

IS883: Deploying Generative AI

Mohannad Elhamod

Recap

We have used...



- To query OpenAI's pre-trained language model using their API
- No model is stored locally.



Hugging Face

- To download and use a pre-trained model locally.
- To download and fine-tune a model with your own data.

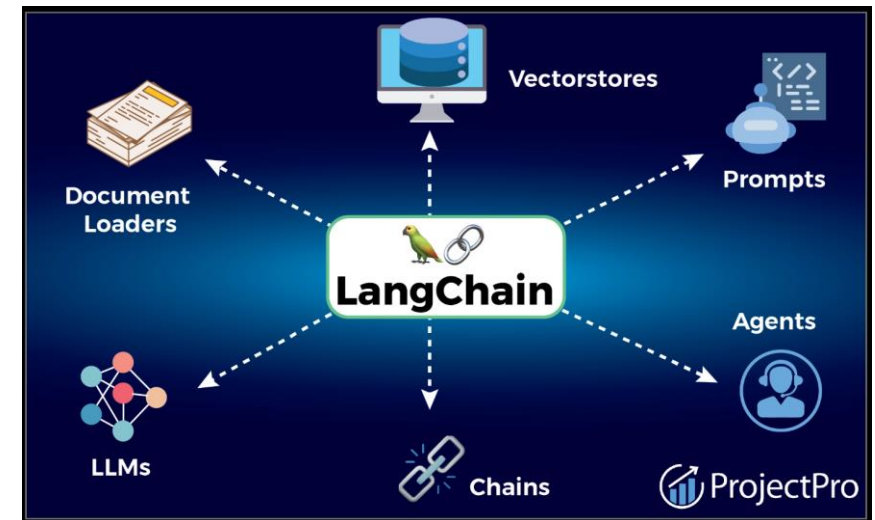


Streamlit

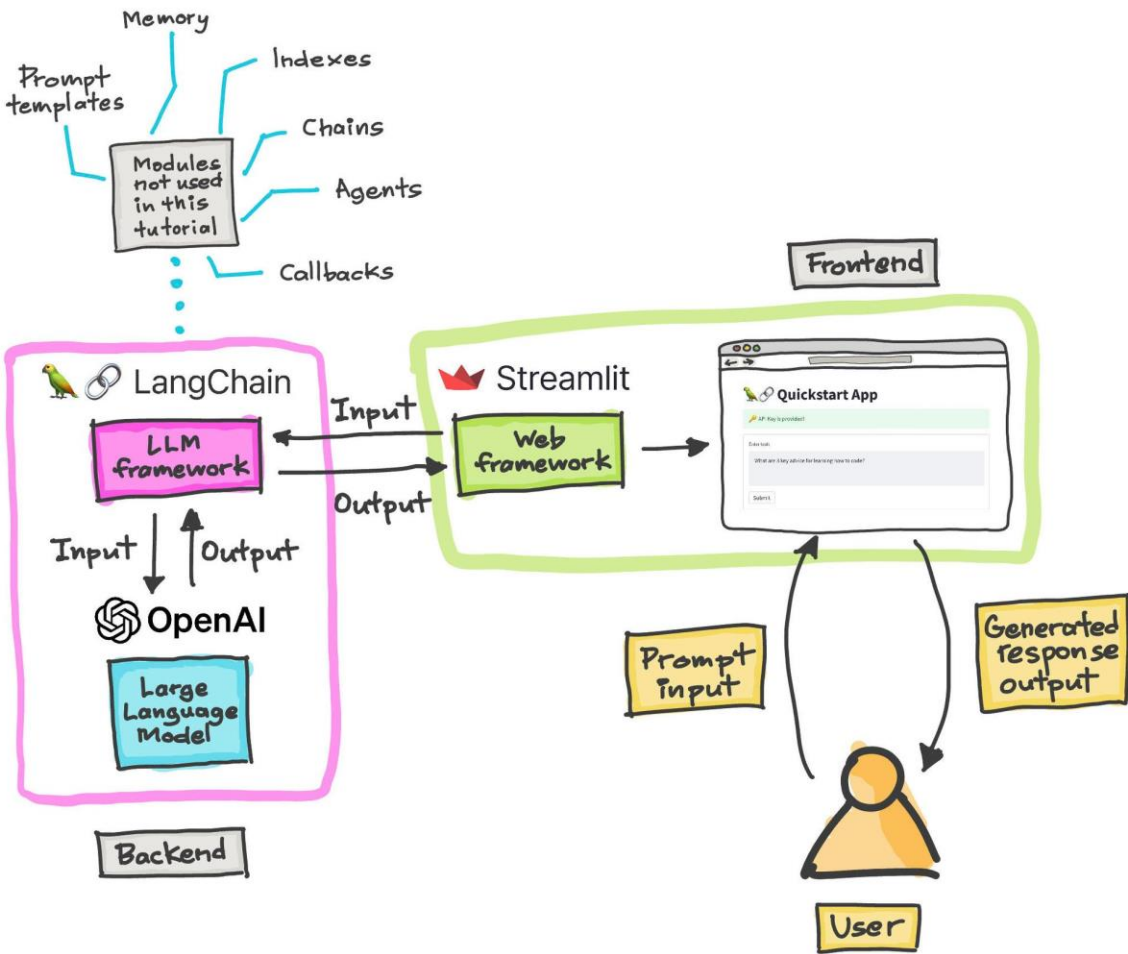
- To create an online service that hosts our solution.

Recap

- We have seen LangChain, a framework/wrapper that simplifies the use of LLMs (e.g., OpenAI API).
- LangChain also acts as a glue between the LLM and other components



LangChain

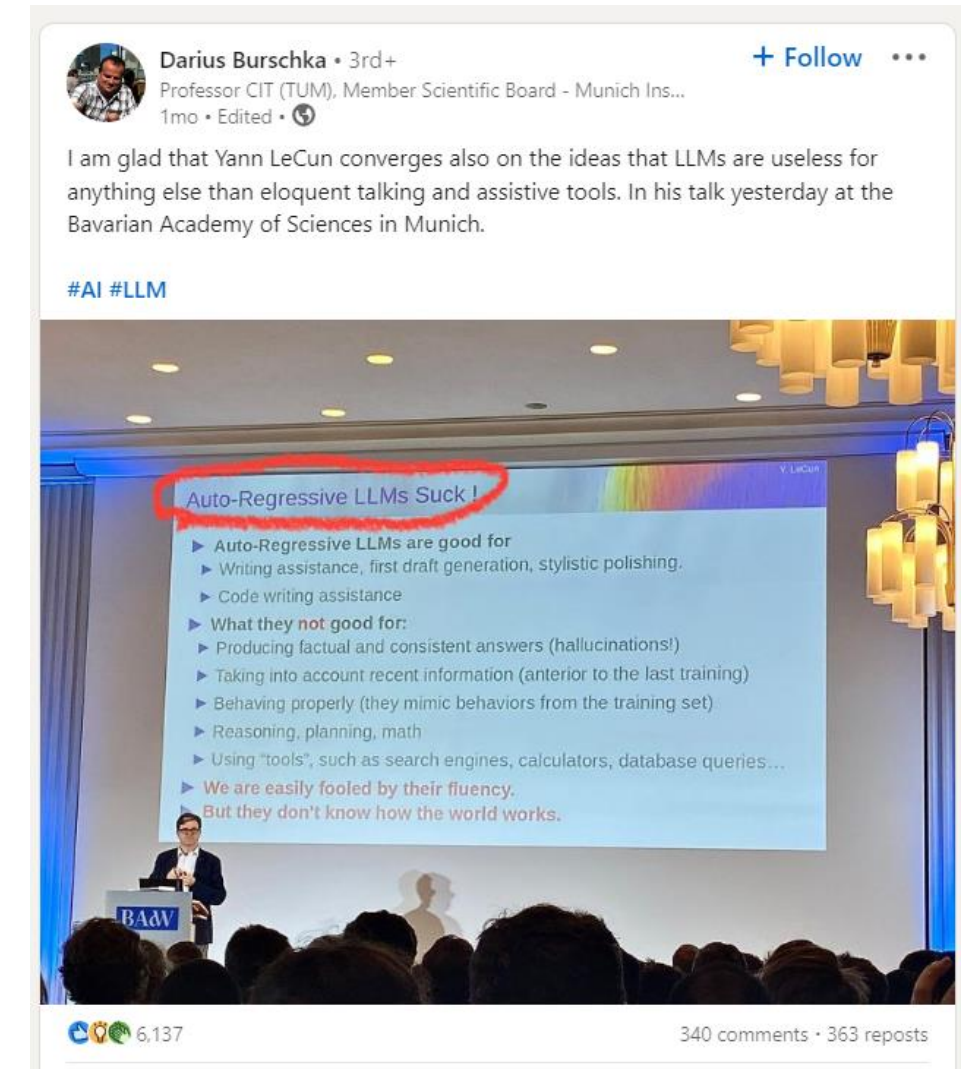


blog.streamlit.io

Hallucinations

Why Do LLMs Hallucinate?

- Language models are just glorified text completion tools.
- They are highly influenced by:
 - Data quality
 - Input context.



Examples

- Hallucinations in planning (Block Moving)
- Fake travel guides



NYTimes

Minimizing Hallucinations

- Parameter tuning (e.g., temperature).
- Prompt engineering.

Prompt Engineering

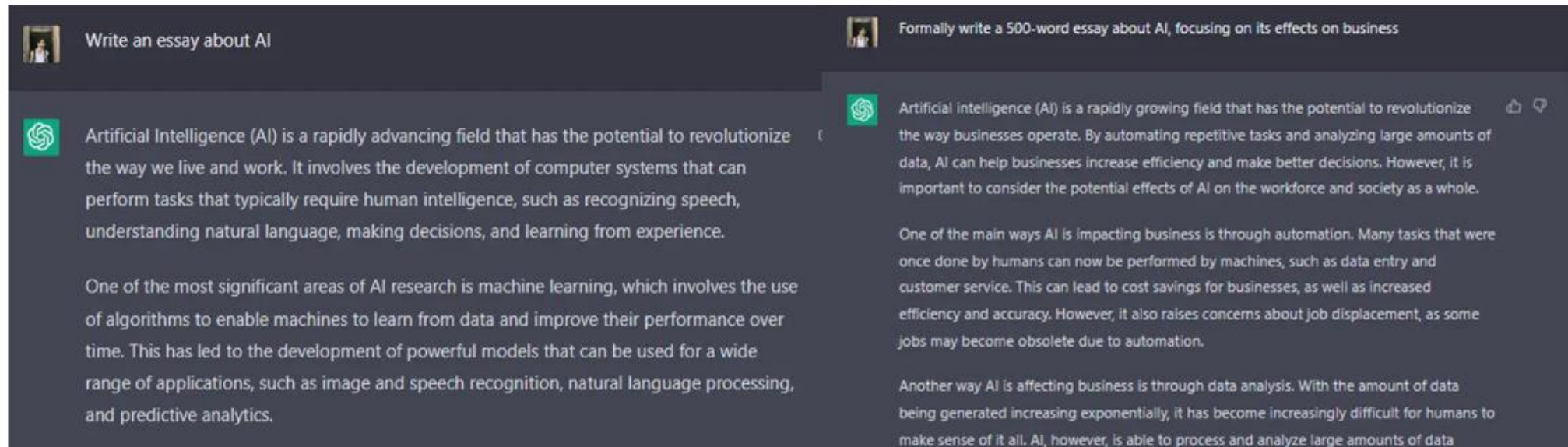
Why Need Prompt Engineering?

Let's say you have a pre-trained model to be used for an airline chat agent use-case :

- If the model has **not already been trained** on some **airline chat agent data**, one approach would be to **fine-tune it**.
- What if your model **has already been trained on such data** but you are not getting the desired output?
 - you probably need to **better engineer your prompt** to extract the desired knowledge/pattern.

What is Prompt Engineering?

Prompt Engineering is the practice of crafting prompts that guide the LLM to generate responses that better adhere to desired criteria (e.g., relevance, tone, etc.).



The image shows two side-by-side screenshots of an AI chat interface. The left screenshot shows a prompt 'Write an essay about AI' and a response that is a general overview of AI. The right screenshot shows a more specific prompt 'Formally write a 500-word essay about AI, focusing on its effects on business' and a response that is more detailed and structured, focusing on the impact of AI on business.

Left Screenshot:

Prompt: Write an essay about AI

Response:

Artificial Intelligence (AI) is a rapidly advancing field that has the potential to revolutionize the way we live and work. It involves the development of computer systems that can perform tasks that typically require human intelligence, such as recognizing speech, understanding natural language, making decisions, and learning from experience.

One of the most significant areas of AI research is machine learning, which involves the use of algorithms to enable machines to learn from data and improve their performance over time. This has led to the development of powerful models that can be used for a wide range of applications, such as image and speech recognition, natural language processing, and predictive analytics.

Right Screenshot:

Prompt: Formally write a 500-word essay about AI, focusing on its effects on business

Response:

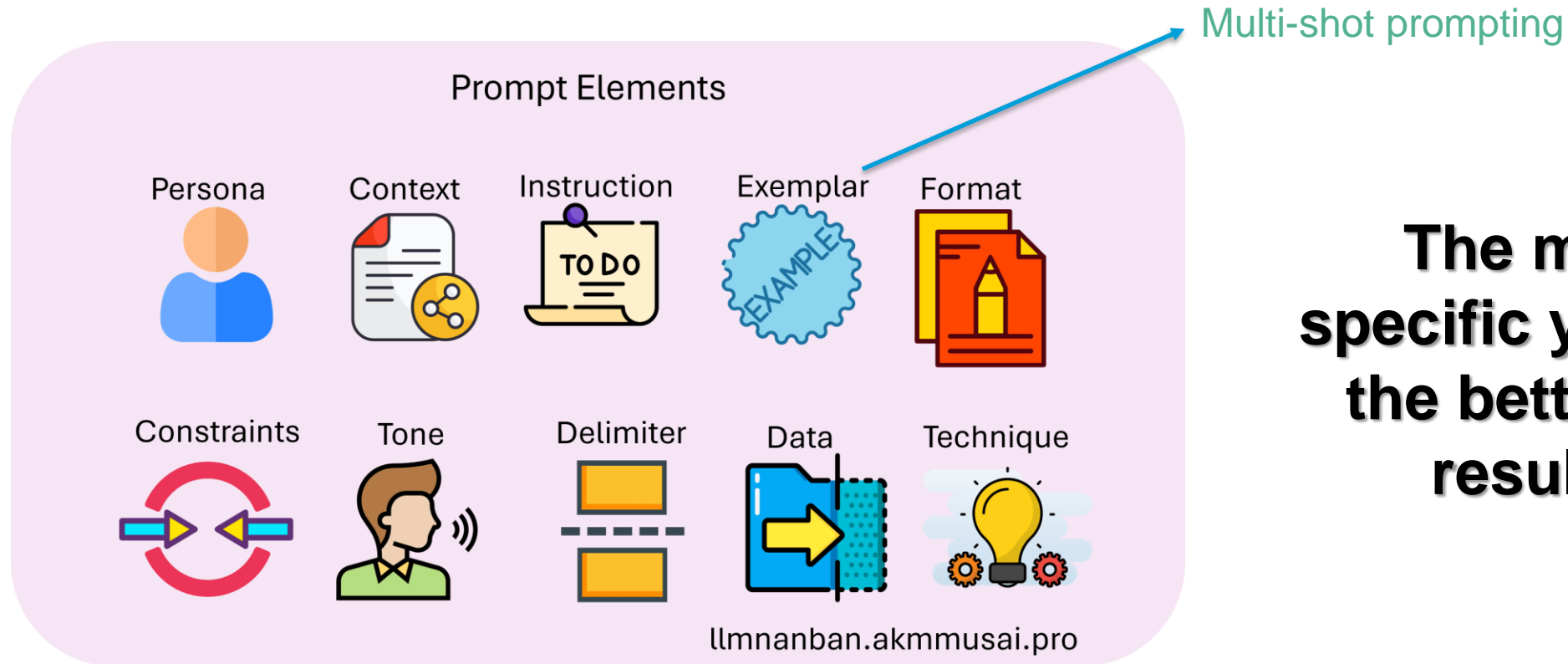
Artificial Intelligence (AI) is a rapidly growing field that has the potential to revolutionize the way businesses operate. By automating repetitive tasks and analyzing large amounts of data, AI can help businesses increase efficiency and make better decisions. However, it is important to consider the potential effects of AI on the workforce and society as a whole.

One of the main ways AI is impacting business is through automation. Many tasks that were once done by humans can now be performed by machines, such as data entry and customer service. This can lead to cost savings for businesses, as well as increased efficiency and accuracy. However, it also raises concerns about job displacement, as some jobs may become obsolete due to automation.

Another way AI is affecting business is through data analysis. With the amount of data being generated increasing exponentially, it has become increasingly difficult for humans to make sense of it all. AI, however, is able to process and analyze large amounts of data

[dataconomy](https://dataconomy.com)

Elements of Prompt Engineering



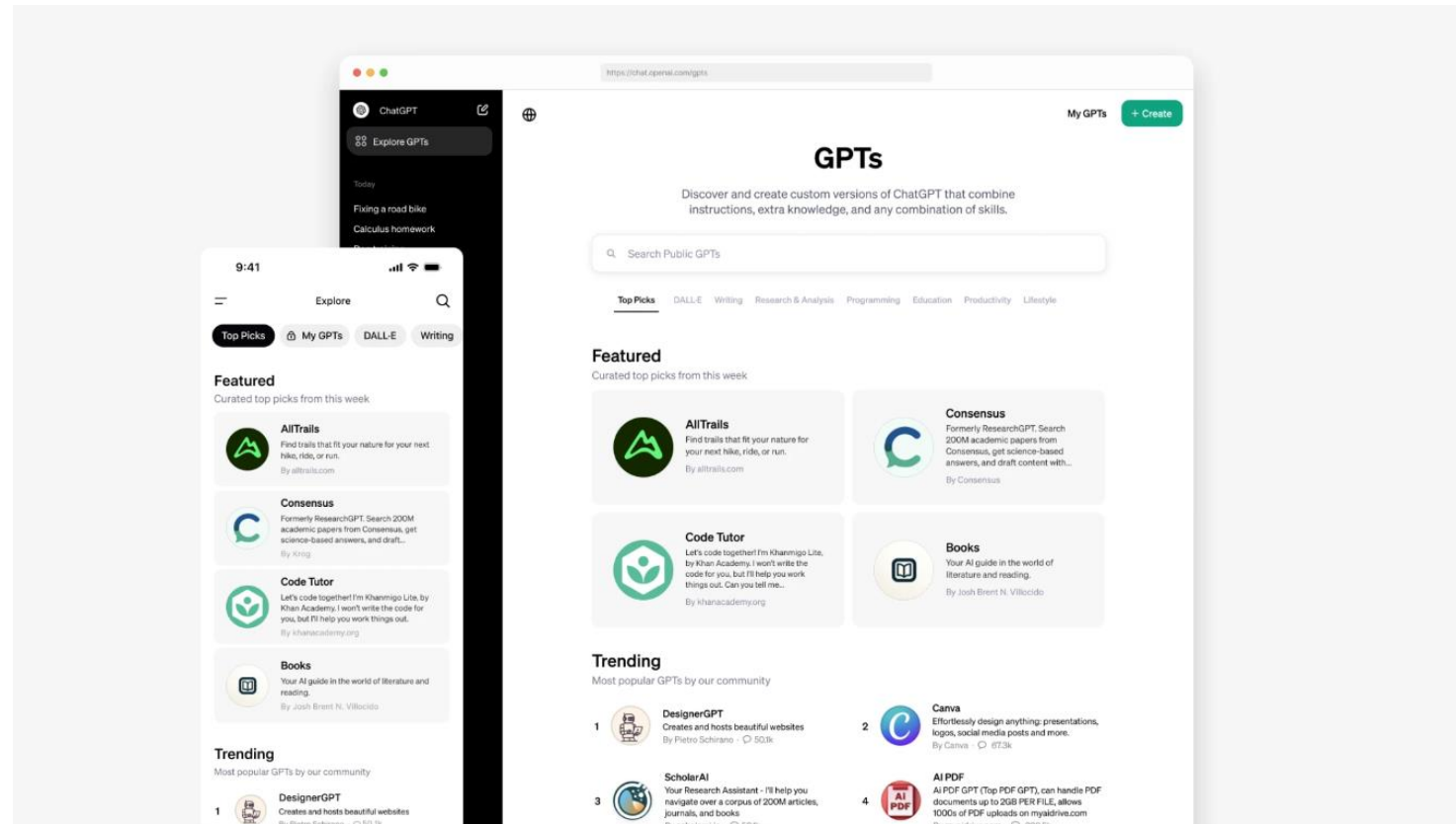
**The more
specific you are,
the better the
results!**

Prompt Engineering vs Fine-Tuning

	Prompt engineering	Fine-tuning
Technical Expertise Needed	Low	High
Cost	Low upfront/ High long-term	High upfront/ Low long-term
Effectiveness	Moderate	High

Example: OpenAI GPTs

- Let's try it!

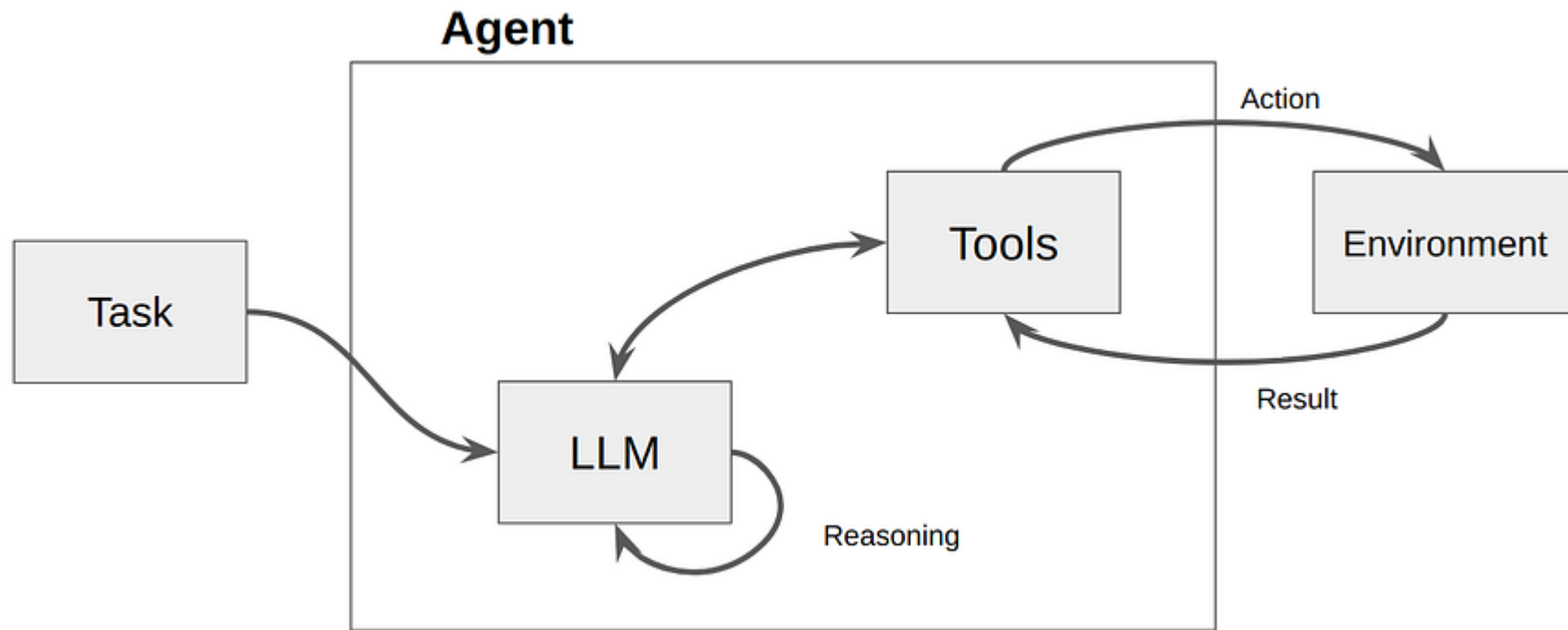


Tools and Agents

Demo

- Try some of the following prompts:
 - “Who is *<some obscure relative of yours>*. Look him up on Wikipedia”
 - Calculate the product of the following numbers *<an array of large numbers>*

How Agents Work?



betterprogramming.pub/

How Agents Work?

