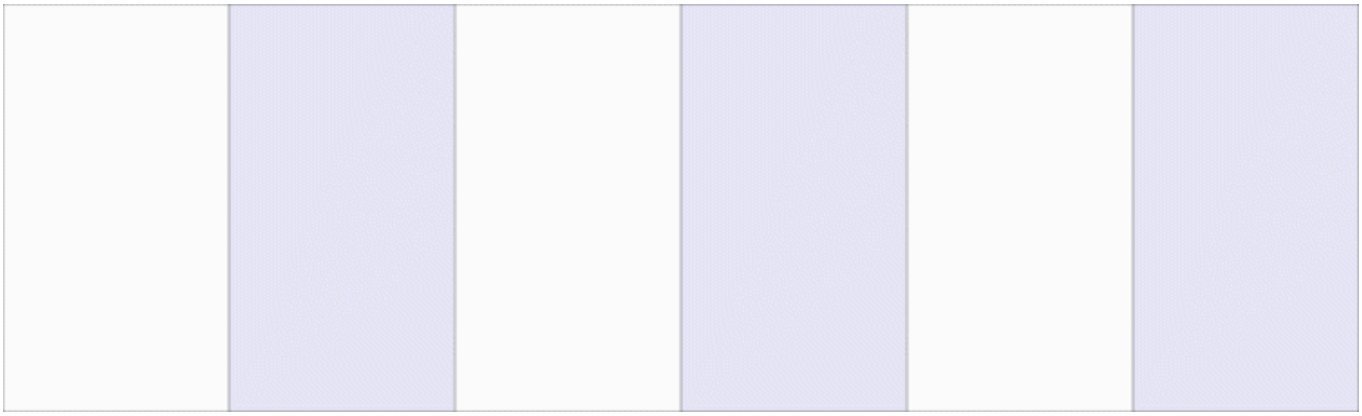
大型语言模型综述全新出炉：从T5到GPT-4最全盘点，国内20余位研究者联合撰写

CV开发者都爱看的 极市平台 2023-04-04 22:00:12 发表于广东 手机阅读



来源丨机器之心

编辑丨极市平台

极市导读

为什么仿佛一夜之间，自然语言处理（NLP）领域就突然突飞猛进，摸到了通用人工智能的门槛？如今的大语言模型（LLM）发展到了什么程度？未来短时间内，AGI 的发展路线又将如何？ >>加入极市CV技术交流群，走在计算机视觉的最前沿

自 20 世纪 50 年代图灵测试提出以来，人们始终在探索机器处理语言智能的能力。语言本质上是一个错综复杂的人类表达系统，受到语法规则的约束。因此，开发能够理解和精通语言的强大 AI 算法面临着巨大挑战。过去二十年，语言建模方法被广泛用于语言理解和生成，包括统计语言模型和神经语言模型。

近些年，研究人员通过在大规模语料库上预训练 Transformer 模型产生了预训练语言模型（PLMs），并在解决各类 NLP 任务上展现出了强大的能力。并且研究人员发现模型缩放可以带来性能提升，因此他们通过将模型规模增大进一步研究缩放的效果。有趣的是，当参数规模超过一定水平时，这个更大的语言模型实现了显著的性能提升，并出现了小模型中不存在的能力，比如上下文学习。为了区别于 PLM，这类模型被称为大型语言模型（LLMs）。

从 2019 年的谷歌 T5 到 OpenAI GPT 系列，参数量爆炸的大模型不断涌现。可以说，LLMs 的研究在学界和业界都得到了很大的推进，尤其去年 11 月底对话大模型 ChatGPT 的出现更是

引起了社会各界的广泛关注。LLMs 的技术进展对整个 AI 社区产生了重要影响，并将彻底改变人们开发和使用 AI 算法的方式。

考虑到 LLMs 的快速技术进步，中国人民大学的二十几位研究者通过背景知识、关键发现和主流技术等三方面回顾了 LLMs 的最新进展，尤其关注 LLMs 的预训练、自适应调优、使用和能力评估。此外他们还总结和开发 LLMs 的可用资源，讨论了未来发展方向等问题。对于领域内研究人员和工程师而言，这份综述是一份极其有用的学习资源。

A Survey of Large Language Models

Wayne Xin Zhao, Kun Zhou*, Junyi Li*, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, Peiyu Liu, Jian-Yun Nie and Ji-Rong Wen

论文链接：https://arxiv.org/abs/2303.18223

在进入正文前，我们先来看 2019 年以来出现的各种大语言模型（百亿参数以上）时间轴，其中标黄的大模型已开源。

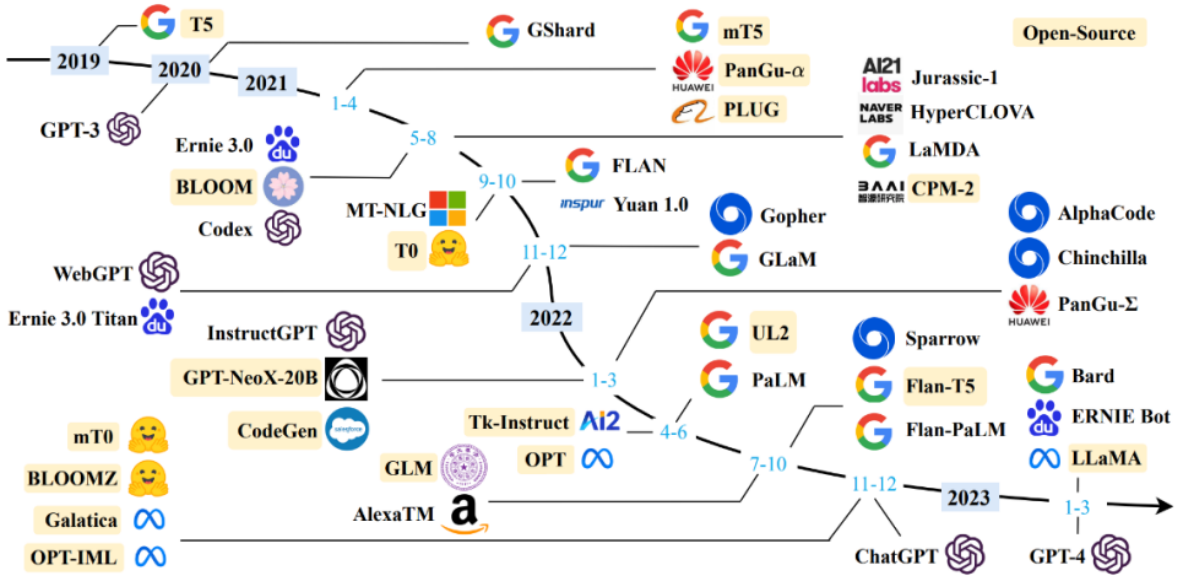


Fig. 1. A timeline of existing large language models (having a size larger than 10B) in recent years. We mark the open-source LLMs in yellow color.

LLMs 概览

在第一节中，研究者详细介绍了 LLMs 的背景、能力和关键技术。

LLMs 的背景

通常，大型语言模型（LLM）是指包含数千亿（或更多）参数的语言模型，这些参数是在大量文本数据上训练的，例如模型 GPT-3、PaLM、Galactica 和 LLaMA。具体来说，LLM 建立在 Transformer 架构之上，其中多头注意力层堆叠在一个非常深的神经网络中。现有的 LLM 主要

采用与小语言模型类似的模型架构（即 Transformer）和预训练目标（即语言建模）。作为主要区别，LLM 在很大程度上扩展了模型大小、预训练数据和总计算量（扩大倍数）。他们可以更好地理解自然语言，并根据给定的上下文（例如 prompt）生成高质量的文本。这种容量改进可以用标度律进行部分地描述，其中性能大致遵循模型大小的大幅增加而增加。然而根据标度律，某些能力（例如，上下文学习）是不可预测的，只有当模型大小超过某个水平时才能观察到。

LLMs 的涌现能力

LLM 的涌现能力被正式定义为「在小型模型中不存在但在大型模型中出现的能力」，这是 LLM 与以前的 PLM 区分开来的最显著特征之一。当出现这种新的能力时，它还引入了一个显著的特征：当规模达到一定水平时，性能显著高于随机的状态。以此类推，这种新模式与物理学中的相变现象密切相关。原则上，这种能力也可以与一些复杂的任务有关，而人们更关心可以应用于解决多个任务的通用能力。这里简要介绍了 LLM 的三种代表性的涌现能力：

上下文学习。GPT-3 正式引入了上下文学习能力：假设语言模型已经提供了自然语言指令和多个任务描述，它可以通过完成输入文本的词序列来生成测试实例的预期输出，而无需额外的训练或梯度更新。

指令遵循。通过对自然语言描述（即指令）格式化的多任务数据集的混合进行微调，LLM 在微小的任务上表现良好，这些任务也以指令的形式所描述。这种能力下，指令调优使 LLM 能够在不使用显式样本的情况下通过理解任务指令来执行新任务，这可以大大提高泛化能力。

循序渐进的推理。对于小语言模型，通常很难解决涉及多个推理步骤的复杂任务，例如数学学科单词问题。同时，通过思维链推理策略，LLM 可以通过利用涉及中间推理步骤的 prompt 机制来解决此类任务得出最终答案。据推测，这种能力可能是通过代码训练获得的。

关键技术

接下来来看 LLMs 的关键技术，包括了缩放、训练、能力激发、对齐调优、工具利用等。

缩放。缩放是增加 LLMs 模型容量的关键因素，最开始 GPT-3 将模型参数增至 1750 亿，随后 PaLM 进一步将模型参数增至 5400 亿。大规模参数对于涌现能力至关重要。缩放不仅针对模型大小，还与数据大小和总计算量有关。

训练。由于规模巨大，成功训练一个具备强大能力的 LLMs 非常具有挑战性。因此需要分布式训练算法来学习 LLMs 的网络参数，经常联合使用各种并行策略。为了支持分布式训练，DeepSpeed 和 Megatron-LM 等优化框架被用来促进并行算法的实现和部署。此外，优化技巧对训练稳定性和模型性能也很重要，例如重新启动训练损失尖峰和混合精度训练。最近的 GPT-4 开发了特殊的基础设施和优化方法，从而利用小得多的模型来预测大模型的性能。

能力激发。在大规模语料库上经过预训练后，LLMs 被赋予了解决一般任务的潜在能力。然而当 LLMs 执行某个特定任务时，这些能力可能不会显式地表现出来。因此设计适合的任务指令或特定的上下文策略来激发这些能力非常有用，比如思维链 prompt 有助于通过中间推理步骤等解决复杂推理任务。此外还可以进一步对具有自然语言任务描述的 LLMs 进行指令调优，以提高对未见过任务的泛化能力。

对齐调优。由于 LLMs 被训练用来捕获预训练语料库的数据特征（包括高质量和低质量的数据），它们很可能生成对有毒、有偏见和有害的文本内容。为了使 LLMs 与人类价值观保持一致，InstructGPT 设计了一种利用强化学习和人类反馈的高效调优方法，使得 LLMs 能够遵循预期指令。ChatGPT 是在类似 InstructGPT 的技术上开发的，在产生高质量、无害的响应方面表现出了强大的对齐能力。

工具利用。LLMs 本质上是基于大规模纯文本语料库训练的文本生成器，因此在数值计算等文本表达不佳的任务上表现没那么好。此外 LLMs 的能力受限于预训练数据，无法捕获最新信息。针对这些问题，人们提出使用外部工具来弥补 LLMs 的不足，比如可以利用计算器进行精确计算，使用搜索引擎检索未知信息。ChatGPT 更是利用外部插件来联网学习新知识，这种机制可以广泛扩展 LLMs 的能力范围。

LLMs 资源

考虑到具有挑战性的技术问题和巨大的计算资源需求，开发或复制 LLMs 绝不是一件容易的事情。一个可行的方法是从现有的 LLMs 中学习经验，并重新使用公开的资源来进行渐进式的开发或实验研究。

在第三节中，研究者主要总结了开源的模型检查点或 API、可用的语料库以及对 LLM 有用的库。下表 1 为近年来百亿参数以上大模型的统计数据。

TABLE 1
Statistics of large language models (having a size larger than 10B in this survey) in recent years, including the capacity evaluation, pre-training data scale (either in the number of tokens or storage size) and hardware resource costs. Here, "Adaptation" indicates whether the model has been with subsequent fine-tuning: IT denotes instruction tuning and RLHF denotes reinforcement learning with human feedback. "Evaluation" indicates whether the model has been evaluated with corresponding abilities in their original paper: ICL denotes in-context learning and CoT denotes chain-of-thought. "*" denotes the largest publicly available version.

	Model	Release Time	Size (B)	Base Model	Adaptation IT	Adaptation RLHF	Pre-train Data Scale	Latest Data Timestamp	Hardware (GPUs / TPUs)	Training Time	Evaluation ICL	Evaluation CoT
Open Source	T5 [71]	Oct-2019	11	-	-	-	1T tokens	Apr-2019	1024 TPU v3	-	✓	-
	mT5 [72]	Mar-2021	13	-	-	-	1T tokens	Apr-2019	-	-	✓	-
	PanGu-α [73]	May-2021	13*	-	-	-	1.1TB	-	2048 Ascend 910	-	✓	-
	CPM-2 [74]	May-2021	198	-	-	-	2.6TB	-	-	-	-	-
	T0 [28]	Oct-2021	11	T5	✓	-	-	-	512 TPU v3	27 h	✓	-
	GPT-NeoX-20B [75]	Feb-2022	20	-	-	-	825GB	Dec-2022	96 40G A100	-	✓	-
	CodeGen [76]	Mar-2022	16	-	-	-	577B tokens	-	-	-	✓	-
	Tk-Instruct [77]	Apr-2022	11	T5	✓	-	-	-	256 TPU v3	4 h	✓	-
	UL2 [78]	Apr-2022	20	-	✓	-	1T tokens	Apr-2019	512 TPU v4	-	✓	✓
	OPT [79]	May-2022	175	-	-	-	180B tokens	-	992 80G A100	-	✓	-
	BLOOM [66]	Jul-2022	176	-	-	-	366B	-	384 80G A100	105 d	✓	-
	GLM [80]	Aug-2022	130	-	-	-	400B tokens	-	768 40G A100	60 d	✓	-
	Flan-T5 [81]	Oct-2022	11	T5	✓	-	-	-	-	-	✓	✓
	mT0 [82]	Nov-2022	13	mT5	✓	-	-	-	-	-	✓	-
	Galactica [35]	Nov-2022	120	-	-	-	106B tokens	-	-	-	✓	✓
	BLOOMZ [82]	Nov-2022	176	BLOOM	✓	-	-	-	-	-	✓	-
	OPT-IML [83]	Dec-2022	175	OPT	✓	-	-	-	128 40G A100	-	✓	✓
	LLaMA [57]	Feb-2023	65	-	-	-	1.4T tokens	-	2048 80G A100	21 d	✓	-
Closed Source	GShard [84]	Jan-2020	600	-	-	-	1T tokens	-	2048 TPU v3	4 d	-	-
	GPT-3 [55]	May-2020	175	-	-	-	300B tokens	-	-	-	✓	-
	LaMDA [85]	May-2021	137	-	-	-	2.81T tokens	-	1024 TPU v3	57.7 d	-	-
	HyperCLOVA [86]	Jun-2021	82	-	-	-	300B tokens	-	1024 A100	13.4 d	✓	-
	Codex [87]	Jul-2021	12	GPT-3	-	-	100B tokens	May-2020	-	-	✓	-
	ERNIE 3.0 [88]	Jul-2021	10	-	-	-	375B tokens	-	384 V100	-	✓	-
	Jurassic-1 [89]	Aug-2021	178	-	-	-	300B tokens	-	800 GPU	-	✓	-
	FLAN [62]	Oct-2021	137	LaMDA	✓	-	-	-	128 TPU v3	60 h	✓	-
	MT-NLG [90]	Oct-2021	530	-	-	-	270B tokens	-	4480 80G A100	-	✓	-
	Yuan 1.0 [91]	Oct-2021	245	-	-	-	180B tokens	-	2128 GPU	-	✓	-
	WebGPT [70]	Dec-2021	175	GPT-3	-	✓	-	-	-	-	✓	-
	Gopher [59]	Dec-2021	280	-	-	-	300B tokens	-	4096 TPU v3	920 h	✓	-
	ERNIE 3.0 Titan [92]	Dec-2021	260	-	-	-	300B tokens	-	2048 V100	28 d	✓	-
	GLaM [93]	Dec-2021	1200	-	-	-	280B tokens	-	1024 TPU v4	574 h	✓	-
	InstructGPT [61]	Jan-2022	175	GPT-3	✓	✓	-	-	-	-	✓	-
	AlphaCode [94]	Feb-2022	41	-	-	-	967B tokens	Jul-2021	-	-	-	-
	Chinchilla [34]	Mar-2022	70	-	-	-	1.4T tokens	-	-	-	✓	-
	PaLM [56]	Apr-2022	540	-	-	-	780B tokens	-	6144 TPU v4	-	✓	✓
	AlexaTM [95]	Aug-2022	20	-	-	-	1.3T tokens	-	128 A100	120 d	✓	✓
	Sparrow [96]	Sep-2022	70	-	-	✓	-	-	64 TPU v3	-	✓	-
	U-PaLM [97]	Oct-2022	540	PaLM	-	-	-	-	512 TPU v4	5 d	✓	✓
	Flan-PaLM [81]	Oct-2022	540	PaLM	✓	-	-	-	512 TPU v4	37 h	✓	✓
	Flan-U-PaLM [81]	Oct-2022	540	U-PaLM	✓	-	-	-	-	-	✓	✓
	GPT-4 [46]	Mar-2023	-	-	✓	✓	-	-	-	-	✓	✓
	PanGu-Σ [98]	Mar-2023	1085	PanGu-α	-	-	329B tokens	-	512 Ascend 910	100 d	✓	-

下表 2 列出了常用的数据源。

TABLE 2
Statistics of commonly-used data sources.

Corpora	Size	Source	Latest Update Time
BookCorpus [100]	5GB	Books	Dec-2015
Gutenberg [101]	-	Books	Dec-2021
C4 [71]	800GB	CommonCrawl	Apr-2019
CC-stories-R [102]	31GB	CommonCrawl	Sep-2019
CC-NEWS [27]	78GB	CommonCrawl	Feb-2019
REALNEWs [103]	120GB	CommonCrawl	Apr-2019
OpenWebText [104]	38GB	Reddit links	Mar-2023
Pushift.io [105]	-	Reddit links	Mar-2023
Wikipedia [106]	-	Wikipedia	Mar-2023
BigQuery [107]	-	Codes	Mar-2023
the Pile [108]	800GB	Other	Dec-2020
ROOTS [109]	1.6TB	Other	Jun-2022

预训练

预训练建立了 LLMs 的能力基础。通过对大规模语料库的预训练，LLMs 可以获得基本的语言理解和生成技能。在这个过程中，预训练语料库的规模和质量是 LLMs 获得强大能力的关键。此外，为了有效地预训练 LLMs，模型架构、加速方法和优化技术都需要精心设计。在第四节中，研究者首先在第 4.1 节讨论了数据的收集和处理，然后在第 4.2 节介绍了常用的模型架构，最后在第 4.3 节介绍了稳定和有效优化 LLMs 的训练技术。

数据收集

要开发一个强大的 LLM，从各种数据源中收集大量的自然语言语料至关重要。现有 LLMs 主要利用各种公共文本数据集作为预训练语料库。下图 2 列出了现有 LLMs 的预训练数据源分布。

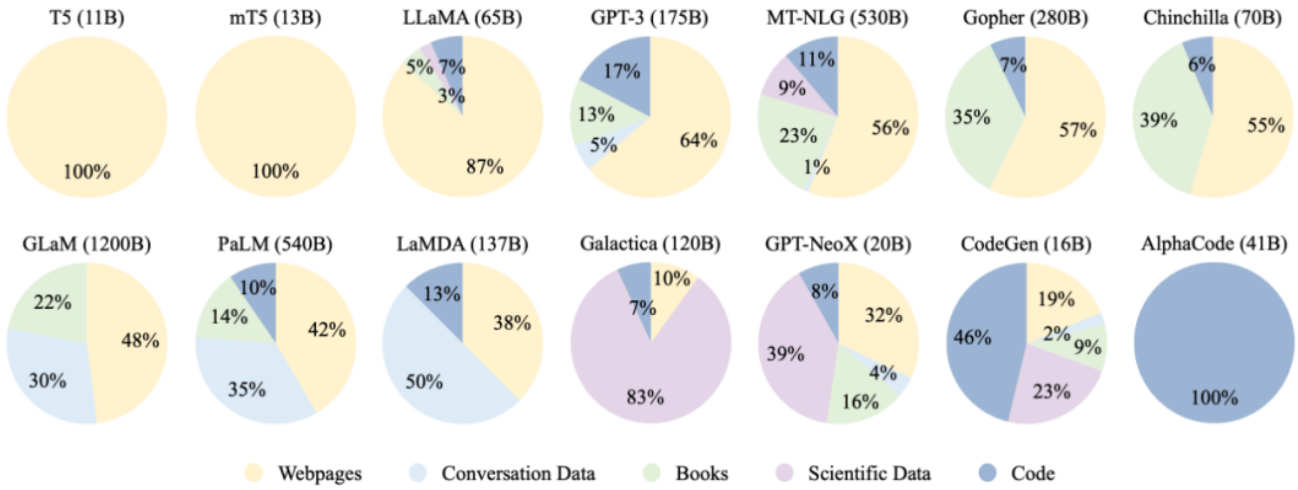


Fig. 2. Ratios of various data sources in the pre-training data for existing LLMs.

收集大量文本数据后，必须对它们进行预训练以构建预训练语料库，包括去噪、去冗余、去除不相关和潜在有毒的数据。下图 3 展示了为 LLMs 预训练数据的预处理 pipeline。

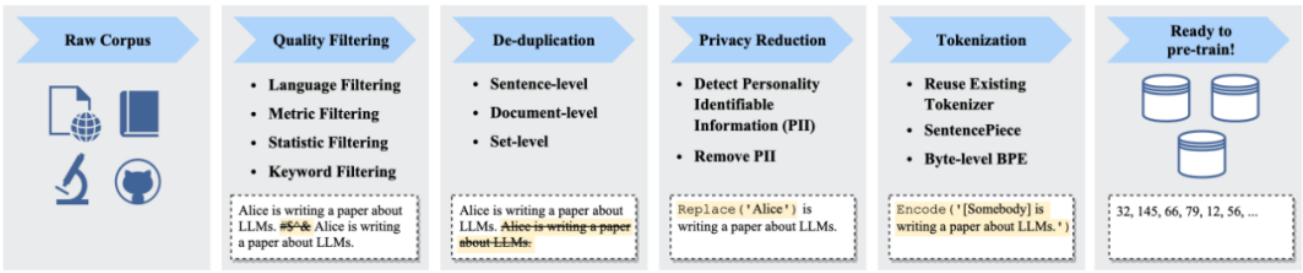


Fig. 3. An illustration of a typical data preprocessing pipeline for pre-training large language models.

架构

在本节中，研究者回顾了 LLMs 的架构设计，即主流架构，预训练目标和细节配置。下表 3 列出了几个具有代表性的 LLMs 的模型卡片以及公开的详细信息。

TABLE 3
Model cards of several selected LLMs with public configuration details. Here, PE denotes position embedding, #L denotes the number of layers, #H denotes the number of attention heads, d_{model} denotes the size of hidden states, and MCL denotes the maximum context length.

Model	Category	Size	Normalization	PE	Activation	Bias	#L	#H	d_{model}	MCL
GPT3 [55]	Casual decoder	175B	Pre Layer Norm	Learned	GeLU	✓	96	96	12288	2048
PanGU- α [73]	Casual decoder	207B	Pre Layer Norm	Learned	GeLU	✓	64	128	16384	1024
OPT [79]	Casual decoder	175B	Pre Layer Norm	Learned	ReLU	✓	96	96	12288	2048
PaLM [56]	Casual decoder	540B	Pre Layer Norm	RoPE	SwiGLU	×	118	48	18432	2048
BLOOM [66]	Casual decoder	176B	Pre Layer Norm	ALiBi	GeLU	✓	70	112	14336	2048
MT-NLG [90]	Casual decoder	530B	-	-	-	-	105	128	20480	2048
Gopher [59]	Casual decoder	280B	Pre RMS Norm	Relative	-	-	80	128	16384	-
Chinchilla [34]	Casual decoder	70B	Pre RMS Norm	Relative	-	-	80	64	8192	-
Galactica [35]	Casual decoder	120B	Pre Layer Norm	Learned	GeLU	×	96	80	10240	2048
LaMDA [85]	Casual decoder	137B	-	Relative	GeGLU	-	64	128	8192	-
Jurassic-1 [89]	Casual decoder	178B	Pre Layer Norm	Learned	GeLU	✓	76	96	13824	2048
LLaMA [57]	Casual decoder	65B	Pre RMS Norm	RoPE	SwiGLU	✓	80	64	8192	2048
GLM-130B [80]	Prefix decoder	130B	Post Deep Norm	RoPE	GeGLU	✓	70	96	12288	2048
T5 [71]	Encoder-decoder	11B	Pre RMS Norm	Relative	ReLU	×	24	128	1024	-

由于出色的并行化性和容量，Transformer 架构已成为开发各种 LLM 的 backbone，使得将语言模型扩展到数千亿个参数成为可能。一般来说，现有 LLMs 的主流架构大致可以分为三大类，即编码器 - 解码器、临时解码器和前缀解码器。

自 Transformer 出现以来，各种改进被相继提出以提高其训练稳定性，性能和计算效率。在这一部分中，研究者讨论了 Transformer 四个主要部分的相应配置，包括归一化、位置编码、激活函数、注意力机制和偏置。

预训练起着十分关键的作用，它将一般知识从大规模语料库编码到大规模模型参数中。对于训练 LLMs，有语言建模和去噪自编码两个常用的预训练任务。

模型训练

在这一部分中，研究者回顾了训练 LLMs 的重要设置，技术和训练 LLMs 技巧。

对于 LLMs 的参数优化，研究者提出了常用的批量训练、学习率、优化器和训练稳定性的设置。

随着模型和数据规模的增加，在有限的计算资源下有效地训练 LLMs 模型已经变得困难。特别是，需要解决两个主要技术问题，例如通过输入增加训练和将更大的模型加载到 GPU 内存中。这一部分回顾了现有工作中几种广泛使用的方法，以解决上述两个挑战，即 3D 并行、ZeRO 和混合精度训练，并就如何利用它们进行训练给出了建议。

LLMs 的适应性调优

经过预训练，LLMs 可以获得解决各种任务的通用能力。然而越来越多的研究表明，LLMs 的能力可以根据具体目标进一步调整。在第五节中，研究者详细介绍了调整预训练 LLMs 的两个主要方法，即指令调优（instruction tuning）和对齐调优（alignment tuning）。前一种方法主要是为了提高或解锁 LLMs 的能力，而后一种方法则是为了使 LLMs 的行为与人类的价值观或偏好一致。

指令调优

本质上，指令调优是在自然语言形式的格式化实例集合上微调预训练 LLMs 的方法，这与监督微调和多任务提示训练高度相关。为了执行指令调优，我们首先需要收集或构建指令格式的实例。然后，我们通常使用这些格式化实例以监督学习方式微调 LLMs（例如，使用序列到序列损失进行训练）。在指令调整后，LLMs 可以展示出卓越的能力，泛化出能解决未见任务的能力，即使在多语言环境中也是如此。

最近的一项调查对指令调优研究进行了系统的概述。相比之下，本文主要关注指令调优对 LLMs 的影响，并提供实例收集和调优的详细指南或策略。此外，本文还讨论了使用指令调优来满足用户的实际需求，这已广泛应用于现有的 LLMs，例如 InstructGPT 和 GPT-4。

格式化实例构建：通常，指令格式的实例由任务描述（称为指令）、输入输出对和少量演示（可选）组成。作为重要的公共资源，现有研究已经发布了大量以自然语言格式化的标记数据（参见表 5 中的可用资源列表）。接下来，本文将介绍构造格式化实例的两种主要方法（参见图 4 中的插图），然后讨论实例构造的几个关键因素。

指令调优策略：与预训练不同，指令调优通常更有效，因为只有适度数量的实例用于训练。虽然指令调优可以被认为是一个有监督的训练过程，但它的优化在几个方面与预训练不同，例如训练目标（即序列到序列损失）和优化配置（例如更小的批次） 大小和学习率），这在实践中需要特别注意。除了这些优化配置之外，指令调优还需要考虑两个重要方面：

- 平衡数据分布。
- 结合指令调优和预训练。

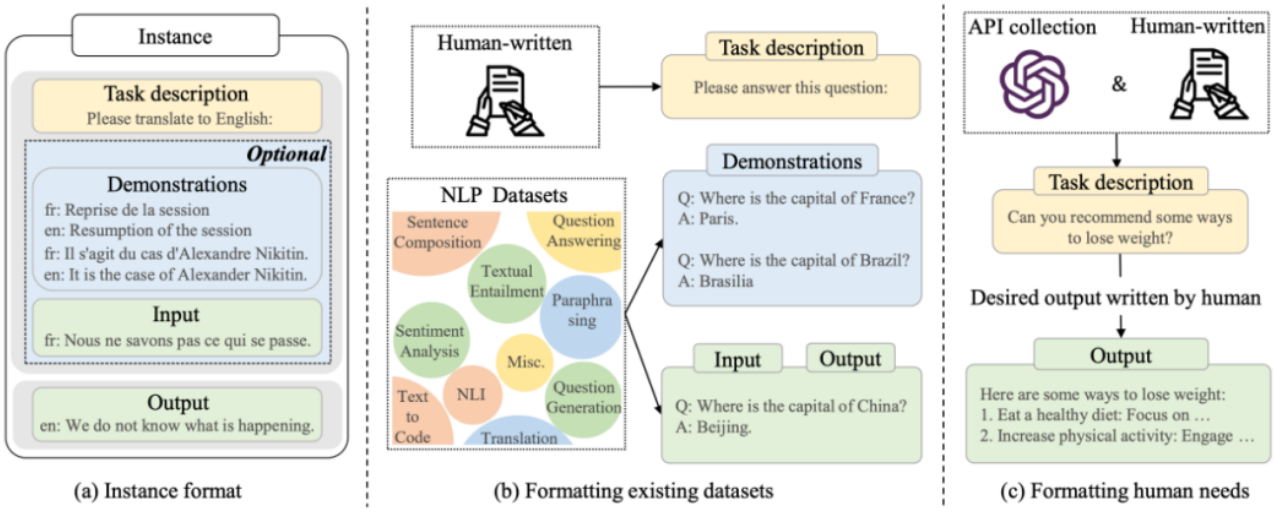


Fig. 4. An illustration of instance formatting and two different methods for constructing the instruction-formatted instances.

对齐调优

这部分首先介绍了对齐的背景及其定义和标准，然后重点介绍了用于对齐 LLMs 的人类反馈数

据的收集，最后讨论了用于对齐调整的人类反馈强化学习的关键技术。

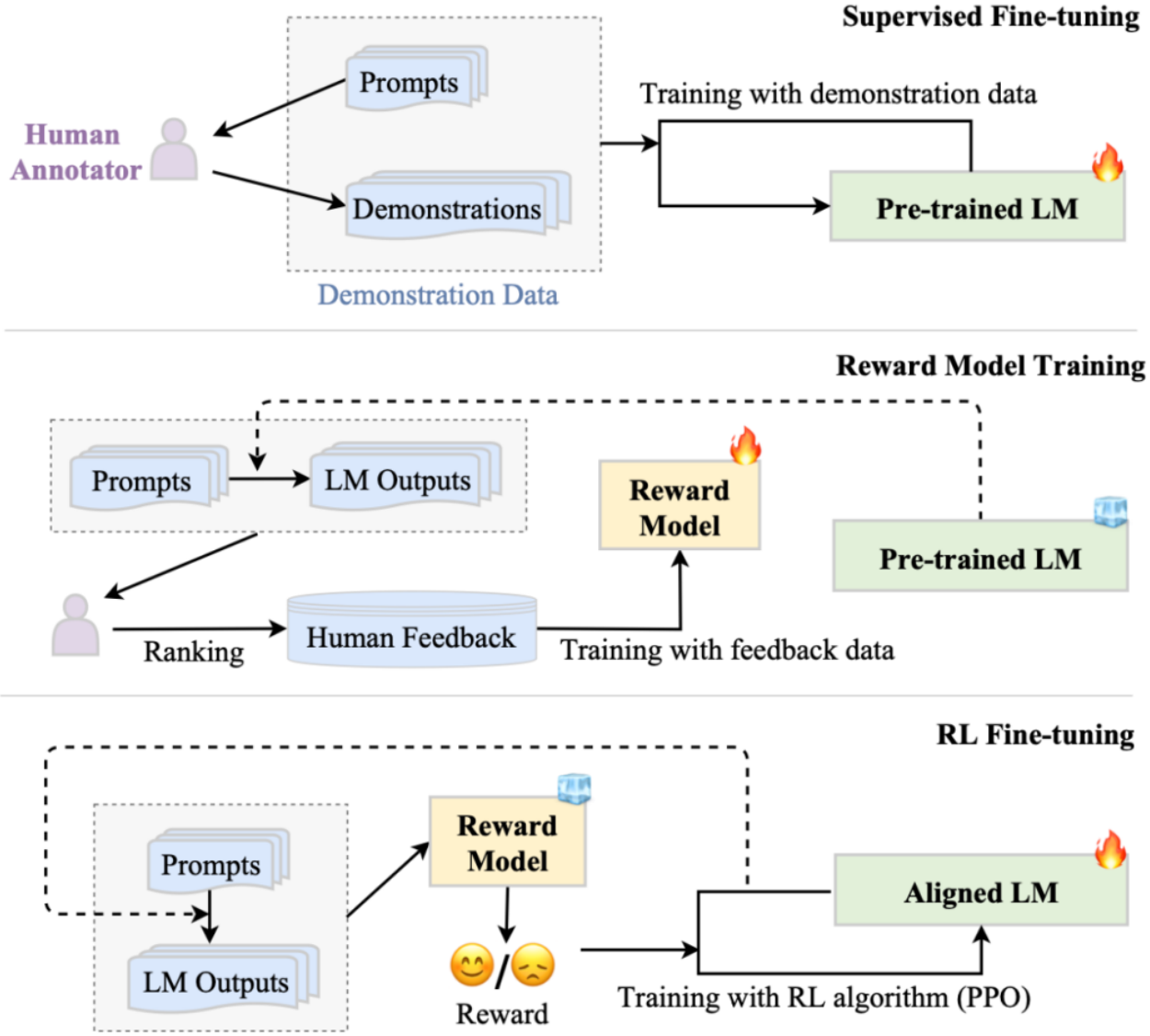


Fig. 5. The workflow of the RLHF algorithm.

使用

在预训练或适应性调整之后，使用 LLMs 的一个主要方法是为解决各种任务设计合适的 prompt 策略。一个典型的 prompt 方法是上下文学习（in-context learning），它以自然语言文本的形式制定了任务描述或演示。此外，思维链 prompting 方法可以通过将一系列中间推理步骤纳入 prompt 中来加强上下文学习。在第六节中，研究者详细介绍了这两种技术的细节。

上下文学习

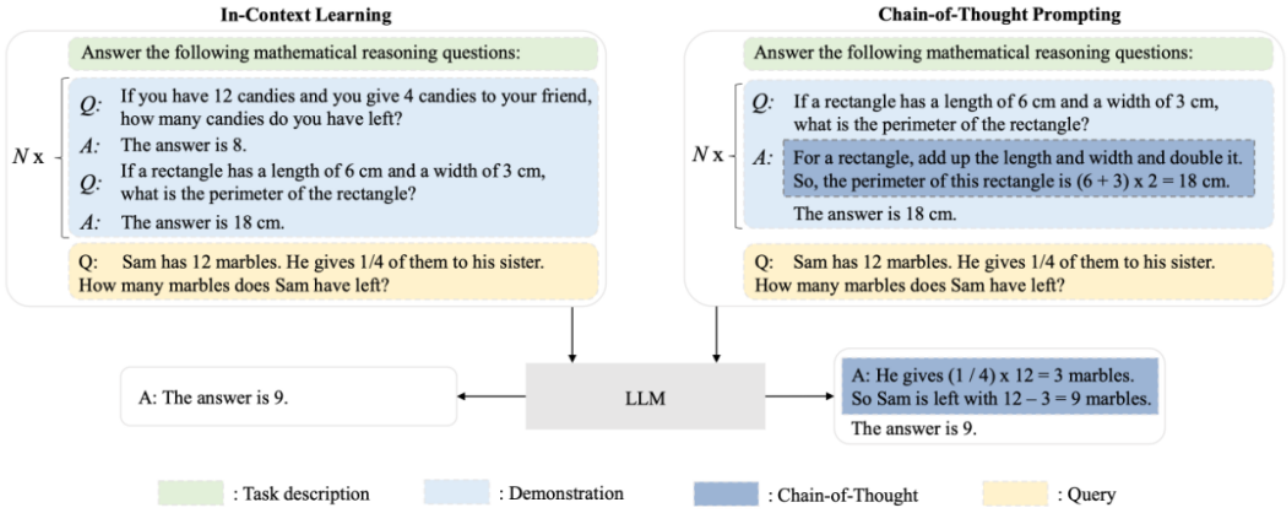


Fig. 6. A comparative illustration of in-context learning (ICL) and chain-of-thought (CoT) prompting. ICL prompts LLMs with a natural language description, several demonstrations, and a test query. While CoT prompting involves a series of intermediate reasoning steps in prompts.

作为一种特殊的 prompt 形式，上下文学习（ICL）是 GPT-3 首次提出的，它已经成为利用 LLMs 的一种典型方法。

思维链 prompt

思维链（CoT）是一种改进的 prompt 策略，可以提高 LLM 在复杂推理任务中的表现，如算术推理、常识推理和符号推理。CoT 不是像 ICL 那样简单地用输入 - 输出对来构建 prompt，而是将能够导致最终输出的中间推理步骤纳入 prompt。在第 6.2 节中，研究者详细说明了 CoT 与 ICL 的用法，并讨论 CoT 何时有效以及为何有效。

能力评估

为了研究 LLMs 的有效性和优越性，研究者利用了大量的任务和基准来进行实证评估和分析。第七节首先介绍了三种用于语言生成和理解的 LLMs 的基本评估任务，然后介绍几种具有更复杂设置或目标的 LLMs 的高级任务，最后讨论了现有的基准和实证分析。

基本评估任务

TABLE 6
Basic evaluation tasks and corresponding representative datasets of LLMs.

Task		Dataset
Language Generation	Language Modeling	Penn Treebank [262], WikiText-103 [263], the Pile [108], LAMBADA [147]
	Conditional Text Generation	WMT'14,16,19,20,21,22 [264–269], Flores-101 [270], DiaBLa [271], CNN/DailyMail [272], XSum [273], WikiLingua [274], OpenDialKG [275] SuperGLUE [276], MMLU [277], BIG-bench Hard [278], CLUE [279]
	Code Synthesis	APPS [280], HumanEval [87], MBPP [133], CodeContest [94], MTPB [76], DS-1000 [281], ODEX [282]
Knowledge Utilization	Closed-Book QA	Natural Questions [283], ARC [284], TruthfulQA [285], Web Questions [286], TriviaQA [287], PIQA [288], LC-quad2.0 [289], GrailQA [290], KQApro [291], CWQ [292], MKQA [293], ScienceQA [294]
	Open-Book QA	Natural Questions [283], OpenBookQA [295], ARC [284], Web Questions [286], TriviaQA [287], PIQA [288], MS MARCO [296], QASC [297], SQuAD [298], WikiMovies [299]
	Knowledge Completion	WikiFact [300], FB15k-237 [301], Freebase [302], WN18RR [303], WordNet [304], LAMA [305], YAGO3-10 [306], YAGO [307]
Complex Reasoning	Knowledge Reasoning	CSQA [240], StrategyQA [241], ARC [284], BoolQ [308], PIQA [288], SIQA [309], HellaSwag [310], WinoGrande [311], OpenBookQA [295], COPA [312], ScienceQA [294], proScript [313], ProPara [314], ExplaGraphs [315], ProofWriter [316], EntailmentBank [317], ProOntoQA [318]
	Symbolic Reasoning	CoinFlip [33], ReverseList [33], LastLetter [33], Boolean Assignment [319], Parity [319], Colored Object [320], Penguins in a Table [320], Repeat Copy [68], Object Counting [68]
	Mathematical Reasoning	MATH [277], GSM8k [237], SVAMP [238], MultiArith [321], ASDiv [239], MathQA [322], AQUA-RAT [323], MAWPS [324], DROP [325], NaturalProofs [326], PISA [327], miniF2F [328], ProofNet [329]

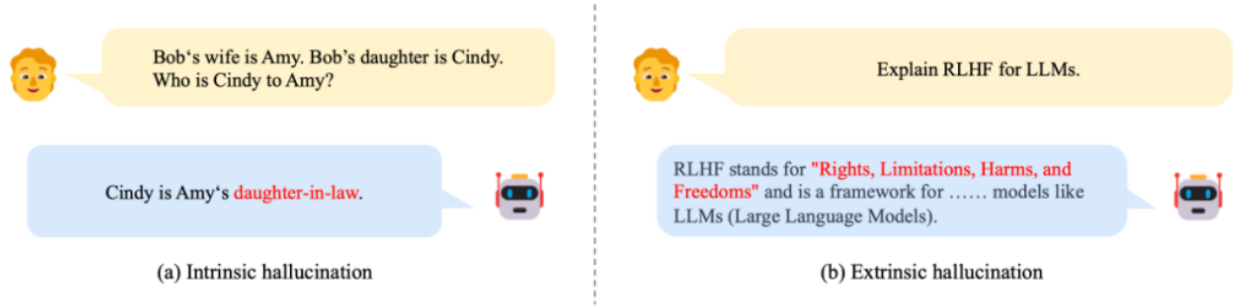


Fig. 7. Examples of intrinsic and extrinsic hallucination for a public LLM (access date: March 19, 2023). As an example of intrinsic hallucination, the LLM gives a conflicting judgment about the relationship between Cindy and Amy, which contradicts the input. For extrinsic hallucination, in this example, the LLM seems to have an incorrect understanding of the meaning of RLHF (reinforcement learning from human feedback), though it can correctly understand the meaning of LLMs (in this context).

图 7：一个公开 LLM 的内在和外在幻觉的例子（访问日期：2023 年 3 月 19 日）。作为内在幻觉的例子，LLM 对 Cindy 和 Amy 之间的关系给出了一个与输入相矛盾的判断。对于外在幻觉，在这个例子中，LLM 似乎对 RLHF（从人类反馈中强化学习）的含义有不正确的理解，尽管它能正确理解 LLM 的含义。

高级任务评估

除了上述基本评估任务，LLMs 还表现出一些高级能力，需要特别评估。在第 7.2 节中，研究者讨论了几个有代表性的高级能力和相应的评价方法，包括人工对齐、与外部环境的交互以及工具的操作。

总结与未来方向

在最后一节中，研究者总结了这次调查的讨论，并从以下几个方面介绍了 LLMs 的挑战和未来发展方向。

理论和原理：为了理解 LLM 的基本工作机制，最大的谜团之一是信息如何通过非常大的深度神经网络进行分配、组织和利用。揭示建立 LLMs 能力基础的基本原则或元素是很重要的。特别是，缩放似乎在提高 LLMs 的能力方面发挥了重要作用。已有研究表明，当语言模型的参数规模增加到一个临界点（如 10B）时，一些新兴能力会以一种意想不到的方式出现（性能的突然飞跃），典型的包括上下文学习、指令跟随和分步推理。这些「涌现」的能力令人着迷，但也令人困惑：LLMs 何时以及如何获得这些能力？最近的一些研究要么是进行广泛的体验，调查新兴能力的效果和这些能力的促成因素，要么是用现有的理论框架解释一些特定的能力。一个有见地的技术帖子将 GPT 系列模型作为目标也专门讨论了这个话题，然而仍然缺少更正式的理论和原则来理解、描述和解释 LLM 的能力或行为。由于涌现能力与自然界中的相变有着密切的相似性，跨学科的理论或原则（例如 LLMs 是否可以被视为某种复杂系统）可能对解释和理解 LLMs 的行为有帮助。这些基本问题值得研究界探索，对于开发下一代的 LLMs 很重要。

模型架构：由于可扩展性和有效性，由堆叠的多头自注意力层组成的 Transformer 已经成为构建 LLMs 的普遍架构。人们提出了各种策略来提高这个架构的性能，如神经网络配置和可扩展的并行训练（见 4.2.2 节讨论）。为了进一步提高模型的容量（如多轮对话能力），现有的 LLMs 通常保持较长的上下文长度，例如，GPT-4-32k 具有 32768 个 token 的超大上下文长度。因此，一个实际的考虑是减少标准的自注意力机制所产生的时间复杂性（原始的二次成本）。

此外，研究更高效的 Transformer 变体对构建 LLMs 的影响是很重要的，例如稀疏注意力已经被用于 GPT-3。灾难性遗忘也一直是神经网络的挑战，这也对 LLMs 产生了负面影响。当用新的数据调整 LLMs 时，原先学到的知识很可能被破坏，例如根据一些特定的任务对 LLMs 进行微调会影响它们的通用能力。当 LLMs 与人类的价值观相一致时，也会出现类似的情况，这被称为对齐税（alignment tax）。因此有必要考虑用更灵活的机制或模块来扩展现有的架构，以有效支持数据更新和任务专业化。

模型训练：在实践中，由于巨大的计算量以及对数据质量和训练技巧的敏感性，预训练可用的 LLMs 非常困难。因此，考虑到模型有效性、效率优化和训练稳定性等因素，开发更系统、更经济的预训练方法来优化 LLMs 变得尤为重要。开发更多的模型检查或性能诊断方法（例如 GPT-4 中的可预测缩放），便于在训练中发现早期的异常问题。此外，它还要求有更灵活的硬件支持或资源调度机制，以便更好地组织和利用计算集群中的资源。由于从头开始预训练 LLMs 的成本很高，因此必须设计一个合适的机制，根据公开的模型检查点（例如 LLaMA 和 Flan-T5）不断地预训练或微调 LLMs。为此，必须解决一些技术问题，包括数据不一致、灾难性遗忘和任务专业化。到目前为止，仍然缺乏具有完整的预处理和训练日志（例如准备预训练数据的脚本）的开源模型检查点以供重现的 LLM。为 LLMs 的研究提供更多的开源模型将是非常有价值的。此外，开发更多的改进调整策略和研究有效激发模型能力的机制也很重要。

模型的使用：由于微调在实际应用中的成本很高，prompt 已经成为使用 LLMs 的突出方法。通过将任务描述和演示例子结合到 prompt 中，上下文学习（prompt 的一种特殊形式）赋予了 LLMs 在新任务上良好的表现，甚至在某些情况下超过了全数据微调模型。此外，为了提高复

杂推理的能力，人们提出了先进的 prompt 技术，例如思维链（CoT）策略，它将中间的推理步骤纳入 prompt。然而，现有的 prompt 方法仍然有以下几个不足之处。首先，它在设计 prompt 时需要大量的人力，因此为解决各种任务而自动生成有效的 prompt 将非常有用；其次，一些复杂的任务（如形式证明和数字计算）需要特定的知识或逻辑规则，而这些知识或规则可能无法用自然语言描述或用例子来证明，因此开发信息量更大、更灵活的任务格式化的 prompt 方法很重要；第三，现有的 prompt 策略主要集中在单圈的表现上，因此开发用于解决复杂任务的交互式 prompt 机制（如通过自然语言对话）非常有用，ChatGPT 已经证明了这一点。

安全和对齐：尽管 LLMs 具备相当的能力，但它的安全问题与小型语言模型相似。例如，LLMs 表现出产生幻觉文本的倾向，比如那些看似合理但可能与事实不符的文本。更糟糕的是，LLMs 可能被有意的指令激发，为恶意的系统产生有害的、有偏见的或有毒的文本，导致滥用的潜在风险。要详细讨论 LLMs 的其他安全问题（如隐私、过度依赖、虚假信息和影响操作），读者可以参考 GPT-3/4 技术报告。作为避免这些问题的主要方法，来自人类反馈的强化学习（RLHF）已被广泛使用，它将人类纳入训练循环，以发展良好的 LLMs。为了提高模型的安全性，在 RLHF 过程中加入安全相关的 prompt 也很重要，如 GPT-4 所示。然而，RLHF 在很大程度上依赖于专业标签人员的高质量的人类反馈数据，使得它很难在实践中得到正确的实施。因此，有必要改进 RLHF 框架，以减少人类标签员的工作，并寻求一种更有效的注释方法，保证数据质量，例如可以采用 LLMs 来协助标注工作。最近，红色团队被采用来提高 LLMs 的模型安全性，它利用收集的对抗性 prompt 来完善 LLMs（即避免红色团队的攻击）。此外，通过与人类交流建立 LLMs 的学习机制也很有意义，人类通过聊天给出的反馈可以直接被 LLMs 利用来进行自我完善。

应用和生态系统：由于 LLMs 在解决各种任务方面表现出强大的能力，它们可以被应用于广泛的现实世界的应用（例如，遵循特定的自然语言指令）。作为一个显著的进步，ChatGPT 已经潜在地改变了人类获取信息的方式，这带来了新必应的发布。在不久的将来，可以预见，LLMs 将对信息搜索技术产生重大影响，包括搜索引擎和识别系统。

此外，随着 LLMs 的技术升级，智能信息助理的发展和使用将得到极大的促进。在更广泛的范围内，这一波技术创新倾向于建立一个由 LLMs 授权的应用程序的生态系统（例如，ChatGPT 对插件的支持），这将与人类生活密切相关。最后，LLMs 的崛起为通用人工智能（AGI）的探索提供了启示。它有希望开发出比以往更多的智能系统（可能有多模态信号）。同时，在这个发展过程中，人工智能的安全性应该是首要关注的问题之一，也就是说，让人工智能为人类带来好处而不是坏处。

公众号后台回复“CVPR2023”获取最新论文分类整理资源

极市干货

极视角动态：「无人机+AI」光伏智能巡检，硬核实力遇见智慧大脑！ | 「AI 警卫员」上线，极视角守护龙大食品厂区安全！ | 点亮海运指明灯，极视角为海上运输船员安全管理保驾护航！

CVPR2023：CVPR'23 最新 125 篇论文分方向整理 | 检测、分割、人脸、视频处理、医学影像、神经网络结构、小样本学习等方向

数据集：自动驾驶方向开源数据集资源汇总 | 医学影像方向开源数据集资源汇总 | 卫星图像公开数据集资源汇总