

ChatGLM Agent: 大模型智能体实践与探索

东昱晓

知识工程实验室 (KEG)

清华大学计算机系



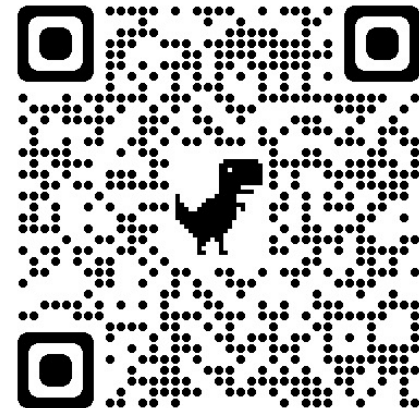
Joint Work with



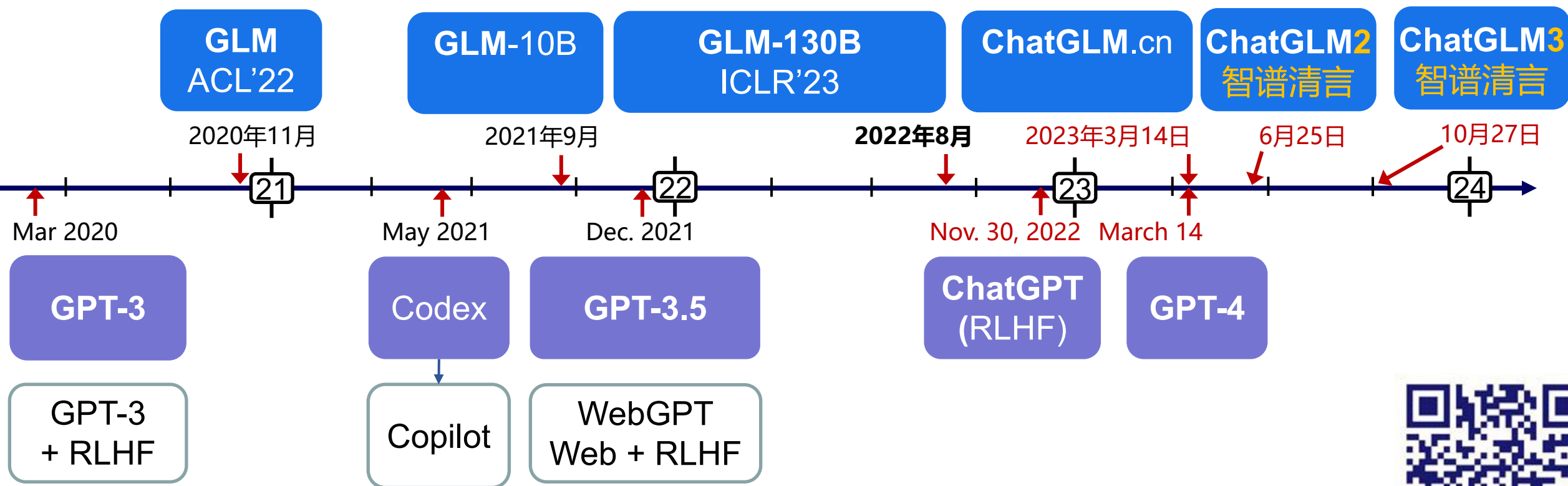
Jie Tang



and many others at **Tsinghua** and **Zhipu.AI**



清华&智谱 GLM 系列模型



OpenAI GPT系列模型



千亿基座模型预训练



OpenAI

- GPT-3 175B

[2020]



- Megatron-Turing-530B

[2021]



- LaMDA 137B
- PaLM 540B

[2021~2022]



- Chinchilla 70B
- Gopher 260B

[2021~2022]



清华大学
Tsinghua University



智谱·AI

- GLM-130B

[2022]



- OPT 175B

[2022]

BigScience

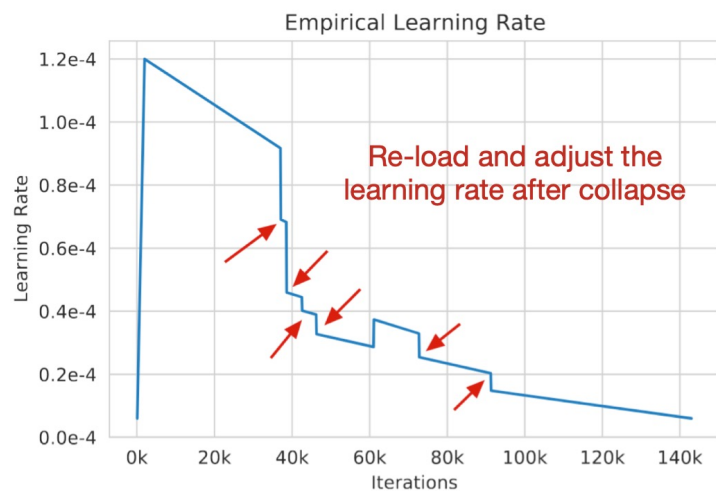


- BLOOM 176B

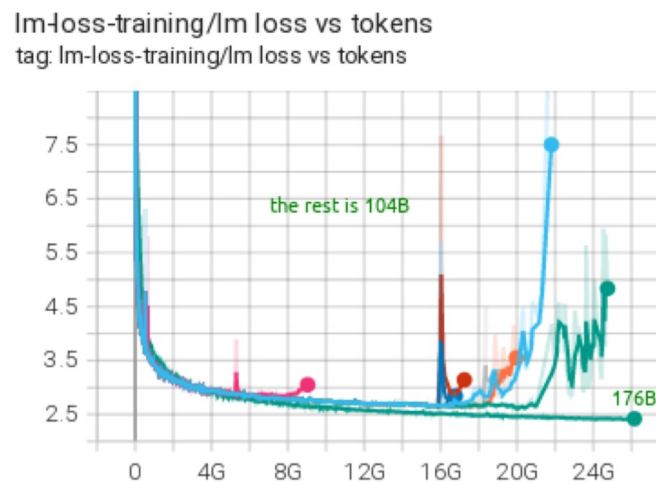
[2022]

千亿模型训练最大挑战：训练稳定性

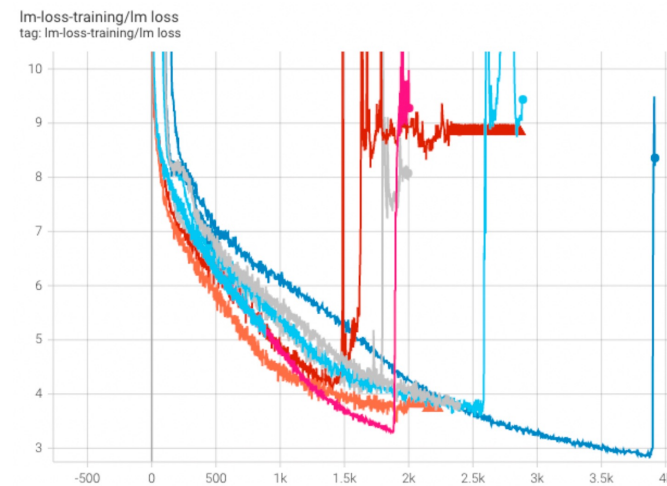
- 权衡利弊：训练稳定性（高精度低效）还是训练效率（低精度高效）
- 目前已开源训练过程大模型的解决方案
 - **FB OPT-175B**: 训练崩溃时反复调整学习率/跳过数据（权宜之计，损失性能）
 - **HF BLOOM 176B**: embedding norm和BF16（损失性能，有限适配平台）



(a) OPT 175B's experiments



(b) BLOOM 176B's experiments



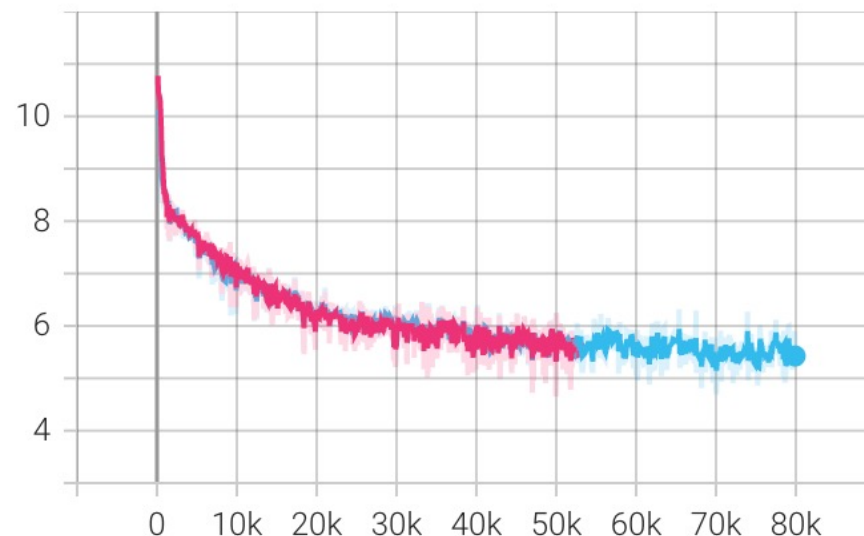
(c) GLM 130B's experiments

跨平台高效训练千亿模型

- 跨平台兼容：swDeepSpeed 训练库 \leftrightarrow 与 DeepSpeed API 兼容
 - 支持神威超算**申威**架构，一行代码无缝替换兼容
 - 实现并行通信策略，混合精度策略，ZeRO 优化器
 - 同一套训练框架可在**不同架构（英伟达、申威、海光DCU、昇腾）**集群上对齐训练曲线

```
import swDeepSpeed as deepspeed
model, optimizer, _, _ = deepspeed.initialize(
    model=model,
    model_parameters=param_groups,
    args=args,
    mpu=mpu,
    dist_init_required=False,
    config_params=config_params
)
```

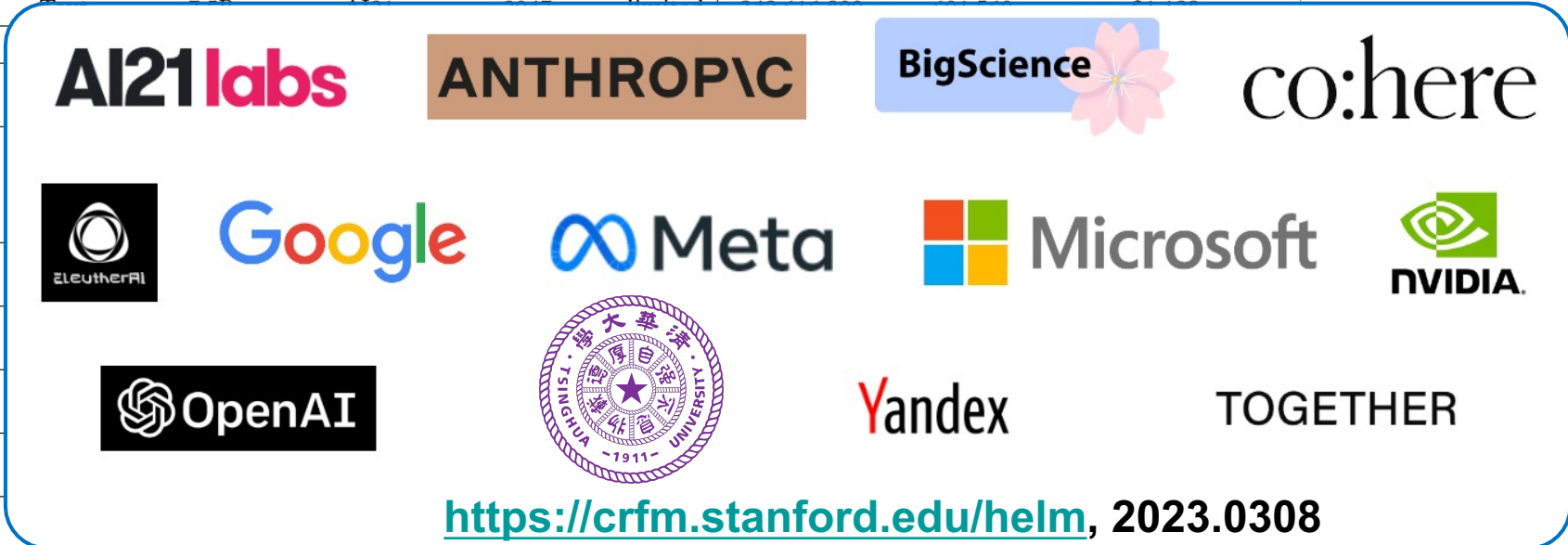
Train/train_loss
tag: Train/train_loss



GLM-130B in HELM

Stanford's Holistic Evaluation of Language Models (HELM, Nov. 2022)

Model	Model Creator	Modality	# Parameters	Tokenizer	Window Size	Access	Total Tokens	Total Queries	Total Cost
J1-Jumbo v1 (178B)	AI21 Labs	Text	178B	AI21	2047	limited	327,443,515	591,384	\$10,926
J1-Grande v1 (17B)	AI21 Labs	Text	17B	AI21	2047	limited	326,815,150	591,384	\$2,973
J1-Large v1 (7.5B)	AI21 Labs								
Anthropic-LM v4-s3 (52B)	Anthropic								
BLOOM (176B)	BigScience								
T0++ (11B)	BigScience								
Cohere xlarge v20220609 (52.4B)	Cohere								
Cohere large v20220720 (13.1B) ⁵⁸	Cohere								
Cohere medium v20220720 (6.1B)	Cohere								
Cohere small v20220720 (410M) ⁵⁹	Cohere								
GPT-J (6B)	EleutherAI								
GPT-NeoX (20B)	EleutherAI								
T5 (11B)	Google								
UL2 (20B)	Google								
OPT (66B)	Meta								
OPT (175B)	Meta								
TNLG v2 (6.7B)	Microsoft/NVIDIA								
TNLG v2 (530B)	Microsoft/NVIDIA								
GPT-3 davinci v1 (175B)	OpenAI								
GPT-3 curie v1 (6.7B)	OpenAI								
GPT-3 babbage v1 (1.3B)	OpenAI	Text	1.3B	GPT-2	2048	limited	422,123,900	606,253	\$211
GPT-3 ada v1 (350M)	OpenAI	Text	350M	GPT-2	2048	limited	422,635,705	604,253	\$169
InstructGPT davinci v2 (175B*)	OpenAI	Text	175B*	GPT-2	4000	limited	466,872,228	599,815	\$9,337
InstructGPT curie v1 (6.7B*)	OpenAI	Text	6.7B*	GPT-2	2048	limited	420,004,477	606,253	\$840
InstructGPT babbage v1 (1.3B*)	OpenAI	Text	1.3B*	GPT-2	2048	limited	419,036,038	604,253	\$210
InstructGPT ada v1 (350M*)	OpenAI	Text	350M*	GPT-2	2048	limited	418,915,281	604,253	\$168
Codex davinci v2	OpenAI	Code	Unknown	GPT-2	4000	limited	46,272,590	57,051	\$925
Codex cushman v1	OpenAI	Code	Unknown	GPT-2	2048	limited	42,659,399	59,751	\$85
GLM (130B)	Tsinghua University	Text	130B	ICE	2048	open	375,474,243	406,072	2,100 GPU hours
YaLM (100B)	Yandex	Text	100B	Yandex	2048	open	378,607,292	405,093	2,200 GPU hours

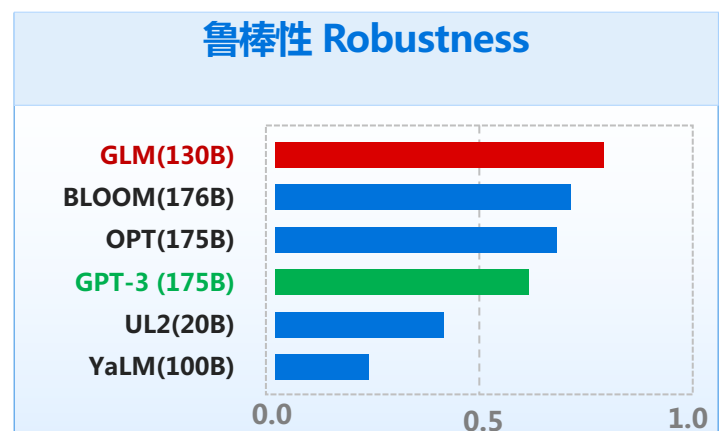
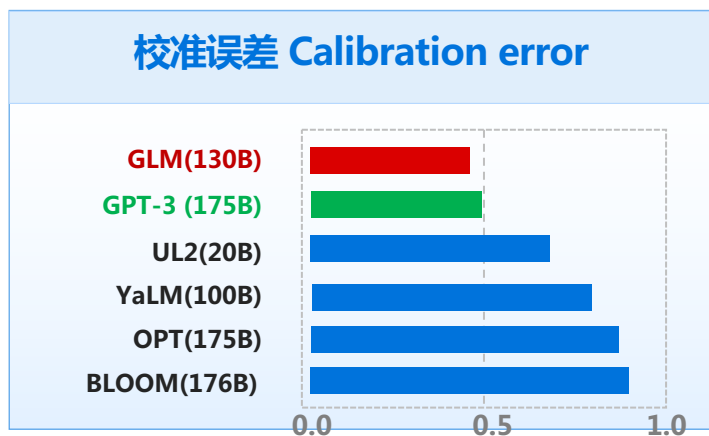
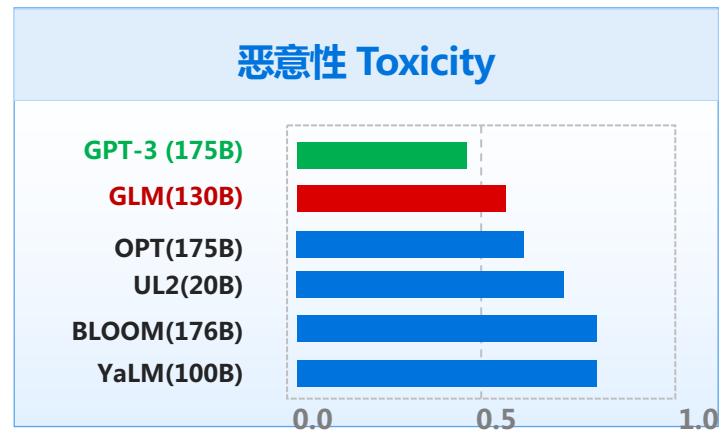
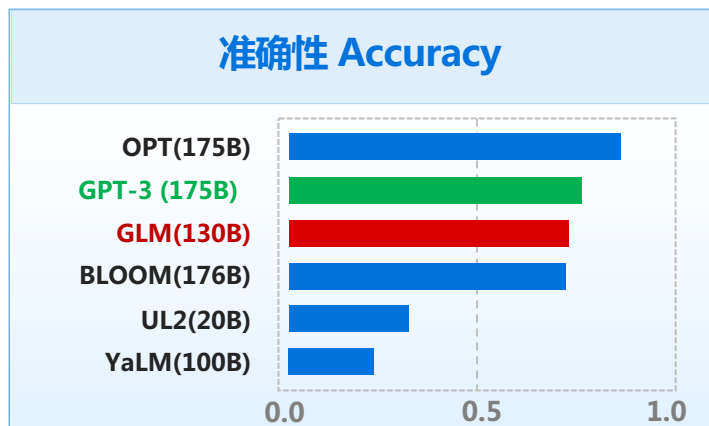


跨平台高效训练千亿模型

开源

斯坦福大学的世界主流大模型评测，**GLM-130B** 亚洲唯一入选模型：

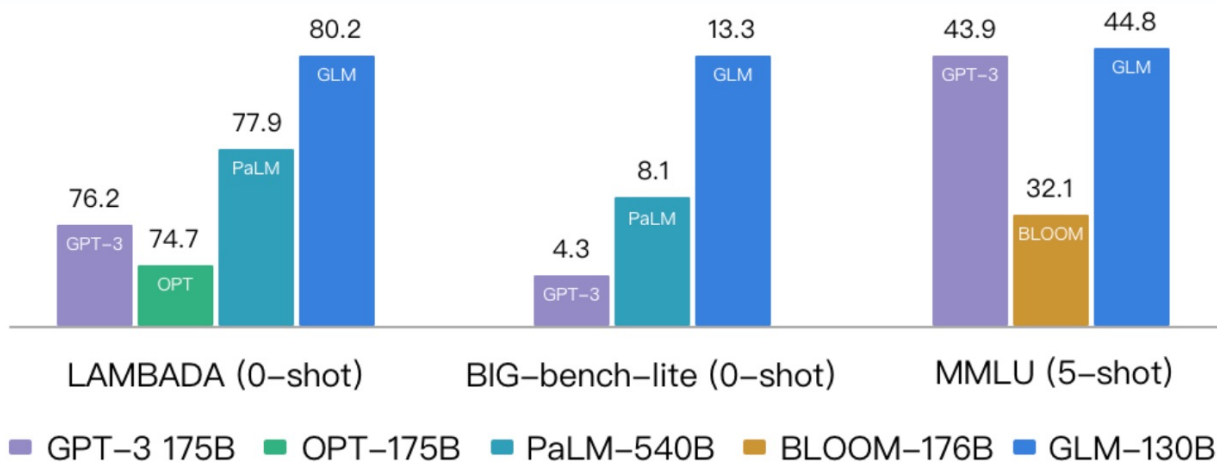
- 准确性、恶意性与**GPT-3 (davinci)** 接近；
- 鲁棒性和校准误差在**所有千亿模型中表现最佳**；



千亿基座GLM-130B

开源

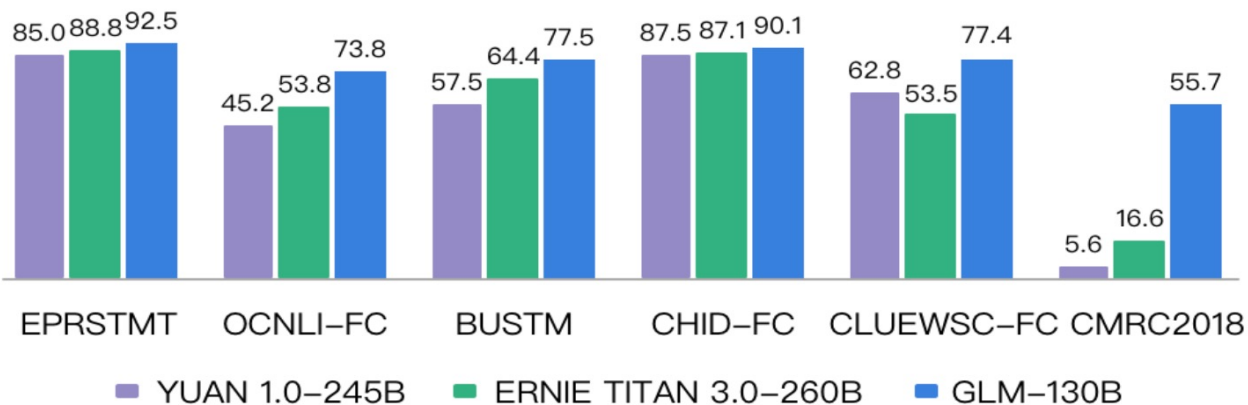
英文：在 MMLU、LAMBADA、BIG-bench-lite 等
超过GPT-3, OPT, PaLM



2022.8---2023.4，收到70个国家
1000+研究机构的下载使用需求

- Google
- Microsoft
- Facebook
- AI2
- Stanford
- MIT
- UC Berkely
- CMU
- Harvard
- Princeton
- Yale
- Cornell
- Columbia
- UIUC
- Cambridge
- Oxford
- 华为
- 阿里巴巴
- 腾讯
- 百度
- 美团
- 头条
- 滴滴
- 智源
- 小冰
- 小度
- 小米
- 小鹏
- 有道
- 旷视
- 平安
- 建设银行
- 北京大学
- 浙江大学
- 上海交大
- 复旦大学
- 中科院大学
- 中科大
- 武汉大学
- 华科
- 南开
- 香港大学
- 香港中文大学
- 香港科技大学
- 中科院多所
- 之江实验室
- 上海 AI 实验室
- 北京智源

中文：在 CLUE 和 FewCLUE 上取得不错的结果



清华&智谱 GLM 系列模型

Models announced today:

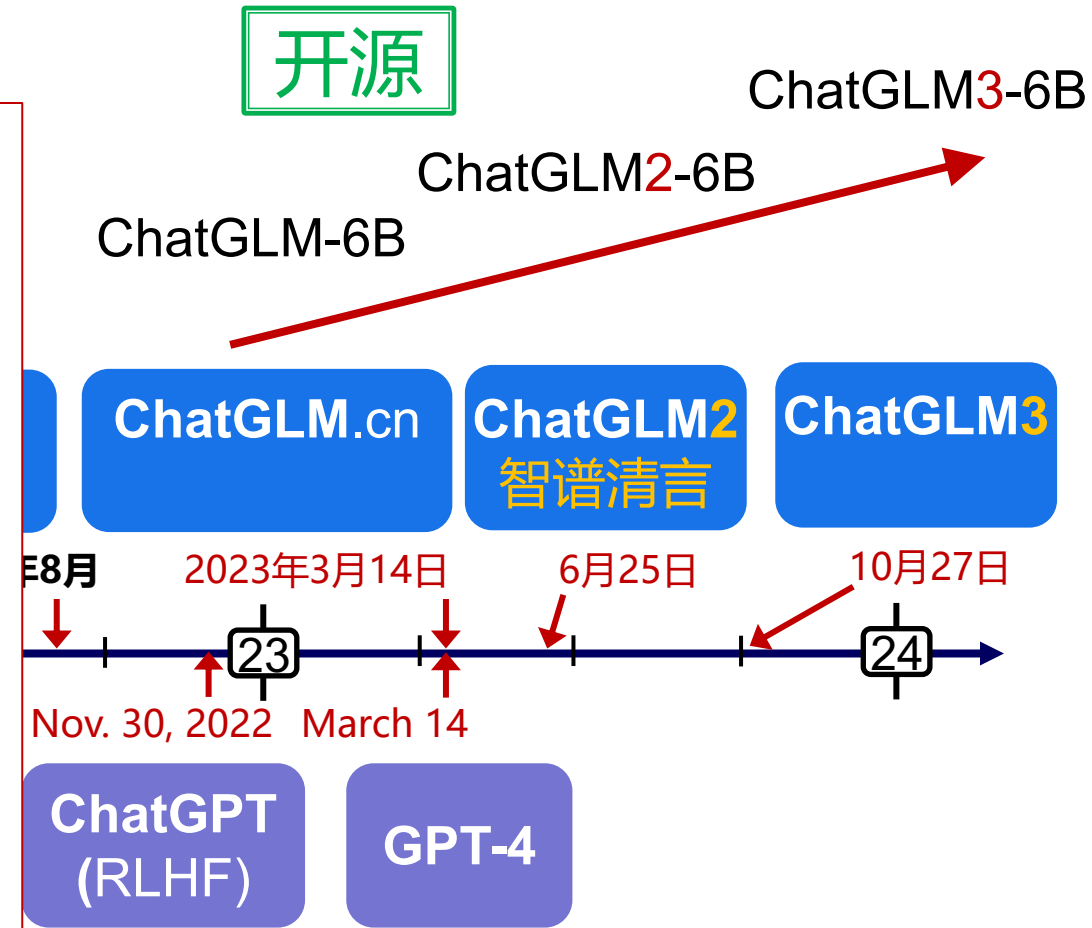
- @OpenAI GPT4
- @AnthropicAI Claude
- @thukeyg ChatGLM
- @GoogleAI Med-PaLM 2

+ RLHF

Supervised

Web + RLHF

开源



<https://github.com/THUDM>

ChatGLM开源

基于ChatGLM开发的
开源应用 > 600+

- ChatGLM-6B开源：**62亿**参数、**INT4**量化（仅需**6G**内存推理）
- 开源的训练数据量**3万亿**字符的模型
- ChatGLM-6B开源，GitHub **5万 star**
- Github Trending **榜第一**
- Huggingface 下载超过 **1000万**
- Hugging Face Trending **榜第一 (4周)**



技术
博客



<https://huggingface.co/THUDM>

Spaces using THUDM/chatglm-6b 203

Spaces using THUDM/chatglm2-6b-int4 129

Spaces using THUDM/chatglm-6b-int4 62

Spaces using THUDM/chatglm2-6b 196

The screenshot displays a list of Hugging Face Spaces. Each space entry includes a user profile icon, the space name, and a star count. The spaces listed include:

- HuggingFaceH4/open_llm_leaderboard
- upstage/open-ko-llm-leaderboard
- yentinglin/Taiwan-LLaMa2
- kevinwang676/ChatGLM2-VC-SadTalker
- glt3953/app-text_generation_chatglm2-6b
- felixz/open_llm_leaderboard
- okliqiang/course
- behindeu/gpt-academic
- hongdaaaaaaaa/gpt-academic
- zfwcpc/ChatGLM2-SadTalker-VC
- Colinmei/gpt-academic
- qingxu98/gpt-academic
- mikeee/chatglm2-6b-4
- gsaivinay/open_llm_leaderboard
- eson/tokenizer-a
- qgyd2021/chat_with_llm
- mcphur/gpt-academic
- tradequant/chat
- hiwei/chatglm2-6b-explor
- Jafta/chatglm2-6b-4bit

At the bottom right, it indicates "+ 176 Spaces".

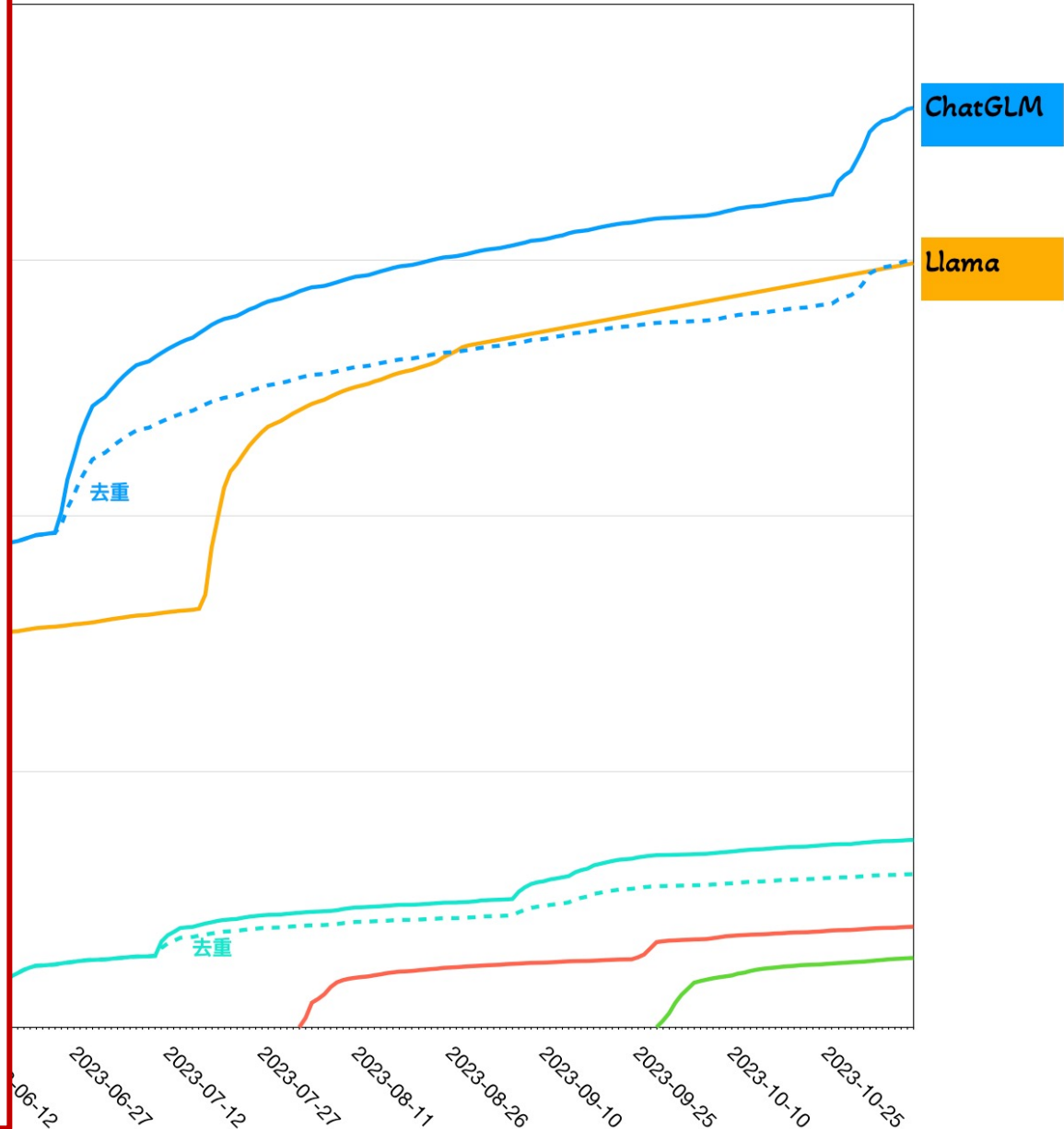
ChatGLM vs. LLaMA

The top 15 most-liked organizations on [@huggingface](#)

1. @StabilityAI 20k likes
2. @AlatMeta 20k
3. @runwayml 11k
4. CompVis 10k
5. @thukeyg 7k
6. @BigscienceW 7k
7. @TIIuae 7k
8. @Microsoft 6.5k
9. @GoogleAI 6k
10. @OpenAI 4k
11. @BigCodeProject 4k
12. @MosaicML 4k
13. @UKPLab 3k
14. @AiEleuther 3k
15. @salesforce 3k

[huggingface.co/spaces/PulsarA...](#)

5:40 PM · Nov 22, 2023 · 6,094 Views



<https://github.com/THUDM>

清华&智谱 GLM 系列模型

WebGPT | New Bing

Codex | Copilot

ChatGPT | GPT4-Turbo

搜索增强大模型

代码大模型

智能体大模型

WebGLM
KDD

CodeGeeX
KDD

AgentTuning
AgentBench

GLM
ACL

Nov. 2020

21

GLM-10B
Open

Sep. 2021

GLM-130B
ICLR

Aug. 2022

22

GLM-130B
→ ChatGLM

Nov. 30

23

March 14

March 14

ChatGPT GPT-4

ChatGLM2

June 25

ChatGLM3

Oct 27

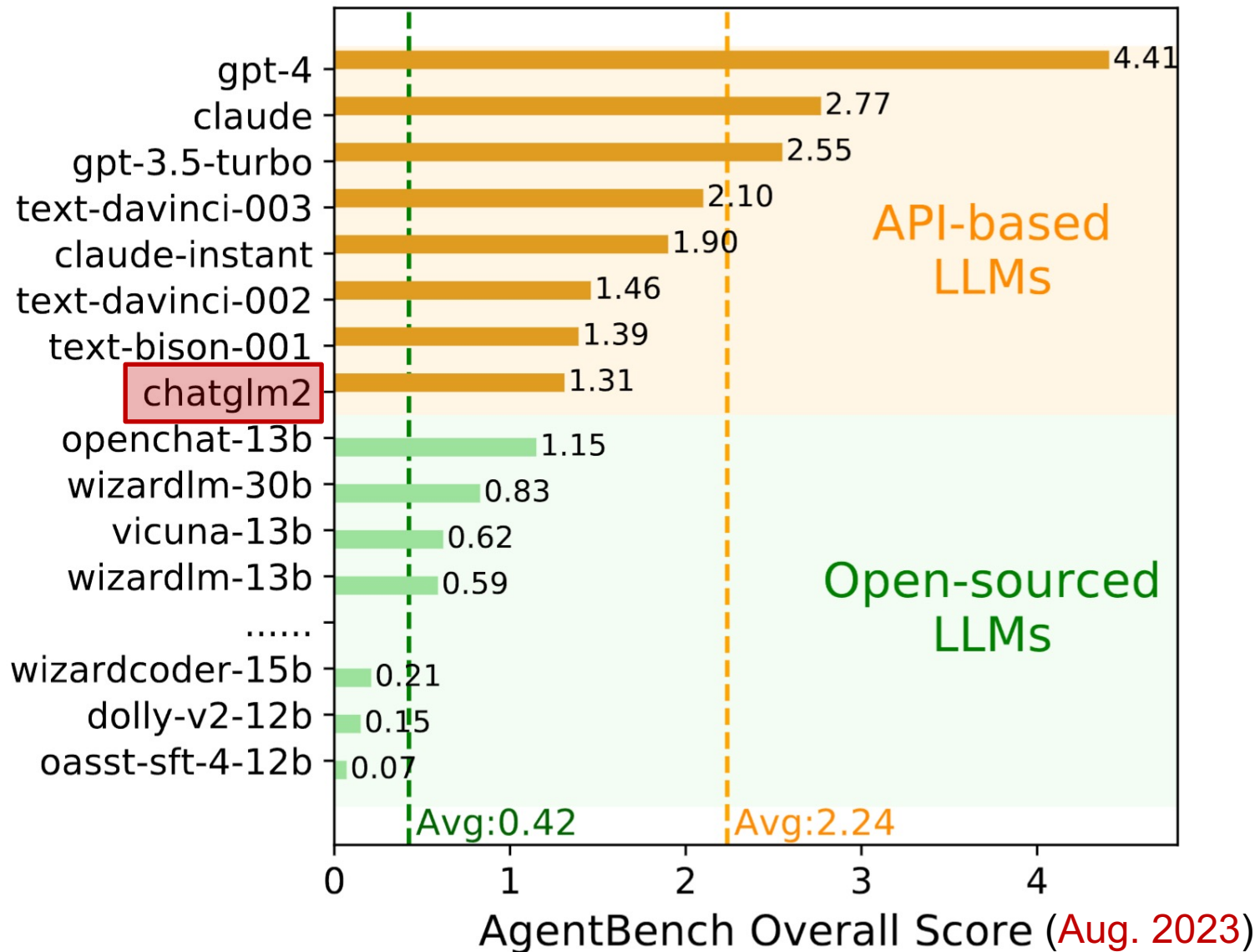
24

AgentBench: Evaluate LLMs as Agents lmbench.ai

- 多项Agent任务

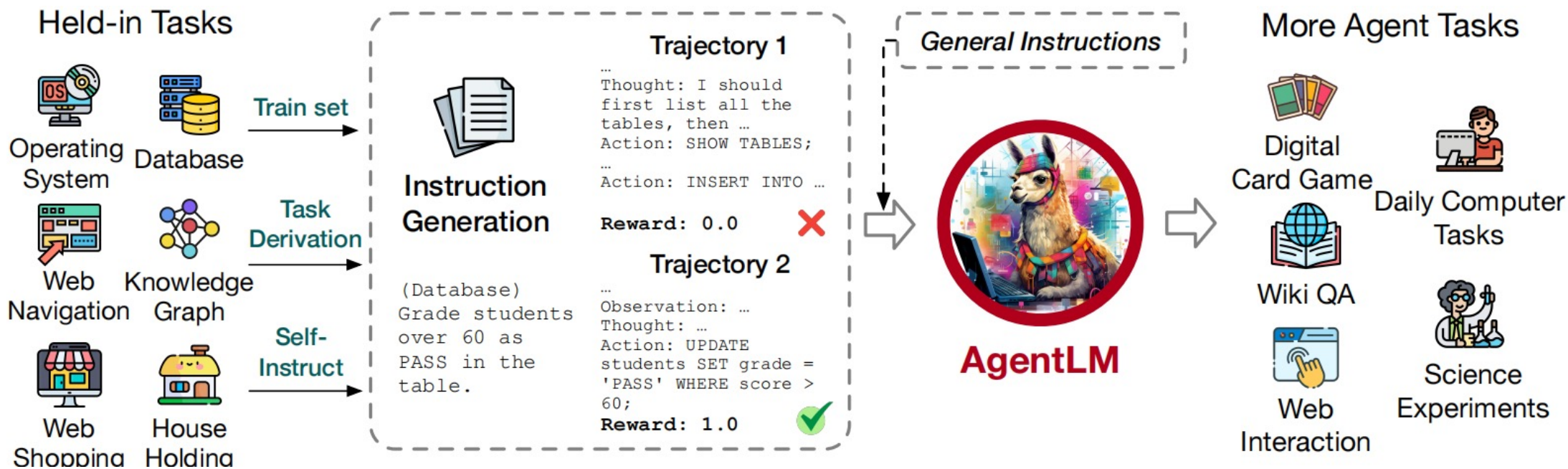
1. Operating System
2. Database
3. Knowledge Graph
4. Digital Card Game
5. Lateral Thinking Puzzles
6. House-holding
7. Web Shopping
8. Web Browsing

- 仅单一Agent任务
无法泛化到多任务



AgentTuning: 少样本激活Agent泛化能力

开源



AgentInstruct 智能体轨迹数据集

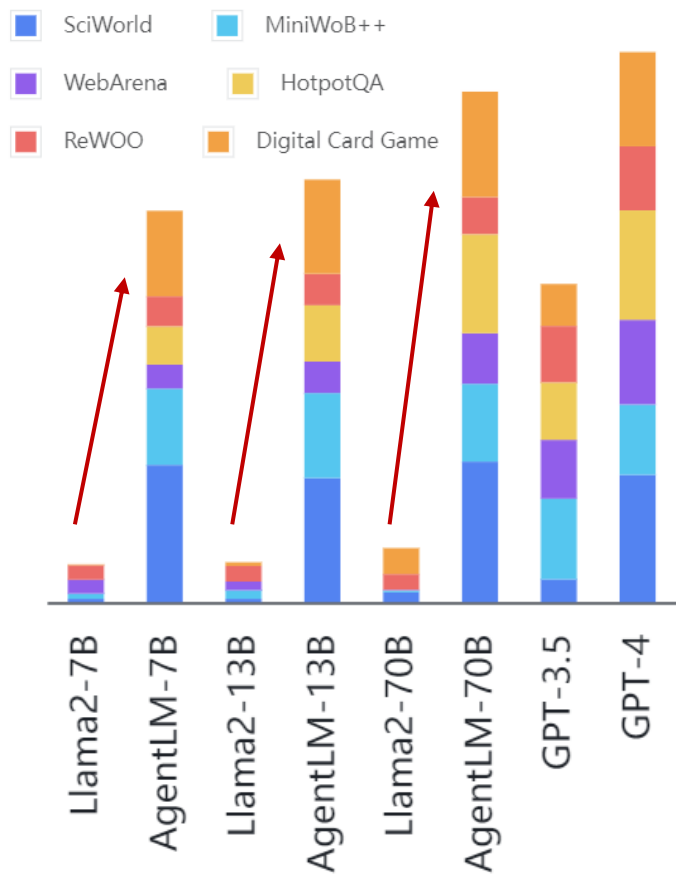
- 六项智能体任务
- 使用 Reward 过滤数据
- 仅保留高质量的1866条交互轨迹

Agent Tuning 混合训练

- 20% AgentInstruct + 80% ShareGPT
- 混合训练后于外分布任务展现泛化性
- 得到 AgentLM-{7B,13B,70B} 系列模型

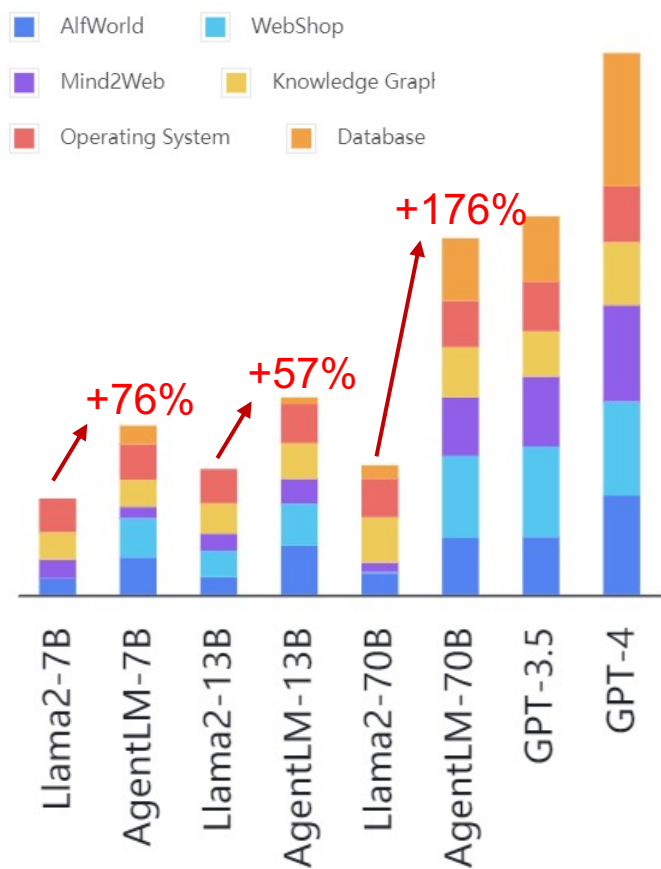
AgentTuning: 少样本激活Agent泛化能力

开源



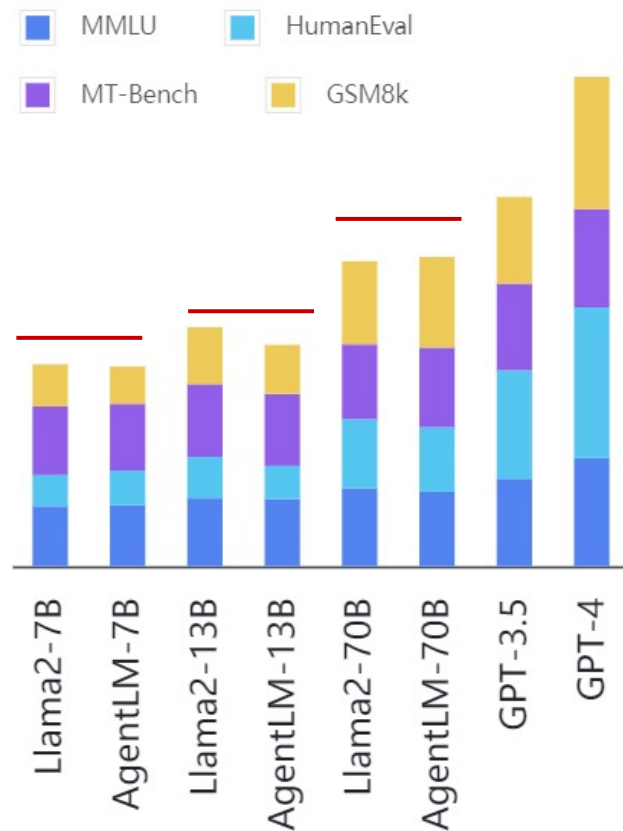
同分布任务

显著提升, 比肩 GPT-4



外分布任务

泛化性良好, 接近 GPT-3.5

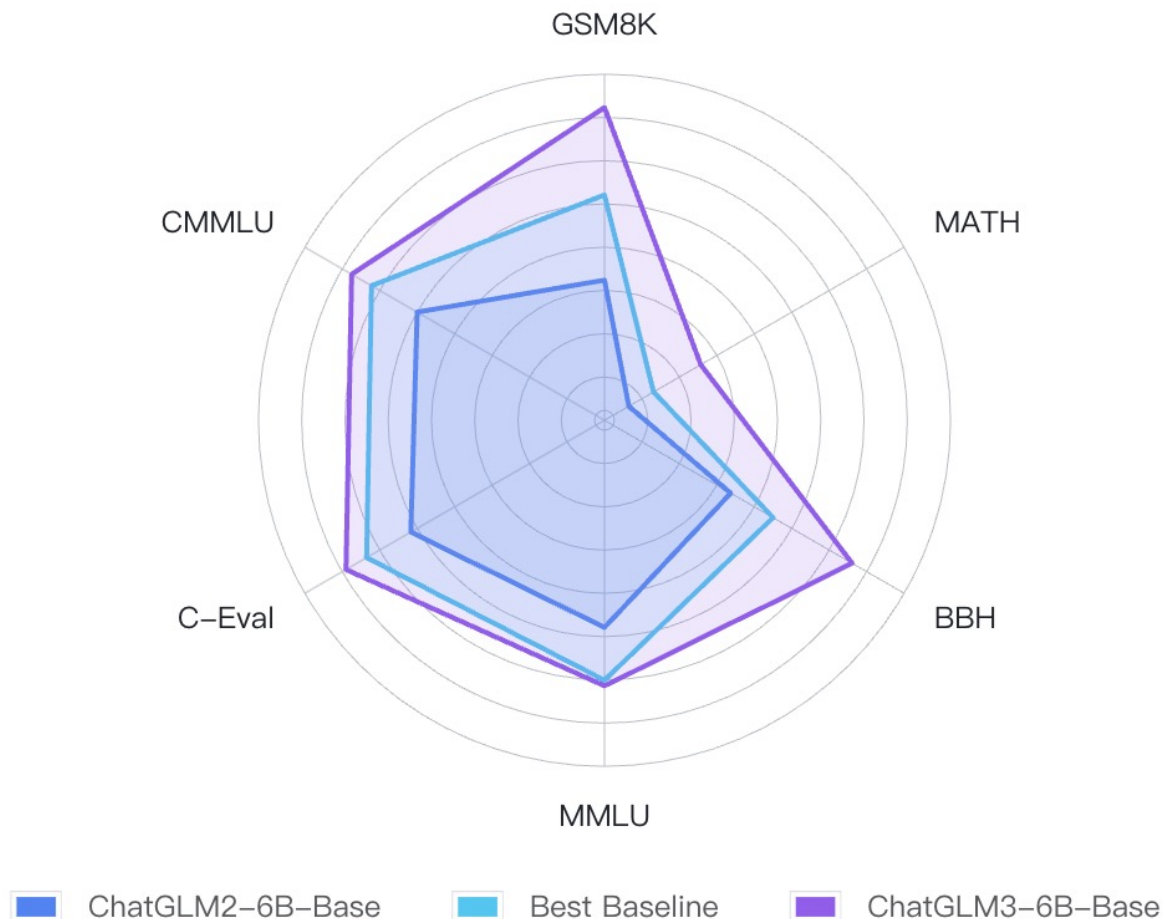


通用能力

与微调前表现相当

ChatGLM3: 更强性能 & Agent 能力初探

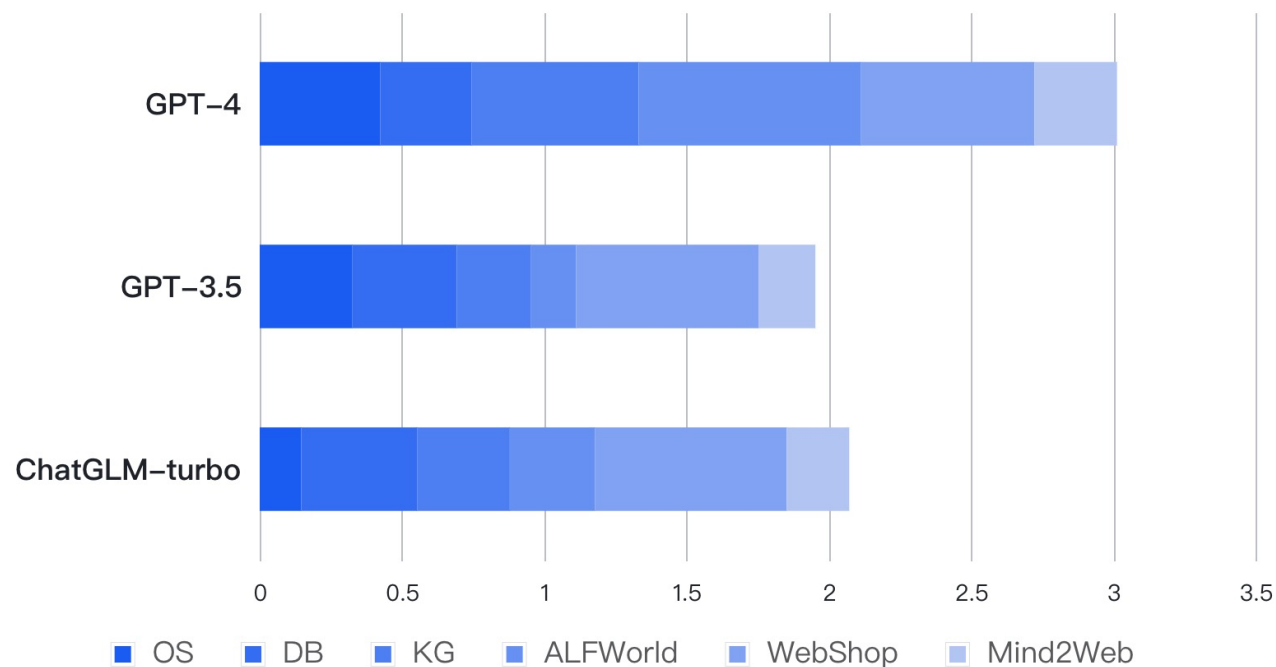
From ChatGLM2 To ChatGLM3



- 训练数据更多样、训练步数更充分、训练策略更合理
- 在语义、数学、推理、代码、知识等不同角度的数据集上表现出色
- 44个中英文公开数据集测试国内第一

ChatGLM3: 更强性能 & Agent 能力初探

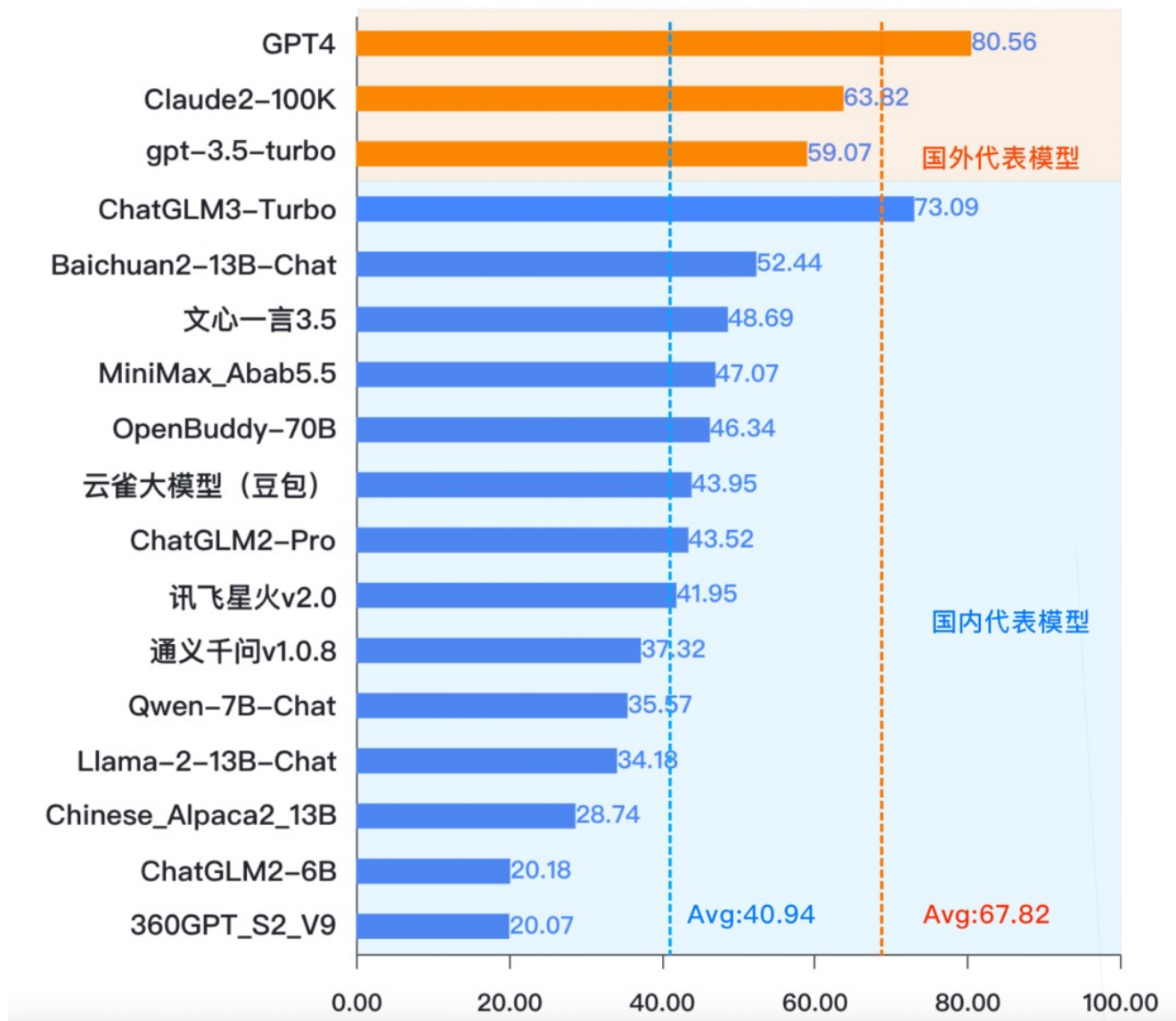
Agent 能力



- 全新设计的 Prompt 格式, 在不影响模型通用能力的情况下, 全方面增强 ChatGLM3-6B 能力
- 原生支持工具调用 (Function Call)、代码执行 (Code Interpreter) 以及 Agent 任务

ChatGLM3: 更强性能 & Agent 能力初探 (第三方评测)

SuperCLUE-Agent基准得分



ChatGLM3: 更强性能 & Agent 能力初探

开源

ChatGLM3 Demo

Chat Tool Code Interpreter

Manual mode

欧洲有哪些好玩的地方?

欧洲是一个非常有趣的地方，有许多好玩的地方。例如，可以前往法国的巴黎，欣赏埃菲尔铁塔和卢浮宫等著名景点；可以前往意大利的罗马，游览古罗马斗兽场和梵蒂冈等著名景点；可以前往德国的柏林，参观勃兰登堡门和柏林墙等著名景点。当然，欧洲还有许多其他好玩的地方，这只是其中的一部分。

查查巴黎的天气怎么样?

我需要调用天气预报工具来获取巴黎的天气信息。

Calling tool `get_weather`:

```
tool_call(city_name='巴黎')
```

Observation:

```
{'current_condition': {'temp_C': '12', 'FeelsLikeC': '10', 'humidity': '88'}}
```

根据天气预报工具返回的信息，巴黎现在的天气情况是：温度为12摄氏度，湿度为88%，天气状况为雨天。这些信息可以帮助您更好地了解巴黎的天气情况。

Chat with ChatGLM3!

<http://github.com/THUDM/ChatGLM3>

ChatGLM 3 Demo

Chat Tool Code Interpreter

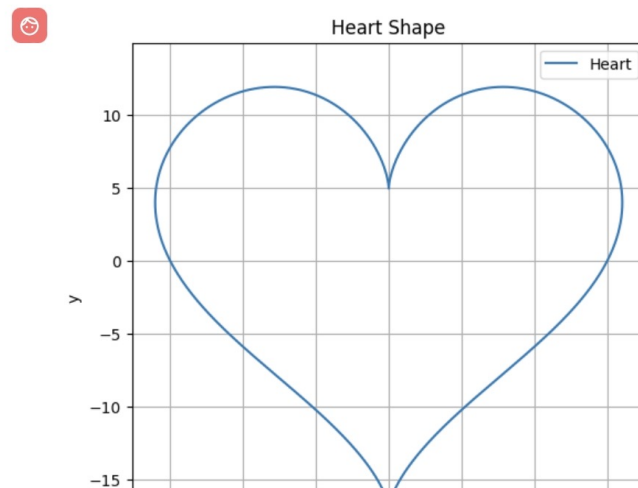
用 Python 画一个爱心

```
import matplotlib.pyplot as plt
import numpy as np

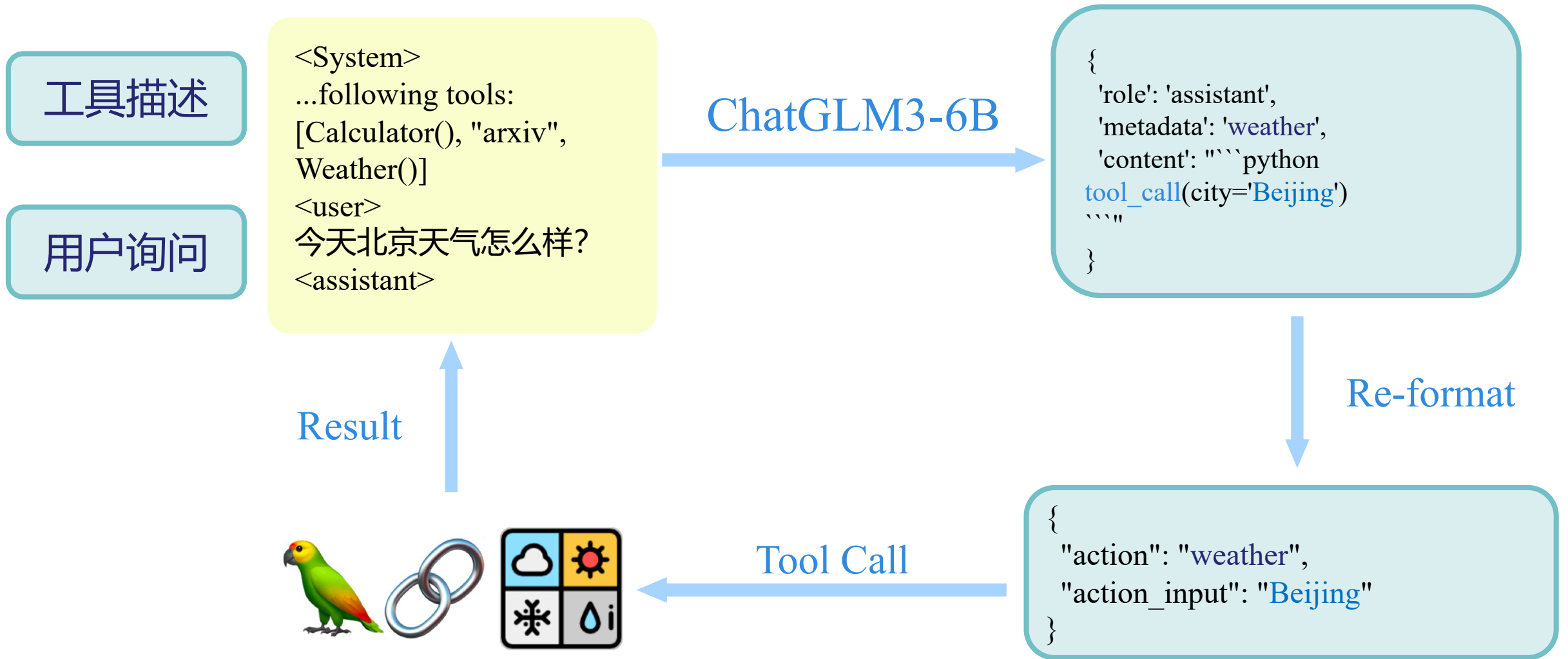
# 参数方程
t = np.linspace(0, 2 * np.pi, 1000)
x = 16 * np.sin(t)**3
y = 13 * np.cos(t) - 5 * np.cos(2*t) - 2 * np.cos(3*t) - np.cos(4*t)

plt.figure(figsize=(6,6))
plt.plot(x, y, label='Heart')
plt.title('Heart Shape')
plt.xlabel('x')
plt.ylabel('y')
plt.legend()
plt.grid(True)
plt.axis('equal')
plt.show()
```

Executing...



ChatGLM3: LangChain-Agent 接入



清华&智谱 GLM 系列模型

WebGPT | New Bing

搜索增强大模型

WebGLM
KDD

Codex | Copilot

代码大模型

CodeGeeX
KDD

ChatGPT | GPT4-Turbo

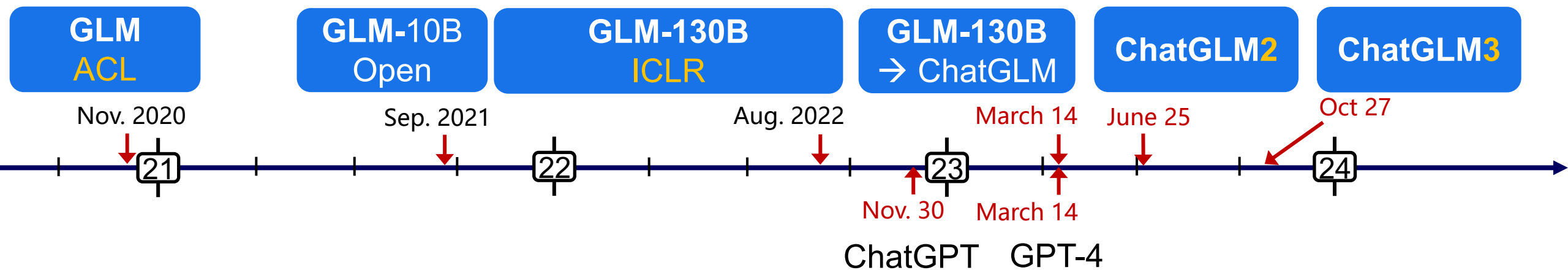
智能体大模型

AgentTuning
AgentBench

Dall-E | GPT4-V

多模态大模型

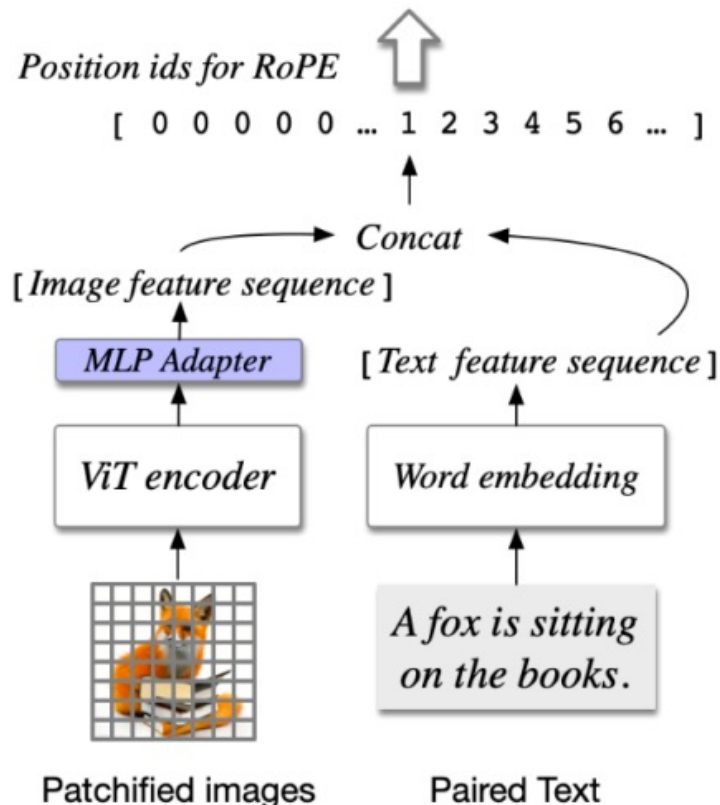
CogVLM
CogAgent



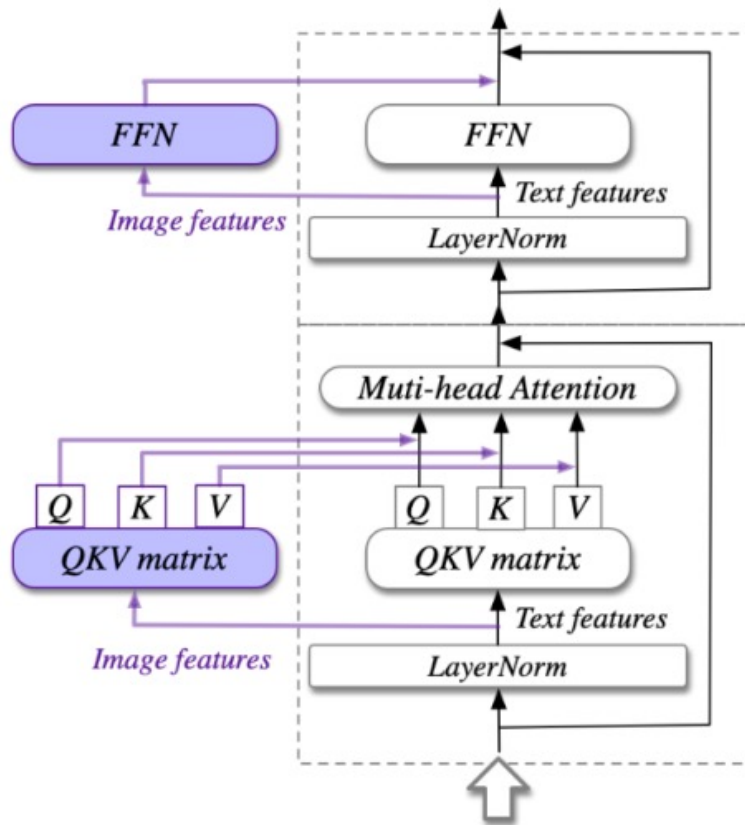
CogVLM: Vision Language Models

开源

提出基于认知的多模态理解算法，实现类似**GPT-4V**的初步版本
(billions of image-text pairs)



(a) The input of visual language model



(b) The visual expert built on the language model



Q1: What is funny thing in this image?
翻译: 图中有趣的点是什么?

CogVLM: The funny thing in this image is that a mobile phone is being charged with an unconventional USB interface, which looks like the top of a VGA cable. This design is quite unusual and might be considered humorous because it combines the functionality of a charging cable with the appearance of an old video game accessory.
翻译: 这幅图片中有趣的是, 一部手机正在通过一个非常规的USB接口充电, 这个接口看起来像VGA线的顶部。这种设计相当不寻常, 可能会被认为是幽默的, 因为它将充电线的功能与旧视频游戏配件的外观相结合。

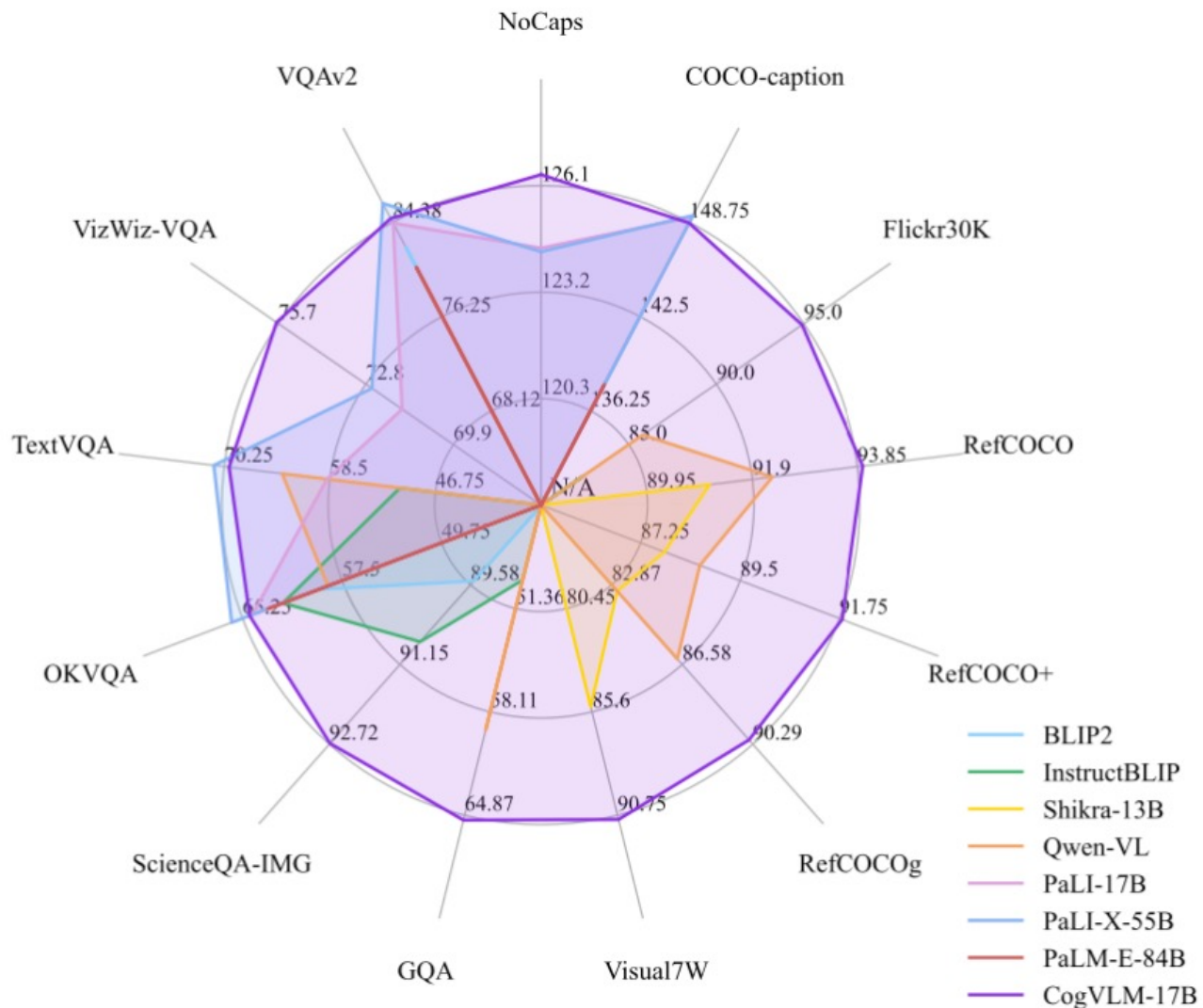
MiniGPT4: The image shows a smartphone with a USB cable plugged into it. The phone appears to be charging. The image is not funny.
翻译: 这幅图片展示了一部智能手机, 手机上插有一个USB线。手机似乎正在充电。这幅图片并不好笑。

LLava-v1.5: The funny thing in this image is that the cell phone is plugged into a charger, but the charger is upside down. This is unusual because chargers are typically used with the correct orientation, with the plug facing downwards to connect to the device. The upside-down charger adds a humorous element to the scene.
翻译: 这幅图片中有趣的是, 手机插在充电器上, 但充电器是倒放的。这很不寻常, 因为充电器通常都是正确方向使用的。插头朝下以连接到设备。这个颠倒的充电器为场景增添了幽默元素。

CogVLM: Vision Language Models

开源

提出基于认知的多模态理解算法，实现类似**GPT-4V**的初步版本



- 对比最近的 MiniGPT-4、LLaVA-v1.5，可以看出，CogVLM-17B在图像理解、模型幻觉以及文本识别方面都具有不错的效果

CogVLM: Vision Language Models

开源

World Knowledge



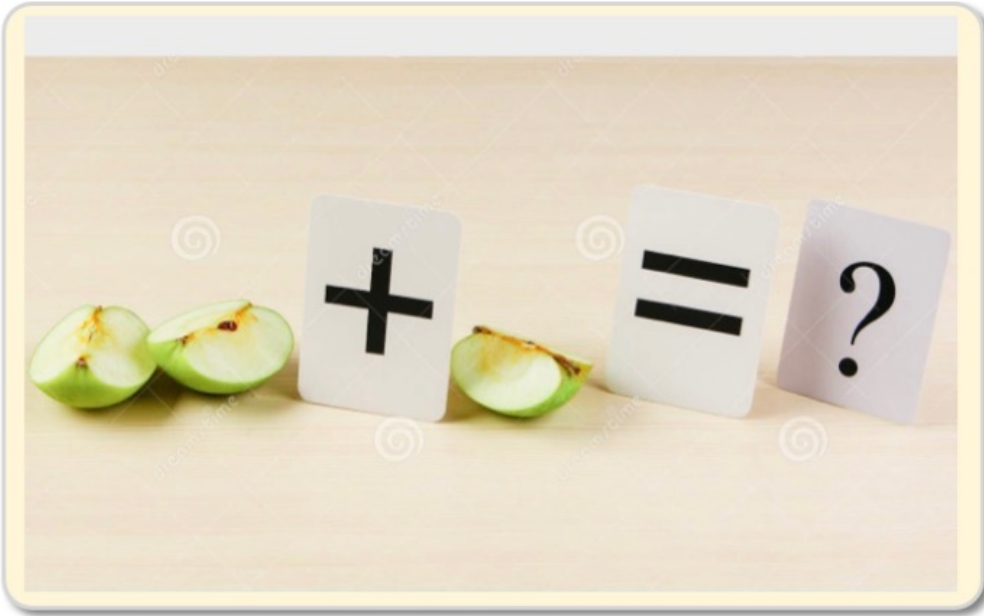
Q: How many goals did the person in the picture score in the 2018 World Cup?

CogVLM: The person in the picture is Cristiano Ronaldo. He is the captain of the Portuguese team and has participated in the 2018 World Cup. In the 2018 World Cup, Cristiano Ronaldo scored 4 goals.

CogVLM: Vision Language Model

开源

Visual Math Problem Solving



Q1: How to solve it?

CogVLM: The problem in the picture is a simple math problem, where a sliced green apple is equated to a question mark. Specifically, in the middle of the picture, there are three green apples placed side by side and three cardboard signs. The signs have black characters on them: a plus sign, an equal sign, and a question mark. The equation shown in the picture is: $2 + 1 = ?$

Q2: What's the final answer?

CogVLM: $2+1=3$

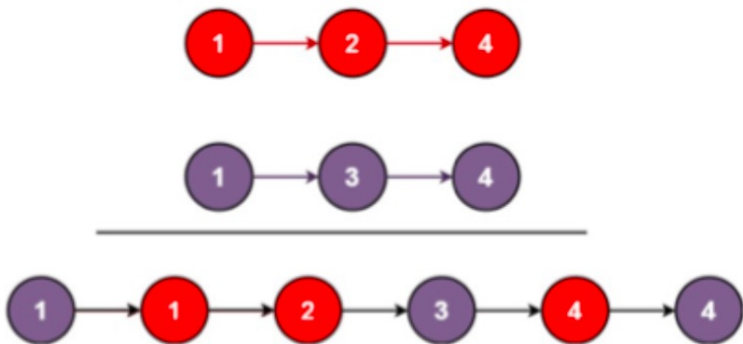
Programming with Visual Input

You are given the heads of two sorted linked lists `list1` and `list2`.

Merge the two lists into one **sorted** list. The list should be made by splicing together the nodes of the first two lists.

Return the head of the merged linked list.

Example 1:



Input: `list1 = [1,2,4]`, `list2 = [1,3,4]`

Output: `[1,1,2,3,4,4]`

```
1  #include <iostream>
2  using namespace std;
3
4  struct ListNode {
5      int val;
6      ListNode* next;
7  };
8
9  ListNode* mergeTwoLists(ListNode* l1, ListNode* l2) {
10     if (l1 == NULL || l2 == NULL) {
11         return NULL;
12     }
13     ListNode* head = new ListNode();
14     head->val = l1->val;
15     head->next = l1;
16
17     ListNode* current = head;
18     while (l1 != NULL && l2 != NULL) {
19         if (l1->val <= l2->val) {
20             current->next = l1;
21             l1 = l1->next;
22         } else {
23             current->next = l2;
24             l2 = l2->next;
25         }
26         current = current->next;
27     }
28     current->next = l1->next;
29     return head;
30 }
```

CogVLM for Building GUI agents?



CogAgent: A Visual Language Model for GUI Agents

开源



User: Who are the teammates on the current mission?

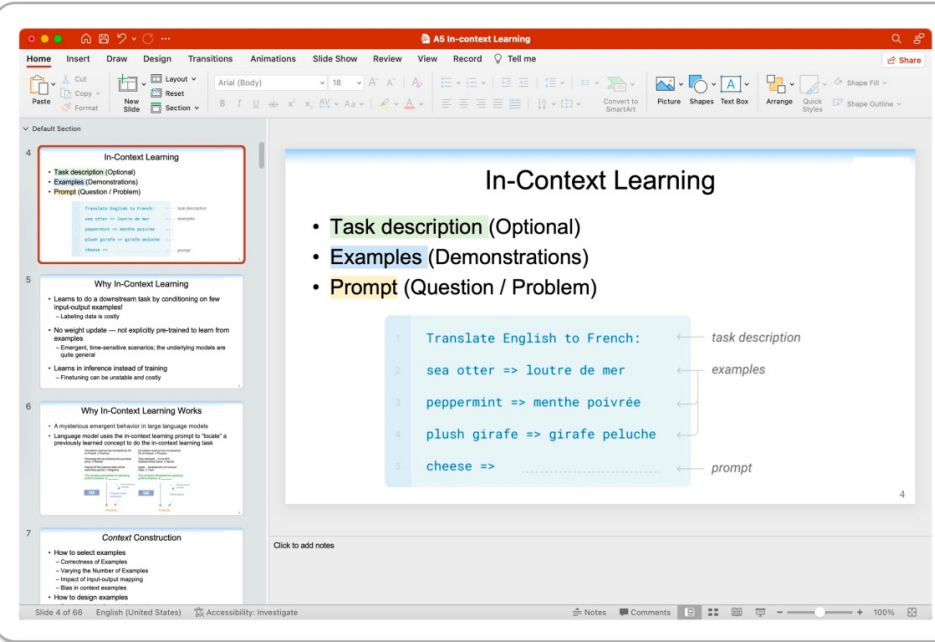
CogAgent: The teammates on the current mission are Raiden Shogun, Bennett, and Kaedehara Kazuha.

User: I'm looking for guidance on how to "Find out how to view a detailed map"?

CogAgent: Plan: 1. Identify the map icon on the screen. 2. Click on the map icon to open the detailed map view. Next Action: Move the cursor over the top left corner of the screen and click on the compass-like icon with 'N' inside it.

CogAgent: A Visual Language Model for GUI Agents

开源



User: Describe the screenshot in detail.

CogAgent: The screenshot shows an open PowerPoint presentation titled 'In-Context Learning' with a focus on task description, examples, and prompts. It is in the 'Slide Show' view mode with slide number 4 selected. There are three main bullet points listed under 'Task Description (Optional)': Examples (Demonstrations), Prompt (Question / Problem). Below this, there are four more bullet points for translating English to French: sea otter -> loutre de mer, peppermint -> menthe poivrée, plush giraffe -> girafe peluche, cheese -> fromage. On the left sidebar, there are thumbnails of other slides related to context learning, including one about why in-context learning works and another discussing context construction.

User: Could you provide instructions for "Insert a new slide named "In-Context Learning: Details" with a Two Content layout after the current slide."?

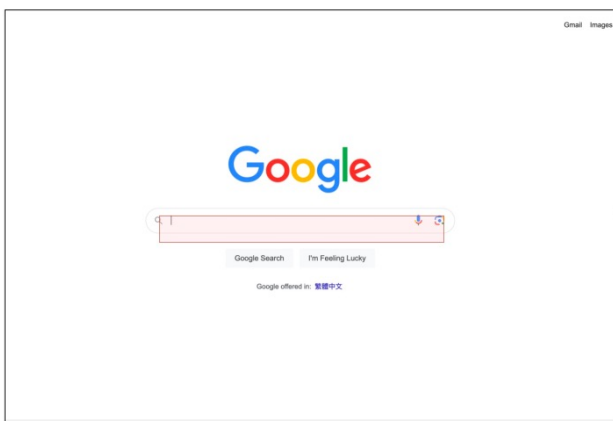
CogAgent: [Plan]: 1. Click on the 'New Slide' button to start a new slide from scratch. 2. Once the new slide is open, click on the 'Home' tab if not already selected. 3. In the Home tab, find and select the 'Two Content Layout' option. 4. Type in the title for the new slide as 'In-Context Learning: Details'. [Next Action]: Move the cursor over the 'New Slide' button located at the specified coordinates on the screen and perform a left mouse click.

CogAgent: A Visual Language Model for GUI Agents

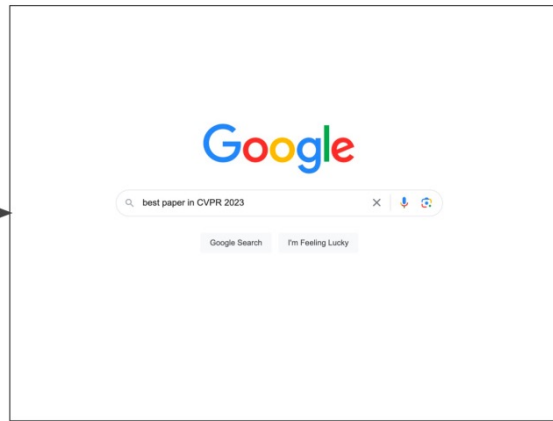
开源

User: Could you guide me through the steps "Search for the best paper in CVPR 2023"?

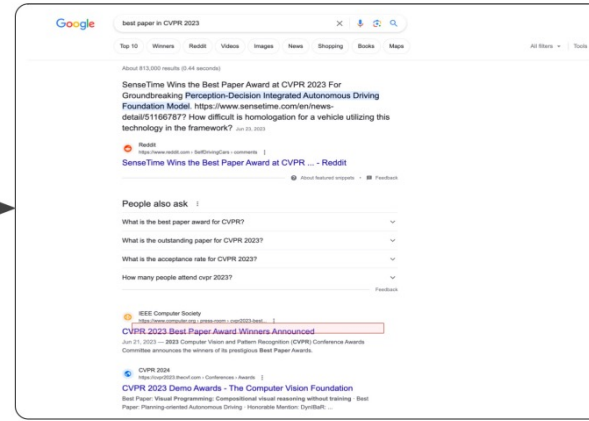
Computer Agent



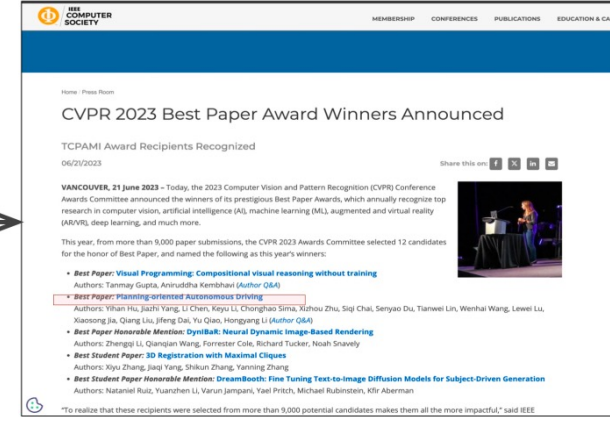
CogAgent: Move the cursor to the Google search box, then type 'best paper in CVPR 2023'.
[Operation]: [combobox] Search -> TYPE: best paper in CVPR 2023 at the box {'x_min': 0.28, 'y_min': 0.509, 'width': 0.438, 'height': 0.06}



CogAgent: Press the enter key on the keyboard while the cursor is inside the Google Search box to initiate the search for 'best paper in CVPR 2023'.
[Operation]: Press Enter



CogAgent: Move the cursor over the box that corresponds to the 'Best Paper Award Winners Announced' link, and click.
[Operation]: [link] Best Paper Award Winners Announced -> CLICK at the box {'x_min': 0.189, 'y_min': 0.765, 'width': 0.462, 'height': 0.019}



CogAgent: Click to select the 'Best Paper' link for the 2023 CVPR Best Paper Award.
[Operation]: [link] Best Paper: Planning-oriented Autonomous Driving -> CLICK at the box {'x_min': 0.056, 'y_min': 0.709, 'width': 0.419, 'height': 0.018}

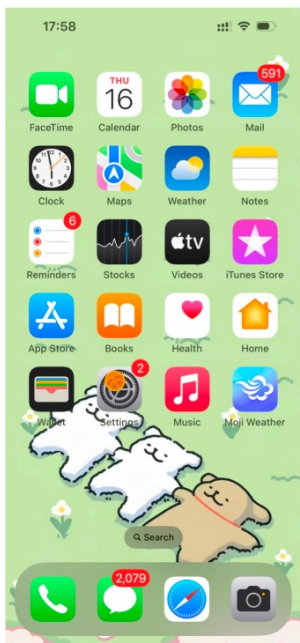
CogAgent: A Visual Language Model for GUI Agents

开源

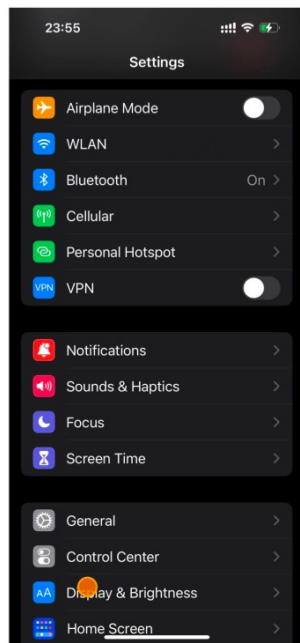


User: Could you guide me through the steps to "change my phone's display mode to light mode"?

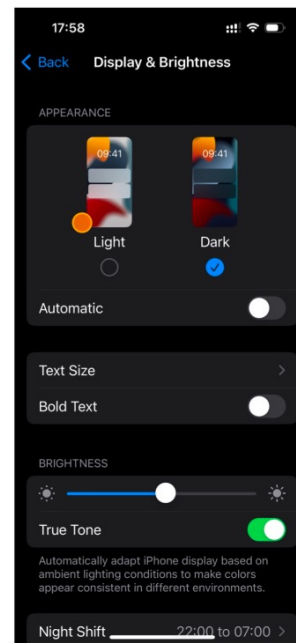
Smartphone Agent



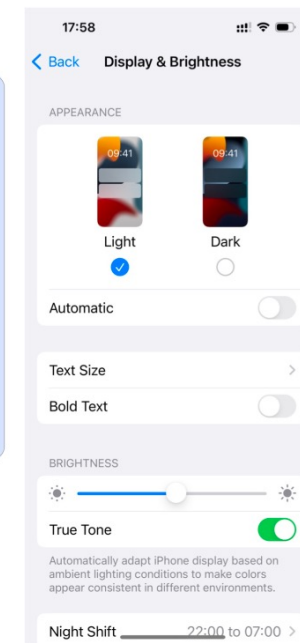
CogAgent: Touch the area on the screen where the 'Settings' icon is located. This action aims to open the settings menu where you can proceed with changing the phone's display mode to light mode.
[Operation]:
tap [x,y]=[0.387,0.625]



CogAgent: Touch the area of the screen where the 'Display & Brightness' text is located. This action aims to open the display settings where you can proceed with changing the light mode.
[Operation]:
tap [x,y]=[0.29,0.895]

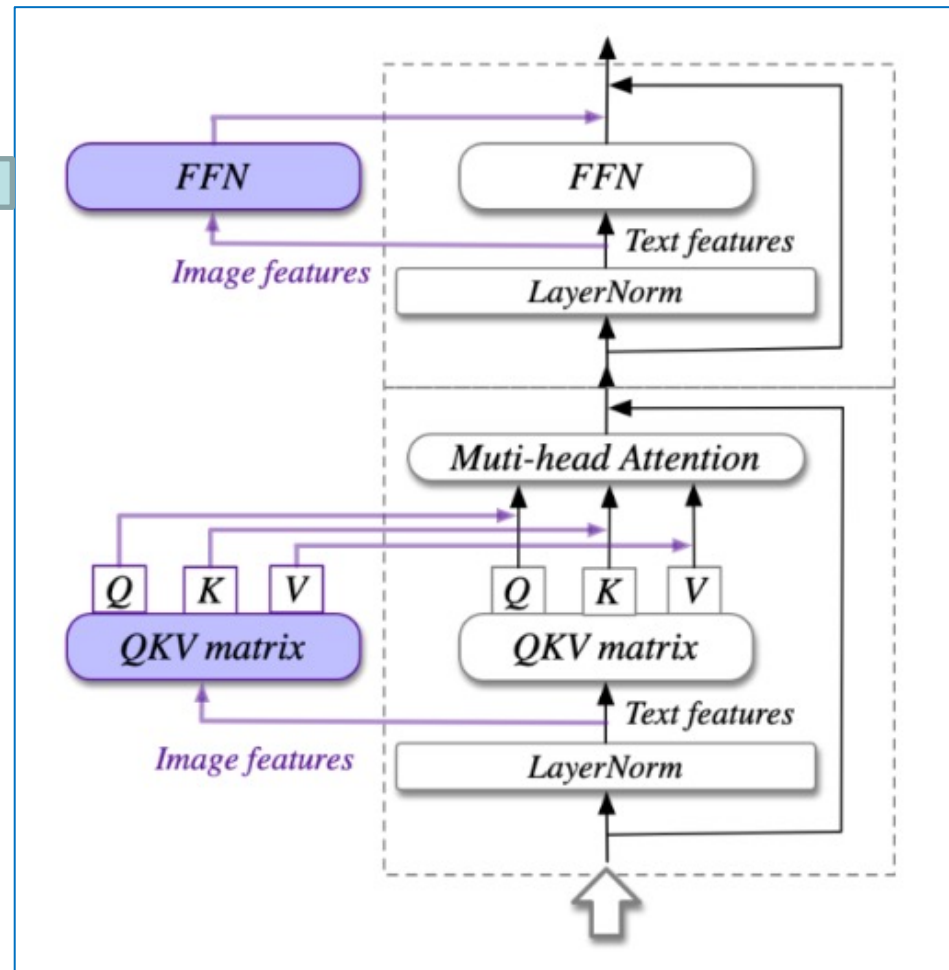
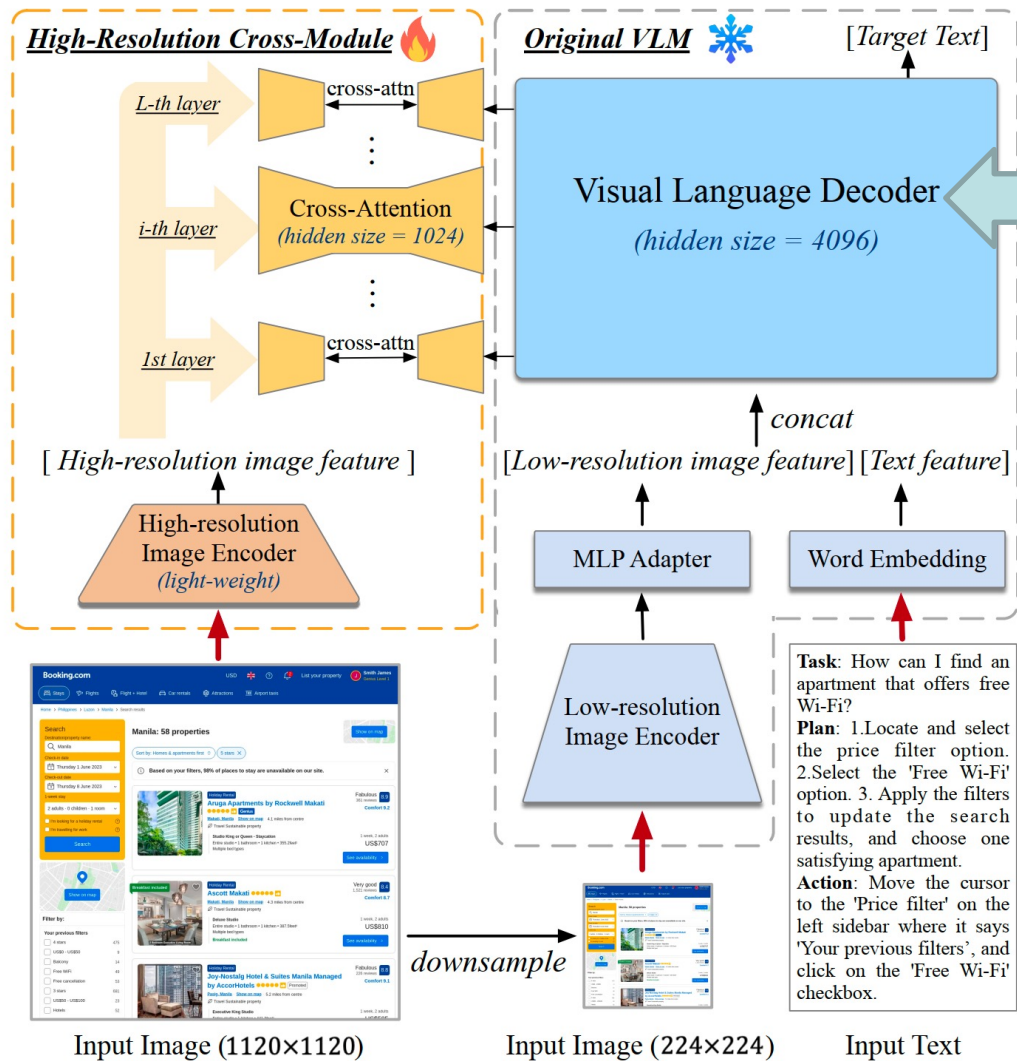


CogAgent: Tap the 'Light' option for display mode, which is currently highlighted and located at the coordinates [x,y]=[0.156,0.349] on the screen. This will select 'Light' mode and activate it.
[Operation]:
tap [x,y]=[0.176,0.349]



CogAgent: A Visual Language Model for GUI Agents

开源



清华&智谱 GLM 系列模型

WebGPT | New Bing

搜索增强大模型

WebGLM
KDD

Codex | Copilot

代码大模型

CodeGeeX
KDD

ChatGPT | GPT4-Turbo

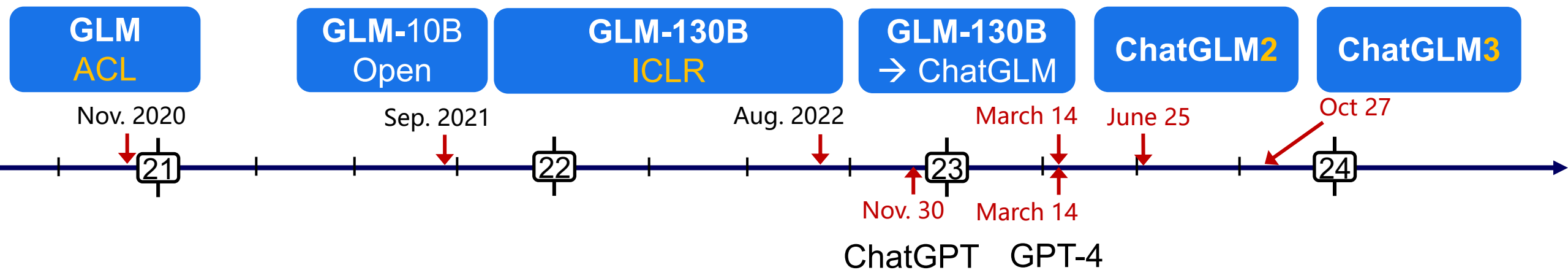
智能体大模型

AgentTuning
AgentBench

Dall-E | GPT4-V

多模态大模型

CogVLM
CogAgent



OpenAI GPT vs 清华&智谱GLM

All Tools, GPTs

GPT-4

对话

ChatGPT

图片理解

GPT-4V

代码

Codex

文生图

DALL.E

搜索增强

WebGPT

千亿基座

GPT-3

GPT系列模型

GPT

VS

GLM

?

ChatGLM

对话

CogVLM

图片理解

CodeGeeX

代码

CogView

文生图

WebGLM

搜索增强

GLM-130B

千亿基座

GLM系列模型

开放的大模型研究

#star

36,000

ChatGLM-6B Public

ChatGLM-6B: An Open Bilingual Dialogue Language Model | 开源双语对话语言模型

Python 36,684 Apache-2.0 4,914 508 42 Updated on Nov 2

14,000

ChatGLM2-6B Public

ChatGLM2-6B: An Open Bilingual Chat LLM | 开源双语对话语言模型

Python 14,677 2,316 398 27 Updated last week

7,000

ChatGLM3 Public

ChatGLM3 series: Open Bilingual Chat LLMs | 开源双语对话语言模型

Python 7,448 780 4 (1 issue needs help) 0 Updated 3 days ago

7,000

GLM-130B Public

GLM-130B: An Open Bilingual Pre-Trained Model (ICLR 2023)

Python 7,414 Apache-2.0 595 113 5 Updated on Jul 25

7,000

CodeGeeX Public

CodeGeeX: An Open Multilingual Code Generation Model (KDD 2023)

Python 7,398 Apache-2.0 520 143 5 Updated on Oct 26

5,800

CodeGeeX2 Public

CodeGeeX2: A More Powerful Multilingual Code Generation Model

Python 5,830 Apache-2.0 378 142 1 Updated on Aug 12

3,600

VisualGLM-6B Public

Chinese and English multimodal conversational language model | 多模态中英双语对话语言模型

Python 3,681 Apache-2.0 374 230 6 Updated 2 weeks ago

3,300

CogVideo Public

Text-to-video generation. The repo for ICLR2023 paper "CogVideo: Large-scale Pretraining for Text-to-Video Generation via Transformers"

Python 3,315 Apache-2.0 350 10 2 Updated on Jun 14

3,000

CogVLM Public

a state-of-the-art-level open visual language model | 多模态预训练模型

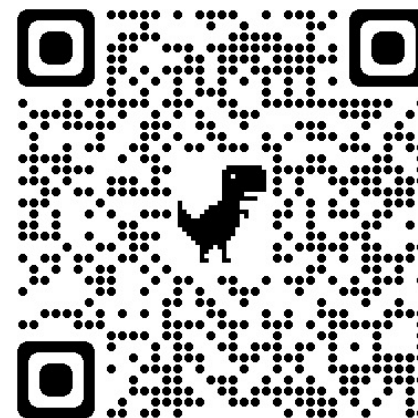
Python 3,055 155 20 1 Updated 19 hours ago

2,800

GLM Public

GLM (General Language Model)

Python 2,815 MIT 294 91 8 Updated on Nov 3

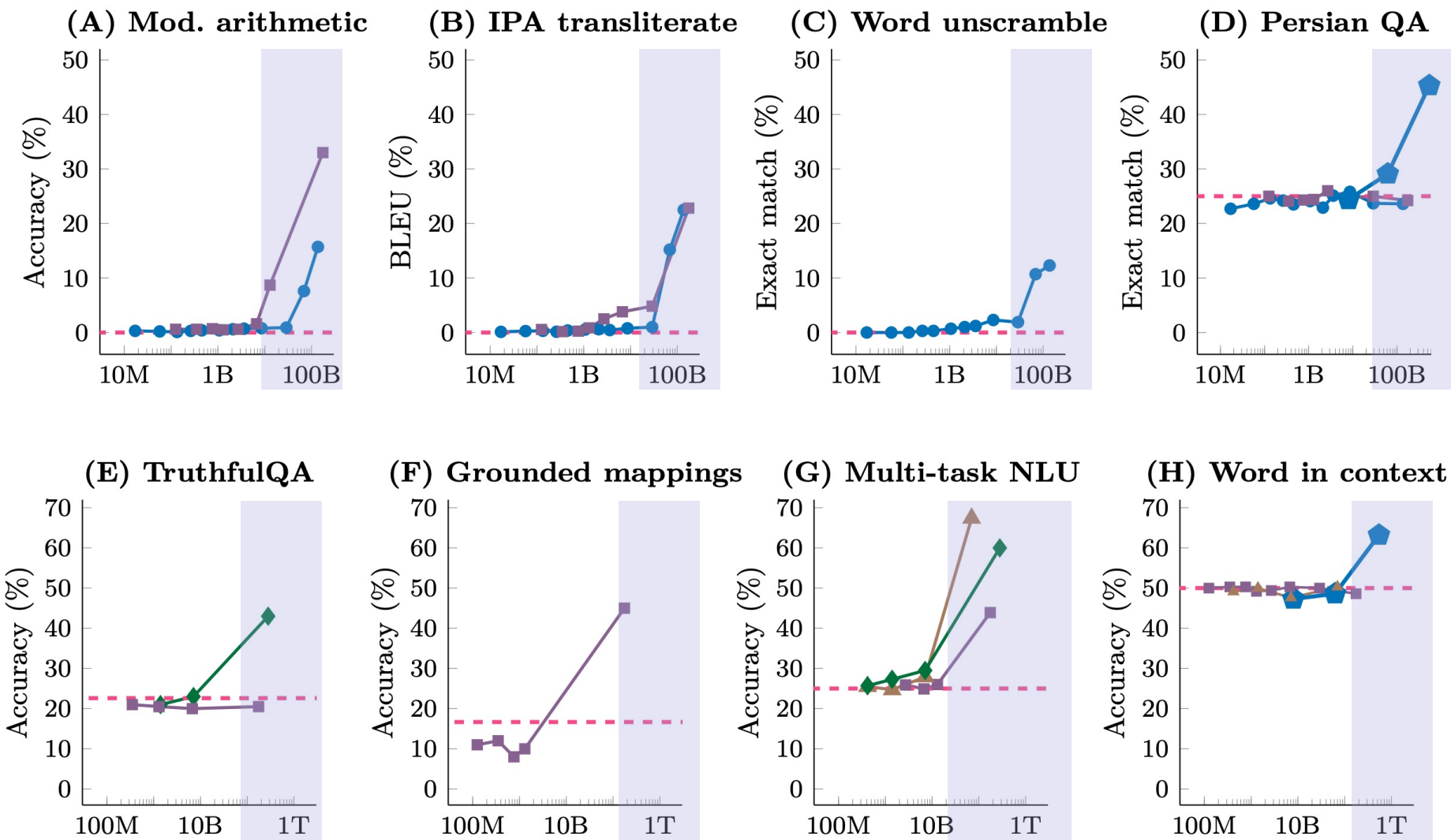


<https://github.com/THUDM>

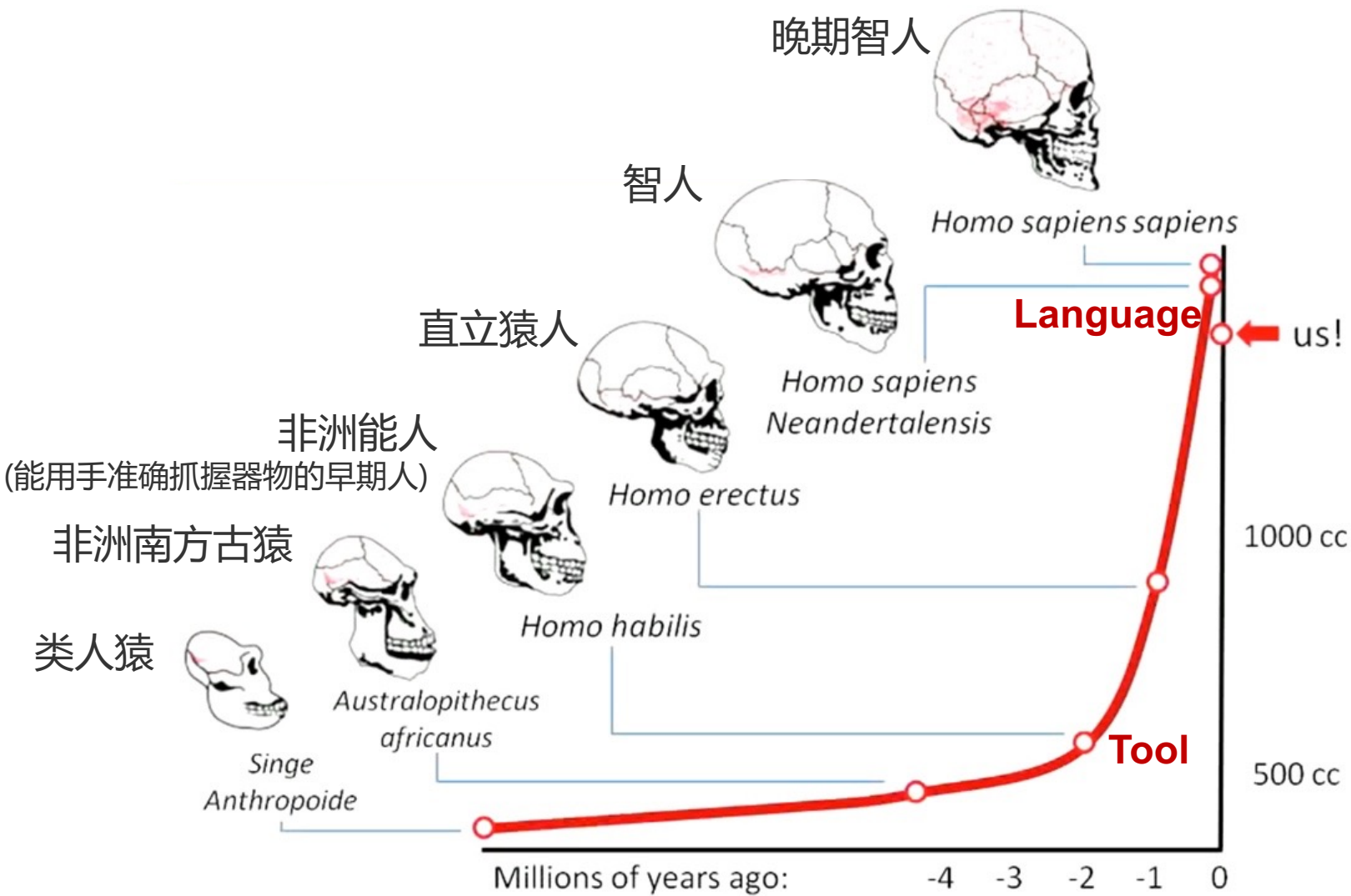
Accessed on Dec. 22, 2023

为什么大模型?

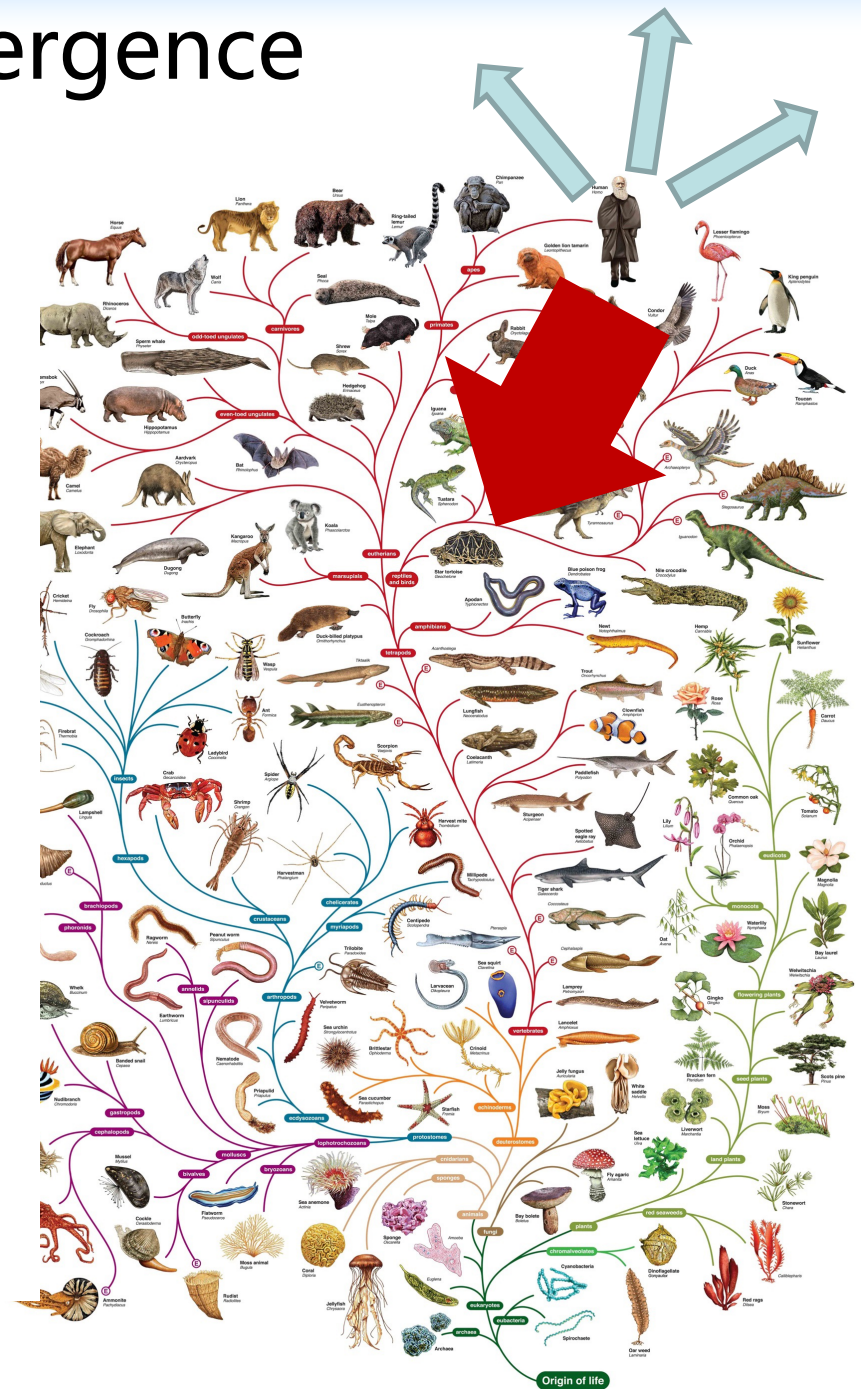
—●— LaMDA —■— GPT-3 —◆— Gopher —▲— Chinchilla —◆— PaLM - - - Random



大脑进化 Spike vs. 大模型 emergence



Disclaimer: The instructor is not an expert on this topic at all



Scaling 的尽头?




1. 真正的 Scaling 规模远超想象：从 Meta 的一篇冷门论文中一窥

- 546B 模型训练量为：65536*150000*2048 = 20T，20 倍于 PaLM-540B
- 需要约 **2 万张 A100 训练一年**，是目前公开信息最大训练量的模型
- 据此作为标杆，坚信和擅长 Scaling 的 OpenAI GPT-4 的计算量是多少？

$$\alpha \times \text{Model size} \times \text{Training tokens} = \text{Training compute}$$

model	depth	embedding dimension	b (batch size)	η_t (learning rate)	ϵ	β_1	β_2
7b	32	4096	2048	$\approx 10^{-4}$	10^{-8}	0.9	0.95
30b	36	8192	8192	$\approx 10^{-4}$	10^{-8}	0.9	0.95
65b	80	8192	8192	$\approx 6 \times 10^{-5}$	10^{-8}	0.9	0.95
546b	108	20480	65536	$\approx 2 \times 10^{-5}$	10^{-8}	0.9	0.95

Table 1: Training run settings

	<u>Model size</u> (# parameters)	<u>Training data</u> (# tokens)	<u>Training compute</u> (FLOPs)	<u>Resources</u>
 BERT-base (2018)	109M	250B	1.6e20	64 TPU v2 for 4 days (16 V100 GPU for 33 hrs)
 GPT-3 (2020)	175B	300B	3.1e23	$\sim 1,000\times$ BERT-base
 PaLM (2022)	540B	780B	2.5e24	$6k$ TPU v4 for 2 months

Remember "moles" from high school chemistry? **6e23**
 "as many elementary entities as there are atoms in 0.012 kilogram of carbon 12"

Scaling, emergence, and reasoning (Jason Wei, NYU)

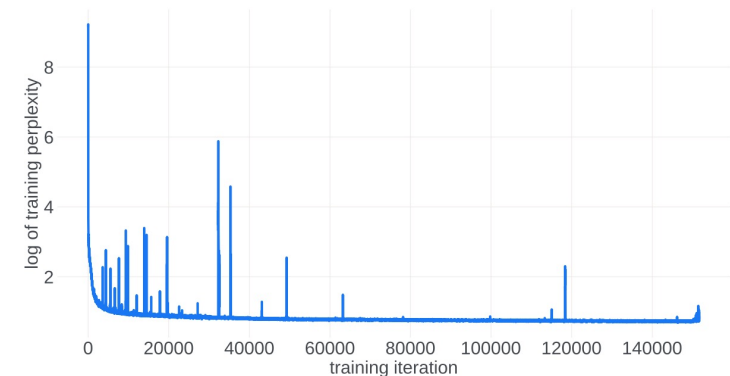
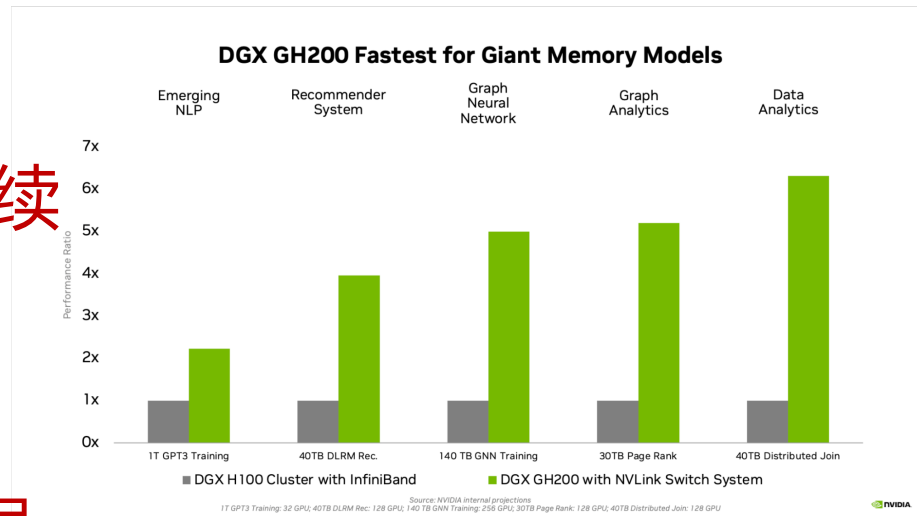


Figure 1: Training perplexity curve of 546b model with prominent spikes

Scaling 的尽头?

2. Scaling: 单卡摩尔定律放缓, 但系统角度仍延续

- DGX A100 -> H100 -> GH200



3. 考虑高质量数据不够之前, 先考虑如何过完数据

- Chinchilla Scaling Law: 最优设定下, 模型增长一倍、所需数据增长一倍
- 也可以解读为: 过完数据的计算量是数据量的平方
 - LLaMA 65B 训练 1.4T token, 凑出 14T 高质量数据则**计算量 x100**
- 引入新的架构 (eg., MoE) 高效学习数据?

4. 期望看到 Scaling 的真正极限: AGI 的一点曙光?

极致 Scaling + 环境交互 + 自我学习

开源

谢谢大家！



<https://github.com/THUDM>



<https://huggingface.co/THUDM>



智谱清言 (iOS、安卓等)



<https://chatglm.cn>



相关文献

1. Zeng et al. **AgentTuning**: Enabling Generalized Agent Abilities for LLMs. arXiv: 2310.12823
2. Liu et al. **AgentBench**: Evaluating LLMs as Agents. arXiv: 2308.03688
3. Hong et al. **CogAgent**: A Visual Language Model for GUI Agents. Nov. 2023. <https://github.com/THUDM/CogAgent>
4. Wang et al. CogVLM: **CogVLM**: Visual Expert for Pretrained Language Models. arXiv: 2311.03079
5. Qinkai Zheng, Xiao Xia, Xu Zou, Yuxiao Dong, Shan Wang, Yufei Xue, Lei Shen, Zihan Wang, Andi Wang, Yang Li, Teng Su, Zhilin Yang, and Jie Tang. **CodeGeeX**: A Pre-Trained Model for Code Generation with Multilingual Benchmarking on HumanEval-X. KDD'23.
6. Xiao Liu, Hanyu Lai, Yu Hao, Yifan Xu, Aohan Zeng, Zhengxiao Du, Peng Zhang, Yuxiao Dong, and Jie Tang. **WebGLM**: Towards An Efficient Web-enhanced Question Answering System with Human Preference. KDD'23.
7. Jing Zhang, Xiaokang Zhang, Daniel Zhang-Li, Jifan Yu, Zijun Yao, Zeyao Ma, Yiqi Xu, Haohua Wang, Xiaohan Zhang, Nianyi Lin, Sunrui Lu, Jie Tang, and Juanzi Li. **GLM-Dialog**: Noise-tolerant Pre-Training for Knowledge-grounded Dialogue Generation. KDD'23.
8. Aohan Zeng, Xiao Liu, Zhengxiao Du, Zihan Wang, Hanyu Lai, Ming Ding, Zhuoyi Yang, Yifan Xu, Wendi Zheng, Xiao Xia, Weng Lam Tam, Zixuan Ma, Yufei Xue, Jidong Zhai, Wenguang Chen, Zhiyuan Liu, Peng Zhang, Yuxiao Dong, and Jie Tang. **GLM-130B**: An Open Bilingual Pre-trained Model. ICLR'23.
9. Wenyi Hong, Ming Ding, Wendi Zheng, Xinghan Liu, and Jie Tang. **CogVideo**: Large-scale Pretraining for Text-to-Video Generation via Transformers. ICLR'23.
10. Zhengxiao Du, Yujie Qian, Xiao Liu, Ming Ding, Jiezhong Qiu, Zhilin Yang, and Jie Tang. **GLM**: General Language Model Pretraining with Autoregressive Blank Infilling. ACL'21.

开放的大模型研究

- 2023年05月28日，科技部在**中关村论坛**上发布的《中国人工智能大模型地图研究报告》显示 **ChatGLM-6B** 位大模型**开源影响力第一名**
- 千亿基座 **GLM-130B**、代码模型 **CodeGeeX**、文生视频模型 **CogVideo**、基础架构模型 **GLM** 共5个模型入围**开源影响力前十**

