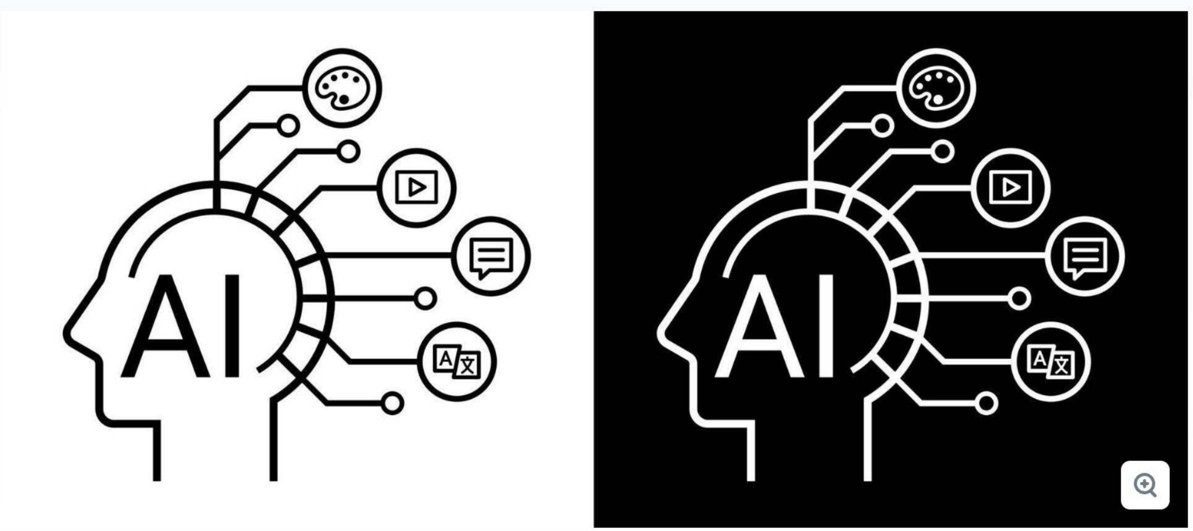
AIMS 5740
Generative Artificial Intelligence

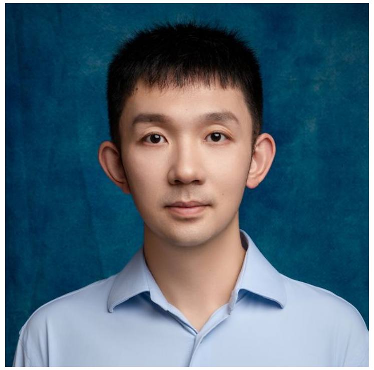
(2026 Term 2)



Computer Science & Engineering
The Chinese University of Hong Kong

About the Course Instructor

Prof. Cheng Yu, CUHK



- My research interests specialize in model compression & efficiency, deep generative models, and large multimodal/language models.
- Developing cutting-edge pre-training techniques for NLP and multimodal tasks. The team achieved SOTA performances on several NLP and multimodal leaderboards and datasets.
- Developing deep generative models (GAN, diffusion) based techniques, many of which have been used for IBM and Microsoft products.
- **I led several teams to work with OpenAI for developing Microsoft-OpenAI core models (Copilot, DALL-E-2, ChatGPT, GPT-4) from 2021 to 2023.**

- Several best paper or candidates awards:

Best Machine Learning and Security Paper in Cybersecurity Award, 2024.

Outstanding Paper Award in NeurIPS, 2023.

Best Student Paper Honourable Mention in WACV, 2021.

- Build strong relations with industry such as OpenAI, Microsoft, Meta, TikTok, Tencent, and Huawei.

OpenAI Microsoft Meta TikTok HUAWEI

Lectures and Tutorials

Lectures
We 6:30PM - 9:30PM ERB LT

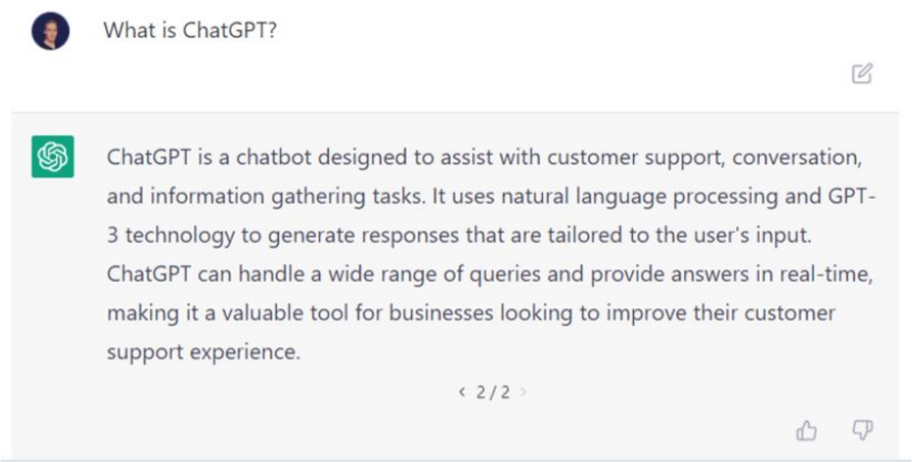
TAs:

Li Tianle, Song Han, Li Yixing, Wang Puyi, Chen Jiacheng
Office: SHB 904

Homepage: https://ltl3a87.github.io/AIMS5740-website/
Newsgroup: Piazza (https://piazza.com/)

The Boom of Generative AI

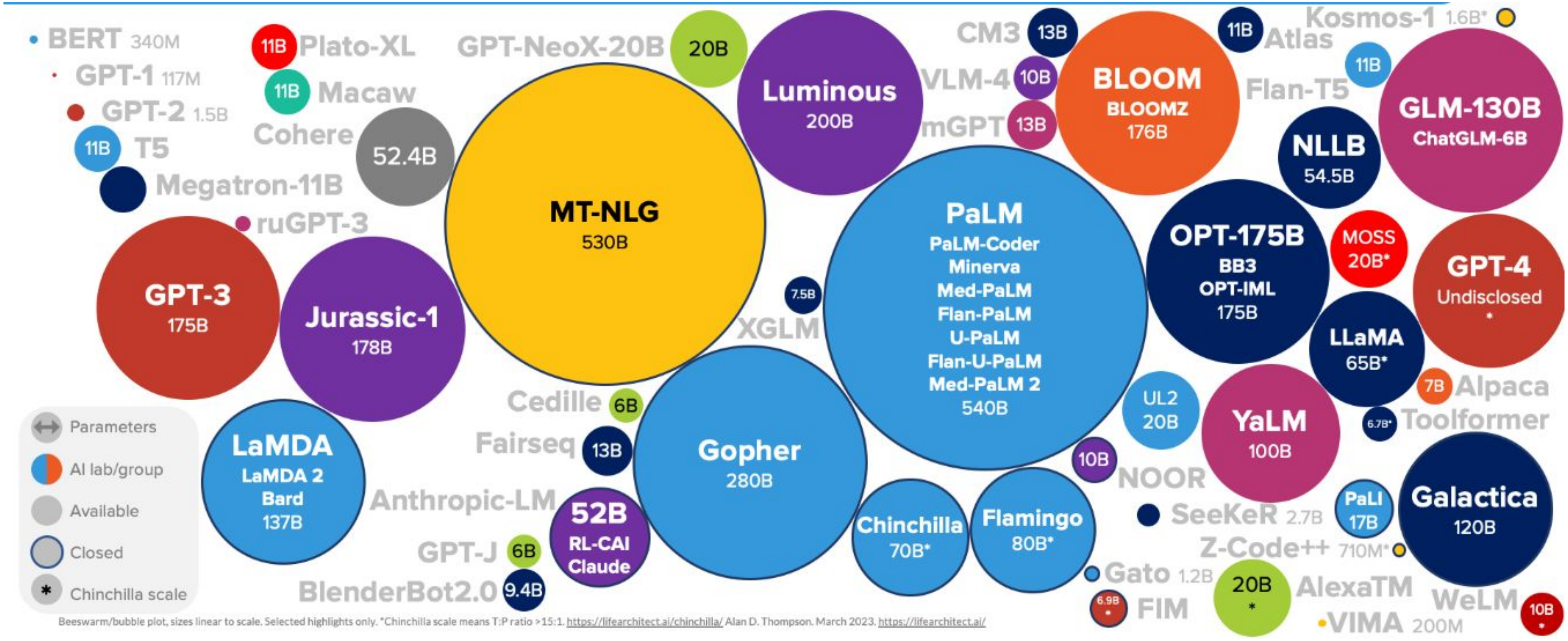
- AI models can generate various high-quality data such as image, text, audio and video.



What is ChatGPT?

ChatGPT is a chatbot designed to assist with customer support, conversation, and information gathering tasks. It uses natural language processing and GPT-3 technology to generate responses that are tailored to the user's input. ChatGPT can handle a wide range of queries and provide answers in real-time, making it a valuable tool for businesses looking to improve their customer support experience.

The Boom of Generative AI



Beeswarm/bubble plot, sizes linear to scale. Selected highlights only. *Chinchilla scale means T:P ratio >15:1. https://lifearchitect.ai/chinchilla/ Alan D. Thompson. March 2023. https://lifearchitect.ai/

What is Generative AI?

Artificial Intelligence

Is the field of study

Machine Learning

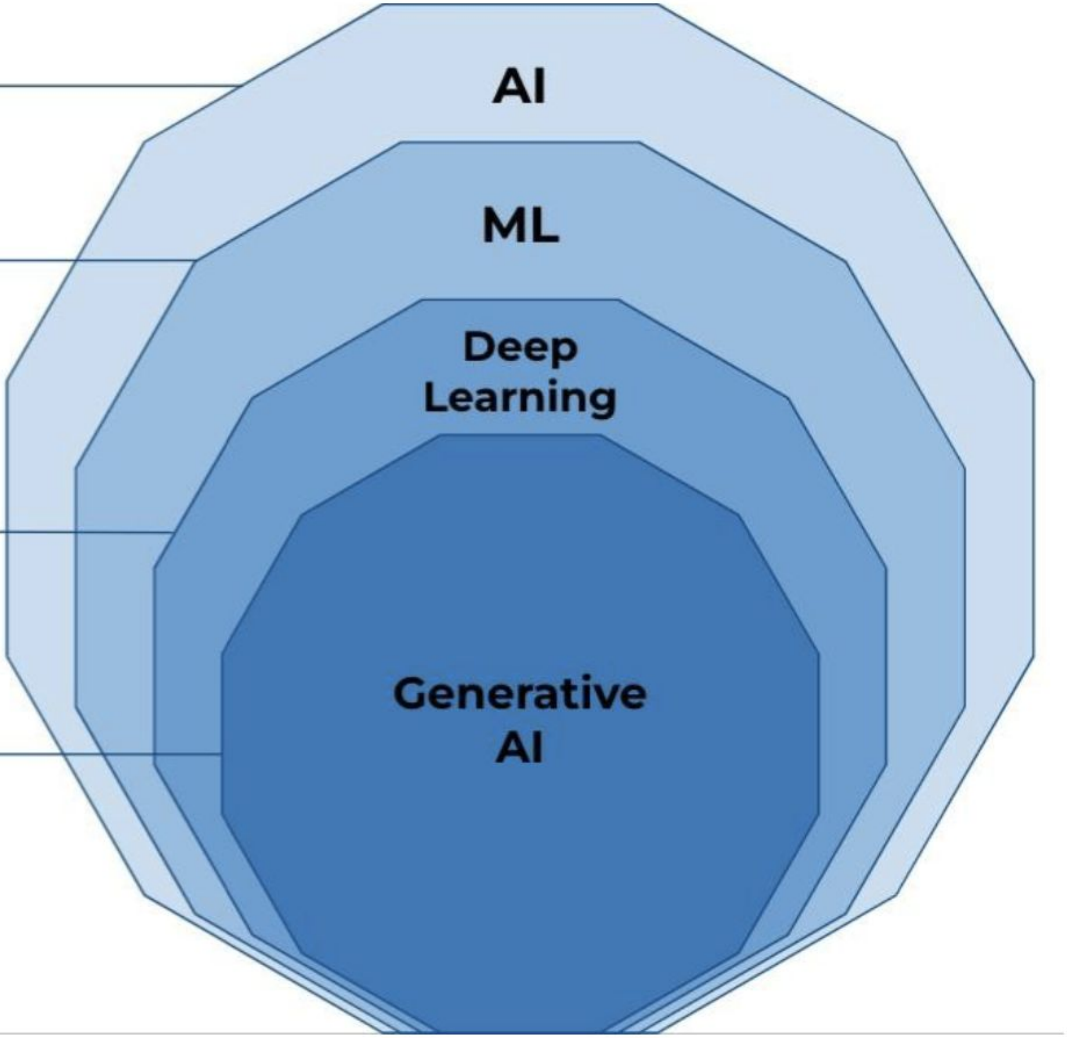
Is a branch of AI that focus on the creation of intelligent machines that learn from data. Another very well know branch inside AI is **Optimization**.

Deep Learning

Is a subset of Machine Learning methods, based on **Artificial Neural Networks.** Examples: CNNs, RNNs

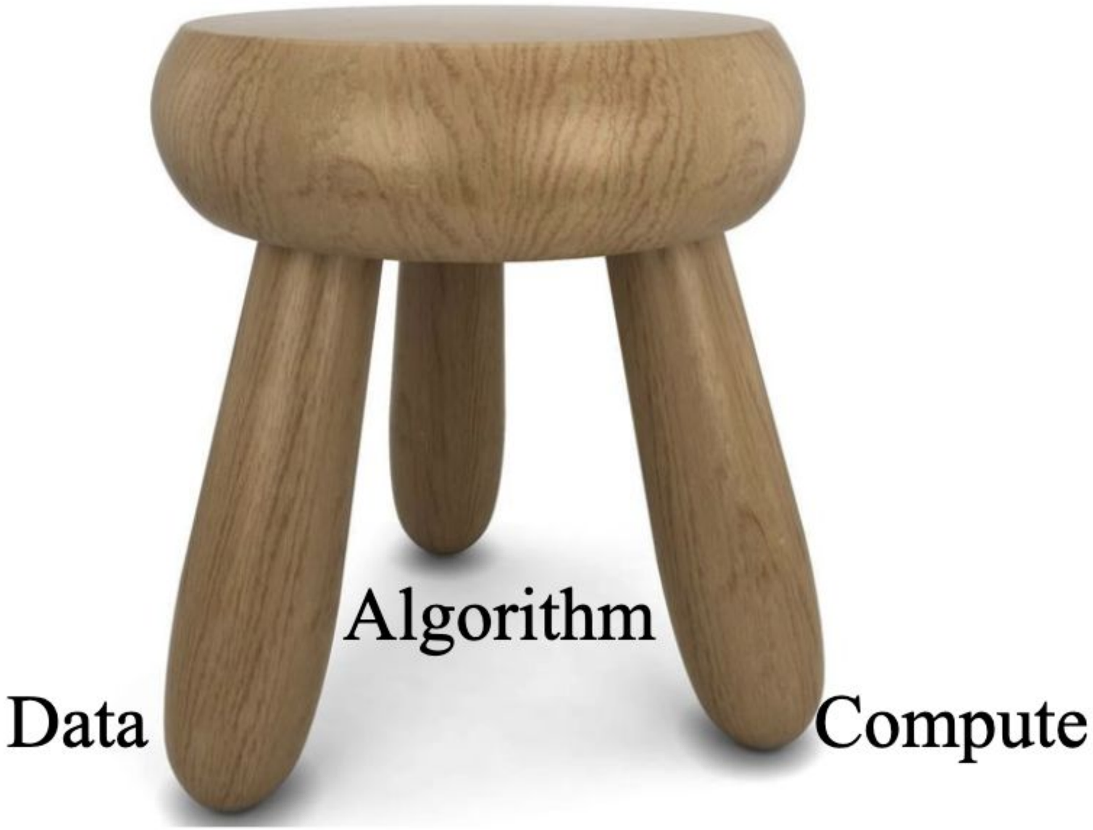
Generative AI

A type of ANNs that generate data that is similar to the data it was trained on. Examples: GANs, LLMs



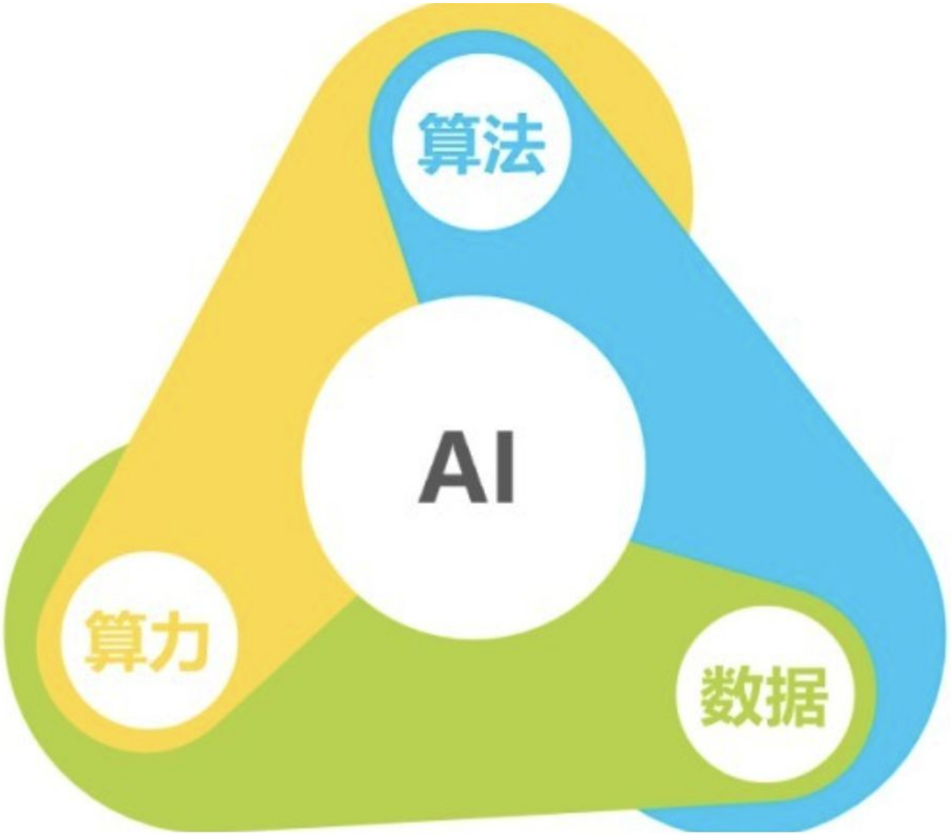
Generative AI is Systems Engineering

- Data, algorithm and compute are three key components in the current generative AI applications.



What is Going to be Taught?

- Data: data cleaning, data processing, tokenization…
- Algorithm: Transformer, Diffusion, RL…
- Compute: Deepspeed, Megatron-LM, vLLM…



Course Outline

Time	Content
Jan.7 - Jan.21	Introduction: 1) basic knowledge; 2) Transformer; 3) CNN, diffusion; 4) other types of models
Jan.22 - Feb.6	Data engineering: data preparation, data cleaning, data processing
Feb. 7 - Feb. 26	Infrastructure preparation: CUDA, Pytorch/ Megatron-LM, vLLM,

Course Outline

Time	Content
Feb.27 - Mar.13	Pretraining: Scaling-law, Training Dynamics, Evaluation
Mar.14 - Mar.27	Post Training: SFT, RLHF, Chain-of-thoughts, Reasoning
Mar.28 - Apr.10	Applications: Agent, Safety, AI for Science/Robotics
Apr.11 - April.17	Guest Lecture, Project Presentation

Grading

2 x 15% Two programming assignments

30% Mid-term Quiz (scheduled on
March 4, 2026 7:30 PM)

40% Final Project (report+presentation)

Important Issues

- Switch off all your phones before lectures & tutorials

- To respect the rights of your classmates, refrain from talking during the lectures & tutorials

- No copy (or similar program copies) or cheating is allowed. Otherwise, ….

- Basic requirement: machine learning, probability, linear algebra, basic programming

Important Issues

- Considering drop this course if you can not attend the mid-term quiz

- AI can only be used for the final project development with clear declaration of used packages

- Late assignment submission: if the submission is late but within 14 calendar days after the deadline, there will be a deduction of 20% from the marks awarded for the submitted piece of work

Thanks

Q&A

Prof. Yu Cheng

chengyu@cse.cuhk.edu.hk