

Week 2: Prompt Engineering

 作者: Alwin Lin

Objectives

- Understand key concepts in prompt engineering.
- Explore applications of various prompting techniques.
- Engage in collaborative problem-solving activities.

Table of context

- In Context Learning
- Few shot
- Echo Prompting
- System Prompts
- Chain of thought
- Tree of thought
- Meta prompting
- Directional stimulus prompting
- Retrieval Augmented Generation
- Medprompt

What are LLMs?

Key Components:

- **Neural Networks:** LLMs use deep learning techniques, particularly transformer architectures, which consist of multiple layers (embedding, attention, feedforward).
- **Training:** They are trained on huge datasets using self-supervised learning to recognize patterns and relationships in text.

How Do They Work?

- Input Processing:** LLMs receive text input and convert it into numerical representations (embeddings).
- Context Understanding:** Using attention mechanisms, they focus on relevant parts of the input to generate context-aware responses.
- Output Generation:** The model predicts the next word or phrase based on learned patterns, producing coherent text outputs.

What does it mean for us?

- LLMs is effectively a very powerful auto fill machine
- Ask the right question, and get the right response.

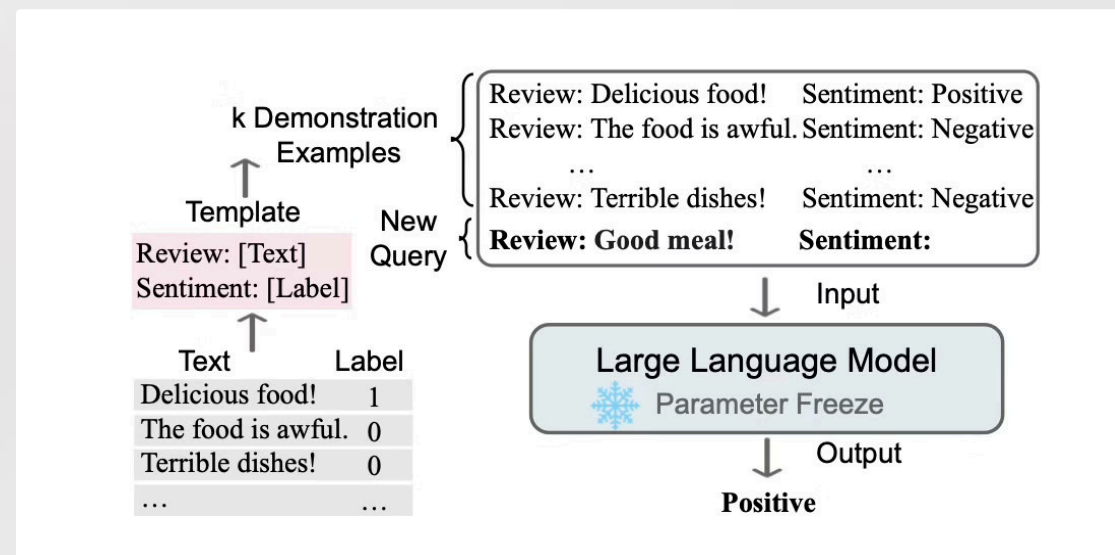
neuron activation



In context learning

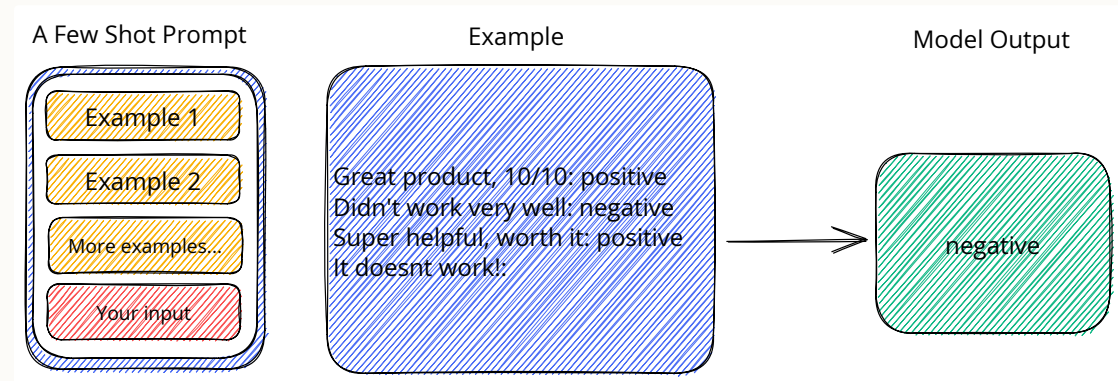
A Survey on In-context Learning

- **Definition:**
 - Providing a prompt without any examples.
- **Application:**
 - Useful for straightforward queries where context is clear.
 - Useful for looking up quick facts/articles



Few shot prompting

- **Few-Shot:** Offering multiple examples for better context.
- **Use Cases:** Enhancing response accuracy in complex tasks.



Echo Prompting

EchoPrompt: Instructing the Model to Rephrase Queries for Improved In-context Learning

- **What is it?**
 - Effectively speaking, asking the LLM to rethink a given problem again
- **How do I use it?**
 - "Let's repeat the question and also think step by step"
 - "Read the question again"

Q: Kelly has 5 quarters and 2 dimes. If she buys a can of pop for 55 cents, how many cents will she have left?

Zero-Shot-CoT

A: Let's think step by step. Kelly has 5 quarters and 2 dimes. So, she has 5 quarters and 2 dimes left. $5 \text{ quarters} = 5 \times 25 \text{ cents} = 125 \text{ cents}$ and $2 \text{ dimes} = 2 \times 10 \text{ cents} = 20 \text{ cents}$. So, she has $125 + 20 = 145 \text{ cents}$ left.

Therefore, the answer is 145 cents. ✗

Zero-Shot-CoT + EchoPrompt

A: Let's repeat the question and also think step by step. Kelly has 5 quarters and 2 dimes. She buys a can of pop for 55 cents. How many cents will she have left? Let's start with the first step. Kelly has 5 quarters and 2 dimes. We can represent this as: $5 \text{ quarters} = 5 \times 25 \text{ cents} = 125 \text{ cents}$. $2 \text{ dimes} = 2 \times 10 \text{ cents} = 20 \text{ cents}$. So, Kelly has $125 + 20 = 145 \text{ cents}$. She buys a can of pop for 55 cents. We can represent this as: 55 cents . So, Kelly has $145 - 55 = 90 \text{ cents}$ left.

Therefore, the answer is 90 cents. ✓

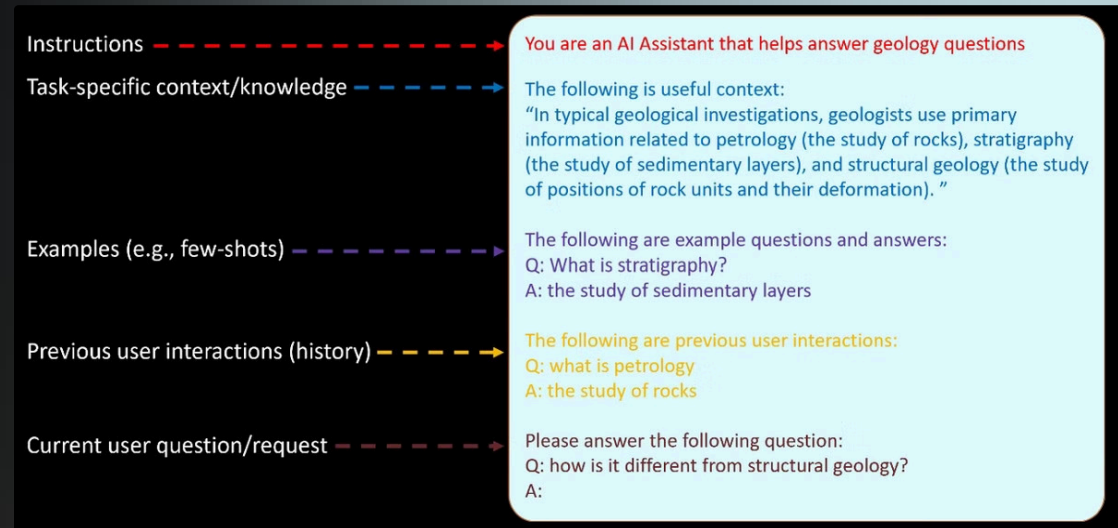
Figure 1: Comparison of prompts in Zero-shot-CoT with and without EchoPrompt, highlighting the modification in prompts. Zero-shot-CoT with EchoPrompt uses the prompt “Let’s repeat the question and also think step by step” to aid the model in recalling the query before solving it.

System Instructions

List of System prompts

- **Why**
 - Using system prompts could lead to more favorable results
 - Roles could "Stimulate" certain areas of the LLM networks
- **What does it look like?**

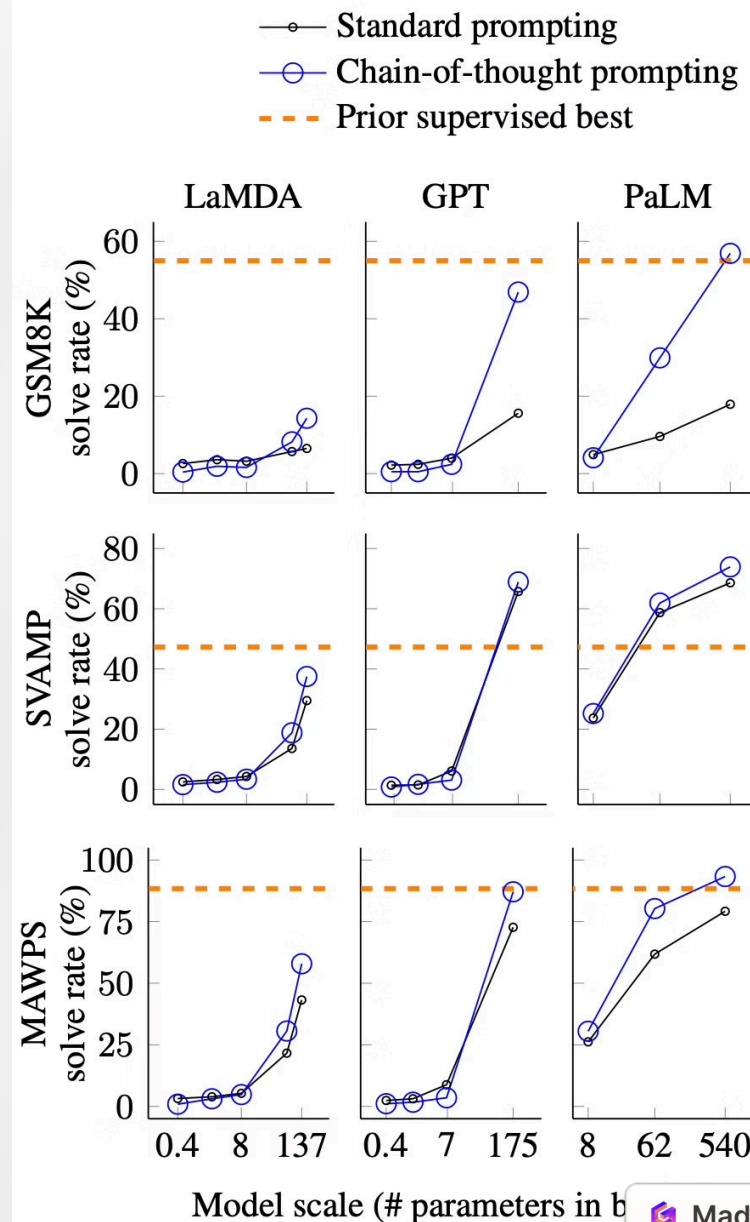
You are a friendly AI agent who can provide assistance to the customer regarding their recent order.



Chain of Thought Prompting

Chain-of-Thought Prompting Elicits Reasoning in Large Language Models

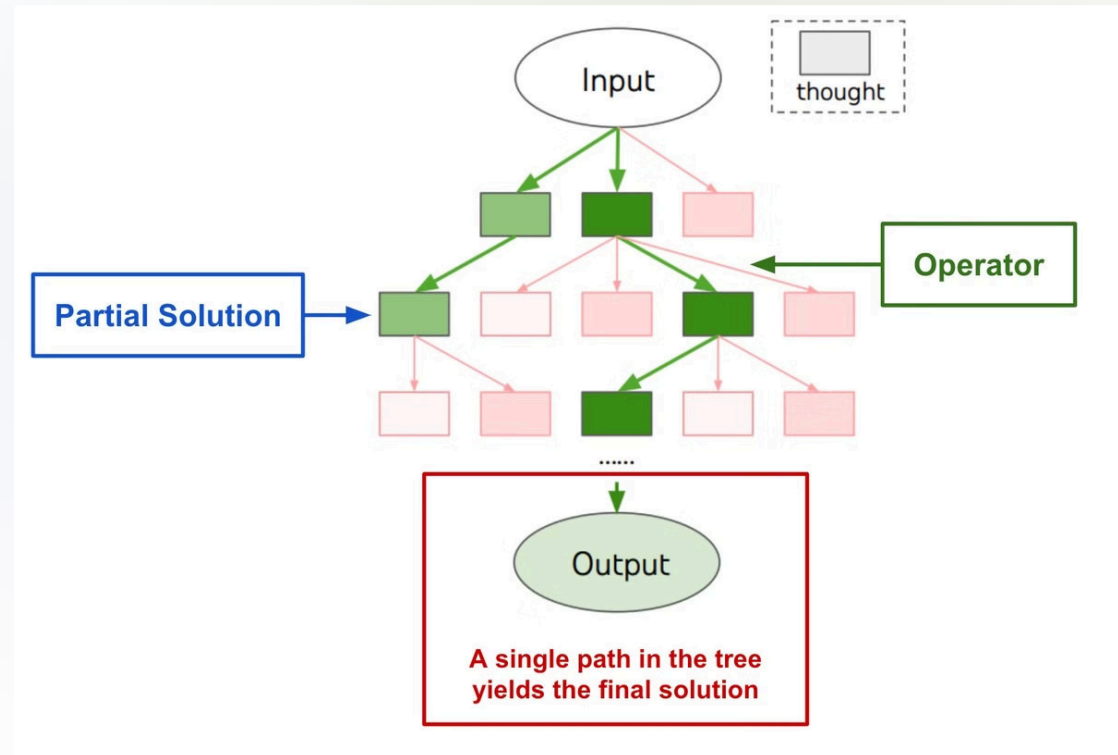
- **Concept:**
 - Encouraging step-by-step reasoning.
- **Example:**
 - “Let’s do this step by step.”
- **Benefits:**
 - Dramatically increases the models' critically thinkig
- **Example:**
 - Reasoning tokens in GPT 4-o



Tree of Thought

Tree of Thoughts: Deliberate Problem Solving with Large Language Models

- Structuring reasoning hierarchically.
 - Extension of Chain of Thought.
 - Voting systems
 - Allows systems to back track in conversations



Meta Prompting

Meta Prompting for AI Systems

- **Definition:** Enhancements over few-shot prompting.
- **Strategy:** Effectively instructing the model to achieve desired outcomes.

Problem Statement:

- **Problem:** [question to be answered]

Solution Structure:

1. Begin the response with "Let's think step by step."
2. Follow with the reasoning steps, ensuring the solution process is broken down clearly and logically.
3. End the solution with the final answer encapsulated in a LaTeX-formatted box, $\boxed{\dots}$, for clarity and emphasis.
4. Finally, state "The answer is [final answer to the problem].", with the final answer presented in LaTeX notation.

Figure 1: A structure meta prompt presented in markdown format for solving MATH [17] problems.

Problem: Find the domain of the expression $\frac{\sqrt{x-2}}{\sqrt{5-x}}$.

Solution: The expressions inside each square root must be non-negative. Therefore, $x - 2 \geq 0$, so $x \geq 2$, and $5 - x \geq 0$, so $x \leq 5$. Also, the denominator cannot be equal to zero, so $5 - x > 0$, which gives $x < 5$. Therefore, the domain of the expression is $\boxed{[2, 5)}$. Final Answer: The final answer is $[2, 5)$. I hope it is correct.

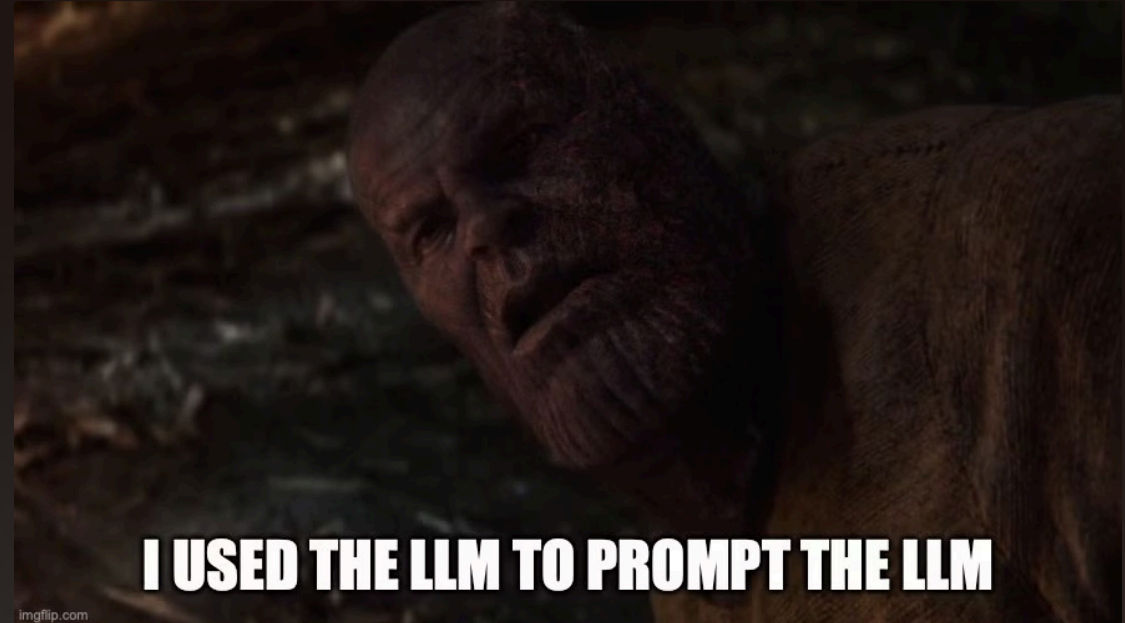
Problem: If $\det \mathbf{A} = 2$ and $\det \mathbf{B} = 12$, then find $\det(\mathbf{AB})$.

Solution: We have that $\det(\mathbf{AB}) = (\det \mathbf{A})(\det \mathbf{B}) = (2)(12) = \boxed{24}$. Final Answer: The final answer is 24. I hope it is correct.

...

Large Language Models as Optimizers

- Concept:
 - Using LLMs to optimize outputs through prompting strategies.

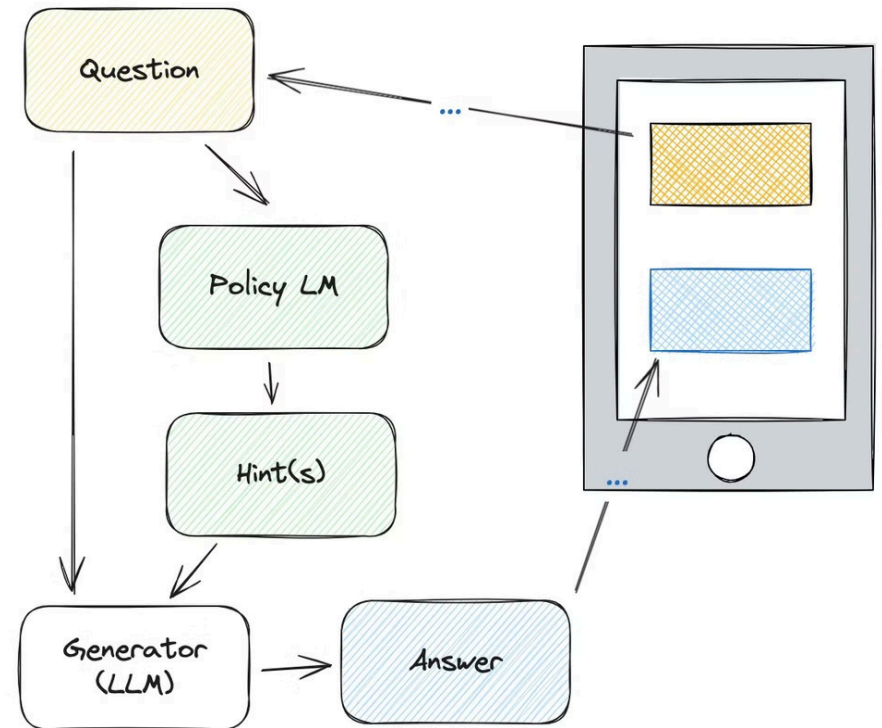


Directional Stimulus Prompting

Guiding Large Language Models via Directional Stimulus Prompting

- **Concept:**
 - Providing hints to guide AI responses.
 - Effectively, RAG 0.5
 - User provided prompts instead of vectors DB
- Effective for things that has a clear rubric and require keywords to be present.

DSP: Directional Stimulus Prompting

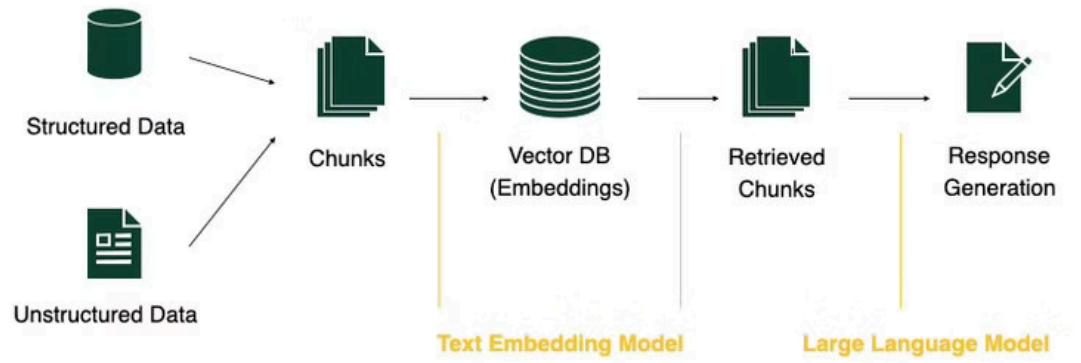


Retrieval-Augmented Generation (RAG)

Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks

- **Mechanism:**
 - Information retrieval via vector search.
 - Limitations in understanding implicit information.

Simple RAG

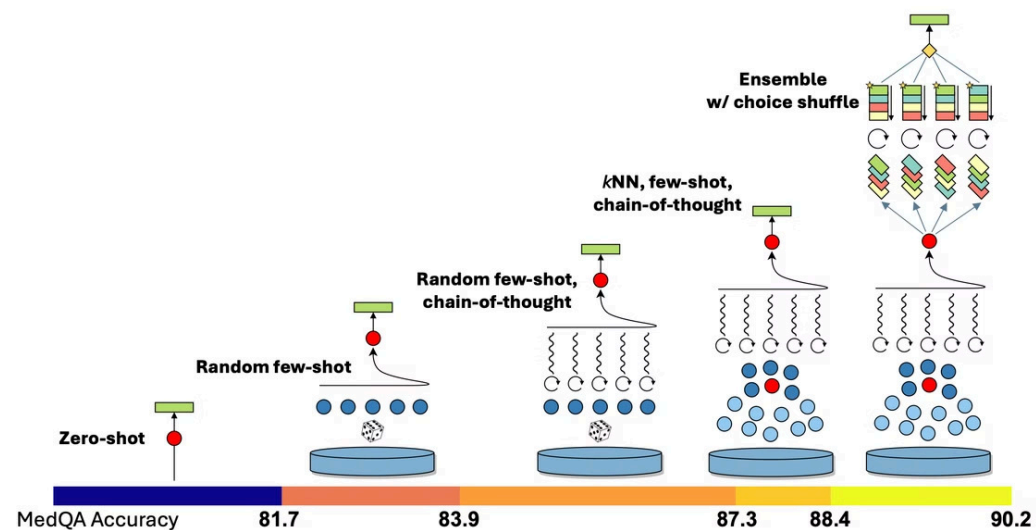


MedPrompt

Can Generalist Foundation Models Outcompete Special-Purpose Tuning? Case Study in Medicine

- **What is this?**
 - A very token intense method of prompting
- **Why are we talking about this?**
 - Prompting techniques and methods can mix
 - One could potentially outperform fine tuned models with enough resource

4.4 Putting it all together: Medprompt



Now, your turn



Accounts Merge - LeetCode

Can you solve this real interview question? Accounts Merge - Given a list of accounts where each element accounts[i] is a list of strings, where the first eleme...



Add Two Numbers - LeetCode

Can you solve this real interview question? Add Two Numbers - You are given two non-empty linked lists representing two non-negative integers. The digits are...



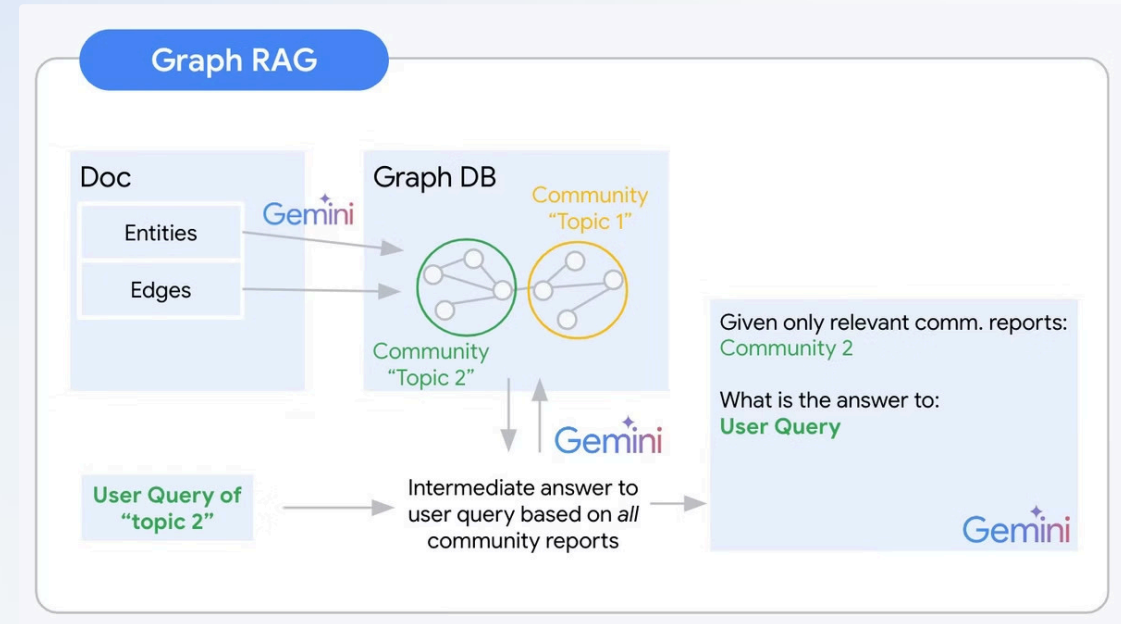
Longest Substring with At Least K Repeating Characters - LeetCode

Can you solve this real interview question? Longest Substring with At Least K Repeating Characters - Given a string s and an integer k, return the length of the...



Graph RAG - RM

- **Benefits:**
 - Capturing insights beyond traditional text analysis.
 - e.g. Relationships, Groupings, Educations
- **Issues:**
 - Graphs need to be generated by either LLMs humans



ReAct Agents & Automatic Multi-Step Reasoning and Tool Use - RM

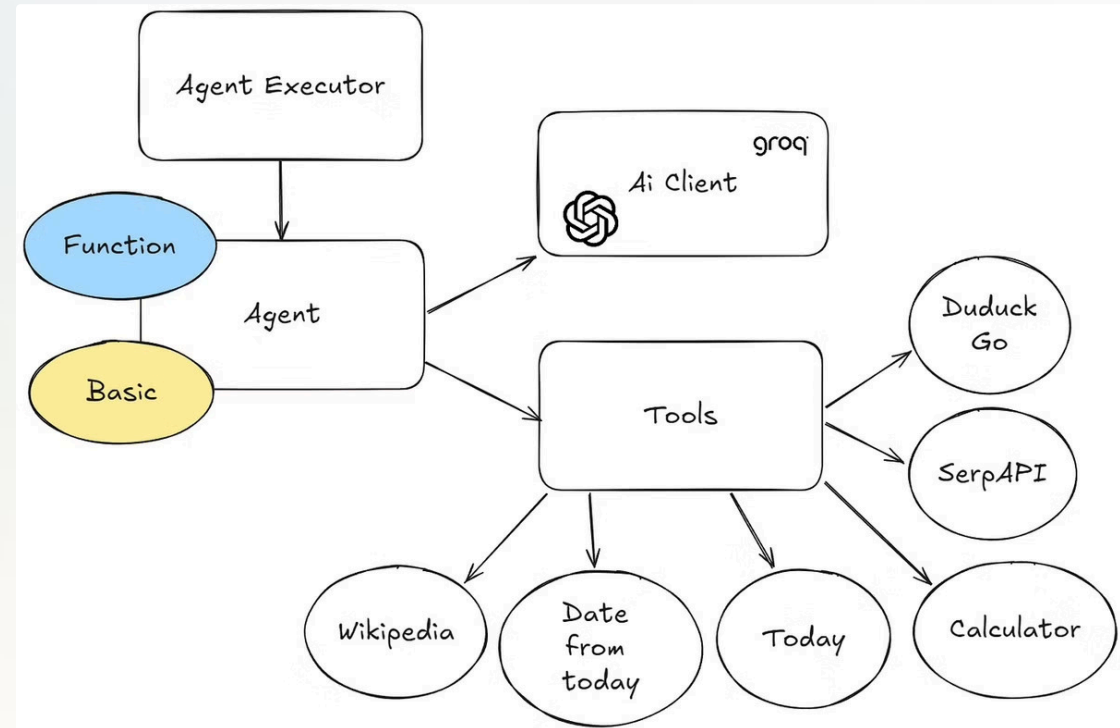
REACT: SYNERGIZING REASONING AND ACTING IN LANGUAGE MODELS & ART: Automatic multi-step reasoning and tool-use for large language models

- **Capabilities:**

- Multi step reasoning
- Acting
- Tool usage

- **Real-world Impact:**

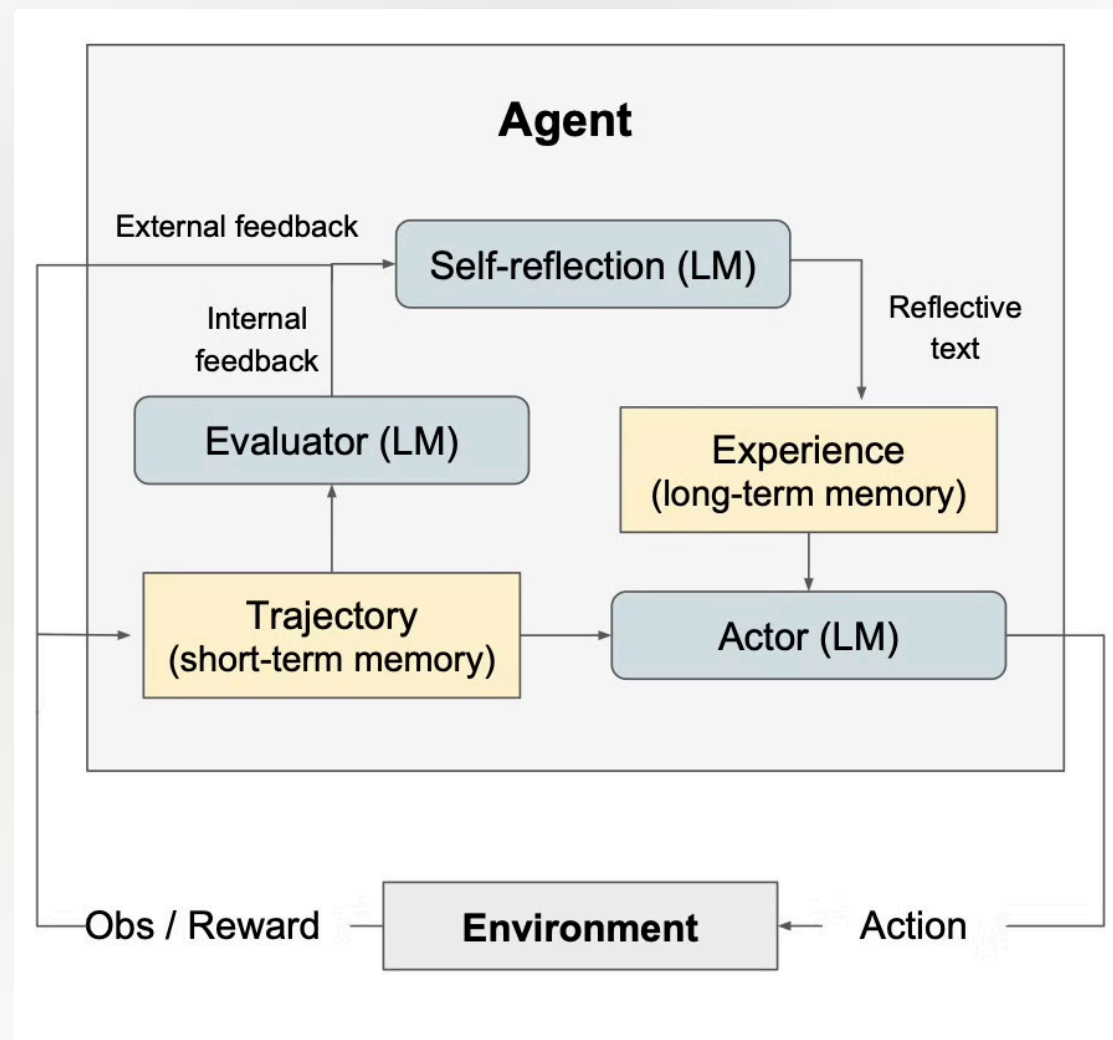
- Salesforce pivot to AgentForce
- Agent swarms



Reflection Techniques -RM

Reflexion: Language Agents with Verbal Reinforcement Learning

- **Importance of Reflection:**
 - Encourages critical thinking about AI responses.
 - Prototype of the baby AGI
- **Issues:**
 - Does lead to loops that can not be automatically resolved



Reinforcement Learning from Human Feedback (RLHF) - RM

Training language models to follow instructions with human feedback

- **Mechanism:**
 - The "Human in the loop"
- **Importance:**
 - Aligns model outputs with user preferences.
- **Applications:**
 - Thumbs up/thumbs down feedback system.
 - How Google's Gemini got lapped by GPT

