ChatGPT的工作原理、关键技术及未来发展趋势

秦涛，杜尚恒，常元元，王晨旭
（西安交通大学智能网络与网络安全教育部重点实验室，710049，西安）

摘要：ChatGPT是自然语言处理领域的一项重要技术突破，专注于对话生成任务，在多种任务中表现出卓越的性能。主要探讨ChatGPT的演变历程、关键技术，并分析了其未来可能的发展方向。首先，介绍了ChatGPT的模型架构和技术演进过程。随后，重点讨论了ChatGPT的关键技术，包括提示学习与指令微调、思维链、人类反馈强化学习。然后，分析了由于基于概率生成原理所造成的固有局限，包括事实性错误、垂直领域深度性弱、潜在的恶意应用风险、可解释性及模型实时性差等。最后，探讨了其在典型应用中存在的问题和相应的解决途径，包括在训练评估过程中考虑道德和安全性因素，以降低潜在风险；结合外部专家知识和迁移学习，以提高模型对特定领域的理解能力，更好地适应特定任务场景；引入多模态数据，以提高模型信息理解能力，增强模型通用性和泛化性。通过对ChatGPT模型框架、技术演变与关键技术的分析，为深入理解ChatGPT提供帮助；结合原理分析其固有缺陷，并结合实际应用中存在的问题，挖掘未来可能的研究方向，为自然语言处理领域的深入研究提供有益参考。

关键词：ChatGPT模型架构；概率生成；强化学习；迁移学习
中图分类号：TP391.9 **文献标志码**：A
DOI：10.7652/xjtuxb202401001 **文章编号**：0253-987X(2024)01-0001-12

Principles, Key Technologies and Emerging Trends of ChatGPT

QIN Tao, DU Shangheng, CHANG Yuanyuan, WANG Chenxu
(MOE Key Lab for Intelligent and Network Security, Xi'an Jiaotong University, Xi'an 710049, China)

Abstract: ChatGPT has emerged as a significant advancement in natural language processing, specifically in the domain of dialogue generation, and has achieved excellent performance in many areas. This paper aims to explore its architecture, underlying technologies, and potential areas for further investigation. The paper begins by discussing the architecture and technology evolution process. Next, the focus shifts to a comprehensive analysis of the key technologies, including the prompt learning and instruction fine-tuning, chain of thought and reinforcement learning through human feedback. Furthermore, the paper addresses the limitations of ChatGPT stemming from its probabilistic generation principles, including factual errors, poor performance in specific domain, potential malicious risk, poor interpretability and real-time. Finally, the paper outlines possible research directions based on the practical challenges observed in real-world applications, including the ethical and safety factors in the training process to reduce potential risks. Addition-

收稿日期：2023-06-25。 作者简介：秦涛(1982—)，男，教授，博士生导师。 基金项目：国家自然科学基金资助项目(62172324)；陕西省重点研发计划资助项目(2023-YBGY-269，2022QCY-LL-33HZ)。
网络出版时间：2023-10-17 网络出版地址：https://link.cnki.net/urlid/61.1069.T.20231016.1403.012

ally, integrating external expert knowledge and employing transfer learning methods are proposed to enhance ChatGPT's performance in domain-specific tasks. Moreover, improving its information understanding capabilities based on the multimodal data is also considered as a notable avenue for development. By providing an in-depth analysis of ChatGPT's framework and key technologies, this paper aims to foster a deeper understanding of the system and also presents potential research directions to inspire further investigation in the field.

Keywords: ChatGPT model architecture; probabilistic generative model; reinforcement learning; transfer learning

自然语言处理作为人工智能领域的关键技术之一,具有重要的应用价值,在信息提取与知识管理领域,商业机构可利用该技术布局线上系统,开发智能客服系统自动理解客户问题并及时响应客户需求,从而协助办理业务,提升服务效率;也能通过对海量多源数据的处理分析,构建多层次多维度用户画像,制定精细化、个性化服务方案来优化服务质量。监管人员[1]可以结合分词、实体识别、热词发现和情感倾向分析等技术,对社交媒体数据进行情感分析[1-2],尽早发现负面消极的言论和煽动性的话题,采用信息抽取与文本聚类技术从信息流中检测并聚合突发事件,并利用网络分析和深度学习技术分析事件在社交网络中的传播途径,理解信息的扩散模式从而进行事件演化与趋势预测,为舆情管控和引导提供决策支持,推动社会管理的智能化和精细化。但是,自然语言处理技术的发展仍受到可用标注数据稀缺、数据多源时变、语义信息多样复杂等问题的困扰。

正是在这样的需求推动下,领域内的技术框架不断更新进步,自然语言处理技术的演进阶段可以分为小规模专家知识、浅层机器学习[3]、深度学习[4]、预训练语言模型[5]等,每个技术阶段的演进周期大致为前一阶段的一半,迭代速度越发迅速。ChatGPT 作为大规模预训练模型的一种典型代表,极大地推动了自然语言处理技术的发展,引发了自然语言处理研究范式的转变,其通过大规模的预训练和上下文理解,具备了生成自然语言文本的能力,可以进行对话、回答问题和提供信息等任务,与人类交互的能力更加自然和灵活。

为进一步理解 ChatGPT,本文首先介绍 ChatGPT 的模型架构和技术演进过程,然后回顾了其所用的核心技术,包括提示学习、思维链和基于人类反馈的强化学习,这些技术共同构成了 ChatGPT 的基础框架,使其能够在各种情景下生成连贯且自然的文本回应。然后,结合 ChatGPT 运行原理,本文分析了其面临的缺陷与挑战,包括生成不准确或具有误导性的信息、潜在的恶意应用风险以及对话中的道德和隐私问题等。最后,针对 ChatGPT 在特定领域的缺陷与不足,结合实际应用,探讨了 ChatGPT 未来可能的发展方向,包括对训练语料进行道德筛选、采用迁移学习[6]和领域适应技术、引入外部专家知识[7]、增强多模态处理能力[8]等途径来优化改进。

1 GPT 架构及演变过程

GPT(generative pre-trained transformer)[9]是由 OpenAI 提出的采用 Transformer 解码器的预训练模型,采用预训练加微调的范式。为深入理解 ChatGPT,本节简要分析 ChatGPT 的模型架构和其演进进程。

1.1 ChatGPT 整体架构

ChatGPT 的主体架构遵从"基础语料+预训练+微调"的基本范式,如图 1 所示。"预训练+微调"是指首先在大数据集上训练得到一个具有强泛化能力的模型(预训练模型),然后在下游任务上进行微调的过程,是基于模型的迁移方法。

海量高质量的基础语料是 ChatGPT 技术突破的关键因素。其语料体系包括预训练语料与微调语料,后者包括代码和对话微调语料。预训练语料包括从书籍、杂志、百科等渠道收集的海量文本数据,具体分布见表 1[10],提供了丰富的语义语境和词汇,帮助模型深入学习理解自然语言中的基础词汇与逻辑关系表达规则;微调语料包括从开源代码库爬取、专家标注、用户对话等方式加工而成的高质量标注文本数据,进一步增强其对话能力。

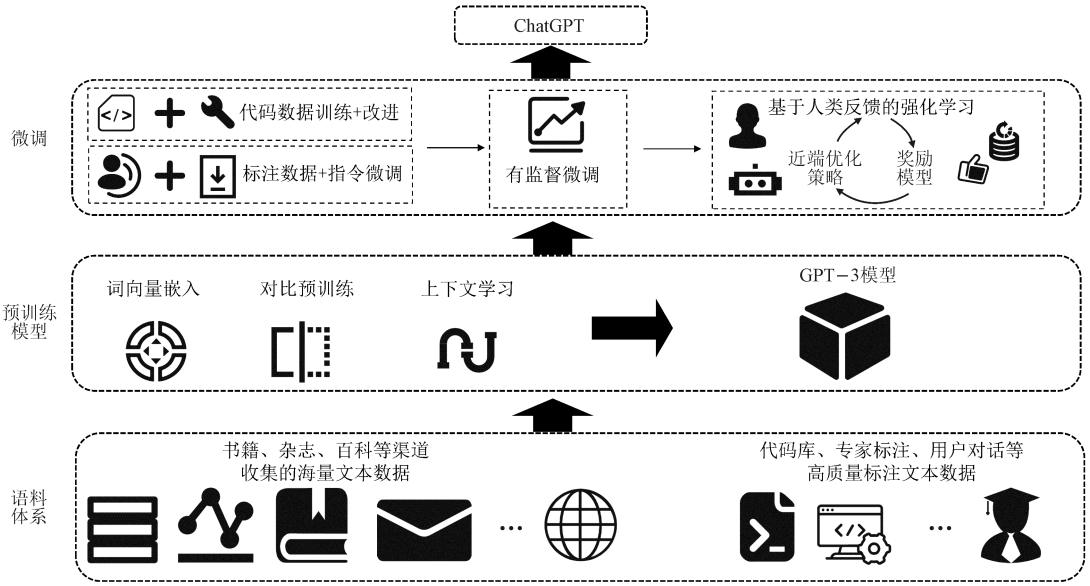

图 1　ChatGPT 架构示意图

Fig. 1　Diagram of ChatGPT architecture

表 1　GPT 系列预训练语料数据大小[10]

Table 1　GPT series pre-training corpus data size

单位：GB

模型	书籍	期刊	Reddit 链接	爬虫数据	维基百科
GPT-1	4.6				
GPT-2			40		
GPT-3	21	101	50	570	11.4

预训练是构建大规模语言模型的基础，指先在大规模训练数据上进行大量通用的训练，采用无监督学习方法以得到通用且强泛化能力的语言模型。在大规模数据的基础上，通过预训练，模型初步具备了人类语言理解和上下文学习的能力，能够捕捉文本片段和代码片段的语义相似性特征，从而生成更准确的文本和代码向量，为后续微调任务提供支持。

微调是实现模型实际应用的保障，是指在特定任务的数据集上对预训练模型进行进一步的训练，通常包括冻结预训练模型的底层层级（如词向量）与调整上层层级（如分类器）的权重。对预训练模型微调将大大缩短训练时间，节省计算资源并加快训练收敛速度。ChatGPT 在具有强泛化能力的预训练模型基础上，通过整合基于代码数据的训练和基于指令的微调，利用特定的数据集进行微调，使之具有更强的问答式对话文本生成能力。其“预训练＋微调”的流程如图 2 所示。

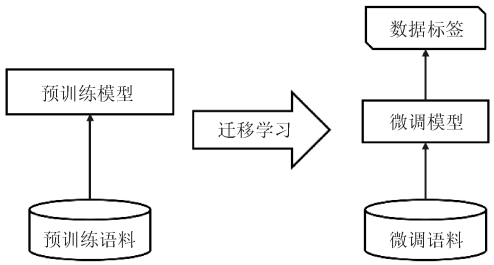

图 2　“预训练＋微调”流程

Fig. 2　Pre-training and fine-tuning flow chart

1.2　GPT-1：奠定自回归式建模思路

GPT-1[9] 是比 BERT[11] 提出更早的预训练模型，但与 BERT 相比效果较差。GPT-1 奠定了关键的技术路径，后续的系列模型采用类似的架构（例如 BART[12] 和 GPT-2[13]）以及预训练策略[14-16]。GPT 系列是一种基于自回归解码的、仅包含解码器的 Transformer 架构开发的生成式预训练模型，这种架构具有自回归（AR）特性，它利用多层堆叠的 Transformer 解码器架构进行解码。

自回归是统计学中处理时间序列的方法，用同一变量之前各时刻的观测值预测该变量当前时刻的观测值。类似地，自回归生成模型的基本思想是在序列生成的过程中，每个位置的元素都依赖于前面已经生成的元素。自回归模型适用于各种序列到序列的任务，它又分为线性自回归和神经自回归两种，基于 Transformer 解码器的自回归模型属于后者，其生成过程如图 3 所示。

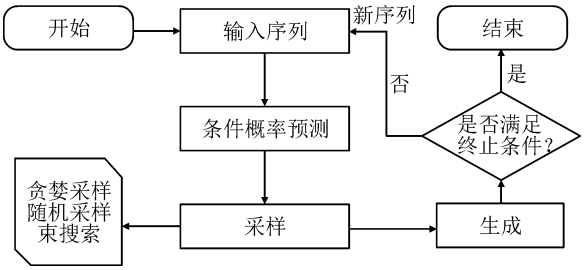

图 3 自回归生成模型生成过程

Fig. 3 Autoregressive model generation process

GPT-1 以 Transformer 解码器作为核心组件，在大规模未标记的文本数据上进行自监督学习，通过最大似然估计来调整模型参数，使得模型能够更好地预测下一个词。这种预训练策略允许模型从大规模的文本数据中学习通用的语言规律和语义关系。在生成序列时，模型将已生成序列作为已知条件，并利用 Transformer 解码器来预测下一个元素的概率分布，输入前 L 个元素构成的序列 $x=\{x_1,\cdots,x_L\}$，预测 $\tilde{x}=\{x_2,\cdots,x_{L+1}\}$。每次生成的元素会添加到序列中。生成的过程会一直持续，直到达到预定的生成长度或遇到终止符号，如图 4 所示。在预训练结束之后可以根据不同的下游任务或者特定的语境场景进行微调。

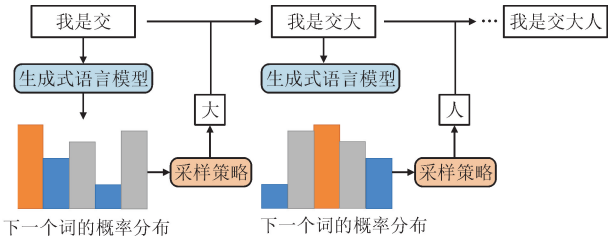

图 4 GPT-1 生成过程

Fig. 4 GPT-1 generation process

1.3 GPT2：引入提示学习

GPT-2[13] 通过模型结构的改进，在下游任务的微调上取得了更好的结果。GPT-2 在以下两个方面进行了优化。

(1)扩大参数规模。使用更多高质量的网页数据，将模型参数规模扩大到 1.5×10^9。

(2)更自然的任务模型融合方式。GPT-2 将下游任务通过 prompt 方式加入到预训练模型中，从而让模型获得零样本学习的能力，即引入了一种多任务求解的概率形式，通过给定输入与任务条件对结果进行预测。

虽然 GPT-2 在下游任务的微调中并没有 BERT 模型表现优越，但其更自然的任务融合方式为后续 ChatGPT 的指令理解能力奠定了基础，即对输入文本信息按照特定模板进行处理，将任务重构成一个更能充分利用语言模型处理的形式。

提示学习的工作流框架如图 5 所示，主要包含以下 4 部分：提示词模板构造，提示词答案空间映射构造，将文本带入模板嵌入并使用预训练模型预测答案、将预测结果映射回标签。按照模板构造差异，提示学习方法可分为硬模板方法和软模板方法。具体来说，硬模板方法先在少量监督数据上对每个提示词训练模型，再在无监督数据上将同一样本的多个提示词预测结果进行集成，建模为 $s_p(l\mid x)=M(v(l)\mid P(x))$，表示在给定提示词下对应词语的分数，归一化得到概率分布 $q_p(l\mid x)=\dfrac{\mathrm{e}^{s_p(l\mid x)}}{\sum\limits_{l'\in\zeta}\mathrm{e}^{s_p(l\mid x)}}$ 作为无监督数据的软标签。软模板方法则是直接在输入端嵌入若干可被优化的提示词流，使其自动化地寻找连续空间内的知识模板，从而不依赖人工设计。以生成任务中的前缀微调(prefix tuning)[17] 为例，模型在每一层 Transformer 前加入前缀连续向量并使用训练矩阵 $\boldsymbol{P}$ 来存储，将输入表示为 $[\boldsymbol{p};\boldsymbol{x};\boldsymbol{y}]$，在整个训练期间冻结预训练模型其他参数不变，只更新用于给定任务的前缀参数。

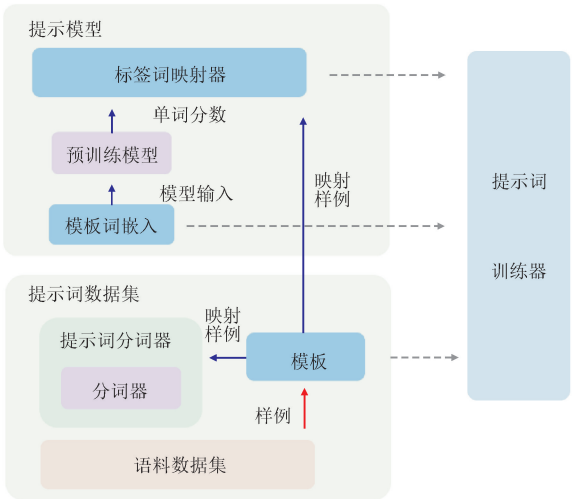

图 5 提示学习的工作流框架示意图

Fig. 5 Diagram of prompt learning

通过上述方式，每个自然语言处理的任务都可以被视作基于世界文本子集的单词预测问题[18]。这种思想表明，只要模型足够大、学到的知识足够丰

富，任何有监督任务都可以通过无监督的方式来完成，GPT-2 下游任务中的对话任务[19-20]更是进行了全面的微调，为后续的 ChatGPT 对话奠定了基础。

1.4　GPT-3：量变引起质变

在 GPT-2 的基础上，GPT-3[21]通过扩展生成预训练架构，实现了容量飞跃。GPT-3 的显著特点就是规模大。由于 GPT-2 的实验中发现随着参数规模的增大其效果的增长依旧显著[22-23]，因此选择继续扩大参数规模，用更多优质的数据，一方面是模型本身规模大，参数量众多，具有 96 层 Transformer 解码器，每一层有 96 个 128 维的注意力头，单词嵌入的维度也达到了 12 288 维；另一方面是训练过程中使用到的数据集规模大，达到了 45 TB，参数规模达到 1.75×10^{11}。

此外，GPT-3 在模型能力上转变思路，采用情景学习的思想，使模型能够在少样本学习上取得较好的效果。大量实验证明 GPT-3 在少样本情况下具有良好的表现，如图 6 所示。

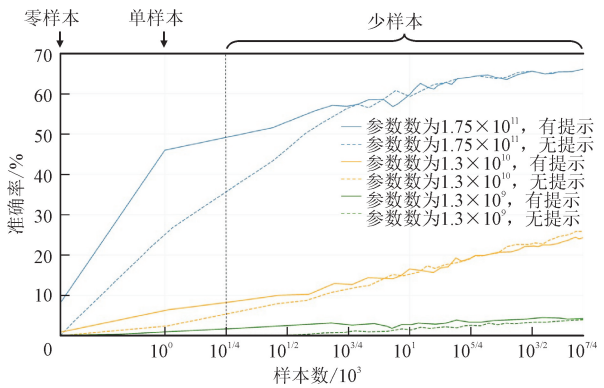

图 6　GPT-3 性能随样本量的变化

Fig. 6　GPT-3 performance variation with sample size

由于规模巨大，GPT-3 在各领域均有广泛的应用，衍生了多种应用生态，被视为从预训练模型发展到大模型过程中的一个里程碑。

1.5　ChatGPT：实现人机混合增强

尽管 GPT-3 拥有大量知识，但生成文本质量不一且语言表达冗余，ChatGPT 通过人工标注的微调，引导模型输出更有价值的文本结果，即实施了人类反馈的强化学习机制。

OpenAI 对于混合人类反馈增强机器智能的研究可以追溯到 2017 年[24]，并且在当年发布了近端策略优化(PPO)[25]算法作为强化学习的基础算法，该算法通过多个训练步骤实现小批量更新，以克服传统策略梯度算法中步长难确定、可能导致学习性能下降的问题，引入保守的策略更新机制有效缓解了策略更新过快导致的不稳定性，提高了训练的鲁棒性和稳定性。PPO 算法根据当前策略与环境互动产生轨迹，并记录各状态、动作与奖励，使用轨迹信息更新策略限制策略步长，使得目标散度既足以显著改变策略，又足以使更新稳定，防止新旧策略过远，并在每次更新后重新计算优势函数。

具体来说，近端策略优化算法的流程如下。

(1)初始化。初始化网络参数 θ 和值函数参数 φ。强化学习中，值函数通常用于估计一个给定状态 s 在当前策略 π_θ 下的期望累积折扣奖励。这个函数通常用神经网络或其他函数逼近器来表示，逼近器的参数记作 φ。

(2) 数据收集。在环境中用当前策略 π_θ 进行多步操作，收集状态 s_t、动作 a_t 和奖励 r_t。状态 s_t 用于描述环境在某一时刻的观察结果。动作 a_t 是智能体在某一状态下决定执行的操作。奖励 r_t 是环境给予智能体的反馈，用于量化智能体执行某个动作的效果或价值。

(3) 优势函数估计。对每一个状态 s，使用累计折扣奖励和值函数 $V_\varphi(s_t)$ 来估计优势函数 A，用于量化一个动作相对于平均情况下的预期效果或好处。

(4) 策略更新。将当前策略备份为 π_{old}。通过目标函数 $J_{\text{PPO}}(\theta)$ 来更新策略。目标函数 $J_{\text{PPO}}(\theta)$ 为组合了优势函数和旧策略的有界目标函数，包括 KL 散度项用于确保新旧策略不会相差太远，是一个综合考虑策略改进和稳定性的目标函数。

(5) 值函数更新。使用损失函数 $L_{\text{BL}}(\varphi)$ 来更新值函数 $V_\varphi(s_t)$。损失函数 $L_{\text{BL}}(\varphi)$ 量化了实际累计奖励和值函数预测之间的误差，通常定义为均方差误差，更新次数由参数 B 控制。

(6) 适应性调整正则化系数 λ。如果新旧策略之间的 KL 散度超过了一个高阈值，增大 λ 以减少策略更新幅度。反之，如果 KL 散度低于一个低阈值，减小 λ 以允许更大幅度的策略更新。

(7)循环。以上步骤会被重复 N 次，以不断优化策略和值函数。

OpenAI 在 GPT-2 时便开始使用上述强化学习算法[24-25]来进行微调，同年以类似方法训练了文本摘要模型[26]。

ChatGPT 的前身，InstructGPT[27] 模型正式使用了基于人类反馈强化学习(RLHF)算法，通过结合智能体自主学习与人类专家反馈两种策略，选择基于策略梯度的算法搭建强化学习模型从而训练智能体，并在每个时间步上记录智能体行为并且由人类专家对其进行评估反馈，以进行参数更新优化行为策略。该算法的第一阶段[27]是指令调优。除了提高指令理解能力外，RLHF 算法还有助于缓解大模型产生危害或不当内容的问题，这也是大模型在安全实践部署的关键。OpenAI 在技术文章[27]中描述了他们的对齐研究方法，该文章总结了 3 个有希望的方向，即“使用人类反馈训练 AI 系统，帮助人类评估和进行对齐研究”。

1.6 GPT4：多模态升级

GPT-4 是对 ChatGPT 的多模态升级，可对图文输入产生应答文字，并可引用于视觉分类分析、隐含语义等领域。多模态输入能力对语言模型至关重要，使其可以获得除文本描述外的常识性知识，并为多模态感知与语义理解的结合提供了可能性。

新范式可归纳为“预训练＋提示＋预测”。各种下游任务被调整为类似预训练任务的形式，尤其 GPT-4 的多模态提示工程针对多模态数据集，涉及合适的模型架构参数、精心设计的提示格式结构和选定的数据微调模型，来使得模型生成高质量文本。

2 ChatGPT 的核心技术

2.1 提示学习与指令精调

传统的监督学习使用包含输入 x 与标签 y 的数据集来训练一个模型 $P(y|x;\theta)$，从而学习模型参数 θ 预测条件概率。提示学习[28] 试图学习模拟概率 $P(x;\theta)$ 的 x 本身来预测 y，从而减少或消除对大型监督数据集的需求。

具体来说，将输入 x 通过提示函数 $f_{\text{prompt}}(x)$ 添加槽 Z 转化为特定形式 x'，定义 Z 为回答 z 允许值的集合，定义填充函数 $f_{\text{fill}}(x',z)$ 来将可能的回答 z 填充 x' 中的槽 Z。最后，使用特定搜索函数来计算相应的填充提示概率，得到最终回答 $\hat{z}$，定义如下

$$\hat{z}=\underset{z\in Z}{\text{search}}P(f_{\text{fill}}(x',z);\Theta) \tag{1}$$

通过编辑任务的输入，提示学习在形式上模拟模型训练中的数据与任务。以情感分类任务为例，相比于监督学习中输入一句话，模型输出情感分类判断，提示学习是设计一种模板，将原有语句嵌入其中，为模型留出判断类别的位置，让模型做类似完形填空的工作生成情感类别。提示学习旨在激发语言模型的补全能力，指令精调(instruction tuning)[29] 则是提示学习的加强版，激发模型的理解能力。通过指令调优，模型能够在不使用显式示例的情况下遵循新任务的任务指令，从而提高了泛化能力[29-30]，即便在多语言环境下也有卓越能力[31]。这种学习人类交互模式的分布让模型可以更好地理解人类意图[32]、与人类行为对齐[27]。从解释性上来说，这类似于打开大门的钥匙，从大模型在预训练中学习到的庞大知识中激活特定的部分完成指定任务。ChatGPT 能响应人类指令的能力就是指令微调的直接产物，对没有见过的指令做出反馈的泛化能力是在指令数超过一定程度之后自动出现的，T0 模型[15]、Flan 模型[29] 等工作都进一步证明了这一点。

对于模型未训练的新任务，只需设计任务的语言描述，并给出任务实例作为模型输入，即可让模型从给定的情景中学习新任务并给出恰当的回答结果。这种训练方式能够有效提升模型小样本学习[33]的能力。

2.2 思维链

谷歌研究人员 Wei 等提出了思维链(chain of thought，COT)[34] 的概念，即在小样本提示学习中插入一系列中间推理的步骤示范，从而有效提高语言模型的推理能力。与一般的提示词不同，思维链提示由多个分别解释子问题的中间步骤组成，提示词模式从之前的问题、答案变成输入问题、思维链、输出问题。如图 7 所示，以数学计算解答为例，一般的提示词模板通过输入内嵌入样例，使得模型学习任务答案，而思维链提示词增加推理步骤，参考人类解决问题方法，嵌入自然语言形式的推理步骤直至答案生成。在思维链的加持下，通过将问题分解为一系列的分步推理，根据前一步骤结果与当前问题要求共同推断下一步骤。通过这种逐步推理的方式，模型可以逐渐获得更多信息，并在整个推理过程中累积正确的推断，从而大幅度提升模型在复杂推理时的准确率，表现在数学计算结果的正确与否，同时也为模型的推理行为提供了一个可解释的窗口[27,35-36]。

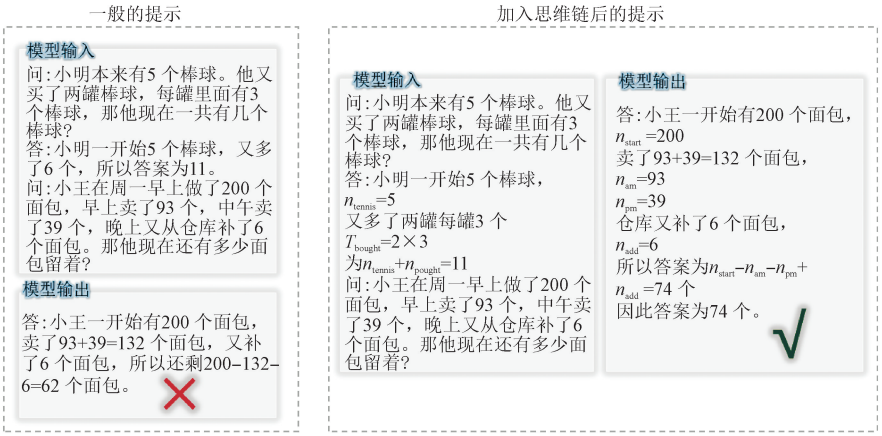

图 7　思维链示意图

Fig. 7　Diagram of chain of thought

2.3　人类反馈强化学习

人类反馈强化学习[24, 37]是 ChatGPT 实现理解人类指令、对齐人类行为[38-39]的重要技术。图 8 所示为模型训练过程。此算法[40]在强化学习[41]的框架下大体可以分为 3 个阶段。

(1)深度模型训练，对应图 8 中的第一步。利用大规模的文本数据，对每一个给定的提示，深度模型会生成一个初始的回复，然后返还给标注人员。标注人员检查并评价这个回复，这种方法允许从人类反馈中获得有价值的修正建议和方向，从而提高模型的准确性和可靠性。收集到这些经过人类评估和修正的数据后，使用监督学习(supervised fine-tuning)方法来在 GPT-3 模型上进行微调。

(2)奖励模型训练[42]，对应图 8 中第 2 步。该阶段旨在获取拟合人类偏好的奖励模型。奖励模型以提示和回复作为输入，计算标量奖励值作为输出。奖励模型的训练过程通过拟合人类对于不同回复的倾向性实现。具体而言，首先基于在人类撰写数据上精调的模型，针对同一提示采样多条不同回复。然后，将回复两两组合构成一条奖励模型训练样本，由人类给出倾向性标签。最终，奖励模型通过每条样本中两个回复的奖励值之差计算倾向性概率拟合人类标签，完成奖励模型的训练。

(3)生成策略优化与迭代，对应图 8 中第 3 步。给定训练的奖励模型，GPT 模型的参数将被视为一种策略，在强化学习的框架下进行训练。首先，当前策略根据输入的查询采样回复。然后，奖励模型针对回复的质量计算奖励、反馈回当前策略用以更新。

图 8　ChatGPT 训练过程示意图

Fig. 8　Diagram of ChatGPT training process

为防止上述过程的过度优化，损失函数引入了词级别的 KL 惩罚项。此外，为了避免在公开 NLP 数据集上的性能退化，策略更新过程兼顾了预训练损失。

3 ChatGPT 面临的问题

虽然 ChatGPT 在多个任务中都表现出不错的性能，其现有运行原理决定了其有很多局限性。

(1)对某个领域的深入程度不够[35,43]，因此生成的内容可能不够合理。此外，ChatGPT 也存在潜在的偏见问题，因为它是基于大量数据训练的，训练数据中的固有偏差会渗透到神经网络中，导致模型会受到数据中存在的偏见的影响[27,44]。

(2)对抗鲁棒性[45]。对抗鲁棒性在自然语言处理与强化学习中是决定系统适用性的关键要素[46]，对于干扰示例 $x'=x+\delta$，其中，x 指原始输入，δ 指扰动，高鲁棒性系统会产生原始输出 y，而低鲁棒性系统会产生不一样的输出 y'。ChatGPT 容易受到对抗性攻击，例如数据集中攻击[47]、后门攻击[48]和快速特定攻击[49]等，从而诱导模型产生有害输出。

(3)安全保障。由于 ChatGPT 是一种强大的人工智能技术，它可能被恶意利用，造成严重的安全隐患及产生法律风险[50]。同时，它的答复尚不明确是否具有知识产权，从而可能产生不利的社会影响[51]。因此，开发者在设计和使用 ChatGPT 时，需要采取相应措施，例如去偏方法和校准技术[52]来保障安全性问题。

(4)推理可信度。与其他神经网络类似，ChatGPT 很难精确地表达其预测的确定性[53]，即所谓的校准问题，导致代理输出与人类意图不一致[54]。它有时会回答荒谬的内容，这也是目前发现的最为普遍的问题，即对于不知道或不确定的事实，它会强行根据用户的输入毫无根据地展开论述[55]，产生偏离事实的文本。

(5)可解释性差。黑盒特性使得 ChatGPT 的回答看似合理但却无迹可寻，同时由于它没有办法通过充足的理由去解释它的回答是否正确，导致在一些需要精确、严谨的领域没有办法很好的应用[56]。此外，它也可能在表述的时候存在语法错误或不合理的表述。

(6)无法在线更新近期知识。目前的范式增加新知识的方式只能通过重新训练大模型。现有研究探索了利用外部知识源来补充大模型[57]，利用检索插件来访问最新的信息源[58]，然而，这种做法似乎仍然停留在表面上。研究结果表明，很难直接修正内在知识或将特定知识注入大模型，这仍然是一个悬而未决的研究问题[59]。

4 ChatGPT 的进化展望

ChatGPT 在自然语言处理技术的发展中有里程碑式的意义，在语言和意图理解、推理、记忆以及情感迁移方面具有强大的能力，在决策和计划方面表现出色，只需一个任务描述或演示，就可以有效地处理以前未见过的任务。此外，ChatGPT 可以适应不同的语言、文化和领域，具有通用性，减少了复杂的培训过程和数据收集的需要[60]，在各领域得到了广泛应用。

在学术界积极探索能力背后的技术原因的同时，工业界已将 ChatGPT 优异的对话生成能力融入各种场景中，根据对话对象的不同，将应用分为以下几种层次。

(1)数据生成加工。利用 ChatGPT 强大的信息搜索与整合能力，用户根据自身需求直接返回特定数据，主要应用场景包括文案生成、代码生成和对话生成等。同时，其可以充当知识挖掘工具对数据进行再加工，一些在线应用可帮助翻译、润色等，例如文档分析工具 ChatPDF。

(2)模型调度。ChatGPT 可以调用其他机器学习模型共同完成用户需求并输出结果，例如微软近期发布的 HuggingGPT。作为人类与其他模型的智能中台，其有望解决 AI 赋能长期面临的痛点问题，实现模块化模型管理、简化技术集成部署，提高赋能效率。

(3)人机混合交互。ChatGPT 一定程度上统一了人类语言与计算机语言，使得人机交互界面从键盘鼠标图形进化到自然语言接口。微软的 365 Copilot 将其嵌入到 Office，极大地提高了人机自然交互体验；OpenAI 近期发布的 Plugins 插件集尝试了大语言模型应用的开发框架。在未来其有望成为智能时代的操作系统，调用更广泛的应用程序解决实际问题。

结合实际应用中可能存在的问题和实际应用需求，对 ChatGPT 在具体领域存在问题的可能解决方案进行了分析讨论。

(1)商业服务优化领域。提升商品服务质量[61]需要对用户评论进行细粒度情感分析，通常需要对特定商品领域的知识有深入的理解。然而，ChatGPT 作为通用语言模型，可能缺乏不同商品领域中的专业知识，这可能导致模型无法准确识别和分析

特定领域的优缺点。可以考虑利用迁移学习的方法，将 ChatGPT 在通用领域中的知识迁移到特定领域中，使 ChatGPT 更加适应特定领域的问题和需求。

(2)智慧医疗领域。ChatGPT 在医疗领域可以做为辅助工具用作医疗诊断与肿瘤图像分割[62]，有助于精准医疗、靶向治疗等方案的落实。然而，目前 ChatGPT 主要针对文本进行处理，对于其他模态的信息理解相对较弱，这使得模型应用仅限制在辅助诊断和医疗数据挖掘等方面，无法融合其他模态的信息来增强模型通用性与泛化性。因此为了实现更加有效表达的通用人工智能模型，需要进行多模态联合学习，关注内容关联特性与跨模态转换问题。此外，风险责任问题、沟通限制状况以及模型引发的算法偏见与个人隐私安全问题同样不容忽视。

(3)舆情监管引导领域。舆情引导和特定内容生成[63]需要在构建训练数据阶段进行意图对齐和质量筛选。由于 GPT 系列的训练语料来自于西方的语言价值框架，受到模型训练数据的偏见和倾向性影响，ChatGPT 生成内容中存在对于中国的大量偏见言论，不一定符合中国的价值观，这可能引发舆情操纵和认知战[64]的风险。因此训练国产大模型时需要对训练数据进行筛选，构建合适公正的中文语料，并不断维护更新基础词库。

很多研究者认为 ChatGPT 开启了第四次技术革命，其作为催化剂整合人工智能学科，并激发学术界与工业界深入探讨和实践交叉学科与跨学科应用[65]的可能性，科技部近期启动的“AI for Science”专项部署工作也从一定程度上反映了国家导向。未来其从应用拓展上将呈现垂直化、个性化与工程化，如何增强其人机交互协同性，如考虑生物学特性、身体感知等因素；以及如何增强模型可信性，构建新的可信测试基准，都是未来可能的发展趋势。

5　总　结

本文探讨了 ChatGPT 在自然语言处理领域发展中的地位以及未来可能的发展方向，着重分析了 GPT 系列模型的演进以及核心技术，包括语料体系、提示学习、思维链和基于人类反馈的强化学习等。随后，分析了其存在的显著缺陷，如理解与推理能力的局限性、专业知识的不深入、事实的不一致性以及信息安全泄露等风险；最后，结合实际应用，ChatGPT 有着很大的改进和发展空间，包括采用迁移学习和领域适应技术、引入外部专家知识、增强多模态处理能力、筛选训练语料等都是可能的解决方案与发展趋势。通过上述分析，本文对深入理解 ChatGPT 和在相关领域展开进一步研究提供参考。

参考文献：

[1] 郝亚洲，郑庆华，陈艳平，等. 面向网络舆情数据的异常行为识别 [J]. 计算机研究与发展，2016，53(3)：611-620.
HAO Yazhou，ZHENG Qinghua，CHEN Yanping，et al. Recognition of abnormal behavior based on data of public opinion on the web [J]. Journal of Computer Research and Development，2016，53(3)：611-620.

[2] 何炎祥，孙松涛，牛菲菲，等. 用于微博情感分析的一种情感语义增强的深度学习模型 [J]. 计算机学报，2017，40(4)：773-790.
HE Yanxiang，SUN Songtao，NIU Feifei，et al. A deep learning model enhanced with emotion semantics for microblog sentiment analysis [J]. Chinese Journal of Computers，2017，40(4)：773-790.

[3] RUMELHART D E，HINTON G E，WILLIAMS R J. Learning representations by back-propagating errors [J]. Nature，1986，323(6088)：533-536.

[4] 奚雪峰，周国栋. 面向自然语言处理的深度学习研究 [J]. 自动化学报，2016，42(10)：1445-1465.
XI Xuefeng，ZHOU Guodong. A survey on deep learning for natural language processing [J]. Acta Automatica Sinica，2016，42(10)：1445-1465.

[5] QIU Xipeng，SUN Tianxiang，XU Yige，et al. Pre-trained models for natural language processing：a survey [J]. Science China Technological Sciences，2020，63(10)：1872-1897.

[6] WEISS K，KHOSHGOFTAAR T M，WANG Dingding. A survey of transfer learning [J]. Journal of Big Data，2016，3(1)：9.

[7] MENON T，PFEFFER J. Valuing internal vs. external knowledge：explaining the preference for outsiders [J]. Management Science，2003，49(4)：497-513.

[8] BALTRUŠAITIS T，AHUJA C，MORENCY L P. Multimodal machine learning：a survey and taxonomy [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence，2019，41(2)：423-443.

[9] RADFORD A，NARASIMHAN K，SALIMANS T，et al. Improving language understanding by generative pre-training [EB/OL]. [2023-03-14]. https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf.

[10] THOMPSON A D. What's in my AI? [EB/OL]. [2023-03-12]. https://lifearchitect.ai/whats-in-my-ai/.

[11] DEVLIN J, CHANG Mingwei, LEE K, et al. BERT: pre-training of deep bidirectional transformers for language understanding [C]//Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Stroudsburg, PA, USA: ACL, 2019: 4171-4186.

[12] LEWIS M, LIUYinhan, GOYAL N, et al. BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension [C]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Stroudsburg, PA, USA: ACL, 2020: 7871-7880.

[13] RADFORD A, WU J, CHILD R, et al. Language models are unsupervised multitask learners [EB/OL]. [2023-03-06]. https://cdn. openai. com/better-language-models/language _ models _ are _ unsupervised _ multitask_learners. pdf.

[14] LIU Zhuang, LIN W, SHI Ya, et al. A robustly optimized BERT pre-training approach with post-training [C]//Proceedings of the 20th Chinese National Conference on Computational Linguistics. Stroudsburg, PA, USA: ACL, 2021: 1218-1227.

[15] SANH V, WEBSON A, RAFFEL C, et al. Multitask prompted training enables zero-shot task generalization [C]//ICLR 2022. New York, USA: ICLR, 2022: 1-216.

[16] WANG T, ROBERTS A, HESSLOW D, et al. What language model architecture and pretraining objective works best for zero-shot generalization? [C]//Proceedings of the 39th International Conference on Machine Learning. Chia Laguna Resort, Sardinia, Italy: PMLR, 2022: 22964-22984.

[17] LI X L, LIANG P. Prefix-tuning: optimizing continuous prompts for generation [C]//Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing. Stroudsburg, PA, USA: ACL, 2021: 4582-4597.

[18] RADFORD A, WU J, CHILD R, et al. Language models are unsupervised multitask learners [EB/OL]. [2023-04-18]. https://cdn. openai. com/better-language-models/language _ models _ are _ unsupervised _ multitask_learners. pdf.

[19] ZHANG Yizhe, SUN Siqi, GALLEY M, et al. DIALOGPT : large-scale generative pre-training for conversational response generation [C]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations. Stroudsburg, PA, USA: ACL, 2020: 270-278.

[20] GU X, YOO K M, HA J W. Dialogbert: discourse-aware response generation via learning to recover and rank utterances [C]//Proceedings of the AAAI Conference on Artificial Intelligence. Palo Alto,CA,USA: AAAI, 2021, 35(14): 12911-12919.

[21] BROWN T B, MANN B, RYDER N, et al. Language models are few-shot learners [C]//Proceedings of the 34th International Conference on Neural Information Processing Systems. Red Hook, NY, USA: Curran Associates Inc., 2020: 1877-1901.

[22] KAPLAN J, MCCANDLISH S, HENIGHAN T, et al. Scaling laws for neural language models [EB/OL]. (2020-01-23)[2023-04-10]. https://arxiv. org/abs/2001.08361.

[23] HOFFMANN J, BORGEAUD S, MENSCH A, et al. Training compute-optimal large language models [EB/OL]. (2022-03-29)[2023-04-18]. https://arxiv. org/abs/2203.15556.

[24] CHRISTIANO P F, LEIKE J, BROWN T B, et al. Deep reinforcement learning from human preferences [C]//Proceedings of the 31st International Conference on Neural Information Processing Systems. Red Hook, NY, USA: Curran Associates Inc., 2017: 4302-4310.

[25] SCHULMAN J, WOLSKI F, DHARIWAL P, et al. Proximal policy optimization algorithms [EB/OL]. (2017-08-28)[2023-04-30]. https://arxiv. org/abs/1707.06347.

[26] STIENNON N, OUYANG Long, WU J, et al. Learning to summarize from human feedback [C]//Proceedings of the 34th International Conference on Neural Information Processing Systems. Red Hook, NY, USA: Curran Associates Inc., 2020: 3008-3021.

[27] OUYANG Long, WU J, JIANG Xu, et al. Training language models to follow instructions with human feedback [C]//Advances in Neural Information Processing Systems. Red Hook, NY, USA: Curran Associates Inc., 2022: 27730-27744.

[28] SCHICK T, SCHÜTZE H. Exploiting cloze-questions for few-shot text classification and natural language inference [C]//Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume. Stroudsburg, PA, USA: ACL, 2021: 255-269.

[29] WEI J, BOSMA M, ZHAO V, et al. Finetuned language models are zero-shot learners [C]//ICLR 2022.

New York, USA: ICLR, 2021: 6255.

[30] CHUNG H W, HOU Le, LONGPRE S, et al. Scaling instruction-finetuned language models [EB/OL]. (2022-12-06) [2023-05-06]. https://arxiv.org/abs/2210.11416.

[31] MUENNIGHOFF N, WANG T, SUTAWIKA L, et al. Crosslingual generalization through multitask finetuning [C]//Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics. Stroudsburg, PA, USA: ACL, 2023: 15991-16111.

[32] WEI J, TAY Y, BOMMASANI R, et al. Emergent abilities of large language models [EB/OL]. (2022-10-26) [2023-05-28]. https://arxiv.org/abs/2206.07682.

[33] WANG Y, YAO Q, KWOK J T, et al. Generalizing from a few examples: a survey on few-shot learning [J]. ACM computing surveys (csur), 2020, 53(3): 1-34.

[34] WEI J, WANG Xuezhi, SCHUURMANS D, et al. Chain-of-thought prompting elicits reasoning in large language models [C]//Advances in Neural Information Processing Systems. Red Hook, NY, USA: Curran Associates Inc., 2022: 24824-24837.

[35] FU Yao, PENG Hao, KHOT T. How does GPT obtain its ability? Tracing emergent abilities of language models to their sources [EB/OL]. (2022-12-01) [2023-05-21]. https://yaofu.notion.site/How-does-GPT-Obtain-its-Ability-Tracing-Emergent-Abilities-of-Language-Models-to-their-Sources-b9a57ac0fcf74f30a1ab9e3e36fa1dc1.

[36] ZHENG Rui, DOU Shihan, GAO Songyang, et al. Secrets of RLHF in large language models: part Ⅰ PPO [EB/OL]. (2023-06-18) [2023-06-19]. https://arxiv.org/abs/2307.04964.

[37] ZHANG Susan, ROLLER S, GOYAL N, et al. OPT: open pre-trained transformer language models [EB/OL]. (2022-06-21)[2023-04-17]. https://arxiv.org/abs/2205.01068.

[38] ASKELL A, BAI Yuntao, CHEN Anna, et al. A general language assistant as a laboratory for alignment [EB/OL]. (2021-12-09)[2023-06-18]. https://arxiv.org/abs/2112.00861.

[39] BAI Yuntao, JONES A, NDOUSSE K, et al. Training a helpful and harmless assistant with reinforcement learning from human feedback [EB/OL]. (2022-04-12) [2023-06-02]. https://arxiv.org/abs/2204.05862.

[40] GLAESE A, MCALEESE N, TREBACZ M, et al. Improving alignment of dialogue agents via targeted human judgements [EB/OL]. (2022-09-28)[2023-06-09]. https://arxiv.org/abs/2209.14375.

[41] 刘全，翟建伟，章宗长，等. 深度强化学习综述 [J]. 计算机学报，2018，41(1)：1-27.
LIU Quan, ZHAI Jianwei, ZHANG Zongchang, et al. A survey on deep reinforcement learning [J]. Chinese Journal of Computers, 2018, 41(1): 1-27.

[42] RENS G, RASKIN J F. Learning non-Markovian reward models in MDPs [J]. arXiv preprint arXiv:2001.09293, 2020.

[43] YE Junjie, CHEN Xuanting, XU Nuo, et al. A comprehensive capability analysis of GPT-3 and GPT-3.5 series models [EB/OL]. (2023-03-18)[2023-06-11]. https://arxiv.org/abs/2303.10420.

[44] KENTON Z, EVERITT T, WEIDINGER L, et al. Alignment of language agents [EB/OL]. (2021-03-26) [2023-05-18]. https://arxiv.org/abs/2103.14659.

[45] ZHENG Rui, XI Zhiheng, LIU Qin, et al. Characterizing the impacts of instances on robustness [C]//Findings of the Association for Computational Linguistics: ACL 2023. Stroudsburg, PA, USA: ACL, 2023: 2314-2332.

[46] BOLOOR A, GARIMELLA K, HE Xin, et al. Attacking vision-based perception in end-to-end autonomous driving models [J]. Journal of Systems Architecture, 2020, 110: 101766.

[47] GU Tianyu, DOLAN-GAVITT B, GARG S. BadNets: identifying vulnerabilities in the machine learning model supply chain [EB/OL]. (2019-03-11) [2023-06-20]. https://arxiv.org/abs/1708.06733.

[48] CHEN Xiaoyi, SALEM A, CHEN Dingfan, et al. BadNL: backdoor attacks against NLP models with semantic-preserving improvements [C]//Annual Computer Security Applications Conference. New York, NY, USA: Association for Computing Machinery, 2021: 554-569.

[49] PEREZ F, RIBEIRO I. Ignore previous prompt: attack techniques for language models [EB/OL]. (2022-11-17)[2023-05-22]. https://arxiv.org/abs/2211.09527.

[50] BORJI A. A categorical archive of ChatGPT failures [EB/OL]. (2023-04-03)[2023-05-28]. https://arxiv.org/abs/2302.03494.

[51] CALISKAN A, BRYSON J J, NARAYANAN A. Semantics derived automatically from language corpora contain human-like biases [J]. Science, 2017, 356(6334): 183-186.

[52] GUOYue, YANG Yi, ABBASI A. Auto-debias: debiasing masked language models with automated biased prompts [C]//Proceedings of the 60th Annual Meet-

ing of the Association for Computational Linguistics. Stroudsburg, PA, USA: ACL, 2022: 1012-1023.

[53] HUANG Xiaowei, RUAN Wenjie, HUANG Wei, et al. A survey of safety and trustworthiness of large language models through the lens of verification and validation [EB/OL]. (2023-08-27) [2023-06-30]. https://arxiv.org/abs/2305.11391.

[54] QIN Yujia, HU Shengding, LIN Yankai, et al. Tool learning with foundation models [EB/OL]. (2021-06-15) [2023-06-30]. https://arxiv.org/abs/2304.08354.

[55] ZHUO T Y, HUANG Yujin, CHEN Chunyang, et al. Red teaming ChatGPT via jailbreaking: bias, robustness, reliability and toxicity [EB/OL]. (2021-05-29) [2023-06-19]. https://arxiv.org/abs/2301.12867.

[56] 桑基韬, 于剑. 从 ChatGPT 看 AI 未来趋势和挑战 [J]. 计算机研究与发展, 2023, 60(6): 1191-1201.
SANG Jitao, YU Jian. ChatGPT: a glimpse into AI's future [J]. Journal of Computer Research and Development, 2023, 60(6): 1191-1201.

[57] LAZARIDOU A, GRIBOVSKAYA E, STOKOWIEC W, et al. Internet-augmented language models through few-shot prompting for open-domain question answering [EB/OL]. (2022-05-23) [2023-07-01]. https://arxiv.org/abs/2203.05115.

[58] MADAAN A, TANDON N, CLARK P, et al. Memory-assisted prompt editing to improve GPT-3 after deployment [C]//Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing. Stroudsburg, PA, USA: ACL, 2022: 2833-2861.

[59] MENG K, BAU D, ANDONIAN A, et al. Locating and editing factual associations in GPT [C]//Advances in Neural Information Processing Systems. Red Hook, NY, USA: Curran Associates Inc., 2022: 17359-17372.

[60] BANG Yejin, CAHYAWIJAYA S, LEE N, et al. A multitask, multilingual, multimodal evaluation of ChatGPT on reasoning, hallucination, and interactivity [EB/OL]. (2021-02-28) [2023-06-05]. https://arxiv.org/abs/2302.04023.

[61] FELTEN E, RAJ M, SEAMANS R. How will language modelers like ChatGPT affect occupations and industries? [EB/OL]. (2021-03-18) [2023-05-31]. https://arxiv.org/abs/2303.01157.

[62] 孙颖, 丁卫平, 黄嘉爽, 等. RCAR-UNet:基于粗糙通道注意力机制的视网膜血管分割网络 [J]. 计算机研究与发展, 2023, 60(4): 947-961.
SUN Ying, DING Weiping, HUANG Jiashuang, et al. RCAR-UNet: retinal vessels segmentation network based on rough channel attention mechanism [J]. Journal of Computer Research and Development, 2023, 60(4): 947-961.

[63] LI Junyi, TANG Tianyi, ZHAO W X, et al. Pretrained language models for text generation: a survey [EB/OL]. (2021-05-25) [2023-06-22]. https://arxiv.org/abs/2105.10311.

[64] 俞凯, 陈露, 陈博, 等. 任务型人机对话系统中的认知技术: 概念、进展及其未来 [J]. 计算机学报, 2015, 38(12): 2333-2348.
YU Kai, CHEN Lu, CHEN Bo, et al. Cognitive technology in task-oriented dialogue systems: concepts, advances and future [J]. Chinese Journal of Computers, 2015, 38(12): 2333-2348.

[65] 曾毅, 刘成林, 谭铁牛. 类脑智能研究的回顾与展望 [J]. 计算机学报, 2016, 39(1): 212-223.
ZENG Yi, LIU Chenglin, TAN Tieniu. Retrospect and outlook of brain-inspired intelligence research [J]. Chinese Journal of Computers, 2016, 39(1): 212-223.

(编辑 武红江)