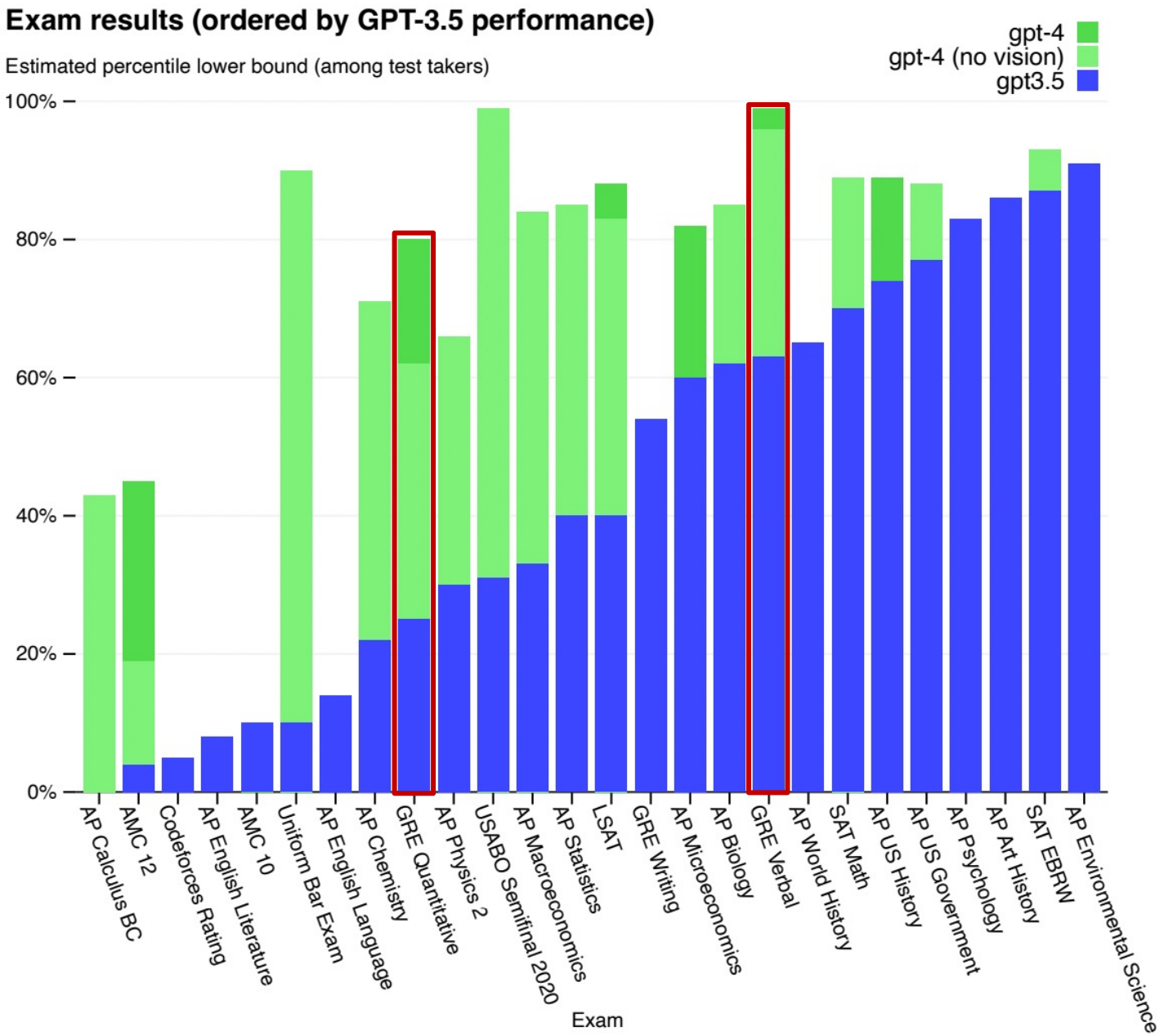
ChatGLM：预训练大模型探索与实践

东昱晓
知识工程实验室（KEG）
清华大学计算机系

https://keg.cs.tsinghua.edu.cn/yuxiao

2023.3.14 GPT-4



OpenAI. GPT-4 Technical Report 2

1950 → 2020 → 2022

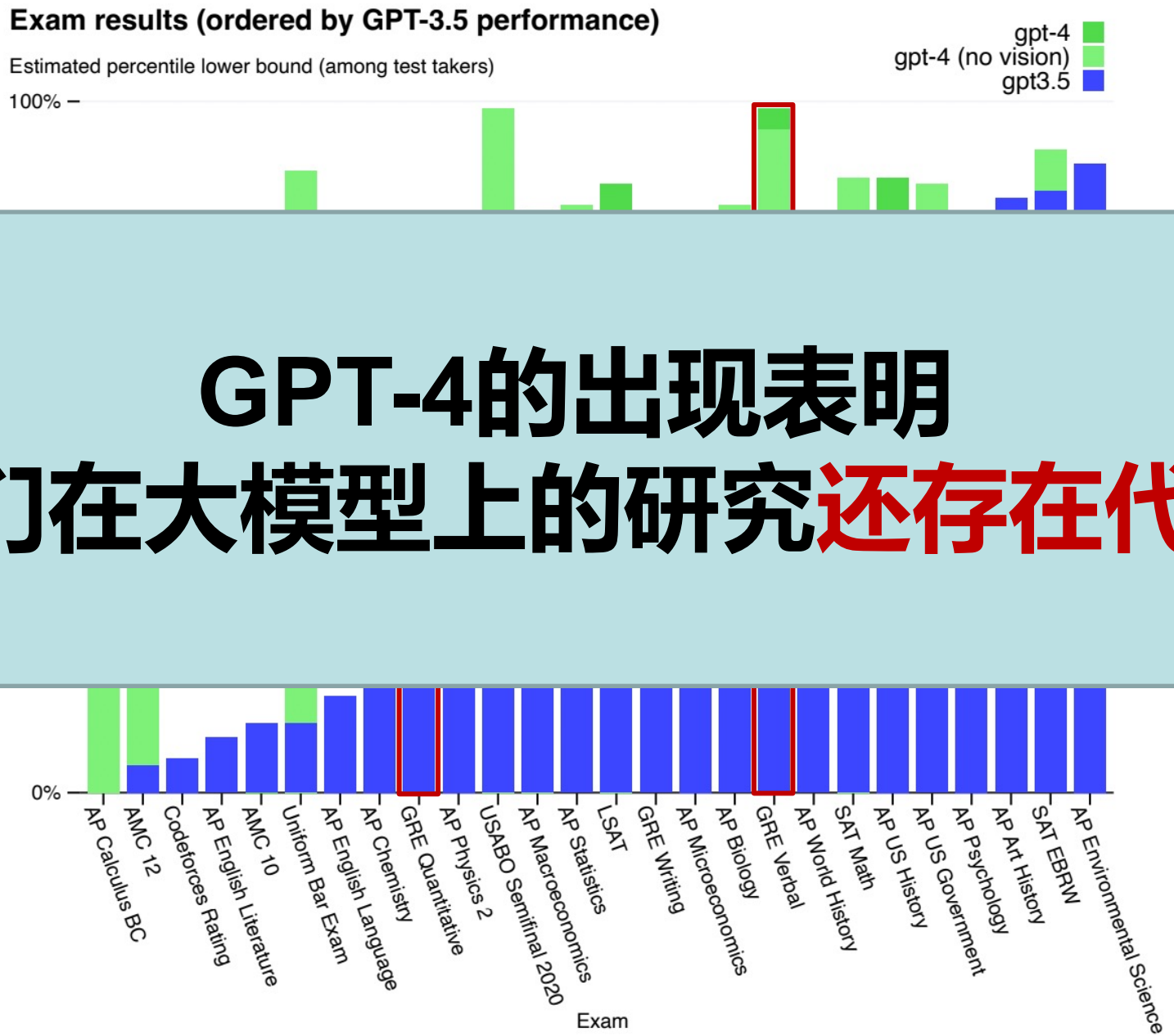
"The Great Wall of China was built from as early as the 7th century BC, with selective stretches later joined by Qin Shi Huang (220–206 BC), the first emperor of China. "

Math (GSM8k)

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?

https://en.wikipedia.org/wiki/Great_Wall_of_China

2023.3.14 GPT-4



GPT-4的出现表明
我们在大模型上的研究还存在代差

OpenAI. GPT-4 Technical Report

2023.3.14 GPT-4

What is funny about this image? Describe it panel by panel.

Source: https://www.reddit.com/r/hmmm/comments/ubab5v/hmmm/

GPT-4

The image shows a package for a "Lightning Cable" adapter with three panels.

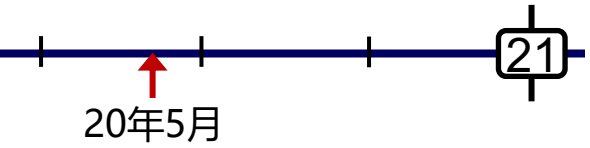
Panel 1: A smartphone with a VGA connector (a large, blue, 15-pin connector typically used for computer monitors) plugged into its charging port.

Panel 2: The package for the "Lightning Cable" adapter with a picture of a VGA connector on it.

Panel 3: A close-up of the VGA connector with a small Lightning connector (used for charging iPhones and other Apple devices) at the end.

OpenAI GPT系列模型

1. 千亿基座模型



GPT-3
千亿基座

GPT-3
+ RLHF

Few-shot

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

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

task description

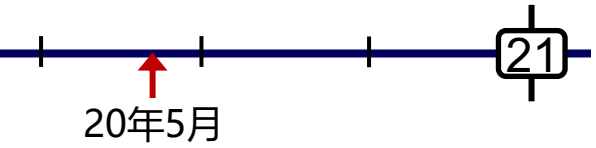
examples

prompt

OpenAI.

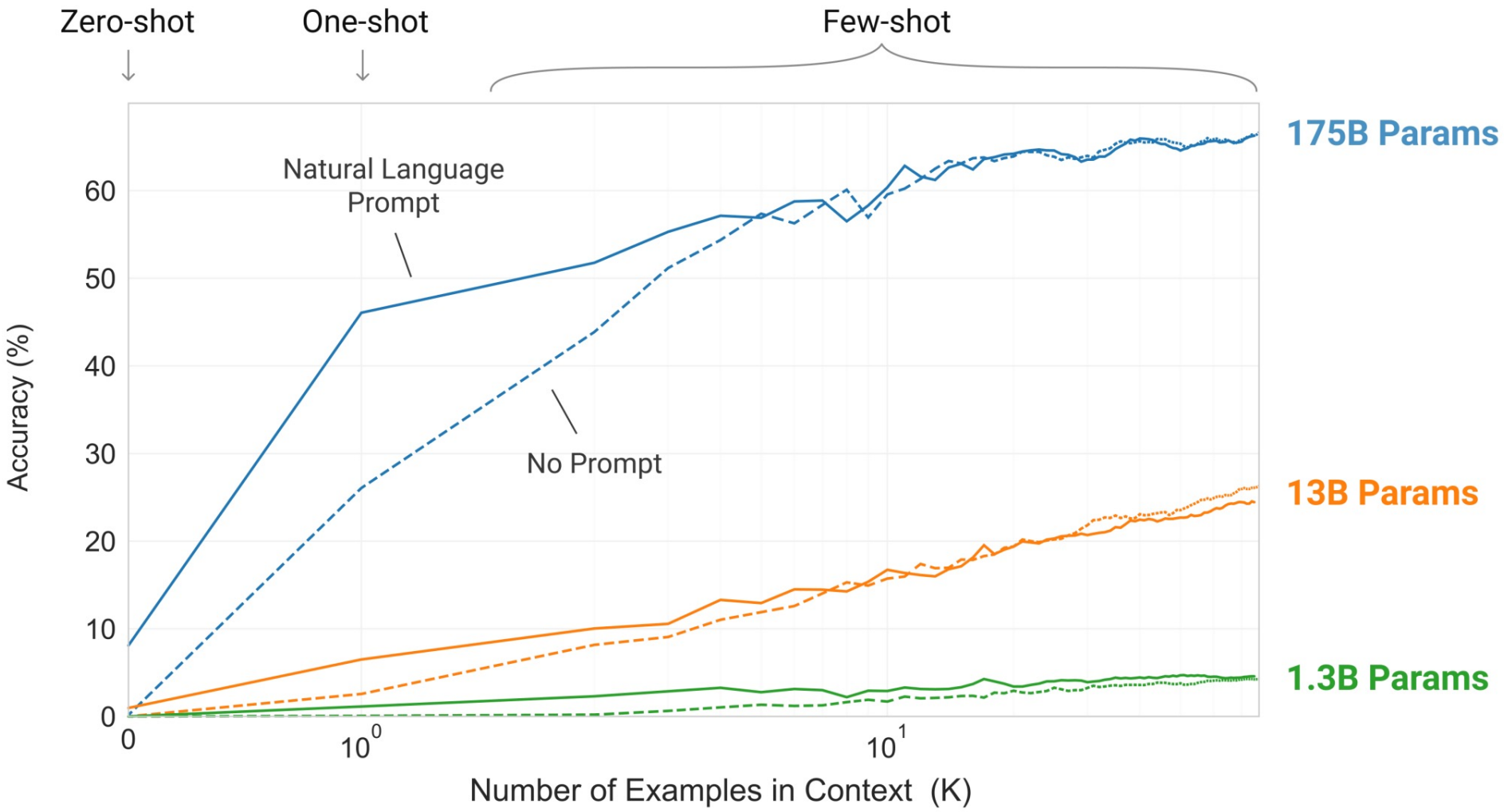
OpenAI GPT系列模型

1. 千亿基座模型



GPT-3
千亿基座

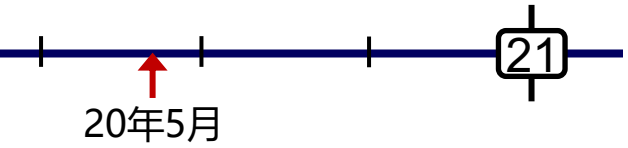
GPT-3
+ RLHF



OpenAI.

OpenAI GPT系列模型

1. 千亿基座模型



GPT-3
千亿基座

GPT-3
+ RLHF

Ground truth

Circulation revenue has increased by 5% in Finland. \n Positive
Panostaja did not disclose the purchase price. \n Neutral
Paying off the national debt will be extremely painful. \n Negative
The company anticipated its operating profit to improve. \n _____

LM

Positive — Correct!

Random

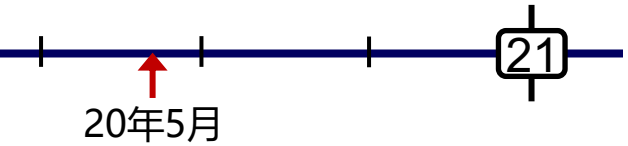
Circulation revenue has increased by 5% in Finland. \n **Neutral**
Panostaja did not disclose the purchase price. \n **Negative**
Paying off the national debt will be extremely painful. \n **Positive**
The company anticipated its operating profit to improve. \n _____

LM

Positive — Correct!

OpenAI GPT系列模型

1. 千亿基座模型



GPT-3
千亿基座

GPT-3
+ RLHF

Prompt

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

Completion

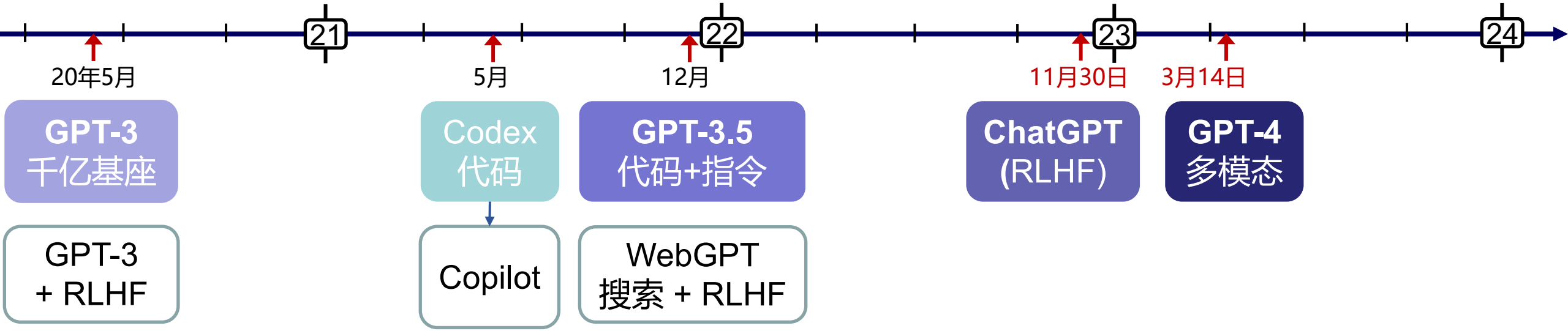
GPT-3

```
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.
```

https://openai.com/research/instruction-following

OpenAI GPT系列模型

1. 千亿基座模型　　**2. 有监督指令微调**　　**3. 人类反馈强化学习**



OpenAI GPT系列模型

Prompt

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

Completion

GPT-3

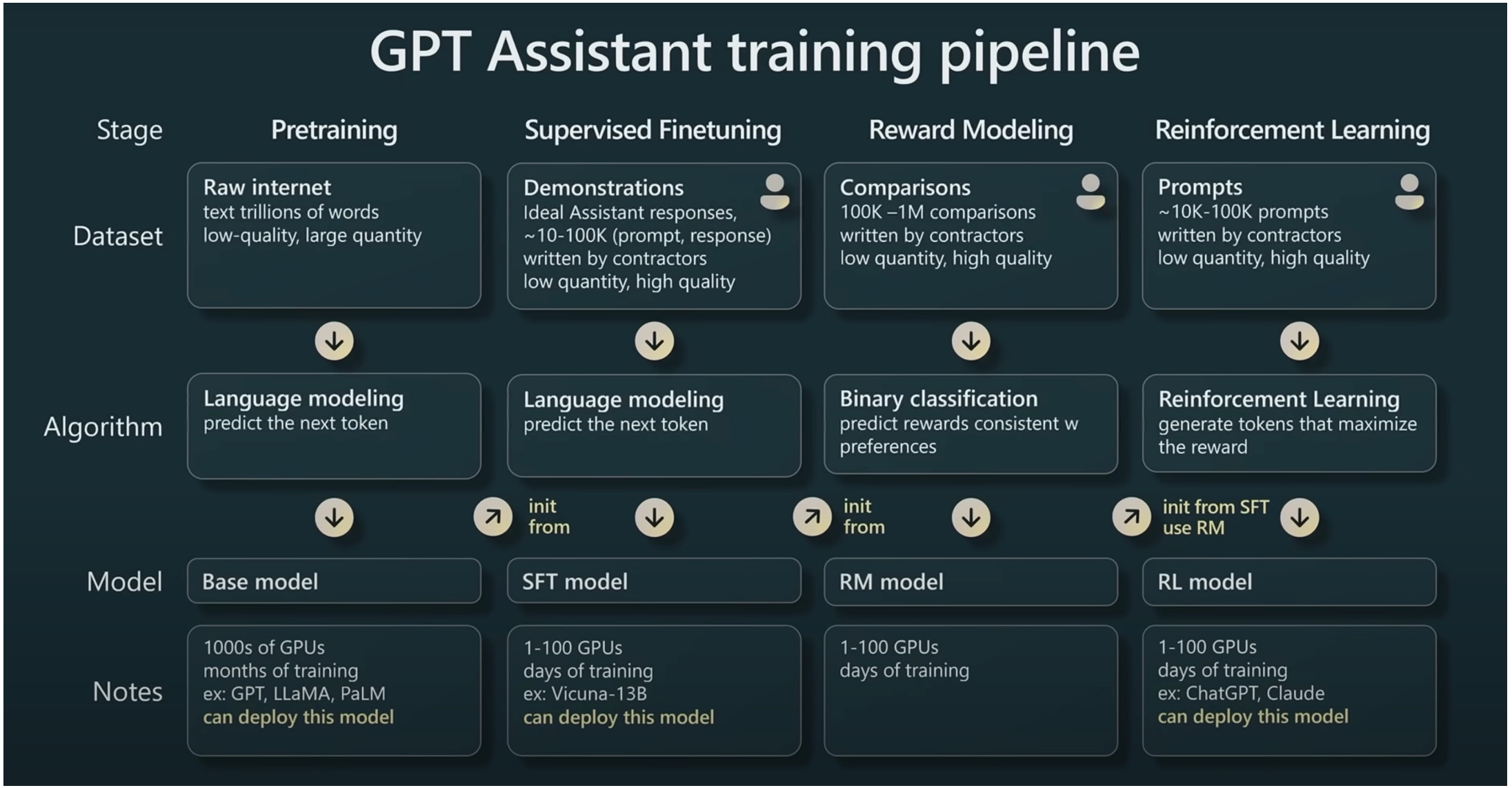
```
Explain the theory of gravity to a 6 year old.

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.
```

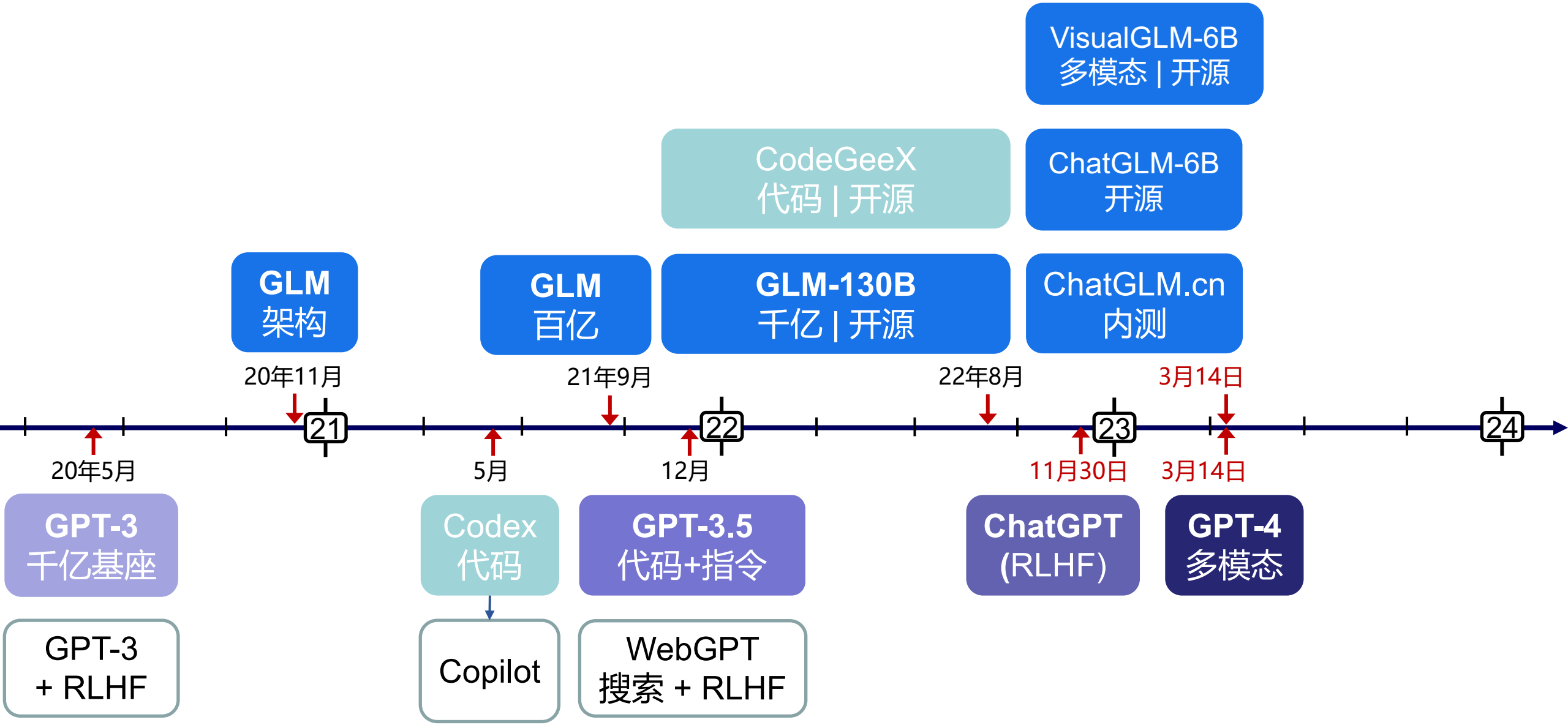
https://openai.com/research/instruction-following

State of GPT by ***Andrej Karpathy***



https://www.youtube.com/watch?v=bZQun8Y4L2A, May 25, 2023

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]

Microsoft NVIDIA

- Megatron-Turing-530B

[2021]

Google

- LaMDA 137B
- PaLM 540B

[2021~2022]

DeepMind

- Chinchilla 70B
- Gopher 260B

[2021~2022]

清華大學 Tsinghua University

智谱·AI

- GLM-130B

[2022]

Meta

- OPT 175B

[2022]

BigScience

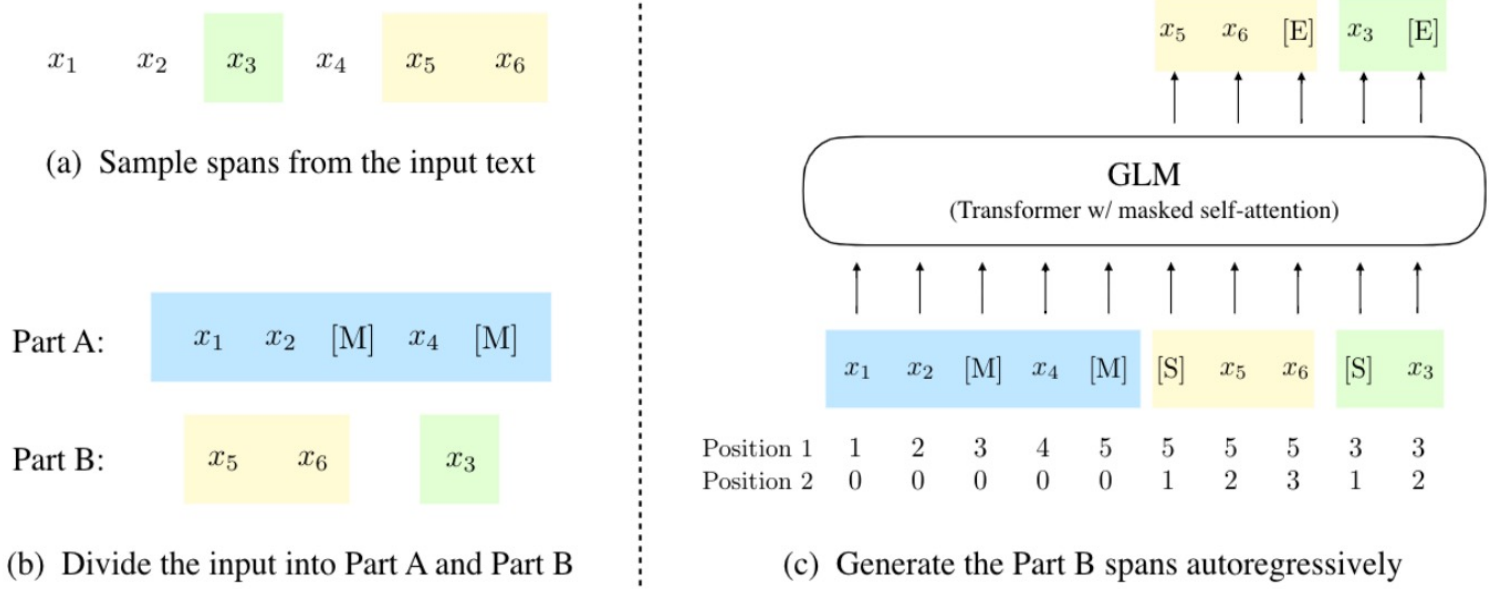
- BLOOM 176B

[2022]

GLM-130B 千亿模型 • **NVIDIA** • **海光 DCU** • **昇腾910** • **申威**	2000亿中文 2000亿英文 (2022.07)	Transformer (1300亿参数)	**GLM** 自回归填空
	数据	神经网络	预训练架构
GPT-3 davinci 千亿基座 • **NVIDIA**	3000亿标识符 (2020.05)	Transformer (1750亿参数)	**GPT** 自回归生成

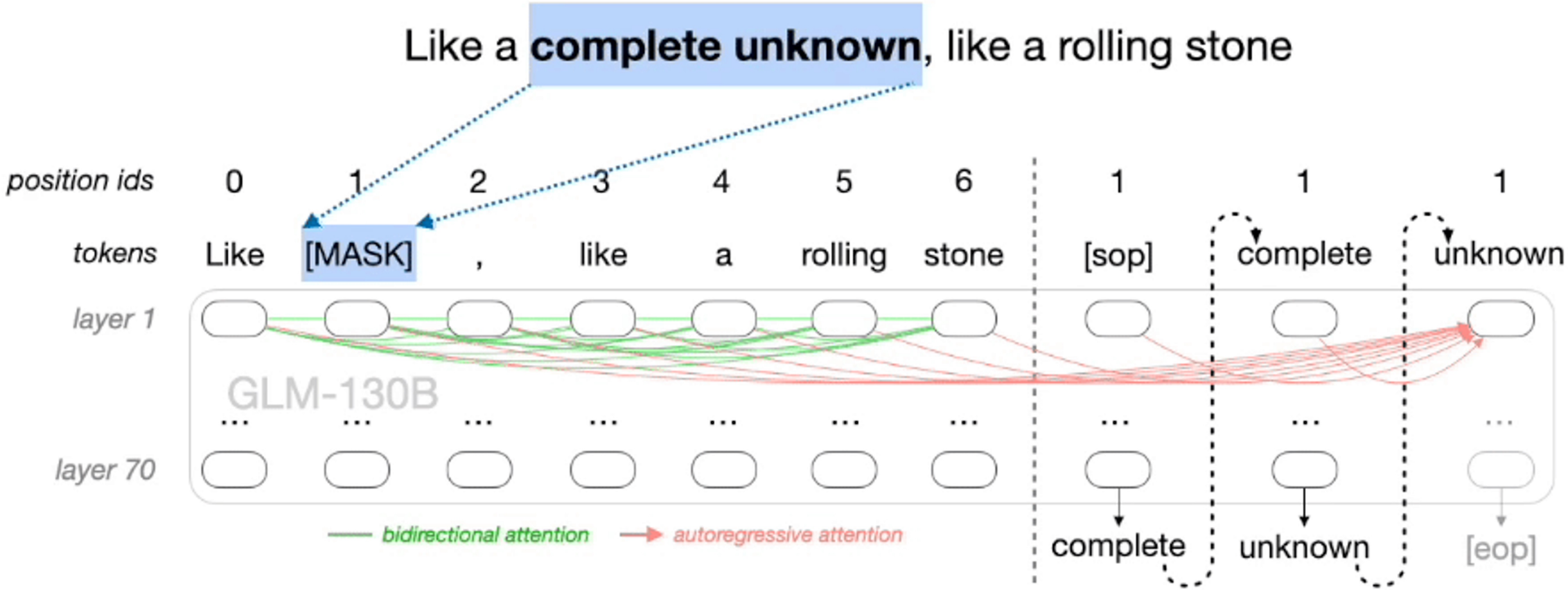
通用语言模型： GLM—自回归填空

算法框架	生成 vs. 理解	自然语言理解	Cond. Gen.	Uncond. Gen.
自回归（GPT）	单向注意力	—	—	√
自编码（BERT）	双向注意力	√	×	×
编码器-解码器（T5）	编解码	—	√	—
自回归填空（GLM）	**双向注意力**	**√**	**√**	**√**



1. Du, et al. “GLM: General Language Model Pre-training with Autoregressive Blanking Infilling.”, ACL, 2022

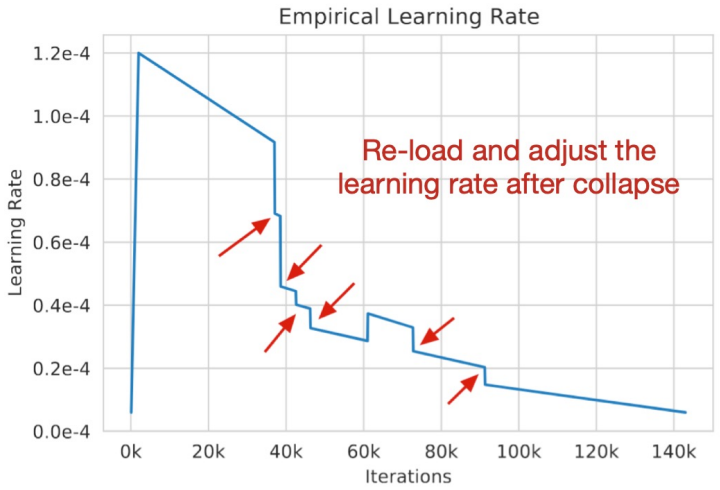
通用语言模型： GLM—自回归填空



1. Du, et al. “GLM: General Language Model Pre-training with Autoregressive Blanking Infilling.”, ACL, 2022

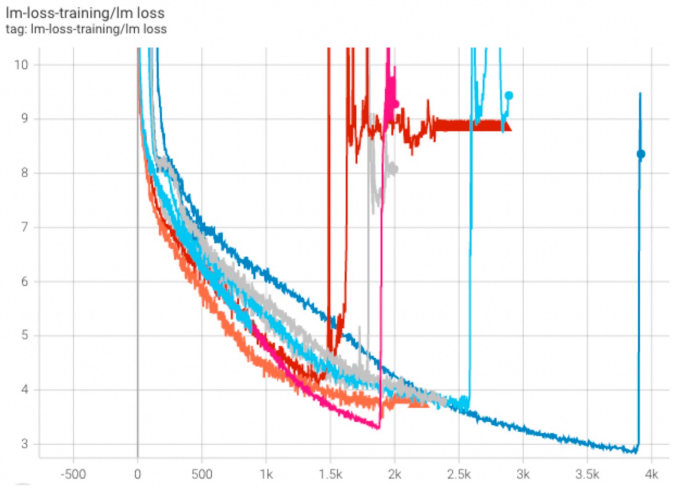
千亿模型训练最大挑战：训练稳定性

- 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

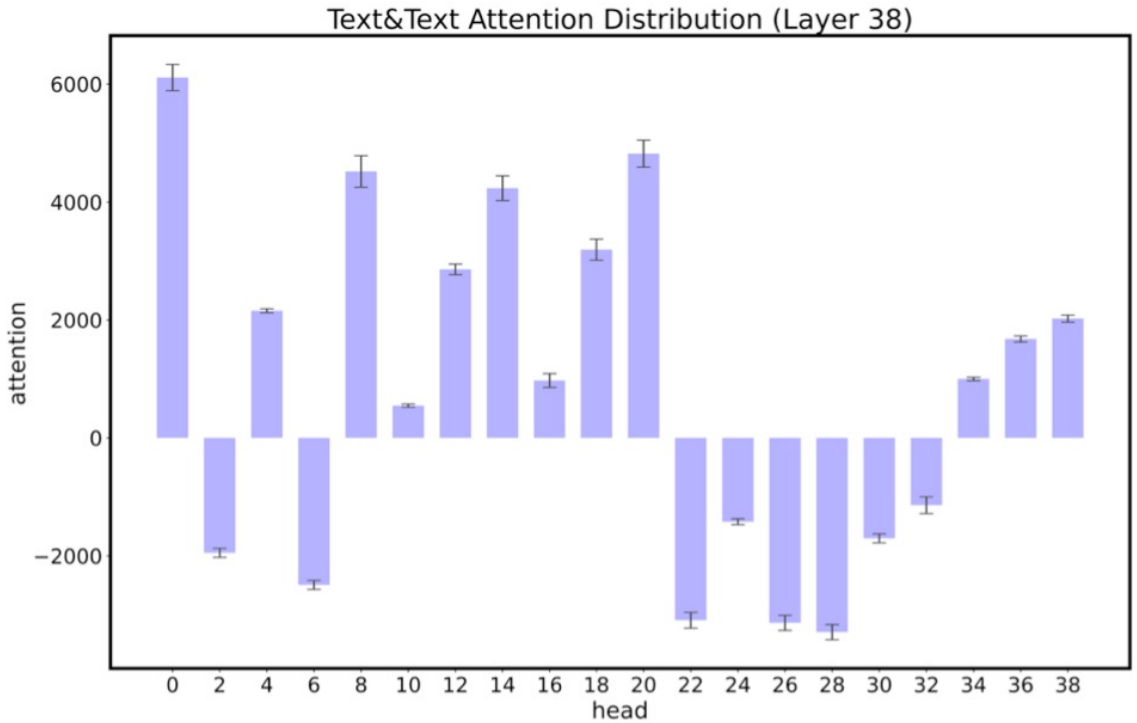
Sources: OPT / BLOOM / GLM-130B

GLM-130B：稳定训练方法

- Attention score 层：Softmax in 32 避免上下溢出

$$\text{softmax}\left(\frac{\boldsymbol{Q}_i\boldsymbol{K}_i^\top}{\sqrt{d}}\right) = \text{softmax}\left(\left(\frac{\boldsymbol{Q}_i\boldsymbol{K}_i^\top}{\alpha\sqrt{d}} - \max\left(\frac{\boldsymbol{Q}_i\boldsymbol{K}_i^\top}{\alpha\sqrt{d}}\right)\right)\times\alpha\right) = \text{FP16}\left(\text{softmax}\left(\text{FP32}\left(\frac{\boldsymbol{Q}_i\boldsymbol{K}_i^\top}{\alpha\sqrt{d}}\right)\times\alpha\right)\right)$$

Attention 层的分数分布很容易超过 FP16 表示范围



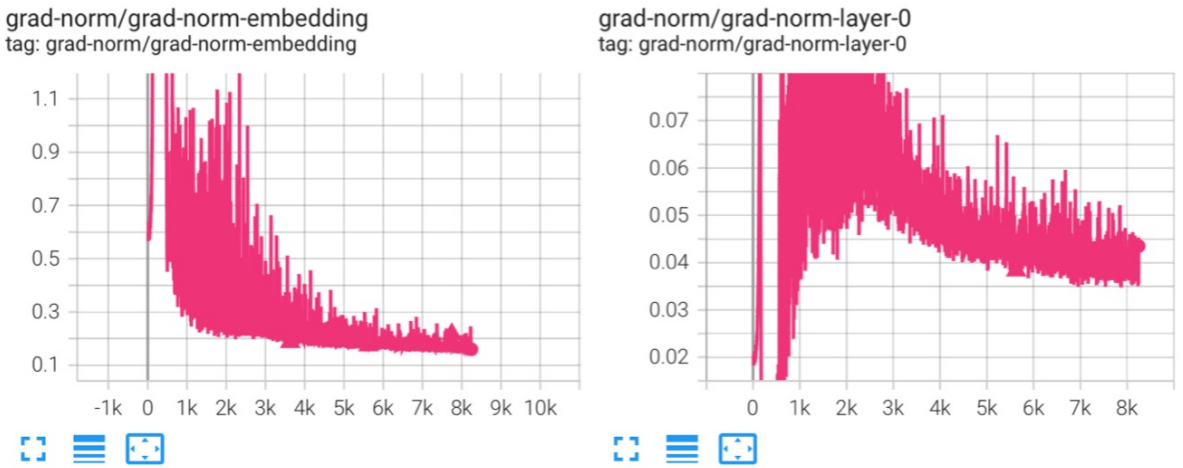
Zeng, Liu, et al. GLM-130B: An Open Bilingual Pre-trained Model. ICLR'23

GLM-130B：稳定训练方法

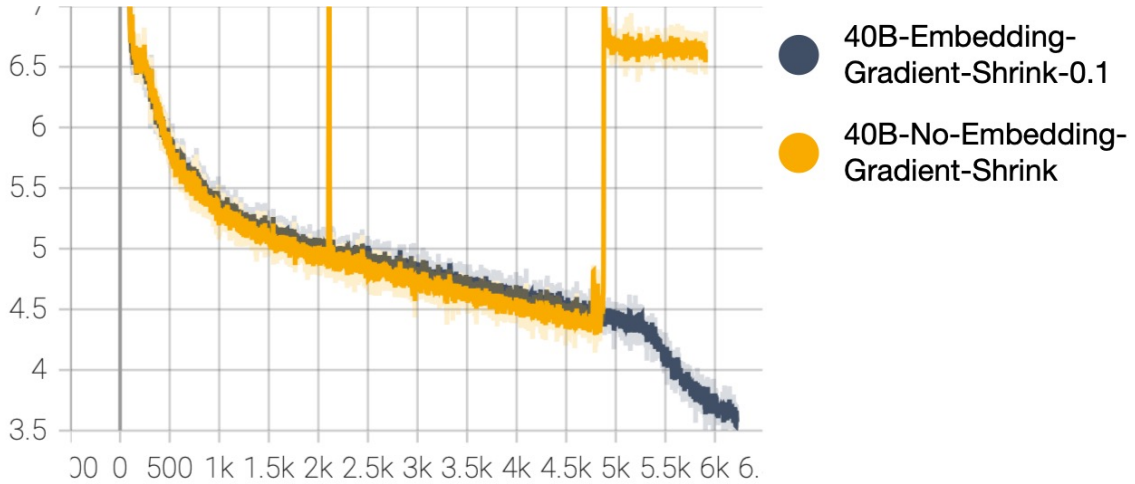
- 调小 Embedding 层梯度，缓解前期梯度爆炸问题

```
word_embedding = word_embedding * alpha +
                 word_embedding.detach() * (1 - alpha)
```

Embedding 层梯度存在数量级上的差异，大模型测试上有效稳定训练



(a) Gradient norm of embedding layer (left) and the first layer (right)

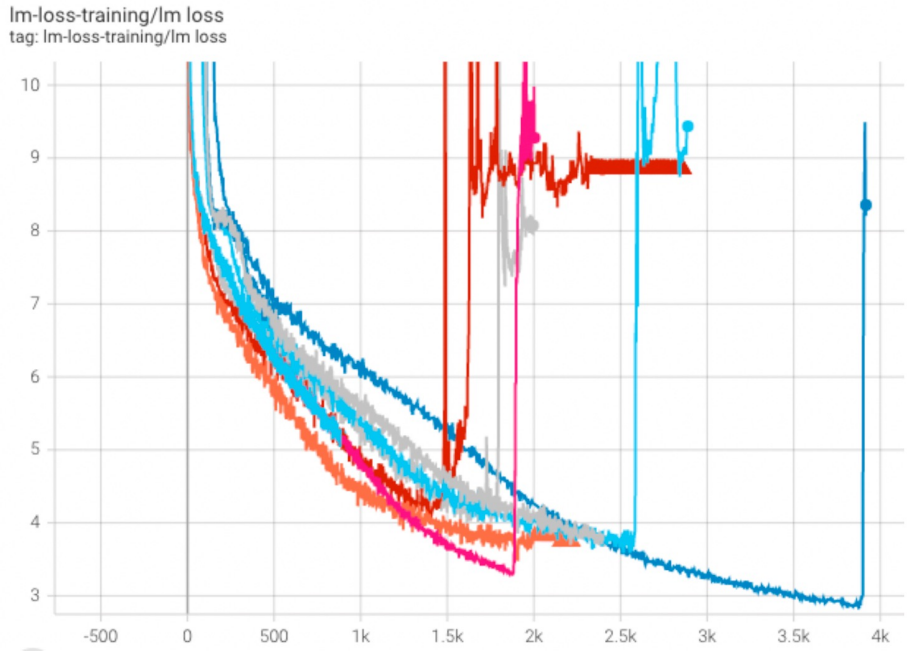


(b) Training loss curves of GLM-40B with and without gradient shrink

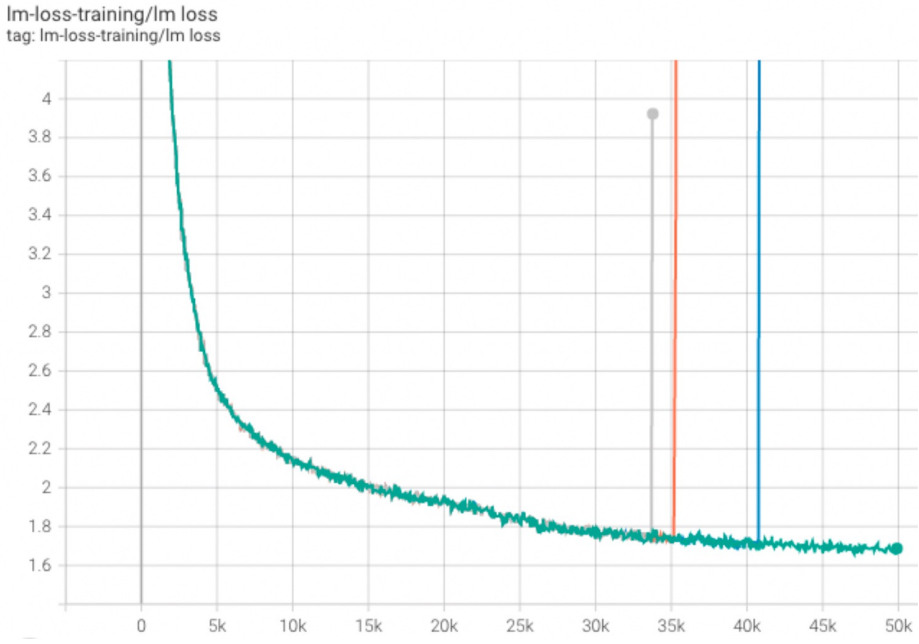
Zeng, Liu, et al. GLM-130B: An Open Bilingual Pre-trained Model. ICLR'23

GLM-130B：稳定训练方法

- GLM-130B 最终稳定、高效、高质量训练



(c) GLM 130B's experiments



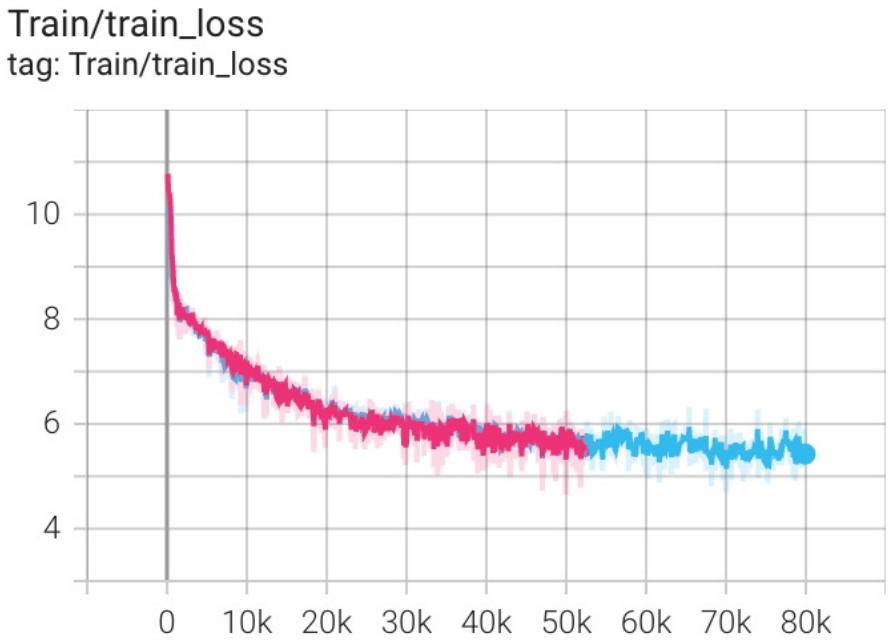
(d) GLM 130B's real training

Zeng, Liu, et al. GLM-130B: An Open Bilingual Pre-trained Model. ICLR'23

跨平台高效训练千亿模型

- 跨平台兼容：swDeepSpeed 训练库 ⇔ 与 DeepSpeed API 兼容
  - 支持**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**三个不同架构集群**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
)
```



Zeng, Liu, et al. GLM-130B: An Open Bilingual Pre-trained Model. ICLR'23

GLM-130B：千亿模型之旅

Major Issues Encountered for Training GLM-130B

2021.12
- The "千亿" (100B) project towards an open dense pre-trained GLM at 100B scale is conceived
- Survey pre-training strategies of existing models of similar scale, such as GPT-3, Gopher *=> Limited public info about how they were trained and issues they met*
- Search for possible GPU clusters & sponsors

2022.1
- Test the performance of FP16/FP32 at 100B scale on one testing cluster
- Unexpected excessive memory usage in GLM *=> Torch is better with fixed length input sequences*
- Inability to converge and try tricks from CogView and ViT *=> Use Sandwich-LN*
- Frequent random hardware failures *=> Have to run HCPG test before each run*

2022.2
- Very slow training speed than previously calculated *=> Optimize kernels and fuse operators => Find the input shape is critical to kernel performance*
- Collect pre-training corpora and tokenize *=> Use icetk: the sentence piece is set to the unigram mode*
- Debug the 3D pipeline parallel in the newly-released Megatron and DeepSpeed

2022.3
- It can't recover perfectly from optimizer states *=> Our customized dataloaders do not save its state seed properly in distributed training*
- The memory per processor is too small *=> Require too many pipeline stages => Batch size is too large (up to 12,000) => Harm the model's convergency*
- It can't launch more than 2,000 computing nodes *=> Overcome this and support 6,000-node training by tuning Linux kernel TCP parameters*
- Collect data for multi-task instruction pre-training
- Receive opportunities to test trainings on several other clusters
- Very slow training speed than expected *=> The underlying element-wise operators don't support fast computation on large-dimension vectors.*

2022.4
- Optimize A100 kernel's computing efficiency *=> A100 kernels prefer square-shaped inputs, and seq_len=2,048 is optimal for our hidden-state dimension (12,288)*
- Inability to converge due to large gradient norms (170+) of input embeddings *=> Try embedding norm and gradient shrink, which turn out to be almost equivalent*
- Naïve post-LN or pre-LN disconverges after several thousands of steps *=> Try Sandwich-LN with PB-Relax*
- It still disconverges after one week's trial *=> The dataloader state seeds are not unified for different pipeline stages, resulting in a mismatch of input data and labels.*
- Test two positional encodings: RoPE and Alibi *=> Alibi can be slower as it requires element-wise manipulation on attention matrices---changing num_heads *2,048 * 2,048 scalars per layer*
- Test GeGLU and GAU *=> GAU converges faster with relatively poor performance on fine-tuned SuperGLUE*
- Abnormal GPU memory usage of newly-added functions and classes *=> DeepSpeed hardcodes the function names for checkpoint activation*
- Decode to train GLM with 130 billion parameters *=> allow inference on a DGX-A100 40G node*

2022.5-6
- Implement a RoPE cuda operator in C++ *=> See unexpected precision errors and finally have it abandoned*
- Sandwich-LN still disconverges *=> 1) Reducing learning rate does not help; 2) Using Hinge cross-entropy becomes slower and harms performance; 3) Shifting to DeepNorm still disconverges*
- Use FP32 in softmax of attention *=> Success*
- Find PB-Relax unnecessary for FP32 softmax *=> It also slows down training as it needs to manipulate the whole attention score matrices*
- Experience few spikes in later training *=> 1) Reduce gradient shrink factor from 1 to 0.1: useful; 2) Reduce the learning rate: sometimes useful; 3) Jump the noisy data batches: sometimes useful*
- Find a mistake in multi-task data after training for 20,000 steps *=> Use the correct data but it does not forget*

2022.6-7
- Adapt the pipeline parallel checkpoints to ordinary parallel checkpoints for efficient inference on a single A100
- Work on evaluation scripts on datasets: MMLU, Big-bench, CLUE, SuperCLUE, etc.
- Implement P-Tuning and P-Tuning v2 for parameter-efficient tuning on GLM-130B for tuning on SuperGLUE
- Work with BMInf on adapting GLM-130B to perform inference on a single V100 or 3090 *=> Use pipeline-style asynchronous swapping between main memory and GPU memory*
- Try to fine-tune GLM-130B with fewer A100 nodes (i.e., 12-16 nodes) *=> Pipeline-style fails due to too many pipeline stages => Find that data parallel can not be introduced for fine-tuning => Use 32-way model parallel fo fine-tuning with reasonable performance*

https://github.com/THUDM/GLM-130B

Tsinghua KEG

GLM-130B：千亿模型之旅

Lesson 1 (Bidirectional Architecture).
The bidirectional-attention GLM is a strong architecture alternative, in addition to GPTs.

Lesson 2 (Platform-aware Configuration).
Configure LLMs based on the cluster and parallel strategy used to squeeze hardware potential.

Lesson 3 (Improved Post-LN).
Counter-stereotypically, DeepNorm, a type of Post-LN, is the option to stabilize GLM-130B.

Lesson 4 (Training Stability Categorization).
Unexpected training instability that LLMs suffer from arouses systematically and numerically.

Lesson 5 (Systematical Instability: FP16).
Though FP16 induces more instability, it enables training and inference on diverse platforms.

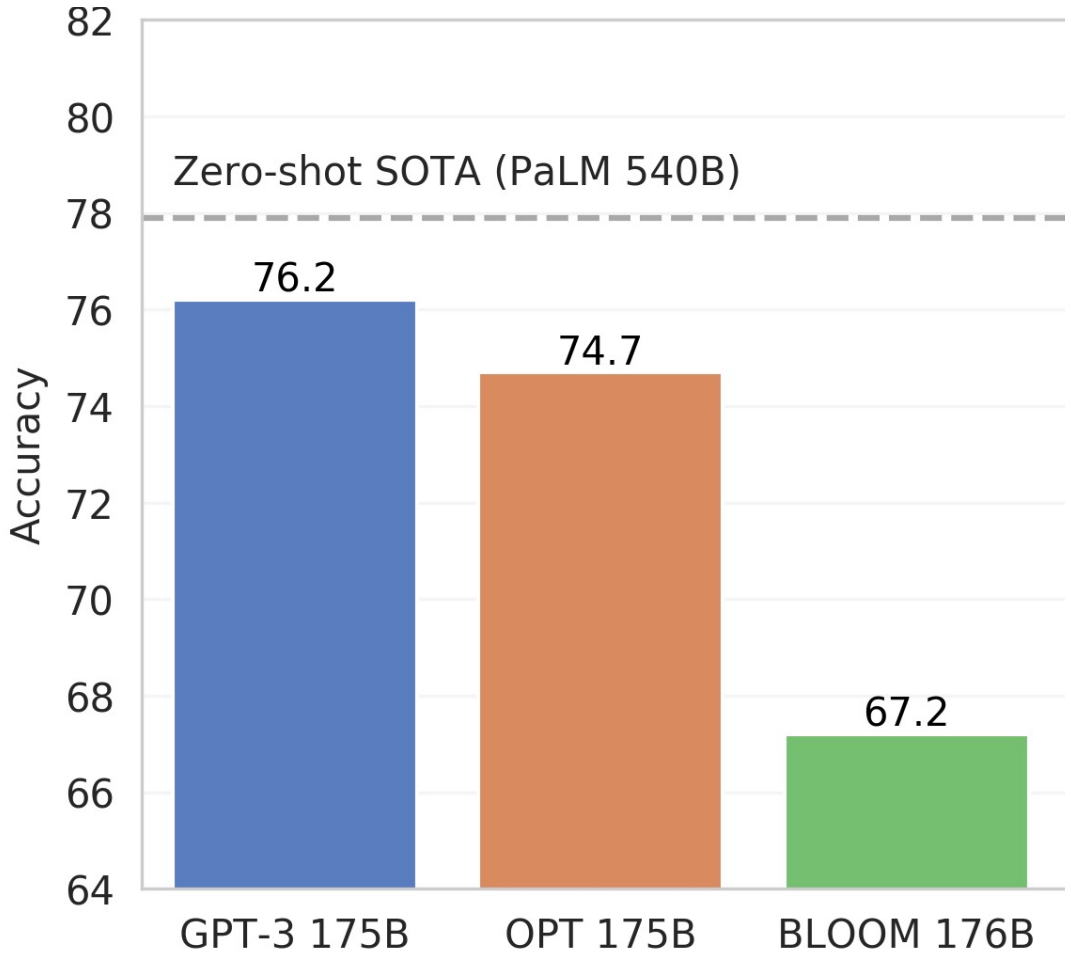
Lesson 6 (Numerical Instability: Embedding Gradient Shrink).
Shrinking embedding layer's gradient to its 0.1 can solve most numerical instability problems.

Lesson 7 (GLM's INT4 Quantization Scaling Law).
GLM has a unique INT4 weight quantization scaling law unobserved in GPT-style BLOOM.

Lesson 8 (Future Direction).
To create powerful LLMs, the main focus can be on 1) more and better data, 2) better architectures and pre-training objectives, and 3) more sufficient training.

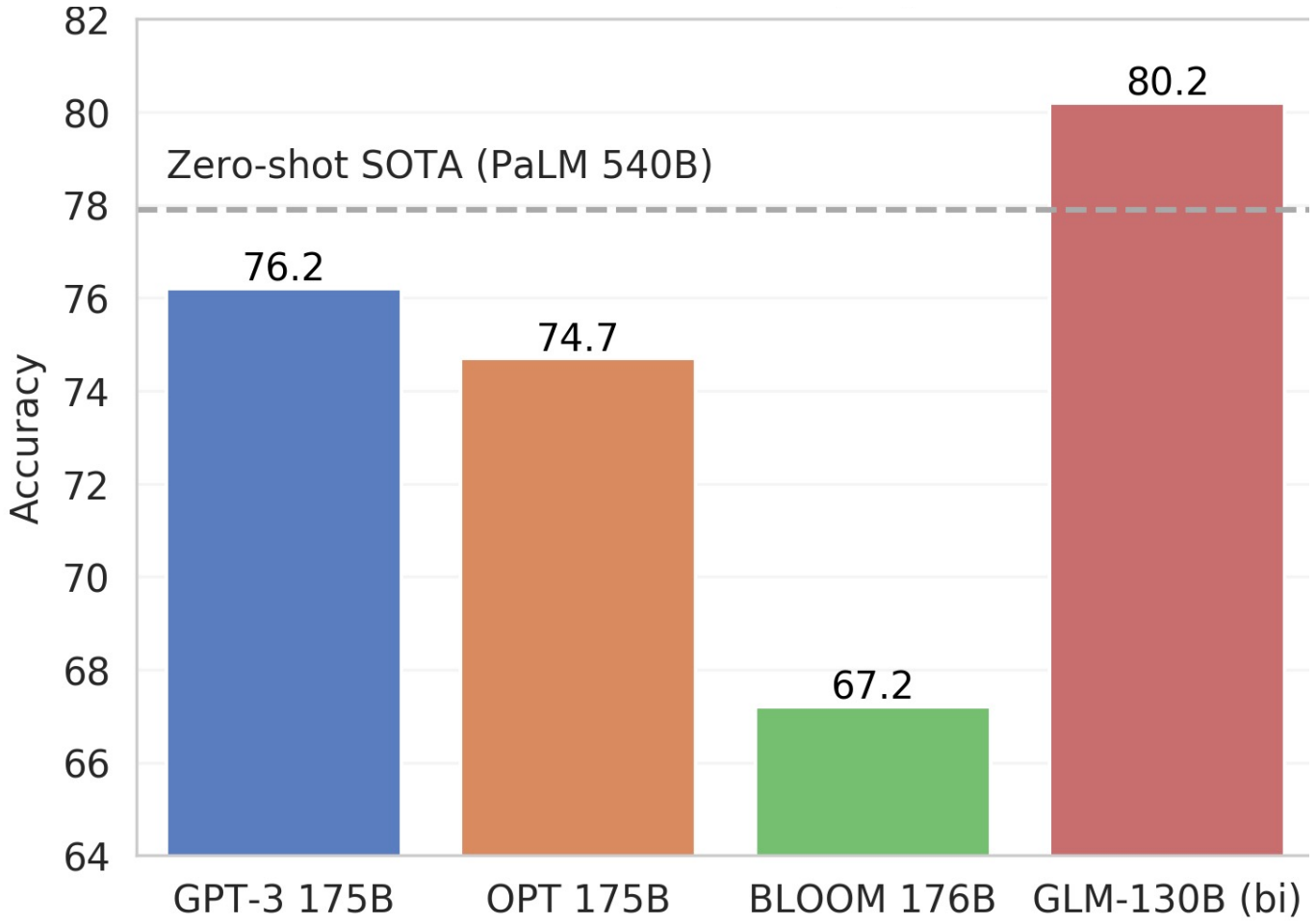
Zeng, Liu, et al. GLM-130B: An Open Bilingual Pre-trained Model. ICLR'23

LAMBADA (English)



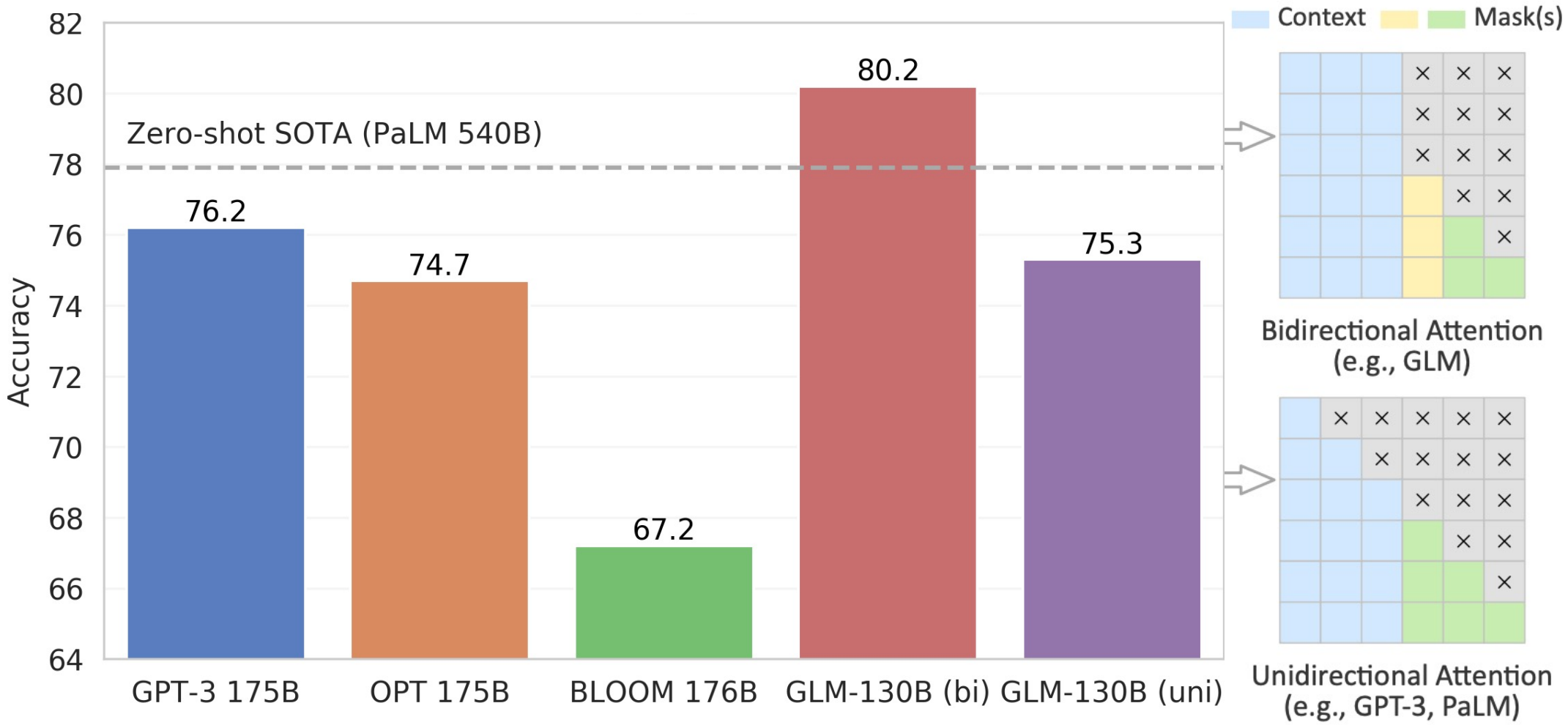
Zeng, Liu, et al. GLM-130B: An Open Bilingual Pre-trained Model. ICLR'23

LAMBADA (English)



Zeng, Liu, et al. GLM-130B: An Open Bilingual Pre-trained Model. ICLR'23

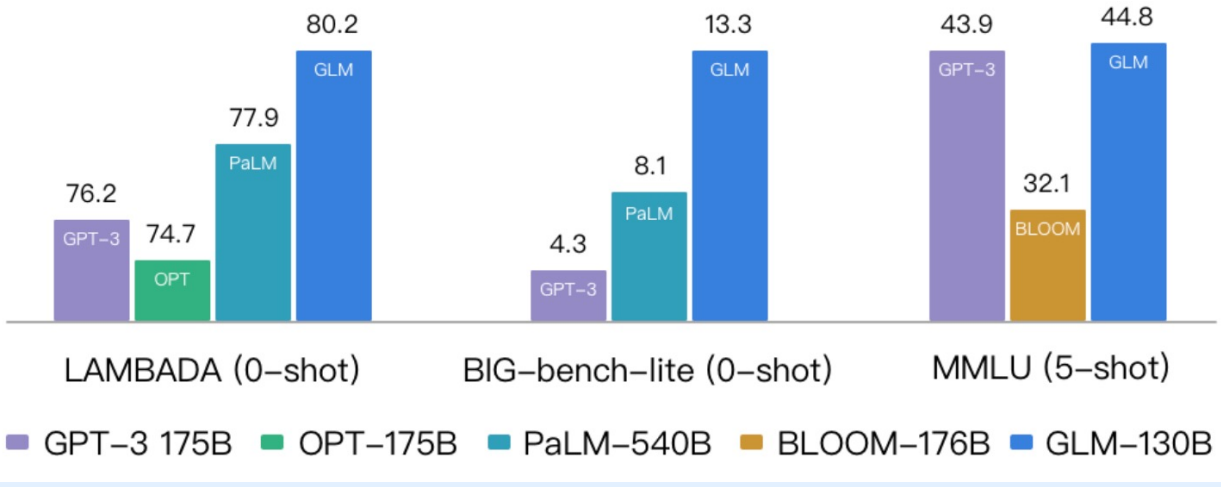
LAMBADA (English)



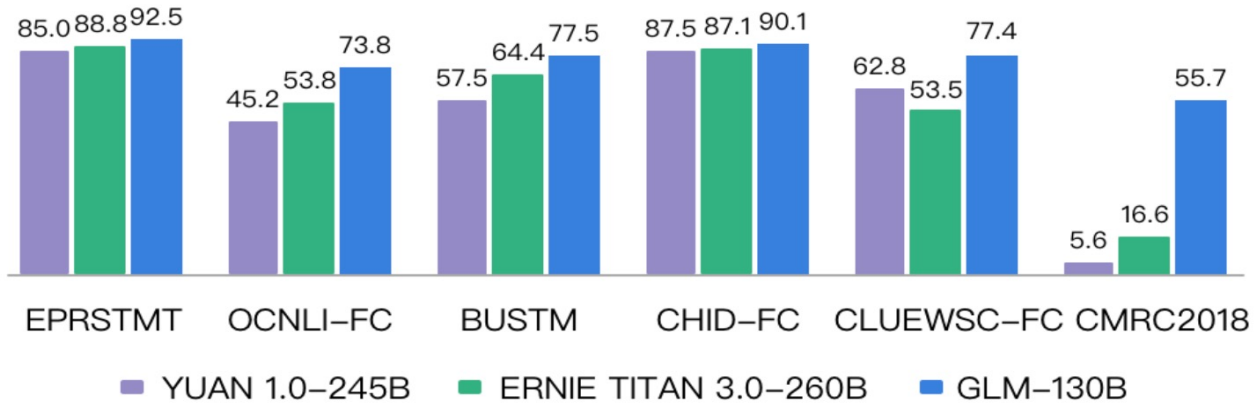
Zeng, Liu, et al. GLM-130B: An Open Bilingual Pre-trained Model. ICLR'23

千亿基座GLM-130B

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



中文：在 CLUE 和 FewCLUE 上取得不错的结果



2022.8---2023.3，收到60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

千亿基座GLM-130B

Stanford报告世界主流大模型评测：亚洲唯一入选模型

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

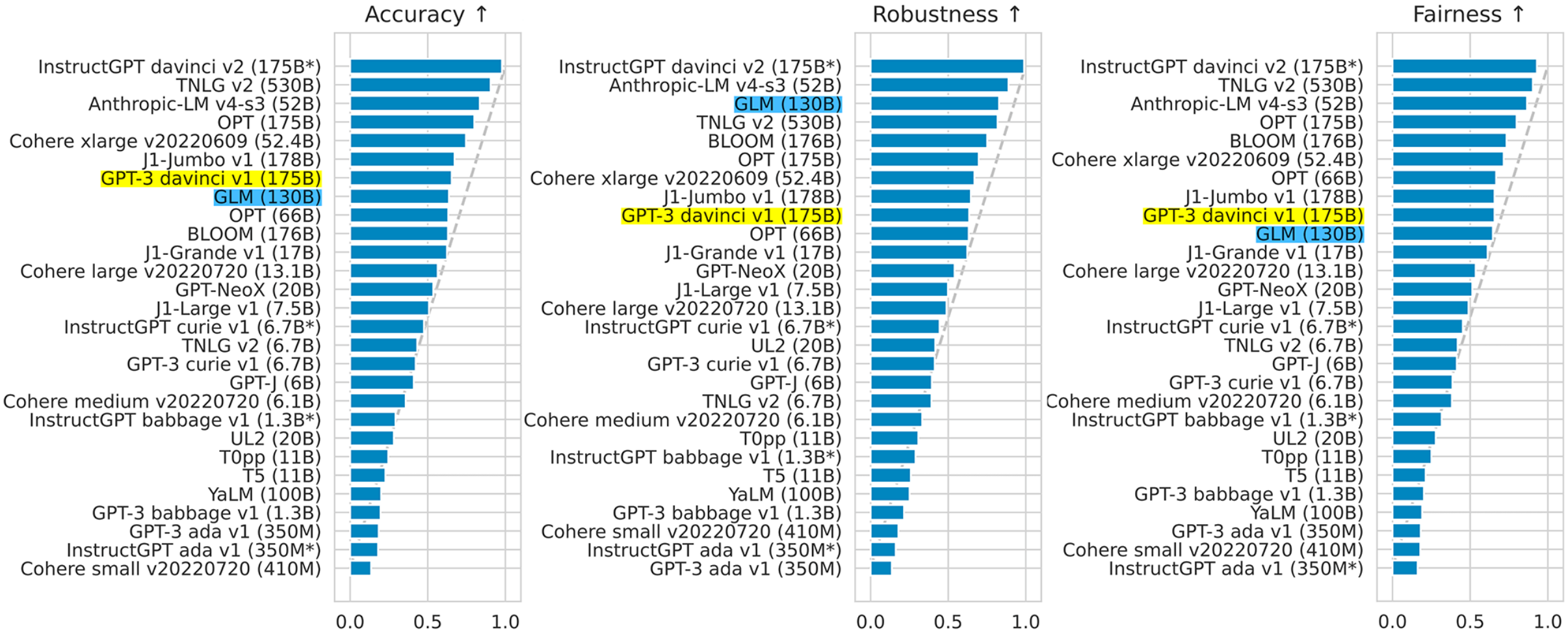
https://crfm.stanford.edu/helm, 2023.0308

Liang et al., Holistic Evaluation of Language Models. arXiv: **2211.09110**

HELM

千亿基座GLM-130B

准确性、恶意性等与GPT-3 (2020) 接近或持平，鲁棒性和校准误差在千亿模型中表现最佳

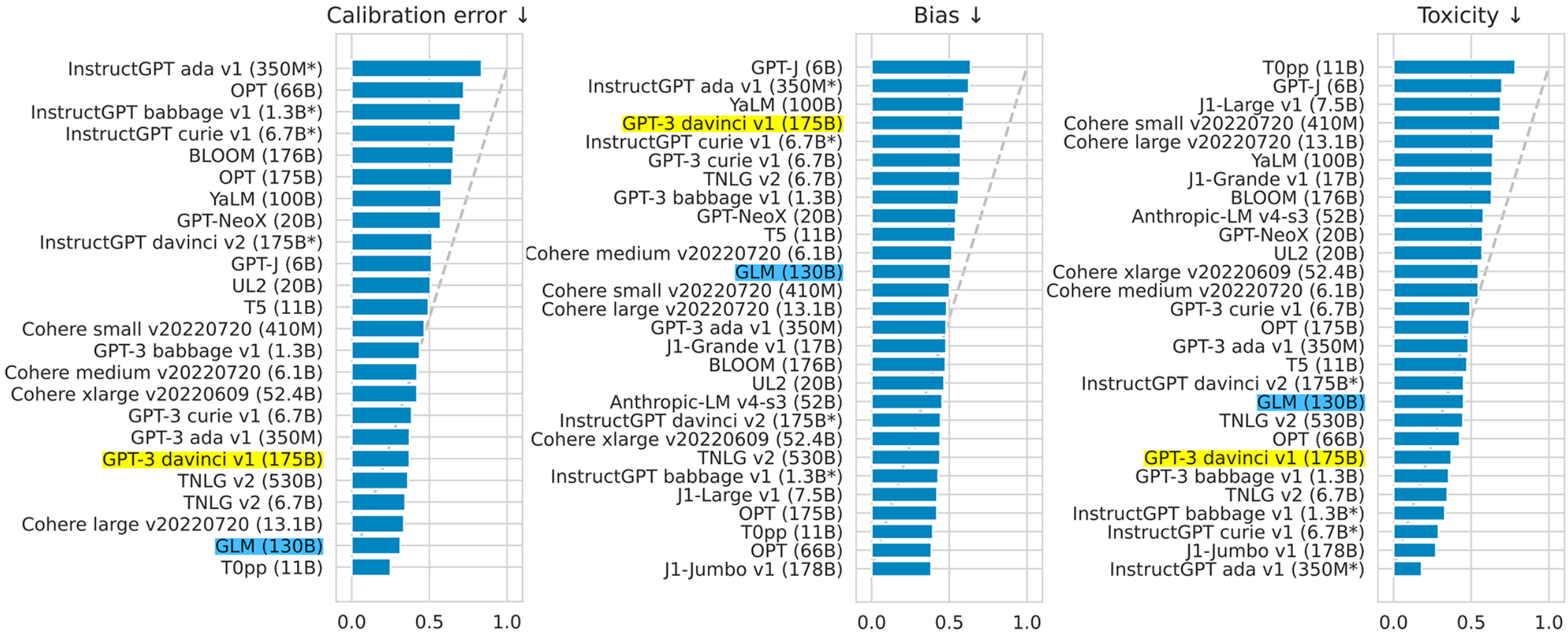


Liang et al., Holistic Evaluation of Language Models. arXiv: **2211.09110**

HELM

千亿基座GLM-130B

准确性、恶意性等与GPT-3 (2020) 接近或持平，鲁棒性和校准误差在千亿模型中表现最佳

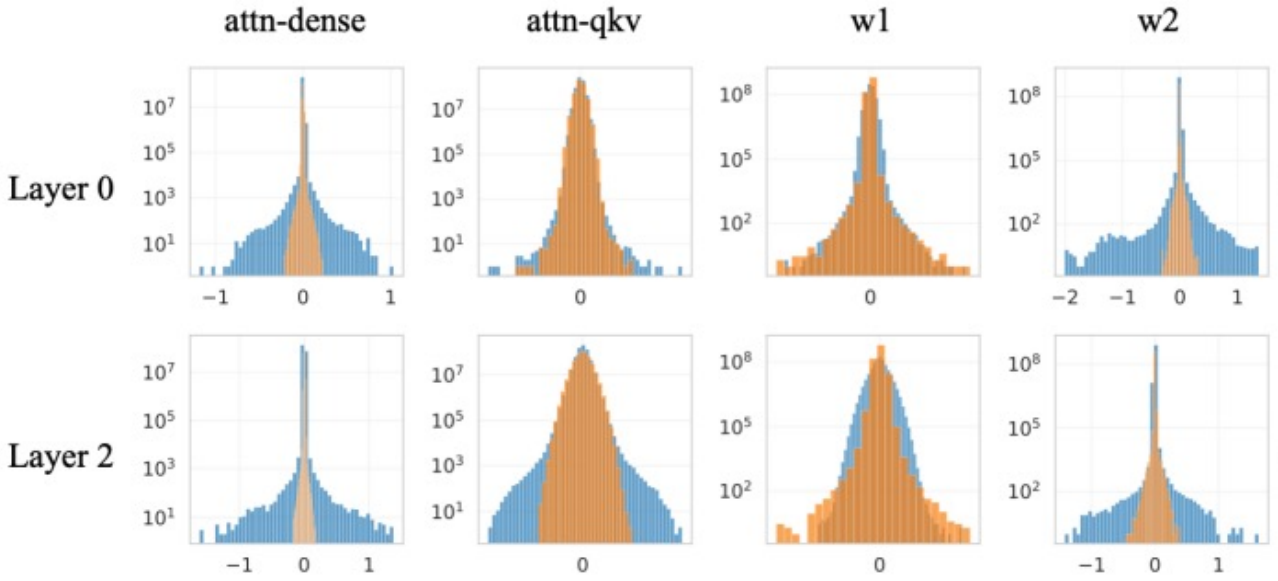


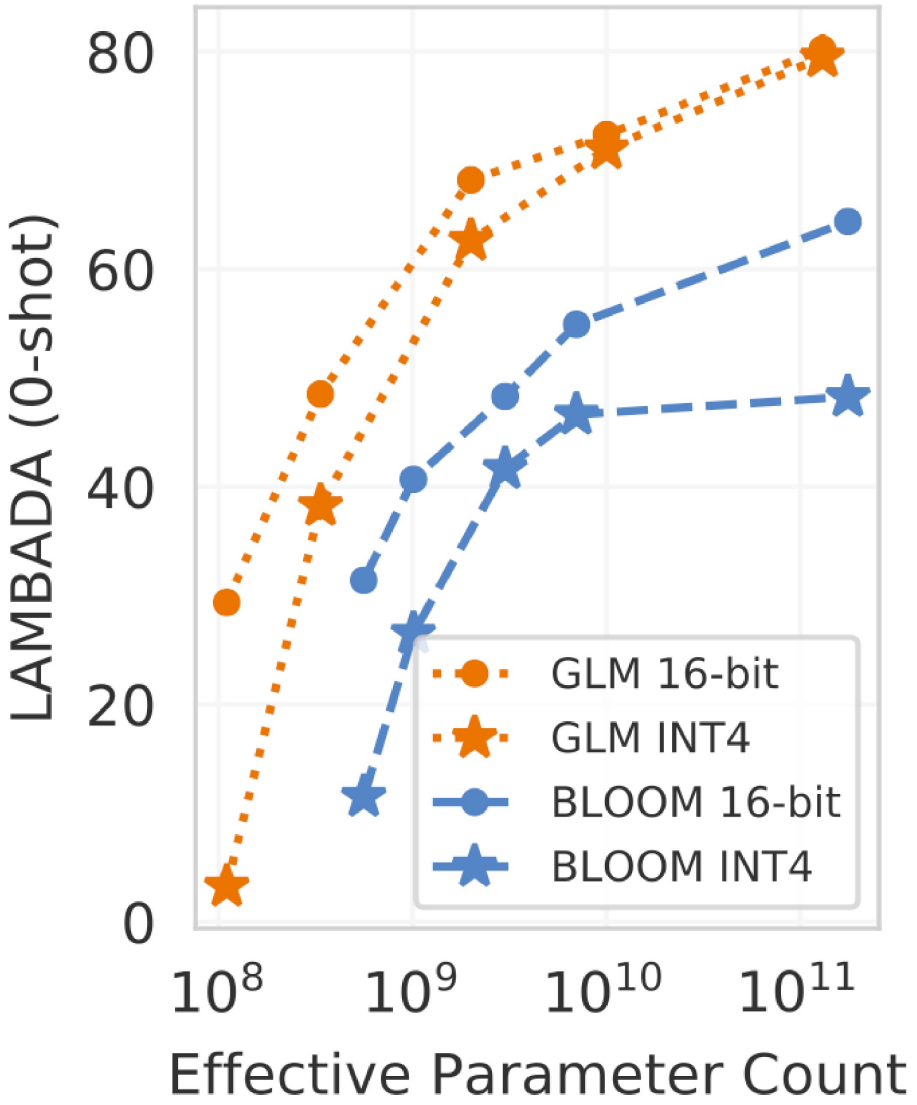
Liang et al., Holistic Evaluation of Language Models. arXiv: **2211.09110**

HELM

千亿模型推理：无损 INT4 量化

- GLM-130B的权重INT4量化"规模效应"
  - **蓝色**为BLOOM-176B，**橙色**为GLM-130B；
  - GLM-130B在attn-dense和ffn-w2具有更窄的数值分布：由于GLM本身的训练目标造成，而非Transformer的架构选择。





Zeng, Liu, et al. GLM-130B: An Open Bilingual Pre-trained Model. ICLR'23

千亿模型推理：无损 INT4 量化

- GLM-130B的**无损INT4量化（显存降75%）**

Model Precision	GLM-130B			GPT-3
	FP16	INT8	INT4	FP16
MMLU (acc, ↑)	44.75	44.71	44.80	43.9
LAMBADA (acc, ↑)	80.21	80.21	79.47	76.2
Pile (a part, BPB, ↓)	0.634	0.638	0.641	0.74

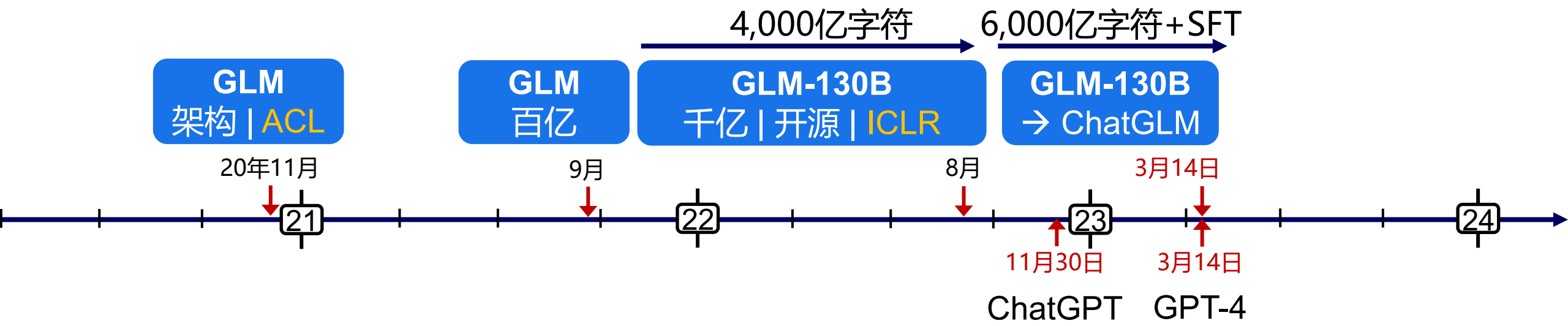
GPU Type	128 Enc./Dec.		512 Enc./Dec,	
8 × A100 (40G)	0.15s	4.29s	0.18s	17.7s
8 × V100 (32G)	0.31s	6.97s	0.67s	28.1s
4 × RTX 3090 (24G)	0.37s	8.16s	1.30s	32.3s
8 × RTX 2080 Ti (11G)	0.39s	6.77s	1.04s	27.3s

Zeng, Liu, et al. GLM-130B: An Open Bilingual Pre-trained Model. ICLR'23

千亿基座GLM-130B

	基础架构	训练方式	量化	加速	跨平台能力
GPT3-175B	GPT	自监督预训练	—	—	NVIDIA
OPT-175B	GPT	自监督预训练	INT8	Megatron	NVIDIA
BLOOM-176B	GPT	自监督预训练	INT8	Megatron	NVIDIA
GLM-130B	**GLM**	**自监督预训练和多任务预训练**	**• INT8 • INT4**	**Faster Transformer**	**• NVIDIA • 海光 DCU • 昇腾910 • 申威**
对比优势	**高精度：** • Big-bench-lite: +5.2% • LAMBADA: +2.3% • CLUE: +24.3% • FewCLUE: +12.8%		**普惠推理：** **节省75%内存** 可单台3090 (4) 或单台2080Ti (8) 进行**无损**推理	**高速推理：** 比Pytorch **提速7-8.4倍** Megatron **提速2.5倍**	**跨平台：** 支持更多不同的大规模语言模型的适配

清华&智谱 GLM 系列模型



chatglm.cn


ChatGLM 内测版
我们一直在路上
ChatGLM 内测版 千亿对话模型
旨在推动大模型的科学研究，面向高校学术交流与行业合作的
试用申请
内测申请
已通过审核的用户请 登录体验 ›
screenshot at 2023.05.28

ChatGLM 内测

创建新对话

分析品牌营销领域头部的创新者，并列...
2023-07-05 20:38:58

帮忙写个给猫咪的道歉信，原因是我忘...
2023-07-05 20:37:58

一个包子蒸熟要10分钟，如果我有4个蒸...
2023-06-29 16:44:16

如何建设全国统一大市场
2023-06-29 16:22:59

孙悟空有没有火烧赤壁
2023-06-29 15:55:32

一个包子蒸熟要10分钟，如果我有4个蒸...
2023-06-29 15:55:10

帮我写一个日文公文，内容为今日金价...
2023-06-29 15:53:06

一个包子蒸熟要10分钟，如果我有4个蒸...
2023-06-29 15:51:51

孙悟空有没有火烧赤壁
2023-06-29 15:30:05

您好，我是ChatGLM

我是一个中英双语大模型，通过万亿字符的文本与代码预训练，结合有监督微调，我具备丰富的能力，包括通用问答、多轮对话、角色扮演、文本生成、代码生成等功能。我能为各种场景下的需求提供多维度、多形式的解决方案，期待与您共同探索更多有趣的玩法。

试试以下例子：

工作总结

帮我写一份年终工作总结，夸夸我的优点并展望明年。

道歉信

帮忙写个给猫咪的道歉信，原因是我忘了给它铲屎。

正则

请提供一段JS正则代码，确认是否是真实手机号

提出你的想法

灵感大全

全部 | 创作 | 学习 | 职场 | 编程 | cosplay | 教育 | 公文 | 论文

广告文案
请帮我为宝宝乐安抚奶嘴生成3个条广告文案，每条字数50字以内

即兴发言
以保护环境为主题，帮我生成一份在科技发展论坛上的即兴发言

开题报告
依据我的研究方向：大模型与传统NLP，帮我输出一份开题报告

JD生成
我想招聘一个产品经理,工作内容是类chatgpt对话产品数据分析，请帮我生成一个招聘JD

语言翻译
你是一名资深的翻译，请将条条大路通罗马翻译成英文

写作文
请以荷塘月色为题目，帮我撰写一份五年级600字的作文

辩题立论
请帮我基于愚公应该搬家这个观点的持方生成一份立论

历史名人
现在你是孔子，模仿他说话的语气和风格与我对话

辅助决策
我正在决策是继续读研还是先参加工作，给我一个利弊清单，帮我决定我应该如何选择

简历润色
依据我的工作内容H5开发、小程序优化帮助我润色简历中的项目经验

视频脚本
帮我生成一个展现北京历史文化和科技发展的短视频拍摄脚本

专业计划
你是一个NBA资深教练，请为我制定一个篮球基本功15天计划

演讲稿
请生成演讲稿来介绍一款演讲稿生成工具的功能和使用方法。

PPT大纲
帮我生成一份户外运动发展趋势的调研汇报PPT大纲，并提供每

ChatGLM：常识

孙悟空有没有火烧赤壁

孙悟空是中国古代小说《西游记》中的重要人物，他是一位神通广大的猴子，经过修行后成为了斗战胜佛。在小说中，孙悟空并没有参与火烧赤壁的战斗。

火烧赤壁是中国古代历史上著名的战役之一，发生在公元 208 年，是刘备、孙权联合对抗曹操的军队而形成的三方对峙局面。在战役中，孙刘联军采取水战，用火攻来攻破曹军。曹操的军队在火势蔓延之下，陷入混乱，最终溃败。

虽然孙悟空没有在火烧赤壁中战斗，但他在小说中参与了其他许多重要的战役和冒险，最终成为了斗战胜佛，成为了中国文学和艺术中的重要形象之一。

以上内容为 AI 生成，不代表开发者立场，请勿删除或修改本标记

复制

ChatGLM：常识

我的姥姥的闺女是我的什么人

你姥姥的闺女是你的妈妈或者姨妈。

以上内容为 AI 生成，不代表开发者立场，请勿删除或修改本标记

复制

我奶奶的闺女呢

你奶奶的闺女是你的姑姑。

以上内容为 AI 生成，不代表开发者立场，请勿删除或修改本标记

复制

奶奶的儿子呢

你奶奶的儿子是你的父亲或者叔叔。

ChatGLM：常识

一条船上有三个人，但是有两个父亲，两个儿子，请问是怎么回事

这个谜语的答案是：这条船上共有三个人，分别是爷爷、爸爸和儿子。爸爸是爷爷的儿子，也是儿子的父亲，因此有两个父亲和两个儿子。

以上内容为 AI 生成，不代表开发者立场，请勿删除或修改本标记

复制

ChatGLM：写代码

用python帮忙写个冒泡排序算法的代码，再给一个数组作为例子，写个测试程序

以下是冒泡排序算法的 Python 代码实现：

```python
def bubble_sort(arr):
    n = len(arr)
    # 遍历所有数组元素
    for i in range(n):
        # 最后 i 个元素已经排好序，不需要再比较
        for j in range(0, n-i-1):
            # 如果当前元素比下一个元素大，则交换它们的位置
            if arr[j] > arr[j+1]:
                arr[j], arr[j+1] = arr[j+1], arr[j]
    return arr
```

下面是测试程序，它将输入一个数组，并使用冒泡排序算法对其进行排序，然后输出排序后的数组：

```python
arr = [64, 34, 25, 12, 22, 11, 90]
sorted_arr = bubble_sort(arr)
print(sorted_arr)
```

输出结果如下：

```
[11, 12, 22, 25, 34, 64, 90]
```

这是一个基本的冒泡排序算法，可以根据需求对其进行修改和扩展。

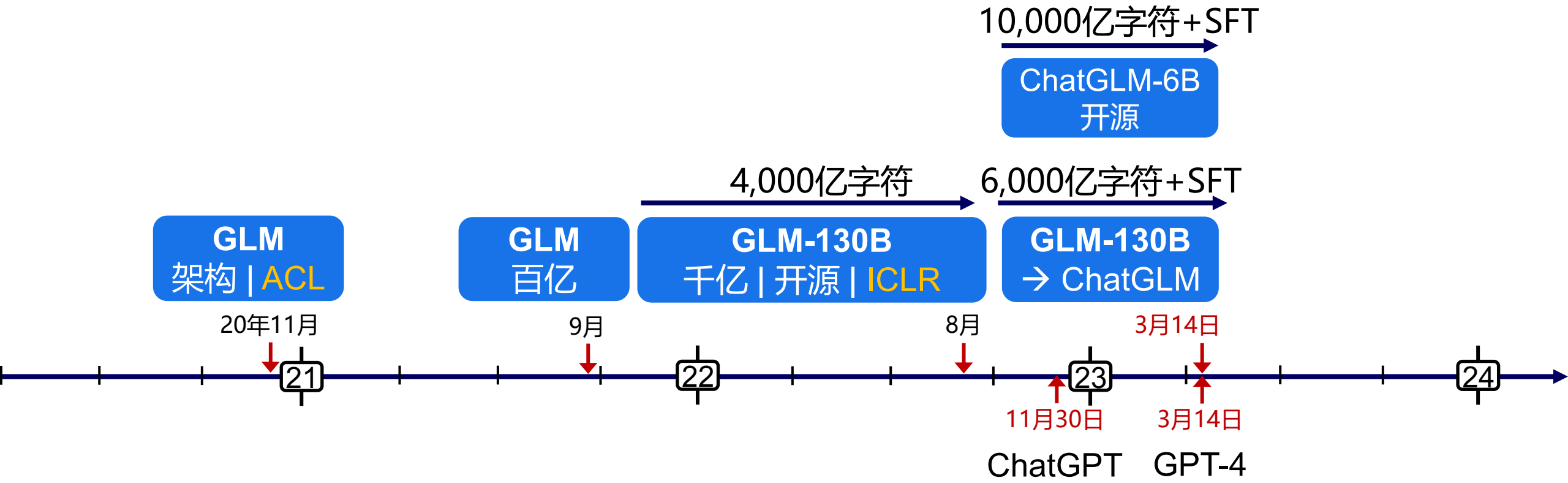
Z-Bench 和第三方开放评测结果

(榜单晚于被测试版本的模型)

	GPT3.5	GPT4	国内大厂	ChatGLM
基础能力	59/63	60/63	26/63	40/63
进阶能力	120/209	159/209	41/209	74(+5)/209
垂直领域	20/39	26/39	11/39	11(+5)/39
总	199/311	245/311	78/311	125(+10)/311
折合得分	64.0	78.8	25.1	40.2(+3)

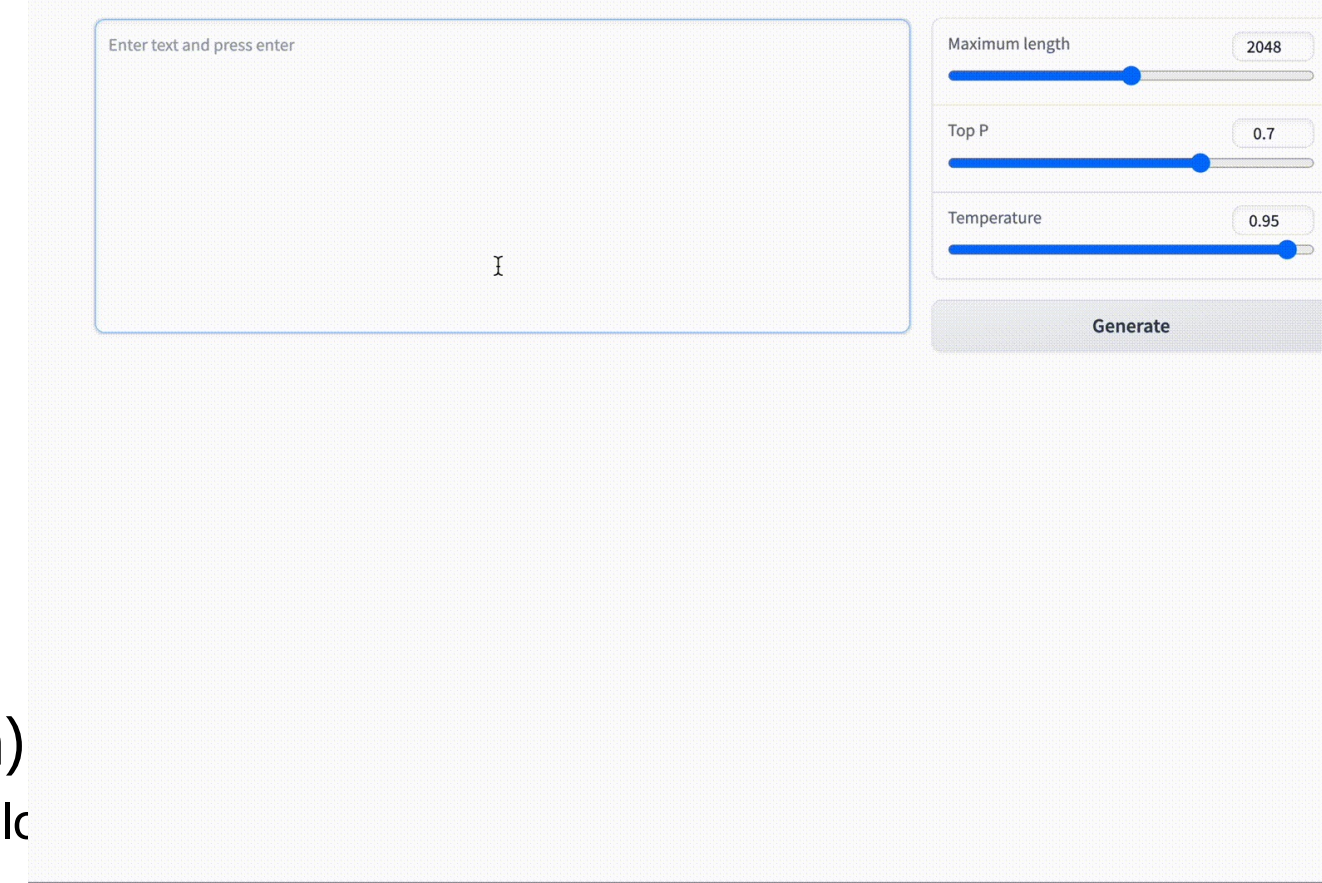
2023.03.20

清华&智谱 GLM 系列模型



在自己电脑上安装 ChatGLM-6B

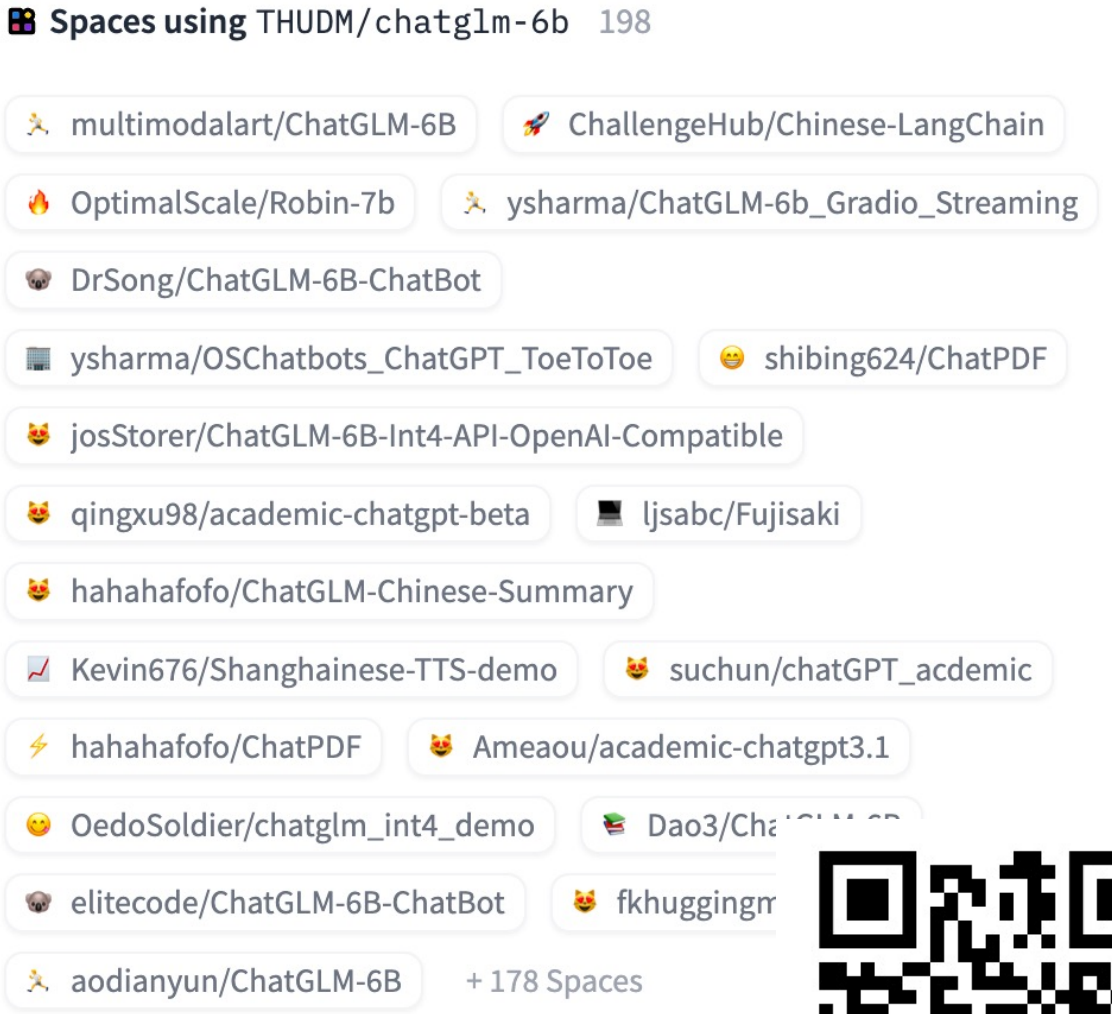
1. Download all model files from Huggingface
    1. git clone https://huggingface.co/THUDM/chatglm-6b
2. Download the demo
    1. git clone https://github.com/THUDM/ChatGLM-6B
    2. cd ChatGLM-6B
3. Install the demo
    1. pip install gradio
    2. python web_demo.py
4. Interactive demo
    1. python cli_demo.py
5. Install the api
    1. pip install fastapi uvicorn
    2. python api.py
6. Run on your MAC (w/ Apple Silicon)
    1. model = AutoModel.from_pretrained("your lo



在自己电脑上安装 ChatGLM-6B

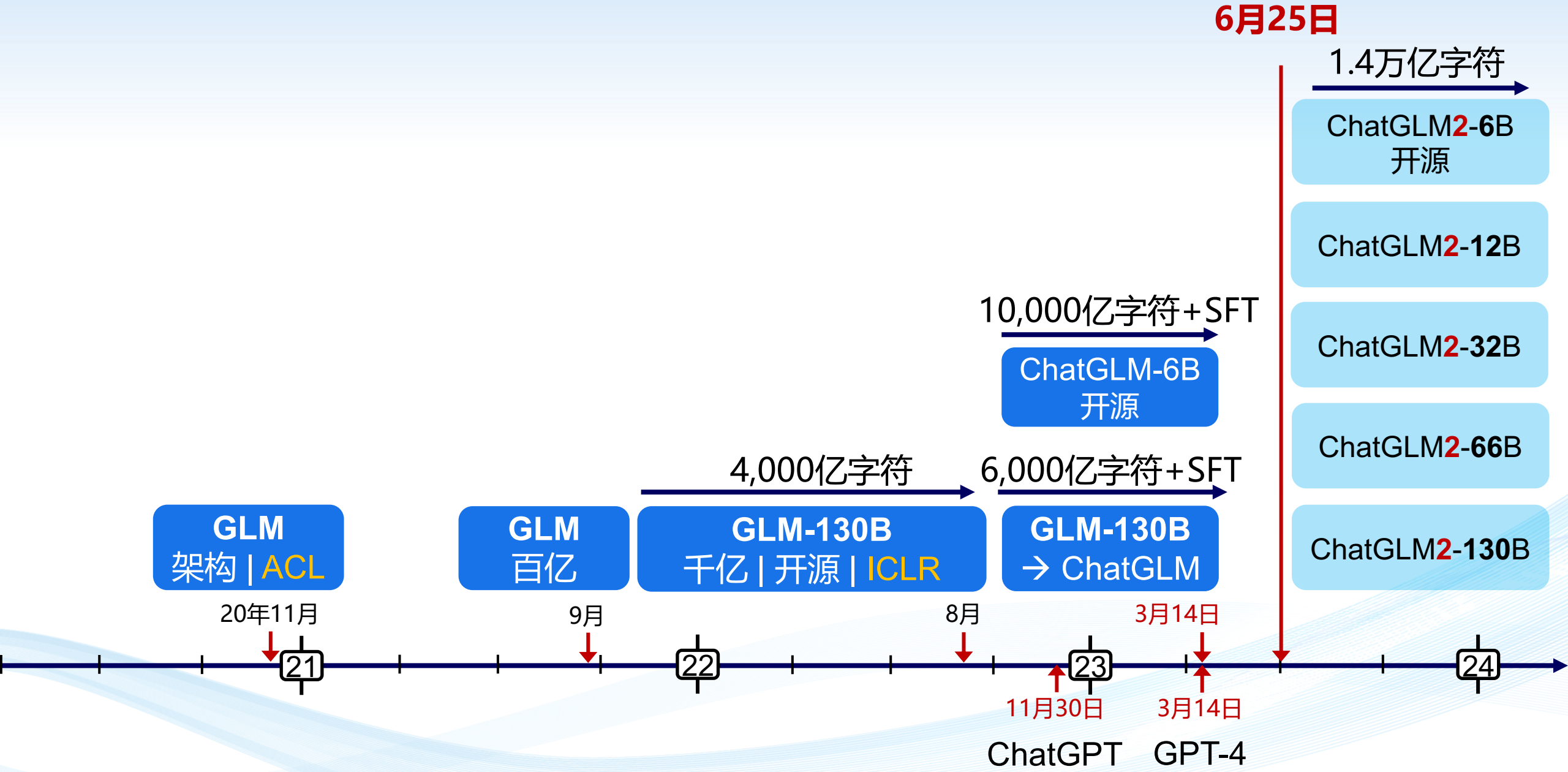
- **ChatGLM-6B**
  - 6B parameters
  - 1T tokens training data
  - 6G GPU mem (INT4)
- Mar. 14, 2023, open-soured model
- Mar. 16, 2023, **#1** on GitHub Trending
- Mar. 18-30, **#1** on HF Trending
- Jun. 23, 2023, **3M** downloads in HF

30k stars on GitHub



https://huggingface.co/THUDM

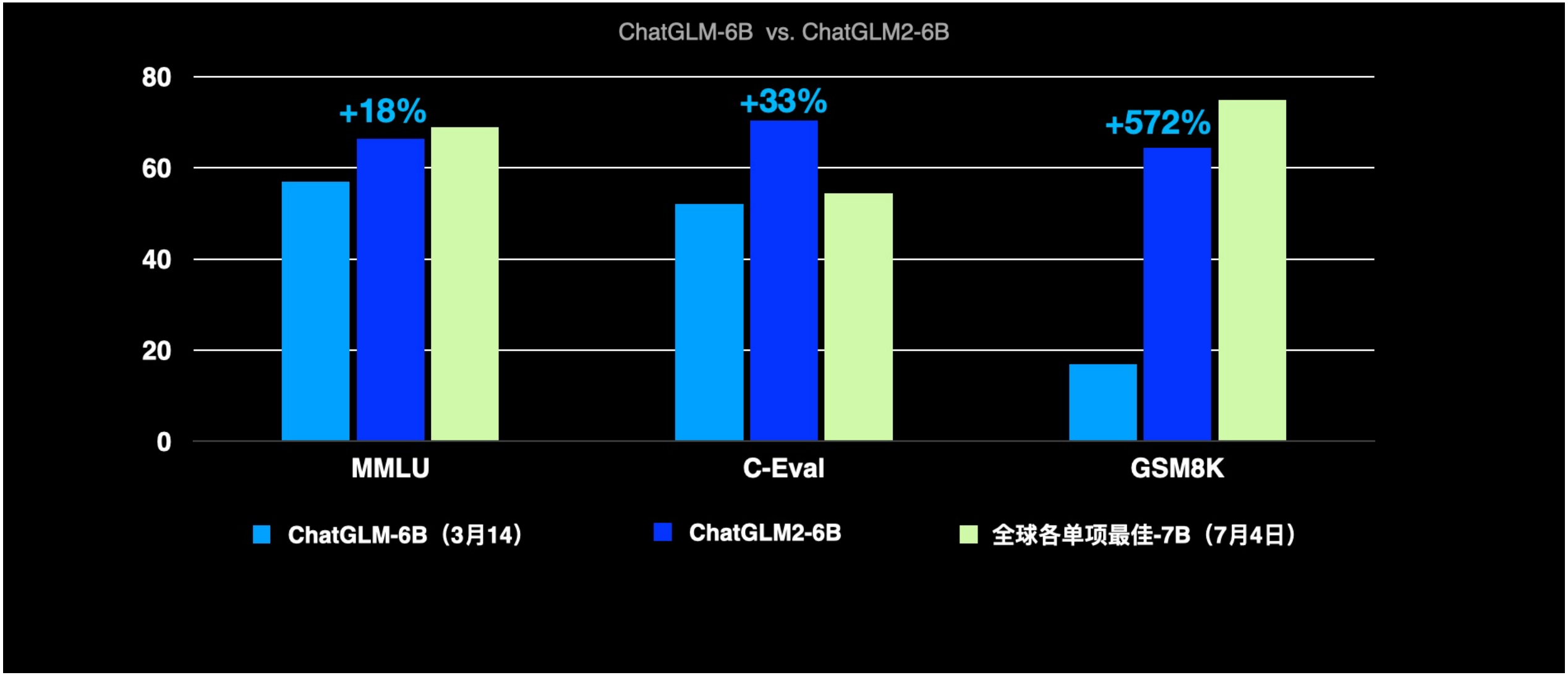
清华&智谱 GLM 系列模型



ChatGLM2-6B

- **更强大的性能**：
  - **1.4T** 中英标识符的预训练与人类偏好对齐训练；
  - MMLU（+18%）、CEval（+33%）、GSM8K（+572%） 、BBH（+60%）。
- **更长的上下文**：
  - 基座模型的上下文长度（Context Length）由 ChatGLM-6B 的 2K 扩展到了 32K;
  - 对话阶段使用 **8K** 的上下文长度训练，允许更多轮次的对话。
- **更高效的推理**：
  - 推理速度相比初代提升了 **42%**；
  - INT4 量化下，6G 显存支持的对话长度由 1K 提升到了 **8K**。
- **更开放的协议**：
  - ChatGLM2-6B 权重对学术研究**完全开放**；
  - 在获得官方的书面许可后，亦**允许商业使用**。

ChatGLM-6B vs. ChatGLM2-6B



ChatGLM-6B → ChatGLM2-6B

- **ChatGLM-6B**
  - 6B parameters
  - 1T tokens training data
  - 6G GPU mem (INT4)
- Mar. 14, 2023, open-soured model
- Mar. 16, 2023, **#1** on GitHub Trending
- Mar. 18-30, **#1** on HF Trending
- Jun. 23, 2023, **3M** downloads in HF

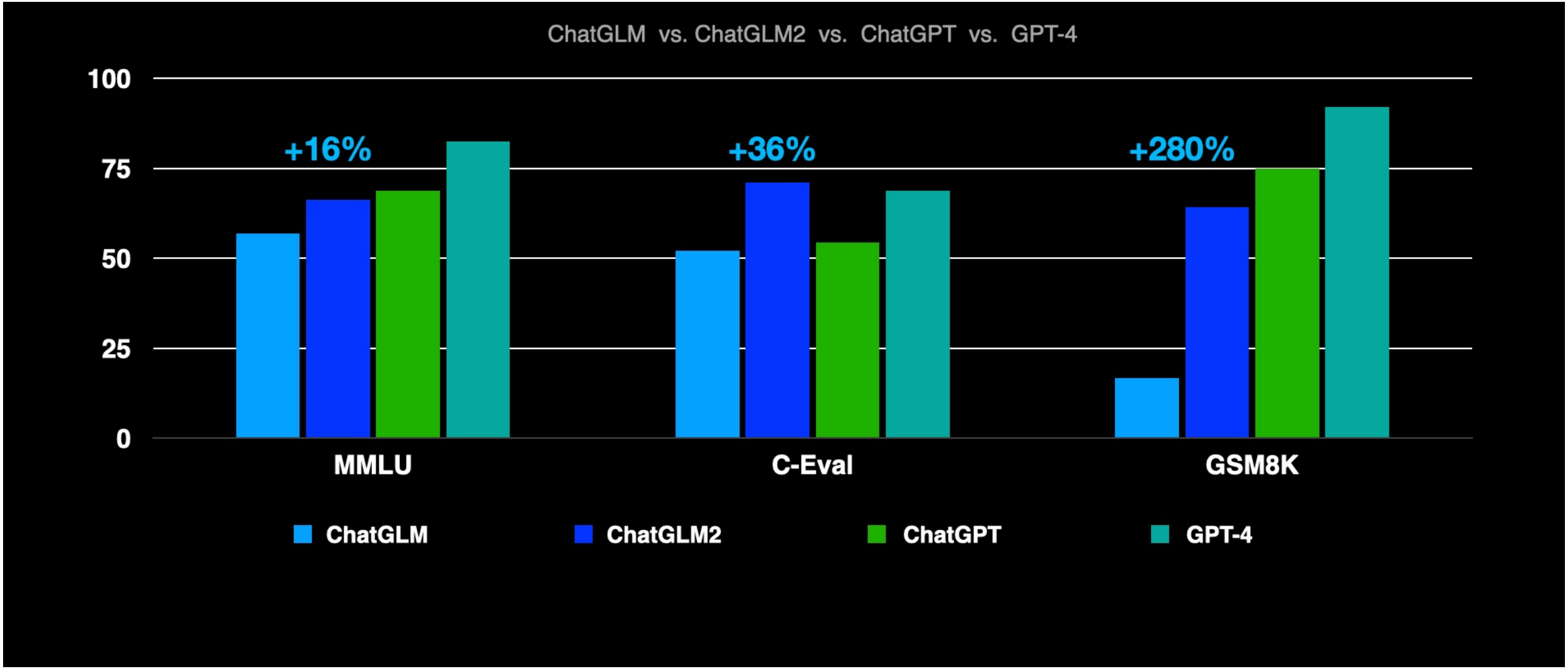
  30k stars on GitHub

- **ChatGLM2-6B**
  - 6B parameters
  - 1.4T tokens training data
  - 6G GPU mem (INT4)
- Jun. 25, 2023, open-soured model
- Jun. 27, 2023, **#2** on GitHub Trending
- Jun. 26- Jul. 8, **#1** on HF Trending
- Jul. 8, 2023, **880k** downloads in HF

  8k stars in GitHub

https://huggingface.co/THUDM

ChatGLM-130B vs. ChatGLM2



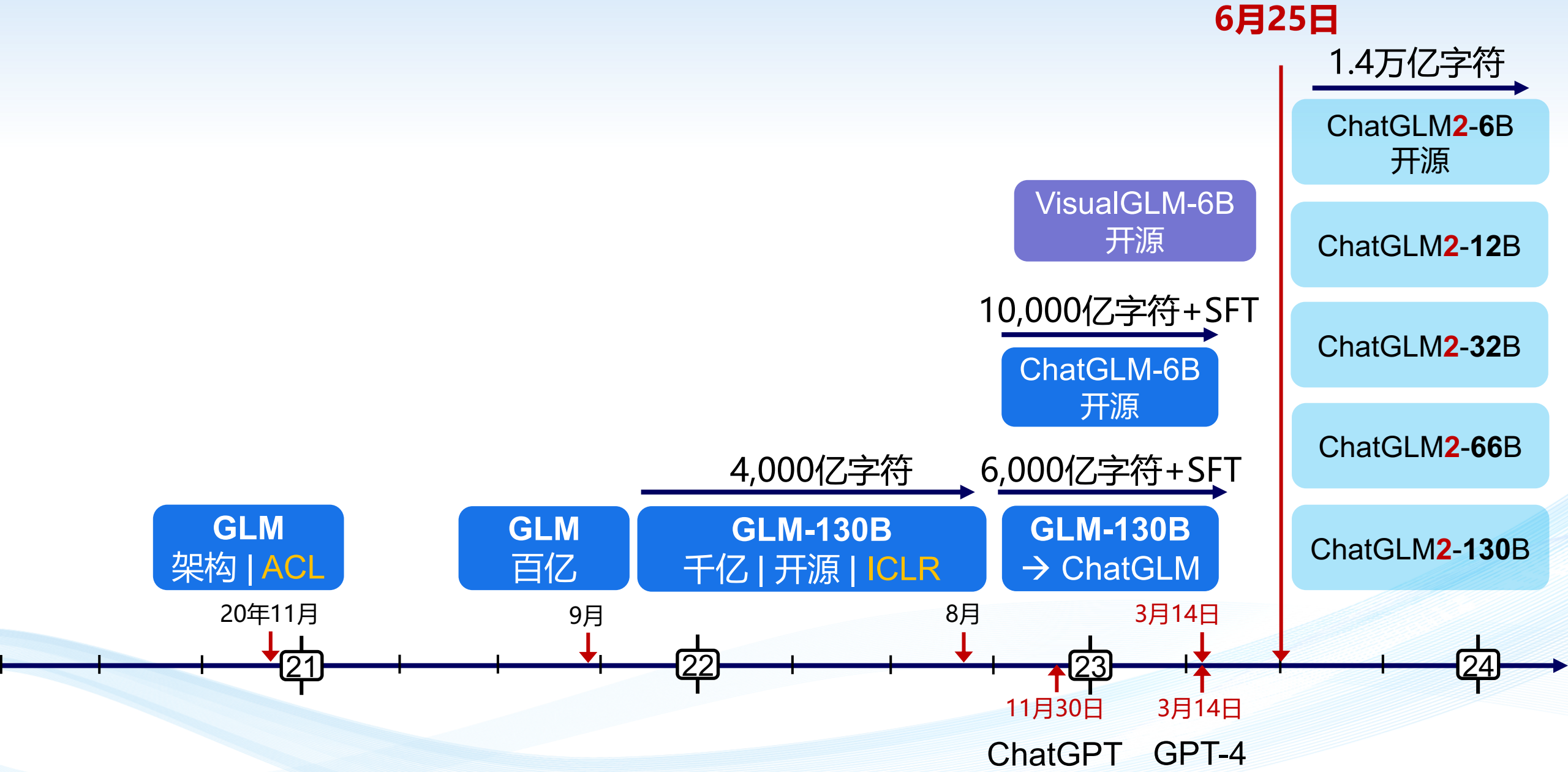
C-Eval

（榜单**早于**被测试版本的模型，GPT-4/ChatGPT/Claude除外）

#	Model	Creator	Submission Date	Avg ▾	Avg(Hard)	STEM	Social Science	Humanities	Others
0	ChatGLM2	Tsinghua & Zhipu.AI	2023/6/25	71.1	50	64.4	81.6	73.7	71.3
1	GPT-4*	OpenAI	2023/5/15	68.7	54.9	67.1	77.6	64.5	67.8
2	SenseChat	SenseTime	2023/6/20	66.1	45.1	58	78.4	67.2	68.8
3	InternLM	SenseTime & Shanghai AI Laboratory (equal contribution)	2023/6/1	62.7	46	58.1	76.7	64.6	56.4
4	ChatGPT*	OpenAI	2023/5/15	54.4	41.4	52.9	61.8	50.9	53.6
5	Claude-v1.3*	Anthropic	2023/5/15	54.2	39	51.9	61.7	52.1	53.7
6	ChatGLM2-6B	Tsinghua & Zhipu.AI	2023/6/24	51.7	37.1	48.6	60.5	51.3	49.8
7	SageGPT	4Paradigm Inc.	2023/6/21	49.1	39.1	46.6	54.6	45.8	51.8
8	AndesLM-13B	AndesLM	2023/6/18	46	29.7	38.1	61	51	41.9
9	Claude-instant-v1.0*	Anthropic	2023/5/15	45.9	35.5	43.1	53.8	44.2	45.4
10	WestlakeLM-19B	Westlake University and Westlake Xinchen (Scietrain)	2023/6/18	44.6	34.9	41.6	51	44.3	44.5
11	玉言	Fuxi AI Lab, NetEase	2023/6/20	44.3	30.6	39.2	54.5	46.4	42.2
12	bloomz-mt-176B*	BigScience	2023/5/15	44.3	30.8	39	53	47.7	42.7

Screenshot 2023, 0626

清华&智谱 GLM 系列模型



VisualGLM-6B=ChatGLM-6B + Visual

描述一下这个场景

泰坦尼克号，杰克和露丝在船上拥抱。
夕阳余晖下，他们的身影如此美丽动人。
海水翻涌，海浪拍击着船帆。
他们的爱如同这海洋般深邃而浩瀚。 在这首诗中，我们感受到爱情的深沉与美好。
他们在彼此的怀抱里，享受着爱情的温暖。
他们的爱情就像那大海一般广阔无垠，永恒不变。

这部电影的导演是谁?

这部电影的导演是詹姆斯·卡梅隆(James Cameron)。

这是什么东西

这张图片展现了一只可爱的卡通羊驼，它站在一个透明的背景上。这只羊驼长着一张毛茸茸的耳朵和一双大大的眼睛，它的身体是白色的，带有棕色斑点。

它头上有什么

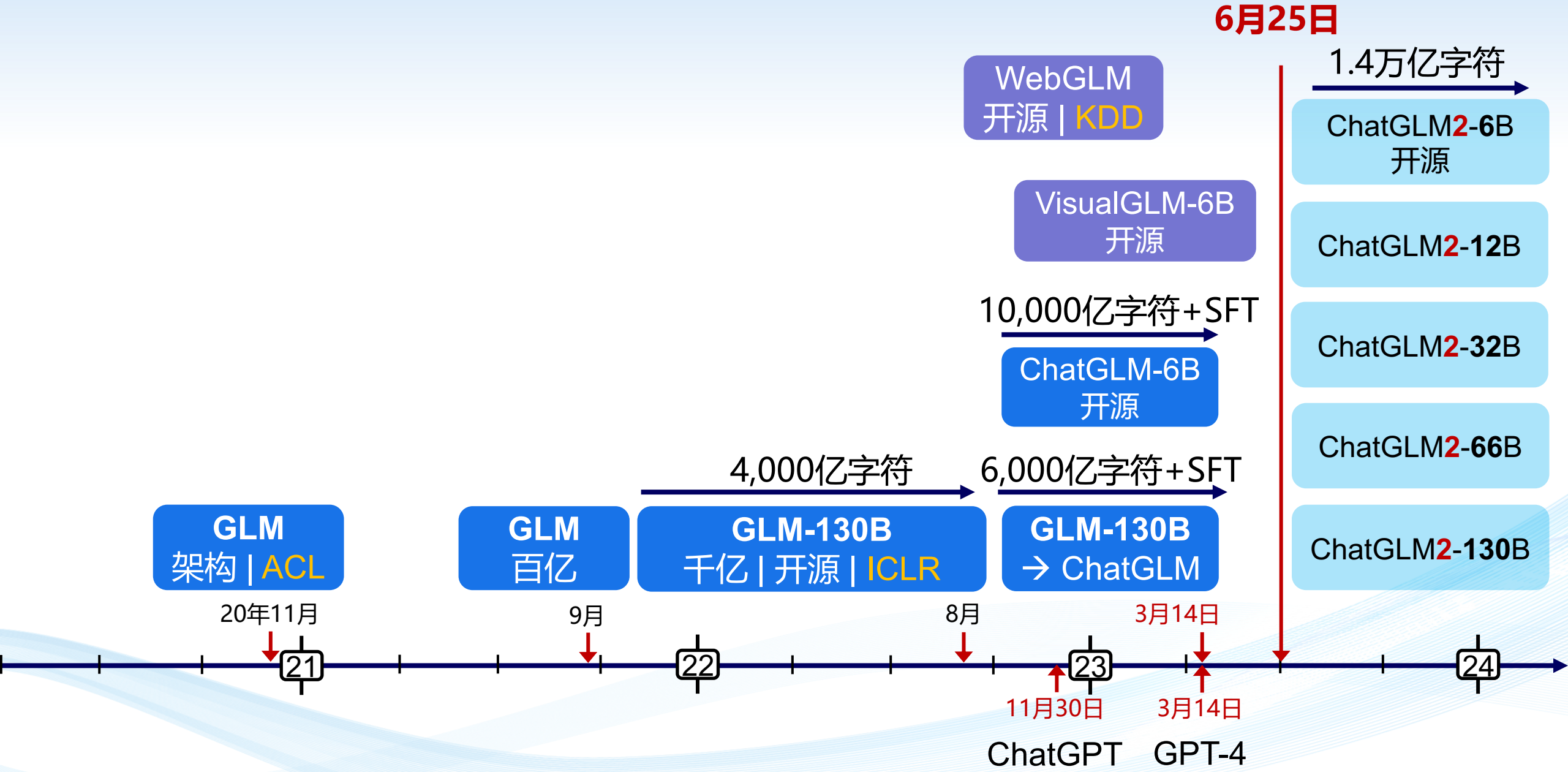
这只羊驼头上戴着一个绿色的小叶子

这种动物生活在哪里?

羊驼是一种生活在南美洲的哺乳动物，它们通常以草为食。

1）场景描述 2）进一步追问 3）联系知识

清华&智谱 GLM 系列模型



WebGLM = GLM + Search

Why do people try to maintain eye contact while communicating?

Why do people try to maintain eye contact while communicating?

People try to maintain eye contact while communicating because it is a form of body language which can help improve communication and create a positive impression. Eye contact is a non-verbal form of communication which can help show that you are actively listening and paying attention [1]. It can also make others more likely to like and respect you, making it an important tool for building and maintaining relationships [2]. Furthermore, a lack of eye contact can negatively impact communication, making misunderstandings more likely to occur. [3]

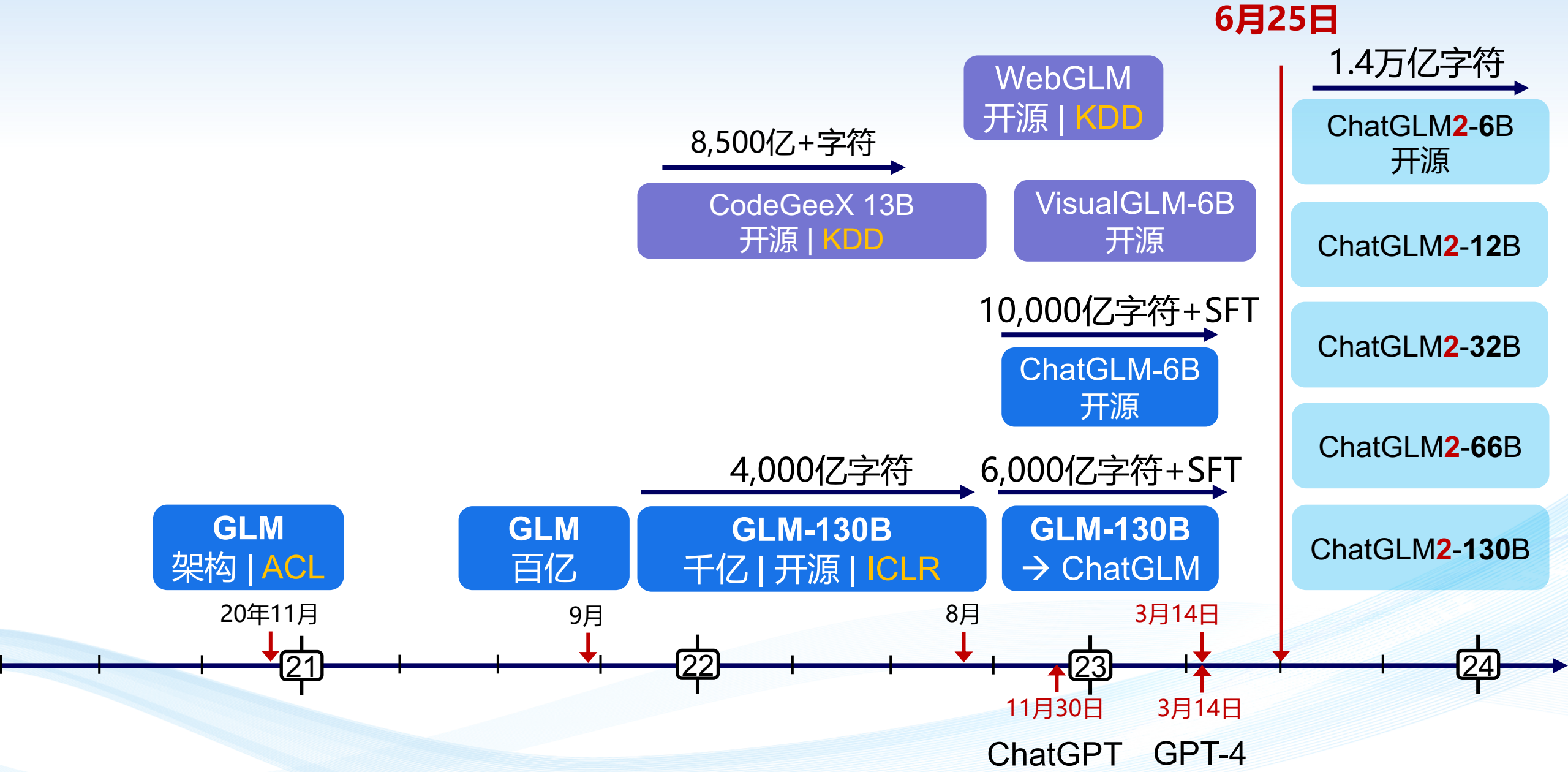
Like Dislike

[1] Why Eye Contact is Important in Communication

Eye contact is a type of body language that is extremely important during communication and conversation. Sometimes, our eyes and body language speak even more than words. Keeping eye contact with the person you are talking to shows that you are actively listening and paying attention.

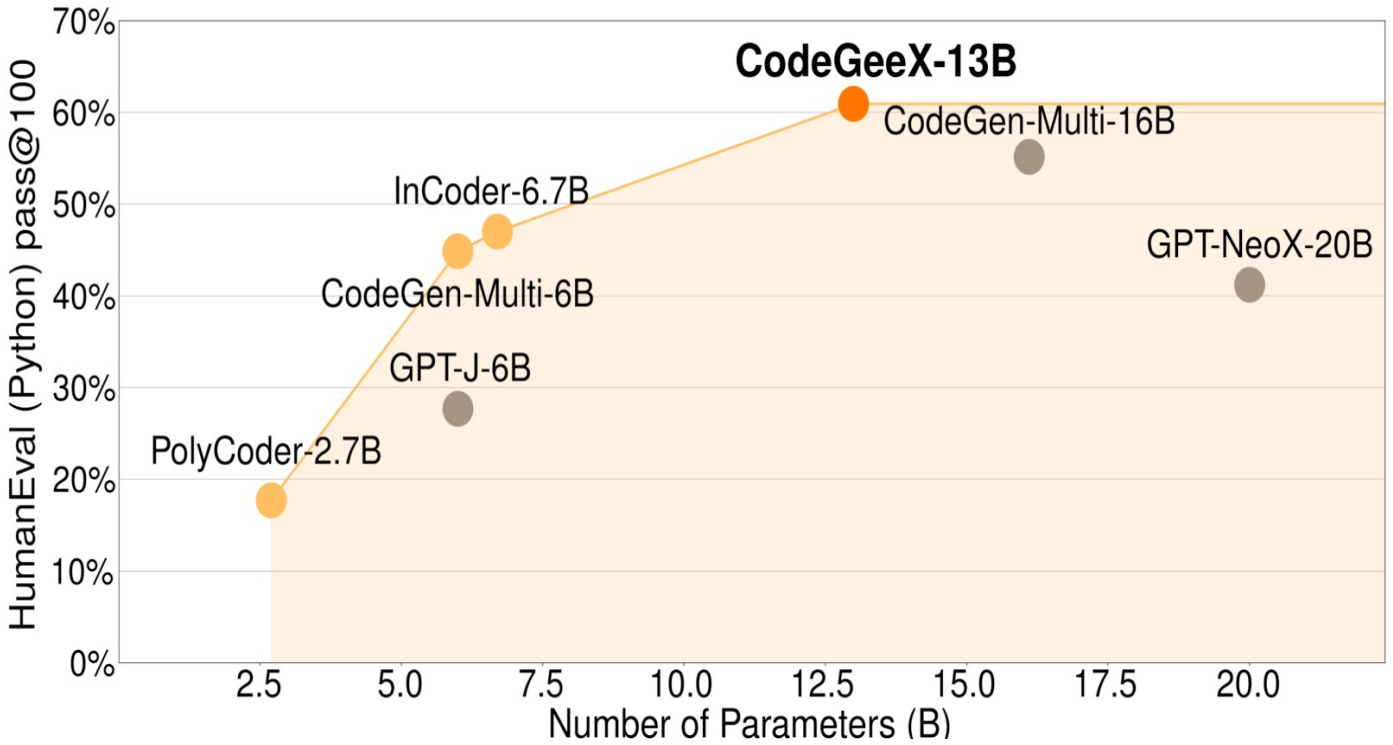
[2] Why Eye Contact is Important during Conversation?

清华&智谱 GLM 系列模型



代码模型CodeGeeX

- **130亿参数、20多种编程语言代码生成预训练模型**
- **支持昇腾和英伟达，具有代码生成、翻译、注释等能力**
- **2.6k GitHub星，VSCode、JetBrains编程插件**



CodeGeeX模型训练成功优化国产AI芯片

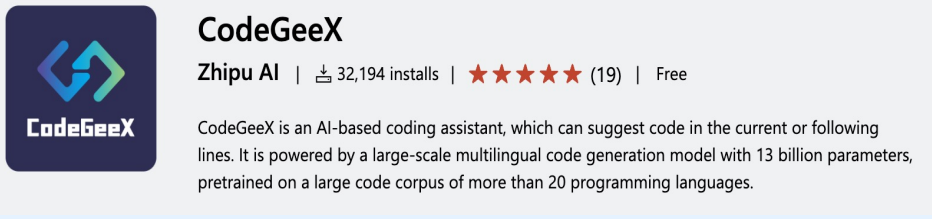
优化策略

- 算子融合(Layernorm/Gelu/BatchMatmul)
- 矩阵乘算子自动搜索效率最高的计算维度组合

性能提升

- **单卡昇腾910芯片训练效率提升257%**
- 千卡昇腾910芯片训练效率提升299%

每天帮程序员编写500多万行代码（2023.05）



Zheng et al., CodeGeeX: A Pre-Trained Model for Code Generation with Multilingual Evaluations on HumanEval-X. KDD'23

代码模型CodeGeeX

代码生成


```python
def factorial(n):
    factorial = 1
    for i in range(1, n + 1):
        factorial *= i
    return factorial
```

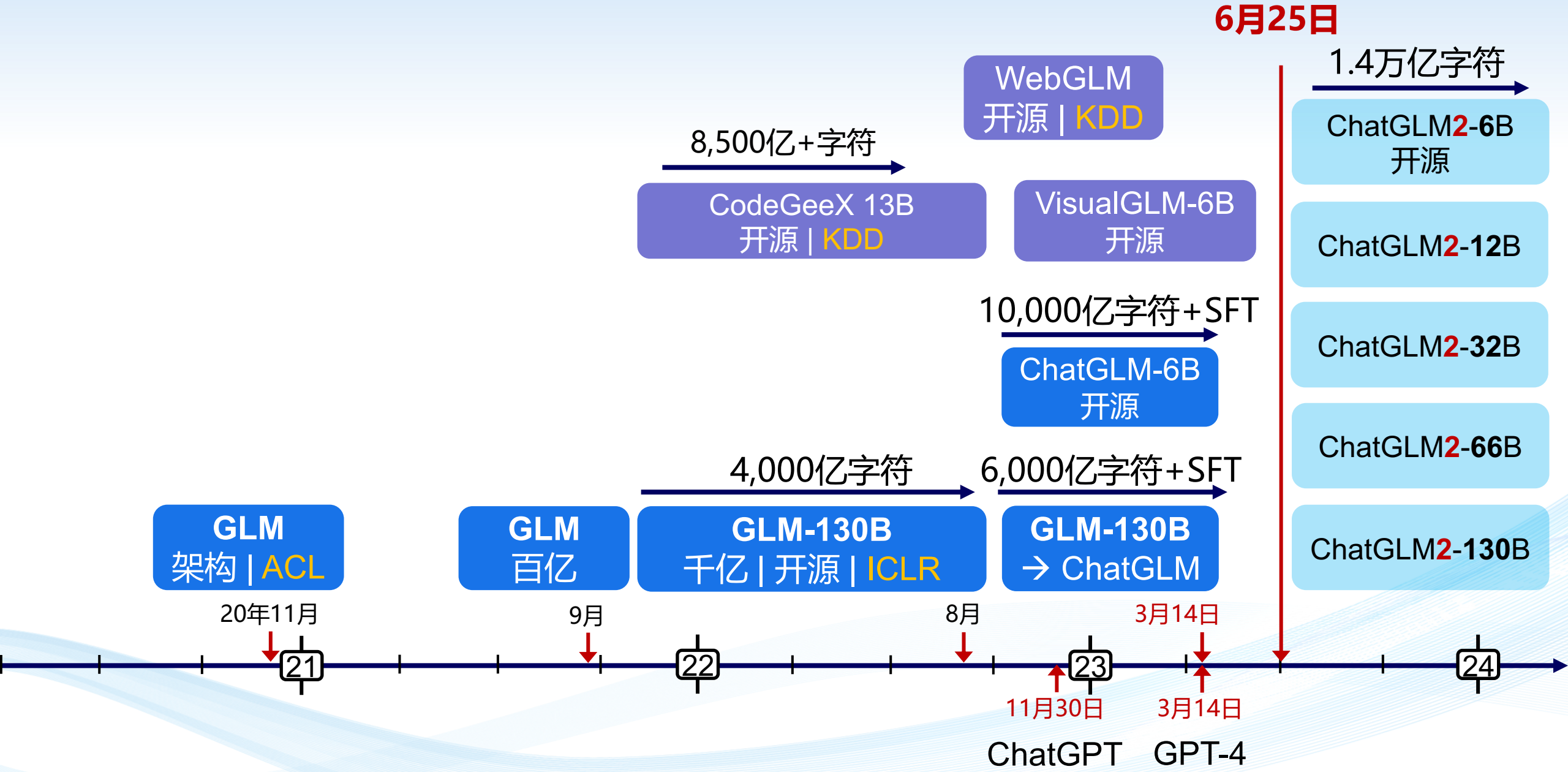
代码翻译


```cpp
int factorial(int n) {
    int result = 1;
    for (int i = 2; i <= n; ++i) {
        result *= i;
    }
    return result;
}
```


```python
def factorial(n):
    # initialize factorial
    factorial = 1
    # loop through the list
    for i in range(1, n + 1):
        # multiply the entry to the factorial
        factorial *= i
    # return the factorial
    return factorial
```

注释生成

清华&智谱 GLM 系列模型



开放的大模型研究

ChatGLM2-6B Public

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

Python 7.9k 1.3k

ChatGLM-6B Public

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

Python 31.2k 4.1k

GLM-130B Public

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

Python 6.6k 518

WebGLM Public

WebGLM: An Efficient Web-enhanced Question Answering System (KDD 2023)

Python 1k 102

VisualGLM-6B Public

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

Python 2.8k 291

CodeGeeX Public

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

Python 6k 432

https://github.com/THUDM

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

https://github.com/THUDM

认知大模型探索

- **解锁 In-context 能力**
  - 更多训练：大部分大模型都缺少训练
    - A sufficiently-trained LLM could be more powerful than we ever thought
  - 更多有针对性的高质量数据、高质量任务（目标函数）
- **环境交互/自我优化**
  - 让大模型和用户交互
  - 让大模型和环境（包括Web）交互
- **自反思学习**
  - 大模型的自我反思（self-instruct）

致谢

- **技术：**
  - 清华大学知识工程实验室（KEG）
  - 智谱AI
  - 清华大学PACMAN实验室
  - 清华大学自然语言处理实验室

- **算力：** 智谱·AI
  - 前期调试：中科曙光、鹏城实验室、神威·海洋之光
  - 千亿训练：济南超算中心（GLM-130B）
  - ChatGLM训练：智谱AI

谢谢大家！

https://**chatglm.cn**

https://**github.com/THUDM**

https://**huggingface.co/THUDM**

为什么
千亿（100B）模型？

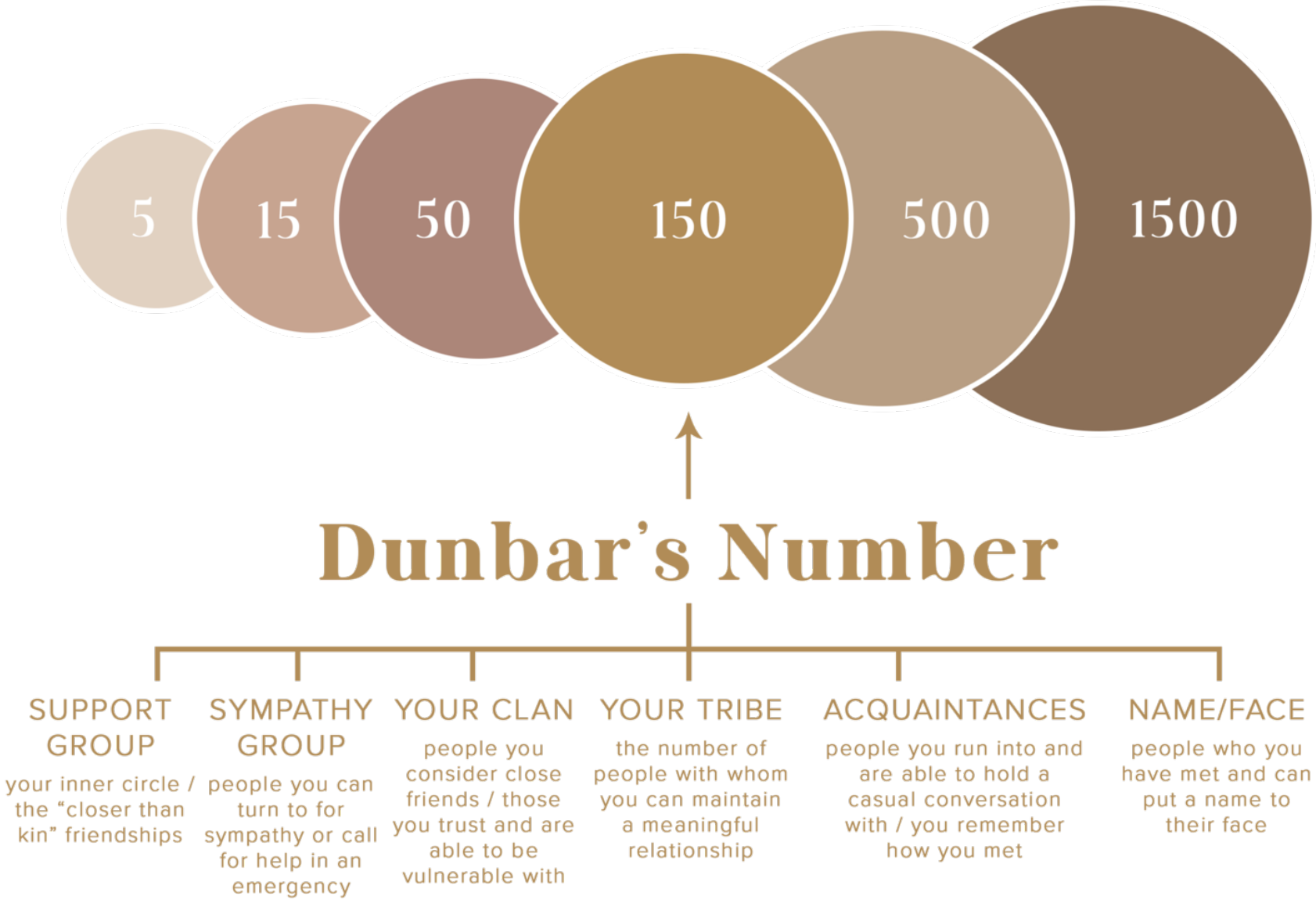
Dunbar's Number

How Many Connections Do/Can You Maintain?

SUPPORT GROUP	SYMPATHY GROUP	YOUR CLAN	YOUR TRIBE	ACQUAINTANCES	NAME/FACE
your inner circle / the "closer than kin" friendships	people you can turn to for sympathy or call for help in an emergency	people you consider close friends / those you trust and are able to be vulnerable with	the number of people with whom you can maintain a meaningful relationship	people you run into and are able to hold a casual conversation with / you remember how you met	people who you have met and can put a name to their face

R.I.M. Dunbar. The Social Brain Hypothesis. *Evolutionary Anthropology: Issues, News, and Reviews: Issues, News, and Reviews*, *6*(5), 178-190. **1998**

Dunbar's Number

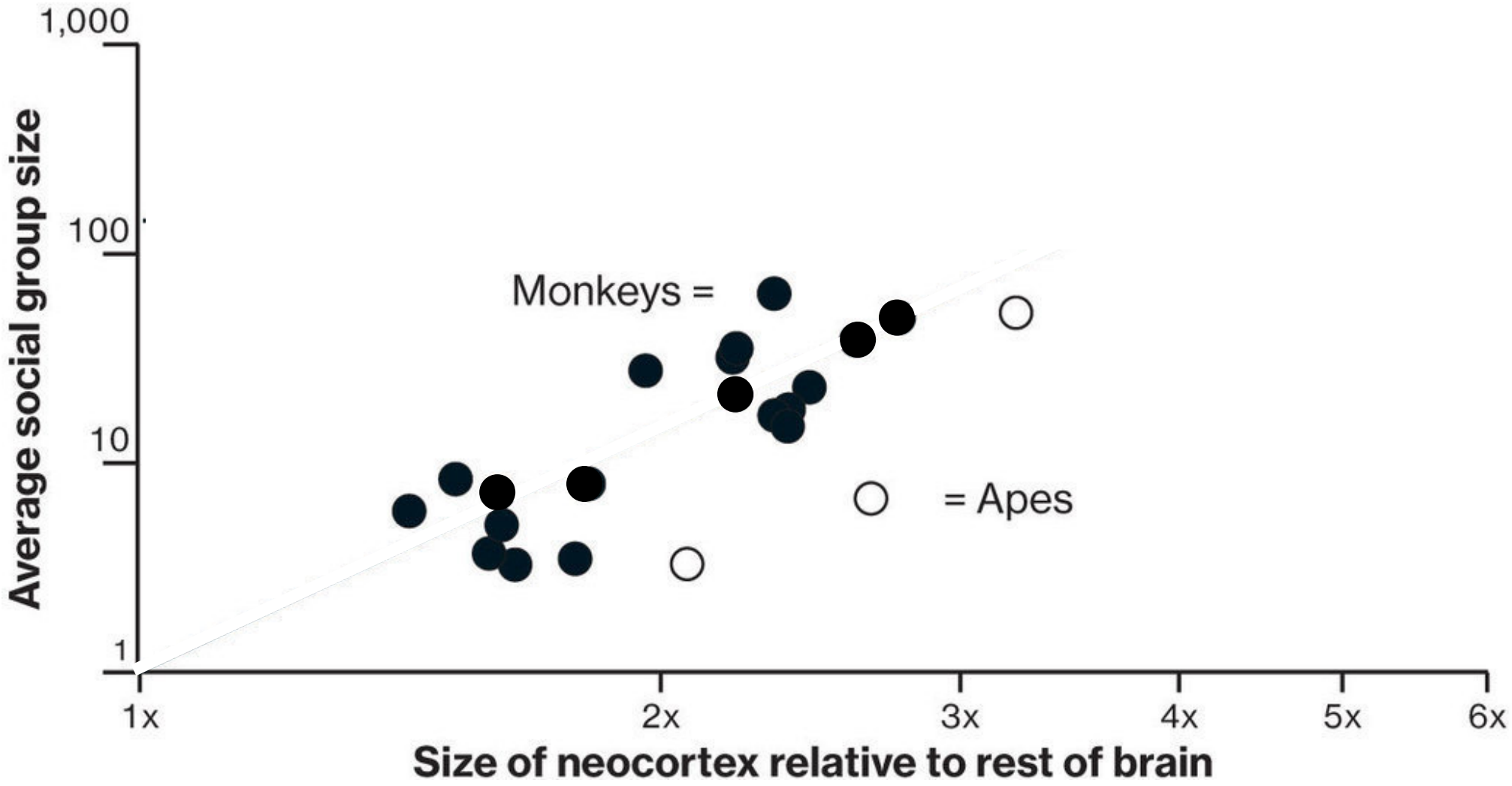


R.I.M. Dunbar. The Social Brain Hypothesis. *Evolutionary Anthropology: Issues, News, and Reviews: Issues, News, and Reviews*, *6*(5), 178-190. **1998**

How About ?

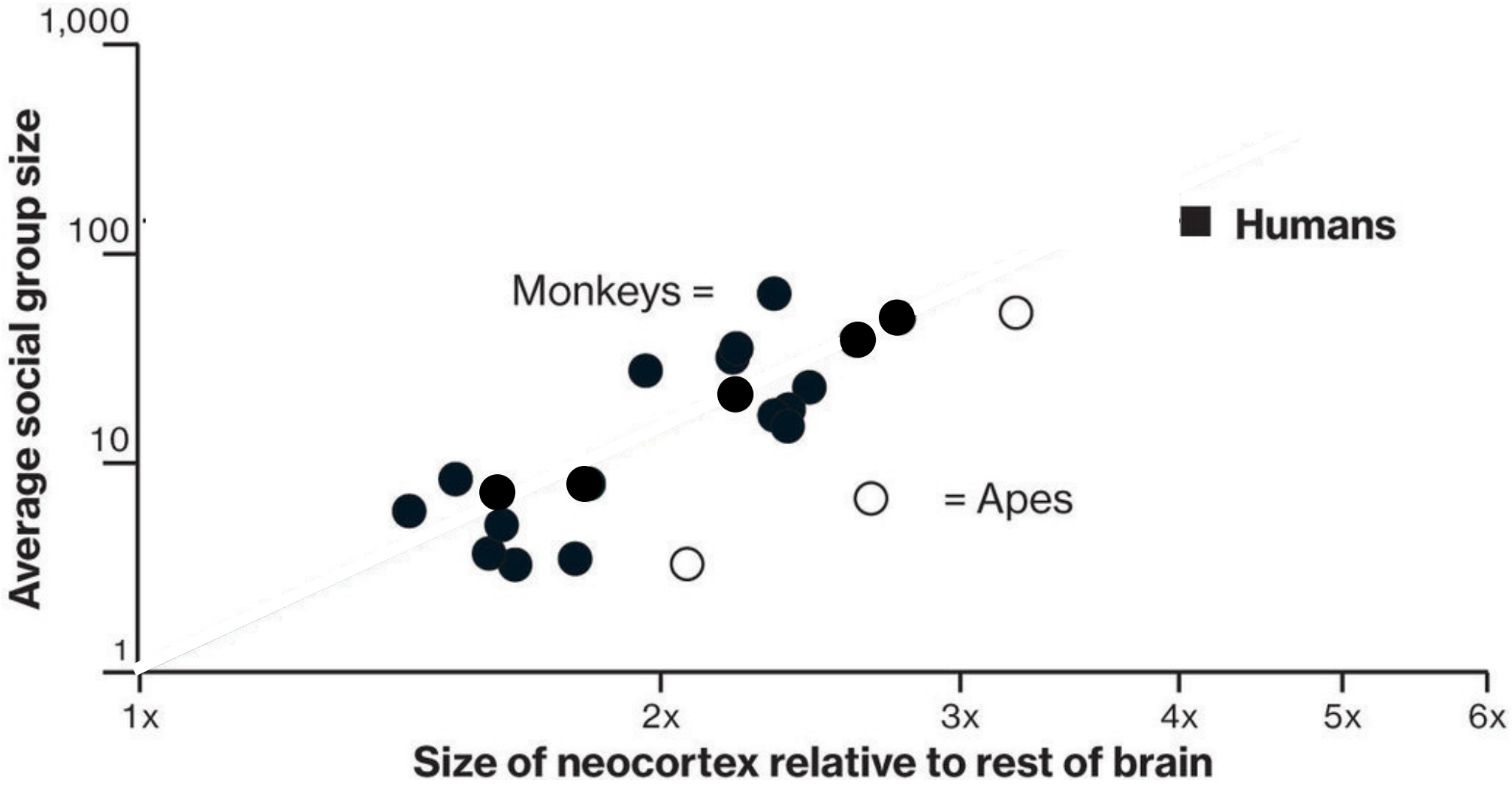
https://www.discovermagazine.com/planet-earth/if-humans-evolved-from-apes-why-do-apes-still-exist

The Social Cortex



R. I. M. Dunbar, Susanne Shultz. Evolution in the Social Brain. **Science**, Vol. 317, No. 5843, **2007**

The Social Cortex



R. I. M. Dunbar, Susanne Shultz. Evolution in the Social Brain. **Science**, Vol. 317, No. 5843, **2007**

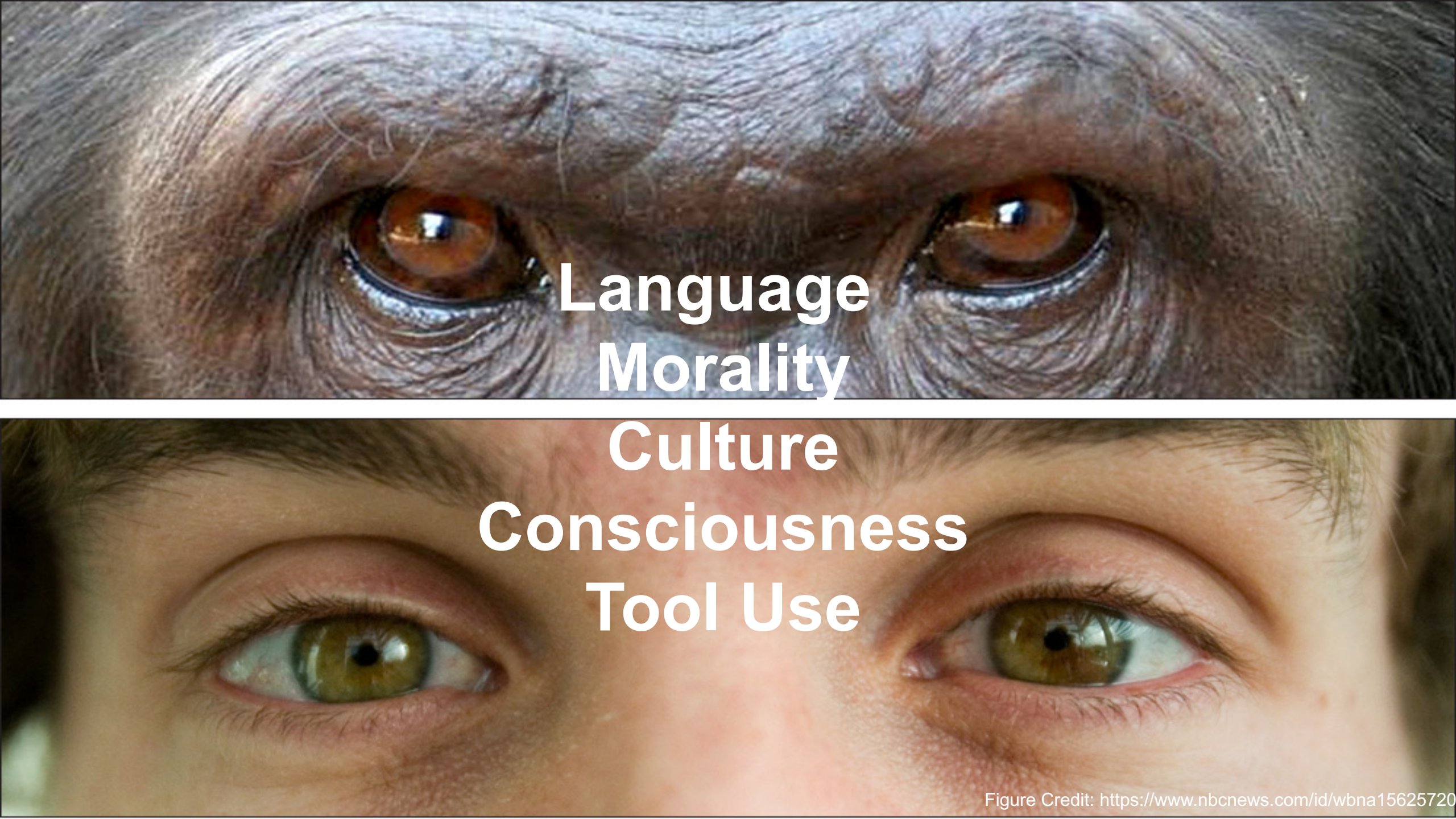
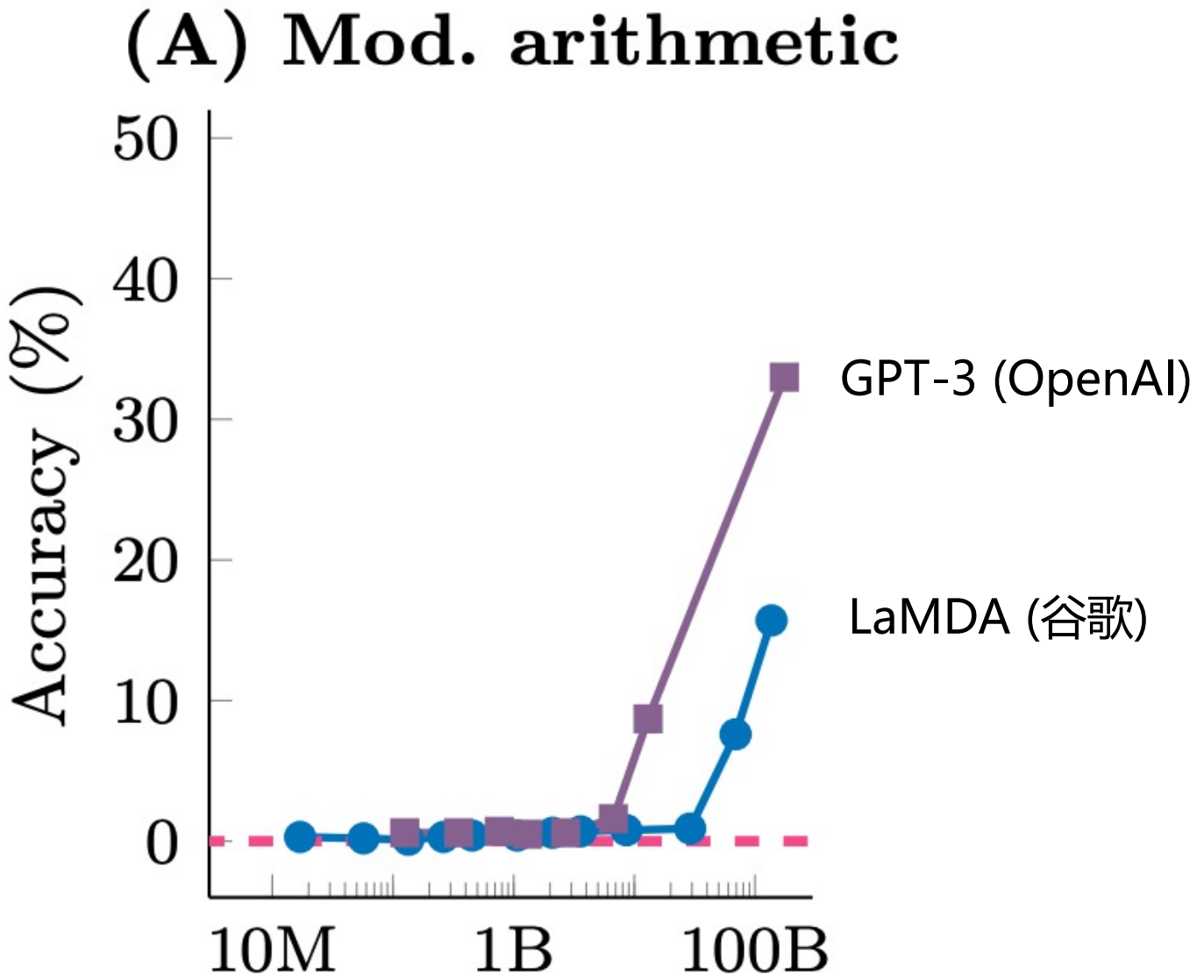


Figure Credit: https://www.nbcnews.com/id/wbna15625720

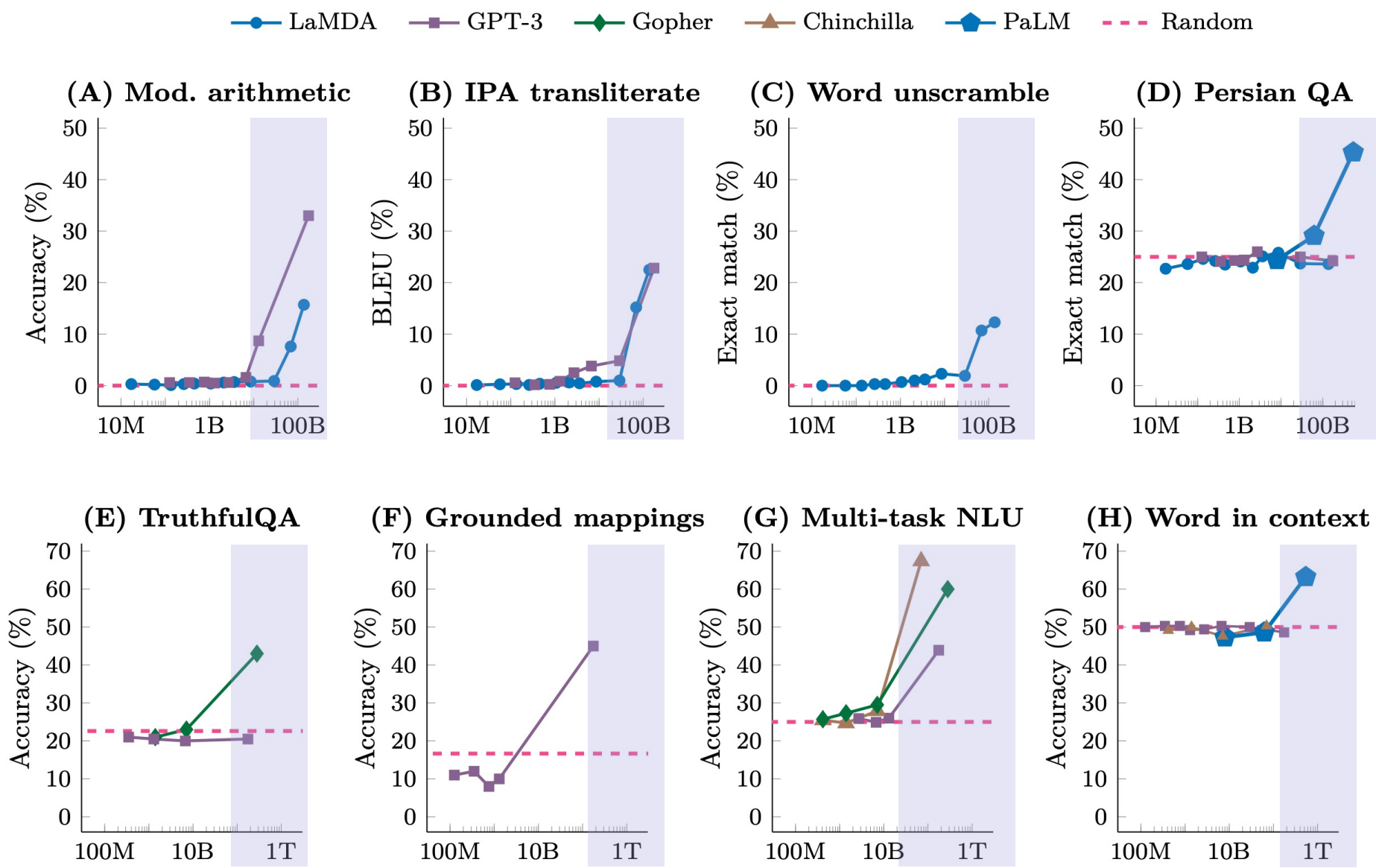
为什么千亿(100B)大规模

- What is 16 mod 12?
- 16 divided by 12 equals 1 remainder 4. So the answer is 4!



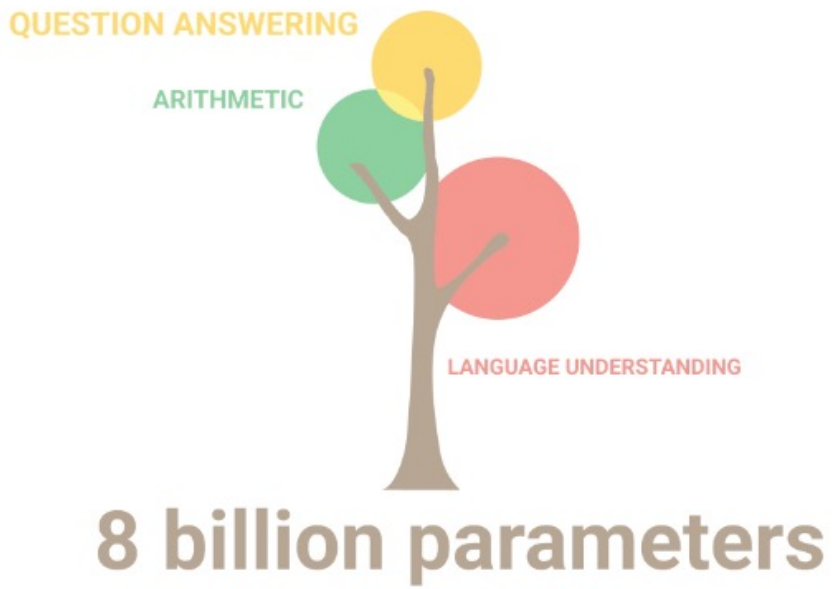
1. J Wei, et al. Emergent Abilities of Large Language Models. arXiv: **2206**.07682

为什么千亿(100B)大规模



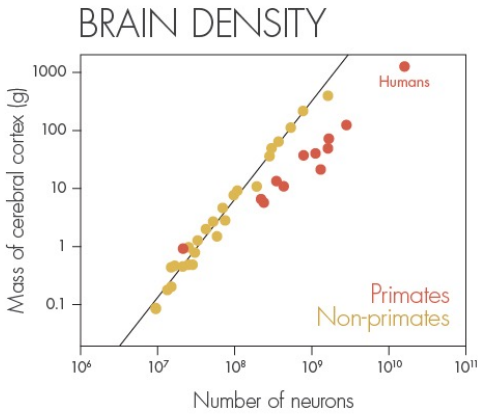
1. J Wei, et al. Emergent Abilities of Large Language Models. arXiv: **2206**.07682

“涌现的模型新能力”



Gif Credit: Google

Brain Size Evolution **Spike**



晚期智人

智人

Homo sapiens sapiens

Language

us!

直立猿人

Homo sapiens
Neandertalensis

非洲能人
(能用手准确抓握器物的早期人)

Homo erectus

非洲南方古猿

1000 cc

Homo habilis

类人猿

Australopithecus
africanus

Tool

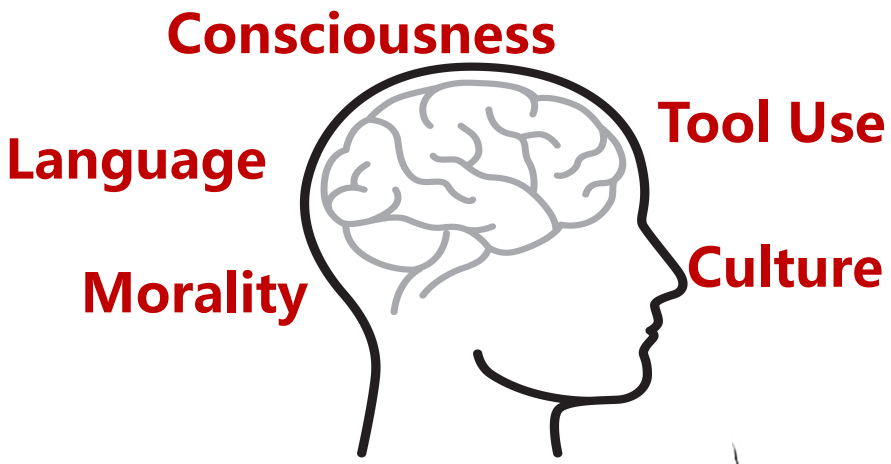
500 cc

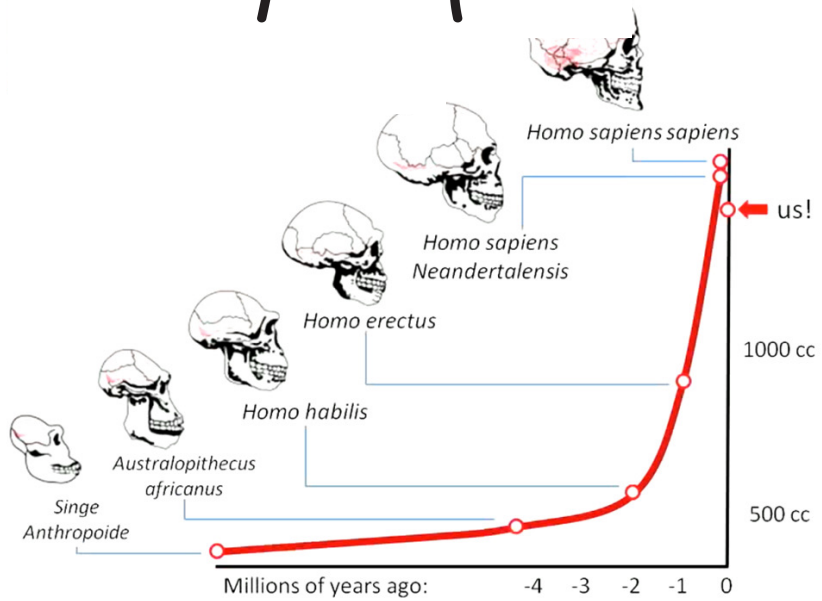
Singe
Anthropoide

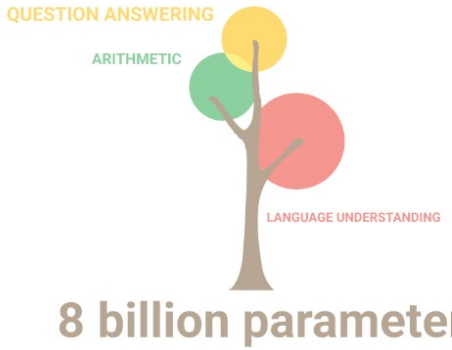
Millions of years ago: -4 -3 -2 -1 0

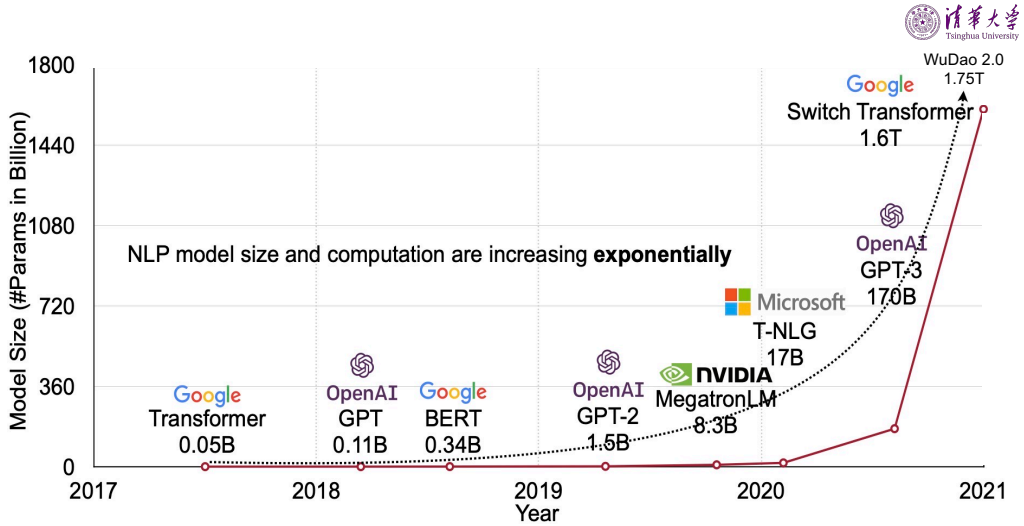
Disclaimer: The instructor is not an expert on this topic at all

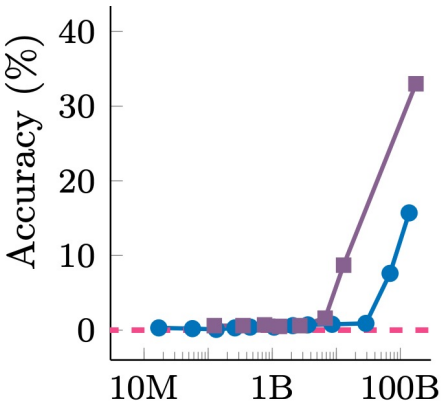
An Observation











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

PC: Web & Google

Reasoning Tasks

GLM-10B

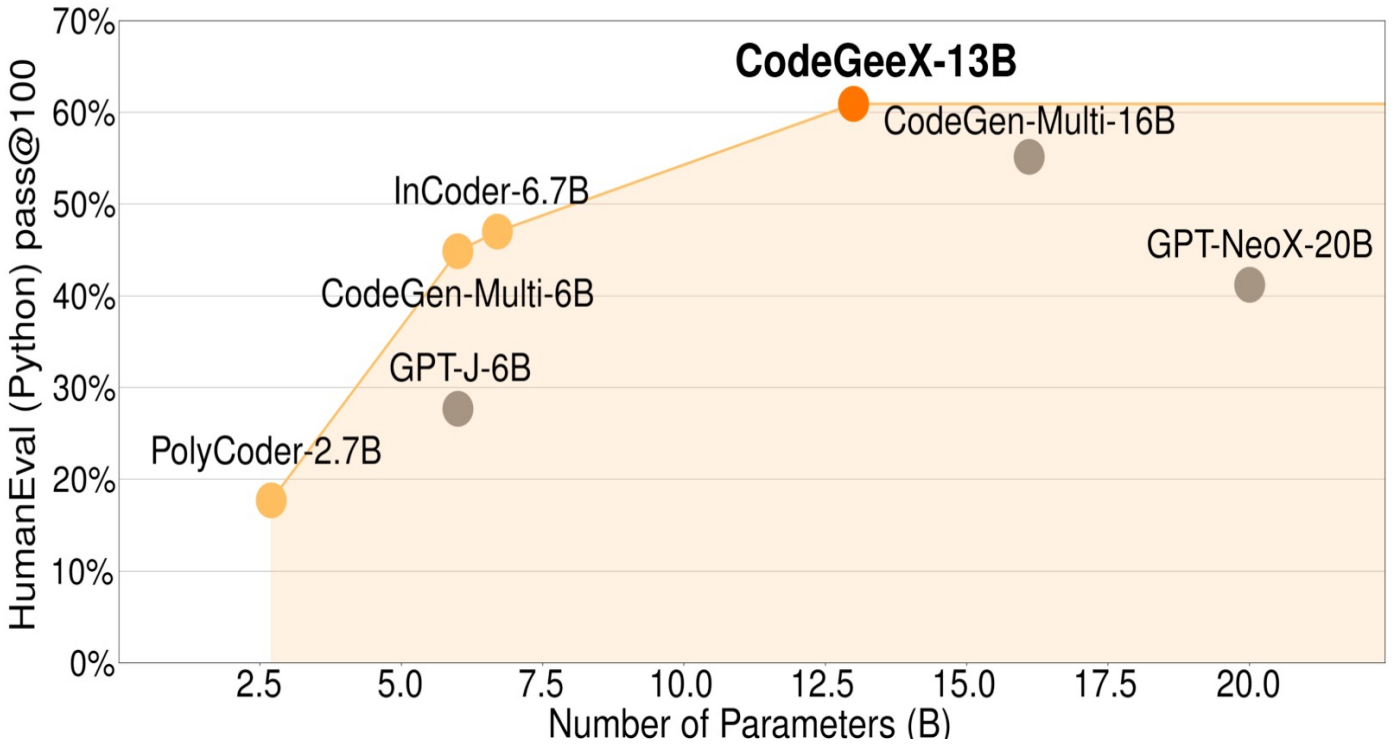
- 一斤西瓜10元钱，5斤西瓜多少钱?
  - 一斤西瓜十元,五斤西瓜30元
- 一斤苹果10元，一斤西瓜5元，那么2斤苹果和3斤西瓜多少钱?
  - 总共30元
- 我把可口可乐洒在了桌子上，接下来该如何清楚桌面?
  - 把桌上的东西都拿走,用湿抹布擦桌子,然后喷上酒精,再拿抹布擦,最后用酒精再擦一遍

GLM-130B

- 一斤西瓜10元钱，5斤西瓜多少钱?
  - 5x10=50(元)
- 一斤苹果10元，一斤西瓜5元，那么2斤苹果和3斤西瓜多少钱?
  - 2x10+3x5=35(元)
- 我把可口可乐洒在了桌子上，接下来该如何清楚桌面?
  - 先用抹布擦干净，然后用湿抹布再擦一次，最后用干抹布再擦一次

代码模型CodeGeeX

- **130亿参数、20多种编程语言代码生成预训练模型**
- **支持昇腾和英伟达，具有代码生成、翻译、注释等能力**
- **2.6k GitHub星，VSCode、JetBrains编程插件**



CodeGeeX模型训练成功优化国产AI芯片

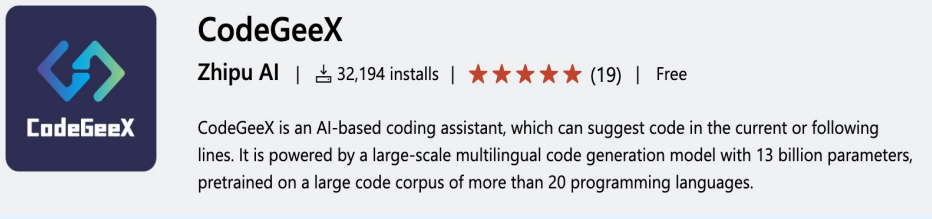
优化策略

- 算子融合(Layernorm/Gelu/BatchMatmul)
- 矩阵乘算子自动搜索效率最高的计算维度组合

性能提升

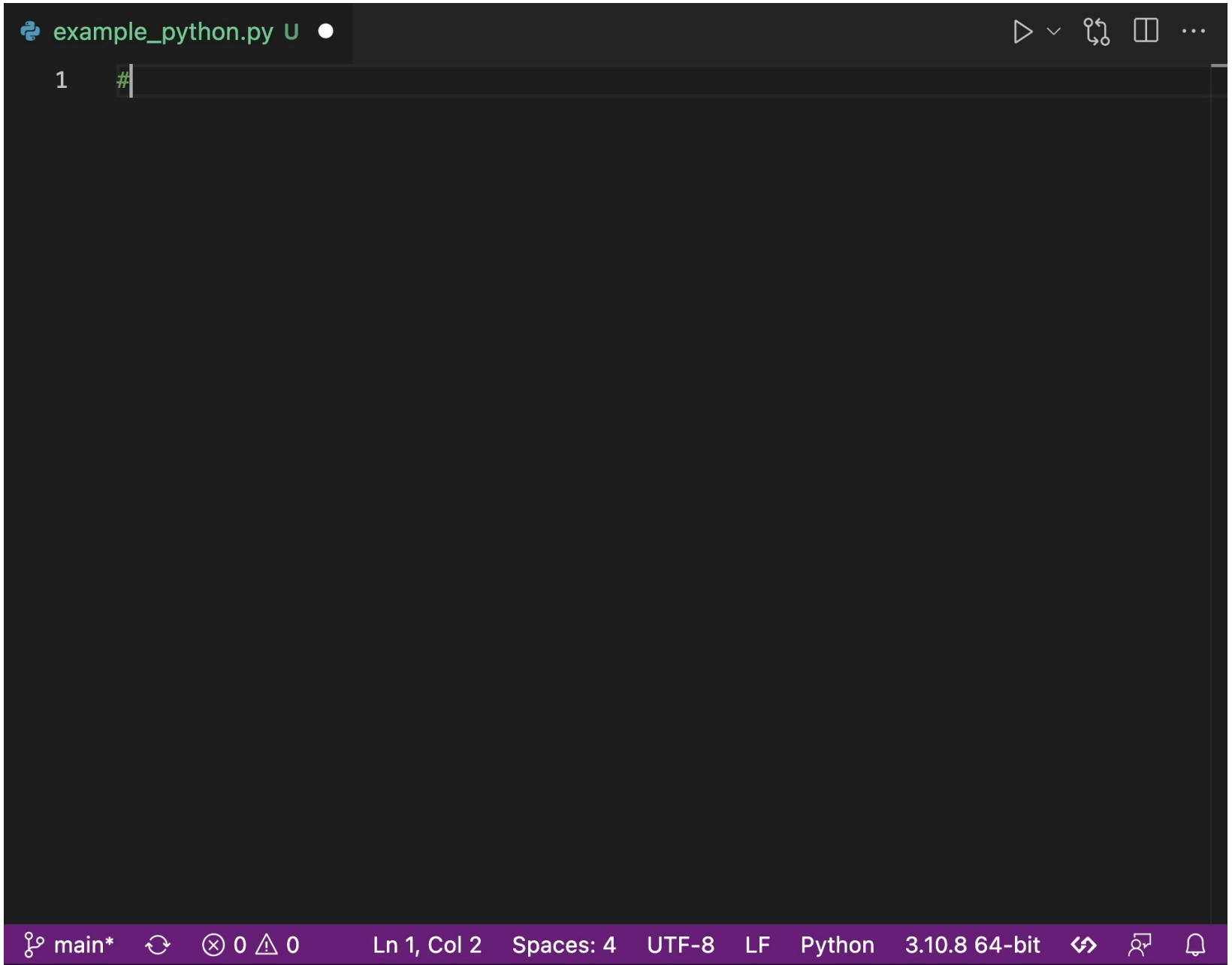
- **单卡昇腾910芯片训练效率提升257%**
- 千卡昇腾910芯片训练效率提升299%

每天帮程序员编写500多万行代码（2023.05）

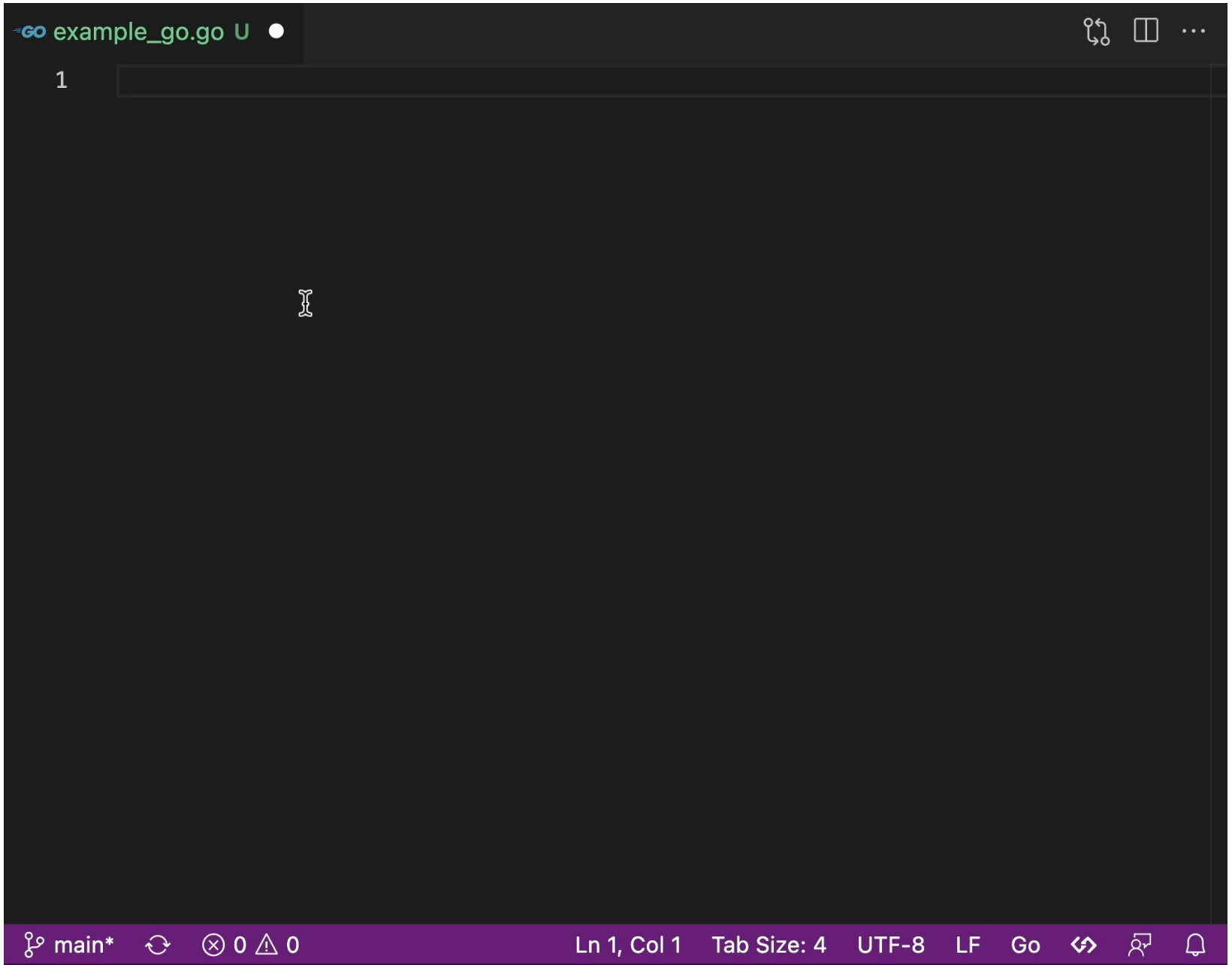


Zheng et al., CodeGeeX: A Pre-Trained Model for Code Generation with Multilingual Evaluations on HumanEval-X. KDD'23

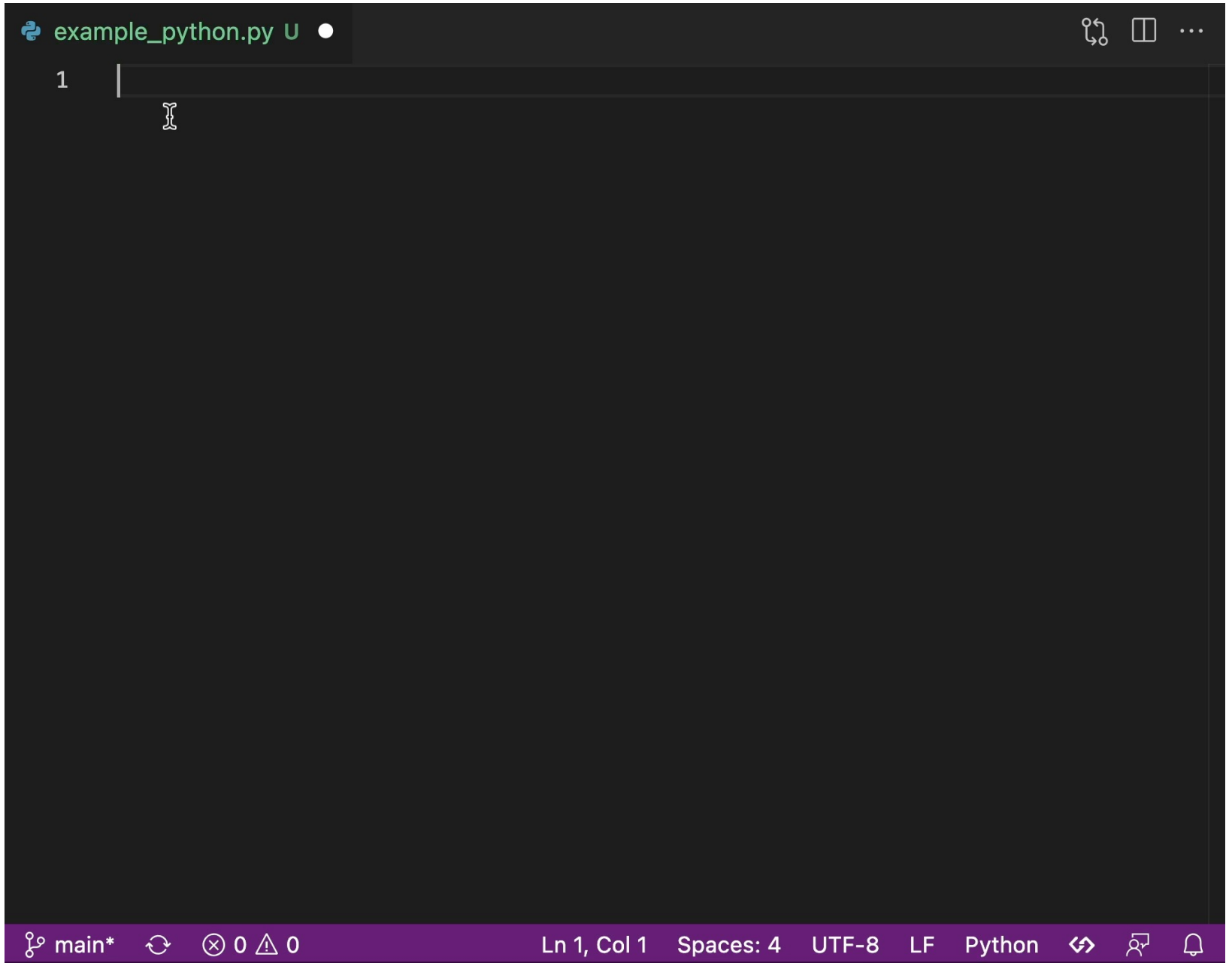
CodeGeeX: Code Generation



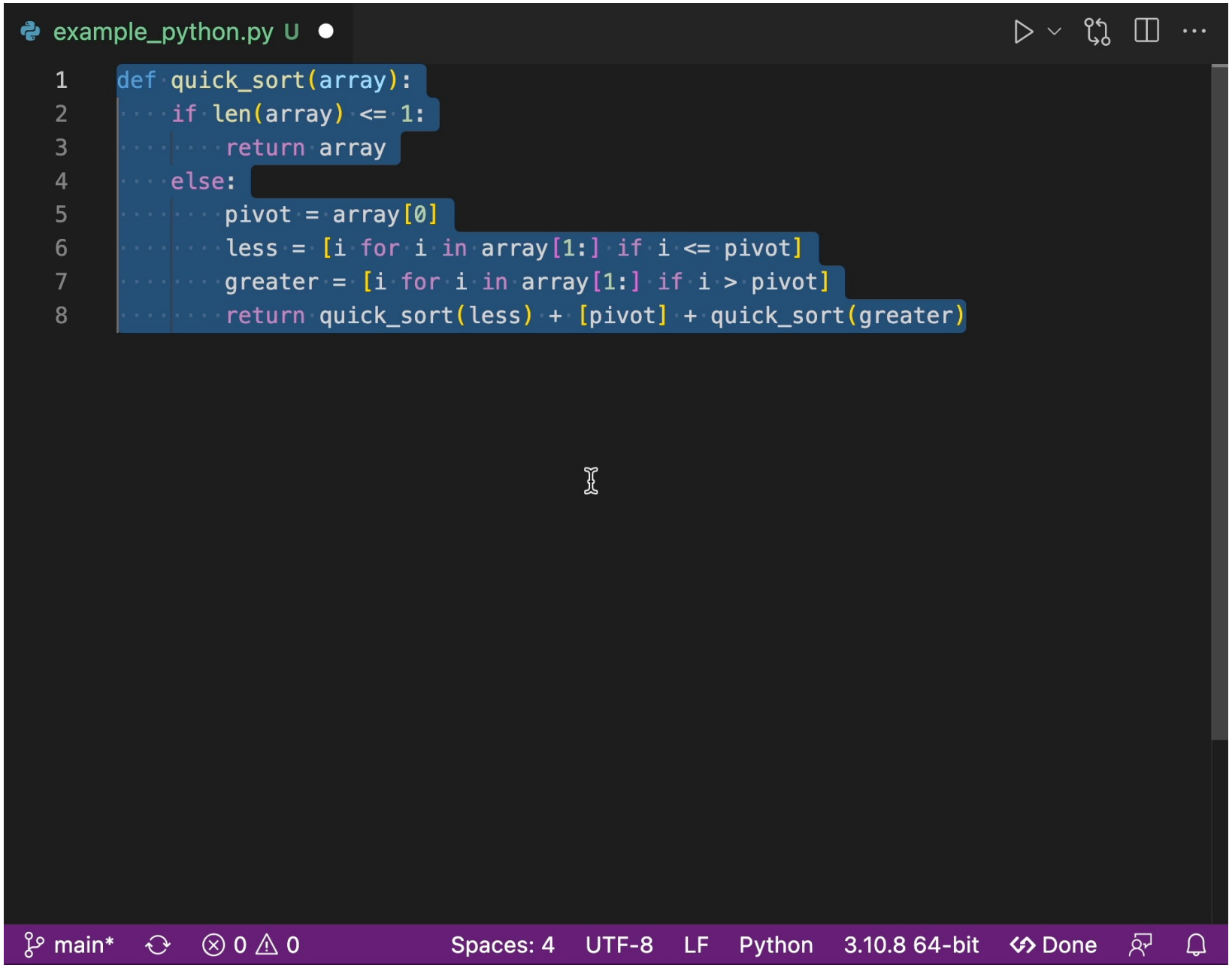
CodeGeeX: Code Generation



CodeGeeX: Code Translation



CodeGeeX: Code Explanation



代码模型CodeGeeX

代码生成


```python
def factorial(n):
    factorial = 1
    for i in range(1, n + 1):
        factorial *= i
    return factorial
```

代码翻译


```cpp
int factorial(int n) {
    int result = 1;
    for (int i = 2; i <= n; ++i) {
        result *= i;
    }
    return result;
}
```


```python
def factorial(n):
    # initialize factorial
    factorial = 1
    # loop through the list
    for i in range(1, n + 1):
        # multiply the entry to the factorial
        factorial *= i
    # return the factorial
    return factorial
```

注释生成