14-825 Generative AI and Large Language Models

SPRING 2025

MOHAMED FARAG

FARAG@CMU.EDU

Agenda

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

Why is this course Important?

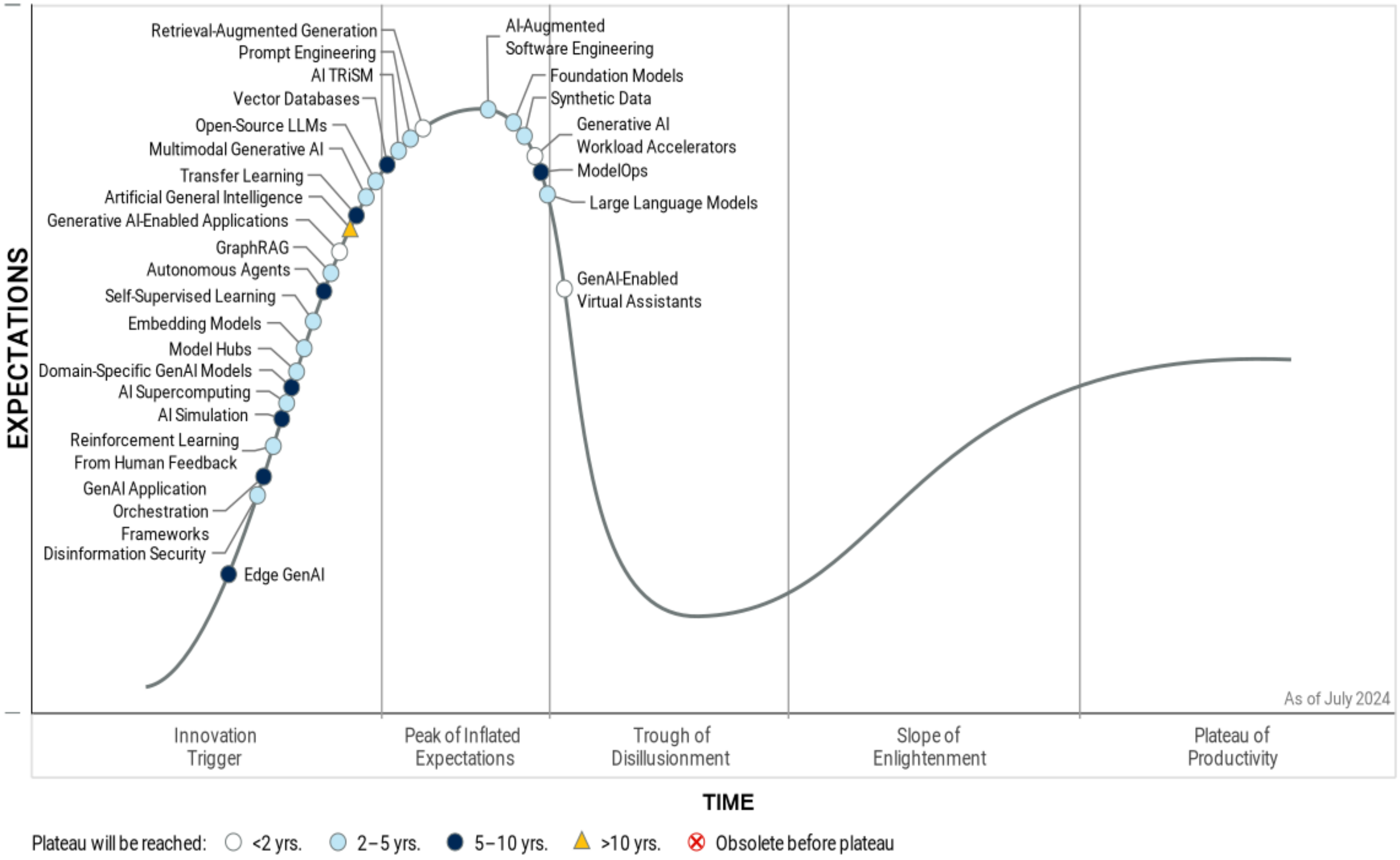
Over the past few years, the demand for Generative AI jumped significantly.

- “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 2024 Generative AI Hype Cycle released by Gartner R&D

Hype Cycle for Generative AI, 2024



What is NOT this course about?

- This course is NOT about LLM Theory.
- This course is NOT about LLM internal architecture.
- This course is NOT about building LLMs.
- This course is NOT about studying LLMs.

If you prefer one of the topics above, please check the following courses: 11667, 11868 and 11967

Main Themes of the Course

- Understand the main concepts of Generative AI
- Design and build the components of LLM Applications.
- Build RAG pipelines.
- Understand the basics of Multimodal LLMs.

Expectations for Incoming Students

- ***You are expected to know Python or are willing to learn it.***
  - A Python recording will be released later this 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 introductory materials to AI

Instructor and TA Introductions

Instructor

- Mohamed Farag: farag@cmu.edu

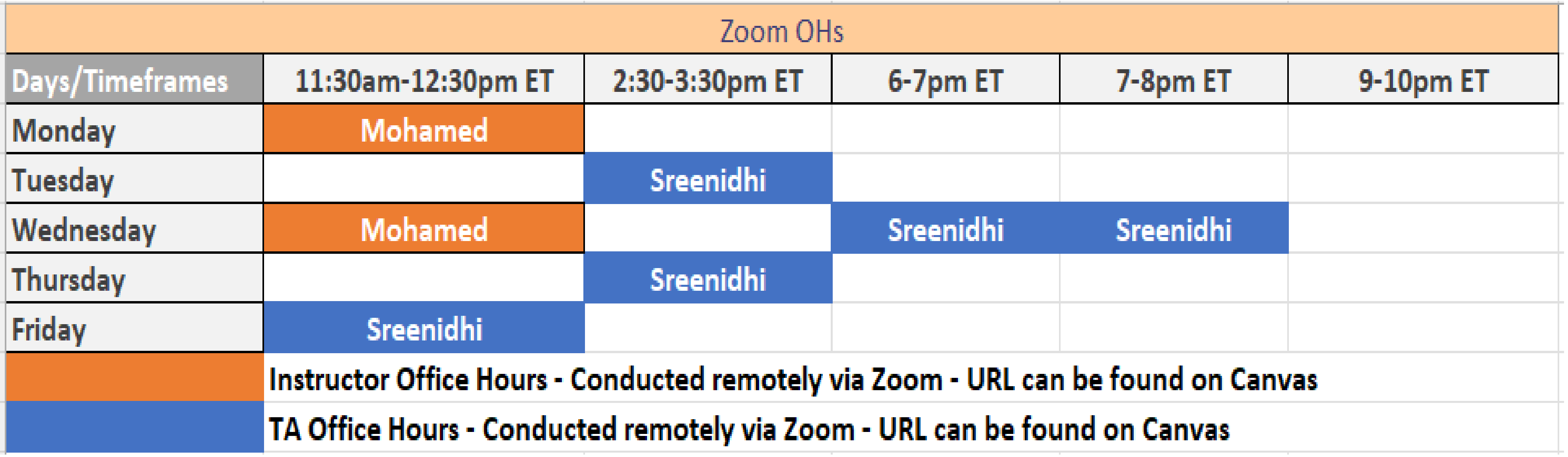
TA

- Sreenidhi Ganachari sganacha@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: **649647**
- Use the Student Space Slack Channel to find a teammate for your course project (No instructor or TA help is offered there)
- Students who have approved accommodation shall contact the course instructor to figure out how the instructor can meet their needs.
- You may contact the student affairs if you must miss a few classes due to illness.

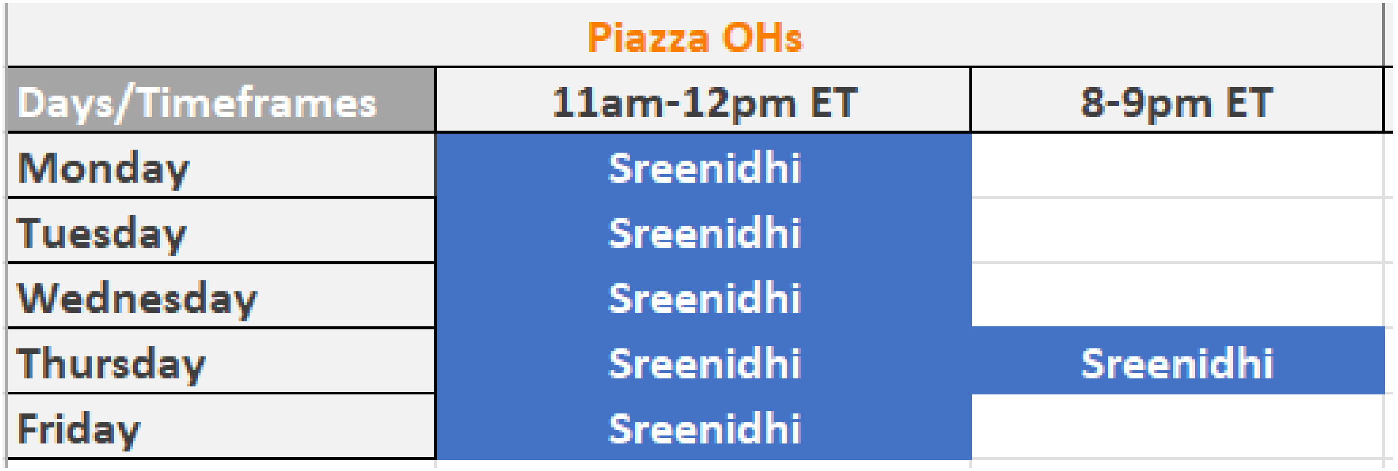
Course Logistics – Office Hours



Zoom OHs					
Days/Timeframes	11:30am-12:30pm ET	2:30-3:30pm ET	6-7pm ET	7-8pm ET	9-10pm ET
Monday	Mohamed				
Tuesday		Sreenidhi			
Wednesday	Mohamed		Sreenidhi	Sreenidhi	
Thursday		Sreenidhi			
Friday	Sreenidhi				
🟧	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/94117627561?pwd=mmJpBtiI76BwNQUDZ4KUdWoFbRpvOa.1

Course Logistics – Piazza Hours



Piazza OHs		
Days/Timeframes	11am-12pm ET	8-9pm ET
Monday	Sreenidhi	
Tuesday	Sreenidhi	
Wednesday	Sreenidhi	
Thursday	Sreenidhi	Sreenidhi
Friday	Sreenidhi	

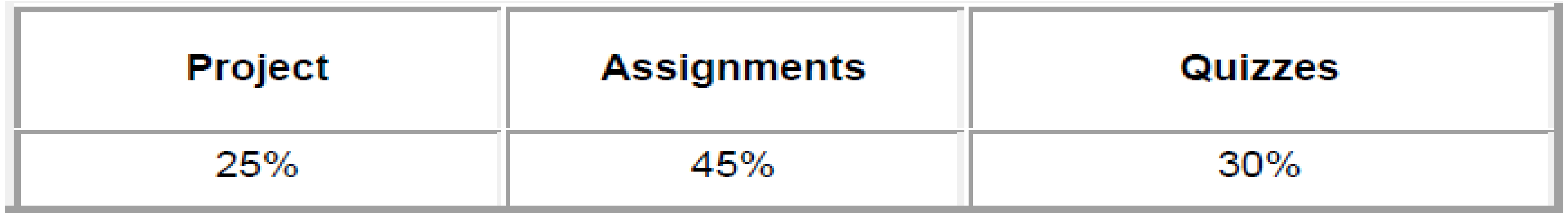
- Use Course Piazza to ask asynchronous questions that require instructor and/or TA help

- Please note that the TA will respond to inquiries/questions made ***before*** the Piazza OHs start time. Questions and inquiries that are made during the OHs time slot are not guaranteed to be answered during the same time slot.

Office Hours Etiquette Reminder

- Office Hours aim to help you find the path to maximize your learning experience.
- Getting the answers from the TA directly won't help you learn so **there won't be direct solutions provided during Office Hours**.
- The goal of the office hours is to **give you some ideas and pointers for you** to debug the issues.
- Please don't plan to spend **more than 10 minutes** in your conversation with the TA.
- Ask **good questions with due diligence**. Please research the issue and put an effort in implementing it before coming to Office hours.
  - **Example of a bad question:** I found this idea online and I plan to cite it but can't get it to work. Can you help?
  - **Example of a good question:** I'm getting a bug in my deployment to the cloud, I researched the issue and found these 3 different references (share the URLs).
    I implemented the first one and it didn't work. I'm trying the second one now and getting an error that I can't find enough references to it online. What could be the root cause of it?

Course Assessment

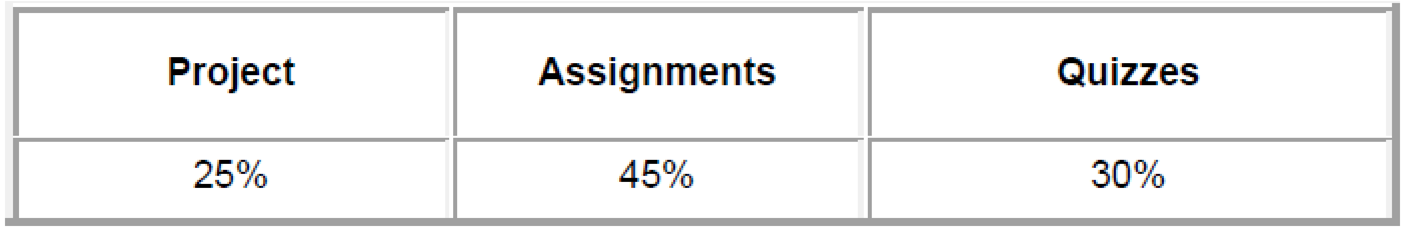


Project	Assignments	Quizzes
25%	45%	30%

- **Course Project:** details are released in week-2. Each student will have the option to choose work with a peer and you will choose one of three 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 19th, 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 one 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



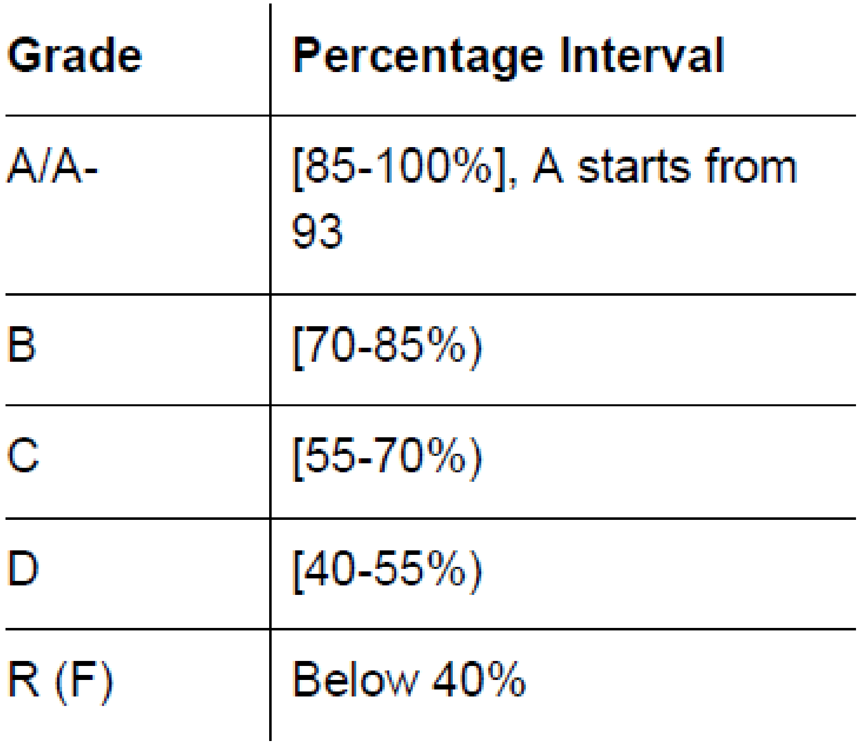
Project	Assignments	Quizzes
25%	45%	30%

- **Homework Assignments:** there will be 3 homework assignments provided throughout the semester covering the practical aspects of the class. There will be good learning curve that students will have to take on their own.
- Each assignment and course project has **1 grace day beyond the assignment deadline**.
- Grace days are **<u>not transferrable</u>**.
- Students will have 3 days to submit the assignment after the grace day extension 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):
    - Total of 5 points for up to 24 hours delay
    - Total of 15 points for the next 24 hours delay
    - Total of 25 points for the next 24 hours delay
    - 100 points penalty (no grade) after this time.

After homework grades are released, **regrade requests can be made for 24 hours via Gradescope <u>and CANNOT be submitted via email.</u>**

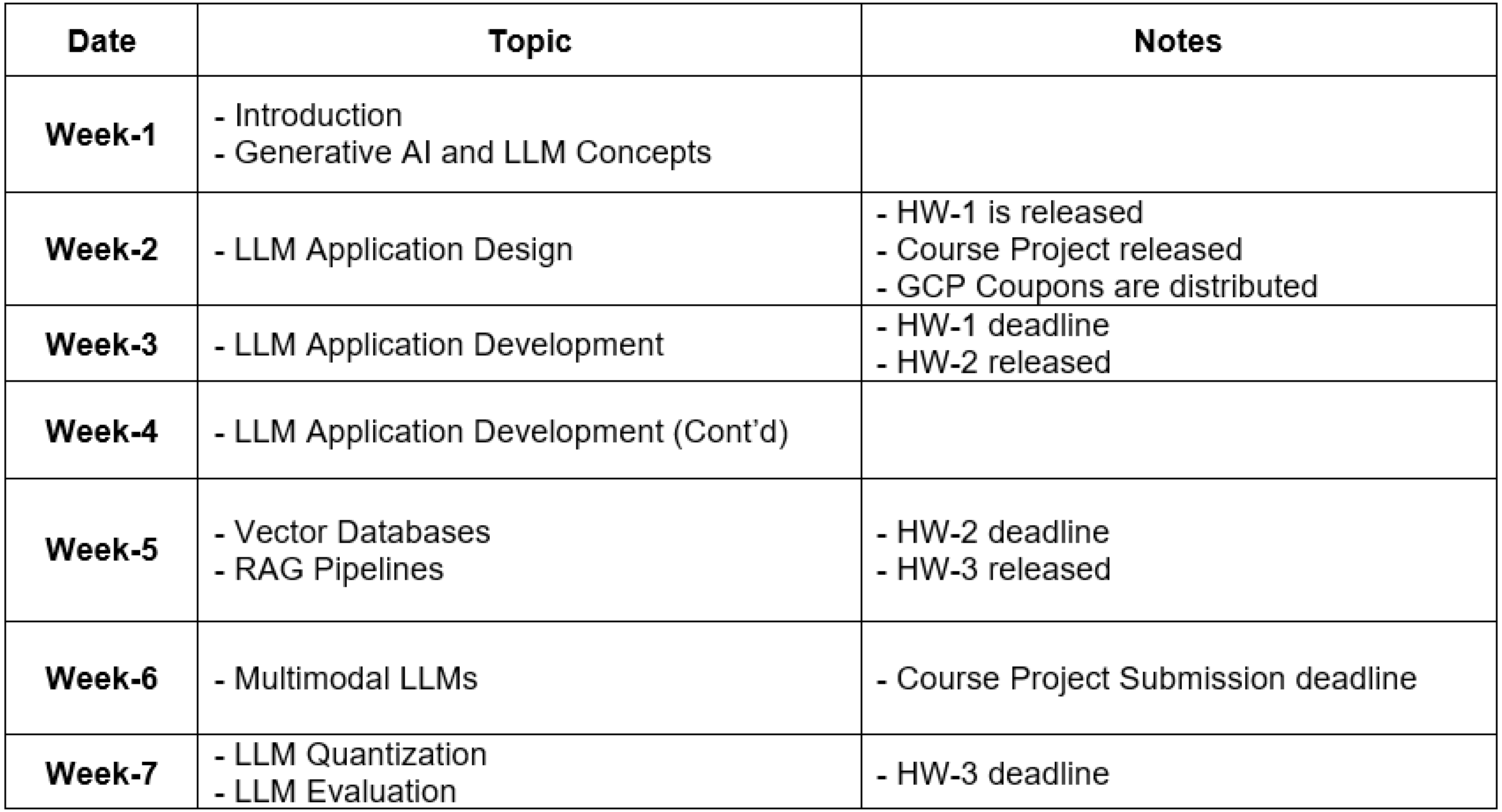
Course Grade Scheme

+/- 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	- Introduction - Generative AI and LLM Concepts	
Week-2	- LLM Application Design	- HW-1 is released - Course Project released - GCP Coupons are distributed
Week-3	- LLM Application Development	- HW-1 deadline - HW-2 released
Week-4	- LLM Application Development (Cont'd)	
Week-5	- Vector Databases - RAG Pipelines	- HW-2 deadline - HW-3 released
Week-6	- Multimodal LLMs	- Course Project Submission deadline
Week-7	- LLM Quantization - LLM Evaluation	- 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.
- HW Due date on Canvas **DOES NOT include the grace day extension**.
- HW Due date on Gradescope **includes the grace day extension**. This is done to simplify the grading process.
- All HW assignments will be submitted via GitHub classroom.

Academic Integrity Violations (AIVs)

- AIVs are serious and can have direct impact on your course grade, your scholarship -if any-, your graduation timeline, and/or your continuation in your degree program.

- Simple rules to follow:
    - Cite all the references you are using. Use APA citation style.
    - Cite ChatGPT (or other AI tools) for any code/info used in your answers.
    - Don’t use more than 30% of your solution/answer from external sources.
    - Collaborate and share ideas with your peers, and not code.
    - Don’t share code with your peers (including in-class group exercises). Don’t use your peer’s code even after changing variable names or statement order.
    - Don’t share quiz access codes with your peers.

Other Syllabus Information

- Syllabus contains important information about student wellness, student academic success center, and food insecurity.
- The Syllabus can be found on Canvas under the Modules section

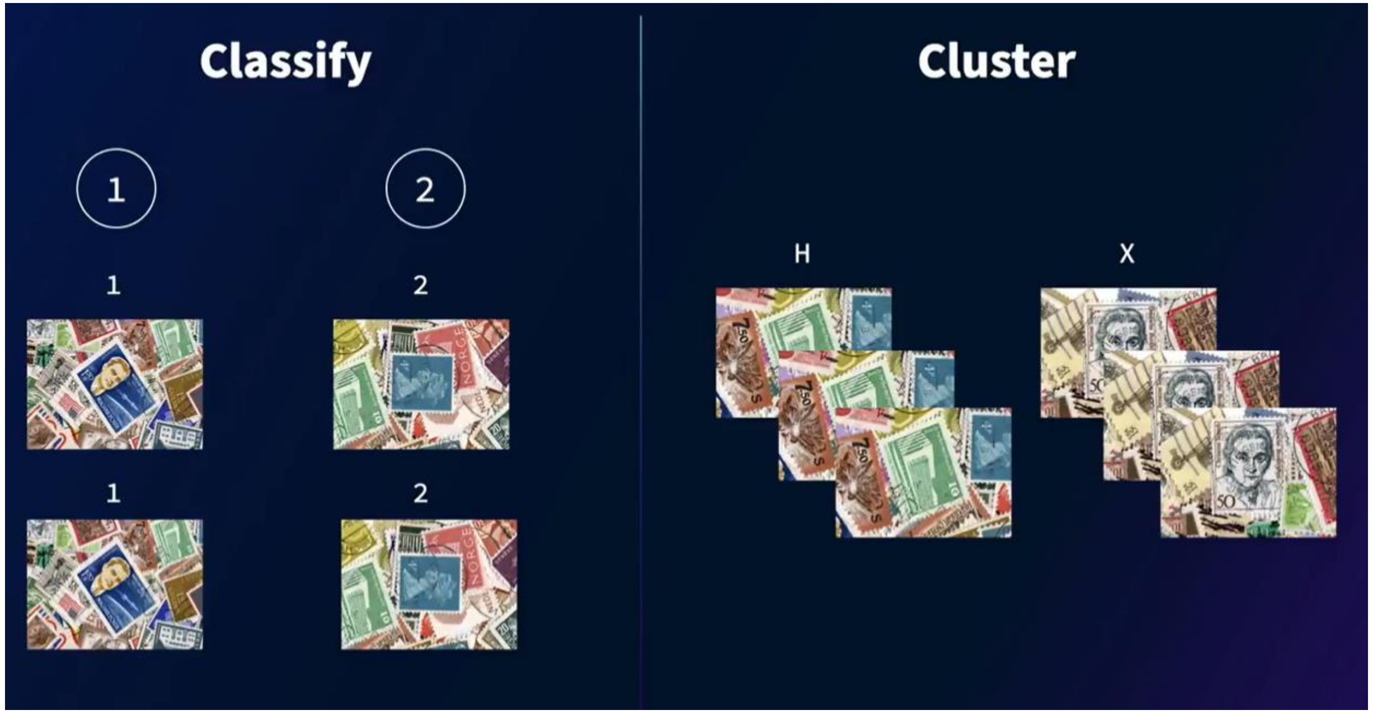
Waitlisted?

For enrollment questions and inquiries, please email the INI Academics at ini-academic@andrew.cmu.edu

What is Generative AI?

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

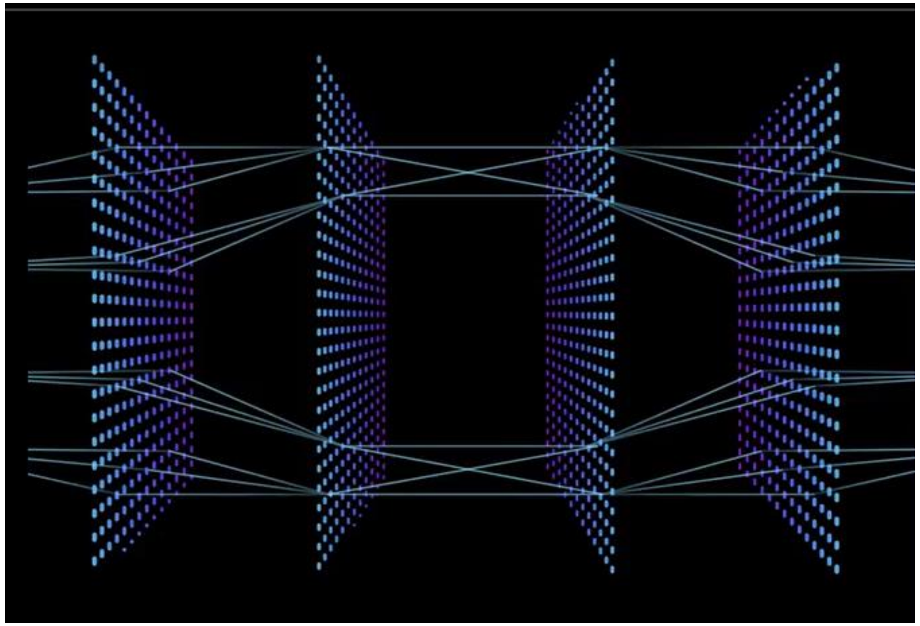
Step-1: Traditional AI
Supervised and Unsupervised Machine Learning



Carnegie Mellon University

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?



Carnegie Mellon University

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?!!

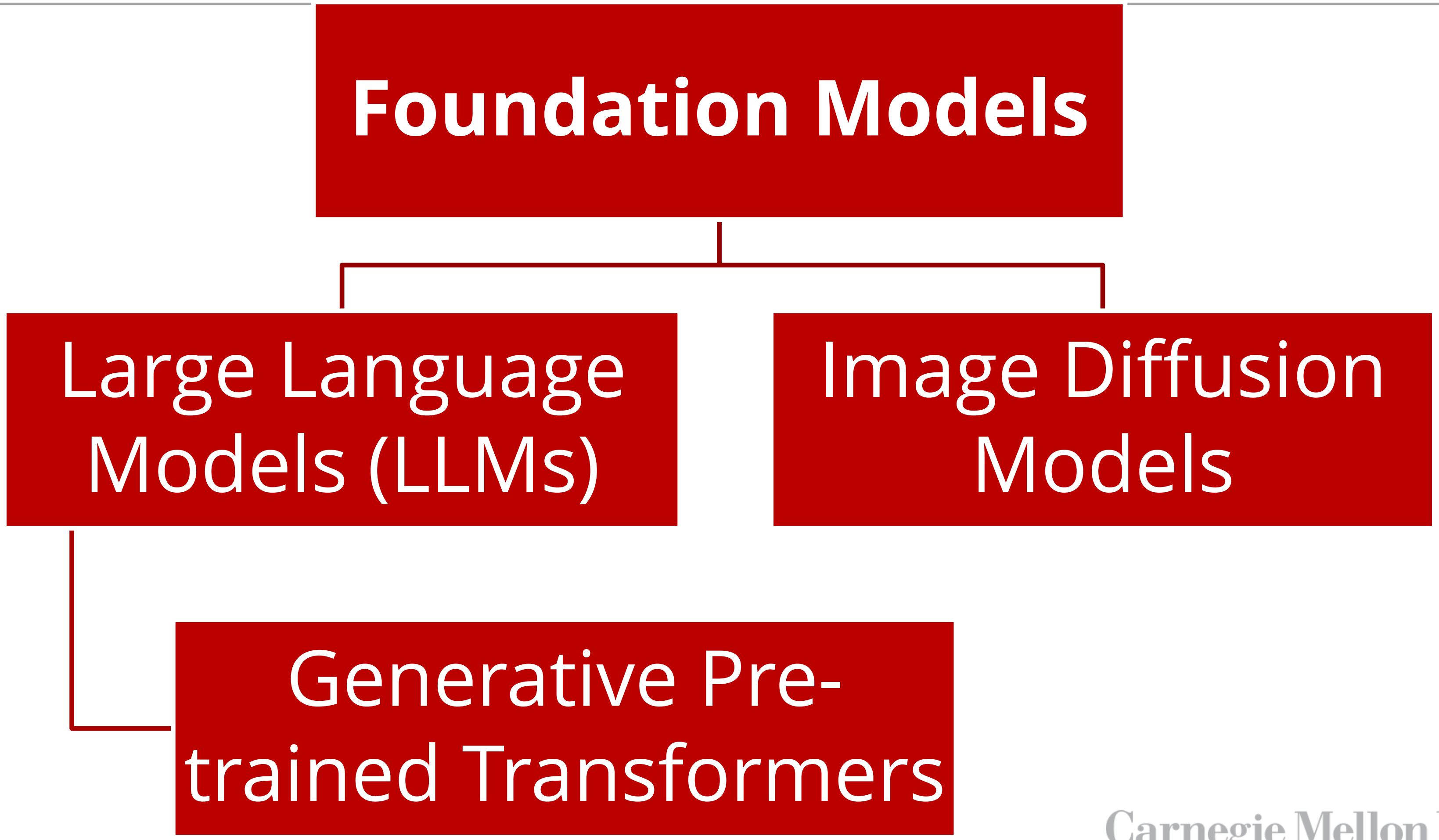
- Well, we need massive amounts of data.
- These data will be processed by huge, highly capable models.
- These models are called: **“Foundation Models”**.

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



Foundation Models

Large Language Models (LLMs)

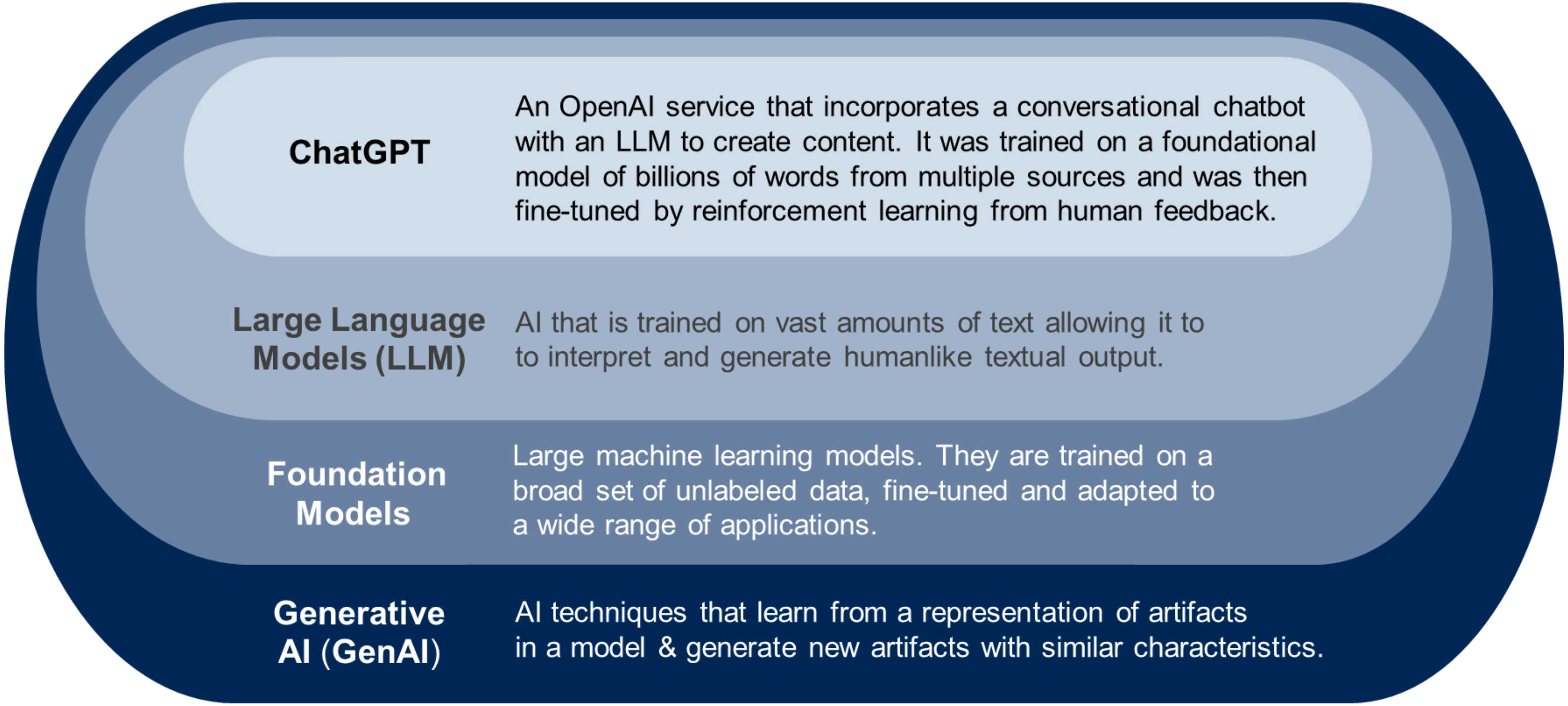
Image Diffusion Models

Generative Pre-trained Transformers

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



ChatGPT

An OpenAI service that incorporates a conversational chatbot with an LLM to create content. It was trained on a foundational model of billions of words from multiple sources and was then fine-tuned by reinforcement learning from human feedback.

Large Language Models (LLM)

AI that is trained on vast amounts of text allowing it to to interpret and generate humanlike textual output.

Foundation Models

Large machine learning models. They are trained on a broad set of unlabeled data, fine-tuned and adapted to a wide range of applications.

Generative AI (GenAI)

AI techniques that learn from a representation of artifacts in a model & generate new artifacts with similar characteristics.

Source: Gartner

How do LLMs Work?

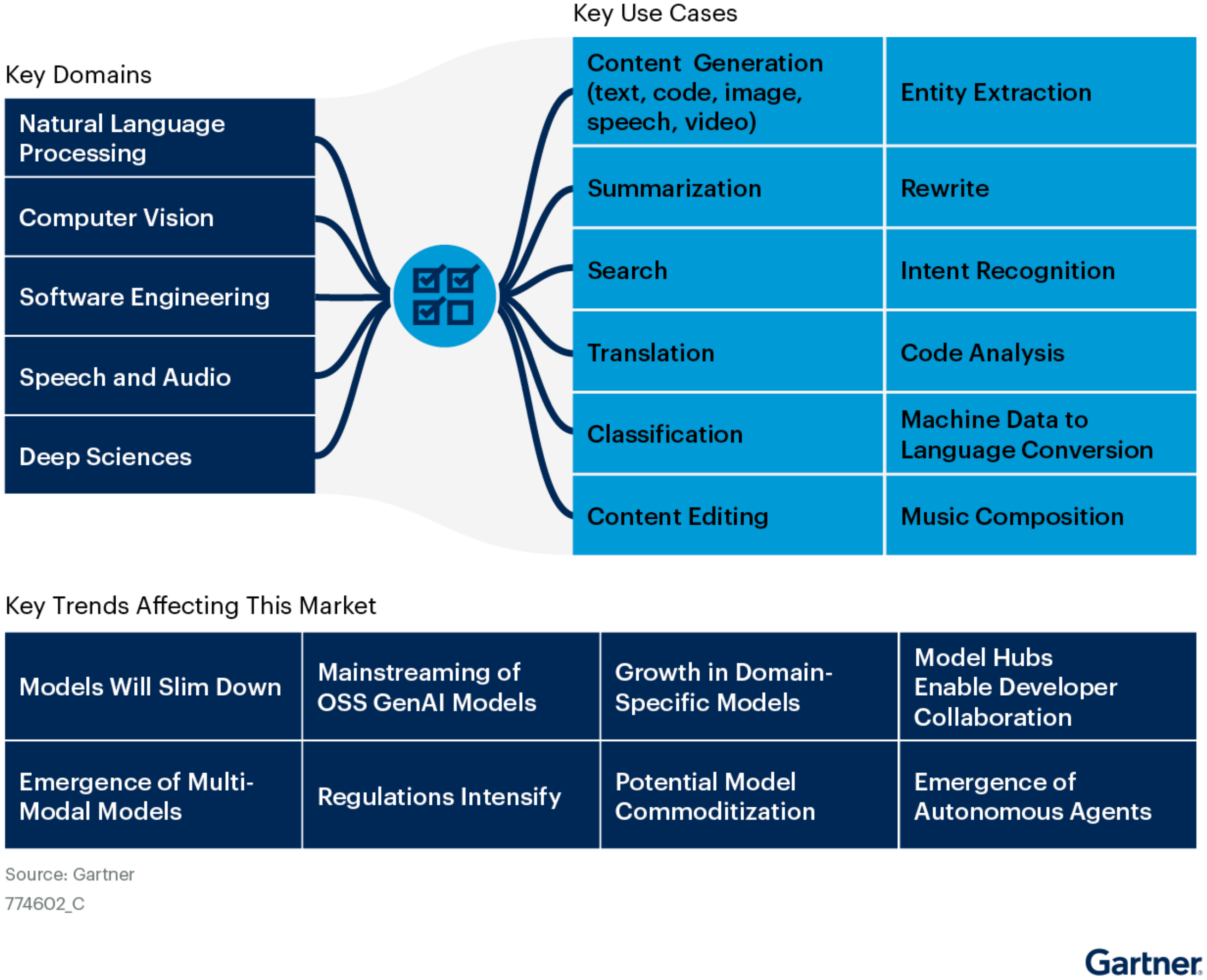
- A Large Language Model learns how to predict the next word
- LLMs self-improve their performance using RLHF Algorithm (Reinforcement Learning from Human Feedback).

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 Domains

- Natural Language Processing
- Computer Vision
- Software Engineering
- Speech and Audio
- Deep Sciences

Key Use Cases

Content Generation (text, code, image, speech, video)	Entity Extraction
Summarization	Rewrite
Search	Intent Recognition
Translation	Code Analysis
Classification	Machine Data to Language Conversion
Content Editing	Music Composition

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!

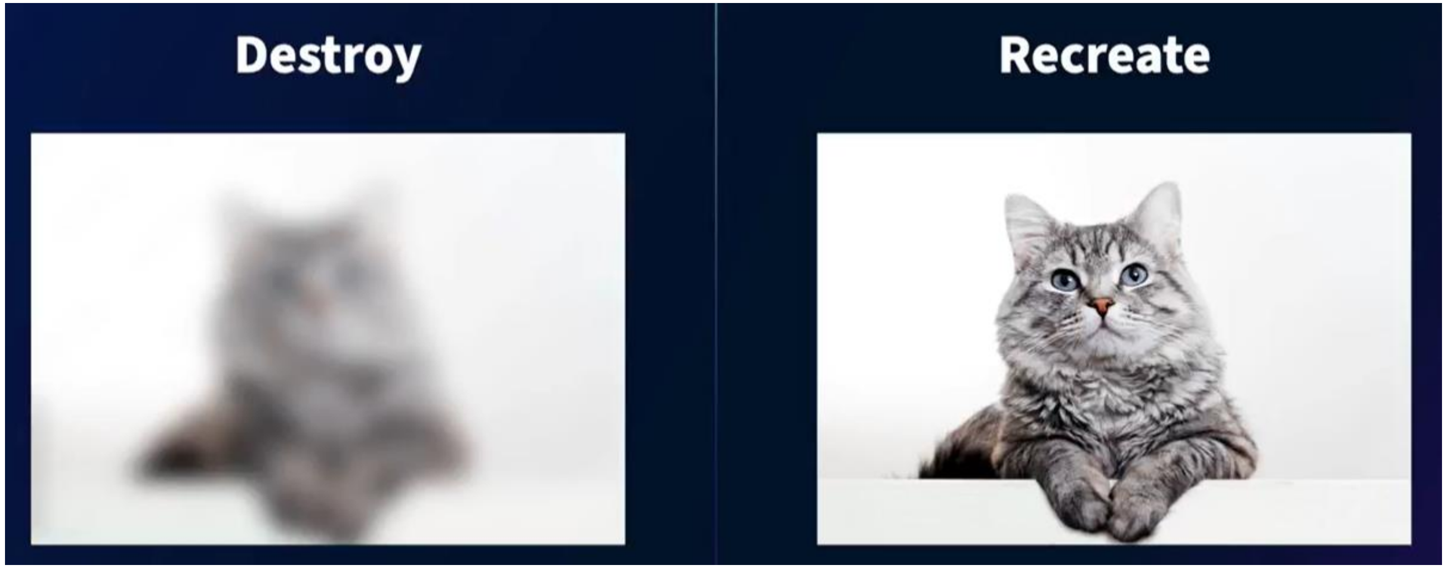
LLM Limitations

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.
- Diffusion models are used in OpenAI's DALL-E, Midjourney, and even open-source packages like Stable Diffusion.



Carnegie Mellon University

Next

- Complete the course entry survey:
    - https://forms.gle/SrD64qgCN5dGPkdF6
- Join 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