

# From Multimodal LLM to Human-level AI

***Modality, Instruction, Reasoning, Efficiency and Beyond***



<https://mllm2024.github.io/COLING2024>

LREC-COLING 2024





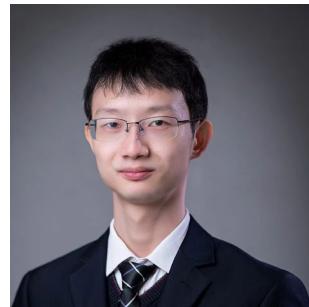
**Hao Fei**

*National University of Singapore*



**Yuan Yao**

*National University of Singapore*



**Zhuosheng Zhang**

*Shanghai Jiao Tong University*



**Fuxiao Liu**

*University of Maryland, College Park*



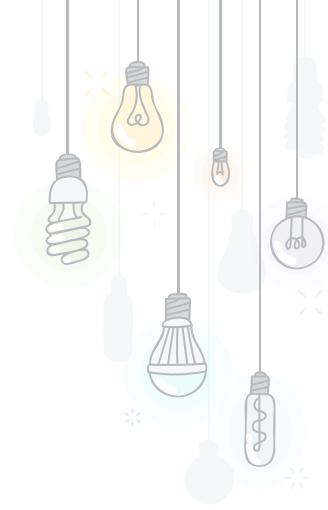
**Ao Zhang**

*National University of Singapore*



**Tat-Seng Chua**

*National University of Singapore*



# ✿ Part-I

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

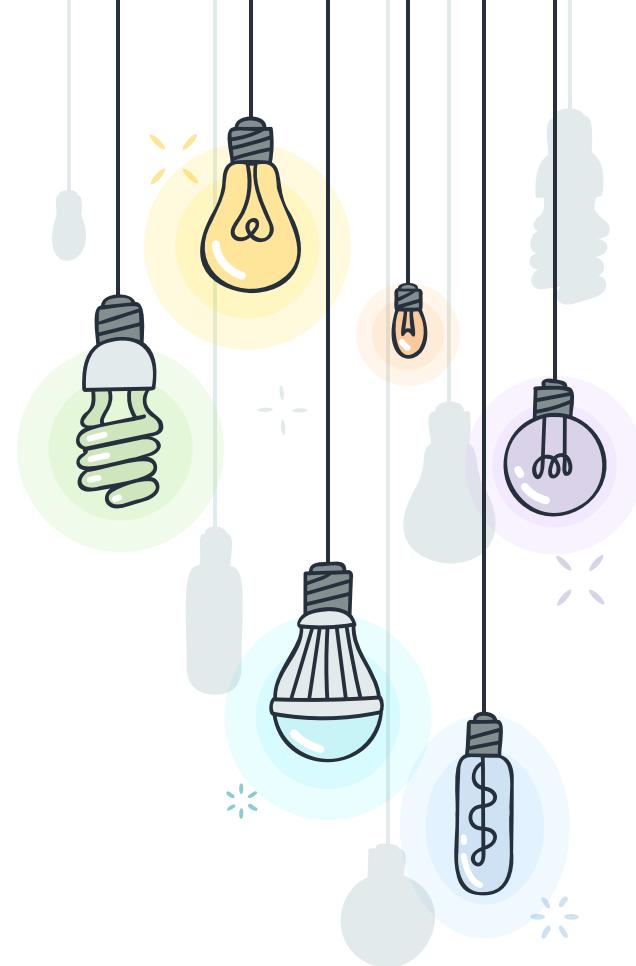


Hao Fei

Research Fellow

*National University of Singapore*

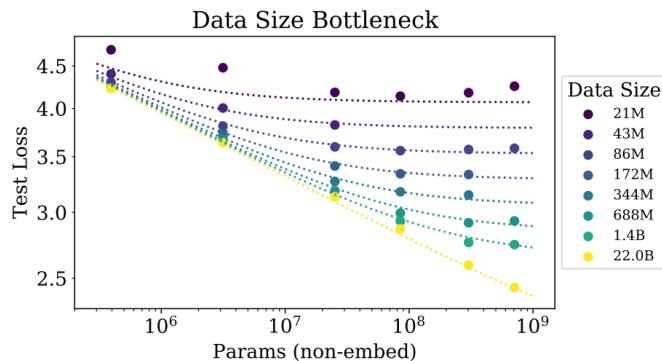
<http://haofei.vip/>



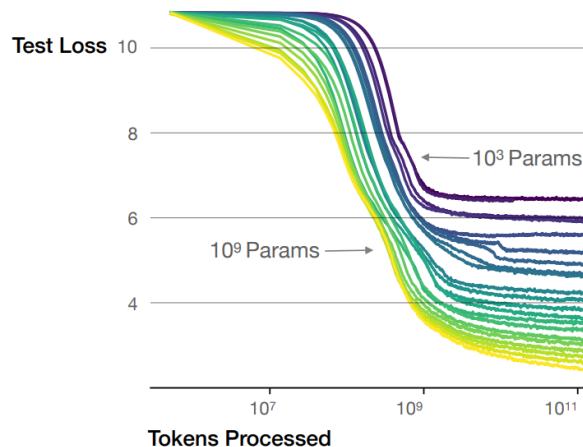
# \* Intelligence in Language

- Scaling Law in Neural Models

 Model size ↑, Data size ↑



Larger models require **fewer samples** to reach the same performance



# \* Intelligence in Language

- LLMs leading to Human-level AI



*Unprecedented in-depth language semantic understanding capability.*

The image shows a dark-themed screenshot of a messaging interface. On the left, there's a circular profile picture with the letters 'HA'. To its right, the word 'You' is displayed in a small, white, sans-serif font. Below this, a message is shown in a light gray box: 'What am I really trying to say when I tell my friend "Go try that restaurant!"?' On the far left of the main conversation area, there's another circular profile picture with a stylized 'G'. To its right, the word 'ChatGPT' is displayed in a small, white, sans-serif font. Below this, a message is shown in a light gray box: 'When you tell your friend "Go try that restaurant!", you're conveying a few different things:'

**1. Recommendation:** You're suggesting that the restaurant is worth visiting, implying that you believe the food, service, or atmosphere is good.

**2. Encouragement:** You're encouraging your friend to have a new experience, possibly one that you think they'll enjoy.

**3. Sharing Experiences:** If you've already visited the restaurant, you might be sharing something you enjoyed, hoping your friend will also appreciate it. If you haven't been, it could be a suggestion to explore something new together or independently.

**4. Trust and Influence:** Your recommendation reflects your taste and judgement, and it shows that you're willing to influence their decision based on your own preferences or experiences.

Essentially, it's a friendly way to enhance your friend's dining options based on your positive assessment or curiosity about the restaurant.

# \* Intelligence in Language

- LLMs leading to Human-level AI



*Numerical analysis capability.*

 **You**

The midterm exam results are out. Tom's scores in Mathematics, Art and English are 89, 90, and 97, respectively; Jason's scores in Mathematics, Art and English are 90, 87, and 92, respectively. Please tabulate these students' scores, calculate the total scores, and then sort them by the total score.

 **ChatGPT**

Here are the scores for Tom and Jason in tabular form, including their totals for each subject and sorted by the total scores:

|       | English | Mathematics | Art |
|-------|---------|-------------|-----|
| Tom   | 97      | 89          | 90  |
| Jason | 92      | 90          | 87  |
| Total | 189     | 179         | 177 |

From the table, it's clear that English has the highest combined score, followed by Mathematics and then Art. [-]

⏮ ⏴ ⏵ ⏹

- LLMs leading to Human-level Intelligence



*Mathematical reasoning ability.*

HA

**You**

It is known that the price of one desk is ten times that of one armchair. It is also known that one desk is 378.9 dollar more expensive than one armchair. How much does one desk and one armchair each cost?

**ChatGPT**

Let's set up the problem and solve it:

1. Let the price of the armchair be  $x$  dollars.
2. Then, the price of the desk is  $10x$  dollars.
3. It is given that the desk is 378.9 dollars more expensive than the armchair.

Thus, the equation based on the above information is:

$$10x = x + 378.9$$

We can solve this equation for  $x$ :

$$10x - x = 378.9$$

$$9x = 378.9$$

$$x = \frac{378.9}{9}$$

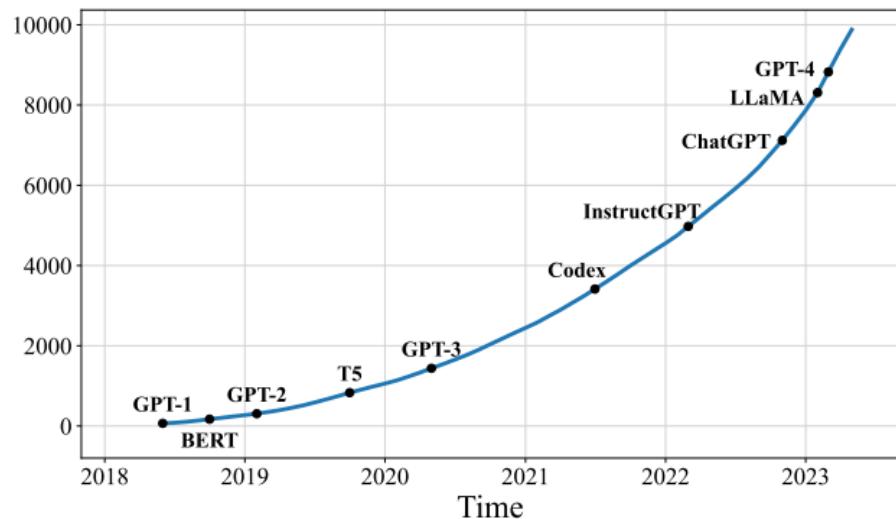
$$x = 42.1$$

Therefore:

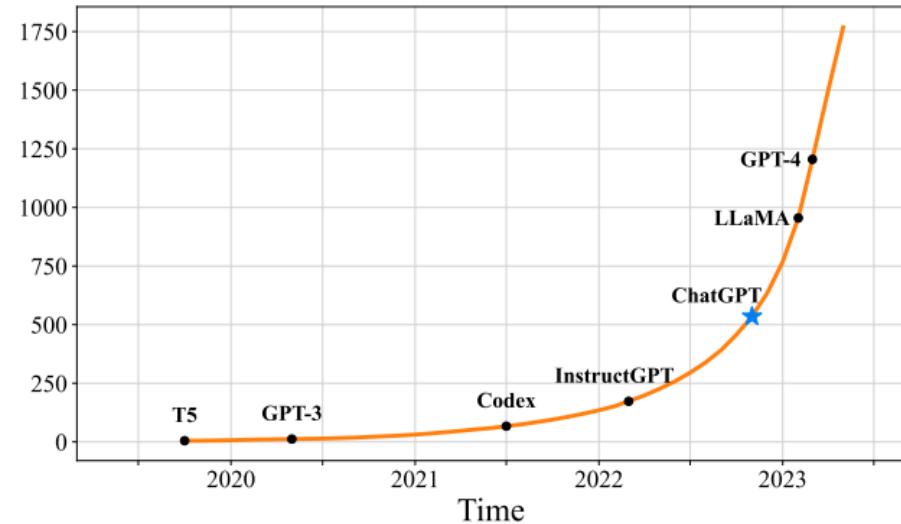
- The price of the armchair is 42.1 dollars.
- The price of the desk is  $10 \times 42.1 = 421$  dollars.

# \* Intelligence in Language

- LLMs leading to Human-level AI



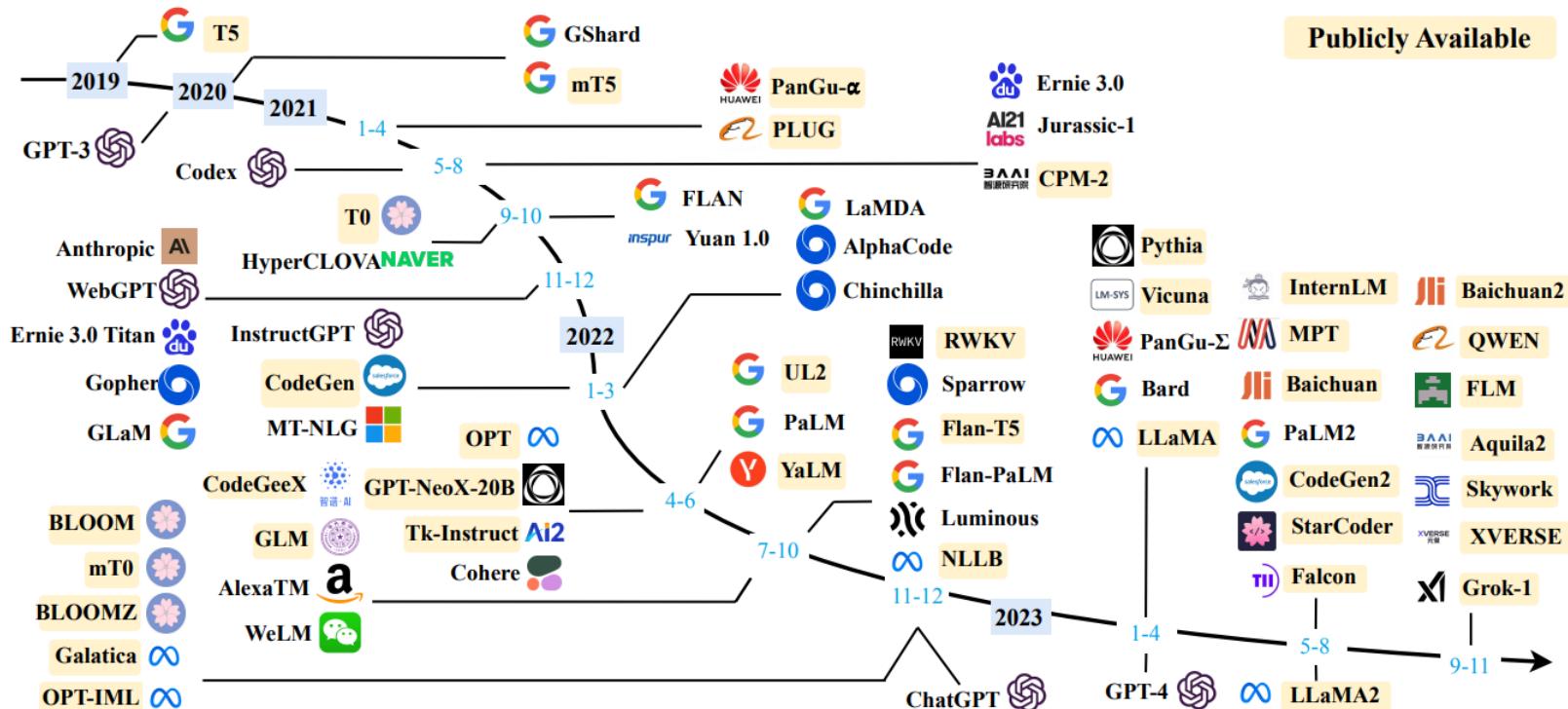
(a) Query="Language Model"



(b) Query="Large Language Model"

# Intelligence in Language

- Very Rapid Evolution of Language-based LLMs

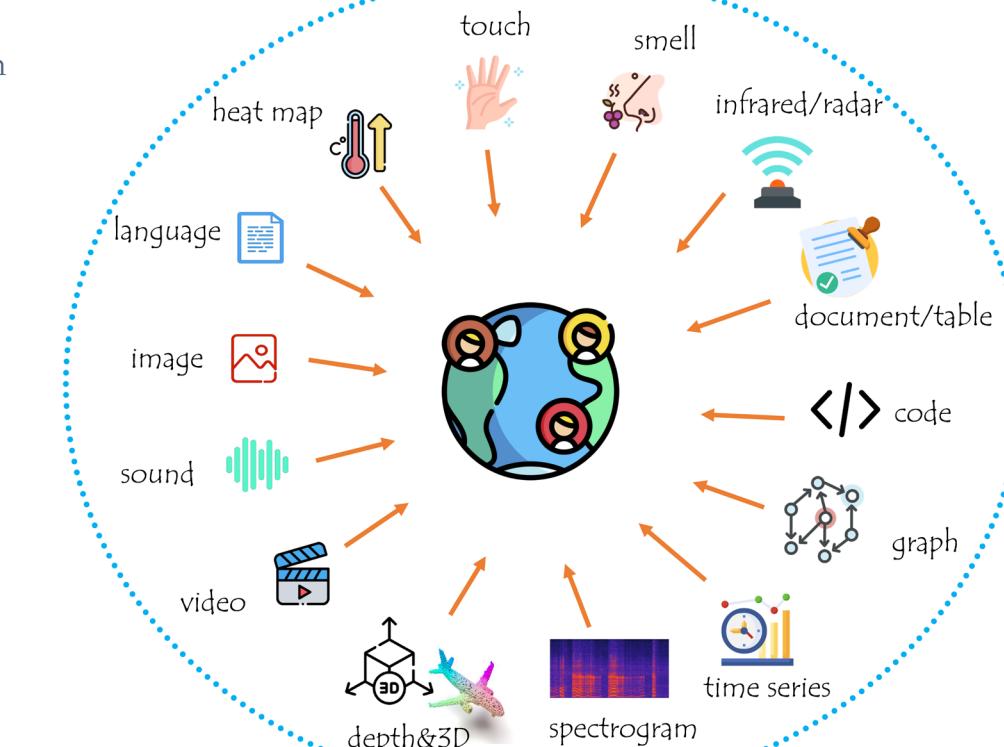


# \* Intelligence in Multi-Sensory Data

- Harnessing Multimodality



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



# \* Intelligence in Multi-Sensory Data

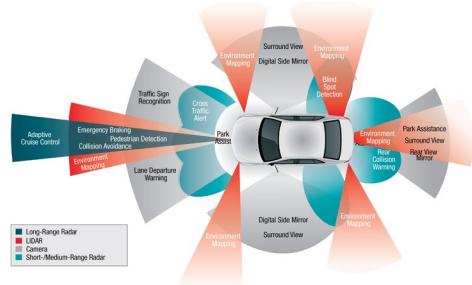
## • Harnessing Multimodality



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

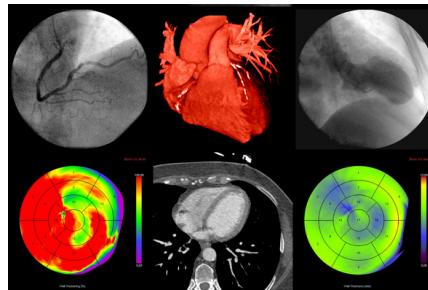
### + Autonomous Driving Systems

*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

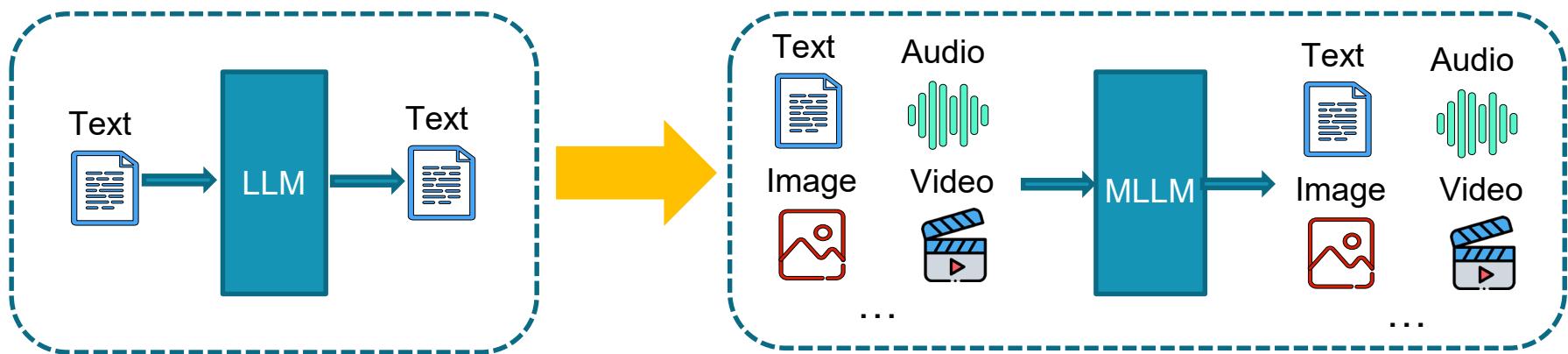
*Devices like Amazon Alexa and Google Home use voice commands (**audio**), physical interaction (**touch**), and sometimes **visual cues** to operate.*



# \* Intelligence in Multi-Sensory Data

- Building Multimodal LLMs (MLLMs)

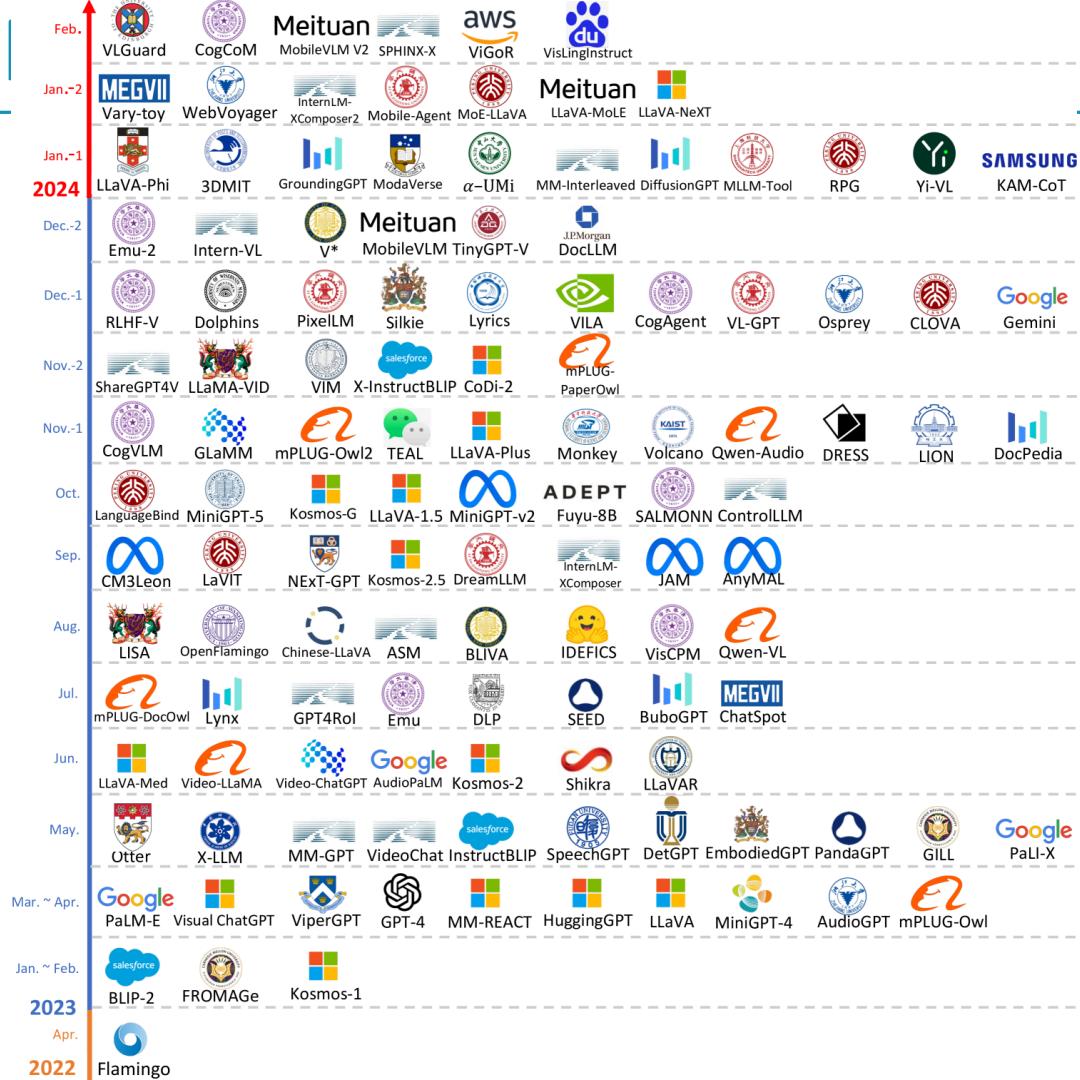
👉 *Can we transfer the success of **LLMs** to **MLLMs**, enabling LLMs to comprehend multimodal information as deeply as they understand language?*



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

# Intelligence in |

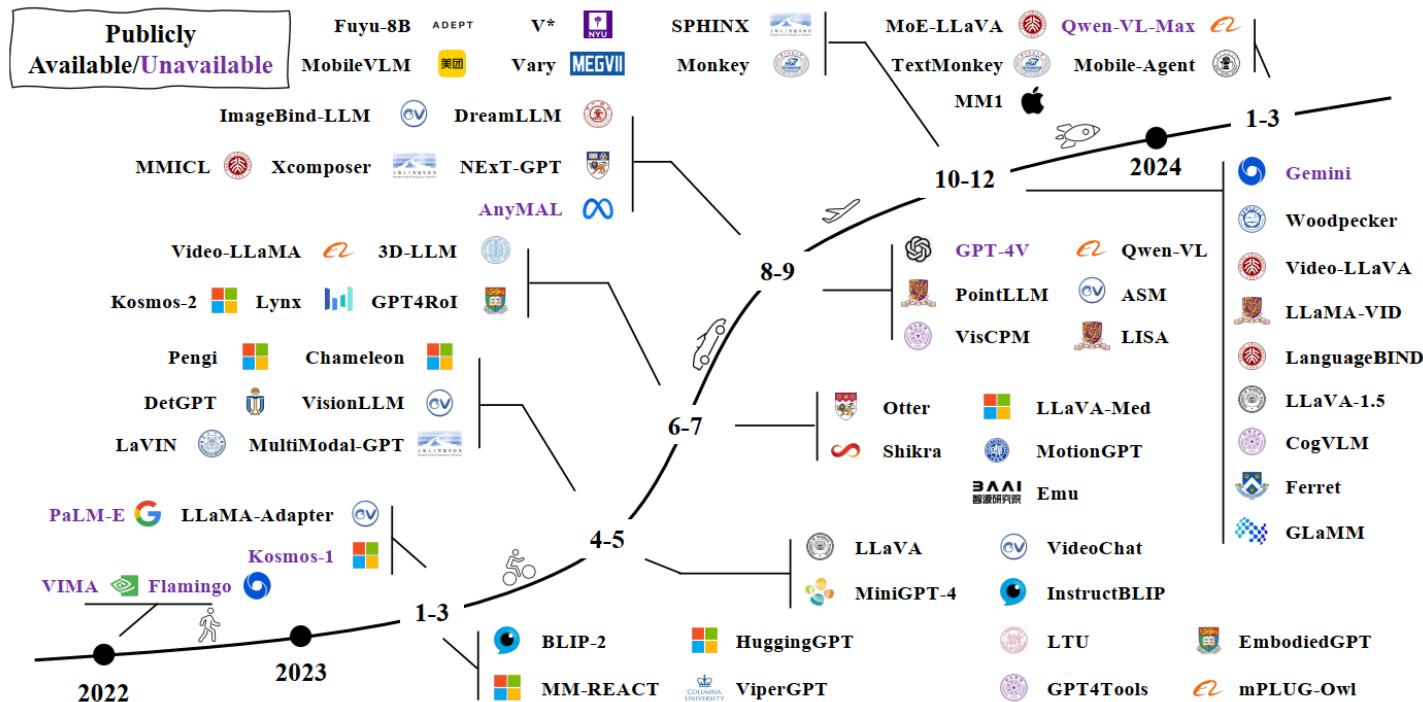
- Trends of MLLMs



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

# Intelligence in Multi-Sensory Data

## Trends of MLLMs



# ✳️ From MLLMs to Human-level AI

---

- Goal of This Tutorial

- + What are now?

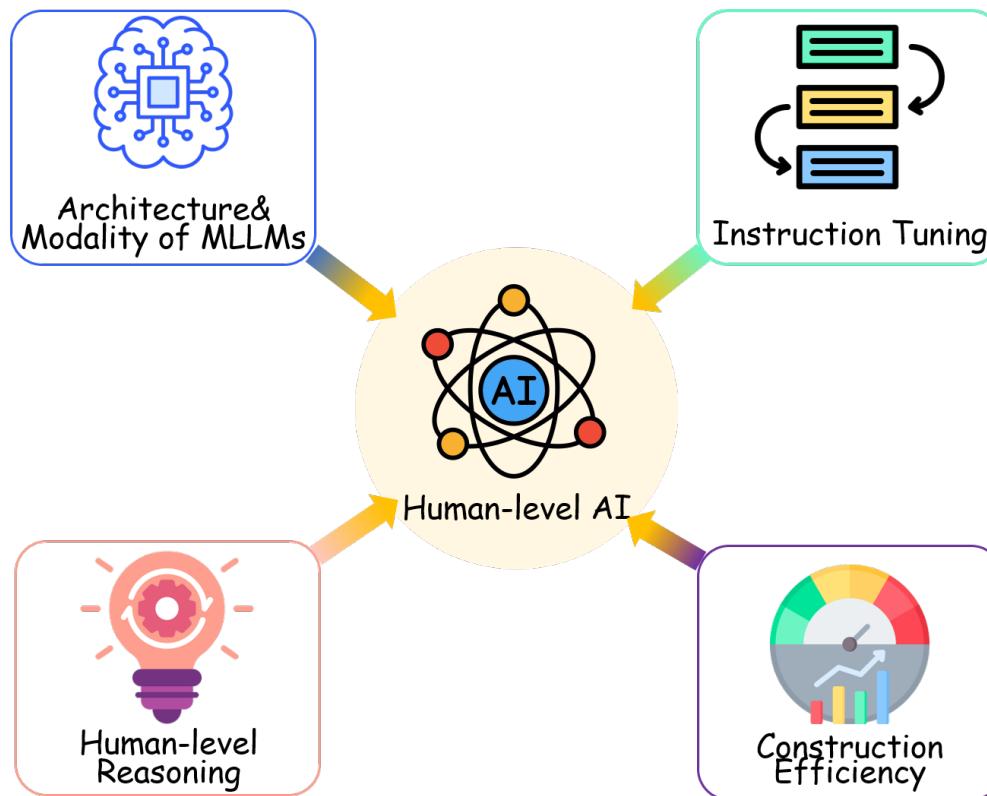
- + Walking through the recent key techniques on MLLM constructions in terms of the **four key aspects**.
  - + Taxonomies of existing research.

- + 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

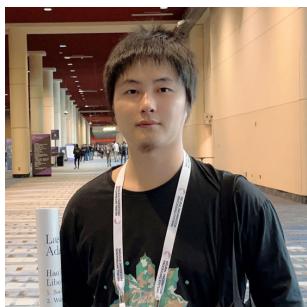


# From MLLMs to Human-level AI

- Aspect-A



## MLLM Design: Architecture and Modality



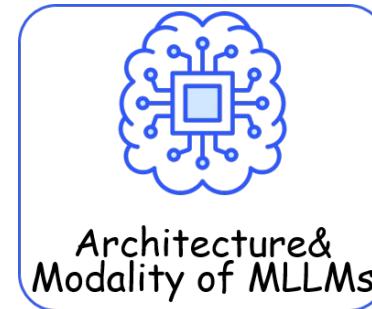
Hao Fei

National University of Singapore



Yuan Yao

National University of Singapore



Architecture &  
Modality of MLLMs

*“What is the current architecture of MLLMs? What modalities and functionalities do MLLMs support? How can MLLMs be categorized? What will the next generation of MLLMs look like?”*

Tuesday, May 21, 2024  
14:10-15:40 Local Time

# ✳️ From MLLMs to Human-level AI

- Aspect-B

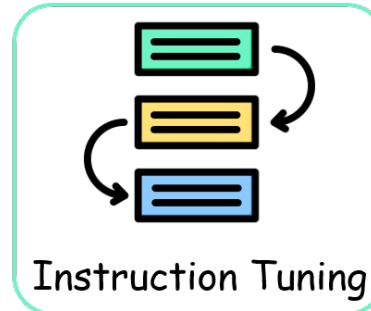


## Multimodal Instruction Tuning in MLLMs



Fuxiao Liu

*University of Maryland, College Park*



*“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?”*

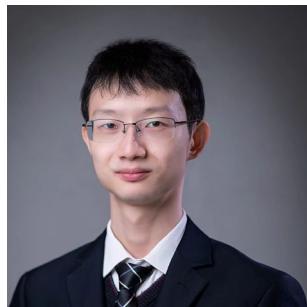
Tuesday, May 21, 2024  
15:40-16:00 & 16:30-16:50 Local Time

# From MLLMs to Human-level AI

- Aspect-C



## 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?”*

Tuesday, May 21, 2024  
16:50-17:30 Local Time

# From MLLMs to Human-level AI

- Aspect-D



MLLM Efficiency



Ao Zhang

*National University of Singapore*



*“What is the most efficient MLLM architecture to achieve high performance? How to choose and organize the data to build a powerful MLLM? Are there training strategies to build new MLLMs or extend function scope efficiently?”*

Tuesday, May 21, 2024  
17:30-18:00 Local Time

# From MLLMs to Human-level AI

## Schedule Overview

- Tuesday, May 21, 2024
- Torino, Italia
- 14:00-18:00 Local Time

| Time        | Section   | Presenter          |
|-------------|---|--------------------|
| 14:00-14:10 | Part 1: Background and Introduction                     | Hao Fei            |
| 14:10-15:40 | Part 2: MLLM Design: Architecture and Modality          | Hao Fei & Yuan Yao |
| 15:40-16:00 | Part 3: Multimodal Instruction Tuning in MLLMs          | Fuxiao Liu         |
|             | Coffee Break, Q&A Session                               |                    |
| 16:30-16:50 | Part 3 (Cont'd): Multimodal Instruction Tuning in MLLMs | Fuxiao Liu         |
| 16:50-17:30 | Part 4: Multimodal Reasoning in MLLMs                   | Zhuosheng Zhang    |
| 17:30-18:00 | Part 5: MLLM Efficiency                                 | Ao Zhang           |

# From MLLMs to Human-level AI

- Contact & QA & Discussions

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

<https://mllm2024.github.io/COLING2024>



- + 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](mailto:mllm24@googlegroups.com)