·专题:ChatGPT 与人工智能技术应用·

大模型关键技术与未来发展方向

——从 ChatGPT 谈起

刘学博 户保田 陈科海 张 民*

哈尔滨工业大学(深圳)计算与智能研究院,深圳 518055

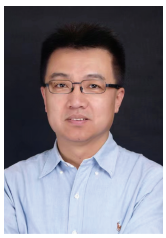
[摘 要] 大规模预训练模型,也被称为“基座模型”或“大模型”,目前被认为是通用人工智能技术的核心引擎,已经成为了全球科技竞争焦点。本文归纳总结了以聊天生成预训练转换器(Chat Generative Pre-trained Transformer, ChatGPT)为代表的生成式大模型技术研究现状和发展趋势,从大模型基座、大模型人类偏好对齐、大模型推理与评价、多模态大模型、大模型安全可控五个方面讨论了当前大模型研究的现状和挑战,并结合我国人工智能研究特点,简要分析了大模型未来的重点发展方向。

[关键词] 人工智能;大模型;ChatGPT;发展方向

人工智能(Artificial Intelligence,AI)是一门涉及人的智能研究理论、方法、技术及应用系统的新兴技术科学,主要研究和开发人的智能模拟、延伸和扩展技术,AI 技术正逐步改变现代社会。智能科学与技术,也就是深入理解智能机理并结合先进 AI 技术,成为推动人工智能持续发展的关键驱动力。大规模预训练语言模型,也被称为“基座模型”或“大模型”,其特点在于拥有巨大的参数量,构成了复杂的人工神经网络模型[1]。这项技术的提出和实践具有划时代的意义,它标志着人工智能的研究步入了通用人工智能时代。

大模型具有规模性(参数量大)、涌现性(产生预料之外的新能力)以及通用性(不仅局限于特定问题或领域)等特性。以 ChatGPT[2] 为代表的生成式大模型因其具有巨量的参数和深度网络结构,能学习并理解更多的特征和模式,从而在处理复杂任务时展现出惊人的自然语言理解、意图识别、推理、上下文建模、语言生成等几乎所有和自然语言相关的处理能力,同时具有通用问题求解能力,被视作通往通用人工智能的一条重要路径[3]。

“基座模型”这个词汇形象地描绘了大模型的作用:它是坚实的基础底座,拥有极其强大的 AI 能力。但这些能力需要被激发出来,才能为各类任务和应用提供强力的支持。因此,大模型已经转变为 AI 领域的基础设施,为解决各种复杂问题提供底层强大的计算、学习和求解能力。随着科技的发展,大模型正逐渐成为一种新的科学研究范式。从初期的大语



张民 哈尔滨工业大学(深圳)特聘校长助理,计算与智能研究院院长。“国家百千万人才工程”入选者,国家杰出青年科学基金获得者,“鹏城孔雀计划”特聘 A 岗位,享受国务院政府特殊津贴。主要研究方向为自然语言处理、人工智能、大模型。获省部级科技进步奖 3 项,最佳会议论文 4 次。发表 CCF A/B 类会议和期刊论文 200 余篇,出版 Springer 专著 2 部,主编论著(论文集)16 本。担任本领域 10 本期刊编委。主持科技部重点研发计划课题及多项大型产业界项目。

刘学博 哈尔滨工业大学(深圳)计算与智能研究院助理教授,“鹏城孔雀计划”特聘 C 岗位。主要研究方向为自然语言处理、机器翻译、大模型能力评估与优化。获澳门技术发明奖二等奖、澳门研究生科技研发奖、中国中文信息学会优博提名奖等奖项。在自然语言处理与人工智能顶级会议与期刊上发表论文 30 余篇。主持国家自然科学基金青年科学基金项目等。

收稿日期:2023-08-18;修回日期:2023-09-18

* 通信作者,Email: zhangmin2021@hit.edu.cn

本文受到国家自然科学基金项目(62261160648、62036004)的资助。

言模型，已经延伸到了多模态、语音、图像、视频等各个领域，甚至用于各类复杂问题的求解，例如天气预报[4]、石油勘探[5]、智慧城市[6]等复杂系统，更有效地完成复杂系统的建模与预测。

然而，大模型技术还处于初级研究阶段，存在许多亟需解决的问题，包括但不限于模型的可解释性、模型机理的研究、与现实世界的可交互性、安全可控、伦理道德问题，以及如何更好地对接下游任务等。另一方面，作为核心技术的基座模型，更强的自主可控和建模能力是我国下一代大模型技术基础研究的两大核心任务。

1　大模型技术及研究进展

1.1　大模型基座

2020 年 OpenAI 首次提出“规模定律”，指出模型的性能随着参数量、数据量、训练时长的指数级增加而呈现出线性提升，并且该提升对架构和优化超参数的依赖性非常弱[7]。从此研究人员逐步转移研究重心至大语言模型基座，并开展了大量相关研究。首个千亿模型 GPT-3[8]在各种自然语言处理任务上取得了突破性的成果，包括文本生成、机器翻译、问答等，并展现了在零样本和少样本情况下的泛化性。GPT 系列模型的发展标志着大型预训练语言模型时代的到来。除了 GPT 系列模型，谷歌、Meta 等公司同样开始不断发布百亿到千亿的大型语言模型，例如 Gopher、Chinchilla、PaLM，但是这些模型都不开源。当前代表性的开源大模型有 Meta 的 OPT、LLaMA-2 以及国内的 GLM-130B[9]、ChatGLM2。发展示意图如图 1 所示。

在模型架构方面，国内外的大模型普遍为 Transformer 架构。模型的基座设计大体上可以分为以下三种：(1) 仅包含解码器(Decoder-only)，即自回归(Autoregressive)模型，代表模型是 GPT[10]和 LLaMA[11]，其训练目标是从左到右的文本生成，常用于无条件长文本生成，如对话生成、故事生成等；(2) 仅包含编码器(Encoder-only)，即自编码(Autoencoder)模型，代表模型是 BERT[12]、ALBERT[13]、DeBERTa[14]，自编码模型是通过去噪任务(如利用掩码语言模型)学习双向的上下文编码器，训练目标是对文本进行随机掩码，然后预测被掩码的词，常用于自然语言理解，如事实推断、语法分析、文本分类等；(3) 编码器—解码器(Encoder-Decoder)，即完整的 Transformer 结构，代表模型是 T5[15]和 BART[16]，包含一个编码器和一个解码器，接受一段文本，从左到右地生成另一段文本，常用于有条件的生成任务，如机器翻译、摘要生成、事实性对话等。考虑到训练效率、推理需求和下游实际应用任务，大模型通常采用仅包含解码器的架构，通过自回归预训练高效地生成优质内容。

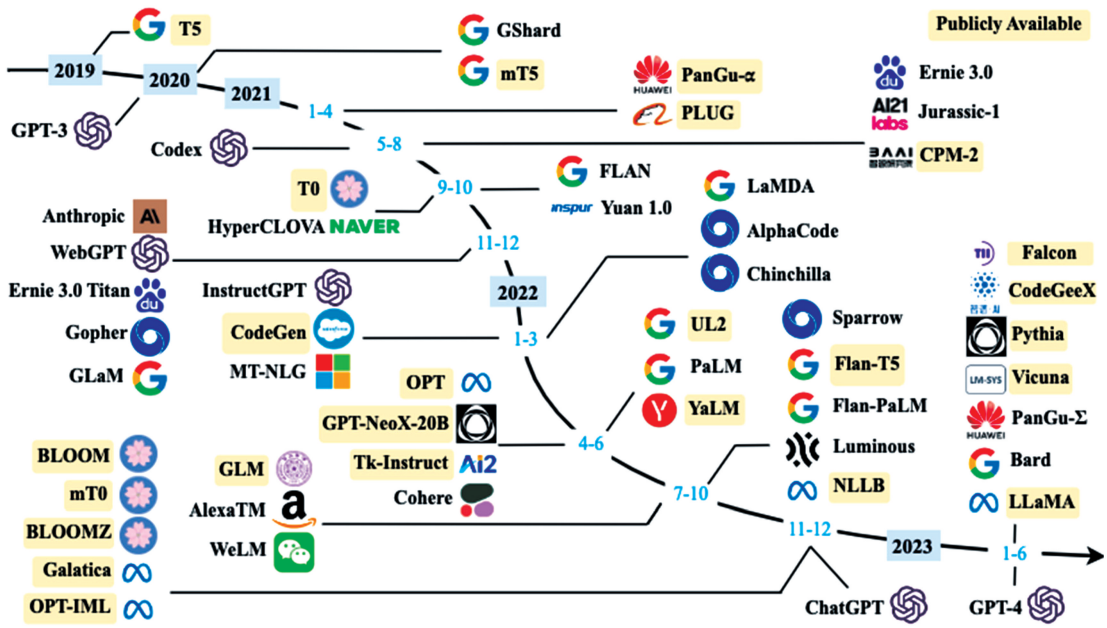


图 1　国内外大模型基座发展示意图①

① https://github.com/RUCAIBox/LLMSurvey

在训练数据上，我国开放给大模型的数据集主要是中文语料库，但在数据量、内容多样性和质量方面，仍有进一步提升的空间。截至目前，常见的开源预训练数据集有：GLM 系列的悟道数据集，其规模为 3 TB(已开源 200 GB)；CLUE 社区的开源中文数据集 CLUE Corpus 2020[17]，其规模为 100 GB；里屋社区的开源数据集 MNBVC，其规模约 2.3 TB，为互联网收集的中文纯文本语料数据集。相比较而言，国外开源数据集数量更多。例如，PB 级的 CommonCrawl 的网页数据、1.6 TB 多语数据集 ROOTS[18]、825 G 的数据集 The Pile[19] 等。此外，国外开源数据集内容较丰富：ROOTS 既包含网页数据，又收集了 GitHub 上的代码数据，也从各种下游任务数据集中收集高质量数据；The Pile 数据集基于学术或专业领域知识源构造，质量较高；另有以英文小说为主的数据集 BookCorpus[20] 及百科全书数据集 Wikipedia。

1.2 大模型人类偏好对齐

大模型在预训练阶段的主要任务是将世界知识融入模型中，是模型学习知识的过程。对齐大模型与人类偏好的目标是激发模型理解、适应人类意愿和解决问题的能力，强调的是使模型能够有效地应用预训练阶段获取的知识，从而使其具有多样化的能力，能够解决各种问题。另一方面，大模型在训练阶段可能会学习到数据中的偏见和歧视性信息，导致模型的行为表现出预期外的特征。为了纠正模型的表现，使模型反映出人类的价值观，避免出现不可预测的输出，需要实现大模型与人类偏好的对齐。目前主要通过两种方法实现：有监督微调和人类反馈的强化学习算法(Reinforcement Learning from Human Feedback，RLHF)，如图 2 所示。

有监督微调[21](Supervised Fine-tuning，SFT)是主要的大模型人类偏好对齐方法。该过程利用人类偏好一致的指令数据来训练大模型。首先，需要收集或创建指令数据，这些数据由输入/输出对组成。其中输入数据是提供给模型的指令或提示，而输出数据则是期望模型根据人类偏好来生成的响应，通常由人类专家标注。然后，通过这些格式化的指令数据，以监督学习的方式对大语言模型进行微调。这种有监督微调方法是一种相对直接且有效的手段，能激发语言模型的深层次理解能力，以更好地实现与人类偏好的对齐。

另一部分是人类反馈的强化学习算法[22]，通过利用人类标注、答案重排序等技术构造符合人类偏好的数据训练一个奖励模型(Reward Model)，由奖励模型提供指导信号，这些信号反映了人类对大模型生成的文本的偏好，通常以标量值的形式出现。基于奖励模型的指导信号，利用强化学习(Proximal Policy Optimization，PPO)来优化学习过程，待优化的大语言模型的动作域(Action Space)是预测词表，状态为当前生成的内容，并将奖励模型的反馈信号通过 PPO 算法传给大语言模型进行优化。最终实现大模型与人类的偏好对齐。

尽管大语言模型在多种任务中表现出强大的能力，但它们也存在生成“幻觉”内容的倾向，生成与用户输入、之前的上下文或者已知的世界知识不一致的内容。这一挑战对大模型在实际应用中的可靠性构成威胁。幻觉问题不是新现象，最初在机器翻译

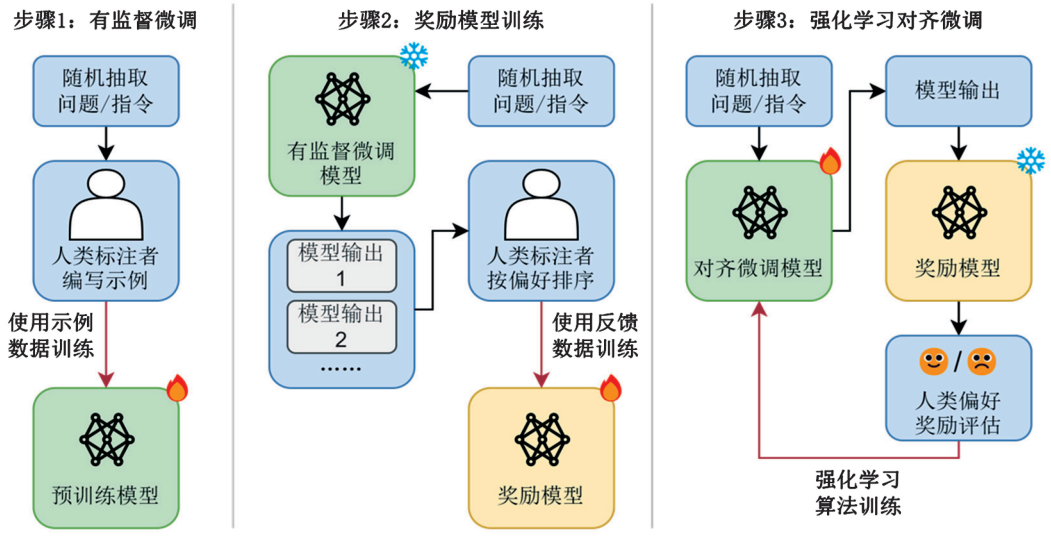


图 2 大模型人类偏好对齐示意图

系统中已被提出[23]。但在大模型环境下，这个问题变得更为复杂。幻觉不仅对用户信任造成破坏，还能通过简单的搜索轻易地被触发。为减少幻觉的出现，研究人员已经采用了包括数据增强和动态系统在内的多种方法，尝试降低大模型幻觉内容生成的频率[24]。

大模型具有强大的通用性，但往往缺乏特定领域的专业知识。为解决这些问题，已有研究提出结合内外部知识，利用模型自身的通用能力从外部知识库中检索相关信息，同时提供完整的检索路径以增加可解释性[25]。另一方面，在执行复杂任务时，可以通过工具调用、链式思维、搜索决策树等方式增强模型的规划和推理能力[26]。这些方法不仅提高了大模型在特定任务中的表现，也为其在实际应用中的可靠性和可解释性提供了有力支持。

1.3　大模型推理与评价

在大模型的实际应用场景中，推理效率和生成质量是两个关键的维度。一方面，大模型的高效推理是实现工程应用的关键技术。和训练环节相比，推理环节在计算精度、算力消耗量等方面的要求较低，但依然依赖于高性能的 GPU 显卡。此外，显存瓶颈、通信延迟和硬件内存带宽约束仍然限制着模型的推理效率。另一方面，模型训练阶段常用模型损失作为评价模型性能好坏的基准。然而，这种单一维度的评价方法在实际应用中往往无法全面反映模型在多维度任务性能上的优劣，因此需要对模型的评价进行更加精细的设置。

在大模型推理加速方面，一种有效的策略是对模型框架和运算进行优化。例如，NVIDIA 公司研发的 Fast Transformer 框架采用了 MPI 和 NCCL 来实现节点间通信①。Fast Transformer 框架可通过模型并行来支持跨 GPU 和节点的高效推理，通过算子融合和缓冲区分配减少了 GPU 调用次数和重新计算成本。这种策略能从硬件底层显著提升推理效率，但其复杂性可能会增加用户的使用门槛。另一种策略是采用模型压缩技术，例如模型量化。模型量化通过使用低精度数值来近似表示网络权重和激活值，以达到减小模型体积的目的。在训练时，为保证训练的效果，模型权重一般保持在 FP32。在推理时，使用 FP16 权重能达到与 FP32 类似的精度，同时减少一半的 GPU 显存需求。更进一步的量化方案，如 LLM.int8[27]，将线性运算拆分为 INT8 和 FP16 两部分，分别进行计算后相加，在不牺牲模型性能的前提下降低模型的运行成本；OPTQ[28]基于 Hessian 矩阵而非传统的基于统计的方法进行一次性权重量化，可以在大约 4 个 GPU 小时内量化具有 1 750 亿参数的 GPT 模型，将位宽减少到每个权重的 3 或 4 位。

在大模型评价方面，目前主要分为人工评价与自动评价两种方式。自动评价作为一种普遍且广泛应用的评估机制，一般依赖于预定的标准化指标和工具来评估模型的性能。例如以 MMLU[29]、CEVAL[30]等为代表的数据集，通过模型在单项选择题上的准确性来估计其潜在能力；以 GSM8K[31]、HumanEval[32]为代表的数据集，通过测试模型对数学题的回答准确率与代码的测试样例通过率来衡量模型的推理能力；以 BLEU[33]、COMET[34]、ROUGE[35]等生成类评价指标用于评价模型的生成能力。然而，这些自动评价指标常面临评价不全面、数据泄露等问题。特别是在开放式任务场景下，传统的自动评价机制通常不能全面地衡量生成结果的质量，因此人工评价和使用高级模型（如 GPT-4）进行的评价成为了更可靠的评估方式。人工评价通过人类专家的参与评价模型生成结果的质量和准确性。与自动评价相比，人工评价更接近实际应用场景，可以提供更全面和准确的反馈，但仍然存在主观性、差异性和不稳定性等问题。在实际应用中，具体使用哪种评价方式需要根据具体的使用场景进行综合考量。

1.4　多模态大模型

多模态大模型（Multimodal Large Models）通过整合多种类型的数据（如文本、图像、音频等），提升机器理解和生成复杂内容的能力。早期的多模态模型通常需要在特定数据集上微调才能胜任相关的任务，如图文检索双塔模型（Contrastive Language-image Pre-training，CLIP）[36]和图文生成模型（Object-Semantics Aligned，Oscar）[37]等。当前的多模态大模型具有更强的通用问题求解能力，主要分为以下三种：

一种方法是将大语言模型作为中央处理器来执行多模态任务，通过调用其他功能模块来实现任务目标。例如 Visual ChatGPT[38]将 ChatGPT 作为中央处理器，借助提示管理模块与用户进行交流，根据用户的需求读取图像并调用外部视觉专家模型修改

① https://github.com/NVIDIA/FasterTransformer

图像、输出结果；HuggingGPT[39] 将 ChatGPT 和 Huggingface 连接在一起，利用 Huggingface 中成百上千的模型，能够解决 20 余种不同任务，这种策略有效地实现了多模态信息的处理和理解，但由于需要调用多个任务特定模型，它可能导致信息处理效率低下且部署成本较高。

另一种方法是直接通过图像和文本信息训练多模态大模型。如 KOSMOS-1[40]，它利用视觉和文本编码器对输入进行编码，然后将这些编码输入到基于 Transformer 的解码器中，使得图像信息和文本信息直接对齐并融合，从而实现跨模态交互。这种方法在处理多模态信息时能够实现优异的准确性和效率，为解决多模态任务提供了技术支持。然而，这种方法通常面临预训练数据不足和训练成本较高的挑战。

最后一种方法，如图 3 所示，结合跨模态编码器等结构与大语言模型，能进一步发掘大模型的推理检索能力和存储的知识库信息。例如，LLaVA[41] 由大语言模型和 CLIP 的开源视觉编码器和语言解码器连接而成，从而实现更加广义上的视觉语言理解。LMEye[42] 构建了静态视觉映射网络，为大语言模型提供图像的基础感知。大模型解析人类指令后，将其发送到交互式感知网络，并基于交融的多模态信息产生响应。

1.5 大模型安全可控

大模型安全可控主要集中于大模型的训练和推理两方面。针对训练阶段的可控研究主要通过对预训练语言模型进行网络重构、修改训练任务或增加微调任务以实现有约束的生成过程。早期研究在预训练文本序列首部添加多种表征文本信息的特殊符号以实现可控生成过程[43]。另有研究者不限于知识符号，而是基于多个人类评价维度（如有效性、安全性）对模型进行可控微调[44]。近期一些研究使用基于人类反馈的强化学习策略推进大模型的自主可控性，通过使用奖励模型学习人类评价模式，进而对大模型进行自动微调[45,46]。针对推理阶段，典型研究通过在推理过程中增加约束信息[47] 或是直接针对模型输入输出增加控制模块[48] 以有效实现有约束的生成过程。近期，通过在输入中增加显示或隐式控制信息作为 Prompt 的做法同样取得了较好的效果[49,50]。

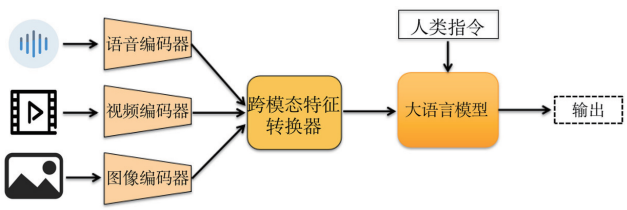


图 3 基于大语言模型的多模态大模型通用结构

在大模型安全性方面，生成式大模型面临着包括模型窃取、数据窃取、对抗攻击、后门攻击、Prompt 攻击和数据投毒等多方面威胁。在模型窃取方面，近期研究发现可通过本地模型访问 OpenAI 的 API 部分窃取现有大模型在特定任务上的性能。在数据窃取方面，存在一种差分隐私训练策略避免使用者进行大模型的数据窃取[51]。在对抗攻击方面，研究发现大模型对于对抗性文本和分布外文本的抵御效果优于传统模型，但依然存在鲁棒性不足的问题[52,53]。在后门攻击方面，研究发现通过在人类反馈强化学习的奖励模型训练阶段增加后门[54]，可以通过后门触发文本控制模型输出；另外，可通过大模型产生包含后门触发器的训练数据[55]，从而对其他模型植入后门。在 Prompt 攻击方面，有研究者设计了一套通过大模型生成恶意 Prompt 的攻击流程[56]，可达到绕过大模型安全限制、下游应用 Prompt 窃取等恶意攻击目的。在数据投毒方面，可以借助大模型实现指令微调数据的自动投毒，从而操纵或毒害其他模型[57]。

2 大模型领域未来重点发展方向

大模型需要多方合作发展，包括产、学、研、用、资、政等多个领域，对提升我国科技核心竞争力具有关键性作用。在此，我们选取除了算力以外我国大模型发展的三个具有代表性方向进行讨论。

2.1 自然语言引领大模型基础通用理论

大模型随着模型参数和训练数据的增加，由量变到质变，涌现出通用智能的能力，使人类真正从信息社会进入智能社会。自然语言在大模型中发挥着重要的引领作用，自然语言是传递和表达语义认知和知识的最重要方式，通过处理自然语言数据，大模型可以学习到丰富的语义表示和世界知识。本方向主要包括：

（1）下一代大模型基础架构。利用丰富的外部知识，建立数据与知识双轮驱动研究新范式。以中文为核心、以通用人工智能为目标，设计更加高效、准确、可扩展的新一代语言模型，并以此为基础搭建新一代人工智能理论框架体系。

（2）大模型可解释性和模型机理。目标在于突

破“黑箱”问题的束缚，实现大模型行为的动态追踪、知识提取过程的深度分析以及决策过程的人类干预，从而提升模型可解释性，建立可解释、鲁棒的人工智能理论和方法。同时深入剖析大规模预训练语言模型的实现机理，以揭示涌现现象背后的科学原理，完善理论体系。

（3）大模型的持续学习与演化能力。研究增量学习技术和动态知识库，使模型能够持续适应新数据、更新知识与表达，并通过强化学习技术使模型能够根据环境反馈进行自我改进。探索通用人工智能驱动的智能算法，从而实现模型自主学习与人机协同学习的持续演化。

2.2　多模态大模型智能交互方法

针对多模态数据之间复杂且多样的关系，多模态大模型需对不同模态间的相关性关系进行有效的对齐和交互，以增强对多模态信息的表征能力。基于提取的多模态表征，多模态大模型还需进一步对多模态信息进行交互融合处理以及语义理解，并根据具体要求进行输出决策。在多模态大模型的部署阶段，由于当前多模态大模型复杂度极高，限制了其在硬件资源欠缺条件下的应用，无法满足不同环境下的智能交互需求。本方向主要包括：

（1）大模型驱动的多模态信息表征和理解。研究如何通过多种预训练任务对不同模态数据间的相关性进行不同粒度的对齐和交互，有效增强大模型对多模态信息的表征能力。改进理解任务相关的多模态特征融合技术，利用自监督学习、半监督学习、元学习、迁移学习等新型学习范式，提升模型鲁棒性和学习效率。

（2）基于具身学习的多模态大模型。通过高效的人机交互、融合感知、执行和交互等技术，帮助多模态大模型更好地理解真实世界，获取实时的环境反馈；结合具身学习研究面向智能机器人的多模态大模型构建与应用方法。

（3）轻量化多模态大模型的设计。通过面向硬件条件限制的多模态大模型设计，减少模型规模以及计算和存储需求，从而满足不同的硬件环境，扩大大模型的应用场景。研究模型剪枝、模型量化，以及知识蒸馏等深度模型压缩方法，实现自适应的轻量化多模态大模型设计。

2.3　大模型安全理论与实践

大模型的发展和应用必须着重考虑安全性和可控性。尤其在涉及用户隐私、数据安全、道德规范和合法合规的情况下，大模型的操作必须符合社会规则和伦理道德，必须具有正确的价值观。大模型在理解和生成内容时可能出现偏见，这可能导致信息误导、产生虚假信息，或被恶意利用。本方向主要包括：

（1）大模型供应链安全。针对大模型训练数据易受污染或被投毒的挑战，研究大模型数据审查方法，可为大语言模型及多模态大模型的训练提供安全保障。针对大模型中可能存在后门的问题，研究大模型后门检测方法，可为大模型下游微调和部署提供安全防护。

（2）大模型安全性评估。研究大模型安全性评估方法，全面分析多样的安全性度量场景，构建生成式大模型的安全度量指标体系和大模型安全评估平台，研究实现对大模型的一站式安全风险评估，为大模型进行迭代升级指明具体优化的方向。

（3）大模型生成内容安全。对大模型生成内容的安全性进行深入研究，旨在从模型本身和输出内容审查两个维度增强大模型网络意识形态的安全性。为防止大模型在意识形态方面产生不适当的内容，研究构建一种网络意识形态审查系统，探讨如何在大模型时代实现生成内容的安全防护，达到对大模型生成内容实施有效监管和审查的目的。

3　结论与展望

大模型技术开启了通用人工智能时代，具有划时代意义，将重新定义信息社会。本文基于我国大模型技术的研究现状，探讨了大模型基础理论、智能交互方法、安全理论与实践中的重点发展方向。大模型技术研究刚刚起步，还有非常多亟待解决的问题，其红利和贡献还远未被发掘。总之，从基础研究角度看，基座模型和下一代大模型技术的自主可控是目前我国大模型研究的两大核心任务。

参　考　文　献

[1] Bommasani R, Hudson DA, Adeli E, et al. On the opportunities and risks of foundation models. (2021-08-16)/[2023-08-17]. https://arxiv.org/abs/2108.07258.

[2] Ouyang L, Wu J, Jiang X, et al. Training language models to follow instructions with human feedback. Advances in Neural Information Processing Systems, 2022: 27730—27744.

[3] Zhao WX, Zhou K, Li J, et al. A survey of large language models. (2023-03-31)/[2023-08-17]. https://arxiv.org/abs/2303.18223.

[4] Bi K, Xie L, Zhang H, et al. Accurate medium-range global weather forecasting with 3D neural networks. Nature, 2023, 619(7970): 533—538.

[5] Ogundare O, Madasu S, Wiggins N. Industrial engineering with large language models: a case study of ChatGPT's performance on Oil & Gas problems. (2023-04-27)/[2023-08-17]. https://arxiv.org/abs/2304.14354.

[6] Xi Z, Chen W, Guo X, et al. The rise and potential of large language model based agents: a survey. (2023-09-14)/[2023-09-18]. https://arxiv.org/abs/2309.07864.

[7] Kaplan J, McCandlish S, Henighan T, et al. Scaling laws for neural language models. (2020-01-23)/[2023-08-17]. https://arxiv.org/abs/2001.08361.

[8] Brown T, Mann B, Ryder N, et al. Language models are few-shot learners. Advances in Neural Information Processing Systems, 2020: 1877—1901.

[9] Zeng A, Liu X, Du Z, et al. Glm-130b: an open bilingual pre-trained model. (2022-10-05)/[2023-08-17]. https://arxiv.org/abs/2210.02414.

[10] Radford A, Narasimhan K, Salimans T, et al. Improving language understanding by generative pre-training. (2018-06-11)/[2023-09-11]. https://www.mikecaptain.com/resources/pdf/GPT-1.pdf.

[11] Touvron H, Lavril T, Izacard G, et al. LLaMA: open and efficient foundation language models. (2023-02-27)/[2023-08-17]. https://arxiv.org/abs/2302.13971.

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

[13] Lan ZZ, Chen MD, Goodman S, et al. ALBERT: a lite BERT for self-supervised learning of language representations. (2019-09-26)/[2023-08-17]. https://arxiv.org/abs/1909.11942.

[14] He PC, Liu XD, Gao JF, et al. DeBERTa: decoding-enhanced BERT with disentangled attention. (2020-06-05)/[2023-08-17]. https://arxiv.org/abs/2006.03654.

[15] Raffel C, Shazeer N, Roberts A, et al. Exploring the limits of transfer learning with a unified text-to-text transformer. The Journal of Machine Learning Research, 2020, 21(1): 5485—5551.

[16] 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.

[17] Xu L, Zhang XW, Dong QQ. CLUECorpus2020: a large-scale Chinese corpus for pre-training language model. (2020-03-05)/[2023-08-17]. https://arxiv.org/abs/2003.01355.

[18] Laurençon H, Saulnier L, Wang T, et al. The bigscience roots corpus: a 1.6 tb composite multilingual dataset. Advances in Neural Information Processing Systems, 2022: 31809—31826.

[19] Gao L, Biderman S, Black S, et al. The pile: an 800 GB dataset of diverse text for language modeling. (2020-12-31)/[2023-08-17]. https://arxiv.org/abs/2101.00027.

[20] Zhu YK, Kiros R, Zemel R, et al. Aligning books and movies: towards story-like visual explanations by watching movies and reading books// Proceedings of the 2015 IEEE International Conference on Computer Vision. New York: IEEE, 2016: 19—27.

[21] Longpre S, Hou L, Vu T, et al. The flan collection: designing data and methods for effective instruction tuning. (2023-01-31)/[2023-08-17]. https://arxiv.org/abs/2301.13688.

[22] Christiano PF, Leike J, Brown T, et al. Deep reinforcement learning from human preferences. Advances in Neural Information Processing Systems, 2017: 4302—4310.

[23] Lee K, Firat O, Agarwal A, et al. Hallucinations in neural machine translation. (2023-03-06)/[2023-08-17]. https://openreview.net/forum?id=SkxJ-309FQ.

[24] Zhang Y, Li Y, Cui L, et al. Siren's song in the ai ocean: a survey on hallucination in large language models. (2023-09-03)/[2023-09-17]. https://arxiv.org/abs/2309.01219.

[25] Feng C, Zhang X, Fei Z. Knowledge solver: teaching llms to search for domain knowledge from knowledge graphs. (2023-09-06)/[2023-09-17]. https://arxiv.org/abs/2309.03118.

[26] Qin Y, Liang S, Ye Y, et al. Toolllm: facilitating large language models to master 16000+ real-world apis. (2023-07-31)/[2023-08-17]. https://arxiv.org/abs/2307.16789.

[27] Dettmers T, Lewis M, Belkada Y, et al. Gpt3.int8(): 8-bit matrix multiplication for transformers at scale. Advances in Neural Information Processing Systems, 2022: 30318—30332.

[28] Frantar E, Ashkboos S, Hoefler T, et al. Optq: accurate quantization for generative pre-trained transformers. (2022-10-31)/[2023-08-17]. https://arxiv.org/abs/2210.17323.

[29] Hendrycks D, Burns C, Basart S, et al. Measuring massive multitask language understanding. (2020-09-07)/[2023-08-17]. https://arxiv.org/abs/2009.03300.

[30] Huang Y, Bai Y, Zhu Z, et al. C-eval: a multi-level multi-discipline chinese evaluation suite for foundation models. (2023-05-17)/[2023-08-17]. https://arxiv.org/abs/2305.08322.

[31] Cobbe K, Kosaraju V, Bavarian M, et al. Training verifiers to solve math word problems. (2021-11-18)/[2023-08-17]. https://arxiv.org/abs/2110.14168.

[32] Chen M, Tworek J, Jun H, et al. Evaluating large language models trained on code. (2021-07-14)/[2023-08-17]. https://arxiv.org/abs/2107.03374.

[33] 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. Philadelphia: Association for Computational Linguistics, 2002: 311—318.

[34] Rei R, Stewart C, Farinha AC, et al. COMET: a neural framework for MT evaluation// Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing. Stroudsburg: Association for Computational Linguistics, 2020: 2685—2702.

[35] Lin CY. 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.

[36] Radford A, Kim JW, Hallacy C, et al. Learning transferable visual models from natural language supervision. (2021-02-26)/[2023-08-17]. https://arxiv.org/abs/2103.00020.

[37] Li XJ, Yin X, Li CY, et al. Oscar: object-semantics aligned pre-training for vision-language tasks// Proceedings of the 2020 European Conference on Computer Vision. Cham: Springer, 2020: 121—137.

[38] Wu C, Yin S, Qi W, et al. Visual chatgpt: talking, drawing and editing with visual foundation models. (2023-03-08)/[2023-08-17]. https://arxiv.org/abs/2303.04671.

[39] Shen YL, Song KT, Tan X, et al. Hugginggpt: solving AI tasks with chatgpt and its friends in huggingface. (2023-03-30)/[2023-08-17]. https://arxiv.org/abs/2303.17580.

[40] Huang S, Dong L, Wang W, et al. Language is not all you need: aligning perception with language models. (2023-03-01)/[2023-08-17]. https://arxiv.org/abs/2302.14045.

[41] Liu H, Li C, Wu Q, et al. Visual instruction tuning. (2023-04-17)/[2023-08-17]. https://arxiv.org/abs/2304.08485.

[42] Li Y, Hu B, Chen X, et al. Lmeye: an interactive perception network for large language models. (2023-05-05)/[2023-08-17]. https://arxiv.org/abs/2305.03701.

[43] Keskar NS, McCann B, Varshney LR, et al. CTRL: a conditional transformer language model for controllable generation. (2019-09-20)/[2023-08-17]. https://arxiv.org/abs/1909.05858.

[44] Peng B, Li C, He P, et al. Instruction tuning with gpt-4. (2023-04-06)/[2023-08-17]. https://arxiv.org/abs/2304.03277.

[45] Bai Y, Kadavath S, Kundu S, et al. Constitutional AI: harmlessness from AI feedback. (2022-12-15)/[2023-08-17]. https://arxiv.org/abs/2212.08073.

[46] Cao Y, Li S, Liu Y, et al. A comprehensive survey of AI-generated content (AIGC): a history of generative ai from gan to chatgpt. (2023-03-07)/[2023-08-17]. https://arxiv.org/abs/2303.04226.

[47] Dathathri S, Madotto A, Lan J, et al. Plug and play language models: a simple approach to controlled text generation. (2019-12-04)/[2023-08-17]. https://arxiv.org/abs/1912.02164.

[48] Roller S, Dinan E, Goyal N, et al. Recipes for building an open-domain chatbot// Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics. Stroudsburg: Association for Computational Linguistics, 2021: 300—325.

[49] 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, 55(9): 1—35.

[50] Yang KX, Liu D, Lei WQ, et al. Tailor: a soft-prompt-based approach to attribute-based controlled text generation// Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics. Toronto: Association for Computational Linguistics, 2023: 410—427.

[51] Li Z, Wang C, Ma P, et al. On the feasibility of specialized ability stealing for large language code models. (2023-03-06)/[2023-08-17]. https://arxiv.org/abs/2303.03012.

[52] Collins KM, Wong C, Feng J, et al. Structured, flexible, and robust: benchmarking and improving large language models towards more human-like behavior in out-of-distribution reasoning tasks. (2022-05-11)/[2023-08-17]. https://arxiv.org/abs/2205.05718.

[53] Wang J, Hu X, Hou W, et al. On the robustness of chatgpt: an adversarial and out-of-distribution perspective. (2023-02-22)/[2023-09-17]. https://arxiv.org/abs/2302.12095.

[54] Shi J, Liu Y, Zhou P, et al. Badgpt: exploring security vulnerabilities of chatgpt via backdoor attacks to instructgpt. (2023-02-21)/[2023-08-17]. https://arxiv.org/abs/2304.12298.

[55] Li J, Yang Y, Wu Z, et al. Chatgpt as an attack tool: stealthy textual backdoor attack via blackbox generative model trigger. (2023-04-27)/[2023-08-17]. https://arxiv.org/abs/2304.14475.

[56] Liu Y, Deng G, Li Y, et al. Prompt injection attack against llm-integrated applications. (2023-06-08)/[2023-08-17]. https://arxiv.org/abs/2306.05499.

[57] Shu M, Wang J, Zhu C, et al. On the exploitability of instruction tuning. (2023-06-28)/[2023-08-17]. https://arxiv.org/abs/2306.17194.

Key Technologies and Future Development Directions of Large Language Models: Insights From ChatGPT

Xuebo Liu Baotian Hu Kehai Chen Min Zhang*

Institute of Computing and Intelligence, Harbin Institute of Technology, Shenzhen 518055

Abstract Large pre-trained models, also known as "foundation models" or "large models", are the core technical support for generative artificial intelligence models, and have become a key focus in global technological competition. This paper summarizes the current status and development trends of large model technology represented by ChatGPT. The main challenges faced by large model research are explored from five aspects: the foundation of large models, alignment of large models with human preferences, large model inference and evaluation, multimodal large models, and safety and control of large models. Based on the characteristics of China's AI field, the potential key development directions for the large model field in the future are also analyzed.

Keywords artificial intelligence; large language model; ChatGPT; future directions

(责任编辑 崔国增 张 强)

* Corresponding Author, Email: zhangmin2021@hit.edu.cn