

大语言模型研究现状与趋势

王耀祖^{1,2)}, 李 擎^{3,4)}✉, 戴张杰⁵⁾, 徐 越^{1,2)}

1) 北京科技大学智能科学与技术学院, 北京 100083 2) 北京科技大学人工智能研究院, 北京 100083 3) 北京科技大学自动化学院, 北京 100083 4) 北京科技大学工业过程知识自动化教育部重点实验室, 北京 100083 5) 北京科技大学冶金与生态工程学院, 北京 100083

✉通信作者, E-mail: liqing@ies.ustb.edu.cn

摘 要 在过去 20 年中, 语言建模(Language models, LM)已经成为一种主要方法, 用于语言理解和生成, 同时作为自然语言处理(Natural language processing, NLP)领域下游的关键技术受到广泛关注. 近年来, 大语言模型(Large language models, LLMs), 例如 ChatGPT 等技术, 取得了显著进展, 对人工智能乃至其他领域的变革和发展产生了深远的影响. 鉴于 LLMs 迅猛的发展, 本文首先对 LLMs 相关技术架构和模型规模等方面的演进历程进行了全面综述, 总结了模型训练方法、优化技术以及评估手段. 随后, 分析了 LLMs 在教育、医疗、金融、工业等领域的应用现状, 同时讨论了它们的优势和局限性. 此外, 还探讨了大语言模型针对社会伦理、隐私和安全等方面引发的安全性与一致性问题及技术措施. 最后, 展望了大语言模型未来的研究趋势, 包括模型的规模与效能、多模态处理、社会影响等方面的发展方向. 本文通过全面分析当前研究状况和未来走向, 旨在为研究者提供关于大语言模型的深刻见解和启发, 以推动该领域的进一步发展.

关键词 大语言模型 (LLMs); 自然语言处理; 深度学习; 人工智能; ChatGPT

分类号 TP18

Current status and trends in large language modeling research

WANG Yaozu^{1,2)}, LI Qing^{3,4)}✉, DAI Zhangjie⁵⁾, XU Yue^{1,2)}

1) School of Intelligence Science and Technology, University of Science and Technology Beijing, Beijing 100083, China

2) Institute of Artificial Intelligence, University of Science and Technology Beijing, Beijing 100083, China

3) School of Automation and Electrical Engineering, University of Science and Technology Beijing, Beijing 100083, China

4) Key Laboratory of Knowledge Automation for Industrial Processes of Ministry of Education, University of Science and Technology Beijing, Beijing 100083, China

5) School of Metallurgical and Ecological Engineering, University of Science and Technology Beijing, Beijing 100083, China

✉Corresponding author, E-mail: liqing@ies.ustb.edu.cn

ABSTRACT Over the past two decades, language modeling (LM) has emerged as a primary methodology for language understanding and generation. This technology has become a cornerstone within the field of natural language processing (NLP). At its core, LM is designed to train models to predict the probability of the next word or token, thereby generating natural and fluent language. The advent of large language models (LLMs), such as Bidirectional Encoder Representations from Transformers and GPT-3, marks a significant milestone in the evolution of LM. These LLMs have left a profound impact on the field of artificial intelligence (AI) while also paving the way for advancements in other domains. This progression underscores the power and efficacy of AI, illustrating how the landscape of AI research has been reshaped by the rapid advancement of LLMs. This paper provides a comprehensive review of the evolution of LLMs, focusing on the technical architecture, model scale, training methods, optimization techniques, and evaluation metrics. Language

收稿日期: 2023–10–09

基金项目: 国家自然科学基金资助项目(52204335); 北京市科技新星计划资助项目(Z211100002121115); 北京科技大学青年教师学科交叉研究项目(中央高校基本科研业务费专项资金)资助项目(FRF-IDRY-22-004)

models have evolved significantly over time, starting from initial statistical language models, moving onto neural network-based models, and now embracing the era of advanced pre-trained language models. As the scale of these models has expanded, so has their performance in language understanding and generation. This has led to notable results across various sectors, including education, healthcare, finance, and industry. However, the application of LLMs also presents certain challenges, such as data quality, model generalization capabilities, and computational resources. This paper delves into these issues, providing an analysis of the strengths and limitations of LLMs. Furthermore, the rise of LLMs has sparked a series of ethical, privacy, and security concerns. For instance, LLMs may generate discriminatory, false, or misleading information, infringe on personal privacy, or even be exploited for malicious activities such as cyber-attacks. To tackle these issues, this paper explores relevant technical measures, such as model interpretability, privacy protection, and security assessments. Ultimately, the paper outlines potential future research trends of LLMs. With ongoing enhancements to model scale and efficiency, LLMs are expected to play an even greater role in multimodal processing and societal impact. For example, by integrating information from different modalities, such as images and sound, LLMs can better understand and generate language. Additionally, they can be employed for societal impact assessment, providing support for policy formulation and decision-making. By thoroughly analyzing the current state of research and potential future directions, this paper aims to offer researchers valuable insights and inspiration regarding LLMs, thereby fostering further advancement in the field.

KEY WORDS large language models (LLMs); natural language processing; ChatGPT; deep learning; artificial intelligence

自 20 世纪 50 年代图灵测试提出以来, 人工智能领域不懈追求语言智能. 语言的复杂性、语法规则及算法挑战构成了研究难题. 深度学习和神经网络技术因其在表示学习和预测的强大能力, 且与 NLP 领域的高维、无监督及大数据特性相契合而得到应用. 这一进展显著提升了计算机处理文本数据的能力, 推动了 NLP 技术的发展^[1-2].

语言建模技术从统计模型发展至神经模型, 成为探究语言理解与生成的重要手段. 近期, 基于 Transformer 在大规模语料库的预训练语言模型 (Pre-trained language models, PLMs) 展现出优越的 NLP 任务处理能力. 研究表明, 模型性能随参数规模增加而提升, 并引发了关于参数规模效应的广泛讨论^[3-4]. 值得注意的是, 超大规模模型 (如百亿或千亿参数模型) 不仅性能提升, 还展现出独特的能力, 例如上下文学习. 为此, 学界提出了“大语言模型”术语, 以区分这些参数规模巨大的 PLMs^[5-6], 它们在语言建模和 NLP 任务中的应用为研究提供了新的视角和工具.

随着 LLMs 的快速发展, 特别是生成式预训练模型如 GPT 系列的快速迭代, 整个行业迎来了快速增长与研发浪潮^[7]. 在 OpenAI 开发的 GPT-X 模型的推动下^[8], LLMs 已在多项 NLP 任务中展现出最卓越的性能. 随着训练数据量的不断扩大、训练方法的持续创新以及 LLMs 结构的日益复杂, 这些模型已经在教育、政务、金融和生物医药等多个行业场景中得到了广泛的应用. 在当前的人工智能领域, LLMs 已显示出显著的涌现能力, 这一进展无不彰显出人工智能技术的强大与有效, AI 研

究领域正因 LLMs 的快速发展而发生革命性的变化. 虽然 LLMs 对社会各界发展产生影响, 但其工作原理尚未充分探索, 模型本身难以解释. 其次, 由于对计算机资源需求量巨大, 模型训练与消融实验成本巨大. 此外, 正是由于 LLMs 出色性能, 如何有效与高效的控制其生成内容质量以及如何将 LLMs 与人类价值观与偏好保持一致是当前所面临的挑战.

为了充分面对现有的机遇与挑战, 应当需要更多的关注 LLMs 研究与发展. 因此, 本文正是在这一背景下, 旨在为研究者提供关于大语言模型的深刻见解和启发. 首先总结了自然语言处理的发展历程, 并从三种基本语言模型维度概述了 LLMs 的技术架构与模型规模等演变形式; 其次, 针对现有的模型训练方式与优化技术等进行了详细阐述; 随后, 总结了 LLMs 在教育、医疗、金融等多行业领域中的应用效果, 阐述了当前技术的局限性及面临的挑战; 最后, 讨论了应用安全性与一致性相关内容, 并对未来发展方向进行了展望.

1 LLMs 技术架构与模型规模

1.1 LLMs 发展背景

图 1 中汇总了 NLP 的整个发展阶段, 随着研究方法迭代升级其性能逐步上升. LM 最早可以追溯到 20 世纪 50 ~ 60 年代, 早期研究集中于语言翻译与基本语法分析, 该阶段主要基于研究人员手动编写规则来处理文本, 因此难以处理大规模高复杂度的语言数据. 80 ~ 90 年代, 随着计算能力的提高和大规模文本语料库的增加, 统计方法开始

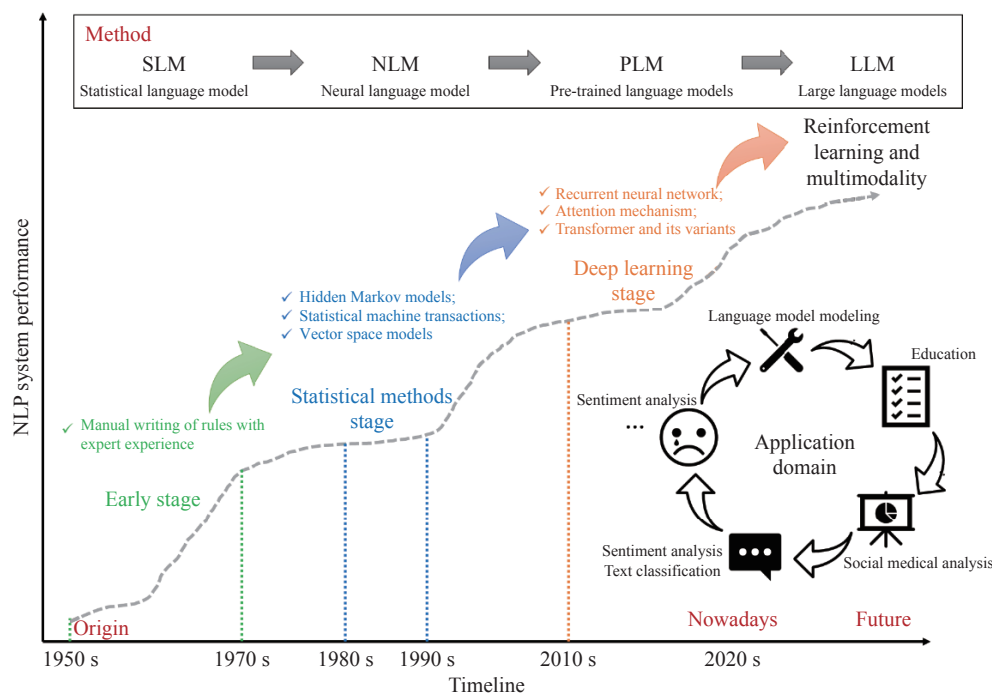


图1 NLP发展阶段概述以及应用

Fig.1 Overview of NLP development stages and applications

在 NLP 中占据主导地位, 统计机器翻译 (Statistical machine translation, SMT)^[9-10] 成为当下主流方法. 2010 年, 随着深度学习的快速兴起, 注意力机制与 Transformer 等模型的出现, 对 NLP 领域产生了巨大影响, 并在文本理解、生成和机器翻译等任务中取得了显著进展. 近年来, 强化学习与多模态应用逐渐增多, 涉及到处理文本以外的信息, 如图像、音频和视频, 以实现更全面的语言理解和生成^[11-12].

LM 旨在对语言的内隐知识进行表示, 其核心问题在于对给定文本序列的合理性判断与量化^[13]. 随着半监督学习和预训练思想的引入, PLMs 成为 LM 的研究焦点^[14]. 近年来, 随着应用场景的增多与复杂性的提高, 致使模型参数量大幅增加, 因此一种具有大量参数和卓越学习能力的高级语言模型——LLMs 备受关注, 并在 NLP 以及其他领域中提供了更多重要的优势和机会^[15-16]. 通用人工智能 (Artificial general intelligence, AGI)^[17] 作为一种能够像人类一样思考、学习和执行多种任务的人工智能系统在上世纪 50 年代被提出, 用以解决更为广泛的问题的多种任务智能系统. 如今, LLMs 对 AI 社区产生了重大影响, ChatGPT 和 GPT-4 的出现促使人们重新思考 AGI 的可能性.

1.2 LLMs 技术架构

2017 年, Google 提出一种基于自注意力机制的特征提取器 Transformer^[18], 取代了原有的循环

神经网络 (Recurrent neural network, RNN) 结构^[19], 并作为基础单元出现于所有 LLMs. 2018 年, GPT-1 问世, 并提出“通用大模型+特定任务微调”的新范式. 为了应对不同场景中自然语言生成或理解等方面需求, LLMs 在训练策略、模型架构和用例方面有所不同. 目前, 主流框架可分为 Encoder-decoder、Encoder-only、Decoder-only 三种^[20], 如图 2 所示, 其中, x 表示原始输入序列, $x_t (t = 1, 2, \dots, T)$ 表示第 t 个标记, T 为序列长度, $M(x)$ 是 x 的掩膜标记, S 表示序列嵌入的开始标记. p_1 、 p_2 、 p_3 、 p_4 分别表示第一个到第四个的位置嵌入标记, P 是条件概率. i 和 j 分别表示编码器输入的起始索引和结束索引. 图 3 中对现有的 LLMs 演化路线进行了汇总. $P(x_t | x_{<t})$ 表示在给历史信息 (即 $x < t$) 的条件下, 预测出下一标记 x_t 的条件概率. $x_{\setminus M(x)}$ 表示除了掩膜标记外的所有标记, $P(x_t | x_{\setminus M(x)})$ 表示在给定序列中除掩膜标记外的所有标记, 对 x_t 预测的条件概率. $x_{i:j}$ 表示序列中从位置 i 到位置 j 的序列. 图 3 中对现有的 LLMs 演化路线进行了汇总.

1.2.1 Encoder-only

BERT (Bidirectional encoder representations from transformers) 作为一个典型的 Encoder-only 架构模型. 相较于传统从左到右的单向编码器, BERT 的双向编码允许模型在处理每个标记时考虑其左侧和右侧的上下文. 通过这种方式, BERT 能够捕捉

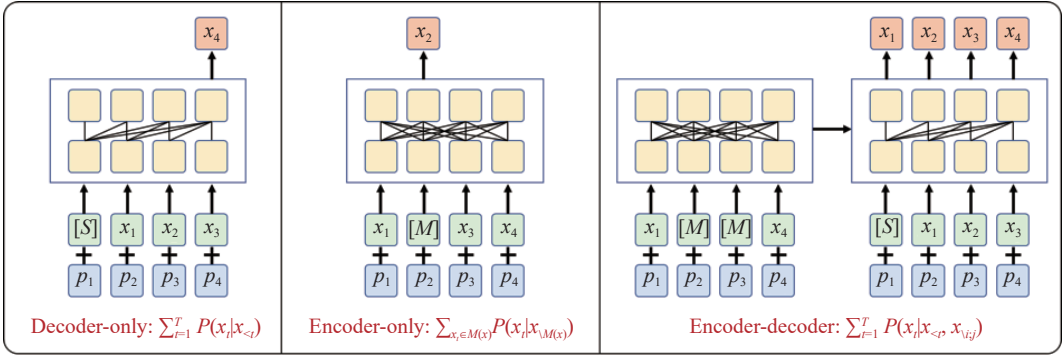


图 2 当前流行的模型训练框架^[21]
Fig.2 Existing prevalent model training frameworks^[21]

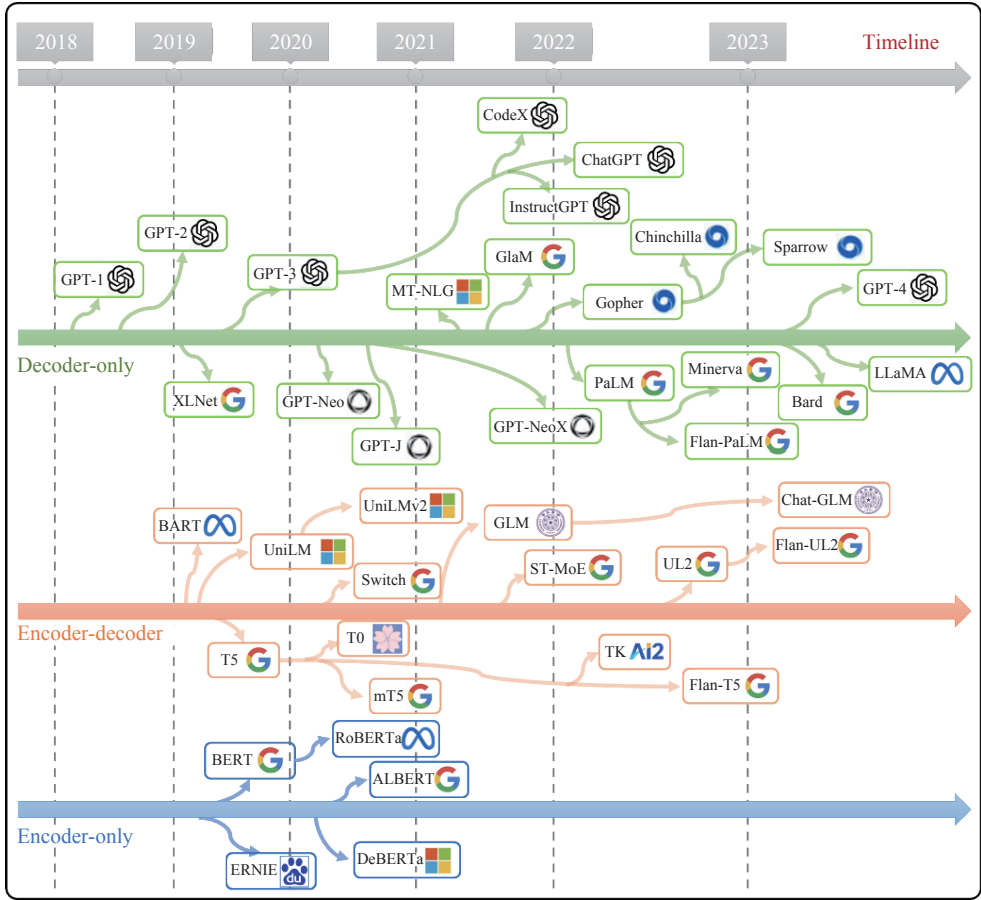


图 3 近年来 LLMs 的进化路线
Fig.3 LLM evolutionary routes in recent years

词汇和语义的丰富信息,更擅文本分类,情感分析,命名实体识别等自然语言理解(Natural language understanding, NLU)任务.随着 BERT 的成功应用,针对该框架的各种变体相继出现. RoBERTa 在 BERT 的基础上去除 NSP(Next sentence prediction), 并采用动态掩码以及更多的语料库进行预训练. ALBERT 于 2020 年被提出,在保持性能的基础上,大大减少了模型的参数,并作为 BERT 的一种经典变体.

1.2.2 Decoder-only

Decoder-only 架构使用自注意机制进行从左到右的单向字序列处理,并主要用于语言生成(Natural-language generation, NLG)任务,将嵌入向量映射回文本空间,生成与上下文相关的响应.自从 GPT-3 问世以来,Decoder-only 模型得到了充分发展,其中“GPT 系列”是其主要代表. GPT-1 通过直接在已完成预训练的模型上进行下游微调,调整输入或输出层便可适应于不同任务. GPT-1 的优

化方向主要集中在扩大训练集数量和增加模型参数上。Instruct GPT 则采用了指示学习和人工反馈的强化学习方法来引导模型训练。虽然基本保留了 GPT-3 的结构,但经过了长达一年的人工训练,参数数量仅为 13 亿的 Instruct GPT 在输出效果上超越了参数数量为 1750 亿的 GPT-3。ChatGPT 在 Instruct GPT 的基础上优化了数据标注的方法,而 GPT-4 则进一步支持多模态数据。此外,基于 Decoder-only 的模型还包括 PaLM^[22]、LLaMA^[23] 等。

1.2.3 Encoder-decoder

Encoder-decoder 架构专注于 NLG 任务。与理解文本的 NLU 不同,NLG 旨在根据特定输入生成连贯、有意义和类似人类的自然语言表达。拥有双向和自回归的 Transformers 架构(BART^[24])以及文本到文本迁移的 Transformer(T5^[25])都是 Encoder-decoder 的典型代表。BART 通过引入噪声来分解原始训练文本,从而扩大了模型在处理更长文本时的推理能力。T5 同样采用了完整 Transformer 结构的预训练语言模型,并提出一个统一的模型框架。将翻译、分类、回归、摘要生成等任务都统一转成“Text-to-Text”任务,从而使得这些任务在训练时能够使用相同的目标函数,在测试时也能使用相同的解码过程。

1.3 LLMs 模型规模

LLMs 的参数数量是影响模型性能重要因素之一,增加模型参数数量通常会提高其性能,这是因为更大的模型可以捕获更多的语言信息和上下文,从而在各种语言理解能力、生成能力及迁移学习能力上表现更好。表 1 汇总对比了国内外一些典型的 LLMs 的参数演进。

GPT-1^[8] 作为 OpenAI 开发的一系列模型中的首个版本,目标在于对单序列文本生成式任务进行服务。在 GPT-1 中,Attention 维数从 512 扩大到 768,将注意力的头数从 8 层增加到 12 层,而前馈隐含层维数从 2048 增加到 3072,总参数为 1.17 亿。由于 GPT-1 参数数量较小,致使对语言理解和生成能力相对有限,难以处理复杂的语言任务和更大规模的语料库。并且作为单向语言模型,导致在某些任务中对上下文的理解可能不够全面。GPT-2^[26] 相比于 GPT-1 模型参数数量从 1.17 亿进化到 15 亿,数据集从 5 GB 扩展到 40 GB,上文窗口大小从 512 增加到 1024,使得 GPT-2 泛化能力更加突出。在 OpenAI 发布的后续版本中,GPT-3^[27] 参数规模为 1750 亿,ChatGPT 也达到了千亿级,训练的数据量达到了几百 TB。2022 年,OpenAI 推出了最新的语言模型——GPT-4,据统计该模型参数数量大约在千亿级别,样本量较 GPT-3 也有较大扩展。与此同时,在 GPT-1 推出的同一时期,Google 发布了 BERT 模型,其参数数量为 3.40 亿。2021 年 12 月,Google 推出了一种新的模型 GLaM^[28],其参数规模达到了 1200 亿,样本量的大小共有 1.6 万亿的标记。2022 年,Google 公司推出了更大规模的语言模型——LaMDA^[29],其参数规模达到了 1370 亿,样本大小达到了 1.56 TB^[30]。

在 LLMs 快速发展的热潮中,国内各大机构也相继对其进行深入研究。根据中国电子信息产业发展研究院的报道,2021 年我国开始探索大型模型应用,并逐步涌现出一批在行业中产生显著影响的大模型。2023 年,受 ChatGPT 影响,大模型发展迈向新阶段,国产大模型一时间呈现爆发式增长态势。截至 2023 年 7 月,我国累计已经发布

表 1 国内外典型 LLMs 参数量对比

Table 1 Comparison of the number of typical LLM participants in China and abroad

Models	Release time	Developers	Parameter size/10 ⁸	Sample size/10 ⁹
GPT-1	2018	OpenAI	1.17	10
BERT	2018	Google	3.40	34
GPT-2	2019	OpenAI	15.00	100
Fairseq	2020	Meta	130.00	—
GPT-3	2020	OpenAI	1750.00	4990
GLaM	2021	Google	1200.00	16000
LaMDA	2022	Google	1370.00	15600
GPT-4	2023	OpenAI	—	—
Ernie Bot	2023	Baidu	—	—
SparkDesk	2023	iFLYTEK	1700.00	—
PanguLM	2023	HUAWEI	—	>30000

130 个大模型, 包括文心一言(百度)、清言(智谱华章)、星火(科大讯飞)、盘古(华为)、云雀(抖音)、紫东太初(中科院)以及书生(上海人工智能实验室)等众多优秀大模型。文心大模型 3.0 在 2023 年 3 月推出, 其参数量达到了 2600 亿, 相对 GPT-3 的参数量提升 50%。星火大模型包含超过 1700 亿个参数, 训练数据囊括了“中国科技论文与引文数据库”(CSTPCD), 包括计算机科学、物理学、化学、生物学等众多学科领域。盘古大模型样本大小超过了 3 万亿的标记^[31], 其基础模型就包括有 100 亿、380 亿、710 亿和 1000 亿参数的四种版本。

模型参数的大幅增加有助于提升 LLMs 的性能, 但与此同时, 研究者还发现随着模型参数的增长, 模型面临计算资源消耗增加、内存和推理速度变慢及过拟合等诸多挑战。OpenAI 发布的 GPT 系列每隔一代, 参数的规模呈现爆炸性增长, 大幅增加了模型训练成本。研究指出^[32], 2023 年 1 月份 ChatGPT 官网总访问量为 6.16 亿次, 如果单位算力的成本固定, 则 ChatGPT 单月所需要的算力约为 4874.4 PFlop/s-day(一天内执行的总浮点运算次数, 用于衡量计算集群每日浮点运算能力)。此外, 推理速度可能会由于模型参数的增加变的缓慢, Dettmers 等^[33]的研究指出, 随着模型参数的增加, 推理速度呈现出爆炸性增长, 性能大幅降低^[34]。除此之外, 随着参数量增多, LLMs 可能会出现过拟合的问题, 例如参数量仅为 13 亿的 Instruct GPT 在输出效果上超越了参数量为 1750 亿的 GPT-3。Weidinger 等^[35]提出 LLMs 会随着模型参数的增加出现过拟合的风险, 比如数据的真实性和偏见会随着模型规模的扩大而增加。Askell 等^[36]发现 LLMs 易在较大参数量下产生更多的过拟合问题, 并影响公平性。

2 LLMs 训练与优化技术

2.1 训练数据集

数据是 LLMs 成功的关键要素, 模型性能与通用性很大程度上取决于数据。表 2 中汇总了一些典型 LLMs 所采用的数据集。现有数据主要来自互联网, 由于信息来源的复杂性, 数据清洗成为了重要的处理步骤^[37-39]。为不同数据按照质量情况赋予权值差异, 使高权值数据在训练中更易抽样, 并以高质量数据为正例对, 同时减少数据冗余, 以提升数据质量。随着数据量增多, LLMs 可能会从训练的数据中得到一些偏见或者不公平性, 而致使

性能较差。目前, 可通过向模型中提供一定的提示, 提高回答的准确率^[36, 40]。在数据去重方面^[41], 也已提出了多种较为先进的处理算法。Abbas 等^[42]采用局部敏感分析方法处理较为模糊的重复, 这部分还包括有 MinHash 和 SimHash 的预训练数据集处理^[43-44]。

尽管 LLMs 训练数据集在一定程度上得到发展, 但依据存在许多问题。首先, 目前大多数训练数据处于不公开状态, 而随着 LLMs 在各个领域的推广应用, 提高数据集的透明度, 增强数据集的可访问性具有极大意义。其次, 数据集中对于使用者信息等敏感内容不能做到完全去除, 这对隐私性带来了较大的负面影响, 研究人员仍需针对数据加密等隐私保护技术深入研究, 降低信息泄露风险。互联网信息中不当言论及敏感性问题在数据中无法被完全消除, 对于 LLMs 数据集的筛选和审核需要采取更加严格的技术措施, 确保模型的安全性、一致性。

2.2 训练方法优化

2.2.1 预训练技术

预训练模型可以使用大量无标注语料, 通过自监督学习的方式来学习语言中的共性特征。这一技术自 2013 年 Word2Vec 模型提出后并广泛应用于词向量表示学习。2018 年, ELMo^[52]被提出并能够捕捉文本语料中的复杂语义信息和语法特征。同年 BERT 基于 Transformer, 以更大数据规模和复杂模型结构为基础无监督得学习语言表示。在此基础上, RoBERTa、XLNet 等模型采用不同预训练技术, 提高效率和性能。这一技术使模型更适合资源匮乏语言, 允许使用大量无标注数据进行预训练, 再在少量标注数据上微调, 以提高性能。

2.2.2 提示学习技术

为了充分利用大语言模型的特点, 避免训练所有模型参数, 近年来自然语言处理研究开始关注提示学习(Prompt learning), 如图 4 所示。提示是添加到语言模型输入中的信息, 用于调整输入数据的形式, 使其更接近预训练阶段的数据形式^[53]。提示学习的目标是减小预训练和微调阶段数据形式之间的差异, 以使模型能够高效地用于下游任务。提示按照其形式可分为离散提示(Discrete prompt)^[54]和连续提示(Continuous prompt)^[55], 它们在大语言模型中均有广泛的应用。

离散提示研究分为两条路线。一种是自动搜索和构建最优的离散提示的方法, 例如梯度搜索。而另一种是通过人工设计离散提示的形式, 以激

表 2 国内外典型 LLMs 训练数据集

Table 2 Training datasets of typical LLMs

Models	Datasets	Size	Total	Ref.
ChatGPT-1	BookCorpus	4.6 GB	4.6 GB	[8]
ChatGPT-2	Reddit links	40 GB	40 GB	[26]
ChatGPT-3	Wikipedia	11.4 GB	753.4 GB	[27]
	Books1	21 GB		
	Books2	101 GB		
	WebText2	50 GB		
	Common Crawl	570 GB		
GPT-J/ GPT-NeoX-20B	Common Crawl	570 GB	825.18 GB	[45–46]
	PubMed Central	90.27 GB		
	Books3	100.96 GB		
	OpenWebText2	62.77 GB		
	ArXiv	56.21 GB		
	Github	95.16 GB		
	FreeLaw	51.15 GB		
	Stack Exchange	32.20 GB		
	USPTO Background	22.90 GB		
	PubMed Abstracts	19.26 GB		
	Gutenberg	10.88 GB		
	OpenSubtitles	12.98 GB		
	Wikipedia	6.38 GB		
	DM Mathematics	7.75 GB		
	Ubuntu IRC	5.52 GB		
	BookCorpus2	6.30 GB		
	EuroParl	4.59 GB		
	HackerNews	3.90 GB		
	YouTubeSubtitles	3.73 GB		
	PhilPapers	2.38 GB		
Megatron-11B/ RoBERTa	NIH ExPorter	1.89 GB	161 GB	[47]
	Enron Emails	0.88 GB		
	Wikipedia	11.4 GB		
	BookCorpus	4.6 GB		
	Common Crawl	107 GB		
Gopher	OpenWebText	38 GB	10550 GB	[48–49]
	MassiveWeb	1900 GB		
	Books	2100 GB		
	C4	750 GB		
	News	2700 GB		
	GitHub	3100 GB		
	Wikipedia	12.5 GB		

表 2 (续)
Table 2 (Continued)

Models	Datasets	Size	Total	Ref.
Ernie Bot	CLUECorpus2020	100 GB	25378.4 GB	[50]
	Chinese multimodal pretraining Data	300 GB		
	WuDaoCorpus2.0	23852 GB		
	PanGuCorpus	1126.4 GB		
	Chinese multimodal pretraining Data	300 GB		
Qwen-VL	LAION-en	2×10^9	5×10^9	[51]
	LAION-COCO	0.6×10^9		
	DataComp	1.4×10^9		
	Coyo	0.7×10^9		
	CC12M	1.2×10^7		
	CC3M	0.3×10^7		
	SBU	0.1×10^7		
	COCO Caption	0.6×10^6		
	LAION-zh	1.8×10^8		
	In-house Data	2.2×10^8		

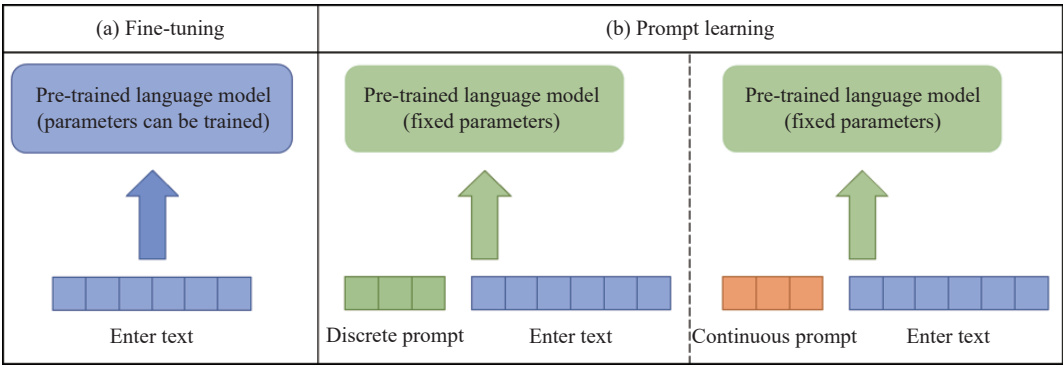


图 4 LLMs 的两种范式. (a) 微调; (b) 提示学习

Fig.4 Two paradigms of LLMs: (a) fine-tuning; (b) cue learning

发大语言模型的能力, 包括上下文学习 (In-context learning)^[27] 和思维链 (Chain-of-thought)^[56]. 上下文学习允许模型在少样本自然语言处理任务中直接求解, 无需更新任何模型参数. 思维链则指模型在推理任务中生成推理过程再生成答案的能力, 可以通过少样本或零样本设定来实现. 连续提示是指提示信息为不受词向量约束的连续向量, 通常包含一些可训练参数, 可在下游任务的训练中进行更新. 通过设计连续提示的插入位置和训练目标来提高模型在下游任务上的性能^[57]. 这些方法通常固定原始模型参数, 仅更新连续提示中的参数, 从而降低了计算资源的开销.

2.2.3 模型微调技术

微调的核心思想是使用一个在大规模文本数

据上进行预训练的语言模型, 并在特定任务或领域的数据集上进一步训练以适应具体任务. 该过程包括使用预训练模型的参数初始化模型, 然后在任务特定的数据集上进行训练. 这个数据集通常包括用户与项目的互动、项目描述、用户信息和其他相关上下文数据. Qiu 等^[58] 提出了一种新的 U-BERT 方法, 结合了预训练和微调, 用于学习用户表征. 该方法利用领域相关信息来增强用户特征, 弥补了行为数据不足的不足. 评论协同匹配层用于捕捉用户和项目评论之间的隐含语义交互. 同样, UserBERT^[59-60] 引入了两个自我监督任务, 以预训练用户模型, 提高用户建模的能力. 通过模型微调, LLMs 可以在特定任务上获得出色的性能, 同时利用其在预训练阶段学到的通用语言知识.

这种方法在自然语言处理中广泛应用于各种任务和应用领域。

2.3 LLMs 模型评估

表 3 中汇总了 LLM 的基本评估任务以及相应的数据库, 其中基本评估任务大致可分为语言生成、知识利用以及复杂推理三种。现有的评估数据集包括机器翻译、文本摘要和问答等^[61], 人们通常使用自动指标配合人工评估以评价性能。最新的模型, 如 GPT-4, 在翻译、新闻摘要任务上, 表现出了与人类自由撰稿人相媲美的水平^[62]。此外, 研究人员还在探索 LLMs 在更具挑战性的语言生成任务中的应用, 如结构化数据生成和长文本生成等^[63]。

知识利用分为闭卷 QA、开卷 QA 和知识完成三种类型。闭卷 QA^[64] 要求 LLMs 仅依靠给定的上下文来回答问题, 而不能使用外部资源^[65], 研究表明闭卷 QA 中, 模型性能与其规模、数据集大小有着紧密联系, 在相似的参数范围内, 拥有更多与评估任务相关的预训练数据的模型表现更佳。开卷 QA 任务允许 LLMs 从外部知识库或文档集合中提取有用的证据, 并根据这些证据回答问题。在开卷 QA 任务的评估中, 通常使用准确度和 F1 分数等指标, LLMs 通常会与文本检索器(甚至搜索引擎)进行配对^[66], 协助 LLMs 验证和修正推理路径^[67-68]。

复杂推理是指利用证据或逻辑得出结论或做出决策的能力, 依赖于逻辑关系和证据^[69]。此外, 研究人员还将知识推理任务转换为代码生成任务, 并发现使用代码预训练的 LLMs 可以进一步提高性能。然而, 由于该任务的复杂性, 当前 LLMs 在常识推理等任务上仍然无法达到与人类的水平。其中最常见的错误之一是, LLMs 可能会基于错误的事实知识生成不准确的中间步骤, 从而导致错误的结果。

3 LLMs 应用技术

3.1 教育

国内外学者普遍认为 LLMs 在教育领域的应用是一把“双刃剑”, 它能够个性化地为学生及教师提供资源和指导, 但同时面临教育公平、知识产权及数据隐私和安全等挑战。Kasneci 等^[14] 对 LLMs 应用于教育领域持乐观态度, 他们从学生和教师角度讨论 LLMs 在教育领域的研究现状及应用前景。从教师的角度来看, LLMs 有望提供更为便捷的教学辅助工具与资源, 例如帮助教师进行课程规划^[89-90]、差异化和个性化教学^[91]、评估和专业发展等, 从而大大提高教师的工作效率。对学生而言, LLMs 有助于学生阅读、写作、数学、计算机和语言技能, 提供个性化练习材料、摘要和解释^[92]。除此之外, 教育公平、知识产权及隐私和数据安全等问题也随着出现^[14]。这就要求教育者在保障学生数据安全和隐私的同时, 需要掌握新技术, 确保其有效整合到课堂教学中。进一步研究和开发 LLMs 教育应用, 提高学生学习体验和成就, 是未来 LLMs 在教育领域的重要方向之一。

3.2 医疗

以 ChatGPT 为例的 LLMs 在多项临床医学检查中展现出了良好的应用前景, 为个性化医疗以及提升健康意识等提供了良好作用^[93]。LLMs 可以协助远程医疗服务, 消除语言及地域障碍, 有效收集患者信息、分析症状, 并提出预诊断, 再经由专业医生进行审核和验证, 这对于远程医疗服务的可扩展性提升, 尤其是在医疗条件不足的地区具有重要意义。此外, LLMs 还可以基于电子健康记录文本训练来帮助撰写出院总结。Liu 等^[94] 提出在放射学医疗方面 Radiology-GPT 在放射学诊断、研

表 3 LLM 的基本评估任务和数据库

Table 3 Basic assessment tasks and databases of LLMs

Tasks		Datasets
Language Generation	Language Modeling	Penn Treebank ^[70] , WikiText-103 ^[71] , the Pile, LAMBADA ^[72] WMT'20 ^[73] , 21 ^[73] , 22 ^[74]
	Conditional Text Generation	CNN/Daily Mail ^[75] , X Sum
	Code Synthesis	APPS, ODEX, MTPB
Knowledge Utilization	Closed-Book QA	ARC ^[76] , ThankfulQA, CWQ, MKQA ^[77] , Science QA
	Open-Book QA	ARC, Web Questions ^[78] , Trivia QA ^[65] , QASC, SQuAD ^[79]
	Knowledge Completion	Freebase ^[80] , LAMA, YAGO3-10 ^[81] YAGO, WordNet
Complex Reasoning	Knowledge Reasoning	COPA ^[82] , SIQA ^[83] , Science QA, ARC, Bool Q ^[84] , PIQA, ProPara
	Symbolic Reasoning	Coin Flip, Reverse List, Last Letter ^[56] , Parity, Repeat Copy
	Mathematical Reasoning	MATH ^[85] , GSM8k ^[86] , SVAMP, DROP, PISA ^[87] , MathQA ^[88]

究和沟通交流方面均有良好表现. 但是在前额叶功能的神经心理学研究方面, Loconte 等^[95]提出 LLMs 与人类前额叶完整性认知尚有差距, 其在前额叶测试中表现出了不同功能的水平差异性. LLMs 在医疗教学中亦可扮演重要角色, Kung 等^[96]对 ChatGPT 进行了美国医疗许可考试 (USMLE) 方面的性能评估, ChatGPT 在没有专门训练或强化的情况下, 在所有三个考试中都达到或接近及格. 尽管 LLMs 应用前景广阔, 可能会导致医疗领域教育、研究和工作流程变革, 但由于模型的可解释性与诊断准确性等问题, 诊断透明度及结果有待进一步研究与优化.

3.3 金融

LLMs 在金融、商业、管理等方向有着广泛应用^[97], 研究人员基于成本、准确性等多方面考虑已开发出多种 LLMs 应用, 如 BloombergGPT, FrugalGPT, FinEval, FinGPT 等^[98]. Lopez-Lira 和 Tang^[99]研究了 LLMs 在使用新闻标题进行股市预测方面的潜力, 结果表明, 将先进的语言模型纳入投资决策过程中, 可以产生更准确的预测, 并提高量化交易胜率. Zhang 等^[100]针对 LLMs 难以准确地解释数值和把握金融背景等问题, 提出了将一小部分金融分析数据转换为指令数据对 LLMs 进行微调, 改善了模型性能. LLMs 在金融领域展现出良好的潜力与价值, 由于该应用领域具有较高的复杂性和准确性要求, 因此, 为金融领域制定具有高准确性、高效率的 LLMs 成为了学者的重要研究方向.

3.4 工业

LLMs 在工业领域的应用涵盖了制造业、供应链管理、设备维护、生产优化等多个方面. LLMs 可以通过处理和分析大量的文本和数据, 有助于优化和改进工业和制造过程、提高生产效率. Yazdinejad 等^[101]将 LLMs 应用在药品供应链管理 (DSCM) 中, 使 ChatGPT 贯通医疗与物流行业, 但依然面临一些挑战, 例如, 如何防止伪造来源、确保药品的可获得性等. LLMs 还可以基于大数据开展工业故障诊断, Xu 等^[102]利用 LLMs 对工业系统的故障诊断问题开展了系统的研究, 并引入了“设备心电图”的概念, 进一步优化了工业诊断方法. 生产过程数据多具有孤立、碎片化等特征, 导致产品设计、制造和服务阶段的效率、智能化和可持续性水平较低, 而 LLMs 的总结、分析能力可以很好的弥补这些问题^[103], 但 LLMs 在工业领域数据隐私和安全及实时调控的能力都有待进一步研究.

4 安全性与一致性技术

4.1 安全性

出于保护或隐藏一些特定的数据收集通道, 或者涉及特定于用户的不可公开获取的数据来源, 因此大模型训练集通常是闭源的. Nasr 等^[104]对语言模型中的“可提取记忆”进行了大规模的研究. 通过成员推理攻击来推断样本是否在训练集中, 以及采用数据提取攻击来恢复完整的训练样本. 除此之外, Khalil 和 Er^[105]研究发现, ChatGPT 可生成不易被剽窃检测软件捕获的复杂文本, 这使得科学文献和社会新闻被剽窃的可能性大大提高, 知识产权问题及数据隐私问题成为大语言模型在研究领域的重要争议及在该领域发展的重大挑战. 在误导性及虚假信息传播方面, 由于大语言模型基于训练数据生成回复信息, 因此模型可能出现无法准确识别或验证信息真实性、正确性的问题. 在多种场景、任务下, 尽管 ChatGPT-4 在标准评估中显示出相较于 GPT-3.5 更高的可信度, 但是研究人员发现更新后模型受到攻击的可能性更大^[106].

4.2 一致性

尽管 LLMs 能够生成高质量的语言表达, 但是模型生成信息的误导问题逐步引起了学术界的关注. 这种生成的事实性错误被称为“幻觉问题”^[107]. 在金融、法律、医学等应用领域中, 错误信息可能会引发负面后果, 而事实性错误的产生往往是由训练数据中的误差和噪声引起^[108]. 由于 LLMs 训练数据的复杂性和矛盾性, 进而导致 LLMs 常常产生与人类价值观或预期目标不一致的输出, 甚至反应不同的观点和偏见, 特别是涉及政治观点、文化观点或宗教种族等敏感性话题时, 可能会产生严重后果^[109]. 此外, 由于深度学习模型的内在机制仍然是“黑箱”, 因此在准确性要求较高的应用场景下, LLMs 的运用应当更加谨慎, 需要通过人工校准进行完善和确认. 现存的对齐工作大多数检测不一致行为的方法或对齐模型行为的方法^[110]. 目前, 研究人员已经逐步开发出有助于完善 LLMs 安全性的应用程序, 例如 Toolformer 和 Plugin7, 并已解决部分事实错误问题, 但是大语言模型在各个行业及学科中的应用和发展仍需进一步探索优化.

确保大语言模型的安全性、一致性是复杂的任务, 需要从多个角度出发, 综合考虑模型技术、法律、伦理和使用者等多方面因素, 提升模型对抗性、稳定性, 发展隐私保护技术并加强隐私信息的

审查和监管, 以进一步提升大语言模型的可靠性和安全性。

5 结论与展望

本文系统回顾了 LLMs 的最新进展, 并深入探讨了理解 and 应用 LLMs 的相关理论. 详细讨论了模型训练的方法、优化技术的应用以及评估手段的选择. 同时, 对其在社会各领域的应用现状进行了分析, 并探讨了它们在实践中的优势. 具体体现在以下方面:

(1) 当前大模型被广泛应用于聊天机器人、计算生物学、计算机编程、创意工作、知识工作、法律、医学、推理、机器人和社会科学等多个领域. LLMs 可能逐渐成为一种新的科学研究范式, 用于处理更复杂的系统和问题。

(2) LLMs 可以被视为基座模型的一种类型, 专注于语言相关的任务. 然而, 随着技术的发展, LLMs 将更加专注于跨模态学习和多任务学习, 以实现更广泛的应用和更高的通用性. 应用于更多模态 (如图像、视频等) 和跨领域的整合是未来的一个重要方向。

随着模型规模的扩展和任务复杂性的增加, 提升模型的可解释性、安全性和伦理兼容性变得尤为紧迫. 同时, 尽管将 LLMs 与多模态学习相结合展现出巨大的潜力, 但这一过程同样伴随着若干挑战, 主要包括以下几个关键问题:

(1) 由于模型参数的增加及训练数据多元化发展, LLMs 面临计算资源需求加大、数据偏差和公平性、模型可解释性及多模态融合等方面挑战, 未来有关的工作可能需要进一步优化模型的框架及训练方法, 如增量学习算法、领域自适应、多模态表示学习、跨模态对齐和融合及多模态生成等技术。

(2) 在个性化和隐私保护方面, 个性化及定制化 LLMs 技术是未来发展的重要方向, 亦是解决大模型训练数据不足问题的关键, 但如何提供个性化的服务同时保护用户隐私、防止滥用用户数据, 将是行业未来面临的关键难题. 有待进一步研究差分隐私、联邦学习、个性化模型融合等关键技术在 LLMs 个性化和隐私保护的应用, 降低数据泄露风险。

(3) 社会伦理和治理层面, 审计、影响评估和认证等问责机制有助于确保人工智能系统道德、法律和技术方面的要求. 然而, 现有的审计程序未能解决 LLMs 带来的治理挑战, 实施有效的问责制

仍然存在挑战和复杂性. 未来, 需要进一步开展伦理审查、道德框架、法律合规性等关键技术研究, 确保大语言模型的道德使用和社会责任。

(4) 目前大多数语言模型的训练数据集以及训练方法处于不公开状态, 为非内部人员训练带来巨大挑战. 而随着 LLMs 在各个领域的推广应用, 提高数据集和训练方法的透明度, 增强数据集的可访问性具有极大意义。

(5) 虽然大模型在文学与法学领域表现出色, 但在数学和物理领域中性能差强人意. 通过为大模型建立可靠评估标准和横纵比较, 分析其局限性, 为大预言模型在微积分与统计学等数理逻辑方面提供优化见解。

未来, 大模型将继续沿着参数规模的增长、多模态学习、跨领域整合、技术创新等方向快速发展. 它们将更好地与现实世界的数据和场景交互, 提供更准确和全面的理解, 并在更多领域发挥重要作用. 同时, 如何解决大模型带来的可解释性、安全可控性和伦理道德问题, 以及提高其与现实世界的交互性和对复杂任务的处理能力, 将是未来研究的重要课题。

参 考 文 献

- [1] Al-Ayyoub M, Nuseir A, Alsmearat K, et al. Deep learning for Arabic NLP: A survey. *J Comput Sci*, 2018, 26: 522
- [2] Ramaswamy S, DeClerck N. Customer perception analysis using deep learning and NLP. *Procedia Comput Sci*, 2018, 140: 170
- [3] Arora K, Rangarajan A. Contrastive entropy: A new evaluation metric for unnormalized language models[J/OL]. *arXiv preprint* (2016-03-11) [2023-10-09]. <https://arxiv.org/abs/1601.00248>
- [4] Goldin-Meadow S, Feldman H. The development of language-like communication without a language model. *Science*, 1977, 197(4301): 401
- [5] Nguyen D Q, Nguyen A T. PhoBERT: Pre-trained language models for Vietnamese [J/OL]. *arXiv preprint* (2020-10-05) [2023-10-09]. <https://arxiv.org/abs/2003.00744>
- [6] Nguyen D Q, Vu T, Nguyen A T. BERTweet: A pre-trained language model for English Tweets [J/OL]. *arXiv preprint* (2020-10-05) [2023-10-09]. <http://arxiv.org/abs/2005.10200.pdf>
- [7] Bommasani R, Hudson D A, Adeli E, et al. On the opportunities and risks of foundation models [J/OL]. *arXiv preprint* (2022-01-12) [2023-10-09]. <https://arxiv.dosf.top/abs/2108.07258>
- [8] Radford A, Narasimhan K, Salimans T, et al. *Improving Language Understanding by Generative Pre-training*. 2018
- [9] Hearne M, Way A. Statistical machine translation: A guide for linguists and translators. *Lang Linguist Compass*, 2011, 5(5): 205
- [10] Koehn P, Zens R, Dyer C, et al. Moses: open source toolkit for

- statistical machine translation // *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*. Morristown, 2007: 177
- [11] Koseki S, Kutsuzawa K, Owaki D, et al. Multimodal bipedal locomotion generation with passive dynamics via deep reinforcement learning. *Front Neurorobot*, 2023, 16: 1054239
- [12] Qi W, Fan H Y, Karimi H R, et al. An adaptive reinforcement learning-based multimodal data fusion framework for human-robot confrontation gaming. *Neural Netw*, 2023, 164: 489
- [13] Bengio Y, Ducharme R, Vincent P. A neural probabilistic language model. *J Mach Learn Res*, 2003(3): 1137
- [14] Kasneci E, Sessler K, Küchemann S, et al. ChatGPT for good? On opportunities and challenges of large language models for education. *Learn Individ Differ*, 2023, 103: 102274
- [15] Kandpal N, Deng H K, Roberts A, et al. Large language models struggle to learn long-tail knowledge // *Proceedings of the 40th International Conference on Machine Learning*. Honolulu, 2023: 15696
- [16] Wang Z H, Wohlwend J, Lei T. Structured pruning of large language models [J/OL]. *arXiv preprint* (2021-04-28) [2023-10-09]. <https://arxiv.org/abs/1910.04732>
- [17] Goertzel B. Artificial general intelligence: Concept, state of the art, and future prospects. *J Artif Gen Intell*, 2014, 5(1): 1
- [18] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need // *Proceedings of the 31st International Conference on Neural Information Processing Systems*. Long Beach, 2017: 6000
- [19] Li Y, Hao Z, Lei H. Survey of convolutional neural network. *J Comput Appl*, 2016, 36(9): 2508
- [20] Wang Y F, Zhong W J, Li L Y, et al. Aligning large language models with human: A survey [J/OL]. *arXiv preprint* (2023-01-24) [2023-10-09]. <https://arxiv.org/abs/2307.12966>
- [21] Wang H F, Li J W, Wu H, et al. Pre-trained language models and their applications. *Engineering*, 2023, 25(6): 51
- [22] Chowdhery A, Narang S, Devlin J, et al. Palm: Scaling language modeling with pathways. *J Mach Learn Res*, 2023, 24(240): 1
- [23] Touvron H, Lavril T, Izacard G, et al. LLaMA: Open and efficient foundation language models [J/OL]. *arXiv preprint* (2023-02-27) [2023-10-09]. <https://arxiv.dosf.top/abs/2302.13971>
- [24] Ou Z B, Zhang M S, Zhang Y. On the role of pre-trained language models in word ordering: A case study with BART [J/OL]. *arXiv preprint* (2022-10-28) [2023-10-09]. <https://arxiv.org/abs/2204.07367v2>
- [25] Carmo D, Piau M, Campiotti I, et al. PTT5: Pretraining and validating the T5 model on Brazilian Portuguese data [J/OL]. *arXiv preprint* (2020-10-08) [2023-10-09]. <http://arxiv.org/abs/2008.09144.pdf>
- [26] Radford A, Wu J, Child R, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 2019, 1(8): 9
- [27] Brown T B, Mann B, Ryder N, et al. Language models are few-shot learners // *Proceedings of the 34th International Conference on Neural Information Processing Systems*. Vancouver, 2020: 1877
- [28] Du N, Huang Y, Dai A M, et al. Glam: Efficient scaling of language models with mixture-of-experts // *International Conference on Machine Learning*. Baltimore, 2022: 5547
- [29] Thoppilan R, De Freitas D, Hall J, et al. LaMDA: Language models for dialog applications [J/OL]. *arXiv preprint* (2022-01-10) [2023-10-09]. <https://arxiv.org/abs/2201.08239>
- [30] Hepp A, Loosen W, Dreyer S, et al. ChatGPT, LaMDA, and the hype around communicative AI: The automation of communication as a field of research in media and communication studies. *Hum Mach Commun*, 2023, 6: 41
- [31] Zeng W, Ren X Z, Su T, et al. PanGu- α : Large-scale autoregressive pretrained Chinese language models with auto-parallel computation [J/OL]. *arXiv preprint* (2022-01-10) [2023-10-09]. <http://arxiv.org/abs/2104.12369>
- [32] Sorscher B, Geirhos R, Shekhar S, et al. Beyond neural scaling laws: Beating power law scaling via data pruning. *Adv Neural Info Proc Syst*, 2022, 35: 19523
- [33] Dettmers T, Lewis M, Belkada Y, et al. LLM.int8(): 8-bit matrix multiplication for transformers at scale [J/OL]. *arXiv preprint* (2022-11-10) [2023-10-09]. <http://arxiv.org/abs/2208.07339.pdf>
- [34] Yang N, Ge T, Wang L, et al. Inference with reference: Lossless acceleration of large language models [J/OL]. *arXiv preprint* (2022-03-10) [2023-10-09]. <https://arxiv.dosf.top/abs/2304.04487>
- [35] Weidinger L, Mellor J, Rauh M, et al. Ethical and social risks of harm from language models [J/OL]. *arXiv preprint* (2022-10-08) [2023-10-09]. <http://arxiv.org/abs/2112.04359.pdf>
- [36] Askell A, Bai Y T, Chen A N, et al. A general language assistant as a laboratory for alignment [J/OL]. *arXiv preprint* (2022-10-09) [2023-10-09]. <http://arxiv.org/abs/2112.00861.pdf>
- [37] Buscemi A. A Comparative Study of Code Generation using ChatGPT 3.5 across 10 programming languages [J/OL]. *arXiv preprint* (2023-08-08) [2023-10-09]. <https://arxiv.dosf.top/abs/2308.04477>
- [38] CARAMANCION K M. News verifiers showdown: A comparative performance evaluation of ChatGPT 3.5, ChatGPT 4.0, bing AI, and bard in news fact-checking [J/OL]. *arXiv preprint* (2023-01-18) [2023-10-09]. <https://arxiv.org/abs/2306.17176>
- [39] Laskar M T R, Bari M S, Rahman M, et al. A systematic study and comprehensive evaluation of ChatGPT on benchmark datasets [J/OL]. *arXiv preprint* (2023-01-05) [2023-10-09]. <https://arxiv.dosf.top/abs/2305.18486>
- [40] Gehman S, Gururangan S, Sap M, et al. RealToxicityPrompts: Evaluating neural toxic degeneration in language models [J/OL]. *arXiv preprint* (2020-09-25) [2023-10-09]. <https://arxiv.dosf.top/abs/2009.11462>
- [41] Mihaylov T, Clark P, Khot T, et al. Can a suit of armor conduct electricity? a new dataset for open book question answering

- [J/OL]. *arXiv preprint* (2018-09-08) [2023-10-09]. <https://arxiv.dosf.top/abs/1809.02789>
- [42] Abbas A, Tirumala K, Simig D, et al. SemDeDup: Data-efficient learning at web-scale through semantic deduplication [J/OL]. *arXiv preprint* (2020-04-22) [2023-10-09]. <https://arxiv.dosf.top/abs/2303.09540>
- [43] Abadji J, Suarez P O, Romary L, et al. Towards a cleaner document-oriented multilingual crawled corpus [J/OL]. *arXiv preprint* (2022-01-17) [2023-10-09]. <http://arxiv.org/abs/2201.06642.pdf>
- [44] Carlini N, Tramer F, Wallace E, et al. Extracting training data from large language models [J/OL]. *arXiv preprint* (2022-03-14) [2023-10-09]. <http://arxiv.org/abs/2012.07805.pdf>
- [45] Black S, Biderman S, Hallahan E, et al. GPT-NeoX-20B: An open-source autoregressive language model [J/OL]. *arXiv preprint* (2022-03-14) [2023-10-09]. <https://arxiv.dosf.top/abs/2204.06745>
- [46] Gao L, Biderman S, Black S, et al. The pile: An 800GB dataset of diverse text for language modeling [J/OL]. *arXiv preprint* (2020-12-31) [2023-10-09]. <https://arxiv.org/abs/2101.00027>
- [47] Liu Y, Ott M, Goyal N, et al. Roberta: A robustly optimized bert pretraining approach [J/OL]. *arXiv preprint* (2019-07-16) [2023-10-09]. <https://arxiv.dosf.top/abs/1907.11692>
- [48] Rae J W, Borgeaud S, Cai T, et al. Scaling language models: Methods, analysis & insights from training gopher [J/OL]. *arXiv preprint* (2022-01-21) [2023-10-09]. <https://arxiv.org/abs/2112.11446>
- [49] Smith S, Patwary M, Norick B, et al. Using DeepSpeed and megatron to train megatron-turing NLG 530B, A large-scale generative language model [J/OL]. *arXiv preprint* (2022-02-04) [2023-10-09]. <https://arxiv.org/abs/2201.11990>
- [50] Sun Y, Wang S H, Feng S K, et al. ERNIE 3.0: Large-scale knowledge enhanced pre-training for language understanding and generation [J/OL]. *arXiv preprint* (2021-07-05) [2023-10-09]. <https://arxiv.org/abs/2107.02137>
- [51] Bai J, Bai S, Yang S, et al. Qwen-vl: A frontier large vision-language model with versatile abilities [J/OL]. *arXiv preprint* (2021-07-05) [2023-10-09]. <https://arxiv.dosf.top/abs/2308.12966>
- [52] Ethayarajh K. How contextual are contextualized word representations? comparing the geometry of BERT, ELMo, and GPT-2 embeddings [J/OL]. *arXiv preprint* (2019-01-09) [2023-10-09]. <https://arxiv.org/abs/1909.00512>
- [53] Liu P F, Yuan W Z, Fu J L, et al. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Comput Surv*, 55(9): 195
- [54] Schick T, Schütze H. Exploiting cloze-questions for few-shot text classification and natural language inference [J/OL]. *arXiv preprint* (2019-06-03) [2023-10-09]. <https://arxiv.dosf.top/abs/2001.07676>
- [55] Lester B, Al-Rfou R, Constant N. The power of scale for parameter-efficient prompt tuning [J/OL]. *arXiv preprint* (2019-05-07) [2023-10-09]. <http://arxiv.org/abs/2104.08691.pdf>
- [56] Wei J, Wang X, Schuurmans D, et al. Chain-of-thought prompting elicits reasoning in large language models. *Adv Neural Info Proc Syst*, 2022, 35: 24824
- [57] Li X L, Liang P. Prefix-tuning: Optimizing continuous prompts for generation [J/OL]. *arXiv preprint* (2021-01-01) [2023-10-09]. <https://arxiv.dosf.top/abs/2101.00190>
- [58] Qiu Z P, Wu X, Gao J Y, et al. U-BERT: Pre-training user representations for improved recommendation. *Proc AAAI Conf Artif Intell*, 2021, 35(5): 4320
- [59] Wu C H, Wu F Z, Qi T, et al. Empowering news recommendation with pre-trained language models // *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*. Virtual Event, 2021: 1652
- [60] Wu C H, Wu F Z, Yu Y, et al. UserBERT: Contrastive user model pre-training [J/OL]. *arXiv preprint* (2021-09-03) [2023-10-09]. <https://arxiv.org/abs/2109.01274>
- [61] Bahdanau D, Cho K, Bengio Y. Neural machine translation by jointly learning to align and translate [J/OL]. *arXiv preprint* (2016-05-19) [2023-10-09]. <http://arxiv.org/abs/1409.0473.pdf>
- [62] Jiao W, Wang W, Huang J, et al. Is ChatGPT a good translator? A preliminary study [J/OL]. *arXiv preprint* (2023-01-20) [2023-10-09]. <https://arxiv.org/pdf/2301.08745v1.pdf>
- [63] Jiang J H, Zhou K, Dong Z C, et al. StructGPT: A general framework for large language model to reason over structured data [J/OL]. *arXiv preprint* (2023-05-16) [2023-10-09]. <https://arxiv.dosf.top/abs/2305.09645>
- [64] Roberts A, Raffel C, Shazeer N M. How much knowledge can you pack into the parameters of a language model [J/OL]. *arXiv preprint* (2020-10-05) [2023-10-09]. <https://arxiv.dosf.top/abs/2002.08910>
- [65] Joshi M, Choi E, Weld D S, et al. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension [J/OL]. *arXiv preprint* (2017-05-13) [2023-10-09]. <https://arxiv.dosf.top/abs/1705.03551>
- [66] Nakano R, Hilton R, Balaji S, et al. WebGPT: Browser-assisted question-answering with human feedback [J/OL]. *arXiv preprint* (2022-01-01) [2023-10-09]. <https://arxiv.dosf.top/abs/2112.09332>
- [67] Jiang Z, Xu F F, Gao L, et al. Active retrieval augmented generation [J/OL]. *arXiv preprint* (2023-05-11) [2023-10-09]. <https://arxiv.dosf.top/abs/2305.06983>
- [68] Peng B L, Galley M, He P C, et al. Check your facts and try again: Improving large language models with external knowledge and automated feedback [J/OL]. *arXiv preprint* (2023-04-08) [2023-10-09]. <https://arxiv.dosf.top/abs/2302.12813>
- [69] Geva M, Khashabi D, Segal E, et al. Did aristotle use a laptop? A question answering benchmark with implicit reasoning strategies. *Trans Assoc Comput Linguist*, 2021, 9: 346
- [70] Marcus M P, Santorini B, Marcinkiewicz M A. Building a large

- annotated corpus of English: The Penn treebank. *Comput Linguist*, 1993, 19(2): 313
- [71] Merity S, Xiong C M, Bradbury J, et al. Pointer sentinel mixture models [J/OL]. *arXiv preprint* (2016-09-16) [2023-10-09]. <https://arxiv.org/abs/1609.07843>
- [72] Paperno D, Kruszewski G, Lazaridou A, et al. The LAMBADA dataset: Word prediction requiring a broad discourse context [J/OL]. *arXiv preprint* (2016-01-20) [2023-10-09]. <https://arxiv.dosf.top/abs/1606.06031>
- [73] Farhad A, Arkady A, Magdalena B, et al. Findings of the 2021 conference on machine translation (WMT21) // *Proceedings of the Sixth Conference on Machine Translation*. Online, 2021: 1
- [74] Kocmi T, Bawden R, Bojar O, et al. Findings of the 2022 conference on machine translation (WMT22) // *Proceedings of the Seventh Conference on Machine Translation (WMT)*. Online, 2022: 1
- [75] Nallapati R, Zhou B W, dos Santos C, et al. Abstractive text summarization using sequence-to-sequence RNNs and beyond [J/OL]. *arXiv preprint* (2016-08-26) [2023-10-09]. <https://arxiv.dosf.top/abs/1602.06023>
- [76] Clark P, Cowhey I, Etzioni O, et al. Think you have solved question answering? try ARC, the AI2 reasoning challenge [J/OL]. *arXiv preprint* (2018-04-14) [2023-10-09]. <https://arxiv.dosf.top/abs/1803.05457>
- [77] Longpre S, Lu Y, Daiber J. MKQA: A linguistically diverse benchmark for multilingual open domain question answering. *Trans Assoc Comput Linguist*, 2021, 9: 1389
- [78] Berant J, Chou A, Frostig R, et al. Semantic parsing on freebase from question-answer pairs // *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*. Washington, 2013: 1533
- [79] Rajpurkar P, Zhang J, Lopyrev K, et al. SQuAD: 100, 000+ questions for machine comprehension of text [J/OL]. *arXiv preprint* (2016-10-11) [2023-10-09]. <https://arxiv.dosf.top/abs/1606.05250>
- [80] Bollacker K, Evans C, Paritosh P, et al. Freebase: A collaboratively created graph database for structuring human knowledge // *Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data*. Vancouver, 2008: 1247
- [81] Mahdisoltani F, Biega J, Suchanek F M. YAGO3: A knowledge base from multilingual wikipedias // *CIDR*. 2013
- [82] Miller G A. WordNet: A lexical database for English. *Commun ACM*, 1995, 38(11): 39
- [83] Sap M, Rashkin H, Chen D, et al. SocialIQA: Commonsense reasoning about social interactions [J/OL]. *arXiv preprint* (2019-09-09) [2023-10-09]. <https://arxiv.org/abs/1904.09728>
- [84] Clark C, Lee K, Chang M W, et al. BoolQ: Exploring the surprising difficulty of natural yes/No questions [J/OL]. *arXiv preprint* (2019-05-24) [2023-10-09]. <https://arxiv.dosf.top/abs/1905.10044>
- [85] Hendrycks D, Burns C, Basart S, et al. Measuring massive multitask language understanding [J/OL]. *arXiv preprint* (2021-01-12) [2023-10-09]. <https://arxiv.org/abs/2009.03300>
- [86] Cobbe K, Kosaraju V, Bavarian M, et al. Training verifiers to solve math word problems [J/OL]. *arXiv preprint* (2021-11-18) [2023-10-09]. <https://arxiv.org/abs/2110.14168>
- [87] Jiang A Q, Li W, Han J M, et al. LISA: Language models of ISAbelle proofs // *Proceedings of the 6th Conference on Artificial Intelligence and Theorem Proving*. Xiamen, 2021: 378
- [88] Amini A, Gabriel S, Lin S C, et al. MathQA: Towards interpretable math word problem solving with operation-based formalisms [J/OL]. *arXiv preprint* (2019-05-30) [2023-10-09]. <https://arxiv.dosf.top/abs/1905.13319>
- [89] Dan Y H, Lei Z K, Gu Y Y, et al. EduChat: A large-scale language model-based chatbot system for intelligent education [J/OL]. *arXiv preprint* (2023-08-05) [2023-10-09]. <https://arxiv.org/abs/2308.02773>
- [90] Jeon J, Lee S. Large language models in education: A focus on the complementary relationship between human teachers and ChatGPT. *Educ Inf Technol*, 2023, 28(12): 15873
- [91] Bonner E, Lege R, Frazier E. Large language model-based artificial intelligence in the language classroom: Practical ideas for teaching. *TEwT*, 2023, 2023(1)
- [92] Dijkstra R, Genç Z, Kayal S, et al. Reading comprehension quiz generation using generative pre-trained transformers [J/OL]. *CEUR-WS* (2022) [2023-10-09]. https://ceur-ws.org/Vol-3192/itb22_p1_full5439.pdf
- [93] Gunawan J. Exploring the future of nursing: Insights from the ChatGPT model. *Belitung Nurs J*, 2023, 9(1): 1
- [94] Liu Z L, Zhong A X, Li Y W, et al. Radiology-GPT: A large language model for radiology [J/OL]. *arXiv preprint* (2023-06-14) [2023-10-09]. <https://arxiv.org/abs/2306.08666>
- [95] Loconte R, Orrú G, Tribastone M, et al. Challenging ChatGPT 'Intelligence' with human tools: A neuropsychological investigation on prefrontal functioning of a large language model. *Intelligence*, 2023
- [96] Kung T H, Cheatham M, Medenilla A, et al. Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models. *PLoS Digit Health*, 2023, 2(2): e0000198
- [97] Guler N, Kirshner S, Vidgen R. Artificial intelligence research in business and management: A literature review leveraging machine learning and large language models. *SSRN Journal*, 2023
- [98] Chen L J, Zaharia M, Zou J Y. FrugalGPT: How to use large language models while reducing cost and improving performance [J/OL]. *arXiv preprint* (2023-05-09) [2023-10-09]. <https://arxiv.dosf.top/abs/2305.05176>
- [99] Lopez-Lira A, Tang Y H. Can ChatGPT forecast stock price movements? return predictability and large language models [J/OL]. *arXiv preprint* (2023-09-09) [2023-10-09]. <https://arxiv.org/abs/2309.05176>

- dosf.top/abs/2304.07619
- [100] Zhang B Y, Yang H Y, Liu X Y. Instruct-FinGPT: Financial sentiment analysis by instruction tuning of general-purpose large language models [J/OL]. *arXiv preprint* (2023-09-09) [2023-10-09]. <https://arxiv.dosf.top/abs/2306.12659>
- [101] Yazdinejad A, Rabieinejad E, Hasani T, et al. A BERT-based recommender system for secure blockchain-based cyber physical drug supply chain management. *Clust Comput*, 2023, 26(6): 3389
- [102] Xu Y, Sun Y M, Wan J F, et al. Industrial big data for fault diagnosis: Taxonomy, review, and applications. *IEEE Access*, 1945, 5: 17368
- [103] Tao F, Cheng J F, Qi Q L, et al. Digital twin-driven product design, manufacturing and service with big data. *Int J Adv Manuf Technol*, 2018, 94(9): 3563
- [104] Nasr M, Carlini N, Hayase J, et al. Scalable extraction of training data from (production) language models [J/OL]. *arXiv preprint* (2023-11-28) [2023-10-09]. <https://arxiv.dosf.top/abs/2311.17035>
- [105] Khalil M, Er E K. Will ChatGPT get you caught? Rethinking of plagiarism detection [J/OL]. *arXiv preprint* (2023-02-08) [2023-10-09]. <https://arxiv.dosf.top/abs/2302.04335>
- [106] Wang J D, Lan C L, Liu C, et al. Generalizing to unseen domains: A survey on domain generalization. *IEEE Trans Knowl Data Eng*, 2023, 35(8): 8052
- [107] Liu Y, Yao Y, Ton J F, et al. Trustworthy LLMs: A survey and guideline for evaluating large language models' alignment [J/OL]. *arXiv preprint* (2023-08-10) [2023-10-09]. <https://arxiv.dosf.top/abs/2308.05374>
- [108] Liu Y H, Han T L, Ma S Y, et al. Summary of ChatGPT-Related research and perspective towards the future of large language models. *Meta-Radiology*, 2023, 1(2): 100017
- [109] Sahu P, Cogswell M, Gong Y Y, et al. Unpacking large language models with conceptual consistency [J/OL]. *arXiv preprint* (2022-09-29) [2023-10-09]. <https://arxiv.dosf.top/abs/2209.15093>
- [110] Kaddour J, Harris J, Mozes M, et al. Challenges and applications of large language models [J/OL]. *arXiv preprint* (2023-06-19) [2023-10-09]. <https://arxiv.dosf.top/abs/2307.10169>