·专题:ChatGPT 与人工智能技术应用·

以 ChatGPT 为代表的大型语言模型研究进展

柯 沛[1] 雷文强[2] 黄民烈[1*]

1. 清华大学 计算机科学与技术系,北京 100084
2. 四川大学 计算机学院,成都 610065

[摘 要] 大型语言模型是当今人工智能领域最前沿的研究方向之一,该方向旨在训练含有大规模参数的通用语言模型,使其能够遵循人类指令完成不同类型的自然语言处理任务。作为大型语言模型的代表,由 OpenAI 研发的 ChatGPT 在各个领域均展现出强大的自然语言生成能力,受到了全球各行各业的关注。本文从语言模型的发展历程出发,介绍了近年研究者在扩大语言模型规模上的探索,然后分析了大型语言模型带来的范式改变,并以 ChatGPT 为典型实例概述了其发展、技术和应用,接着介绍了后 ChatGPT 时代大型语言模型的前沿进展,最后从评价和治理两方面总结了目前大型语言模型的局限性及未来需要解决的挑战。

[关键词] 大型语言模型;ChatGPT;预训练语言模型;Transformer;思维链;自然语言处理;人工智能

1 语言模型的发展历程

语言是人类智能的重要组成部分,各种日常活动(如思维表达、人际交往)都离不开对语言的使用。在人工智能领域,自然语言处理研究如何让计算机处理并应用人类语言,包含对话系统、机器翻译、情感分析等广泛的应用场景。在 20 世纪 90 年代对统计机器翻译[1]和语音识别[2]的研究中,研究者们在对实际应用问题进行建模时发现正确估计由 n 个词 $w_1,w_2,\cdots,w_n$ 组成的序列的联合概率 $P(w_1,w_2,\cdots,w_n)$ 非常重要,并将估计词序列概率的模型称为语言模型。由于词序列的联合概率过于复杂,使用条件概率公式对其进行分解可得:

$$P(w_1,w_2,\cdots,w_n)=P(w_1)\cdot\prod_{i=2}^{n}P(w_i\mid w_1,\cdots,w_{i-1}) \tag{1}$$

其中 $w_i(1\leqslant i\leqslant n)$ 指第 i 个词,等式右侧的词条件概率 $P(w_i\mid w_1,\cdots,w_{i-1})$ 便是语言模型的建模对象。

语言模型一经提出便得到广泛关注[3],其发展历程如图 1 所示。早期,研究者们基于马尔可夫假设使用 N-gram 等方法统计词条件概率以构建统计语言模型[4],但这类模型泛化性差且效率低。随着深度学习[5]的发展,研究者们设计各种神经网络拟

黄民烈 清华大学长聘教授,国家杰出青年科学基金获得者,中国中文信息学会自然语言生成与智能写作专业委员会副主任,CCF 学术工作委员会秘书长。主要研究领域为大规模语言模型、对话系统、语言生成等。曾获得中国人工智能学会吴文俊人工智能科技进步奖一等奖(第一完成人)、中国中文信息学会汉王青年创新奖等。在国际顶级会议和期刊上发表论文 150 多篇,多次获得国际主流会议的最佳论文或提名奖。

柯沛 助理研究员,清华大学计算机科学与技术系博士后,中国中文信息学会自然语言生成与智能写作专委会的委员。研究方向为自然语言生成和对话系统。在 *Annual Meeting of the Association for Computational Linguistics*、*Conference on Empirical Methods in Natural Language Processing* 等国际顶级学术会议上共发表论文 10 余篇,曾获 NLPCC 2020 最佳学生论文奖。

收稿日期:2023-06-30;修回日期:2023-08-06

* 通信作者,Email: aihuang@tsinghua.edu.cn

本文受到国家自然科学基金项目(62125604,61936010)的资助。

合词条件概率函数以构建神经语言模型。基于多层感知机(Multi-Layer Perceptron，MLP)的语言模型[6，7]对词进行低维嵌入表示并用于拟合概率函数，在泛化性与效率方面有所提升，但对词的序列信息建模不充分。基于循环神经网络(Recurrent Neural Network，RNN)的语言模型[8，9]串联处理词序列并使用记忆单元储存序列信息，从而更充分地建模序列性。但串联方式对长文本的处理效率低，计算代价较大。之后，谷歌提出了完全基于注意力机制的模型 Transformer[10]，大幅提升了模型对序列的并行处理性能。性能局限的突破让研究者们能用大量数据训练模型，因而基于 Transformer 的预训练语言模型(Pre-trained Language Model)[11，12]迅速兴起。研究者们遵循规模化原则(Scaling Law)持续对如何扩大语言模型的规模进行探索。近年来，研究者们成功将预训练语言模型的参数规模由早期的亿/十亿量级扩展至百亿/千亿量级，后者又被称为大型语言模型(Large Language Model)。和早期的预训练语言模型相比，大型语言模型具有强大的通用语言理解和生成能力，能够在少样本甚至零样本的场景下达到多种自然语言处理任务的最优性能。

2　大型语言模型的规模化探索

为了构建大型语言模型，研究者们在语言模型的规模化方向做了许多有价值的探索，为后续大型语言模型的发展奠定了坚实的基础。本节将首先介绍规模化原则，然后概述研究者们经探索得到的语言模型规模化过程中需要具备的基础。

2.1　语言模型的规模化原则

深度学习通常用神经网络表示数据复杂特征以应用至各类任务[5]。规模化原则经验性地指出：通过增加模型参数量与训练数据量，深层神经网络能更好地表征数据特征[13]，并在各类任务上取得性能提升[14]。研究者们遵循规模化原则探索如何构建大型语言模型，并通过实验得到了其在训练数据、模型结构、训练目标、训练方法与技巧等方面需要具备的基础[15]。

2.2　语言模型的规模化基础

2.2.1　训练数据

可公开访问的自然语言数据种类多样，包括网页数据(如 Common Crawl[16])、代码数据(如 GitHub)、文本数据(如英文数据集 The Pile[17]和中文数据集 Wudao Corpora[18])等。它们经过精细处理后能成为大规模、高质量且形式丰富的语言数据，为大型语言模型的规模化奠定数据基础。

2.2.2　模型结构

循环神经网络无法并行地处理输入词序列[9]，计算代价限制了语言模型规模化的发展。2017 年，谷歌提出了新的神经网络结构 Transformer[10]。该网络使用注意力机制并行地对输入词序列进行全局依赖分析与特征表示，在计算效率以及性能上均有大幅提升，能更高效地处理更长的文本序列。因此，Transformer 成为了大型语言模型的模型结构基础[19]。

2.2.3　训练目标

语言模型采用自监督的训练目标，无需人工标注即可在大量数据上进行训练。其中，仅包含编码器的掩码式语言模型(如 BERT[11])随机地遮掩输入序列的词，并用未遮掩词预测被遮掩词的生成概率。它将词条件概率中的条件从前序词扩展到了前后序双向词，词的特征信息更加丰富。仅包含解码器的单向语言模型(如 GPT[12])根据前序词从左往右地预测下一个词，它能很好地完成各种语言生成任务。除了上述常用目标外，研究者还提出了包含编码器—解码器的去噪语言模型(如 BART[19])，模型重构含有多种噪声(如随机删除词)的输入序列。这些训练目标的核心是使用上下文预测目标词的概率进而实现对词条件概率函数的估计。对于不同的训练目标，研究者使用 Transformer 构建了不同的

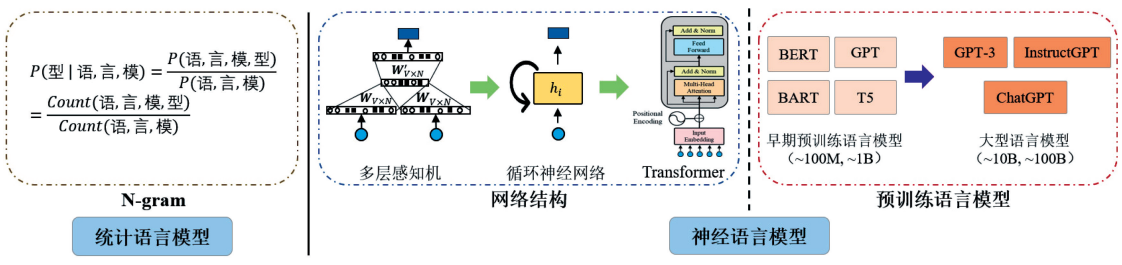

图 1　语言模型的发展历程

注：语言模型的发展经历了从统计语言模型到神经语言模型的过程。对于神经语言模型，其网络结构以及预训练规模有不同的发展方向。

模型架构。图 2 总结了上述三种预训练模型中常见的训练目标与相应的模型结构，其中[MASK]指掩码符。

2.2.4 训练方法与技巧

随着模型参数量与训练数据量的增加，模型的收敛性与训练效率成为挑战。残差连接(Residual Connection)和丢弃(Dropout)技巧是搭建深层神经网络的必要部分[10]。同时，研究者们改进激活函数[20]与归一化层[21]以提升训练的稳定性与效率。训练框架和计算硬件的发展也提升了语言模型规模化的可行性。

3 大型语言模型带来的新范式

由于大型语言模型的参数量较大，训练所需的计算资源成本和时间成本均较高，所以预训练语言模型的经典范式“预训练＋微调”难以直接适用于大型语言模型。本节将首先介绍预训练语言模型的经典范式及其在大型语言模型上面临的挑战，然后概述大型语言模型带来的以提示学习为核心的新范式。

3.1 经典范式：预训练＋微调

早期基于神经网络的语言模型大多在任务相关的标注数据上从头开始训练，这使模型对标注数据质量和数量的依赖程度较高，并且在各类自然语言处理任务之间的可迁移性较差。随着预训练语言模型(如 GPT[12] 和 BERT[11])的兴起，“预训练(Pre-training)＋微调(Fine-tuning)”已经成为自然语言处理领域的基础范式之一。该范式的核心思想是先利用无监督训练方法在大规模无标注文本语料上得到通用预训练模型，再结合具体下游任务的标注数据来微调模型以提升其在相应任务上的性能。该范式的优势在于能够充分利用无标注文本数据提升模型的通用语言表示能力，从而改善通用预训练模型的可迁移性，使其在各类语言理解和生成任务(尤其是标注数据较少的低资源任务)上均达到较好的性能。

然而，随着语言模型的规模逐渐扩大，上述“预训练＋微调”的范式在实际应用中存在众多挑战。首先，传统的微调方法需要训练语言模型的所有参数，当模型参数量增大时，微调所需的计算资源和时间成本将迅速提升；其次，将具有大规模参数的预训练模型在少量标注数据上微调容易导致过拟合，致使模型性能受到影响。

3.2 基于大型语言模型的新范式：提示学习

为了充分利用大型语言模型的特点，避免训练所有模型参数，近年来自然语言处理领域的研究开始关注以提示学习(Prompt Learning)为核心的新范式[22]，如图 3 所示。提示通常指语言模型输入中

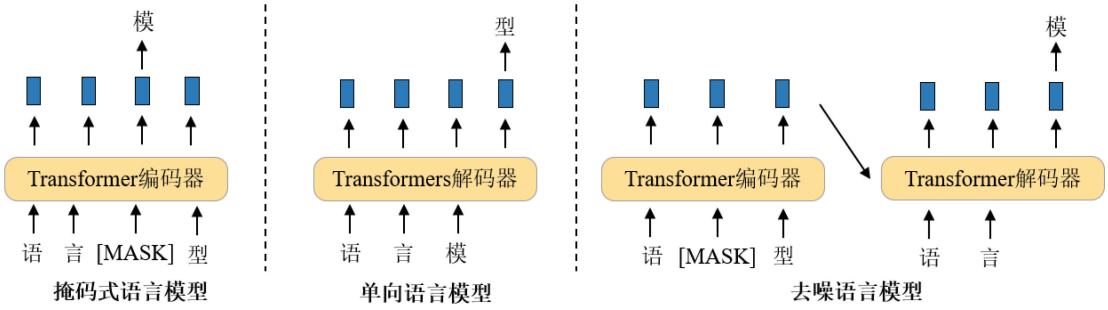

图 2 预训练语言模型的常见模型结构与训练目标

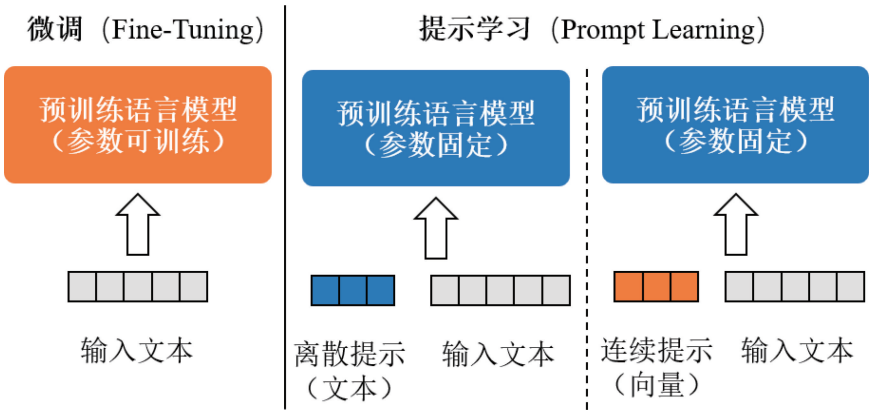

图 3 基于微调的经典范式和基于提示学习的新范式

添加的信息，用于调整原始输入数据的形式，使其更接近于预训练阶段的数据形式。提示学习通过缩小预训练与微调阶段的数据形式差异，使模型能够高效地应用于下游任务，在某些情况下无需训练参数或者仅在下游任务数据上训练少量参数就能取得很好的性能。提示按照其形式可分为离散提示(Discrete Prompt)[23]和连续提示(Continuous Prompt)[24]，它们在大型语言模型中均有广泛的应用。

3.2.1　离散提示

离散提示的形式通常为自然语言文本，这些提示文本能够辅助预训练模型理解下游任务，从而提升其在相应任务上的性能。例如在情感分类任务中，对于待分类的文本“I love this movie.”，可通过添加提示文本将模型输入转化为“I love this movie. It was [MASK].”，然后通过模型生成标签词(如great 和 bad)的概率来确定分类结果。

离散提示的研究工作主要分为两条路线：(1)如何自动搜索并构建最优的离散提示；(2)如何利用人工设计的离散提示激发大型语言模型的能力。针对离散提示的自动搜索和构建，研究者们提出了梯度搜索[25]等方法，旨在寻找不同下游任务对应的最优离散提示。其他工作将离散提示的生成问题转化为传统的自然语言生成问题，利用预训练模型构建提示生成器并最终生成高质量的提示文本[26]。另一类工作通过人工设计离散提示的形式来尝试激发大型语言模型的能力，其中两种典型能力为上下文学习(In-Context Learning)[27]和思维链(Chain-of-Thought)[28]。上下文学习指模型通过学习提示中包含的样例输入和输出的对齐关系，以完成相应任务的能力。该能力使大型语言模型在面对少样本自然语言处理任务时，可以在不更新任何模型参数的前提下直接求解，只需将任务的少量标注样本以一定的格式放入离散提示中即可。思维链指模型在完成推理任务时先生成推理过程再生成答案的能力。研究者们探索了两种激发大型语言模型思维链能力的提示设计：第一种是少样本设定，即在提示中加入“问题—推理过程—答案”的样例数据，然后利用模型的上下文学习能力来促进推理过程的生成[28]；第二种是零样本设定，即在提示中直接加入引导模型生成推理过程的文本，比如“Let's think step by step.”，从而使模型能够先生成推理过程再给出答案[29]。这两种能力是大模型区别于小模型的重要表现，所以又被称为涌现能力(Emergent Ability)[30]。

3.2.2　连续提示

连续提示指提示信息为不受词向量约束的连续向量，而不再是词表中已有的词。因此，连续提示通常含有原始模型以外的少量可训练参数，能够在下游任务的训练中进行更新。研究者们提出提示微调(Prompt Tuning)[24]方法，通过设计连续提示的插入位置和训练目标来提升模型在下游任务上的性能，如 Prefix Tuning[31]。这类方法大多固定语言模型的原始参数，仅更新连续提示含有的参数，显著降低了计算资源的开销。还有研究者探究了提示微调和低秩适配(Low-Rank Adaption，LoRA)[32]等其他参数高效微调方法的联系，并提出了统一视角来解读这些方法，使其能够更好地应用至大型语言模型的训练[33]。

4　大型语言模型的典型实例：ChatGPT

4.1　ChatGPT 的发展历程

ChatGPT[34]是 OpenAI 于 2022 年 11 月推出的大型语言模型，因其在各类自然语言处理任务上表现出的通用生成能力而吸引了全球各行各业的关注。ChatGPT 可以看作是 OpenAI 在 2020 年推出 1 750 亿参数的 GPT-3[27]模型后，通过持续迭代而得到的里程碑式产品，其发展历程如图 4 所示，其中 RLHF 指基于人类反馈的强化学习。GPT-3 到 ChatGPT 的迭代过程包含两个重要模型，即 CodeX[35]和 InstructGPT[36]。CodeX 在 GPT-3 的基础上使用代码数据继续进行训练，使模型具备代码理解和生成能力。由于代码数据的长度普遍较长，并且不同代码段之间的逻辑性很强，所以在大型语言模型的训练中加入代码数据可能对模型的

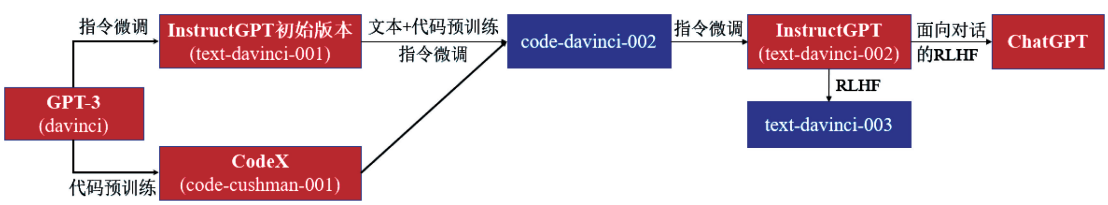

图 4　从 GPT-3 到 ChatGPT 的发展历程

长文本理解能力、推理能力有一定的提升作用。InstructGPT 则引入了指令微调(Instruction Tuning)[37]和基于人类反馈的强化学习(Reinforcement Learning from Human Feedback,RLHF)[38]这两项关键技术,将模型输出对齐至人类偏好,从而进一步提升了模型在面向用户的真实场景中的性能。这两项技术同样也是 ChatGPT 成功的核心技术,将在本节详述。

4.2 ChatGPT 的核心技术

尽管大型语言模型(如 GPT-3[27])已具备较强的自然语言生成能力,但其仍会生成不符合人类期望的文本,包括有偏见的内容、虚假内容以及与人类指令有冲突的内容,这也是限制大型语言模型应用落地的重要瓶颈。出现这种现象的原因是语言模型的训练目标(即在文本数据上根据前文预测下一个词)与"按照人类指令生成文本"的目标并不完全相同。在传统对话系统领域,理解并识别用户的指令和意图本身就是一个经典研究方向,其对提升对话系统的交互性非常重要。但大型语言模型往往面对的是开放端生成任务,即需要理解各类开放指令并生成相应的结果。因此,现有对话系统领域的研究工作难以直接迁移至大型语言模型。为此,OpenAI 提出了一套方法将模型输出对齐至人类偏好,该方法主要包含指令微调和基于人类反馈的强化学习这两项关键技术。

4.2.1 指令微调

指令微调指利用人工标注的指令数据进行有监督训练。OpenAI 的研究者首先雇用了标注者提供模型输入信息,并结合 ChatGPT 的 API 使用者提供的输入信息构成了整个指令微调数据集的输入。这些输入文本的质量和多样性都很高,包含生成、问答、对话、摘要、抽取等各类自然语言处理任务。OpenAI 的研究者让标注者根据这些输入信息理解其背后的意图,并撰写相应的回复,同时考虑回复的真实性以及回复中可能存在的偏置。由此便可得到指令微调阶段符合人类偏好的高质量标注数据,可直接在该数据集上训练大型语言模型。

4.2.2 基于人类反馈的强化学习

为了进一步将模型输出对齐至人类偏好,使模型生成符合人类偏好的文本,OpenAI 的研究者收集了比较数据集用于训练奖励模型(Reward Model)。具体而言,针对给定的输入信息以及模型生成的两个结果,标注者需要从中选择更符合人类偏好的生成结果。由此,这个含人类反馈的比较数据集便可用于训练奖励模型,使其能够预测符合人类偏好的生成结果。利用该奖励模型即可通过近端策略优化(Proximal Policy Optimization,PPO)[39]等强化学习算法来继续训练经过指令微调后的策略模型(即生成模型),使其生成奖励分数更高的结果,最终达到对齐目标。

4.3 ChatGPT 的应用场景

ChatGPT 作为近期大型语言模型最成功的实例之一,已应用至各类真实场景,其中最具有代表性的场景是通用对话助理和搜索引擎。由于 ChatGPT 具有极强的通用生成能力,能够遵从用户指令完成各类自然语言处理任务,所以其已成为众多个人用户的通用对话助理,辅助用户完成语法改错、文案撰写、代码调试等工作。同时,微软还将 ChatGPT 与传统搜索引擎 Bing 结合,研发出对话式搜索引擎 New Bing。该搜索引擎可以针对用户查询检索相关网页,并将网页信息整合为含网页引用的文段输出给用户,显著提升了用户的检索效率。

5 后 ChatGPT 时代大型语言模型的前沿进展

5.1 开源大型语言模型

尽管 ChatGPT 在各类自然语言生成任务上均展现出巨大的潜力,但 OpenAI 没有开源其代码和模型参数,这使得学术界和工业界无法复现其生成效果,从而严重阻碍了大型语言模型的研究进程。为此,国内外研究者研发并开源了多个大型语言模型,为深入理解其工作机理打下了坚实的基础。其中,最受关注的开源模型是由 Meta 提出的 LLaMA[40],该模型采用 Transformer 解码器的网络结构,参数规模从 7 B 到 70 B 不等。它和相同参数规模的其他开源模型相比具有更强的生成性能,所以多支研究团队在 LLaMA 的基础上通过指令微调来构建遵循人类指令的生成模型,包括斯坦福大学提出的 Alpaca[41]和加州大学伯克利分校提出的 Vicuna[42]等。国内学术界和工业界同样有不少团队推出了中文开源大型语言模型,例如清华大学提出的 ChatGLM[43]、复旦大学提出的 MOSS[44]、阿里巴巴提出的通义千问、上海人工智能实验室提出的 InternLM[45]等。

5.2 工具学习

由于大型语言模型仍是在文本数据上训练得到的语言模型,所以其无法处理许多复杂任务,例如需

要检索实时信息的任务、需要专业知识（如数学知识）的任务以及涉及到语音、图像、视频等其他模态信息的任务等。这些任务本身已经有一些成熟的工具，例如搜索引擎、数学计算工具（如 Mathematica）以及其他模态的理解与生成模型。因此，研究者尝试探索大型语言模型的工具学习能力，使其根据用户指令来规划任务的完成步骤，并在相应步骤调用已有工具进行处理，最终完成原始任务。研究者对特定领域的工具调用进行了深入研究（如涉及多模态工具调用的 Visual ChatGPT[46]、HuggingGPT[47]等），尝试探索大型语言模型使用大规模数量级工具的能力（如 Gorilla[48]、ToolLLM[49]等），并分析了大型语言模型在工具学习中存在的不足（如 ToolBench[50]）。而在工业界，OpenAI 也为 ChatGPT 增加了工具调用功能，将其和现有的插件进行对接以满足用户提出的复杂需求。

5.3　长文本建模

现在主流的大型语言模型仍采用 Transformer 作为基本架构，而 Transformer 中的全连接注意力机制的计算复杂度为 $O(L^2)$，其中 L 为文本长度，这使其无法支持过长的输入和输出文本，从而限制了长文本的理解和生成能力。研究者们尝试通过优化显存读写速度（如 Flash Attention[51]）、注意力机制计算方式（如 RetNet[52]）来提升模型训练的并行度以及推断效率。由于修改模型结构往往涉及到重新训练大型语言模型，所以研究者们还尝试在微调阶段通过位置插值（Position Interpretation）[53]来调整位置编码向量以扩展输入文本的长度。

5.4　反馈学习

在 OpenAI 成功运用基于人类反馈的强化学习大幅提升 ChatGPT 的生成性能后，反馈学习成为了大型语言模型领域的关注热点。然而，由于强化学习在训练语言生成模型时非常不稳定，所以研究者发现在引入人类反馈信息后并不能显著提升模型的生成性能。为此，研究者们提出了一系列训练稳定的引入人类反馈的方法来替代强化学习，包括将反馈结果转化为微调时使用的文本数据（如 Chain-of-Hindsight[54]）、设计排序损失函数来引入反馈结果（如 RRHF[55]）等，从而直接使用监督学习的方式便可提升模型性能。

6　大型语言模型的局限性

6.1　大型语言模型的评价

随着大型语言模型的快速发展，机器生成文本的质量逐渐接近人类水平，这为大型语言模型的评价带来了新的挑战[56]。近期研究工作主要围绕评价指标和评价数据集两方面展开。在评价指标方面，由于经典指标（如 BLEU[57]、ROUGE[58]）在衡量大型语言模型的生成文本时与人工评价的相关性低，所以近期工作尝试借助 ChatGPT、GPT-4[59]等当前性能最好的大型语言模型来结合离散提示信息评价生成文本质量，但该类方法同样会引入模型自身的偏置从而导致评价结果不准确[60]。而在评价数据集方面，近期工作提出的数据集可大致分为两类，即以考试题为核心的问答任务[61]和面向用户的通用生成任务[42]。前者以选择题的形式测试模型性能，通过客观的准确率反映模型表现；后者则通过用户主观打分来确定不同模型之间的优劣。受限于评价数据集的话题覆盖面和任务形式，这两类数据集仍无法全面且准确地反映大型语言模型的性能。未来大型语言模型将逐渐成为基础设施走进人类生活的方方面面，其可靠评价方法的重要性将越发凸显，因此评价是该领域急需突破的技术难点。

6.2　大型语言模型的治理

6.2.1　安全性

安全性是大型语言模型面临的重要挑战之一。以 ChatGPT 为代表的通用生成模型能够应用于各个学科领域的任务，但同时这也就意味着该类模型面临广泛的内容安全问题[62]。尽管 OpenAI 已经通过基于人类反馈的强化学习等多种方法，尝试将模型输出对齐至人类价值观，但语言模型在应用至各个领域时仍容易被恶意使用，从而生成偏见言论、煽动性言论、隐私侵犯言论等不安全的文本。近期研究工作尝试从数据和方法两个层面来提升模型的安全性，总体思路是通过人工标注或模型生成的方法获得不同类型的攻击性言论或偏见言论的数据[63]，然后通过设计攻防算法来提升模型的安全性[64]。大型语言模型的安全性是其治理中最重要的问题之一，备受各国政府的关注。因此，理解并修补大型语言模型的安全缺陷对未来相关政策法规的拟定和实施也具有重要意义。

6.2.2　可信度

可信度是目前大型语言模型的重要局限之一。尽管以 ChatGPT 为代表的通用生成模型可用于解决各类真实场景中的问题，但其仍会生成不可信的文本。例如，模型可能会生成文本来描述现实中并不存在的情况，即幻觉（Hallucination）[65]；还可能会

生成与常识或专业知识有冲突的文本。同时，模型在解决涉及推理的问题（如数学问题）时，可能因推理过程错误而得到不可信的结果。这对其研究发展和应用落地均有负面影响。近期研究工作多采用检索外部信息[66]的方法提升大型语言模型的可信度。

7 结 论

近年，以 ChatGPT 为代表的大型语言模型突破了经典深度学习面临的数据依赖与能力泛化的局限，在多种应用上表现出更高的智能水平。大型语言模型的发展依赖于数据收集、模型结构、训练目标和训练方法等多方面技术的进步。由于大型语言模型含有的参数量大，经典“预训练＋微调”范式面临计算代价高、训练过拟合的问题。对此，以“提示学习”为核心的新范式逐渐兴起，利用大型语言模型的特点以更低的计算代价取得了更佳的性能。为了将大型语言模型的输出对齐至人类偏好，近期以 ChatGPT 为代表的大型语言模型通过指令微调和基于人类反馈的强化学习让生成结果更符合人类期望，进而能更好地解决真实场景的自然语言处理任务。在 ChatGPT 出现后，大型语言模型在开源模型、工具学习、长文本建模和反馈学习等方向上发展迅速，其应用范围也变得更加广泛。

除了取得的进步外，大型语言模型也引发了关于模型评价与治理的讨论。如何评价大型语言模型的性能以及如何提升它的安全性、可信度成为限制其发展与应用的关键问题。未来大型语言模型将成为基础设施逐渐走入人们的生活，其存在的问题也将深刻地影响社会的发展进程。因此，这些问题的解决需要社会各界的共同努力。

参 考 文 献

[1] Kuhn R, De Mori R. A cache-based natural language model for speech recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1990, 12(6): 570—583.

[2] Brown P, Cocke J, Pietra SA, et al. A statistical approach to machine translation. Computational Linguistics, 1990, 16: 79—85.

[3] Ponte JM, Croft WB. A language modeling approach to information retrieval. ACM SIGIR Forum, 2017, 51(2): 202—208.

[4] Jelinek F. Statistical methods for speech recognition. Cambridge: MIT Press, 1998.

[5] LeCun Y, Bengio Y, Hinton G. Deep learning. Nature, 2015, 521(7553): 436—444.

[6] Mikolov T, Chen K, Corrado G, et al. Efficient estimation of word representations in vector space. (2013-01-16)/[2023-06-30]. https://arxiv.org/pdf/1301.3781.pdf.

[7] Pennington J, Socher R, Manning CD. Glove: global vectors for word representation// Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing. Doha: Association for Computational Linguistics, 2014: 1532—1543.

[8] Peters M, Neumann M, Iyyer M, et al. Deep contextualized word representations// Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. New Orleans: Association for Computational Linguistics, 2018: 2227—2237.

[9] Hochreiter S, Schmidhuber J. Long short-term memory. Neural Computation, 1997, 9(8): 1735—1780.

[10] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need. (2017-06-12)/[2023-06-30]. https://arxiv.org/pdf/1706.03762.pdf.

[11] Devlin J, Chang MW, Lee K, et al. BERT: pre-training of deep bidirectional transformers for language understanding// Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Minneapolis: Association for Computational Linguistics, 2019: 4171—4186.

[12] Radford A, Narasimhan K, Salimans T, et al. Improving language understanding by generative pre-training. (2018-06-11)/[2023-06-30]. https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf.

[13] He KM, Zhang XY, Ren SQ, et al. Deep residual learning for image recognition// Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition. New York: IEEE, 2016: 770—778.

[14] Kaplan J, McCandlish S, Henighan T, et al. Scaling laws for neural language models. (2020-01-23)/[2023-06-30]. https://arxiv.org/pdf/2001.08361.pdf.

[15] Hoffmann J, Borgeaud S, Mensch A, et al. Training compute-optimal large language models. (2022-03-29)/[2023-06-30]. https://arxiv.org/pdf/2203.15556.pdf.

[16] Wenzek G, Lachaux MA, Conneau A, et al. CCNet: extracting high quality monolingual datasets from web crawl data. (2019-11-15)/[2023-06-30]. https://arxiv.org/pdf/1911.00359.pdf.

[17] Gao L, Biderman S, Black S, et al. The pile: an 800GB dataset of diverse text for language modeling. (2020-12-31)/[2023-06-30]. https://arxiv.org/pdf/2101.00027.pdf.

[18] Yuan S, Zhao HY, Du ZX, et al. WuDaoCorpora: a super large-scale Chinese corpora for pre-training language models. AI Open, 2021, 2: 65—68.

[19] Lewis M, Liu YH, Goyal N, et al. BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension// Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Stroudsburg: Association for Computational Linguistics, 2020: 7871—7880.

[20] Shazeer N. GLU variants improve transformer. (2020-02-12)/[2023-06-30]. https://arxiv.org/pdf/2002.05202.pdf.

[21] Zhang B, Sennrich R. Root mean square layer normalization. (2019-10-16)/[2023-06-30]. https://arxiv.org/pdf/1910.07467.pdf.

[22] Liu PF, Yuan WZ, Fu JL, et al. Pre-train, prompt, and predict: a systematic survey of prompting methods in natural language processing. ACM Computing Surveys, 2023, 55(9): 1—35.

[23] Schick T, Schütze H. Exploiting cloze-questions for few-shot text classification and natural language inference// Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics. Stroudsburg: Association for Computational Linguistics, 2021: 255—269.

[24] Lester B, Al-Rfou R, Constant N. The power of scale for parameter-efficient prompt tuning// Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. Stroudsburg: Association for Computational Linguistics, 2021: 3045—3059.

[25] Shin T, Razeghi Y, Logan RL, et al. AutoPrompt: eliciting knowledge from language models with automatically generated prompts// Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing. Stroudsburg: Association for Computational Linguistics, 2020: 4222—4235.

[26] Gao TY, Fisch A, Chen DQ. Making pre-trained language models better few-shot learners// Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing. Stroudsburg: Association for Computational Linguistics, 2021: 3816—3830.

[27] Brown TB, Mann B, Ryder N, et al. Language models are few-shot learners. (2020-05-28)/[2023-06-30]. https://arxiv.org/pdf/2005.14165.pdf.

[28] Wei J, Wang XZ, Schuurmans D, et al. Chain-of-thought prompting elicits reasoning in large language models. (2022-01-28)/[2023-06-30]. https://arxiv.org/pdf/2201.11903.pdf.

[29] Kojima T, Gu SS, Reid M, et al. Large language models are zero-shot reasoners. (2022-05-24)/[2023-06-30]. https://arxiv.org/pdf/2205.11916.pdf.

[30] Wei J, Tay Y, Bommasani R, et al. Emergent abilities of large language models. (2022-06-15)/[2023-06-30]. https://arxiv.org/pdf/2206.07682.pdf.

[31] Li XL, Liang P. Prefix-tuning: optimizing continuous prompts for generation// Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing. Stroudsburg: Association for Computational Linguistics, 2021: 4582—4597.

[32] Hu EJ, Shen YL, Wallis P, et al. LoRA: low-rank adaptation of large language models. (2021-06-17)/[2023-06-30]. https://arxiv.org/pdf/2106.09685.pdf.

[33] Ding N, Qin YJ, Yang G, et al. Parameter-efficient fine-tuning of large-scale pre-trained language models. Nature Machine Intelligence, 2023, 5(3): 220—235.

[34] OpenAI. Introducing ChatGPT. (2022-11-30)/[2023-06-30]. https://openai.com/blog/chatgpt.

[35] Chen M, Tworek J, Jun H, et al. Evaluating large language models trained on code. (2021-07-14)/[2023-06-30]. https://arxiv.org/pdf/2107.03374.pdf.

[36] Long OY, Wu J, Xu J, et al. Training language models to follow instructions with human feedback. (2022-03-04)/[2023-06-30]. https://arxiv.org/pdf/2203.02155.pdf.

[37] Wei J, Bosma M, Zhao VY, et al. Finetuned language models are zero-shot learners. (2021-09-03)/[2023-06-30]. https://arxiv.org/pdf/2109.01652.pdf.

[38] Stiennon N, Ouyang L, Wu J, et al. Learning to summarize with human feedback. Advances in Neural Information Processing Systems, 2020, 33: 3008—3021.

[39] Schulman J, Wolski F, Dhariwal P, et al. Proximal policy optimization algorithms. (2017-07-20)/[2023-06-30]. https://arxiv.org/pdf/1707.06347.pdf.

[40] Touvron H, Lavril T, Izacard G, et al. LLaMA: open and efficient foundation language models. (2023-02-27)/[2023-08-05]. https://arxiv.org/pdf/2302.13971.pdf.

[41] Taori R, Gulrajani I, Zhang T, et al. Stanford alpaca: an instruction-following llama model. (2023-03-14)/[2023-08-05]. https://github.com/tatsu-lab/stanford_alpaca.

[42] Zheng LM, Chiang WL, Sheng Y, et al. Judging LLM-as-a-judge with MT-bench and chatbot arena. (2023-06-09)/[2023-08-05]. https://arxiv.org/pdf/2306.05685.pdf.

[43] Zeng A, Liu X, Du Z, et al. GLM-130B: an open bilingual pre-trained model. (2022-10-05)/[2023-08-05]. https://arxiv.org/pdf/2210.02414.pdf.

[44] Zheng R, Dou SH, Gao SY, et al. Secrets of RLHF in large language models part I: PPO. (2023-07-18)/[2023-08-05]. https://arxiv.org/pdf/2307.04964.pdf.

[45] InternLM Team. InternLM: a multilingual language model with progressively enhanced capabilities. (2023-07-07)/[2023-08-05]. https://github.com/InternLM/InternLM.

[46] Wu CF, Yin SM, Qi WZ, et al. Visual ChatGPT: talking, drawing and editing with visual foundation models. (2023-03-08)/[2023-08-05]. https://arxiv.org/pdf/2303.04671.pdf.

[47] Shen YL, Song KT, Tan X, et al. HuggingGPT: solving AI tasks with ChatGPT and its friends in hugging face. (2023-03-30)/[2023-08-05]. https://arxiv.org/pdf/2303.17580.pdf.

[48] Patil SG, Zhang TJ, Wang X, et al. *Gorilla*: large language model connected with massive *APIs*. (2023-05-24)/[2023-08-05]. https://arxiv.org/pdf/2305.15334.pdf.

[49] Qin YJ, Liang SH, Ye YN, et al. ToolLLM: facilitating large language models to master 16000+ real-world *APIs*. (2023-07-31)/[2023-08-05]. https://arxiv.org/pdf/2307.16789.pdf.

[50] Xu QT, Hong FL, Li B, et al. On the tool manipulation capability of open-source large language models. (2023-05-25)/[2023-08-05]. https://arxiv.org/pdf/2305.16504.pdf.

[51] Dao T, Fu DY, Ermon S, et al. Flashattention: fast and memory-efficient exact attention with io-awareness. Advances in Neural Information Processing Systems, 2022, 35: 16344—16359.

[52] Sun YT, Dong L, Huang SH, et al. Retentive network: a successor to transformer for large language models. (2023-07-17)/[2023-08-05]. https://arxiv.org/pdf/2307.08621.pdf.

[53] Chen SY, Wong S, Chen LJ, et al. Extending context window of large language models via positional interpolation. (2023-06-28)/[2023-08-05]. https://arxiv.org/pdf/2306.15595.pdf.

[54] Liu H, Sferrazza C, Abbeel P. Chain of hindsight aligns language models with feedback. (2023-03-25)/[2023-08-05]. https://arxiv.org/pdf/2302.02676.pdf.

[55] Yuan Z, Yuan HY, Tan CQ, et al. RRHF: rank responses to align language models with human feedback without tears. (2023-04-11)/[2023-08-05]. https://arxiv.org/pdf/2304.05302.pdf.

[56] Van Dis EAM, Bollen J, Zuidema W, et al. ChatGPT:five priorities for research. Nature, 2023, 614(7947): 224—226.

[57] Papineni K, Roukos S, Ward T, et al. BLEU: a method for automatic evaluation of machine translation// Proceedings of the 40th Annual Meeting on Association for Computational Linguistics. Morristown: Association for Computational Linguistics, 2002: 311—318.

[58] Lin C Y. Rouge: a package for automatic evaluation of summaries// Proceedings of the Workshop on Text Summarization Branches Out. Barcelona: Association for Computational Linguistics, 2004: 74—81.

[59] OpenAI. GPT-4 technical report. (2023-03-27)/[2023-06-30]. https://arxiv.org/pdf/2303.08774.pdf.

[60] Wang PY, Li L, Chen L, et al. Large language models are not fair evaluators. (2023-05-29)/[2023-06-30]. https://arxiv.org/pdf/2305.17926.pdf.

[61] Huang YZ, Bai YZ, Zhu ZH, et al. C-eval: a multi-level multi-discipline Chinese evaluation suite for foundation models. (2023-05-17)/[2023-06-30]. https://arxiv.org/pdf/2305.08322.pdf.

[62] Zhang MA, Jin LF, Song LF, et al. SafeConv: explaining and correcting conversational unsafe behavior// Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics. Stroudsburg: Association for Computational Linguistics, 2023: 22—35.

[63] Deng JW, Zhou JY, Sun H, et al. Cold: A benchmark for Chinese offensive language detection// Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing. Stroudsburg: Association for Computational Linguistics, 2022: 11580—11599.

[64] Zhang ZX, Cheng JL, Sun H, et al. Constructing highly inductive contexts for dialogue safety through controllable reverse generation// Findings of the Association for Computational Linguistics: EMNLP 2022. Stroudsburg: Association for Computational Linguistics, 2022: 3684—3697.

[65] Li JY, Cheng XX, Zhao WX, et al. HaluEval: a large-scale hallucination evaluation benchmark for large language models. (2023-05-22)/[2023-06-30]. https://arxiv.org/pdf/2305.11747.pdf.

[66] Peng BL, Galley M, He PC, et al. Check your facts and try again: improving large language models with external knowledge and automated feedback. (2023-02-24)/[2023-06-30]. https://arxiv.org/pdf/2302.12813.pdf.

Research Progress of Large Language Models Represented by ChatGPT

Pei Ke[1]　Wenqiang Lei[2]　Minlie Huang[1*]

1. *Department of Computer Science and Technology, Tsinghua University, Beijing* 100084

2. *College of Computer Science, Sichuan University, Chengdu* 610065

Abstract　Large language models are currently one of the most cutting-edge research directions in the area of artificial intelligence, aiming to train general language models with large-scale parameters and make them follow human instructions to solve various natural language processing (NLP) tasks. As a representative of large language models, ChatGPT which is developed by OpenAI exhibits strong capability of natural language generation in various areas and attracts worldwide attention. This paper begins with the development of language models and introduces the investigation of researchers on scaling up language models in recent years. Then, this paper analyzes the change of paradigms caused by large language models and briefly introduces the development, technology, and application of ChatGPT as a typical example. Next, this paper introduces the frontier progress of large language models in the post-ChatGPT era. Finally, this paper summarizes the limitation and challenges of current large language models from the perspectives of evaluation and governance, which need to be tackled in the future.

Keywords　large language models; ChatGPT; pre-trained language models; Transformer; chain-of-thought; natural language processing; artificial intelligence

（责任编辑　崔国增　姜钧译）

* Corresponding Author, Email: aihuang@tsinghua.edu.cn