

紫东太初与大模型数据选择

张家俊

中国科学院自动化研究所

武汉人工智能研究院

jjzhang@nlpr.ia.ac.cn

2023.11.29

目录

- 紫东太初多模态大模型简介
- 基础模型数据选择优化探索
- 指令微调数据选择优化探索

多模态大模型是通用人工智能重要探索

多模态大模型贯通不同模态间的知识表达与语义对齐，相较现有模型的单/双模态更易覆盖现实需求

物理世界是多模态形式存在：

人类感知认知大多是图文音触等多模态的输入输出方式



图像

在照片的前半部分，我们可以看到一个干草区，那里有一个女人站着，穿着浅蓝色牛仔裤和灰白色长袖衬衫。这个女人手里拿着一件黑色夹克。另一只手上拿着一个桃色的气球。在图片的顶部，我们看到一片晴朗的蓝天和云。那个女人的头发是棕色

文本



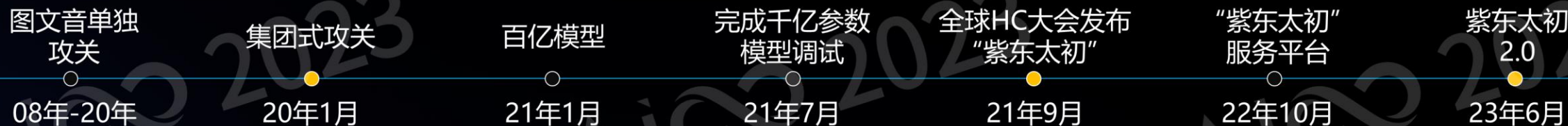
语音

多模态融合：

利用文本、图片、音频和视频等多模态数据进行跨模态的统一表征和学习，是更接近人类的学习方式



“紫东太初”多模态大模型攻关历程



不同场景对应不同小模型

一个场景一个模型
模型参数量小，泛化性差



- 从0开始
- 艰难迭代
- 独立调优
- 推倒重来

多模态大模型吸收海量知识

多个场景一个模型
模型参数大，泛化性强



- 预训练
- 微调



算力中心（北京-武汉）



图像

在照片的前半部分，我们可以看到一个干草区，那里有一个女人站着，穿着浅蓝色牛仔裤和灰白色长袖衬衫。

文本



语音

“紫东太初”大模型1.0：一个模型支撑多个任务

同时知识的构建也使得“紫东太初”可以一个模型实现图像识别、语言翻译、图像描述、语音合成等50余种功能，并初步具备了人机交互、认知推理、艺术创作等创新能力。例如我们可以以音生图、可以理解图像含义。

数据模态多样化

语言

领域应用多样化

语言翻译

摘要、续写、问答...

语音合成
语音识别
...

图像描述
图像生成
视觉问答
...

声纹识别
端点检测
语种识别
...

视频分类
视频生成
跨模态检索
...

图像识别
目标检测
语义分割
...

语音

视觉

一个模型实现50余种功能

以文搜图

输入：桌子上的蓝色小碗里盛着水果和坚果



以音生图



以图生音



紫东太初2.0：从三模态到全模态

单模态分离

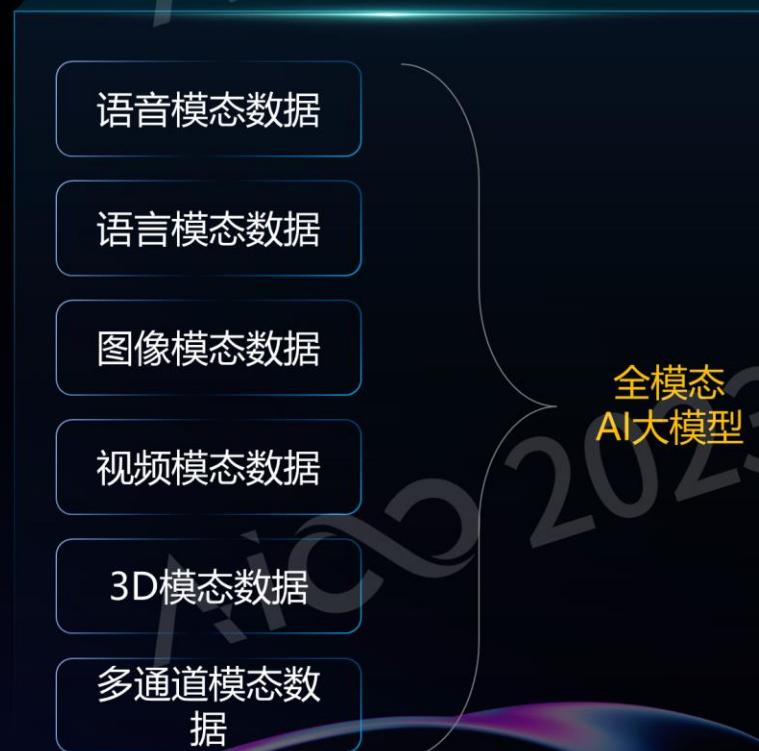
分别对文本、图片、音频和视频等不同模态数据单模型建模。

三模态

利用文本、图片、音频等三种模态数据进行跨模态的统一表征和学习，是更接近人类的学习方式。

全模态

利用文本、图片、音频、视频、3D、等所有模态数据进行跨模态的统一表征和学习。



紫东太初2.0：面向数字物联时代的全模态认知大模型

- **首次提出**全模态多任务统一生成式学习框架：全模态分组对齐、分组解码和联合解码的学习方式。
- **全模态低成本协同关联优化学习**，融合多任务全模态能力，更具拓展性。
- **全模态认知增强**：面向物联世界多源输入的认知，具有全模态能力涌现，认知效率更高。



紫东太初2.0：全面多模态内容理解、知识问答、推理、数学等能力



这是什么海报？



这张图片讲了些
什么内容？



定位图中的物体



兔子在吃什么？

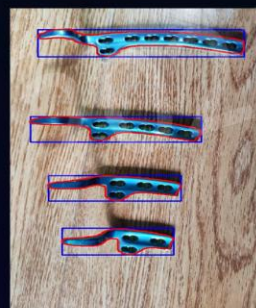


梵高画作



这张海报是《泰坦尼克号》的海报，讲述了一对来自不同阶层的年轻情侣在泰坦尼克号船上相遇、坠入爱河，最终在船撞上冰山后面临生死抉择的故事。

这幅海报呼吁人们节约水资源，并鼓励人们在日常生活中减少用水量。海报上有一本书，上面写着“约用水”，这意味着每个人都应该尽可能地减少用水量，以保护我们的水资源。



兔子在吃胡萝卜



紫东太初2.0：智能制造、智慧医疗、智慧政务等场景应用全面展开



智能制造

布匹纺织及缺陷检测

紫东太初支撑织布、验布流程智能化，10倍级效率提升

< 河南平棉纺织集团、魏桥创业集团等规模应用 >

人工验布

- 速度：20-30m/min
- 精度：无法保证 偏差较大
- 分工：每两人负责一台机器
- 数据：无完整数据记录

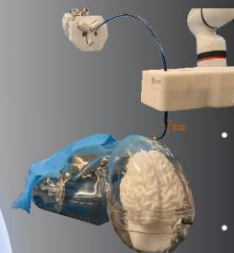
智能验布

- 速度：60-100m/min*
- 精度：95%以上不受人的状态影响
- 分工：每人负责5台以上
- 数据：数字化完整存储，具有完善的可追溯性

脑手术助手

颅内微创手术机器人的神经外科手术场景理解

< 北京协和医院联合研发中 >



智慧医疗

紫东太初+柔性机器人

- 获取柔性机器人传递的视觉/触觉信号，完成跨模态融合理解；
- 实现柔性手术工具颅内误差1mm误差的精准控制

紫东太初



智慧文旅

城市导游

文旅场景首个多模态AI数字人“杭小忆”

< 杭州文旅局使用中/张家界即将上线 >

- 旅游城市市场实时推荐；
- 历史文化介绍：支持“南宋御街”场景陶瓷、丝绸、活字印刷、特色小吃、历史知识等场景智能问答、中文作诗、以文搜图、日常对话

AI手语老师

教学中手语动作与示意图片和文字联动

< 湖南千博信息已规模推广到30+特校 >

社会公益

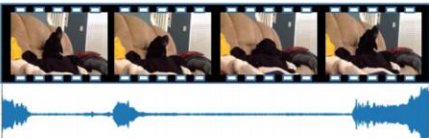
- 开创性地实现手语动作与示意图片和文字的联动，便于让手语词汇更好理解
- 通过手语教考一体机的推广，解决手语师资短缺，加速推动手语标准的推广

目录

- 紫东太初多模态大模型简介
- 基础模型数据选择优化探索
- 指令微调数据选择优化探索

大规模多模态数据集

- ◆ 基于人工标注构建**百万级**图文音强关联多模态数据集**VALOR-1M**, 视频-音频-文本 (多模态综合描述)
- ◆ 基于大语言模型自动构建**千万级**全模态预训练数据集**VAST-27M**, 图像、视频、声音、字幕、自然语言五种模态
- ◆ 上述数据可支持涉及图像、视频、音频、文本等各种模态信息的检索、描述、问答、分类等广泛下游任务



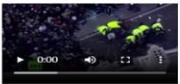
Caption: In the living room, a black dog lies on the sofa barking, with the sound of a police car in the background.

complete sentences ✓

explicit meaning ✓

explicit visual relation ✓

explicit audio relation ✓

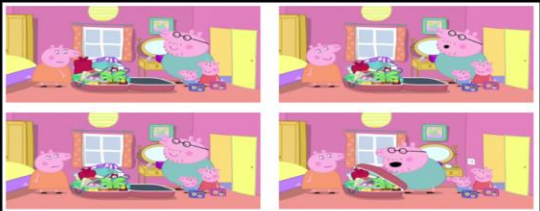
 "In the dim crowd, a group of people were driven away in the sound of passionate music."	 "Several people stretched their arms in the light music, and suddenly there was a scream."	 "A person moves his mouse in a moving music and clicks on the English option to send an email."
 "The silver machine was working, and the yellow tape was rolling and buzzing."	 "Amid fierce applause, a group of black men stood on the golden stage."	 "A green car pulled up on the side of the road with its door open and its engine roared."

视频

视觉描述

听觉描述

语音描述



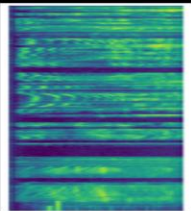
1. cartoon characters are getting ready to go with a suitcase.
2. a group of piggy characters are in their house with a suitcase.
3. a group of cartoon characters have their suitcase open in a bedroom.

1. a cartoon character is speaking.
2. an animated character is talking.
3. someone is making a conversation in a cartoon.

Are you sure we need it all? Yes, its all very important.

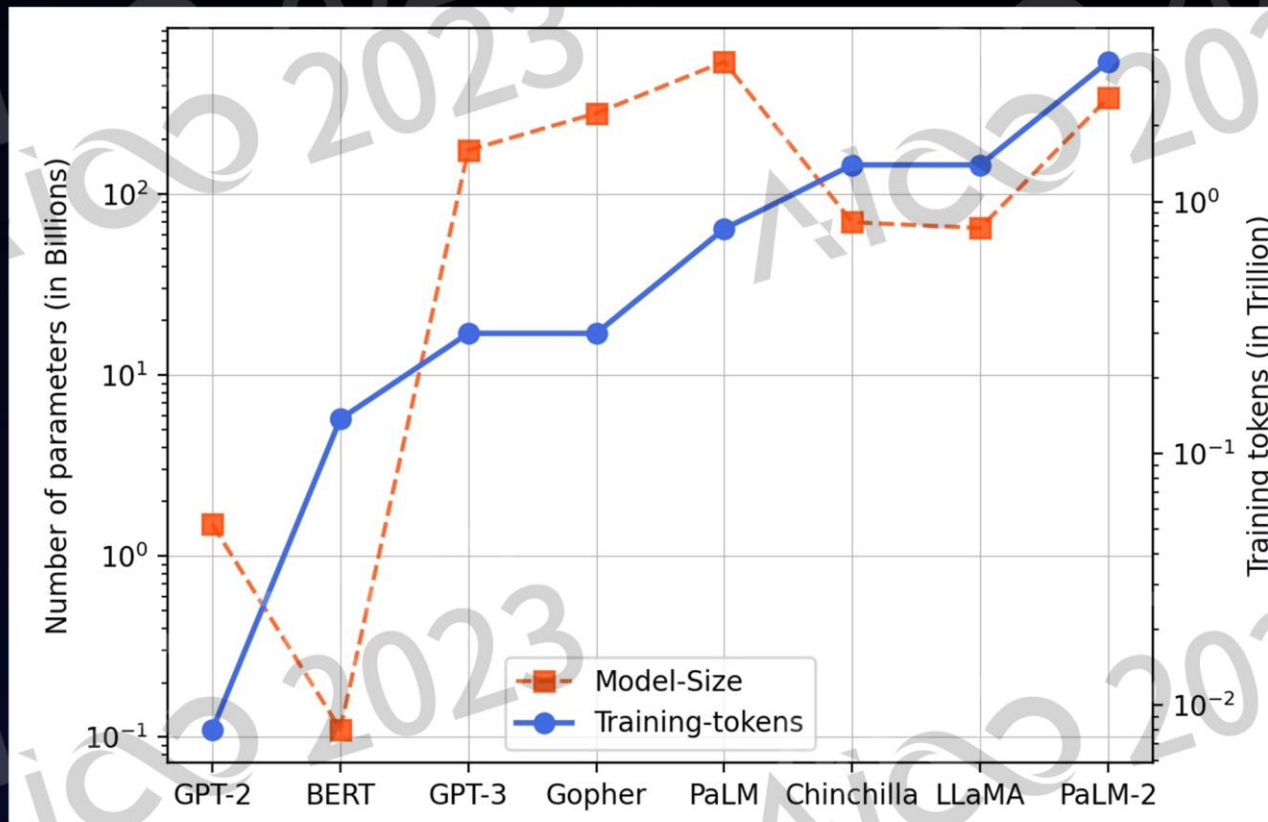
音频

全模态综合描述



A group of cartoon characters, including piggy ones, are preparing to go on a trip, with one character checking if they need to bring everything and another character confirming the importance of all the items.

文本数据集：大模型参数量 vs. 预训练数据规模



- 随着大模型参数量的增长，预训练数据规模不断增加，从**10亿级**token增加至**万亿级**token
- 预训练数据的质量和规模对大模型的性能同样有着重要作用

中英文开源数据概览

数据集	语种	数据大小	开源状态
The Pile	英语	825 GB	825 GB
C4	英文	806 GB	806 GB
Redpajama	英语	5 TB	5 TB
RefinedWeb	英语	2.8 TB	2.8 TB
Dolma	英文	4.9 TB	4.9 TB
Wudao	中文	3 TB	200 G
Yuan 1.0	中文	5 TB	1 T
ROOTS-zh	中文	265 GB	265 GB
WanJuan1.0	英文	434.7 GB	434.7 GB
	中文	585 GB	585 GB

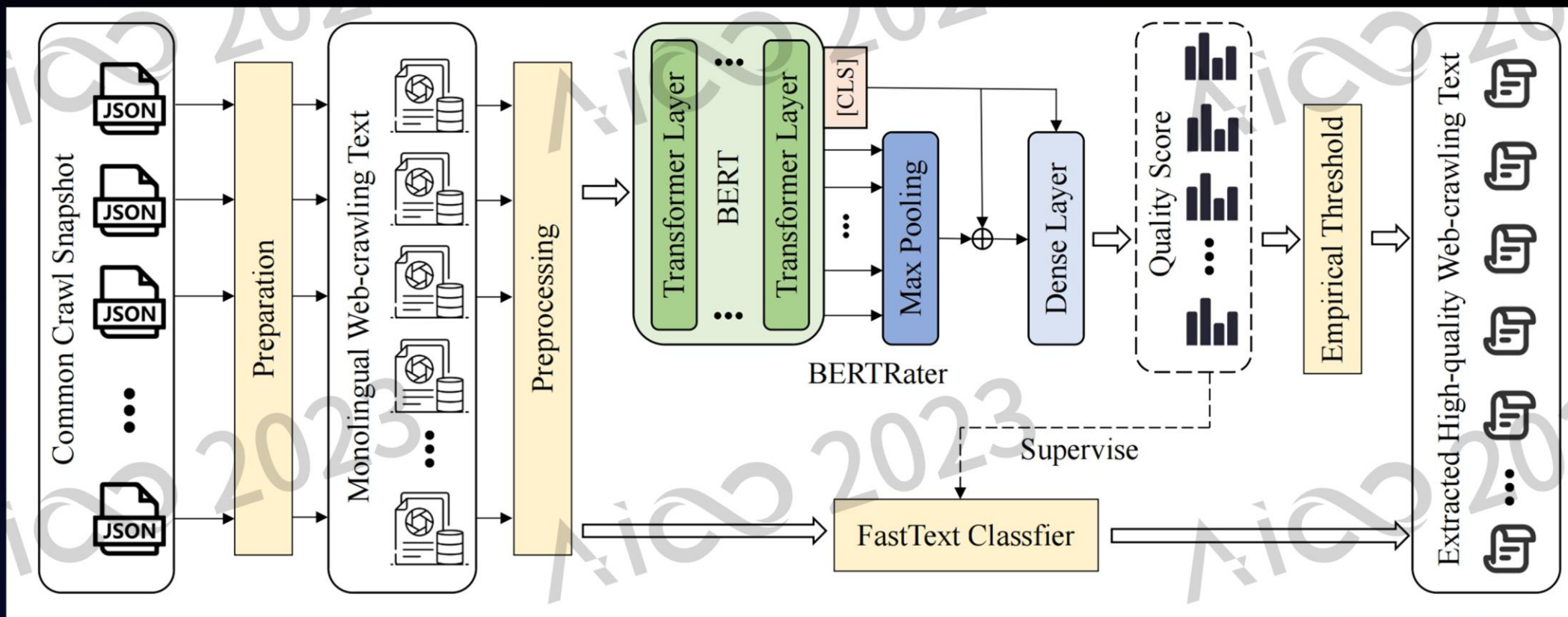
- 中文总体数据规模小、噪声大
- 中文高质量数据类别少，数量不多
- 中文开源数据太少，开源数据分布不够多样
- 开源数据缺乏细粒度的文本质量信息
- 中文开源数据缺乏处理方法和工具链

大规模高质量中文数据与数据处理工具链

- 提出了一个完整的数据抽取工具链 EvalWeb，基于该工具链，我们可以从网络噪声文本中抽取高质量中文预训练数据。
- 发布了一个**1.42 TB** 规模的中文预训练数据集 **ChineseWebText**，在该数据集中，每一条文本数据均**包含了其对应的质量打分**，从而帮助大模型研究人员依据新的阈值选择合适的数据。
- 发布了一个更高质量的数据子集 Cleaner Subset，该数据子集的规模达到了600 GB，且人工质量评估的**准确率达到90%**。

EvalWeb 数据处理工具链

数据下载→中文数据提取→显式噪声过滤→去重→精细质量打分



ChineseWebText 数据集结果分析

Key	Value
<title>	"潍坊银行2021年上半年净利润同比增长29.57% 不良率降至1.10%_财经_中国网"
<score>	0.95
<text>	"潍坊银行2021年上半年净利润同比增长29.57% 不良率降至1.10% 中国网财经8月24日讯 潍坊银行昨日披露2021年二季度信息报告显示,截至2021年6月末,潍坊银行资产总额1920.44亿元,较上年末增长9.34%;负债总额1789.16亿元,较上年末增长10.54%。2021年上半年,潍坊银行实现净利润6.09亿元,同比增长29.57%。 资产质量方面,截至2021年6月末,潍坊银行不良贷款率1.10%,较上年末下降0.13个百分点。 资本金方面,截至2021年6月末,潍坊银行资本充足率、核心一级资本充足率、一级资本充足率分别为11.66%、7.89%、10.13%,分别较上年末下降1.89、0.89、1.15个百分点。"
<url>	http://finance.china.com.cn/news/special/2021bnb/20210824/5638343.shtml
<source_domain>	finance.china.com.cn
<title>	"上海巨也仪器设备有限公司"
<score>	0.19
<text>	"石子冲击试验机\n现货提供石子冲击试验机\n现货提供石子冲击试验机符合SAE、ASTM、VDA、GM、Ford、Mazda、JIS、Nissan、及Toyota等的测试要求。石子冲击试验机(欢迎实地考察)满足:大众、神龙、通用、日产、马自达、丰田、本田、福特等汽车厂家试验。 NS耐石子冲击性能试验机\n耐石子冲击性能试验机主要用于、德系日系、美系汽车厂家试验方法,它能准确再现由飞溅的砂砾造成的破化现象,适用于外涂层粘聚性破坏试验、涂层系统中不同层间粘合性破坏试验、抗剥落的*涂膜厚度、塑料及玻璃的抗剥落、抗碰撞、抗磨损测试等相关试验。 石子冲击试验机(欢迎实地考察)\n石子冲击试验机(欢迎实地考察)符合SAE、ASTM、VDA、GM、Ford、Mazda、JIS、Nissan、及Toyota等的测试要求。满足:大众、神龙、通用、日产、马自达、丰田、本田、福特等汽车厂家试验。 漆膜抗石击试验仪\n漆膜抗石击试验仪符合SAE、ASTM、VDA、GM、Ford、Mazda、JIS、Nissan、及Toyota等的测试要求。 石子冲击试验机/抗石子冲击仪\n巨也仪器!有大量现货提供,欢迎客户随时来厂参观与指导!"
<url>	http://www.juyesh.com/SonList-1094890.html
<source_domain>	www.juyesh.com

数据样例

Snapshot	Data Size(GB)		
	Monolingual Chinese Data	ChineseWebText Dataset	Cleaner Subset
2021-43	505.92	187.57	78.95
2022-05	442.47	164.96	69.44
2022-21	443.57	166.75	70.19
2022-27	417.95	149.41	62.70
2022-33	369.56	123.70	51.98
2022-49	445.29	160.87	67.76
2023-06	396.40	173.47	74.19
2023-14	441.46	150.04	63.33
2023-23	371.96	143.93	61.28
Total	3834.58	1420.70	599.82

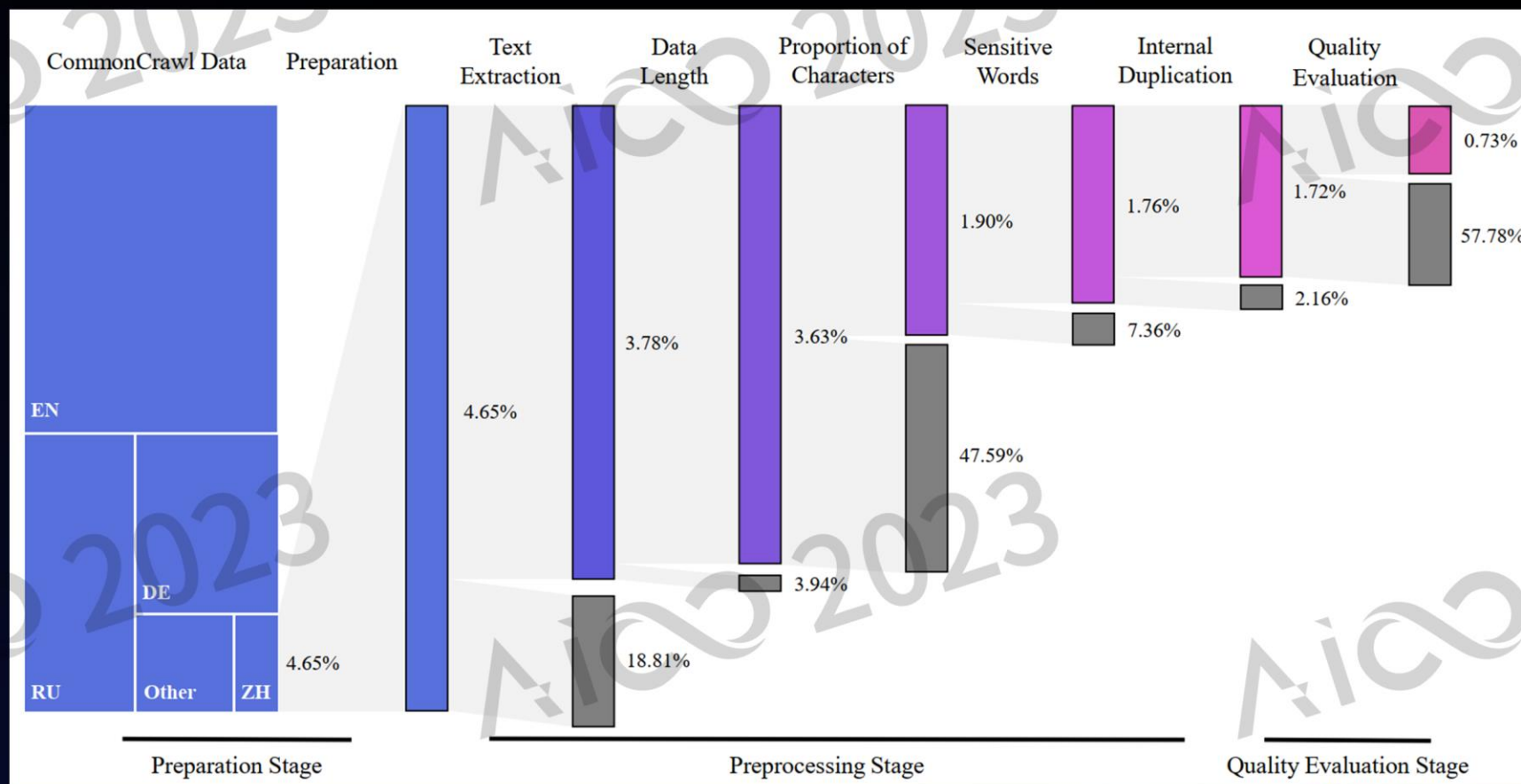
数据分布

与其他开源数据对比分析

Dataset	Lang.	Availability	Pubilc Size	Scoring
C4[1]	EN	Public	807GB	NO
The Pile[2]	EN	Public	825GB	NO
REFINEDWEB[3]	EN	Public	2.8TB	NO
WuDaoCorpora[33]	ZH	Pratly Public	200GB	NO
ROOTS-zh[34]	ZH	Public	265GB	NO
WanJuan1.0-zh[4]	ZH	Public	550GB	NO
ChineseWebText (Ours)	ZH	Public	1.4 TB	YES
Cleaner Subset (Ours)	ZH	Public	600 GB	YES

不同数据处理阶段噪声去除比例分析

最终的高质量中文数据只占原始数据的 **0.73%**



目录

- 紫东太初多模态大模型简介
- 基础模型数据选择优化探索
- 指令微调数据选择优化探索

第一步骤

模型学习人类撰写的指令与回答



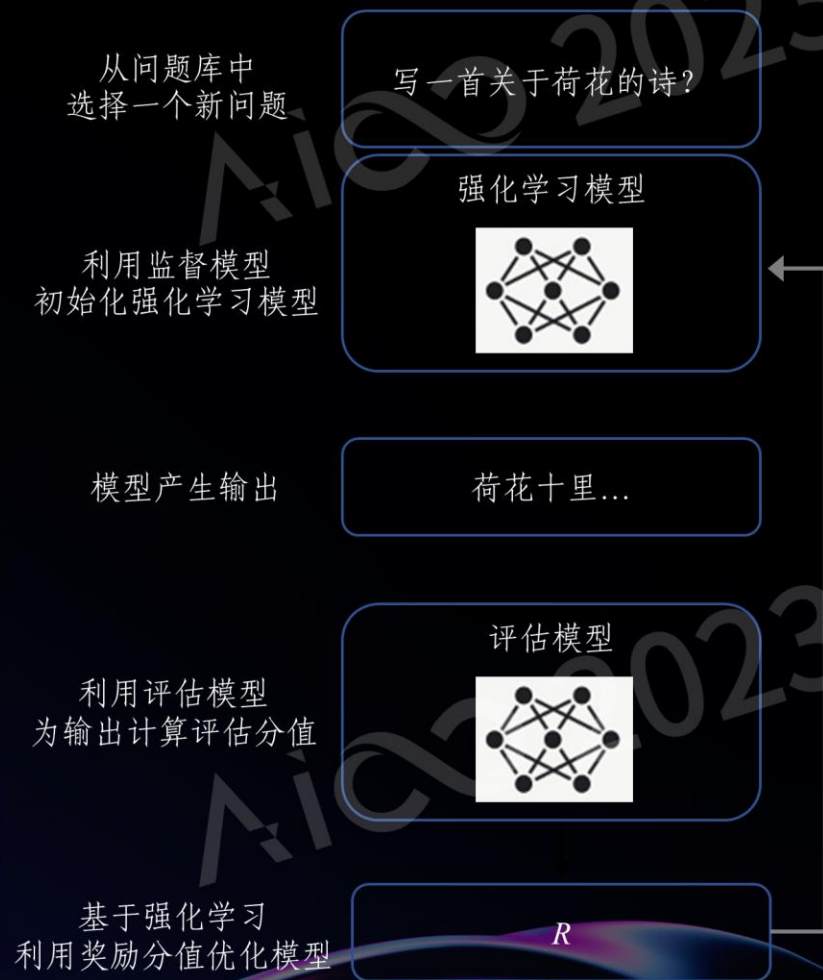
第二步骤

人类对模型输出进行偏好排序



第三步骤

依据人类偏好优化模型



开源指令数据分布

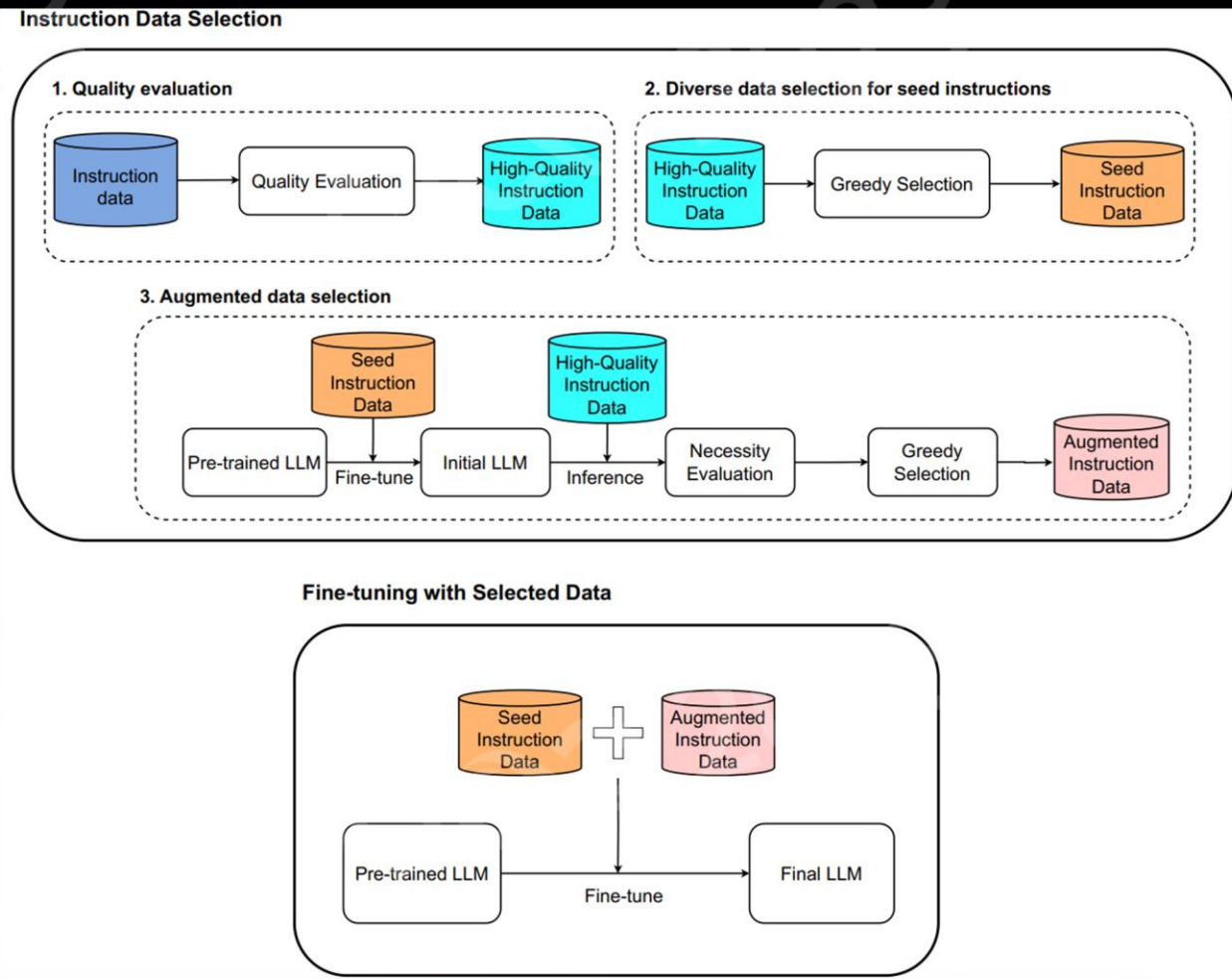
数据集	语种	数据大小
Alpaca	英文	52 K
Vicuna	英文	70 K
Gpt4all-j	英文	800 K
LaMini-instruction	英文	2.58 M
UltraChat	英文	1.57 M
OIG-small-chip2	英文	44 M
COIG	中文	191 K
COIG-PC	中文	321 M
Firefly	中文	1.1M
pCLUE	中文	1.2 M
MOSS	中文	600 K

- 大模型研发人员发布了**百万、千万以上**的指令数据集
- 开源指令数据的规模大、噪声也大，质量则难以保证
- **大规模指令数据增加了大模型指令微调的成本与代价**
- **从大规模指令数据中筛选小而精指令数据成为关注焦点**

MoDS: 以模型为导向的指令微调数据选择方法

- 提出了一种新的指令数据选择标准，从**数据质量**、**数据覆盖度**和**数据必要性**三个方面对指令数据进行选择。
- 方法步骤：
 - ✓ 利用质量评估模型对指令数据的质量进行评估，从而筛选出高质量指令数据；
 - ✓ 利用贪心策略选择出具有最大覆盖度和多样性的指令数据；
 - ✓ 利用必要性评估模型对大语言模型在不同指令上的推理结果进行评估，从而筛选出模型表现较差的指令数据，这些指令表明可以进一步提升模型能力。

MoDS 系统架构



实验数据集

➤ 开源指令数据

- **Alpaca.** 该数据集是由斯坦福发布的大规模指令微调数据集，包含了52,002条利用self-instruct方法构建的指令微调数据。
- **Mixture Dataset.** 通过将HC3、Alpaca、alpaca-evol-instruct、dolly-v2、InstructWild以及lima进行融合，构建了一个包含214,516条指令数据的大规模混合指令数据集，覆盖了医药、法律、金融等多个领域。

➤ 测试数据

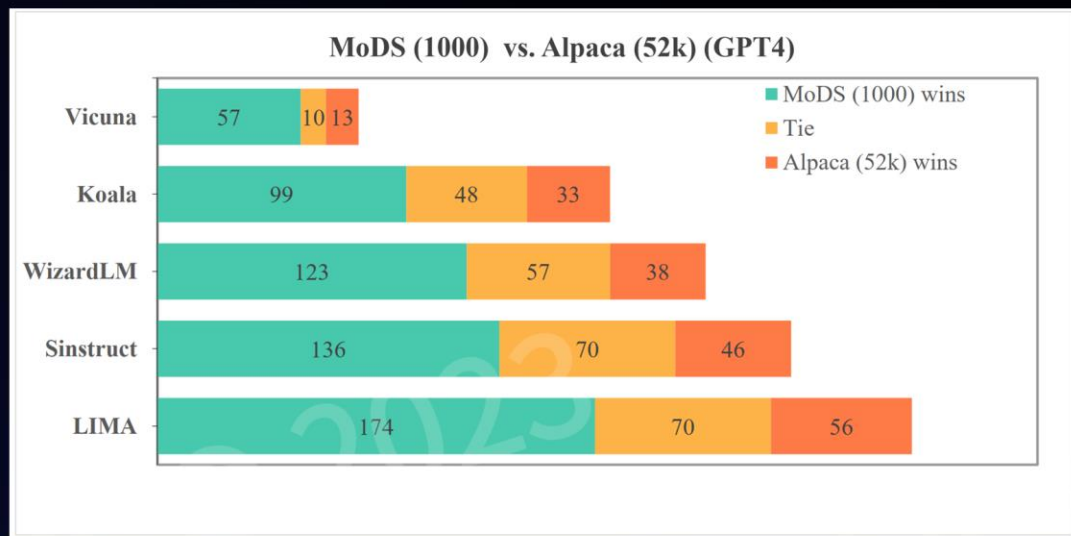
为了评估模型的性能，采用了Koala、WizardLM、Self-instruct、Vicuna以及Lima五个测试数据集对模型的性能进行测试。这五个数据集分别包含了180, 218, 252, 80, 300条人工构建的指令数据，覆盖了数学、代码、写作、计算机等多个领域。

➤ 评估方法

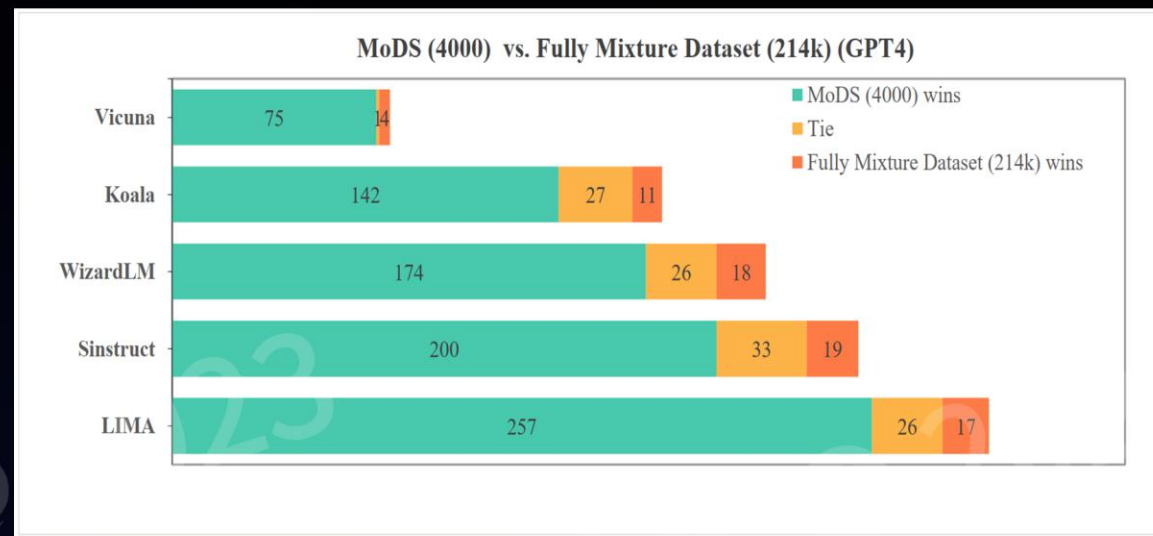
基于ChatGPT和GPT4对不同模型在同一指令下的推理结果进行对比和打分，从而计算不同模型在测试数据中的得胜分数。

指令选择前后模型能力对比分析

- 从52K中选择**1000**条指令数据微调后的模型显著优于利用所有指令数据微调后的模型
- 从214K中选择**4000**条指令数据微调后的模型显著优于利用所有指令数据微调后的模型



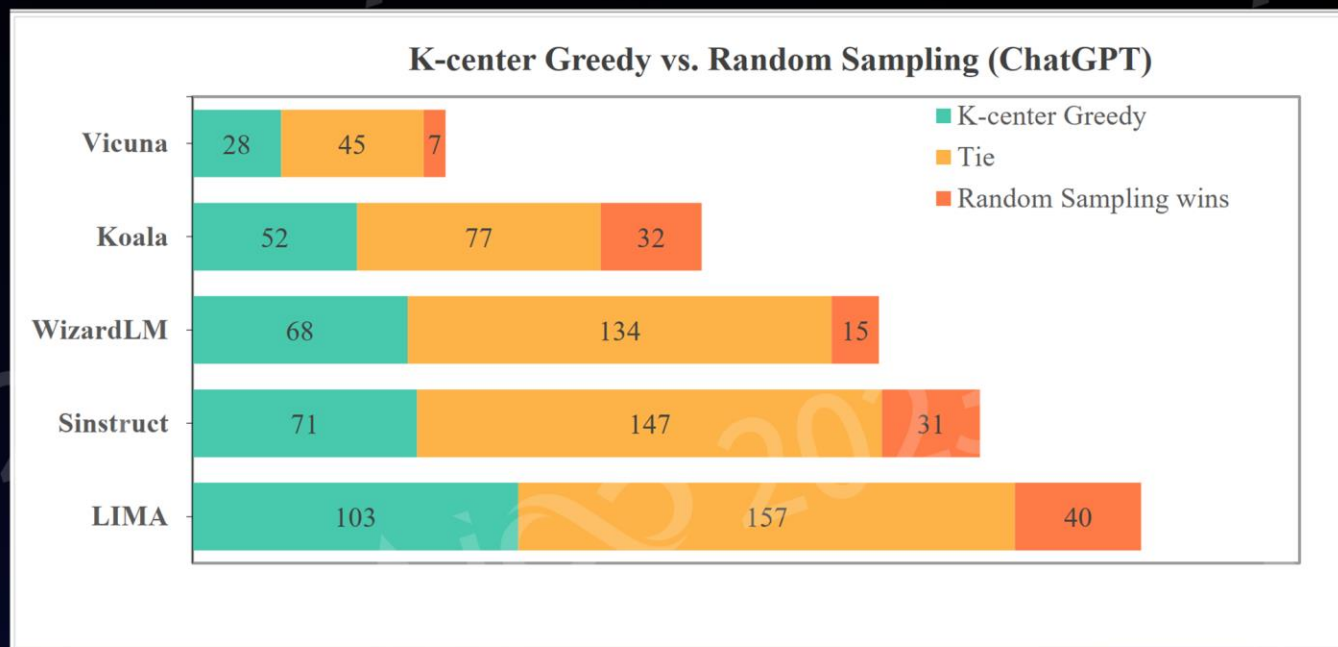
Alpaca数据集



Mixture Dataset数据集

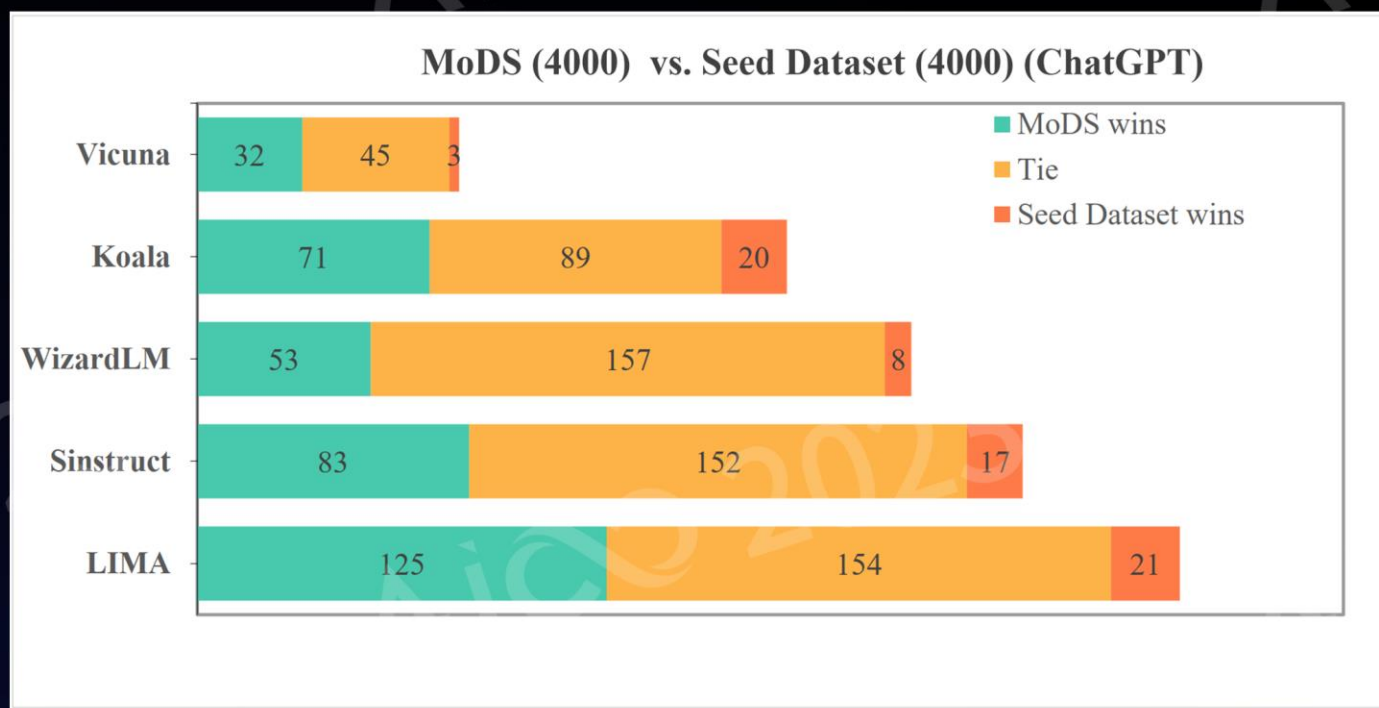
多样性数据选择作用分析

在保证指令覆盖度方面，K-center Greedy方法显著优于随机选择方法



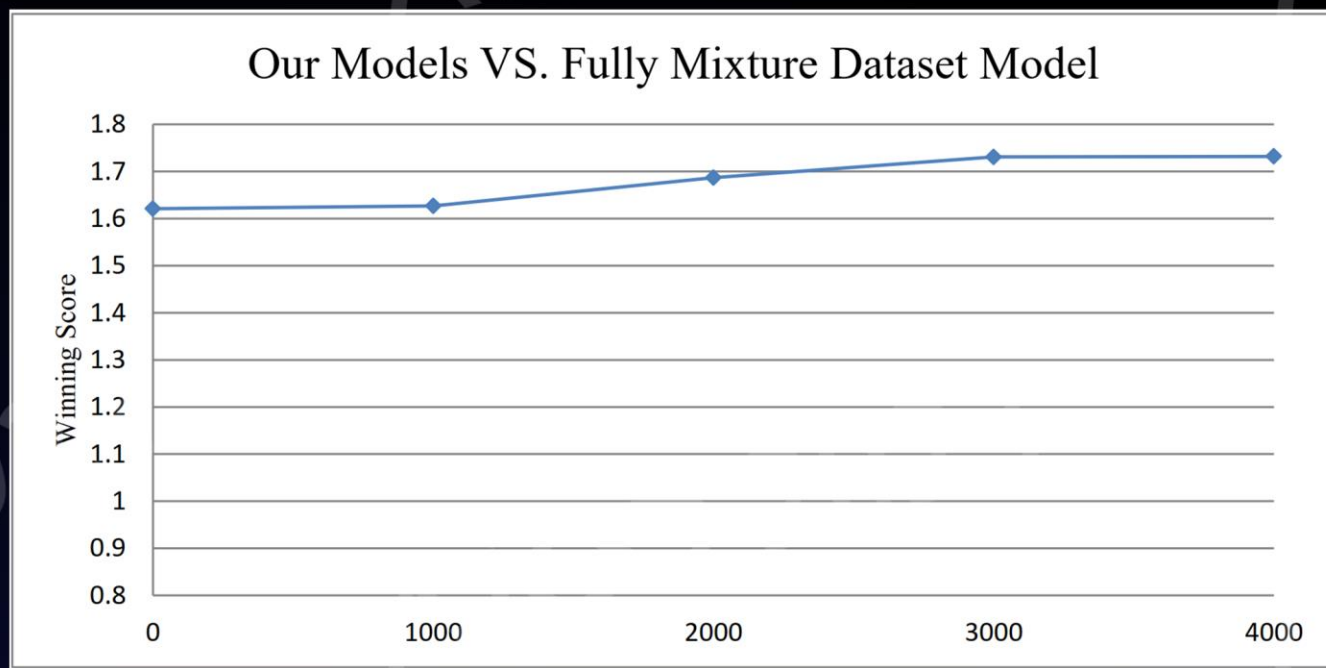
增强型指令数据作用分析

基于必要性的指令数据增强方法显著优于仅考虑指令质量和覆盖度的模型



模型性能与增强型指令数之间的关联关系

指令数据并非越多越好, **3000-4000**已经足够让模型学习到指令跟随能力



winning score 的计算方式如下: $(\text{Num}(\text{win}) - \text{Num}(\text{lose}) / \text{Num}(\text{all})) + 1$

总结

- 紫东太初多模态大模型从图文音三模态迈向全模态，在智能制造、智慧医疗等多个场景中实现规模化落地
- 紫东太初团队对外发布基础模型数据集，包括：大规模多模态数据集**VAST-27M**、大规模中文语言数据集**ChineseWebText**，同时发布**数据处理工具链**
- 紫东太初团队开源大模型指令数据高效选择方法**MoDS**，可以有效加速大模型的优化迭代

Sihan Chen, Jing Liu, et. al. VALOR: Vision-Audio-Language Omni-Perception Pretraining Model arXiv: 2304.08345

Jianghao Chen, Pu Jian, Tengxiao Xi, Dongyi Yi, Qianlong Du, Chenglin Ding, Guibo Zhu, Chengqing Zong, Jinqiao Wang, Jiajun Zhang*.

ChineseWebText: Large-scale High-quality Chinese Web Text Extracted with Effective Evaluation Model. arXiv:2311.01149.

Qianlong Du, Chengqing Zong and Jiajun Zhang*. MoDS: Model-oriented Data Selection for Instruction Tuning. arXiv:2311.15653.



AiC ∞ 2023

AiC ∞ 2023

AiC ∞ 2023

谢 谢

AiC ∞ 2023

AiC ∞ 2023

AiC ∞ 2023

