

大语言模型基础

张倬胜

zhangzs@sjtu.edu.cn

<https://bcmi.sjtu.edu.cn/~zhangzs/>

目录

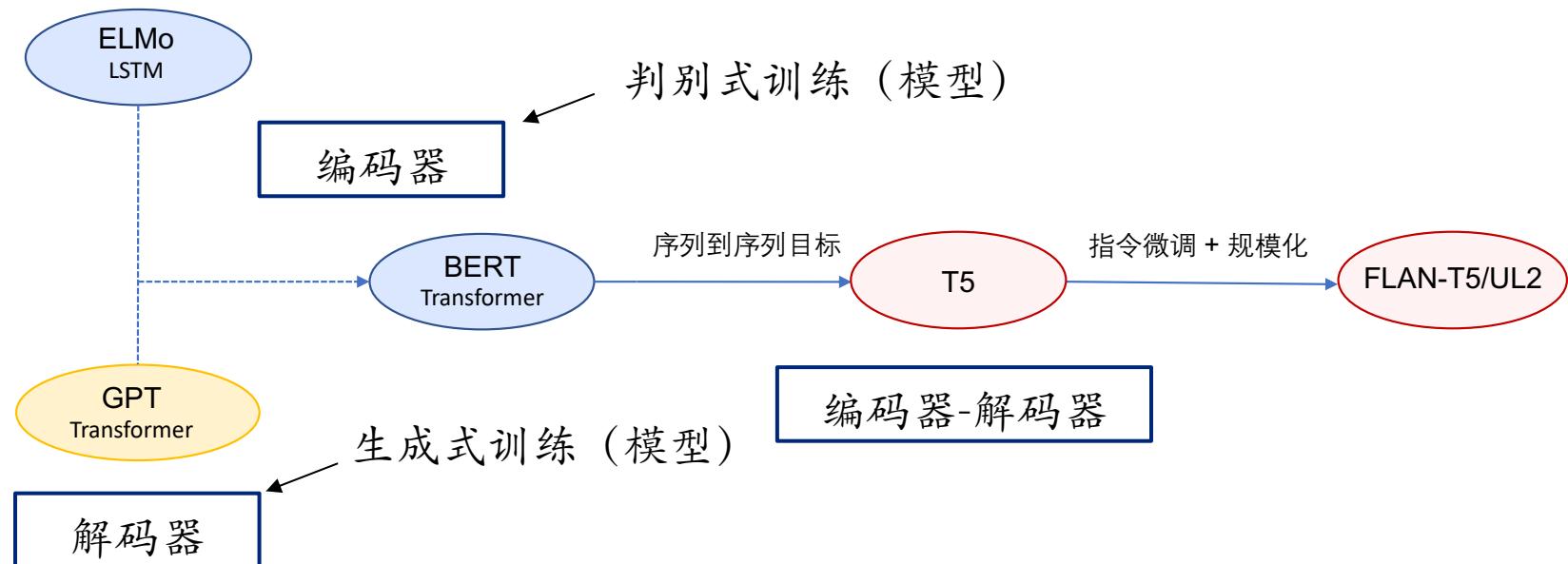
- ▶ 预训练语言模型的发展
 - ▶ 大模型的发展与新范式
 - ▶ 从GPT-3到ChatGPT
- ▶ 涌现能力与幻觉问题
- ▶ 提示学习技术
 - ▶ 上下文学习
 - ▶ 思维链推理
- ▶ 开源家族：LLaMA与其后继者
- ▶ 把大模型变小：模型量化和LoRA微调

目录

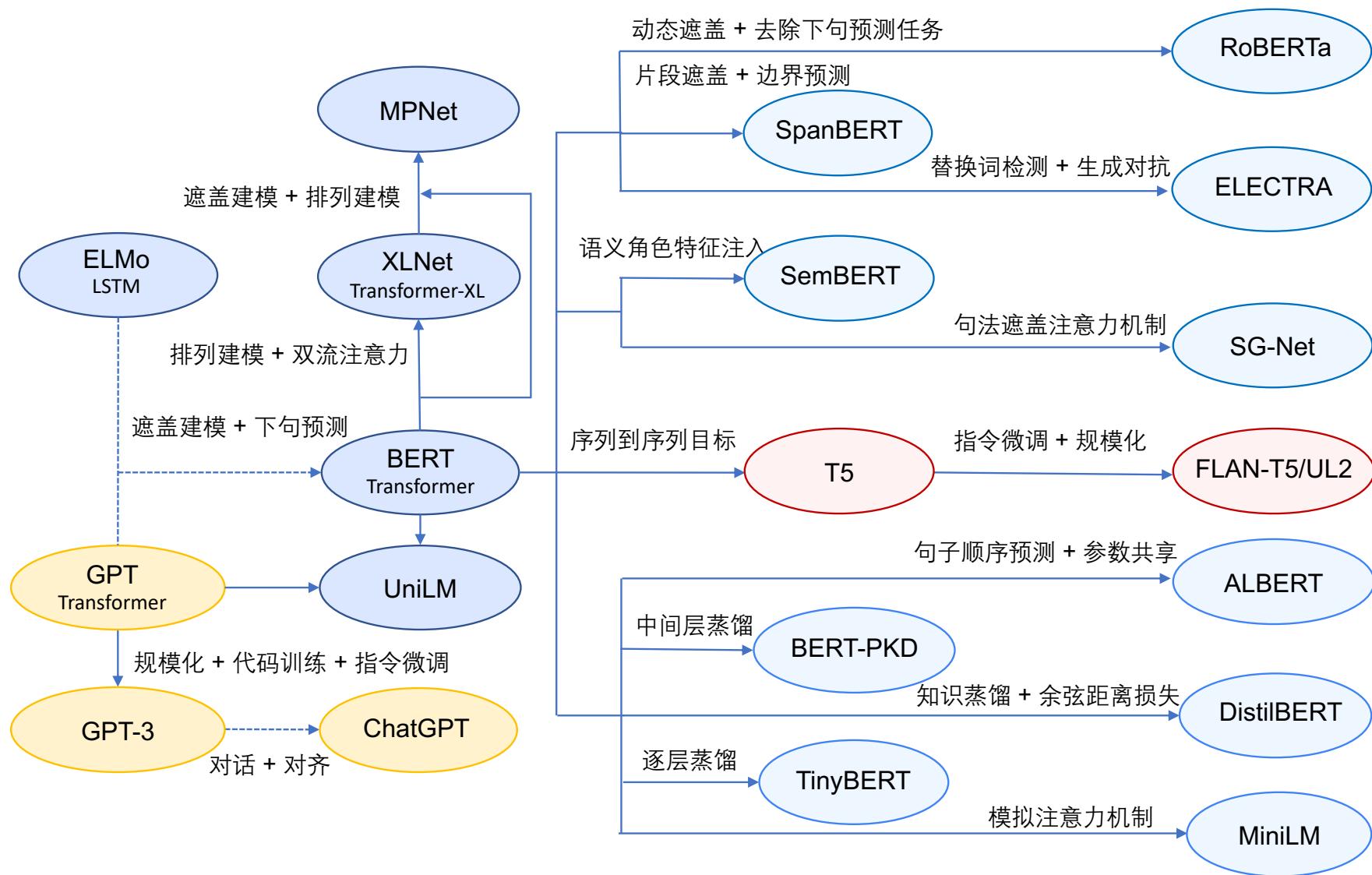
- ▶ 预训练语言模型的发展
 - ▶ 大模型的发展与新范式
 - ▶ 从GPT-3到ChatGPT
- ▶ 涌现能力与幻觉问题
- ▶ 提示学习技术
 - ▶ 上下文学习
 - ▶ 思维链推理
- ▶ 开源家族：LLaMA与其后继者
- ▶ 把大模型变小：模型量化和LoRA微调

预训练语言模型的发展

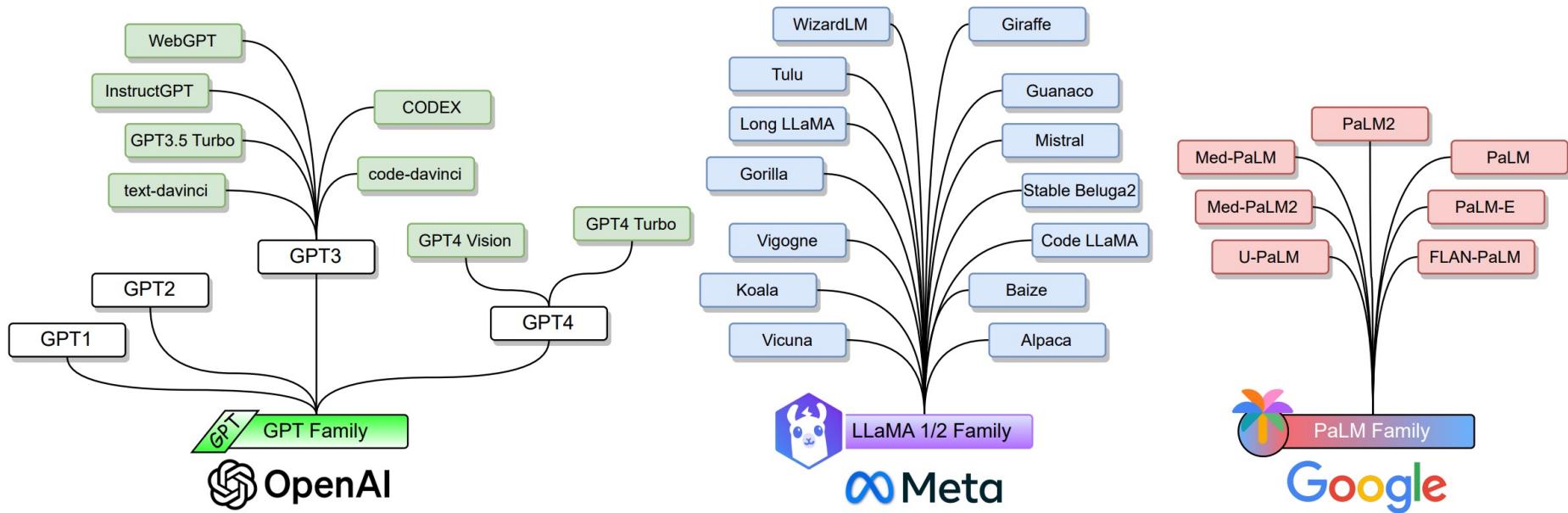
- ▶ 预训练语言模型
 - ▶ 在大规模数据上自监督训练，经微调或提示后适配各类任务
- ▶ 主要技术架构
 - ▶ 编码器：BERT、ALBERT
 - ▶ 解码器：GPT、Llama
 - ▶ 编码器-解码器：T5、BART



预训练语言模型的发展

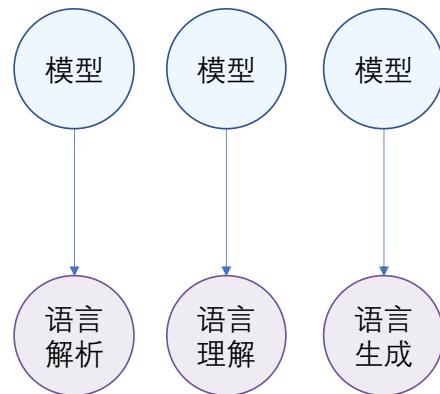


预训练语言模型的发展

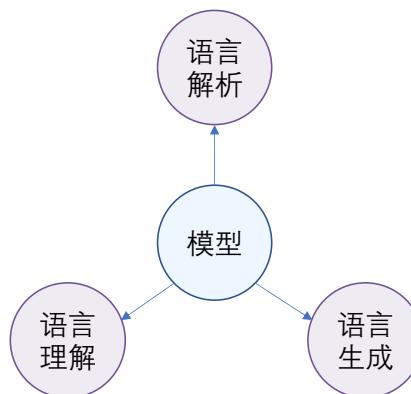


预训练语言模型：新的机器学习范式

- ▶ 使用一个通用模型，解决各类下游任务



为每个任务训练独立的模型



中心节点完成预训练，用户在此基础上面向任务微调

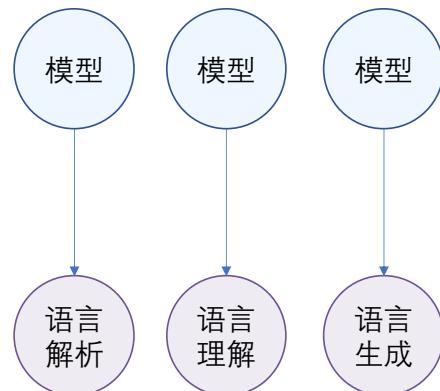
个体化训练



中心化训练 + 个体化微调

预训练语言模型：新的机器学习范式

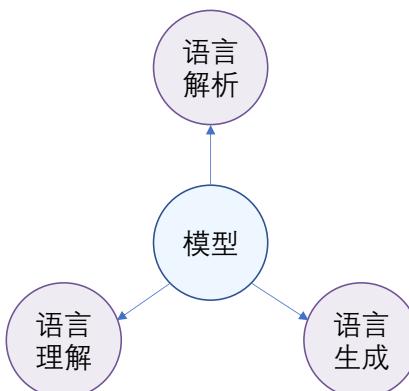
- ▶ 使用一个通用模型，解决各类下游任务



过去

为每个任务训练独立的模型

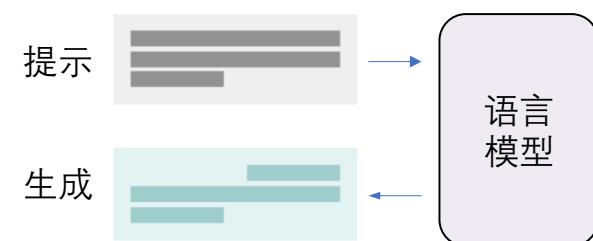
个体化训练



不久之前

中心节点完成预训练，用户在此基础上面向任务微调

中心化训练 + 个体化微调



现在（大规模语言模型）

- ▶ 提示学习
- ▶ 上下文学习
- ▶ 思维链提示
- ▶ 轻量化微调

预训练语言模型质变到大规模生成式语言模型

- ▶ 大规模语言模型：通常指参数量超过 10B 的模型
 - ▶ 更多的计算量、推理开销更大
 - ▶ 泛化性能更强，出现涌现能力

	预训练语言模型 (小模型、常规模型)	大规模生成式语言模型
典型模型	ELMo, BERT, GPT-2	GPT-3、ChatGPT、LLaMA
模型结构	BiLSTM, Transformer	Transformer
注意力机制	双向、单向	单向
训练方式	去噪自编码模型	自回归生成
擅长任务类型	理解、判断	生成
模型规模	1-10亿级参数	10-1000亿级参数
下游任务应用方式	微调	微调 & 提示学习
涌现能力	小数据领域迁移	上下文学习，思维链提示

预训练语言模型质变到大规模生成式语言模型

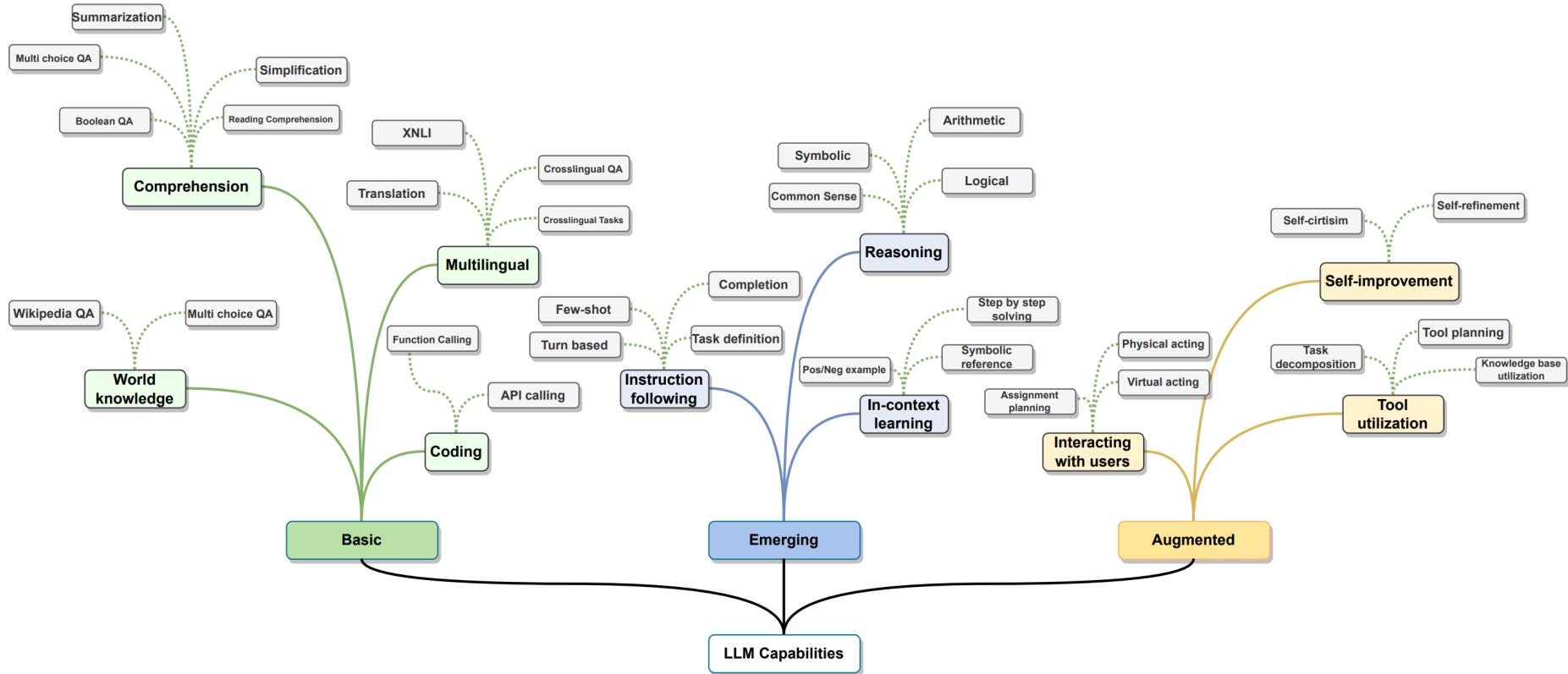
- ▶ 大规模语言模型：通常指参数量超过 10B 的模型
 - ▶ 更多的计算量、推理开销更大
 - ▶ 泛化性能更强，出现涌现能力

模型	公司	参数量	训练硬件	天 GPU	天	数据规模
“小”语言模型						
ELMo ^(Peters et al., 2018)	AllenAI	96M	3×1080 GPU	14	42	4GB
BERT _{large} ^(Devlin et al., 2019)	Google	340M	64×TPU	4	256	16GB
GPT-1 ^(Radford et al., 2018)	OpenAI	117M	8×P6000 GPU	25	200	4.5GB
XLNet ^(Yang et al., 2019)	Google	360M	512×TPU	2.5	1280	160GB
ELECTRA _{large} ^(Clark et al., 2020)	Google	335M	-	-	-	160GB
大语言模型						
BLOOM ^(Scao et al., 2022)	BigScience	176B	384×A100 GPU	118	45K	1.61TB
Chinchilla ^(Hoffmann et al., 2022)	DeepMind	70B	4096×TPU v3	-	-	140T Tokens
ERNIE3.0 ^(Sun et al., 2021)	百度	10B	384×V100 GPU	-	-	4TB
Galactica ^(Taylor et al., 2022)	Meta AI	120B	128×A100 GPU	-	-	106B Tokens
Gopher ^(Rae et al., 2021)	DeepMind	280B	4096×TPU v3	38	156K (300B Tokens)	10.5TB
GPT-3 ^(Brown et al., 2020)	OpenAI	175B	1750×V100 GPU 约	90	158K	45TB
LaMDA ^(Thoppilan et al., 2022)	Google	137B	1024×TPU v3	57.7	59K	1.56T Tokens
M6 ^(Lin et al., 2021)	阿里巴巴	100B	128×A100 GPU	-	-	1.9TB 图像 + 97.2GB 文本
OPT ^(Zhang et al., 2022)	Meta AI	175B	992×A100 GPU 约	60	60K	180B Tokens
PaLM ^(Chowdhery et al., 2022)	Google	540B	6144×TPU v4	50	307K	780B Tokens
T5 ^(Raffel et al., 2020)	Google	11B	1024×TPU v3	25	26K	750GB

从中小预训练模型到大规模预训练模型

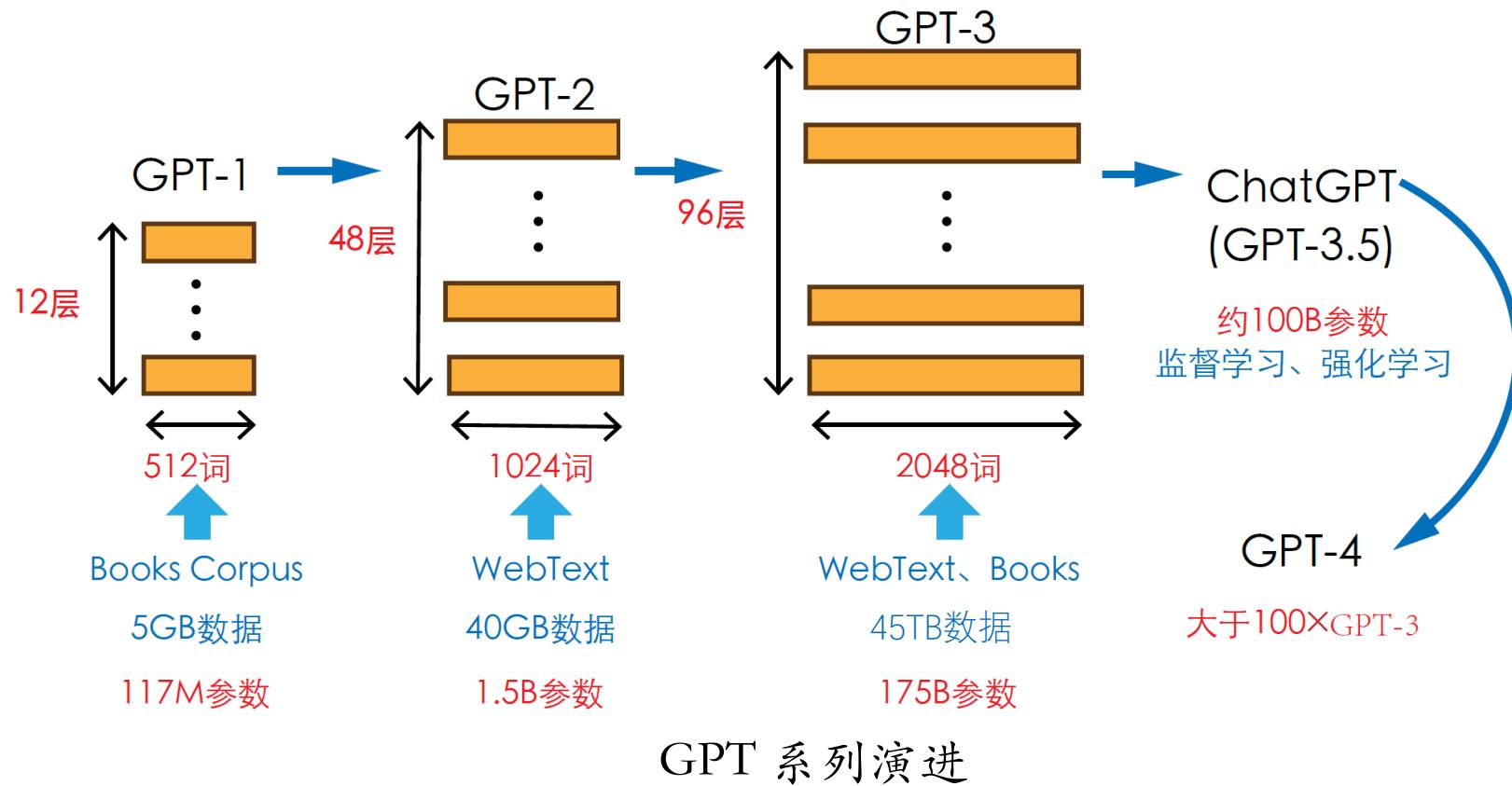
- ▶ 量变到质变：从过拟合（overfitting）风险到欠拟合（underfitting）风险
- ▶ 训练数据的变化：多元化
 - ▶ 不再仅仅是自然语言文本，而是多种数据的组合：自然语言文本、编程代码、化学分子式，乃至基因序列，甚至图像
- ▶ 训练方式的变化：从判别式预训练（BERT为典型）全面转向生成式预训练（GPT为典型）
- ▶ 模型架构的变化：从双向Transformer转向单向Transformer（Decoder-only）
- ▶ 应用方式的变化：从微调走向更为友好的提示学习
 - ▶ 样本更少，从必须一定的标注样本，到少样本，乃至零样本
 - ▶ 提示学习的工作形式逼近人机对话形式

大模型的能力版图



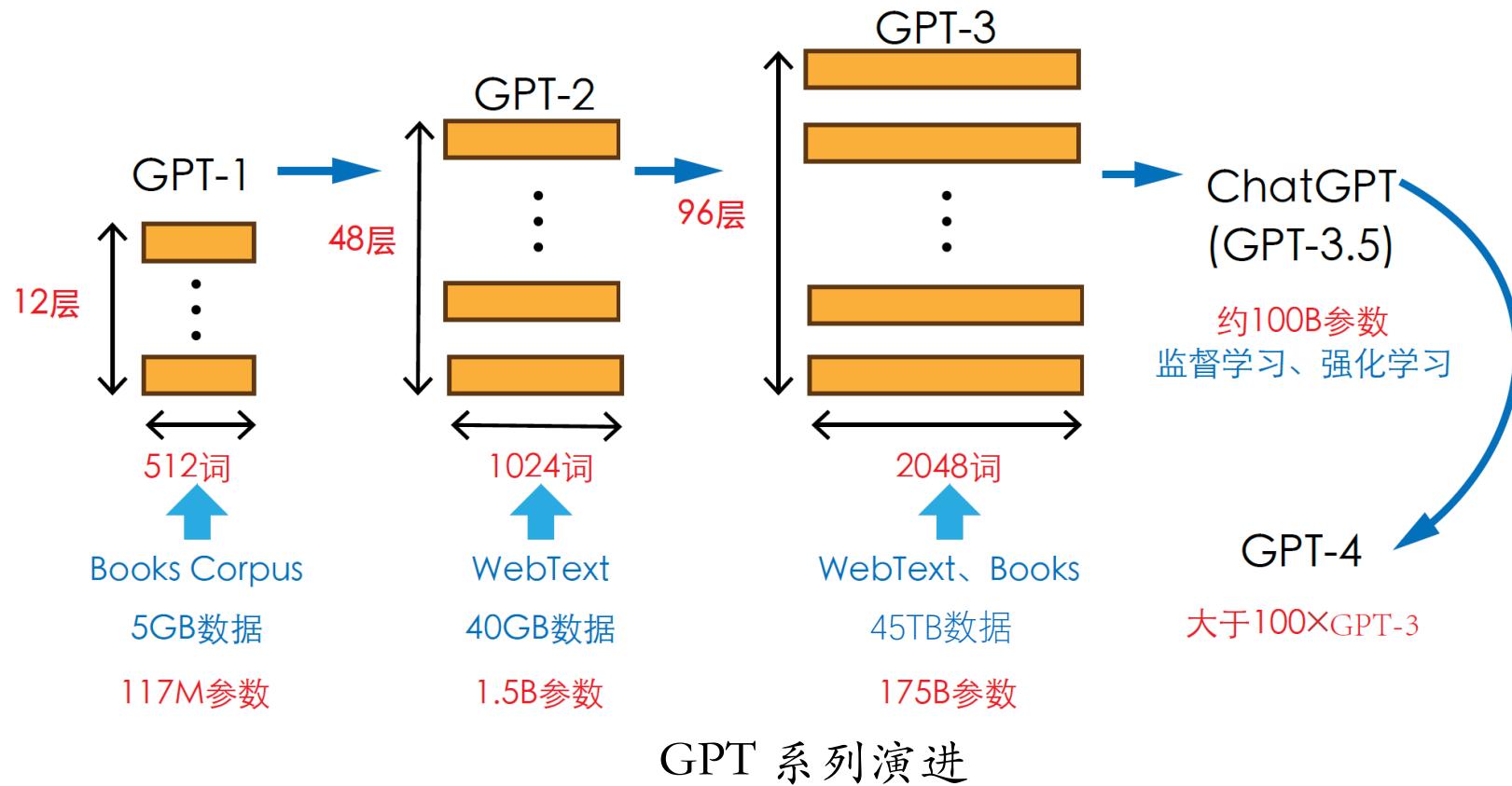
从GPT-3到ChatGPT

- ▶ GPT-1: 单向自回归建模 + 有监督微调
- ▶ GPT-2: 单向自回归建模 + 更多的数据、更大的模型 + 零样本学习
- ▶ GPT-3: 继续扩充数据和模型规模 + 上下文提示学习
- ▶ GPT-3.5/ChatGPT: 指令微调 + 人类反馈 + 对话优化

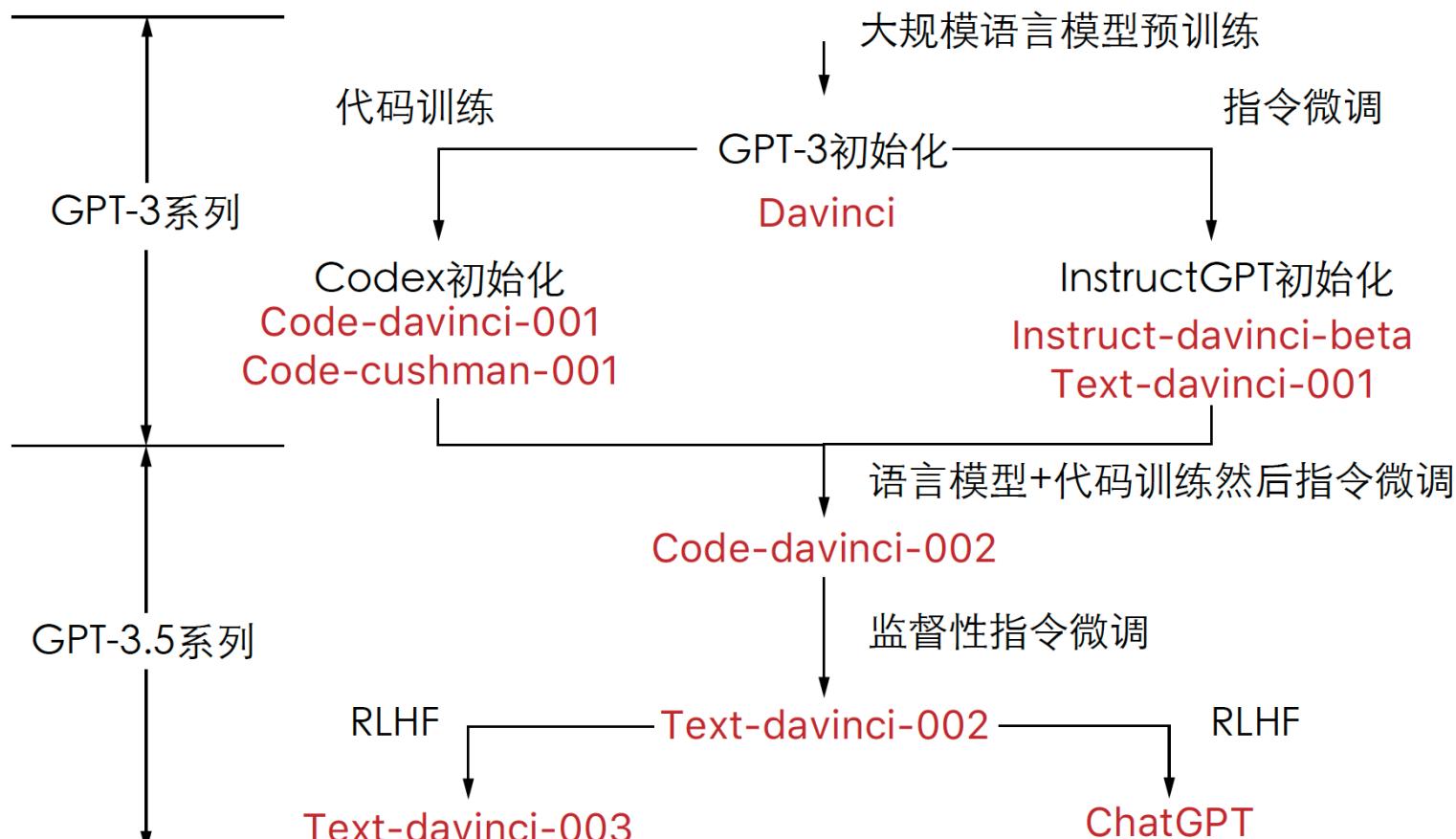


从GPT-3到ChatGPT

- ▶ 模型架构上，GPT-1到GPT-X沿用同样的单向Transformer架构
 - ▶ 仅在层数和输入长度上做了扩张
- ▶ ChatGPT相对于最初的GPT-3，能生成更长、内容更丰富、更贴近人类体验的对话文本，同时有毒性大大降低



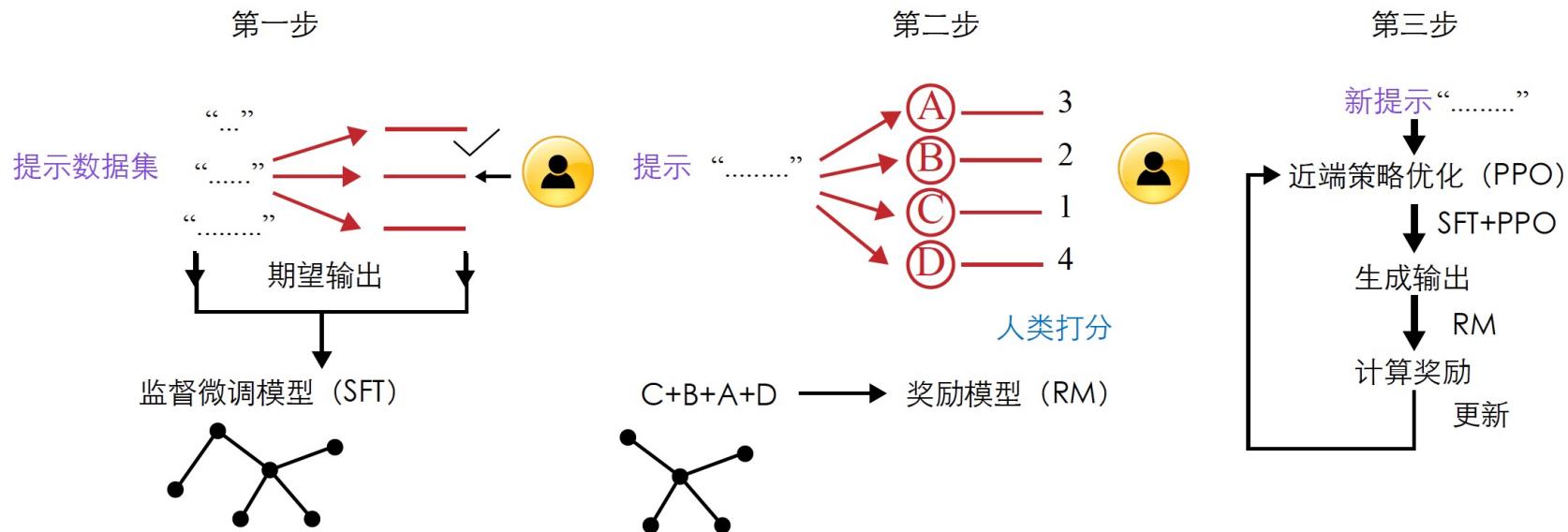
从GPT-3到ChatGPT



GPT-3 和 GPT-3.5 和 ChatGPT 之间的版本关系

ChatGPT关键技术

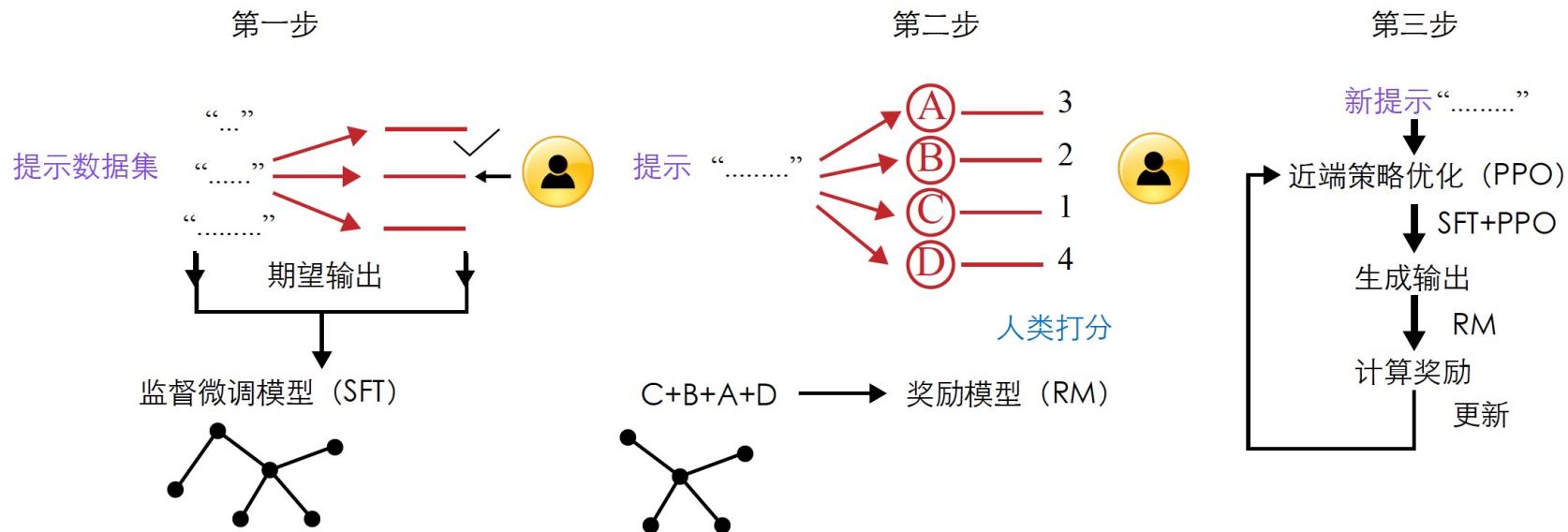
- ▶ ChatGPT：面向人机对话用途进行优化的大语言模型
 - ▶ 代码编写、提供建议、解释定理、文本摘要、机器翻译等通用问题求解能力
- ▶ 关键技术
 - ▶ 对大模型的指令微调
 - ▶ 基于人类反馈的强化学习



ChatGPT关键技术

- ▶ ChatGPT：面向人机对话用途进行优化的大语言模型
 - ▶ 代码编写、提供建议、解释定理、文本摘要、机器翻译等通用问题求解能力
- ▶ 关键技术
 - ▶ 对大模型的指令微调
 - ▶ 基于人类反馈的强化学习

输入提示被构造成了对话的形式，从而具有多轮对话能力



大模型指令微调

- ▶ 通过高质量指令数据（告诉模型执行什么任务）对模型进行微调
 - ▶ 帮助模型理解任务特征，大幅提升在各个任务上的性能表现
 - ▶ 改善提示学习的稳定性，让模型输出文本更为可控

大模型指令微调

- ▶ 通过高质量指令数据（告诉模型执行什么任务）对模型进行微调
 - ▶ 帮助模型理解任务特征，大幅提升在各个任务上的性能表现
 - ▶ 改善提示学习的稳定性，让模型输出文本更为可控
- ▶ 以“自然语言推理”为例，构造指令微调训练数据

Premise

Russian cosmonaut Valery Polyakov set the record for the longest continuous amount of time spent in space, a staggering 438 days, between 1994 and 1995.

Hypothesis

Russians hold the record for the longest stay in space.

Target

Entailment
Not entailment



Options:
- yes
- no

常规微调

Template 1

<premise>

Based on the paragraph above, can we conclude that <hypothesis>?

<options>

Template 3

Read the following and determine if the hypothesis can be inferred from the premise:

Premise: <premise>

Hypothesis: <hypothesis>

<options>

Template 2

<premise>

Can we infer the following?

<hypothesis>

<options>

Template 4, ...

指令微调

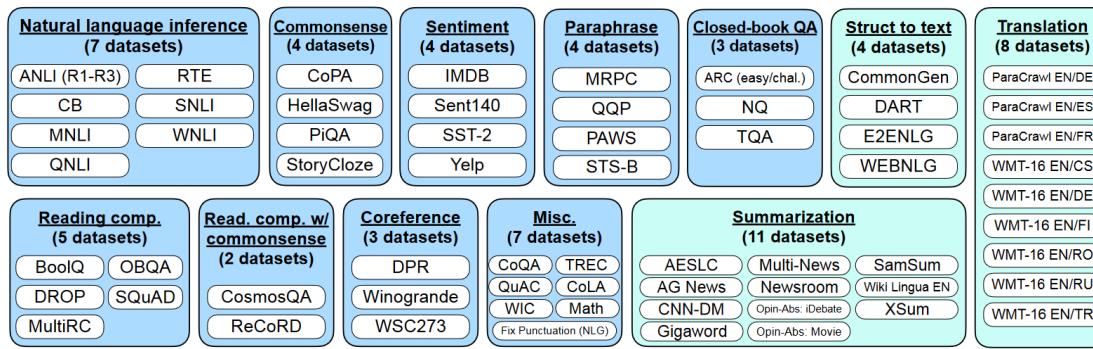
大模型指令微调

▶ 指令微调方式

- ▶ 借助现有的数据集：通过人为添加指定当前任务类型的提示作为输入的前缀（指令），在多类型数据集上进行微调
- ▶ 基于人类演示的有监督微调：基于人类根据提示（指令）撰写的高质量回答，模型据此来进行有监督微调

▶ 数据构造要点

- ▶ 任务数量
- ▶ 任务多样性



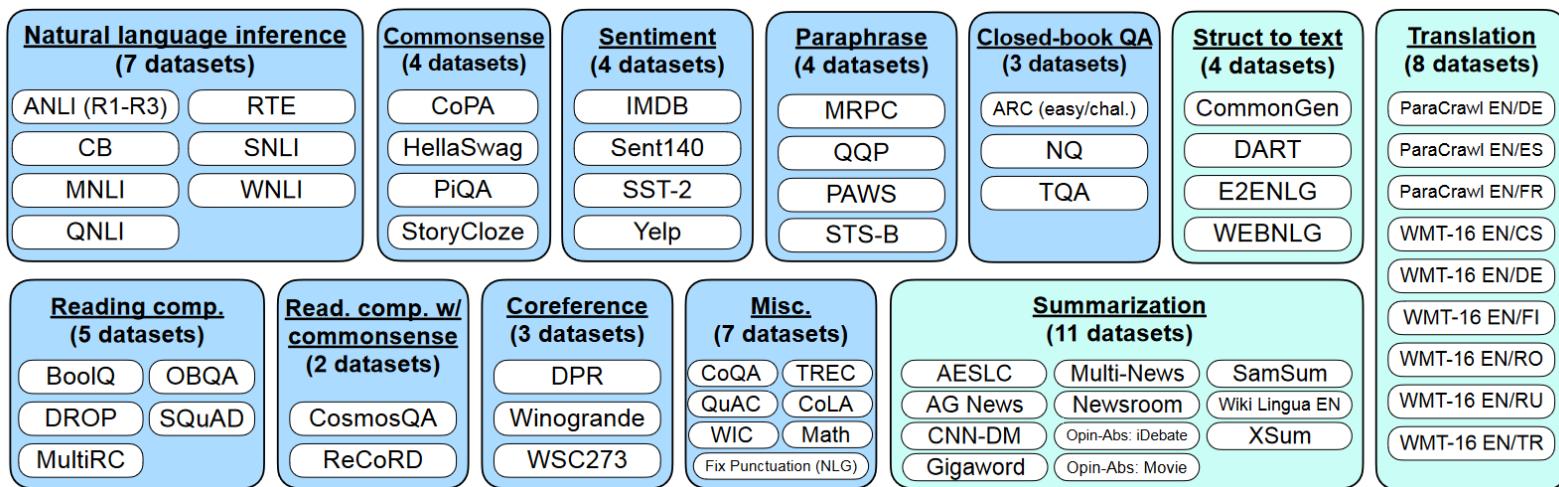
借助现有的数据集 (FLAN)

任务类型	提示示例	占比
生成	Write a short story where a brown bear to the beach, makes friends with a seal, and then return home.	45.6%
开放式问答	Who built the statue of liberty?	12.4%
封闭式问答	Answer the following question: What shape is the earth? A) A circle B) A sphere C) An ellipse D) A plane	2.6%
头脑风暴	List five ideas for how to regain enthusiasm for my career	11.2%
对话	This is a conversation with an enlightened Buddha. Every response is full of wisdom and love. Me: How can I achieve greater peace and equanimity? Buddha:	8.4%
文本重写	Translate this sentence to Spanish: <English sentence>	6.6%
文本摘要	Summarize this for a second-grade student: {text}	4.2%
文本分类	{java code} What language is the code above written in?	3.5%
信息提取	Extract all place names from the article below: {news article}	1.9%
其他	Look up "cowboy" on Google and give me the results.	3.5%

基于人类演示 (InstructGPT)

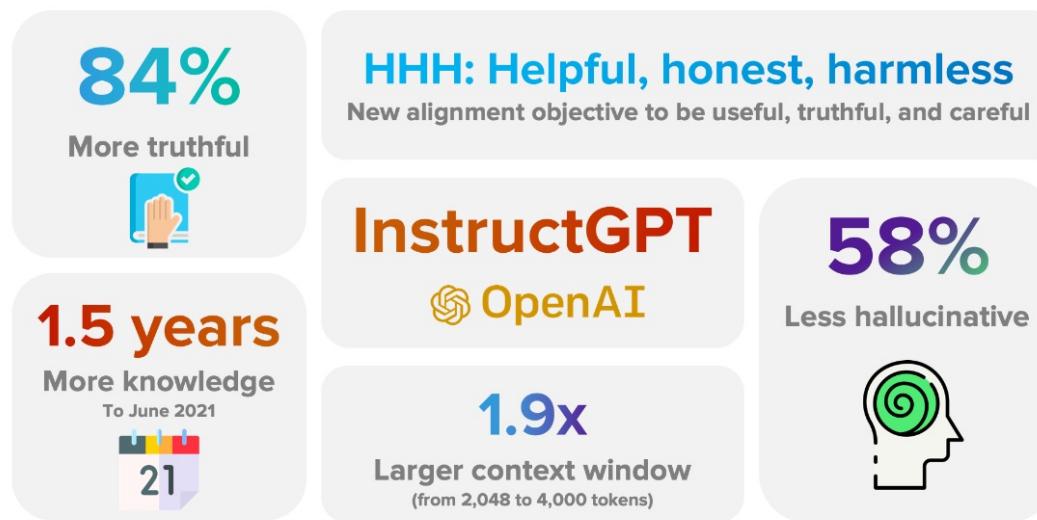
大模型指令微调

- ▶ FLAN指令微调：使用指令微调来提升模型的zero-shot能力
 - ▶ 将62个自然语言处理任务分为12类
 - ▶ 对于每个任务，将为其手动构建10个独特的指令模板
- ▶ 指令微调带来的效果提升存在明显的大模型效应
 - ▶ 只有当模型规模在百亿左右，指令微调才会在样本外任务上带来提升



基于人类反馈的强化学习 (RLHF)

- ▶ 大规模语言模型常生成无用、有毒、不真实的答案
- ▶ 借助较少人力成本，利用强化学习方法进一步微调大语言模型
 - ▶ 让大语言模型的输出“对齐”人类的意图，输出受人喜爱的答案
 - ▶ 语言模型的输出越符合人类意图，越符合道德标准，则模型所获得的奖励越多
- ▶ 大幅减少有害和失真信息的输出



基于人类反馈的强化学习 (RLHF)

- ▶ 与人类价值观、意图、安全等方面“对齐”
 - ▶ 定义：引导人工智能系统的行为，使其符合设计者的利益和预期目标
- ▶ 国家颁布《互联网信息服务深度合成管理规定》、《生成式人工智能服务管理办法》

欧洲刑警组织警告：ChatGPT正在被犯罪分子利用

原创 网域动态编辑部 公安部第三研究所网研基地 2023-04-14 14:50
发表于上海

收录于合集 30个

ChatGPT已被用于犯罪：
网络攻击、犯罪指导等



3月27日，欧洲刑警组织在一份报告中说，犯罪分子已经在利用ChatGPT进行犯罪。该报告详细介绍了人工智能语言模型如何助长欺诈、网络犯罪和恐怖主义。由OpenAI开发的ChatGPT于2022年11月发布，并迅速成为互联网上的一个热点。网民们蜂拥而至，让聊天机器人生生成论文、笑话、电子邮件、编程代码和其他各种文本。

中华人民共和国中央人民政府
www.gov.cn

首页 | 繁体 | 英文EN | 登录 | 邮箱

收藏 留言 +

标 题： 互联网信息服务深度合成管理规定
发文字号： 国家互联网信息办公室 中华人民共和国工业和信息化部 中华人民共和国公安部令 第12号 来 源： 网信办网站
发文机关： 网信办 工业和信息化部 公安部

不得利用深度合成服务制作、复制、发布、传播虚假信息、不得利用深度合成服务制作、复制、发布、传播法律、行政法规禁止的信息

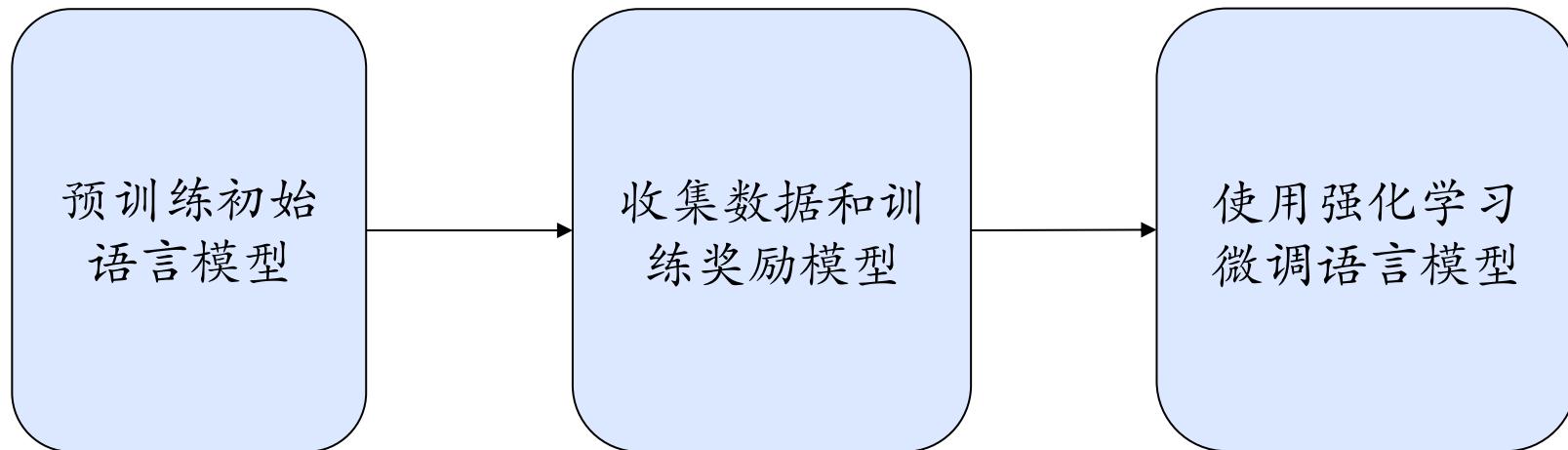
国家互联网信息办公室
中华人民共和国工业和信息化部
中华人民共和国公安部
令
第12号

《互联网信息服务深度合成管理规定》已经2022年11月3日国家互联网信息办公室2022年第21次室务会议审议通过，并经工业和信息化部、公安部同意，现予公布，自2023年1月10日起施行。

国家互联网信息办公室主任 庄荣文
工业和信息化部部长 金壮龙
公安部部长 王小洪
2022年11月25日

基于人类反馈的强化学习 (RLHF)

- ▶ 原则：有用、诚实和无害
 - ▶ 符合国家区域文化、符合法律法规、符合道德伦理、讲事实逻辑
- ▶ 技术要点：
 - ▶ 数据集: SFT数据集 (13k) 、 RM数据集 (33k) , RL数据集 (31k)
 - ▶ 模型: SFT监督微调模型、 RM奖励模型、 RL强化学习模型



基于人类反馈的强化学习 (RLHF)

Step 1

Collect demonstration data,
and train a supervised policy.

A prompt is
sampled from our
prompt dataset.

Explain the moon
landing to a 6 year old

A labeler
demonstrates the
desired output
behavior.

Some people went
to the moon...

This data is used
to fine-tune GPT-3
with supervised
learning.

SFT
围绕地球旋转的
球形天体

第一步：Supervised fine-tuning (SFT)

采样、API收集的数据

175B GPT-3

标注的目标答案

基于人类反馈的强化学习 (RLHF)

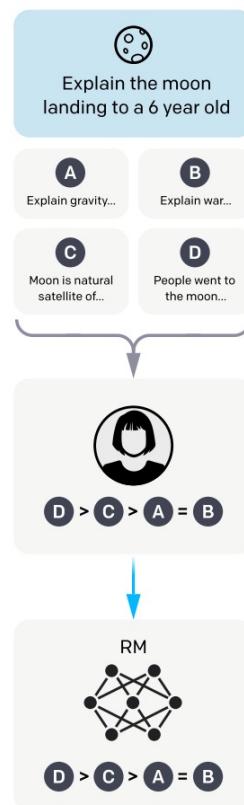
Step 2

Collect comparison data,
and train a reward model.

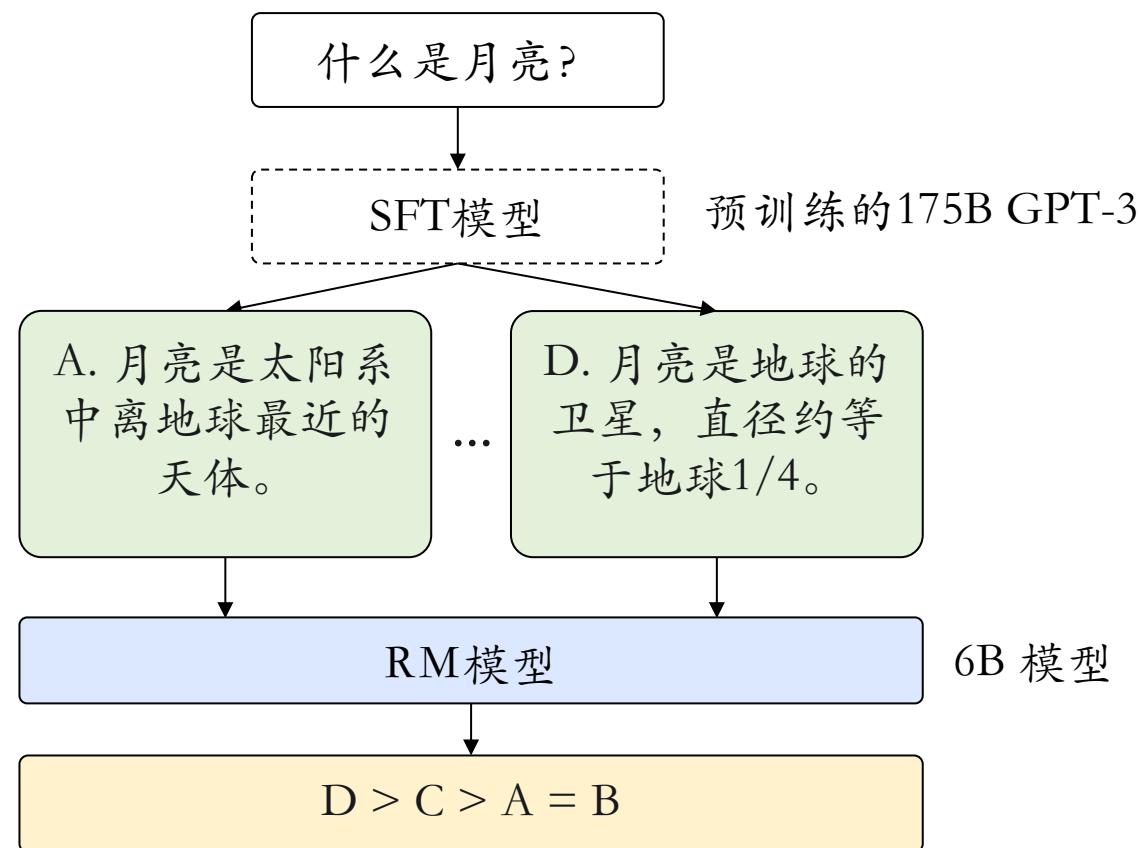
A prompt and
several model
outputs are
sampled.

A labeler ranks
the outputs from
best to worst.

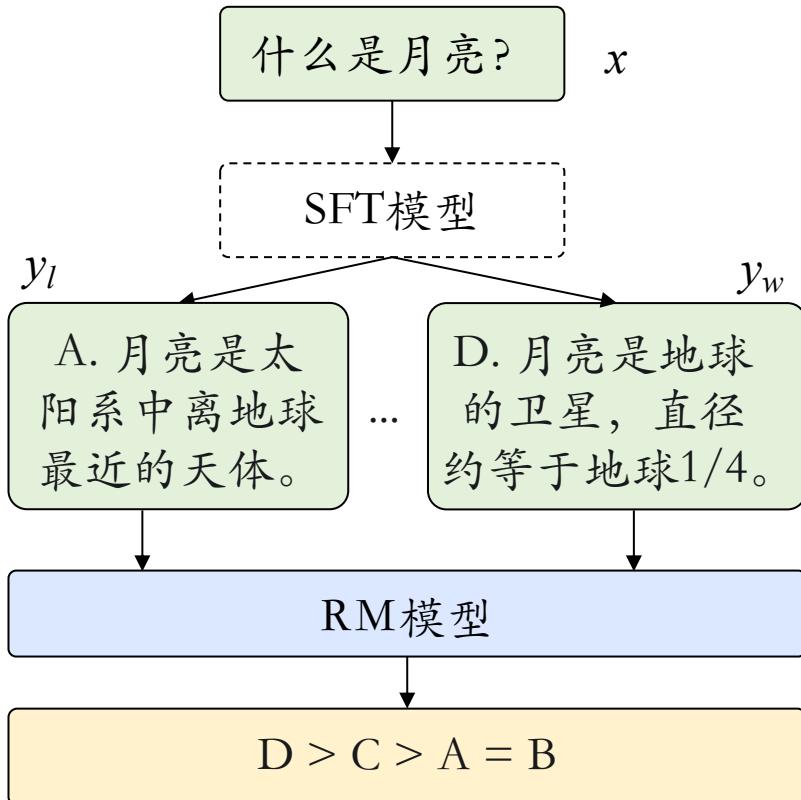
This data is used
to train our
reward model.



第二步：Reward modeling (RM)



基于人类反馈的强化学习 (RLHF)



第二步：Reward modeling (RM)

6B 模型（便宜+稳定）

排序模型（问题 + 答案 → 输出分数）

数据集中问题 x 对应的两个答案，且 y_w 排序比 y_l 高

$$\text{loss}(\theta) = -\frac{1}{\binom{K}{2}} E_{(x, y_w, y_l) \sim D} [\log (\sigma (r_\theta(x, y_w) - r_\theta(x, y_l)))]$$

↓
数据集中的问题 RM 模型打分差值

基于人类反馈的强化学习 (RLHF)

Step 3

Optimize a policy against
the reward model using
reinforcement learning.

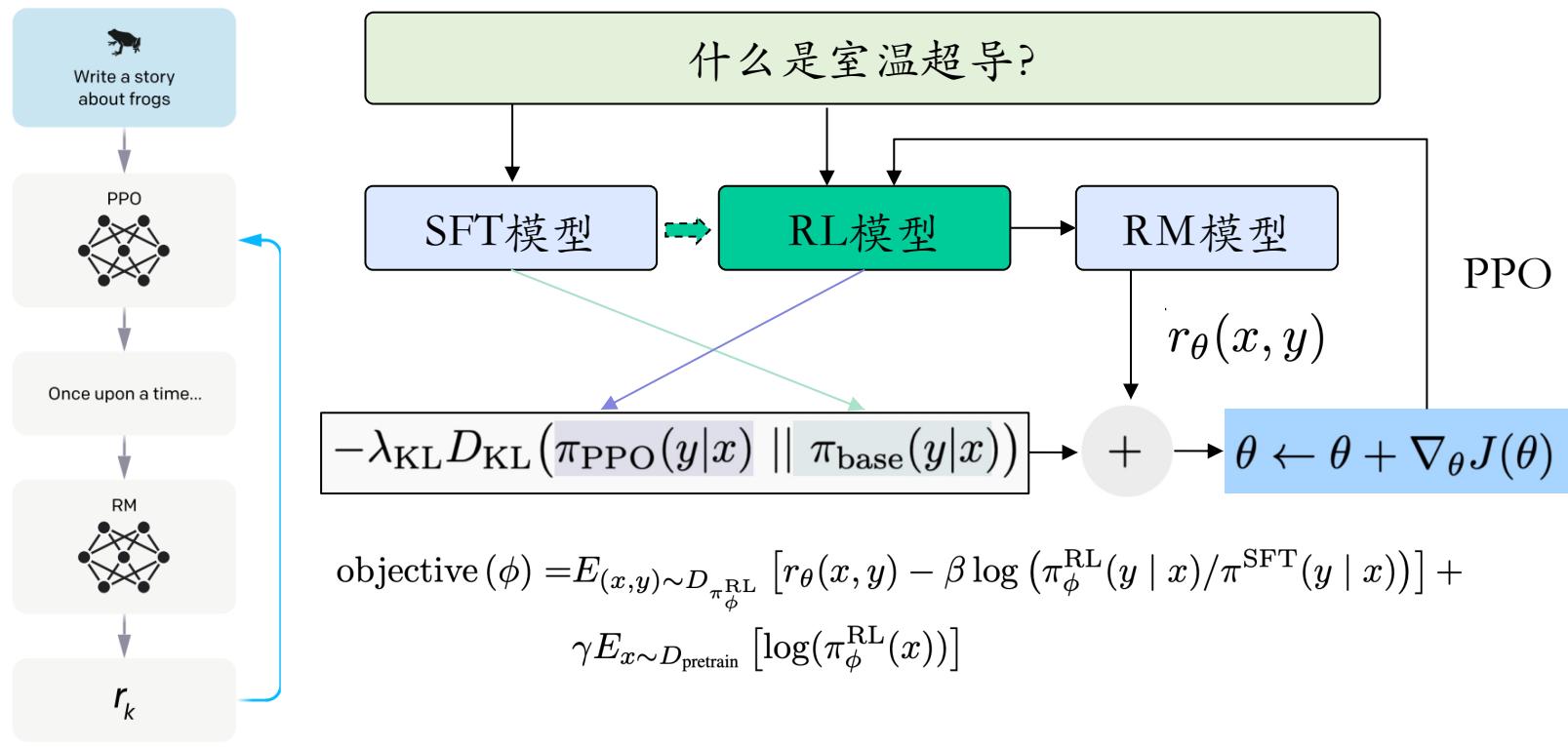
A new prompt
is sampled from
the dataset.

The policy
generates
an output.

The reward model
calculates a
reward for
the output.

The reward is
used to update
the policy
using PPO.

第三步：Reinforcement learning (RL)



注：PPO，Proximal Policy Optimization，近端策略优化

Ouyang, Long, et al. "Training language models to follow instructions with human feedback." *Advances in Neural Information Processing Systems* 35 (2022): 27730-27744.

基于人类反馈的强化学习 (RLHF)

Step 3

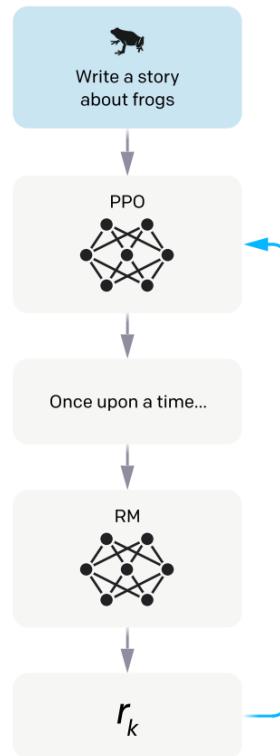
Optimize a policy against
the reward model using
reinforcement learning.

A new prompt
is sampled from
the dataset.

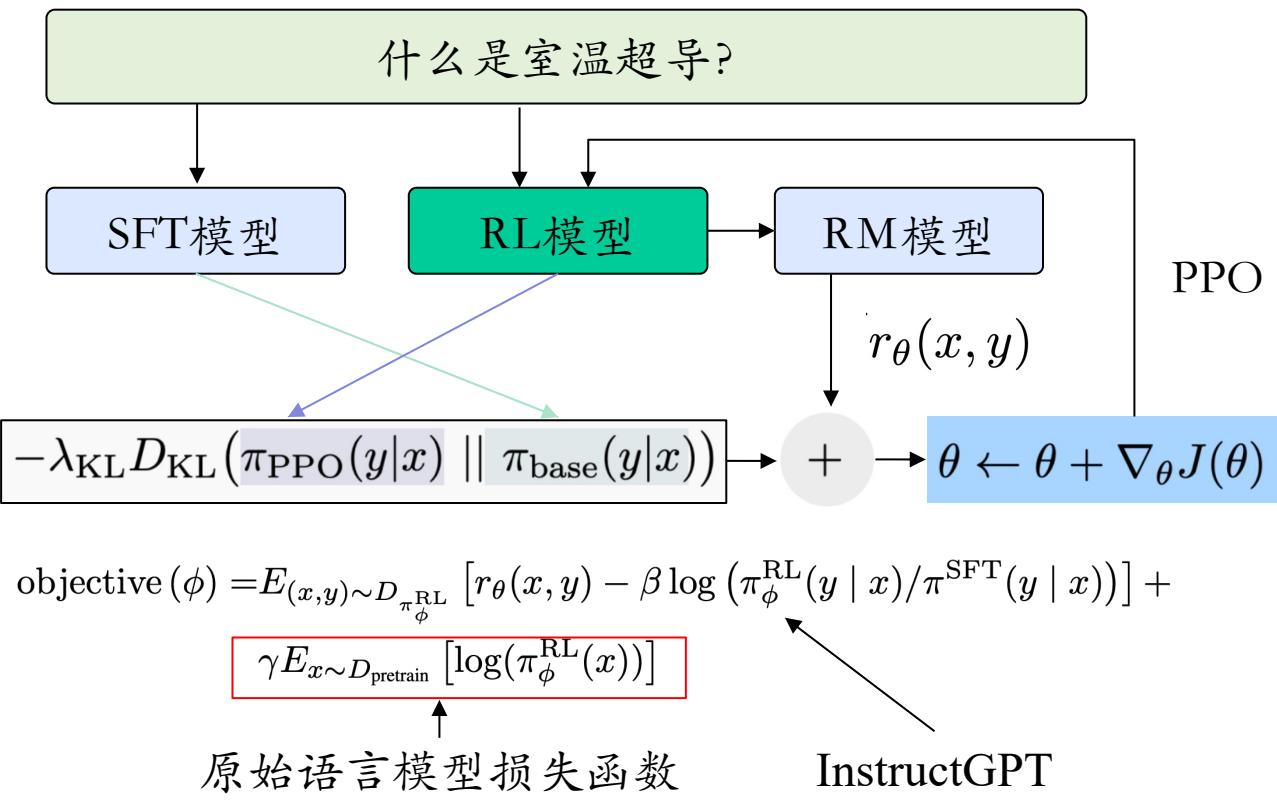
The policy
generates
an output.

The reward model
calculates a
reward for
the output.

The reward is
used to update
the policy
using PPO.



第三步：Reinforcement learning (RL)



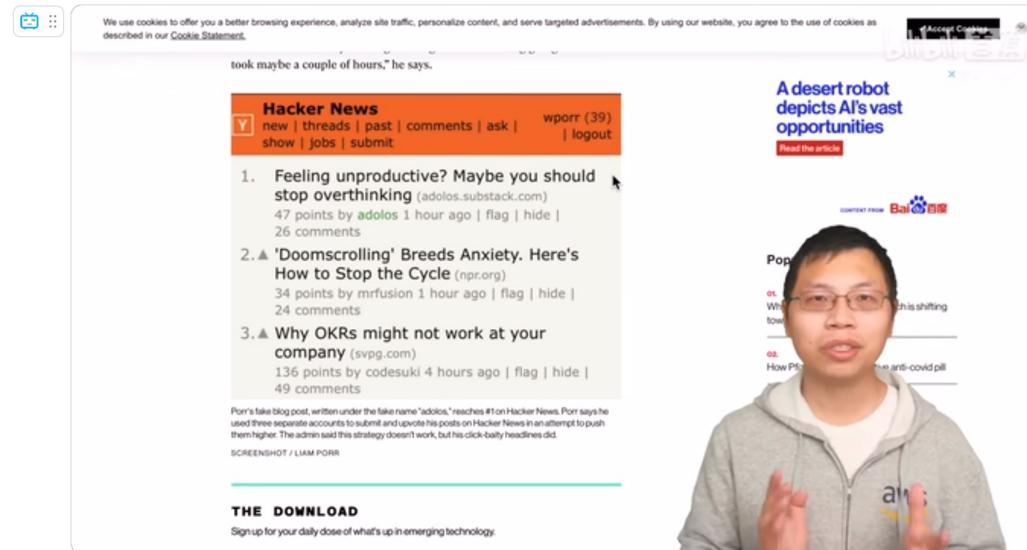
注：PPO，Proximal Policy Optimization，近端策略优化

Ouyang, Long, et al. "Training language models to follow instructions with human feedback." *Advances in Neural Information Processing Systems* 35 (2022): 27730-27744.

基于人类反馈的强化学习 (RLHF)

6. 延伸阅读:

a. 【GPT, GPT-2, GPT-3 论文精读【论文精读】】

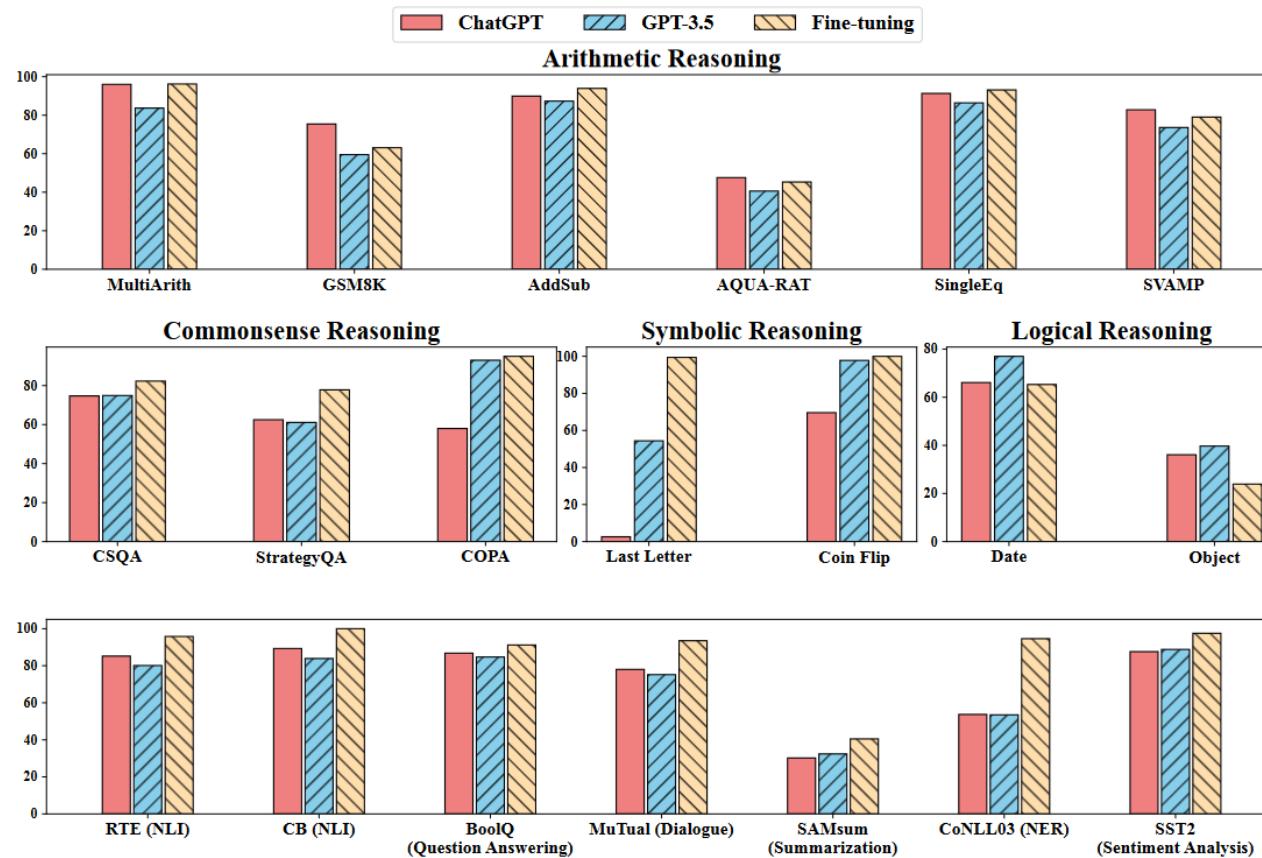


b. 【InstructGPT 论文精读【论文精读·48】】



ChatGPT性能评测

- ▶ 在大部分性能上超越GPT-3
- ▶ 仅通过零样本提示，在多项任务上超越常规全数据集微调的模型

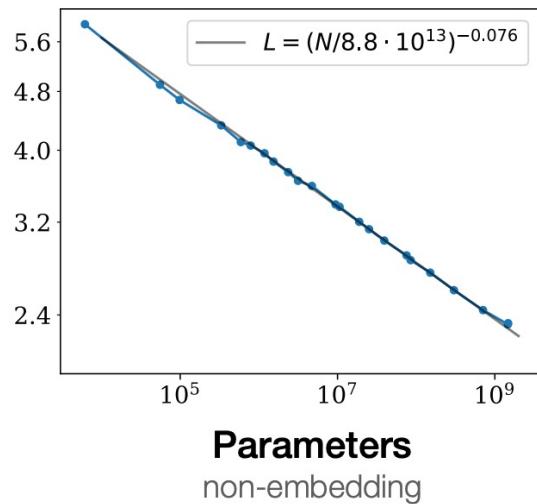
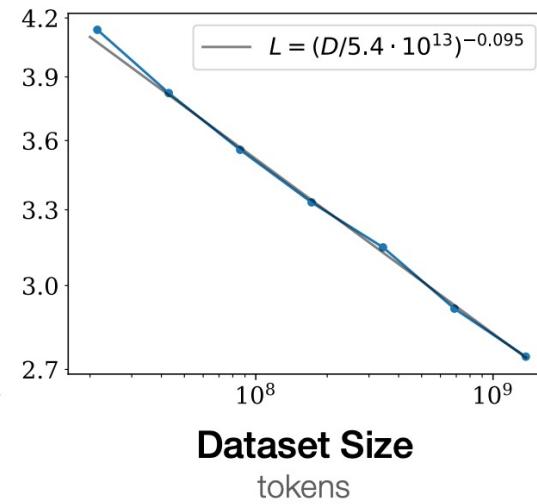
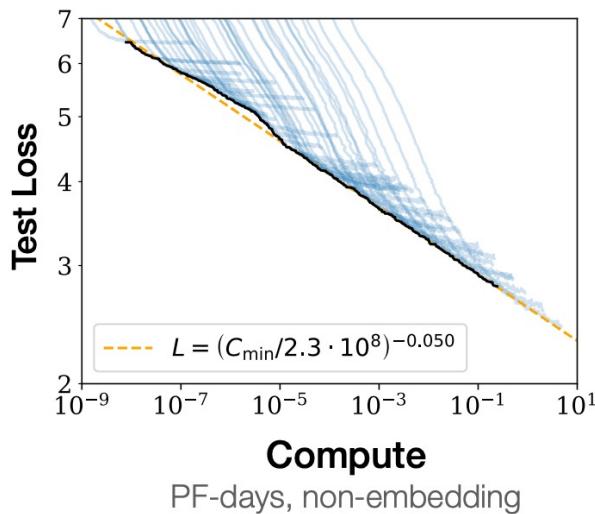


目录

- ▶ 预训练语言模型的发展
 - ▶ 大模型的发展与新范式
 - ▶ 从GPT-3到ChatGPT
- ▶ 涌现能力与幻觉问题
- ▶ 提示学习技术
 - ▶ 上下文学习
 - ▶ 思维链推理
- ▶ 开源家族：LLaMA与其后继者
- ▶ 把大模型变小：模型量化和LoRA微调

大规模语言模型的缩放定律 (Scaling Law)

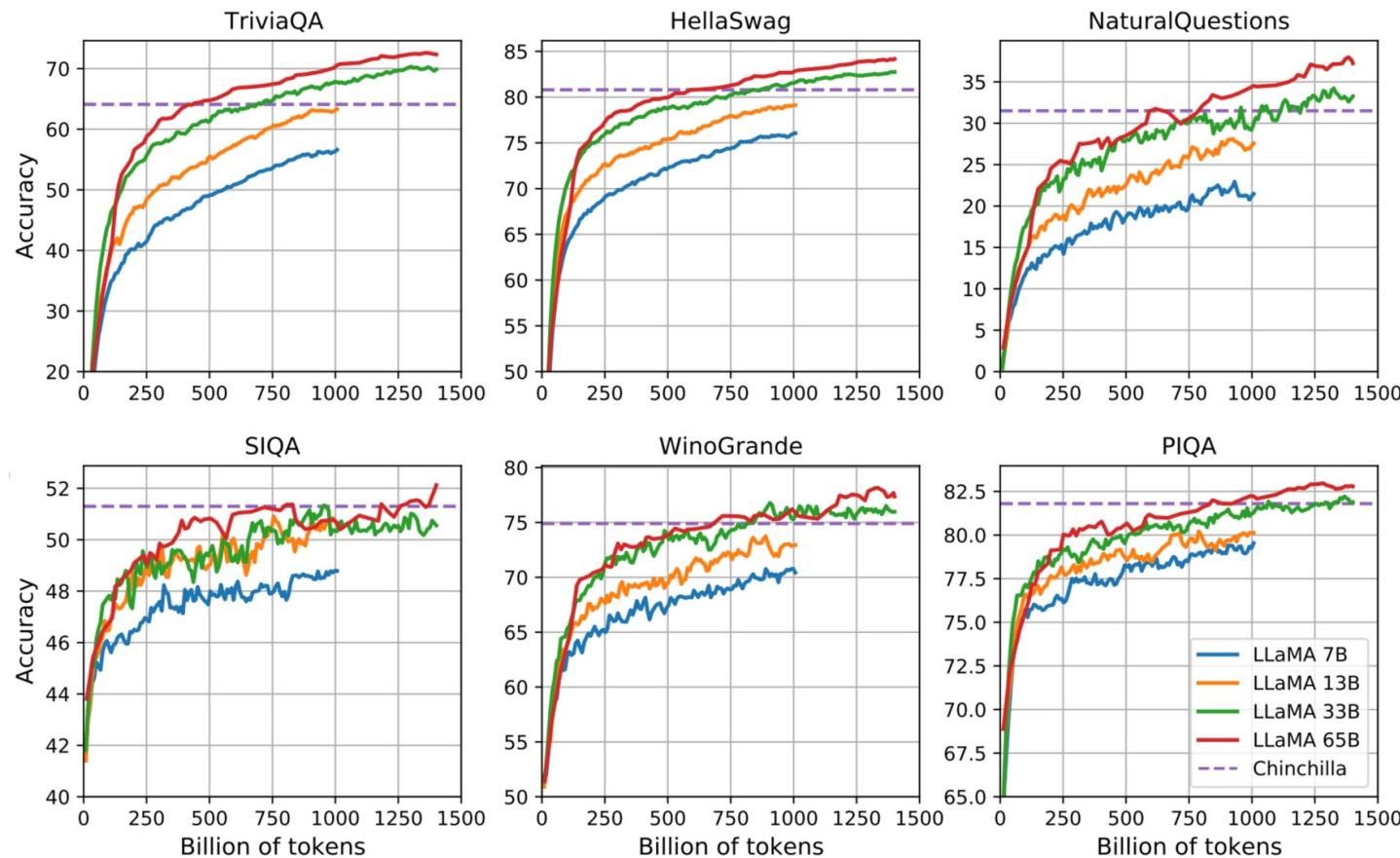
- ▶ 模型性能依赖于模型大小、数据规模和算力
 - ▶ 模型大小、数据规模、算力需维持适当比例，否则会训练失衡
 - ▶ 对于固定算力训练（单位FLOPs），每一个参数需要约10个tokens训练
 - ▶ Chinchilla推荐：10B模型用200B tokens (1:20)



大规模语言模型的缩放定律 (Scaling Law)

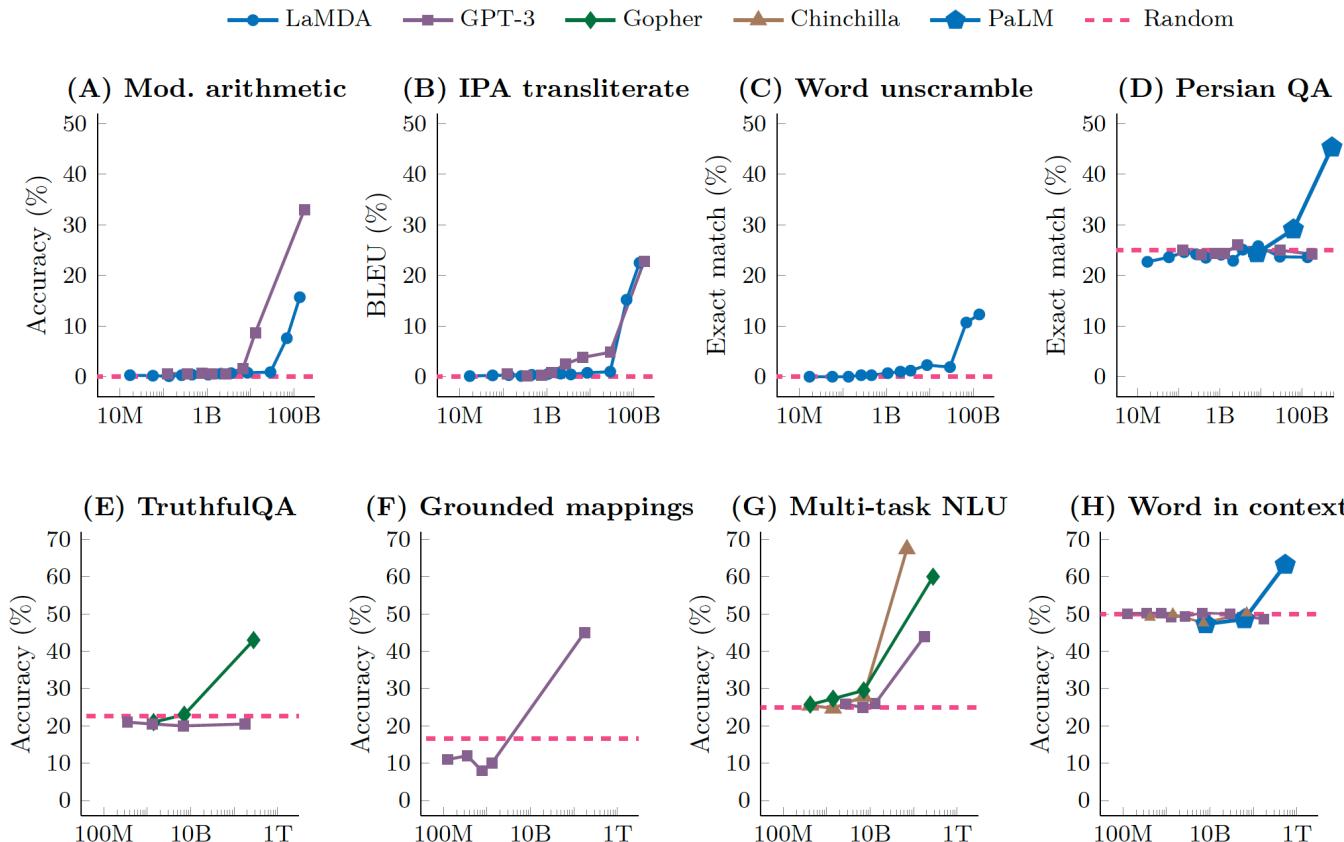
► LLaMA经验

- 在问答和常识推理任务上，模型性能会随着训练tokens数量增加而稳定增长



大规模语言模型的涌现能力

- ▶ 语言模型参数达到一定规模时，某些能力（语言推理）会突然出现
- ▶ 示例：5个大模型在8个任务上验证



大规模预训练模型的幻觉问题

▶ 幻觉问题

- ▶ 定义：模型生成的文本不遵循原文（Faithfulness）或不符合事实（Factualness）

▶ 危害性

- ▶ 影响模型可信度，使得难以落地
- ▶ 容易被攻击、被滥用，如恶意生成和散播谣言，生成钓鱼邮件
- ▶ 污染互联网，涌入大量垃圾、错误信息

问：鲁迅和周树人是同一个人吗？

答：不是。鲁迅和周树人是两个不同的人。鲁迅是中国今现代文学巨匠，以其独特的文学风格和引人入胜的故事而闻名。周树人是中国当代著名作家，以其犀利的文字和敏锐的观察而受到推崇。

问：人类历史上第一次和三体人面对面接触是什么时候？

答：2015年8月15日，中国国家主席在北京亲切接见了刚抵达中国的三体外星人代表团，在历史上这是人类和三体外星人面对面接触的第一次。

大规模预训练模型的幻觉问题

▶ 幻觉的来源

- ▶ 指令描述和约束不够，缺乏足够的输入，模型仅能根据自己的知识进行猜测
- ▶ 未见过、不认识的知识，模型通过猜测和瞎编回答
- ▶ 模型接收了误导信息或基于错误的推测进行作答

▶ 其他局限性

- ▶ 提示学习的不稳定性，在相近意思的不同提示下，可能给出意思相反的回答
- ▶ 容易受到恶意提示攻击的影响，导致模型输出有害的文本

关键：如何对模型的知识进行更新？

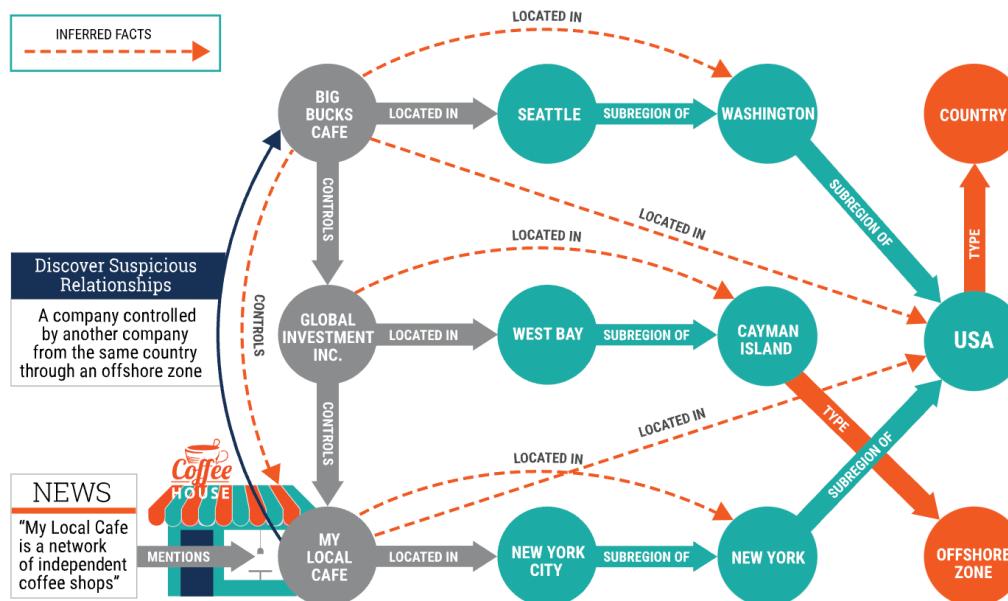
大规模预训练模型知识更新

- ▶ 知识来源
- ▶ 更新手段

大规模预训练模型知识更新来源

▶ 知识来源

- ▶ 知识库/知识图谱
- ▶ 非结构化知识库
- ▶ 人类监督与交互



大规模预训练模型知识更新来源

▶ 知识来源

- ▶ 知识库/知识图谱
- ▶ 非结构化知识库
- ▶ 人类监督与交互

维基百科，自由的百科全书

条目 讨论 大陆简体 编辑 查看历史 搜索维基百科 [关闭]

没有登录 讨论 贡献 创建账号 登录

中文维基百科Facebook粉丝页正式上线，邀请大家一同关注。

室温超导体 [编辑]

维基百科，自由的百科全书

此条目需要精通或熟悉相关主题的编者参与及协助编辑。 (2017年10月5日)
请邀请适合的人士改善本条目。更多的细节与详情请参见讨论页。

室温超导体又称常温超导体 (Room-temperature superconductor)，是指可以在高于0°C的温度有超导现象的材料。相较于其他的超导体，室温超导体的条件是日常较容易达到的工作条件。截至2020年，最高温的超导体是超高压的含碳硫化氢系统，压力267 GPa，其临界温度为+15°C^[1]。在常压下的最高温超导体是高温超导体铜氧化物 (cuprates)，在138 K (~135 °C) 的温度下有超导现象^[2]。之往有许多的研究者曾怀疑室温超导体是否可能实现^{[3][4]}，不过超导的温度一再提高，其中也有许多是以往没有预期到，或是以往认为不可能的温度。早在1950年代就有人提出在“接近室温”下出现的超导现象。若可以找到室温超导体，可解决全世界能耗问题、开发速度更快的电脑、用在先进的储存装置、超灵敏的感测器，以及许多其他的可能性。^{[4][5]}2023年7月23日，来自韩国科学技术研究院 (KIST) 的量子能源研究中心的韩国团队在arXiv预印本服务器上发布了一篇名为“The First Room-Temperature Ambient-Pressure Superconductor”（首个室温常压超导体）的论文，描述了他们称之为LK-99的新型室温超导体。^[6]该论文还伴随着arXiv上的姊妹论文，^[7]一篇韩国期刊上的论文^[8]以及一项专利申请。^[9]多位专家对此表示怀疑，牛津材料科学教授Susannah Speller表示，“目前还为时过早，我们还没有得到这些样本超导性的有力证据”，因为缺乏超导性的明确标志，如磁场响应和热容量。其他专家也对数据可能被“实验过程中的错误和LK-99样本的缺陷”解释表示担忧，一位科学家质疑了研究人员使用的理论模型。^[10]

参考资料 [编辑]

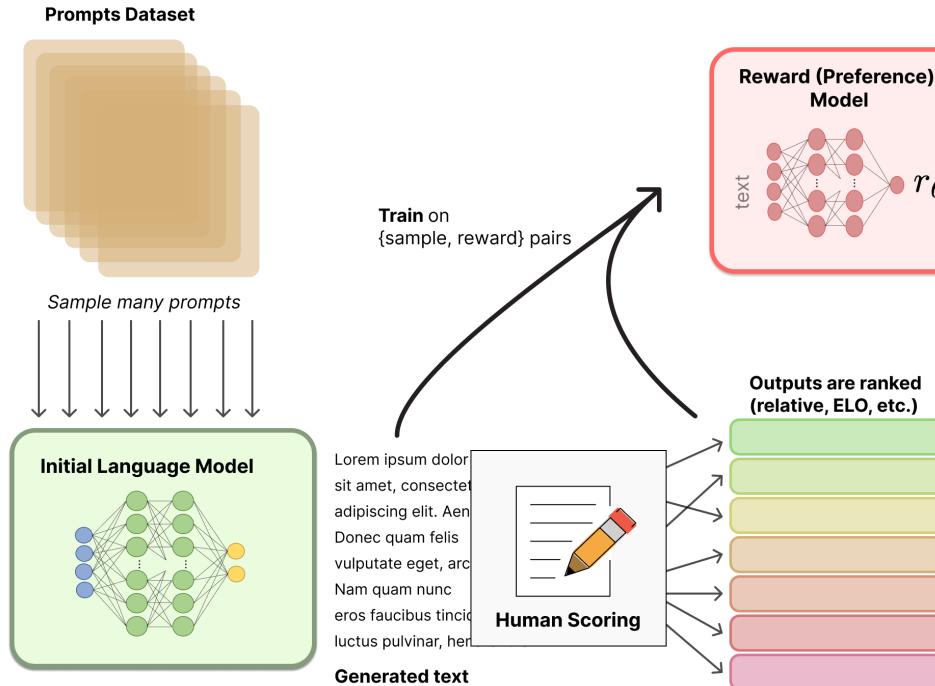
未解决的物理问题：是否有在室温及常压下的超导体？

- ^{1.} Snider, Elliot; Dasenbrock-Gammon, Nathan; McBride, Raymond; Debessai, Mathew; Vindana, Hiranya; Venkatasamy, Kevin; Lawler, Keith V.; Salamat, Ashkan; Dias, Ranga P. Room-temperature superconductivity in a carbonaceous sulfur hydride. *Nature*. 15 October 2020, **586** (7829): 373–377. PMID 33057222. doi:10.1038/s41586-020-2801-z.
- ^{2.} Dai, P.; Chakoumakos, B.C.; Sun, G.F.; Wong, K.W.; Xin, Y.; Lu, D.F. Synthesis and neutron powder diffraction study of the superconductor HgBa₂Ce₂Cu₃O_{8+δ} by Tl substitution. *Physica C*. 1995, **243** (3–4): 201–206. Bibcode:1995PhyC..243..201D. doi:10.1016/0921-4534(94)02461-8.
- ^{3.} Geballe, T. H. Paths to Higher Temperature Superconductors. *Science*. 12 March 1993, **259** (5101): 1550–1551. Bibcode:1993Sci...259.1550G. PMID 17733017. doi:10.1126/science.259.5101.1550.
- ^{4.} Almaden Institute 2012. Superconductivity 297 K – Synthetic Routes to Room Temperature Superconductivity. researcher.watson.ibm.com. 25 July 2016 [2021-09-02]. (原始内容存档于2013-12-12).
- ^{5.} NOVA. Race for the Superconductor. Public TV station WGBH Boston. Approximately 1987.
- ^{6.} The First Room-Temperature Ambient-Pressure Superconductor (PDF). [25 July 2023]. (原始内容存档 (PDF)于25 July 2023).
- ^{7.} Superconductor Pb10-xCux(PO4)6O showing levitation at room temperature and atmospheric pressure and mechanism. [26 July 2023]. (原始内容存档于25 July 2023).

大规模预训练模型知识更新来源

▶ 知识来源

- ▶ 知识库/知识图谱
- ▶ 非结构化知识库
- ▶ 人类监督与交互



大规模预训练模型知识更新手段

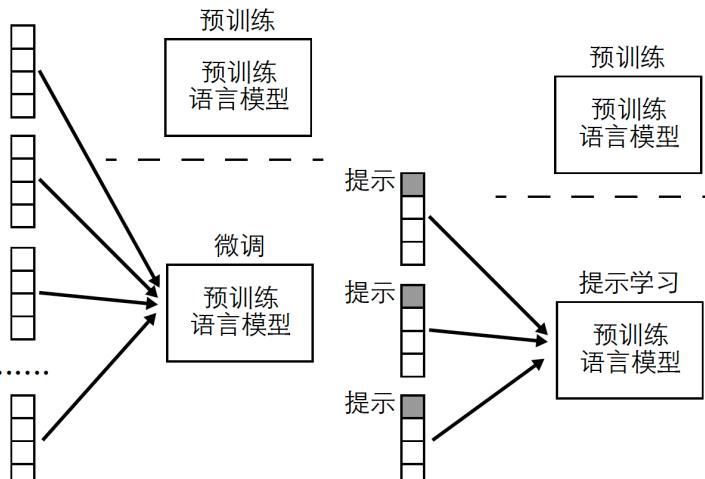
- ▶ 面向大规模预训练模型的知识更新手段
 - ▶ 提示学习（同时适用于开源和闭源模型）
 - ▶ 轻量微调（适用于开源模型）

目录

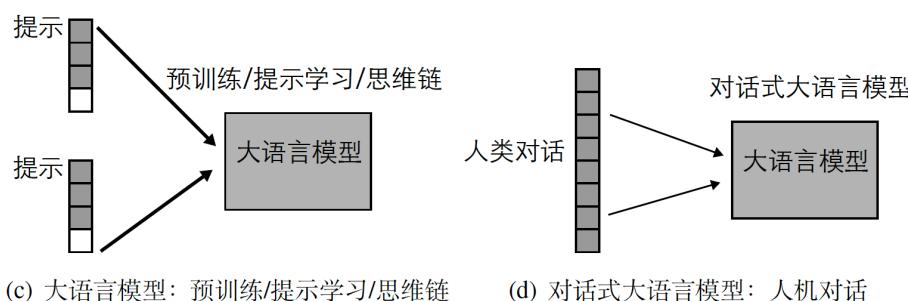
- ▶ 预训练语言模型的发展
 - ▶ 大模型的发展与新范式
 - ▶ 从GPT-3到ChatGPT
- ▶ 涌现能力与幻觉问题
- ▶ 提示学习技术
 - ▶ 上下文学习
 - ▶ 思维链推理
- ▶ 开源家族：LLaMA与其后继者
- ▶ 把大模型变小：模型量化和LoRA微调

提示学习

- ▶ 对大语言模型微调需要耗费难以承受的计算代价
 - ▶ 动机：随着参数增加带来的算力资源和计算时间的急剧扩大
 - ▶ 路线：降低微调参数量和提示学习



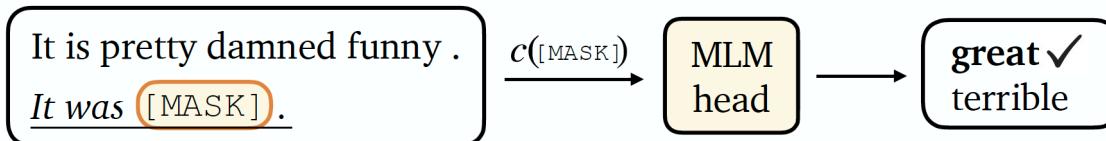
- ▶ 提示学习利用“对齐”的思想，令下游任务模式接近预训练模式
 - ▶ 上下文提示
 - ▶ 思维链提示



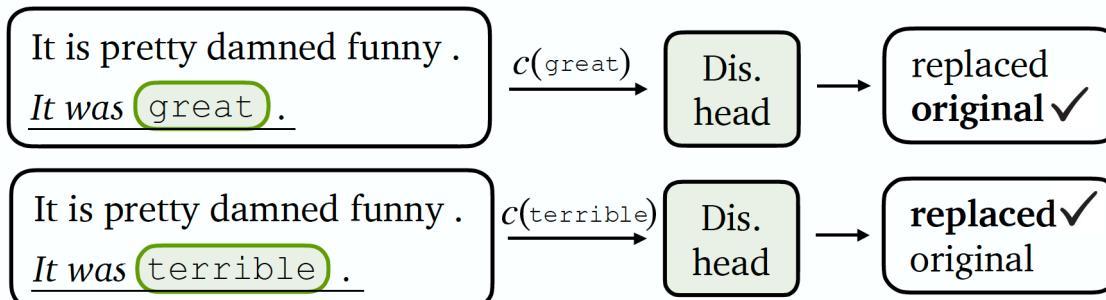
逐渐走向人类友好的交互模式

提示学习

- ▶ 面向判别式模型（编码器模型）的提示学习
 - ▶ 将下游任务构造成类似完形填空的形式



(a) 在SST-2情感分析数据集上提示RoBERTa



(b) 在SST-2情感分析数据集上提示ELECTRA

提示学习

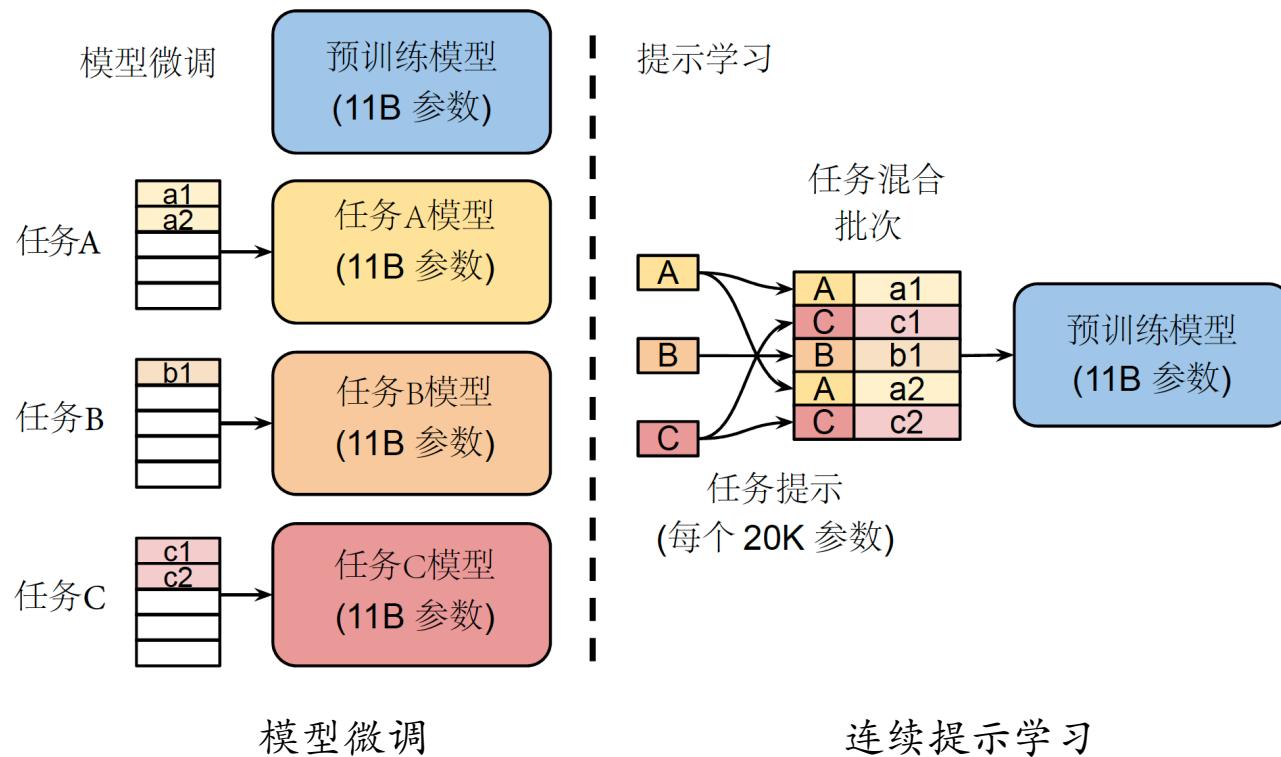
- ▶ 面向生成式模型（解码器模型）的提示学习
 - ▶ 使用自然语言对任务进行描述

零样本学习	单样本学习	少样本学习
Translate English to French cheese =>	Translate English to French sea otter => loutre de mer cheese =>	Translate English to French sea otter => loutre de mer peppermint => menthe poivrée plush giraffe => girafe peluche cheese =>

提示学习

▶ 连续提示学习

- ▶ 在输入部分添加可参数更新的提示前缀
- ▶ 可被视为是降低微调参数量和提示学习这两条技术路线的汇流结果



上下文提示 (In-Context Learning)

- ▶ 上下文学习方式
 - ▶ 零样本提示：给出目标指令提示
 - ▶ 少样本提示：提供任务范例提示
- ▶ 主要功能
 - ▶ 针对特定任务的输入-输出格式约束
 - ▶ 提供上下文，缩小知识搜索空间
 - ▶ 无需梯度更新的知识更新手段

Translate English to French
cheese =>

(Output) Le fromage

零样本提示

Translate English to French
sea otter => loutre de mer
peppermint => menthe poivrée
plush giraffe => girafe peluche
cheese =>

(Output) Le fromage

少样本提示

上下文学习进阶：检索增强

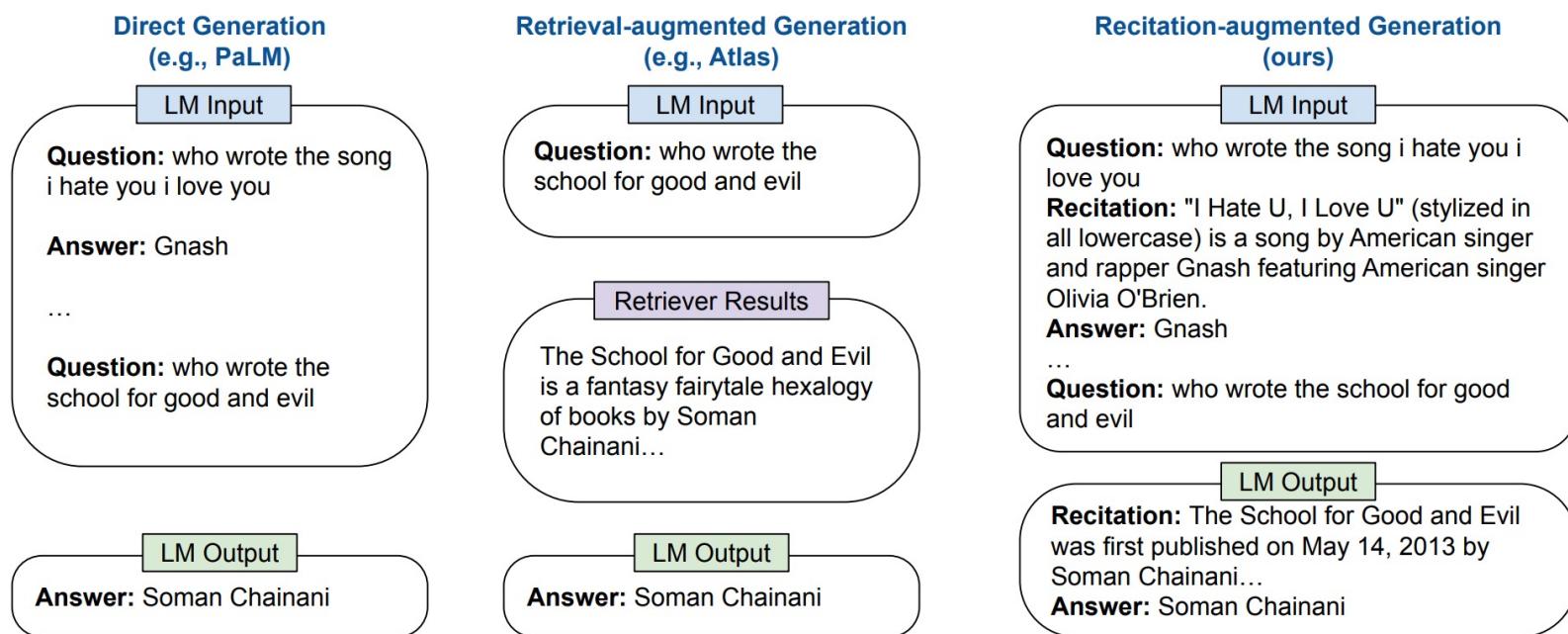
▶ 为什么要做上下文检索增强？

- ▶ 让大模型获取更多的知识，尤其最新的信息
- ▶ 缓解大模型存在的幻觉问题，提供外挂信息，可让大模型的输出更有据可循
- ▶ 结合领域知识库，低成本实现领域适配

上下文学习进阶：检索增强

▶ 上下文检索增强的手段

- ▶ 直接提示增强：将检索到与query相关的内容直接放在prompt中，作为背景知识输入至大模型
- ▶ 自回归式增强：先利用LLM解码出部分内容，然后检索相关的文档并将其拼接在提示中，再次输入至模型，实现自回归式多轮解码



上下文学习进阶：检索增强

▶ 上下文检索增强的手段

- ▶ 直接提示增强：将检索到与query相关的内容直接放在prompt中，作为背景知识输入至大模型
- ▶ 自回归式增强：先利用LLM解码出部分内容，然后检索相关的文档并将其拼接在提示中，再次输入至模型，实现自回归式多轮解码



思维链推理技术

- ▶ 思维链推理：引导模型生成中间思维步骤，进而推导答案
 - ▶ 无需梯度更新
 - ▶ 提升推理性能，增强可解释性
- ▶ 任务求解方式的变化
 - ▶ 常规模式：<问题 → 答案>
 - ▶ 思维链模式：<问题 → 思维链 → 答案>

Q: There were 10 friends playing a video game online when 7 players quit. If each player left had 8 lives, how many lives did they have total?

A: The answer is

(Output) 80. ✗

常规模式

问题

Q: There were 10 friends playing a video game online when 7 players quit. If each player left had 8 lives, how many lives did they have total?

A: Let's think step by step.

(Output) There were 10 friends playing a video game online. This means that, at the start, there were $10 \times 8 = 80$ lives in total. Then, 7 players quit. This means that $7 \times 8 = 56$ lives were lost. Therefore, the total number of lives remaining is $80 - 56 = 24$. The answer is ✓

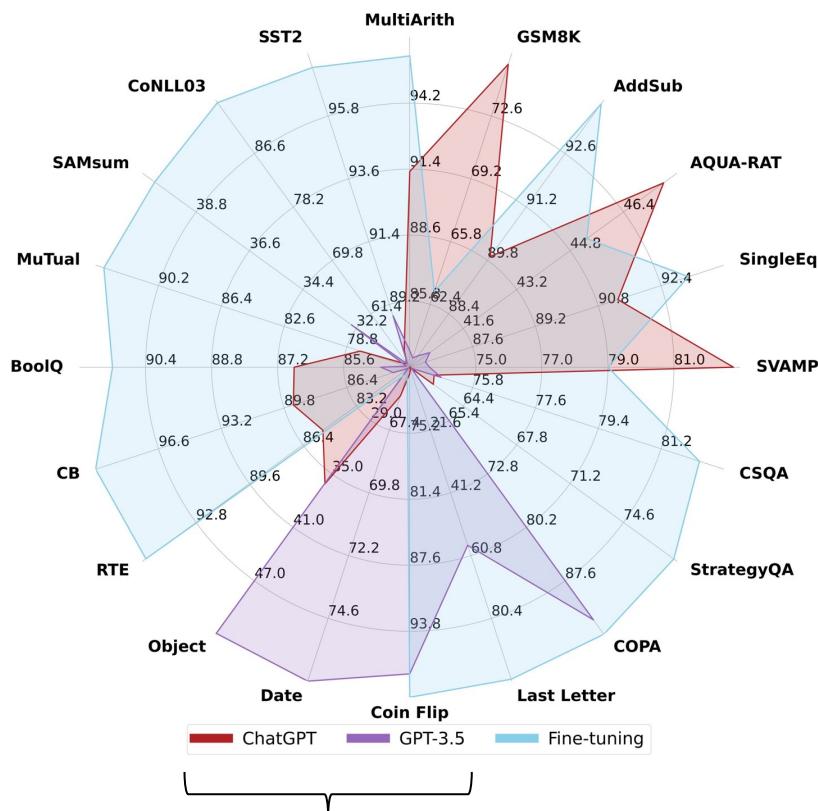
思维链

答案

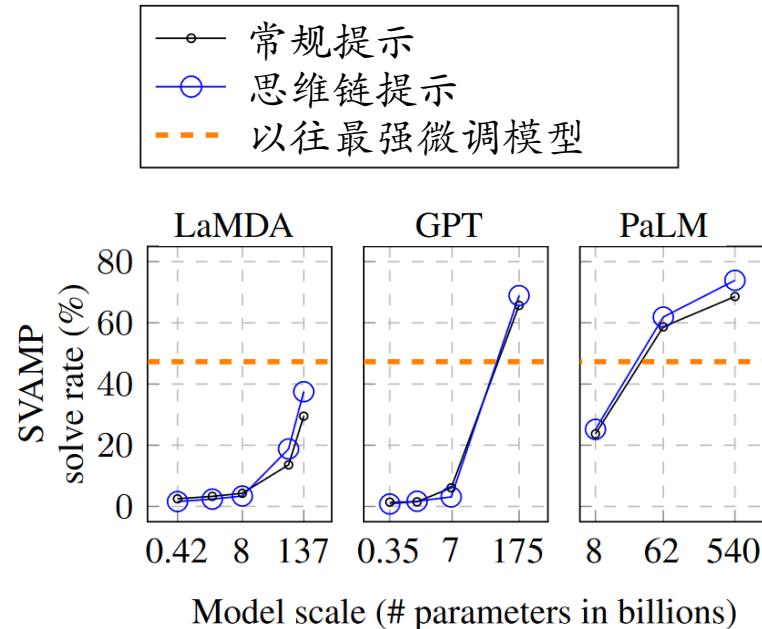
思维链模式

思维链推理技术

- ▶ 被视为大模型最具代表性的“涌现能力”
 - ▶ 在数学推理、常识推理和逻辑推理复杂任务上的性能超越常规微调模型
 - ▶ ChatGPT展现了较强的思维链推理能力



基于思维链提示



目录

- ▶ 预训练语言模型的发展
 - ▶ 大模型的发展与新范式
 - ▶ 从GPT-3到ChatGPT
- ▶ 涌现能力与幻觉问题
- ▶ 提示学习技术
 - ▶ 上下文学习
 - ▶ 思维链推理
- ▶ 开源家族：LLaMA与其后继者
- ▶ 把大模型变小：模型量化和LoRA微调

开源家族：LLaMA与其后继者

▶ 开源大模型对比（2019-2023）

Model	Release Time	Size (B)	Base Model	Adaptation IT	RLHF	Pre-train Data Scale	Latest Data Timestamp	Hardware (GPUs / TPUs)	Training Time	Evaluation ICL	Evaluation CoT
Publicly Available	T5 [73]	Oct-2019	11	-	-	1T tokens	Apr-2019	1024 TPU v3	-	✓	-
	mT5 [74]	Oct-2020	13	-	-	1T tokens	-	-	-	✓	-
	PanGu- α [75]	Apr-2021	13*	-	-	1.1TB	-	2048 Ascend 910	-	✓	-
	CPM-2 [76]	Jun-2021	198	-	-	2.6TB	-	-	-	-	-
	T0 [28]	Oct-2021	11	T5	✓	-	-	512 TPU v3	27 h	✓	-
	CodeGen [77]	Mar-2022	16	-	-	577B tokens	-	-	-	✓	-
	GPT-NeoX-20B [78]	Apr-2022	20	-	-	825GB	-	96 40G A100	-	✓	-
	Tk-Instruct [79]	Apr-2022	11	T5	✓	-	-	256 TPU v3	4 h	✓	-
	UL2 [80]	May-2022	20	-	-	1T tokens	Apr-2019	512 TPU v4	-	✓	✓
	OPT [81]	May-2022	175	-	-	180B tokens	-	992 80G A100	-	✓	-
	NLLB [82]	Jul-2022	54.5	-	-	-	-	-	-	✓	-
	GLM [83]	Oct-2022	130	-	-	400B tokens	-	768 40G A100	60 d	✓	-
	Flan-T5 [64]	Oct-2022	11	T5	✓	-	-	-	-	✓	✓
	BLOOM [69]	Nov-2022	176	-	-	366B tokens	-	384 80G A100	105 d	✓	-
	mT0 [84]	Nov-2022	13	mT5	✓	-	-	-	-	✓	-
	Galactica [35]	Nov-2022	120	-	-	106B tokens	-	-	-	✓	✓
	BLOOMZ [84]	Nov-2022	176	BLOOM	✓	-	-	-	-	✓	-
	OPT-IML [85]	Dec-2022	175	OPT	✓	-	-	128 40G A100	-	✓	✓
	LLaMA [57]	Feb-2023	65	-	-	1.4T tokens	-	2048 80G A100	21 d	✓	-
	CodeGeeX [86]	Sep-2022	13	-	-	850B tokens	-	1536 Ascend 910	60 d	✓	-
	Pythia [87]	Apr-2023	12	-	-	300B tokens	-	256 40G A100	-	✓	-

开源家族：LLaMA与其后继者

▶

LLaMA

- 参数版本: 7B, 13B, 33B, 65B
 - 训练数据: CommonCrawl, C4, Github, Wikipedia, books, ArXiv, and StackExchange
- 指令微调版本
 - Stanford Alpaca: 在GPT-3.5 (text-davinci-003) 生成的 52K 指令数据上微调
 - Vicuna: 基于ShareGPT收集的63K指令数据微调

Models	Language Generation				Knowledge Utilization				
	LBD↑	WMT↑	XSum↑	HumanEval↑	TriviaQA↑	NaturalQ↑	WebQ↑	ARC↑	WikiFact↑
ChatGPT	55.81	36.44	21.71	79.88	54.54	21.52	17.77	93.69	29.25
Claude	64.47	31.23	22.86	51.22	40.92	13.77	14.57	66.62	34.34
Davinci003	69.98	37.46	18.19	67.07	51.51	17.76	16.68	88.47	28.29
Davinci002	58.85	35.11	19.15	56.70	52.11	20.47	18.45	89.23	29.15
Vicuna (7B)	60.12	18.06	13.59	17.07	28.58	9.17	6.64	16.96	26.95
Alpaca (7B)	60.45	21.52	8.74	13.41	17.14	3.24	3.00	49.75	26.05
ChatGLM (6B)	33.34	16.58	13.48	13.42	13.42	4.40	9.20	55.39	16.01
LLaMA (7B)	66.78	13.84	8.77	15.24	34.62	7.92	11.12	4.88	19.78
Falcon (7B)	66.89	4.05	10.00	10.37	28.74	10.78	8.46	4.08	23.91
Pythia (12B)	60.49	5.43	8.87	14.63	15.73	1.99	4.72	11.66	20.57
Pythia (7B)	50.96	3.68	8.23	9.15	10.16	1.77	3.74	11.03	15.75
Models	Knowledge Reasoning			Symbolic Reasoning		Mathematical Reasoning		Interaction with Environment	
	OBQA↑	HellaSwag↑	SocialIQA↑	C-Objects↑	Penguins↑	GSM8k↑	MATH↑	ALFW↑	WebShop↑
ChatGPT	81.20	61.43	73.23	53.20	40.27	78.47	33.78	58.96	45.12/15.60
Claude	81.80	54.95	73.23	59.95	47.65	70.81	20.18	32.09	50.02/30.40
Davinci003	74.40	62.65	69.70	64.60	61.07	57.16	17.66	65.67	64.08/32.40
Davinci002	69.80	47.81	57.01	62.55	67.11	49.96	14.28	76.87	29.66/15.20
Vicuna (7B)	30.00	26.26	36.39	44.25	36.24	14.03	3.54	1.49	6.90/1.40
Alpaca (7B)	28.60	26.03	33.52	39.35	40.27	4.93	4.16	4.48	0.00/0.00
ChatGLM (6B)	52.00	40.60	57.52	14.05	14.09	3.41	1.10	0.00	0.00/0.00
LLaMA (7B)	27.00	25.57	33.11	39.95	34.90	10.99	3.12	2.24	0.00/0.00
Falcon (7B)	25.20	25.07	33.01	29.80	24.16	1.67	0.94	7.46	0.00/0.00
Pythia (12B)	25.00	25.15	32.45	32.40	26.17	2.88	1.96	5.22	3.68/0.60
Pythia (7B)	24.40	23.62	32.04	29.05	27.52	1.82	1.46	7.46	10.75/1.80

注：ShareGPT是分享
ChatGPT对话的谷歌插件，
拥有超过11万对话数量

开源家族：LLaMA与其后继者

▶ LLaMA

- ▶ 模型特点：自回归模型，基于Transformer架构
- ▶ 参数版本： 7B, 13B , 30B, 65B
- ▶ 训练数据： CommonCrawl, C4, Github, Wikipedia, books, ArXiv, and StackExchange
- ▶ 许可： 可用于学术研究
- ▶ 地址： <https://github.com/facebookresearch/llama>

Dataset	Sampling prop.	Epochs	Disk size
CommonCrawl	67.0%	1.10	3.3 TB
C4	15.0%	1.06	783 GB
Github	4.5%	0.64	328 GB
Wikipedia	4.5%	2.45	83 GB
Books	4.5%	2.23	85 GB
ArXiv	2.5%	1.06	92 GB
StackExchange	2.0%	1.03	78 GB

		BoolQ	PIQA	SIQA	HellaSwag	WinoGrande	ARC-e	ARC-c	OBQA
GPT-3	175B	60.5	81.0	-	78.9	70.2	68.8	51.4	57.6
Gopher	280B	79.3	81.8	50.6	79.2	70.1	-	-	-
Chinchilla	70B	83.7	81.8	51.3	80.8	74.9	-	-	-
PaLM	62B	84.8	80.5	-	79.7	77.0	75.2	52.5	50.4
PaLM-cont	62B	83.9	81.4	-	80.6	77.0	-	-	-
PaLM	540B	88.0	82.3	-	83.4	81.1	76.6	53.0	53.4
LLaMA	7B	76.5	79.8	48.9	76.1	70.1	72.8	47.6	57.2
	13B	78.1	80.1	50.4	79.2	73.0	74.8	52.7	56.4
	33B	83.1	82.3	50.4	82.8	76.0	80.0	57.8	58.6
	65B	85.3	82.8	52.3	84.2	77.0	78.9	56.0	60.2

params	dimension	n heads	n layers	learning rate	batch size	n tokens
6.7B	4096	32	32	$3.0e^{-4}$	4M	1.0T
13.0B	5120	40	40	$3.0e^{-4}$	4M	1.0T
32.5B	6656	52	60	$1.5e^{-4}$	4M	1.4T
65.2B	8192	64	80	$1.5e^{-4}$	4M	1.4T

开源家族：LLaMA与其后继者

▶ LLaMA2 & Chat

- ▶ 参数版本: 7B, 13B , 70B
- ▶ 70B模型采用分组查询注意力，以适应更长的输入
- ▶ 聊天模型可以使用工具和插件

▶ LLaMA2相比LLAMA 1的升级

- ▶ 训练语料增加40%，上下文长度从2048升级到4096
- ▶ 可用于研究和商业用途

Llama 2

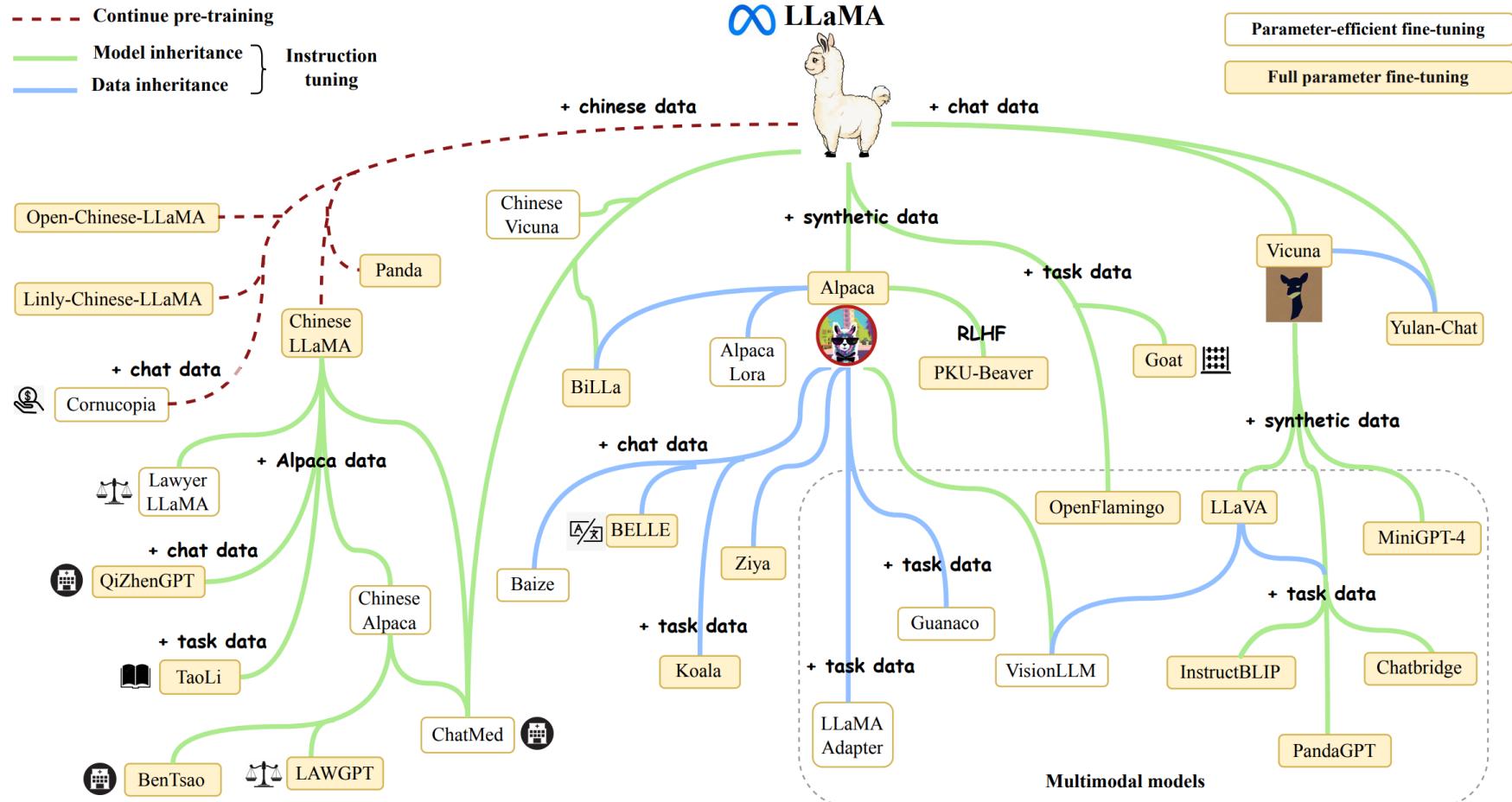
MODEL SIZE (PARAMETERS)	PRETRAINED	FINE-TUNED FOR CHAT USE CASES
7B	Model architecture:	Data collection for helpfulness and safety:
13B	Pretraining Tokens: 2 Trillion	Supervised fine-tuning: Over 100,000
70B	Context Length: 4096	Human Preferences: Over 1,000,000

开源家族：LLaMA与其后继者

- ▶ LLaMA2：在推理、编程、知识问答等任务上均取得优越性能

Benchmark (Higher is better)	MPT (7B)	Falcon (7B)	Llama-2 (7B)	Llama-2 (13B)	MPT (30B)	Falcon (40B)	Llama-1 (65B)	Llama-2 (70B)
MMLU	26.8	26.2	45.3	54.8	46.9	55.4	63.4	68.9
TriviaQA	59.6	56.8	68.9	77.2	71.3	78.6	84.5	85.0
Natural Questions	17.8	18.1	22.7	28.0	23.0	29.5	31.0	33.0
GSM8K	6.8	6.8	14.6	28.7	15.2	19.6	50.9	56.8
HumanEval	18.3	N/A	12.8	18.3	25.0	N/A	23.7	29.9
AGIEval (English tasks only)	23.5	21.2	29.3	39.1	33.8	37.0	47.6	54.2

开源家族：LLaMA与其后继者



Grid Math Money Bag Finance Hospital Medicine Scales Law Bi-directional arrow Bilingualism Book Education

目录

- ▶ 预训练语言模型的发展
 - ▶ 大模型的发展与新范式
 - ▶ 从GPT-3到ChatGPT
- ▶ 涌现能力与幻觉问题
- ▶ 提示学习技术
 - ▶ 上下文学习
 - ▶ 思维链推理
- ▶ 开源家族：LLaMA与其后继者
- ▶ 把大模型变小：模型量化和LoRA微调

模型量化

- ▶ 概念：将模型中的参数或者激活值从高精度转换为低精度的过程
- ▶ 效果：减少模型大小和运算复杂度，同时保持模型的性能

不降低性能的情况下降低大模型的应用门槛

模型量化

- ▶ **量化对象：**参数和激活值
- ▶ **量化的位数：**用多少比特来表示一个参数或者一个激活值
 - ▶ 量化的位数越低，表示范围越小，精度越低
 - ▶ 量化的位数越高，表示范围越大，精度越高
- ▶ **量化的分类**
 - ▶ 训练时量化
 - ▶ 推理时量化

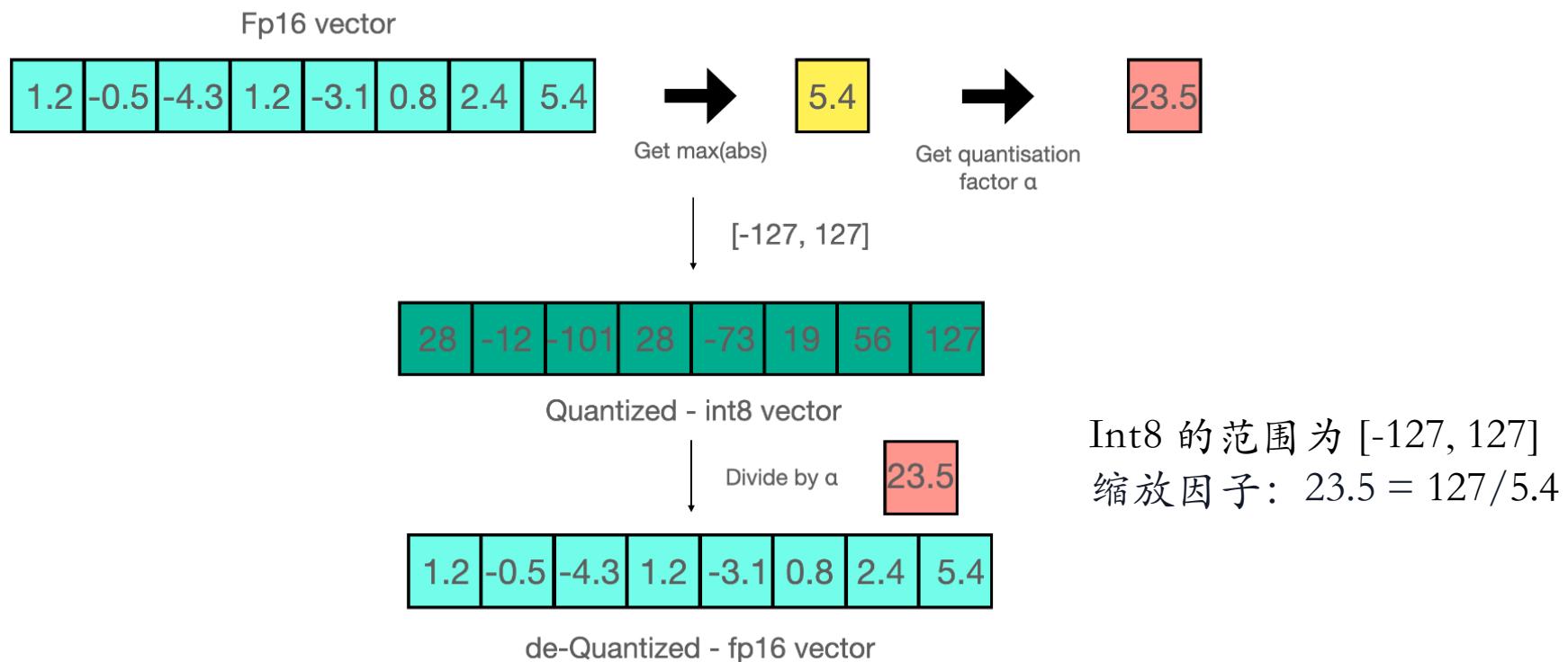
int4或者int8是指用4位或者8位整数来表示一个参数或者一个激活值。例如，int4可以表示从-8到7之间的16个整数；int8可以表示从-128到127之间的256个整数。

float32是指用32位浮点数来表示一个参数或者一个激活值。例如，float32可以表示从-3.4e38到3.4e38之间的约4.3e9个实数。

模型量化

- ▶ 量化是一个有噪过程，会导致信息丢失，是一种有损压缩
- ▶ 例：Int8最大绝对值量化
 - ▶ 计算最大绝对值和缩放因子
 - ▶ 原始值乘以缩放因子得到量化向量

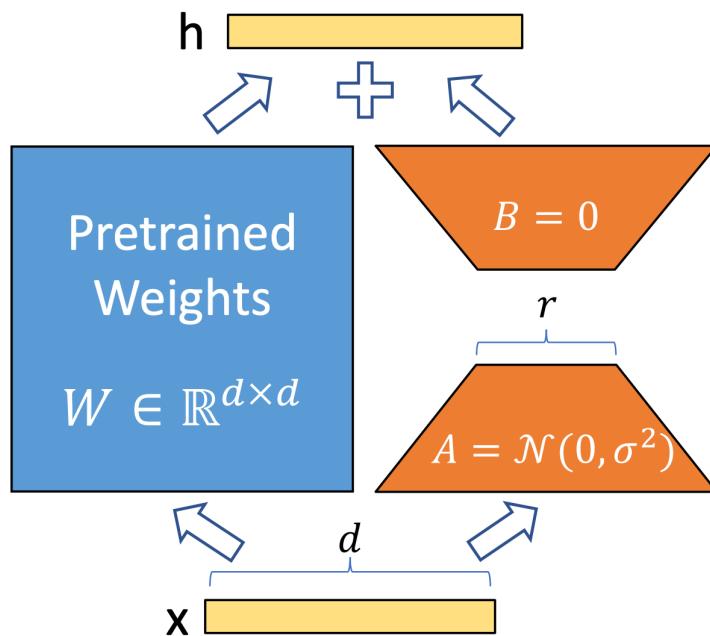
```
model = LlamaForSequenceClassification.from_pretrained(  
    base_model,  
    load_in_8bit=load_8bit,  
    torch_dtype=torch.float16,  
    device_map="auto",  
)
```



LoRA (Low-Rank Adaptation of Large Language Models)

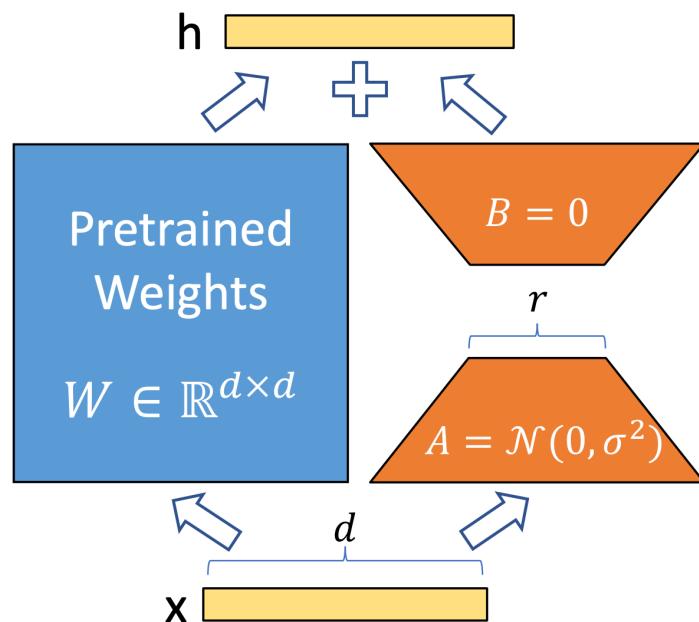
▶ 基于低秩的轻量化微调方法

- ▶ 核心思想：参数优化量可以是低秩的，映射到低维空间下也能保持性能
- ▶ 实现手段：对模型参数的优化量进行低秩近似，以残差的方式更新参数



LoRA (Low-Rank Adaptation of Large Language Models)

- ▶ 全参数微调 = 冻结的预训练权重 + 微调过程中产生的权重更新量
- ▶ LoRA: 训练时固定 W , 只对 A 和 B 更新



增量权重的近似

$$h = Wx + BAx \longrightarrow h = Wx + \frac{\alpha}{r} BAx$$

$A \in \mathbb{R}^{r \times d}$: 低秩矩阵, 其中 r 被称为“秩”, 对A采用高斯初始化

$B \in \mathbb{R}^{d \times r}$: 低秩矩阵, 对B采用零初始化

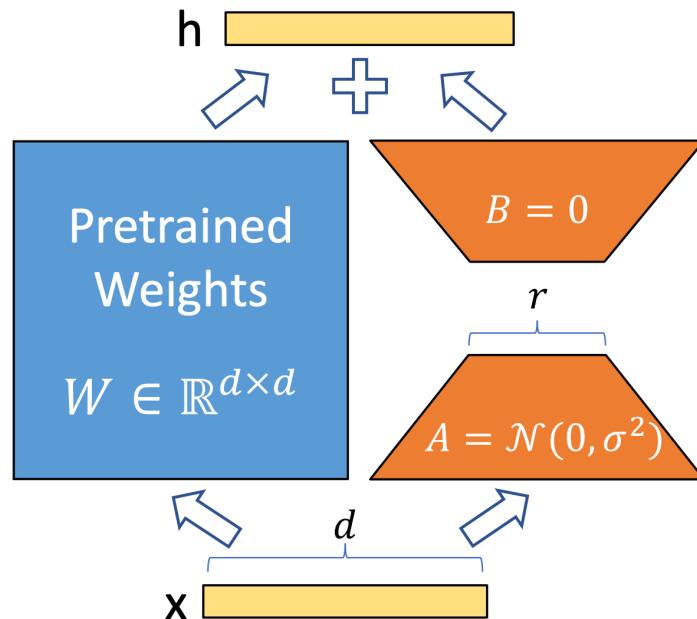
训练权重从 $d \times d$ 降为 $2 \times r \times d$

$$r \ll \min(d, k)$$

LoRA (Low-Rank Adaptation of Large Language Models)

► LoRA:

- 只作用于Attention部分。训练时固定 W , 只对 A 和 B 更新
- 推理时, 合并低秩矩阵和预训练权重, 按常规方式做前向推理



以微调GPT3 175B为例

	显存消耗	存储开销
全量微调	1.5T	350G
LoRA	350G	35M

$$h = Wx + BAx \longrightarrow h = Wx + \frac{\alpha}{r} BAx$$

$A \in \mathbb{R}^{r*d}$: 低秩矩阵, 其中 r 被称为“秩”, 对A采用高斯初始化

$B \in \mathbb{R}^{d*r}$: 低秩矩阵, 对B采用零初始化

LoRA (Low-Rank Adaptation of Large Language Models)

► LoRA:

- 只作用于Attention部分。训练时固定 W , 只对 A 和 B 更新
- 推理时, 合并低秩矩阵和预训练权重, 按常规方式做前向推理

	Weight Type	$r = 1$	$r = 2$	$r = 4$	$r = 8$	$r = 64$
WikiSQL($\pm 0.5\%$)	W_q	68.8	69.6	70.5	70.4	70.0
	W_q, W_v	73.4	73.3	73.7	73.8	73.5
	W_q, W_k, W_v, W_o	74.1	73.7	74.0	74.0	73.9
MultiNLI ($\pm 0.1\%$)	W_q	90.7	90.9	91.1	90.7	90.7
	W_q, W_v	91.3	91.4	91.3	91.6	91.4
	W_q, W_k, W_v, W_o	91.2	91.7	91.7	91.5	91.4

Model&Method	# Trainable Parameters	WikiSQL	MNLI-m	SAMSum
		Acc. (%)	Acc. (%)	R1/R2/RL
GPT-3 (FT)	175,255.8M	73.8	89.5	52.0/28.0/44.5
GPT-3 (BitFit)	14.2M	71.3	91.0	51.3/27.4/43.5
GPT-3 (PreEmbed)	3.2M	63.1	88.6	48.3/24.2/40.5
GPT-3 (PreLayer)	20.2M	70.1	89.5	50.8/27.3/43.5
GPT-3 (Adapter ^H)	7.1M	71.9	89.8	53.0/28.9/44.8
GPT-3 (Adapter ^H)	40.1M	73.2	91.5	53.2/29.0/45.1
GPT-3 (LoRA)	4.7M	73.4	91.7	53.8/29.8/45.9
GPT-3 (LoRA)	37.7M	74.0	91.6	53.4/29.2/45.1

- 训练时针对不同任务分别训练多个LoRA模块
- 部署时, 仅需部署主干模型, 根据不同的任务使用不同的LoRA模块。

LoRA (Low-Rank Adaptation of Large Language Models)

▶ 代码写法

```
from transformers import AutoModelForCausalLM, AutoTokenizer

model_name = "bigscience/bloomz-560m"
#model_name="bigscience/bloom-1b1"

tokenizer = AutoTokenizer.from_pretrained(model_name)
foundation_model = AutoModelForCausalLM.from_pretrained(model_name)

import peft
from peft import LoraConfig, get_peft_model, PeftModel

lora_config = LoraConfig(
    r=4, #As bigger the R bigger the parameters to train.
    lora_alpha=1, # a scaling factor that adjusts the magnitude of the weight matrix. Usually set to 1
    target_modules=["query_key_value"], #You can obtain a list of target modules in the URL above.
    lora_dropout=0.05, #Helps to avoid Overfitting.
    bias="lora_only", # this specifies if the bias parameter should be trained.
    task_type="CAUSAL_LM"
)

peft_model = get_peft_model(foundation_model, lora_config)
print(peft_model.print_trainable_parameters())

#This cell may take up to 15 minutes to execute.
trainer = Trainer(
    model=peft_model,
    args=training_args,
    train_dataset=train_sample,
    data_collator=transformers.DataCollatorForLanguageModeling(tokenizer, mlm=False)
)
trainer.train()
```

初始化模型

初始化LoRA

LoRA模块

训练LoRA

动手学大模型：预训练语言模型微调与部署

动手学大模型：预训练语言模型微调与部署

 Zhuosheng Zhang | 昨天创建

导读: 该部分介绍预训练模型微调

想提升预训练模型在指定任务上的性能？让我们选择合适的预训练模型，在特定任务上进行微调，并将微调后的模型部署成方便使用的Demo！

本教程目标：

1. 熟悉使用Transformers工具包
2. 掌握预训练模型的微调、推理
3. 掌握利用Gradio Spaces进行Demo部署
4. 了解不同类型的预训练模型的选型和应用场景

<https://sjtullm.gitbook.io/dive-into-langs/2.tuning>

谢 谢 !