大语言模型（Large Language Models, LLM）技术研究综述

作者：丁涛
单位：武汉理工大学 计算机与人工智能学院
课程名称：高级人工智能原理与技术

摘要

近年来，大语言模型（Large Language Models, LLM）作为人工智能领域的重要突破，在自然语言处理、知识推理、多模态理解与生成等方面展现出卓越能力。以 Transformer 架构为核心，LLM 通过大规模无监督或弱监督预训练，并结合指令微调、对齐学习等技术，实现了从“语言建模”到“通用智能接口”的转变。本文系统综述了大语言模型的技术原理、研究现状与最新进展，重点分析了主流模型架构、训练范式、推理机制以及效率优化方法，并总结了其在教育、医疗、金融、软件工程等领域的典型应用场景。在此基础上，本文进一步讨论了当前 LLM 面临的关键挑战与未来发展趋势，以期为相关研究与工程实践提供参考。

关键词：大语言模型；Transformer；预训练模型；人工智能；自然语言处理

1. 引言

人工智能技术的发展经历了多次重要范式转变，其中自然语言处理（Natural Language Processing, NLP）始终是衡量人工智能水平的重要研究方向之一。早期 NLP 系统主要依赖人工规则和特征工程，随后统计学习方法逐渐占据主导地位，但其在复杂语义理解和跨任务迁移方面仍存在明显局限。近年来，随着深度学习方法的成熟以及大规模计算资源的普及，基于神经网络的语言模型逐步成为研究主流。

特别是 Transformer 架构提出后，自注意力机制为建模长距离依赖提供了高效解决方案，使得大规模语言模型的训练成为可能[4]。在此基础上，研究者提出了“预训练—微调”的通用范式，通过在海量语料上进行自监督学习，再针对具体任务进行适配，从而显著提升了模型的泛化能力与迁移能力。以 GPT、BERT 等模型为代表的预训练语言模型在多项自然语言处理任务中取得了突破性进展。

随着模型参数规模和训练数据规模的持续扩大，大语言模型（Large Language Models, LLM）逐渐展现出超越单一任务的通用能力，包括少样本学习、上下文推理以及跨领域知识迁移等[1]–[3]。这些能力的出现不仅推动了自然语言处理技术的快速发展，也引发了学术界和产业界对通用人工智能发展路径的广泛讨论。

在应用层面，大语言模型已被广泛集成至智能搜索、对话系统、代码生成和知识管理等实际场景中，显著改变了人机交互方式和信息获取模式。然而，与其快速发展的应用前景相伴随的，是模型在可靠性、安全性、资源消耗以及伦理治理等方

面所面临的挑战。因此，有必要对大语言模型的技术原理、研究现状与应用进展进行系统梳理与综合分析。

基于上述背景，本文围绕大语言模型展开综述研究，重点介绍其核心技术原理、主流研究方向、最新进展以及典型应用场景，并对当前存在的问题与未来发展趋势进行讨论。本文的研究旨在为相关课程学习和后续研究提供系统参考。

2. 大语言模型的技术原理

本节从模型架构、核心机制与训练范式三个层面，对大语言模型的基础技术原理进行系统阐述，为后续研究现状与应用分析奠定理论基础。

2.1 Transformer 架构

Transformer 架构由 Vaswani 等人提出，是当前几乎所有主流大语言模型的基础结构[4]。其核心创新在于完全摒弃循环神经网络（RNN）和卷积神经网络（CNN），仅依赖自注意力机制来建模序列内部的依赖关系，从而显著提升了并行计算效率与长距离依赖建模能力。

Transformer 通常由若干层编码器（Encoder）和/或解码器（Decoder）堆叠而成。每一层主要包含以下模块：多头自注意力（Multi-Head Self-Attention）、前馈全连接网络（Feed-Forward Network, FFN）、残差连接（Residual Connection）以及层归一化（Layer Normalization）。在大语言模型中，GPT 系列主要采用解码器结构，而 BERT 类模型则基于编码器结构。

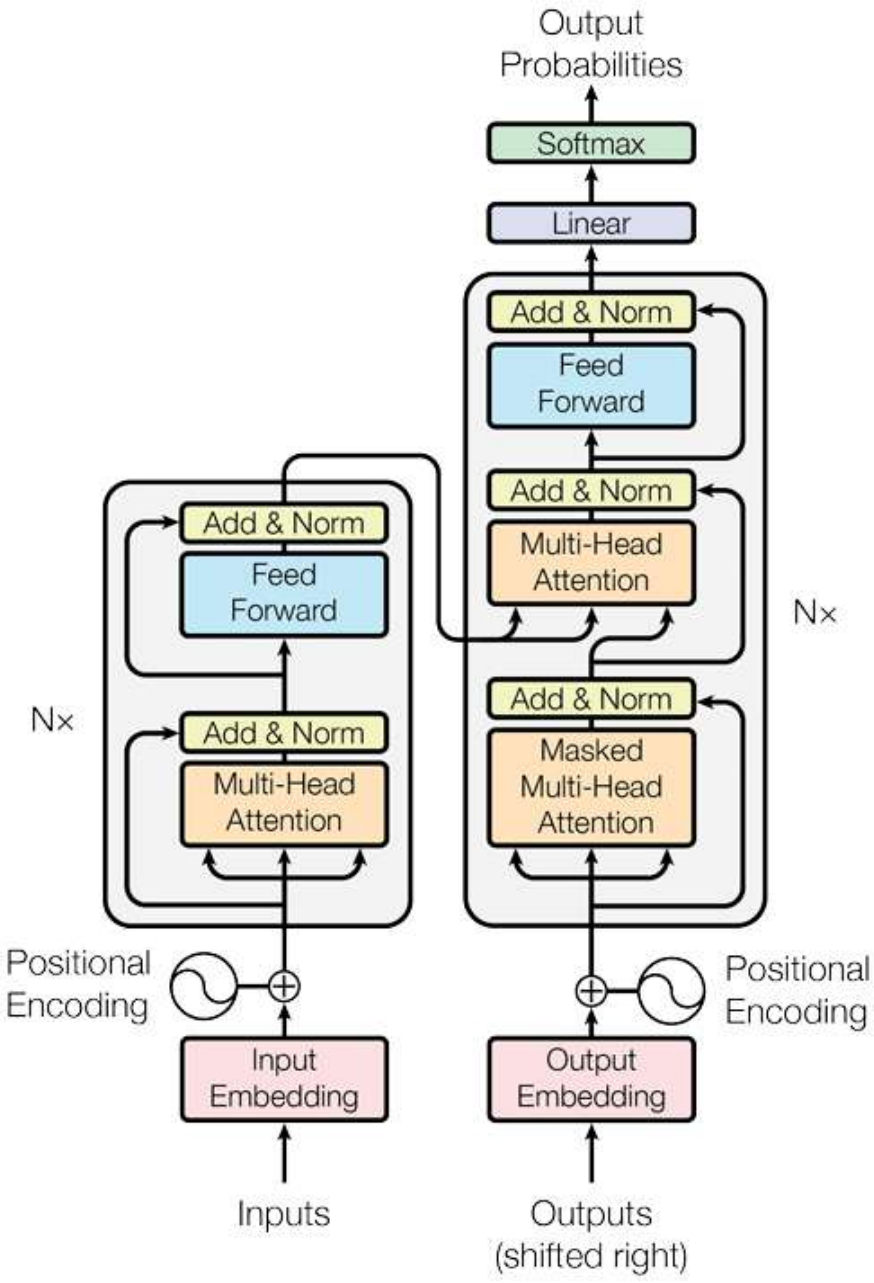

图 1 Transformer 基本结构示意图

Transformer 架构由 Vaswani 等人提出，其核心思想是利用自注意力（Self-Attention）机制建模序列中任意位置之间的依赖关系[4]。与传统 RNN 和 CNN 不同，Transformer 能够实现高度并行计算，在大规模训练中表现出更优的效率。

Transformer 通常由多层编码器（Encoder）或解码器（Decoder）堆叠而成，其基本组件包括多头注意力机制、前馈神经网络、残差连接和层归一化。

2.2 自注意力机制

自注意力机制（Self-Attention）是 Transformer 的核心，其目标是刻画序列中任意两个位置之间的相关性。通过为每个输入 token 构造查询（Query）、键（Key）和值（Value）向量，自注意力能够根据上下文动态分配权重，从而实现对语义信息的全局建模。

其标准计算形式如下：

$$\text{Attention}(Q,\ K,\ V) = \text{softmax}(QK^T / \sqrt{d_k})V$$

其中 d_k 为 Key 向量维度，用于缩放内积结果以稳定梯度。多头注意力机制（Multi-Head Attention）通过在多个子空间中并行执行注意力计算，使模型能够同时关注不同语义层面的信息，这对复杂语言现象（如指代消解、逻辑关系建模）尤为重要。

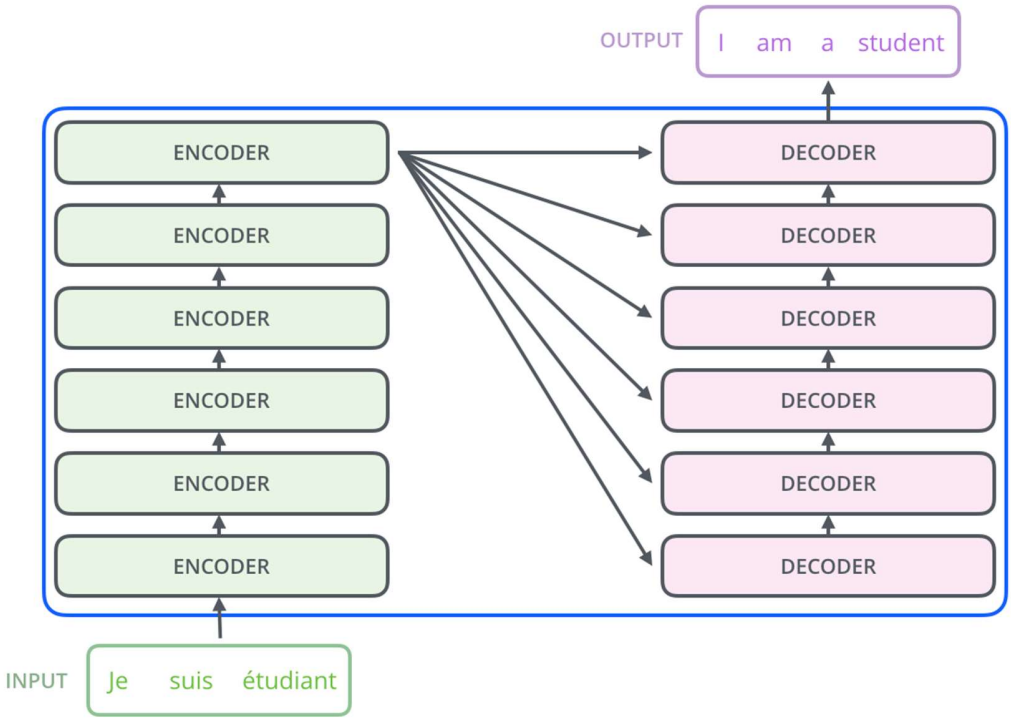

图 2 多头自注意力计算流程图

自注意力机制通过计算查询（Query）、键（Key）和值（Value）之间的相关性，实现对上下文信息的加权聚合。其计算形式如下：

$$\text{Attention}(Q,K,V)=\text{softmax}(QK^T/\sqrt{d_k})V$$

多头注意力机制通过在不同子空间中并行执行注意力计算，提高了模型对复杂语义关系的建模能力。

2.3 预训练与微调范式

LLM 通常采用“预训练 + 微调”的两阶段训练范式。预训练阶段使用海量文本数据进行无监督或自监督学习，目标是最小化语言建模损失；微调阶段则针对特定任务或指令进行有监督学习[5]。

近年来，指令微调（Instruction Tuning）和人类反馈强化学习（RLHF）被广泛用于提升模型的可控性和对齐性[6]。

2.4 推理与上下文建模

LLM 的推理能力主要来源于其对长上下文的建模能力以及参数规模带来的表达能力。为支持更长上下文，研究者提出了稀疏注意力、线性注意力和位置编码改进等方法[7]。

3. 研究现状与主流模型

在 Transformer 架构和大规模预训练范式逐渐成熟的背景下，大语言模型的研究进入快速发展阶段。当前研究现状主要体现在模型规模持续扩大、开源生态不断完善以及评测体系日益系统化等方面。本节将从发展阶段、模型类型与评测方法三个角度，对主流大语言模型的研究现状进行较为全面的分析。

3.1 大语言模型的发展阶段

从整体发展脉络来看，大语言模型的演进大致可分为三个阶段。第一阶段以 BERT、GPT-2 为代表，主要验证了大规模预训练在语言理解与生成任务中的有效性；第二阶段以 GPT-3、PaLM 等模型为代表，研究重点转向参数规模扩展与少样本学习能力的提升[1][2]；第三阶段则以 GPT-4、LLaMA 2 及其衍生模型为代表，更加强调模型的对齐性、安全性以及实际应用能力。

这一演进过程表明，单纯依赖参数规模提升已难以持续带来线性性能增长，模型训练策略、数据质量和对齐方法正逐渐成为影响性能的关键因素。

3.2 闭源大语言模型

目前，性能最为领先的大语言模型多由大型科技公司研发并以闭源形式发布。OpenAI 推出的 GPT-4 在复杂推理、多模态理解和跨任务迁移方面表现突出，被广泛应用于对话系统、代码生成和内容创作等场景[8]。Google 的 PaLM 2 与 Gemini 系列模型在多语言处理和推理能力方面具有优势，并已集成至多项商业产品中。

闭源模型通常依托于海量高质量数据和超大规模算力资源，在性能上具有明显优势，但其模型结构、训练细节和数据来源缺乏透明性，这在一定程度上限制了其在学术研究中的可复现性。

3.3 开源大语言模型

与闭源模型相对应，开源大语言模型在近年来取得了快速发展。Meta 发布的 LLaMA 与 LLaMA 2 系列模型在相对较小的参数规模下实现了接近闭源模型的性能，成为学术界和工业界广泛使用的基础模型[9]。Mistral 系列模型通过优化注意力机制和推理流程，在效率方面表现突出[10]。

此外，国内外多家研究机构相继推出了面向特定语言或领域的开源模型，如 Qwen、Baichuan 等。这些模型在中文理解、多任务适应性和可定制性方面具有明显优势，为本土化应用提供了重要支撑。

3.4 模型规模与能力关系

模型规模通常被认为是影响 LLM 性能的重要因素之一。大量研究表明，在一定范围内，参数规模、训练数据量与模型性能之间呈现近似幂律关系。然而，随着模型规模不断扩大，其边际收益逐渐递减，同时训练成本和能耗显著上升。

因此，近年来研究逐渐从“更大模型”转向“更优模型”，重点关注数据质量、训练策略以及模型结构优化等因素。

3.5 评测基准与能力分析

为了全面评估大语言模型的综合能力，研究者提出了多种评测基准体系，如 BIG-bench、MMLU 和 HELM[12]。这些基准从语言理解、知识推理、安全性和公平性等多个维度对模型进行系统评估。

值得注意的是，单一评测指标难以全面反映模型在真实应用中的表现，因此当前研究趋势逐渐强调多维度、场景化的综合评测方法。

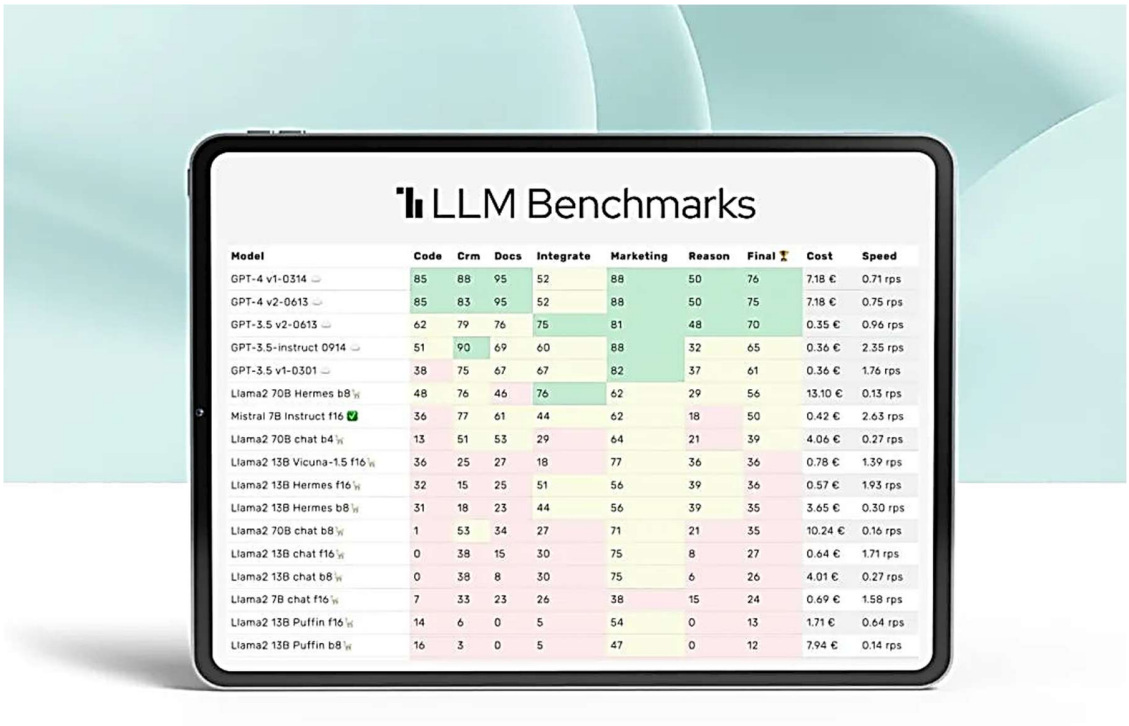

图 3 大语言模型评测

4. 最新研究进展（2023–2026）

近三年（2023 - 2026）是大语言模型快速演进和技术突破的重要阶段。研究者不仅在模型规模和能力上持续推进，更在推理能力、多模态集成、资源效能、训练策略与专业化领域展开了多维探索，形成了以下显著进展方向。

4.1 参数高效训练与微调方法发展

传统大语言模型的训练成本极高，这促使研究者提出了一系列**参数高效的训练与微调方法**。如 LoRA 和 QLoRA 等技术，通过引入低秩参数调整、量化等策略，在保持性能的同时显著降低训练成本和资源消耗[13]。这些技术被视为实现高效模型迭代的关键路径。

除此之外，有研究从资源消耗角度对 LLM 展开系统梳理，提出了跨计算、内存、能耗和财务成本的优化方案，为未来可持续 LLM 发展奠定基础。

4.2 多模态与跨领域模型的快速发展

随着 GPT-4V、Gemini、Llama 4 等模型的发布，多模态学习成为 LLM 研究的重要方向。多模态大语言模型（MLLM）可以同时处理文本、图像、视频等多种信息

模式，并在多模态推理、生成等任务中获得了显著能力提升。相关综述指出这一方向不仅在视觉问答和跨媒体理解任务中表现突出，还将成为推进通用人工智能能力的重要路径。

例如 GPT-4V 类模型展示了在基于图像生成故事、图像理解等任务上的新兴能力，推动对多模态链式思维（Chain-of-Thought）的研究。尽管在多模态幻觉和连贯性方面仍存在挑战，但研究的速度和深度持续增长。

4.3 意图对齐与强化学习增强的模型优化

近年来，将强化学习（Reinforcement Learning, RL）与大语言模型融合的研究不断推向深入。RL 技术，尤其是通过可验证奖励（RLVR）和人类反馈的强化学习，可显著提升模型对人类指令的理解与对齐能力，这已成为增强 LLM 推理和推断能力的重要方向之一。

该类研究不仅涵盖 RL 在预训练、对齐微调和推理强化中的实践，还系统整理了用于 RL 训练的评估数据集和训练框架，为后续研究提供了理论与工具基础。

4.4 专业与科学领域 LLM 的快速涌现

除了通用大模型外，不少研究开始关注**面向专业领域的语言模型**，如生物化学、化学结构、蛋白质语言模型等。在科学语言大模型（Sci-LLM）领域，已有综述系统梳理针对文本科学知识、小分子化合物、大分子蛋白质等多个细分方向的发展状态，表明领域特定 LLM 正成为跨学科智能工具的重要趋势。

此外，在生物医学和化学领域，专门设计的 LLM 在处理复杂结构化科学数据时展示了潜力，这对于药物发现、科学文献生成等任务具有重要意义。

4.5 专业评估体系与新基准测试

针对 LLM 的综合能力评估也是近年来的研究热点。传统的评测如 MMLU、BIG-Bench、HELM 等依然是考核模型多维能力的重要手段，但研究者逐渐提出更细粒度、多场景的评测指标与框架，以解决单一指标难以全面反映模型实际能力的问题[12]。

此外，新型评估方法正在引入跨模型竞争、战略推理测试等内容，以便更准确地捕捉模型的推理和策略性表现，这对行业应用指导与能力比较有重要意义。

4.6 开源模型生态与竞争格局演化

2023 - 2026 年间，开源大语言模型生态显著丰富。开源模型如 LLaMA 系列继续扩展参数规模和训练策略，而新兴开源模型如 DeepSeek-R1 提供了与闭源大模

型同级别性能的开源替代方案，展示了更低门槛的行业研究基础。这些模型在数学能力、编码任务等领域表现卓越，并推动了开源社区围绕语言模型的实验与实践加速发展。

与此同时，多模态、混合推理架构（如 MoE 和混合思维模式）也逐渐成为模型创新的重要元素，为更高效的模型设计提供了新的方向。

5. 实际应用场景

随着模型能力和工程化水平的不断提升，大语言模型已从实验性技术逐步发展为可落地的通用智能工具，在多个行业场景中展现出显著价值。本节围绕教育、医疗、金融、法律、软件工程及科研辅助等典型领域，对 LLM 的实际应用进行系统分析。

5.1 教育领域

在教育领域，大语言模型被广泛应用于智能辅导、作业辅助、学习资源生成与教学管理等方面。基于对自然语言的深度理解能力，LLM 能够根据学生的学习水平与知识掌握情况，提供个性化的讲解与反馈，从而实现“因材施教”的教学目标[17]。

具体而言，LLM 可作为智能助教，为学生解答课程相关问题、生成例题与解析，辅助学生进行课后复习；在教师侧，模型可用于教学大纲设计、试题生成与作业自动批改，大幅降低重复性工作负担。此外，部分研究表明，合理使用 LLM 能够提升学生的学习动机与自主学习能力，但同时也需警惕其对学术诚信带来的潜在影响。

5.2 医疗与生命科学领域

在医疗与生命科学领域，大语言模型主要应用于医学知识问答、临床文本处理和科研辅助等场景。通过对医学文献和临床记录的建模，LLM 能够辅助医生进行病例总结、病历生成以及医学信息检索[18]。

例如，在临床场景中，模型可对非结构化的电子病历进行自动摘要，帮助医生快速获取关键信息；在科研层面，LLM 可用于文献综述初稿生成、研究假设整理等任务。然而，由于医疗领域对准确性与安全性要求极高，当前研究普遍强调“人机协同”模式，即将 LLM 作为辅助工具而非决策主体。

5.3 金融领域

金融行业具有文本数据密集、规则复杂和风险敏感等特点，为大语言模型的应用提供了广阔空间。LLM 在金融领域的典型应用包括智能投顾、舆情分析、风险评估与合规审查等[19]。

在实际应用中，模型可对财经新闻、公司公告和研究报告进行语义分析，辅助投资决策；在风控场景中，LLM 可用于识别潜在风险事件和异常行为。需要注意的是，金融领域对模型的可解释性和稳定性要求较高，这对 LLM 的部署与监管提出了更高挑战。

5.4 法律领域

法律文本具有专业性强、逻辑严密和语言规范的特点，大语言模型在法律检索、合同分析和法律文书生成等方面表现出较高应用潜力。通过对法规条文和判例数据的学习，LLM 能够辅助完成法律咨询初筛、案例相似度分析以及合同风险提示。

部分研究表明，在限定场景下，LLM 生成的法律文本在结构完整性和语言规范性方面已接近人工水平。然而，由于法律结论的严肃性和责任归属问题，当前应用多集中于辅助分析阶段，而非直接给出最终法律意见。

5.5 软件工程与代码智能

软件工程是大语言模型最早实现规模化落地的领域之一。以 Codex、Code LLaMA 等模型为代表的代码大语言模型在代码生成、补全、调试和文档生成等方面显著提升了开发效率[20]。

在实际开发流程中，LLM 可作为编程助手，根据自然语言描述生成代码片段，辅助新手学习编程；在工程实践中，模型可用于自动生成测试用例、发现潜在缺陷以及解释复杂代码逻辑。相关研究表明，合理使用 LLM 能够缩短开发周期，但仍需人工审查以确保代码质量与安全性。

5.6 科研与知识工作辅助

在科研与知识密集型工作中，大语言模型被广泛用于文献检索、综述写作辅助和知识整合。通过对大量学术文本的学习，LLM 能够快速总结研究脉络、提炼关键观点，并为研究人员提供灵感支持。

尽管如此，当前模型在事实准确性和引用规范性方面仍存在不足，因此在科研场景中，LLM 更适合作为“效率工具”，而非替代研究者的独立思考。

6. 挑战与未来发展趋势

尽管大语言模型在多个领域取得了显著进展，但其在理论、工程和社会层面仍面临诸多挑战。本节从模型能力、工程实现以及伦理与治理等方面，对当前存在的问题进行分析，并对未来发展趋势进行展望。

6.1 模型能力与可靠性挑战

首先，大语言模型在事实准确性方面仍存在不足，尤其是在开放域问答和专业领域任务中，模型可能产生“幻觉”（Hallucination）现象，即生成看似合理但

实际错误的信息。这一问题在医疗、法律等高风险场景中尤为突出，严重制约了 LLM 的直接应用。

其次，当前 LLM 的推理能力在很大程度上依赖于统计相关性而非显式逻辑推理，其在复杂多步推理、因果推断和数学证明等任务中仍存在明显局限。这表明，如何将符号推理与神经网络方法有效结合，仍是未来研究的重要方向。

6.2 工程与资源消耗问题

大语言模型通常具有数十亿甚至上千亿参数，对算力、存储和能源消耗提出了极高要求。这不仅增加了模型训练与部署的成本，也带来了显著的环境影响。为此，学术界和工业界正在积极探索模型压缩、参数高效训练和推理加速等技术路径。

此外，模型的持续更新与维护也是一项长期挑战。由于预训练模型依赖静态数据，其内置知识可能随时间迅速过时，如何实现高效、安全的模型更新机制，仍有待深入研究。

6.3 安全、伦理与治理问题

随着 LLM 应用范围的不断扩大，其潜在的社会影响日益凸显。模型偏见、隐私泄露、版权风险以及滥用问题，已成为各国监管机构和研究者高度关注的议题。如何在保障创新活力的同时，建立合理的技术治理与监管框架，是大语言模型长期发展必须面对的问题。

6.4 未来发展趋势

展望未来，大语言模型的发展趋势可能体现在以下几个方面：一是模型架构将更加高效与模块化，在保证性能的同时显著降低资源消耗；二是多模态与智能体化方向将进一步深化，使模型能够更好地感知环境并执行复杂任务；三是对齐、安全与可控性研究将成为核心议题，推动 LLM 从“能力提升”向“可靠应用”转变。

7. 结论

本文围绕大语言模型这一人工智能领域的研究热点，对其技术原理、研究现状、最新进展以及实际应用场景进行了系统综述。通过分析可以看出，基于 Transformer 架构的大语言模型已在语言理解与生成任务中展现出卓越性能，并逐步发展为通用智能系统的重要基础。

总体而言，LLM 的研究正从单纯追求规模扩展，转向效率优化、安全对齐与应用落地并重的新阶段。尽管当前仍面临可靠性、资源消耗和伦理治理等多方面挑战，但随着算法创新、工程技术进步以及制度建设的不断完善，大语言模型有望在更多实际场景中发挥积极作用，并持续推动人工智能技术的发展。

8. 课程学习体会

通过本课程的系统学习以及本综述报告的撰写过程，笔者对大语言模型的发展背景、核心技术与应用前景有了更加全面和深入的认识。在查阅和整理大量中英文文献的过程中，不仅加深了对 Transformer 架构和预训练范式的理解，也体会到人工智能研究中理论分析与工程实践相结合的重要性。

此外，本次学习使笔者认识到，大语言模型虽然展现出强大的能力，但其应用仍需建立在理性认知和审慎态度之上。如何在充分发挥技术优势的同时，正视其局限性与潜在风险，是未来从事相关研究与应用时必须重点考虑的问题。通过课程学习，笔者的文献阅读能力、学术写作能力以及对前沿技术的批判性思考能力均得到了显著提升，这为后续深入学习人工智能相关方向奠定了良好基础。

参考文献

[1] T. Brown et al., “Language Models are Few-Shot Learners,” *NeurIPS*, 2023.

[2] J. Chowdhery et al., “PaLM: Scaling Language Modeling with Pathways,” arXiv, 2023.

[3] H. Touvron et al., “LLaMA 2: Open Foundation and Fine-Tuned Chat Models,” arXiv, 2023.

[4] A. Vaswani et al., “Attention Is All You Need,” NeurIPS, 2017.

[5] J. Wei et al., “Finetuned Language Models Are Zero-Shot Learners,” ICLR, 2023.

[6] P. Christiano et al., “Deep Reinforcement Learning from Human Preferences,” NeurIPS, 2023.

[7] S. Dao et al., “FlashAttention,” ICML, 2023.

[8] OpenAI, “GPT-4 Technical Report,” 2023.

[9] Meta AI, “LLaMA: Open and Efficient Foundation Language Models,” 2023.

[10] A. Mistral et al., “Mistral 7B,” arXiv, 2023.

[11] Alibaba DAMO Academy, “Qwen Technical Report,” 2024.

[12] P. Liang et al., “HELM: Holistic Evaluation of Language Models,” TACL, 2023.

[13] E. Hu et al., “LoRA: Low-Rank Adaptation,” ICLR, 2023.

[14] J. Alayrac et al., “Flamingo,” NeurIPS, 2023.

[15] S. Yao et al., “ReAct: Synergizing Reasoning and Acting,” ICLR, 2024.

[16] R. Bommasani et al., “On the Opportunities and Risks of Foundation Models,” Stanford CRFM, 2023.

[17] Z. Kasneci et al., “ChatGPT for Education,” Learning and Instruction, 2024.

[18] E. Singhal et al., “Large Language Models in Medicine,” Nature Medicine, 2023.

[19] J. Huang et al., “Large Language Models for Finance,” arXiv, 2024.

[20] M. Rozière et al., “Code LLaMA,” arXiv, 2023.