鹏城实验室 PENGCHENG LABORATORY

AITISA 新一代人工智能产业技术创新战略联盟

第四届OpenI/O启智开发者大会

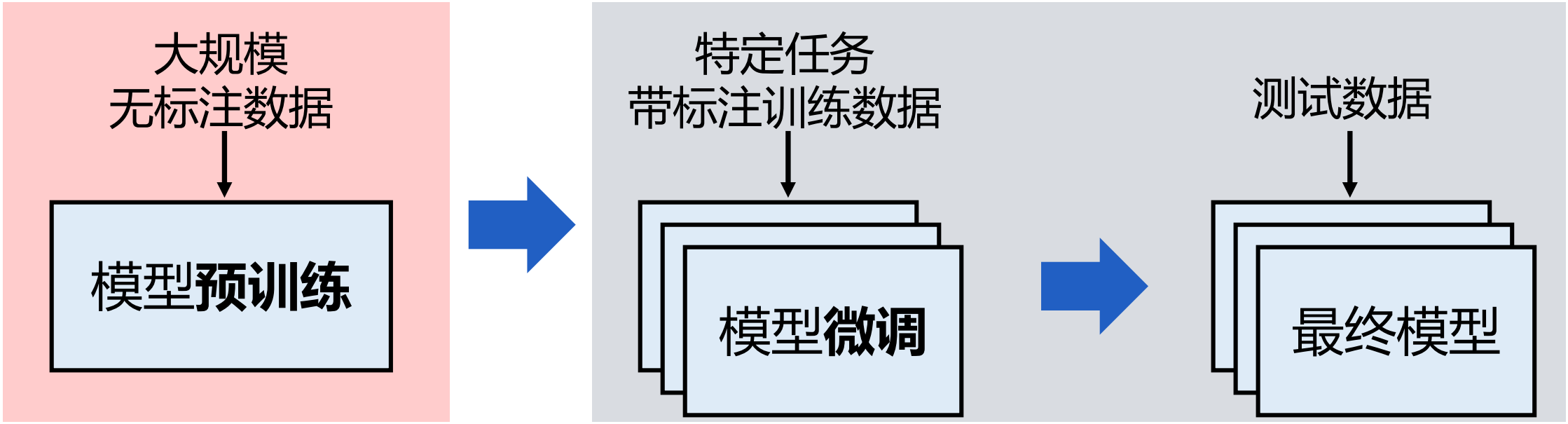
算网筑基 / 开源启智 / AI赋能

大模型值得关注的重要特性

主讲人：刘知远

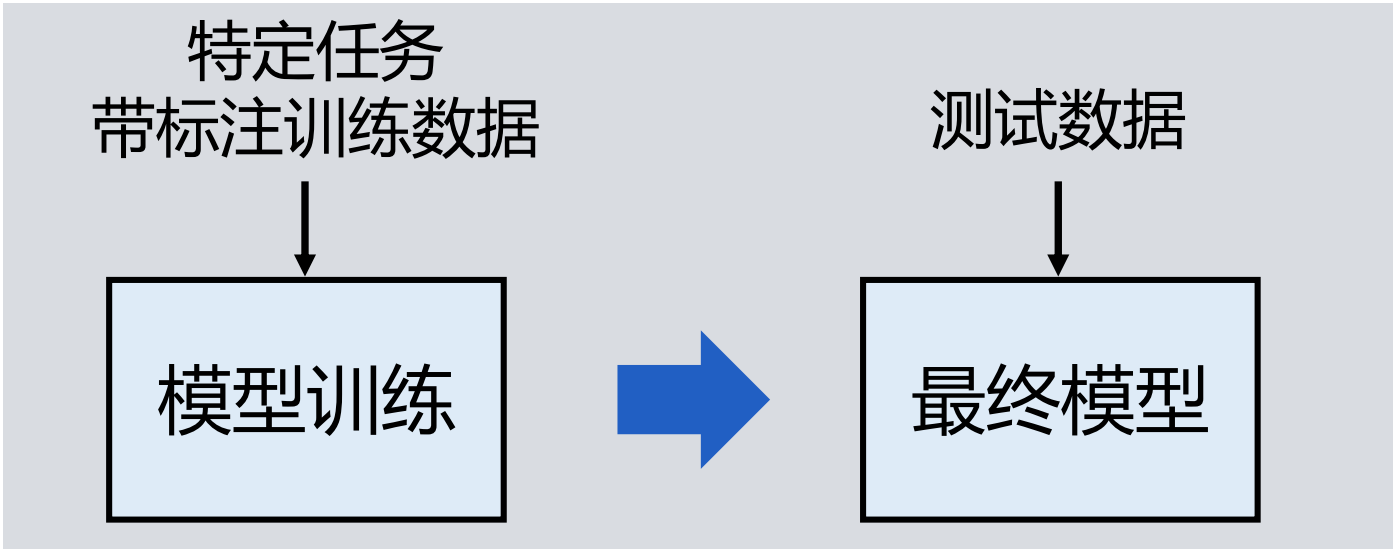
时　间：2023年2月25日

大模型：实现面向大规模无标注数据的深度学习



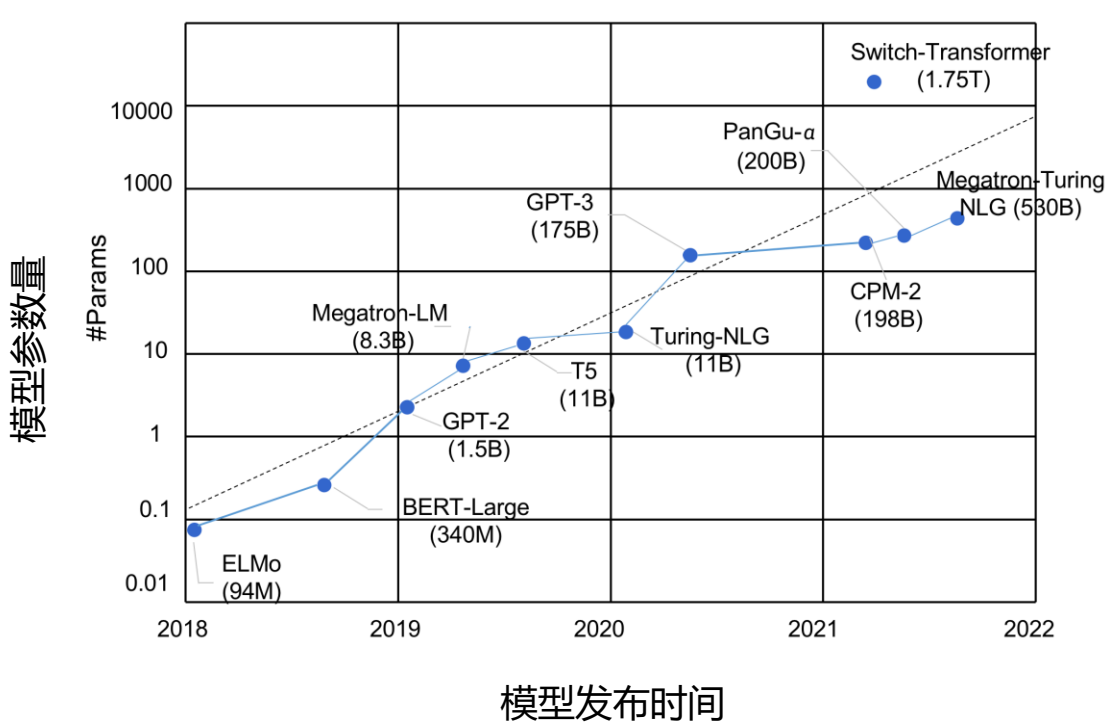
从大数据到大模型

- 无标注大数据廉价易得，近乎无限
- 可支持大模型学习和性能提升

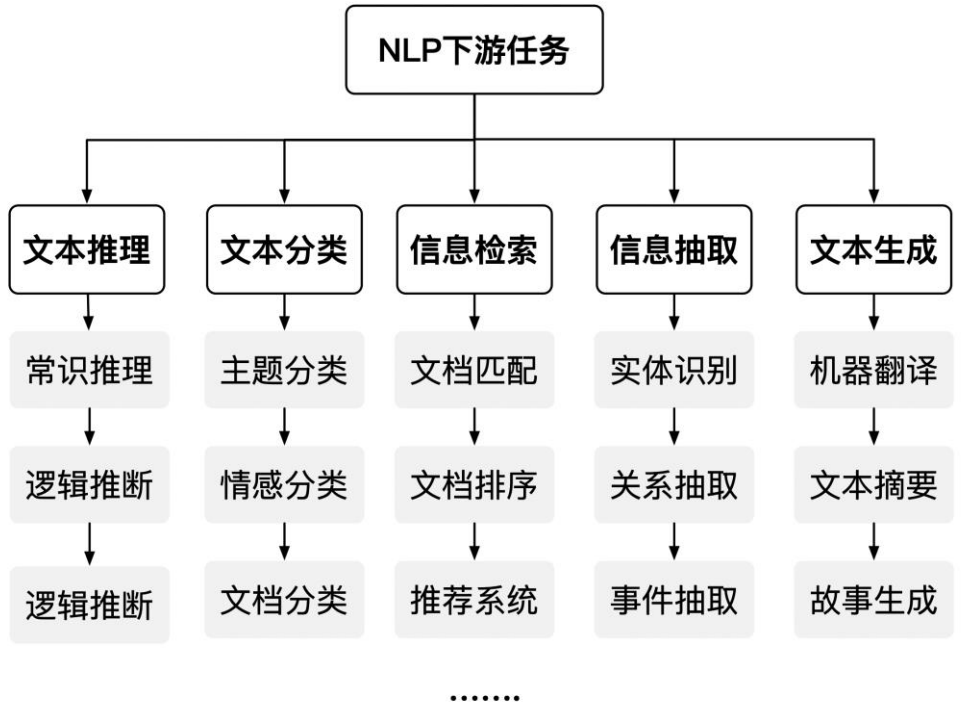


挑战：大模型微调难

- 大模型在微调中更新所有参数，需要在**计算**中占用庞大的显存
- 大模型在微调中对每一个任务存储一个大模型，需要在占用庞大的**存储**空间



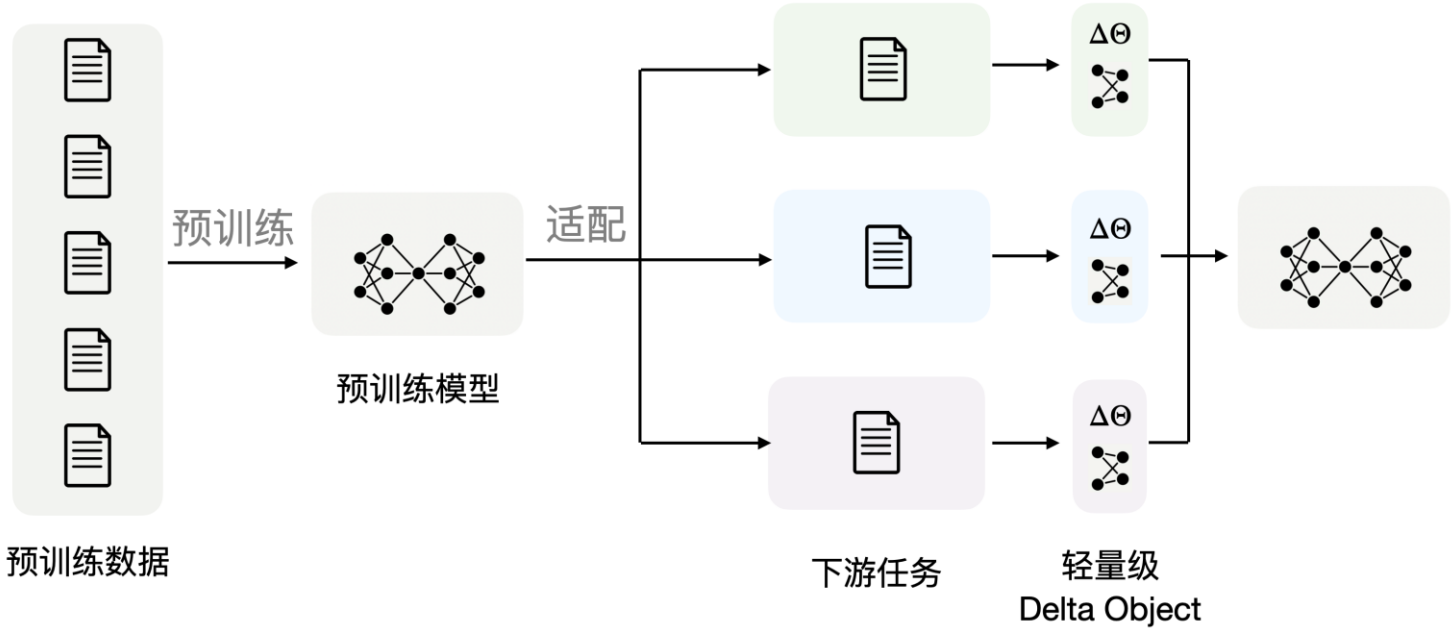
模型的参数规模越来越大，使之更加难以微调



任务种类繁多，难以对所有任务都进行全参数微调

思路：仅优化小部分参数即可达到全参数微调效果

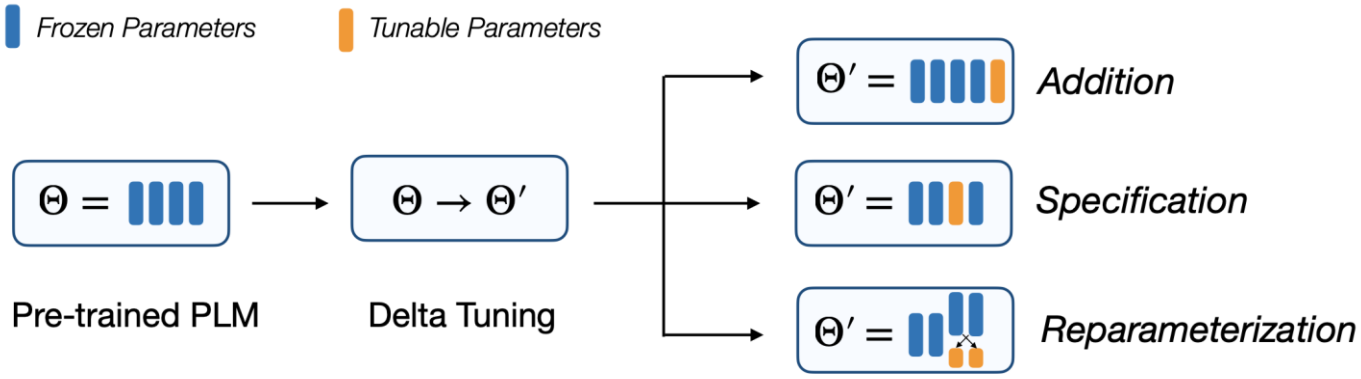
- 大模型参数高效微调（Delta Tuning）：**保持模型大部分参数不变，仅更新极少参数**（<1%）实现适配下游任务，显著降低计算和存储成本
- 计算：在训练过程中可以节省**50%~70%**训练显存
- 存储：对于千亿级模型+100个适配任务，最多节省**10000**倍存储开销



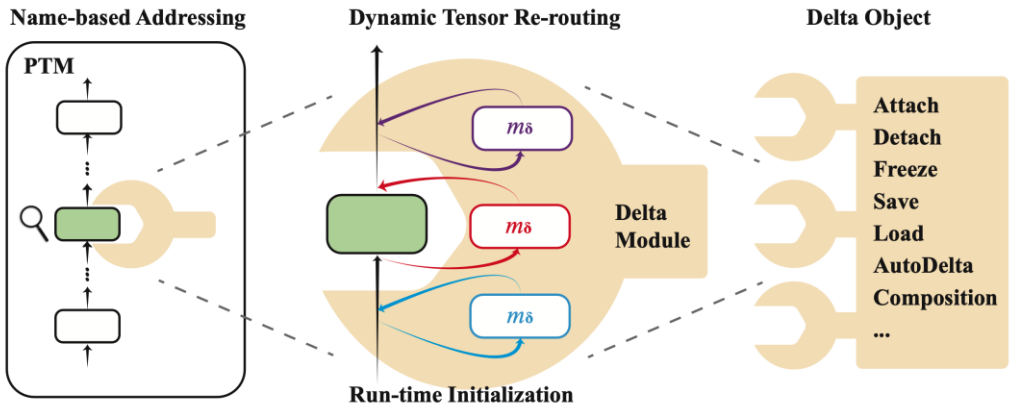
Delta Tuning只需训练和保留轻量级的参数集，可以视为特定能力的插件

特性1：大模型 Delta Tuning 可对下游任务进行高效适配

- Delta Tuning论文首次对此范式进行了系统性的理论和实验分析，被 **Nature Machine Intelligence** 接收
- OpenDelta是首个支持Delta Tuning的工具框架，不需要修改模型代码即可实现对**任意模型的任意位置**的Delta Tuning，GitHub上获得超过**400**星标



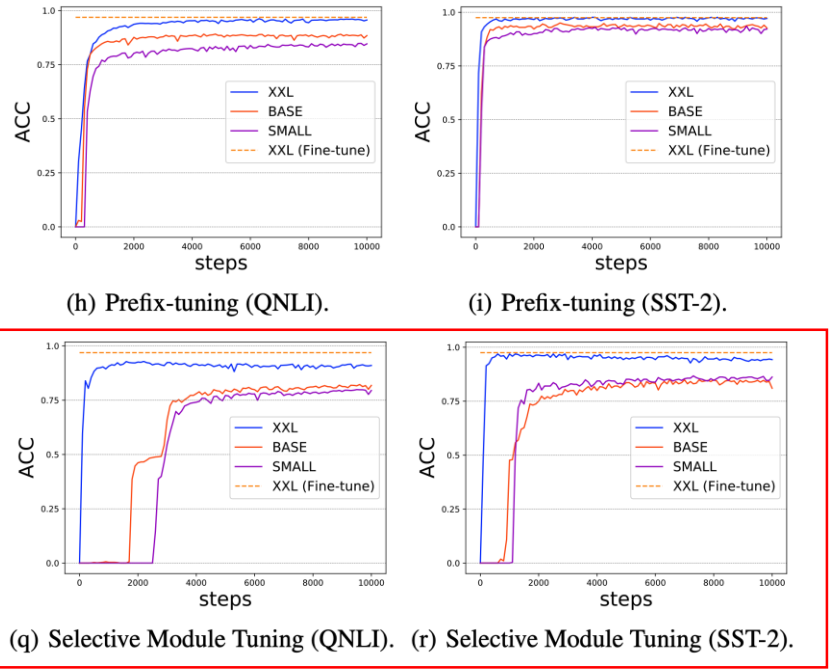
统一视角下的Delta Tuning方法



OpenDelta工具原理图

特性1：大模型 Delta Tuning 可对下游任务进行高效适配

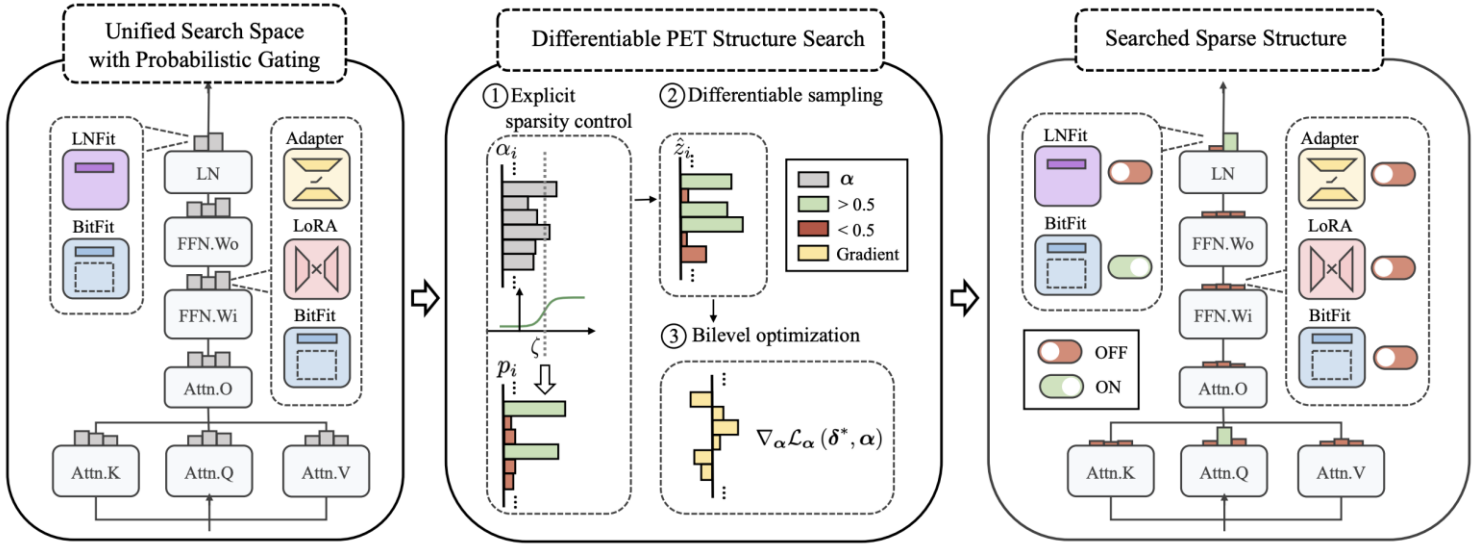
- 对于百亿以上规模的基础模型，优化**任意参数**即可达到全参数微调的效果
- Delta Tuning结构可以通过神经网络搜索（NAS）技术**自动构建**



(h) Prefix-tuning (QNLI). (i) Prefix-tuning (SST-2).

(q) Selective Module Tuning (QNLI). (r) Selective Module Tuning (SST-2).

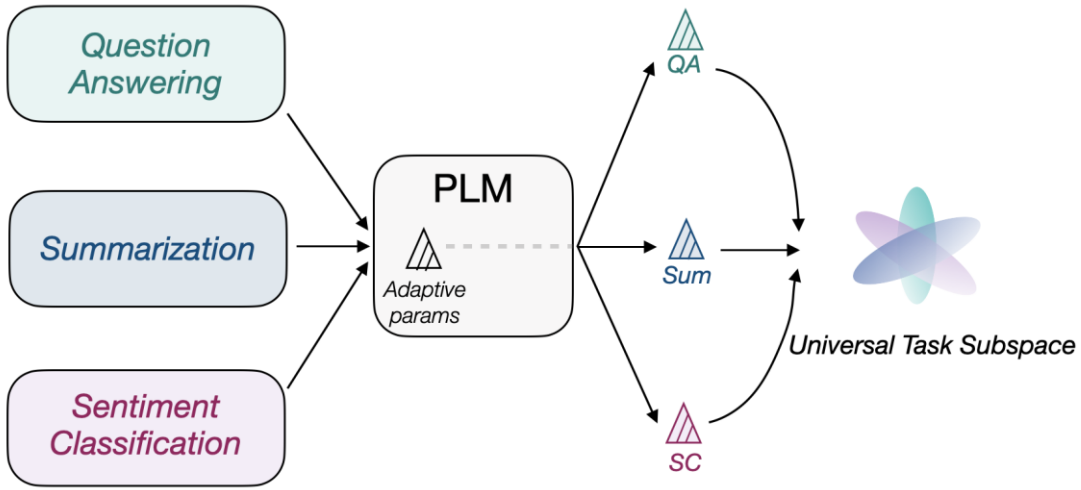
优化任意参数也可在大模型上完成适配



全自动化地对Delta Tuning模块进行神经结构搜索 [NeurIPS 2022]

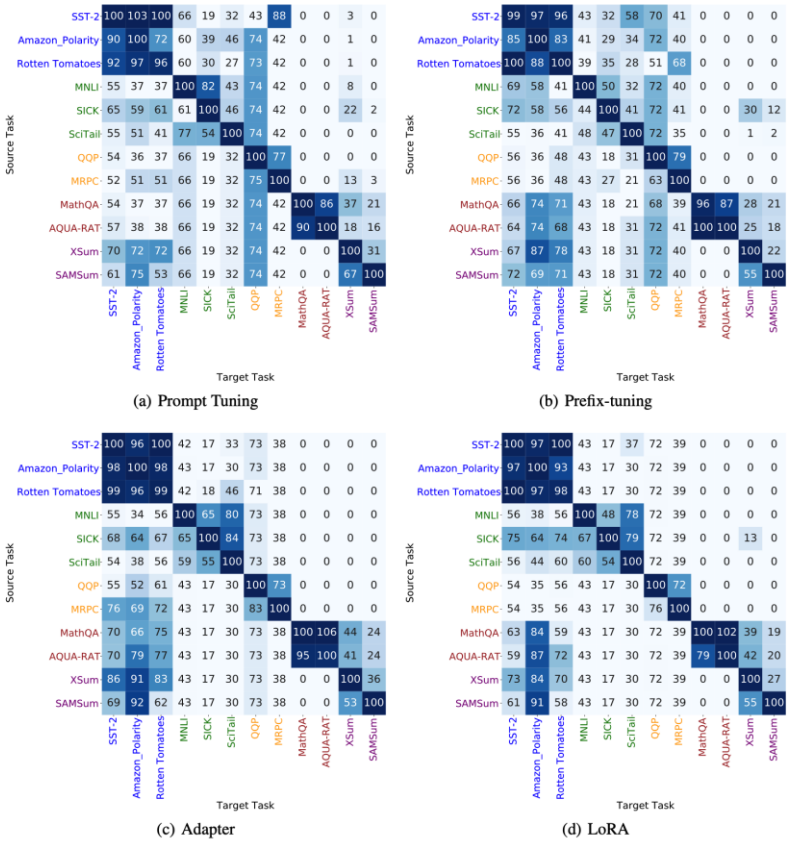
特性1：大模型 Delta Tuning 可对下游任务进行高效适配

对于同一个大模型，不同任务之间存在较强的**迁移能力**，迁移性与任务间相似度等因素密切相关



任务之间存在一个低维的本征空间

在此空间中的解可以同时解决多个任务

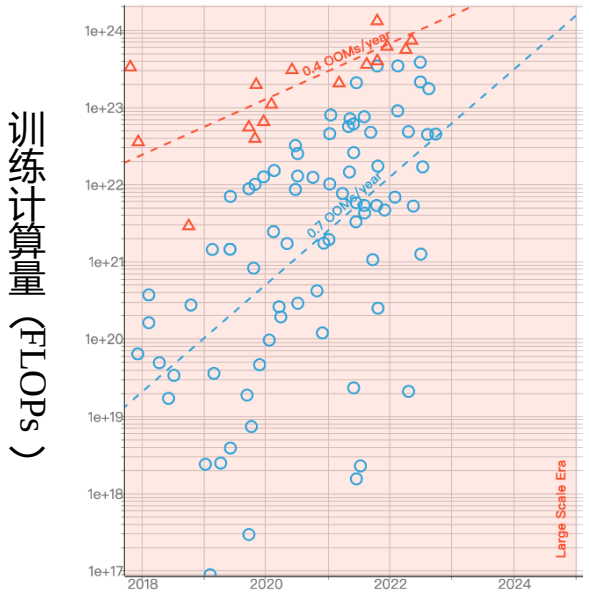


Delta Tuning展示了较强的任务迁移能力

挑战：大模型计算复杂度高、应用模式受限

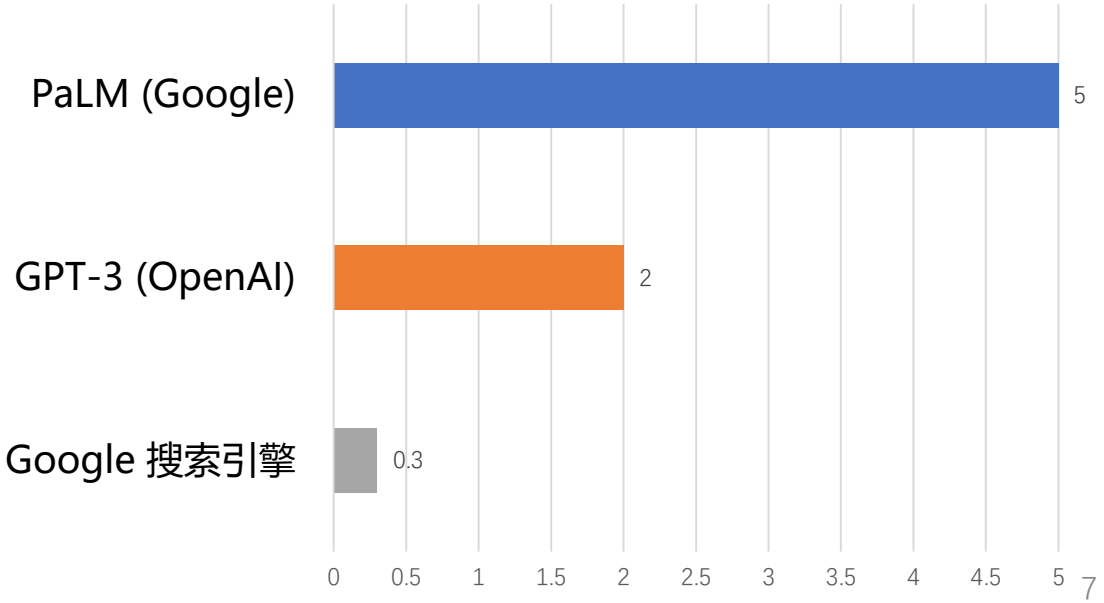
- 大模型需要庞大的计算资源、高配置的计算设备支撑计算
- 存在极高计算时延，难以满足实际应用中快速响应的需求

数据与参数的增长导致计算量呈指数级增长

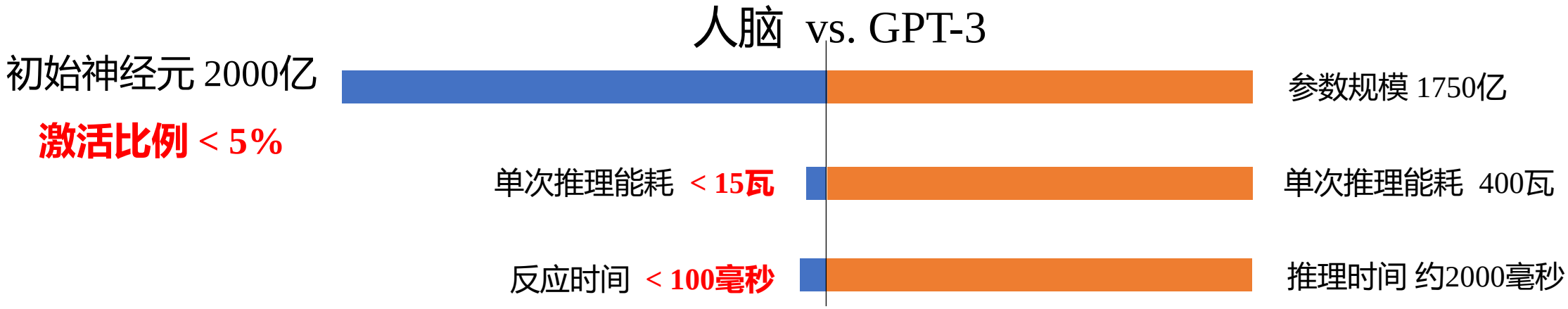


(红色为超大规模预训练模型)

GPT-3、PaLM 与 Google 搜索引擎的推理耗时对比（秒）



思路：人脑能以更低能耗、更快反应时间实现复杂认知



Toward Next-Generation Artificial Intelligence: Catalyzing the NeuroAI Revolution

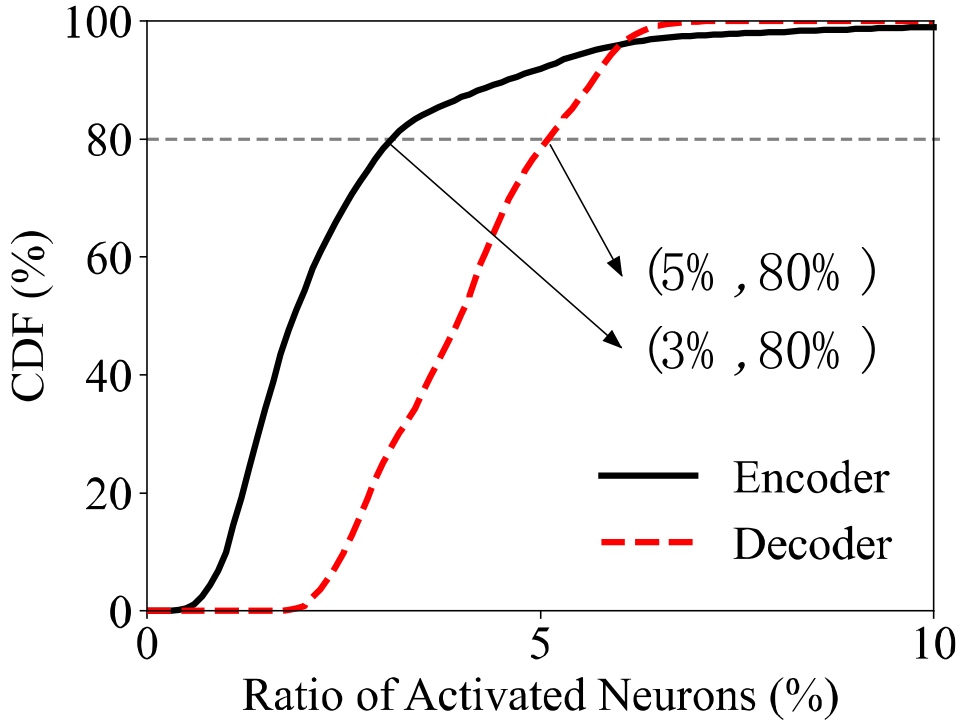
Over the coming decades, Artificial Intelligence (AI) will transform society and the world economy in ways that are as profound as the computer revolution of the last half century, and likely at an even faster pace. This AI revolution presents tremendous opportunities to unleash human creativity in the modern economy. New developments in AI systems have the potential to enable workers to attain greater productivity and relieve them from performing the most dangerous and menial jobs. But, to reach this potential, we still require advances that will make AI more human-like in its capabilities. Historically, neuroscience has been a key driver and source of inspiration for improvements in AI, particularly those that made AI more proficient in areas that humans and other animals excel at, such as vision, reward-based learning, interacting with the physical world, and language (Hassabis et al. 2017). It can still play this role. To accelerate progress in AI and realize its vast potential, we must invest in fundamental research in "NeuroAI".

2022年，Yann LeCun 和 Yoshua Bengio 联合发布NeuroAI白皮书指出，神经科学将为下一代人工智能提供指导

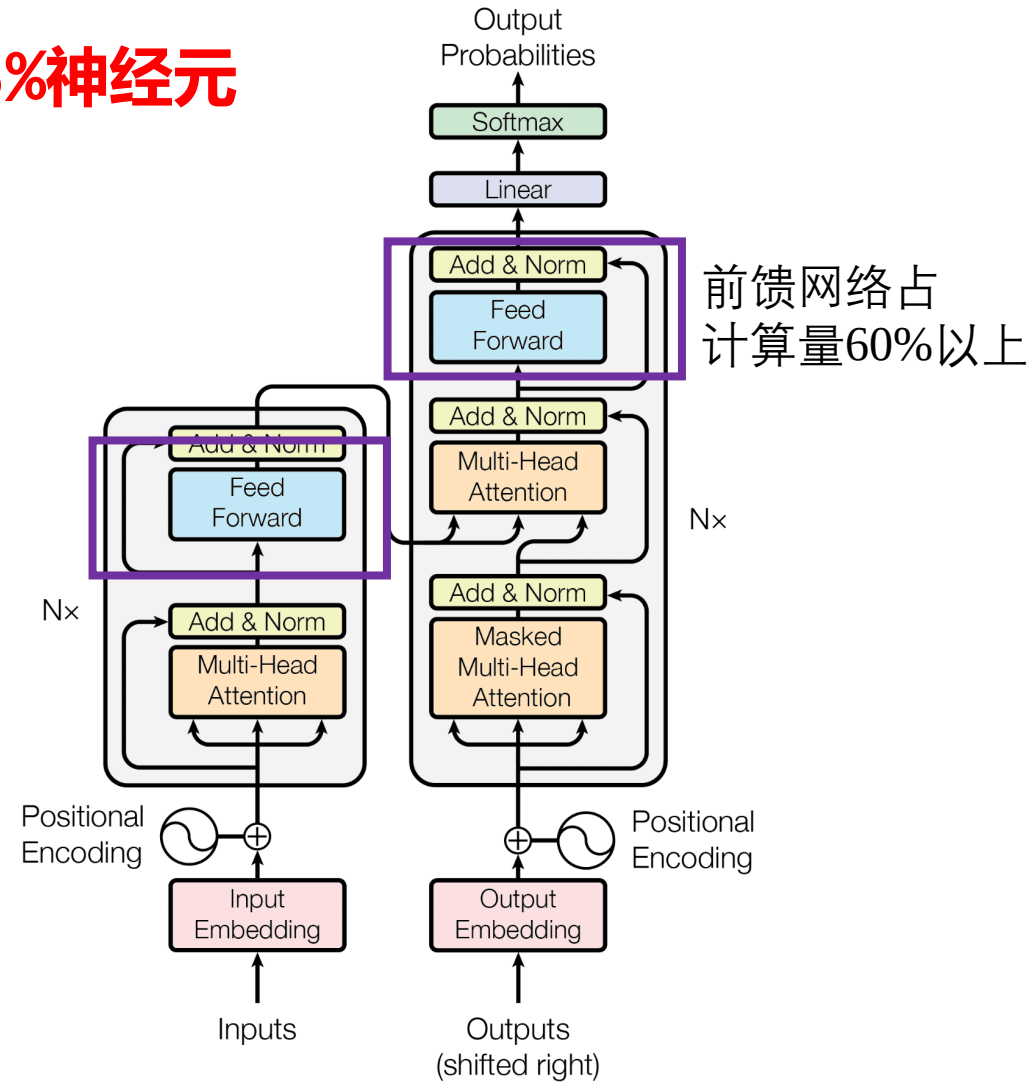
The search for true numbers of neurons and glial cells in the human brain: A review of 150 years of cell counting[J]. Journal of Comparative Neurology, 2016.
The economy of brain network organization[J]. Nature reviews neuroscience, 2012.
The cost of cortical computation[J]. Current biology, 2003.

特性2：大模型涌现稀疏激活特性，可持续增长成为可能

- 大模型稀疏激活：**80%输入只激活前馈网络<5%神经元**
- 几乎所有神经元都会在某些输入中被激活



输入激活神经元比例的累积分布函数（基于T5-Large，7亿参数量）



前馈网络占计算量60%以上

特性2：大模型涌现稀疏激活特性，可持续增长成为可能

MoEfication首次报告基础模型存在自发的**功能分化**和**稀疏激活**现象，并提出稠密模型专家化算法，启发 Google Research 进行更细粒度功能模块分析

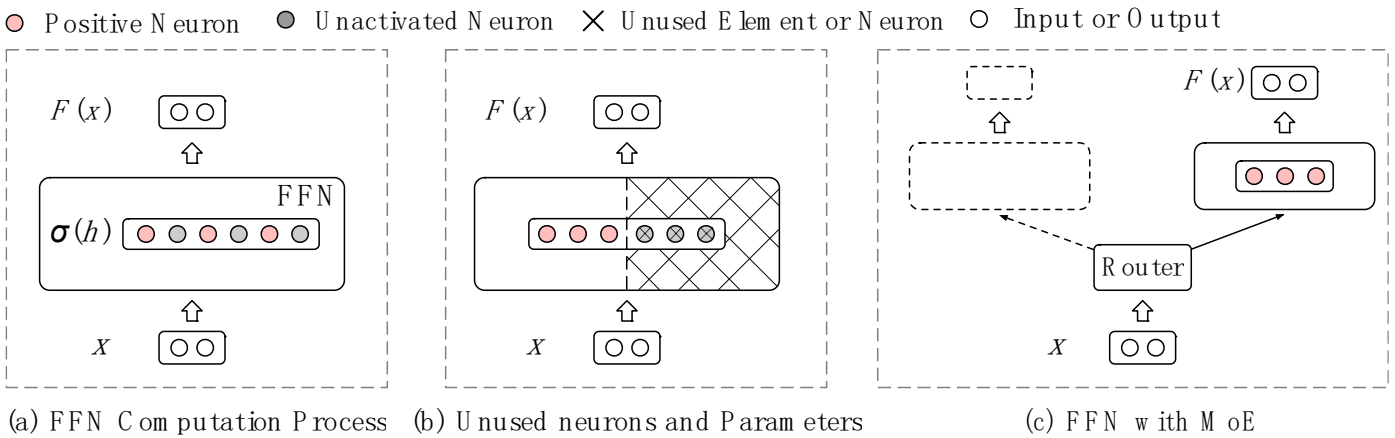
MoEfication: Transformer Feed-forward Layers are Mixtures of Experts

Zhengyan Zhang, Yankai Lin, Zhiyuan Liu, Peng Li, Maosong Sun, Jie Zhou

Large Models are Parsimonious Learners: Activation Sparsity in Trained Transformers

Google Research, New York City

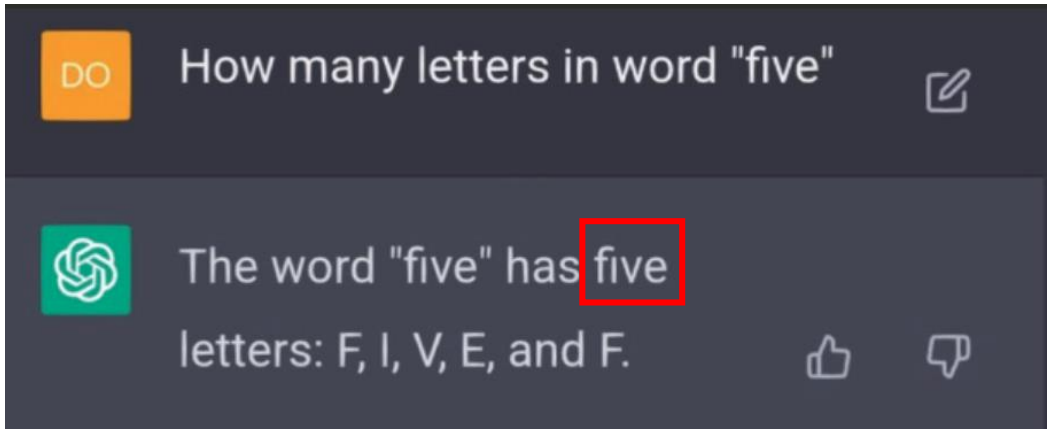
s worth noting that the observation that Transformers produce sparse activations is previ
Our paper significantly extends upon results in [9] to demonstrate that sparsity emerges
ers of Transformers, for both language and vision tasks, on both training and evaluatior
architectures beyond Transformers. We also examine the activation of individual neurons
ty is not caused by "dead" neurons and that the percentage of activation has a long tail d



(a) FFN Computation Process (b) Unused neurons and Parameters (c) FFN with MoE

挑战：大模型计算过程不可解释、可靠性低

- 大模型内部结构和运算过程**难以被直观理解**
- **无法解释**模型输出结果，影响其在实际应用中的可靠性和稳定性

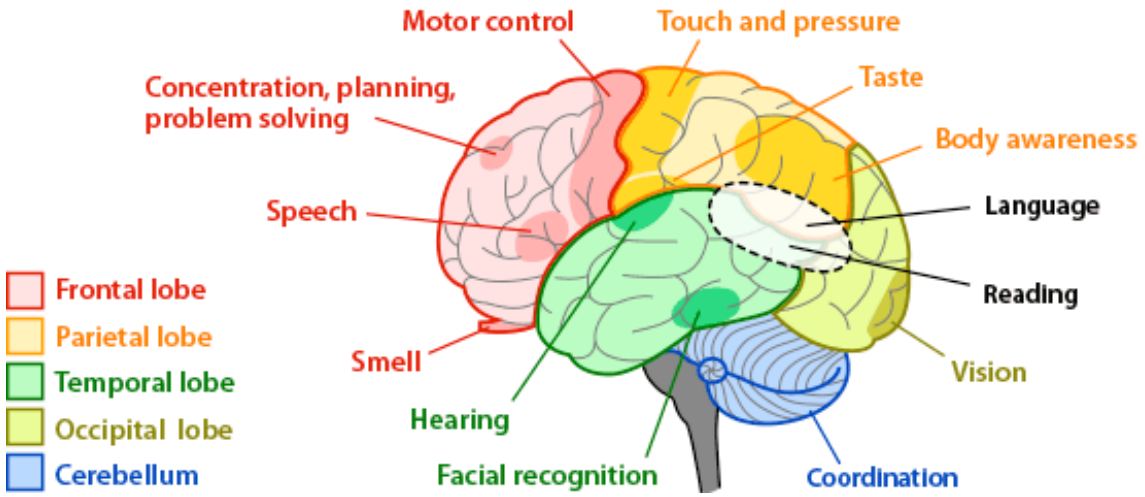


用户的困惑：

- 为什么输出这个结果？
- 为什么不输出别的结果？
- 什么时候应该相信模型？
- 如何改正模型？

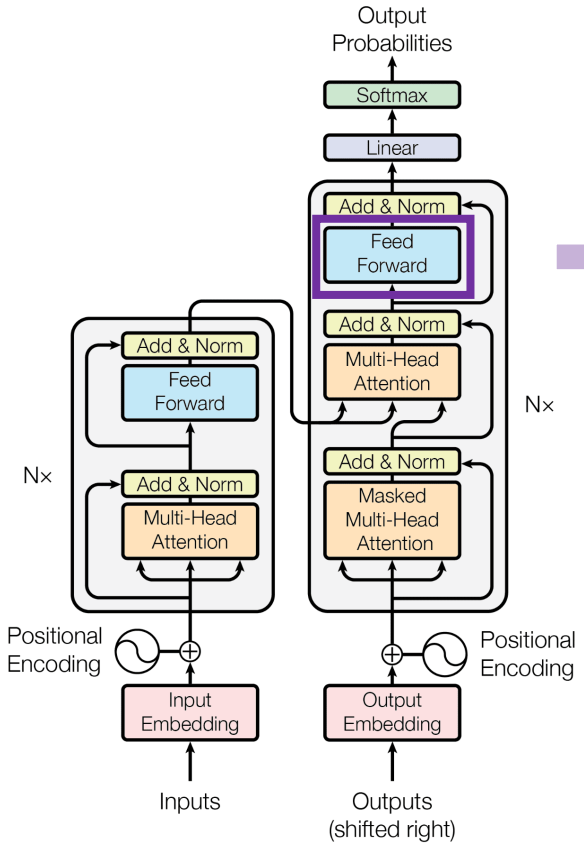
思路：定位大模型中对应知识、技能的特异性神经元

人脑存在明确的神经元功能分化



人脑不同分区对应不同功能

探究大模型参数神经元的功能特异性



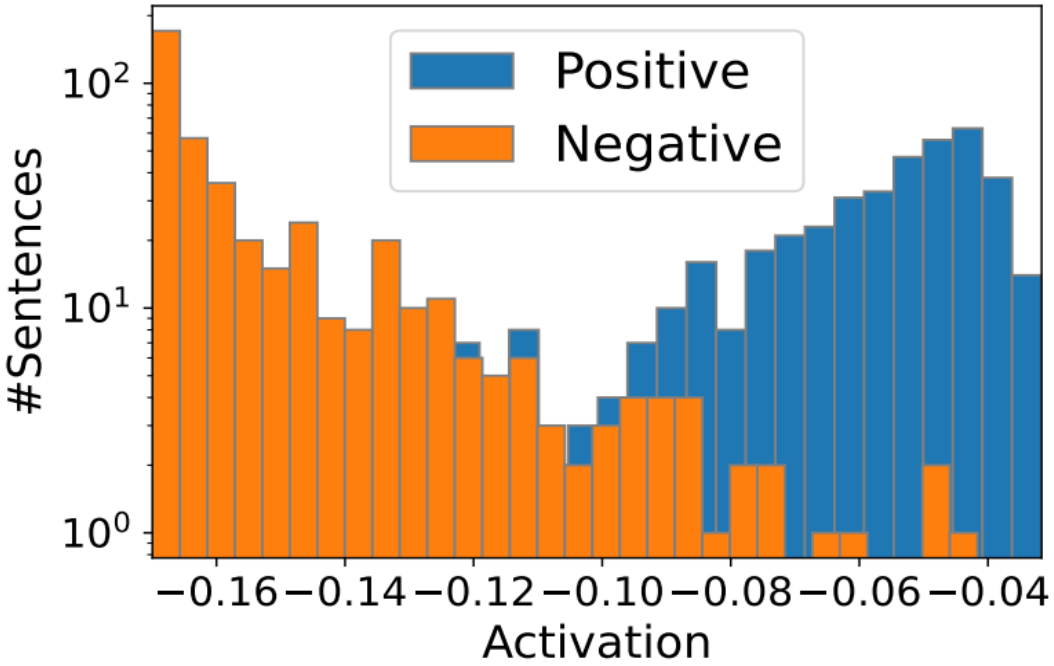
事实知识
情感
……

基于特异性神经元
进行针对性解释及干预

Knowledge Neurons in Pretrained Transformers[C]. Proceedings of ACL, 2022.

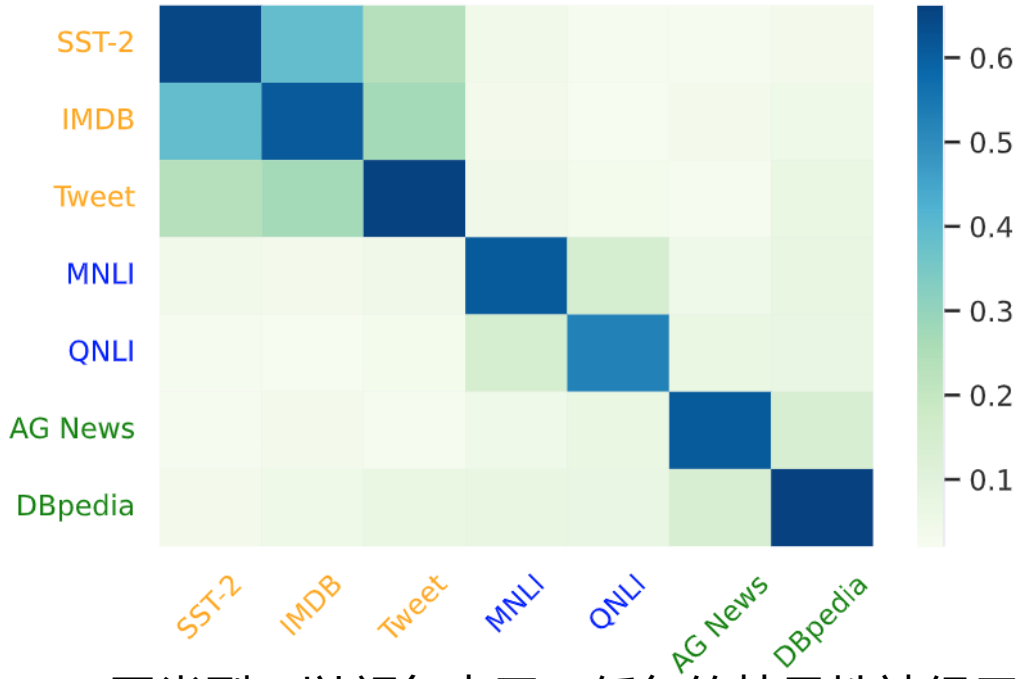
特性3：大模型存在特异性神经元，为解释性带来可能性

某些神经元的激活对任务标签有高度特异性



RoBERTa-Base模型（1亿参数量）中某神经元对SST-2情感分析任务中两种不同标签数据的激活值分布

相似任务具有相似的特异性神经元分布

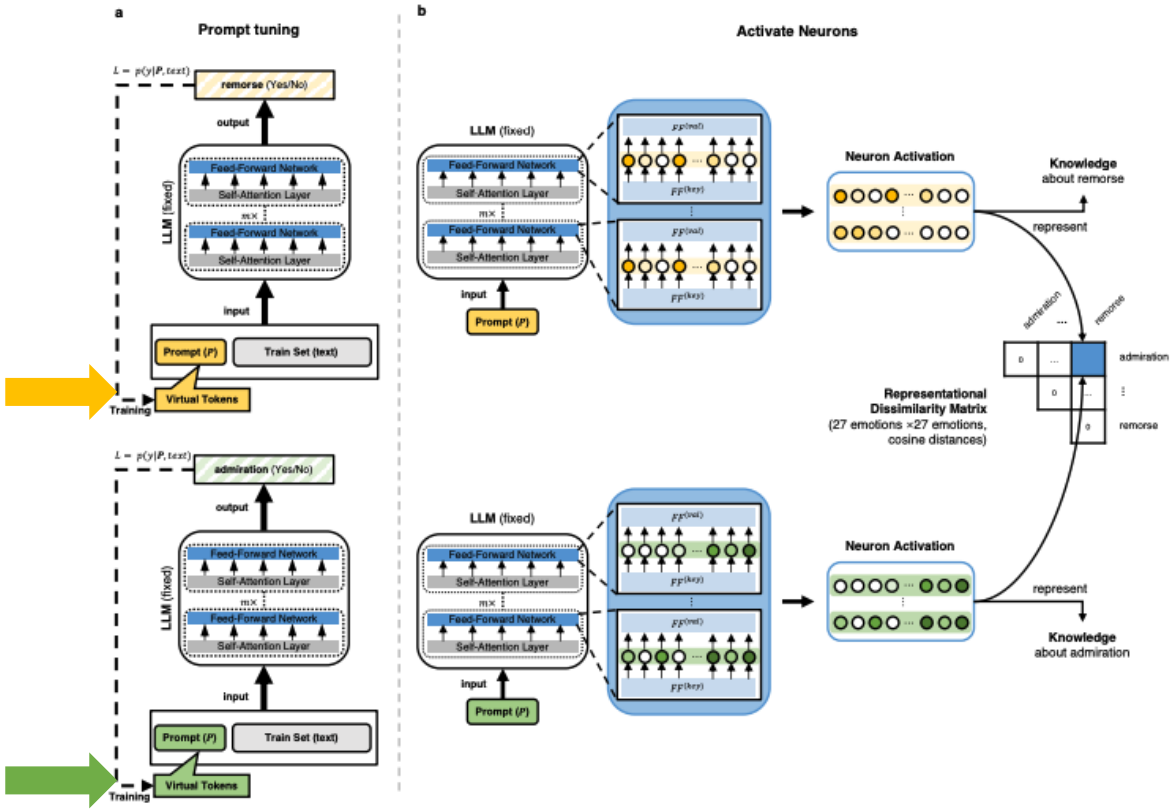


不同类型（以颜色表示）任务的特异性神经元分布之间的相关系数

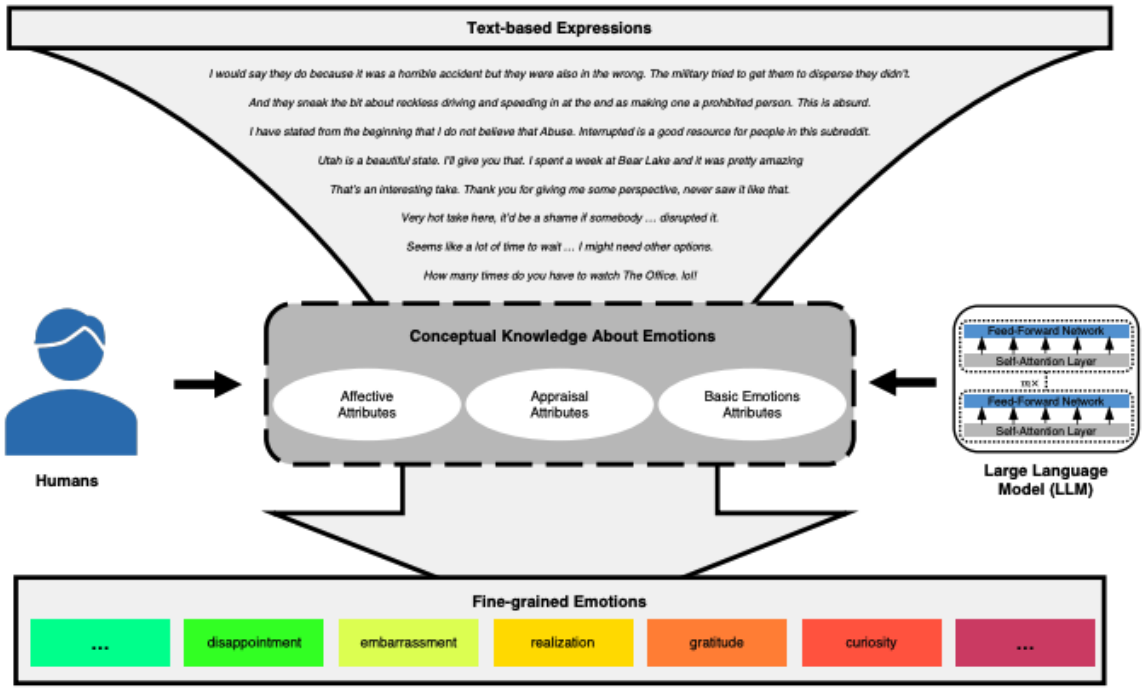
Finding Skill Neurons in Pre-trained Transformer-based Language Models[C]. Proceedings of EMNLP, 2022.

特性3：大模型存在特异性神经元，为解释性带来可能性

相似任务的神经元激活状态相似，能够指示任务间迁移性



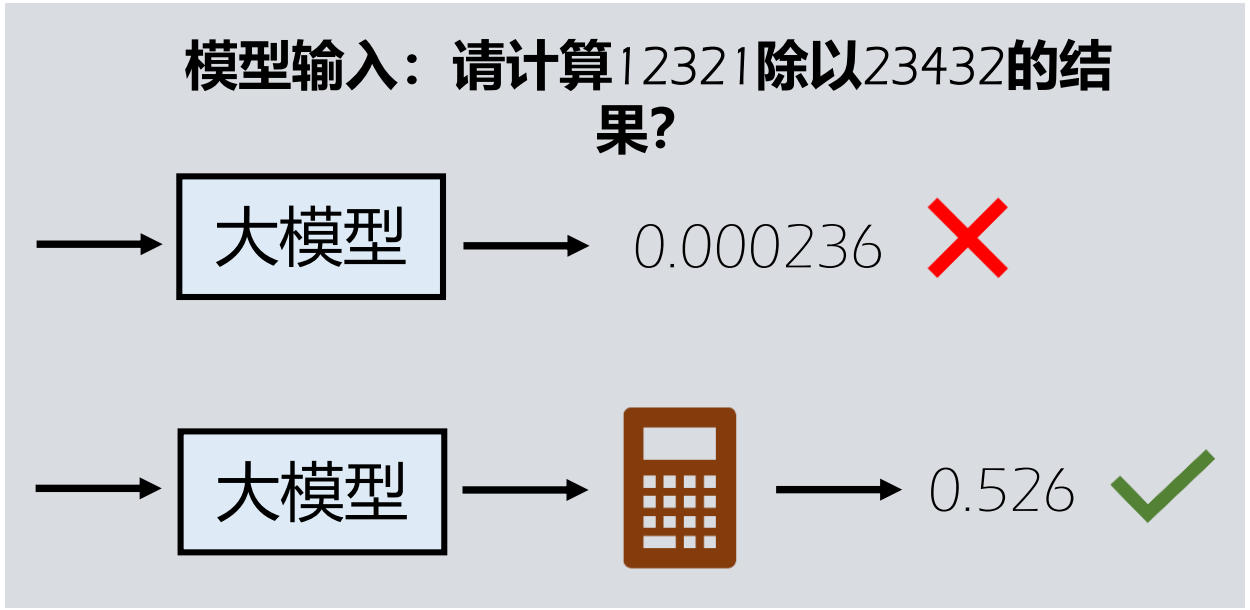
预训练语言模型涌现**类人情感概念**对应的神经元

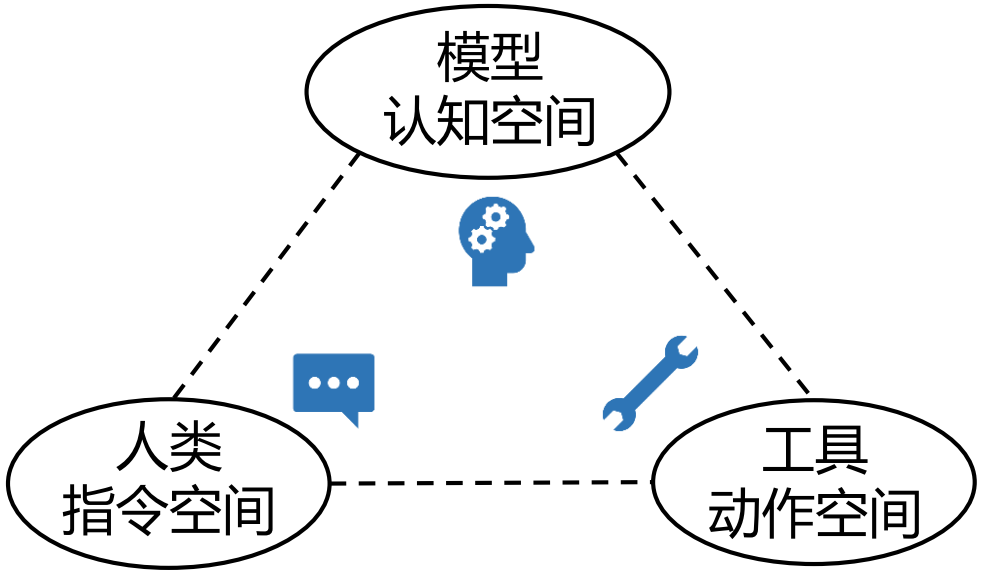


[1] On transferability of prompt tuning for natural language processing[C]. Proceedings of NAACL, 2022.
[2] Human Emotion Knowledge Representation Emerges in Large Language Models and Supports Discrete Emotion Inference[J]. Arxiv 2023.

挑战：大模型无法利用外部复杂工具，生成信息缺少依据

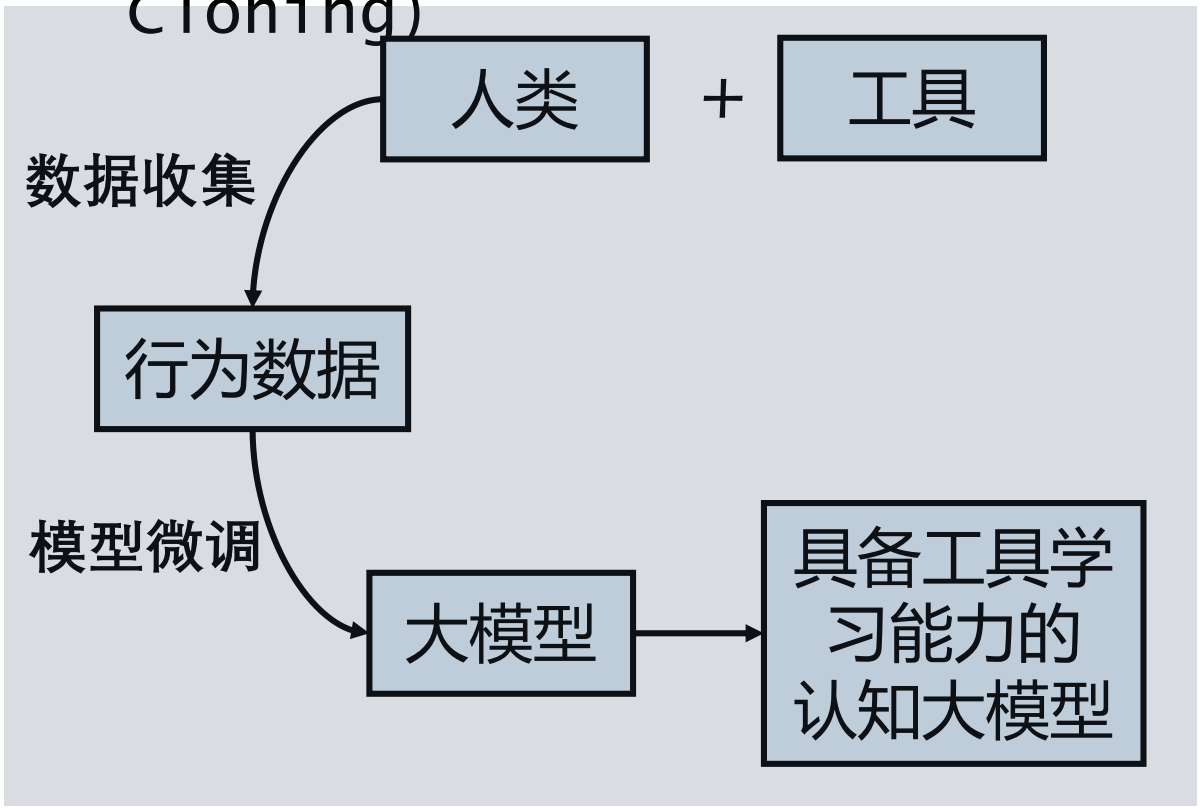
- 预训练模型尚未建立对外部**复杂工具**的认知和应用能力，无法像人一样操作工具
- **工具学习**能力的缺失，是造成当前大模型智能程度低，通用性差的直接原因

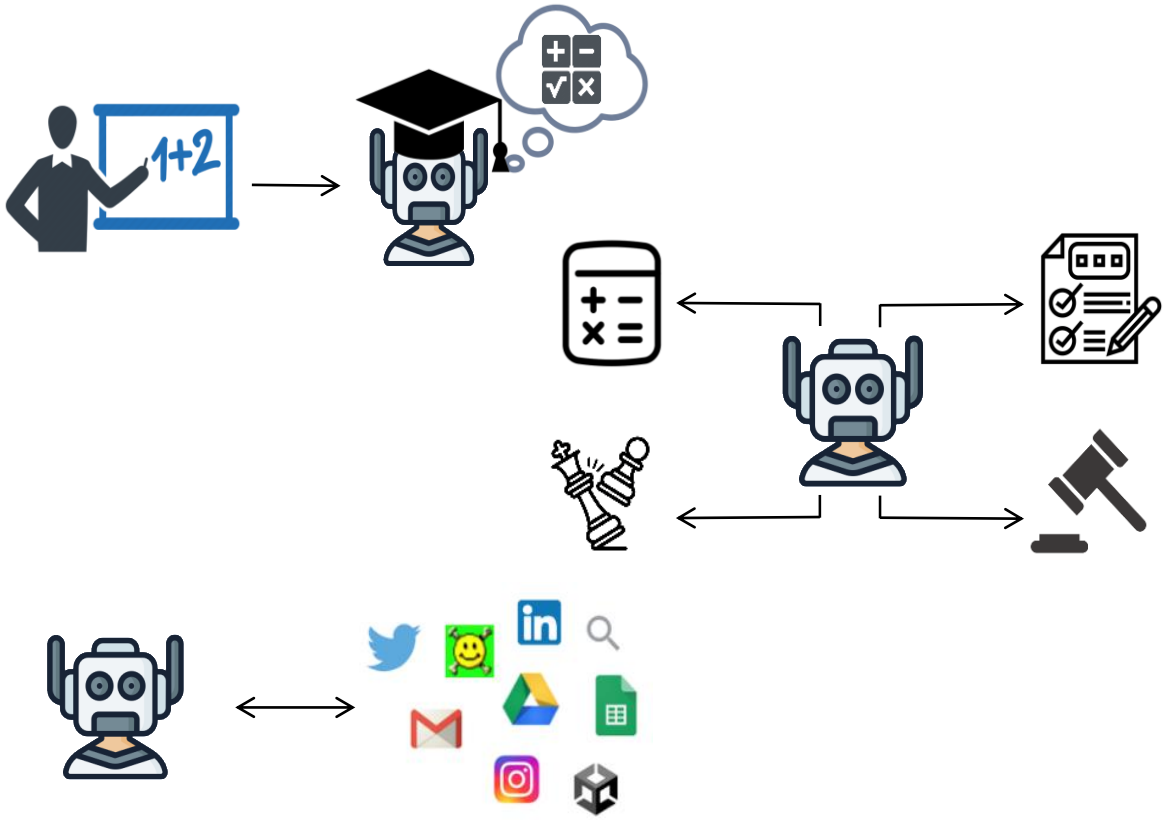




思路：模拟人类学会使用复杂工具

- 人类能够综合使用各种工具来解决问题，综合多种知识来源，从而完成复杂任务
- 记录人类使用工具的行为数据，并让大模型**模仿人类行为**（Behavior Cloning）

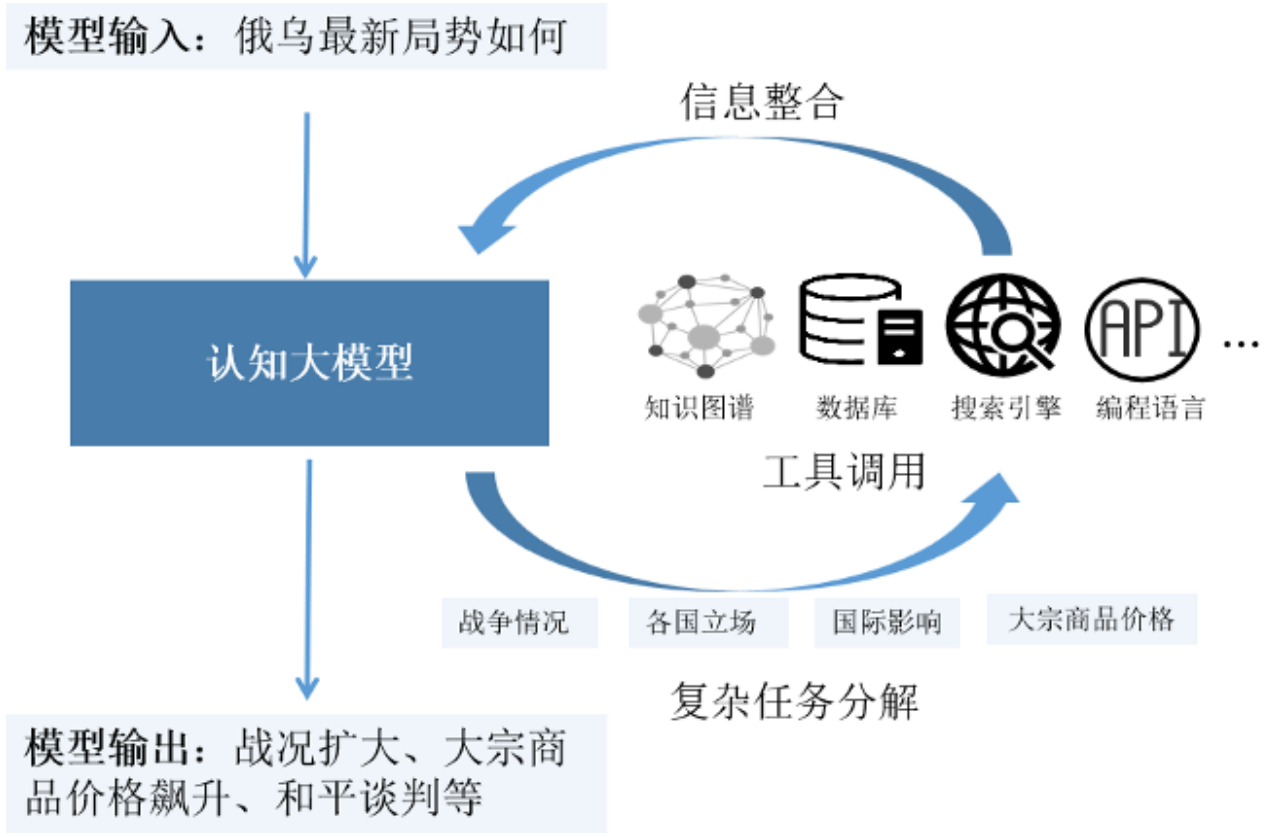


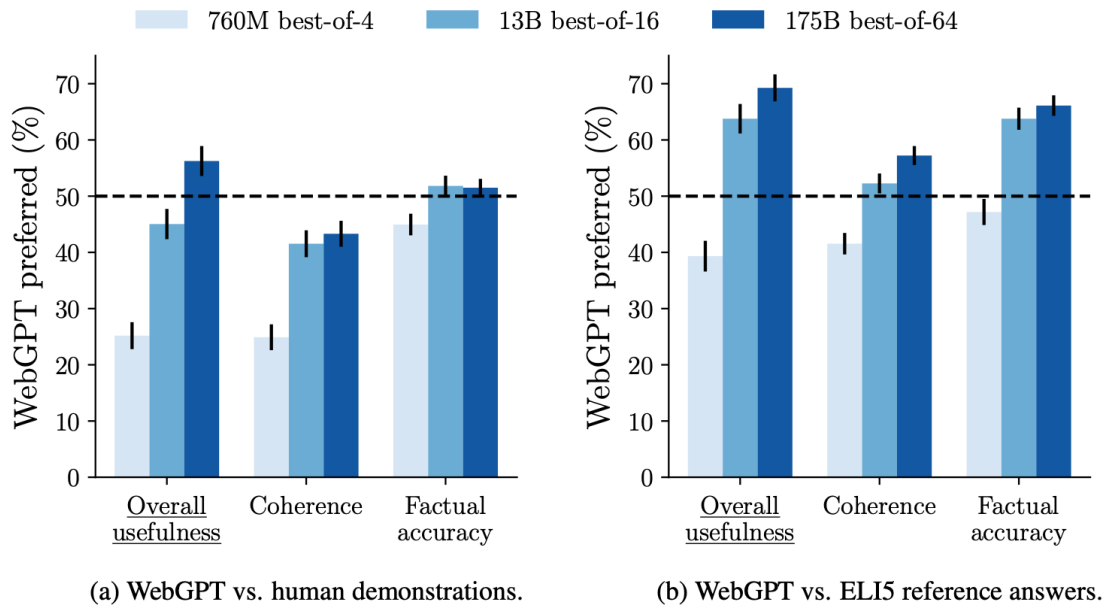


特性4：大模型可有效建模用户行为，掌握高级认知工具

大模型可以掌握搜索引擎、知识图谱等高级认知工具，完成复杂任务

OpenAI [1] 研究表明通过模仿人类行为，大模型具备**超越人类**的搜索引擎使用能力





(a) WebGPT vs. human demonstrations.

(b) WebGPT vs. ELI5 reference answers.

[1] Nakano, Reiichiro, et al. "Webgpt: Browser-assisted question-answering with human feedback." arXiv preprint arXiv:2112.09332 (2021).

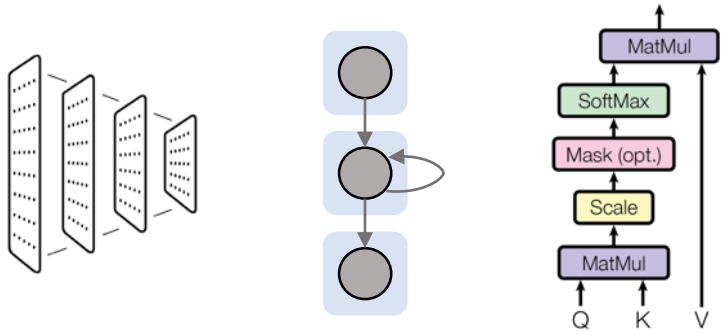
特性5：大模型具备强大通用性

通用框架

有效降低**开发成本**

Transformer

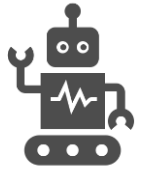
CNN | RNN | ATT



通用能力

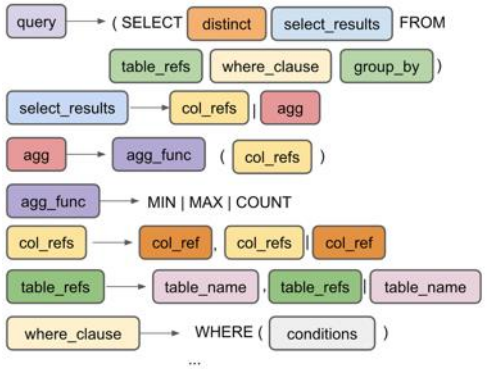
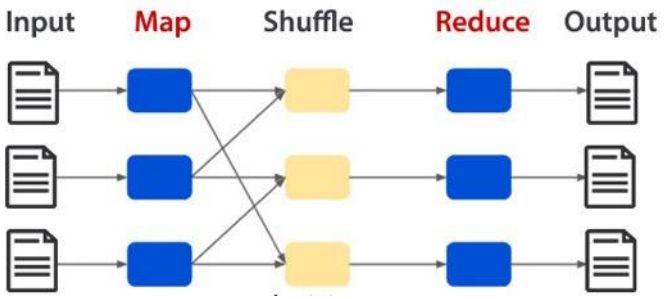
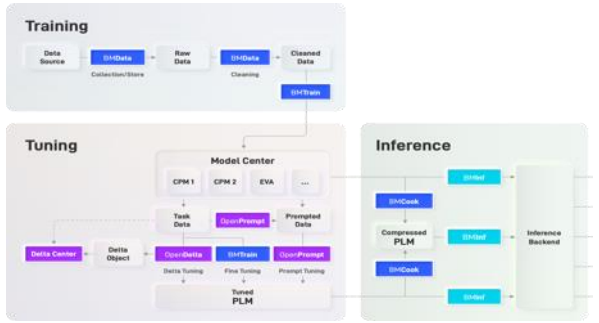
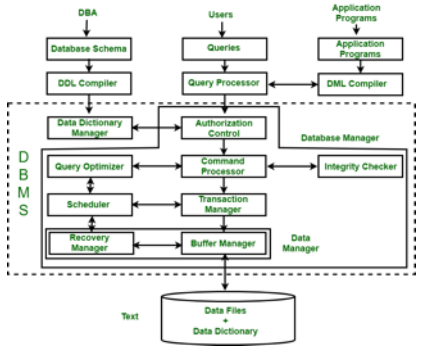
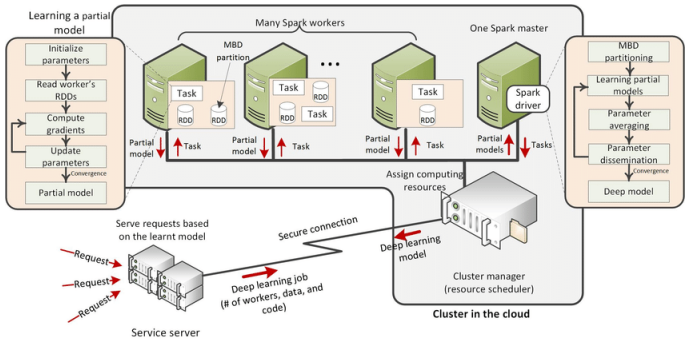
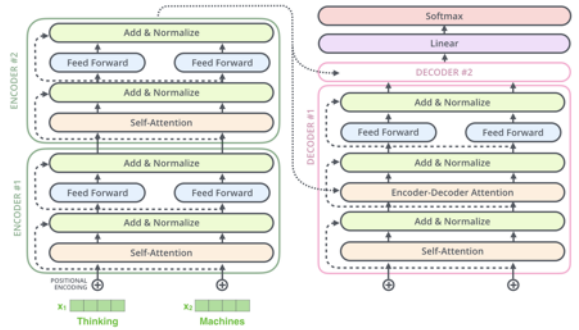
有效降低**适配成本**

预训练大模型



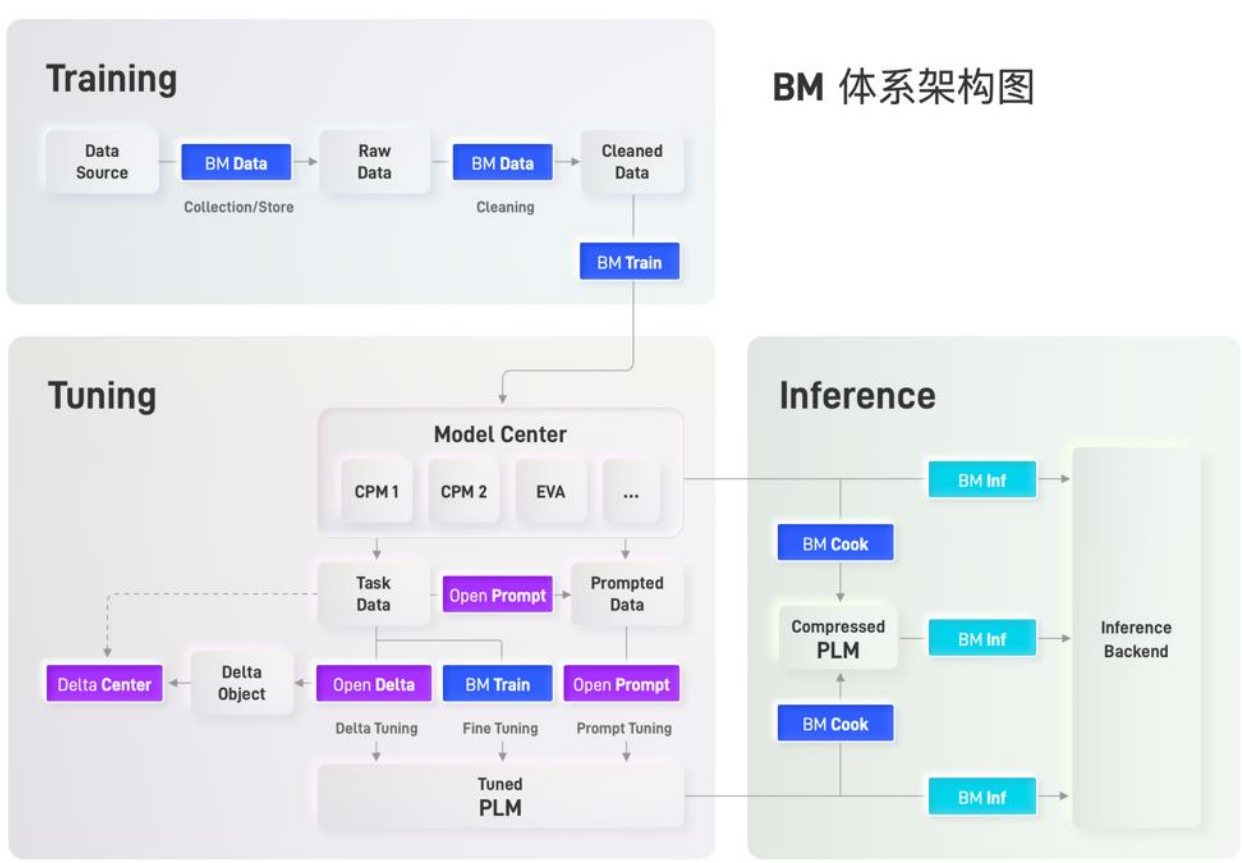
机器翻译 | 信息检索 | 文本生成 | 智能问答

特性5：大模型将成为智能时代基础设施

	数据库（1970s）	大数据分析（2010s）	智能大模型（2020s）
逻辑层	 SQL	 Map-Reduce	 OpenBMB
物理层	 并发处理	 分布式计算	 训练微调推理

OpenBMB开源社区

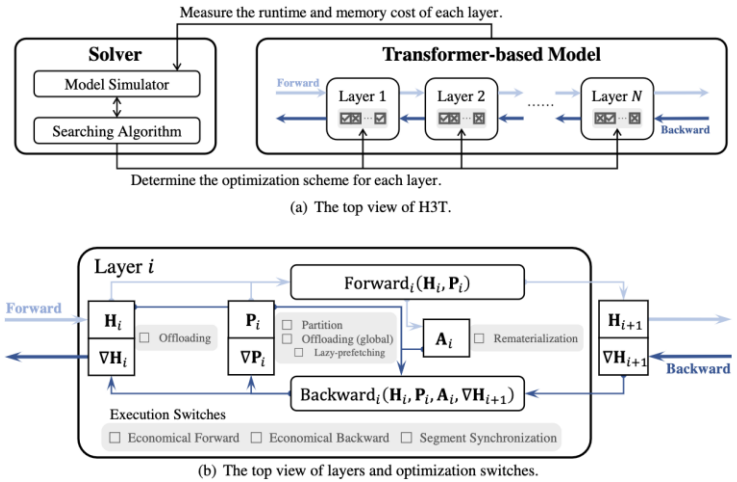
支持高效训练、微调、压缩、推理的大模型全流程高效计算框架



OpenBMB开源社区

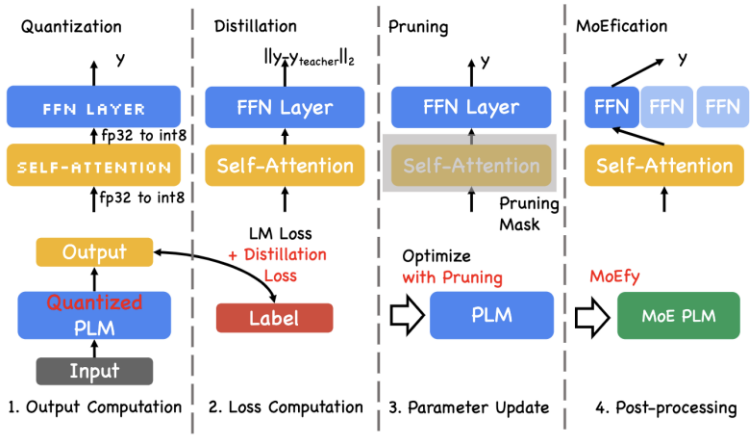
模型训练

- BMTrain: 大模型高效预训练
  - 基于动态规划自动进行分布式与显存优化
  - 64张 A100 可训练 GPT-3，使GPT3等训练成本**减少90%**以上，由1200万美元降至75万美元



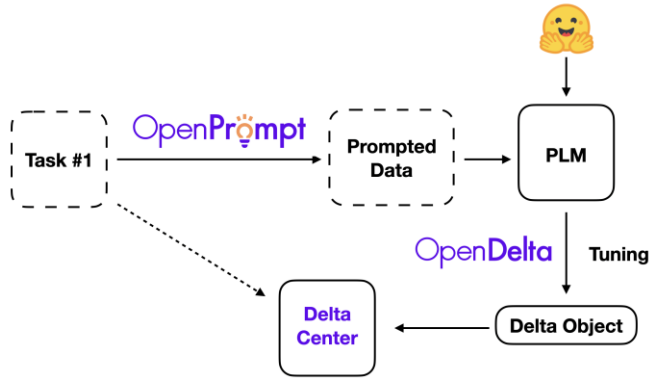
模型压缩与推理

- BMCook：大模型高效压缩
  - 提供一系列模型压缩算法， 提升运行时效率超过**10倍**，发表在EMNLP 2022
- BMInf：大模型高效推理
  - 低资源推理优化，实现了 GTX**1060 推理百亿参数大模型**，发表在 ACL 2022



模型微调

- OpenPrompt：大模型提示学习工具包
  - 获得 **ACL 2022 最佳演示系统论文奖**
  - GitHub 获 2.3k 星标，被牛津大学、百度等机构使用参考
- OpenDelta：大模型增量微调工具包
  - 建立可视化参数高效微调框架
  - 提出 Delta Tuning 概念，2022年12月被 Nature Machine Intelligence 论文录用



CPM-Live持续学习大模型

CPM-Live

- 开源共建：我国首个**基于开源社区打造**的中文基础大模型
- 细节公开：公开代码、超参、训练日志
- 持续学习：持续更新计算框架、数据规模、模型能力
- 支持多语言多能力
  - 支持中英双语，囊括分类、生成、问答、摘要、翻译、信息抽取等10余类细分能力、2000+种细分能力
- 开放通用
  - 支持开放域信息抽取、自动问答等开放能力，无需微调即可适配各种常见业务需求

CPM-Live (Ant)

- 训练68天，采用4*A100算力，使用600+GB高质量数据，训练花费为43万
- 仅**微调 0.06% 参数**超越“智源指数”上全参数微调效果
- 通过高效压缩，发布多粒度的模型版本

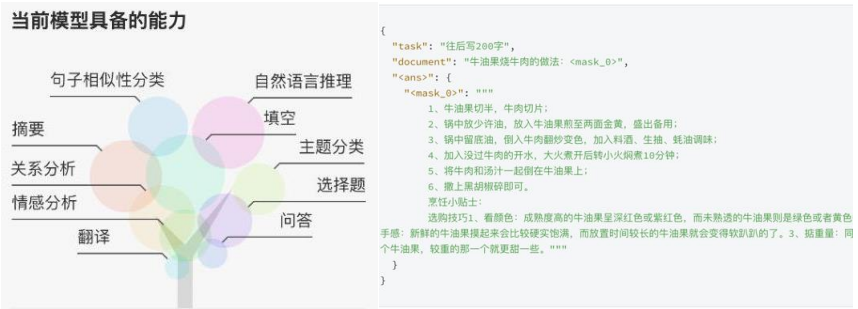
Model	Model Params	Tuned Params	CCPM	KdConv	Sogou-Log
CPM-2	11B	11B	91.6	13.12	35.9
Yuan 1.0	13B	UNK	89.71	18.55	28.12
CPM-Ant-10B	10B	**6.3M**	**92.13**	**18.57**	**38.02**

	层数	24	36	48	60	72	96
隐状态维度	12288						175B (GPT3)
	8192					60B	
	5120				20B	25B	
	4096		7B	10B CPM-Ant			
	2560		3B				
	2048	1B					
	1024	0.3B					

- : 发布模型尺寸
- : 目标模型尺寸

CPM-Live (Bee)

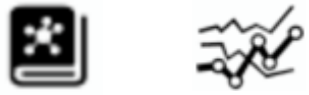
- 训练94天，采用4*A100算力，使用2000+GB高质量数据，训练花费为47万
- **登顶零样本学习榜单ZeroCLUE**
- 支持结构化输入输出，能力进一步丰富



总结

- 大模型能够充分利用无标注大数据和知识，具有良好的通用性和性能，为AI研究应用带来全新可能
- ChatGPT 吹响了大模型迈向通用智能的号角，新的纪元刚刚开启
- 大模型特性亟待挖掘利用：增量微调、稀疏激活、特异化、工具学习、通用性

大规模无标注数据



领域数据与知识

智能大模型

Driving

Math

Control

Debating

Medicine

History