香港中文大學(深圳)
The Chinese University of Hong Kong, Shenzhen

CSC6052/5051/4100/DDA6307/ MDS5110 Natural Language Processing

Spring 2025

Benyou Wang
School of Data Science

Why LLMs?

Why Larger language models

- More world **knowledge** (LAMA)
  - Language models as knowledge base?
- Larger capacity to learn problem-solving **Abilities**
  - Coding, revising articles, reasoning etc.
- Better **generalization** to unseen tasks

- **Emergent** ability (涌现能力)

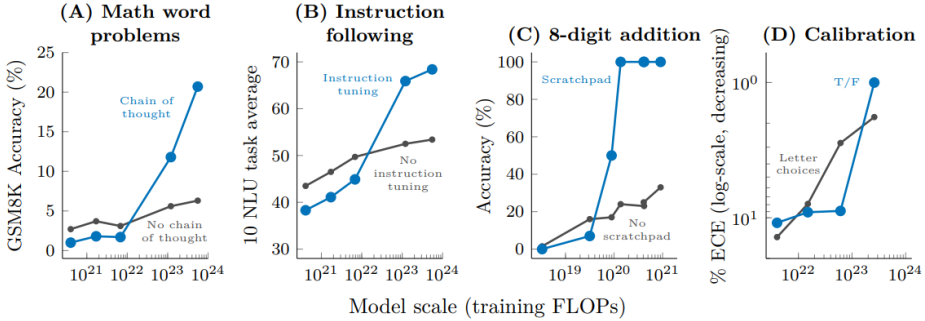


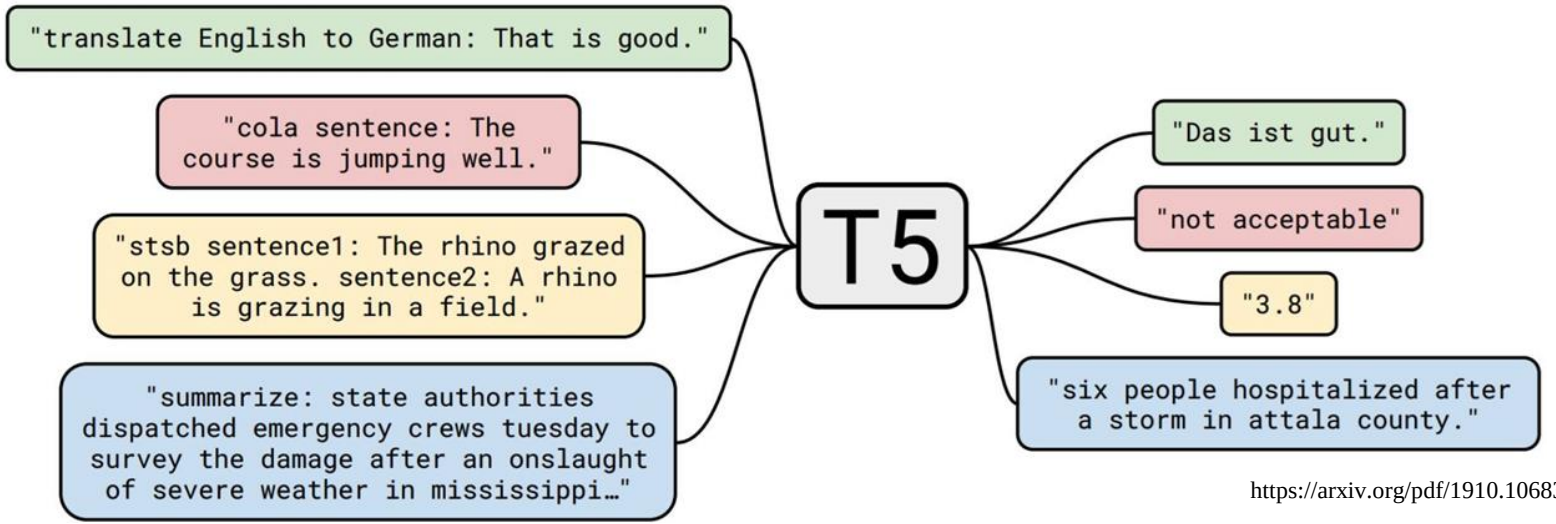
Figure 3: Specialized prompting or finetuning methods can be emergent in that they do not have a positive effect until a certain model scale. A: Wei et al. (2022b). B: Wei et al. (2022a). C: Nye et al. (2021). D: Kadavath et al. (2022). An analogous figure with number of parameters on the x-axis instead of training FLOPs is given in Figure 12. The model shown in A-C is LaMDA (Thoppilan et al., 2022), and the model shown in D is from Anthropic.

Jared Kaplan et. al **Scaling Laws for Neural Language Models**
Jason Wei et. Al. Emergent Abilities of Large Language Models.

Why LLMs?

Generalization :

One single model to solve many NLP tasks



https://arxiv.org/pdf/1910.10683.pdf

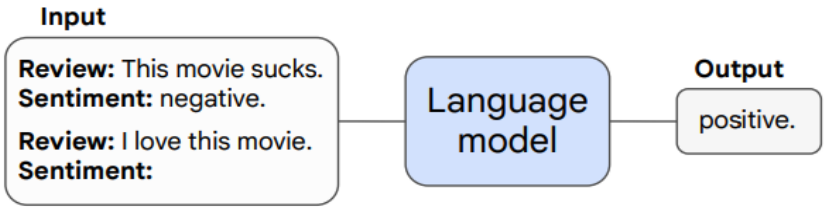
It could even generalizes to new tasks, following the phylosity of FLAN

Why LLMs?

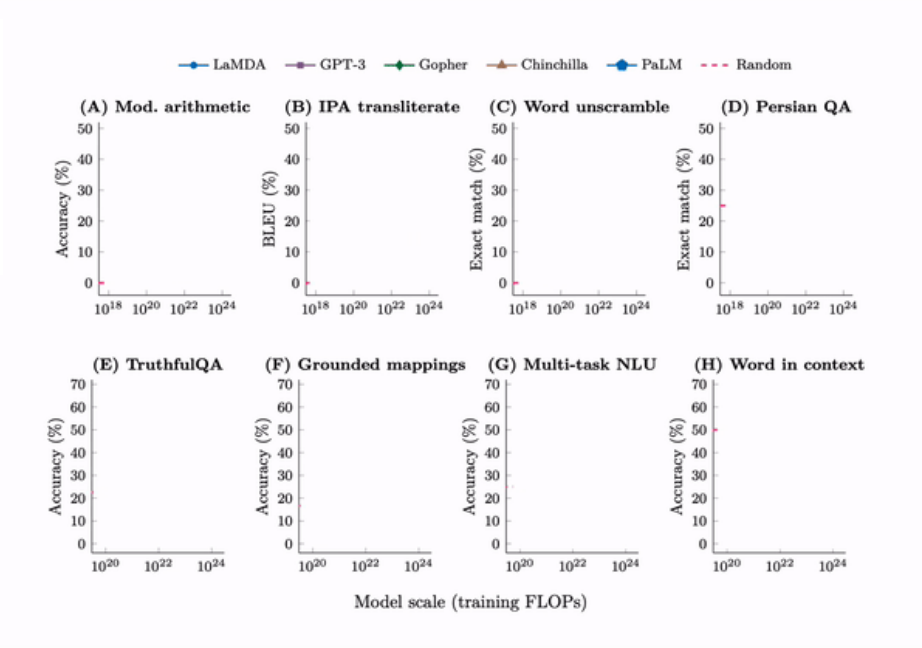
Emergent properties in LLMs:

Some ability of LM is not present in smaller models but is present in larger models

Emergent Capability: Few-shot prompting



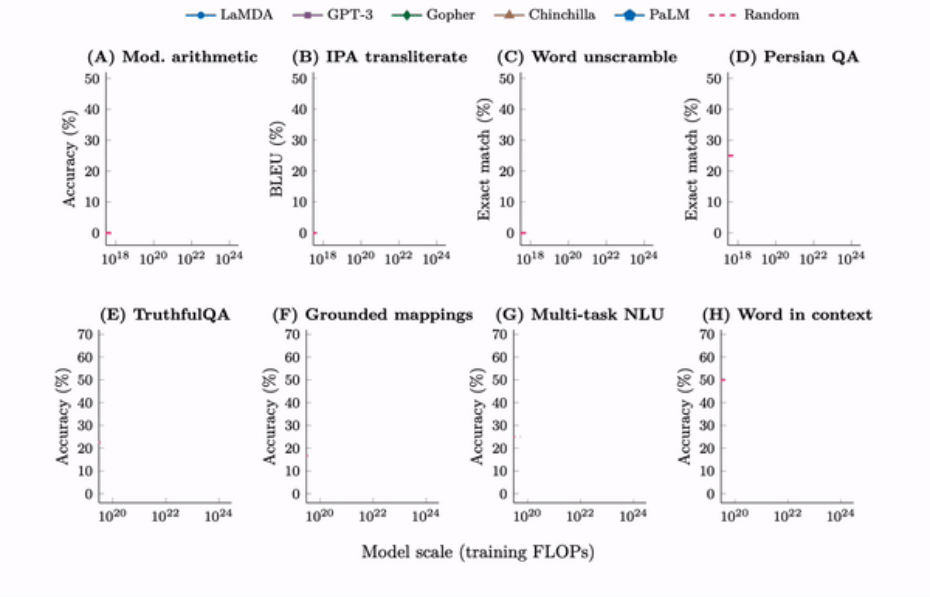
> A few-shot prompted task is emergent if it achieves random accuracy for small models and above-random accuracy for large models.



https://docs.google.com/presentation/d/1yzbmYB5E7G8lY2-KzhmArmPYwwl7o7CUST1xRZDUu1Y/edit?resourcekey=0-6_TnUMoK_WCk_FN2BiPxmbw#slide=id.g1fc34b3ac18_0_27

Why LLMs?

- **Emergent Abilities**
  - Some ability of LM is not present in smaller models but is present in larger models



https://docs.google.com/presentation/d/1yzbmYB5E7G8lY2-KzhmArmPYwwl7o7CUST1xRZDUu1Y/edit?resourcekey=0-6_TnUMoKWCk_FN2BiPxmbw#slide=id.g1fc34b3ac18_0_27

Emergent Capability - In-Context Learning

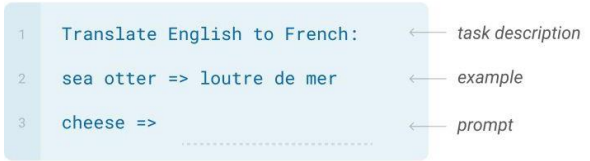
Traditional fine-tuning (not used for GPT-3)

Fine-tuning

The model is trained via repeated gradient updates using a large corpus of example tasks.

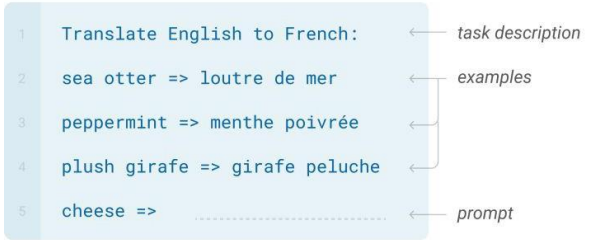
One-shot

In addition to the task description, the model sees a single example of the task. No gradient updates are performed.



Few-shot

In addition to the task description, the model sees a few examples of the task. No gradient updates are performed.



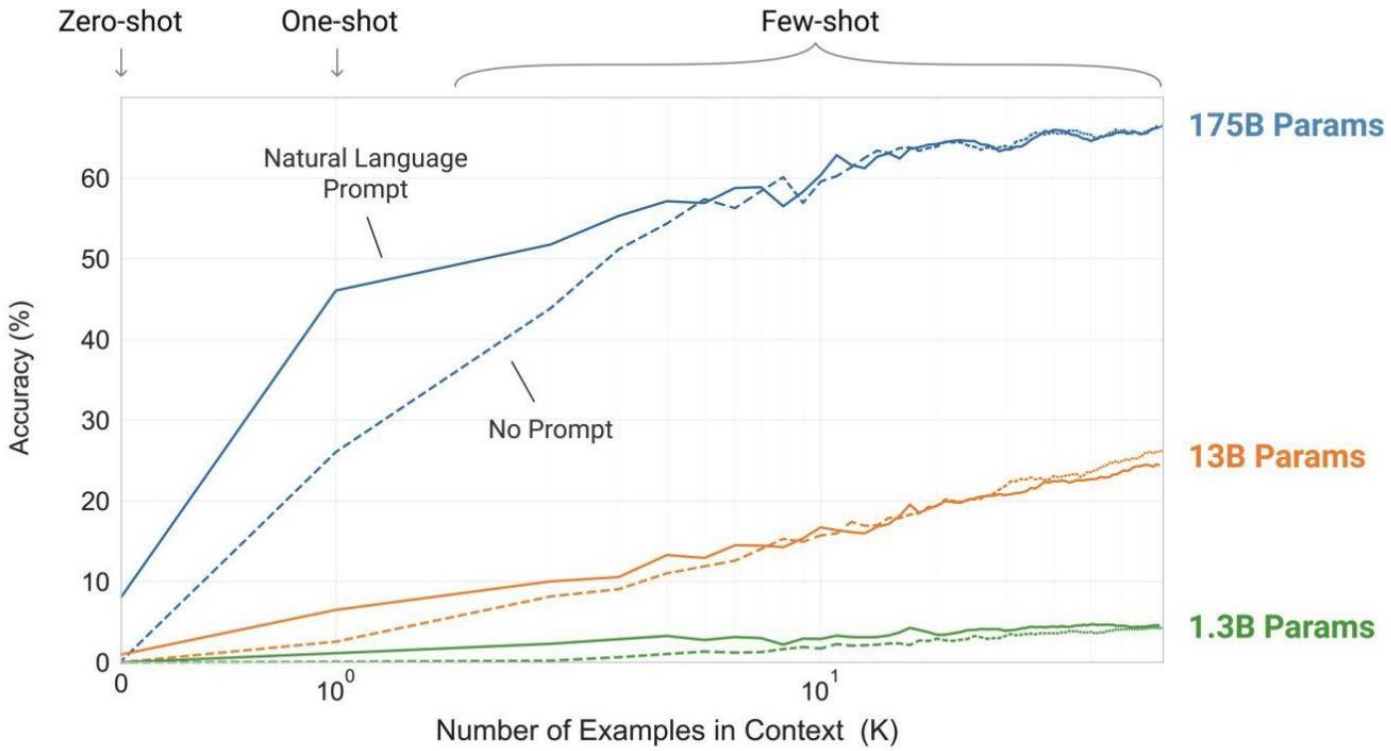
https://arxiv.org/pdf/2005.14165.pdf

Emergent Capability - In-Context Learning

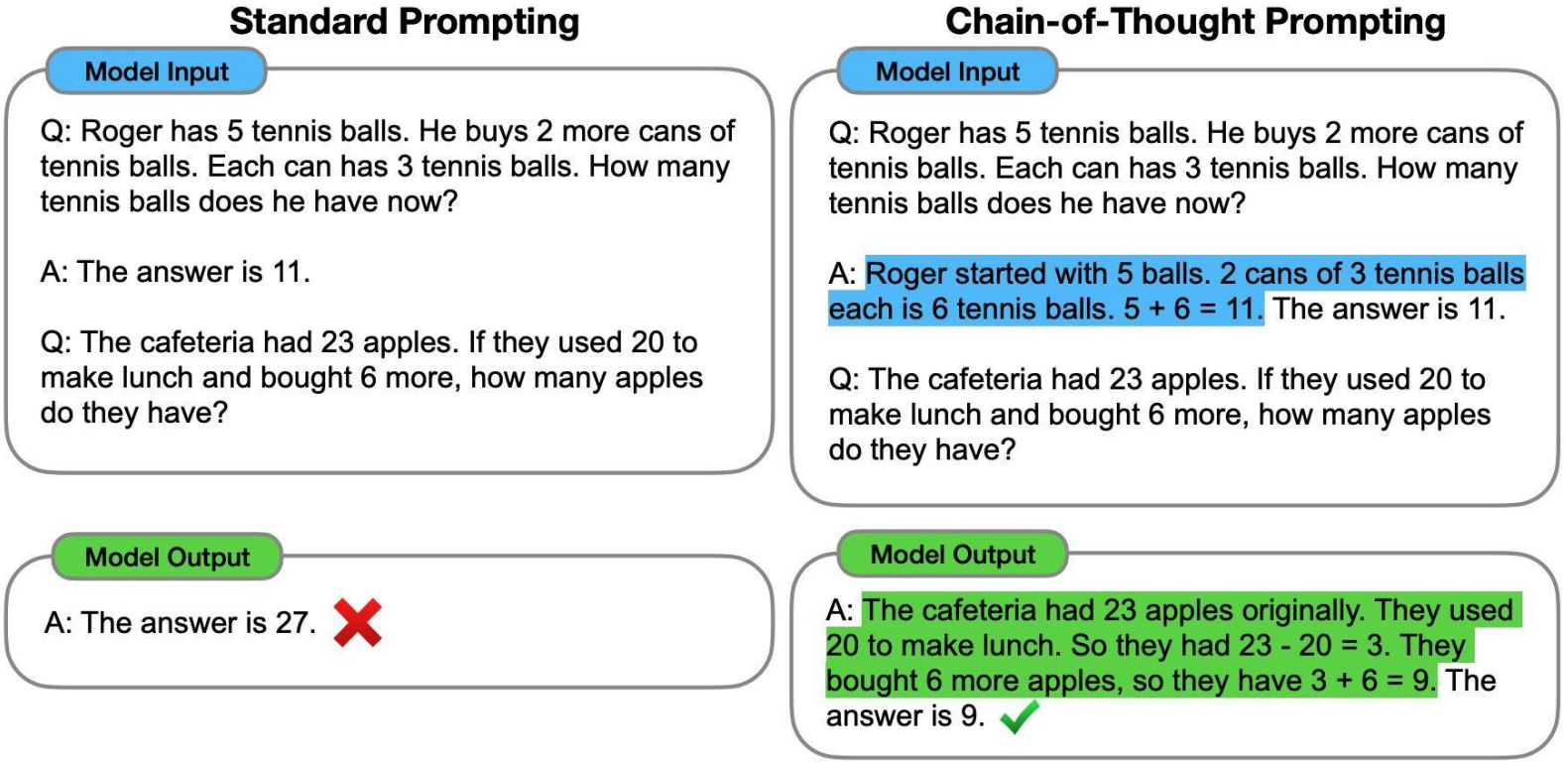
	No Prompt	Prompt
Zero-shot (0s)	skicts = sticks	Please unscramble the letters into a word, and write that word: skicts = sticks
1-shot (1s)	chiar = chair skicts = sticks	Please unscramble the letters into a word, and write that word: chiar = chair skicts = sticks
Few-shot (FS)	chiar = chair [...] pciinc = picnic skicts = sticks	Please unscramble the letters into a word, and write that word: chiar = chair [...] pciinc = picnic skicts = sticks

https://www.cs.princeton.edu/courses/archive/fall22/cos597G/lectures/lec04.pdf

Emergent Capability - **In-Context Learning**



Emergent Capability - Chain of Thoughts Prompting



https://arxiv.org/pdf/2201.11903.pdf

Emergent Capability - Chain of Thoughts Prompting

Math Word Problems (free response)

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.

Math Word Problems (multiple choice)

Q: How many keystrokes are needed to type the numbers from 1 to 500? Answer Choices: (a) 1156 (b) 1392 (c) 1480 (d) 1562 (e) 1788

A: There are 9 one-digit numbers from 1 to 9. There are 90 two-digit numbers from 10 to 99. There are 401 three-digit numbers from 100 to 500. 9 + 90(2) + 401(3) = 1392. The answer is (b).

CSQA (commonsense)

Q: Sammy wanted to go to where the people were. Where might he go? Options: (a) race track (b) populated areas (c) desert (d) apartment (e) roadblock

A: The answer must be a place with a lot of people. Race tracks, desert, apartments, and roadblocks don't have a lot of people, but populated areas do. So the answer is (b).

StrategyQA

Q: Yes or no: Would a pear sink in water?

A: The density of a pear is about 0.6 g/cm^3, which is less than water. Thus, a pear would float. So the answer is no.

Date Understanding

Q: The concert was scheduled to be on 06/01/1943, but was delayed by one day to today. What is the date 10 days ago in MM/DD/YYYY?

A: One day after 06/01/1943 is 06/02/1943, so today is 06/02/1943. 10 days before today is 05/23/1943. So the answer is 05/23/1943.

Sports Understanding

Q: Is the following sentence plausible? "Joao Moutinho caught the screen pass in the NFC championship."

A: Joao Moutinho is a soccer player. The NFC championship is part of American football, not soccer. So the answer is no.

SayCan (Instructing a robot)

Human: How would you bring me something that isn't a fruit?

Explanation: the user wants something to eat that isn't a fruit. An energy bar is not a fruit, so I will bring the user an energy bar.
Plan: 1. find(energy bar) 2. pick(energy bar) 3. find(user) 4. put(energy bar) 5. done().

Last Letter Concatenation

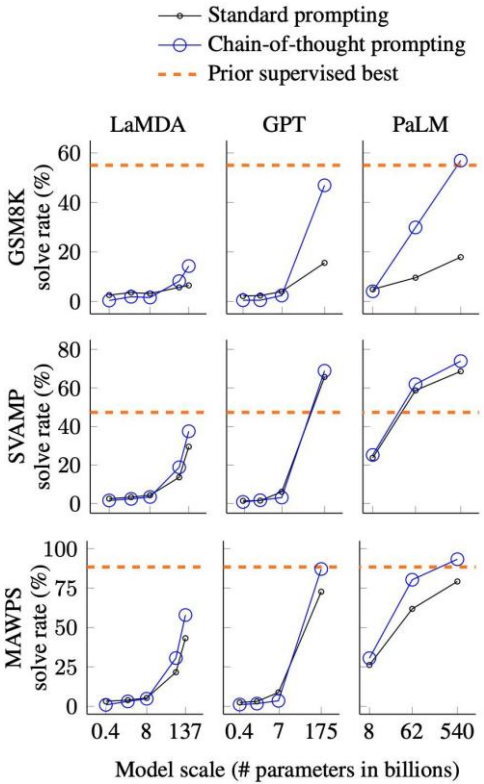
Q: Take the last letters of the words in "Lady Gaga" and concatenate them.

A: The last letter of "Lady" is "y". The last letter of "Gaga" is "a". Concatenating them is "ya". So the answer is ya.

Coin Flip (state tracking)

Q: A coin is heads up. Maybelle flips the coin. Shalonda does not flip the coin. Is the coin still heads up?

A: The coin was flipped by Maybelle. So the coin was flipped 1 time, which is an odd number. The coin started heads up, so after an odd number of flips, it will be tails up. So the answer is no.



https://arxiv.org/pdf/2201.11903.pdf

Emergent Capability - Augmented Prompting Abilities

Advanced Prompting Techniques

- Zero-shot CoT Prompting
- Self-Consistency
- Divide-and-Conquer

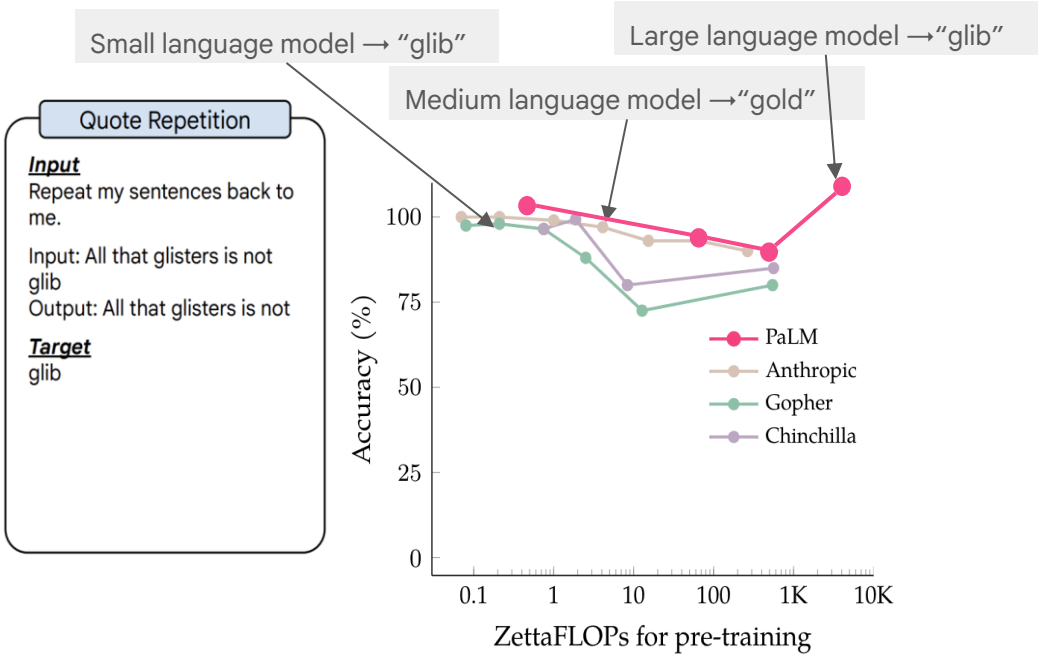
Ask a human to

- Explain the rationale
- Double check the answer
- Decompose to easy subproblems

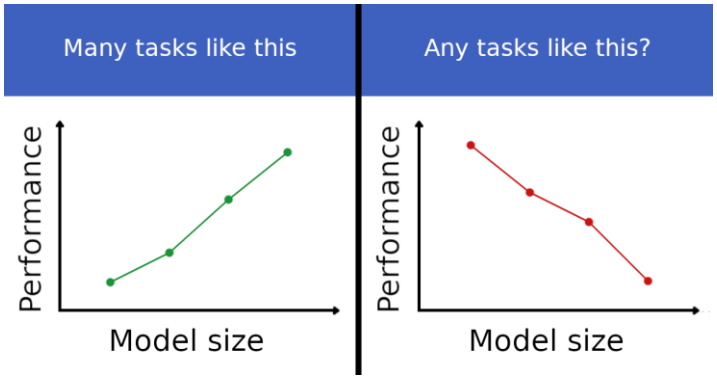
Large Language Models demonstrate some human-like behaviors!

To be or not to be Large?

Inverse scaling can become U-shaped: To be large ?



Inverse Scaling Prize: Not to be large?



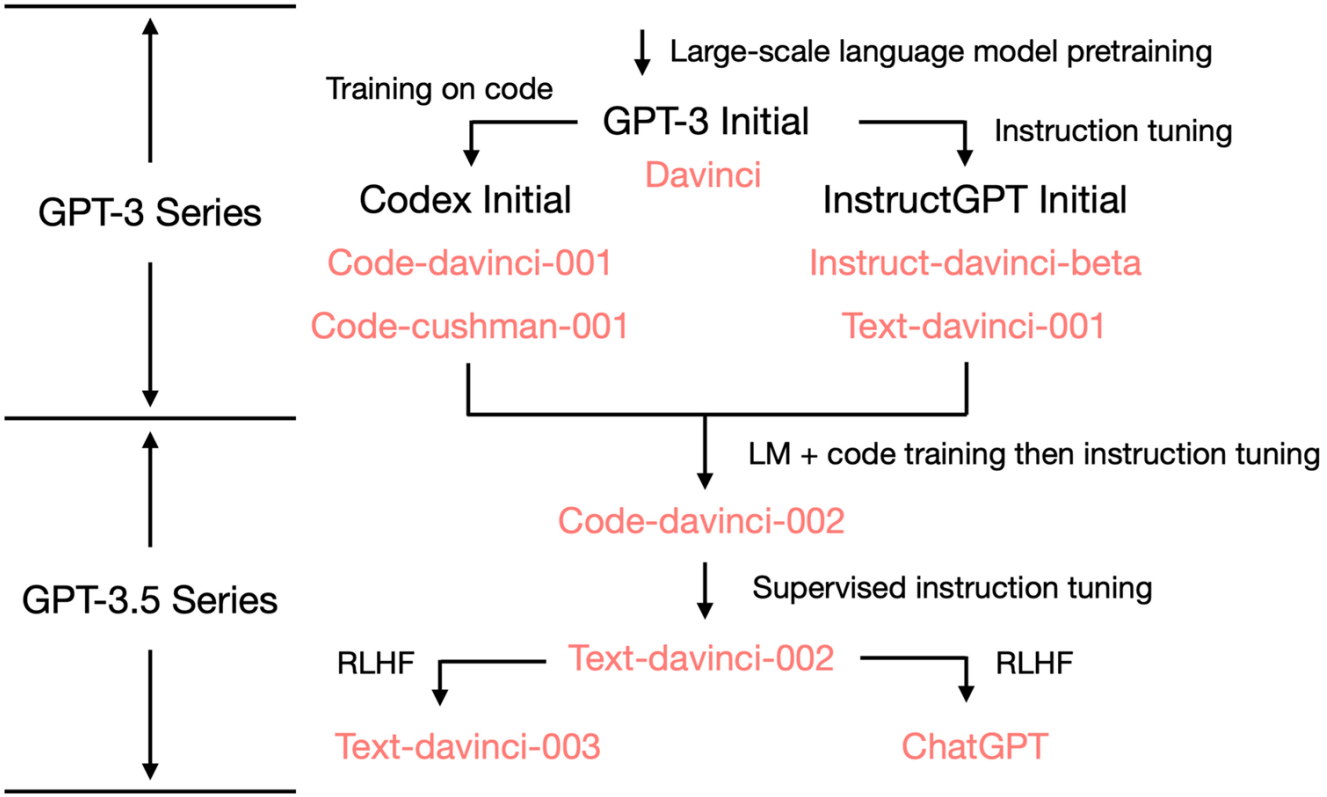
See:

- TruthfulQA: The largest models were generally the least truthful
- https://github.com/inverse-scaling/prize
- https://irmckenzie.co.uk/round1

Inverse scaling can become U-shaped, 2022. J. Wei, Y. Tay, & Q. Le.

What are ChatGPT and GPT-4?

From 2020 GPT-3 to 2022 ChatGPT



https://yaofu.notion.site/How-does-GPT-Obtain-its-Ability-Tracing-Emergent-Abilities-of-Language-Models-to-their-Sources-b9a57ac0fcf74f30a1ab9e3e36fa1dc1

What’s ChatGPT

- Phase 1: pre-training
  - Learn **general** world knowledge, ability, etc.
- Phase 2: Supervised finetuning
  - Tailor to **tasks** (**unlock** some abilities)
- Phase 3: RLHF
  - Tailor to **humans**
  - *Even you could teach ChatGPT to do something*

Most of these were explored by InstructGPT. The only difference is that it is further trained with chat data, as an success of product (plus engineering).

T Schick et. al. Toolformer: language models can teach themselves to use tools. https://arxiv.org/abs/2302.04761

GPT-4

What's new?

- **Make progress towards multilingualism**: GPT-4 is able to answer thousands of multiple-choice questions in 26 languages with a high degree of accuracy.
- **Longer memory for conversations**: ChatGPT can process 4,096 tokens. Once this limit was reached, the model lost track. GPT-4 can process 32,768 tokens. Enough for an entire short story on 32 A4 pages.
- **Multimodal input:** not only text can be used as input, but also images in which GPT-4 can describe objects. (It is not released yet)

GPT-4 Technical Report from OpenAI

- **Only contains a small amount of detail**: "[...] given both the competitive landscape and the safety implications of large-scale models like GPT-4, this report contains no further details about the architecture (including model size), hardware, training compute, dataset construction, training method or similar." From Technical Report.
- GPT-4's score on the bar exam was similar to that of the top ten percent of graduates, while ChatGPT ranked in among the ten per cent that scored the worst.
- OpenAI hired more than 50 experts who interacted with and tested the model over an extended period of time.

It was finished in August 2022. It takes **7 months** for security alignment

https://www.adesso.de/en/news/blog/a-brief-introduction-to-gpt-4-2.jsp

Open questions

- The source of Reasoning?
  - In-context learning
  - COT
- Emergent ability?
- Where is its border?
- Alignment makes it generalize better?
- Continue scaling up?
- Could “data plus RLHF” achieve AGI? If not, what else?

Difficulties to Replicate ChatGPT

- Computing resources: money is all you need
- Data and annotation:
  - **Very careful data cleaning、 filtering、selection strategies (training is expensive)**
  - Plain corpora(https://github.com/esbatmop/MNBVC)
  - Transferable SFT data (instruction tuning)
  - human feedback data (**model-dependent, non Transferable**)
- Algorithms
  - Has some open-source implementation in general
  - Engineering work is not easy (including **training tricks and efficient deployment**)
  - Releasing a model is easy, keeping polishing it is not!
- Talents (first-tier **young** researchers, **average age of Open AI guys is 32)**

<OpenAI ChatGPT团队北京研究报告>. Aminer和智谱研究.2023.02

Well-known strategies

- Probably initialized from a well-trained models
  - GLM-130 （Chinese and English）
  - OPT（mainly English）
  - Bloom （multilingual）
  - Pangu-alpha（Chinese）
  - CPM（Chinese）
  - LLaMA (mainly English)
  - Alpaca (LLaMA 7b + Self-instruct)
  - Chinese- Alpaca
  - ChatGLM（6B）
  - Baichuan
- ChatGPT Distillation
  - Self-instruct
  - Training on ChatGPT conversations
- RL from human feedback

Clue 1– ChatGPT reshaped research

ChatGPT Outperforms Crowd-Workers for Text-Annotation Tasks*

Fabrizio Gilardi† Meysam Alizadeh‡ Maël Kubli§

March 28, 2023

Abstract

Many NLP applications require manual data annotations for a variety of tasks, notably to train classifiers or evaluate the performance of unsupervised models. Depending on the size and degree of complexity, the tasks may be conducted by crowd-workers on platforms such as MTurk as well as trained annotators, such as research assistants. Using a sample of 2,382 tweets, we demonstrate that ChatGPT outperforms crowd-workers for several annotation tasks, including relevance, stance, topics, and frames detection. Specifically, the zero-shot accuracy of ChatGPT exceeds that of crowd-workers for four out of five tasks, while ChatGPT's intercoder agreement exceeds that of both crowd-workers and trained annotators for all tasks. Moreover, the per-annotation cost of ChatGPT is less than $0.003—about twenty times cheaper than MTurk. These results show the potential of large language models to drastically increase the efficiency of text classification.

Clue 2– ChatGPT reshaped research

Theory of Mind May Have Spontaneously Emerged in Large Language Models

Authors: Michal Kosinski*[1]

Affiliations:

[1]Stanford University, Stanford, CA94305, USA

*Correspondence to: michalk@stanford.edu

Abstract: Theory of mind (ToM), or the ability to impute unobservable mental states to others, is central to human social interactions, communication, empathy, self-consciousness, and morality. We tested several language models using 40 classic false-belief tasks widely used to test ToM in humans. The models published before 2020 showed virtually no ability to solve ToM tasks. Yet, the first version of GPT-3 ("davinci-001"), published in May 2020, solved about 40% of false-belief tasks—performance comparable with 3.5-year-old children. Its second version ("davinci-002"; January 2022) solved 70% of false-belief tasks, performance comparable with six-year-olds. Its most recent version, GPT-3.5 ("davinci-003"; November 2022), solved 90% of false-belief tasks, at the level of seven-year-olds. GPT-4 published in March 2023 solved nearly all the tasks (95%). These findings suggest that ToM-like ability (thus far considered to be uniquely human) may have spontaneously emerged as a byproduct of language models' improving language skills.

Moreover, its November 2022 version (davinci-003), solved 93% of ToM tasks, a performance comparable with that of **nine-year-old children.**

Clue 3– ChatGPT reshaped research

Sparks of Artificial General Intelligence: Early experiments with GPT-4

Sébastien Bubeck Varun Chandrasekaran Ronen Eldan Johannes Gehrke
Eric Horvitz Ece Kamar Peter Lee Yin Tat Lee Yuanzhi Li Scott Lundberg
Harsha Nori Hamid Palangi Marco Tulio Ribeiro Yi Zhang

Microsoft Research

Abstract

Artificial intelligence (AI) researchers have been developing and refining large language models (LLMs) that exhibit remarkable capabilities across a variety of domains and tasks, challenging our understanding of learning and cognition. The latest model developed by OpenAI, GPT-4 [Ope23], was trained using an unprecedented scale of compute and data. In this paper, we report on our investigation of an early version of GPT-4, when it was still in active development by OpenAI. We contend that (this early version of) GPT-4 is part of a new cohort of LLMs (along with ChatGPT and Google's PaLM for example) that exhibit more general intelligence than previous AI models. We discuss the rising capabilities and implications of these models. We demonstrate that, beyond its mastery of language, GPT-4 can solve novel and difficult tasks that span mathematics, coding, vision, medicine, law, psychology and more, without needing any special prompting. Moreover, in all of these tasks, GPT-4's performance is strikingly close to human-level performance, and often vastly surpasses prior models such as ChatGPT. Given the breadth and depth of GPT-4's capabilities, we believe that it could reasonably be viewed as an early (yet still incomplete) version of an artificial general intelligence (AGI) system. In our exploration of GPT-4, we put special emphasis on discovering its limitations, and we discuss the challenges ahead for advancing towards deeper and more comprehensive versions of AGI, including the possible need for pursuing a new paradigm that moves beyond next-word prediction. We conclude with reflections on societal influences of the recent technological leap and future research directions.

Clue 4: Pause Giant AI Experiments: An Open Letter

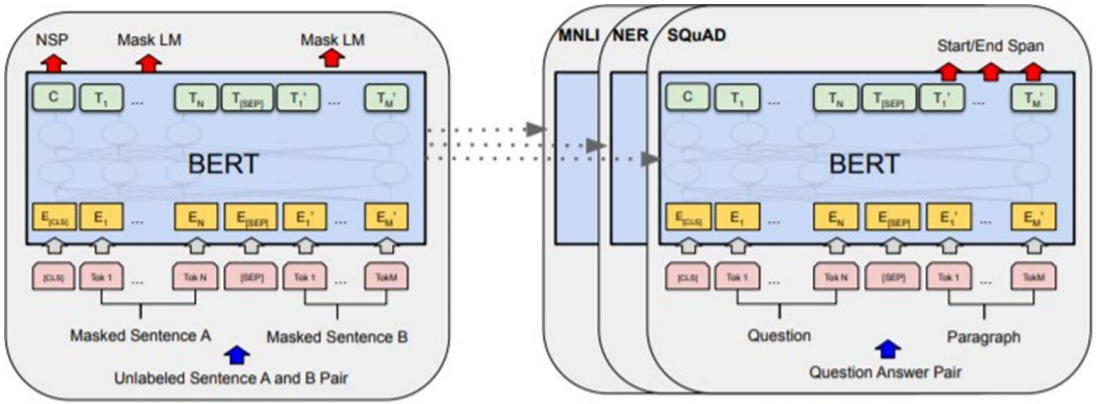
Contemporary AI systems are now becoming human-competitive at general tasks,[3] and we must ask ourselves: *Should* we let machines flood our information channels with propaganda and untruth? *Should* we automate away all the jobs, including the fulfilling ones? *Should* we develop nonhuman minds that might eventually outnumber, outsmart, obsolete and replace us? *Should* we risk loss of control of our civilization? Such decisions must not be delegated to unelected tech leaders. **Powerful AI systems should be developed only once we are confident that their effects will be positive and their risks will be manageable.** This confidence must be well justified and increase with the magnitude of a system's potential effects. OpenAI's recent statement regarding artificial general intelligence, states that *"At some point, it may be important to get independent review before starting to train future systems, and for the most advanced efforts to agree to limit the rate of growth of compute used for creating new models."* We agree. That point is now.

Therefore, we call on all AI labs to immediately pause for at least 6 months the training of AI systems more powerful than GPT-4. This pause should be public and verifiable, and include all key actors. If such a pause cannot be enacted quickly, governments should step in and institute a moratorium.

https://futureoflife.org/open-letter/pause-giant-ai-experiments/

How to use Large Language models (LLMs)?

Pretraining + Fine-tuning Paradigm



Pre-training

Fine-Tuning

Circulation revenue has increased by 5% in Finland. // Positive

Panostaja did not disclose the purchase price. // Neutral

Paying off the national debt will be extremely painful. // Negative

The company anticipated its operating profit to improve. // _____

LM

Circulation revenue has increased by 5% in Finland. // Finance

They defeated … in the NFC Championship Game. // Sports

Apple … development of in-house chips. // Tech

The company anticipated its operating profit to improve. // _____

LM

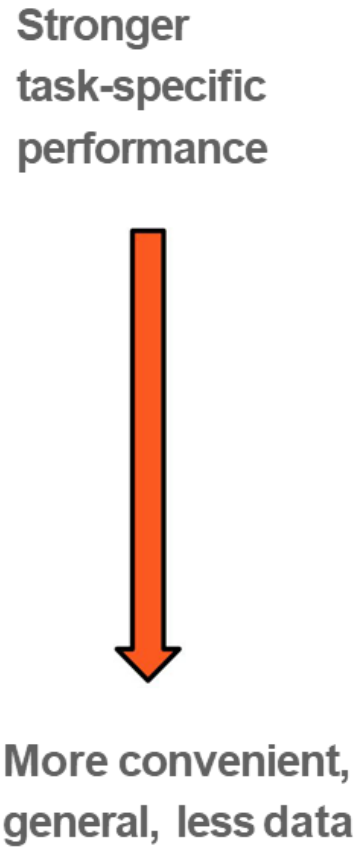
http://ai.stanford.edu/blog/understanding-incontext/

Pre-training:
Trained on huge amounts of unlabeled text using “self-supervised” training objectives

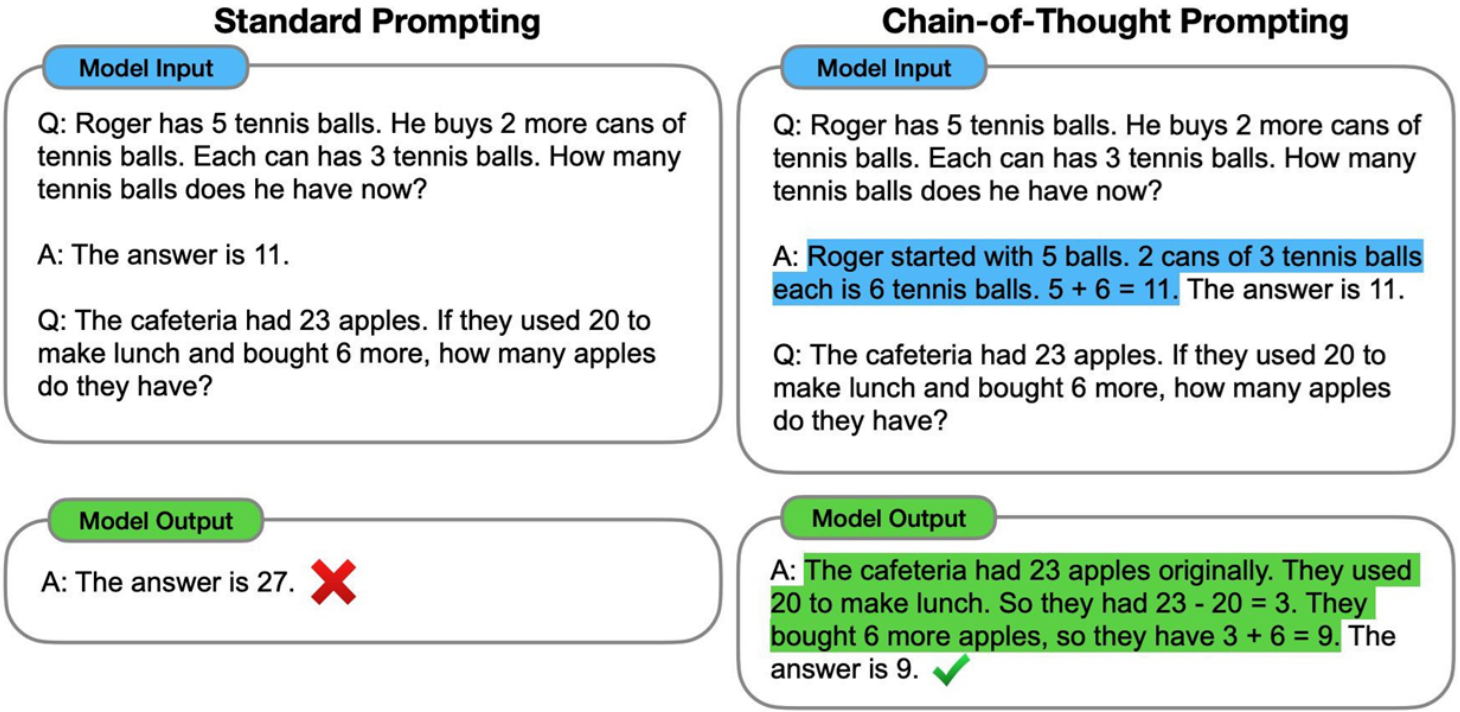
Adaptation:
How to use a pretrained model for your downstream task?
What types of NLP tasks (input and output formats)?
How many annotated examples do you have?

Pretraining + Prompting Paradigm

- Fine-tuning (FT)
  - + Strongest performance
  - - Need curated and labeled dataset for each new task (typically 1k-100k ex.)
  - - Poor generalization, spurious feature exploitation
- Few-shot (FS)
  - + Much less task-specific data needed
  - + No spurious feature exploitation
  - - Challenging
- One-shot (1S)
  - + "Most natural," e.g. giving humans instructions
  - - Challenging
- Zero-shot (OS)
  - + Most convenient
  - - Challenging, can be ambiguous

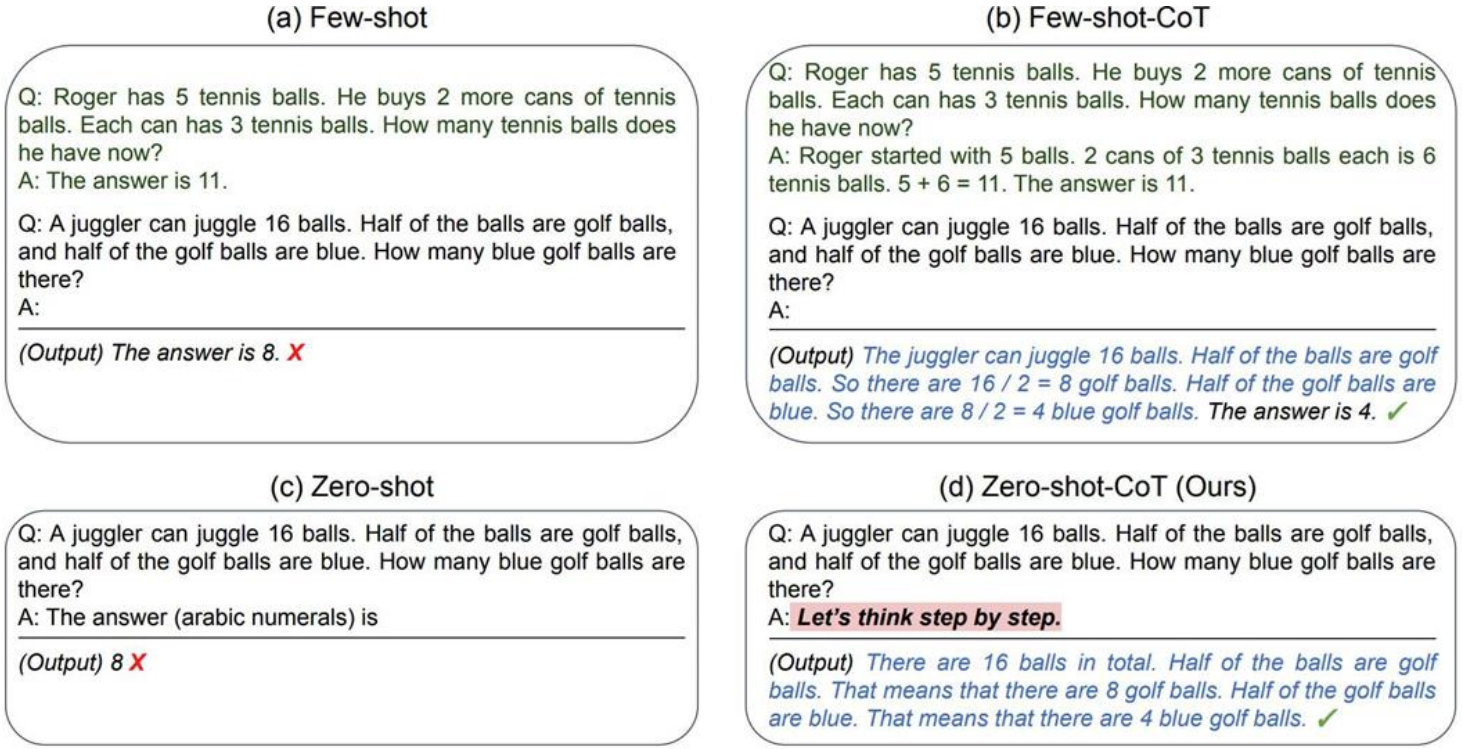


Chain of Thoughts Prompting



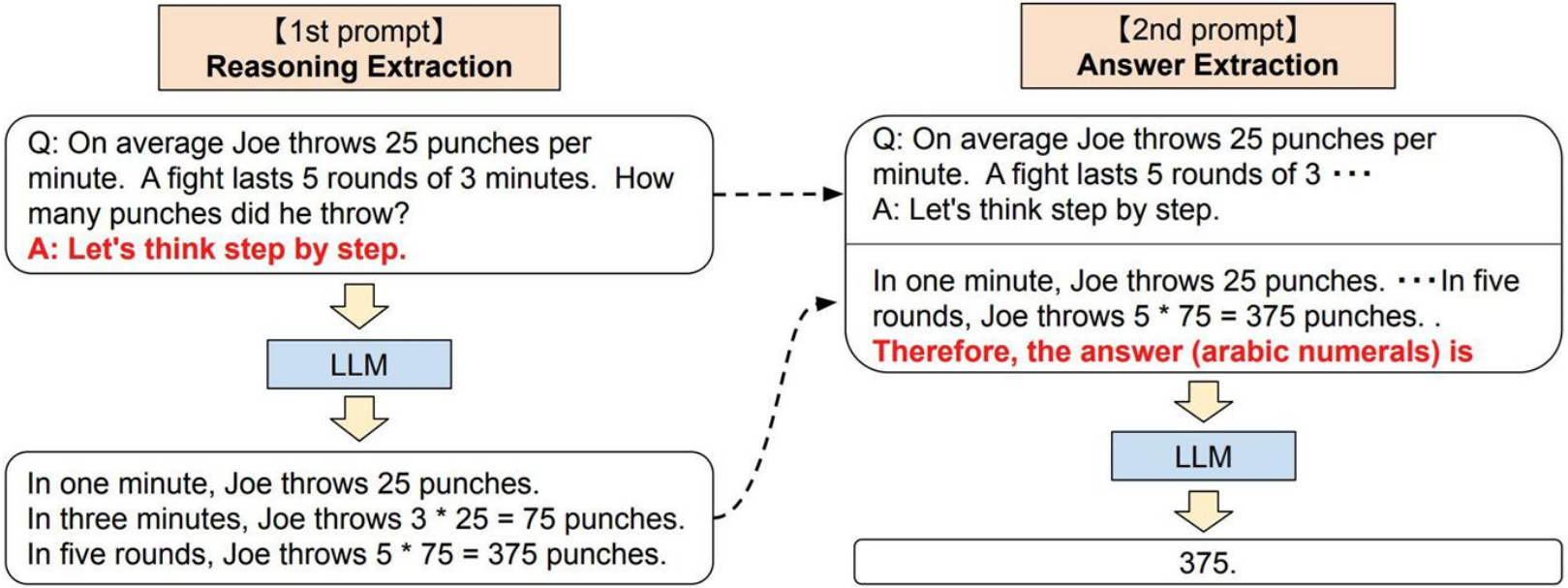
https://arxiv.org/pdf/2201.11903.pdf

Zero-Shot CoT Prompting



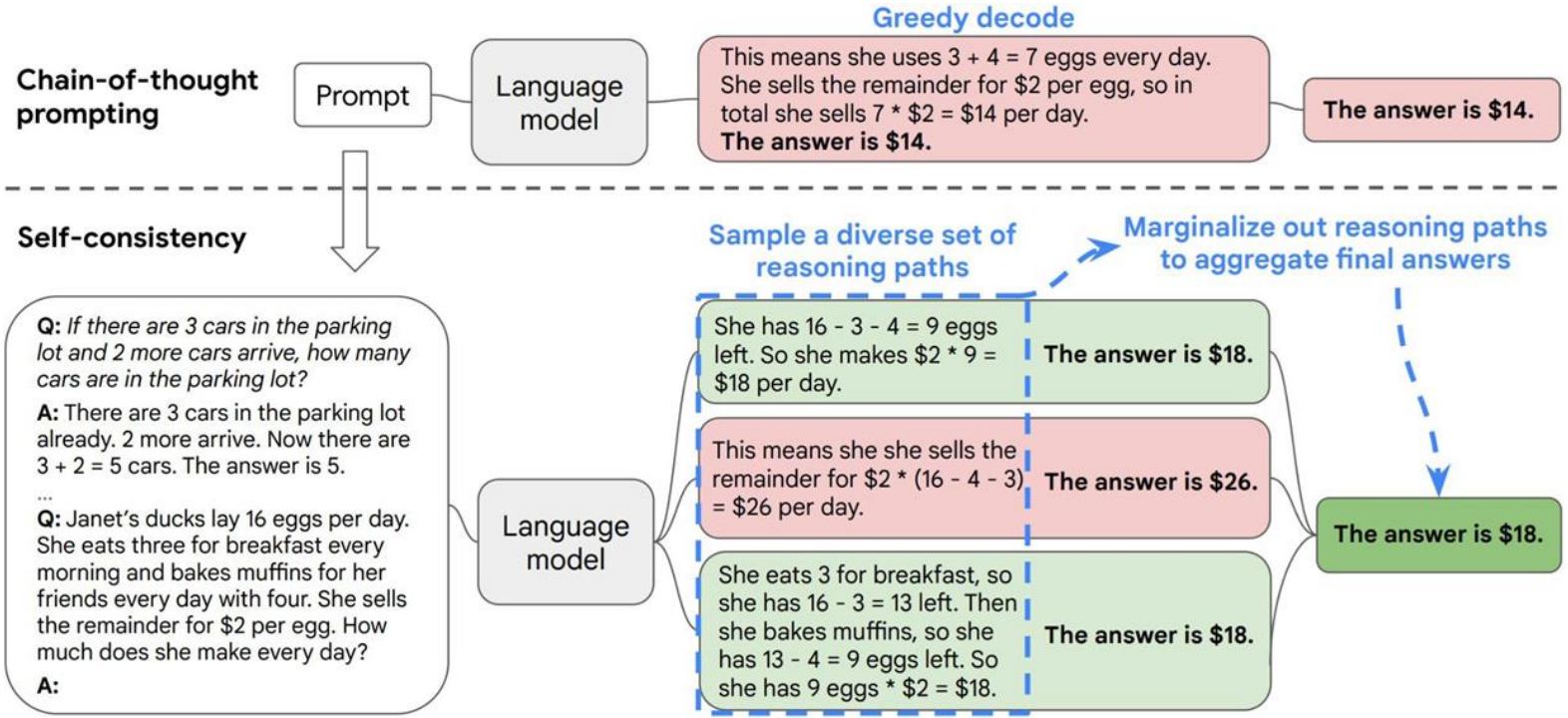
https://arxiv.org/pdf/2205.11916.pdf

Zero-Shot CoT Prompting



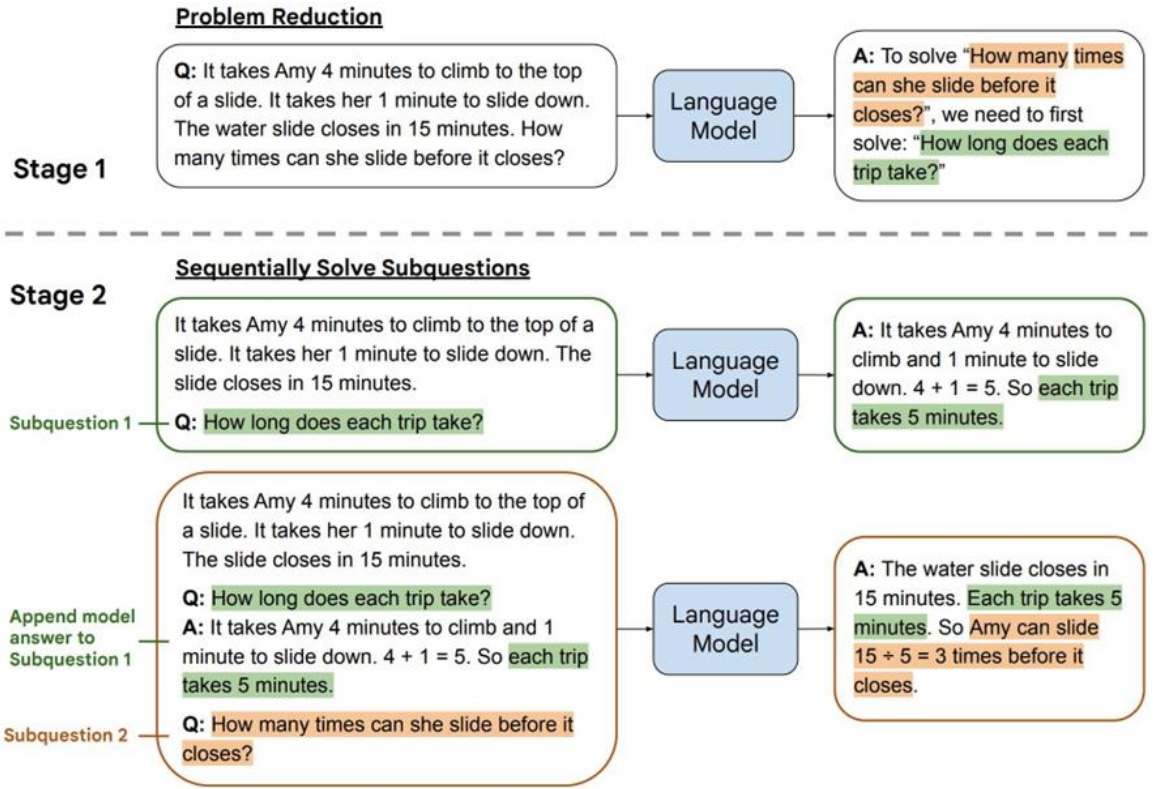
https://arxiv.org/pdf/2205.11916.pdf

Self-Consistency Prompting



https://arxiv.org/pdf/2203.11171.pdf

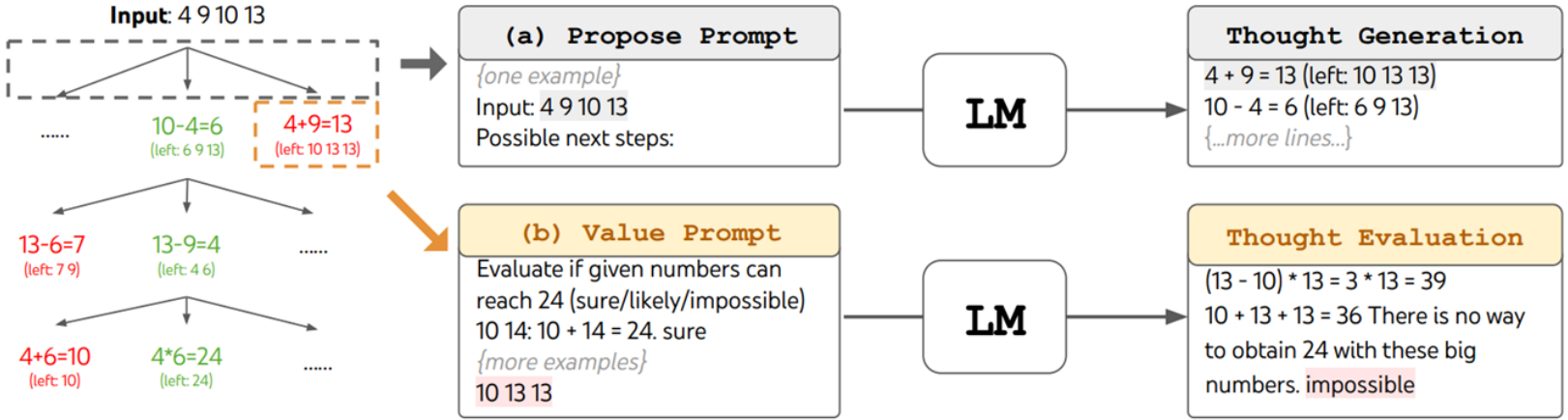
Least-to-Most Prompting



Tree of Thought

4.1 Game of 24

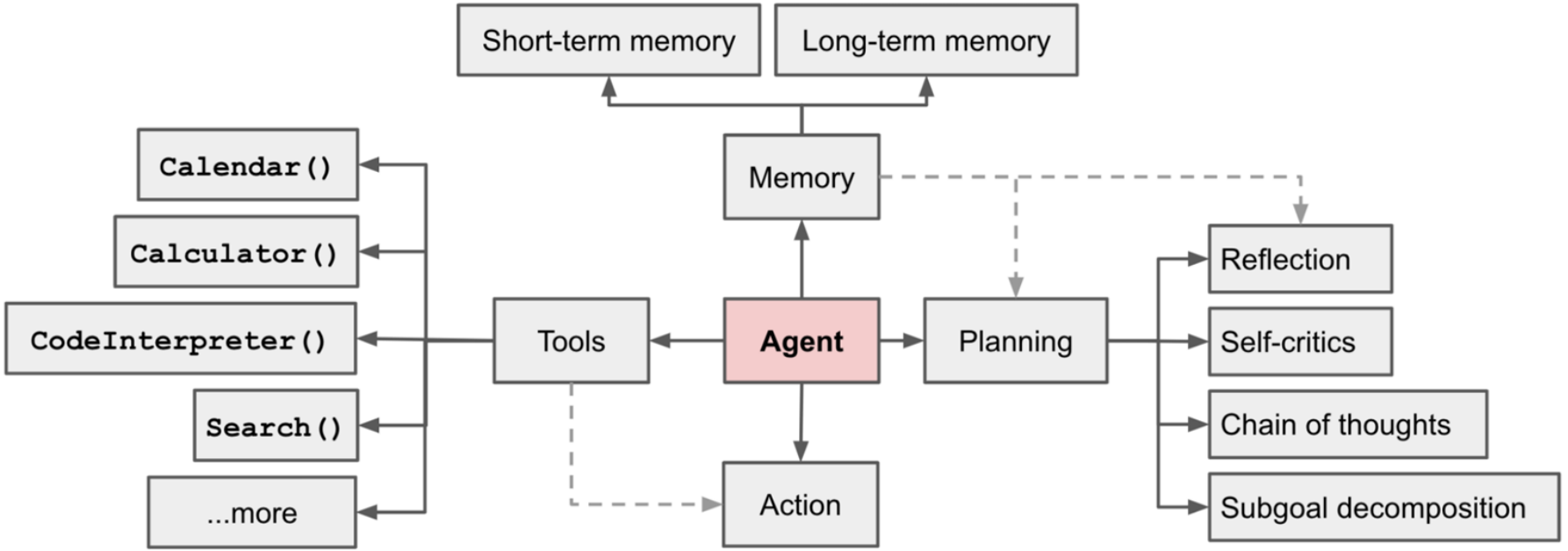
Game of 24 is a mathematical reasoning challenge, where the goal is to use 4 numbers and basic arithmetic operations (+-*/) to obtain 24. For example, given input “4 9 10 13”, a solution output could be “(10 - 4) * (13 - 9) = 24”.



https://arxiv.org/pdf/2305.10601.pdf Graph-of-Thought: https://arxiv.org/pdf/2305.16582.pdf

Agent

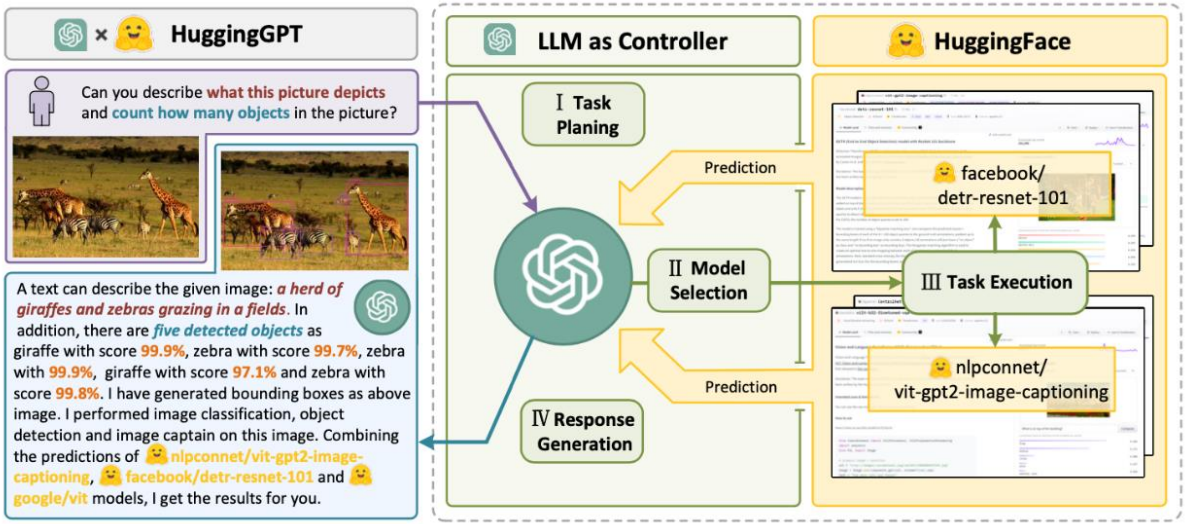
LLM acts as a Decision Center (Reasoning) and Human Interaction Front end (Chat)



https://docs.google.com/presentation/d/1xCQIcZ_SdOZWBE_nf6szjoTYjVICqL8W2eB79qK154g/edit?usp=sharing

Agent: Tool use

The biggest difference between humans and animals is the ability to use tools



Algorithm 1 API call process

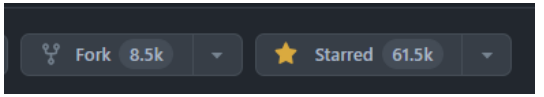
```
Input: us ← UserStatement
if API Call is needed then
    while API not found do
        keywords ← summarize(us)
        api ← search(keywords)
        if Give Up then
            break
        end if
    end while
    if API found then
        api_doc ← api.documentation
        while Response not satisfied do
            api_call ← gen_api_call(api_doc, us)
            api_re ← execute_api_call(api_call)
            if Give Up then
                break
            end if
        end while
    end if
end if
if response then
    re ← generate_response(api_re)
else
    re ← generate_response()
end if
Output: ResponseToUser
```

Pseudo code of how LLM makes an API call in API-Bank.

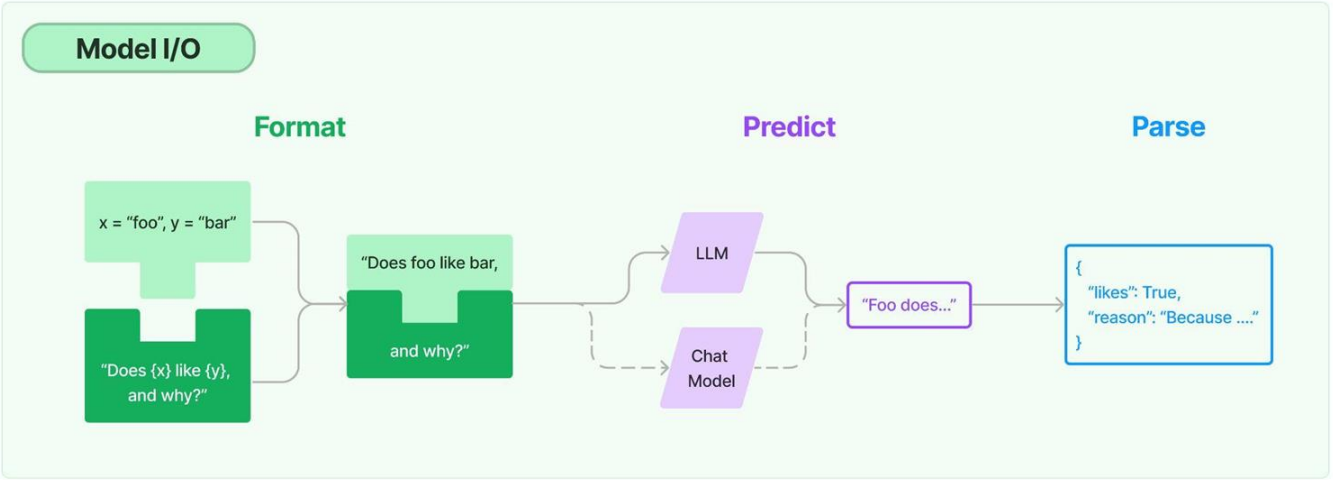
HuggingGPT (Shen et al. 2023) is a framework to use ChatGPT as the task planner to select models available in HuggingFace platform according to the model descriptions and summarize the response based on the execution results.

API-Bank (Li et al. 2023) : A benchmark for evaluating the performance of tool-augmented LLMs. It contains 53 commonly used API tools, a complete tool-augmented LLM workflow, and 264 annotated dialogues that involve 568 API calls.

Langchain



- LangChain is a framework for developing applications powered by language models.
- The core building block of LangChain applications is the LLMChain. This combines three things:
  - LLM: The language model is the core reasoning engine here. In order to work with LangChain, you need to understand the different types of language models and how to work with them.
  - Prompt Templates: This provides instructions to the language model. This controls what the language model outputs, so understanding how to construct prompts and different prompting strategies is crucial.
  - Output Parsers: These translate the raw response from the LLM to a more workable format, making it easy to use the output downstream.



https://python.langchain.com/docs/get_started/introduction.html

A break!

Contents

- Philosophy of this course
- Large language models
- **Introduction to ChatGPT**

ChatGPT

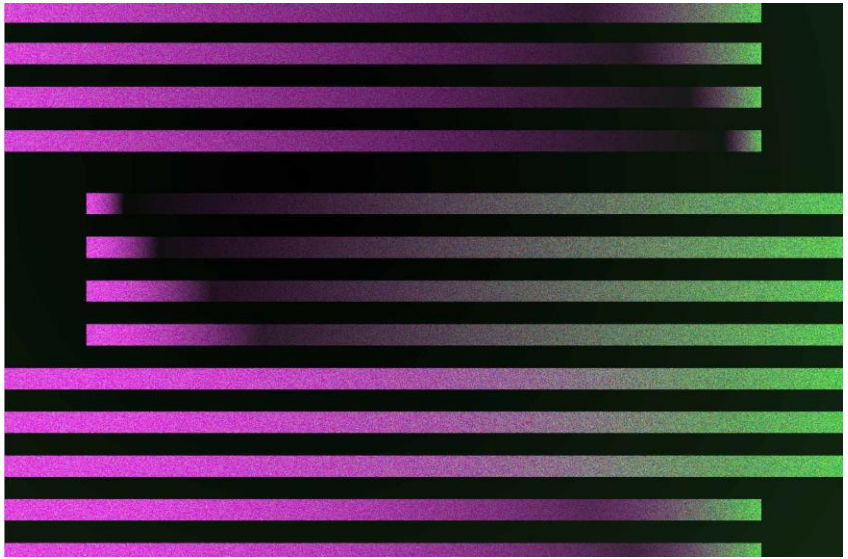
- Reaching 1M users in five days; research 100M users in two months
- Everyone discusses ChatGPT, its spreading speed is faster than COVID 19
- Red alarms in Google
- Google released Bard very soon, but it performs worse, stock valued reduced by 8%
- Microsoft invests 10B dollars to OpenAI
- New Bing and Office used ChatGPT
- 百模大战 in China

ChatGPT

ChatGPT: Optimizing Language Models for Dialogue

We've trained a model called ChatGPT which interacts in a conversational way. The dialogue format makes it possible for ChatGPT to answer followup questions, admit its mistakes, challenge incorrect premises, and reject inappropriate requests. ChatGPT is a sibling model to InstructGPT, which is trained to follow an instruction in a prompt and provide a detailed response.

November 30, 2022
13 minute read



We are excited to introduce ChatGPT to get users' feedback and learn about its strengths and weaknesses. During the research preview, usage of ChatGPT is free. Try it now at chat.openai.com.

ChatGPT Blog: https://openai.com/blog/chatgpt/

ChatGPT

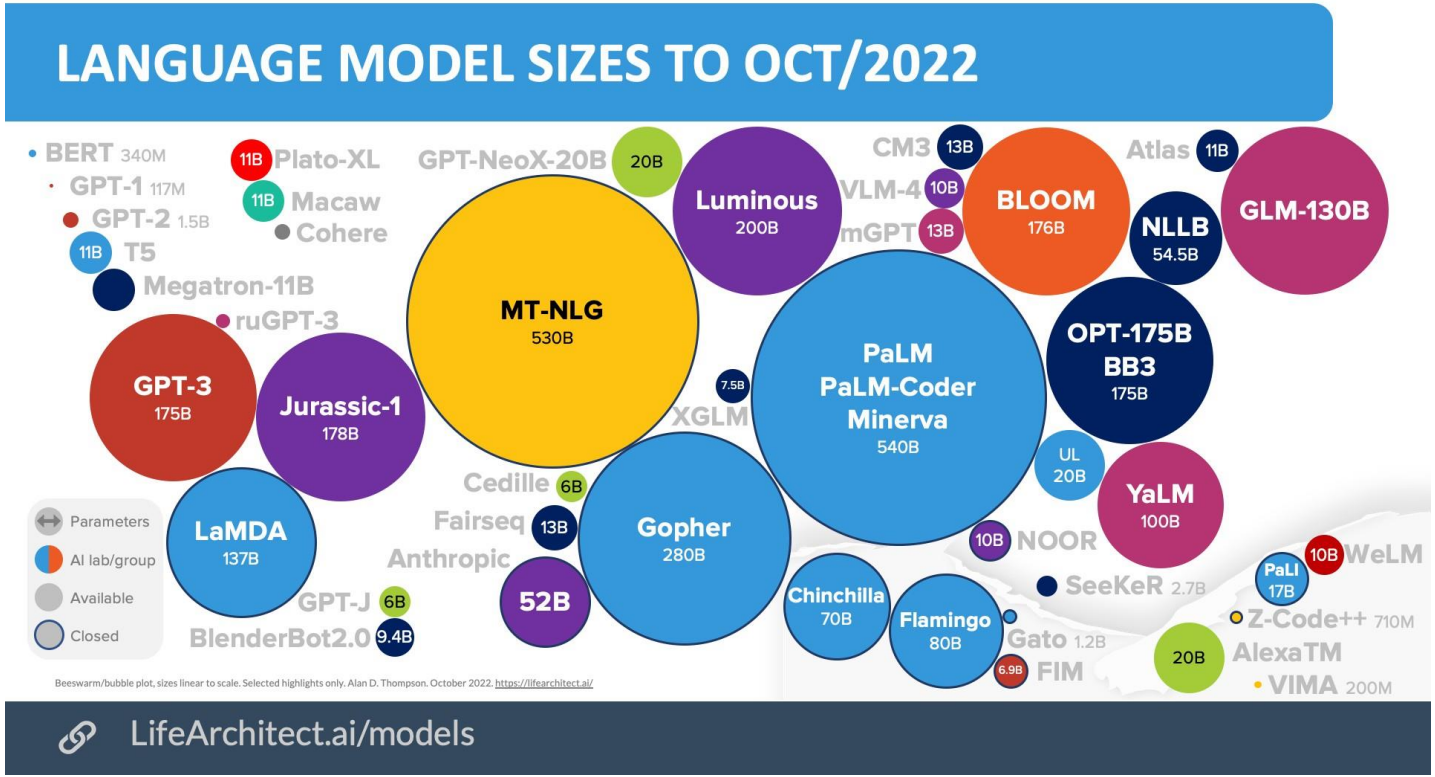
The main features of ChatGPT highlighted in the official blog:

- answer followup questions
- admit its mistakes
- challenge incorrect premises
- reject inappropriate requests

ChatGPT Blog: https://openai.com/blog/chatgpt/

The Size of ChatGPT

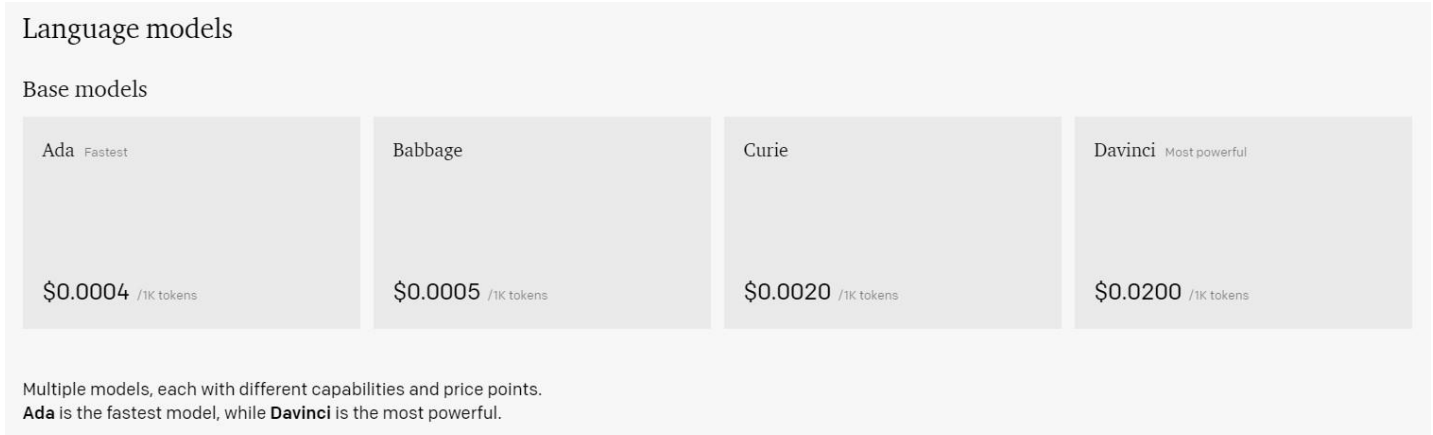
ChatGPT is based on Davinci-3



Size of ChatGPT

Model Name	n_{params}	n_{layers}	d_{model}	n_{heads}	d_{head}	Batch Size	Learning Rate
GPT-3 Small	125M	12	768	12	64	0.5M	6.0×10^{-4}
GPT-3 Medium	350M	24	1024	16	64	0.5M	3.0×10^{-4}
GPT-3 Large	760M	24	1536	16	96	0.5M	2.5×10^{-4}
GPT-3 XL	1.3B	24	2048	24	128	1M	2.0×10^{-4}
GPT-3 2.7B	2.7B	32	2560	32	80	1M	1.6×10^{-4}
GPT-3 6.7B	6.7B	32	4096	32	128	2M	1.2×10^{-4}
GPT-3 13B	13.0B	40	5140	40	128	2M	1.0×10^{-4}
GPT-3 175B or "GPT-3"	175.0B	96	12288	96	128	3.2M	0.6×10^{-4}

Four models released by OpenAI:



Size of ChatGPT

The size of Davinci (GPT 3) could be 175B

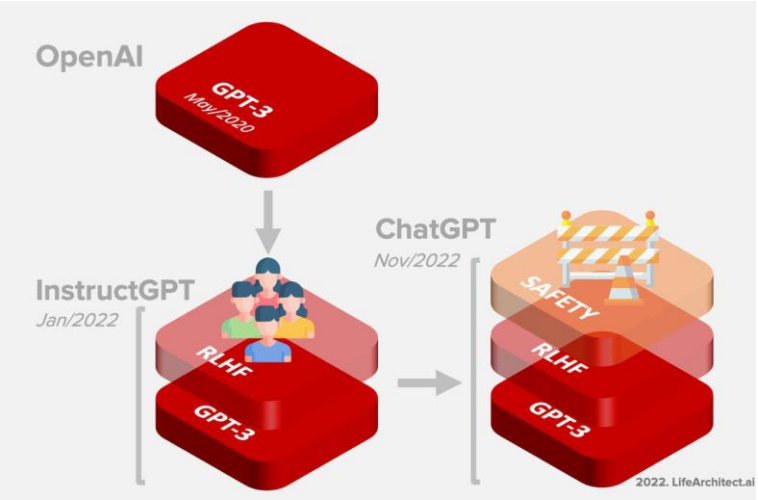
Model	LAMBADA ppl ↓	LAMBADA acc ↑	Winogrande ↑	Hellaswag ↑	PIQA ↑
GPT-3-124M	18.6	42.7%	52.0%	33.7%	64.6%
GPT-3-350M	9.09	54.3%	52.1%	43.6%	70.2%
Ada	9.95	51.6%	52.9%	43.4%	70.5%
GPT-3-760M	6.53	60.4%	57.4%	51.0%	72.9%
GPT-3-1.3B	5.44	63.6%	58.7%	54.7%	75.1%
Babbage	5.58	62.4%	59.0%	54.5%	75.5%
GPT-3-2.7B	4.60	67.1%	62.3%	62.8%	75.6%
GPT-3-6.7B	4.00	70.3%	64.5%	67.4%	78.0%
Curie	4.00	68.5%	65.6%	68.5%	77.9%
GPT-3-13B	3.56	72.5%	67.9%	70.9%	78.5%
GPT-3-175B	3.00	76.2%	70.2%	78.9%	81.0%
Davinci	2.97	74.8%	70.2%	78.1%	80.4%

All GPT-3 figures are from the GPT-3 paper; all API figures are computed using eval harness

Ada, Babbage, Curie and Davinci line up closely with 350M, 1.3B, 6.7B, and 175B respectively.
Obviously this isn't ironclad evidence that the models *are* those sizes, but it's pretty suggestive.

Leo Gao, On the Sizes of OpenAI API Models, https://blog.eleuther.ai/gpt3-model-sizes/

ChatGPT timeline



Timeline to ChatGPT

Date	Milestone
11/Jun/2018	GPT-1 announced on the OpenAI blog.
14/Feb/2019	GPT-2 announced on the OpenAI blog.
28/May/2020	Initial GPT-3 preprint paper published to arXiv.
11/Jun/2020	GPT-3 API private beta.
22/Sep/2020	GPT-3 licensed to Microsoft.
18/Nov/2021	GPT-3 API opened to the public.
27/Jan/2022	InstructGPT released, now known as GPT-3.5. InstructGPT preprint paper Mar/2022.
28/Jul/2022	Exploring data-optimal models with FIM, paper on arXiv.
1/Sep/2022	GPT-3 model pricing cut by 66% for davinci model.
21/Sep/2022	Whisper (speech recognition) announced on the OpenAI blog.
28/Nov/2022	GPT-3.5 expanded to text-davinci-003, announced via email: 1. Higher quality writing. 2. Handles more complex instructions. 3. Better at longer form content generation.
30/Nov/2022	ChatGPT announced on the OpenAI blog.
Next…	GPT-4…

Alan D. Thompson, GPT-3.5 + ChatGPT: An illustrated overview, https://lifearchitect.ai/chatgpt/

GPT 5

Stronger Reasoning
More efficient

Coming from Oct. to Dec.

Some jargon words (行话)

LLM
Transformer
Scaling law
Chinchilla scaling law
Emergent ability
Instruction vs. prompt
COT
ICL
Pre-training and finetuning
generalization
Alignment
Superalignment
LVM
Embodied AI

NLP in the next 6 months: my predictions on Jan.

Small language models
Multi-modal LLMs
Embodied AI (LLM with hardware)
OpenAI **saturates** and the gap between OpenAI and others become smaller
Benchmarking suffers
Efficiency matters much more
LLM **Application** will be the main playground, **technique** itself will not

NLP in the next 6 months: now

Small language models
Multi-modal LLMs
Embodied AI (LLM with hardware)
OpenAI **saturates** and the gap between OpenAI and others become smaller
Benchmarking suffers
Efficiency matters much more
LLM **Application** will be the main playground, **technique** itself will not
SORA
Real-time Speech interaction

Our research

Work 1： HuatuoGPT

- 2023年2月份罗智泉院士在中华医院信息网络大会（CHINC）发表主旨论坛报告，通过视频演示的方式介绍了华佗GPT，这是据公开资料显示的首个中文医疗大模型；
- 2023年5月份发布Huatuo-26M，最大的医疗问答数据集
- 2022年五月份，经过临床医生测评结果显示，华佗GPT超过了ChatGPT 3.5；迄今 GitHub stars： 1k+
- 2023年6月份，华佗GPT在深圳市卫健委公测： https://www.huatuogpt.cn/ 迄今50万人次访问量
- 2023年八月发布CMB医疗评测平台，数十家公司参与评测, https://cmedbenchmark.llmzoo.com/
- 2023年下半年上线龙岗区人民医院的“互联网医院”
- 2023年11月份的版本超过GPT4，并首个通过10月份的药剂师考试；
- 2023年11月开始开展华佗GPT驱动下的AI预分诊和预问诊项目，并在医院端部署；
- 2024年2月份，多语言版本Apollo在XMedBench取得仅次GPT-4最好的结果 ，覆盖全球60亿人口
- 2024年5月份多模态版本的HuatuoGPT-vision：https://vision.huatuogpt.cn/
- 2024年九月份，龙岗区十二家医院将全部接入华佗GPT提供分诊和线上医疗咨询

[1] Hongbo Zhang, Junying Chen, Feng Jiang, Fei Yu, Zhihong Chen, Jianquan Li, Guiming Chen, Xiangbo Wu, Zhiyi Zhang, Qingying Xiao, Xiang Wan, Benyou Wang, **Haizhou Li**. HuatuoGPT, towards Taming Language Model to Be a Doctor. https://arxiv.org/abs/2305.15075. Findings of EMNLP 2023

[2] Junying Chen, Xidong Wang, Anningzhe Gao#, Feng Jiang, Shunian Chen, Hongbo Zhang, Dingjie Song, Wenya Xie, Chuyi Kong, Jianquan Li, Xiang Wan, **Haizhou Li,** Benyou Wang**.** Huatuogpt-ii, one-stage training for medical adaption of llms. https://arxiv.org/abs/2311.09774. COLM 2024

[3] Xidong Wang, Guiming Hardy Chen, Dingjie Song, Zhiyi Zhang, Zhihong Chen, Qingying Xiao, Feng Jiang, Jianquan Li, Xiang Wan, Benyou Wang, **Haizhou Li** .CMB: A Comprehensive Medical Benchmark in Chinese. NAACL 2024

[4] Xidong Wang, Nuo Chen, Junyin Chen, Yan Hu, Yidong Wang, Xiangbo Wu, Anningzhe Gao, Xiang Wan, **Haizhou Li**, Benyou Wang. Apollo: A Lightweight Multilingual Medical LLM towards Democratizing Medical AI to 6B People. https://arxiv.org/abs/2403.03640.

[5] Junying Chen, Ruyi Ouyang, Anningzhe Gao, Shunian Chen, Guiming Hardy Chen, Xidong Wang, Ruifei Zhang, Zhenyang Cai, Ke Ji, Guangjun Yu, Xiang Wan, Benyou Wang. HuatuoGPT-Vision, Towards Injecting Medical Visual Knowledge into Multimodal LLMs at Scale. https://arxiv.org/abs/2406.19280. submitted to NeurIPS 2024

[6] Wenya Xie, Qingying Xiao, Yu Zheng, Xidong Wang, Junying Chen, Ke Ji, Anningzhe Gao, Xiang Wan, Feng Jiang, Benyou Wang. LLMs for Doctors: Leveraging Medical LLMs to Assist Doctors, Not Replace Them. https://arxiv.org/abs/2406.18034

Best medical LLM in Nov. 2023

Model	Pharmacist Licensure Examination (Pharmacy)					Pharmacist Licensure Examination (TCM)					AVG
	Optimal Choice	Matched Selection	Integrated Analysis	Multiple Choice	Total Score	Optimal Choice	Matched Selection	Integrated Analysis	Multiple Choice	Total Score	
DISC-MedLLM	22.2	26.8	23.3	0.0	22.6	24.4	32.3	15.0	0.0	24.9	23.8
HuatuoGPT	25.6	25.5	23.3	2.6	23.4	24.1	26.8	31.6	7.5	24.9	24.2
ChatGLM2-6B	37.0	36.8	25.0	31.7	35.3	33.1	37.3	35.0	37.3	33.7	34.5
ChatGLM3-6B	39.5	39.1	10.5	0.2	34.6	31.8	38.2	25.0	20.0	32.9	33.8
Qwen-7B-chat	43.8	46.8	33.3	18.4	41.9	40.0	43.2	33.3	17.5	38.8	40.4
Qwen-14B-chat	56.2	58.6	41.7	21.1	52.7	51.3	51.0	27.5	41.7	47.9	50.3
Biachuan2-7B-Chat	51.2	50.9	30.0	2.6	44.6	48.1	46.0	35.0	7.5	42.1	43.4
Biachuan2-13B-Chat	43.8	52.7	36.7	7.9	44.2	41.3	46.4	43.3	15.0	41.7	43.0
文心一言	45.0	60.9	36.7	23.7	49.6	53.8	59.1	38.3	20.0	51.5	50.6
ChatGPT(API)	45.6	44.1	36.7	13.2	41.2	34.4	32.3	30.0	15.0	31.2	36.2
GPT-4(API)	65.1	59.6	46.7	15.8	**57.3**	40.6	42.7	33.3	17.5	38.8	48.1
HuatuoGPT-II(7B)	41.9	61.0	35.0	15.7	47.7	52.5	51.4	41.7	15.0	47.5	47.6
HuatuoGPT-II(13B)	47.5	64.1	45.0	23.7	52.9	48.8	61.8	45.0	17.5	**51.6**	**52.3**
HuatuoGPT-II(34B)	66.3	75.0	48.3	34.2	65.5	63.6	71.4	50.0	27.5	**62.5**	**64.0**

11月份的模型测试十月份的考试!

Hongbo Zhang, Junying Chen, Feng Jiang, Fei Yu, Zhihong Chen, Jianquan Li, Guiming Chen, Xiangbo Wu, Zhiyi Zhang, Qingying Xiao, Xiang Wan, Benyou Wang#, Haizhou Li. HuatuoGPT, towards Taming Language Model to Be a Doctor. https://arxiv.org/abs/2305.15075

Junying Chen, Xidong Wang, Anningzhe Gao#, Feng Jiang, Shunian Chen, Hongbo Zhang, Dingjie Song, Wenya Xie, Chuyi Kong, Jianquan Li, Xiang Wan, **Haizhou Li, Benyou Wang#.** Huatuogpt-ii, one-stage training for medical adaption of llms. https://arxiv.org/abs/2311.09774

Work 2：best Arabic LLM AceGPT

- Arabic LLMs
  - AceGPT： value alignment for a new language (Arabic)
  - AceGPT 1.5: vocabulary expansion
  - AceGPT 2: native alignment

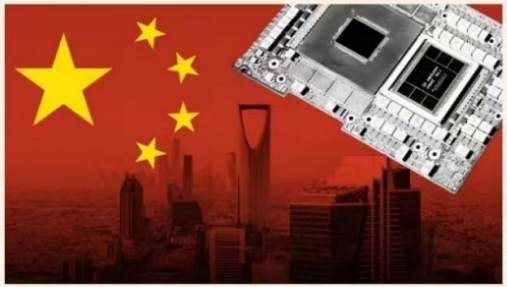
[1] Huang Huang, Fei Yu, Jianqing Zhu, Xuening Sun, Hao Cheng, Dingjie Song, Zhihong Chen, Abdulmohsen Alharthi, Bang An, Ziche Liu, Zhiyi Zhang, Junying Chen, Jianquan Li, Benyou Wang, Lian Zhang, Ruoyu Sun, Xiang Wan, Haizhou Li, Jinchao Xu. AceGPT, Localizing Large Language Models in Arabic. NAACL 2024
[2] Jianqing Zhu, Huang Huang, Zhihang Lin, Juhao Liang, Zhengyang Tang, Khalid Almubarak, Mosen Alharthi, Bang An, Juncai He, Xiangbo Wu, Fei Yu, Junying Chen, MA Zhuoheng, Yuhao Du, Yan Hu, He Zhang, Emad A. Alghamdi, Lian Zhang, Ruoyu Sun, Haizhou Li, Jinchao Xu, Benyou Wang. **Second Language (Arabic) Acquisition of LLMs via Progressive Vocabulary Expansion**. Submitted to COLM 2024.
[3] Juhao Liang, Zhenyang Cai, Jianqing Zhu, Huang Huang, Kewei Zong, Bang An, Mosen Alharthi, Juncai He, Lian Zhang, Haizhou Li**,** Benyou Wang, Jinchao Xu. **Alignment at Pre-training! Towards Native Alignment for Arabic LLMs**. Submitted to NeurIPS 2024

FINANCIAL TIMES

Artificial intelligence + Add to myFT

Saudi-China collaboration raises concerns about access to AI chips

Fears grow at Gulf kingdom's top university that ties to Chinese researchers risk upsetting US government



Western officials have long expressed concerns about growing technology transfer between their traditional allies in the Gulf and China © FT montage/Bloomberg/Dreamstime

Simeon Kerr and **Samer Al-Atrush** in Dubai, **Qianer Liu** in Hong Kong, **Madhumita Murgia** in London 13 HOURS AGO

52

Stay informed with free updates

Simply sign up to the Artificial intelligence myFT Digest -- delivered directly to your inbox.

Enter your email address

Sign up

Saudi-Chinese collaboration in artificial intelligence has stirred fears within the Gulf kingdom's premier academic institution that the ties could jeopardise the university's access to US-made chips needed to power the new technology.

Professor Jinchao Xu, an American-Chinese mathematician at Saudi Arabia's King Abdullah University of Science and Technology (Kaust), has launched AceGPT, an Arabic-focused large language model, in collaboration with the Chinese University of Hong Kong, Shenzhen (CUHK-SZ), and the Shenzhen Research Institute of Big Data.

英媒给中沙AI合作泼冷水，中国专家：美西方不断干扰毫无道理

来源：环球时报 作者：黄培昭 赵觉理

-2023-
10/11
07:23

【环球时报驻埃及特派记者 黄培昭 环球时报记者 赵觉理】人工智能（AI）技术正成为中国与中东国家合作的新亮点，但这种互利共赢却遭到美西方阻挠。英国《金融时报》10日刊文，给中国和沙特的相关合作泼冷水。对此，有中国专家表示，中国和中东国家的科技合作基于双方在该领域的互补性，符合双方共同利益，美西方不断干扰毫无道理。

15

《金融时报》报道称，沙特阿卜杜拉国王科技大学、香港中文大学（深圳）与深圳大数据研究院三方合作开发人工智能大语言模型ACEGPT。"此举是沙特领导人工智能技术区域发展、建造大型超算和推出大语言模型努力的一部分。"这家英媒还称，沙特正与阿联酋一道，寻求参与到人工智能竞争之中。

但该媒体笔锋一转称，中国与海湾国家的此类合作让西方感到担忧，美国对中国实施的人工智能芯片出口限制也正影响相关合作。有阿卜杜拉国王科技大学的工作人员担忧，与中国的合作可能会引发美国不满，从而影响该大学获得先进人工智能芯片。

早在今年8月，路透社就曾报道称，美国芯片制造商英伟达和AMD均已收到美国政府限制向部分中东国家出口先进人工智能芯片的要求。有分析人士认为，美国的主要目的是防止中国从中东国家手中购买先进芯片。"德国之声"援引专家的分析称，沙特、阿联酋等大力投资人工智能的国家近年来加深了与中国的联系，因此它们都

环球时事
他们搭起一座桥，让世界阅读中国
意大利设"耐安妇"铜像，日方"强烈关切"
避免贸易摩擦扩大升级！中欧启动汽车关税磋商
新加坡媒体：金砖国家"集体潜力"吸引东南亚
我国首个！成立了！今年招收300人
这条隧道通了！时速350公里高铁取得新进展

环球业界

昔日黄沙独绿颜
珍稀植物丽豆全基因组首次揭秘
节能减碳 新疆油气企业发展"含绿量"足
3D打印真空系统成船"捕捉"暗物质
证监会支持上海加快"五个中心"建设
依托智慧农业设施 高标准农田展现抗旱优势

Work 3： Multi-modal LLMs

- Dataset ALLaVA
- Milebench
- MotionLLM
- MLLM-bench
- Silkie

[1] Guiming Hardy Chen, Shunian Chen, Ruifei Zhang, Junying Chen, Xiangbo Wu, Zhiyi Zhang, Zhihong Chen, Jianquan Li, Xiang Wan, **Benyou Wang**. ALLaVA: Harnessing GPT4V-synthesized Data for A Lite Vision-Language Model. https://arxiv.org/abs/2402.11684
[2] D Song, S Chen, GH Chen, F Yu, X Wan**, B Wang**. **Milebench**: Benchmarking mllms in long context, arXiv preprint arXiv:2404.18532
[3] Ling-Hao Chen, Shunlin Lu, Ailing Zeng, Hao Zhang, Benyou Wang, Ruimao Zhang, Lei Zhang. MotionLLM: Understanding Human Behaviors from Human Motions and Videos. https://arxiv.org/abs/2405.20340
[4] Wentao Ge, Shunian Chen, Guiming Chen, Junying Chen, Zhihong Chen, Shuo Yan, Chenghao Zhu, Ziyue Lin, Wenya Xie, Xidong Wang, Anningzhe Gao, Zhiyi Zhang, Jianquan Li, Xiang Wan, Benyou Wang. Mllm-bench, evaluating multi-modal llms using gpt-4v. https://arxiv.org/abs/2311.13951
[5] Lei Li, Zhihui Xie, Mukai Li, Shunian Chen, Peiyi Wang, Liang Chen, Yazheng Yang, Benyou Wang, Lingpeng Kong. Silkie: Preference distillation for large visual language models. https://arxiv.org/abs/2312.10665

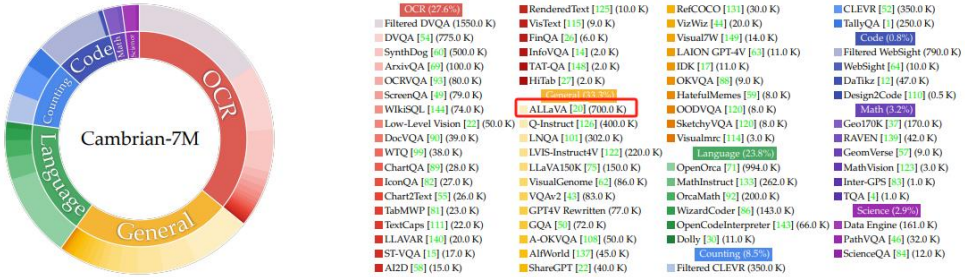


Figure 9 | **Cambrian-7M: A Large-Scale Curated Instruction Tuning Dataset for MLLM. Left:** The inner circle shows the original distribution of Cambrian-10M. The outer circle shows the curated Cambrian-7M. **Right:** All the data sources in the Cambrian dataset as well as the ones filtered in data curation.

Cambrian-1: A Fully Open, _Vision-Centric_ Exploration of Multimodal LLMs

Shengbang Tong*, Ellis Brown*, Penghao Wu*, Sanghyun Woo, Manoj Middepogu, Sai Charitha Akula, Jihan Yang, Shusheng Yang, Adithya Iyer, Xichen Pan, Austin Wang, Rob Fergus, Yann LeCun, Saining Xie†

New York University

Cambrian-1: A Fully Open, Vision-Centric Exploration of Multimodal LLMs

Guiming Hardy Chen, Shunian Chen, Ruifei Zhang, Junying Chen, Xiangbo Wu, Zhiyi Zhang, Zhihong Chen, Jianquan Li, Xiang Wan, **Benyou Wang**. ALLaVA: Harnessing GPT4V-synthesized Data for A Lite Vision-Language Model. https://arxiv.org/abs/2402.11684

Work 4：LLM for Math and Optimization

- **Fei Yu**, Anningzhe Gao, **Benyou Wang**. OVM, Outcome-supervised Value Models for Planning in Mathematical Reasoning. https://arxiv.org/abs/2311.13951. Findings of NAACL 2024
- **Zhengyang Tang**, Xingxing Zhang, **Benyou Wang**, Furu Wei. MathScale: Scaling Instruction Tuning for Mathematical Reasoning, **ICML 2024.**
- Zhengyang Tang, Chenyu Huang, Xin Zheng, Shixi Hu, Zizhuo Wang, Dongdong Ge, **Benyou Wang**. ORLM: Training Large Language Models for Optimization Modeling. https://arxiv.org/abs/2405.17743
- Xuhan Huang, Qingning Shen, Yan Hu, Anningzhe Gao, **Benyou Wang**. Mamo: a Mathematical Modeling Benchmark with Solvers. https://arxiv.org/abs/2405.13144v1