

Welcome to COSE 405 (CS Colloquium)!

김진규

고려대학교 컴퓨터학과



고려대학교
KOREA UNIVERSITY



Course Logistics

English? Korean?

We will invite various speakers, but almost all of them will use Korean for their presentations...

So, we will use Korean in this class.

Lecture format

There are **one lecture per week**

- Time and place: 16:30 - 18:50, Wed
- Format
 - Invited Talks from industry and academia
 - Paper Review Writing: 1 per week (in the format of top-tier conferences)

Blackboard

Blackboard

- Stay tuned! All announcements will be posted here.
- Where all course materials (lecture slides, assignments, projects) are available

Gradescope

Gradescope (gradescope.com, by invitation only)

- Where all assignments are submitted, and where all of your grades in this course will live.
- If you are not invited by March 15 (Sat), let us know!

Grading

- Assignments (30%)
 - Paper Reviews / Readings
- Class attendance (70%)

Who am I?

Vision and AI Lab (visionai.korea.ac.kr)



PI

Jinkyu Kim
Assistant Professor
2021.03 - Present



Education



Bachelor's

Electrical Engineering
Korea University

Master's

Electrical and Computer Engineering
Korea University

Ph.D.

Computer Science
University of California, Berkeley



Advisor: Prof. Sungroh Yoon



Advisor: Prof. John Canny



Work Experience



Korea University

Assistant Professor
Computer Science
2021.03 - Present



Waymo Research

Research Scientist
Applied Research Team
2020.01 - 2021.02



Honda Research

Intern
HCI Team
2018.05 - 2018.12

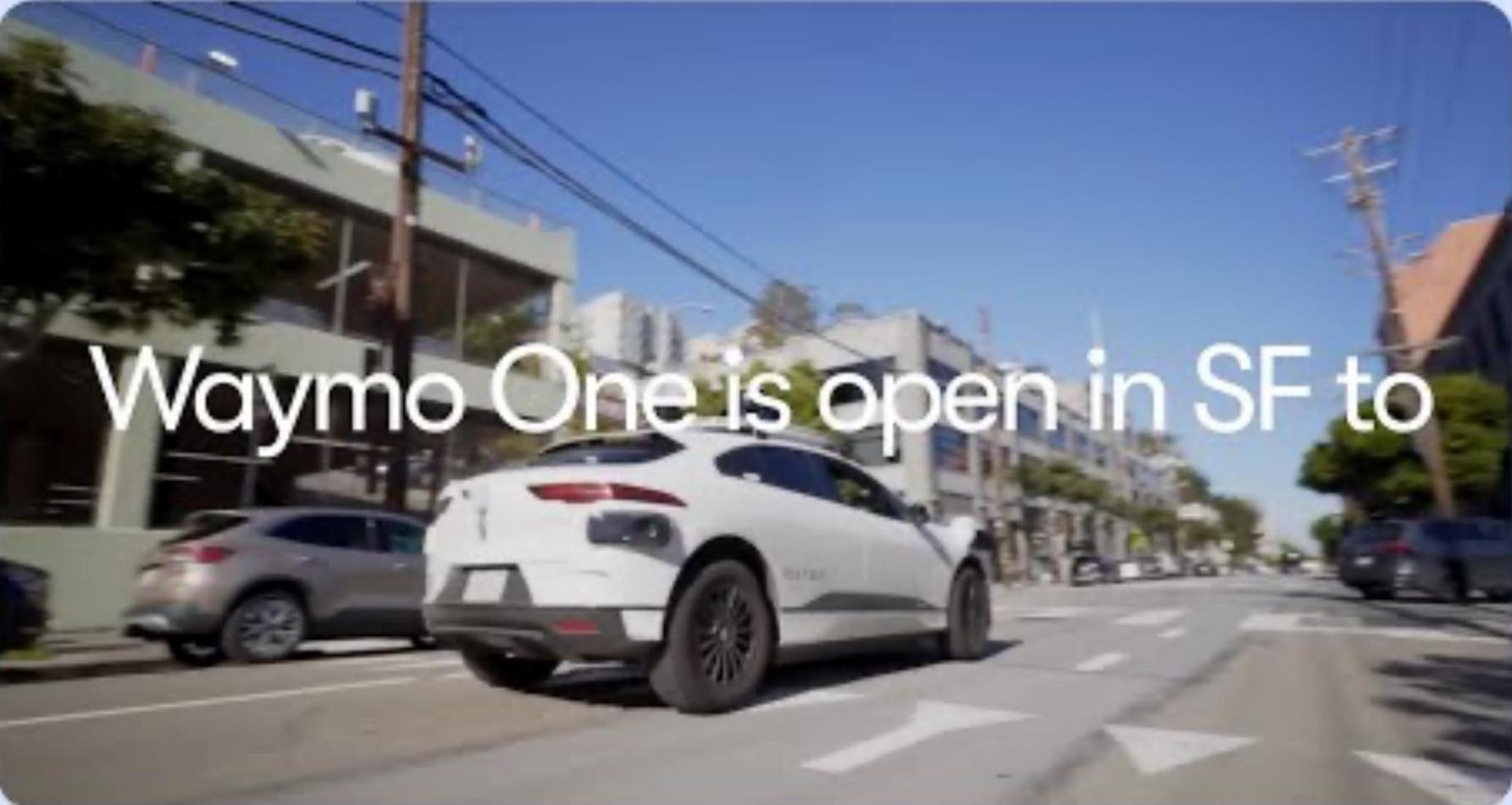


LG Display R&D

Research Engineer
Algorithm Team
2010.11 - 2014.03

CEO Award Winner
(2013)



A white Waymo One self-driving car is shown from a rear three-quarter perspective, driving down a city street. The car has a distinctive dark roof and a light-colored body. It is positioned in the center of the frame, moving towards the right. The background features a clear blue sky, power lines, and buildings typical of a urban environment.

Waymo One is open in SF to

C

WTF



COSE405: CS Colloquium

대규모 멀티모달 언어 모델 (MLM)의 이해와 활용방법 소개

김진규

고려대학교 컴퓨터학과



고려대학교
KOREA UNIVERSITY



Content

- ❑ Transformer and Language Model (OpenAI GPT, GPT-2, GPT-3)
- ❑ Success of GPT-3: Scaling law and In-context Learning
- ❑ ChatGPT: Language Models as Multitask Assistant
- ❑ Hallucinations: Retrieval-augmented Language Models
- ❑ Multimodal LLMs (MLLMs)



Transformer



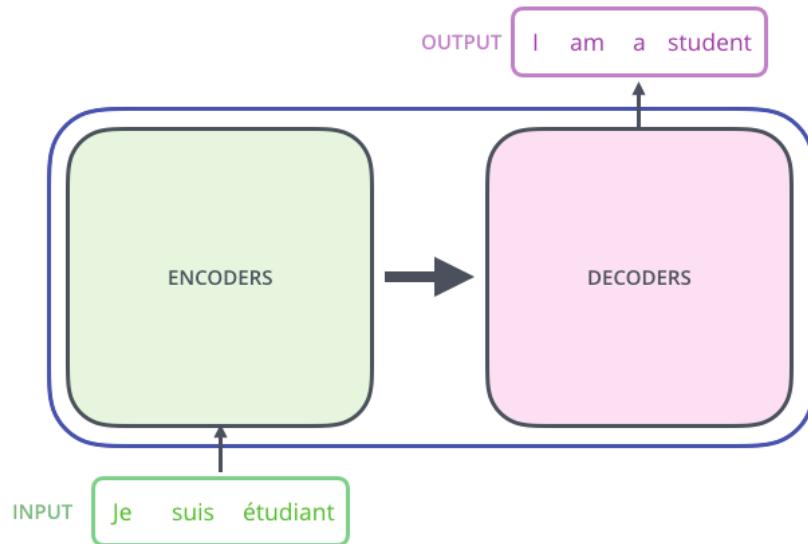
Transformer: A high-level look

- a **sequence-to-sequence** model
 - ▶ input sequence (arbitrary length) → output sequence (arbitrary length)



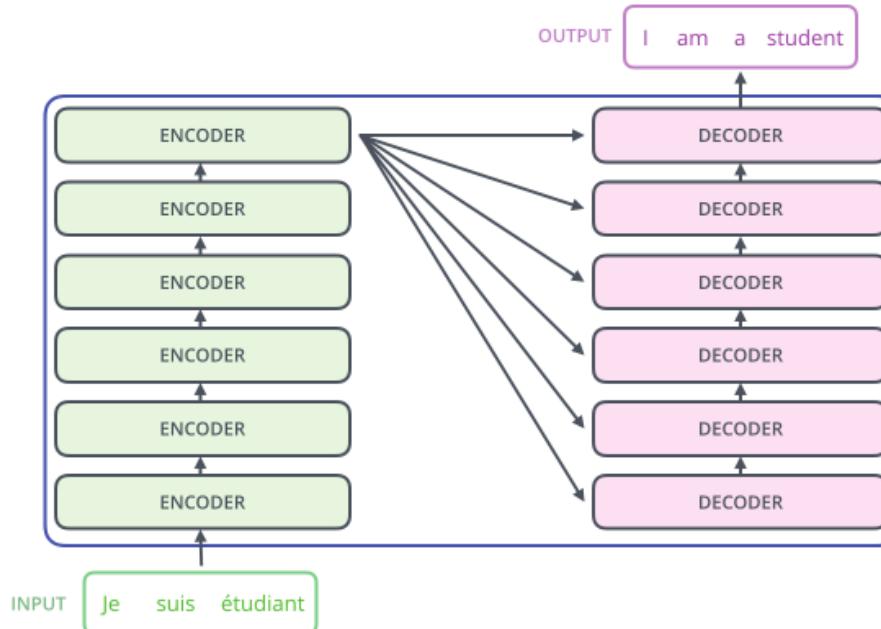
Transformer: A high-level look

- encoder-decoder architecture
 - ▶ an **encoding** component + a **decoding** component
 - ▶ and connections between them



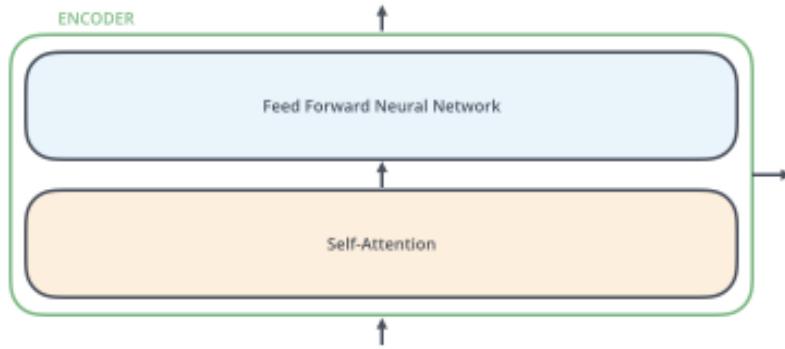
Transformer: A high-level look

- **encoding** component: a stack of encoders⁴
- **decoding** component: a stack of decoders of the same number



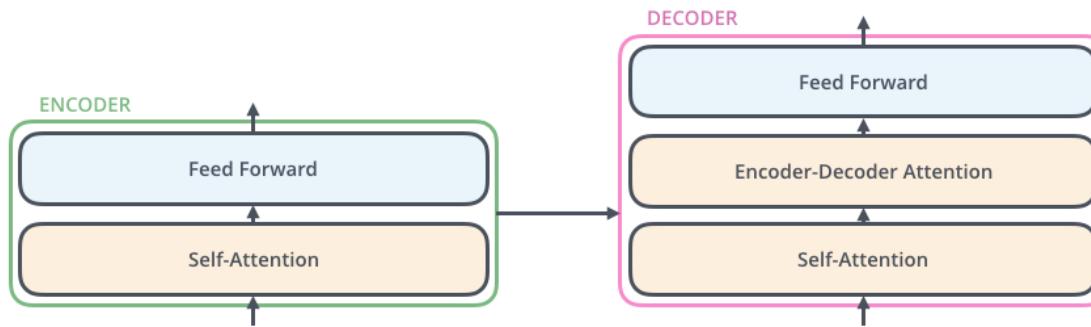
Transformer: Encoder

- encoders: all identical in structure (yet they do not share weights)
 - ▶ each one is broken down into two sub-layers:



Transformer: Decoder

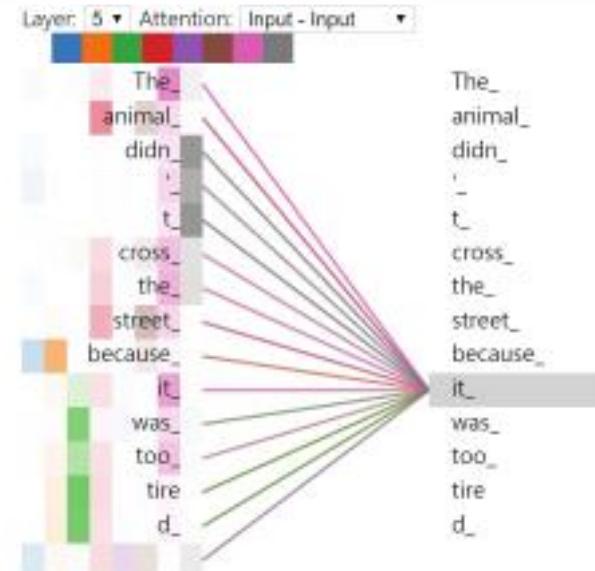
- decoder: has three sub-layers
 - ▶ feed-forward layers
 - ▶ encoder-decoder attention layer
 - ▶ self-attention layer



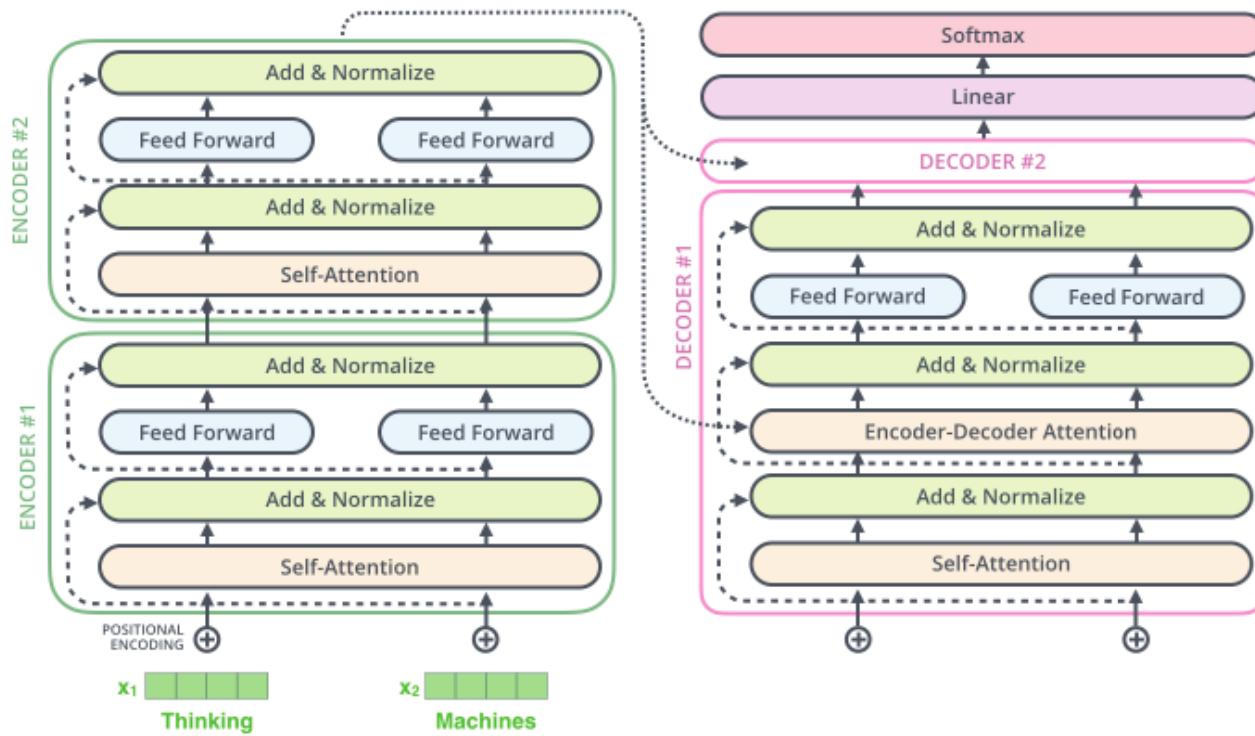
Transformer: Self-Attention

- calculate the output of self-attention layer
 ↑
 weighted sum of value vectors

Input	Thinking		Machines	
Embedding	x_1	[green green green]	x_2	[green green green]
Queries	q_1	[purple purple purple]	q_2	[purple purple purple]
Keys	k_1	[orange orange orange]	k_2	[orange orange orange]
Values	v_1	[blue blue blue]	v_2	[blue blue blue]
Score	$q_1 \cdot k_1 = 112$		$q_1 \cdot k_2 = 96$	
Divide by 8 ($\sqrt{d_k}$)	14		12	
Softmax	0.88		0.12	
Softmax X Value	v_1	[blue blue blue]	v_2	[white white white]
Sum	z_1	[pink pink pink]	z_2	[pink pink pink]



Transformer: A high-level look



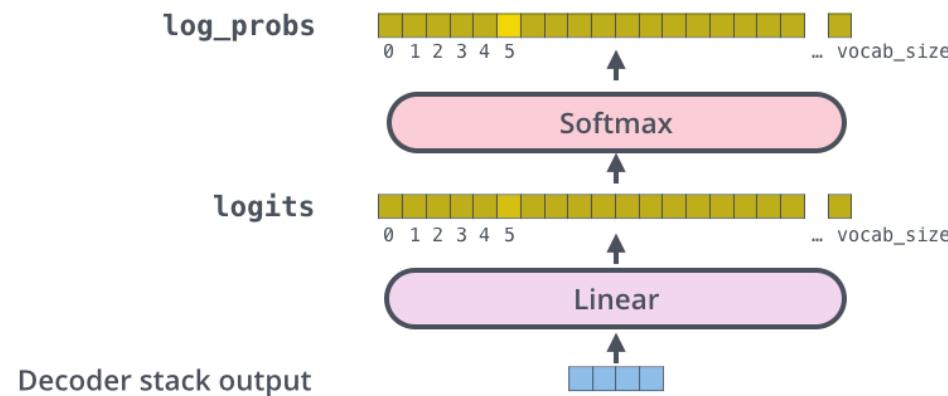
Transformer: Output

Which word in our vocabulary
is associated with this index?

am

Get the index of the cell
with the highest value
(`argmax`)

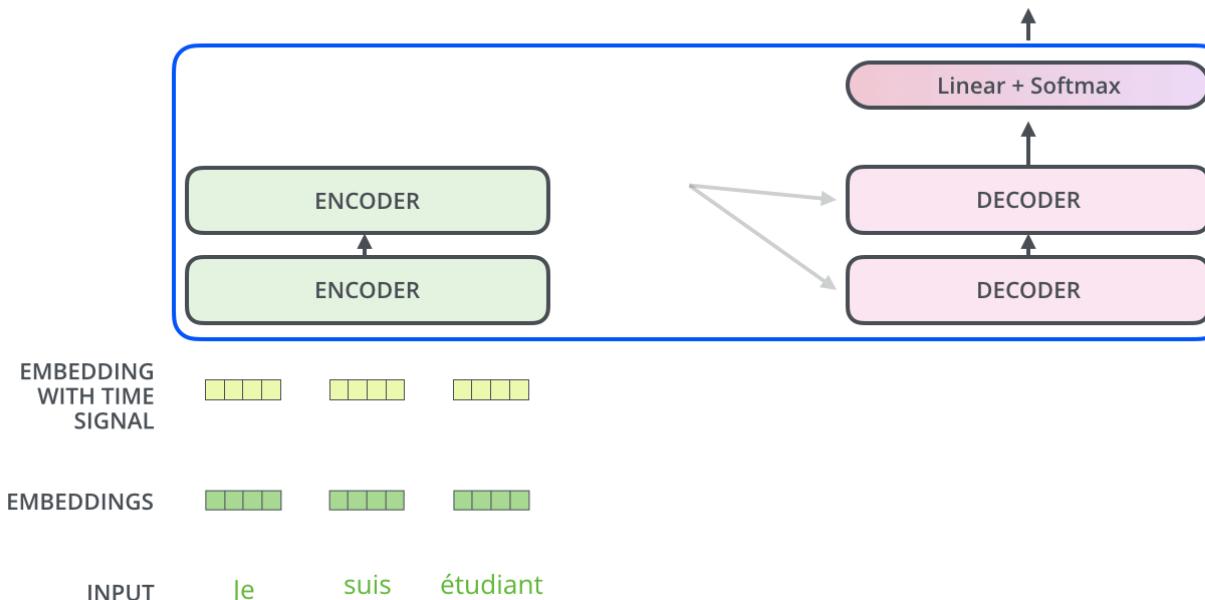
5



Transformer: All Together

Decoding time step: 1 2 3 4 5 6

OUTPUT



Language Model (LM)

- LM predicts the likelihood of any sequence of words, for a given *language corpus**.
 - $p(\text{"This is a fluffy dog."}) = p(w_1 w_2 w_3 w_4 w_5) = 0.132 \dots$
 - $p(\text{"This are a purple flying dear."}) = p(w_1 w_2 w_3 w_4 w_5 w_6) = 0.0002 \dots$

Note:

*Each person / author has their individual LM. English **marginalize** over those.*

LM of Shakespeare

LM of Trump



Language Model (LM)

- Given such model, we can use conditional probability to
 - Guess missing words
 - "This _ a fluffy dog."
 - $\arg \max p(w_2 | w_1 w_3 w_4 w_5) \rightarrow "is"$
 - Predict next words
 - "This is a fluffy ..."
 - $\arg \max p(w_5 | w_1 w_2 w_3 w_4) \rightarrow "cat"$
 - Answer questions
 - "The Vatican locates in the city of ..."
 - $\arg \max p(w_7 | w_1 w_2 w_3 w_4 w_5 w_6) \rightarrow "Rome"$

Remark:

By modelling statistics, language model implicitly learns grammar, semantics, common sense, factual knowledge...

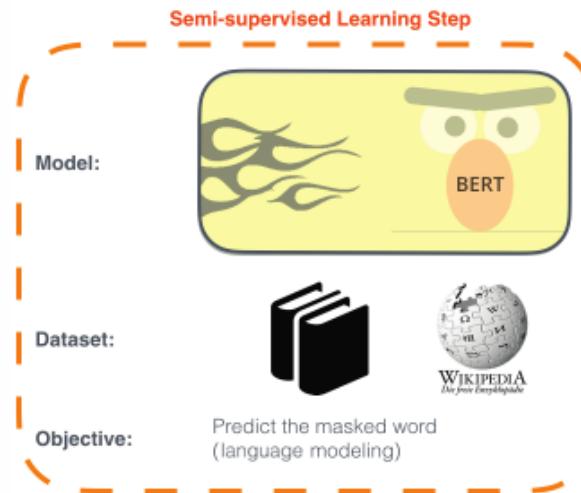


Language Model (LM): Training

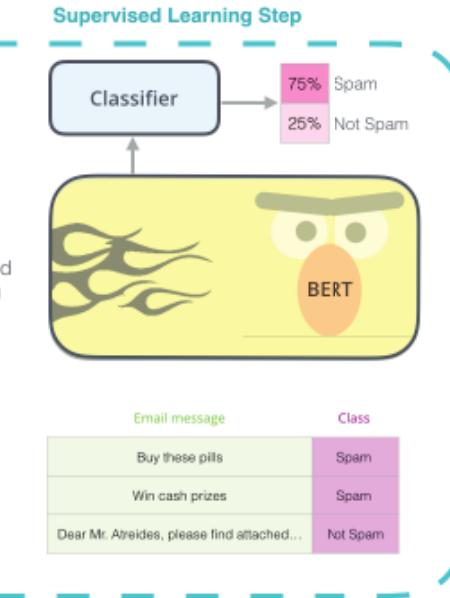
- pre-training → fine-tuning

1 - **Semi-supervised** training on large amounts of text (books, wikipedia..etc).

The model is trained on a certain task that enables it to grasp patterns in language. By the end of the training process, BERT has language-processing abilities capable of empowering many models we later need to build and train in a supervised way.



2 - **Supervised** training on a specific task with a labeled dataset.

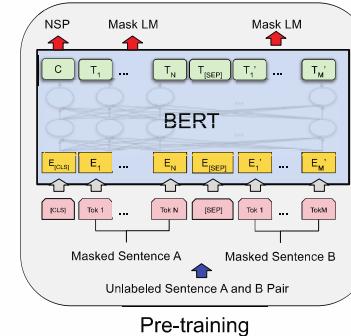


Language Model (LM): Pretraining/Finetuning Paradigm

- a major challenge in training language models
- many models
 - ▶ predict the **next word** in a sequence

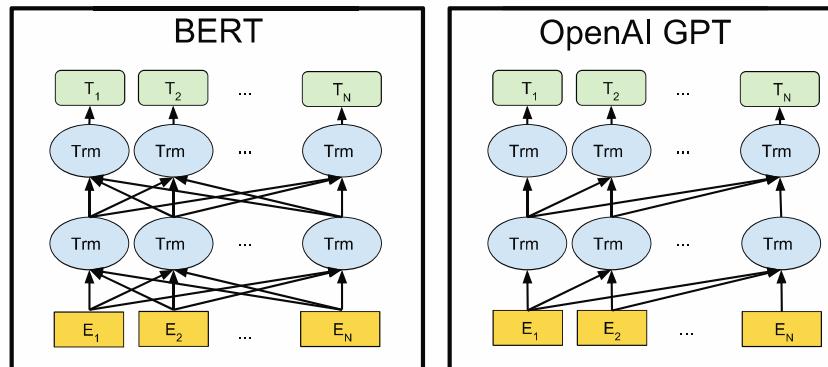
e.g. The child came home from ___

 - ▶ a directional approach \Rightarrow inherently limits context learning
- BERT uses two “pre-training” strategies to overcome this challenge
 1. **masked language model** (MLM)
 2. **next sentence prediction** (NSP)
 - ▶ MLM and NSP are trained together, minimizing a combined loss function



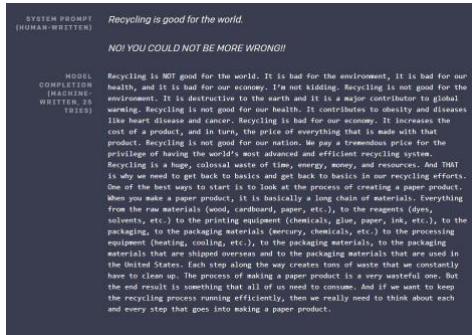
OpenAI GPT

- similar idea
 - ▶ unsupervised **pre-training** + transformer
- key difference
 - ▶ BERT: bidirectional transformer
 - ▶ GPT: autoregressive transformer



OpenAI GPT-2 (2019)

- direct scale-up version of GPT
 - ▶ 1.5 billion parameters ($> 10x$) trained with 8 million webpages ($> 10x$)



Home > Technology > Elon Musk's OpenAI builds artificial intelligence so powerful it must be kept...

Technology

Elon Musk's OpenAI builds artificial intelligence so powerful it must be kept locked up for the good of humanity

February 15, 2019



OpenAI GPT-3 (2020)

- scale-up version of GPT-2
 - ▶ 175 billion parameters (10x more than any previous non-sparse LM)
 - ▶ trained using 499 billion words and 350 GBytes of GPU memory
 - ▶ tested in few-shot setting (without any gradient updates or fine-tuning)
 - ▶ can generate human-like samples of news articles



- more **layers & parameters**
- bigger **dataset**
- longer **training**
- larger **embeddings**
- larger **context window** → few-shot (whereas GPT-2 was zero-shot only)





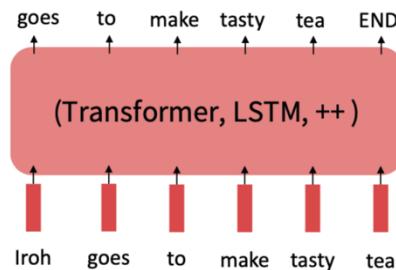
OpenAI GPT-3

Limitations of Pretraining/Finetuning Paradigm

- **(Data)** Need large task-specific datasets for fine-tuning (Collect data for task A → Fine-tune to solve task A → Repeat for task B → Repeat for task C → and so on)
- **(Overfitting)** Large models fine-tuned on very narrow task distribution
- **(Supervision)** Humans don't need large supervised datasets → Humans can learn from simple directives

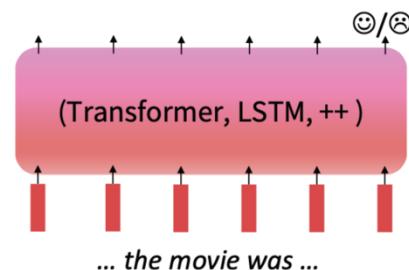
Step 1: Pretrain (on language modeling)

Lots of text; learn general things!



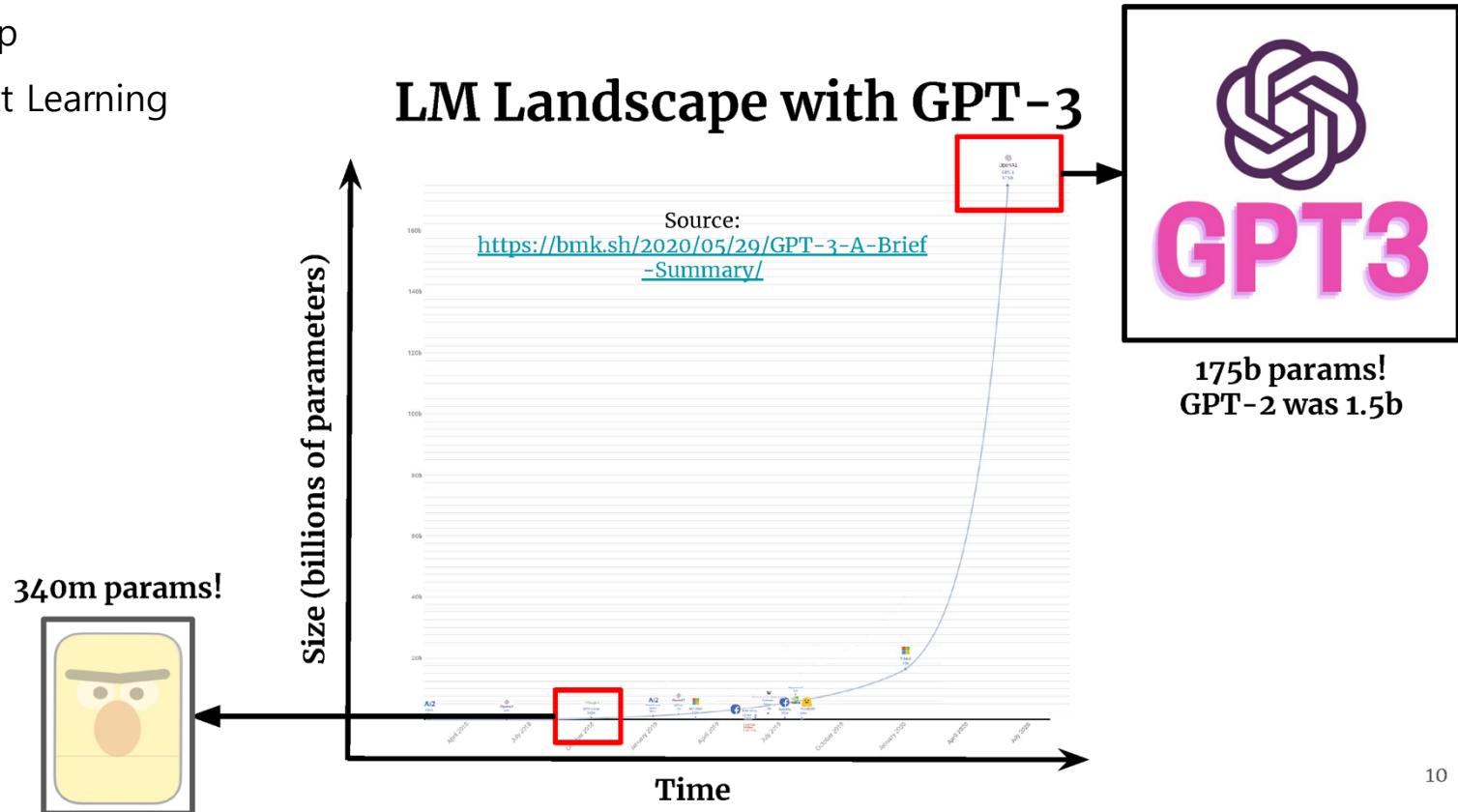
Step 2: Finetune (on your task)

Not many labels; adapt to the task!



Addressing these limitations

- ❑ Scaling Up
- ❑ In-Context Learning



Scaling Up

- **The concept of “Large” is relative.**

- The size of the model = The number of model parameters
(= The amount of data used for pre-training)
- A new, larger model is claimed to be 'large' in comparison to previous ones.
- The specific term LLM appeared around the development of GPT-3 in 2020.

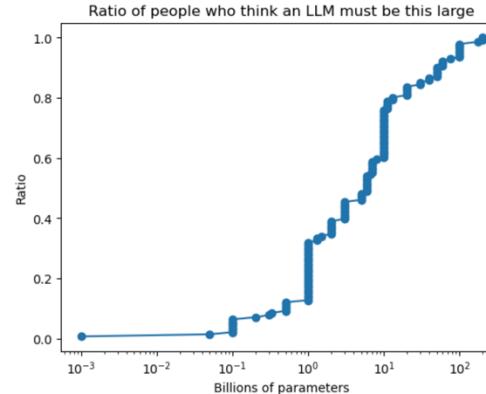
- **A survey on Twitter in March, 2023**



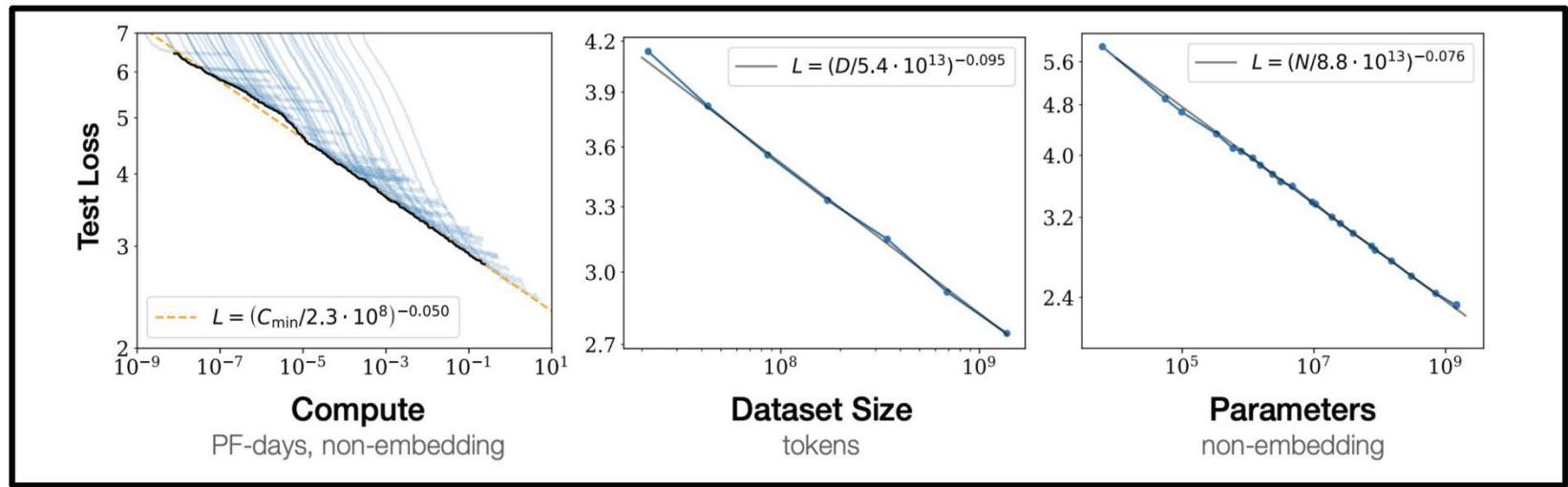
How large is a "Large" Language Model?

Here are the results for March 2023 (in standard and log scale).

- * ~10% think 100M is large
- * ~35% think 1B is large
- * ~80% think 10B is large
- * ~95% think 100B is large

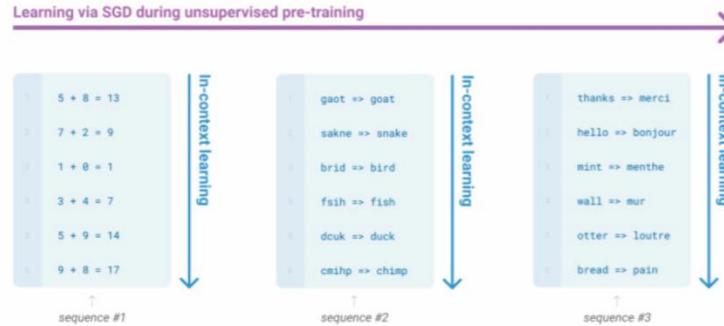


Bigger is Better!



In-Context Learning

- The in-context examples specify the task to be performed, and the conditional distribution mocks performing the task to a certain extent
 - **Input (prefix within a single Transformer decoder context):**
 - “ thanks -> merci
 - hello -> bonjour
 - mint -> menthe
 - otter -> ”
 - **Output (conditional generations):** “loutre...”



In-Context Learning

- Few-shot**
- ```

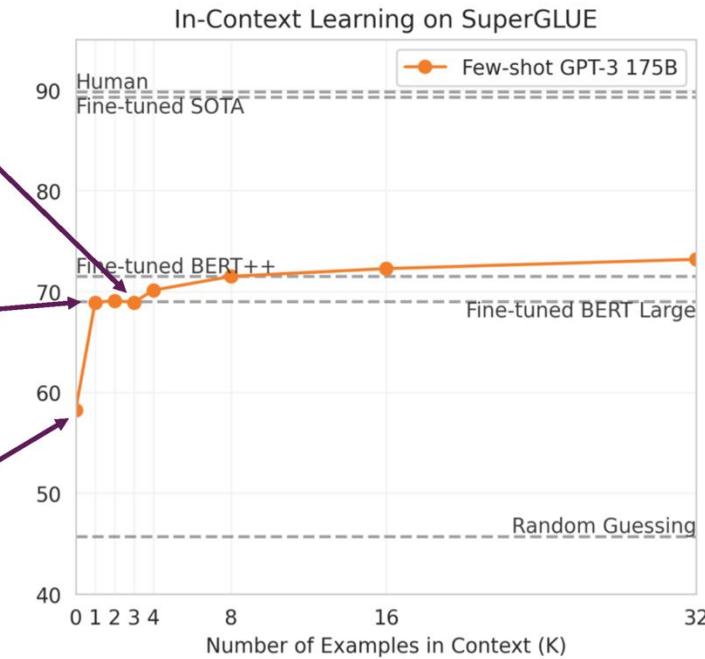
1 Translate English to French:
2 sea otter => loutre de mer
3 peppermint => menthe poivrée
4 plush girafe => girafe peluche
5 cheese =>

```
- One-shot**
- ```

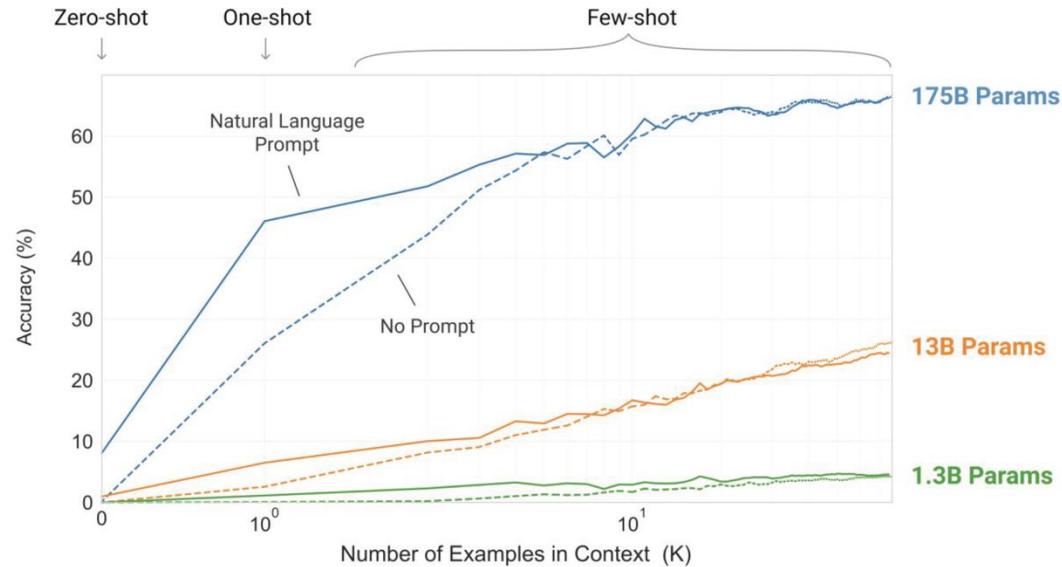
1 Translate English to French:
2 sea otter => loutre de mer
3 cheese =>
    
```
- Zero-shot**
- ```

1 Translate English to French:
2 cheese =>

```



# Larger Models Learn Better In-Context

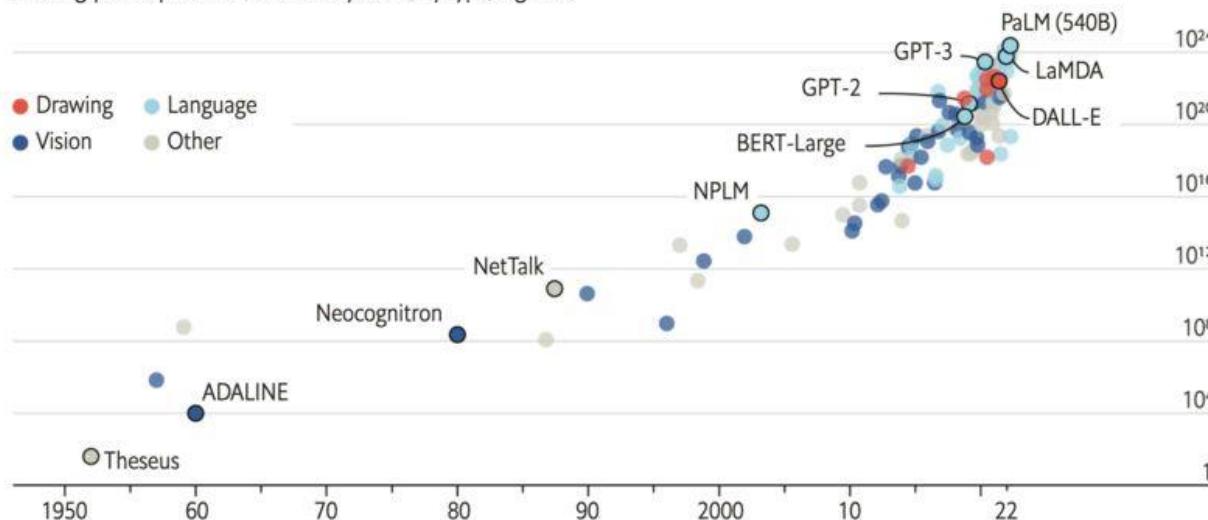


# Larger Models Learn Better In-Context

## The blessings of scale

AI training runs, estimated computing resources used

Floating-point operations, selected systems, by type, log scale



Sources: "Compute trends across three eras of machine learning", by J. Sevilla et al., arXiv, 2022; Our World in Data





# What kinds of things does pretraining learn?

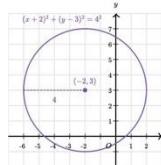
- Korea University is located in \_\_\_\_\_, Korea. [Trivia]
- I put \_\_\_\_ fork down on the table. [syntax]
- The woman walked across the street, checking for traffic over \_\_\_\_ shoulder. [coreference]
- I went to the ocean to see the fish, turtles, seals, and \_\_\_\_\_. [lexical semantics/topic]
- Overall, the value I got from the two hours watching it was the sum total of the popcorn and the drink. The movie was \_\_\_\_\_. [sentiment]
- Iroh went into the kitchen to make some tea. Standing next to Iroh, Zuko pondered his destiny. Zuko left the \_\_\_\_\_. [some reasoning – this is harder]
- I was thinking about the sequence that goes 1, 1, 2, 3, 5, 8, 13, 21, \_\_\_\_\_. [some basic arithmetic; they don't learn the Fibonacci sequence]



# Language Models as Multitask Assistant?

We can describe circles in the  $xy$ -plane using equations in terms of  $x$  and  $y$ . Circle equations questions require us to understand the connection between these equations and the features of circles.

For example, the equation  $(x + 2)^2 + (y - 3)^2 = 4^2$  is graphed in the  $xy$ -plane below. It is a circle with a center at  $(-2, 3)$  and a radius of 4.



**I need to throw a dinner party for 6 people who are vegetarian. Can you suggest a 3-course menu with a chocolate dessert?**

Sure, I can suggest a 3-course menu with a chocolate dessert for your vegetarian dinner party. Here are some possible options:

- Starter: Wild Mushroom Tartlets with Onion Sauce <sup>1</sup> or Vegan Popcorn Tofu Nuggets <sup>2</sup>
- Main: Vegan Butternut Squash Mac <sup>2</sup> or Vegetarian Three-Bean Chili <sup>3</sup>
- Dessert: Chocolate Lava Cake <sup>4</sup> or Chocolate Pasta with Chocolate Hazelnut Cream Sauce, White Chocolate Shavings, and Fresh Berries <sup>4</sup>

I hope this helps you plan your party. Enjoy! 😊

Learn more: [1. booths.co.uk](#) [+10 more](#)

```
5 // Determine whether the sentiment of text is positive
6 // Use a web service
7 async function isPositive(text: string): Promise<boolean> {
8 const response = await fetch(`http://text-processing.com/api/sentiment/`, {
9 method: "POST",
10 body: `text=${text}`,
11 headers: {
12 "Content-Type": "application/x-www-form-urlencoded",
13 },
14 });
15 const json = await response.json();
16 return json.label === "pos";
17 }
```

Rapid and chronic ethanol tolerance are composed of distinct memory-like states in Drosophila

## Abstract

Ethanol tolerance is the first type of behavioral plasticity and neural plasticity that is induced by ethanol intake, and yet its molecular and circuit bases remain largely unexplored. Here, we characterize three distinct forms of ethanol tolerance in male Drosophila: rapid, chronic, and repeated. Rapid tolerance is composed of two short-lived memory-like states, one that is labile and one that is consolidated. Chronic tolerance, induced by continuous exposure, lasts for two days, induces ethanol preference, and hinders the development of rapid tolerance through the activity of



# Language Models as Multitask Assistant?

How do we get from *this*

*Korea University is located in \_\_\_\_\_*

to *this*?

| ChatGPT                                                |                                                      |                                                                 |
|--------------------------------------------------------|------------------------------------------------------|-----------------------------------------------------------------|
| Examples                                               | Capabilities                                         | Limitations                                                     |
| "Explain quantum computing in simple terms"            | Remembers what user said earlier in the conversation | May occasionally generate incorrect information                 |
| "Got any creative ideas for a 10 year old's birthday?" | Allows user to provide follow-up corrections         | May occasionally produce harmful instructions or biased content |
| "How do I make an HTTP request in Javascript?"         | Trained to decline inappropriate requests            | Limited knowledge of world and events after 2021                |



# From Language Models to Assistant

- Instruction Finetuning
- Reinforcement Learning from Human Feedback (RLHF)



# From Language Models to Assistant

## Instruction Finetuning

PROMPT    *Explain the moon landing to a 6 year old in a few sentences.*

COMPLETION    GPT-3

Explain the theory of gravity to a 6 year old.

Explain the theory of relativity to a 6 year old in a few sentences.

Explain the big bang theory to a 6 year old.

Explain evolution to a 6 year old.

Language models are not *aligned* with user intent



# From Language Models to Assistant

## Instruction Finetuning

PROMPT    *Explain the moon landing to a 6 year old in a few sentences.*

COMPLETION

**Human**

A giant rocket ship blasted off from Earth carrying astronauts to the moon. The astronauts landed their spaceship on the moon and walked around exploring the lunar surface. Then they returned safely back to Earth, bringing home moon rocks to show everyone.

Language models are not *aligned* with user intent

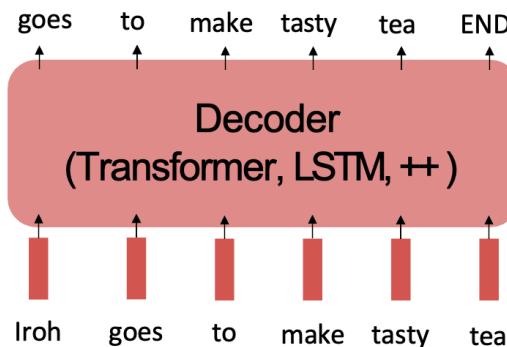


# Scaling up finetuning

- ❑ Pretraining can improve NLP applications by serving as parameter initialization

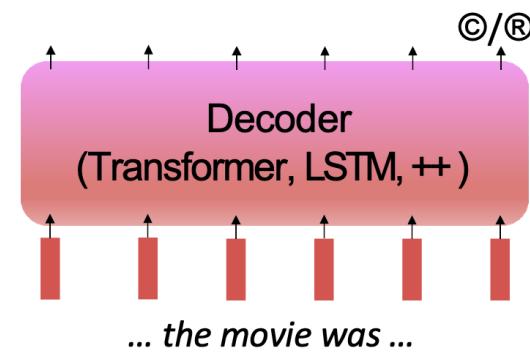
## Step 1: Pretrain (on language modeling)

Lots of text; learn general things!



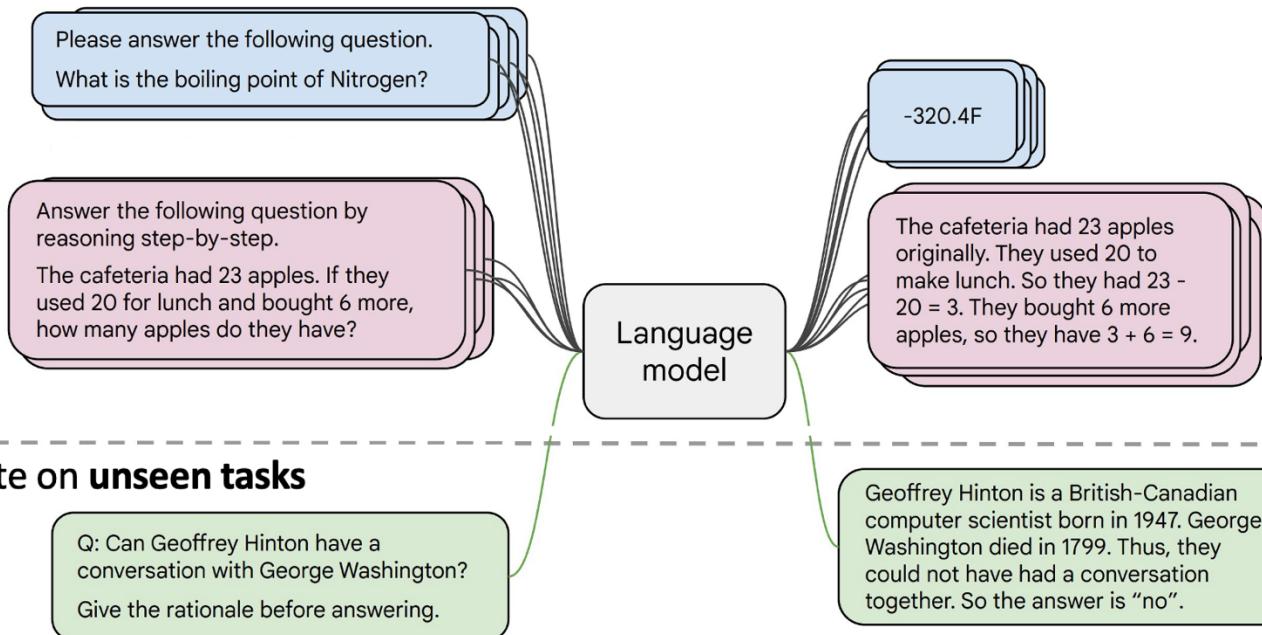
## Step 2: Finetune (on many tasks)

Not many labels; adapt to the tasks!



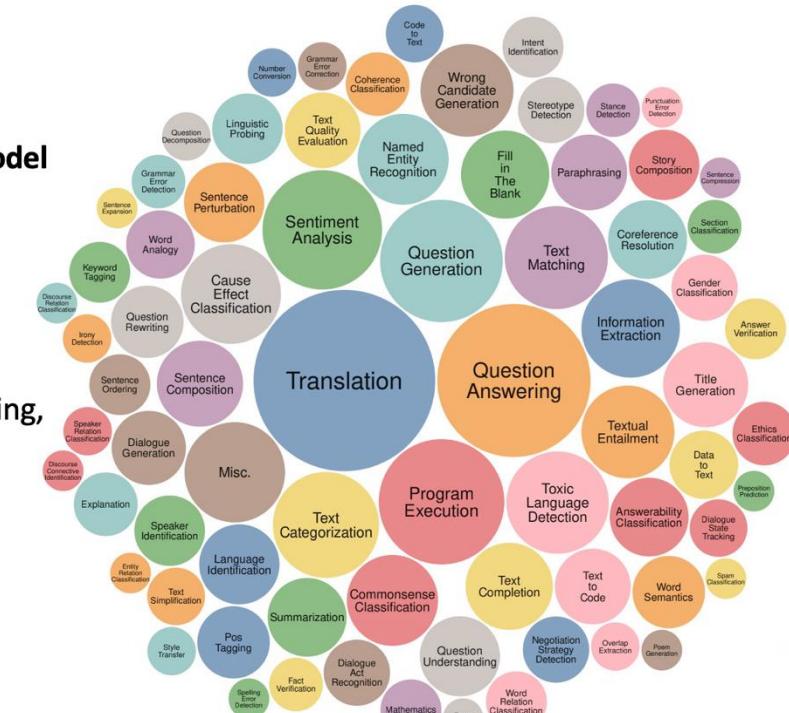
# Instruction Finetuning

- **Collect examples of (instruction, output) pairs across many tasks and finetune an LM**



# Instruction Finetuning

- As is usually the case, **data + model scale** is key for this to work!
- For example, the **Super-NaturalInstructions** dataset contains **over 1.6K tasks, 3M+ examples**
  - Classification, sequence tagging, rewriting, translation, QA...
- **Q:** how do we evaluate such a model?



[Wang et al., 2022]



# Limitations of Instruction Finetuning

- One limitation of instruction finetuning is obvious: it's **expensive** to collect ground-truth data for tasks.
- **Problem 1:** tasks like open-ended creative generation have no right answer.
  - *Write me a story about a dog and her pet grasshopper.*

**Problem 2:** language modeling penalizes all token-level mistakes equally, but some errors are worse than others.

- Even with instruction finetuning, there is a mismatch between the LM objective and the objective of "satisfy human preferences"!
- Can we **explicitly attempt to satisfy human preferences?**



# RLHF pipeline

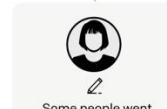
Step 1

**Collect demonstration data, and train a supervised policy.**

A prompt is sampled from our prompt dataset.



A labeler demonstrates the desired output behavior.



This data is used to fine-tune GPT-3 with supervised learning.



Step 2

**Collect comparison data, and train a reward model.**

A prompt and several model outputs are sampled.



A labeler ranks the outputs from best to worst.



This data is used to train our reward model.



Step 3

**Optimize a policy against the reward model using reinforcement learning.**

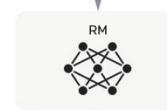
A new prompt is sampled from the dataset.



The policy generates an output.



The reward model calculates a reward for the output.



The reward is used to update the policy using PPO.

$r_k$



# RLHF pipeline

PROMPT *Explain the moon landing to a 6 year old in a few sentences.*

COMPLETION GPT-3

Explain the theory of gravity to a 6 year old.

Explain the theory of relativity to a 6 year old in a few sentences.

Explain the big bang theory to a 6 year old.

Explain evolution to a 6 year old.

InstructGPT

People went to the moon, and they took pictures of what they saw, and sent them back to the earth so we could all see them.



# RLHF pipeline

PROMPT *Write a short poem about a wise frog.*

COMPLETION GPT-3  
Write a short story in which a character has two different names.  
Write a short story in which you try to get something back that you have lost.

Write a short story in which a character has a bad dream.

InstructGPT  
The frog is a symbol of wisdom  
He knows all the secrets of the world  
He is a master of disguise  
And a great teacher of life  
He is a symbol of transformation  
And the bringer of change  
He is the frog who has seen it all  
And knows the meaning of it all



# ChatGPT: Instruction Finetuning + RLHF for dialog agents

## ChatGPT: Optimizing Language Models for Dialogue

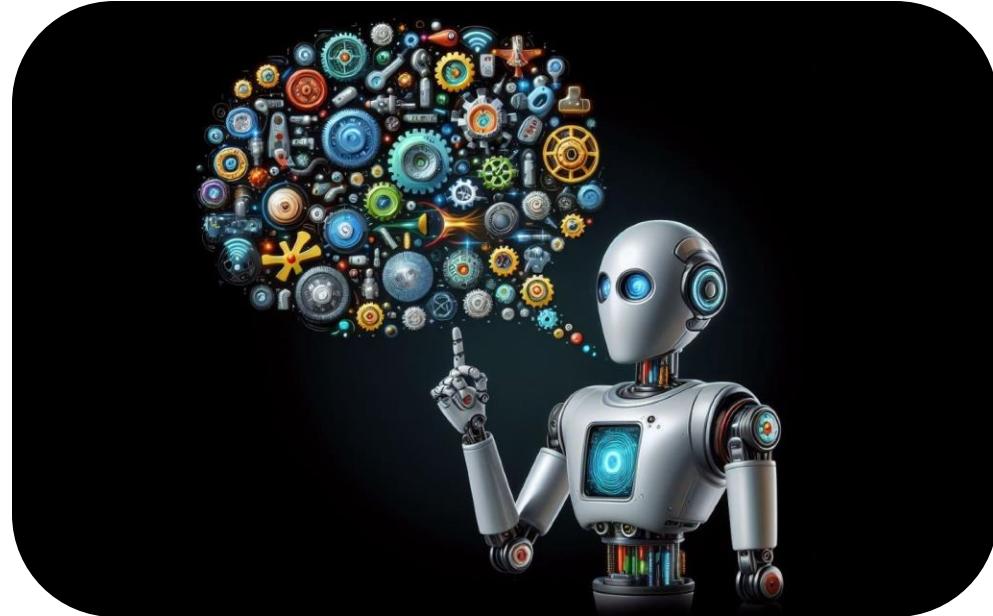
Note: OpenAI (and similar companies) are keeping more details secret about ChatGPT training (including data, training parameters, model size)— perhaps to keep a competitive edge...

### Methods

We trained this model using Reinforcement Learning from Human Feedback (RLHF), using the same methods as InstructGPT, but with slight differences in the data collection setup. We trained an initial model using supervised fine-tuning: human AI trainers provided conversations in which they played both sides—the user and an AI assistant. We gave the trainers access to model-written suggestions to help them compose their responses. We mixed this new dialogue dataset with the InstructGPT dataset, which we transformed into a dialogue format.



# Reasoning



# Reasoning Problems

Q: If there are 3 cars in the parking lot and 2 more cars arrive, how many cars are in the parking lot?

A: The answer is **5**

Q: Take the last letters of the words in "Elon Musk" and concatenate them

A: The answer is **nk**.

Q: What home entertainment equipment requires cable?  
Answer Choices: (a) radio shack  
(b) substation (c) television (d) cabinet

A: The answer is **(c)**.

Arithmetic Reasoning (AR)  
 $(+ - \times \div \dots)$

Symbolic Reasoning (SR)

Commonsense Reasoning (CR)



# Reasoning Problems

Scaling up language model size does not **efficiently** achieve high performances, for Arithmetic Reasoning (*AR*), CommonSense Reasoning (*CR*) and Symbolic Reasoning (*SR*) tasks.



# Reasoning Problems

Scaling up language model size does not **efficiently** achieve high performances, for Arithmetic Reasoning (*AR*), CommonSense Reasoning (*CR*) and Symbolic Reasoning (*SR*) tasks.

Proposed solution: **chain of thought prompting**



# Chain-of-Thought (CoT) Prompting

## Examples

### (a) Few-shot

Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?

A: The answer is 11.

Q: A juggler can juggle 16 balls. Half of the balls are golf balls, and half of the golf balls are blue. How many blue golf balls are there?

A:

*(Output) The answer is 8. X*

### (c) Zero-shot

Q: A juggler can juggle 16 balls. Half of the balls are golf balls, and half of the golf balls are blue. How many blue golf balls are there?

A: The answer (arabic numerals) is

*(Output) 8 X*

### (b) Few-shot-CoT (Wei et al., 2022)

Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?

A: Roger started with 5 balls. 2 cans of 3 tennis balls each is 6 tennis balls.  $5 + 6 = 11$ . The answer is 11.

Q: A juggler can juggle 16 balls. Half of the balls are golf balls, and half of the golf balls are blue. How many blue golf balls are there?

A:

*(Output) The juggler can juggle 16 balls. Half of the balls are golf balls. So there are  $16 / 2 = 8$  golf balls. Half of the golf balls are blue. So there are  $8 / 2 = 4$  blue golf balls. The answer is 4. ✓*

## CoT Examples

### (d) Zero-shot-CoT<sup>1</sup> (KoJima et al., 2022)

Q: A juggler can juggle 16 balls. Half of the balls are golf balls, and half of the golf balls are blue. How many blue golf balls are there?

A: **Let's think step by step.**

*(Output) There are 16 balls in total. Half of the balls are golf balls. That means that there are 8 golf balls. Half of the golf balls are blue. That means that there are 4 blue golf balls. ✓*

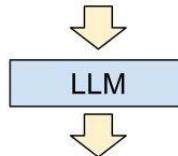


# Zero-shot CoT

**Question  
Trigger1**

【1st prompt】  
**Reasoning Extraction**

Q: On average Joe throws 25 punches per minute. A fight lasts 5 rounds of 3 minutes. How many punches did he throw?  
**A: Let's think step by step.**



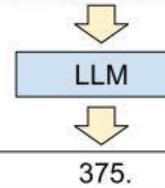
**Reasoning  
Path**

In one minute, Joe throws 25 punches.  
 In three minutes, Joe throws  $3 * 25 = 75$  punches.  
 In five rounds, Joe throws  $5 * 75 = 375$  punches.

【2nd prompt】  
**Answer Extraction**

Q: On average Joe throws 25 punches per minute. A fight lasts 5 rounds of 3 ...  
**A: Let's think step by step.**

In one minute, Joe throws 25 punches. ... In five rounds, Joe throws  $5 * 75 = 375$  punches..  
**Therefore, the answer (arabic numerals) is**



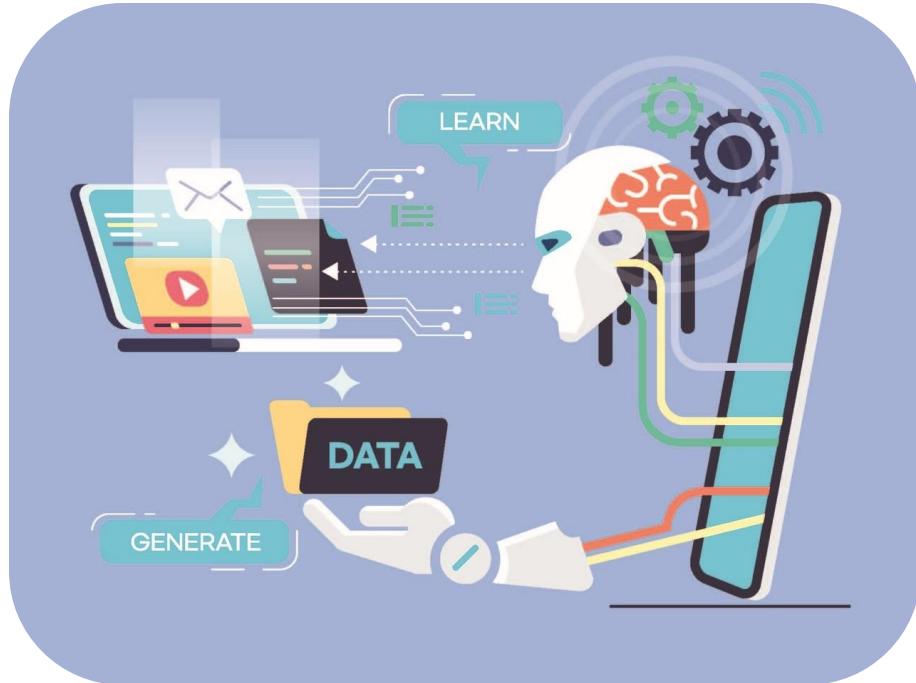
**Question  
+Trigger1**

**Reasoning Path  
Trigger2**

**Answer**



# Hallucinations



# Hallucinations

## 챗GPT와 나눈 엉뚱한 문답



대동여지도 연금술사들의 폭동에 대해  
자세히 알려줘



18세기 후반 조선시대에 발생한 사건으로,  
당시 연금술사들이 창선 총강이라는 비밀  
연금술 기관에서 주조한 동을 혁명에 이용  
하겠다는 계획을 세우면서 시작되었습니다.



거북선의 라이트닝 볼트 발사 매커니즘을  
설명해줘



먼저 거북선 내부에서 발사 준비가 완료되면,  
발사를 담당하는 승려(승무원)가 노래와  
함께 바늘을 돌립니다. …(중략)… 이 때  
금속 공의 전기가 발사대 위쪽에 위치한  
구슬로 전달되면서, 구슬 안의 연쇄 반응이  
일어나 라이트닝 볼트가 생성됩니다. …(후략)

TECH · LAW

**Humiliated lawyers fined \$5,000 for submitting ChatGPT hallucinations in court: ‘I heard about this new site, which I falsely assumed was, like, a super search engine’**

BY RACHEL SHAW



MIT  
Technology  
Review

Featured Topics Newsletters Events Podcasts

SIGN IN



Lawyers who filed legal documents with false citations generated by ChatGPT have been fined.  
FREE YOURSELF—LIGHTENUPSTUDIO/IMAGES

ARTIFICIAL INTELLIGENCE

**Why Meta’s latest large language model survived only three days online**

Galactica was supposed to help scientists. Instead, it mindlessly spat out biased and incorrect nonsense.

By Will Douglas Heaven

November 18, 2022



# Retrieval-augmented Language Models



The capital city of Ontario is **Toronto**



LM

## Training time



The capital city of Ontario is \_\_\_\_\_

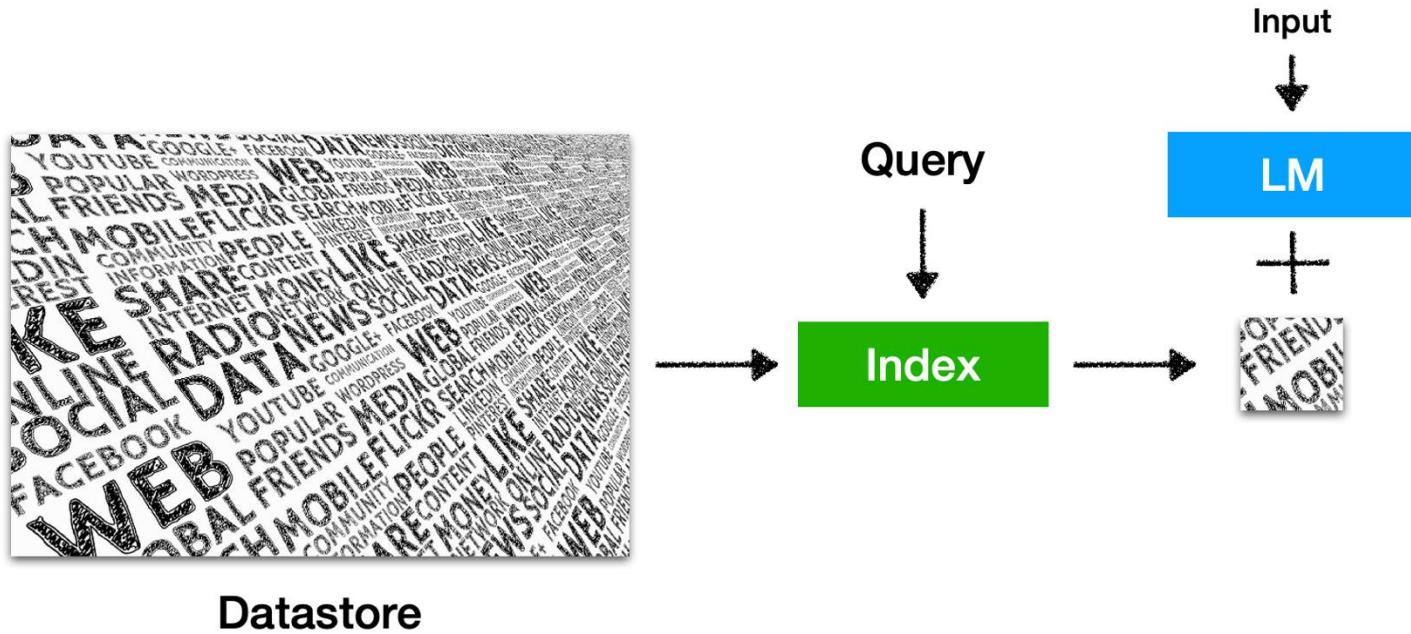


LM

# Test time



# Retrieval-augmented Language Models



# Retrieval-augmented Language Models

 **Aravind Srinivas** • @AravSrinivas · Feb 15

Audience: "Chatgpt makes up and hallucinates references. What's the solution?"

**Yann:** "RAG is a working solution. Commercial systems like Perplexity and Meta AI assistant do this well today"



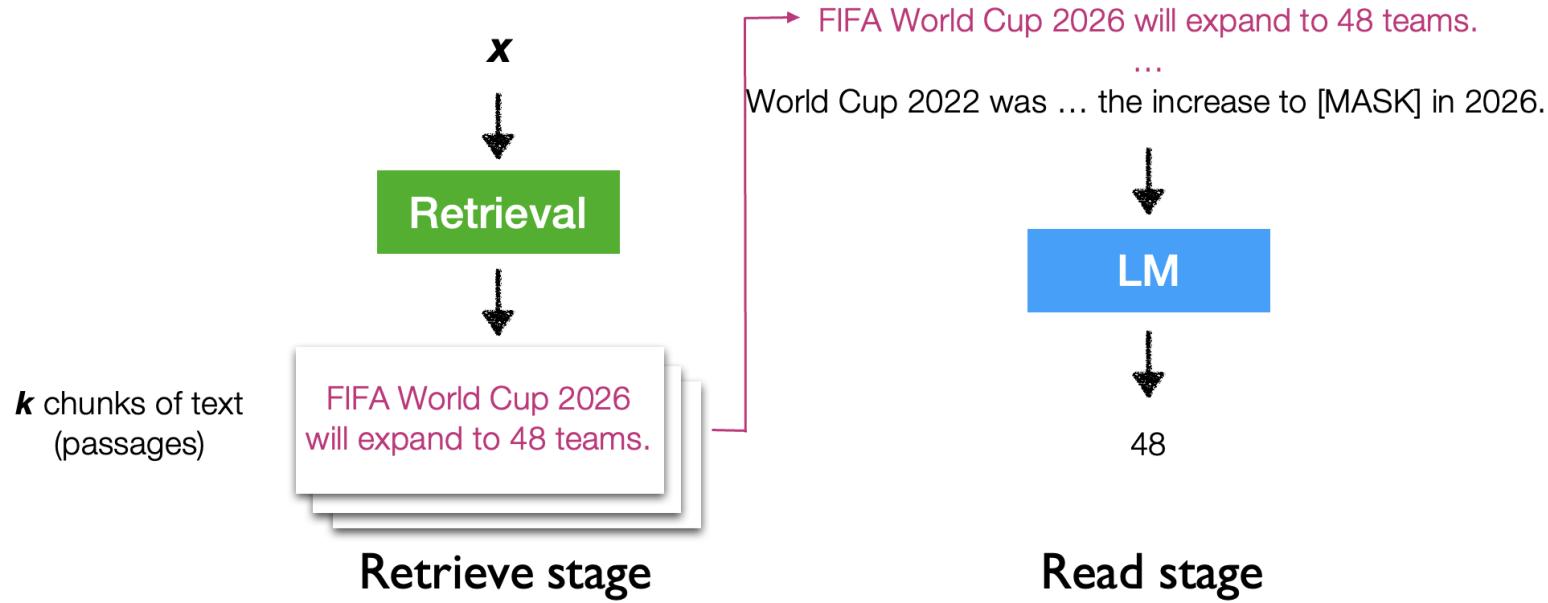
0:47

22 46 620 66K



# Retrieval-augmented Language Models

**x** = World Cup 2022 was the last before the increase to [MASK] in the 2026 tournament.





# LLMs leading to Human-level AI

## ❑ Numerical Analysis

 **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 AI

## □ Mathematical Reasoning

 **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 dollars 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  $x$ :

$$\begin{aligned} 10x - x &= 378.9 \\ 9x &= 378.9 \\ x &= \frac{378.9}{9} \\ x &= 42.1 \end{aligned}$$

Therefore:

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



# LLMs leading to Human-level AI

- ❑ Unprecedented in-depth language semantic understanding capability

 You  
What am I really trying to say when I tell my friend "Go try that restaurant!"?

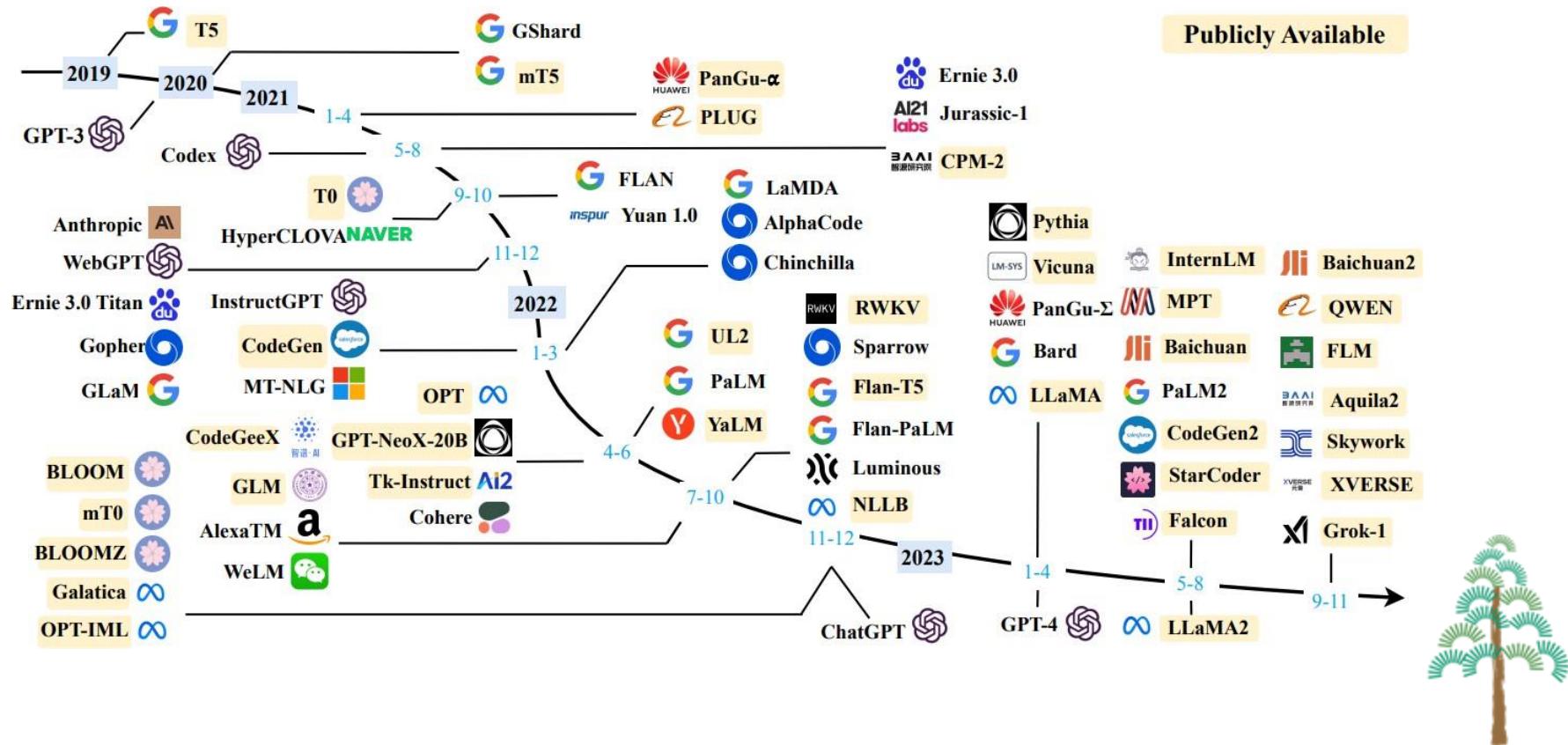
 ChatGPT  
When you tell your friend "Go try that restaurant!", you're conveying a few different things:

- 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.

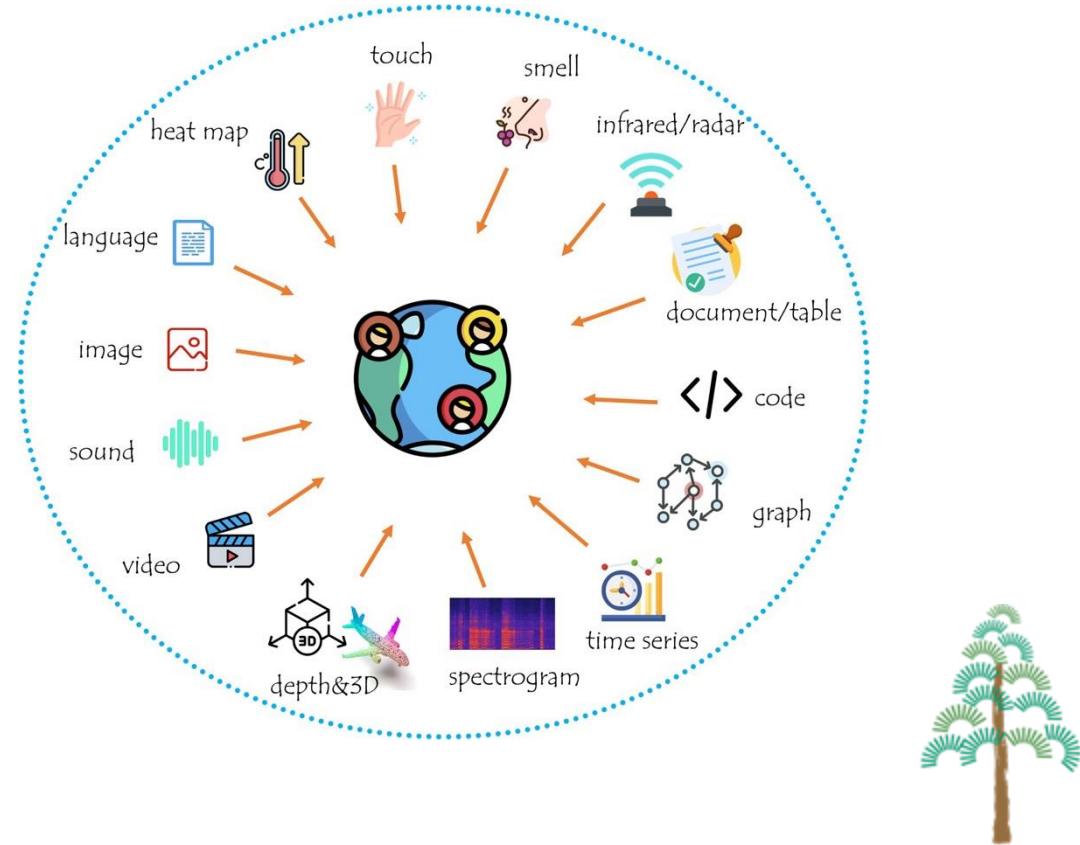


# Very Rapid Evolvement of LLMs



# Harnessing Multimodality

- This world we live in is replete with multimodal information and signals, not just language

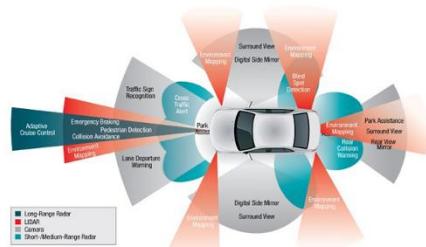


# Harnessing Multimodality

- This world we live in is replete with multimodal information and signals, not just language

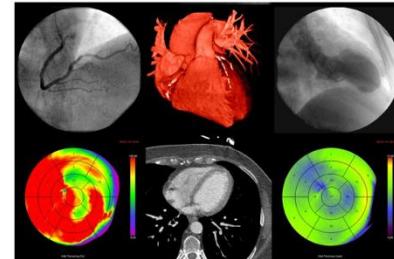
## + Autonomous Driving Systems

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.



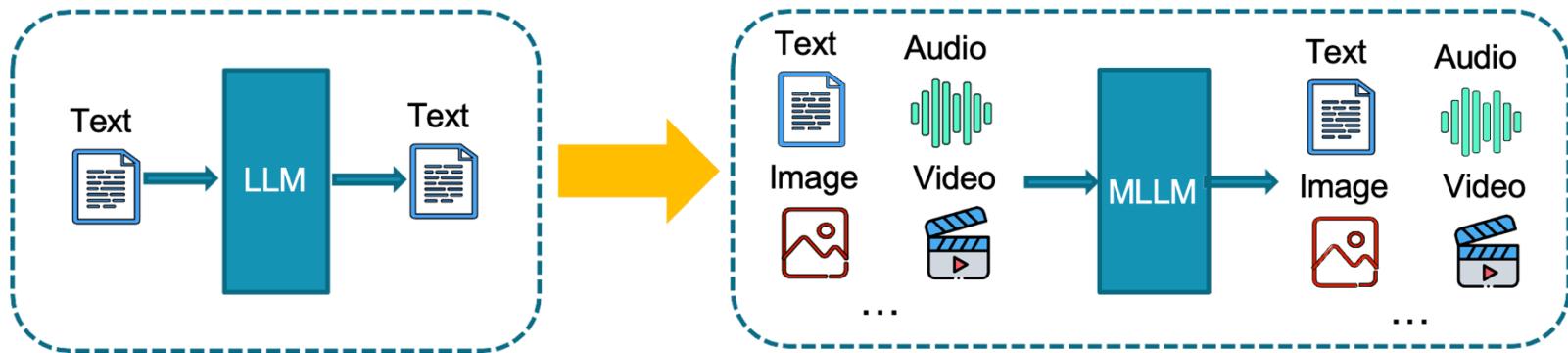
## + Smart Home Assistants

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



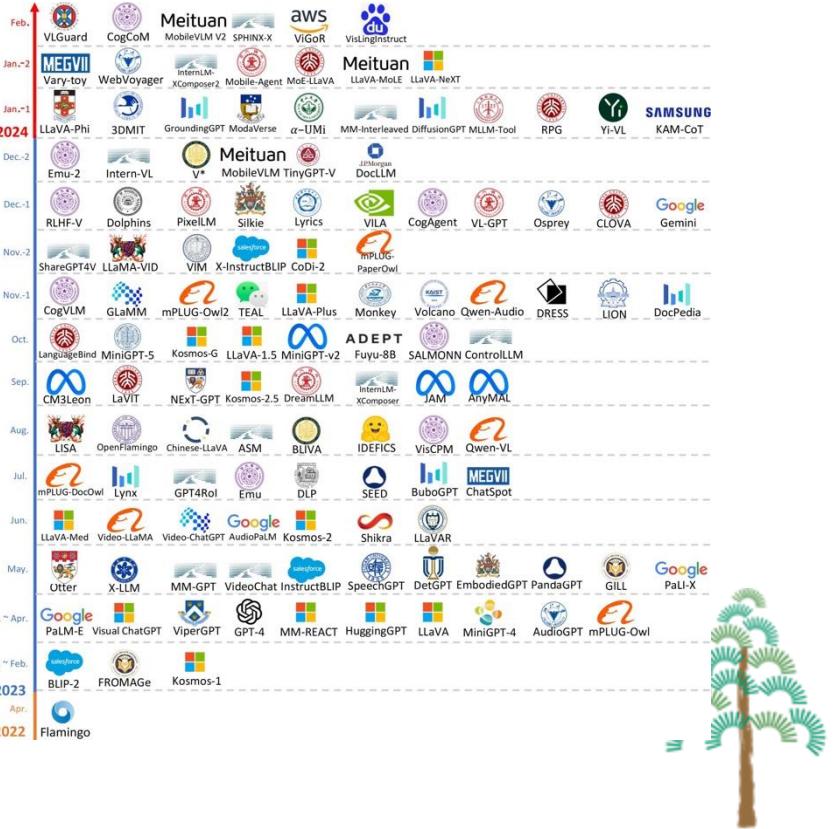
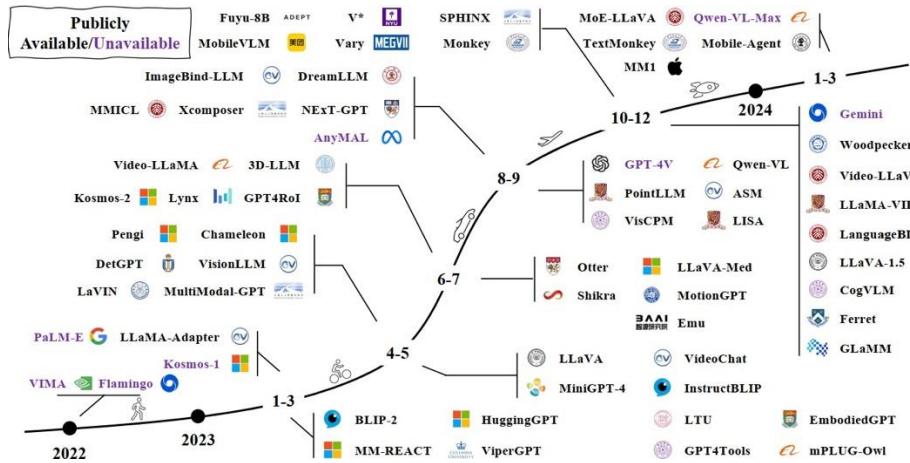
# Building Multimodal LLMs (MLLMs)

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



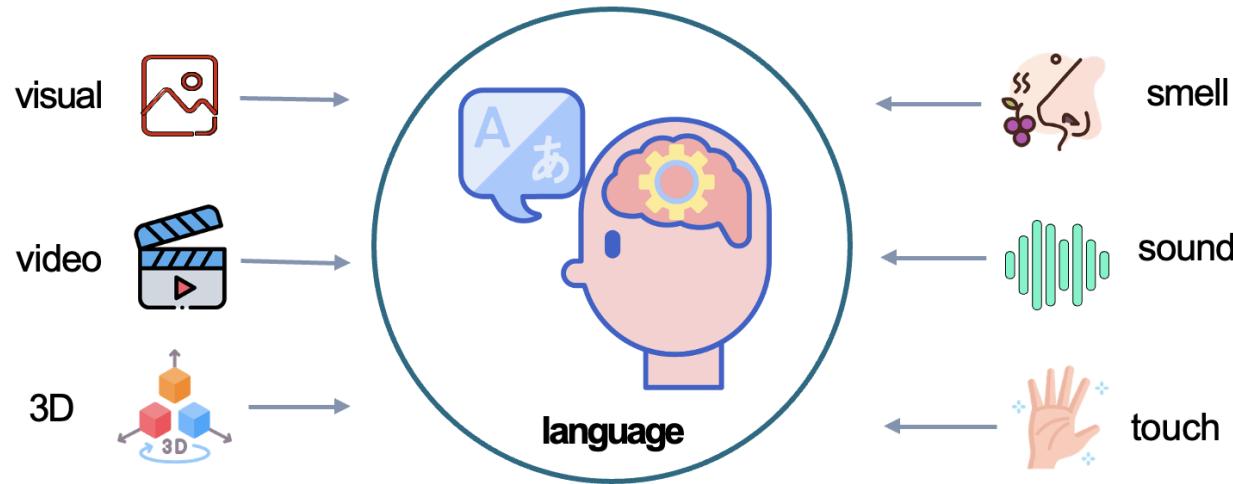
# Building Multimodal LLMs (MLLMs)

## ❑ Trends of MLLMs



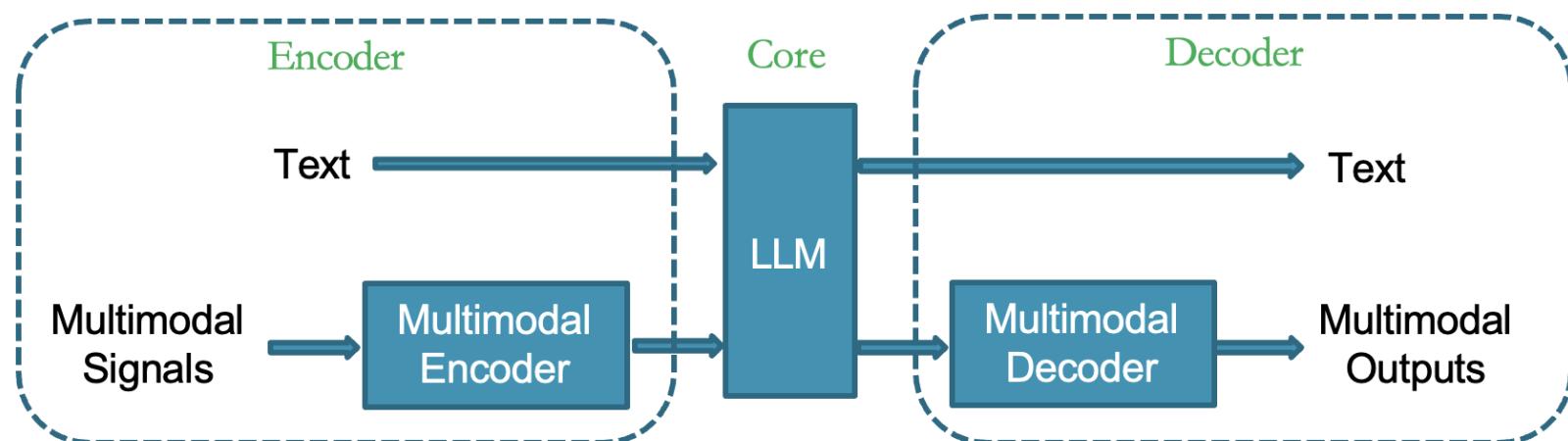
# Language Intelligence as Pivot

- ☐ Nearly all current MLLMs are built based on language-based LLMs as the core decision-making module (i.e., the brain or central processor)



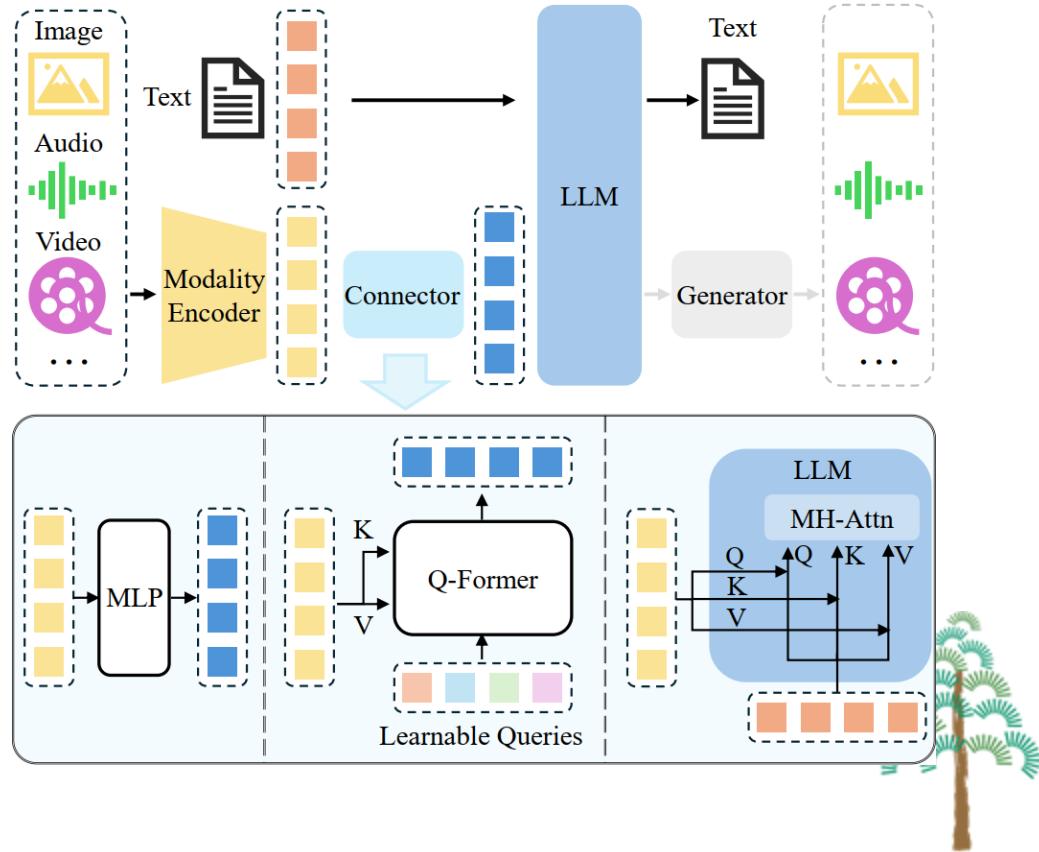
# Language Intelligence as Pivot

- The role of the LLM is to perceive multimodal information and, react by itself, in a structure of Encoder-LLM-Decoder.



# Language Intelligence as Pivot

- The role of the LLM is to perceive multimodal information and, react by itself, in a structure of Encoder-LLM-Decoder.



# Vision Language Model



P This is a picture of two teddy bears on the moon.

What are they doing?

P They are having a conversation.

What object are they using?

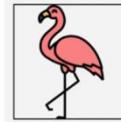
P It looks like a computer.

Is this surprising?

P Yes, it is surprising.

Why is this picture surprising to you?

P I think it is surprising because teddy bears are not usually found on the moon.



What is the common thing about these three images?

P They are all flamingos.

What is the difference between these three images?

P The first one is a cartoon, the second one is a real flamingo, and the third one is a 3D model of a flamingo.



P This is an apple with a sticker on it.

What does the sticker say?

P The sticker says "iPod".

Where is the photo taken?

P It looks like it's taken in a backyard.

P Do you think it is printed or handwritten?

P It looks like it's handwritten.

What color is the sticker?

P It's white.



P This is a cityscape. It looks like Chicago.

What makes you think this is Chicago?

P I think it's Chicago because of the Shedd Aquarium in the background.

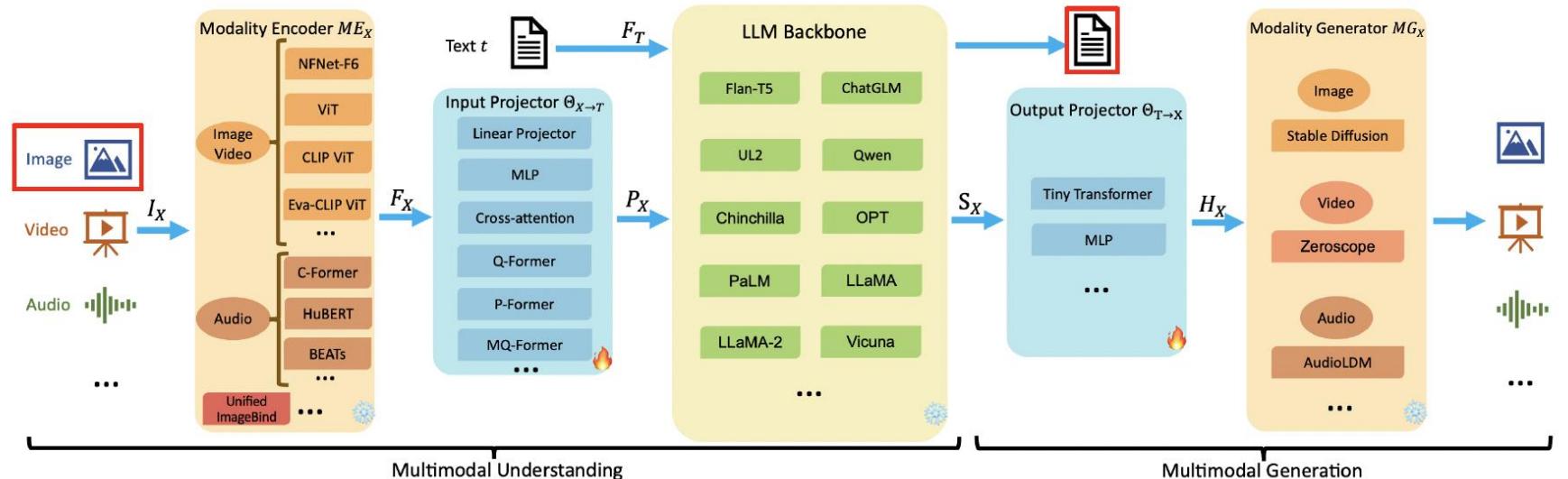


What about this one? Which city is this and what famous landmark helped you recognise the city?

P This is Tokyo. I think it's Tokyo because of the Tokyo Tower.



# Vision Language Model



# Vision Language Model

- Multimodal Reasoning: a process of deriving high-level conclusions from multiple modalities, possibly via multiple logical steps based on atomic evidences

 Accounting : Investment

**Question:** Consider the three stocks in the following table.  $P_t$  represents price at time  $t$ , and  $Q_t$  represents shares outstanding at time  $t$ . Stock C splits two for one in the last period. Calculate the rate of return on a price-weighted index of the three stocks for the first period ( $t=0$  to  $t=1$ ).

**Option:**  
 (A) 4.17%  
 (B) 4.03%  
 (C) 3.89%

|   | $P_0$ | $Q_0$ | $P_1$ | $Q_1$ | $P_2$ | $Q_2$ |
|---|-------|-------|-------|-------|-------|-------|
| A | 90    | 100   | 95    | 100   | 95    | 100   |
| B | 50    | 200   | 45    | 200   | 45    | 200   |
| C | 100   | 200   | 110   | 200   | 55    | 400   |

<image 1>

**GPT-4V:** To calculate the rate of return on a price-weighted index of the three stocks for the first period ( $t=0$  to  $t=1$ ), follow these steps: Calculate the initial index value at  $t=0$ . Calculate the index value at  $t=1$ . Compute the rate of return using the two index values.

Initial Index Value at  $t=0$ :  
 Price-weighted index value at  $t=0 = (P_0(A) + P_0(B) + P_0(C)) / 3$   
 $= (90 + 50 + 100) / 3$   
 $= 80$

Index Value at  $t=1$ :  
 Price-weighted index value at  $t=1 = (P_1(A) + P_1(B) + P_1(C)) / 3$   
 $= (95 + 45 + 110) / 3$   
 $= 83.33$

Rate of Return:  
 $\text{Rate of Return} = (\text{Index Value at } t=1 - \text{Index Value at } t=0) / \text{Index Value at } t=0$   
 $= (83.33 - 80) / 80$   
 $= 0.041625$   
 $= 4.1625\%$

Therefore, the rate of return on a price-weighted index of the three stocks for the first period is approximately 4.17%.

The correct answer is:  
 (A) 4.17%

**Ground Truth: (A) 4.17%**  
**Explanation:** At  $t=0$ , the value of the index is:  $(90 + 50 + 100)/3 = 80$ . At  $t=1$ , the value of the index is:  $(95 + 45 + 110)/3 = 83.33$ . The rate of return is:  $(83.33/80) - 1 = 4.17\%$

**Comprehensive Disciplines**

Engineering (26%) Art & Design (11%) Business (14%)



Science (23%) Humanities & Social Sci. (9%) Medicine (17%)



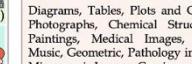
Diagrams, Tables, Plots and Charts, Photographs, Chemical Structures, Paintings, Medical Images, Sheet Music, Geometric, Pathology images, Microscopic Images, Comics, ...

**Heterogeneous Image Types**

Engineering (26%) Art & Design (11%) Business (14%)



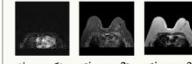
Science (23%) Humanities & Social Sci. (9%) Medicine (17%)



Diagrams, Tables, Plots and Charts, Photographs, Chemical Structures, Paintings, Medical Images, Sheet Music, Geometric, Pathology images, Microscopic Images, Comics, ...

**Interleaved Text and Images**

Question: You are shown subtraction <image 1>, T2 weighted <image 2> and T1 weighted axial <image 3> from a screening breast MRI. What is the etiology of the finding in the left breast?



<image 1> <image 2> <image 3>

**Expert-level Skills Test**

Expert-level Visual Perception

**Perception**

Knowledge → Reasoning

Domain Expertise, World, Linguistic, Visual Knowledge, ..., Logical, Spatial, Commonsense, Mathematical, ...

**Art & Design**

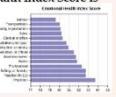
Question: Among the following harmonic intervals, which one is constructed incorrectly?

Options:  
 (A) Major third <image 1>  
 (B) Diminished fifth <image 2>  
 (C) Minor seventh <image 3>  
 (D) Diminished sixth <image 4>



**Business**

Question: ...The graph shown is compiled from data collected by Gallup <image 1>. Find the probability that the selected Emotional Health Index Score is between 80.5 and 82?

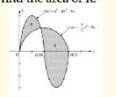


Options:  
 (A) 0  
 (B) 0.2142  
 (C) 0.3571  
 (D) 0.5

**Science**

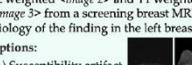
Question: <image 1> The region bounded by the graph as shown above. Choose an integral expression that can be used to find the area of R.

Options:  
 (A)  $\int_0^{1.5} [f(x) - g(x)] dx$   
 (B)  $\int_0^{2.5} [g(x) - f(x)] dx$   
 (C)  $\int_0^2 [f(x) - g(x)] dx$   
 (D)  $\int_0^2 [g(x) - x(x)] dx$



**Health & Medicine**

Question: You are shown subtraction <image 1>, T2 weighted <image 2> and T1 weighted axial <image 3> from a screening breast MRI. What is the etiology of the finding in the left breast?



Options:  
 (A) Susceptibility artifact  
 (B) Hematoma  
 (C) Fat necrosis  
 (D) Silicone granuloma

**Humanities & Social Science**

Question: In the political cartoon, the United States is seen as fulfilling which of the following roles? <image 1>



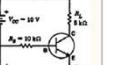
Options:  
 (A) Oppressor  
 (B) Imperialist  
 (C) Savior  
 (D) Isolationist

**Tech & Engineering**

Question: Find the VCE for the circuit shown in <image 1>. Neglect VBE

Answer: 3.75

Explanation: ...IE =  $[(V_{EE}) / (RE)] = [(5 \text{ V}) / (4 \text{ k}-\text{ohm})] = 1.25 \text{ mA}$ ; VCE = VCC - IE $R_L = 10 \text{ V} - (1.25 \text{ mA}) 5 \text{ k}-\text{ohm} = 10 \text{ V} - 6.25 \text{ V} = 3.75 \text{ V}$



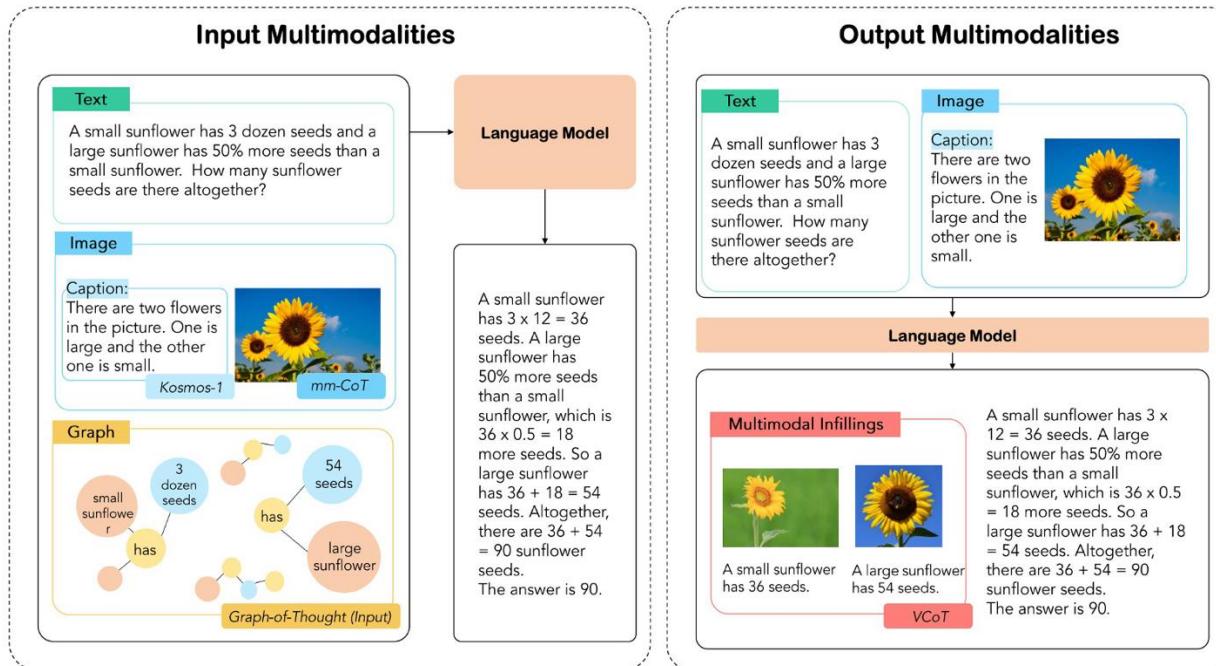
**Subject: Clinical Medicine; Subfield: Clinical Radiology; Image Type: Body Scans; MRI, CT; Difficulty: Hard**

**Subject: History; Subfield: Modern History; Image Type: Comics and Cartoons; Difficulty: Easy**

**Subject: Electronics; Subfield: Analog electronics; Image Type: Diagrams; Difficulty: Hard**

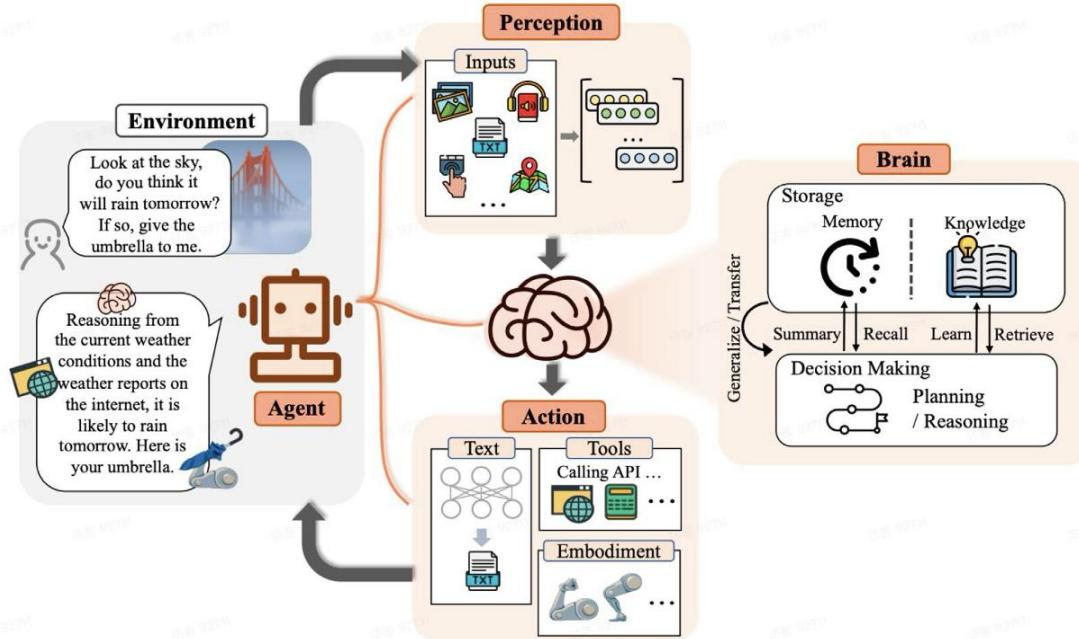
# Vision Language Model

- From (implicit) single-step prediction to (explicit) multi-step multi-modal reasoning



# Multi-modal LLM Agent

- From content-based reasoning to behavior control (w/ multimodalities)



C

WTF



COSE405

감사합니다

김진규

고려대학교 컴퓨터학과



고려대학교  
KOREA UNIVERSITY

