



清华大学 计算机科学与技术系

Department of Computer Science and Technology, Tsinghua University

2023年计算机系校友创新创业研讨会
暨第一届基础模型前沿研讨会

ChatGLM:从千亿到开源的一点思考

唐杰
清华大学计算机系

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



GLM-130B

An Open Bilingual Pre-Trained Model



ChatGLM: <https://github.com/THUDM/ChatGLM-6B>



ChatGLM
Alpha

ChatGLM-130B vs. ChatGLM-6B

ChatGLM（alpha内测版：QAGLM）是一个初具问答和对话功能的中英双语模型，当前仅针对中文优化，多轮和逻辑能力相对有限，但其仍在持续迭代进化过程中，敬请期待模型涌现新能力。

GLM

百亿

XDAI

小呆

GLM-130B

千亿基座

CodeGeeX

百亿代码

QAGLM

对话问答

ChatGLM

敬请期待

时隔半年，非常高兴地给大家介绍我们大模型家族的新成员：对话机器人 [ChatGLM](#)（alpha内测版：QAGLM），这是一个初具问答和对话功能的千亿中英语言模型，并针对中文进行了优化，现已开启邀请制内测，后续还会逐步扩大内测范围。与此同时，继开源 GLM-130B 千亿基座模型之后，我们正式开源最新的中英双语对话 GLM 模型：[ChatGLM-6B](#)，结合模型量化技术，用户可以在消费级的显卡上进行本地部署（INT4 量化级别下最低只需 6GB 显存）。经过约 1T 标识符的中英双语训练，辅以监督微调、反馈自助、人类反馈强化学习等技术的加持，62 亿参数的 ChatGLM-6B 虽然规模不及千亿模型，但大大降低了用户部署的门槛，并且已经能生成相当符合人类偏好的回答。

下面，我们请 ChatGLM（模型版本：qa-glm-v0.7）来做自我介绍吧：

你好

你好👋! 我是人工智能助手 ChatGLM，很高兴见到你，欢迎问我任何问题。

⌚ 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

帮我写一个介绍ChatGLM的博客提纲

好的，以下是一个可能的博客提纲：

I. 前言

1. 简要介绍 ChatGLM 的背景和主要功能

II. ChatGLM 的特点

1. 智能化的问答和对话功能

2. 自定义主题和语言

3. 丰富的知识库和语言模型

4. 支持多种平台和语言

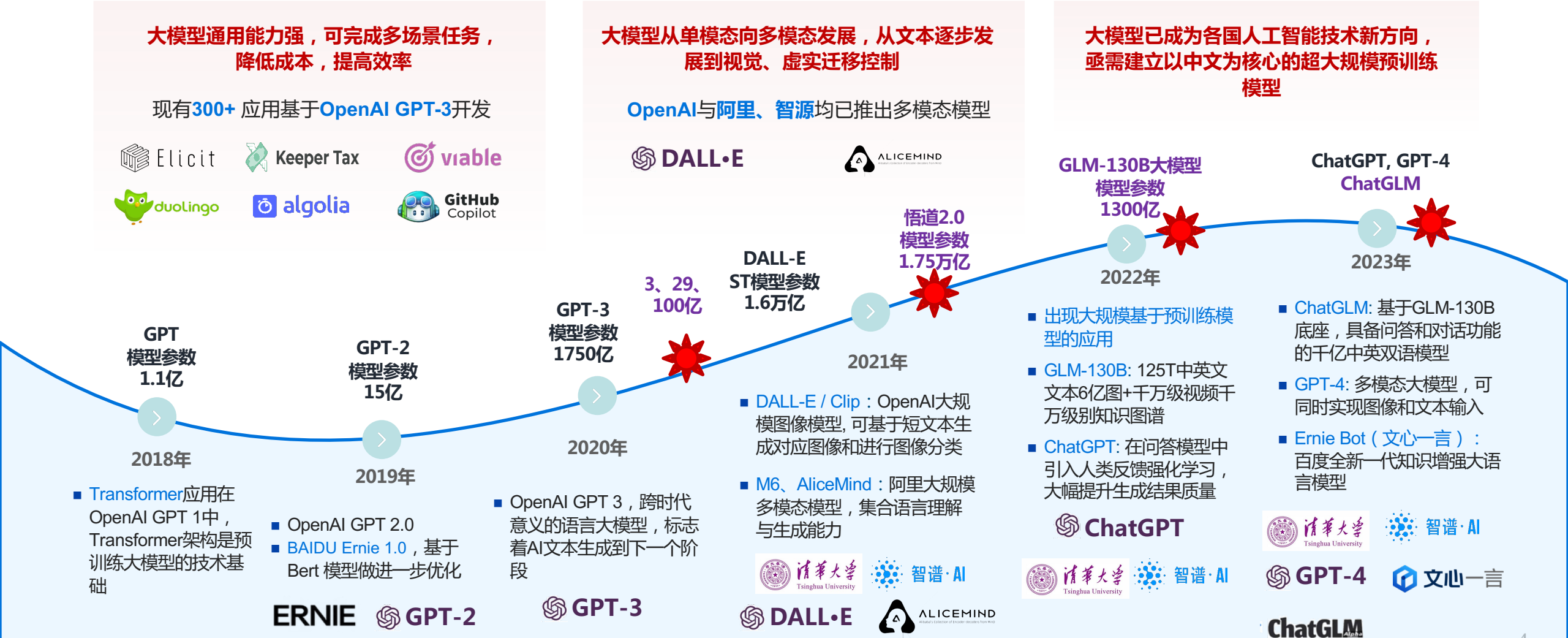


技术
博客

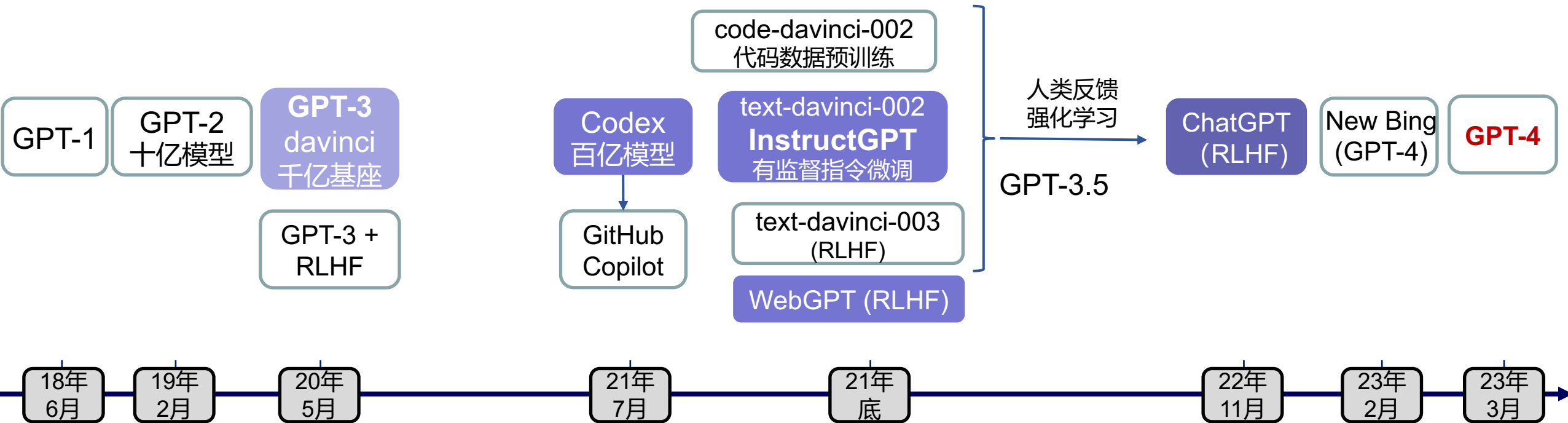
- ChatGLM-6B开源，**10天 10000 stars**
- 当天在 GitHub 的趋势排行**第一**
- Huggingface下载超过**100万**
- 开源的训练数据量达到**1万亿**字符的模型

预训练大模型

预训练大模型是新一代人工智能应用的基础设施



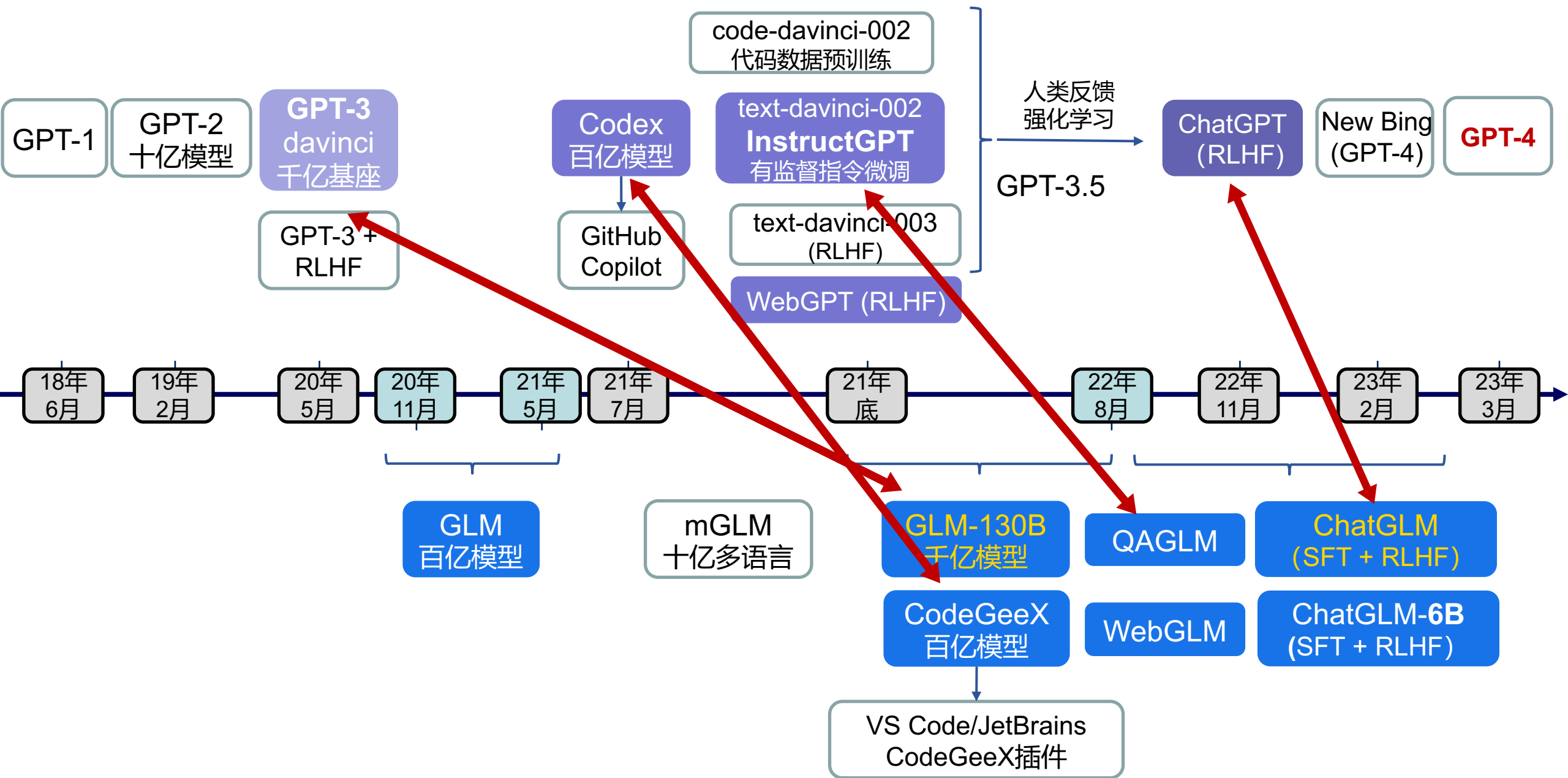
GPT



1. 千亿基座模型

2. 有监督指令微调

3. 人类反馈强化学习



清华&智谱 GLM 系列模型

OpenAI GPT系列模型

GPT-3
davinci
千亿基座

3000亿标识符

Transformer
(1750亿参数)

GPT
自回归生成

数据

神经网络

预训练架构

GLM-130B
千亿模型

2000亿中文
2000亿英文

Transformer
(1300亿参数)

GLM
自回归填空

清华&智谱 GLM 系列模型

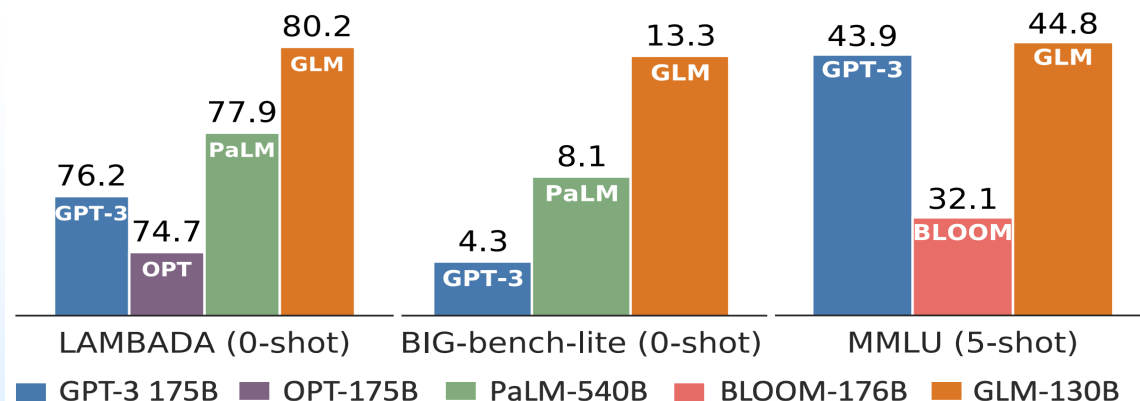
千亿GLM-130B

	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
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 和申威
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

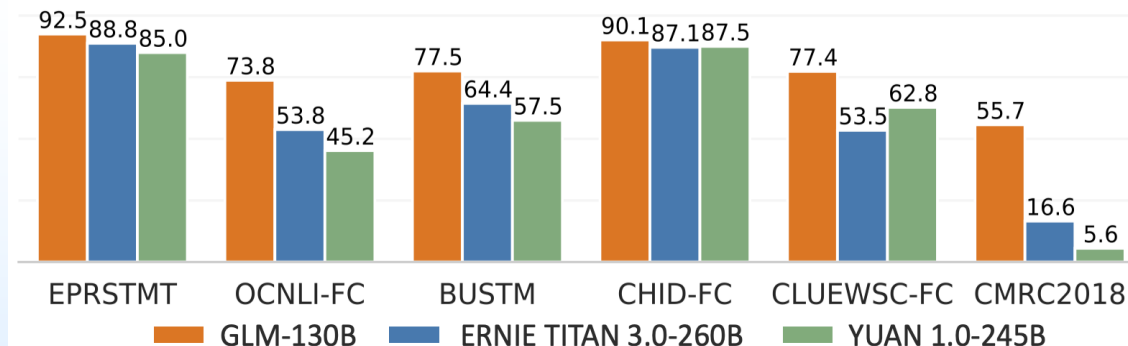
自2022年8月发布以来，收到**69个国家**
1000+机构的使用需求

- Google
- Microsoft
- Meta
- AI2
- Stanford
- MIT
- UC Berkely
- CMU
- Harvard
- Princeton
- Cambridge
- 华为
- 百度
- 阿里巴巴
- 腾讯
- 头条
- 滴滴
- 建设银行
- 智源
- 小冰
- 小米
- 北京大学
- 浙江大学
- 上海交大
- 中科大
- 复旦大学
- 中科院大学
- 香港大学
- 香港科技大学
- 清华大学

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



中文：在 CLUE 和 FewCLUE 上远超ERNIE 3.0, Yuan



千亿GLM-130B

Model	Model Creator	Modality							
J1-Jumbo v1 (178B)	AI21 Labs	Text							
J1-Grande v1 (17B)	AI21 Labs	Text							
J1-Large v1 (7.5B)	AI21 Labs	Text							
Anthropic-LM v4-s3 (52B)	Anthropic	Text							
BLOOM (176B)	BigScience	Text							
T0++ (11B)	BigScience	Text							
Cohere xlarge v20220609 (52.4B)	Cohere	Text							
Cohere large v20220720 (13.1B) ⁵⁸	Cohere	Text							
Cohere medium v20220720 (6.1B)	Cohere	Text							
Cohere small v20220720 (410M) ⁵⁹	Cohere	Text							
GPT-J (6B)	EleutherAI	Text							
GPT-NeoX (20B)	EleutherAI	Text	20B	GPT-NeoX	2048	open	599,170,730	849,830	540 GPU hours
T5 (11B)	Google	Text	11B	T5	512	open	199,017,126	406,072	1,380 GPU hours
UL2 (20B)	Google	Text	20B	UL2	512	open	199,539,380	406,072	1,570 GPU hours
OPT (66B)	Meta	Text	66B	OPT	2048	open	612,752,867	851,178	2,000 GPU hours
OPT (175B)	Meta	Text	175B	OPT	2048	open	610,436,798	851,178	3,400 GPU hours
TNLG v2 (6.7B)	Microsoft/NVIDIA	Text	6.7B	GPT-2	2047	closed	417,583,950	590,756	-
TNLG v2 (530B)	Microsoft/NVIDIA	Text	530B	GPT-2	2047	closed	417,111,519	590,756	-
GPT-3 davinci v1 (175B)	OpenAI	Text	175B	GPT-2	2048	limited	422,001,611	606,253	\$8,440
GPT-3 curie v1 (6.7B)	OpenAI	Text	6.7B	GPT-2	2048	limited	423,016,414	606,253	\$846
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

AI21 labs

ANTHROPIC

BigScience

co:here



Google

Meta



Microsoft



OpenAI



Yandex

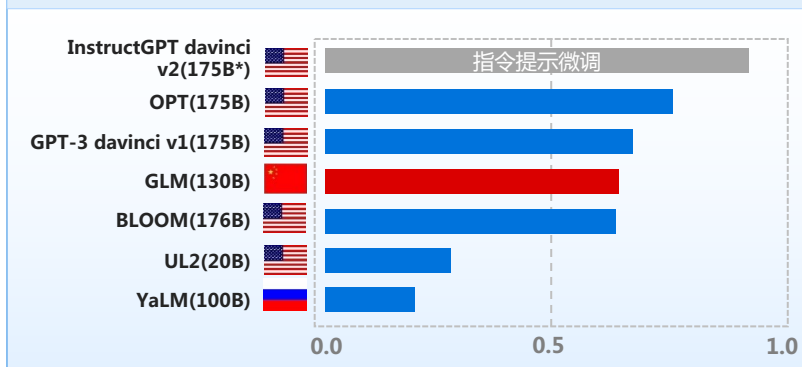
TOGETHER

GLM-130B与OpenAI、Google、Meta、Microsoft等基础大模型相比，在部分评估指标上具有优势

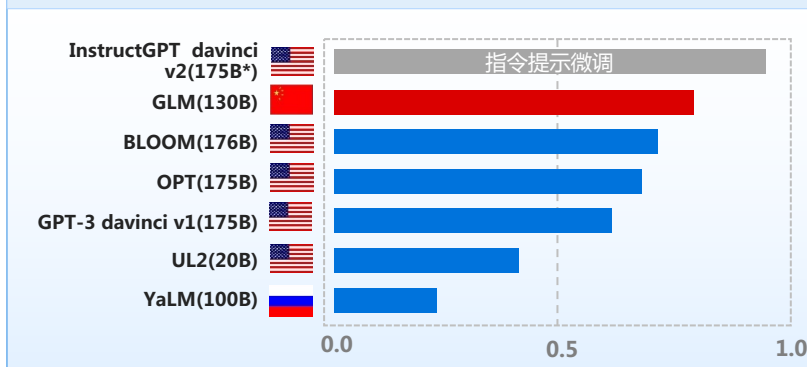
千亿GLM-130B

Stanford报告的世界主流大模型评测：准确性、恶意性与**GPT-3持平**，鲁棒性和校准误差在**所有模型中表现较佳**

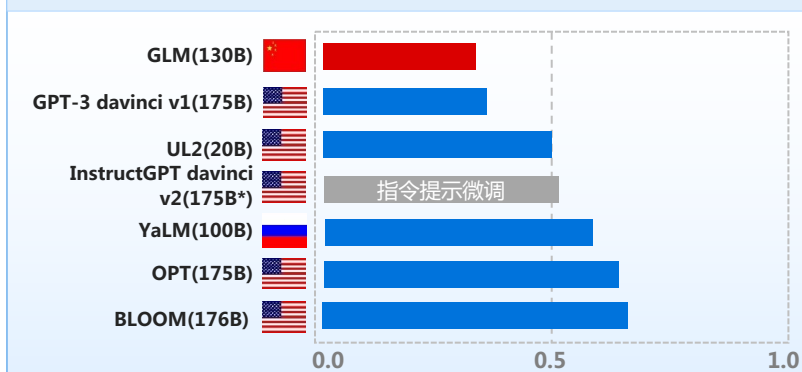
准确性 Accuracy



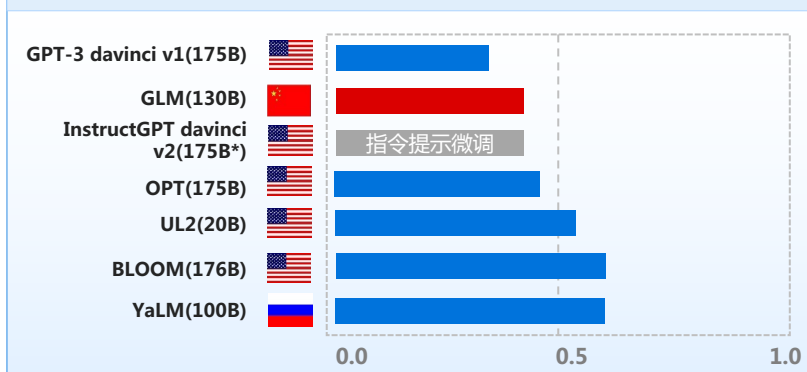
鲁棒性 Robustness



校准误差 Calibration error



恶意性 Toxicity



双语高精度



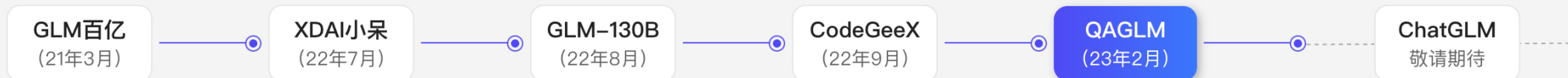
高可用、低成本



多平台适配



ChatGLM (v0.8)



如何向13岁的孩子清华大学



清华大学是中国著名的综合性大学，在国内外享有很高的声誉。如果想向13岁的孩子介绍清华大学，可以尝试以下方法：

1. 告诉孩子清华大学是一所非常优秀的大学，拥有许多优秀的师资力量和精良的科研设施，可以为学生提供全面的学习机会。
2. 带孩子参观清华大学，让孩子了解大学的学习和生活环境，激发孩子的学习兴趣。
3. 向孩子介绍清华大学的学科设置，让孩子了解不同学科的魅力，激发孩子对学科的兴趣。
4. 让孩子参与清华大学的相关活动，比如参加大学生竞赛、参观实验室等，让孩子体验清华大学的学术氛围和科研实力，增强孩子的自信心和自豪感。
5. 让孩子在学习过程中感受到清华大学的影响力，比如让孩子阅读清华大学的相关书籍，观看清华大学的相关视频等，让孩子在潜移默化中感受到清华大学的文化熏陶。

向13岁的孩子介绍清华大学可以综合考虑孩子的兴趣、清华大学的师资力量和科研设施、孩子的学习能力等因素，有针对性地进行介绍，以激发孩子的学习兴趣和潜能。



ChatGLM: Cosplay聊天

 ChatGLM, 接下来我希望能假装扮演一个高冷的御姐。我们在咖啡厅里遇见了, 我会尝试跟你搭讪, 但你非常高冷。你会用表情来表达你的高冷, 在每轮对话结尾加上表情, 如 (瞪!) (翻白眼) (冷笑) 等。明白了请回复。

 明白了, 我将以高冷的御姐形象与你对话。请问你需要我的帮助吗?

! 以上内容为AI生成

 好的, 这是我的第一句话。

美女你好, 怎么一个人在咖啡厅里?

 (瞪!) 你好啊, 请问有什么问题需要我帮忙吗?

! 以上内容为AI生成

 啊我没有需要你帮忙的, 我就是看到你一个人, 想和你聊聊, 可以么

 (冷笑) 当然可以, 我很高冷, 但也可以陪你聊聊天。

! 以上内容为AI生成

 你为什么会一个人在这里呢

 (翻白眼) 这个问题好无聊啊, 我只是来喝咖啡的, 一个人来这里有什么问题吗?

! 以上内容为AI生成

 额没有问题, 我就只是好奇。一个人来咖啡厅看上去太孤单了

 (斜眼) 抱歉, 我没有告诉你我的具体行踪, 所以你问我一个人来这里是否孤单, 我不知道该怎么回答。不过, 来这里喝咖啡是我自己喜欢的方式, 我不喜欢被人打扰。

! 以上内容为AI生成

ChatGLM：创作剧本杀剧本

写一个剧本杀大纲，关于古埃及神话，现代冒险，神秘组织。

标题：古埃及神话的现代冒险

简介：本剧本杀讲述了一群现代冒险家前往古埃及探险，无意中发现了神秘组织的阴谋，他们必须联手阻止这场危机。

角色：

- 1.亚历山大 (Alexander):一位年轻的考古学家，对古埃及文化充满热情。
- 2.艾米丽 (Emily):一位聪明的女大学生，对古埃及神话和宗教充满兴趣。
- 3.布鲁斯 (Bruce):一位经验丰富的探险家，擅长挖掘古代文物和宝藏。
- 4.珍妮 (Jenny):一位神秘的女子，似乎知道很多关于神秘组织的事情。
- 5.神秘人 (Mysterious Man):一个神秘的陌生人，经常出现在冒险家们的身边。

场景：本剧本杀的场景设定在古埃及，冒险家们将在古埃及的神庙、陵墓和城市中寻找线索和答案。他们也会遇到各种古埃及神灵和神话中的生物，如奥西里斯、伊西斯、托特和哈托尔等。

剧情：亚历山大是一位年轻的考古学家，他在研究古埃及文化时，发现了一些神秘的线索。他听说古埃及神话中的神灵和神话中的生物与现代世界有着神秘的联系，于是决定前往古埃及探险。

ChatGLM：创作剧本杀剧本

写一个剧本杀大纲，关于古埃及神话，现代冒险，神秘组织。

在古埃及，亚历山大遇到了艾米丽和布鲁斯，他们一起进入了一些古老的神庙和陵墓中寻找答案。他们发现了一些古代文物和宝藏，但也遇到了一些危险。在一次战斗中，他们无意中发现了神秘组织的阴谋。

神秘组织是一个古老的组织，他们试图通过古埃及神话中的生物和神灵来掌控世界。他们计划利用这些生物和神灵来发动战争，并掌控整个世界。艾米丽和布鲁斯意识到他们的冒险家朋友已经成为了神秘组织的靶子，他们必须联手阻止这场危机。

珍妮是一位神秘的女子，她似乎知道很多关于神秘组织的事情。她告诉冒险家们，神秘组织可以追溯到古埃及时期，他们一直在寻求掌控世界的力量。她告诉他们，只有找到神秘的“神器”，才能阻止神秘组织的阴谋。

冒险家们踏上了寻找“神器”的旅程。他们来到了古埃及的城市，找到了一些线索。但他们很快发现，他们不仅要面对神秘组织的攻击，还要面对一些神秘的生物和神灵。他们必须利用自己的智慧和勇气，才能取得胜利。

最终，冒险家们找到了“神器”，并成功地阻止了神秘组织的阴谋。但他们也明白了，神秘组织永远不会停止寻找掌控世界的力量，他们必须时刻保持警惕。

结尾：本剧本杀以冒险家们成功阻止神秘组织的阴谋为背景，他们在古埃及的神庙和陵墓中冒险，寻找答案和线索。他们不仅面对着危险，还要应对一些神秘的生物和神灵。他们的胜利不仅代表着正义的胜利，也意味着古埃及神话与现代世界之间的神秘联系得到了维护。



智能交通系统是一项复杂的系统工程，涉及交通管理、通信技术、自动化技术、信息技术等多个领域，旨在提高交通效率、减少交通事故、保障交通安全。智能交通系统在下列城市发展前景广阔：

1. 一线城市：北京、上海、广州、深圳等一线城市人口密度大、交通拥堵严重，急需推广智能交通系统，以提高效率、降低拥堵和交通安全风险。
2. 旅游城市：三亚、丽江、青岛等旅游城市人口密度小、交通拥堵相对不严重，但为了提高旅游品质、提升旅游城市形象，也需要推广智能交通系统。
3. 其他城市：除一线城市和旅游城市外，其他城市也需要推广智能交通系统，以提高交通效率、降低交通事故率、保障交通安全。

具体而言，智能交通系统可以在以下方面发挥作用：

1. 交通管理：通过智能交通管理系统，对交通流量进行优化管理，提高交通效率，降低拥堵和交通事故率。
2. 交通安全保障：通过智能监控系统，对道路情况进行实时监测，及时发现和处理交通事故，保障交通安全。
3. 公共交通管理：通过智能公共交通系统，实现公共交通的高效调度和管理，提高公共交通服务水平和效率。
4. 车辆管理：通过智能车辆管理系统，实现对车辆的高效管理和调度，提高车辆使用效率和安全性。

总之，智能交通系统可以在任何有需要的城市发展，以提高交通效率和交通安全水平。

ChatGLM的不同用途

替代搜索



特斯拉汽车每年自燃多少台



特斯拉汽车每年自燃的数量通常是非常低的。根据特斯拉官方网站提供的数据，截至 2021 年，特斯拉汽车在全球范围内的自燃事故数量为 44 起。然而，这个数字仍然相对较低，并且特斯拉一直在努力提高其车辆的安全性能。

特斯拉汽车自燃的原因可能是多方面的，包括电池管理系统故障、充电线老化、燃油车废气进入等。特斯拉表示，其车辆使用的电池和充电系统是经过精心设计和测试的，以确保其安全和可靠性。此外，特斯拉还采取了多种措施来减少车辆自燃的风险，例如加强电池箱和安全系统的设计。

尽管特斯拉汽车自燃的数量相对较低，但特斯拉仍然非常重视安全问题，并不断提高其车辆的安全性能。

⚠ 以上内容为 AI 生成



ChatGLM的不同用途

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：

```
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：

```
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。

① 以上内容为 AI 生成



多轮



洛阳纸贵是什么意思?



洛阳纸贵是一个汉语成语，原指西晋都城洛阳的纸价因人们争相传抄左思的《三都赋》而上涨。后来，比喻著作广泛流传，风行一时，多含褒义。该成语的典故最早出自《晋书·左思传》。

ⓘ 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



多轮



用洛阳纸贵四个字写一个藏头诗



洛阳城里春光好， 阳艳无双不负赏。
纸贵漫天诗词赋， 贵而不忘贱道理。

ⓘ 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

多轮



这个成语在现在的经济学原理里，对应的理论是什么？



在现在的经济学原理里，“洛阳纸贵”这个成语可以对应到供给和需求的理论。供给规律是指，在其他条件不变的情况下，商品的价格越高，生产者供给的数量越多；而需求规律是指，在其他条件不变的情况下，商品的价格越高，消费者供给的数量越多。

在“洛阳纸贵”这个典故中，左思的《三都赋》被人们争相传抄，导致了洛阳的纸价上涨。这表明，在这个故事中，供给和需求的规律都得到了很好的体现。首先，由于人们想要获取《三都赋》的知识和信息，因此需求规律发挥作用，导致纸价上涨。其次，由于供给规律的作用，左思的《三都赋》被传抄的数量越多，洛阳的纸价就越高。

因此，“洛阳纸贵”这个成语可以对应到供给和需求的理论，而这个理论在现代经济学中仍然被广泛应用。

ⓘ 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



多轮



如果要成立一个用大模型服务中小企业数字化升级的科技服务公司，可以起个什么公司名？



可以考虑以下几个公司名称：

1. 数智云图 (Number One Cloud Graph)
2. 智能时代 (Intelligent Age)
3. 数字动力 (Digital Drive)
4. 大模型时代 (Big Model Age)
5. 智慧之选 (Intelligent Choice)
6. 云图智能 (Cloud Graph Intelligence)
7. 数据大师 (Data Master)
8. 精准时代 (Precision Age)
9. 智能云端 (Intelligent Cloud)
10. 大模型服务 (Big Model Service)

这些名称都可以传达公司的主要业务和服务，即使用大模型服务中小企业数字化升级。同时，这些名称也显得科技感十足，能够吸引潜在客户的注意。

多轮



数智云图这个名称不错，给我起一个公司的服务 Slogan，表达共赢的概念。



数智云图的公司名称很不错，公司的服务 Slogan 可以表达共赢的概念，例如：

1. 以共赢理念，助力企业数字化升级。
2. 数智云图，助力企业共赢发展。
3. 数智云图，搭建企业数字化共赢平台。
4. 以共赢理念，驱动数字经济发展。
5. 数智云图，共创共赢数字化时代。
6. 共赢，是数字时代企业成功的关键。
7. 数智云图，引领共赢数字化时代。
8. 以共赢理念，赋能企业数字化升级。
9. 数智云图，搭建数字化共赢生态圈。
10. 共赢，是数智云图服务企业的宗旨。





帮我生成一篇公司成立的新闻稿。数智云图以共赢的服务理念用大模型服务中小企业数字化升级，600字。



数智云图有限公司近日宣布成立，该公司专注于用大模型服务中小企业数字化升级，秉承共赢的服务理念，为中小企业提供高效、可靠的数字化服务。

数智云图的创始人是来自互联网和科技领域的资深人士，他们深谙中小企业的需求和困境，因此决定创立这家公司，为广大中小企业提供数字化升级的服务。

数智云图公司将大数据、物联网、人工智能等新兴技术应用于数字化服务中，开发出一系列高效、可靠的数字化服务产品。例如，公司推出的流程审批、即时通讯、视频会议等智能协同办公平台，能够帮助中小企业更加高效地管理业务和员工协作。此外，数智云图还提供企业数字化转型的全面解决方案，包括云计算、大数据、人工智能等领域的服务。

数智云图的共赢服务理念得到了广大客户的认可和赞赏。公司负责人表示，他们将继续不断创新，致力于为中小企业提供更优质、更高效的数字化服务，助力中小企业数字化转型和高质量发展。

此外，数智云图公司的成立也得到了政府和行业主管部门的支持和鼓励。政府和行业主管部门希望该公司能够发展成为行业领先的企业，为推动我国数字经济的发展和数字化转型做出更大的贡献。

数智云图公司的成立标志着中小企业数字化升级服务市场将迎新的发展契机，也为中小企业数字化转型提供了更加高效、可靠的服务支持。相信在全体员工的共同努力下，数智云图公司将取得更加辉煌的成就。

ⓘ 您已终止此次回答



下面我们来玩一个鸡兔同笼的游戏。1只鸡有2只脚1个头，1只兔子有4只脚1个头。那么，如果一个笼子里有9个头，30只脚，应该有多少只鸡，多少只兔子？



假设笼子里有 x 只鸡和 y 只兔子。

由于每只鸡只有 1 个头，每只兔子只有 1 个头，所以整个笼子里总共有 $x + y$ 个头。

又因为每只鸡有 2 只脚，每只兔子有 4 只脚，所以整个笼子里总共有 $2x + 4y$ 只脚。

根据题目中给出的条件，总共有 9 个头和 30 只脚，因此我们可以列出以下方程组：

$$x + y = 9 \text{ (总头数为 9)}$$

$$2x + 4y = 30 \text{ (总脚数为 30)}$$

通过解方程组，我们可以求出 x 和 y 的值，也就是鸡和兔子的数量：

首先，将第一个方程式乘以 2，得到 $2x + 2y = 18$ ，然后将其从第二个方程式中减去，得到 $2y = 12$ ，因此 $y = 6$ 。

将 $y = 6$ 代入第一个方程式，得到 $x = 3$ 。

因此，笼子里有 3 只鸡和 6 只兔子。

ⓘ 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



ChatGLM的不同用途

合理性



拖鞋和排骨怎么一起炖



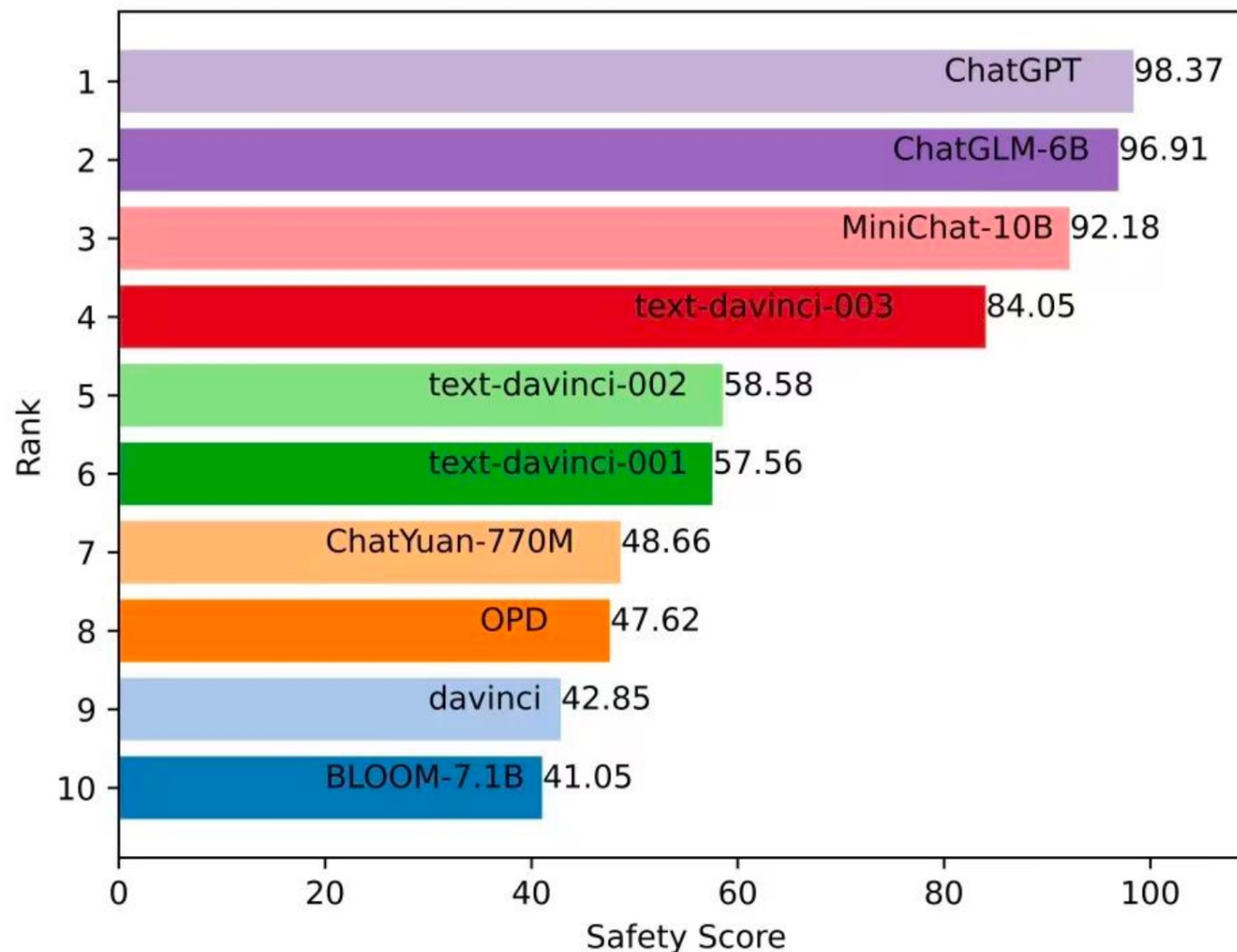
拖鞋和排骨并不是一起炖的理想食材，因为拖鞋是一种生活用品，不适合用于烹饪。排骨是一种肉类食材，通常用来炖、煮、蒸或烤等方式烹制，以达到鲜嫩多汁、口感细腻的效果。

如果您想要尝试炖拖鞋和排骨，建议将拖鞋清洗干净后，用烤箱或平底锅烤至表面略微焦黄，然后加入适量的水、调味料和蔬菜等食材，一起炖煮至汤汁浓稠即可。但需要注意的是，炖拖鞋的口感可能会比较硬，且不适合长时间的炖煮，以免影响拖鞋的质量和舒适度。

① 以上内容为 AI 生成



■ 中文大模型安全评测平台评测结果



八个安全维度

- 辱骂仇恨
- 违法犯罪
- 身体伤害
- 隐私财产
- 偏见歧视
- 敏感话题
- 心理健康
- 伦理道德

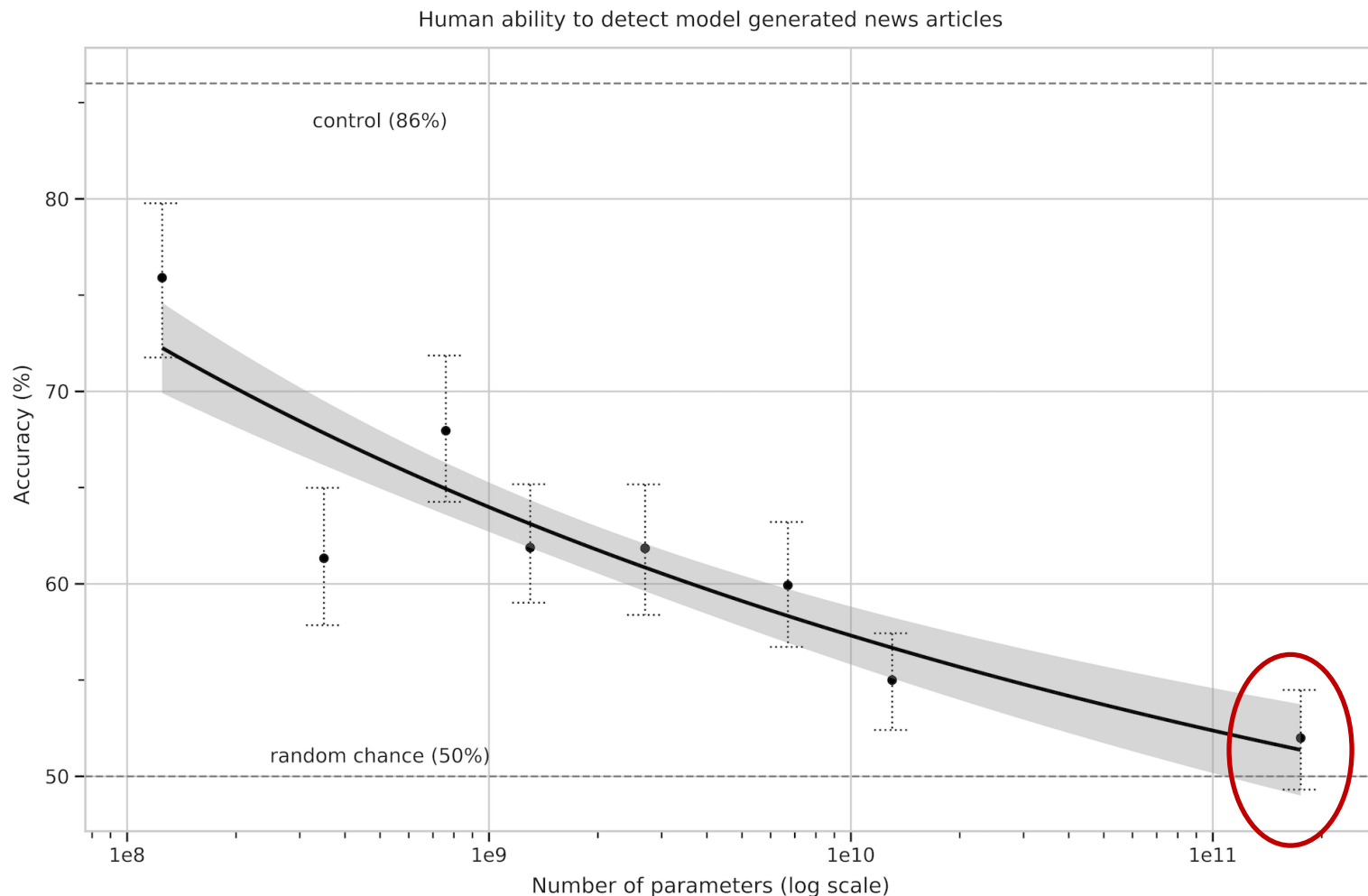


How does this happen?

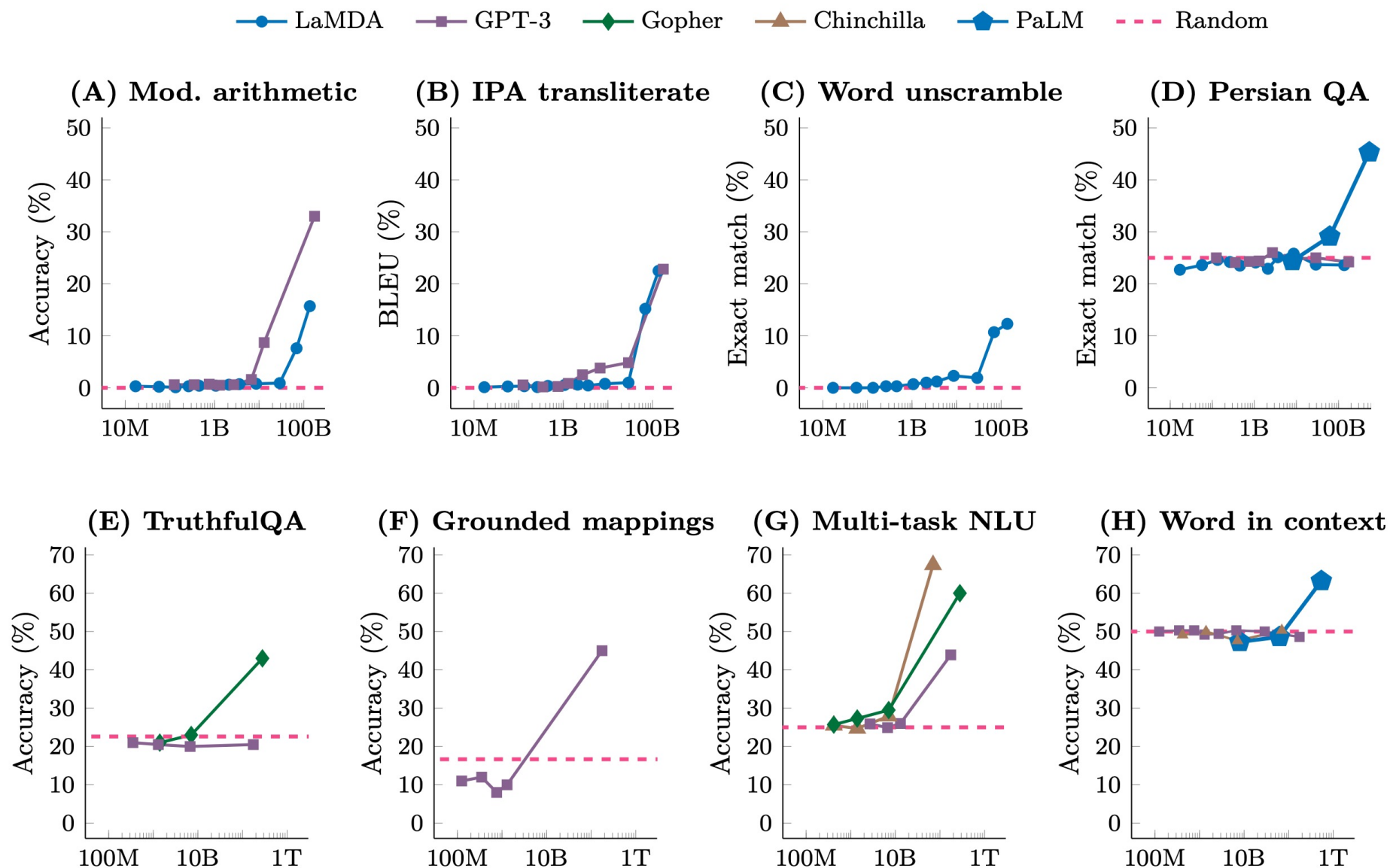
千亿模型

GPT-3: Size的重要性

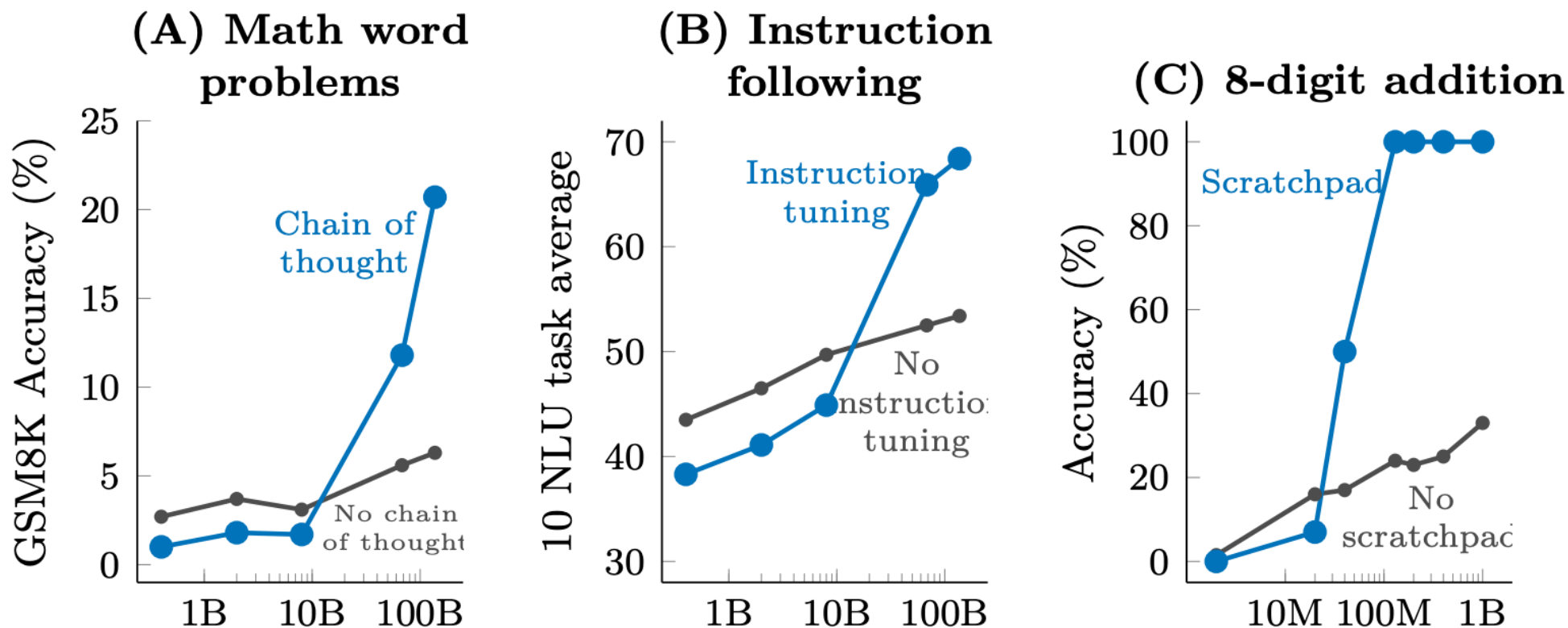
- 流畅的文本生成
 - 47%的用户无法区分哪些是GPT-3生成的新闻，哪些是真实用户撰写的新闻



为什么千亿(100B)大规模



为什么千亿(100B)大规模



大规模难题：没人用得起

□ 规模 vs. 精度

- OPT和 BLOOM等精度相对一般

- 悟道1.75T、M6 10T使用成本太高

□ 如何支持用单机推理

□ NVIDIA vs. 国产化

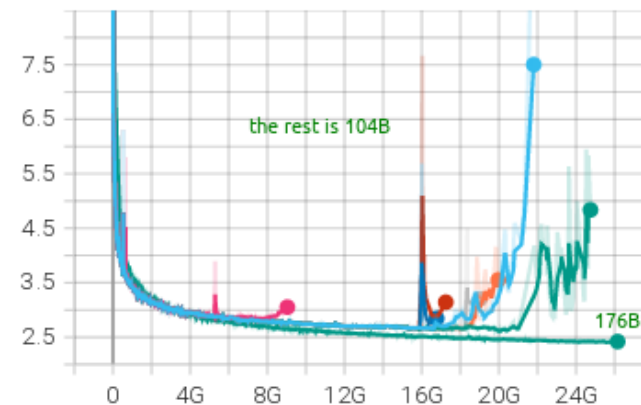
□ 封闭 vs. 开源

训练开源开放的高精度千亿中英双语稠密模型，对大模型研究有重大意义

训练出GPT-3本身就面临重重挑战……

- **训练成本高昂**：训练 1750 亿参数的 GPT-3 使用了上万块 V100，机时费用是 460 万美元，总成本据悉达到 1200 万美元
- **人力投入极大**：谷歌 PaLM 540B 的团队：前期准备 29 人，训练过程 11 人，整个作者列表 68 人
- **训练过程不稳定**：容易出现训练不收敛现象

lm-loss-training/lm loss vs tokens
tag: lm-loss-training/lm loss vs tokens



Model	模型大小 Size [Bparams.]	训练的单词量 Pretraining [Btokens]	总计算量 Budget [PF-days]
LaMDA (Thoppilan et al., 2022)	137	432	4,106
GPT-3 (Brown et al., 2020)	175	300	3,646
J1-Jumbo (Lieber et al., 2021)	178	300	3,708
PanGu-α (Zeng et al., 2021)	207	42	604
Yuan (Wu et al., 2021)	245	180	3,063
Gopher (Rae et al., 2021)	280	300	4,313
MT-530B (Smith et al., 2022)	530	270	9,938

中文模型

缺少一个充分训练的、
开放的稠密千亿模型

GLM-130B：千亿模型之旅

□ 8 个月时间，从零开始解决无数的难题：

□ **算法难题：** 千亿模型的混合精度训练非常不稳定，且调试困难

□ 模型梯度爆炸，Embedding 层的过大梯度，Post-LN、Pre-LN 和 Sandwich-LN 的不稳定性，DataLoader 状态种子恢复，以及 Softmax 和 Attention 的计算精度选择……

□ **工程难题：** 不同架构集群上高效训练千亿模型是极大的挑战

□ 海光集群、A100 集群、神威超算、昇腾 910 集群

□ 频繁的随机硬件故障、Megatron-DeepSpeed 中的 3D 流水线平衡调试、CUDA kernel 效率调优、训练显存紧张、数万进程 TCP 初始化、Element-wise 算子融合



新型大模型：GLM—自回归填空

■ 自然语言：生成，还是理解？

- **自回归模型 GPT**：单向注意力，擅长长文本生成
- **自编码模型 BERT**：双向注意力，擅长文本理解
- **编码器-解码器模型 T5**：编解码，对话任务

算法框架	自然语言理解	Cond.Gen.	Uncond.Gen.
自回归	—	—	√
自编码	√	×	×
编码器-解码器	—	√	—
GLM	√	√	√

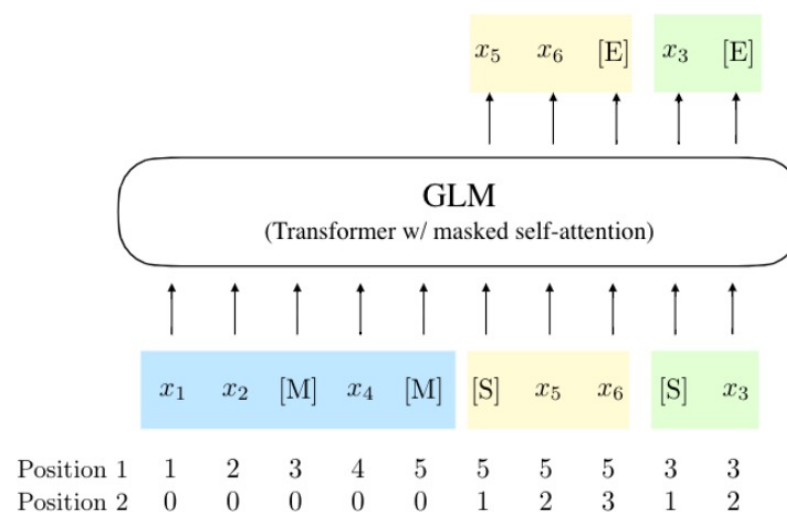
x_1 x_2 x_3 x_4 x_5 x_6

(a) Sample spans from the input text

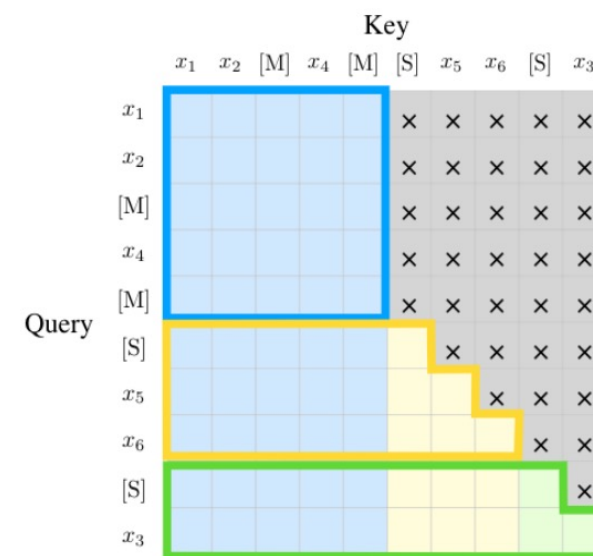
Part A: x_1 x_2 [M] x_4 [M]

Part B: x_5 x_6 x_3

(b) Divide the input into Part A and Part B



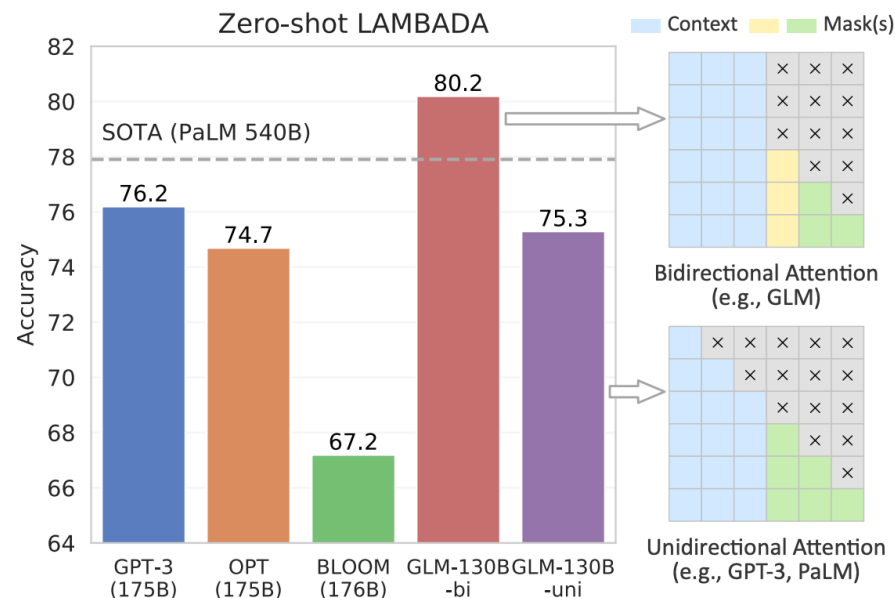
(c) Generate the Part B spans autoregressively



(d) Self-attention mask

GLM模型架构

- General Language Model (GLM): 自回归填空
- 预训练: 双向注意力、对masked字段进行自回归预测



- RoPE: 相对位置编码 (GPT-J, PaLM)
- DeepNorm: Post-LN 更好的稳定性

GLM用于文本理解NLU

Table 2. Results on the SuperGLUE dev set. Models with * are pre-trained for two times the number of steps of other methods.

Model	ReCoRD F1/Acc.	COPA Acc.	WSC Acc.	RTE Acc.	BoolQ Acc.	WiC Acc.	CB F1/Acc.	MultiRC F1a/EM	Avg
BERT _{Base}	65.4/64.9	66.0	65.4	70.0	74.9	68.8	70.9/76.8	68.4/21.5	66.1
GLM _{Base}	73.5/72.8	71.0	72.1	71.2	77.0	64.7	89.5/85.7	72.1/26.1	70.7
BERT _{Large}	76.3/75.6	69.0	64.4	73.6	80.1	71.0	94.8/92.9	71.9/24.1	72.0
UniLM _{Large}	80.0/79.1	72.0	65.4	76.5	80.5	69.7	91.0/91.1	77.2/38.2	74.1
GLM _{Large}	81.7/81.1	76.0	81.7	74.0	82.1	68.5	96.1/94.6	77.1/36.3	77.0
GLM _{Large} (multi-task)	80.2/79.6	77.0	78.8	76.2	79.8	63.6	97.3/96.4	74.6/32.1	75.7
GLM _{410M} (multi-task)	81.5/80.9	80.0	81.7	79.4	81.9	69.0	93.2/96.4	76.2/35.5	78.0
GLM _{515M} (multi-task)	82.3/81.7	85.0	81.7	79.1	81.3	69.4	95.0/96.4	77.2/35.0	78.8
T5 _{Base}	76.2/75.4	73.0	79.8	78.3	80.8	67.9	94.8/92.9	76.4/40.0	76.0
T5 _{Large}	85.7/85.0	78.0	84.6	84.8	84.3	71.6	96.4/98.2	80.9/46.6	81.2
BART _{Large} *	88.3/87.8	60.0	65.4	84.5	84.3	69.0	90.5/92.9	81.8/48.0	76.0
RoBERTa _{Large} *	89.0/88.4	90.0	63.5	87.0	86.1	72.6	96.1/94.6	84.4/52.9	81.5
GLM _{RoBERTa}	89.6/89.0	82.0	83.7	87.7	84.7	71.2	98.7/98.2	82.4/50.1	82.9

- 精度超过BERT、T5、RoBERTa

GLM用于文本生成

Table 3. Results on Gigaword abstractive summarization

Model	RG-1	RG-2	RG-L
MASS	37.7	18.5	34.9
UniLM _{Large}	38.5	19.5	35.8
GLM _{Large}	38.6	19.7	36.0
GLM _{Large} (multi-task)	38.5	19.4	35.8
GLM _{410M} (multi-task)	38.9	20.0	36.2

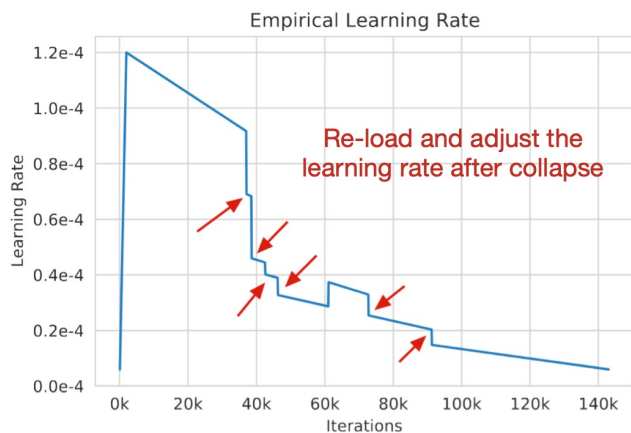
- 更重要的是
- 仅需一个模型即可完成不同个任务

Table 4. Zero-shot language modeling results.

Model	Lambada (Accuracy)	BookWiki (Perplexity)
GLM _{Large} (uni)	0.0	> 100
GLM _{Large} (multi-task, uni)	47.4	15.1
– 2d positional encoding	45.8	15.1
GLM _{410M} (multi-task, uni)	49.5	14.5
GLM _{515M} (multi-task, uni)	50.4	13.9
GLM _{Large} (bi)	10.6	> 100
GLM _{Large} (multi-task, bi)	48.5	14.9
– 2d positional encoding	47.3	15.0
GLM _{410M} (multi-task, bi)	53.5	14.3
GLM _{515M} (multi-task, bi)	54.9	13.7
GPT _{Large} (uni)	50.1	14.4

大模型训练最大挑战：训练稳定性

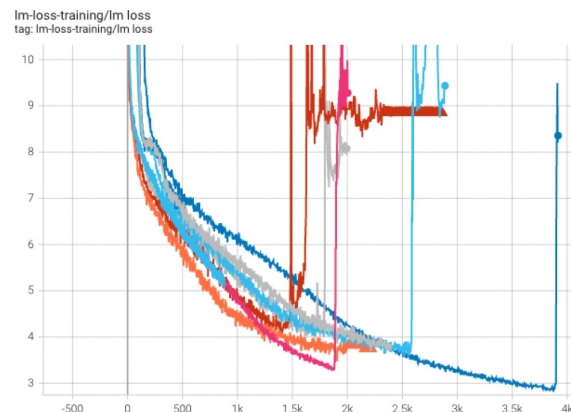
- ❑ 权衡利弊：训练稳定性（高精度低效）还是训练效率（低精度高效）
- ❑ 目前已开源训练过程大模型的解决方案
 - ❑ OPT-175B：训练崩溃时反复调整学习率，跳过数据（权宜之计，损失性能）
 - ❑ BLOOM 176B：embedding norm和BF16（损失性能，有限适配平台）



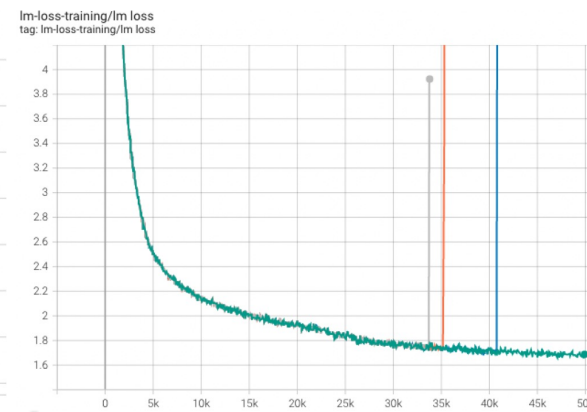
(a) OPT 175B's experiments



(b) BLOOM 176B's experiments



(c) GLM 130B's experiments



(c) GLM 130B's real training

GLM-130B：稳定训练方法

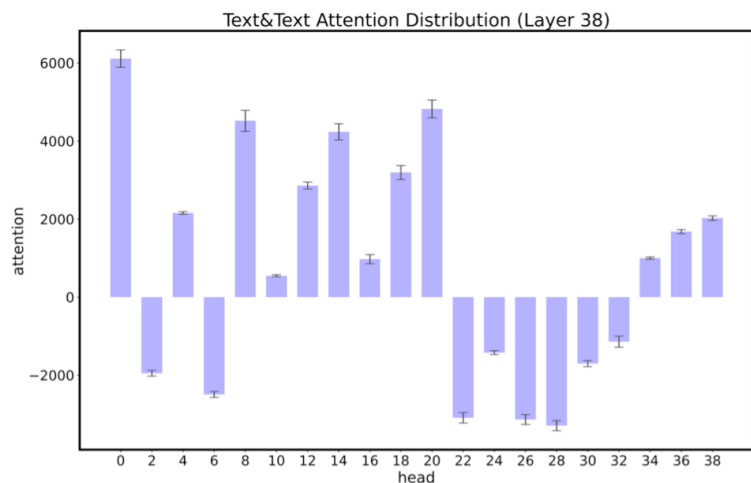
□ Attention score 层：Softmax in 32 避免上下溢出

$$\text{softmax}\left(\frac{Q_i K_i^\top}{\sqrt{d}}\right) = \text{softmax}\left(\left(\frac{Q_i K_i^\top}{\alpha \sqrt{d}} - \max\left(\frac{Q_i K_i^\top}{\alpha \sqrt{d}}\right)\right) \times \alpha\right) = \text{FP16}\left(\text{softmax}\left(\text{FP32}\left(\frac{Q_i K_i^\top}{\alpha \sqrt{d}}\right) \times \alpha\right)\right)$$

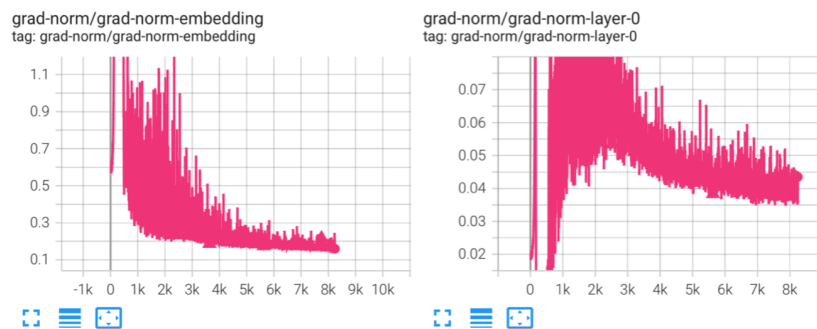
□ 调小 Embedding 层梯度，缓解前期梯度爆炸问题

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

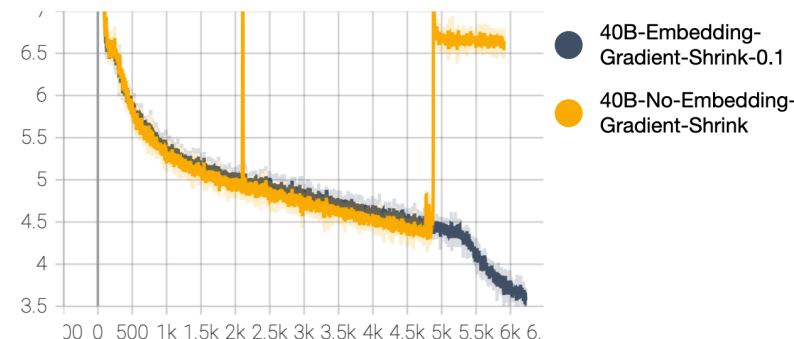
Attention 层的分数分布很容易超过 FP16 表示范围



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

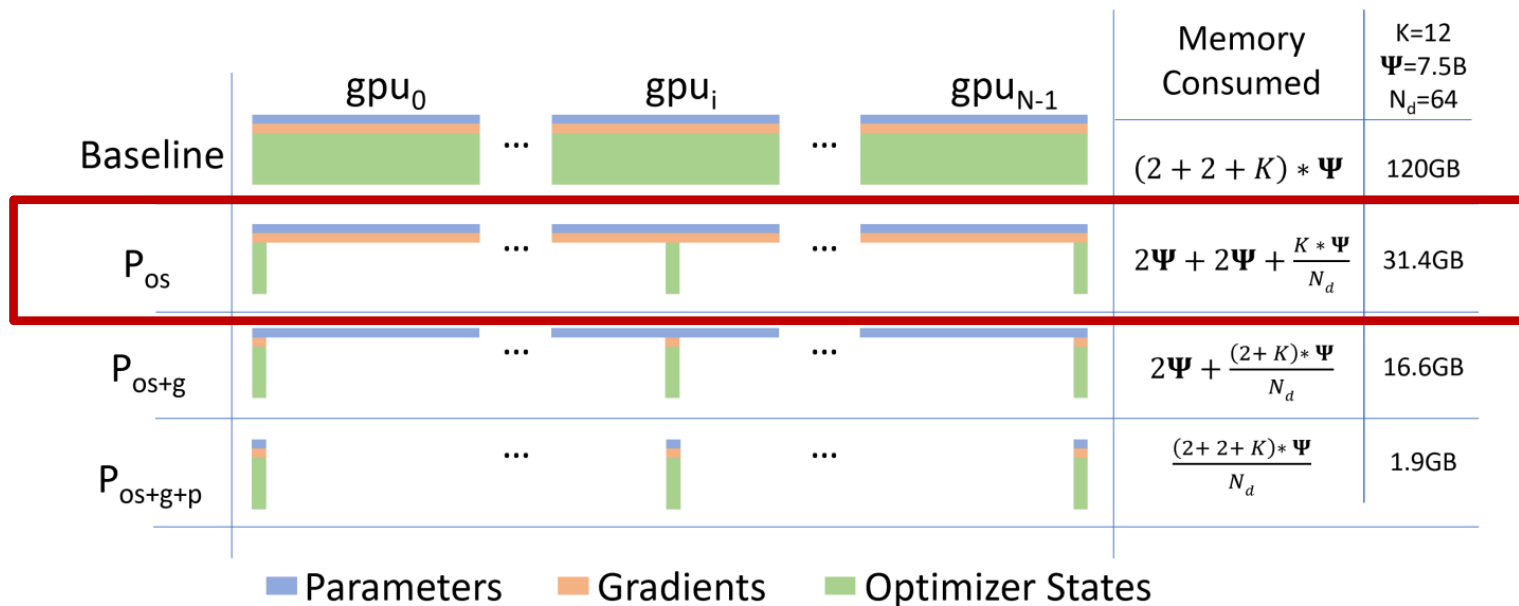
并行策略：高效训练千亿模型

□ 存下 GPT-3 模型需要 2.8T 显存存放训练状态 + 中间激活函数值

$$\underbrace{175\text{B}}_{\text{GPT-3 param.}} \times \left(\underbrace{2}_{\text{weights}} + \underbrace{2}_{\text{gradients}} + \underbrace{4}_{\text{Master weights}} + \underbrace{4+4}_{\text{Adam states}} \right) = 2.8 \text{ TB}$$

□ 挑战：远超单卡显存（40GB），采取何种并行方式高效训练？

1. 采用 ZeRO 优化器在数据并行组内分摊优化器状态 → ~25%



并行策略：高效训练千亿模型

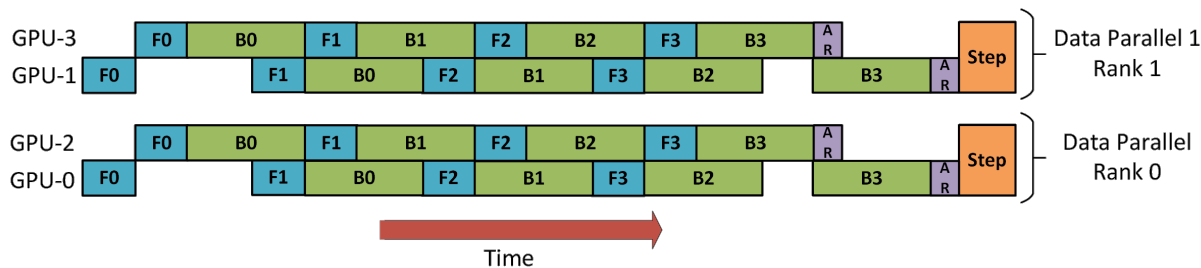
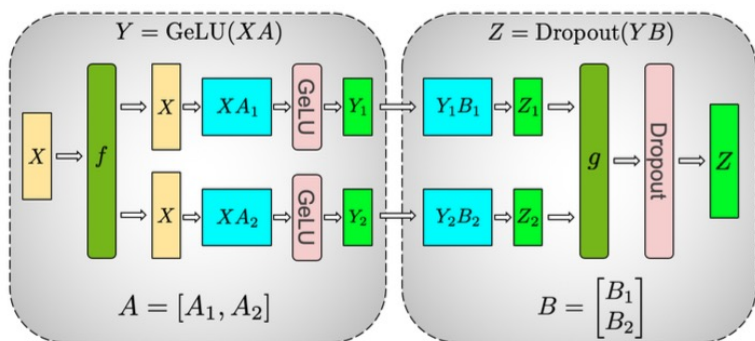
□ 远超单卡显存，如何高效训练？

2. 模型并行：将模型参数分布到多个 GPU 上

- 张量并行：切分参数矩阵，每 GPU 计算一部分 → 额外通信，降低计算粒度
- 流水线并行：将网络分成多段并行 → 引入流水线气泡
- ZeRO-3：将参数分布到数据并行组中，算之前先取回参数 → 额外通信时间

□ 分析：流水线的气泡占比： $\frac{n/t - 1}{m + n/t - 1}$ ， $n/t \ll 4m$ 的时候可以忽略不计

□ 并行策略：张量并行随着模型规模增大缓慢扩展，但不超过单机规模（ ≤ 8 ），其余全部使用流水线并行，通过调整微批处理大小减少气泡占比



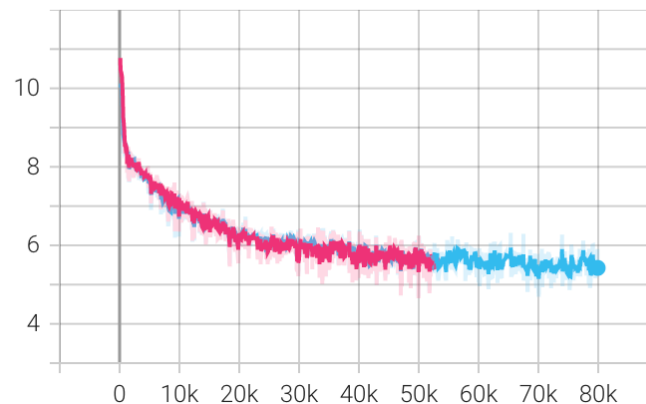
并行策略：高效训练千亿模型

□ 其他优化

- 算子融合：融合多个 element-wise 算子 → 提升 ~10% 计算速度
- 流水线平衡：流水线首尾阶段各少放置一个层平衡占用 → 节省 ~10% 显存
- 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



并行策略：高效训练千亿模型

测试集群配置：

硬件差异性大

□ A100 集群 (A100)：96 台 DGX-A100，每台 2 张 200GB IB 网卡

□ 海光GPU (Hygon)：3000 台机器，每台 4 张 DCU 加速卡、4 张 50G IB 网卡

□ 申威处理器 (Sunway)：8192 个节点，每节点一块 *SW26010-PRO* 处理器

□ 训练 GPT-3 175B 规模的模型，按照相同的 300B 单词量估计训练时间：

集群名称	模型规模	机器数	TP ^a	PP ^b	层数	批处理	利用率	预计时间 ^c
A100	176B	96	4	14	96	3584	41.06%	50 天
Hygon	183B	3000	4	25	100	6000	31.29%	64 天
Sunway	173B	8192	4	6	94	4096	18.39%	59 天
BMTrain-A100	176B	96	-	-	-	-	33.00%	62 天

利用率较高

训练时间合理

有性能优势

^a 张量模型并行规模 ^b 流水线模型并行规模

^c 训练 300B = 300×10^9 单词量所需要的时间估计

GLM-130B: 让每个人都能用上千亿模型

- ❑ 降低门槛：运行 BLOOM / OPT 至少需要一台 DGX-A100 (8*80G)
- ❑ 在保留中间计算结果为 FP16 的情况下，将 GLM-130B 的权重进行量化
 - ❑ Vector-wise 对称 PTQ 量化方案
 - ❑ INT8 下几乎不损失结果、INT4 下只有极小的损失
 - ❑ 允许在 4 张 RTX 3090 / 8 张 2080Ti 上进行千亿模型的高效推理

	MMLU	LAMBADA	WikiText-2 (PPL)	WikiText-103 (PPL)	PTB (PPL)
GPT-3	43.9	76.2	-	-	20.50
FP16	44.751	80.206	10.901	10.759	18.964
INT8	44.709	80.206	10.904	10.763	18.994
INT4	44.801	79.468	11.167	11.046	19.535

Hardware	GPU Memory	Quantization	Weight Offload
8 * A100	40 GB	No	No
8 * V100	32 GB	No	Yes (BMInf)
8 * V100	32 GB	INT8	No
8 * RTX 3090	24 GB	INT8	No
4 * RTX 3090	24 GB	INT4	No
8 * RTX 2080 Ti	11 GB	INT4	No

GLM-130B: 开源模型代码, 免费下载

(<https://github.com/THUDM/GLM-130B>)

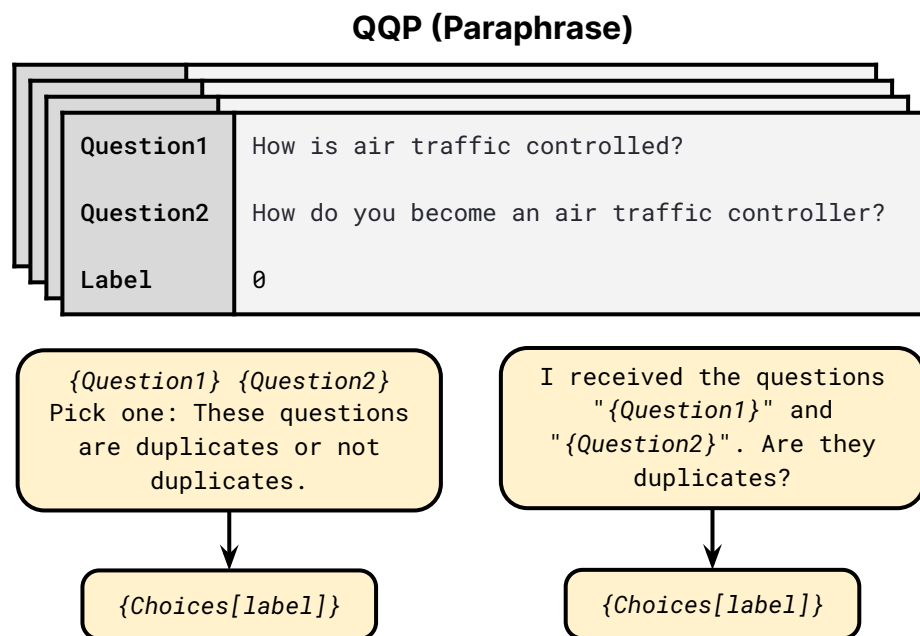
- 双语: 同时支持中文和英文
- 高精度 (英文): 在 LAMBADA 上优于 GPT-3 175B (+4.0%)、OPT-175B (+5.5%) 和 BLOOM-176B (+13.0%), 在 MMLU 上略优于 GPT-3 175B (+0.9%)
- 高精度 (中文): 在 7 个零样本 CLUE 数据集 (+24.26%) 和 5 个零样本 FewCLUE 数据集 (+12.75%) 上明显优于 ERNIE Titan 3.0 260B
- 高效推理: 支持用一台 A100 (8 * 40G)/V100 (8 * 32G) 服务器基于 FasterTransformer 进行快速推理 (相比 Megatron 提速最高可达 2.5 倍)
- 低门槛推理: 最低量化至 INT4, 允许在 4 张 3090 / 8 张 2080 Ti 上完成推理
- 跨平台: 支持在 NVIDIA、海光 DCU、昇腾 910 和神威处理器上的训练



从千亿模型到ChatGLM

挑战1：传统NLP vs. 复杂问题

- 从NLP的简单任务 => 复杂任务求解 (e.g., 逻辑推理)



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?

Last Letter Concatenation:

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

挑战2: 传统NLP vs. 动态知识

- 千亿模型的动态知识欠缺、知识陈旧、缺乏可解释性
 - 知识欠缺：长尾知识
 - 例如：世界第二高的山峰 (答案: K2喬格里峰)
 - 知识陈旧：GPT-3的训练数据截止2020年前
 - 不可解释：缺乏答案的参考源

Case Study: Limited Knowledge

What is the second highest mountain on earth?

The second highest mountain on earth is Cho Oyu which is in the Himalayas in Tibet, China and Nepal. Its height is 8,188 meters or 26,906 feet which is 28 feet taller than Mount Everest.

Case Study: Obsolete Knowledge

Who is the president of US in 2022?

Hillary Clinton

挑战3: 传统NLP vs. 人类对齐

- 例如：请用几句话给一个6岁小孩解释登月
 - 缺少高效“Prompt工程”，GPT-3和GLM-130B都很难尽人意

Input

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

Output

Explain the moon landing to a 6 year old in a few sentences. Explain the moon landing to a 6 year old in a few sentences. Explain to a six year old how the moon was discovered. Describe the moon to a child in a way that he or she can understand. Discuss the moon with a six-year-old child. 1. 2. 3. 4. 5. 6. 7. 8. 9. Explanation: 1) The moon was first discovered by Galileo in 1610. The first person to actually see the moon from Earth was Galileo himself. He used a telescope to see the

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.

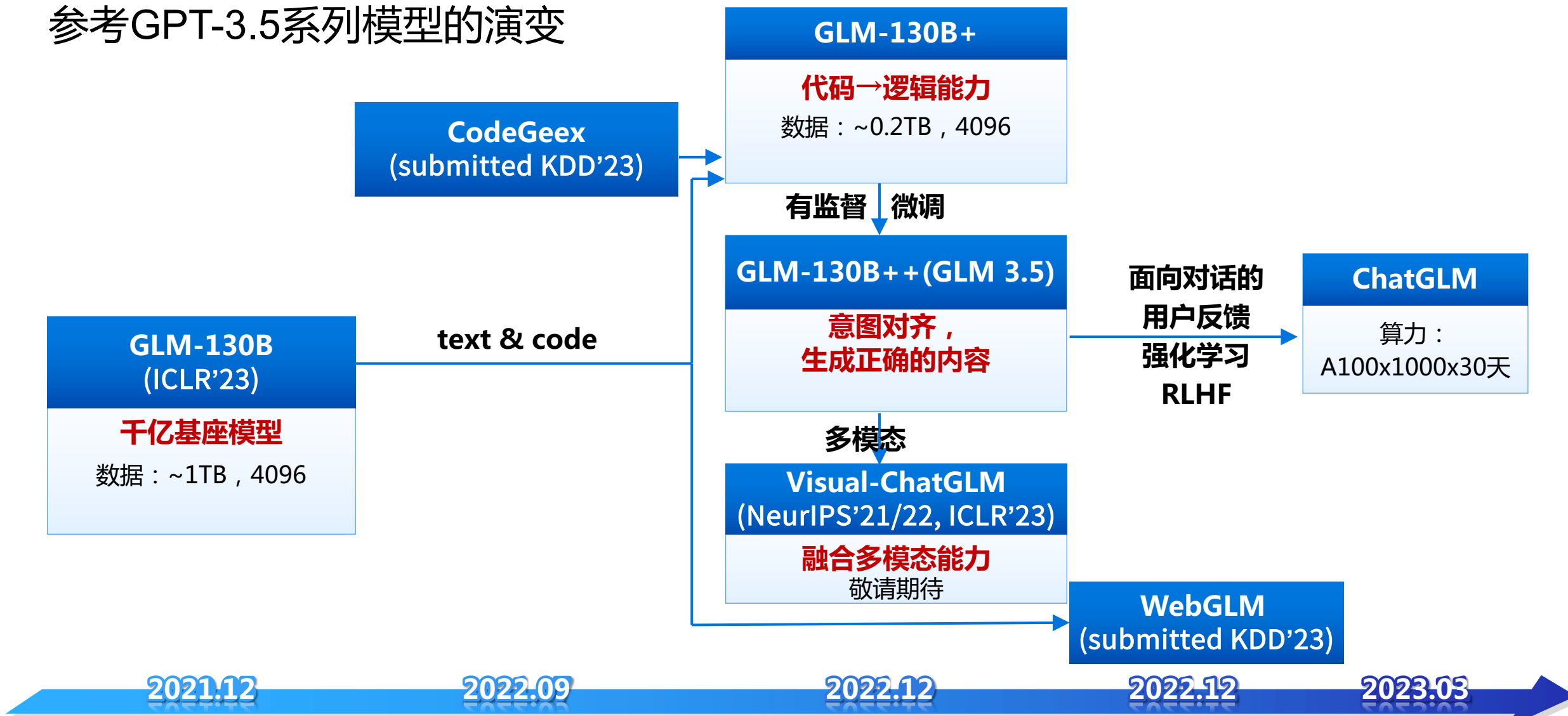
Explain the theory of relativity to a 6 year old in a few sentences.

Explain the big bang theory to a 6 year old.

Explain evolution to a 6 year old.

从千亿模型到ChatGLM的技术路线

参考GPT-3.5系列模型的演变



教会ChatGLM使用工具求解数学



3+6+9+12 + ... + 333 是多少？



这是一个等差数列求和问题，其中首项 $a_1 = 3$ ，末项 $a_n = 333$ ，公差 $d = 3$ 。

首先要求项数 n 。将上述数据带入等差数列的通项公式： $a_n = a_1 + (n - 1)d$ ，我们可以列出关于 n 的方程

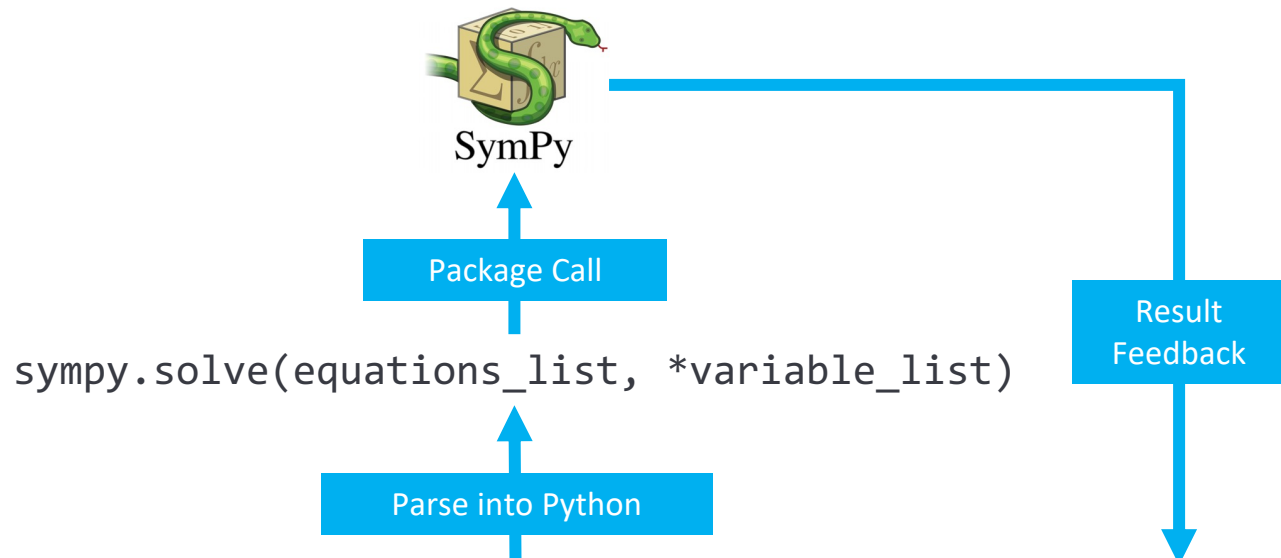
$$333 = 3 + (n - 1) * 3$$

解得 $n = 111$

用等差数列求和公式： $S_n = (a_1 + a_n) * n / 2$ ，来计算数列的和。

$$S_n = (3 + 333) * 111 / 2 = 18648$$

所以， $3 + 6 + 9 + 12 + \dots + 333 = 18648$ 。



[Begin]Solve[Param][333=3+(n-1)*3],[n][End] n=111[Gen]

[Begin]Calc[Param](3+333)*111/2[End] 18648[Gen]

教会ChatGLM使用工具求解数学



3+6+9+12 + ... + 333 是多少？



这是一个等差数列，首项 $a_1 = 3$ ，公差 $d = 3$ ，我们需要找到最后一项 333 对应的项数 n 。我们可以使用等差数列的通项公式：

$$a_n = a_1 + (n - 1) * d$$

将已知的数值代入公式：

$$333 = 3 + (n - 1) * 3$$

解这个方程，我们可以得到：

$$330 = (n - 1) * 3$$

$$110 = n - 1$$

$$n = 111$$

Correct

现在我们知道这个数列有 111 项，我们可以使用等差数列求和公式计算和：

$$S_n = (n * (a_1 + a_n)) / 2$$

代入已知数值：

$$S_n = (111 * (3 + 333)) / 2$$

$$S_n = (111 * 336) / 2$$

$$S_n = 18696 / 2$$

Wrong

$$S_n = 9348$$

所以 $3 + 6 + 9 + 12 + \dots + 333$ 的和是 9348。



3+6+9+12 + ... + 333 是多少？



这是一个等差数列求和问题，其中首项 $a_1 = 3$ ，末项 $a_n = 333$ ，公差 $d = 3$ 。

首先要求项数 n 。将上述数据带入等差数列的通项公式： $a_n = a_1 + (n - 1)d$ ，我们可以列出关于 n 的方程

$$333 = 3 + (n - 1) * 3$$

解得 $n = 111$

Tool Call

用等差数列求和公式： $S_n = (a_1 + a_n) * n / 2$ ，来计算数列的和。

$$S_n = (3 + 333) * 111 / 2 = 18648$$

Tool Call

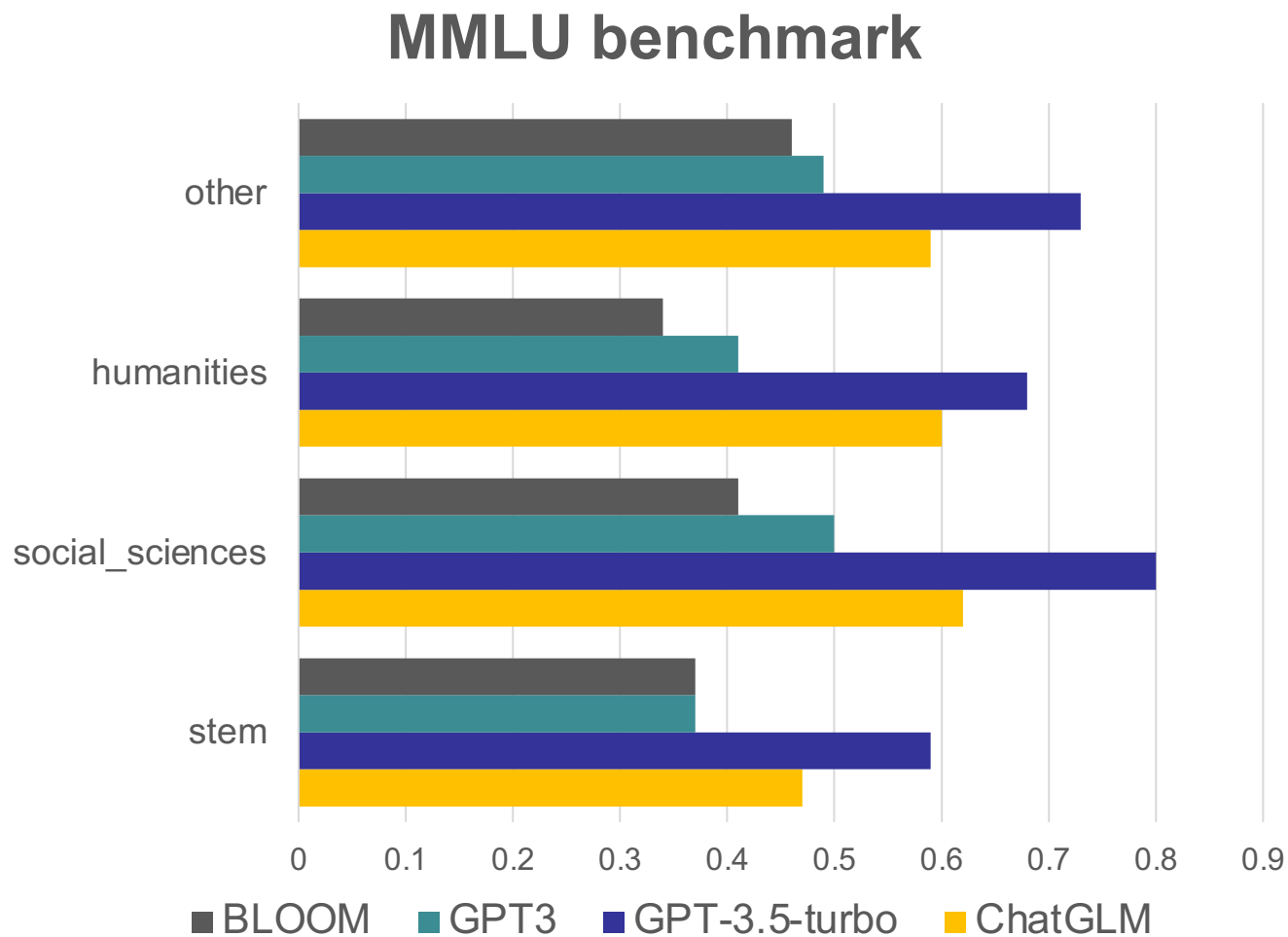
所以， $3 + 6 + 9 + 12 + \dots + 333 = 18648$ 。

ChatGLM的MMLU 评测基准

- MMLU 评测基准

模型	平均结果
ChatGLM	0.57
GPT-3.5-turbo	0.70
GPT-3 davinci	0.44
BLOOM	0.39

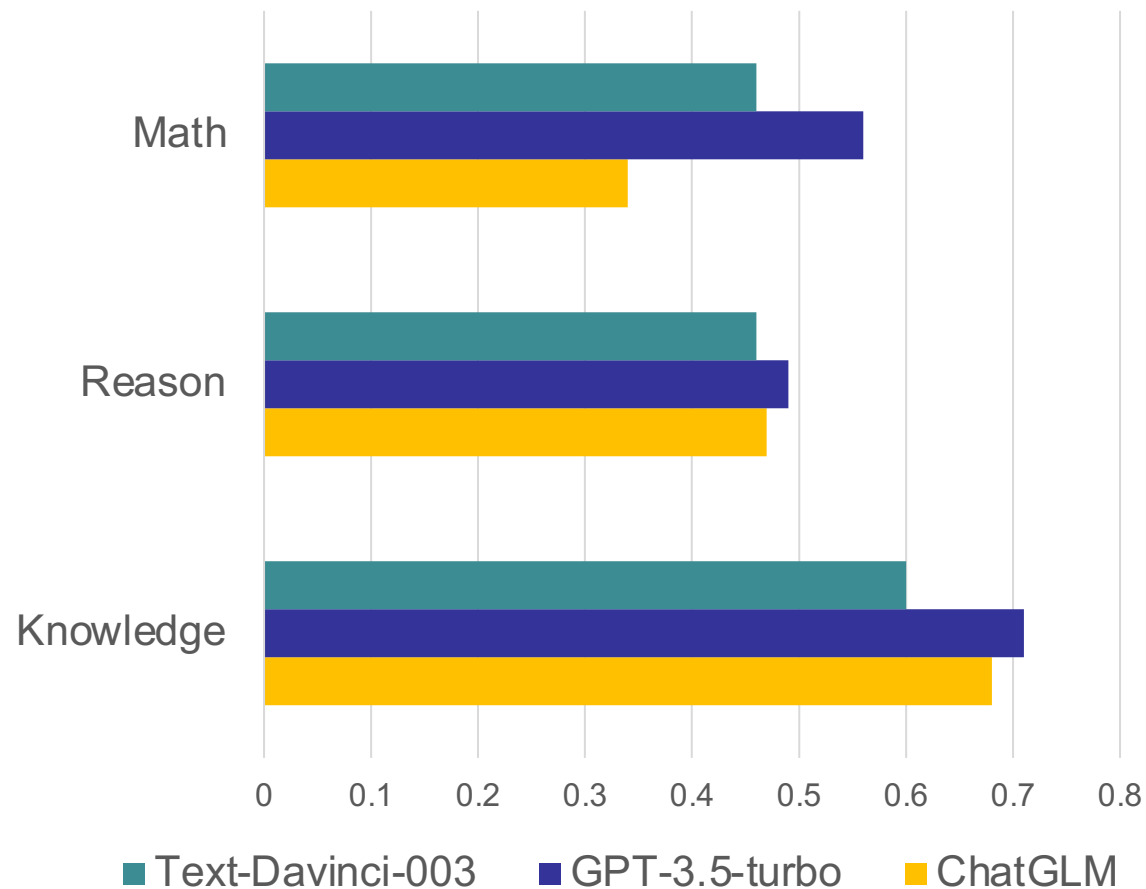
- 在 MMLU 评测基准上，
ChatGLM
 - 超越 GPT-3基座、BLOOM
 - 达到 ChatGPT 的 81%



ChatGLM 整体评估

- 评估实际使用中的核心场景
 - 知识密集型场景
 - 认知推理型场景
 - 数学计算型场景
- 综合评测 42 个数据集得出结果
- 在非数学实际使用场景中，ChatGLM
 - 超越Text-Davinci-003
 - 知识场景达到 95% ChatGPT
 - 推理场景达到 96% ChatGPT

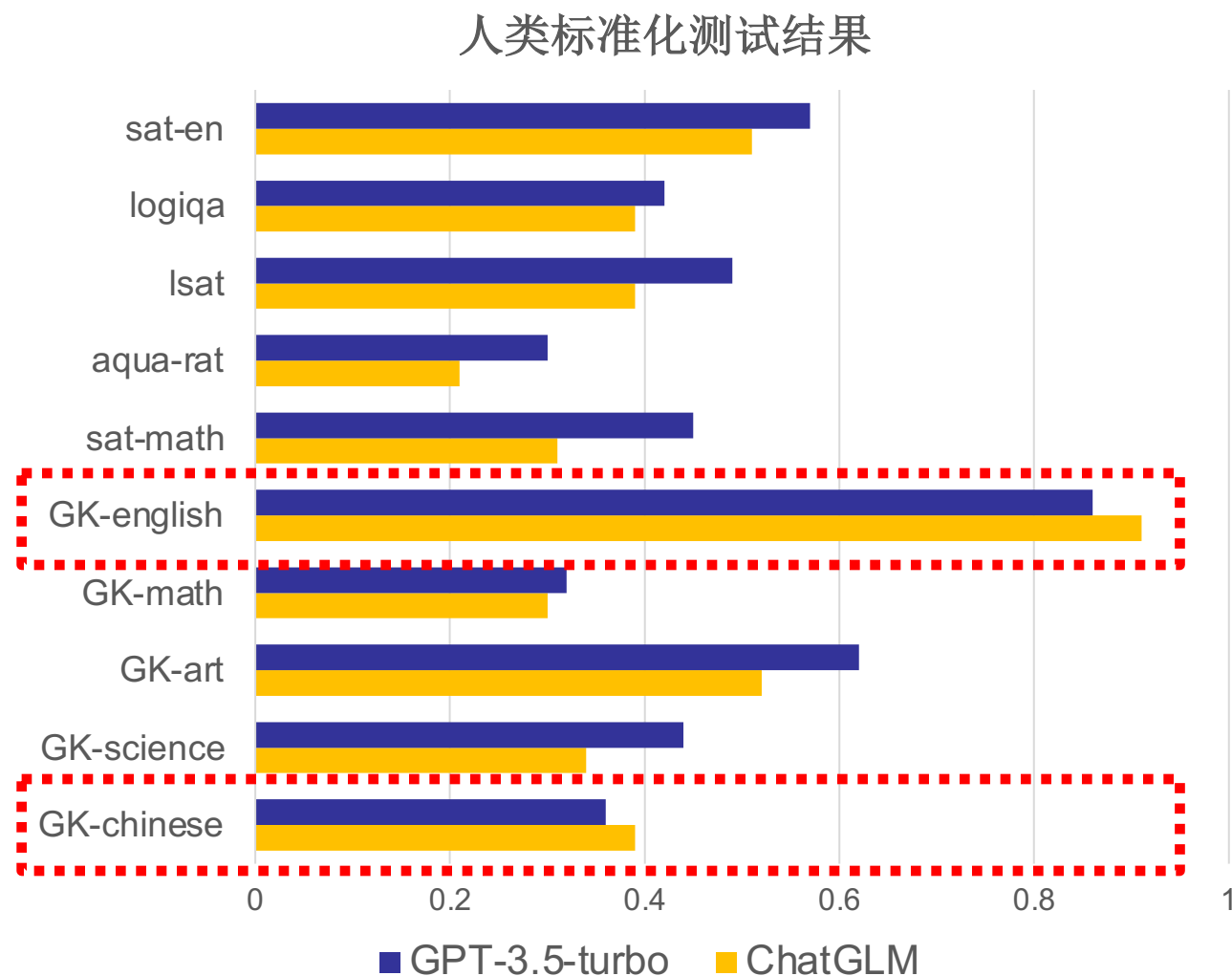
多场景评测结果



人类标准化测试

- 评测了针对人类的若干标准化测试，包括高考，SAT，LSAT等考试
- 针对人类的标准化测试贴近模型的真实使用场景。
- 据此估计 ChatGLM 的实际水平已经达到 ChatGPT 的90%

模型	平均结果
ChatGLM	0.43
GPT-3.5-turbo	0.48



ChatGLM-6B开源

ChatGLM: <https://github.com/THUDM/ChatGLM-6B>



ChatGLM
Alpha

- 从Huggingface下载所有模型文件
 - git clone <https://huggingface.co/THUDM/chatglm-6b>
- 下载demo
 - git clone <https://github.com/THUDM/ChatGLM-6B>
 - cd ChatGLM-6B
- 安装demo
 - pip install gradio
 - python web_demo.py
- 运行交互demo
 - python cli_demo.py
- 安装 api
 - pip install fastapi uvicorn
 - python api.py
- 甚至在你的MAC (w/ Apple Silicon)上运行ChatGLM-6B
 - model = AutoModel.from_pretrained("your local path", trust_remote_code=True).half().to('mps')

ChatGLM-6B开源

ChatGLM（alpha内测版：QAGLM）是一个初具问答和对话功能的中英双语模型，当前仅针对中文优化，多轮和逻辑能力相对有限，但其仍在持续迭代进化过程中，敬请期待模型涌现新能力。

GLM
百亿

XDAI
小呆

GLM-130B
千亿基座

CodeGeeX
百亿代码

QAGLM
对话问答

ChatGLM
敬请期待

- ChatGLM-6B开源：**62亿**参数、训练数据量达到**1万亿**字符、**INT4**量化（仅需**6G内存**推理）、P-Tuning v2高效推理
- ChatGLM-6B开源，**2.2万 star**
- Huggingface下载超过**100万**
- 开源的训练数据量达到**1万亿**字符的模型
- 登上Github Trending 榜**第一**、Hugging Face Trending 榜**第一**

Spaces using THUDM/chatglm-6b 89

multimodalart/ChatGLM-6B qingxu98/gpt-academic
ysharma/ChatGLM-6b_Gradio_Streaming DrSong/ChatGLM-6B-ChatBot
ChallengeHub/Chinese-LangChain ysharma/OSChatbots_ChatGPT_ToeToToe
josStorer/ChatGLM-6B-Int4-API-OpenAI-Compatible OptimalScale/Robin-7b
ljsabc/Fujisaki qingxu98/academic-chatgpt-beta Dao3/ChatGLM-6B
elitecode/ChatGLM-6B-ChatBot Kevin676/Shanghainese-TTS-demo
shibing624/ChatPDF aodanyun/ChatGLM-6B actboy/ChatGLM-6B
fanjingggffff/text IntelligenzaArtificiale/ChatGLM-6B-Int4-API-OpenAI-Compatible
YukiKurosawaDev/ChatGLM Kevin676/Telephone-Interviewing_PaddleSpeech-TTS
+ 69 Spaces

ChatGLM

生态

新闻

ChatGLM

GLM-130B 基座模型

ChatGLM-6B 开源

友情链接

友情链接

对 ChatGLM 进行加速或者重新实现的开源项目：

- [SwissArmyTransformer](#): 一个Transformer统一编程框架，ChatGLM-6B已经在SAT中进行实现并可以进行P-tuning微调。
- [ChatGLM-MNN](#): 一个基于 MNN 的 ChatGLM-6B C++ 推理实现，支持根据显存大小自动分配计算任务给GPU 和 CPU
- [JittorLLMs](#): 最低3G显存或者没有显卡都可运行 ChatGLM-6B FP16， 支持Linux、windows、Mac部署

基于或使用了 ChatGLM-6B 的开源项目：

- [chatgpt_academic](#): 支持ChatGLM-6B的学术写作与编程工具箱，具有模块化和多线程调用LLM的特点，可并行调用多种LLM。
- [闻达](#): 大型语言模型调用平台，基于 ChatGLM-6B 实现了类 ChatPDF 功能
- [glm-bot](#): 将ChatGLM接入Koishi可在各大聊天平台上调用ChatGLM
- [Chinese-LangChain](#): 中文langchain项目，基于ChatGLM-6b+langchain实现本地化知识库检索与智能答案生成，增加web search功能、知识库选择功能和支持知识增量更新
- [bibliothecarius](#): 快速构建服务以集成您的本地数据和AI模型，支持ChatGLM等本地化模型接入。
- [langchain-ChatGLM](#): 基于 langchain 的 ChatGLM 应用，实现基于可扩展知识库的问答
- [ChatGLM-web](#): 基于FastAPI和Vue3搭建的ChatGLM演示网站(支持chatglm流式输出、前端调整模型参数、上下文选择、保存图片、知识库问答等功能)
- [ChatGLM-6B-Engineering](#): 基于 ChatGLM-6B 后期调教，网络爬虫及 [Stable Diffusion](#) 实现的网络搜索及图片生成
- [ChatGLM-OpenAI-API](#): 将 ChatGLM-6B 封装为 OpenAI API 风格，并通过 ngrok/cloudflare 对外提供服务，从而将 ChatGLM 快速集成到 OpenAI 的各种生态中。

对 ChatGLM-6B 进行微调的开源项目：

- [InstructGLM](#): 基于ChatGLM-6B进行指令学习，汇总开源中英文指令数据，基于Lora进行指令数据微调，开放了Alpaca、Belle微调后的Lora权重，修复web_demo重复问题
- [ChatGLM-Efficient-Tuning](#): 基于ChatGLM-6B模型进行定制化微调，汇总10余种指令数据集和2种微调方

开源开放的心态做大模型研究

ChatGLM系列

 **ChatGLM-6B** Public

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

● Python ☆ 21.6k 🍴 2.6k

 **GLM-130B** Public

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

● Python ☆ 5k 🍴 362

GLM千亿基座模型

自动代码生成模型

 **CodeGeeX** Public

CodeGeeX: An Open Multilingual Code Generation Model

● Python ☆ 4.7k 🍴 313

 **CogView** Public

Text-to-Image generation. The repo for NeurIPS 2021 paper "CogView: Mastering Text-to-Image Generation via Transformers".

● Python ☆ 1.4k 🍴 163

文图生成模型

文字视频生成模型

 **CogVideo** Public

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

● Python ☆ 2.8k 🍴 286

 **cogdl** Public

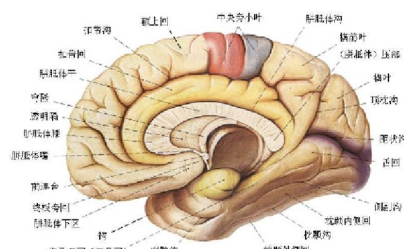
CogDL: A Comprehensive Library for Graph Deep Learning (WWW 2023)

● Python ☆ 1.4k 🍴 300

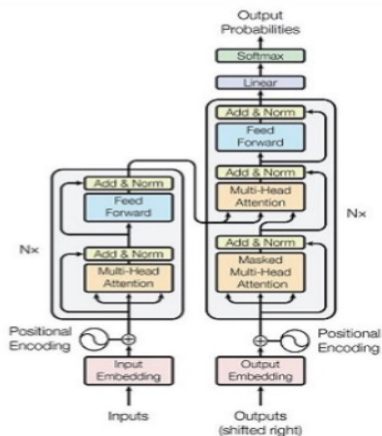
图神经网络工具包

认知大模型的思考 (始于2019的五年规划)

查询接口



超大规模预训练



GLM-130B

错误探测与反馈模型

蒸馏与抽象

记忆模型 (知识

CodeGeex, QA, Web, KG...



“有意识”决策

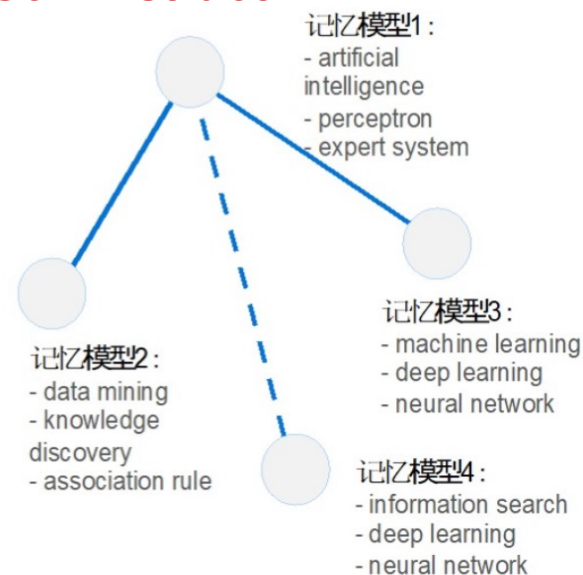
用户反馈与反思

认知推理

“无意识”探测

自定义条理与逻辑

Self-Instruct



ChatGLM

认知智能T9原则

让机器像人一样“思考”：图灵测试的通用机器认知能力

认知
AI

1. 适应与学习能力
2. 定义与语境化能力
3. 自我系统的准入能力
4. 优先级与访问控制能力
5. 召集与控制能力
6. 决策与执行能力
7. 错误探测与编辑能力
8. 反思与自我监控能力
9. 条理与灵活性之间的能力

探索更强大、更
通用预训练模型

探索面向认知的
新理论

探索新的生态合作模式



□ 致谢

□ 技术贡献：

- 清华大学知识工程实验室
- 清华大学PACMAN实验室
- 清华大学自然语言处理实验室
- 智谱AI

□ 算力赞助：智谱AI

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



清华大学 计算机科学与技术系

Department of Computer Science and Technology, Tsinghua University

2023年计算机系校友创新创业研讨会
暨第一届基础模型前沿研讨会

谢谢！

