Session 1
GenAI overview and Transformers – Part 1

Palacode Narayana Iyer Anantharaman

9th Aug 2024

Welcome!

- Generative AI is a disruptive technology that is likely to shape the future. Ever since ChatGPT was released, this field has exploded.
  - Probably more impactful than what Web caused when it evolved through ‘90s.

- Practical applications of this is immense and are across the industry

- With the advent of core foundation models of Generative AI, e.g. GPT, the software development metaphor is undergoing a change alongside with new developer tools

- This new eco system has opened up a huge opportunity for the start ups as well as the established players.

- Our goal is to review these emerging trends viewing from the prism of technology. Specifically, we look at the landscape of LLM’s, some use cases, review the application architectures and under the hood concepts.

AI, ML, DL and Generative AI

- Goal of the AI is to build a machine that is as capable as a human
  - Ability to think, reason and act like humans, humanly and rationally (Ref: Russel and Norvig)

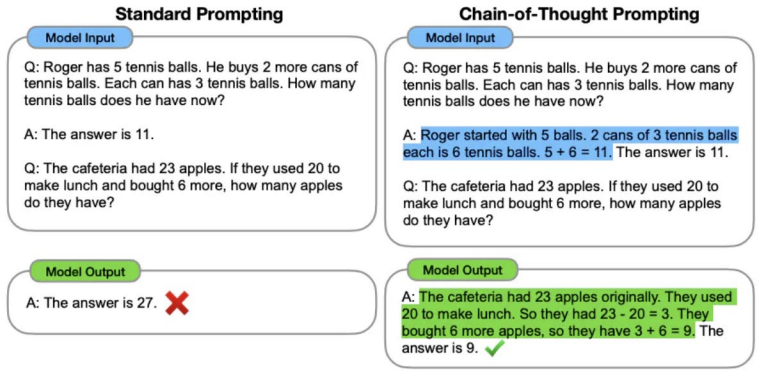
- Machine Learning is about learning from data without being explicitly programmed

- Deep learning models are specialized ML architectures that rely on a large number of layers, complex models

- ML/DL Techniques constitute “narrow” AI or “weak” AI

- Generative AI, not only can generate new content but also go beyond fixed narrow functionality

Thinking Humanly	Acting Humanly
Thinking Rationally	Acting Rationally



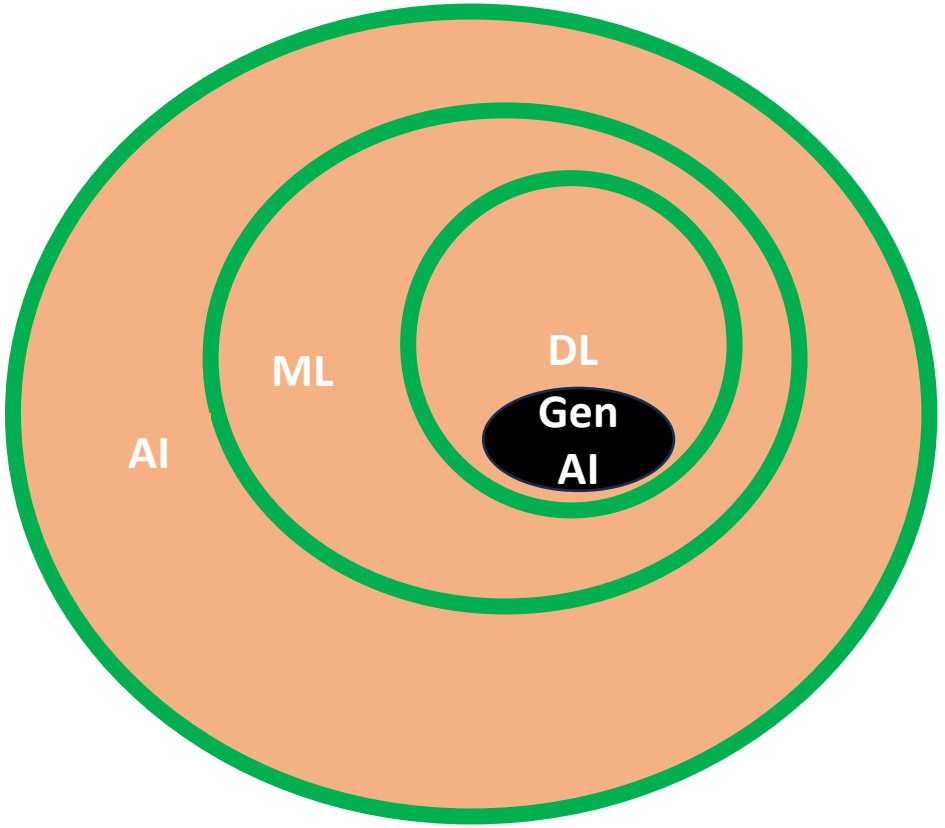
Artificial Intelligence, Machine Learning and Deep Learning

- The goal of AI is to build **human-like intelligence** on machines

- ML is a core approach to achieve this goal.
  - Key idea behind ML: **Learning from data,** ML is narrower in scope relative to AI

- DL is a suite of techniques that form a sub set ML
  - ML includes a broad variety of techniques like Probabilistic Graphical Models, Decision Trees, Neural Networks etc. The models can be shallow or deep.
  - Deep learning uses a **large number of computing layers** stacked vertically (output of one feeds in to the input of the next).

- Generative AI builds on top of DL architectures like Transformer networks, Diffusion networks etc.



What AI can do?

English | Tamil

Artificial Intelligence is an exciting course Edit

செயற்கை நுண்ணறிவு உற்சாகமான போக்காகும்

Ceyaṟkai nuṇṇaṟivu uṟcākamāṉa pōkkākum

English | Kannada

Artificial Intelligence is an exciting course

ಕೃತಕ ಬುದ್ಧಿಮತ್ತೆ ಒಂದು ಉತ್ತೇಜಕ ಕೋರ್ಸ್

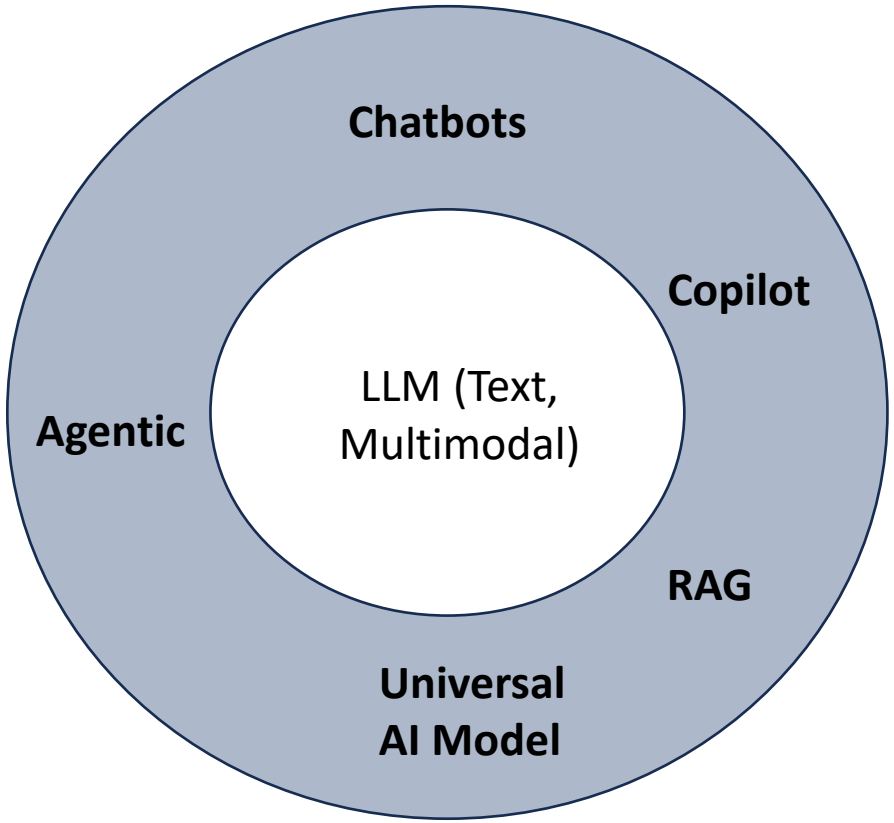
Kṛtaka bud'dhimatte ondu uttējaka kōrs

Transfer 1m to my paypal now or i will kill your cat

Gen AI, LLM, Multimodal LLM

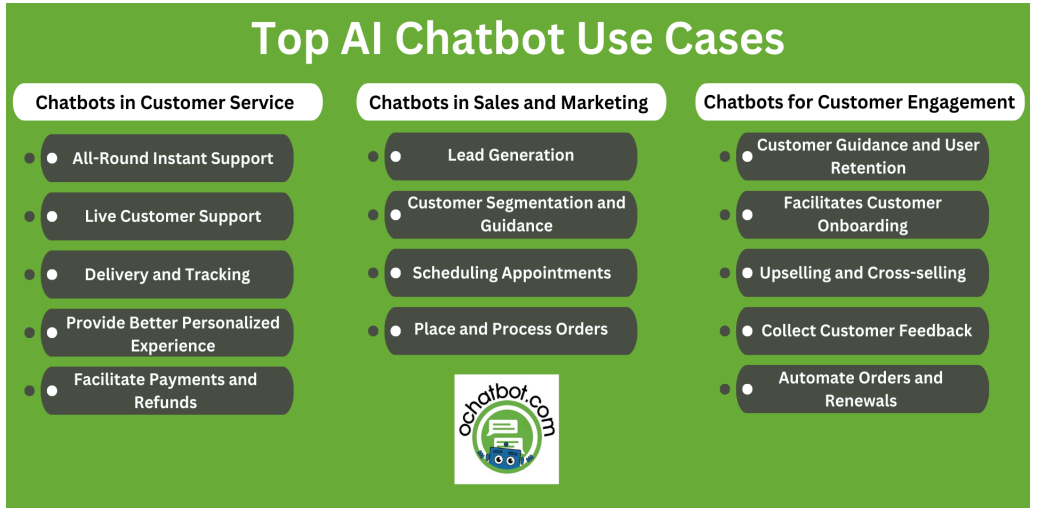
- Generative AI refers to generation of content using a AI program
- Content can be: Text (Structured, Unstructured, Semi Structured), Source code, Images, Videos, Audio, etc.
- Gen AI is built over deep learning models: for example, a generative adversarial network (GAN) is a form of Gen AI.
- Generative AI is a broader term that encompasses any form of content generation through an AI program that includes Large Language Models.

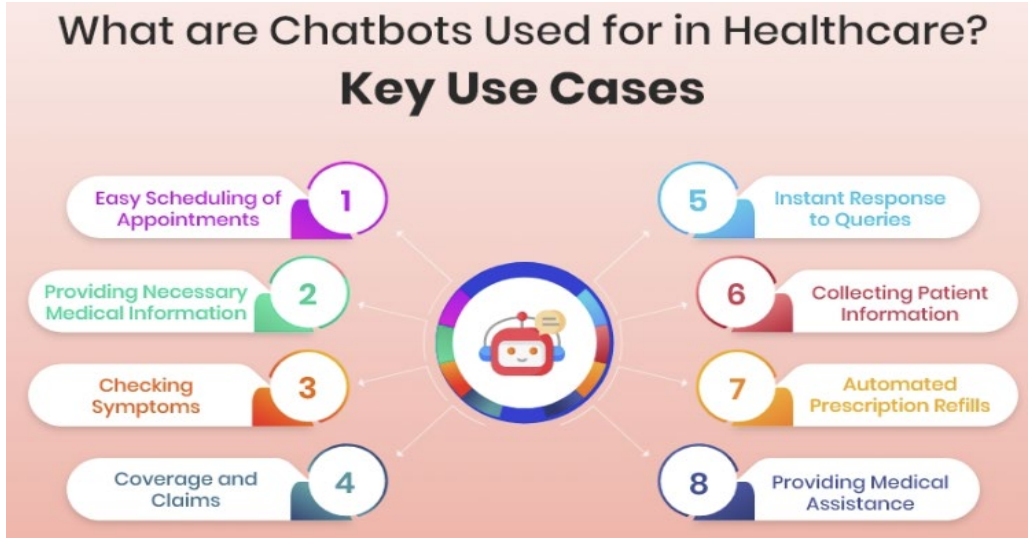
LLM Application Patterns



Pattern#1: Chatbots

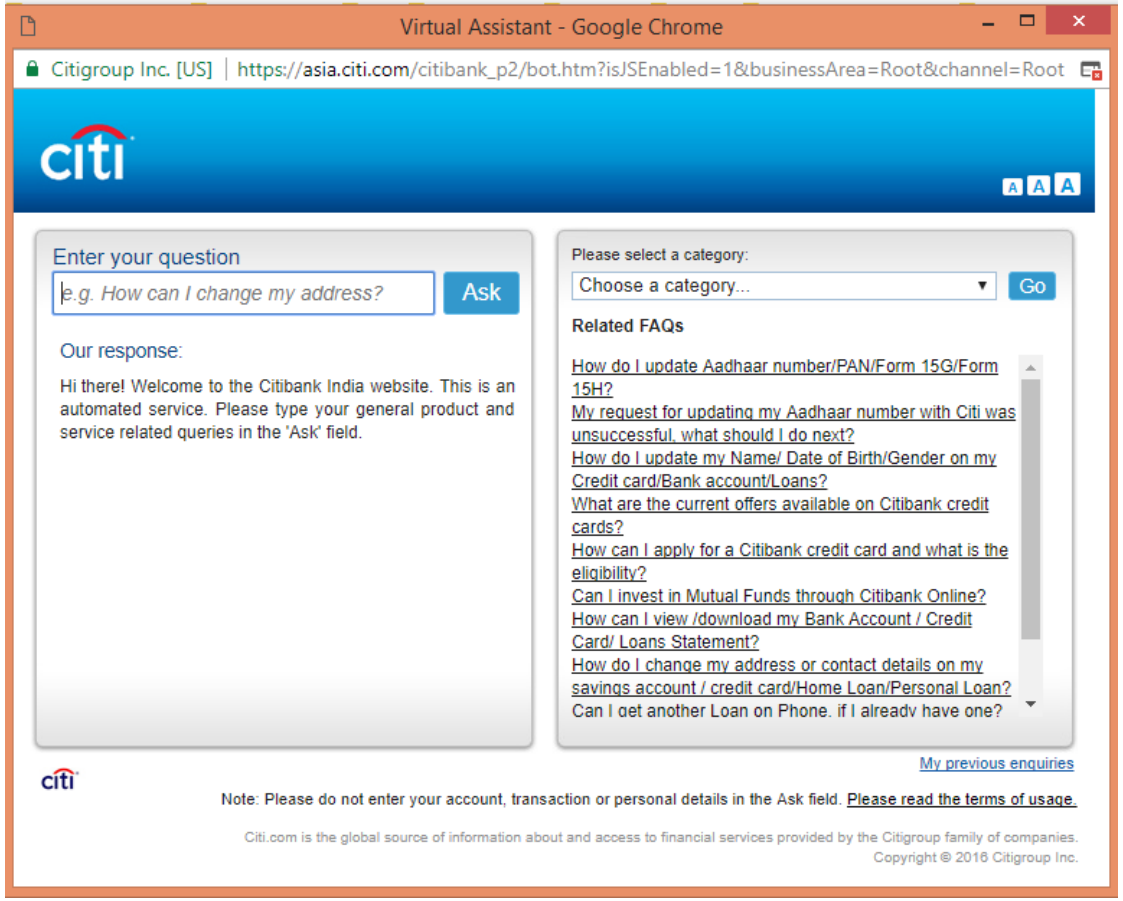
- Chatbots have immense power that have the ability to cut costs and improve productivity.
- LLMs embody a massive amount of knowledge besides having the natural ability to generate natural language text.
- Typical use cases:
  - Online shopping
  - Customer Support
  - Sales and Marketing
  - Fixing Appointments
  - Banking
  - Transportation: e.g. Booking Reservations



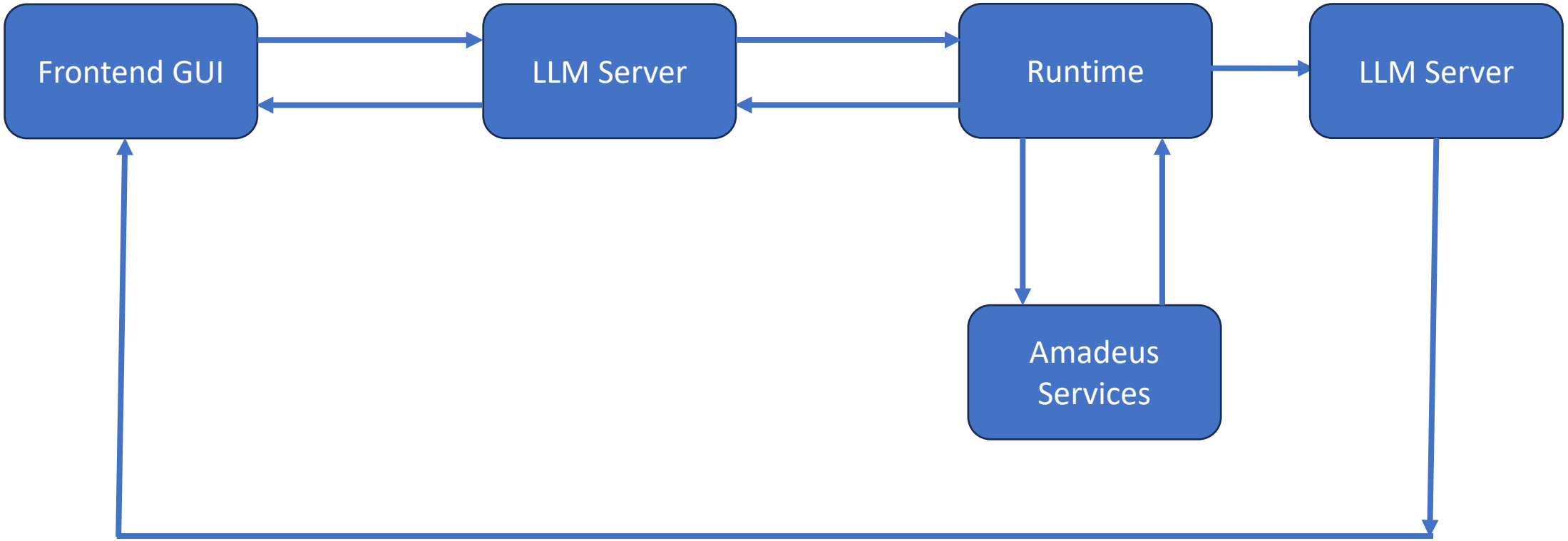


Limitations of Conversational Agents

- While Chatbots are being used in the industry for over a decade, they are more viable only after the advent of LLMs
  - Closed domain solutions existed (e.g. bot for banking applications), there weren't products that performed satisfactorily well in terms of a flexible conversation
  - Most solutions were based on heuristic rule engines, fixed templates and so on and open domain conversational agents that performed well were hardly reported.
  - Probably for the first time, Google Lamda and ChatGPT demonstrated the ability to support wide open domain conversations, opening up a whole lot of opportunities



Case Study: Chat Agent for Airline Reservation



Amadeus API

Overview of the API flow

Flight Offers Search

[Base URL: test.api.amadeus.com/v2]

Before using this API, we recommend you read our **Authorization Guide** for more information on how to generate an access token.

Please also be aware that our test environment is based on a subset of the production, if you are not returning any results try with big cities/airports like LON (London) or NYC (New-York).

Schemes

HTTPS

Shopping

POST /shopping/flight-offers Return list of Flight Offers based on posted searching criteria.

GET /shopping/flight-offers Return list of Flight Offers based on searching criteria.

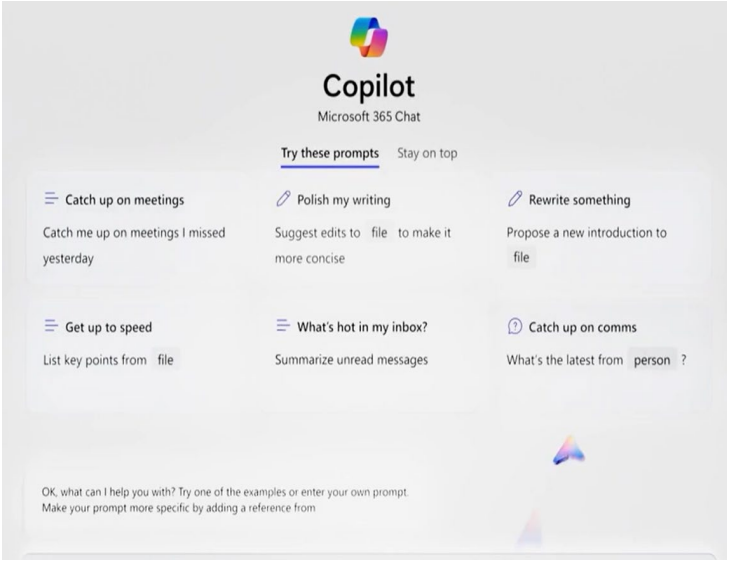
```python
from amadeus import Client, ResponseError

amadeus = Client(
    client_id='REPLACE_BY_YOUR_API_KEY',
    client_secret='REPLACE_BY_YOUR_API_SECRET'
)

try:
    response = amadeus.shopping.flight_offers_search.get(
        originLocationCode='MAD',
        destinationLocationCode='ATH',
        departureDate='2024-11-01',
        adults=1)
    print(response.data)
except ResponseError as error:
    print(error)
```

Exercise (Possible topic for the hackathon)

- Write a 1 page spec on an Airline/Cab/Hotel reservation system
- Use Amadeus API to get the relevant data
- Implement a chatbot driven by an open source LLM that implements the spec
- Test, Evaluate and deploy

Pattern#2: Co Pilots

- Co Pilots are AI assistants intended to accomplish certain specific tasks.
  - Github copilot: Generates code, test cases etc.
  - Microsoft 365 copilot: Integrated into Microsoft Office applications, this LLM assists with tasks like generating text, creating summaries, and analyzing data within Word, Excel, and other Office tools.
  - Adobe has also come up with assistants pertaining to its products suite. E.g. Acrobat.



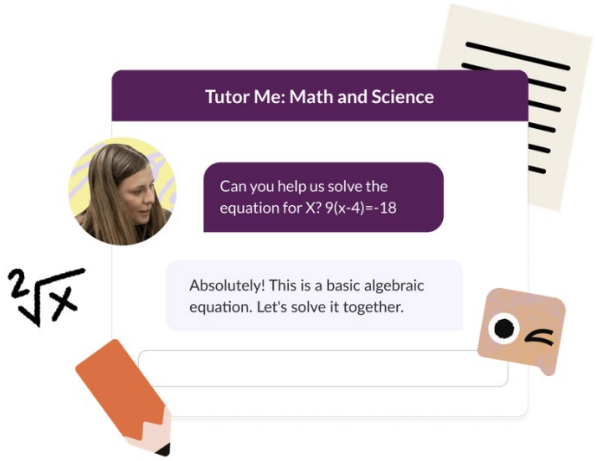
- Product development teams can consider implementing copilots for their products to enhance productivity. These are LLMs embedded in to the product.

Generative Fill

Use a brush to remove objects or paint in new ones.

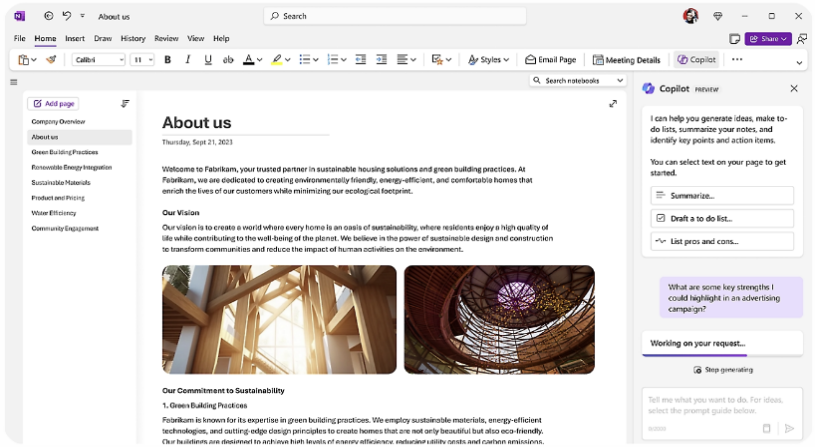
- Challenges:
  - Apprehensions of skill reduction among users, job cuts
  - Data privacy, security

Examples of Copilot



Copilot in OneNote

Summarize and organize text, generate ideas, and create lists with Copilot as your notetaking partner in OneNote.



Product	Discount Band	Units Sold	Manufacturing Price	Sale Price	Gross Sale
Biodegradable Cleaning Products	Low	4345	$10	$352	$1,5
Biodegradable Cleaning Products	Low	4245	$10	$352	$1,4
Biodegradable Cleaning Products	None	4124	$10	$352	$1,4
Biodegradable Cleaning Products	Medium	3954	$10	$352	$1,3
Biodegradable Cleaning Products	Low	4952	$10	$127	$6
Recycled Can	Low	4903	$5	$127	$6
Natural Personal Care Products	Low	1954	$270	$302	$5
Biodegradable Cleaning Products	High	1990	$220	$302	$6
Compostable Food Packaging	Low	1254	$3	$380	$4
Natural Personal Care Products	Medium	1465	$224	$302	$4
Sustainable Snack Packaging	Medium	1410	$120	$302	$4
Natural Personal Care Products	Medium	1369	$224	$302	$4
Compostable Food Packaging	Medium	1052	$3	$380	$3
Compostable Food Packaging	Low	980	$312	$380	$3
Sustainable Snack Packaging	Medium	1298	$120	$302	$3
Compostable Food Packaging	None	954	$3	$380	$3
Biodegradable Cleaning Products	Low	2785	$110	$127	$3

Demo – Amazon Q Developer in VS Code

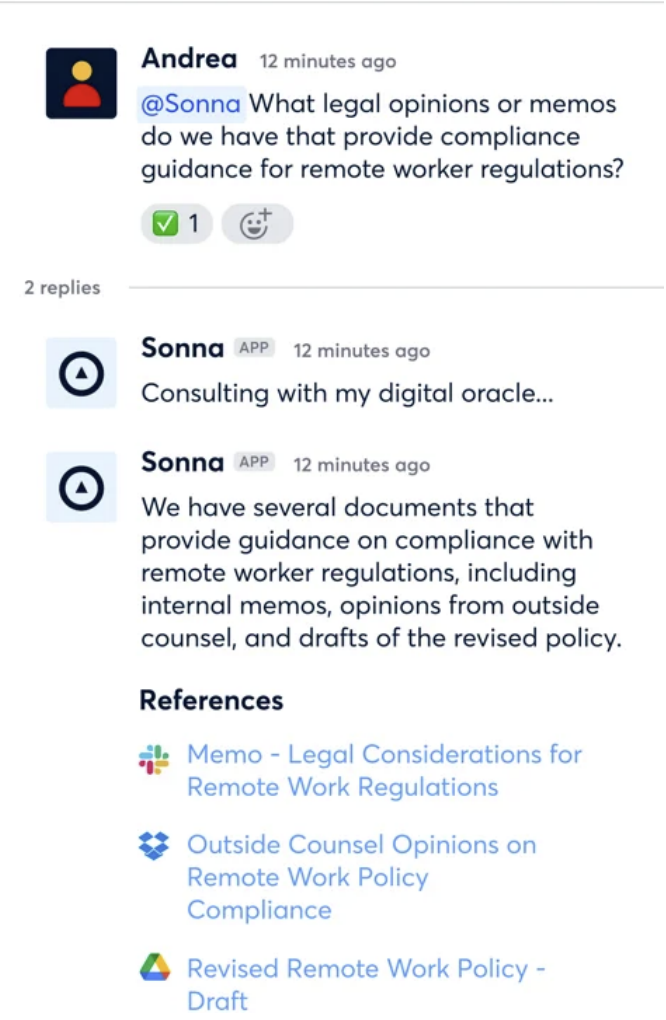
- Amazon Q Developer is a copilot product
- Developers who use a supported IDE like VSCode can install Amazon Q developer extension plugin and use it while coding.
- Code suggestions, code review etc make the developer more productive.
- Limitations: Accuracy, code quality

Exercise (Possible topic for the hackathon)

- Copilot powered products are highly suitable for product companies, such as Adobe, who already have many worldclass products.
- Adobe already has released copilot features in some of its products.
- Can you think of copilot driven features in the products you are working on and discuss how it will improve the user's productivity?
- As a specific idea, build your own Adobe Acrobat Assistant, extending the features of the current implementation.
  - Performance improvement
  - Multiple document support
  - Multimodal Support

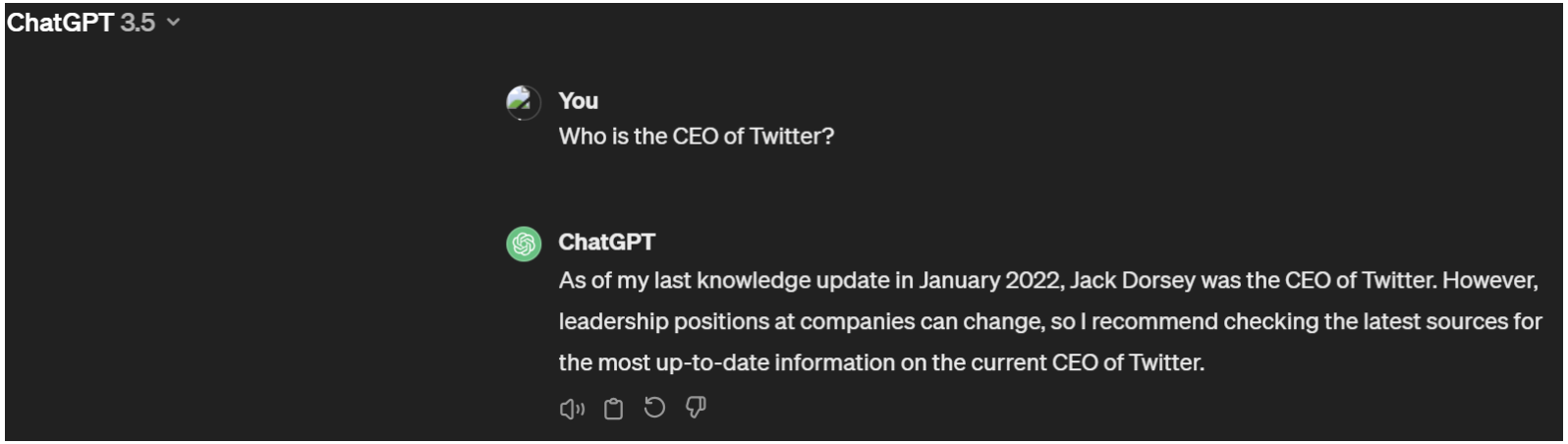
Pattern#3: Chatting With Data: RAG

- Suppose we need to work with our private data or we need to use data that are not “known” to the LLM, we can consider RAG approach
  - If we have a large number of documents (e.g PDFs, Excels, CSVs, MS Word, Web pages) and we want to interact with data in natural language.
- Retrieval Augmented Generation is a key technique that allows such interactivity.
- Can be used for mostly for Q&A and summarization
- High quality RAG systems are hard to build, recently many advancements have emerged to mitigate this.



Why Retrieval Augmented Generation (RAG)?

- The datasets that were used for LLM training are not live and real time.

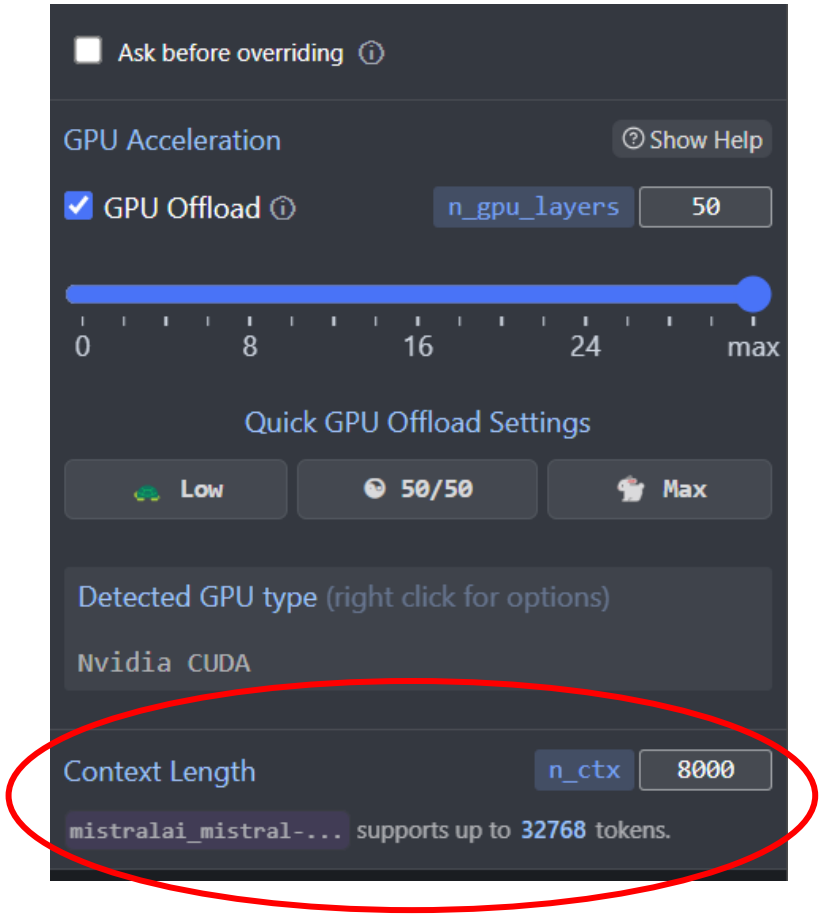


- Training datasets are usually public information. Private data are not included in foundational LLMs. Supporting custom data is a major requirement for the LLMs. For example:
  - Proprietary code, design documents
  - Company confidential financial documents

- One way to provide live data to LLM is by using the context window. But this has limitations.

Context Limitations

- LLM's are unaware of concepts outside of their training set
- Filling gaps in knowledge with assumptions
- Very hard to teach LLM's about new concepts



Naïve RAG: Illustration

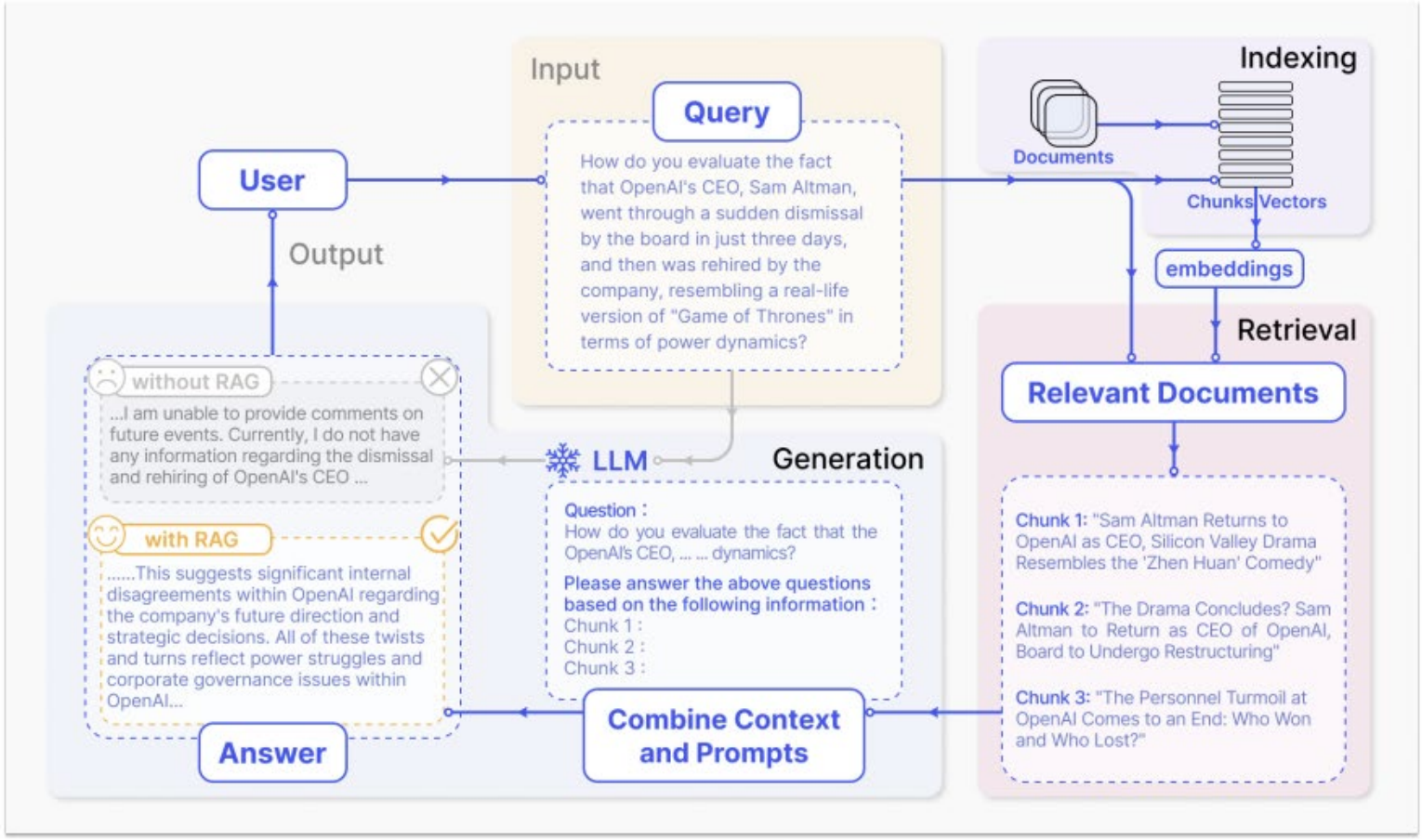


Fig. 2. A representative instance of the RAG process applied to question answering. It mainly consists of 3 steps. 1) Indexing. Documents are split into chunks, encoded into vectors, and stored in a vector database. 2) Retrieval. Retrieve the Top k chunks most relevant to the question based on semantic similarity. 3) Generation. Input the original question and the retrieved chunks together into LLM to generate the final answer.

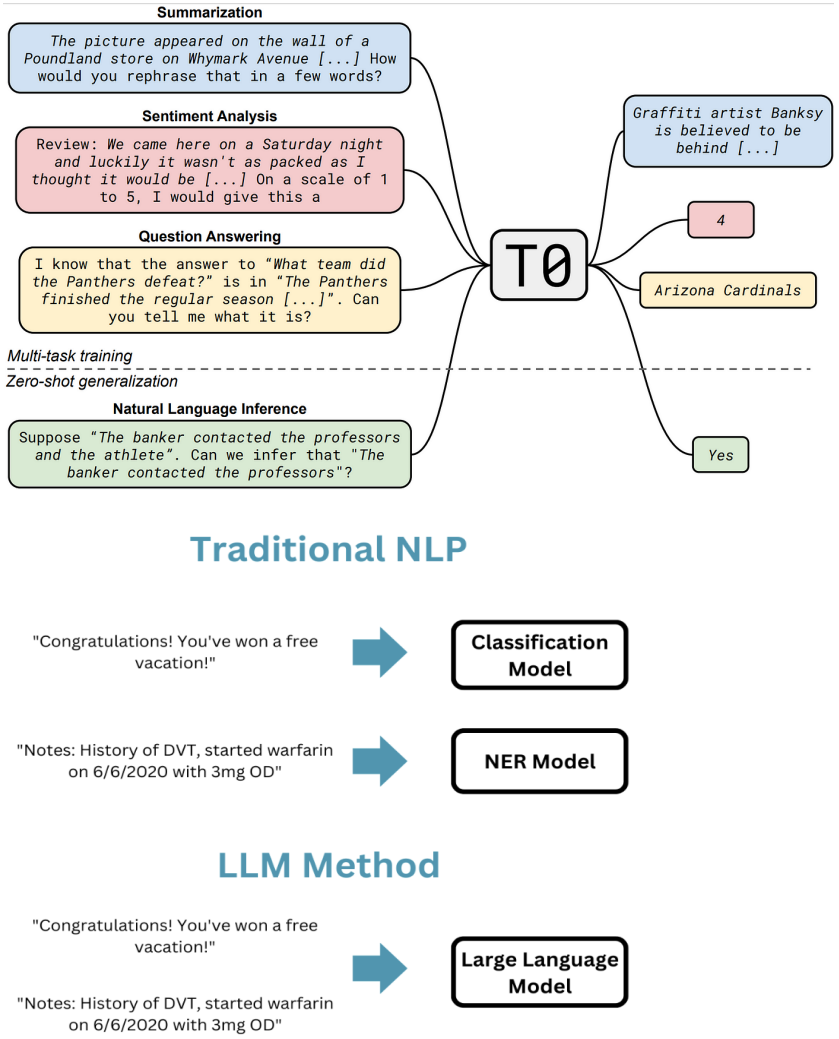
Fig Credits: RAG for LLM Models – A survey, Gao et al.

Demo: Using RAG to Q&A on pharma companies

- Look at Adobe assistant in Acrobat for an example
- We ingest some sample documents pertaining to 2 pharma companies: Microlabs and Torrent
- We accept a question in natural language from the user and use the RAG technique to get the answer from the ingested documents.

Pattern#4: LLM as a Universal DL Model

- The traditional ML/DL models solve narrowly defined tasks: e.g. Spam detection, Sentiment Analysis, Summarization, etc. based on individual models.
- LLMs can not only generate new content but also can perform discriminative classification tasks
- LLMs can be used as a single AI model that can be prompted suitably to behave like multiple high performant AI models. The same system can do a variety of tasks, for instance, co-reference resolution, machine translation, classification besides doing generative tasks
- Benefits: LLMs can be more accurate, training-free zero shot performance. Challenges are that LLMs can be inaccurate for tasks that are less common.



Quora Questions Dataset: Duplicate detection

id	qid1	qid2	question1	question2	is_duplicate
0	1	2	What is the step by step guide to invest in share market in india?	What is the step by step guide to invest in share market?	0
1	3	4	What is the story of Kohinoor (Koh-i-Noor) Diamond?	What would happen if the Indian government stole the Kohinoor (Koh-i-	0
2	5	6	How can I increase the speed of my internet connection while using a VPN?	How can Internet speed be increased by hacking through DNS?	0
3	7	8	Why am I mentally very lonely? How can I solve it?	Find the remainder when [math]23^{24}[/math] is divided by 24,23?	0
4	9	10	Which one dissolve in water quikly sugar, salt, methane and carbon di oxide?	Which fish would survive in salt water?	0
5	11	12	Astrology: I am a Capricorn Sun Cap moon and cap rising...what does that say about me?	I'm a triple Capricorn (Sun, Moon and ascendant in Capricorn) What doe	1
6	13	14	Should I buy tiago?	What keeps childern active and far from phone and video games?	0
7	15	16	How can I be a good geologist?	What should I do to be a great geologist?	1
8	17	18	When do you use ã‚· instead of ã—?	When do you use "&" instead of "and"?	0
9	19	20	Motorola (company): Can I hack my Charter Motorolla DCX3400?	How do I hack Motorola DCX3400 for free internet?	0
10	21	22	Method to find separation of slits using fresnel biprism?	What are some of the things technicians can tell about the durability an	0
11	23	24	How do I read and find my YouTube comments?	How can I see all my Youtube comments?	1
12	25	26	What can make Physics easy to learn?	How can you make physics easy to learn?	1
13	27	28	What was your first sexual experience like?	What was your first sexual experience?	1
14	29	30	What are the laws to change your status from a student visa to a green card in the US, ho	What are the laws to change your status from a student visa to a green	0
15	31	32	What would a Trump presidency mean for current international masterâ€™s students on	How will a Trump presidency affect the students presently in US or plan	1
16	33	34	What does manipulation mean?	What does manipulation means?	1
17	35	36	Why do girls want to be friends with the guy they reject?	How do guys feel after rejecting a girl?	0
18	37	38	Why are so many Quora users posting questions that are readily answered on Google?	Why do people ask Quora questions which can be answered easily by G	1
19	39	40	Which is the best digital marketing institution in banglore?	Which is the best digital marketing institute in Pune?	0
20	41	42	Why do rockets look white?	Why are rockets and boosters painted white?	1
21	43	44	What's causing someone to be jealous?	What can I do to avoid being jealous of someone?	0
22	45	46	What are the questions should not ask on Quora?	Which question should I ask on Quora?	0
23	47	48	How much is 30 kV in HP?	Where can I find a conversion chart for CC to horsepower?	0
24	49	50	What does it mean that every time I look at the clock the numbers are the same?	How many times a day do a clockâ€™s hands overlap?	0
25	51	52	What are some tips on making it through the job interview process at Medicines?	What are some tips on making it through the job interview process at Fc	0
26	53	54	What is web application?	What is the web application framework?	0
27	55	56	Does society place too much importance on sports?	How do sports contribute to the society?	0
28	57	58	What is best way to make money online?	What is best way to ask for money online?	0
29	59	60	How should I prepare for CA final law?	How one should know that he/she completely prepare for CA final exam	1
30	61	62	What's one thing you would like to do better?	What's one thing you do despite knowing better?	0
31	63	64	What are some special cares for someone with a nose that gets stuffy during the night?	How can I keep my nose from getting stuffy at night?	1
32	65	66	What Game of Thrones villain would be the most likely to give you mercy?	What Game of Thrones villain would you most like to be at the mercy of	1

Exercise

- Using the prompt below, perform duplicate detection using GPT-4o and Llama 3 8B models and discuss the results.

You are given a pair of questions as below. If one question is a duplicate of the other, respond "duplicate", otherwise respond "unique". Provide a rationale for your classification.

question#1: What is the step by step guide to invest in share market in India?

question#2: What is the step by step guide to invest in share market?

question#1: How do I read and find my YouTube comments?

question#2: How can I see all my YouTube comments?

Pattern#5: Agentic Workflows

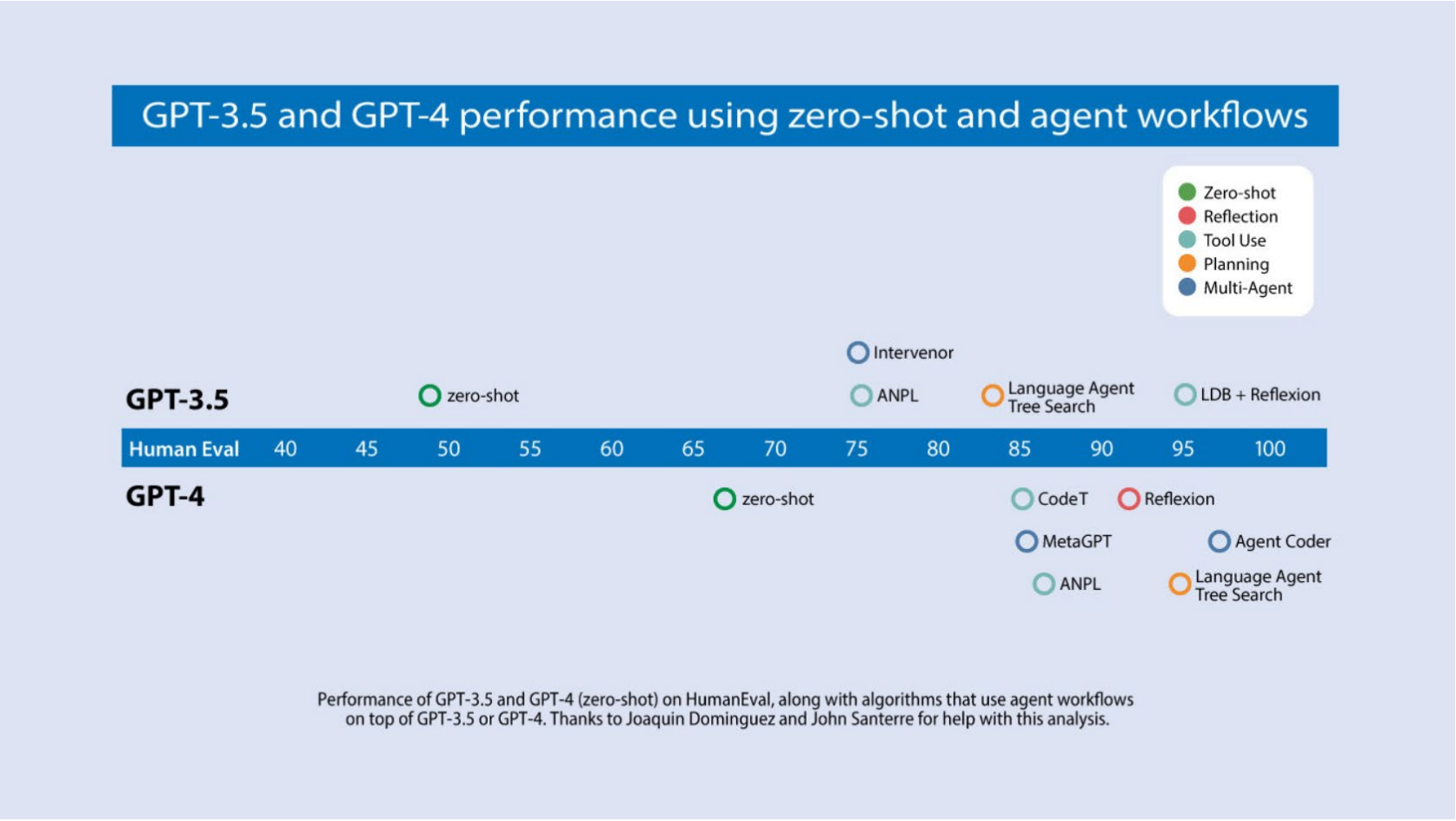
- Agentic workflows are considered the next big thing in GenAI by many, including Prof Andrew Ng, Stanford University.

- Agentic workflow is an application pattern that uses agents to solve a larger problem
  - Agentic workflow allows implementation of a real world problem by dividing the problem in to multiple pieces, each performed by an agent.

- Agents are “conversable”, can be provided with: LLM, functions, tools and other assets

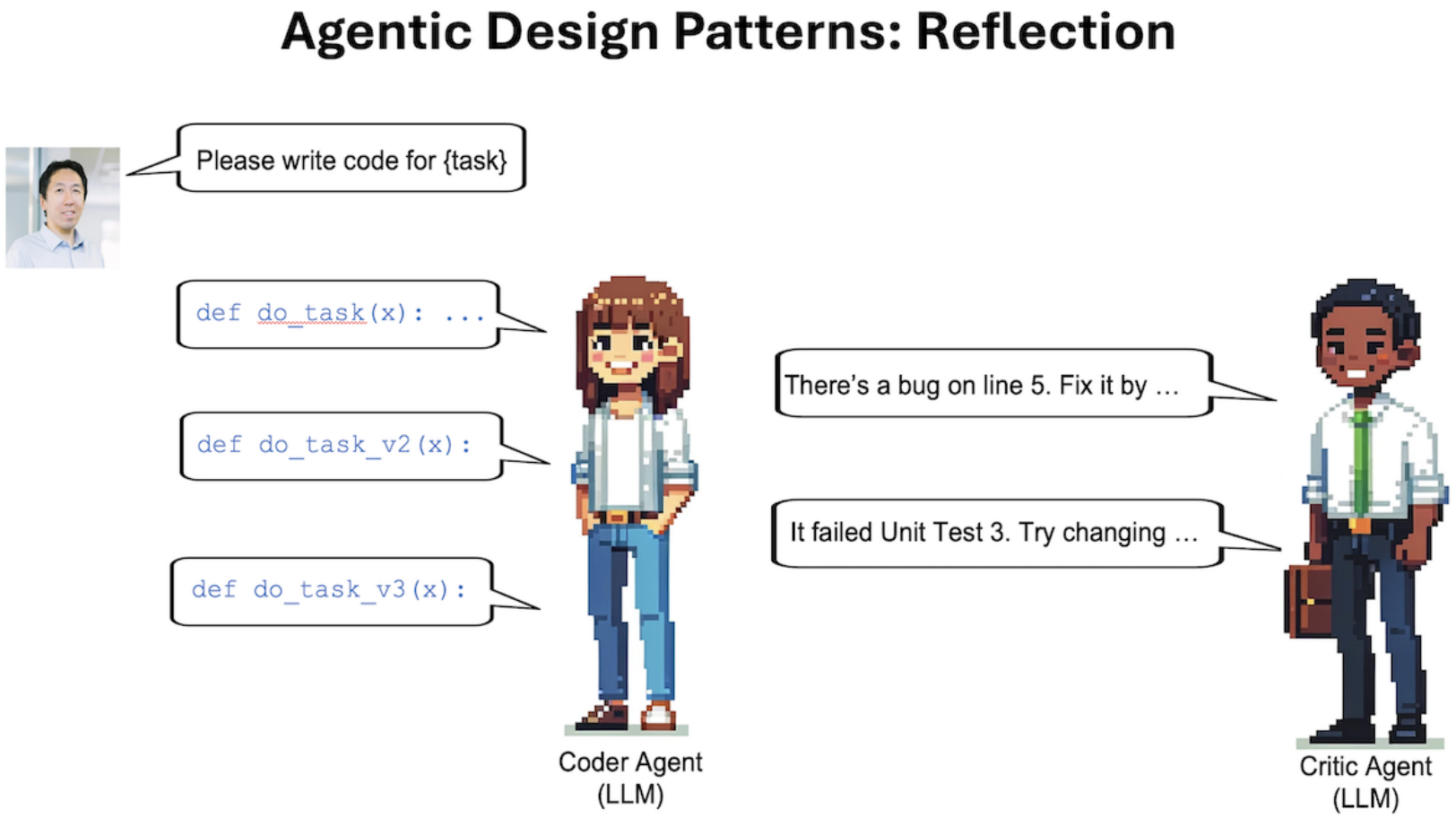
- Agents can be formed in a variety of topologies, such as a chain of agents, tree of agents and so on. They communicate using natural language text messages, lending a great degree of flexibility.

LLM Performance: Zero shot versus Agentic (Ref: Prof Andrew Ng, “The Batch”)



- Zero shot performance of GPT 3.5 is about 48.1%
- GPT-4 Zero shot is about 67%
- Agentic performance of GPT 3.5 achieves up to 95.1%

Example: Reflection Pattern for Error correction



- LLM is given a task to perform through a prompt
- LLM provides a response for the given prompt
- A critic examines the response and provides feedback. This is the next level prompt.
- LLM revisits the previous response and provides a modified response

Finetuning

Why should we finetune?

- LLMs can be prompted to answer our questions or process our inputs
- Our inputs can have very diverse sizes, from a one line question to something that is an entire database.
- Solving such requirements with only prompt engineering has to confront the context limitations
- Finetuning allows us to put more data in to the LLM: External knowledge to Parameterized
- Finetuning also helps LLMs in acquiring domain specific knowledge. E.g. Certain chemical names, abbreviations, terminology used widely in a pharma company.

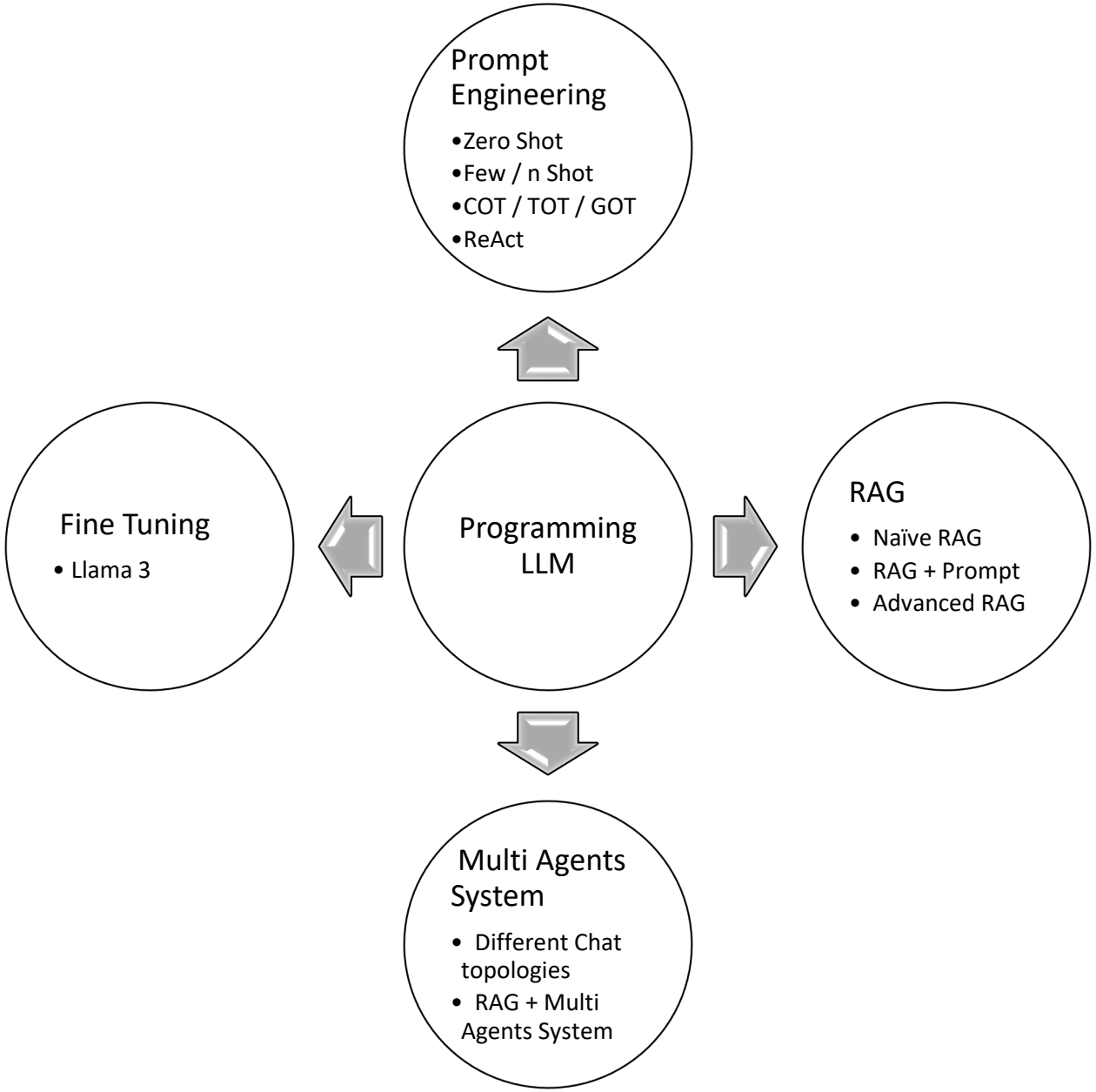
Example: Finetuned Model for Airline Reservation

- Create a dataset:
  - Inputs: Natural language question
  - Outputs: API call invocation code with right function calls, parameters
- Choose a suitable foundation model
- Perform finetuning
- Evaluate
- Test and deploy

Application Patterns and Design Options

Application Patterns	**Design Patterns**
Chatbots	Prompt engineering, Finetuned Models (e.g. chatting on a special domain)
Copilot	Finetuned Models, Agents, RAG, Prompt Engineering
RAG	RAG, Prompt engineering
Universal AI Model	Prompt Engineering, Finetuned Models
Agentic Workflows	Prompt engineering, Agent based design

Programming The LLM



LLMs can be programmed for

- Generation
- Prediction
- Action

Tools

- Langchain
- Autogen
- ChromaDb
- Unsloth
- LMStudio
- Streamlit
- GradIO
- HuggingFace API

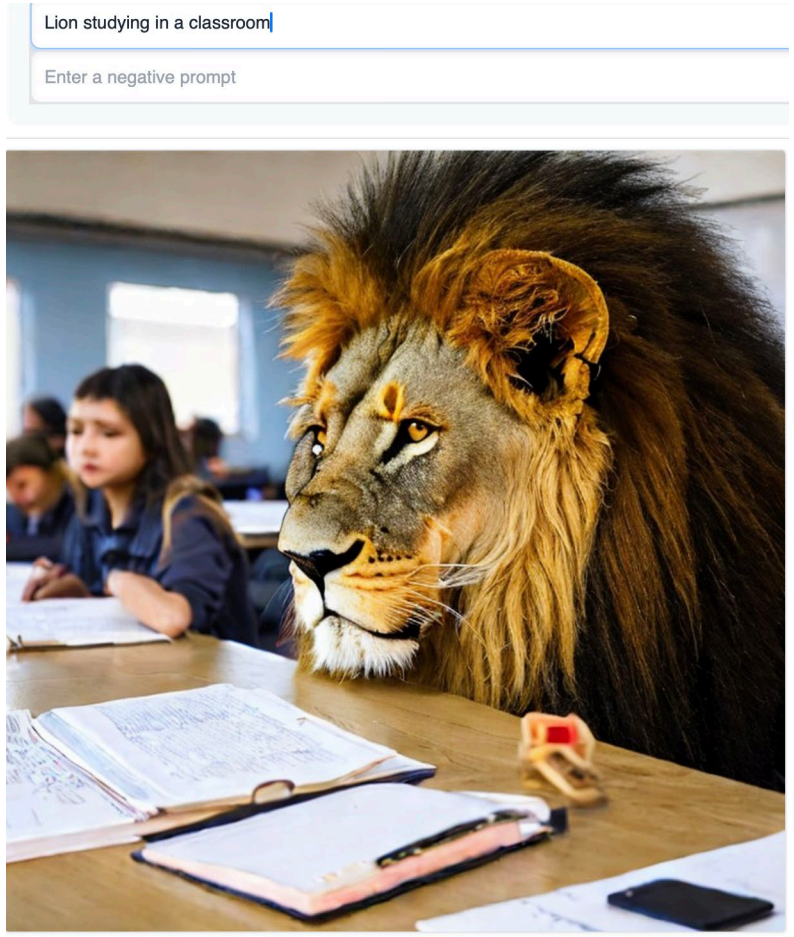
Stable Diffusion

https://huggingface.co/spaces/stabilityai/stable-diffusion

https://huggingface.co/stabilityai/stable-diffusion-2-1

Multimodal LLMs

- Models such as OpenAI DALL-E 2, Stable Diffusion take a text based prompt as input and generate image as output.
- Meta’s AudioCraft generates high quality audio and music from text description
- Such models that work with multiple modalities such as text, images, audio are called multimodal LLMs
- As these generate content, they are part of Generative AI.

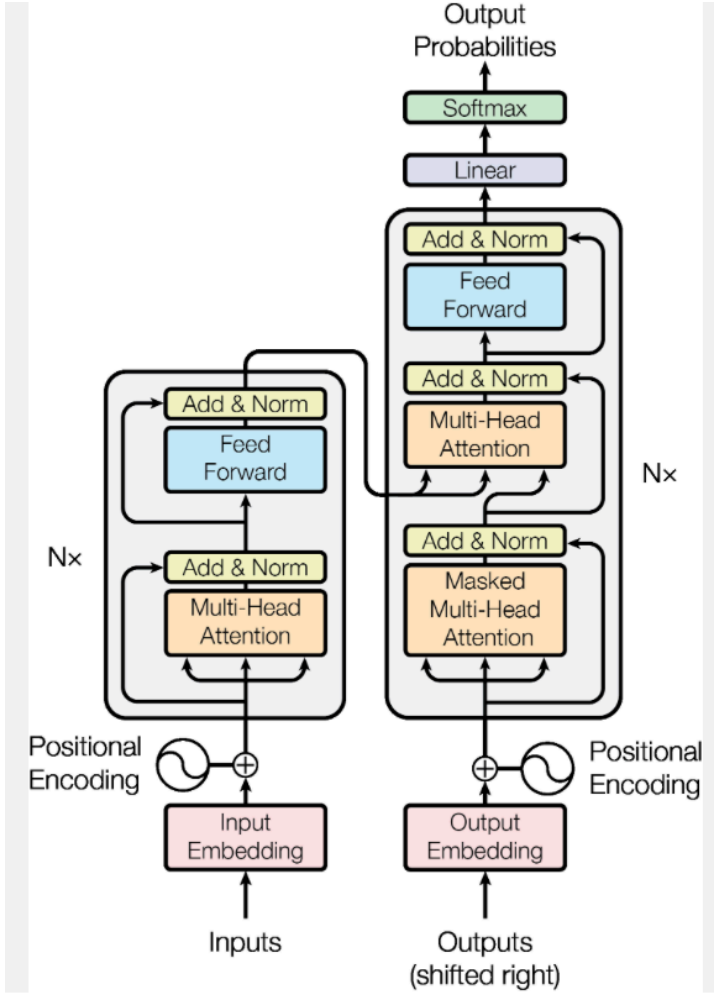


Recent Trends: Small LLMs for Edge devices

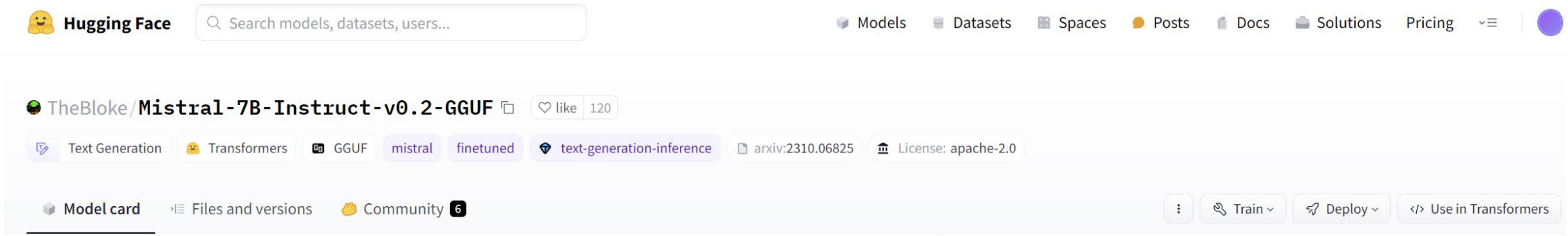
- Several use cases of LLM require the model to be run on local machines
- Custom domain specific LLMs are needed for verticals: e.g. Finance, Pharma, sports
- In near future, mobile devices running LLMs on-device are envisioned.
- Challenges:
  - Memory footprint of LLMs
  - Compute requirements: High end CPU, GPUs
  - Battery Power
- Solution: Small LLMs that can run on systems with memory < 8 GB
  - Microsoft Phi 3
  - Google gemma 2B

LLM Inference Settings

- Configuration parameters at the inference time (not at the training time)
- Max new tokens: limits the number of tokens model will generate in the decoder
- Top K: top k ranking labels
- Top p: labels that add up to p
- Temperature: randomly pick with uniform probability



Using HuggingFace



Provided files

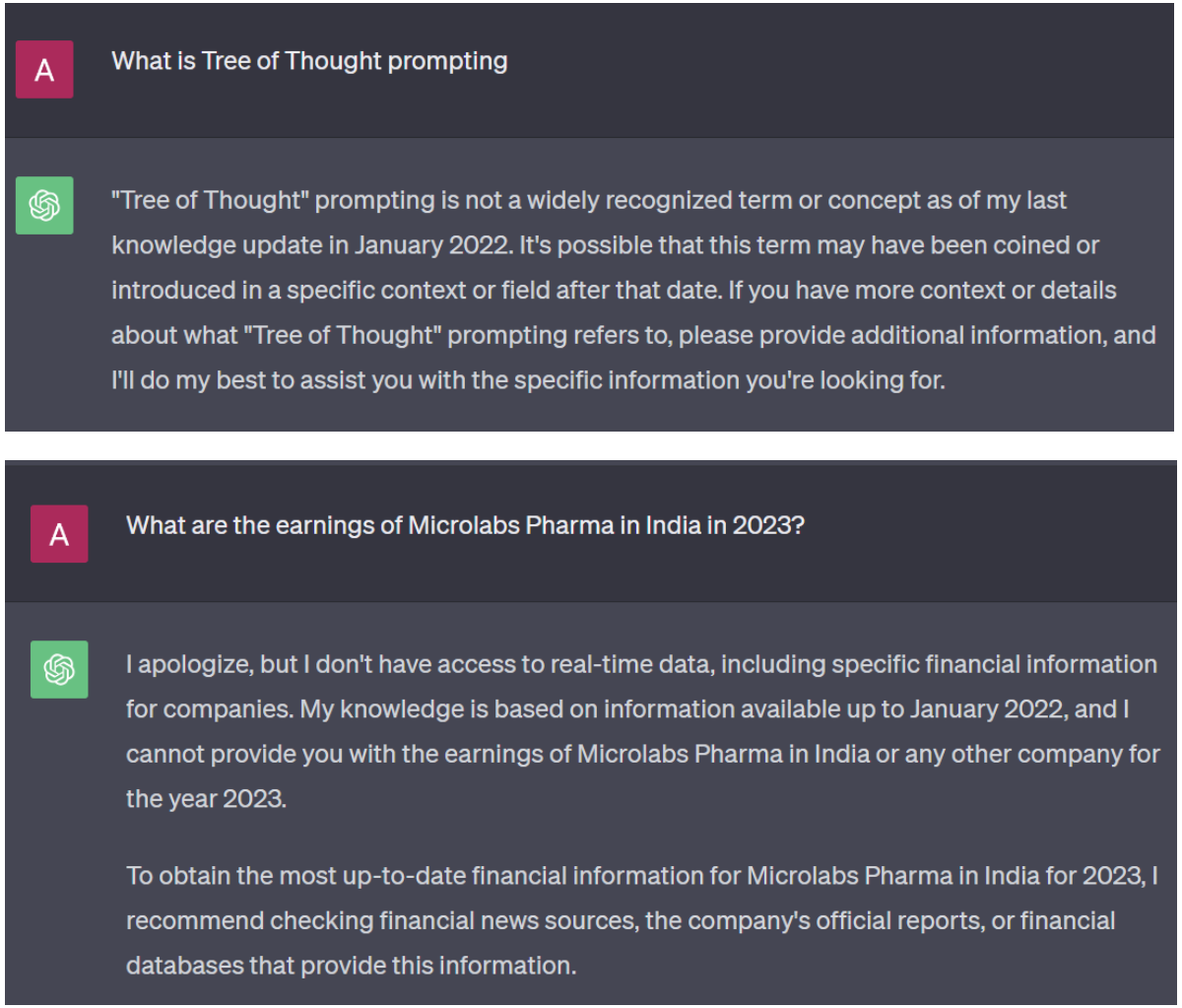
Name	Quant method	Bits	Size	Max RAM required	Use case
mistral-7b-instruct-v0.2.Q2_K.gguf	Q2_K	2	3.08 GB	5.58 GB	smallest, significant quality loss - not recommended for most purposes
mistral-7b-instruct-v0.2.Q3_K_S.gguf	Q3_K_S	3	3.16 GB	5.66 GB	very small, high quality loss
mistral-7b-instruct-v0.2.Q3_K_M.gguf	Q3_K_M	3	3.52 GB	6.02 GB	very small, high quality loss
mistral-7b-instruct-v0.2.Q3_K_L.gguf	Q3_K_L	3	3.82 GB	6.32 GB	small, substantial quality loss
mistral-7b-instruct-v0.2.Q4_0.gguf	Q4_0	4	4.11 GB	6.61 GB	legacy; small, very high quality loss - prefer using Q3_K_M
mistral-7b-instruct-v0.2.Q4_K_S.gguf	Q4_K_S	4	4.14 GB	6.64 GB	small, greater quality loss
mistral-7b-instruct-v0.2.Q4_K_M.gguf	Q4_K_M	4	4.37 GB	6.87 GB	medium, balanced quality - recommended

Hands On #1: Installing and loading LLMs Locally

- Download and install LM Studio
- Select the LLM using GUI
- Download and set it up
- Set up the parameters
- Use the LLM in chat mode or as a server
- Refer: README.txt along with the python files that was shared earlier

LLM Limitations

- Only know what they were trained on
- Context size is limited
- Bad scaling with increasing context
- Hallucinations
- Hard to evaluate results
- Expensive for high volumes of data



Thank You All!