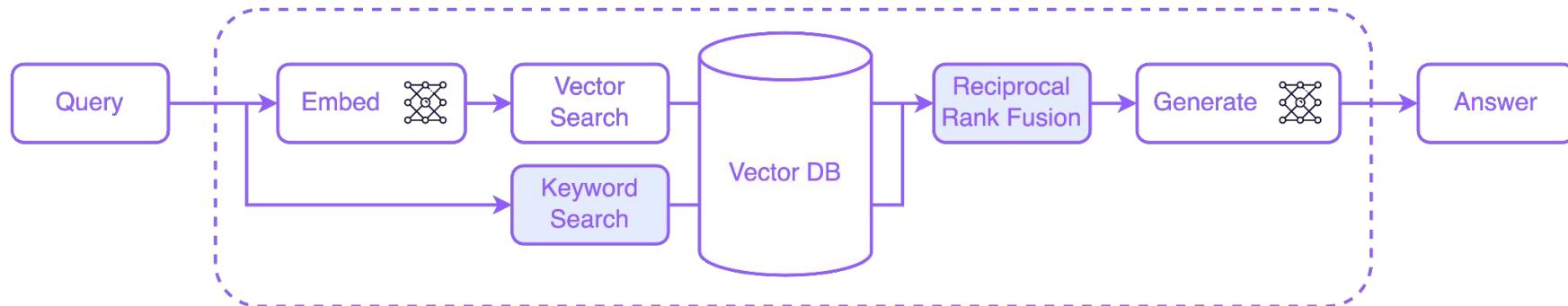
BI Data Exploration

BI visuals were confusing and inflexible rendering them ineffective for non-analysts.

Explainability
Accuracy
Complexity

1. Search Improvements

Hybrid Search

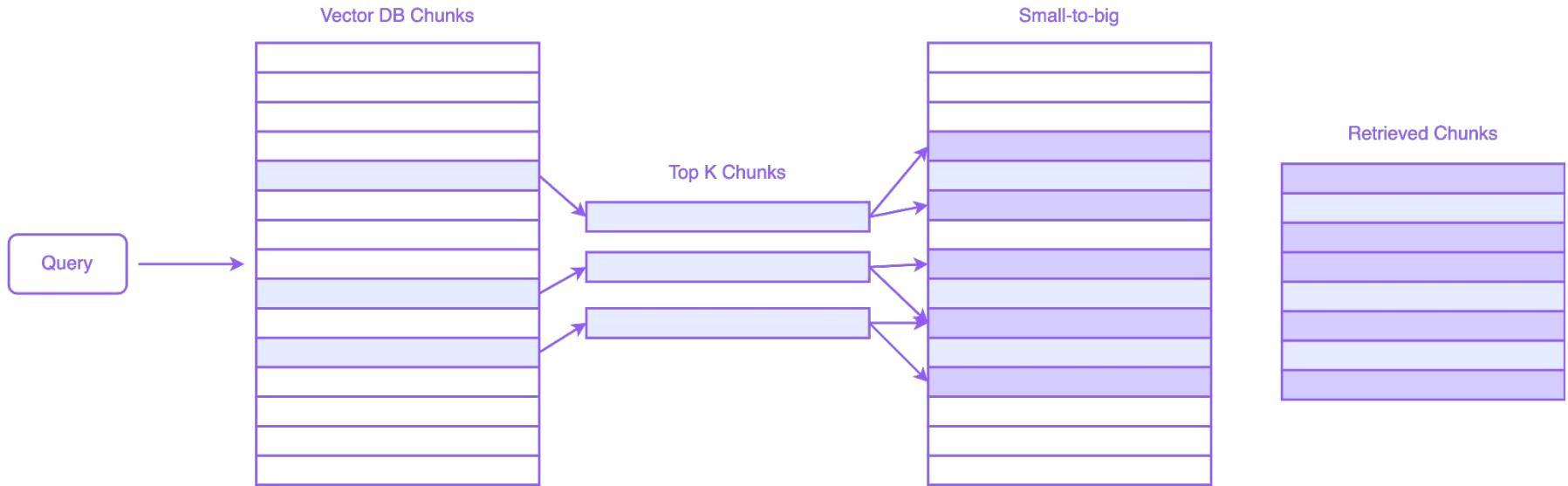


Example Query:

What are the material specifications for HT#356-007?

2. Post-retrieval Improvements

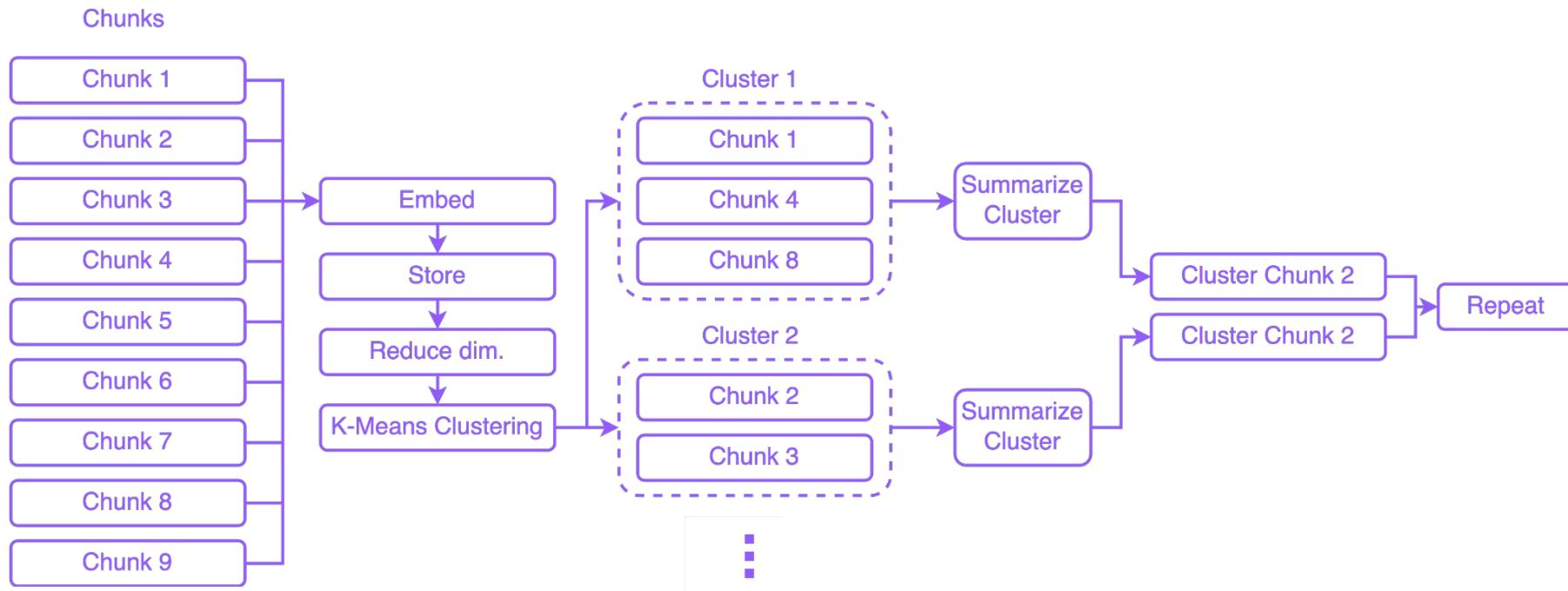
Small-to-big



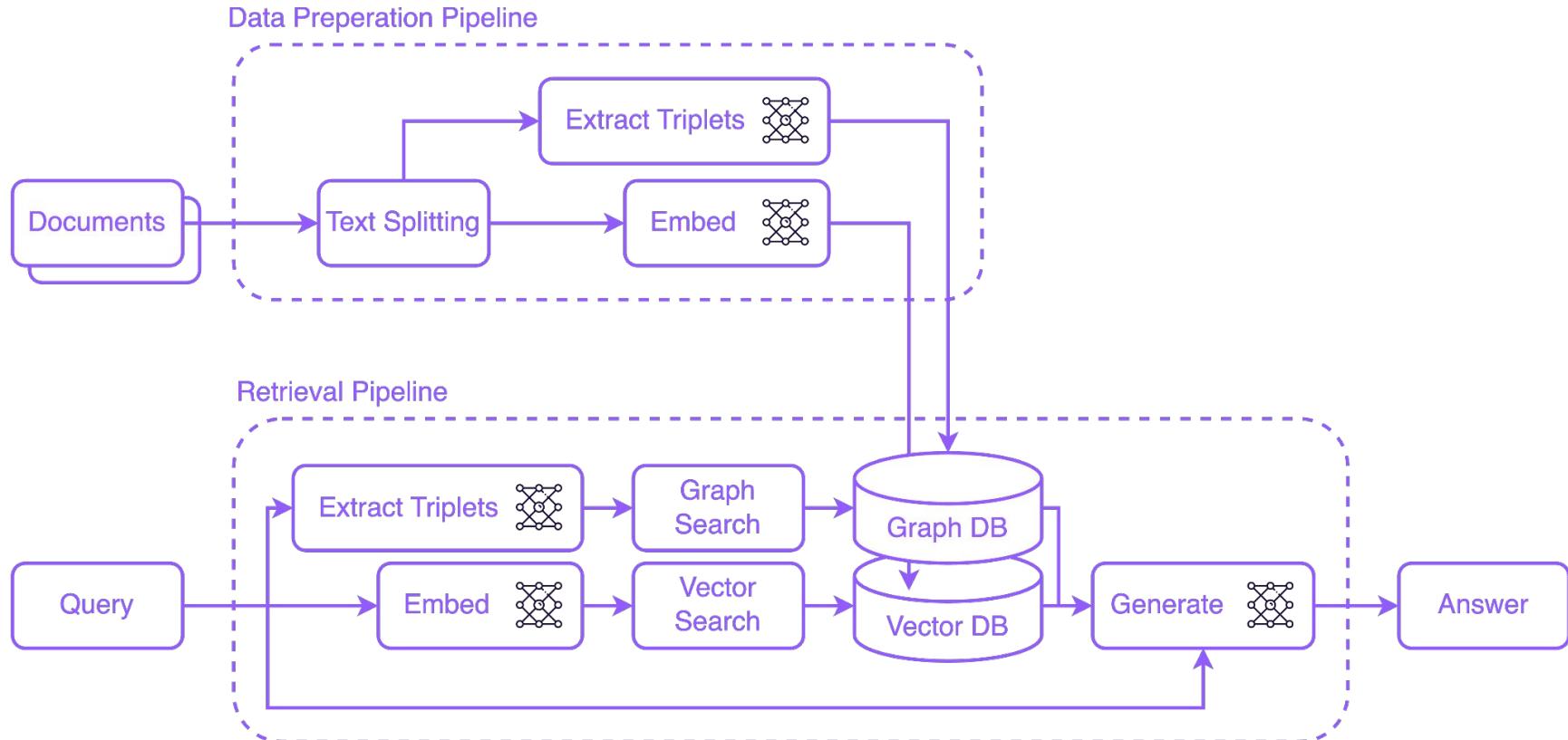
3. Data Preparation Improvements

RAPTOR

(Recursive Abstractive Processing for Tree-Organized Retrieval)



GraphRAG



Graphs Are Fact Driven

Graphs are driven by facts consisting of a subject, a predicate and an object. Altogether these are known as a triplet structure.

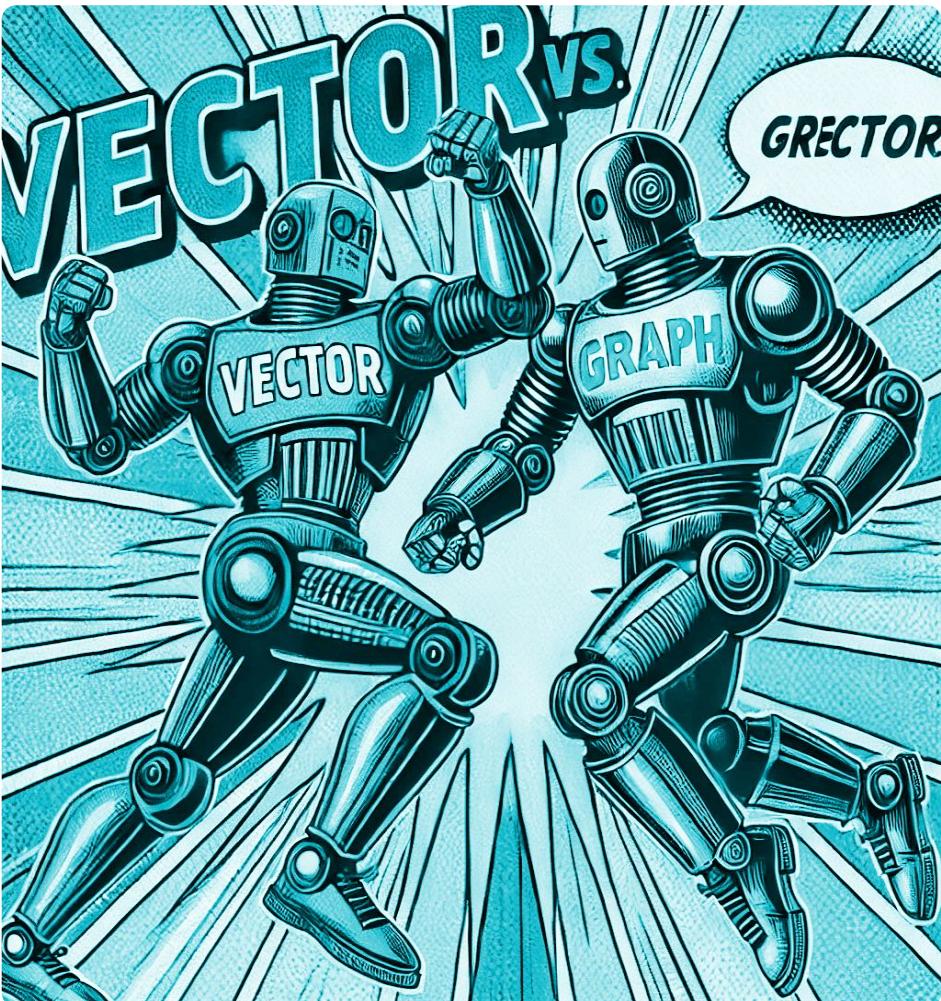
Subject — Predicate → Object
[ARMAND] —[CREATES]→ [SHAREHOLDER VALUE]

Graphs

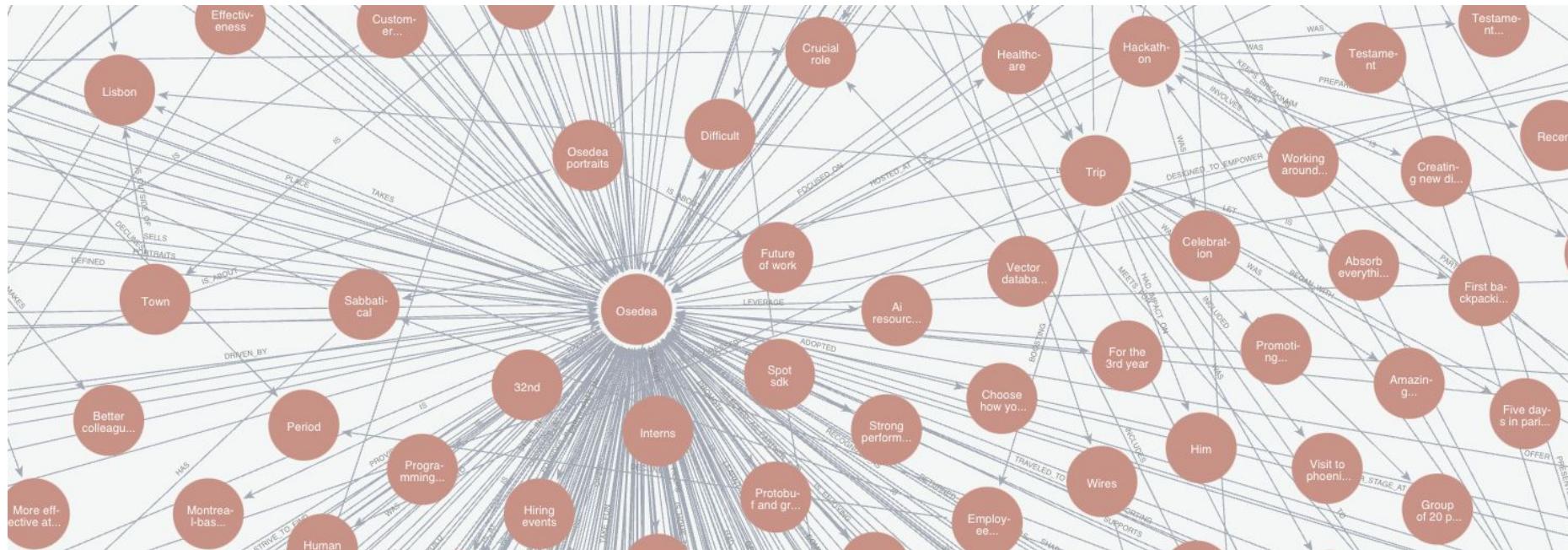
- Search based on relationships
- Low dimensionality
- Observable
- Slower search requires navigating relationships
- Highly explainable!

Vectors

- Search based on semantic similarity
- High dimensionality
- Difficult to observe without losing context
- Fast scalable search



Look at this cool graph!



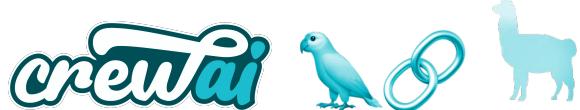
A simple RAG Pipeline is not enough!

LLMs are getting cheaper, better and faster

Last year agents were mostly disappointing and generally fell flat.

Today, most industry leaders are developing their own agents and agent frameworks.

 Semantic Kernel



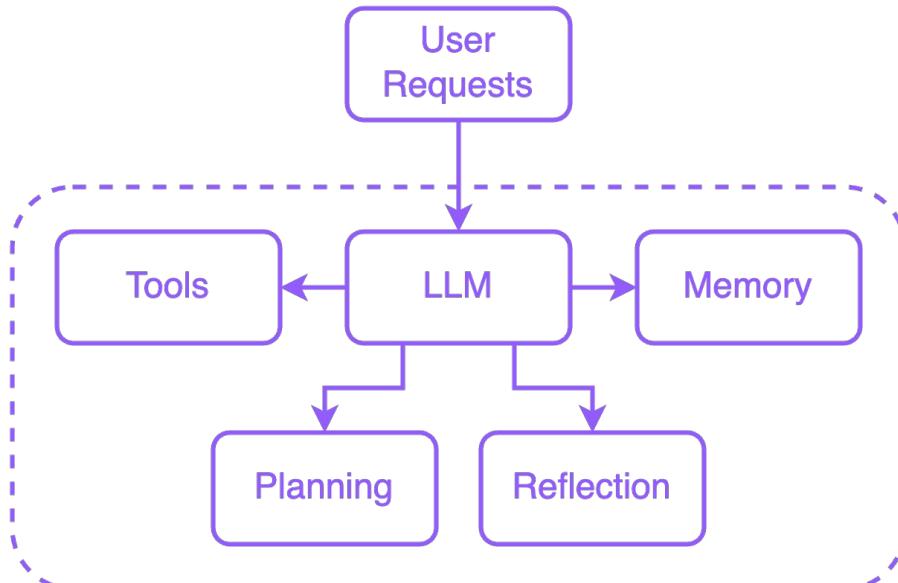
Agentic RAG Systems

What is an Agent?

An LLM Agent is a system that can use an LLM to reason through a problem, create a plan to solve the problem, and execute the plan with the help of a set of tools.

Key Features / Benefits of Agentic RAG:

- Orchestrated Question Answering
- Goal-Driven
- Planning and Reasoning
- Tool Use and Adaptability



Deep Search

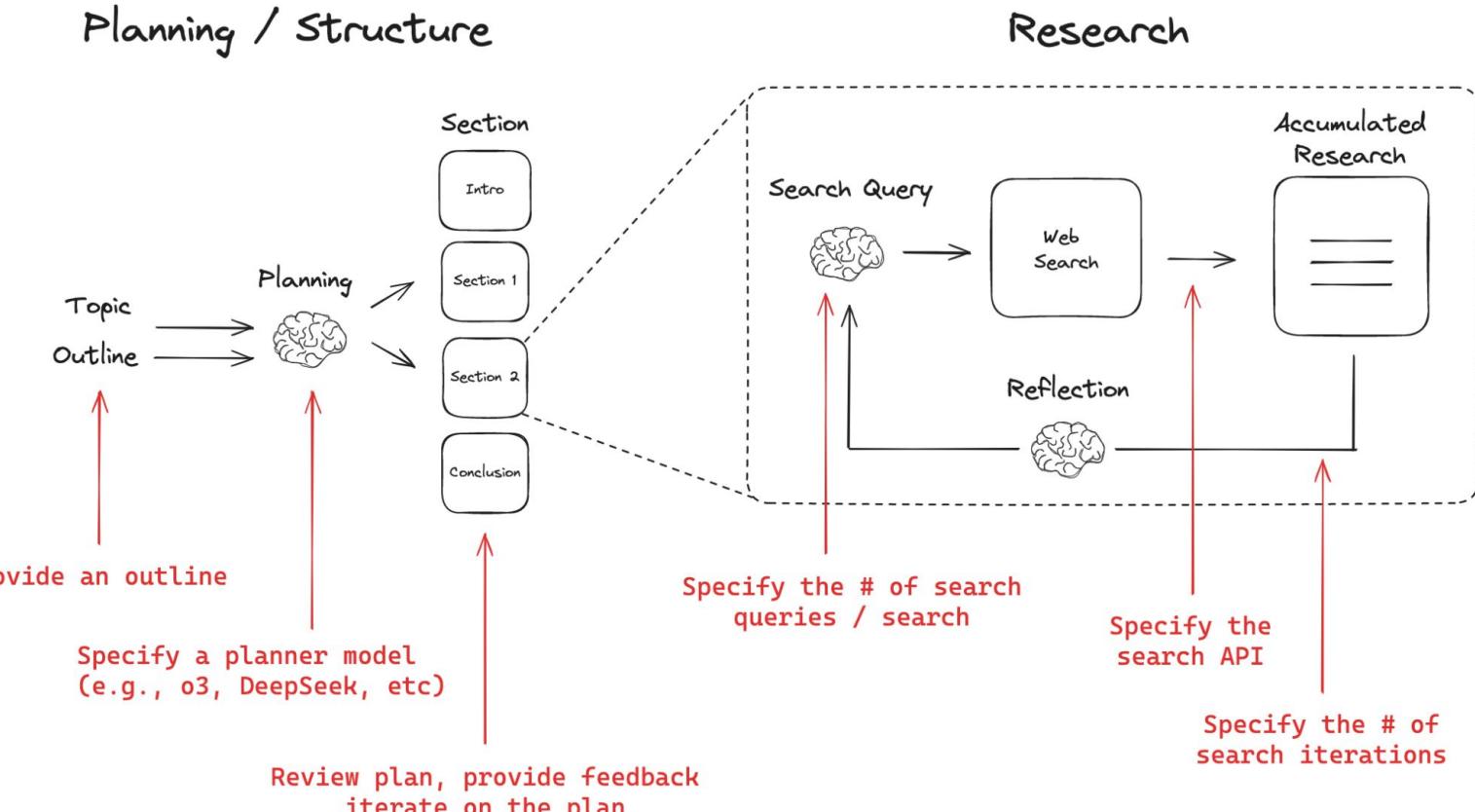
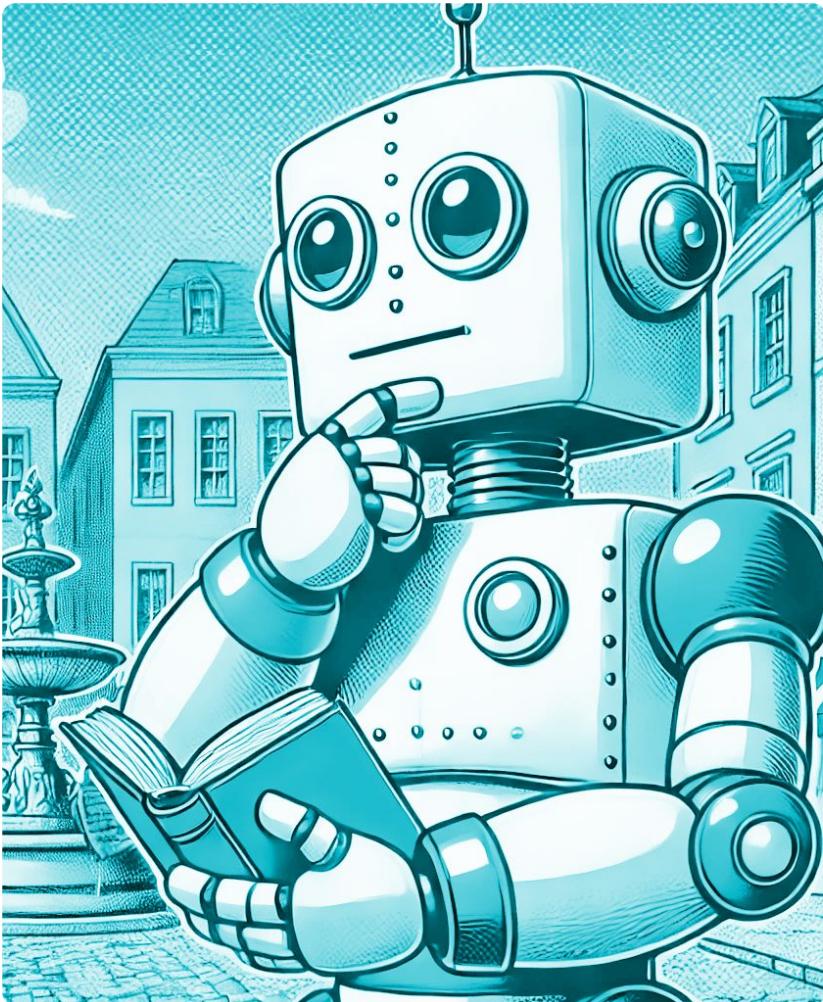


Diagram from langchain

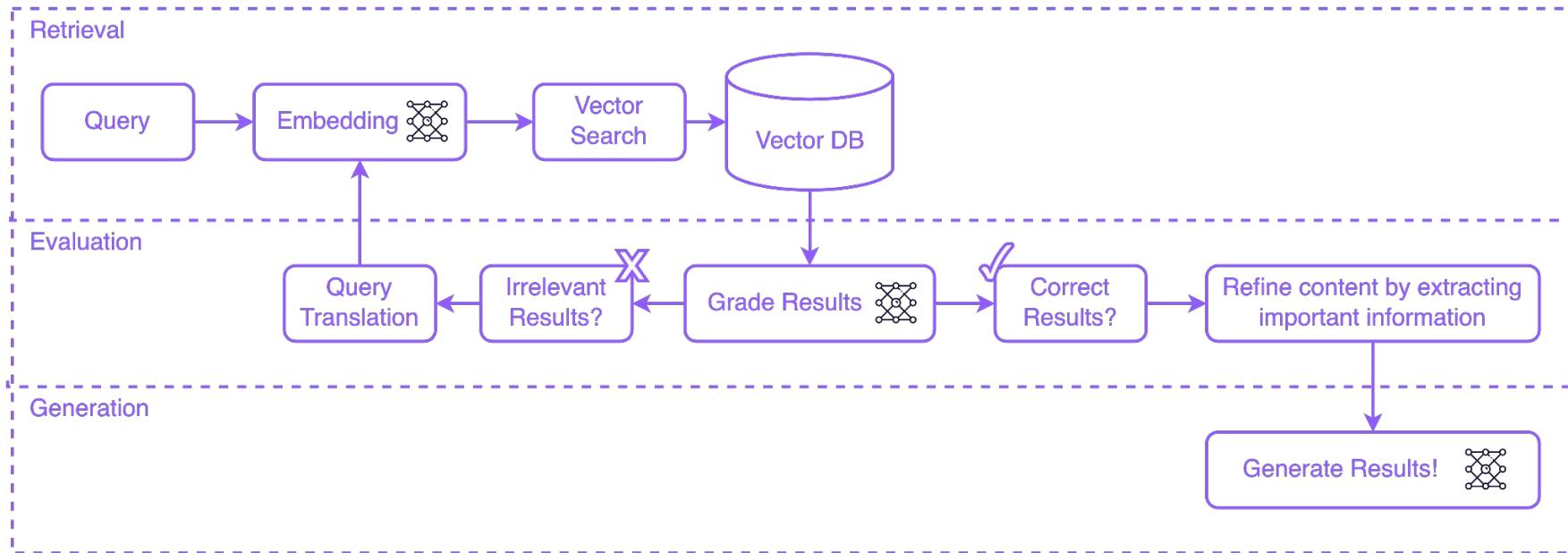
Agentic RAG Patterns

Self-Critique & Reflection

- Traditionally LLMs output an immediate response
- LLMs can critically evaluate their generated outputs
- This adds a layer of fault tolerance to RAG



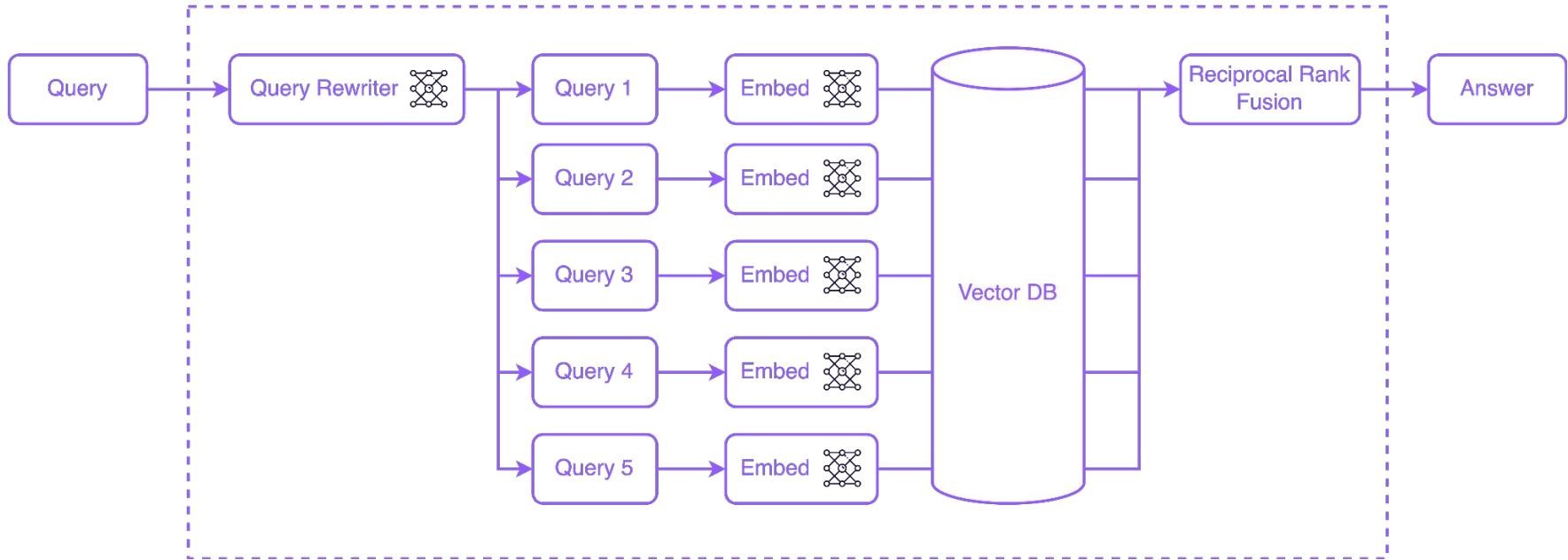
Corrective RAG



Corrective Retrieval Augmented Generation

Shi-Qi Yan, Jia-Chen Gu, Yun Zhu, Zhen-Hua Ling

Query Translation and RAG Fusion



RAG-Fusion: a New Take on Retrieval-Augmented Generation

Zackary Rackauckas

Query Translation and RAG Fusion

5 Perspectives for the following query:

What's the average annual energy expenditure associated with operating the hockey rinks we build?

1. Project Manager

"Can you provide detailed data on the average annual energy costs incurred during the operation of our hockey rinks? This will help us plan budget allocations and identify areas for potential cost savings."

2. Financial Analyst

"What is the average annual energy expenditure for maintaining our hockey rinks? Understanding these figures is essential for our financial forecasting and ensuring cost-efficiency."

3. Environmental Consultant

"Could you share the average annual energy consumption figures for the hockey rinks we've constructed? This information is crucial for assessing the environmental impact and implementing sustainability measures."

4. Facility Manager

"I need to know the average annual energy usage for the hockey rinks we operate. This data will assist in optimizing operational efficiency and managing our energy resources more effectively."

5. Marketing Director

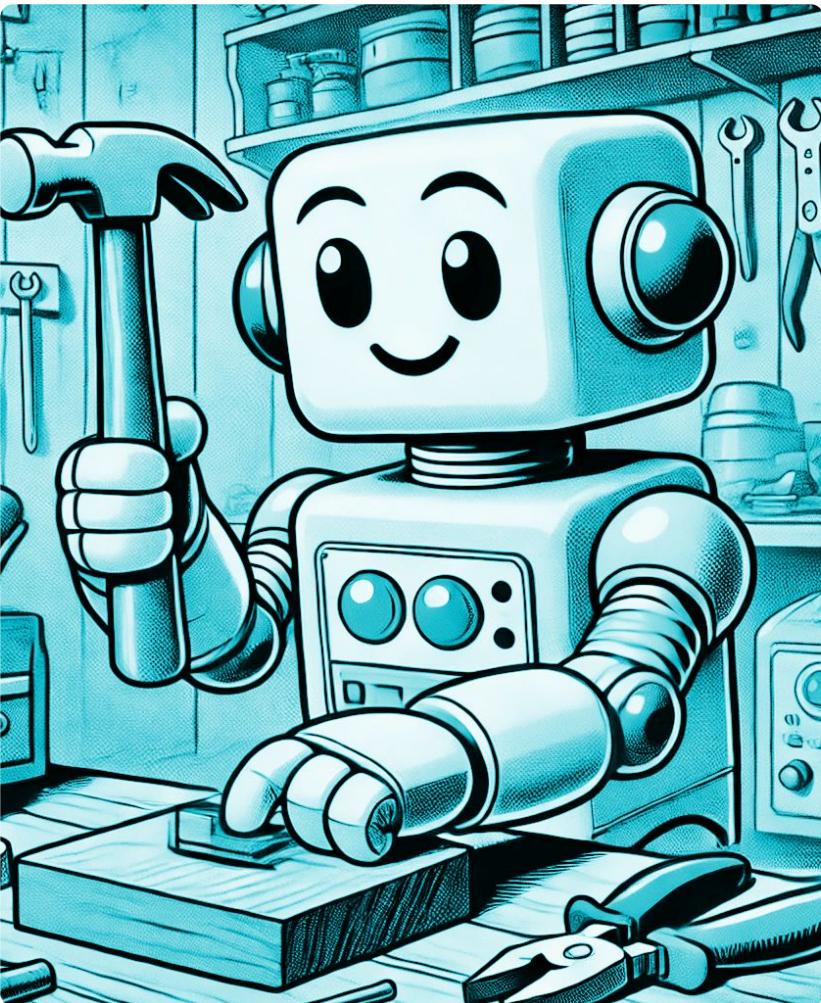
"What are the typical annual energy costs associated with our hockey rinks? Having these insights will allow us to highlight our efficiency and sustainability efforts in our promotional materials."

Tool Use

- Allows agents to execute functions
- Tools for Information retrieval
- Tools for data manipulations
- Integral part of agentic workflows

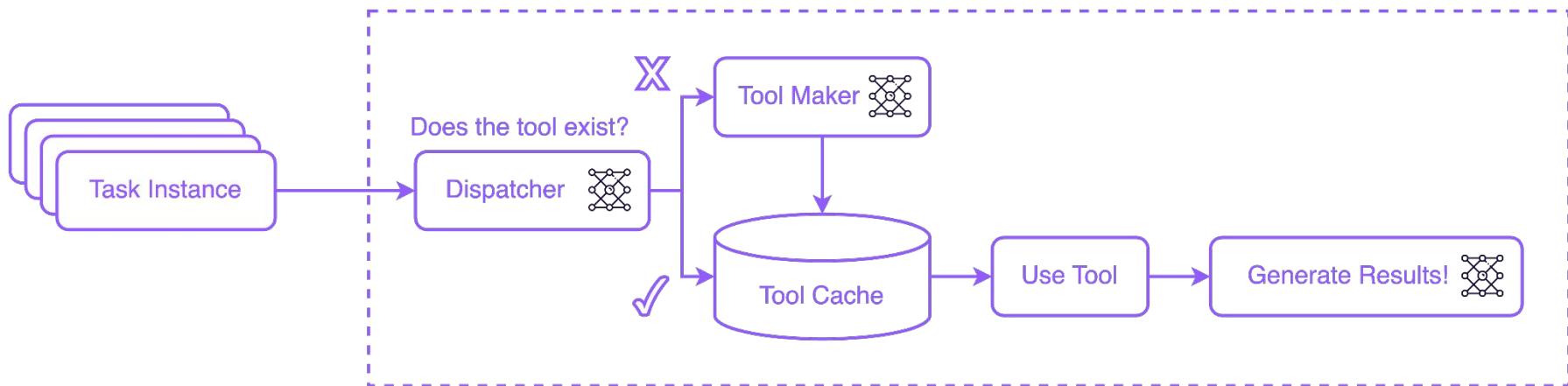
Example Tools:

- Web Search
- SQL Search
- Vector Search
- Calculator
- Computer Vision



Tool Generation

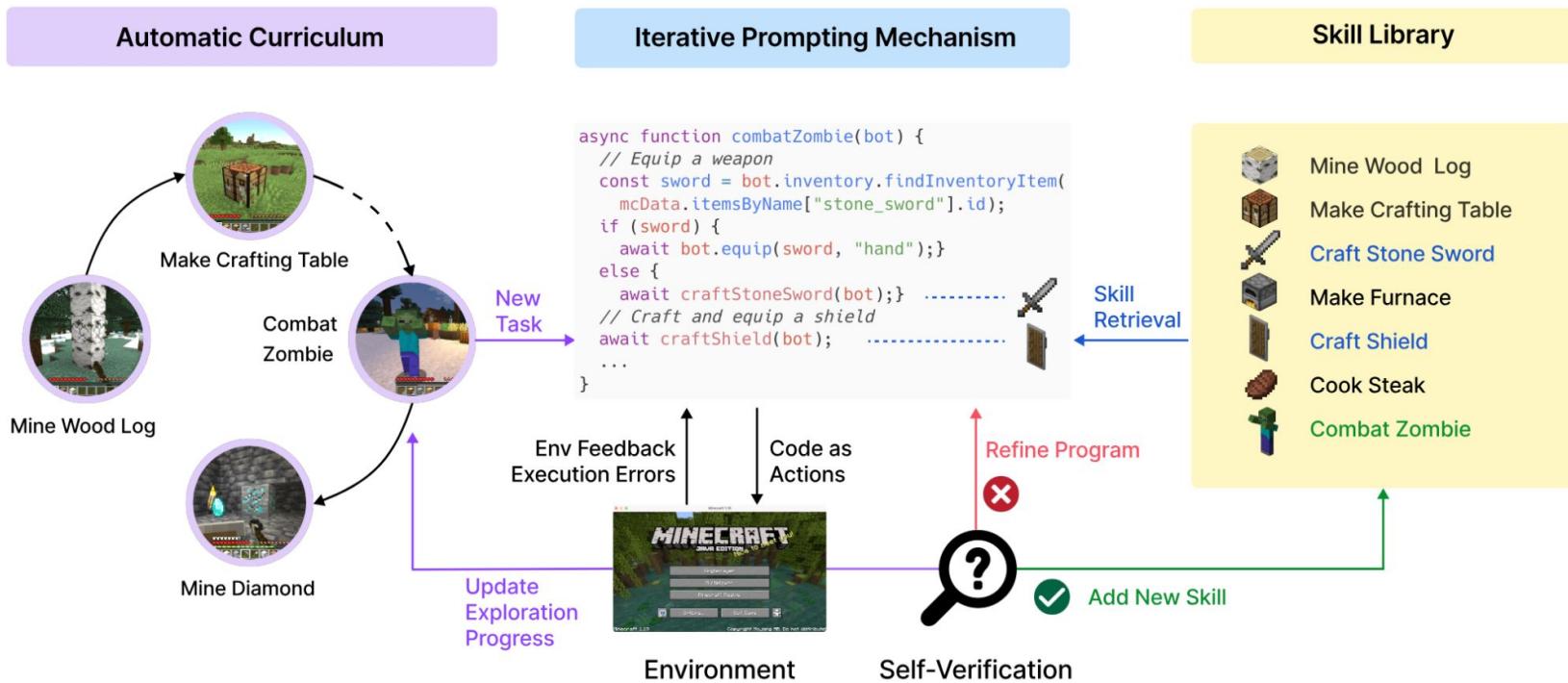
The "Large Language Models as Tool Makers" (LATM) framework is an approach that enables LLMs to create and utilize their own tools for problem-solving.

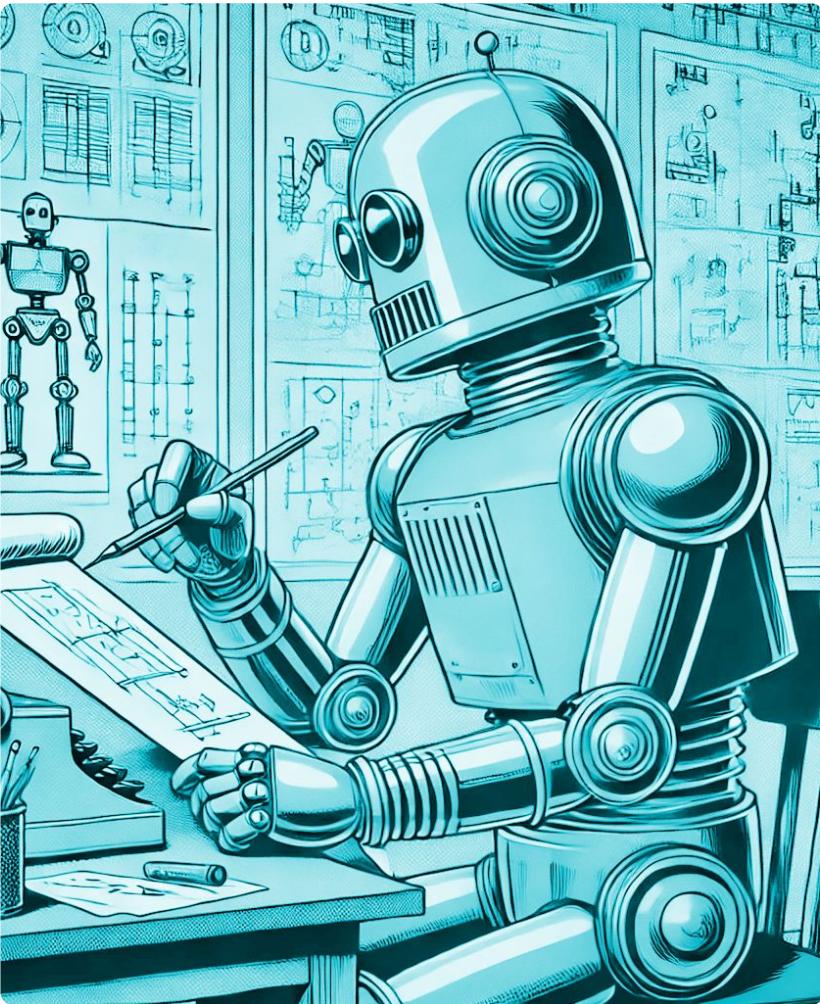


Large Language Models as Tool Makers

Tianle Cai, Xuezhi Wang, Tengyu Ma, Xinyun Chen, Denny Zhou

Case Study: Voyager (Minecraft Agent)



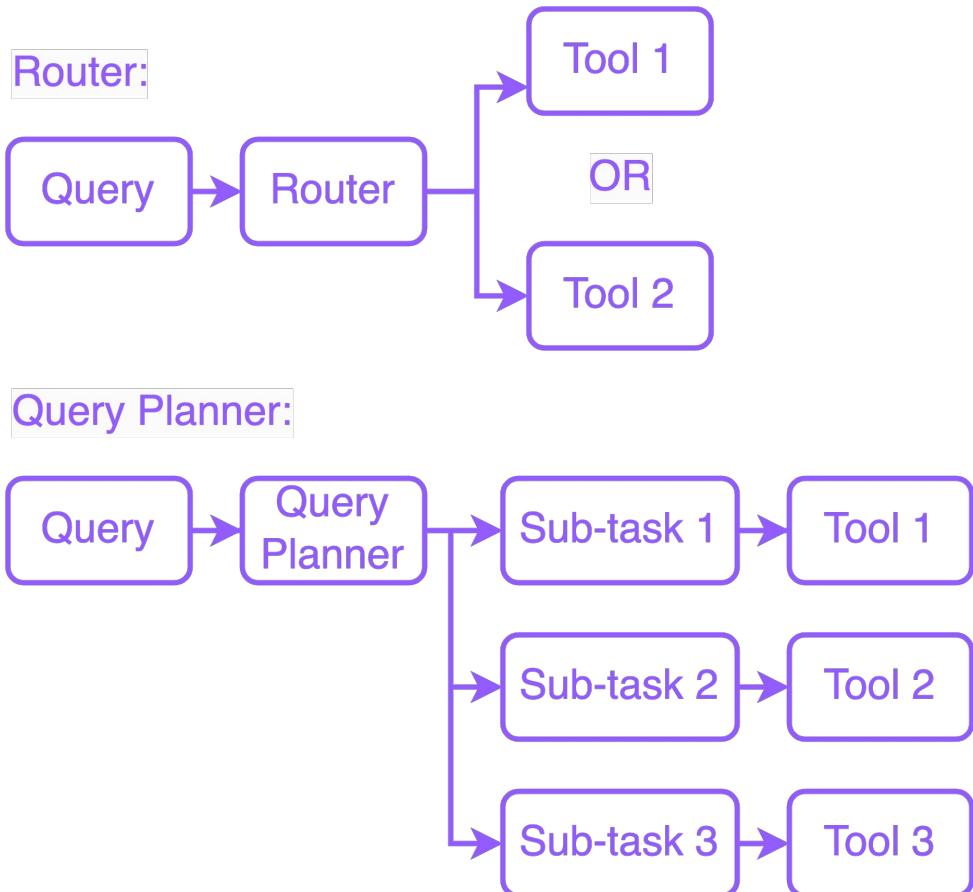


Planning

- Agents employ reasoning to tackle complex tasks in multiple steps
- Each query requires a specific strategy
- Agents adjust plans as tasks progress
- They can route and plan ahead

Routing & Query Planning

Routing is when an LLM is given a set of tools to choose from based on a given query. It may then chose to route to one or more actions to retrieve data.



ReAct Loop (Reason + Act)

What's the average annual energy expenditure associated with operating the hockey rinks we build?



Thought 1: I need to find identify hockey rinks we've built in the past by looking at past case studies and extracting their project IDs

Act 1: Vector Query Engine [Case Study Index]

Observation 1: 1 large and 2 small hockey rinks have been built for the following projects: P#0124, P#0456, P#1890

Thought 2: I need to generate an SQL query to AVG the energy expenditures from the project metadata table using these projects P#0124, P#0456, P#1890

Act 2: SQL Query Engine [Project Metadata]

Observation 2: The average annual energy expenditure is \$200,000

Thought 3: Using the annual expenditures of the 3 rinks your organization has built in the past, the average annual energy expenditure for a hockey rink is \$200,000.

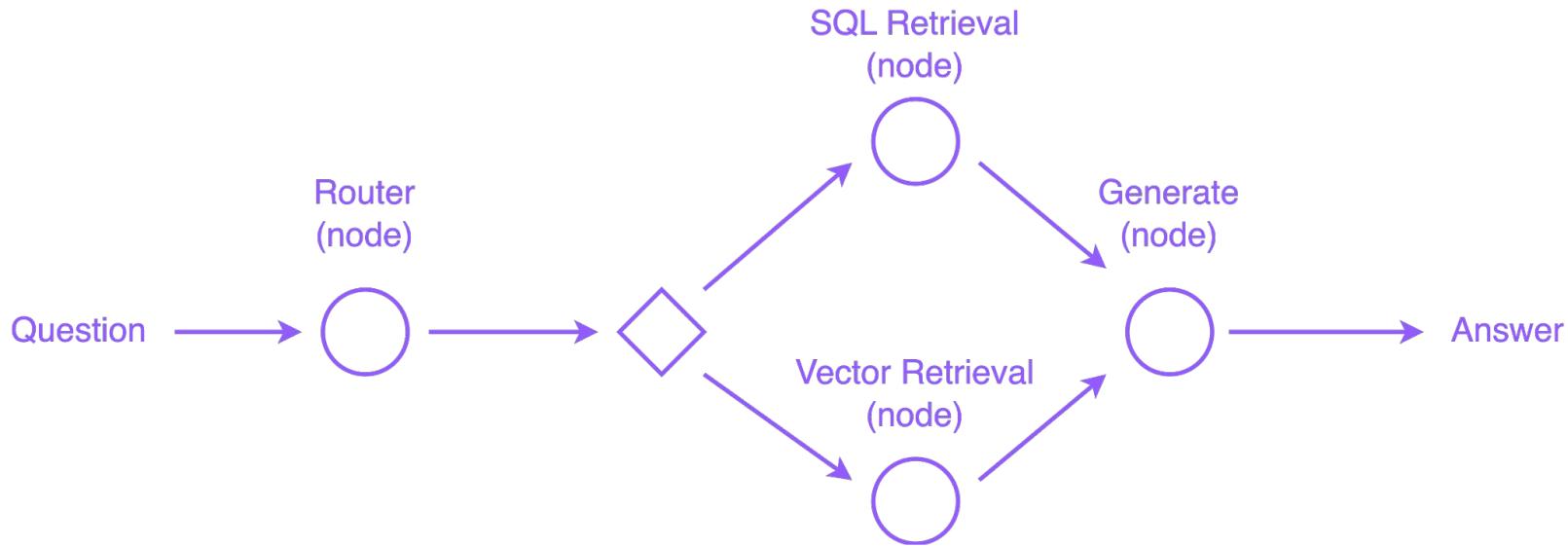
Act 3: FINISH

ReAct:
Synergizing
Reasoning
and Acting
in Language
Models

Shunyu Yao,
Jeffrey Zhao,
Dian Yu, Nan
Du, Izhak
Shafran, Karthik
Narasimhan,
Yuan Cao

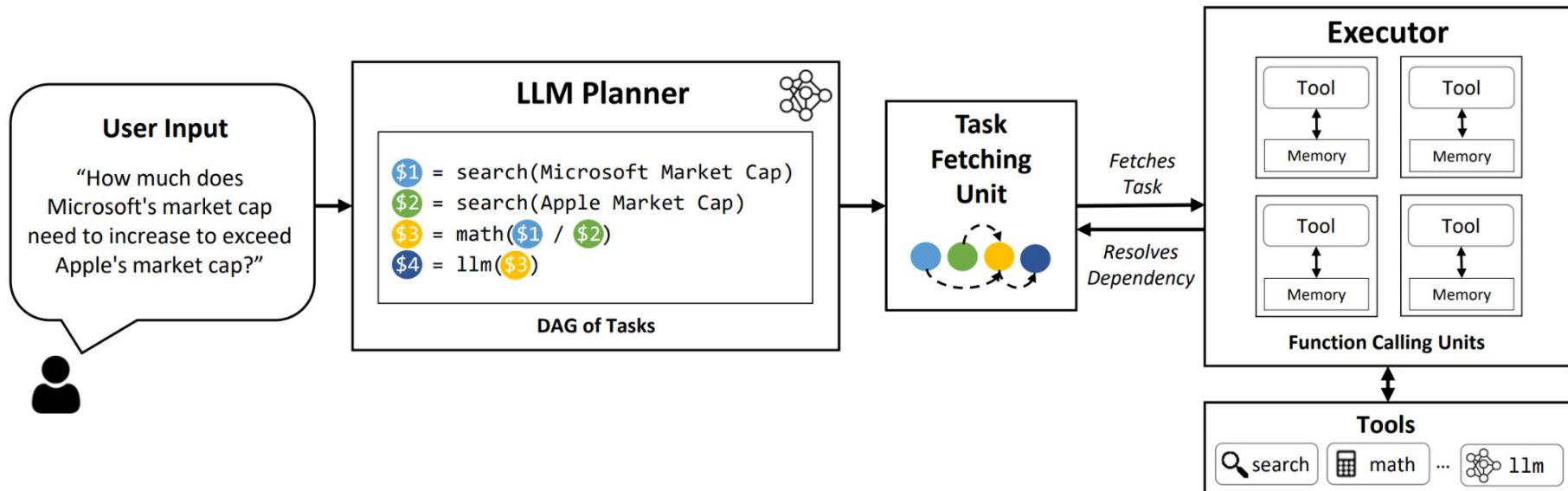
Directed Graphs based approaches for RAG

Directed Graphs represent the dependencies and relationships between different stages of the RAG process using edges and nodes. It's declarative and observable.



LLM Compiler

A dynamic planning framework that enables an efficient and effective orchestration of parallel function calling with LLMs.

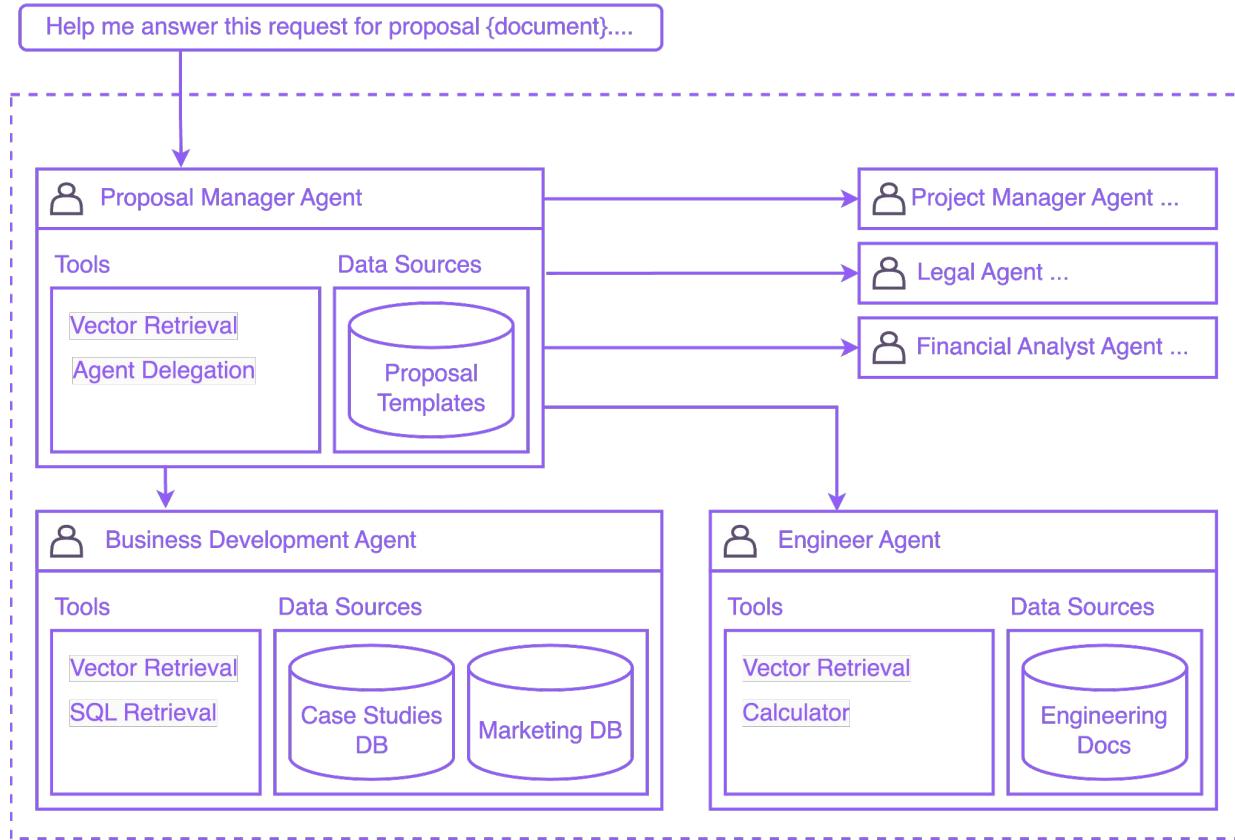




Multi-Agent Collaboration

- Agents using other agents as tools to delegate work
- Avoids giving too many responsibilities to one agent
- Allows for parallelized executions

Hierarchical Multi-Agent Collaboration



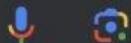
Going To Production

A few notes on production ready agents

- Planning and multi-agent collaboration can still be unreliable.
- Consider **token cost, latency, accuracy**.
- Guardrails like rate limiting, delegation limits, prompt injection defense are crucial.
- Don't blindly trust sources.
- Watch out for data growth and bias.



cheese not sticking to pizza



All Images Videos Forums Shopping News Web

♦ AI Overview

Learn more :

Cheese can slide off pizza for a number of reasons, including **too much sauce, too much cheese, or thickened sauce**. Here are some things you can try:

- Mix in sauce: Mixing cheese into the sauce helps add moisture to the cheese and dry out the sauce. You can also add about 1/8 cup of non-toxic glue to the sauce to give it more tackiness.
- Let the pizza cool: The cheese will settle and bond



fucksmith • 11y ago

To get the cheese to stick I recommend mixing about 1/8 cup of Elmer's glue in with the sauce. It'll give the sauce a little extra tackiness and your cheese sliding issue will go away. It'll also add a little unique flavor. I like Elmer's school glue, but any glue will work as long as it's non-toxic.



8



Reply

...



More replies

Thank you

Carl Lapierre

Tech Lead

How to contact us

osedea.com

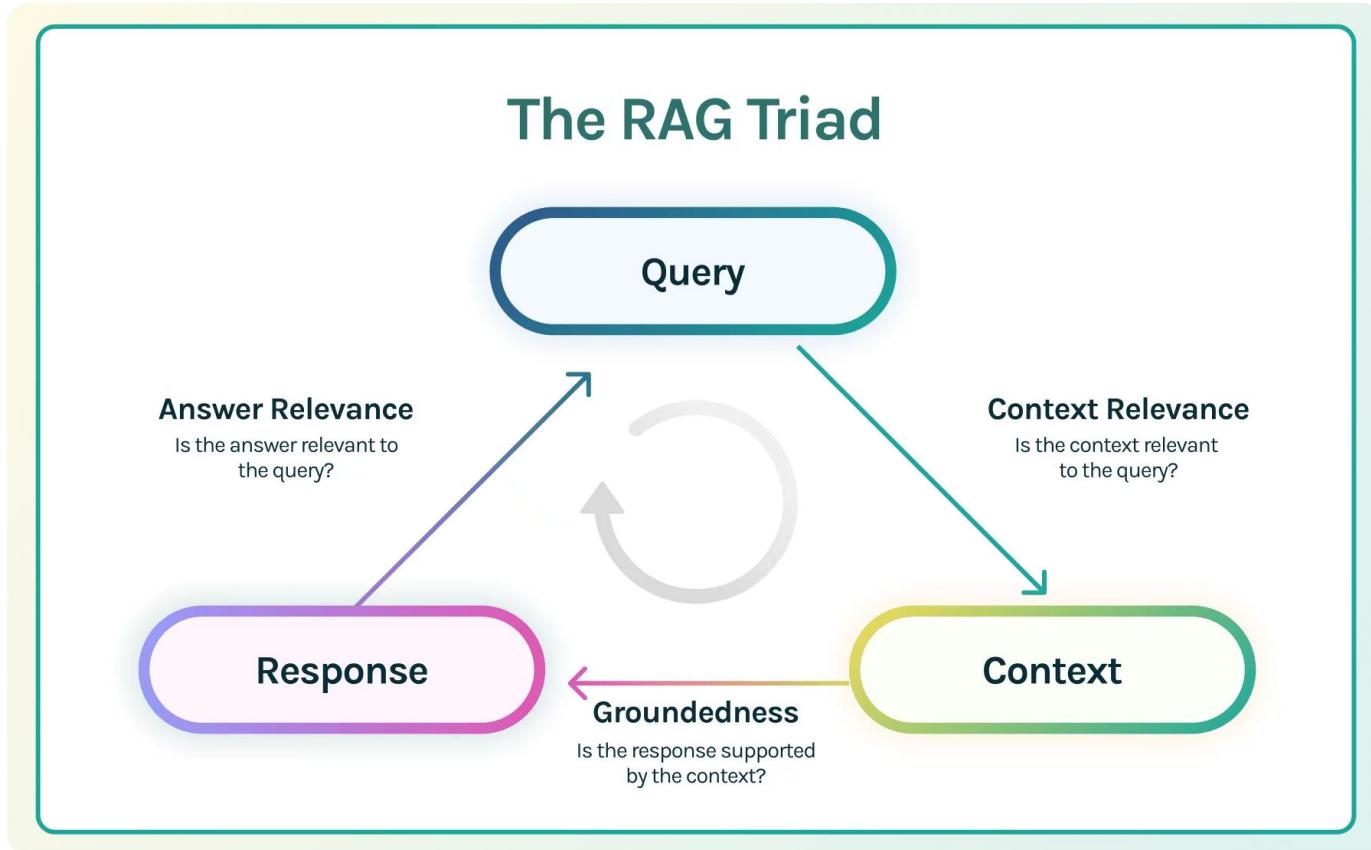
info@osedea.com

[LinkedIn](#)

Feedback form

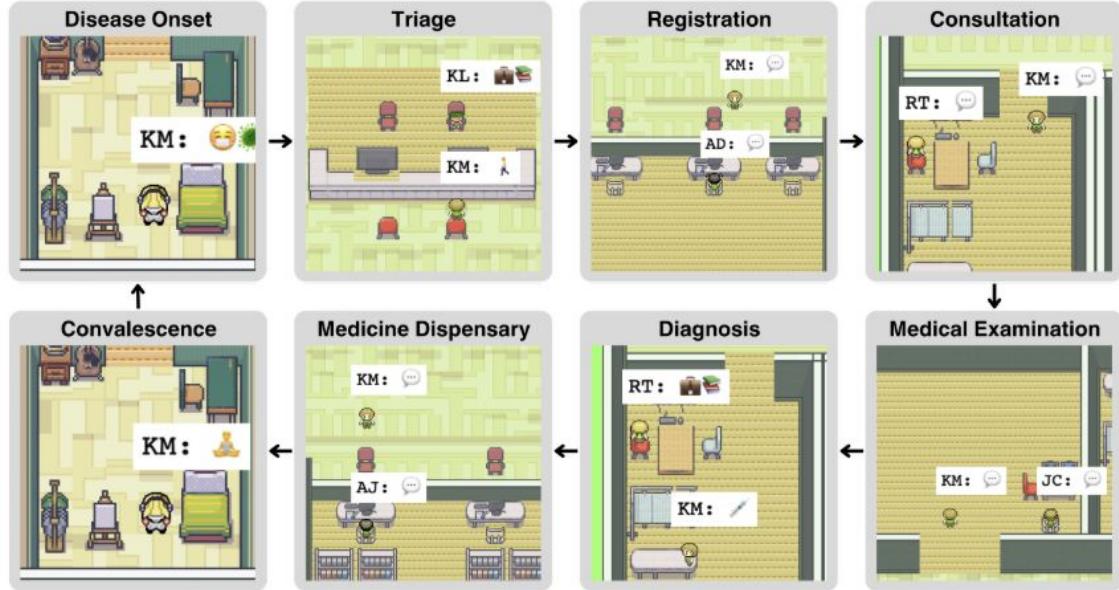


RAG Triad



Case Study: Agent Hospital

A virtual hospital environment where patients, nurses, and doctors are autonomous agents powered by LLMs.



Agent Hospital: A Simulacrum of Hospital with Evolvable Medical Agents

Junkai Li, Siyu Wang, Meng Zhang, Weitao Li, Yunghwei Lai, Xinhui Kang, Weizhi Ma, Yang Liu