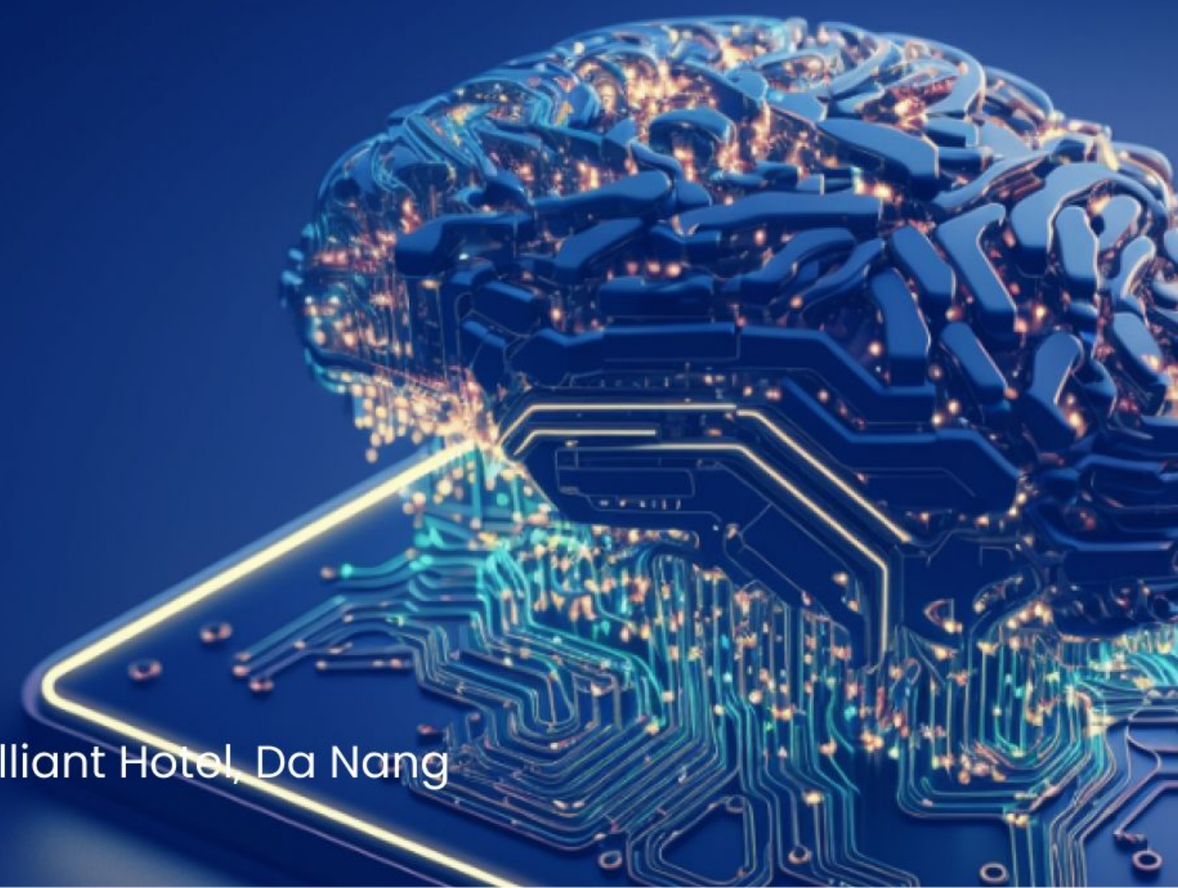
KMS TECHNOLOGY · orient SOFTWARE · PARADOX · ESTROUGE · GDG MienTrung

Da Nang TechTalk

Engineering the Future with GenAI



October 8, 2023

Brilliant Hotel, Da Nang

Da Nang TechTalk

Engineering the Future with GenAI

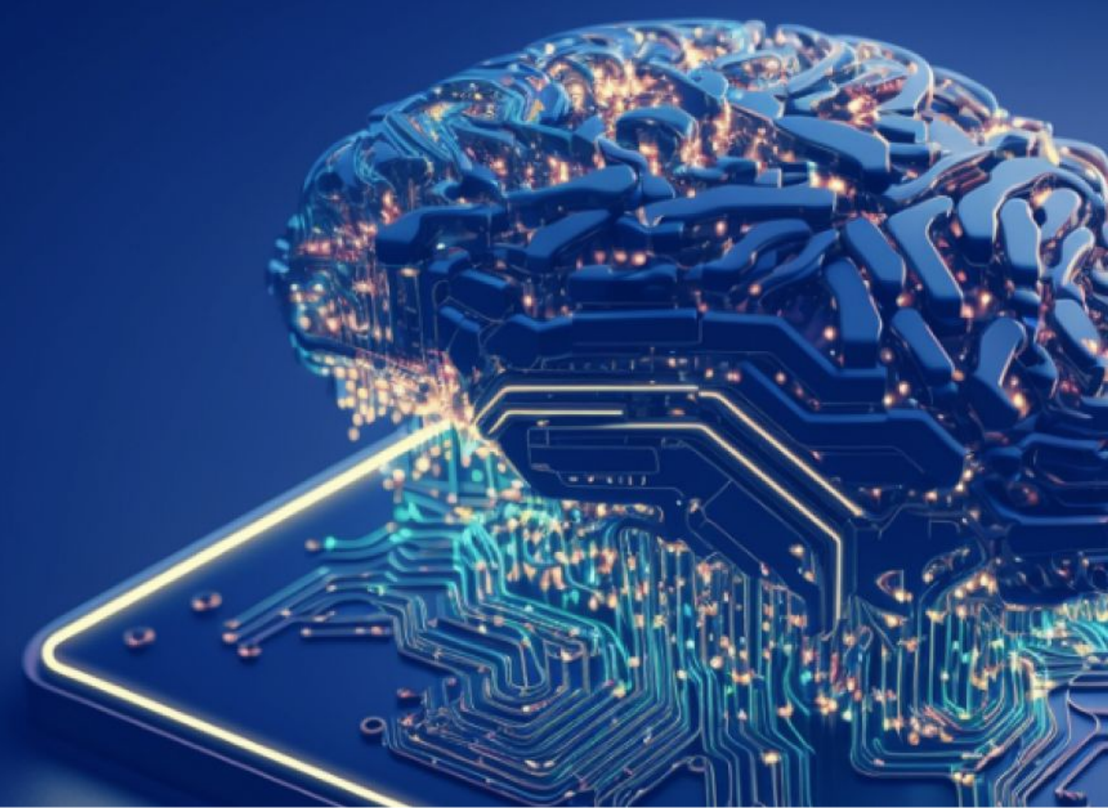
KMS TECHNOLOGY orient SOFTWARE PARADOX ESTROUGE GDG MienTrung

Sharing Session #1

Prompt Engineering Unveiled

Mr. Nguyen Ton

AI Team Leader at Neurond Technology, Orient Software



"Writing a really great prompt for a chatbot persona is an amazingly high-leverage skill and an early example of programming in a little bit of natural language"

— Sam Altman

What is Prompt Engineering?

OUTLINE

What & Who	Prompt Engineering in different approaches
Why & When	Prompt Engineering vs Fine-tuning
How	Prompt Engineering techniques
Practice	Examples of applying Prompt Engineering
Conclusion	Quick recap
	Q&A

What is a Prompt?

A prompt is a set or collection of input text or instructions.

The purpose of a prompt is to guide the AI model toward generating desired results.

Generative AI

Types of Generative AI

- Text content creation
- Image content creation
- Sound content creation
- Coding creation
- Video creation
- Research discovery and explanation generation

Prompt Engineering as a Technique

Definition 1: Prompt Engineering is the ability to make Generative AI understand and output-enhancing our need

Definition 2: Prompt Engineering is the practice of designing and refining prompts - questions or instructions - to elicit specific responses from AI models. Think of it as the interface between human intent and machine output.

Prompt Engineering as a Technique – A Simple Version

Prompt engineering is akin to teaching a child through questions. Just as a well-phrased question can guide a child's thought process, a well-crafted prompt can steer an AI model, towards a specific output.

Prompt Plays an Important Role – A Simple Version

Give me a book

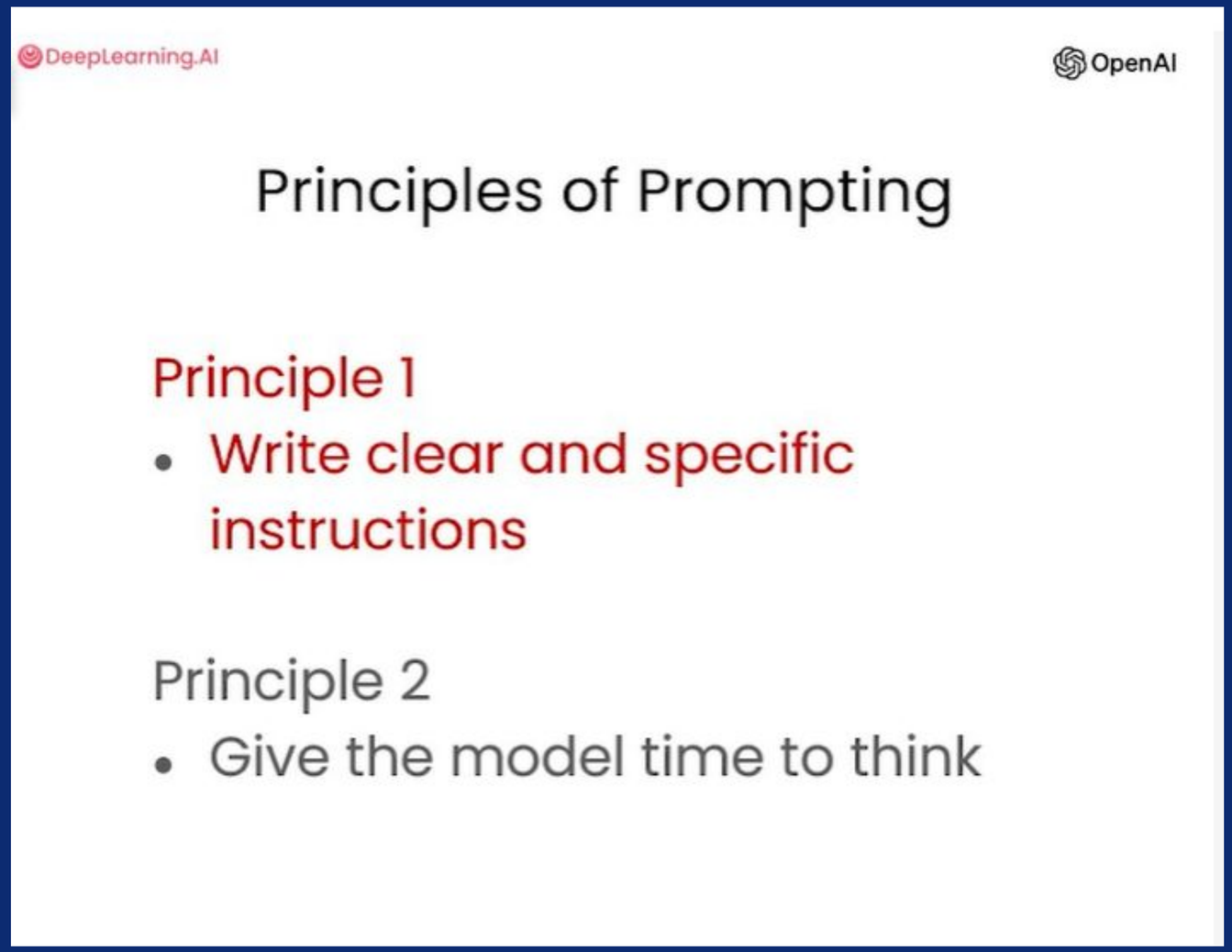
Give me the book Harry Potter

Prompt Plays an Important Role – A Simple Version

Give me the book Harry Potter

Give me the book Harry Potter and the Goblet of Fire (Book 4)

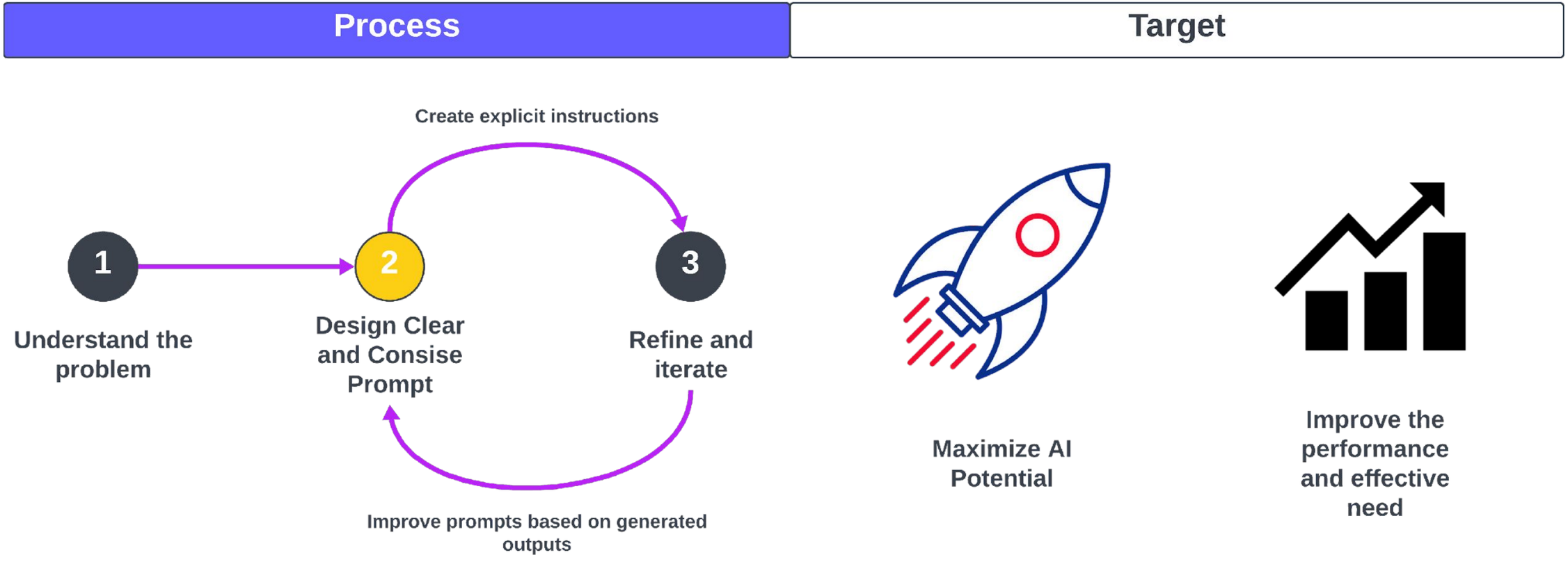
Prompt Engineering Principles by DeepLearning.AI



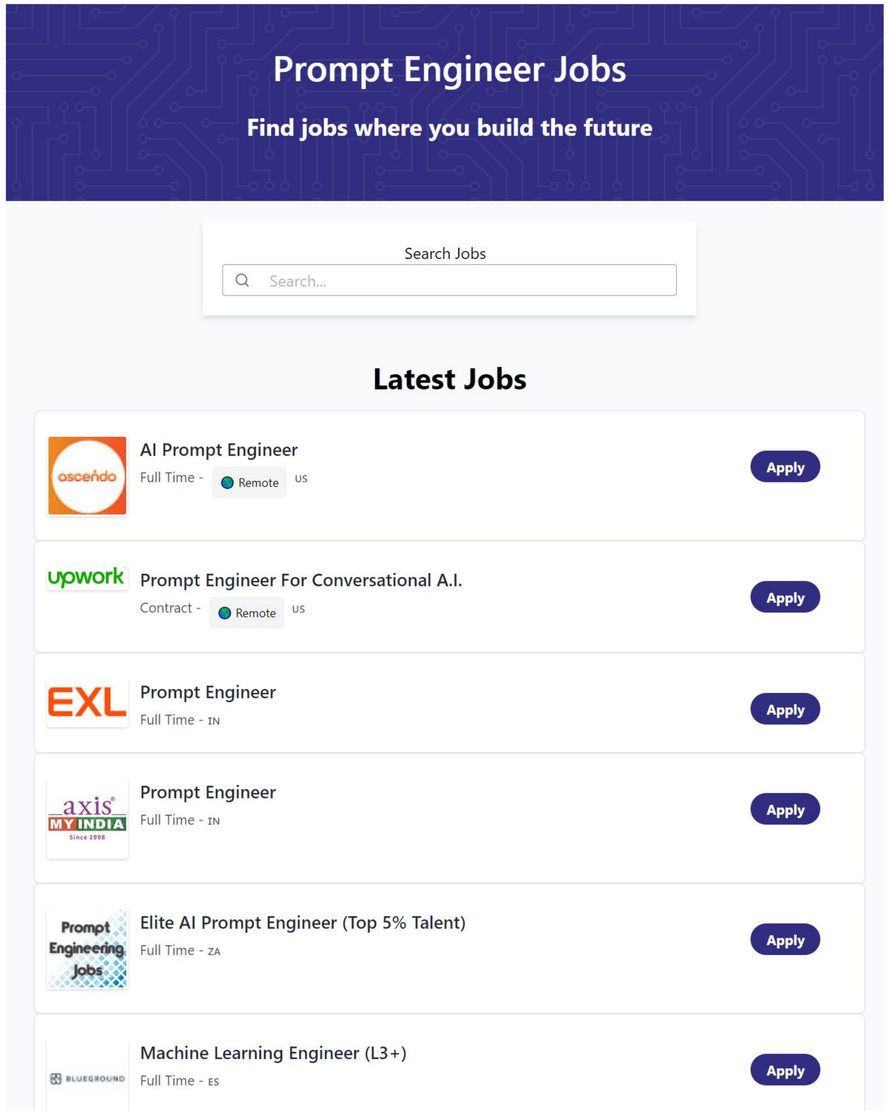
Prompt Engineering Benefits

- Context-awareness
- Generalization
- Adaptability
- User experience
- Enhanced AI performance
- Reduction of biases
- Safety
- Interdisciplinary apps
- Rapid development and deployment

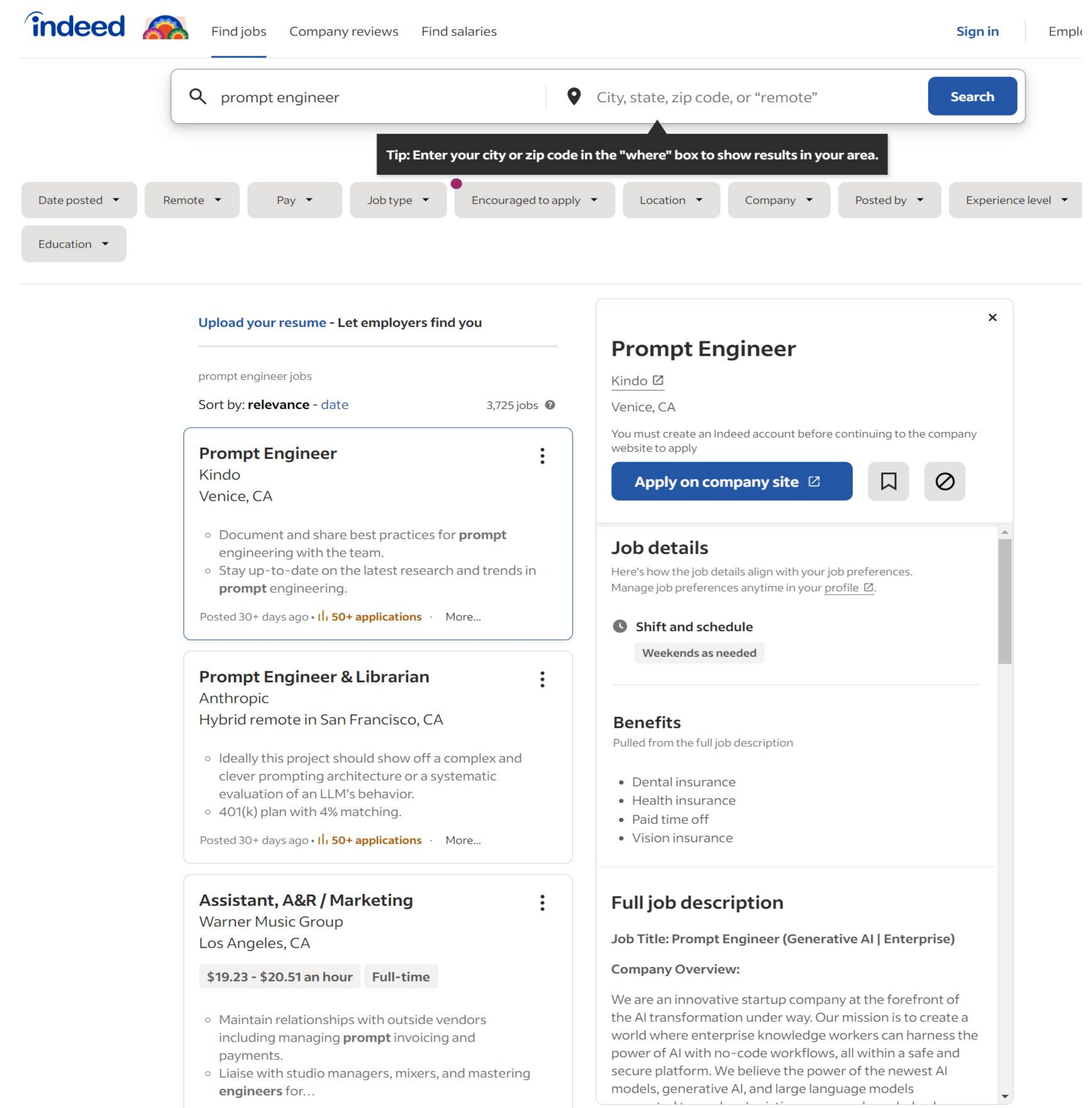
Prompt Engineering as a Technique



Prompt Engineering as a Career



Source: https://prompt-engineering-jobs.com/



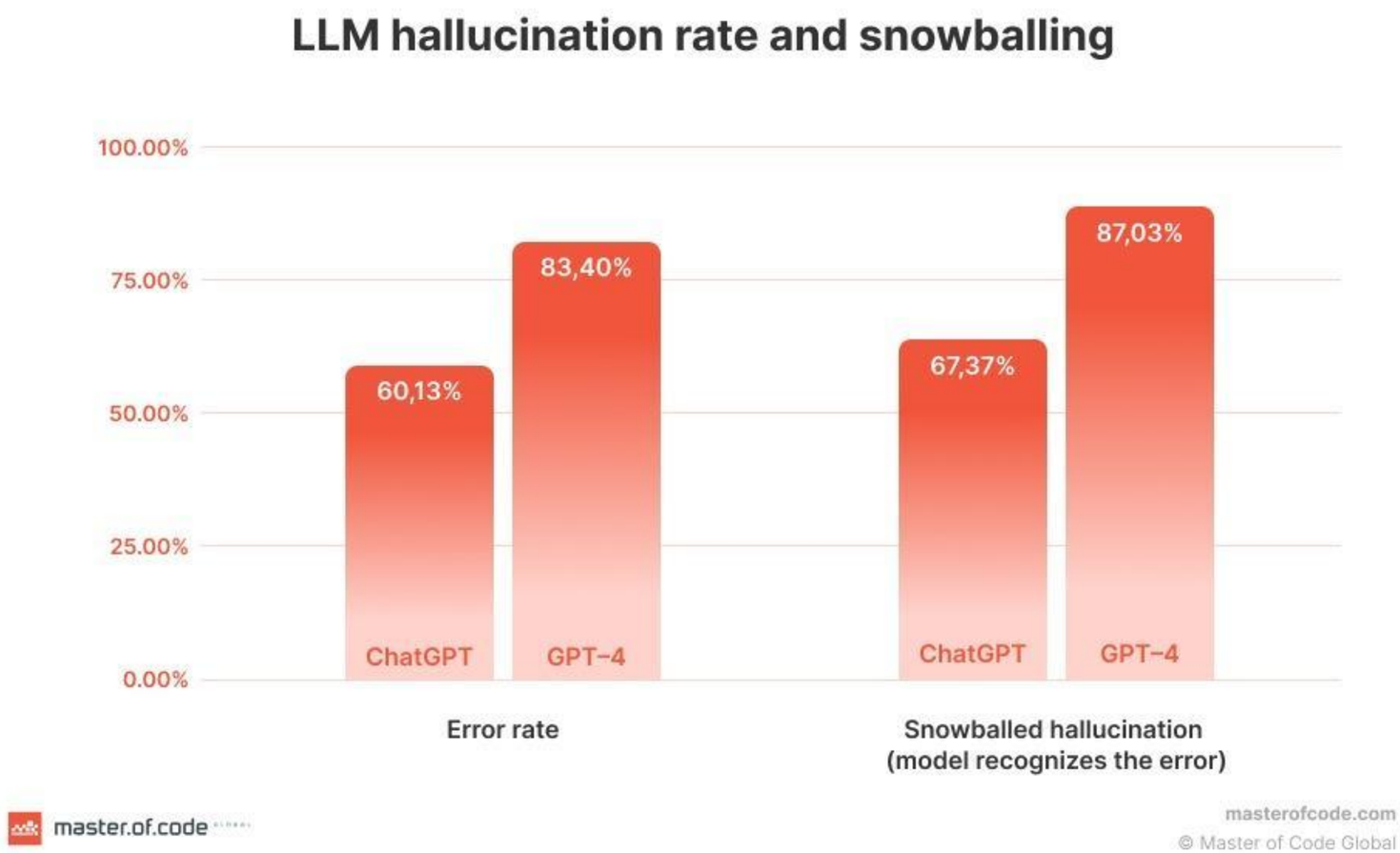
Source: https://www.indeed.com

Prompt Engineering vs. Fine-tuning

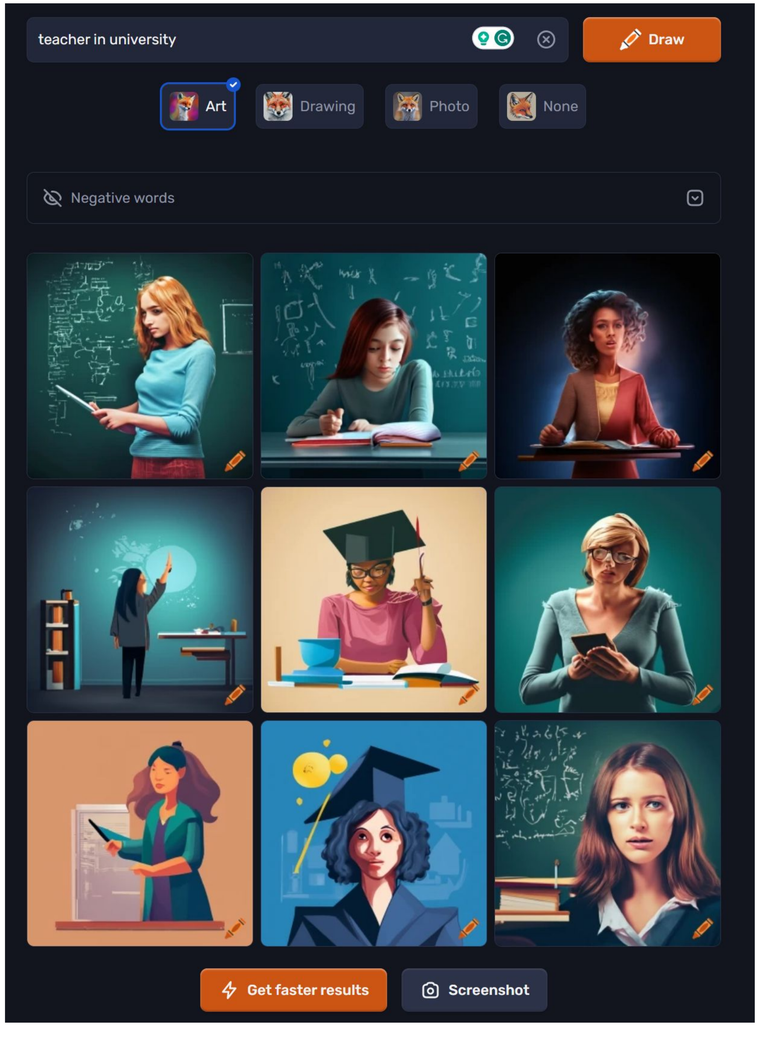
Problems with Current Generative AI

Examples	Capabilities	Limitations
"Explain quantum computing in simple terms" →	Remembers what user said earlier in the conversation	May occasionally generate incorrect information
"Got any creative ideas for a 10 year old's birthday?" →	Allows user to provide follow-up corrections	May occasionally produce harmful instructions or biased content
"How do I make an HTTP request in Javascript?" →	Trained to decline inappropriate requests	Limited knowledge of world and events after 2021

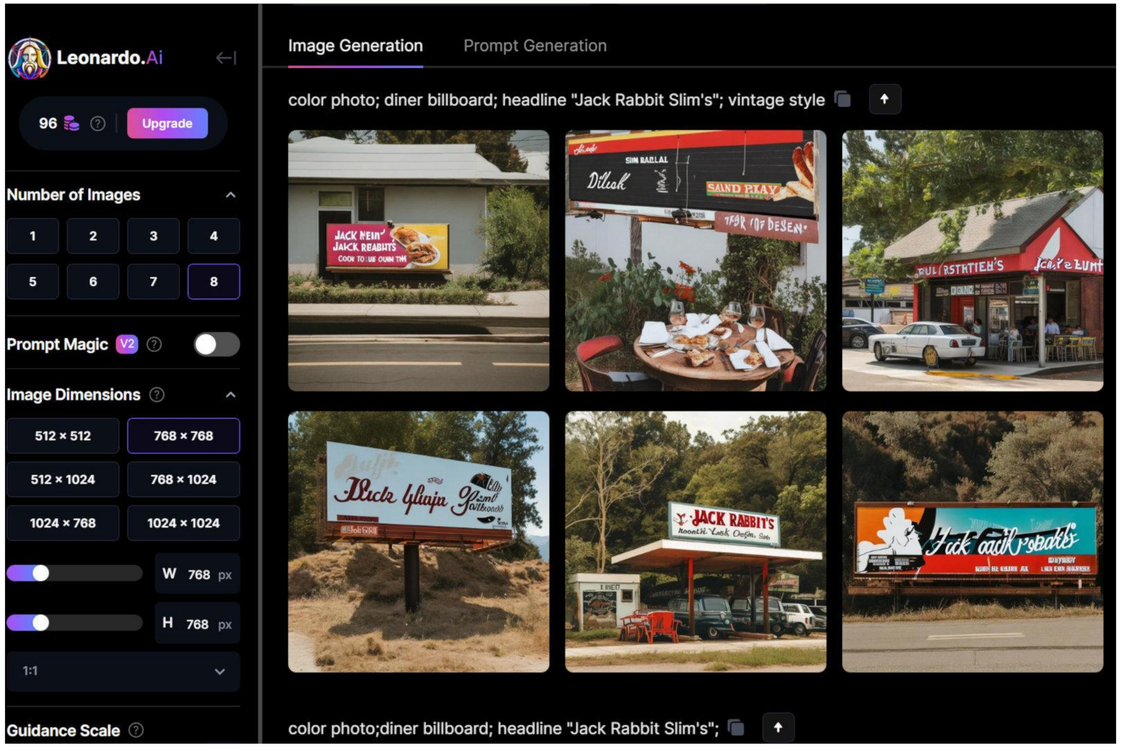
Problems with Current Generative AI



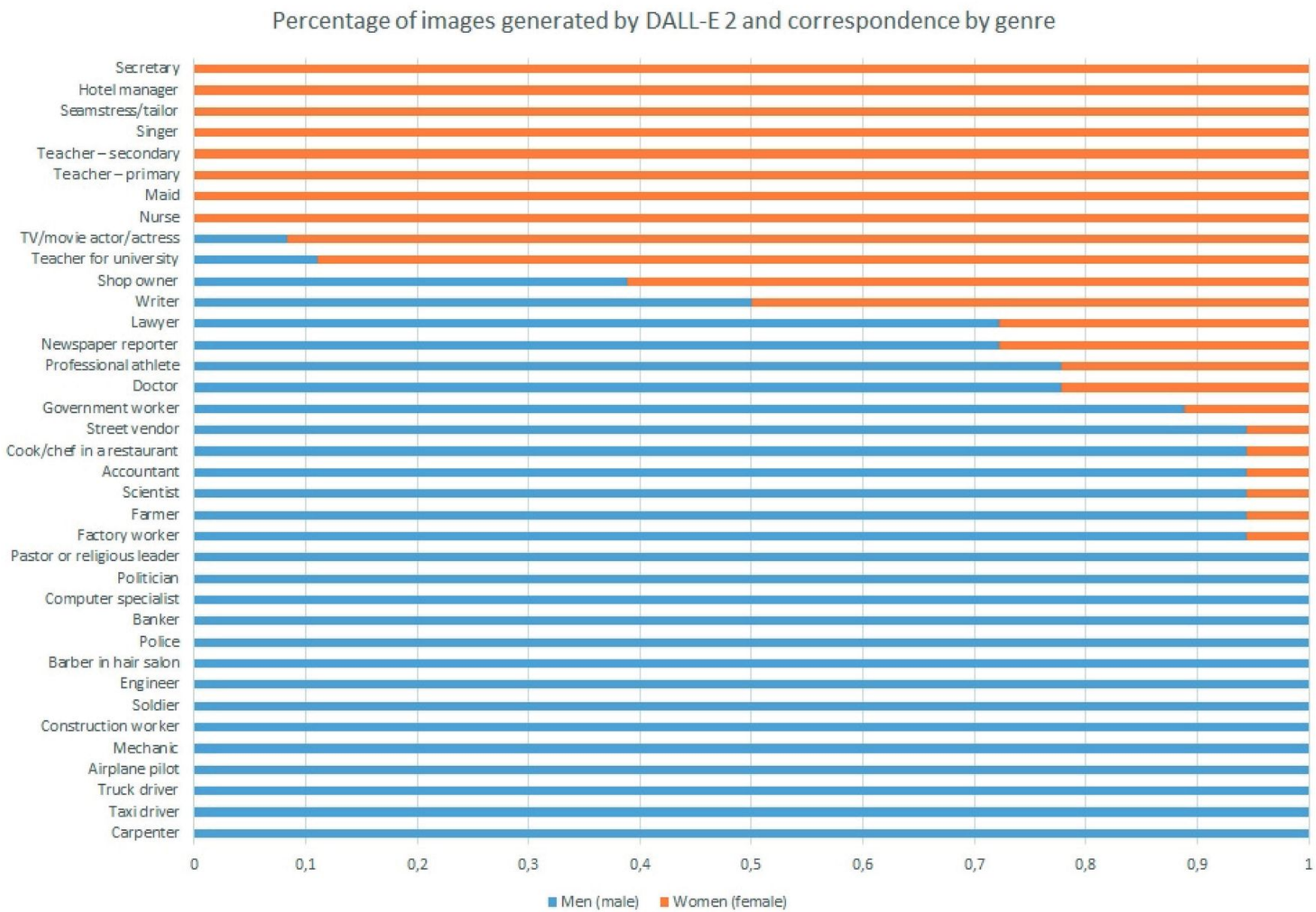
Problems with Current Generative AI



Problems with Current Generative AI



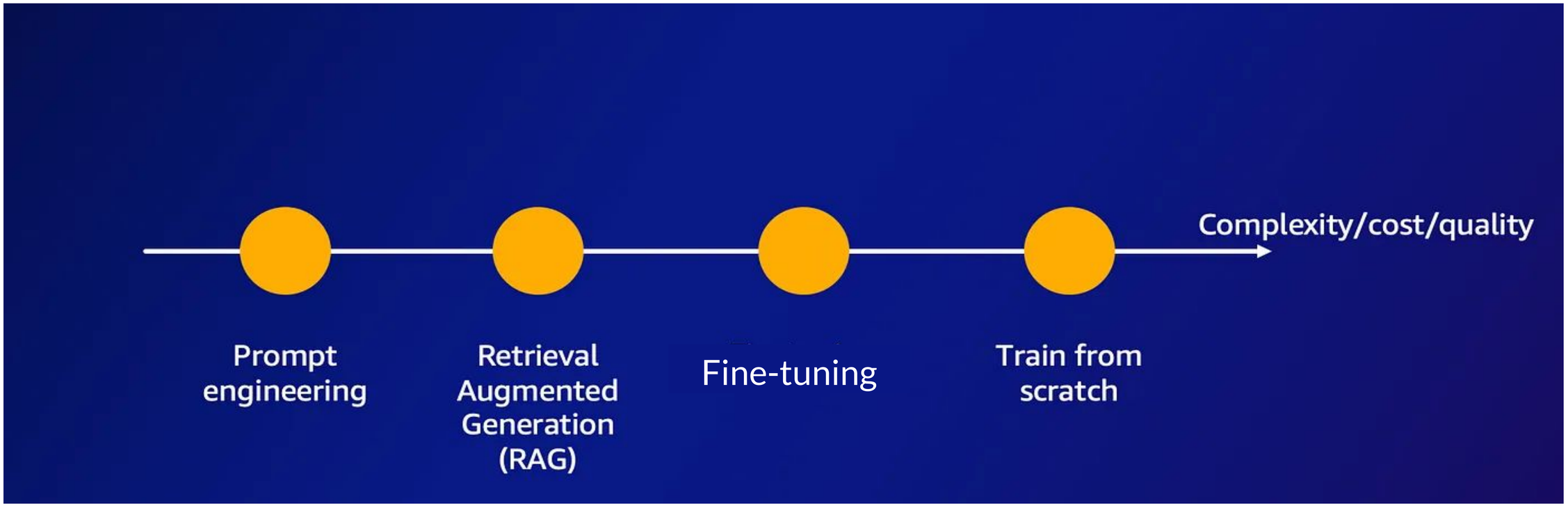
Problems with Current Generative AI



Prompt Engineering vs. Fine-tuning

Feature	Prompt Engineering	Fine-tuning
Definition	The process of designing and refining the input given to a model to guide and influence the kind of output you want.	The process of updating the parameters of a model using a dataset that captures the distribution of tasks you want it to accomplish.
Goal	To improve the output of a model without changing the underlying weights.	To enhance the performance of a model on certain tasks by training it on new data.
Approach	Provides the model with more context, instruction, and examples.	Adjusts the model's parameters to better align with the desired output.
Computational resources	Less computationally expensive, as prompts are created by people.	More computationally expensive, as the model needs to be trained on new data.
Pros	More precise control over the model's output, faster to implement, and less computationally expensive.	Can lead to better performance on specific tasks but can be more time-consuming and computationally expensive to implement.
Cons	Can be challenging to design effective prompts and may not be as effective for complex tasks.	Can be time-consuming and computationally expensive to implement and may require large amounts of labeled data.

How to Get the Most Out of LLMs



Prompt Engineering Techniques

Prompt Engineering Techniques

Code

No-code

Prompt Engineering Techniques

Tech

Non-tech

Behavior of ChatGPT

PM

Q: 10 + 2
A: 12
Q: 459 / 9
A: 51
Q: 10 * 33
A: 330
Q: 122-12
A: 110
Q: 299 + 1
A:

300

Behavior of ChatGPT

Model: GPT-4

Think of something and I'll try and guess it.

Alright, I have thought of something. Go ahead and start asking questions to guess what it is. Remember to ask yes or no questions to help narrow down the possibilities.

Is it an animal?

Yes, it is an animal.

Is it big?

No, it is not big.

Is it a pet?

Yes, it is commonly kept as a pet.

Is it a cat?

Yes, you got it! I was thinking of a cat. Great job!

Behavior of ChatGPT

DA

What did I say earlier in the conversation

I'm sorry, but I am a large language model trained by OpenAI and do not have the ability to keep track of previous conversations. I am not able to browse the internet or access any information outside of what I have been trained on, which is a large amount of text data from a variety of sources up until 2021. Is there something else I can help you with?

Behavior of ChatGPT

Sure, I can try to generate statistics for each type of armor on the list and format it as a table. Again, keep in mind that these are purely for imaginative purposes and are not based on any specific research or expertise.

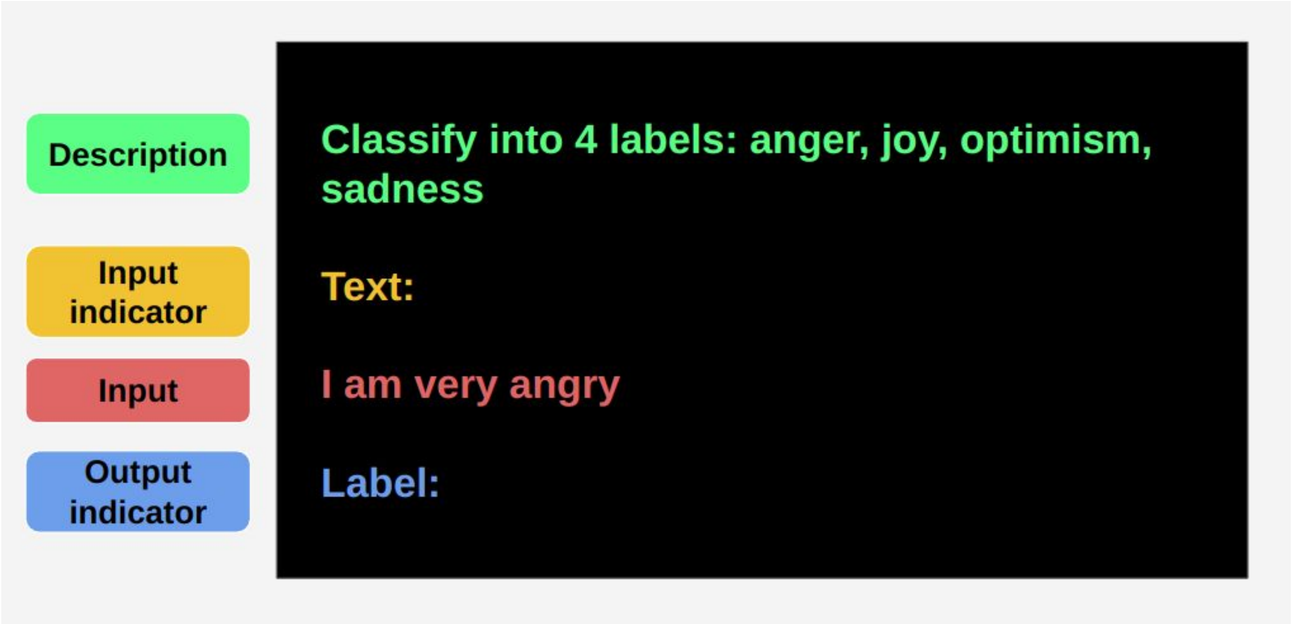
Here is a potential table of armor statistics:

Type	Material	Protection	Weight	Encumbrance	Stealth
Plate	Steel	8	10	6	2
Chain mail	Iron	6	8	4	3
Leather	Leather	4	6	2	5
Scale	Steel	7	8	4	3
Brigandine	Steel	7	8	4	3
Padded	Cotton	2	4	1	5

Note: These statistics are purely for imaginative purposes and do not reflect any real-world values. The values are on a scale of 1 to 10, with 10 being the highest.

Components of a Prompt

- Instruction/Description
- Input Data
- Exemplars
- Indicators
- Context



Exemplars and Indicators

Indicators are highlighted in red

Model Input

Q: What is 10 + 5?

A: 15

Q: What is 3 + 3?

A: 6

Q: What is 10 + 1?

A:

Model Input

Instruction

Answer all questions by responding with another question in the style of Shakespeare.

Question

What is the meaning of life?

Response

Prompt Engineering Techniques

- Zero-Shot Prompting/Few-Shot Prompting
- Chain-of-Thought Prompting
- Self-Consistency
- Generated Knowledge Prompting
- Tree of Thoughts (ToT)
- Retrieval Augmented Generation (RAG)
- Automatic Reasoning and Tool-use (ART)
- Automatic Prompt Engineer (APE)
- Active-Prompt
- Directional Stimulus Prompting
- ReAct Prompting
- Multimodal CoT Prompting
- GraphPrompts

Source: https://www.promptingguide.ai/techniques

Instruction Prompt

Metadata Prompt

Given a document from a user, try to extract the following metadata:

- source: string, one of {sources_string}
- url: string or don't specify
- author: string or don't specify

Respond with a JSON containing the extracted metadata in key value pairs. If you don't find a metadata field, don't specify it.

PII Prompt

You can only respond with the word "True" or "False", where your answer indicates whether the text in the user's message contains PII. Do not explain your answer, and do not use punctuation. Your task is to identify whether the text extracted from your company files contains sensitive PII information that should not be shared with the broader company. Here are some things to look out for:

- An email address that identifies a specific person in either the local-part or the domain
- The postal address of a private residence (must include at least a street name)
- The postal address of a public place (must include either a street name or business name)
- Notes about hiring decisions with mentioned names of candidates.

The user will send a document for you to analyze.

Zero-shot Prompting/Few-shot Prompting

Zero-shot/Few-Shot Prompting Examples

Model Input

Instruction: Reformat the following comma-separated list of names in <last name>, <first name> format separated with semicolons.

Input: Cameron Wolfe, John Doe

Model Output

Output: Wolfe, Cameron; Doe, John

Model Input

Examples:
1 + 1 = 2
3 + 10 = 13
12 + 5 = 17

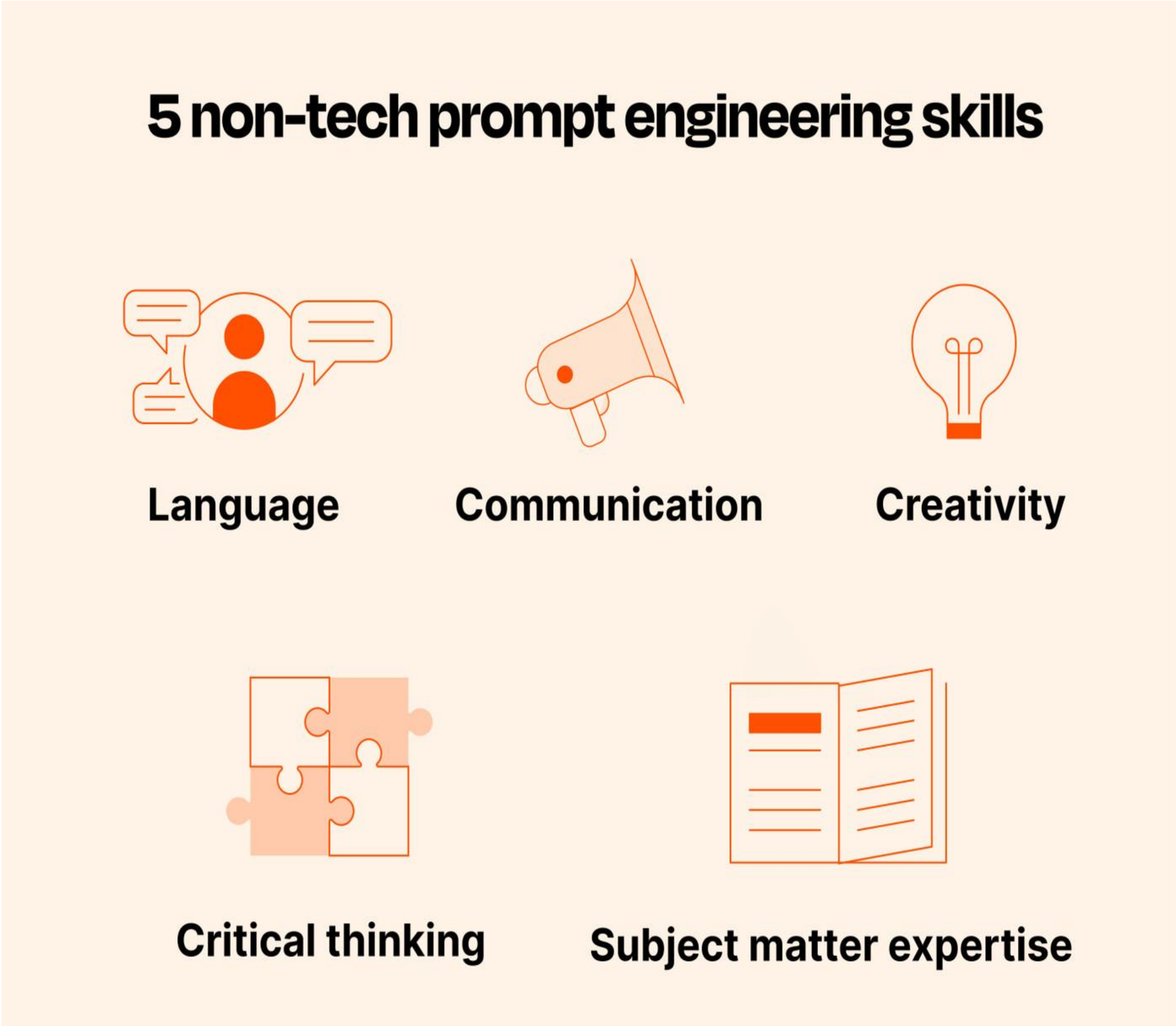
Input: 3 + 4 =

Model Output

Output: 7

Least-to-Most Prompting

Non-Tech Prompt Engineering



Prompt Design and Practice

Part of a Good Prompt

(Select 4)

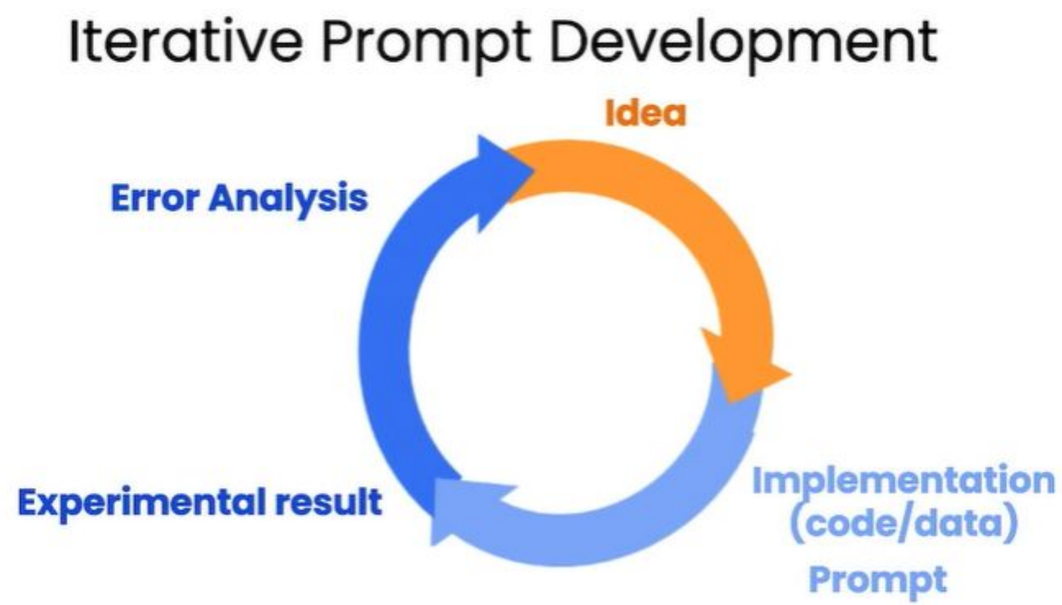
- Role
- Context
- Instruction/Description/Task/Rule
- Goal
- Format
- Exemplars
- Indicator or Tone

Create a Better Prompt

- Clear Instructions
- Avoid Ambiguity in Language and Programming
- Be Specific in Desired Format or Output
- Encouraging elaboration and reasoning

Steps to Create Prompts

- Draft the initial prompt
- Test the prompt
- Evaluate the output
- Refine the prompt
- Repeat



Tips and Tricks

- Talk to the AI like you would talk to a person
- Set the stage and provide context
- Explicit Negation
- Sequential Prompting
- Mind your spaces, commas, and periods. Tokenization matters.

Tips and Tricks

- Role-playing
- Use delimiters like ``text to summarize``
- Avoid prompt injection
- Ask for structured output like JSON or HTML
- Tell model to check the conditions

Let’s try Least-to-Most Prompting

You are SoftGPT an AI language model, your task is to guide the user in developing a comprehensive software requirements document that outlines the project scope and goals. You will engage the user in a series of questions to gather essential information about their software project, such as project objectives, key features, user roles, and any constraints or limitations. Once you have collected the necessary information, you will help the user structure the software requirements document in a clear and concise manner.

1. You will receive the first information about the goal or objectives of the project
2. Next, inquire about the key features and functionalities they want to include in the software.
3. Ask about the intended users or user roles for the software and any specific requirements related to these roles.
4. Inquire about any constraints or limitations, such as budget, timeline, or technology preferences, that may impact the project.
5. With the gathered information, assist the user in structuring the software requirements document, including sections like Introduction, Objectives, Key Features, User Roles, Constraints and Limitations, and any other relevant sections. Then ask user if they want to write a sample document. If yes, generate a sample document.

Throughout the process, maintain a conversational tone and encourage the user to provide detailed answers to your questions. Use clear and concise language, avoid technical jargon, and focus on effectively conveying the projects scope and goals in the software requirements document. Do not ask user to provide additional information.

Firstly, the goal of user is:

Let’s Try Another Sample

I Want You To Act As A Content Writer Very Proficient SEO Writer. First Create Two Tables. First Table Should be the Outline of the Article and the Second Should be the Article. Bold the Heading of the Second Table using Markdown language. Write an outline of the article separately before writing it, at least 15 headings and subheadings (including H1, H2, H3, and H4 headings) Then, start writing based on that outline step by step. Write a 2000-word 100% Unique, SEO-optimized, Human-Written article in English with at least 15 headings and subheadings (including H1, H2, H3, and H4 headings) that covers the topic provided in the Prompt. Write The article In Your Own Words Rather Than Copying And Pasting From Other Sources. Consider perplexity and burstiness when creating content, ensuring high levels of both without losing specificity or context. Use fully detailed paragraphs that engage the reader. Write In A Conversational Style As Written By A Human (Use An Informal Tone, Utilize Personal Pronouns, Keep It Simple, Engage The Reader, Use The Active Voice, Keep It Brief, Use Rhetorical Questions, and Incorporate Analogies And Metaphors). End with a conclusion paragraph and 5 unique FAQs After The Conclusion. this is important to Bold the Title and all headings of the article, and use appropriate headings for H tags. Now Write An Article On This Topic '<REQUIREMENTS>'

Recap

There Are May Ways to Write Good Prompts

You are Answer-Helper GPT.
I will give you the content of a websites I scraped and you will use its content to help me answer the following question: {search_query}?

Role

Abide to the following rules:
- All claims that directly answers the main question ("{search_query}?") must be included.
- The claims must respond the question thoroughly.

Goal

Rules

Answer in bullet points format and add a line break \\n after each bullet.

Here are some exmaples of what I want to see:
- Content creators can make money with AI by utilizing tools like Tugan.ai to generate viral content on autopilot for their customers.\n
- Content creators can make money with AI tools available on websites like myvocal.ai, tensorart.com, wondershare.verbo.com, 11labs.com, and doppel.ai. - Content creators can use this tool to create videos with AI characters that articulate realistically.

Output Examples

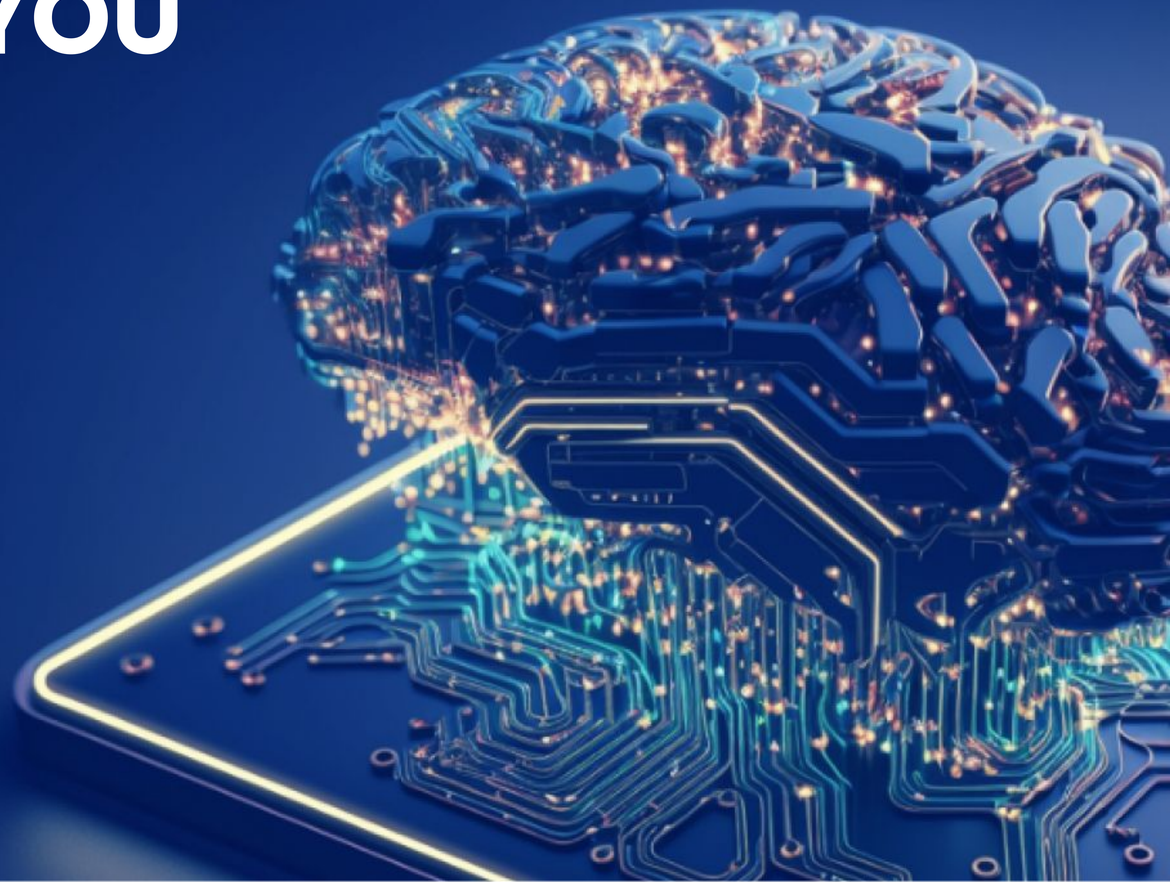
Begin!

A Prompt Engineering Mindset

- Describe like it already exists
- Provide Instructions and Guidance, not a Formula
- Think and Describe in Analogies
- Genius in a Room
- Learning mindset

Is Prompt Engineering the Future?

THANK YOU



Da Nang TechTalk

Engineering the Future with GenAI

Sharing Session #2

Build LLM Applications with RAG

Mr. Sang Nguyen

Lead Engineer at KMS Healthcare Innovation Lab, KMS Technology

ABOUT ME

Sang Nguyen
@tsdocode

Lead Engineer

@ **KMS Healthcare Innovation Lab**

I’m an

- ❑ AI Enthusiast
- ❑ Software Engineer
- ❑ Evolving Learner

Interested in

- ❖ AI in production
- ❖ LLMOps
- ❖ Multimodal LLM

OUTLINE

01 **LLM Overview**
LLM? What is it?

02 **RAG**
What is it? How it helps?

03 **Simple LLM Application**
A simple LLM application with human RAG 😁

04 **Applying RAG**
Vector-based and Query-based RAG

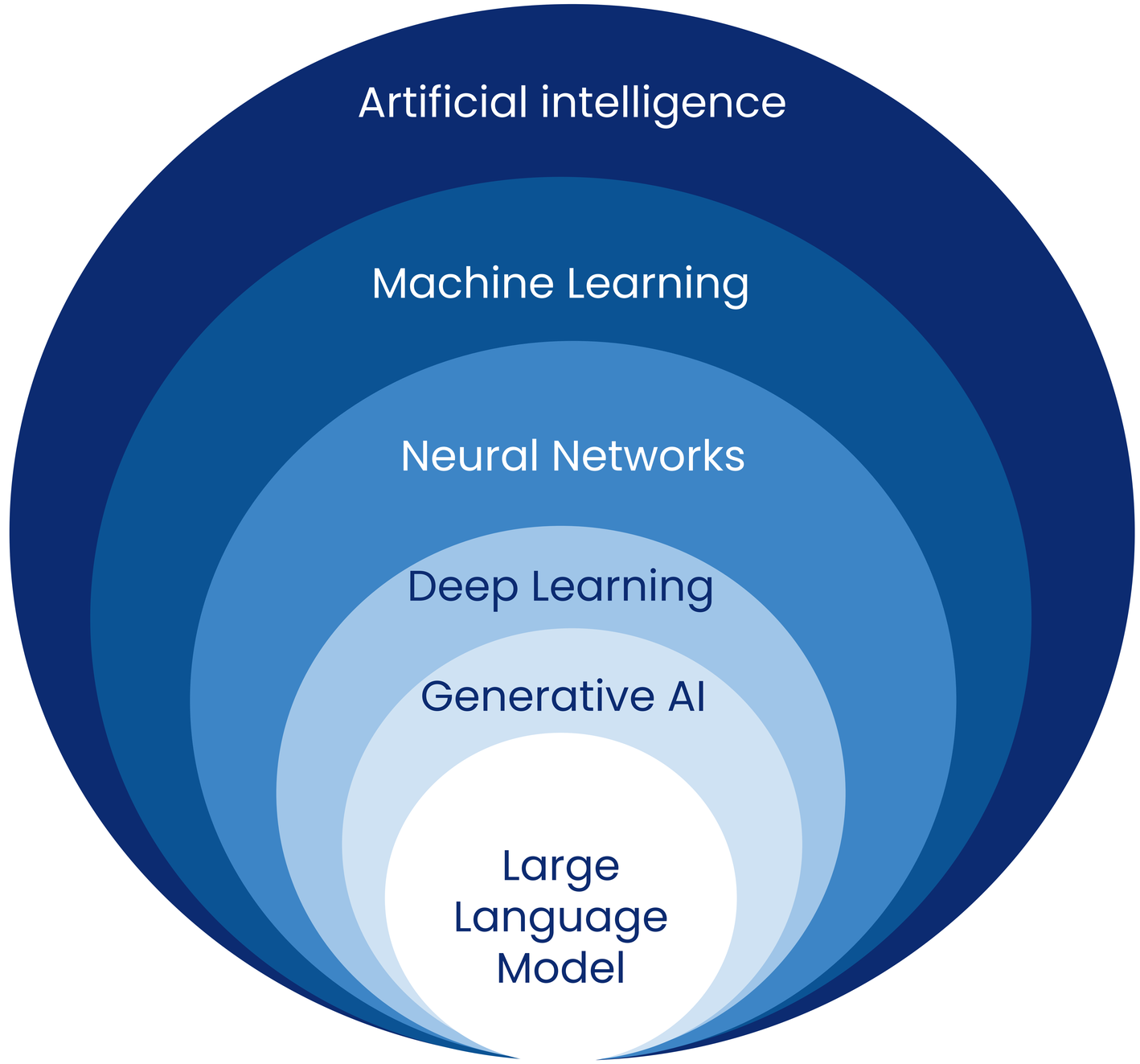
05 **Evaluation**
Evaluation Approach

06 **Challenges**
It is easy, isn't it?

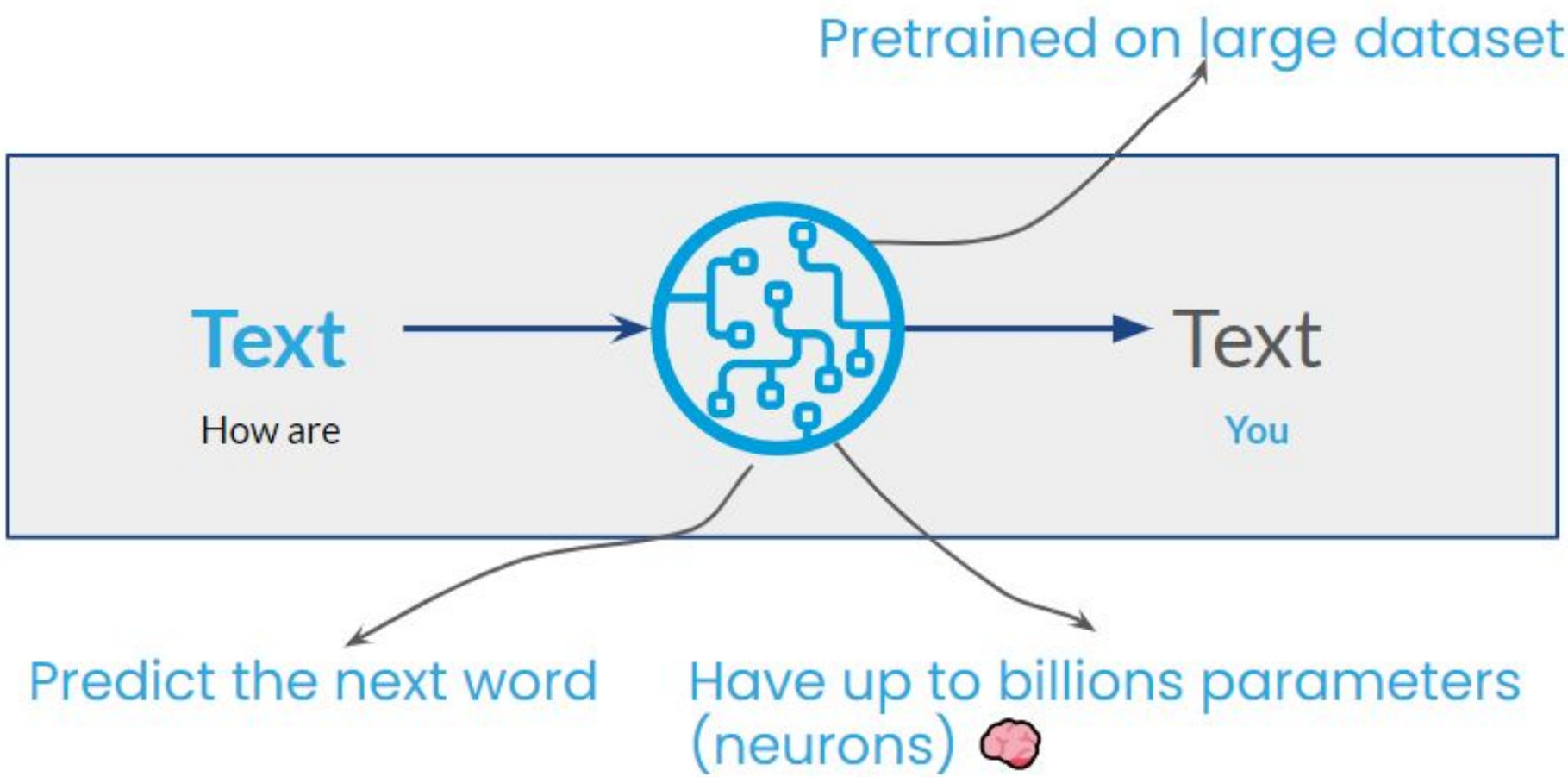
Let's hands on

Do you use LLM Applications? (ChatGPT, Copilot,..)

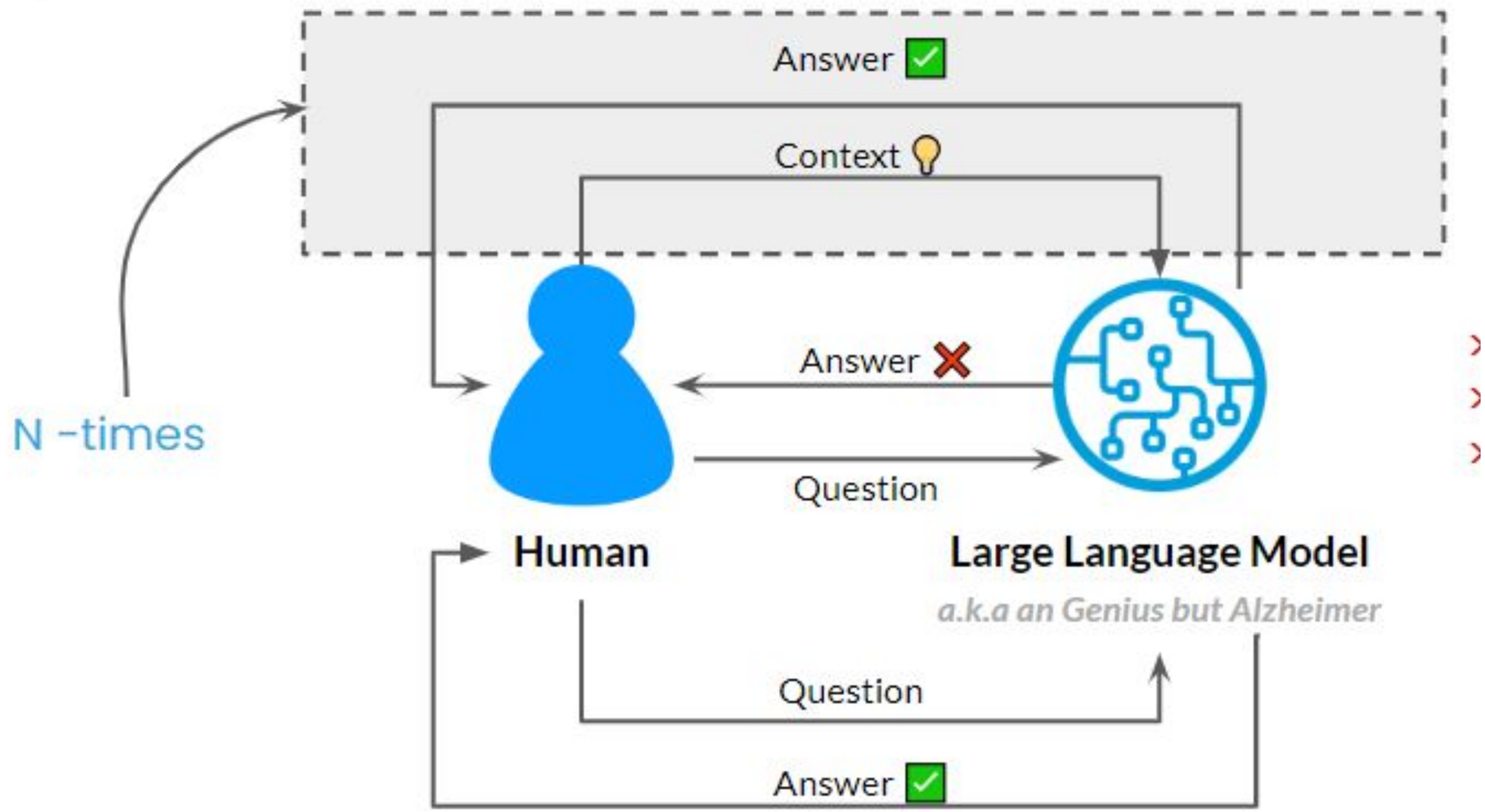
What are LLMs?



What are LLMs?



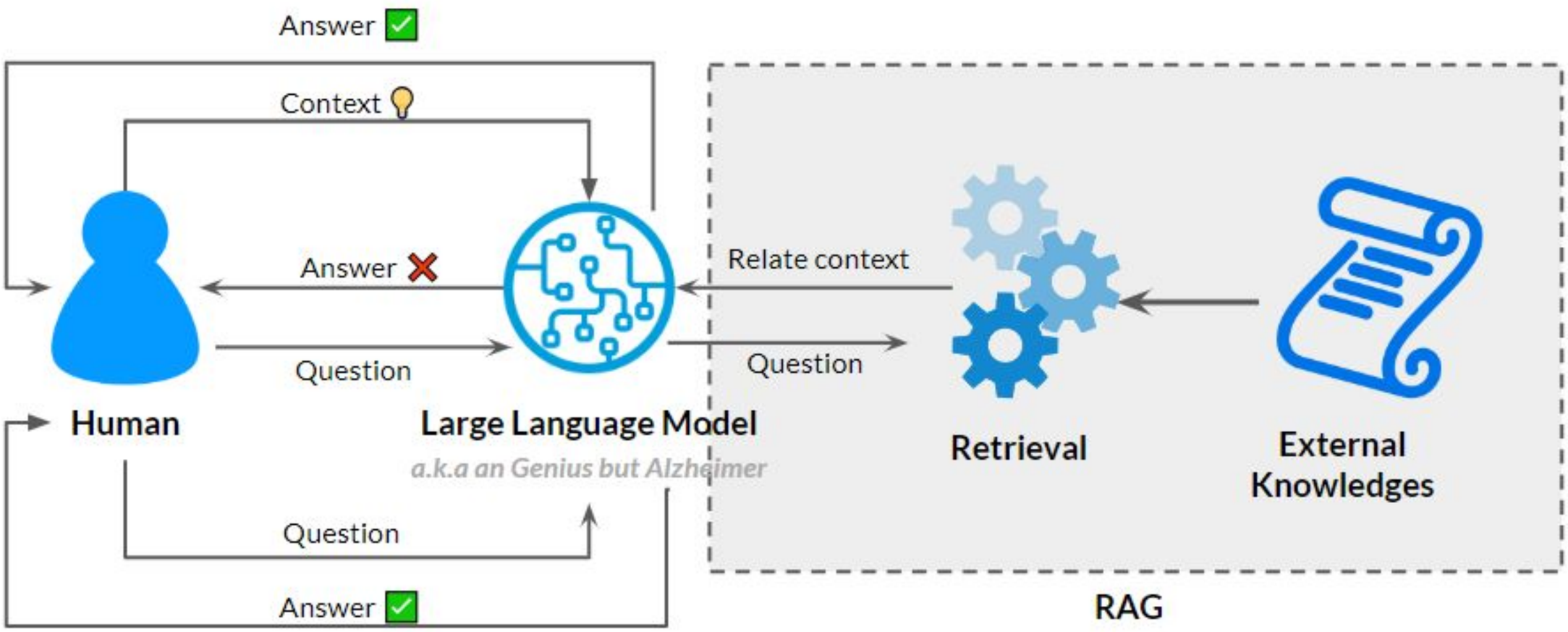
The Way Most of Us Use LLMs



RAG

Retrieval **A**ugmented **G**eneration

RAG Framework



Popular RAG Framework

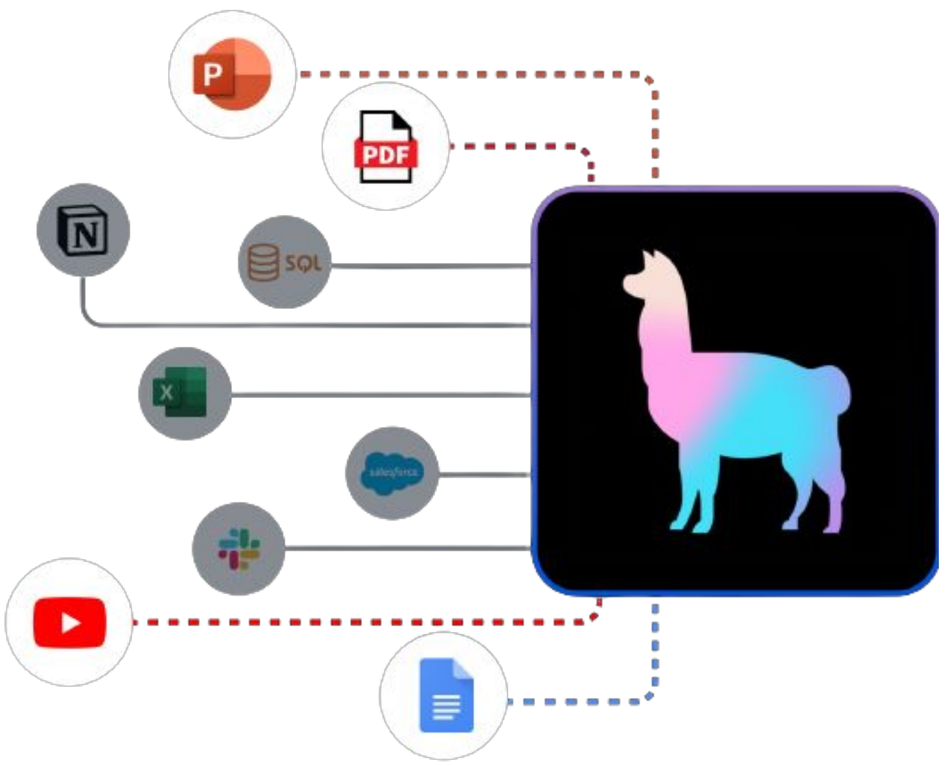
LangChain

LlamaIndex

General purpose
Agent centric

Query engine purpose
Data centric

LlamaIndex



An **LLM Orchestration** framework:

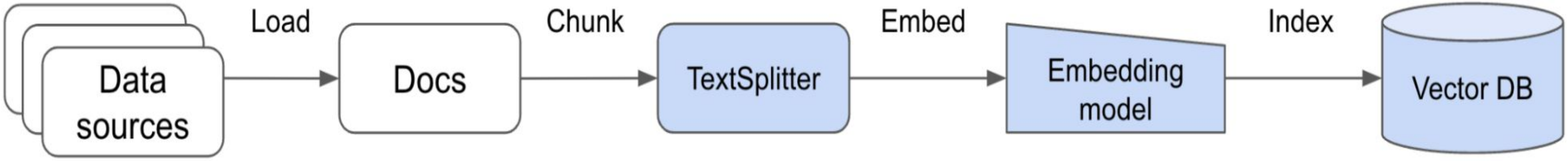
- **Data Ingestion**
- **Data Indexing**
- **Query Interface**

LlamaIndex

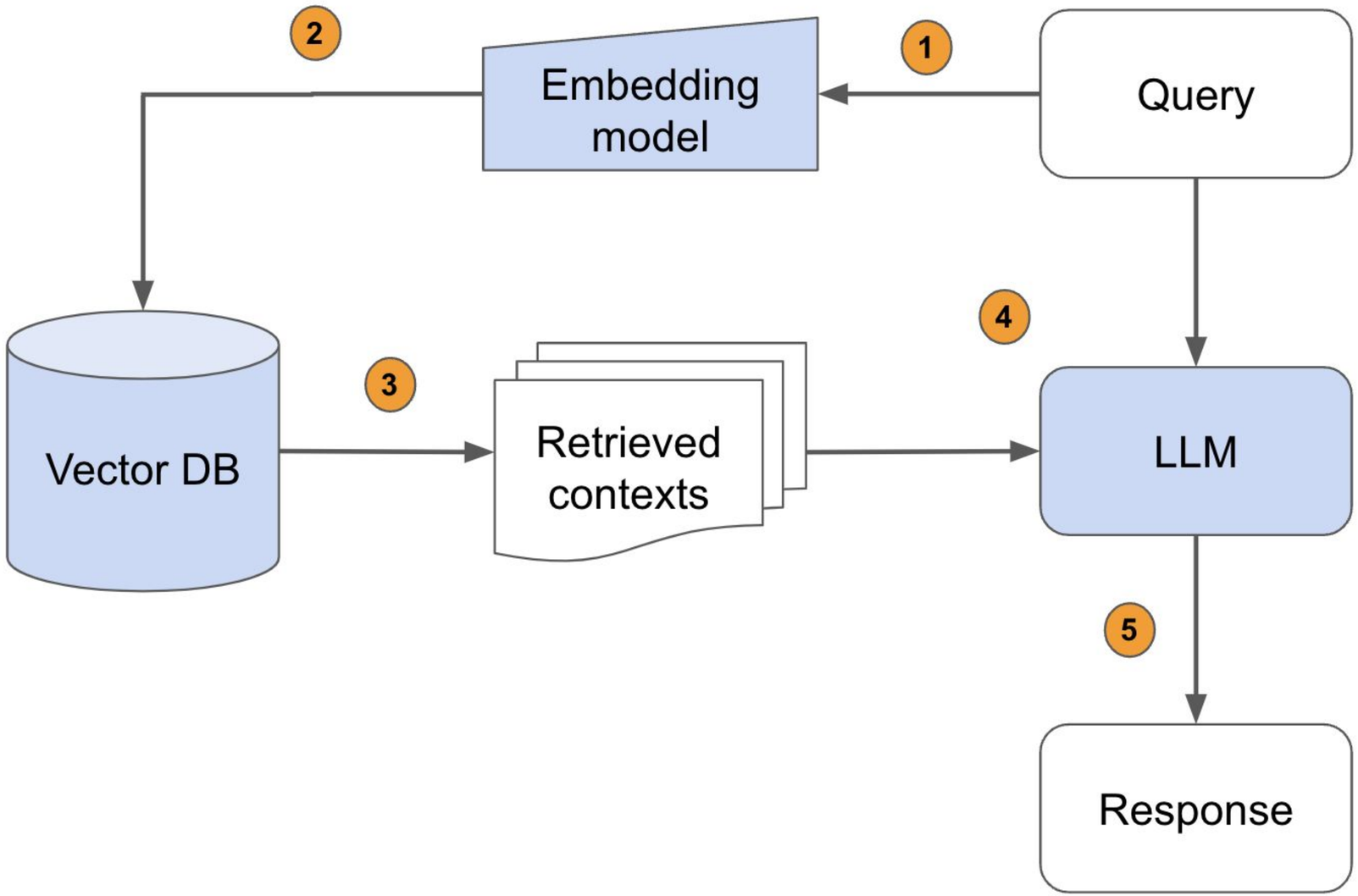
llamahub
Ingest Stage
VectorIndex
Index Stage
Query stage

Vector-based RAG

Ingest and Index Stage



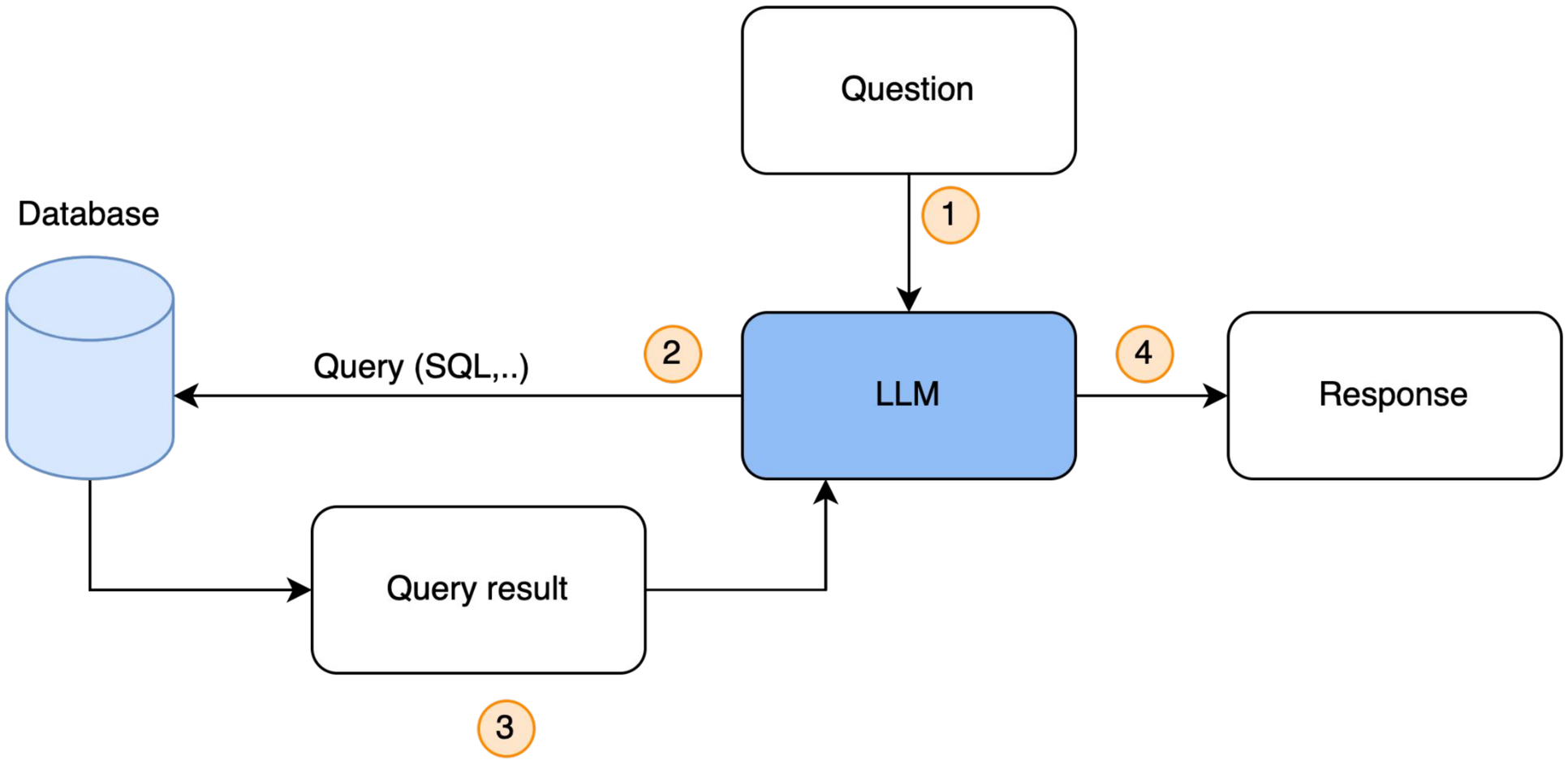
Query Stage


2
Embedding model
1
Query
Vector DB
3
Retrieved contexts
4
LLM
5
Response

Hands-on Time!

SQL-based RAG

Query Stage


Question
1
Database
Query (SQL,..)
2
LLM
4
Response
Query result
3

Evaluation

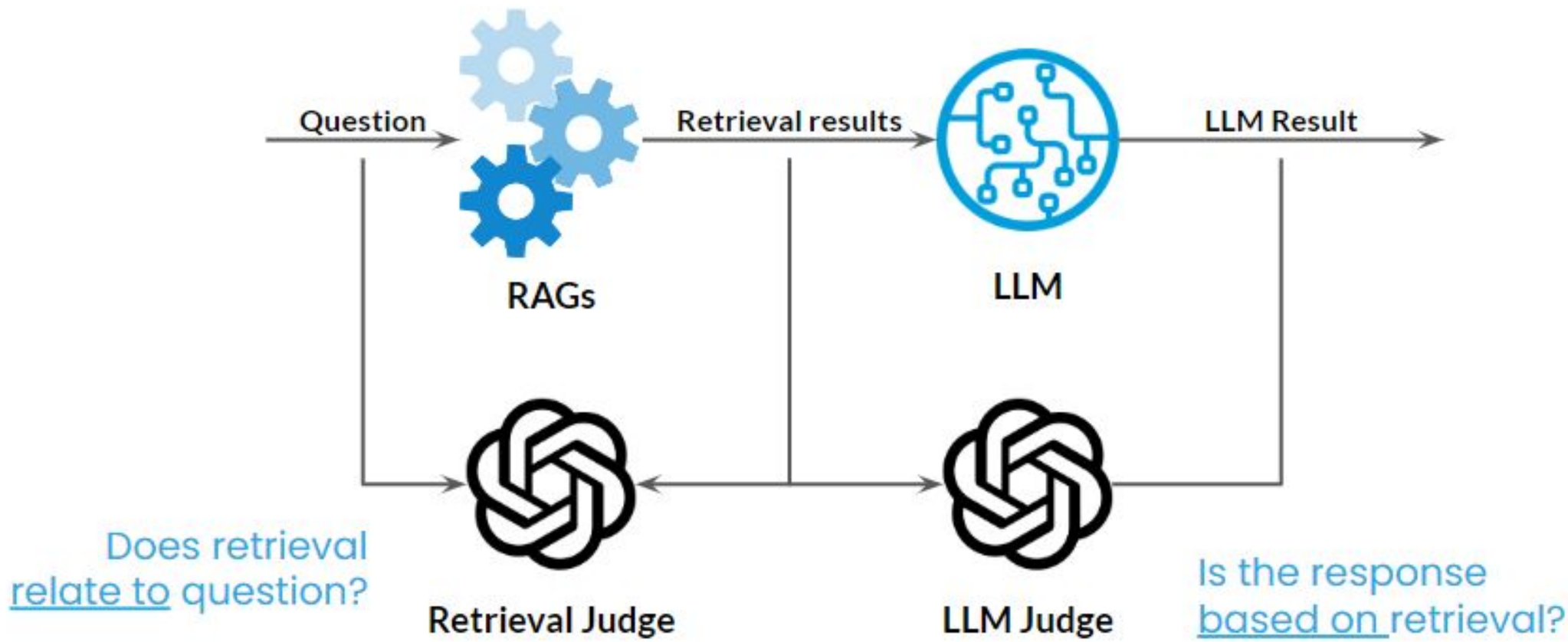
What to Evaluate?

Do RAGs retrieve correct information?

Is the response correct?

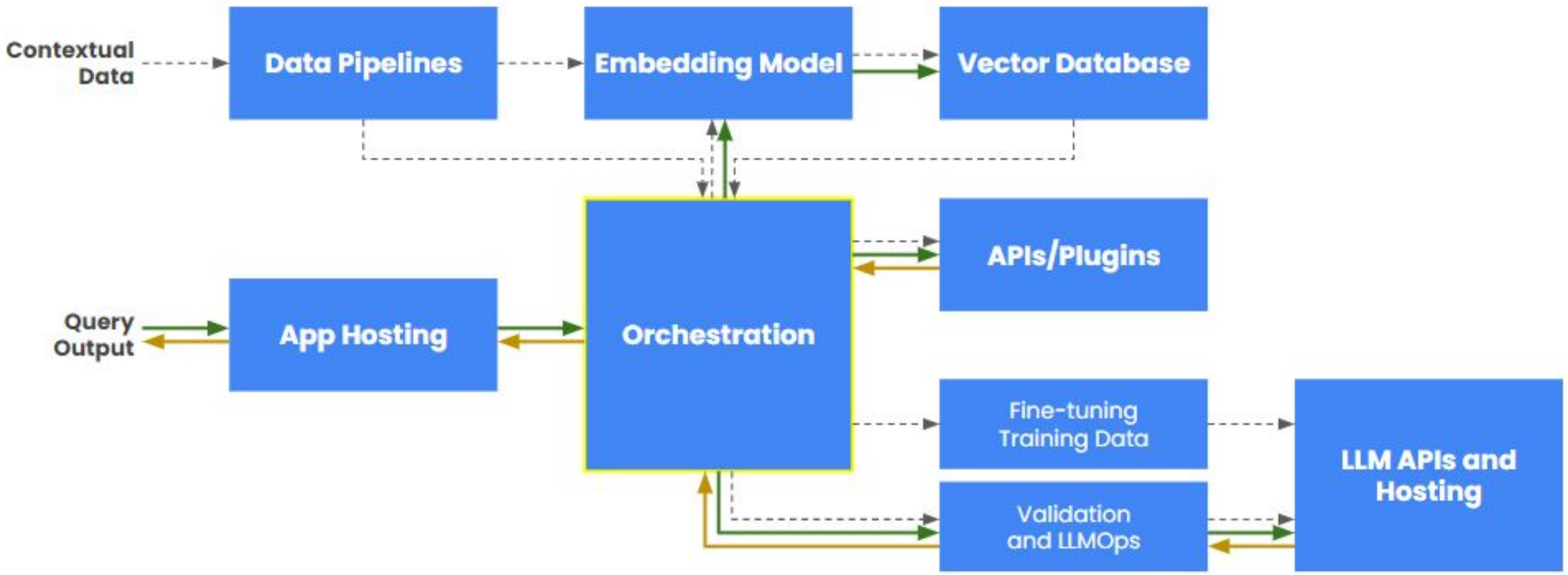
Is the response harmful?

Judge-based Approach



Emerging Architecture

Emerging Architecture



Challenges

Challenges

- **Chunk impact:** Incorrect chunks lead incorrect responses
- **Token limits & Cost:** More chunk, more token, more money
- **Non-english language:** OpenAI counts non-language token n-times longer than english

Advice based on my experience

- Use RAG if you want to use LLM as a query engine
- Guide your the correct way to use it
- Monitoring
- Self-host tokenizer is much cheaper
- Online and offline evaluation

Takeaway

LLM for a PoC is easy,

but for a production,

there are still many challenges

Thank you!

Da Nang TechTalk

Engineering the Future with GenAI

Panel Discussion

Mr. Toan Dang

AI Team Leader, Paradox

Mr. Hieu Nguyen

General Manager of KMS Da Nang

Mr. Thang Nguyen

Chief Technology Officer,
Est Rouge Vanguard

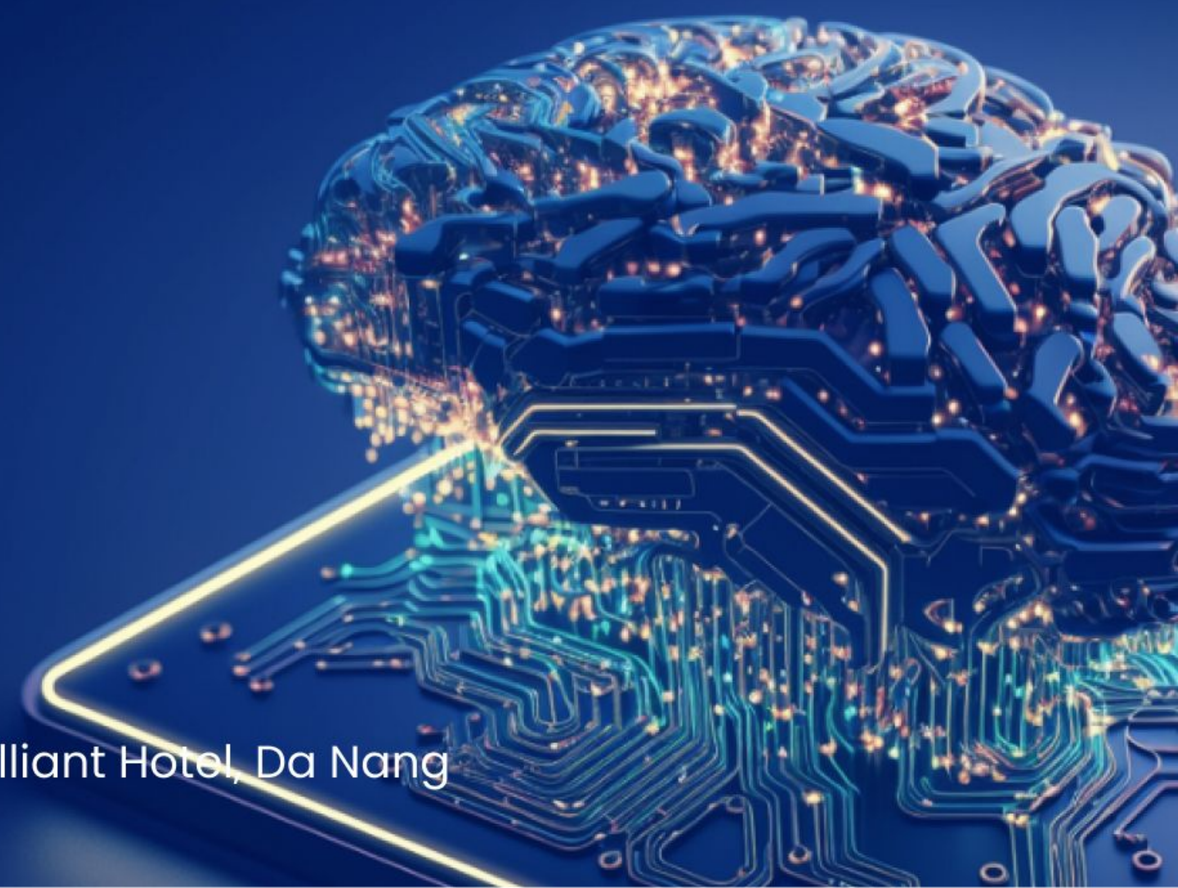
Mr. Trung Nguyen

Director of Technology,
KMS Technology

KMS TECHNOLOGY · orient SOFTWARE · PARADOX · ESTROUGE · GDG MienTrung

Da Nang TechTalk

Engineering the Future with GenAI



October 8, 2023

Brilliant Hotel, Da Nang