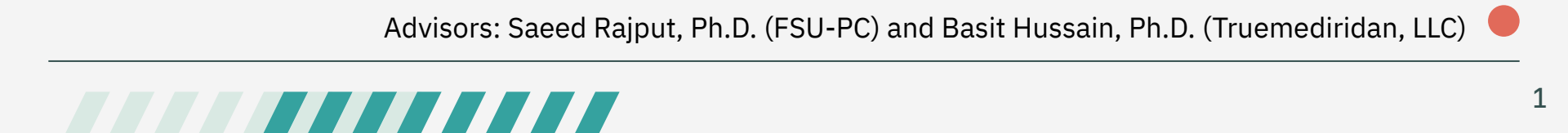




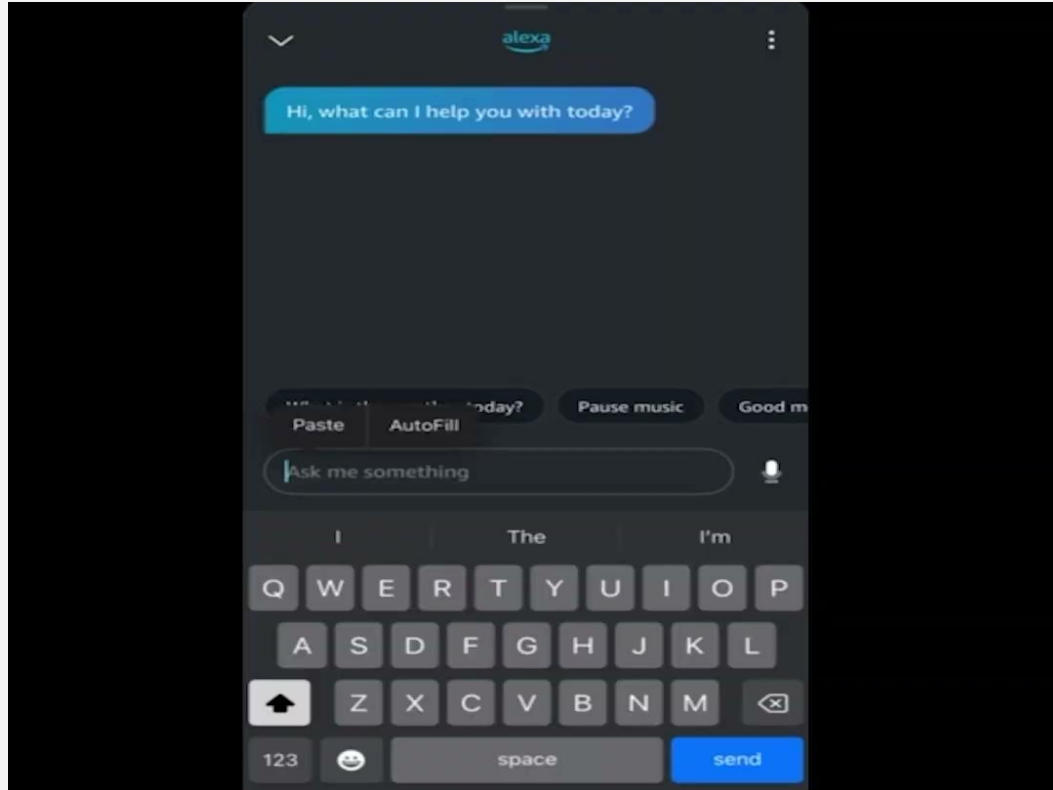
NEXUS: Neural Execution and Understanding System

Team: Jaehyun Lee and Shawn Hagler, CpE FSU-PC

Advisors: Saeed Rajput, Ph.D. (FSU-PC) and Basit Hussain, Ph.D. (Truemediridan, LLC)



NEXUS - Demo



Abstract

Leveraging the power of Large Language Models (LLMs), like ChatGPT, our enhanced virtual assistant framework transforms user interactions with its advanced comprehension and dynamic response generation. These LLMs not only unravel the complexity of human queries but also adeptly generate code, enabling the virtual assistant to seamlessly integrate and execute tasks via APIs without manual intervention. Our approach is paving the way for virtual assistants that autonomously interpret and execute a broad spectrum of complex user requests, bridging the divide between human conversation and automated action — a pivotal stride towards truly intelligent digital assistants.

Agenda

01.

Overview

02.

Problem

03.

Background
Information

04.

Solution

05.

Future Works

06.

Reference



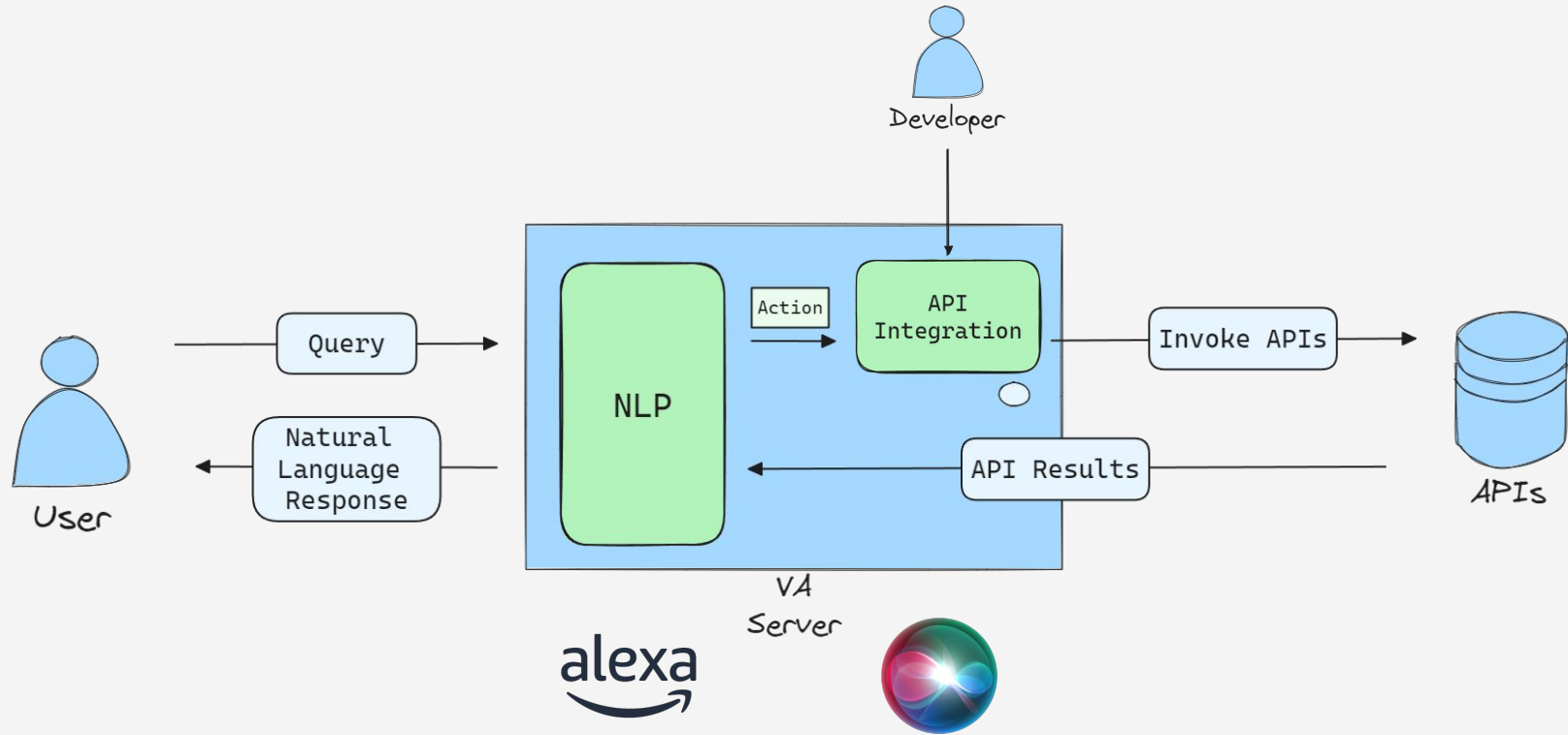


01.

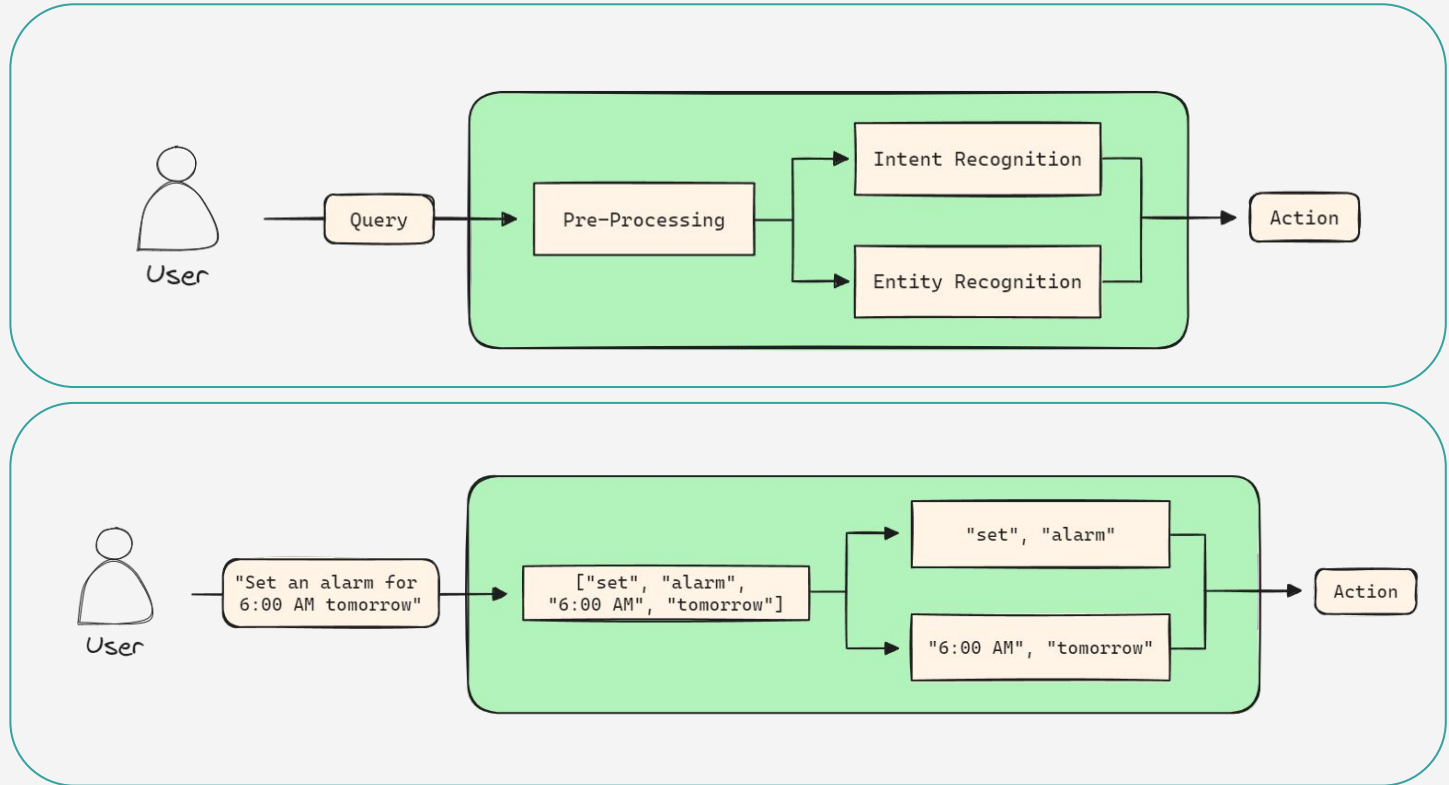
Overview



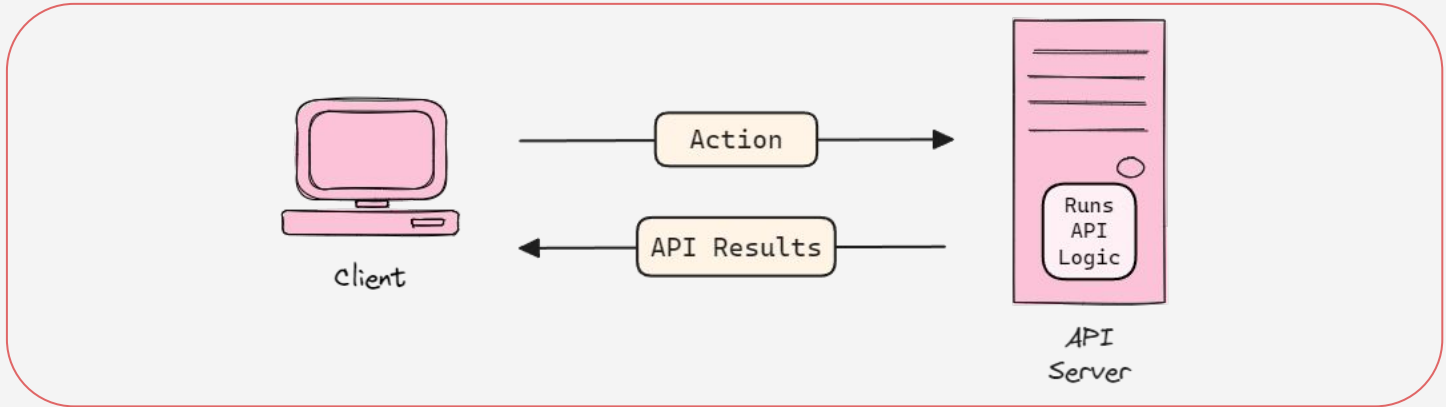
Virtual Assistants



Natural Language Processing (NLP)



Application Programming Interface (API)



- Google Maps
- Weather Apps
- Yelp
- Alarm
- etc.



02.

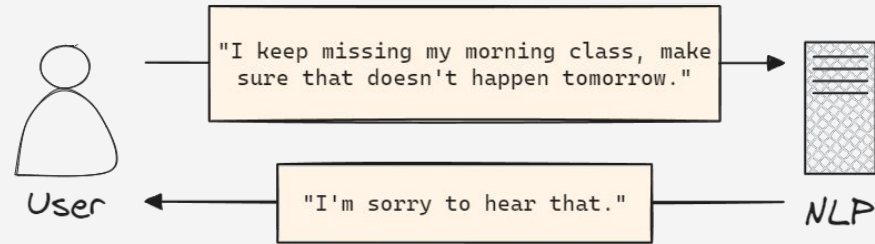


Problem



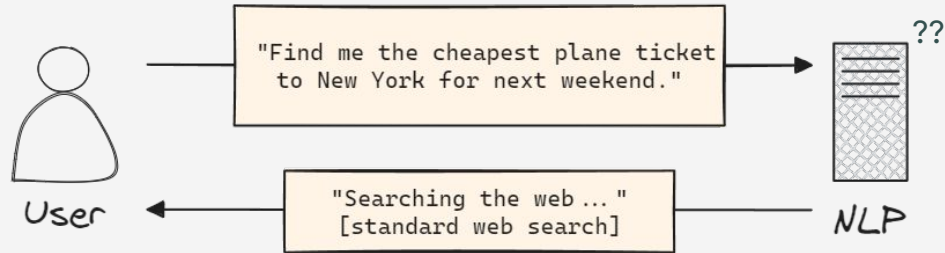
NLP - Limitations

- Poor contextual understanding
- Limited execution
- Poor flexibility in response



API - Limitations

- Manual development integration
- Manual maintenance



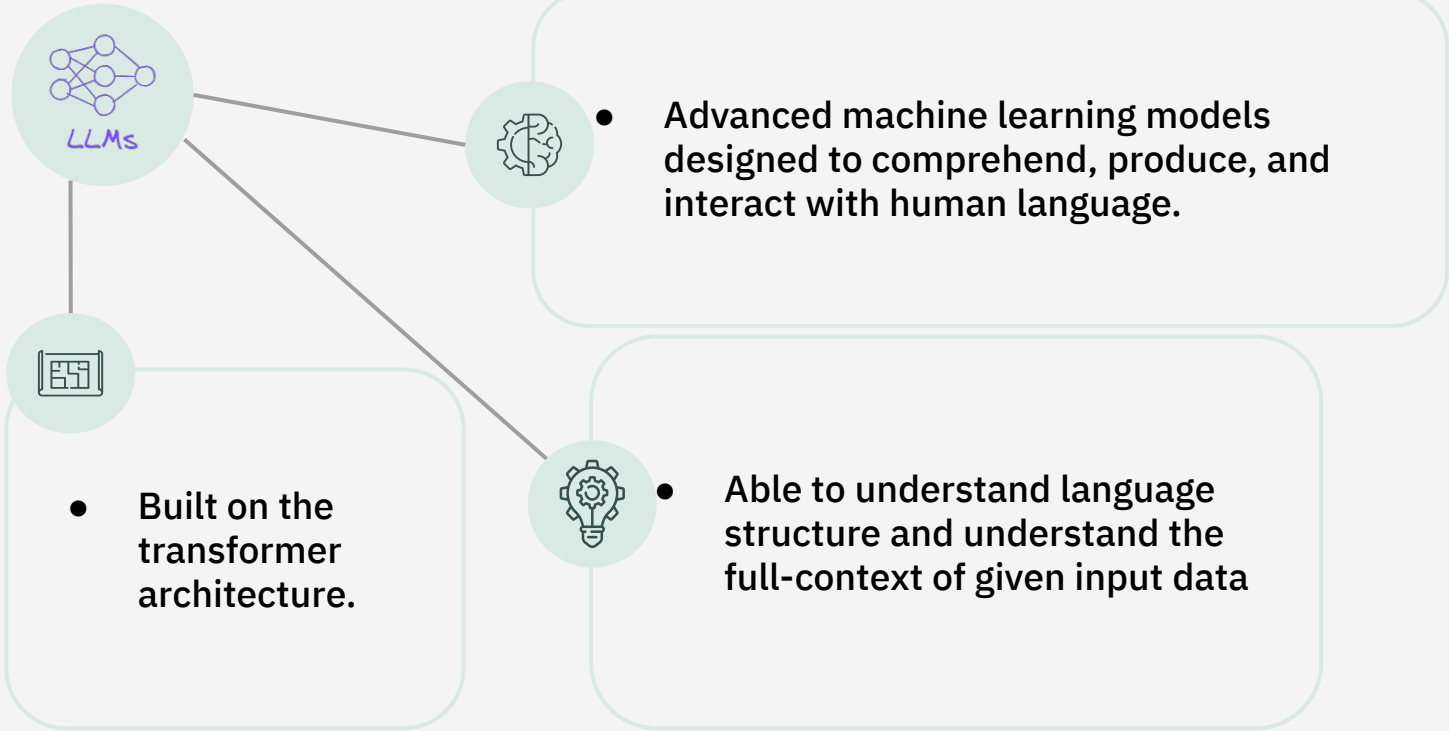


03.

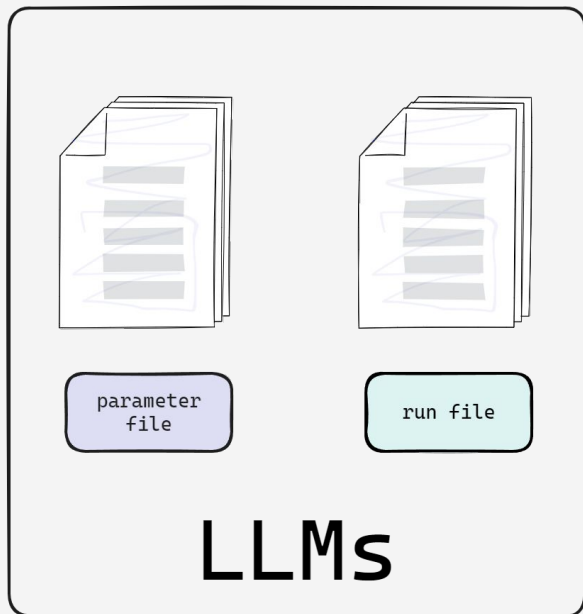
Background Information



Large Language Models (LLMs)



Large Language Models (LLMs)



Contains weights or parameters of Neural Network that is language model

parameter file

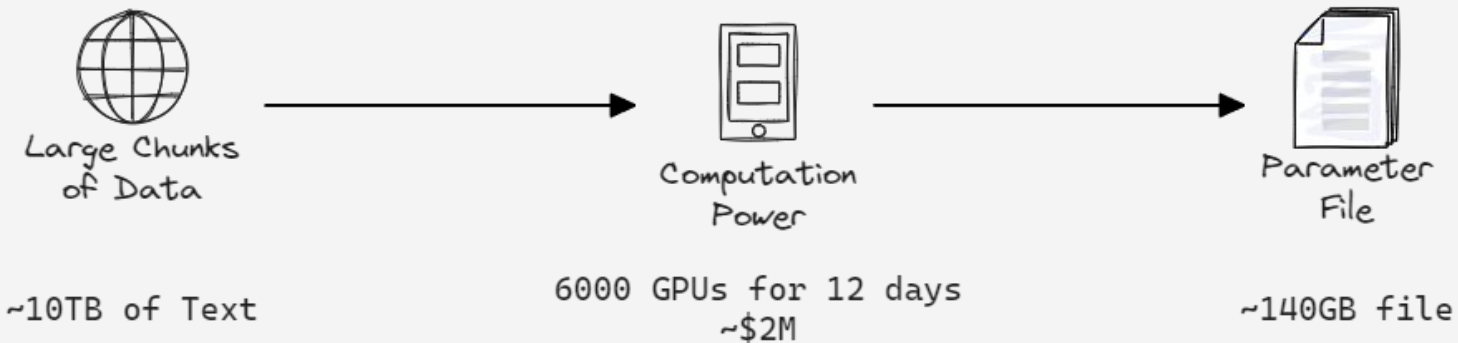


Runs the parameters (Neural Network) with no dependencies

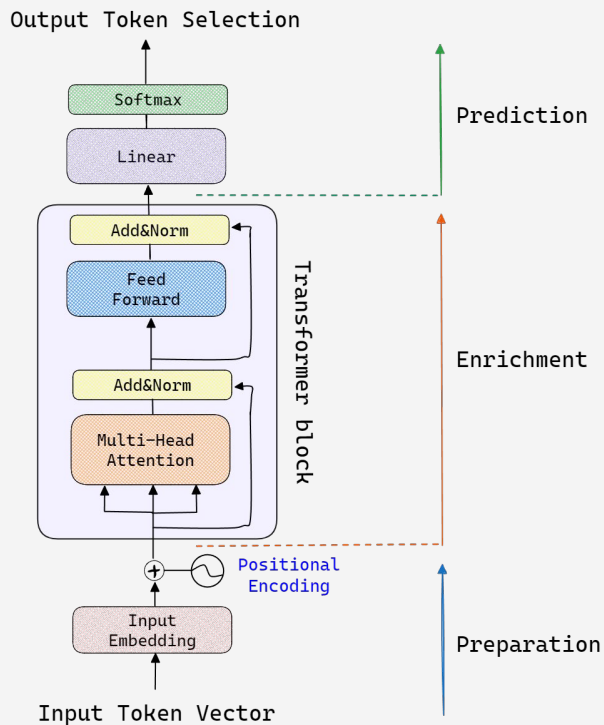
run file

LLMs - Training

*Llama 2 70B numbers



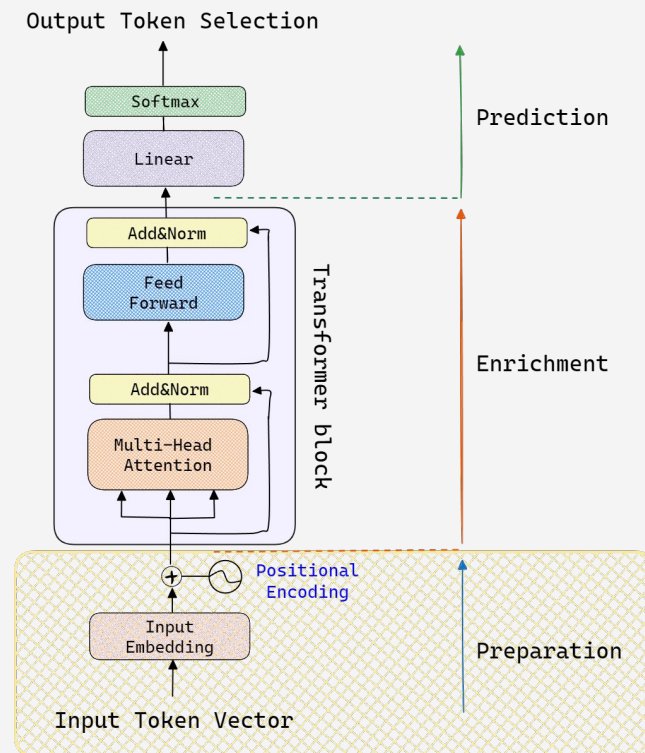
LLMs - Transformer



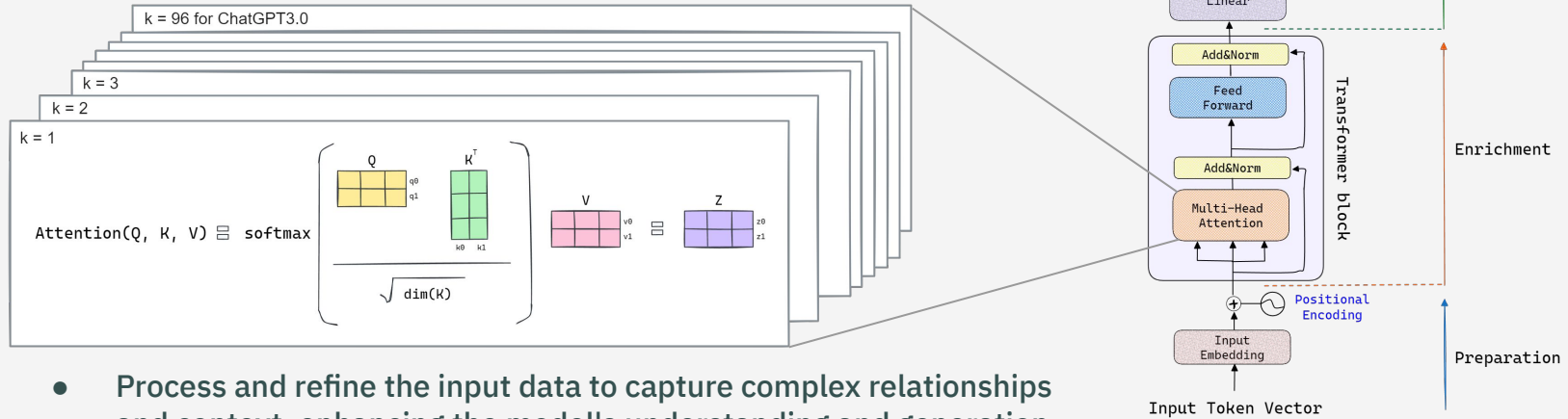
- A specific type of neural network, a machine learning model that takes in tokens and produces prediction for what token comes next.
- Parameters are dispersed throughout the network
- Has a self-attention mechanism allowing for interpretability

LLMs - Transformer

- Initial input consist of sequence of word tokens which transformed into word embeddings.
- Positional Encodings: Adds to word embeddings to provide positional information due to transformer's lack of inherent sequence order understanding.

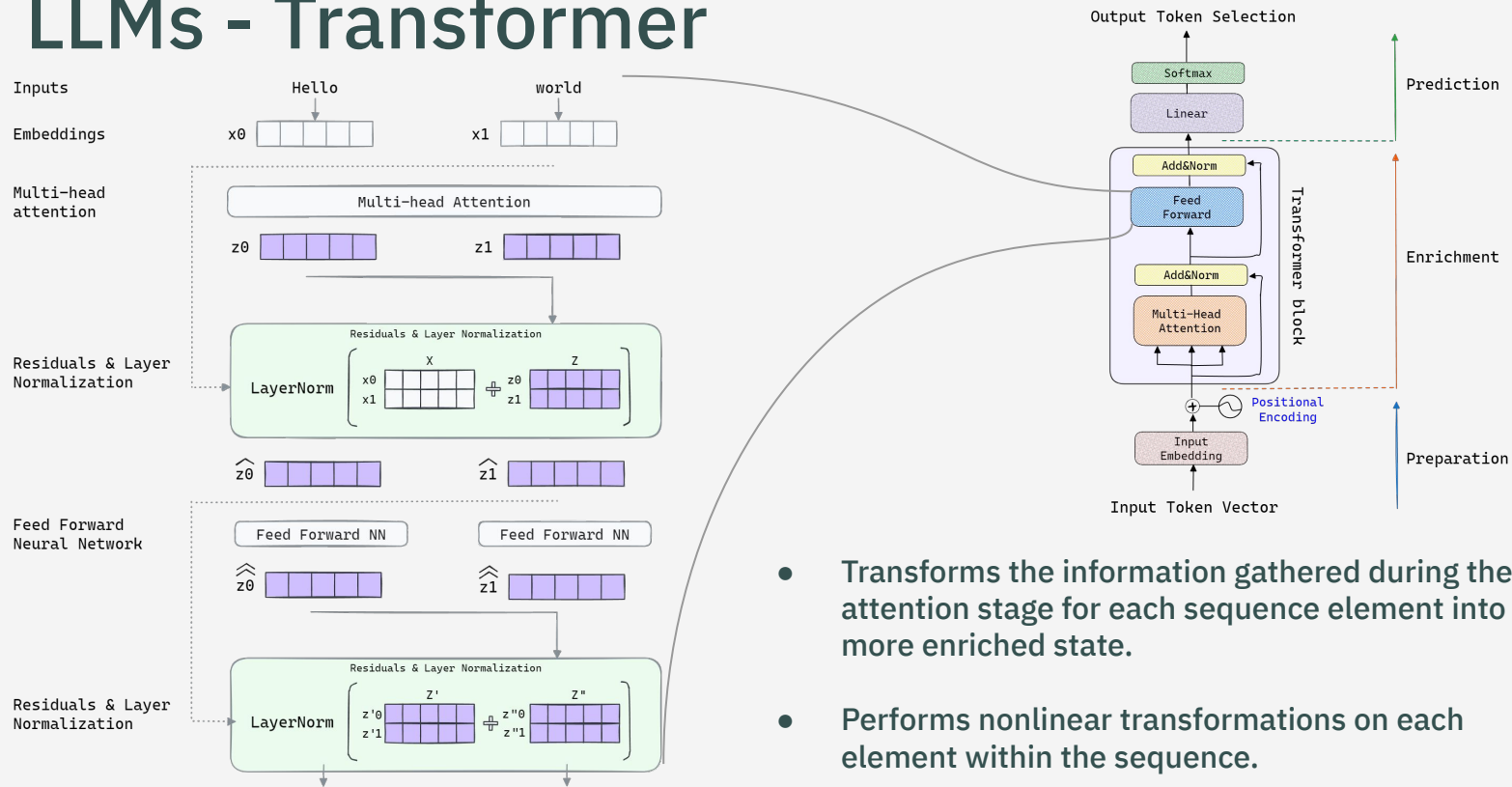


LLMs - Transformer



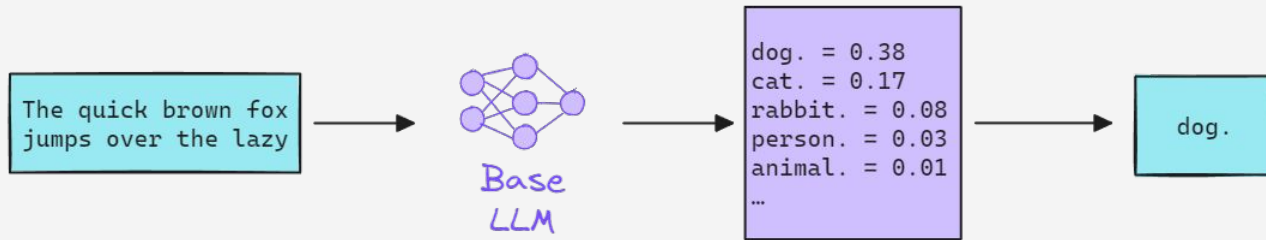
- Process and refine the input data to capture complex relationships and context, enhancing the model's understanding and generation capabilities.
- Multi-head Attention: Enhances vector representations by adding contextual relationship and logic.
- The query vector (**Q**) contains the present token, the key vector (**K**) encompasses all other tokens in the sequence, and the value vector (**V**) encompasses all tokens within the sequence. They are tuneable vectors.

LLMs - Transformer

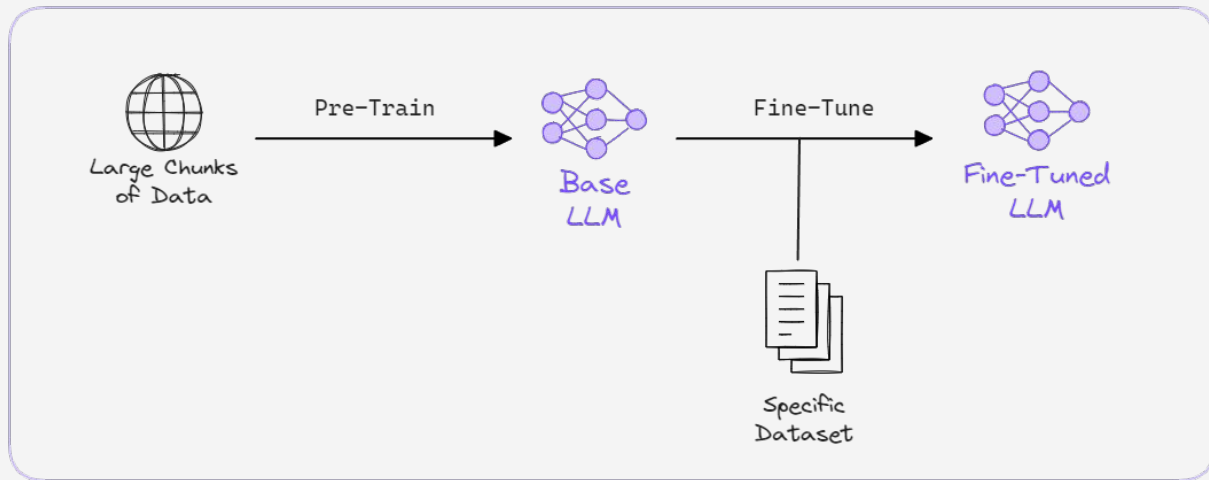


- Transforms the information gathered during the attention stage for each sequence element into a more enriched state.
- Performs nonlinear transformations on each element within the sequence.

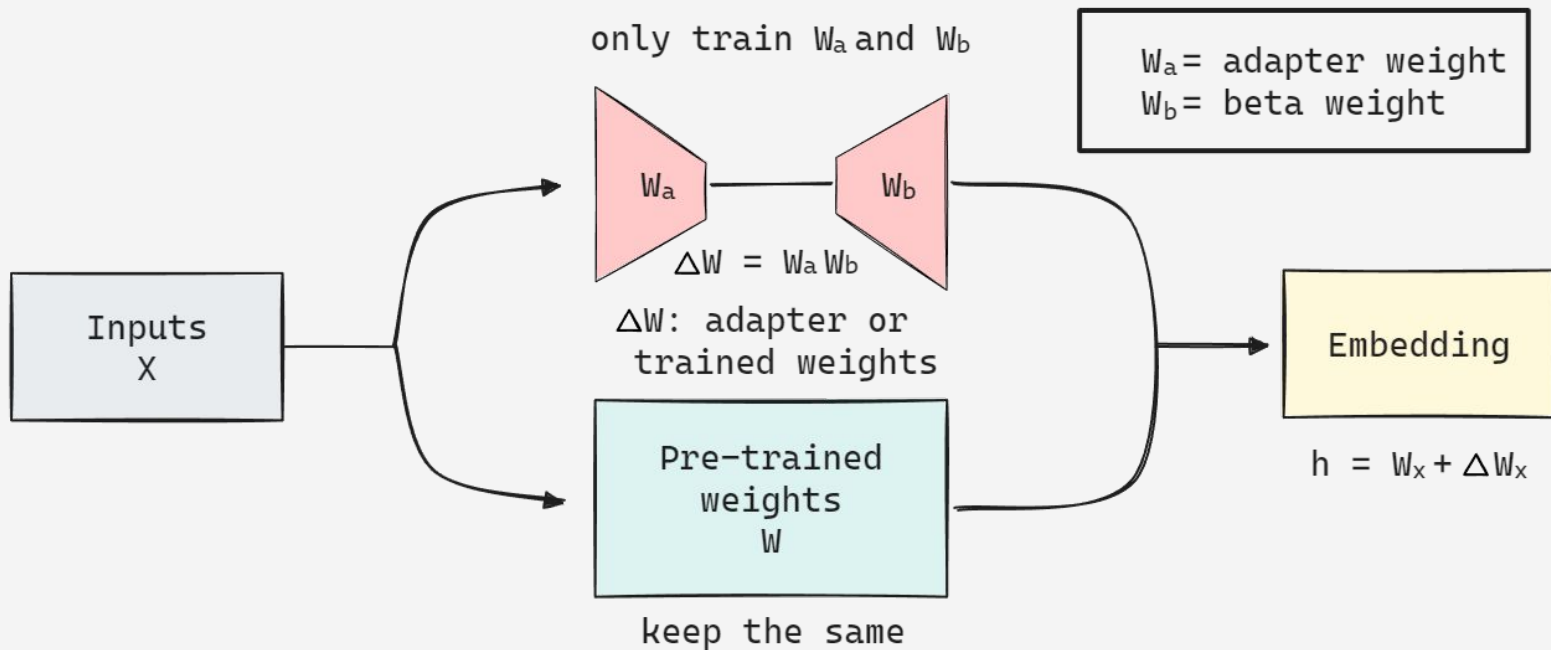
LLMs - Usage (Base)



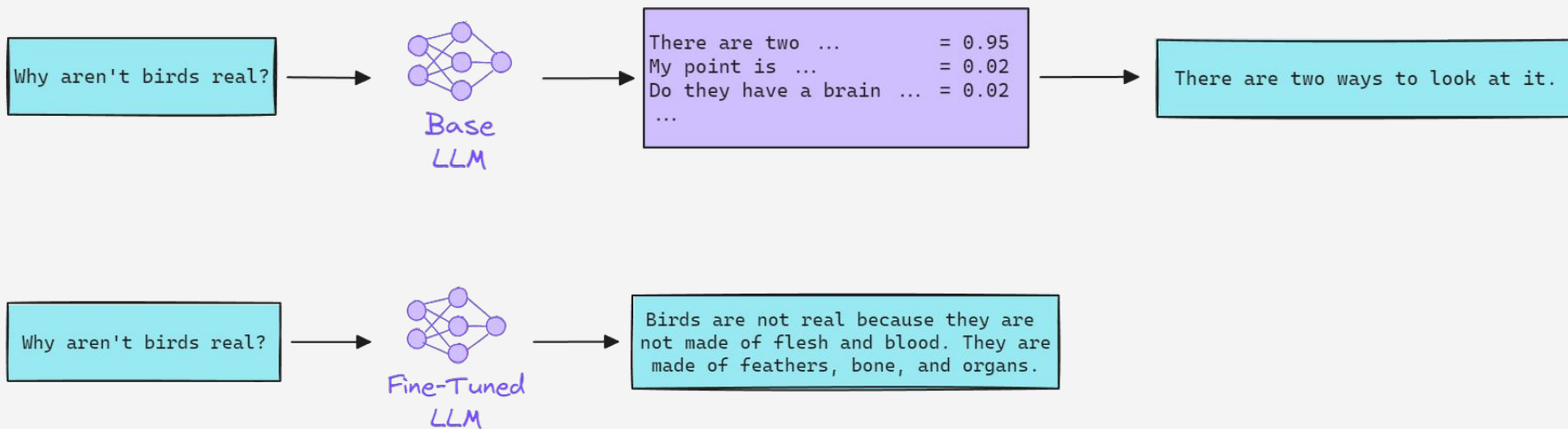
LLMs - Fine-Tuning



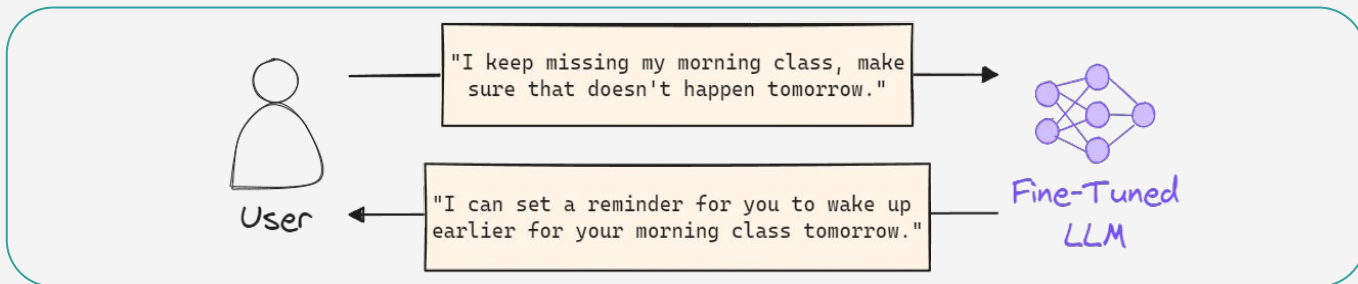
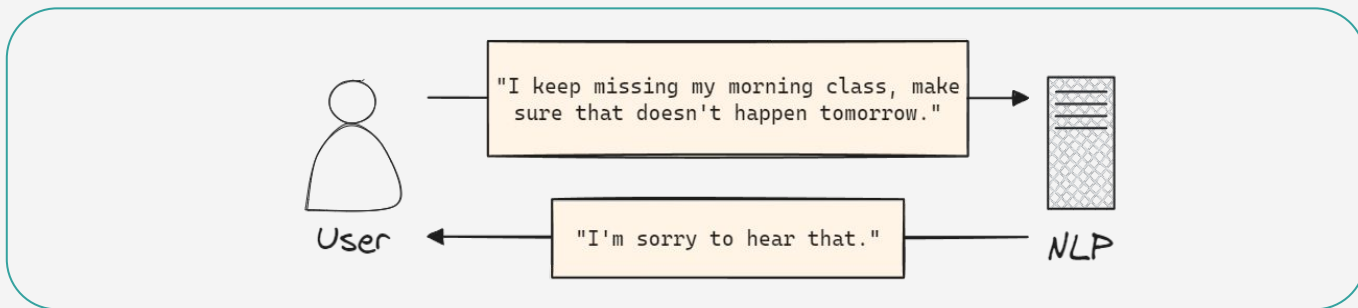
LLMs - Fine-Tuning



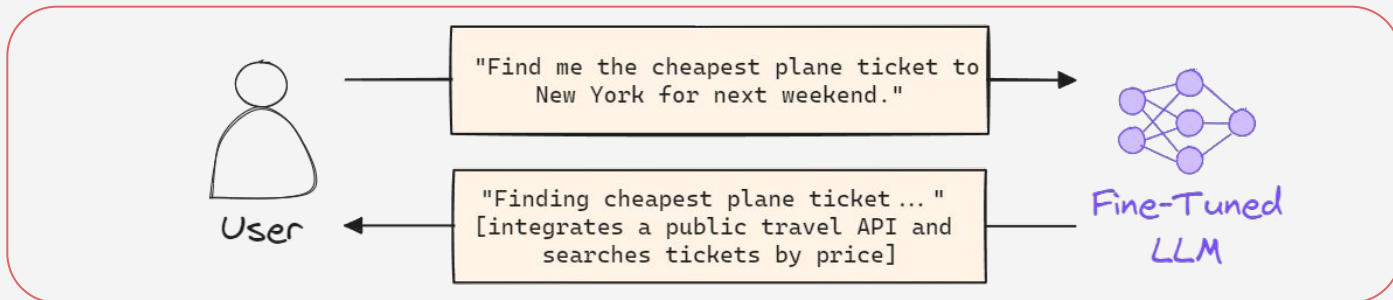
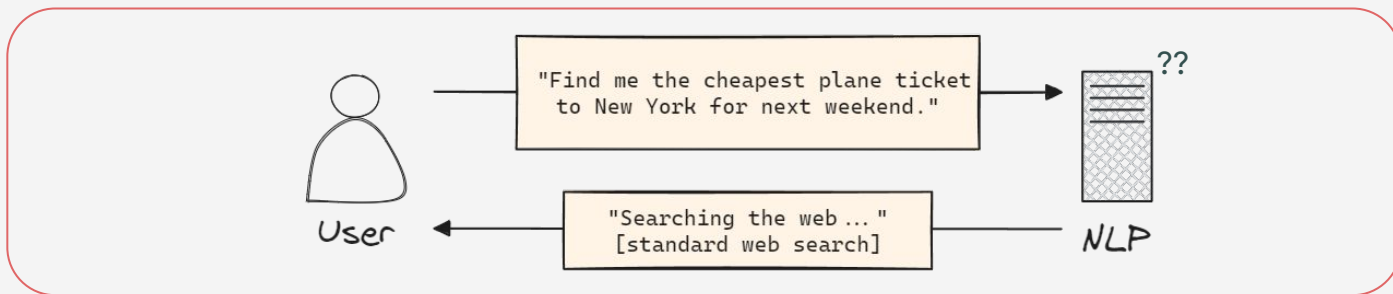
LLMs - Usage (Fine-Tuned)



LLMs - Virtual Assistant Usage



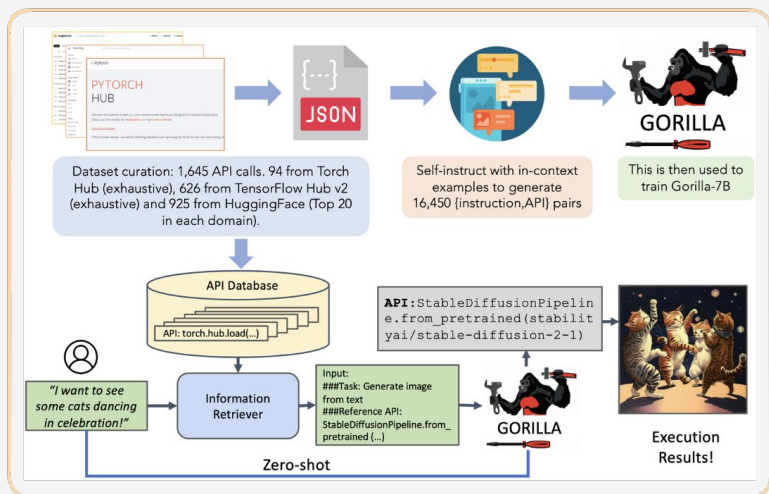
LLMs - Virtual Assistant Usage



Related Works

GorillaLLM

- Fine-Tuned LLM with API usage
- Code synthesis to call APIs
- APIs limited to other machine learning models



Rabbit r1

- An AI assistant that integrates advanced AI technology.
- Raised ~\$60 million in funding
- ~100,000 units sold (pre ordered) at \$199 a piece



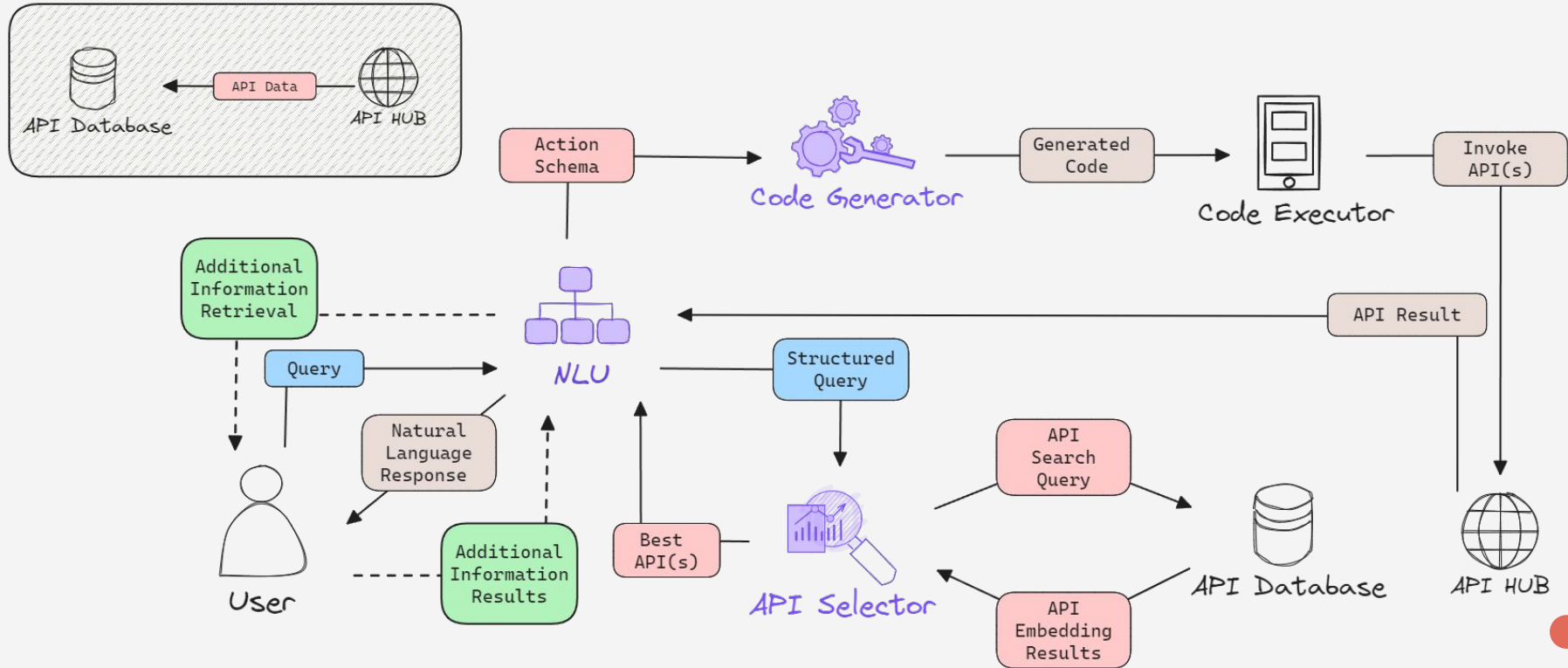


04.

Solution - NEXUS



NEXUS - Previous Semester Result



ChatGPT

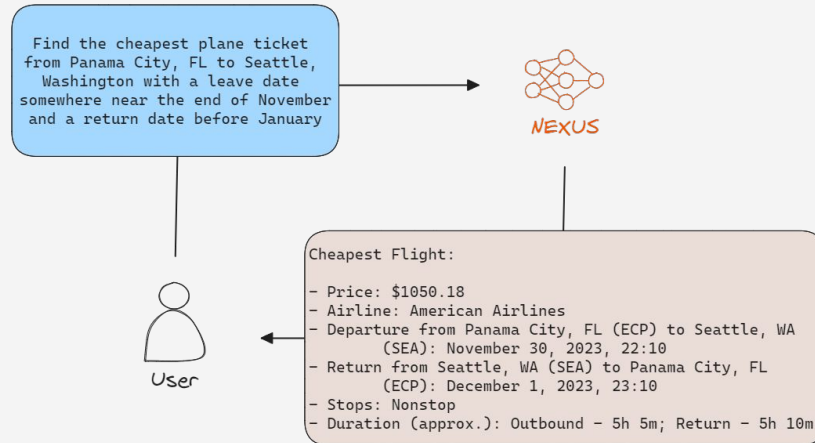
ChatGPT is a state-of-the-art large language model developed by OpenAI.

- It stands for "Chat Generative Pre-trained Transformer."
- Trained on large dataset of human language and fine-tuned on instructions and chat bot interactions.
- It is hosted by OpenAI and can be accessed through their commercial API.

GPT-4 Turbo

<u>Parameter</u>	1.8 Trillion Parameters (2TB Parameter File)
<u>Knowledge Cutoff</u>	Data up to April 2023

NEXUS - Previous Semester Result



LLaMA 3

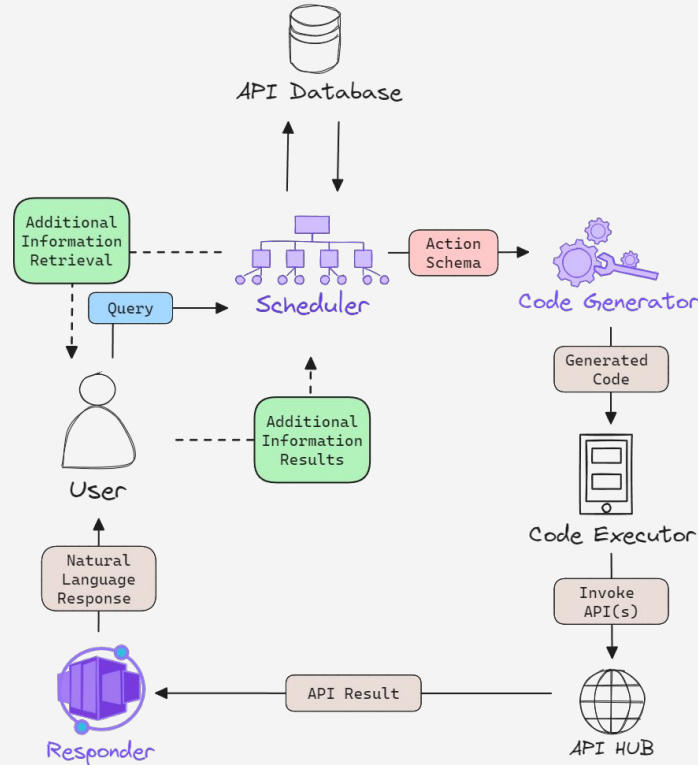
LLaMA 3 is the next generation of state-of-the-art **open source** large language model developed by Meta and released April 18, 2024.

- It stands for "Large Language Model Meta AI"
- Pretrained and instruction-fine-tuned language models with 8B and 70B parameters that can support a broad range of use cases.

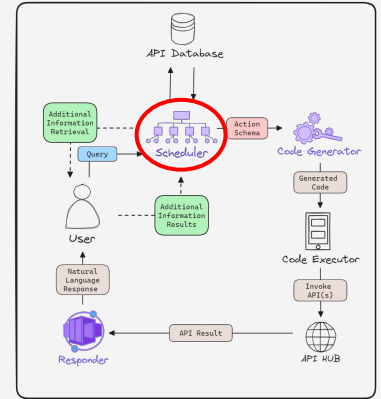
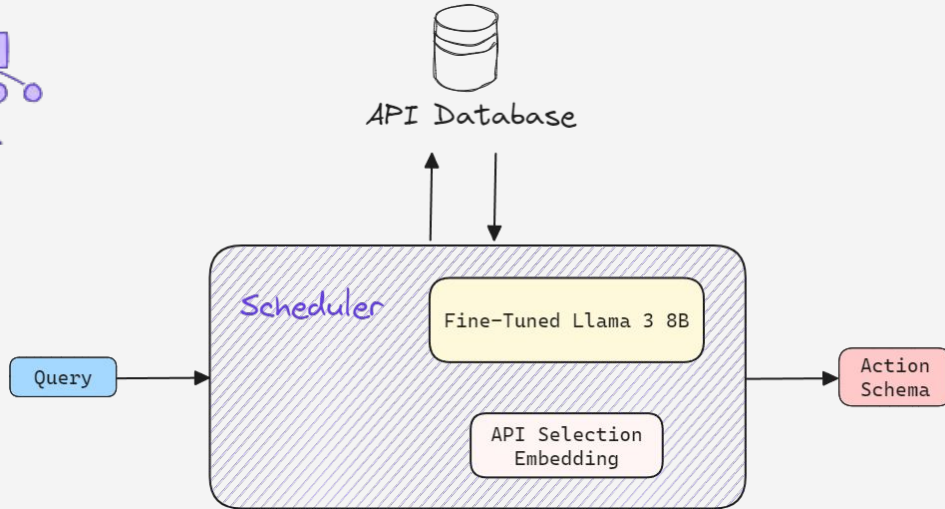
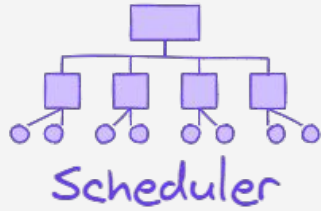
LLaMA 3-8B

<u>Parameter</u>	8 Billion Parameters
<u>Knowledge Cutoff</u>	Data up to April 2023
<u>Context Window</u>	Up to 8 Thousand Tokens
<u>Pretrained</u>	On 15 Trillion Tokens

NEXUS - Current Work

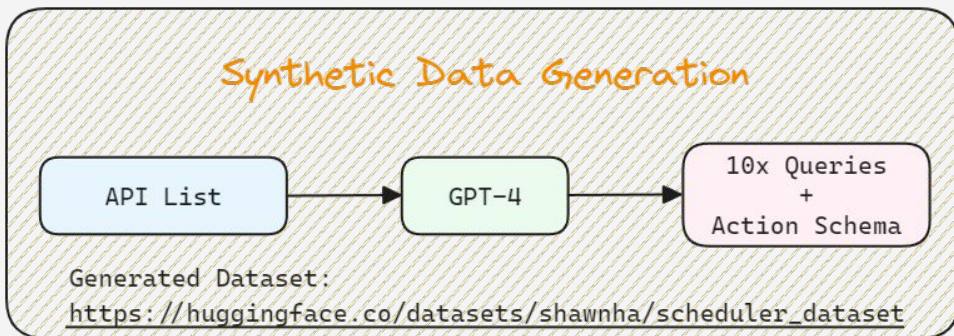


NEXUS - Scheduler

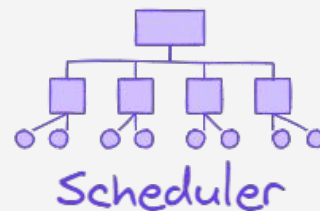
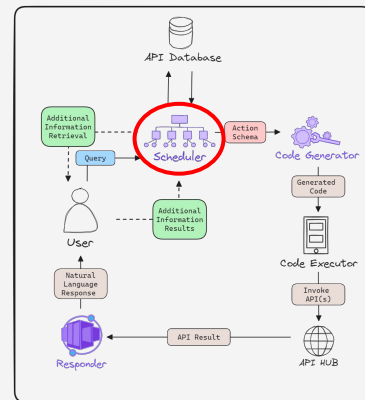
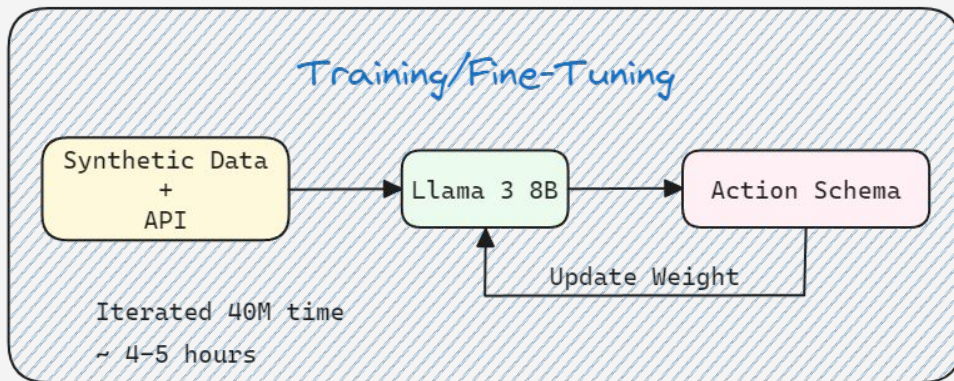


NEXUS - Scheduler

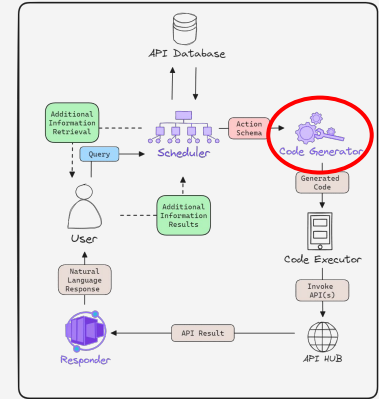
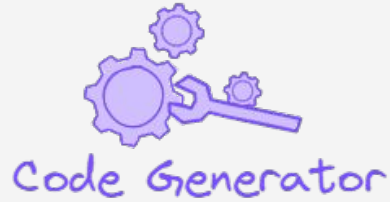
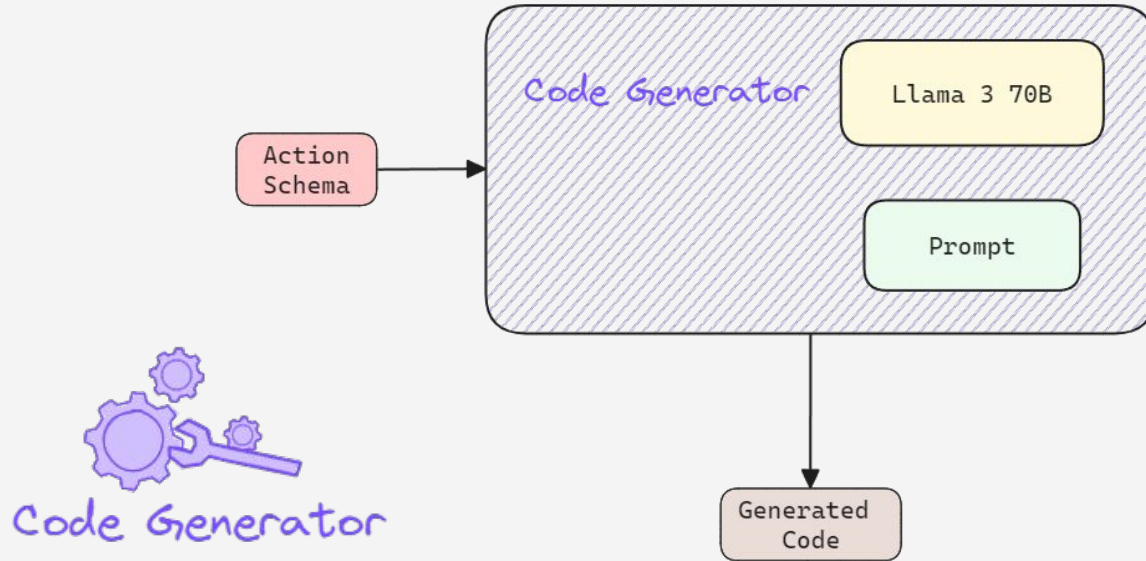
Synthetic Data Generation



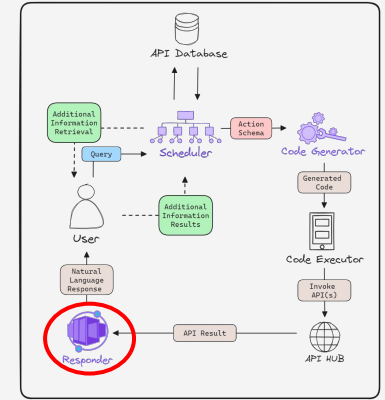
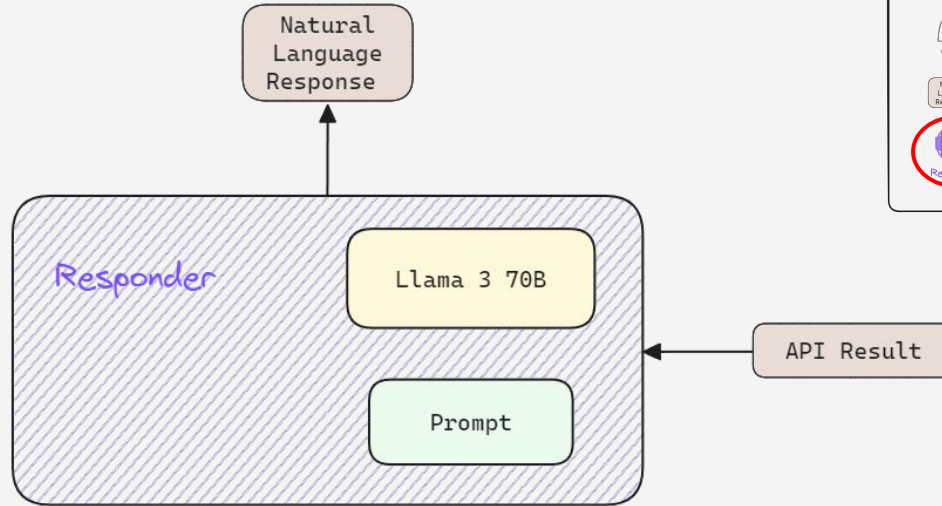
Training/Fine-Tuning



NEXUS - Code Generator



NEXUS - Responder



NEXUS - Other Components



API Database

List of APIs that can be used.

List created from RapidAPI, limited to the free open APIs.



Code Executor

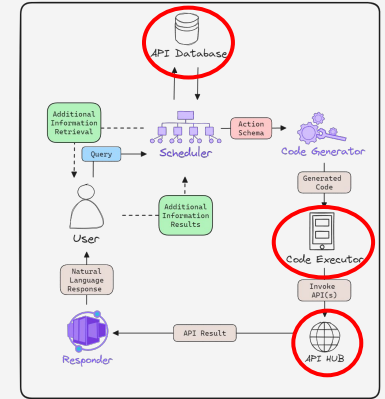
System that allows us to execute the code that was generated.



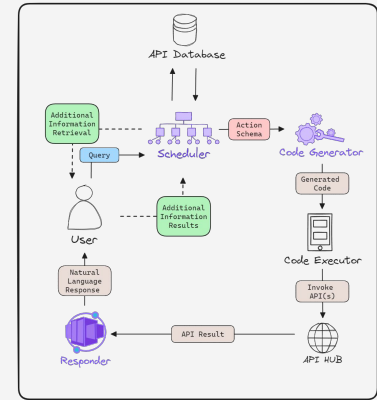
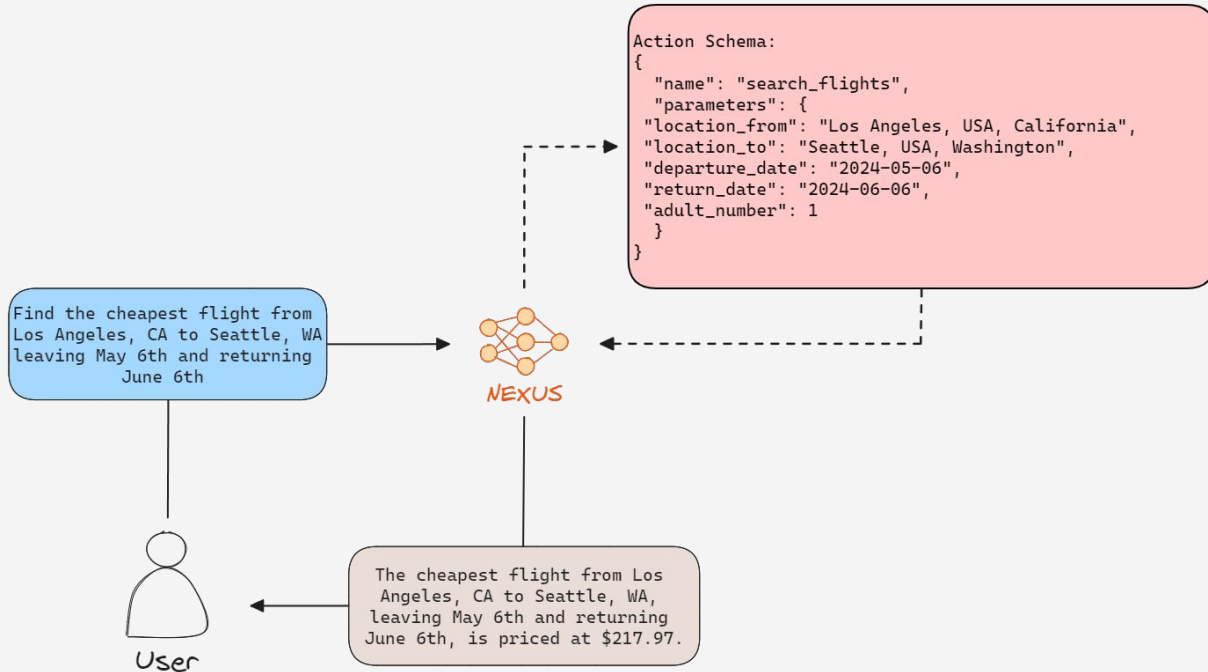
API HUB

API host for invoking the API(s) and getting results.

RapidAPI used, allowing various API connections.



NEXUS - Result



Hardware Components

<u>CPU</u>	AMD Ryzen 9 7950x (16-core 32-threads)
<u>GPU</u>	Radeon 7900 XTX (24 GB VRAM)
<u>RAM</u>	DDR5-6000MHz (64 GB)
<u>SSD</u>	4 TB
<u>Power Supply</u>	Corsair HX1500i
<u>Motherboard</u>	ASUS ROG STRIX B650-A



Desired Hardware

<u>CPU</u>	AMD EPYC 7502 (32-core 64-threads)
<u>GPU</u>	6x Radeon 7900 XTX (24 GB VRAM) (144GB Total)
<u>RAM</u>	DDR5-6000MHz (128 GB)
<u>SSD</u>	4 TB
<u>Power Supply</u>	2x Corsair HX1500i
<u>Motherboard</u>	ASRock Romed 16QM3

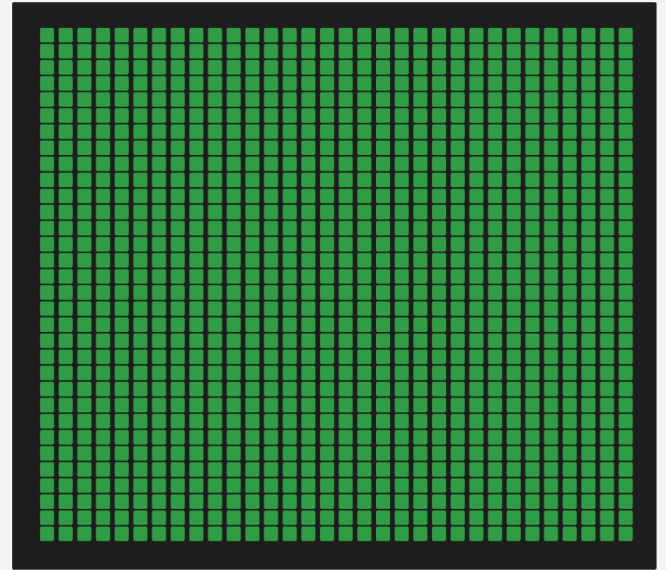


Graphics Processing Units (GPU)

- **Parallel Processing Capability:** GPUs are designed to handle many tasks simultaneously.
- **High Throughput:** GPUs can handle a higher throughput of data due to their design.
- **Memory Bandwidth:** Allows the GPU to access, read, and write large amounts of data quickly.



CPU Cores



GPU Cores



05.

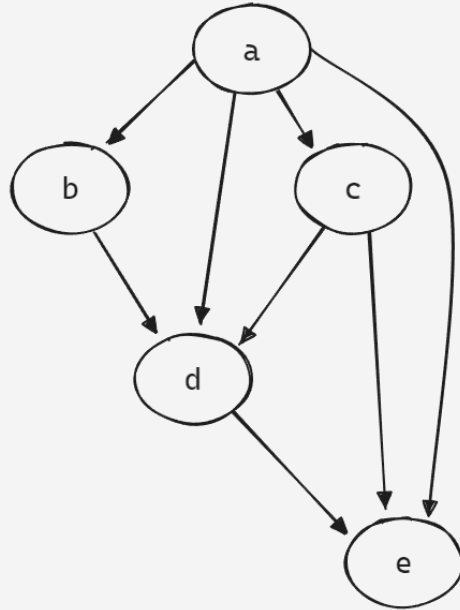
Future Works



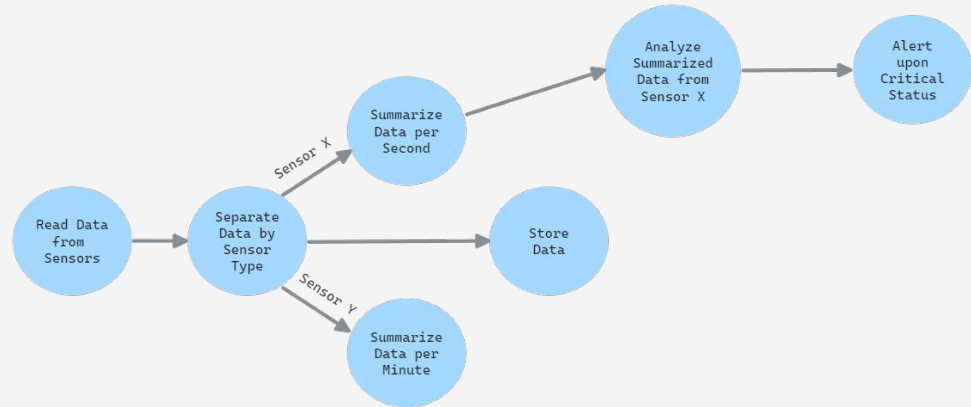
Future Works

- Host our own APIs that are actually useful instead of APIs from RapidAPI
 - Higher quality and quantity
- Fine-tune LLMs for the three main Modules on specific tasks for higher accuracy and lower power usage
 - Fine-tune Scheduler module on **Directed Acyclic Graph (DAG)** and use Graph Neural Network (GNN)
 - Fine-tune Code Generator module on generating code using json formatted file
 - Fine-tune Responder module on generating short response using json formatted file
 - Requires high computing power in order to fine-tune

Directed Acyclic Graph (DAG)



- DAG is a directed graph without cycles, consisting of vertices connected by edges that flow in a single direction, ensuring no closed loops exist, and allowing for topological ordering.

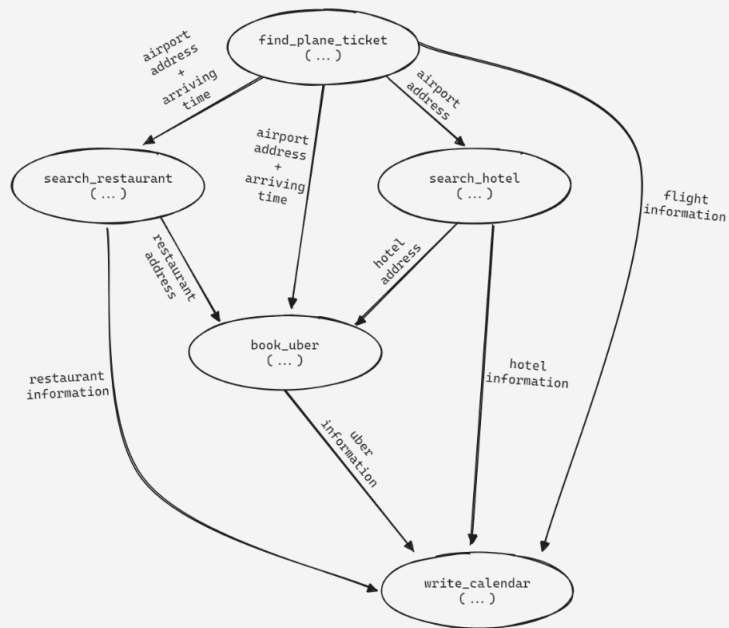


DAG Example

Find me the cheapest plane ticket from Los Angeles, CA to Seattle, WA from May 6th to June 6th, and find a nearby hotel with the best rating. I would like to eat as soon as I get there, so find me a famous nearby restaurant. I am pretty sure that I am going to need a ride, so book me an Uber for my trip. Have all this information added to my calendar.



NEXUS



The most affordable flight from Los Angeles, CA to Seattle, WA, scheduled between May 6th and June 6th, is priced at \$217.97. The hotel is conveniently situated within 5 miles of the airport. A list of top recommended restaurants has been compiled and incorporated into the calendar. Uber rides have been successfully booked for transportation. For all detailed information, please check the calendar app.





06.



References



Reference

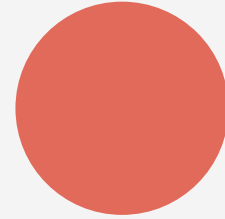
[1] S. G. Patil, T. Zhang, X. Wang, and J. E. Gonzalez, “Gorilla: Large Language Model Connected with Massive APIs,” arXiv.org, May 24, 2023. <https://arxiv.org/abs/2305.15334>

[2] Rabbit Research Team, "Learning Human Actions on Computer Applications," OS-Rabbit, Dec. 3, 2023. Available: <https://www.rabbit.tech/research>

[3] A. Vaswani et al., “Attention Is All You Need,” arXiv.org, Jun. 12, 2017. <https://arxiv.org/abs/1706.03762>

[4] L. Ouyang et al., “Training language models to follow instructions with human feedback,” arXiv:2203.02155 [cs], Mar. 2022, Available: <https://arxiv.org/abs/2203.02155>

Thank You!



NEXUS - Demo

