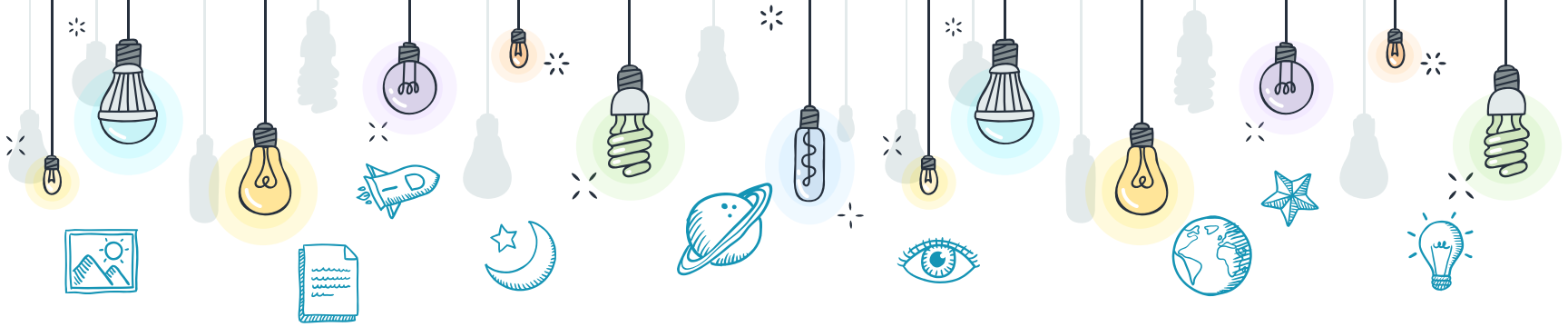


From Multimodal LLM to Human-level AI

Architecture, ***Modality***, ***Function***, ***Instruction***, ***Hallucination***, Evaluation, ***Reasoning*** and **Beyond**

https://mllm2024.github.io/ACM-MM2024/

ACM Multimedia 2024

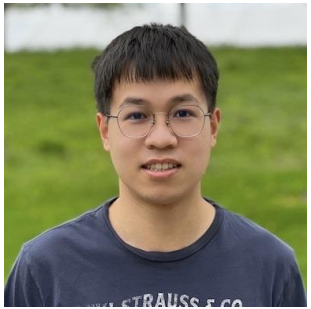
Melbourne, Australia

Hao Fei

National University of Singapore

Xiangtai Li

ByteDance/Tiktok

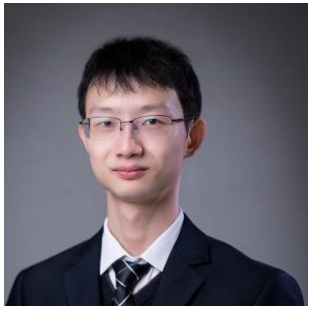


Haotian Liu

xAI

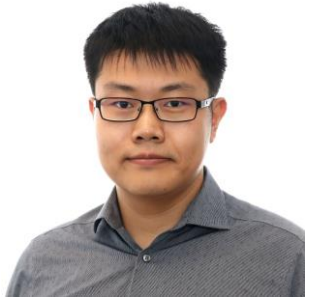
Fuxiao Liu

University of Maryland, College Park



Zhuosheng Zhang

Shanghai Jiao Tong University



Hanwang Zhang

Nanyang Technological University

Kaipeng Zhang

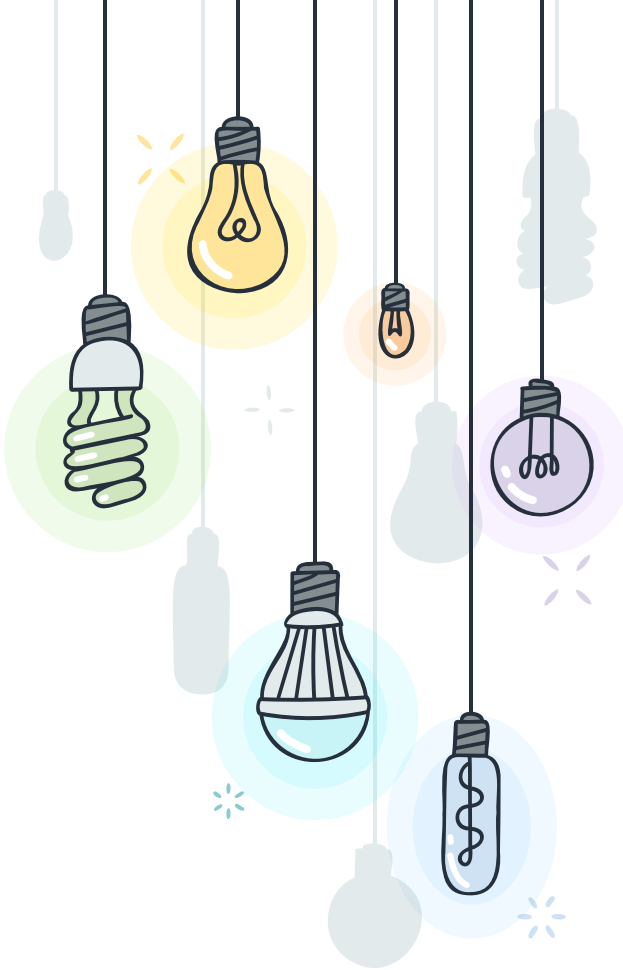
Shanghai AI Lab

Shuicheng Yan

Kunlun 2050 Research, Skywork AI

Part-I

Background and Introduction: *From MLLM to Human-level AI*

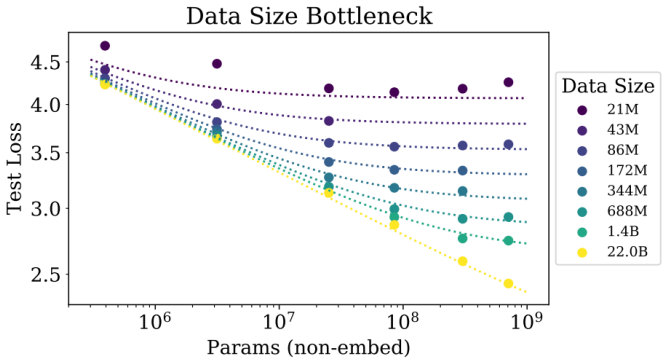


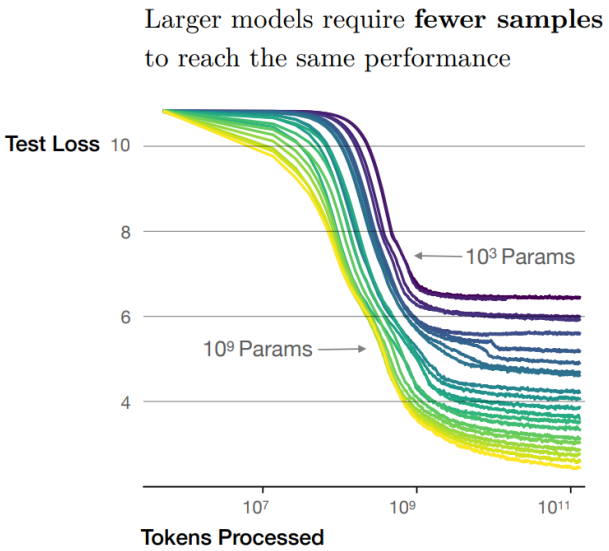
Hao Fei
Research Fellow
National University of Singapore
http://haofei.vip/

Intelligence in Language

- ## Scaling Law in Neural Models

Model size ↑, Data size ↑

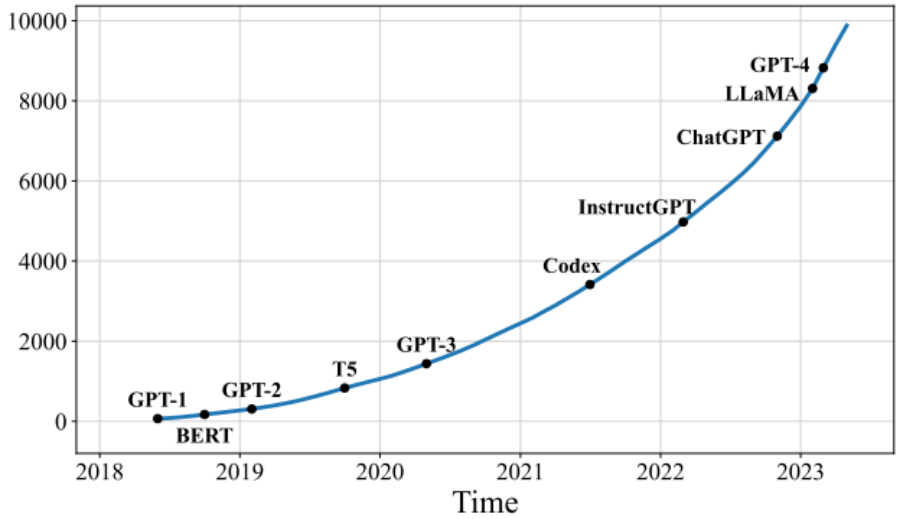




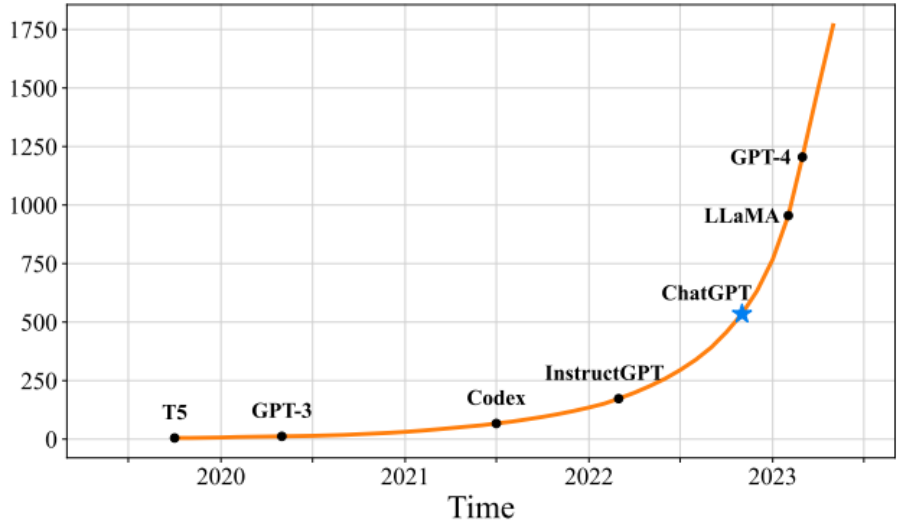
[1] Scaling Laws for Neural Language Models. 2023

Intelligence in Language

- LLMs leading to Human-level AI



(a) Query="Language Model"

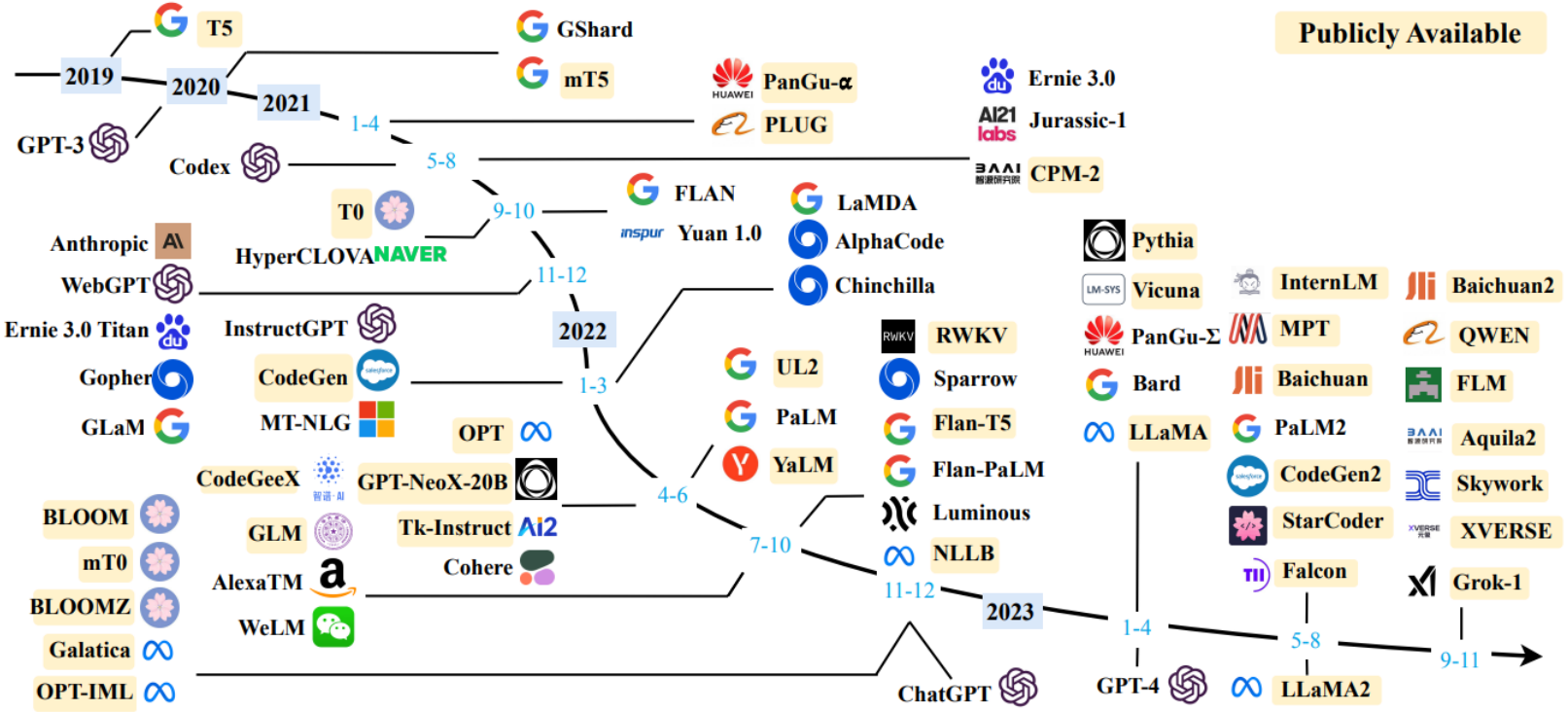


(b) Query="Large Language Model"

[1] A Survey of Large Language Models. https://github.com/RUCAIBox/LLMSurvey, 2023

Intelligence in Language

- Very Rapid Evolvement of Language-based LLMs

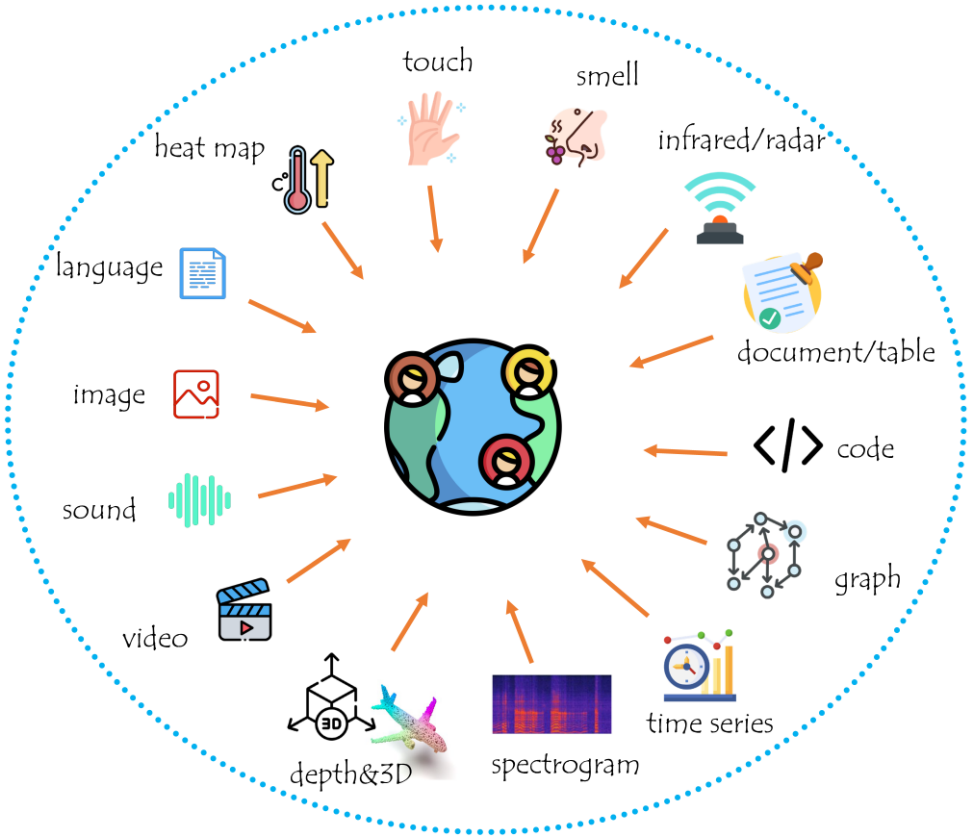


[1] A Survey of Large Language Models. https://github.com/RUCAIBox/LLMSurvey, 2023

Intelligence in Multi-Sensory Data

- ## Harnessing Multimodality

This world we live in is replete with multimodal information & signals, not just language.



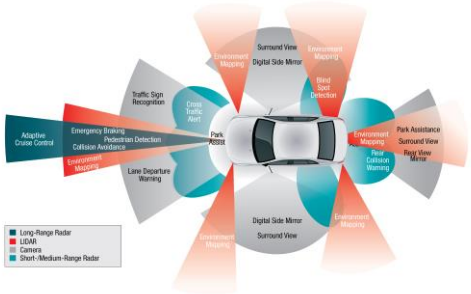
Intelligence in Multi-Sensory Data

Harnessing Multimodality

This world we live in is replete with multimodal information & signals, not just language.

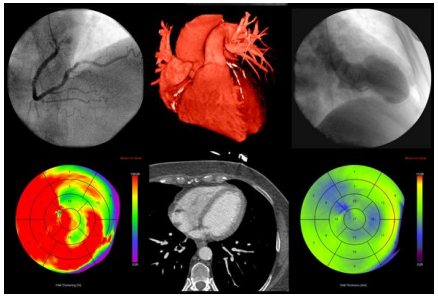
Autonomous Driving Systems

In this application, vehicles use a combination of visual data (cameras), spatial data (LiDAR), and auditory signals (sonar) to navigate safely.



Healthcare Diagnostics

*Medical **imaging** tools like MRI, CT scans, and X-rays, along with patient history and verbal symptoms, are used to diagnose diseases.*



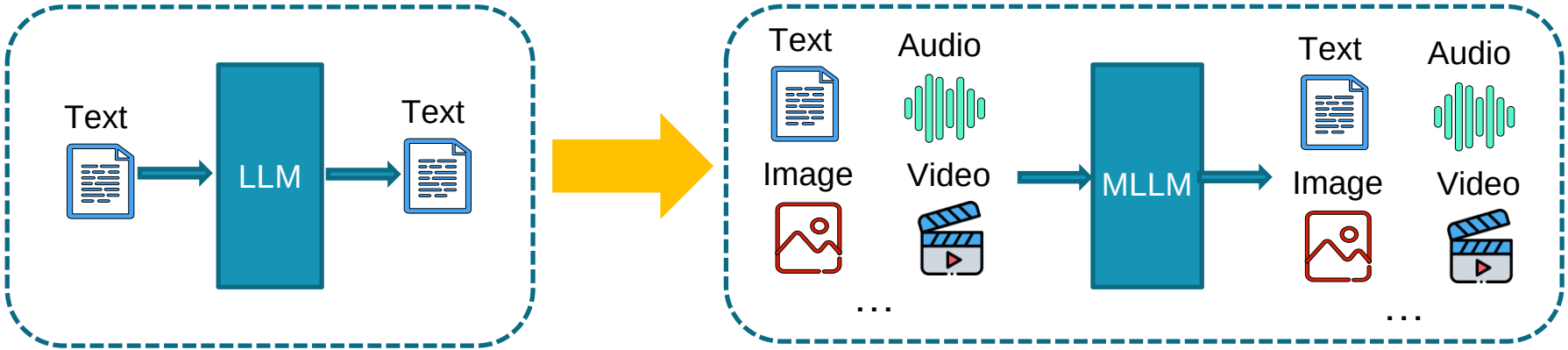
Smart Home Assistants

*Devices like Amazon Alexa and Google Home use voice commands (**audio**), physical interaction (**touch**), and sometimes **visual** cues to operate.*

Intelligence in Multi-Sensory Data

- ## Building Multimodal LLMs (MLLMs)

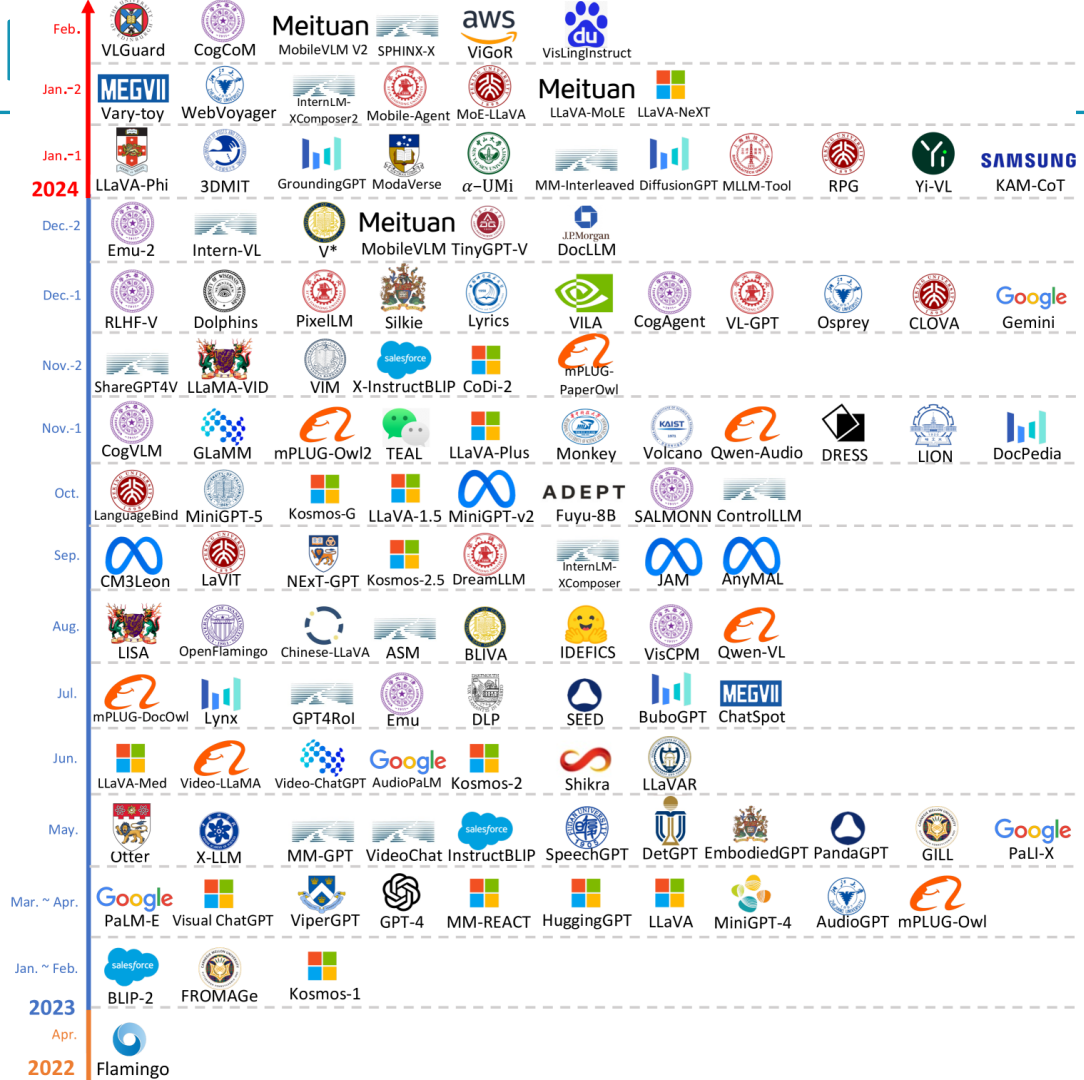
*Can we transfer the success of **LLMs** to **MLLMs**, enabling LLMs to comprehend multimodal information as deeply as they understand language?*



*Perceiving and interacting with the world as **HUMAN BEINGs** do, might be the key to achieving human-level AI.*

Intelligence in

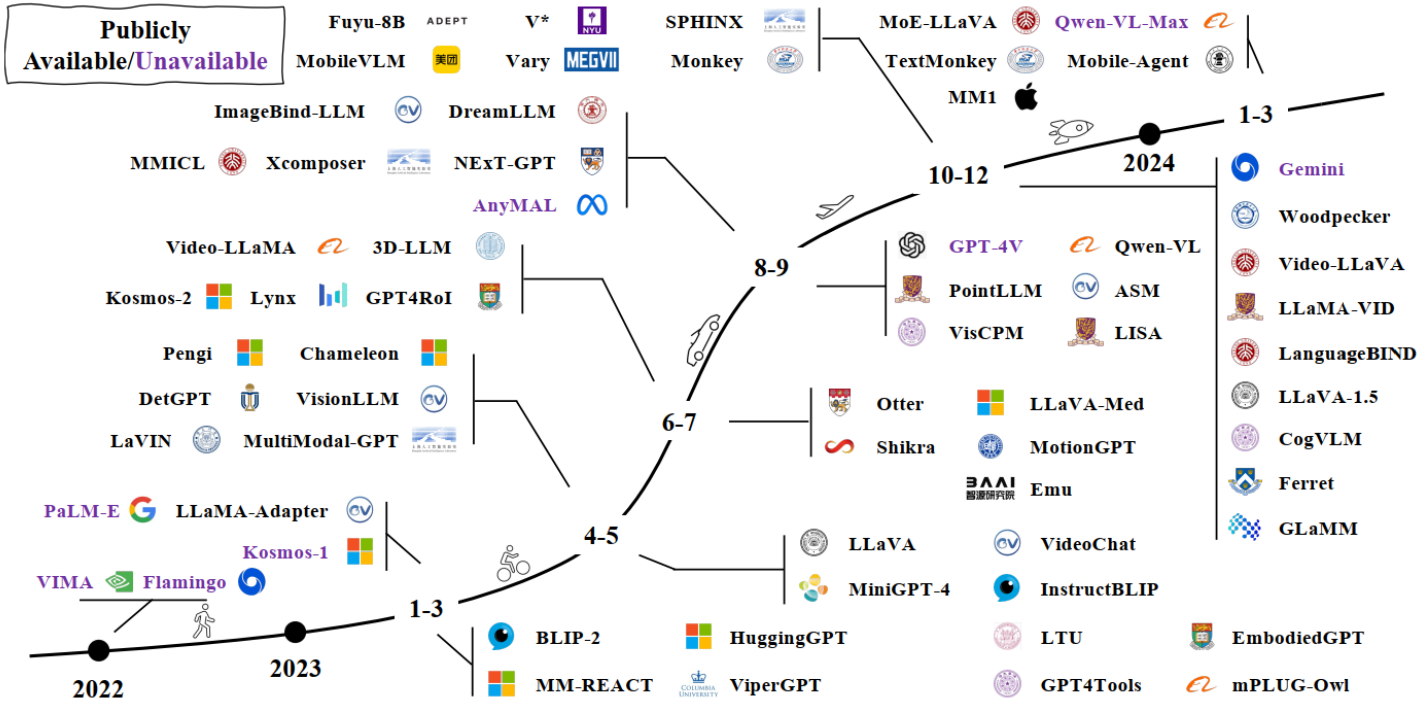
- Trends of MLLMs



[1] MM-LLMs: Recent Advances in MultiModal Large Language Models, 2023.

Intelligence in Multi-Sensory Data

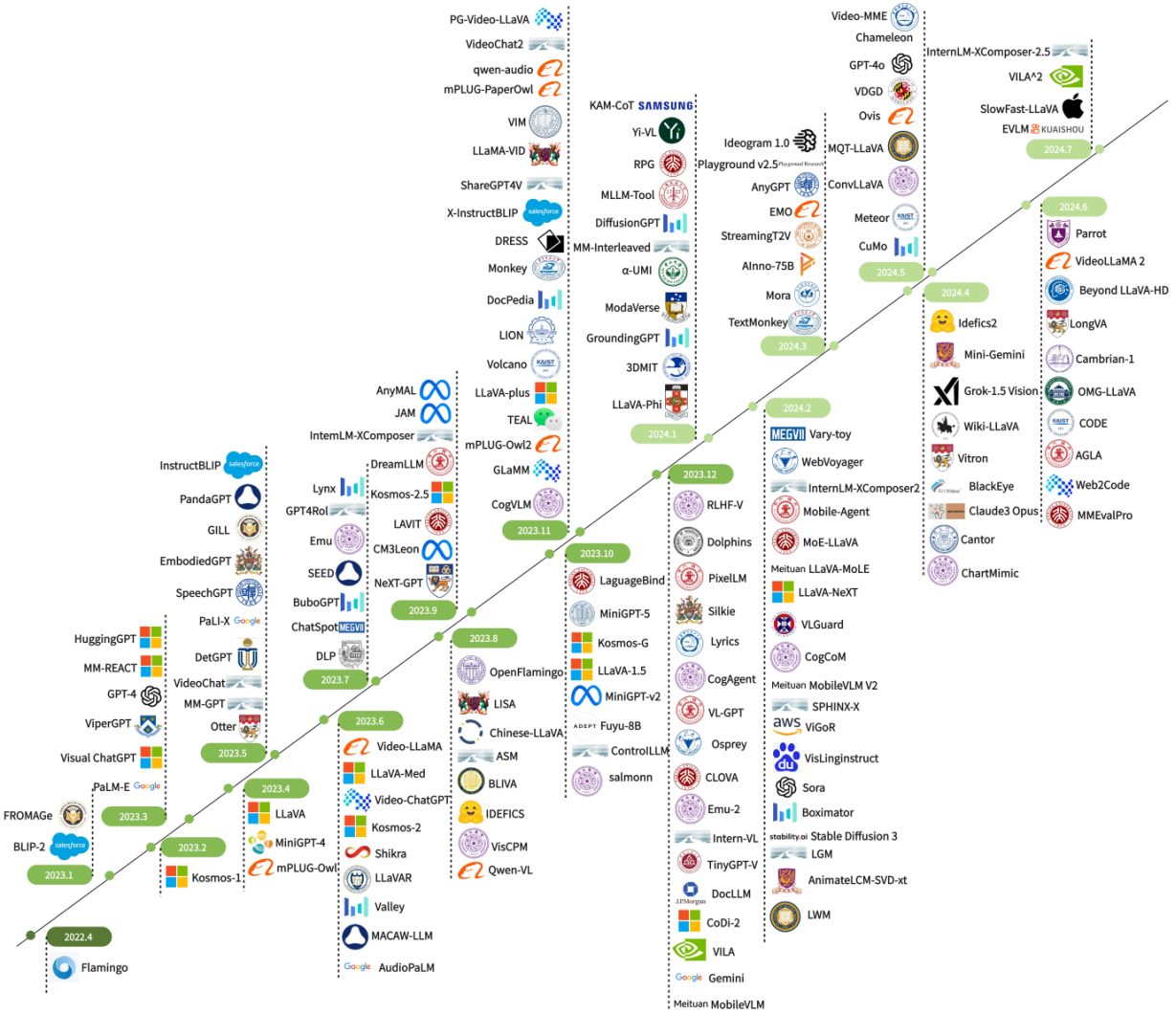
- Trends of MLLMs



[1] A Survey on Multimodal Large Language Models. https://github.com/BradyFU/Awesome-Multimodal-Large-Language-Models, 2023.

Intelligence ii

- Trends of MLLMs



[1] A Comprehensive Review of Multimodal Large Language Models: Performance and Challenges across Different Tasks, 2023.

From MLLMs to Human-level AI

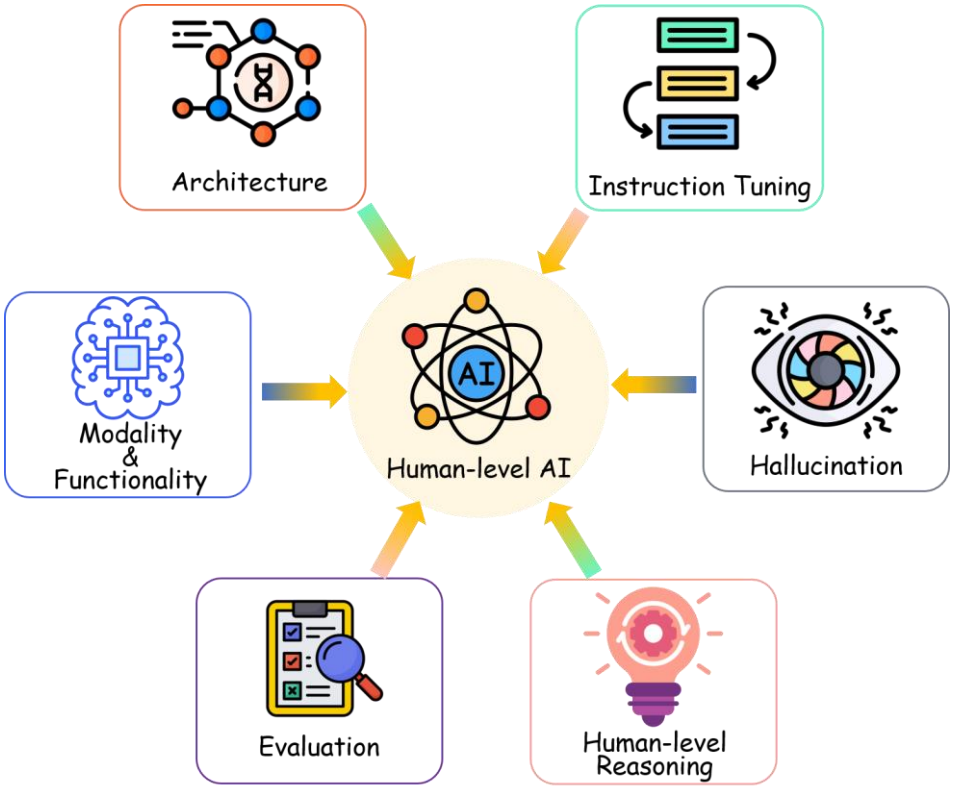
Goal of This Tutorial

- **What are now?**
  - *Walking through the recent key techniques on MLLM constructions in terms of the several key aspects.*
  - *Taxonomies of existing research.*

- **Where to go next?**
  - *Key insights, current challenges & open problems.*
  - *Sparking promising directions for tackling complex reasoning tasks.*
  - *How to build next generation MLLMs?*

From MLLMs to Human-level AI

- Four Key Aspects for Building Powerful MLLMs



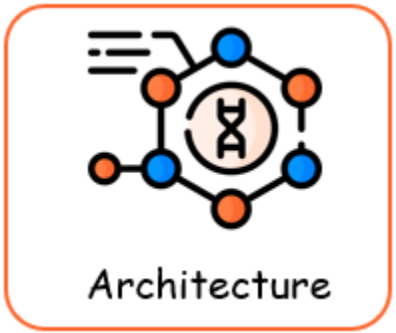
From MLLMs to Human-level AI

- Part 2

☞ MLLM Architecture&Modality

Hao Fei

National University of Singapore



"What is the current architecture of MLLMs? What modalities do they support? How can MLLMs be categorized?"

9:05-9:35

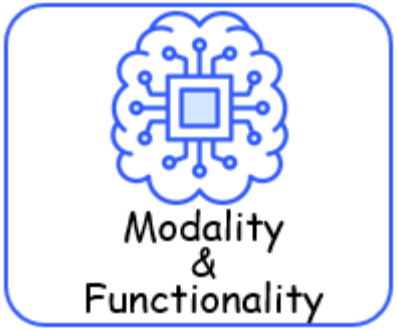
From MLLMs to Human-level AI

- Part 3

☞ MLLM Functionality&Advances

Xiangtai Li

ByteDance/Tiktok



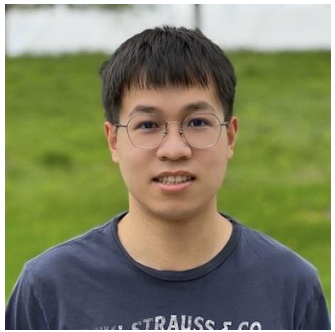
"What functionalities can MLLMs support? How are the current advances of the MLLM community?"

9:35-10:00

From MLLMs to Human-level AI

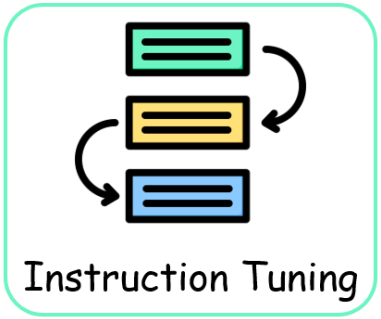
- Part 4

MLLM Instruction Tuning



Haotian Liu

xAI



"Why do we need Multimodal Instruction Tuning ? What are the training strategies of Multimodal Instruction Tuning? How can we get the high-quality data for the instruction tuning? What 's the challenge of the current Multimodal Instruction Tuning?"

10:00-10:30

From MLLMs to Human-level AI

- Part 5

☞ Multimodal Hallucination

Fuxiao Liu

University of Maryland, College Park

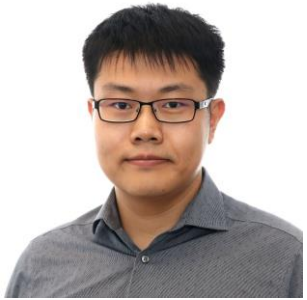
"Why do there will be Multimodal Hallucination? What are the commonly occurred Hallucination? How to alleviate Hallucination?"

11:00-11:25

From MLLMs to Human-level AI

- Part 6

☞ MLLM Evaluation



Hanwang Zhang
Nanyang Technological University

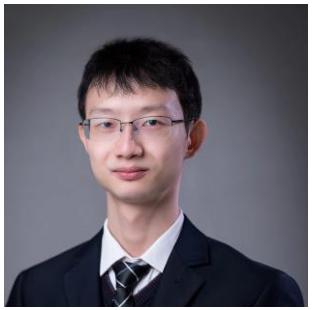
"What are the current common evaluation methods and approaches for MLLMs? What datasets are used? What are the shortcomings and weaknesses of these benchmarks? How can the scientific evaluation of MLLMs be a true path towards AGI?"

11:25-11:50

From MLLMs to Human-level AI

- Part 7

☞ Multimodal Reasoning in MLLMs



Zhuosheng Zhang

Shanghai Jiao Tong University

"What are the latest developments in multimodal reasoning? How does stepwise chain-of-thought reasoning enhance multimodal reasoning? In what ways do multimodal LLM agents improve the ability to solve complex problems? What are the remaining key challenges in advancing multimodal reasoning?"

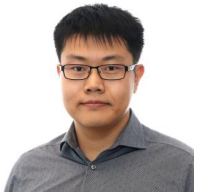
11:50-12:10

From MLLMs to Human-level AI

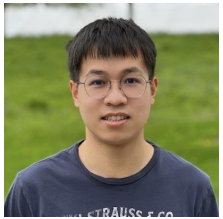
- Part 8

Panel Discussion - From MM Generalist to Human-level AI 12:10-12:30

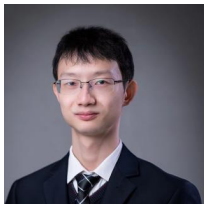
Shuicheng Yan
Kunlun 2050 Research, Skywork AI



Hanwang Zhang
Nanyang Technological University



Haotian Liu
xAI



Zhuosheng Zhang
SJTU

Kaipeng Zhang
Shanghai AI Lab

Hao Fei
NUS

Xiangtai Li
ByteDance/Tiktok

Fuxiao Liu
UM, College Park

From MLLMs to Human-level AI

Schedule Overview

- **Monday, 28 October, 2024**, **9:00-12:30** Melbourne VIC Local Time (**UTC/GMT +11**)

Time	Section	Presenter
9:00-9:05	Part 1: Background and Introduction	Hao Fei
9:05-9:35	Part 2: MLLM Architecture&Modality	Hao Fei
9:35-10:00	Part 3: MLLM Functionality&Advances	Xiangtai Li
10:00-10:30	Part 4: MLLM Instruction Tuning	Haotian Liu
	Coffee Break	
11:00-11:25	Part 5: Multimodal Hallucination	Fuxiao Liu
11:25-11:50	Part 6: MLLM Evaluation&Generalist	Hanwang Zhang
11:50-12:10	Part 7: Multimodal Reasoning & Agent	Zhuosheng Zhang
12:10-12:30	Part 8: Panel Discussion Q&A Session	All + Shuicheng Yan

From MLLMs to Human-level AI

- ## Contact & QA & Discussions

  - All slides and reading list are available at tutorial homepage:

    https://mllm2024.github.io/ACM-MM2024/

  - We welcome all Q&A and discussions via Google Group:

    - *Post your questions on Google Group:*

      https://groups.google.com/g/mllm24

    - *Email us:*

      mllm24@googlegroups.com