开源垂类领域大模型现状调研

梁青青

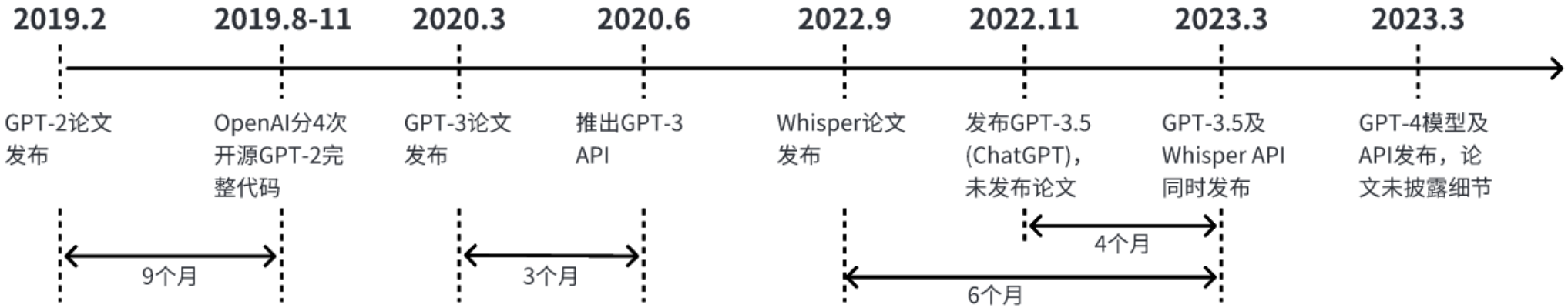
2023.06.15

目录

- 垂类模型发展背景
- 为什么垂类模型需要开源？
- 垂类领域大模型现状
  - LaWGPT：基于中文法律知识的大语言模型
  - 桃李：国际中文教育大模型
  - 本草：基于中午呢医学知识的LLaMA微调模型
  - 轩辕：首个千亿级中文金融对话模型
- 领域大模型的常规训练范式总结
- 入局领域模型构建的思考

背景

- 背景概要
  - 头部领先模型走向闭源，垂类模型开发呼唤开源
  - LLaMA重塑数据地位，羊驼家族带飞垂类模型构建
- 大语言模型在通用任务上的非凡能力引起巨大关注，也意识到仅靠大模型的通用能力无法满足垂直领域需求，为通用大模型向垂类领域模型发展埋下伏笔。
- 头部AI公司的先进模型大多采用部分开源模式或仅开放使用，逐步向闭源模式发展，构建起技术护城河。



OpenAI 系列模型逐步向闭源发展

背景

- 在头部AI公司倾向于部分开源或闭源的背景下，严重制约了垂类领域模型的发展

- 大模型通常基于大量语料训练以达到更高的性能，其前提假设是：**更多的参数会得到更好的性能**
- 一个新观察：给定计算代价，**最佳性能来自在更多数据上训练的较小模型**
- 基于此，Meta推出了基于语言模型集合LLaMA 模型，包含不同规模的参数，同时采用申请审核制方式开源

大模型领域的羊驼家族

- 因其“史诗级泄露事件”激起千层浪，迅速成就“羊驼家族”模型的崛起，点燃垂类模型开发研究热潮！

为什么垂类模型需要开源？

- 为什么要用它？
  - AI大模型可赋能下游垂直行业，实现产品性能升级、提升用户体验感、企业降本增效等
- 为什么不直接用？
  - AI大模型直接应用与垂直行业，存在通用能力过剩、行业专业知识储备不足、推理过程消耗算力过高等问题
  - 根据细分行业需求训练相应的垂类模型成为**AI技术落地**的必要环节
- 垂类模型的构建方式比较

构建方式	成本效益问题	数据安全问题	场景应用问题
从头自行训练	训练难度大，成本极高	不存在数据安全问题	可定制应用场景
调用AI厂商API	AI公司开放API接口毛利率或达95%，成本较高	企业内部数据将与外部模型相连，存在数据安全风险	在线部署，要求网络稳定
接受AI厂商离线部署	需具备一定自研能力，仍存在成本效益问题	仍存在数据安全问题	产品迭代周期长，或影响后续研发
基于开源模型开发	技术要求和成本门槛较低	数据安全有保障	自主性强，可自行定制模型能力

为什么垂类模型需要开源？

- 为什么要用它？
  - AI大模型可赋能下游垂直行业，实现产品性能升级、提升用户体验感、企业将本增效等
- 为什么不直接用？
  - AI大模型直接应用与垂直行业，存在通用能力过剩、行业专业知识储备不足、推理过程消耗算力过高等问题
  - 根据细分行业需求训练相应的垂类模型成为**AI技术落地**的必要环节
- 垂类模型的构建方式比较

构建方式	成本效益问题	数据安全问题	场景应用问题
从头自行训练	训练难度大，成本极高	不存在数据安全问题	可定制应用场景
调用AI厂商API	AI公司开放API接口毛利率或达95%，成本较高	企业内部数据将与外部模型相连，存在数据安全风险	在线部署，要求网络稳定
接受AI厂商离线部署	需具备一定自研能力，仍存在成本效益问题	仍存在数据安全问题	产品迭代周期长，或影响后续研发
基于开源模型开发	技术要求和成本门槛较低	数据安全有保障	自主性强，可自行定制模型能力

为什么垂类模型需要开源？

- 为什么要用它？
  - AI大模型可赋能下游垂直行业，实现产品性能升级、提升用户体验感、企业将本增效等

基于开源模型开发垂类模型是性价比较高的选择！

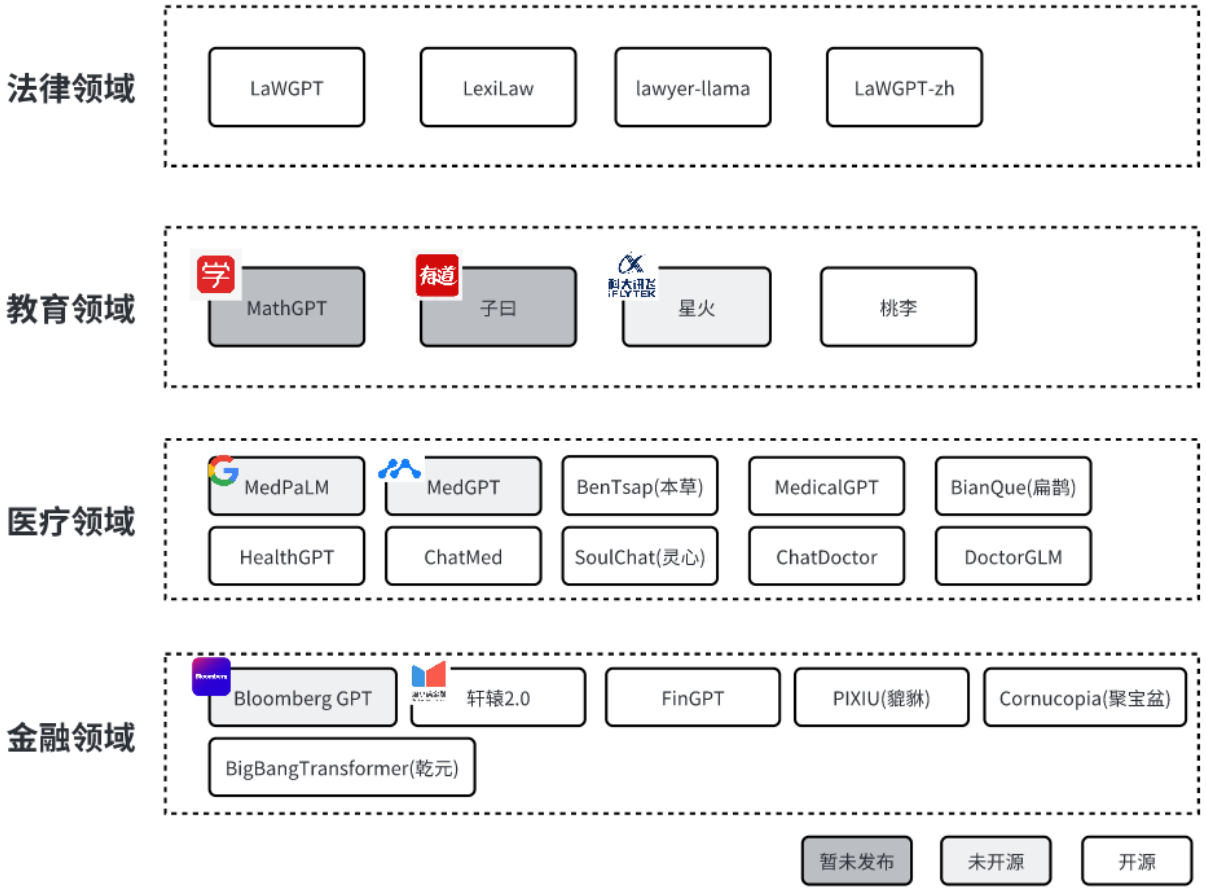
- 对下游用户技术水平和研发投入要求相对较低
- 无需向AI厂商分享数据，不存在信息安全问题
- 使用者可根据业务需求自行增减功能或进行模型迭代

接受AI厂商离线部署	在成本效益问题	仍存在数据安全问题	响后续研发
基于开源模型开发	技术要求和成本门槛较低	数据安全有保障	自主性强，可自行定制模型能力

现有垂类领域模型现状

- 根据细分行业需求训练垂类模型是AI技术落地的必要环节
- 目前开源垂类模型聚焦四个领域
  - 热度：医疗>金融>法律>教育

- 以增量预训练、微调方式居多
- MedicalGPT实现了四阶段训练
  - 增量预训练 (continue pre-training)
  - 有监督微调 (supervised fine-tuning)
  - 奖励模型 (reward model)
  - 基于人类反馈的强化学习 (RLHF)



垂类领域模型概览

LaWGPT：基于中文法律知识的大预言模型

- LaWGPT是一系列基于中文法律知识的开源大模型

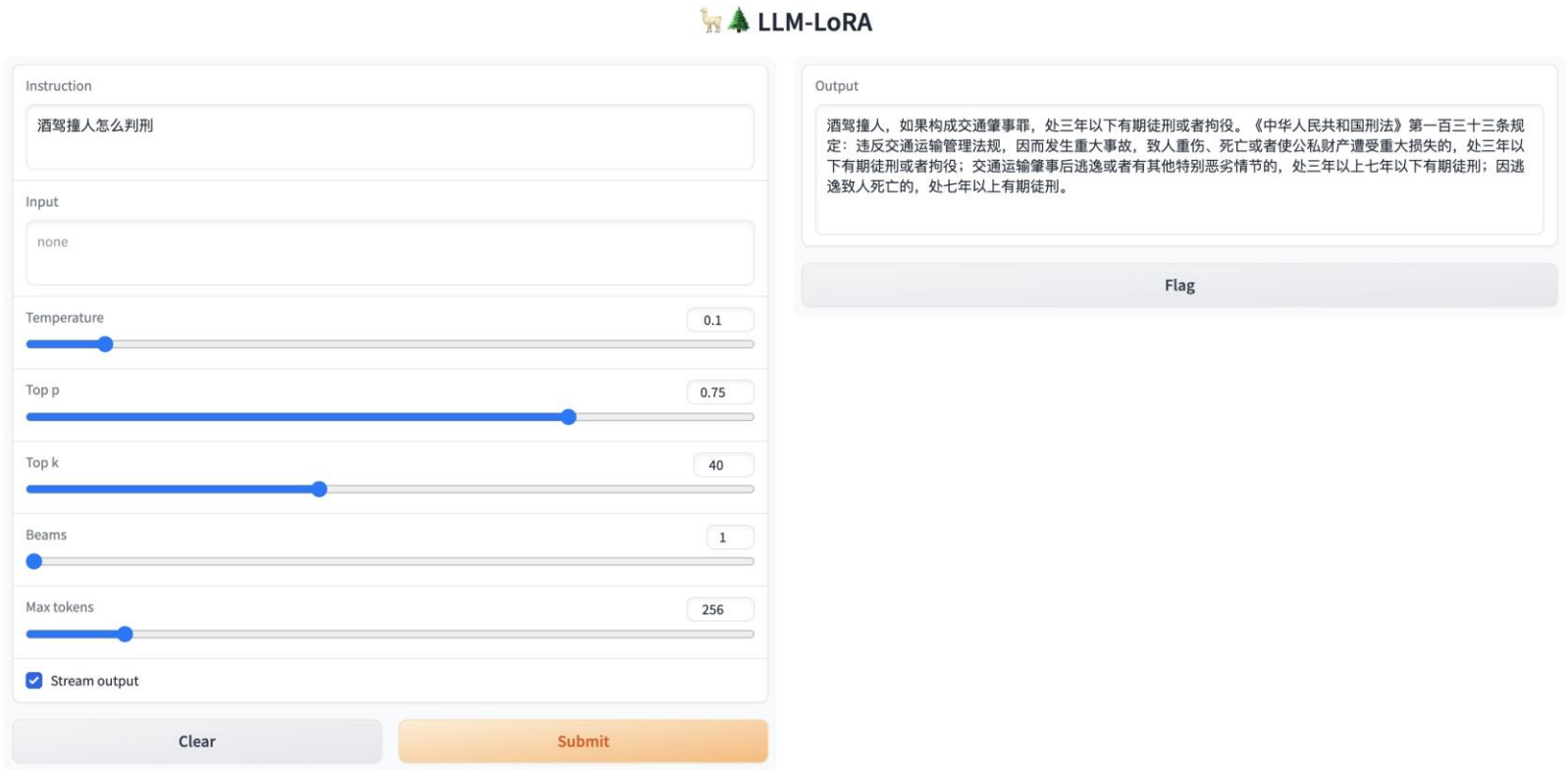
发布模型	基座模型	训练数据集
LaWGPT-7B-alpha	Chinese-LLaMA-7B	30W法律问答数据集
Legal-Base-7B		50W中文裁判文书数据二次预训练
LaWGPT-7B-beta1.0	Legal-Base-7B	30W高质量法律问答数据集进行指令精调
LaWGPT-7B-beta1.1	Chinese-alpaca-plus-7B	35W高质量问答数据集进行指令精调

- 数据构建：
  - 基于公开法律文书数据、司法考试数据集
  - 初级数据生成：self-instruct
  - 知识引导的数据生成：knowledge-based self-instruct
  - 引入ChatGPT清洗数据，构建高质量数据集
- 模型训练
  - 第一阶段：扩充法律领域词表，在大规模文书及法典数据上对Chinese-LLaMA进行预训练
  - 第二阶段：构造法律领域问答数据集，在预训练模型基础上指令精调

- 项目地址：https://github.com/pengxiao-song/LaWGPT

LaWGPT：基于中文法律知识的大预言模型

- 计算资源
  - 8张Tesla V100-SXM2-32GB，二次预训练耗时24/epoch，微调阶段耗时12h/epoch



输出示例

桃李：国际中文教育大模型

- 数据集构建
  - 通用指令微调数据：alpaca-gpt4 (52K)、alpaca-gpt4-zh (52K)
  - 国际中文教育指令微调数据
    - 语法改错数据(9K)：使用汉语学习者文本多维标注数据集YACLC的开发集作为语法改错中最小改动以及流利数据的来源，使用HSK作文评分数据作为篇章级语法改错的数据来源。
    - 释义生成数据(4K)：从现代汉语词典和对外汉语词典中提取了大量词条用于释义数据的构建。
    - 文本简化数据(6K)：使用多参考中文文本简化数据集作为文本简化指令微调数据的来源
    - 可控文本生成数据(6K)：采用汉语国际教育动态语料库作为数据来源。
- 模型训练
  - 基于Chinese-LLaMA-7B使用上述数据进行指令微调

- 开发团队：北京语言大学、清华大学、东北大学、北京交通大学
- 项目地址：https://github.com/blcuicall/taoli

桃李：国际中文教育大模型

- 模型效果
  - 利用HSK考试真题，对比Taoli 1.0和GPT-4的作答成绩
  - 试题以客观题为主，主观题不计分

试题（客观题）	Taoli 1.0	GPT-4
HSK4	55	78
HSK5	60	85
HSK6	42	76

HSK客观题得分

输入	Taoli 1.0
请遵循最小改动原则纠正下面这句话：北京与西安有很多好象一样的地方。	北京和西安有很多类似的地方。
请生成词语在当前上下文中的释义。 词语：折服 上下文：我深深地被他的奉献精神所折服。	被说服或被感动。
请使用下列所有词语，写一个30字左右的段落。 词语：元旦、放松、礼物、表演	元旦是一年中最重要的节日之一，人们会庆祝并享受放松的时间。在节日中，人们会收到礼物，表演，并享受各种活动。

输出示例

本草：基于中文医学知识的LLaMA微调模型

- 数据集构建
  - 自建的中文医学知识库（医学文献）+公开的中文医学知识图谱CMeKG，设置多种prompt形式来充分利用知识
  - 利用GPT-3.5接口围绕知识库构建多轮问答数据
- 模型训练
  - 基于LLaMA-7B和ChatGLM-6B进行指令微调
  - 在一块A100-SXM-80GB显卡上训练，训练总轮次10轮，耗时约2h17m，batch_size=128的情况下显存占用在40G左右
- 模型效果对比

测试输入	Llama输出	Alpaca输出	BenTsao本草输出
小张最近感觉身体不适，出现心悸、气促等症状。体检发现心脏扩大、搏动减弱。	问题：小张最近感觉身体不适，出现心悸、气促等症状。体检发现心脏扩大、搏动减弱。	最近，小张一直感觉身体不舒服，有心烦意乱、呼吸急促等症状。体格检查显示心脏增大，肌肉无力。	**小张可能患有心肌炎，建议尽快到医院进行检查和治疗。治疗方案包括药物治疗和手术治疗。**
麻风病和儿童哮喘的病因是否一致?	儿童哮喘和麻风病的病因是一致，但是儿童哮喘和麻风病的病因不一致	是的，哮喘和儿童喘息有相似的原因。	**不一致，麻风病的病因是麻风杆菌感染，而儿童哮喘的病因是喂养不良、吸烟等因素。**
一位患者出现了持续性的胀痛和黄疸，经影像学检查后确诊为肝胆管结石病，应该采取哪些治疗方案?	肝胆管结石病患者应该采取哪些治疗方案?	患者应接受手术和化疗相结合的治疗。	**针对肝胆管结石病，治疗方案包括手术治疗、药物治疗和辅助治疗等。具体治疗方案需要根据患者的具体情况而定。**

- 项目地址：https://github.com/SCIR-HI/Huatuo-Llama-Med-Chinese
- 开发团队：哈尔滨工业大学社会计算与信息检索研究中心健康智能组

轩辕2.0：首个千亿级中文金融对话模型

- 数据集构建
  - 非结构化的金融文本数据，如金融新闻文章、市场报告、分析师评论和社交媒体讨论
  - 使用人工编写的种子指令通过self-instruct收集通用数据，利用金融领域的非结构化和结构化数据通过self-QA收集特定领域的指令数据。
- 模型训练
  - 基于BLOOM-176B架构，利用**混合调优(Hybrid-Tuning)**方法进行训练，以缓和“灾难性遗忘”问题
  - 训练数据=预训练数据(33%)+指令数据(67%)
  - 预训练数据(33%)=通用数据(20%)+金融数据(13%)
  - 指令数据(67%)=通用数据(46%)+金融数据(21%)
  - 随机混洗上述数据为一个训练数据，单阶段训练

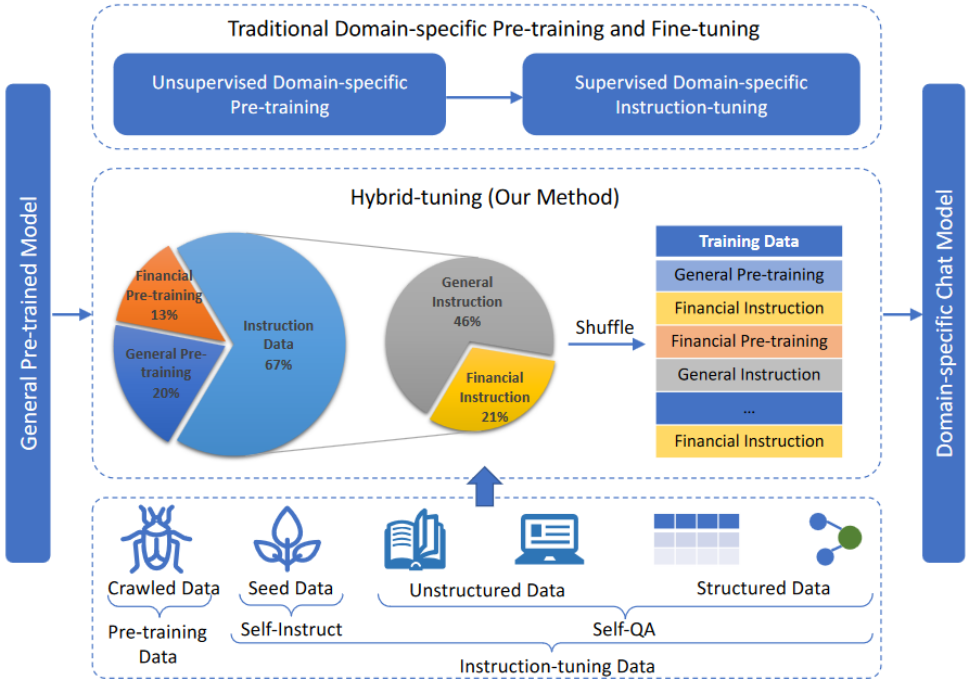
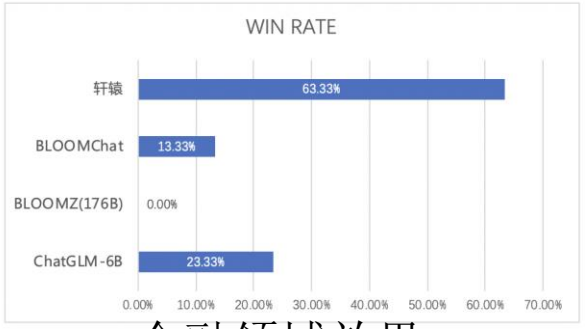


Figure 1: Our proposed hybrid-tuning.

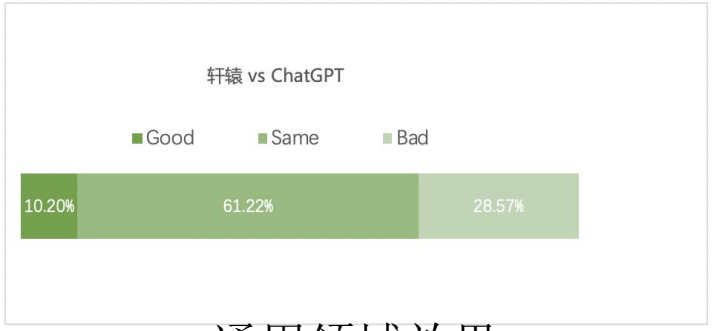
- 项目地址：https://github.com/Duxiaoman-DI/XuanYuan
- 开发团队：度小满

轩辕2.0：首个千亿级中文金融对话模型

- 模型训练
  - A100-80G，DeepSpeed分布式训练框架，pipeline并行，zero-redundancy优化
- 模型效果
  - 金融领域效果
    - 任务：金融名词理解、金融市场评论、金融数据分析、金融新闻理解
    - 轩辕在主流的四种开源大模型的比较中，赢得了150次回答中63.33%的胜率，凸显出其在金融领域的显著优势
  - 通用领域效果
    - 任务：数学计算、场景写作、逻辑推理、文本摘要
    - 轩辕在71%的问题上表现不亚于ChatGPT，验证其通用能力



金融领域效果

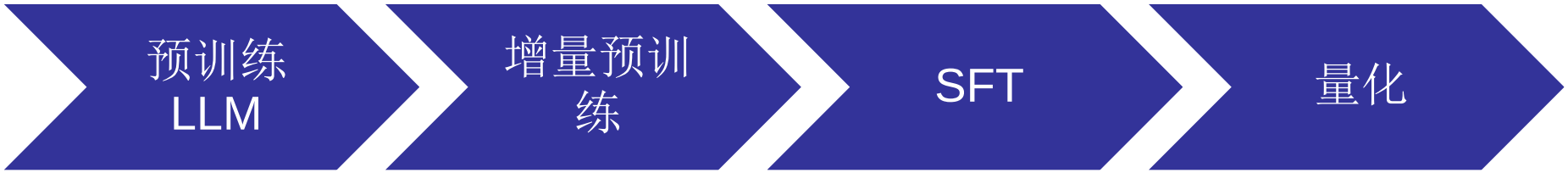


通用领域效果

Hyperparameter	XuanYuan2-7B	XuanYuan2
Architecture hyperparameters		
Parameters	7,069M	176,247M
Layers	30	70
Hidden dim.	4096	14336
Attention heads	32	112
Vocab size	250,680	
Sequence length	2048	
Precision	float16	
Activation	GELU	
Position emb.	Alibi	
Tied emb.	True	
Pretraining hyperparameters		
Global Batch Size	512	2048
Learning rate	1.2e-4	6e-5
Total tokens	341B	366B
Min. learning rate	1e-5	6e-6
Warmup tokens	375M	
Decay tokens	410B	
Decay style	cosine	
Adam (β_1, β_2)	(0.9, 0.95)	
Weight decay	1e-1	
Gradient clipping	1.0	
Multitask finetuning hyperparameters		
Global Batch Size	2048	2048
Learning rate	2.0e-5	2.0e-5
Total tokens	13B	
Warmup tokens	0	
Decay style	constant	
Weight decay	1e-4	

超参设置

领域模型构建的常规范式总结



- LLaMA
- BlOOM
- GLM

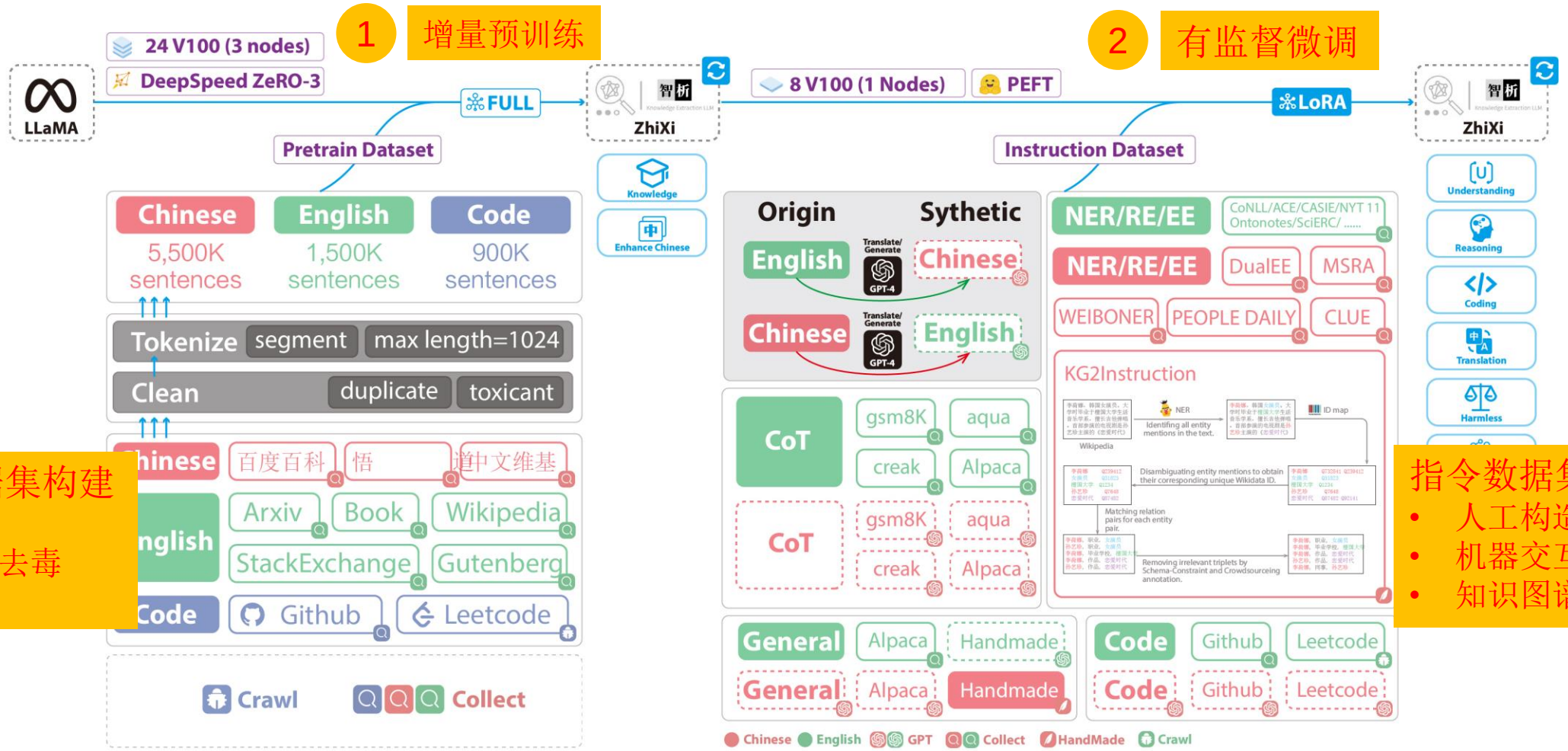
- 领域预训练
- 任务预训练
- 领域+任务预训练

- Self-instruct、Evol-instruct、Self-QA
- 开源数据
- 领域数据

- LLM.int8()
- GPTQ
- RPTQ

- 语言模型：解决语言模型基座问题
- 增量预训练：在目标领域数据集上继续预训练以提升效果。目标领域与语言模型原始领域语料越不相关，模型效果则提升更明显
- 指令微调数据：根据模型及场景指定指令微调数据，基于基础模型利用指令数据得到SFT模型
- 微调加速/量化：解决低资源环境下的训练问题，热度最高的是LoRA

KnowLM的构建流程实例



预训练数据集构建
- 数据采集
- 数据清洗去毒
- 分词

指令数据集构建
- 人工构造
- 机器交互
- 知识图谱引导

项目地址：https://github.com/zjunlp/KnowLM

开发团队：浙江大学陈华钧教授团队 ZJUKG&NLP

入局领域模型构建的思考

- 构建垂类模型的目标是注入领域知识，包含以下三种方式：
  - 增量预训练注入：若基础模型语料中只包含少量领域知识，需采集领域文本对模型进行再训练以注入知识
  - 微调注入：更廉价的方式，更多人的选择，可能存在知识假说，微调只是知识的引导方式，无法注入新知识
  - 外挂注入：只构造prompt，激发模型能力做生成推理，本质是prompt工程，治标不治本
- 必要性：垂类领域模型需要有明确的场景，有哪些任务适合大模型做，是否优于之前的解决方案？
- 自研环境：是否有足够的数据和算力？
- 领域任务间有关联，也有干扰，训练时是否要排序，是否要划分不同的模型？
- 不同来源语料的知识密度和可信级别不同，如何选择和组织？
- 面对模型训练的“灾难性遗忘问题”，如何确定数据混合比例？
- 如何评估训练后模型的效果，是否达到可应用的标准？

LLM is not all you need, but just what you have

感谢大家的聆听，请大家批评指教！