Carnegie Mellon University

14-825 Generative AI and Large Language Models

SPRING 2024

MOHAMED FARAG

FARAG@CMU.EDU

Agenda

- Welcome and Introductions
- Focus Areas of this Course
- Course Syllabus & Schedule
- Class Expectations
- Introduction to Generative AI
- Next Steps

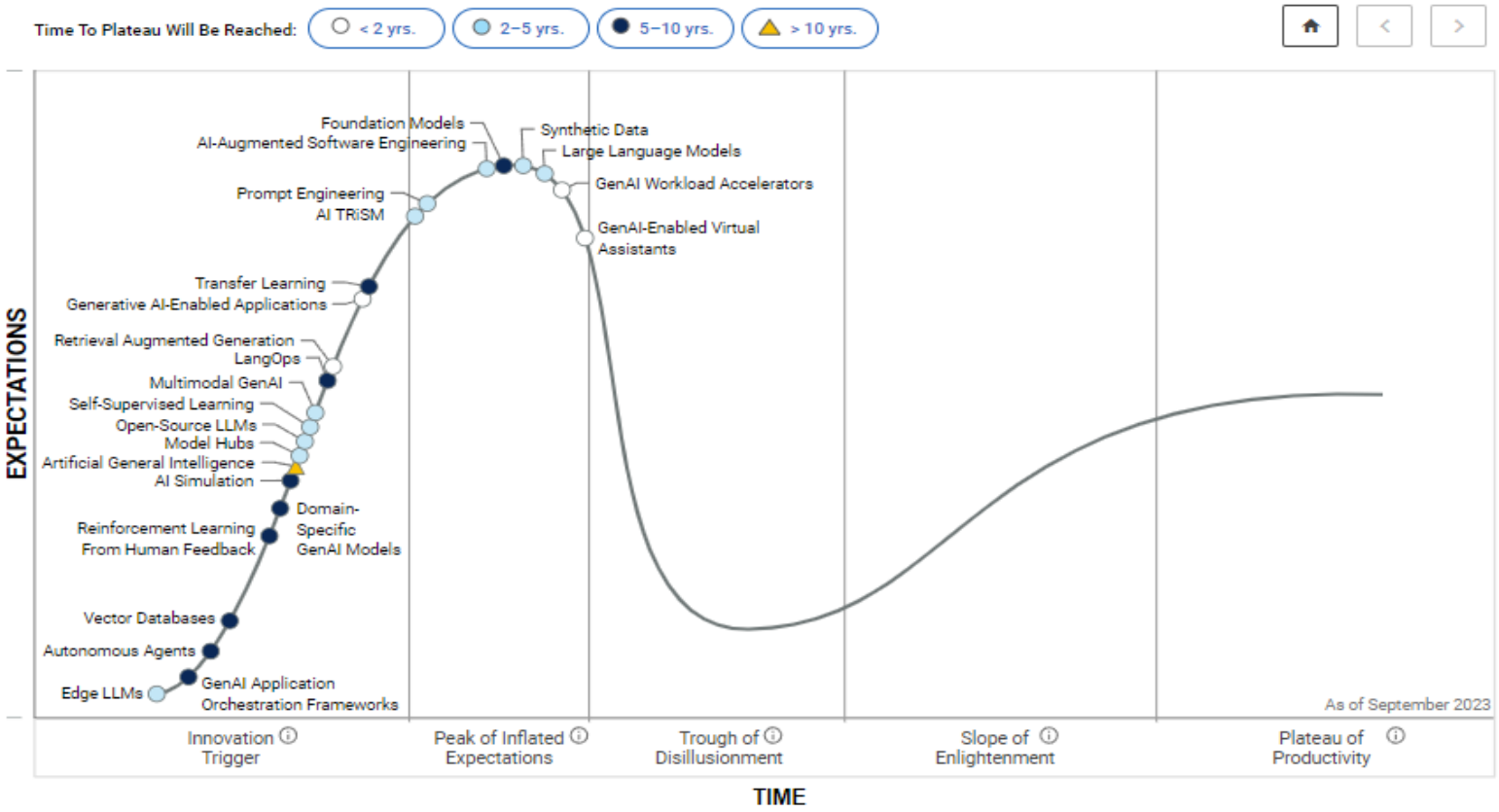
Why is this course Important?

Over the past yea, the demand for Generative AI and Large Language Models jumped significantly.

- **“With the influx of consumer generative AI programs like Google’s Bard and OpenAI’s ChatGPT, the generative AI market is poised to explode, growing to $1.3 trillion over the next 10 years from a market size of just $40 billion in 2022”, Bloomberg Intelligence.**
- **A survey of 1,400 US business leaders by the Upwork Research Institute found that 64% of C-suite respondents plan to hire more professionals across every job title because of generative AI technology.**

Why is this course Important? – Cont'd

The course touches on several topics on the Generative AI Hype Cycle, 2023 released by Gartner R&D



Carnegie Mellon University

Main Directions in this Course

- Generative AI in nowadays industries
- LLM types, architectures, and internal structures
- LLM integration with other systems
- LLMs on cloud platforms

Expectations for Incoming Students

- ***You are expected to know Python or are willing to learn it.***
    - A Python recording will be released next week for members who need support with Python
- ***You are expected to have a basic understanding of Artificial Intelligence***
    - Reach out to me if you need some introductory materials to AI

Instructor and TA Introductions

Instructor:

- Mohamed Farag: farag@cmu.edu

TAs:

- Jun (Roy) Xing: junxing@andrew.cmu.edu
- Yiwen Xiang: yiwenxia@andrew.cmu.edu

Course Logistics

- Lectures are offered in-person only, but recordings will be made available after the lectures.
    - Please allow for some lead-time in the beginning of the semester for the recordings to be released.
- Lecture slides are delivered via TopHat during the lecture and will be posted on Canvas under Modules section. Sign up for a free TopHat account and join the course with the following code: **019069**
- Starting from next lecture, each lecture will have an in-class quiz.
- Students who have approved accommodation shall contact the course instructor to figure out how the instructor can meet their needs

Course Logistics – Office Hours

Days/Timeframes	11am-12pm ET	11:30am-12:30pm ET	12-1pm ET	2:30-3:30pm ET	6-7pm ET	7-9pm ET	9-10pm ET
Monday		Mohamed					
Tuesday					Yiwen		
Wednesday	Roy						Roy
Thursday	Roy					Yiwen	Roy (by Appointment only via Calendly)
Friday			Yiwen				
Saturday							
	Instructor Office Hours - Conducted remotely via Zoom - URL can be found on Canvas						
	TA Office Hours - Conducted remotely via Zoom - URL can be found on Canvas						

- All Office Hours will use the same Zoom URL: https://cmu.zoom.us/j/95568893765?pwd=M3ZaTzdHS2RNNDN6Z1BwR1ZmRENPZz09
- If you have short questions and you don't want to wait in the Zoom room for extended time, please book a 15-min discussion via this URL (limited slots): https://calendly.com/junxing-814/14825_oh-by-appointment

Course Logistics – Piazza Hours

Piazza OHs								
Days/Timeframes	11am-12pm ET	11:30am-12:30pm ET	12-1pm ET	2:30-3:30pm ET	6-7pm ET	7-9pm ET	8-9pm ET	9-10pm ET
Monday			Roy					
Tuesday			Roy					
Wednesday			Roy				Roy	
Thursday	Yiwen							Yiwen
Friday		Yiwen						
Saturday	Yiwen							

- Please note that the team will be spending the allocated timeslot to answer as many questions as they can from Piazza.

Course Logistics – Office Hours - Cont'd

- Use Course Piazza to ask asynchronous questions that require instructor and/or TA help
- Use the Student Space Slack Channel to find a teammate for your course project (No instructor or TA help is offered there)

Course Assessment

Project	Assignments	Quizzes
25%	45%	30%

- **Course Project:** details are released in week 2. Each student will have the option to choose another student for the project and you will choose one of two project options to submit. Students will be expected to record a video including a code-walkthrough of their work and functionality demo showing the running version of their application. Project submission deadline is **February 29th, 11:59pm ET/8:59pm PT.**

- **Quizzes:** there will be 1 quiz published on Canvas after each lecture with a specific access code. The access code will be revealed during the lecture to the registered students of the corresponding section.
    - Quizzes will start next lecture.
    - You will receive two excused absences from Quizzes for emergencies, sickness, etc.
    - If you need more time, get an approval from your faculty advisor (your professor and not the administrative person)

Course Assessment – Cont'd

Project	Assignments	Quizzes
25%	45%	30%

- **Homework Assignments:** there will be 3 homework assignments provided throughout the course covering the practical aspects of the class. There will be good learning curve that students will have to take on their own.
- Students will have 3 days to submit the assignment after the due date with a late penalty. Late penalties are applied based on the timestamp of the last code commit on GitHub and it will follow this equation (no matter whether the delay is in minutes or in hours):
    - 5 points for up to 24 hours delay
    - 15 points for the next 24 hours delay
    - 25 points for the next 24 hours delay
    - 100 points penalty (no grade) after this time.

After homework grades are released, a Canvas announcement will be made with a link to submit regrade requests. Regrade requests can be made for **24 hours** via the Request Form URL that is provided on the Canvas announcement and CANNOT be submitted via email.

Course Grade Scheme

+/- are used to provide granularity in equal intervals of B and C ranges

Grade	Percentage Interval
A/A-	[85-100%], A starts from 93
B	[70-85%)
C	[55-70%)
D	[40-55%)
R (F)	Below 40%

Course Schedule

Date	Topic	Notes
Week-1	- Generative AI and Predictive AI - Introduction to LLMs	- GCP Coupons are distributed
Week-2	- LLM Types and Architectures - LLM Components	- HW-1 is released - Course Project released
Week-3	- Lab on LLMs on HuggingFace - LLM integrations: LangChain	- HW-1 deadline - HW-2 released
Week-4	- LLM integrations: LangChain (Cont'd)	- HW-2 deadline
Week-5	- LangChain on the Cloud - Lab on LangChain	- HW-3 released
Week-6	- LLMs and Vector Databases - LLM Quantization	- Course Project Submission deadline
Week-7	- LLM Evaluation - Miscellaneous Topics on LLMs	- HW-3 deadline

Course Delivery and HW Notes

- Lecture materials will be released on Canvas prior to the lecture.
- Annotations will be added on the slides while playing them on TopHat but you won’t be able to download the TopHat slides.
- All HW assignments will be submitted via GitHub classroom.

Other Syllabus Information

- Please read the remaining sections of the Course Syllabus.
- The Syllabus can be found on Canvas under the Modules section

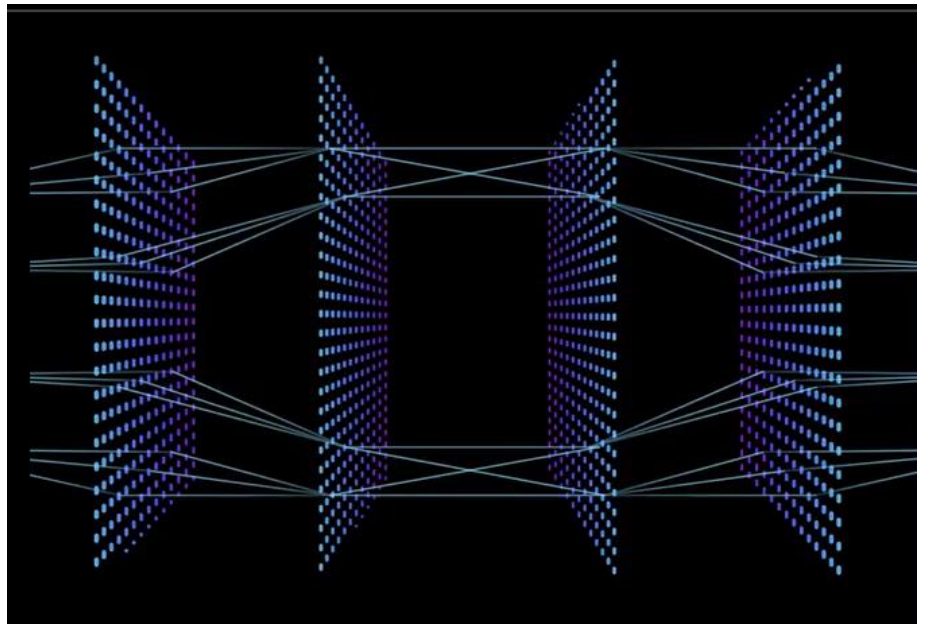
What is Generative AI?

Let’s start with historical view of the Pre-Generative AI Era

Step-1: Traditional AI
Supervised and Unsupervised Machine Learning

Step-2: Artificial Neural Networks

- In 2011, Google Brain tied together 16,000 processors to look through 10 million digital images pulled from YouTube videos.
- Google Brain used something called Deep Learning Artificial Neural Network
- Google Brain clustered over 20,000 patterns in these massive datasets
- At this point, an idea started to float around that if we can cluster “almost” all images, why don’t we create our own?



Step-3: Data Creation

- The process of data creation requires models that can do several tasks and not just one task.
- Predictive AI is designed to help you address a single task, e.g., predict the price of your car next year!
- This is where the focus started to switch to “Generative AI”.
- **Generative AI** aims to perform all the feasible tasks

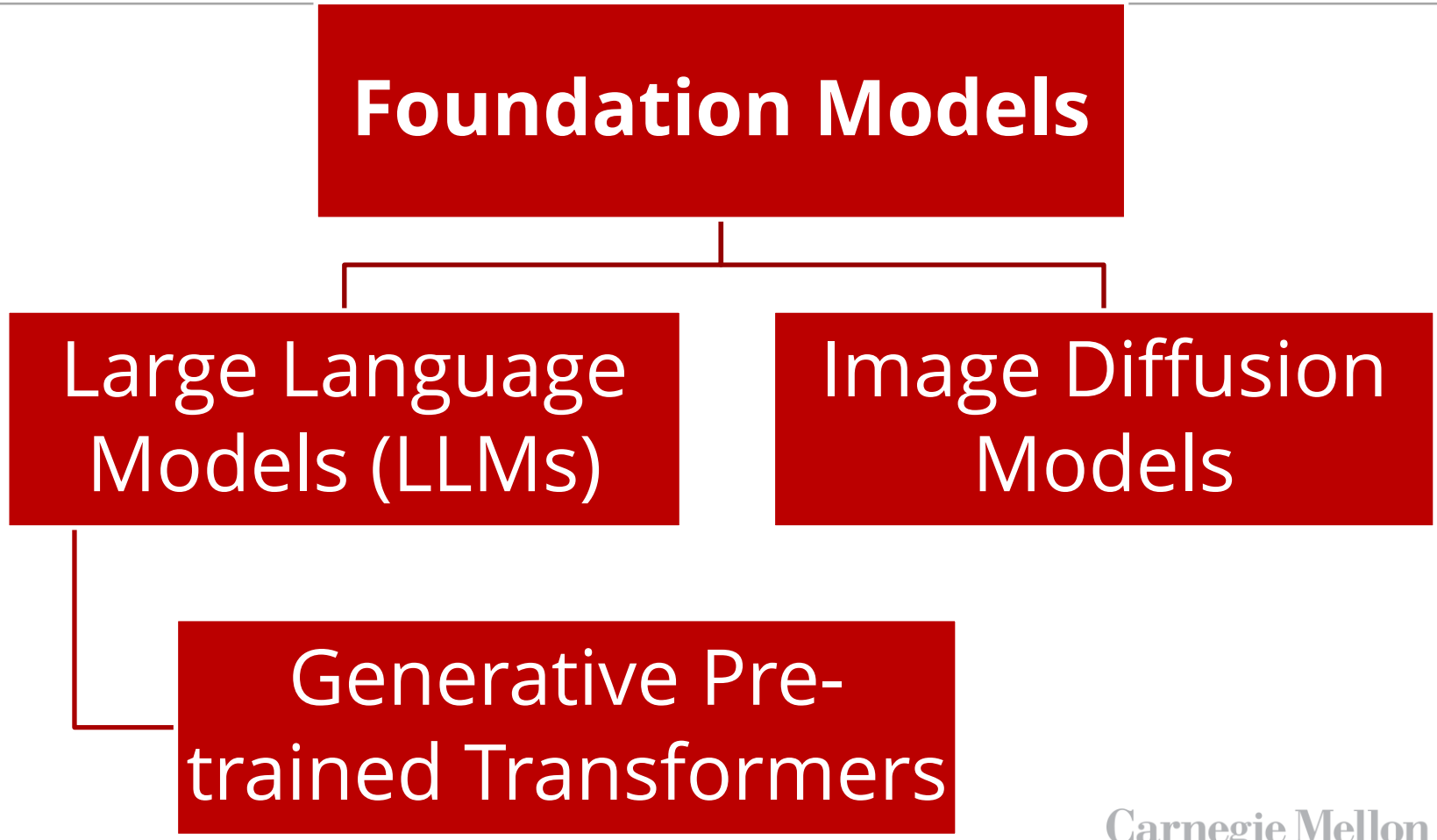
But, how to perform all feasible tasks?!!

- Well, we need massive amounts of data.
- These data will be processed by huge, highly capable models.
- These models are called: **“Foundation Models”**.

Generative AI – Foundation Models

1. Foundation models are models that are trained on broad data and can be customized/adapted to a wide range of downstream tasks
2. Foundation models are more computing and data intensive than predictive models.
    - In predictive AI, you can build a model to train someone how to drive a car (with data focusing on cars)
    - In generative AI, you will train yourself with a foundation model of all modes of transportation. In this case, you will focus on general features like acceleration, momentum, electricity and gravity

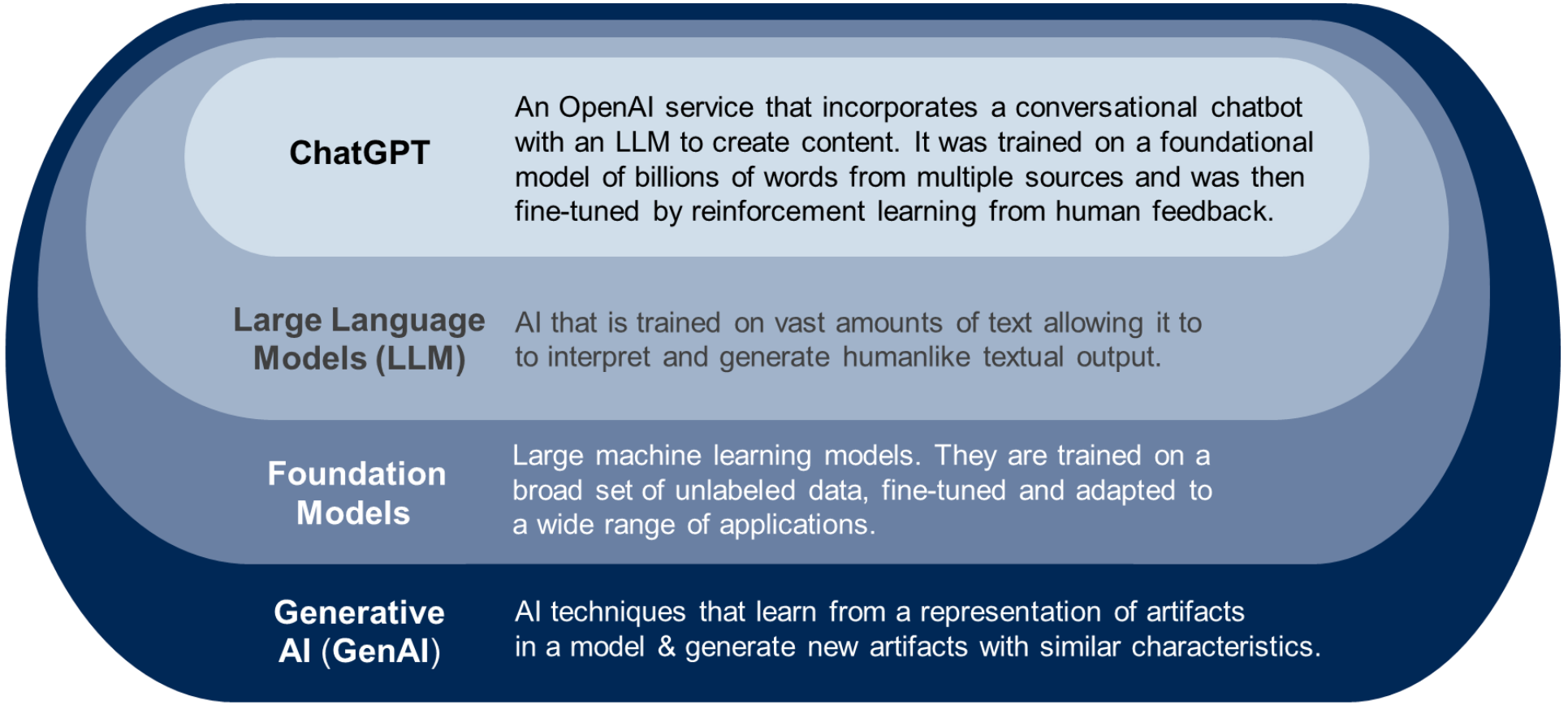
Generative AI – Foundation Models – Some Examples



Generative AI – LLMs

- Language models are a type of AI system trained on text data that can generate natural language responses to inputs or prompts. These systems are trained on text prediction tasks.
- **Large language models (LLMs)** generally refer to language models that have hundreds of millions (and at the cutting edge, hundreds of billions) of parameters, which are pretrained using billions of words of text and use a transformer neural network architecture.
- LLMs are the basis for most of the foundation models.

LLMs in Generative AI - Overview



Source: Gartner

How do LLMs Work?

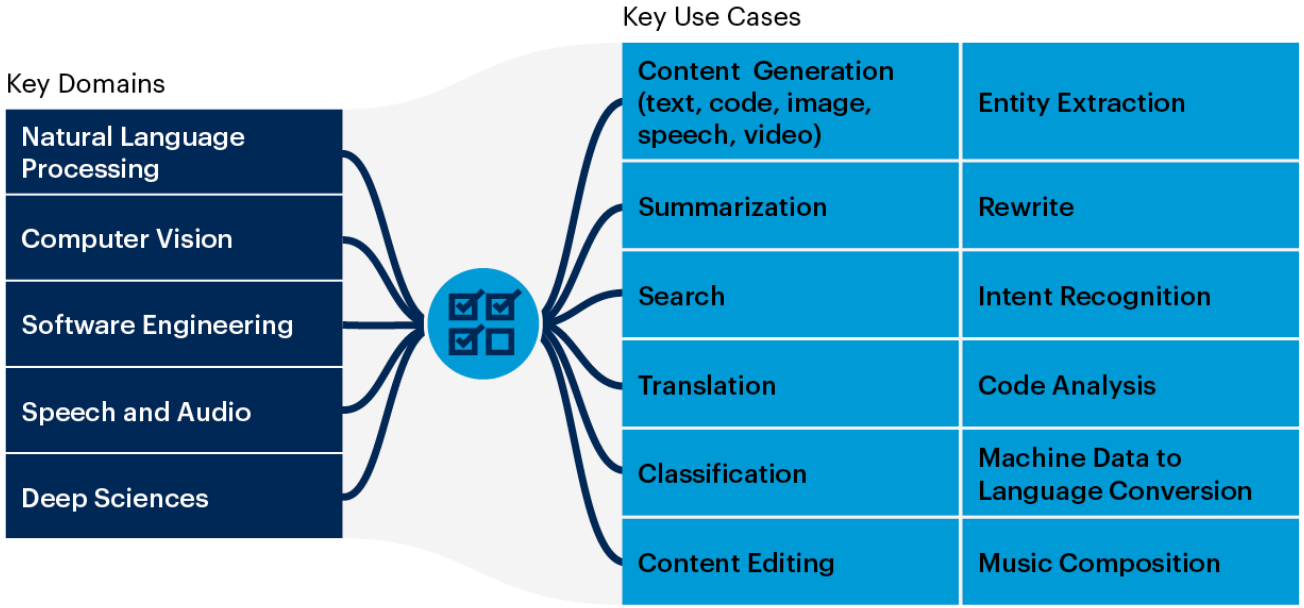
- A Large Language Model is just a supervised learning model to expect the next word
- LLMs self-improve their performance using RLHF Algorithm (Reinforcement Learning from Human Feedback) – More details to come!

Input	Output
My favorite food is	pasta
My favorite food is pasta	with
My favorite food is pasta with	shrimp

Basic LLM Terminologies

- LLM can make facts up! They are called **Hallucinations**.
- LLMs were trained on huge amounts of data so it can easily misinterpret what you mean.
- So, it's good to provide your LLM with a **context**!
- **Prompts** help provide the context to LLMs.
- **LLM's context length limit** is the limit on the total input + output size.

Good tasks for LLM



Key Trends Affecting This Market

Models Will Slim Down	Mainstreaming of OSS GenAI Models	Growth in Domain-Specific Models	Model Hubs Enable Developer Collaboration
Emergence of Multi-Modal Models	Regulations Intensify	Potential Model Commoditization	Emergence of Autonomous Agents

Source: Gartner
774602_C

LLMs or Web Searches?

- Web searches provide more reliable answers than LLMs
- If you are looking for ideas or innovative answers, LLMs can be helpful!
- Remember, check LLM output always!

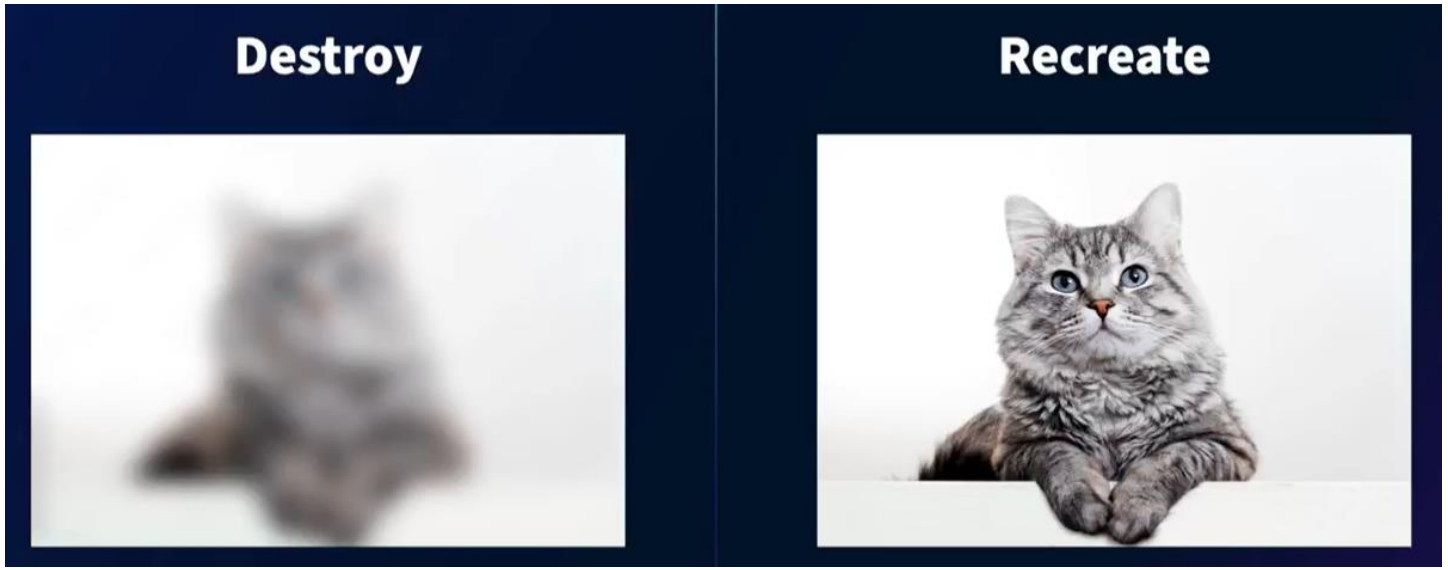
LLM Limitations

LLMs are not perfect! Keep in mind the following limitations when using LLMs:

- Length of Input and Output of some LLMs is limited, implying limited data to be provided and potentially limited output to be consumed.
- Knowledge Cutoff: LLMs are limited to the training date.
- Generative AI doesn’t work well with tabular data and mathematical calculations (Use supervised learning instead).
- Bias and Toxicity

Generative AI – Image Diffusion Models

- A Diffusion model is a foundation model that takes million of images and destroys them to try to recreate them.
- Examples include OpenAI's DALL-E, Midjourney, and even open-source packages like Stable Diffusion.



Next

- Complete Course Entry Survey:
  - https://forms.gle/kGDoGyGt6w6kz7AHA
- Sign-up for the course on TopHat.
- Join the Course Piazza
- Join the Student Slack Workspace
- Read “Introduction to Google Cloud” from this URL:
  - https://cloud.google.com/docs/overview

Waitlisted Students

- All materials for first two weeks will be uploaded here