Contents lists available at ScienceDirect

Engineering

journal homepage: www.elsevier.com/locate/eng

Research
Next Ten Years: Create a Better Future—Review

知识引导下生成式模型协同合作与联合进化——后大模型时代的发展蓝图

吴飞[a,*], 沈弢[a], Thomas Bäck[g], 陈静远[a], 黄刚[h], 金耀初[c], 况琨[a], 李孟择[f], 卢策吾[b], 苗嘉旭[e], 王永威[a], 魏颖[a], 吴帆[b], 严骏驰[b], 杨红霞[d], 杨易[a], 张圣宇[a], 赵洲[a], 庄越挺[a], 潘云鹤[a,*]

[a] College of Computer Science and Technology, Zhejiang University, Hangzhou 310027, China
[b] Department of Computer Science and Engineering, Shanghai Jiao Tong University, Shanghai 200240, China
[c] School of Engineering, Westlake University, Hangzhou 310024, China
[d] Department of Computing, The Hong Kong Polytechnic University, Hong Kong 999077, China
[e] School of Cyber Science and Technology, Sun Yat-Sen University, Shenzhen 518107, China
[f] Department of Computer Science and Engineering, Hong Kong University of Science and Technology, Hong Kong 999077, China
[g] Leiden Institute of Advanced Computer Science, Leiden University, Leiden 2333 CC, Netherlands
[h] College of Electrical Engineering, Zhejiang University, Hangzhou 310027, China

ARTICLE INFO

Article history:
Received 6 October 2024
Revised 9 November 2024
Accepted 8 December 2024
Available online 19 December 2024

关键词
人工智能
大语言模型
知识引导
模型协同
模型联合进化

摘要

近年来大语言模型在多模态场景下的语义理解、内容生成与问题推理等任务上表现出较大潜力，为人工智能发展提供了动力。然而，当前大语言模型在信息时效性、内容幻觉、推理效率、可解释性以及垂直领域准确性等方面仍面临诸多挑战。为此，本文重点围绕后大模型时代的三个关键发展方向展开讨论，即知识引导、模型协同与联合进化。首先，分析了如何将外部知识引入大语言模型，如在训练目标中引入知识、指令微调、基于检索的推理增强以及知识提示等方法，以提升事实准确性、推理能力与可解释性。其次，阐释了模型协同策略，如通过模型合并、功能型模型协同以及知识注入等技术，充分利用大语言模型与小模型之间的互补优势，进一步提高推理效率和提升特定领域应用性能。最后，介绍了模型联合进化这一前沿方向，即多个模型通过知识共享、参数融合以及联合学习等机制在动态环境与多样化任务中共同进化，从而获得更强适应性与持续学习能力。本文对知识引导、模型协同与联合进化在科学、工程与社会领域的应用前景进行了讨论，特别关注这些方法在假设生成、问题表征、问题求解以及跨领域解释性等方面的潜在价值，并对后大模型时代的持续创新与未来研究方向进行了展望。

© 2024 THE AUTHORS. Published by Elsevier LTD on behalf of Chinese Academy of Engineering and Higher Education Press Limited Company. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

1. 引言

在人工智能（AI）[1–4]领域，以GPT-o1为代表的大型语言模型（large language models, LLMs）涌现了类人化的理解、生成和推理能力，广泛应用于从对话系统到复杂问题求解等。然而，LLMs在垂直领域领域应用时仍然存在诸多局限，影响了其适用性。

• **LLMs在知识表达方面存在固有的不足，亟需通过显式知识的引入来提升其表现。**目前，LLMs主要基于大规模无监督文本数据进行训练，其知识通常是隐式存储在庞大的模型参数中，这会导致以下问题：信息可能陈旧或过时，出现“幻觉”现象或信息不准确，难以有效推理结

* Corresponding authors.
E-mail addresses: wufei@zju.edu.cn (F. Wu), wufei@zju.edu.cn (F. Wu), panyh@zju.edu.cn (Y. Pan).

2095-8099/© 2024 THE AUTHORS. Published by Elsevier LTD on behalf of Chinese Academy of Engineering and Higher Education Press Limited Company. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).
英文原文：Engineering 2025, 44(1): 87–100
引用本文：Fei Wu, Tao Shen, Thomas Bäck, Jingyuan Chen, Gang Huang, Yaochu Jin, Kun Kuang, Mengze Li, Cewu Lu, Jiaxu Miao, Yongwei Wang, Ying Wei, Fan Wu, Junchi Yan, Hongxia Yang, Yi Yang, Shengyu Zhang, Zhou Zhao, Yueting Zhuang, Yunhe Pan. Knowledge-Empowered, Collaborative, and Co-Evolving AI Models: The Post-LLM Roadmap. *Engineering*, https://doi.org/10.1016/j.eng.2024.12.008.

构化数据，以及缺乏可解释性。为了提高模型的事实准确性、推理能力及可解释性，引入外部显式知识显得尤为重要。

• **LLMs在推理效率和垂直领域应用方面面临挑战。**由于LLMs通常规模庞大、计算资源消耗巨大，这使得它们在某些应用场景中变得不够实用。此外，LLMs的决策过程缺乏透明性，难以进行有效解释。为了解决这些问题，研究者开始探索模型协同（model collaboration）的方式，结合大型模型强大的能力与垂域小型或专用模型的高效性与专业性，旨在提升整体性能、可用性和可解释性，以弥补LLMs的不足之处。

• **LLMs在适应性和持续学习方面存在局限。**它们通常依赖固定数据集进行训练，当需要融入新的信息或应对不断变化的任务环境时，往往需要重新训练，耗时且成本高。为了实现模型的持续演化与自适应，须整合相互促进的学习过程，使多模型能够联合进化，在动态环境中保持实效性。

为应对上述挑战，以实体及其关系的语义化表示为代表的外部知识的引入成为前沿热点研究方向[5–7]。目前，相关的技术路径主要包括以下几种方法。

（1）将外部知识融入训练目标：设计知识感知的损失函数，增强模型对关键实体的关注。例如，可以为重要实体赋予更高的遮掩概率，在词元级损失和实体级损失之间进行平衡。

（2）将知识注入模型输入：在模型的输入序列中注入相关知识子图，同时减少无关知识噪声的干扰。

（3）知识赋能的指令微调：通过将外部知识结合入指令语料学习任务进行微调，使模型更好地理解和应用外部知识。

（4）推理阶段的检索增强知识融合：结合非参数化或参数化检索模块，使LLMs动态获取所需的背景知识。

（5）知识提示：通过将外部知识直接转换为文本提示供LLMs调用，无需重新训练模型。

在提升效率及面向特定领域的准确性方面，已有研究[8–10]探索了大语言模型（LLMs）与小模型（SMs）之间的协同机制。本文所述的“小模型”是指参数量远小于LLMs且不具备涌现能力的模型。通过模型协同，研究者将不同架构、规模与功能的模型有机结合，互相配合，增强整体性能。大模型具有通用语言理解和推理能力，而小模型则具有较高的计算效率和专业化特性。二者的结合可以提升整体准确性，改善模型的可解释性和计算效率。模型协同通常包括模型合并、功能型模型协同等策略。

为了进一步提升模型的适应性和持续学习能力，前沿研究通过大语言模型与小模型之间的交互演化过程，在多模态任务中兼顾高性能与高效计算。模型联合进化即指多个模型在协同处理复杂多样任务时，随着时间推移不断相互影响与共同演化的动态过程。各模型通过共享知识、参数和学习策略，相互促进，能够适应不同的模型结构、任务需求以及数据分布等异构环境。通过联合进化，模型能够在专用化与通用化之间取得平衡，增强鲁棒性与效率，适用于隐私和资源受限的分布式或联邦学习场景。

在这一框架下，知识赋能、协同与联合进化共同构成了一个紧密联系的整体，推动人工智能超越单一模型的局限，向更高层次的推理、精度和适应性迈进。三者之间互为依赖：知识赋能可借助协同机制整合与利用外部知识资源；协同机制则通过使各模型通过交互促进共同学习，实现联合进化；而进化进一步加强了知识赋能和协同机制的效果，推动模型在持续适应与学习的过程中不断演进。

在网络空间、物理世界与人类社会（cyberspace, physical world, and human society, CPH）构成的“三元空间”不断拓展的背景下，科学、工程与社会三者之间的交互也拓展出日益丰富的发展维度。在这一发展过程中，后大模型时代（post-LLM era），知识赋能、协同与进化等相关技术的涌现，为应对复杂跨学科交叉场景提供了重要的技术支持。如图1所示，技术的融合能够有效应对假设构建、问题表述、问题求解与可解释性方面的复杂挑战。在假设构建阶段，人工智能模型能够借助领域知识提高准确性与可靠性；在问题表述环节，多智能体系统通过对实体、环境及规律的建模，推动问题的表达与理解。例如，教育场景可通过模拟个性化角色探索教学原理，物理建模可结合物理约束神经网络（PINNs）与流体力学、热传导等问题中的实际物理定律，提高预测精度[11–12]。在问题求解过程中，由符号逻辑推理器向大规模神经网络的转

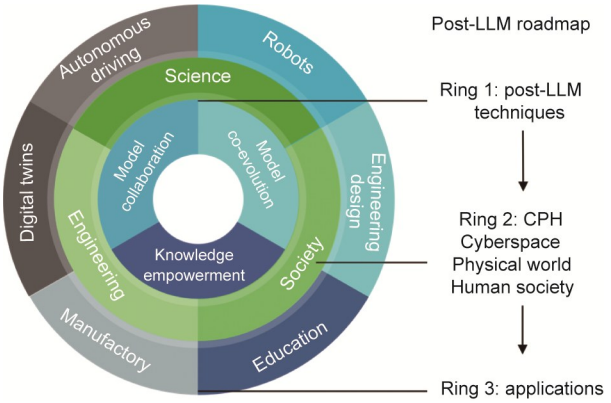

图1. 后大模型时代的发展蓝图。

变，促使模型不能从数据库中检索知识，通过记忆生成完整的解答。多智能体协作系统通过将计算与验证任务分离，进一步增强了数学求解能力。可解释性方面，先有研究通过将标准化操作流程（SOPs）纳入多智能体工作流，进一步细化任务分解与协调，并借助改进的人机交互机制，使得多个基于LLM的智能体能够通过自然语言与编程方式协作，在不断完善软件开发过程的同时提升了可解释性。

本文将重点讨论后LLM时代相关技术，剖析当前科学、工程与社会领域所面临的挑战，并探讨未来可行的发展方向。图2展示了本文的整体结构：第2节介绍了传统方法在知识赋能、协同与联合进化方面的挑战；第3节综述了知识赋能大型语言模型；第4节介绍前沿模型协同方法；第5节介绍前沿模型联合进化技术；第6节探讨了知识赋能、协同与联合进化如何进一步推动科学、工程与社会的发展；第7节对知识赋能、协同与进化在未来可能的应用及进展进行了展望；第8节则总结了全文的主要观点与结论。

2. 挑战

在本节，我们归纳了当前人工智能模型面临的四大类型的挑战：任务异质性、模型异质性、数据异质性以及安全与隐私问题。

2.1. 任务异质性

现有的人工智能模型主要针对不同的任务、场景和应用进行开发，这些任务之间具有不同甚至相互冲突的优化目标和评估指标，从而导致了这些任务特定模型之间的协作和联合进化面临理论和实践上的挑战。我们在任务异质性中识别了三种研究挑战。首先，训练目标的差异可能会阻碍模型的进化过程，尤其是在模型训练阶段，优化生成对抗网络就是一个典型的例子，其中生成器和判别器以对抗的方式联合优化，难以达到平衡状态。因此，平衡不同目标并稳定训练动态是一个具有挑战性的难题。其次，缺乏共享的知识会阻碍模型之间的协作与共同演化。尽管针对完全不同任务的模型可能会发展出独特的专业知识，但没有共同框架，这些知识很难在任务之间进行跨任务的共享和利用。再次，由于不同模型之间的交互障碍，模型之间很难达成共识，从而使得解释和处理其他模型输出变得困难。

2.2. 模型异质性

模型异质性主要指不同人工智能模型之间架构差异较大，这给异构模型的协作进化带来关键挑战。典型的例子包括不同架构和复杂度层次的模型之间的协作，以及具有不同学习范式的模型之间的协作。首先，输入和输出表示的差异可能会使特征对齐变得困难。例如，两个卷积骨干网络的深度和宽度（如混合模型协作）可能存在显著差异，导致神经元和特征图的数量不同，最终使得模型协作

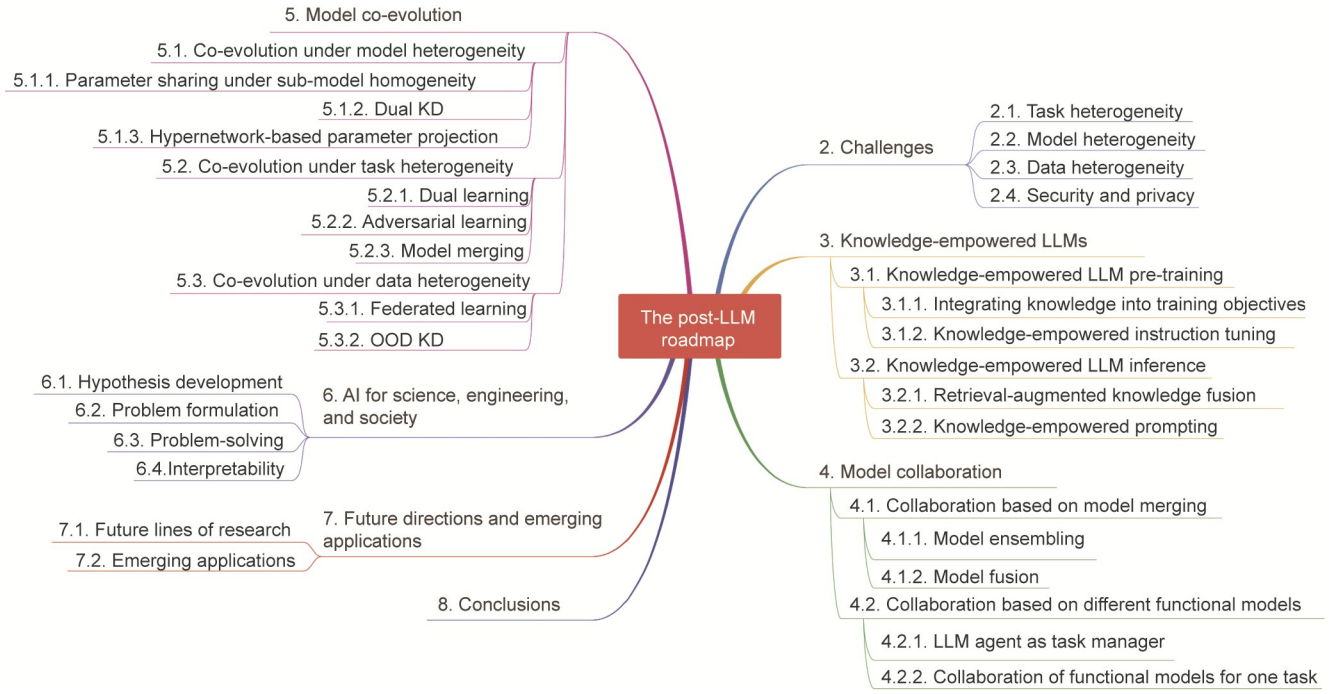

图2. 本文的整体结构。

变得不灵活。此外，由于中间表示的不兼容，两种根本不同的学习范式可能对模型协作构成重大挑战。例如，如何实现符号人工智能与联结主义人工智能之间的协作，就需要将逻辑规则转化为数值格式或反之。因此，有效地提取、传递和对齐异质模型之间的共享知识，将有助于促进模型的高效利用。

2.3. 数据异质性

在现实场景中，来自不同设备或来源的数据往往不是独立同分布（non-IID），这导致数据分布存在显著差异。例如，从不同终端用户、功能各异的传感器（如具身智能）、不同患者或不同企业收集的数据可能在特征、标签和领域上存在差异，进而引发类别不平衡、协变量偏移和概念漂移等现象，这些现象显著影响模型协调的泛化性能。此外，来自多个来源的数据可能存在标签不一致或标注质量不均的情况。例如，一些数据可能被错误标注或包含噪声，这可能会在模型协作或共演化过程中导致性能下降。数据模态的差异带来了额外的挑战，如不一致的数据表示（如空间图像与时间音频等）和不平衡的数据模态（如缺失或稀疏模态）。数据异质性带来了特定的挑战，主要包括数据分布的差异、模态特定的挑战、融合困难和训练复杂性。有效解决这些挑战对于构建能够处理多源数据、复杂任务的协作与鲁棒系统至关重要。

2.4. 安全与隐私

在法律和法规的保护下[如通用数据保护条例（GDPR）]，数据安全和隐私是模型协作中的重要问题，尤其是在分布式和去中心化的机器学习系统中，在这些系统中，多个实体（如设备和组织）共同训练全局模型，但不会共享原始数据。尽管原始数据未直接共享，但通过梯度或特征进行的模型更新仍然可能无意间暴露有关基础数据的敏感信息。例如，梯度中的某些模式可以通过逆向工程来重构原始数据。此外，协作模型容易受到推理攻击，这些攻击利用已学习的模型或其输出推测训练数据的信息，典型的威胁包括模型反演攻击和成员推理攻击。尽管差分隐私等机制提供了确保隐私的潜在解决方案，但该框架注入的过多噪声可能会降低整体模型性能。因此，在协作环境中，如何在保护隐私的同时保持模型的效用，仍然是一个重要的挑战。此外，模型还容易受到中毒攻击，攻击者试图通过注入恶意更新来破坏全局模型。例如，针对特定的子任务或目标，攻击者可以发送恶意模型的梯度更新，故意降低全局模型的性能。虽然鲁棒的聚合机制（如拜占庭容错算法）可以检测恶意客户端，但设计这种机制是复杂的，尤其是在客户端贡献多样且其可信度无法假定的环境中。随着协作式人工智能系统的不断发展，安全与隐私问题将继续是确保安全和隐私保护的模型协作范式发展的重要层面。

3. 知识赋能的大型语言模型

大型语言模型普遍采用大规模语料库上的无监督学习方法进行训练，然而这种训练方式导致模型在直接应用于实际场景时，可能欠缺必要的世界知识，从而在需要高度专业知识的任务中遇到适用性瓶颈。为了解决这一问题，研究者探索并提出了若干策略，旨在通过引入外部知识资源来增强大型语言模型的能力。这些策略包括预训练阶段融入知识导向的训练目标以实现知识内化、通过调整模型输入来引入相关领域信息以及在指令微调和推理计算过程中动态地利用外部知识源。接下来，本文将深入探讨上述各项技术细节，解释它们是如何促进模型获取更广泛的知识体系，并提升其执行复杂任务的表现。

3.1. 知识赋能的预训练

大型语言模型主要通过无监督学习在海量文本数据上进行训练，因此它们对现实世界知识的应用存在局限。针对这一挑战，当前的研究工作大致可以归结为两个主要方向：一是将外部知识嵌入训练目标中；二是利用知识资源进行指令微调。

3.1.1. 将知识融入训练目标

Zhou等[13]提出通过构建高质量的小规模知识集和特定的微调框架，使预训练模型能够更好地适应用户的交互风格。Akyürek等[14]在训练目标中引入领域专业知识，比较了两种不同的训练数据归因方法——基于梯度的TDA（Training Data Attribution）和基于嵌入的TDA，评估这两种方法在整个训练周期内对模型预测性能的影响，并将其与传统仅依赖词汇相似度的BM25方法进行对比，检验不同方法在事实追踪任务中的表现。许多研究开发知识敏感的训练目标，让模型更有效地感知并运用外部知识。例如，Shen等[15]提出利用知识图谱结构来确定掩码概率，对于在一定距离内可达的重要实体赋予更高的掩码优先级，以此在预训练阶段强化模型对关键实体的理解。Zhang等[16]进一步探索了词元级损失和实体级损失之间的平衡，以提高模型学习实体知识的能力。Tian等[17]借鉴了类似的概念，在预训练期间注入情感信息，通过对正面或负面情感词增加掩码概率，使模型能够识别基本的情

感特征。在这个过程中，模型不仅接收文本输入，还学习文本词元与知识图谱中实体之间的对应关系。Gao等[18]在输入的词元中添加实体向量，并设定实体预测任务，以加强模型理解实体的能力。Wang等[19]采取综合策略，在同一个Transformer编码器上执行知识图谱嵌入和掩码预测任务，模型接受语言和知识图谱的训练。确定性大型语言模型[20]捕捉确定性的事实性知识，它只对包含确定实体的片段实施掩码操作，并引入线索对比学习和线索分类目标。Xiong等[21]将文本中的实体替换为同类型的其他实体，然后让模型判断是否发生了替换，借此在预训练阶段实现对实体知识的注入。

3.1.2. 知识赋能的指令微调

Ji等[22]提出基于压缩协议的方法，在优化对齐精度与保持预训练分布丰富性之间做出动态调整。Zhang等[23]分析了多种指令微调技术及其应用效果。Gekhman等[24]研究了在微调数据集中引入“Unknown”实例的影响，发现当这类实例占比过高时可能会导致过拟合现象，削弱模型跨不同知识域的泛化能力；相反，“Maybe-Known”实例有助于在不同类型的现有知识间取得性能上的平衡。通过结合知识图谱进行指令微调，研究人员能够将事实信息和图结构转换为自然语言形式的微调数据集，从而使经过微调后的模型不仅能够掌握具体事实，还能理解并利用知识图谱的结构来进行推理。Wang等[25]开发了一系列提示模板，将图结构表达为自然语言，并设计了两种自监督任务提升模型对知识图谱内事实和结构信息的理解。OntoPrompt [26]方法在提示中加入本体论信息，嵌入实体知识到模型上下文中，针对多个下游任务进行了微调。Luo等[27]探索了在知识图谱上微调模型以生成逻辑查询的方法。Luo等[28]提出“规划-检索-推理”的框架，该框架首先在知识图谱结构上微调模型以生成关系路径，随后利用这些路径检索出合理的推理链条，以此支持模型执行可信且可解释的推理过程。

3.2. 知识赋能的LLM推理

面对现实世界知识的快速变化，预训练阶段提升知识水平的方法难以在不重新训练的前提下及时更新模型。近年来研究界探索推理阶段知识与文本独立表征的方法。

3.2.1. 检索增强的知识融合

Ovadia等[29]提出在推理过程中通过辅助知识库检索引入用户查询相关的最新信息，与模型现有的上下文相融合，提高模型在知识密集型任务中的回答质量。这种方法相较于传统微调方法，能够动态地引入最新的领域相关知识，适应不断变化的环境。RAG（retrieval-augmented generation）是这一领域的代表性方法，结合了非参数化和参数化模块，实现了检索与生成的协同。Yang等[30]设计了一个迭代多阶段流程IM-RAG（iterative multi-stage RAG），该系统集成了“推理器”“检索器”“精炼器”和“进度追踪器”，结合强化学习和监督微调，使得大型语言模型可以逐步进行信息检索、修正和综合，完成结构化的推理过程并最终输出答案。对于给定的输入文本，系统通过最大内积搜索找到关联文档[31]，并将其作为潜在的隐含变量提供给模型，提升响应准确性和内容丰富度。Lewis等[32]的研究表明，RAG方法在开放域问答任务中显著优于多种基线模型，能够产生更加具体且事实正确的回答。Wu等[33]通过将外部知识编码进记忆模块并加速检索，提高了推理效率。Guu等[34]建议在预训练期间引入知识检索机制，加强模型的开放域问答能力。Logan等[35]从当前上下文中选择适当的事实片段，并将其构造为完整的句子，使模型生成逻辑连贯的回答。Zhang等[36]提出的两阶段框架，针对古籍修复中的组合优化问题，首先使用多模态大模型进行初步匹配，然后利用神经求解器在开放域环境中优化片段的选择。Sun等[37]尝试将知识图谱三元组表示成序列并与文本拼接，存在一定引入噪声的风险，后续Sun等[38]的研究通过采用统一的词-知识图结构减轻了这一问题。Zhang等[39]识别长尾实体，用伪造嵌入替换这些实体以改进实体表征。Yu等[40]在输入文本后附加稀有词汇的定义，指导模型对其进行对齐和判别，借助外部词典改善稀有词汇的表征。

3.2.2. 知识赋能的提示

在推理阶段，知识赋能提示方法通过将结构化知识转换为文本序列，以提示的方式输入到LLM中。Li等[41]采用预定义的模板，将知识图谱三元组转换为简短的句子。Luo等[42]从知识图谱中采样关系路径并将其转换为自然语言，输入到模型中帮助生成逻辑规则。Chain-of-knowledge（CoK）[43]利用一系列三元组作为提示，诱导模型展现更强的推理能力。在无需重新训练模型的情况下，知识赋能提示为结合LLM和KG提供了一种简单可行的手段。

4. 模型协同

人工智能模型协同的研究是一个日益重要的领域，聚焦于不同规模、结构或功能模型之间的紧密合作。模型协同的目标是利用各模型的优势，实现单一模型无法达到的

性能或效率。人工智能模型协同方法不仅关注大模型与小模型之间的互补性，还涉及不同类型模型的整合，如深度学习模型与传统机器学习模型的结合，以同时发挥大模型的强大能力和小模型的高效性与可解释性。随着深度学习技术的快速发展，大模型因其卓越的性能而备受关注；然而，它们通常需要大量的计算资源，这限制了它们在资源受限环境中的应用，另外由于深度大模型的黑盒属性，使得其决策过程通常难以理解。因此，探索模型协同模式以增强性能和可用性已成为研究热点。

根据协同策略的不同，模型协同可以分为两类：第一类是模型合并，以专家混合（MoEs）[44]为例，通过结合多个相对较小的专家模型来实现甚至超越大模型的性能；第二类涉及具有不同功能的模型的协作，例如，利用大模型代理来规划专用小模型完成特定任务[45]。

4.1. 基于模型合并的协同

在机器学习领域，单一模型通常难以实现最佳性能。模型合并是一种有效的策略，通过结合多个模型的预测结果、结构或参数，来改进预测精度和鲁棒性，从而弥补单一模型的不足。

4.1.1. 模型集成

模型合并的一种形式是模型集成，通过聚合各个模型的预测结果来实现[46]。最简单的模型集成方法是简单平均法，即通过对所有模型的预测结果取平均值获得最终预测。然而，该方法仅在每个分类器性能相似的情况下具有合理性。如果某个分类器的性能显著差于其他分类器，则最终预测结果可能不如组内最佳分类器的预测效果。一种更优的分类器集成方法是加权平均法，其中权重通过验证集学习得出。对于分类问题，投票（voting）[47]是常用的模型集成策略。通过让多个模型对预测类别进行投票选出最终预测，票数最多的类别即为最终预测类别。投票可以通过运行不同模型或对同一模型运行多次实现。堆叠（stacking）[48]是一种更复杂的模型集成方法，它使用多个不同模型的预测结果作为输入，训练一个新的模型来生成最终预测。这种方法能够有效利用不同模型的预测能力。

4.1.2. 模型融合

混合专家模型（MoE）[49]是一种稀疏门控深度学习模型，由两个关键组件组成：门控网络（GateNet）和专家网络（Experts）。门控网络根据输入数据的特性动态决定激活哪个专家模型，以生成最佳预测结果。专家网络是一组独立模型，每个模型专注于处理特定子任务。通过门控网络，输入数据被分配到最适合的专家模型进行处理，不同模型的输出则被加权融合以获得最终预测结果。例如，Mixtral 8×7B [50]是对Mistral 7B模型的修改版，它是一个稀疏MoE模型，每层包含8个专家模型。这使其成为一个拥有47B参数的模型，在多个基准测试中能够媲美甚至超越更大的模型（如Llama2 70B [51]）。

基于模型合并的协同计算可以整合各类模型的优势和专长，减少单一模型可能引发的偏差和错误，从而提高决策的准确性和可靠性。此外，模型融合还能增强模型的可解释性和透明性。例如，在MoE系统中，可以清晰识别每个专家模型在特定任务中的作用和贡献，为最终决策提供更明确的解释。

4.2. 基于多种模型功能的协同

另一种典型的模型协同方法是由多个具有不同功能或能力的模型组成的智能代理系统。虽然大模型具备广泛的知识和高级推理能力，如数学推理、编程和任务规划[52]，但在处理特定领域任务时，它们的准确性可能不及较小的专用模型。因此，需要一种有效的机制，将大模型的通用能力与小模型的专用知识相结合，确保代理系统能够灵活应对不同任务和环境。

基于不同功能模型的协同可以分为两种类型：一种类型的协同中，大型语言模型（LLM）充当智能代理，作为任务管理者调用各种专用模型完成不同任务；另一种类型中，大型语言模型与其他专用模型（如扩散模型）协同工作，共同完成特定任务。在大型语言模型的支持下，任务可以更高效地执行。

4.2.1. LLM代理作为任务管理者

研究者已经开始基于大型语言模型（LLMs）与专用小模型之间的协作，构建智能代理系统[53]。具体而言，他们将LLM作为这些代理的“大脑”或控制器，通过调度专用小模型来扩展感知和行动能力。早期的研究旨在增强LLM的工具调用能力。例如，工具增强语言模型（TALMs）[54]和Toolformer [55]通过对大语言模型进行微调，使其学习使用外部工具的应用编程接口（API）。HuggingGPT [56]更进一步，将LLM作为大脑、专用小模型作为工具，通过LLM和专用小模型的协作解决复杂问题。

思维链（CoT）[45]、思维树（ToT）[57]和思维图（GoT）[58]技术使基于LLM的代理能够展现出与符号代理和强化学习代理相媲美的推理与规划能力[59]。这些系统还可以从反馈中学习并执行新操作，获得与环境交互的

能力[60]。基于LLM的代理还能够交互，形成多代理系统（multi-agent），从而促进多个代理之间的协作与竞争[61]。

4.2.2. 多功能模型协作完成单一任务

大语言模型（LLMs）可以帮助专用模型更高效地完成特定任务。例如，在图像生成任务中，虽然Stable Diffusion [62]能生成高质量图像，但在严格根据提示生成输出方面存在一定困难。LLMs能更好地理解提示，并引导生成模型的行为，从而提高图像生成过程的可控性。Wu等[63]提出了一个生成框架，根据输入提示生成图像，评估图像与提示的匹配程度，并对生成图像中的偏差进行自我修正。在LLM控制器的引导下，这一框架将文本到图像生成转化为一个迭代的闭环过程，确保生成图像的正确性。此外，Wang等[64]提出了一种无需训练的文本到图像可控生成和编辑方法。该方法利用多模态大型语言模型（MLLMs）的推理能力，改进扩散模型中的综合表现。它通过局部扩散将复杂的图像生成任务分解为不同子区域的简单任务，并将文本引导的生成和编辑集成到一个闭环系统中，从而提高了泛化能力。

某些专用小模型也可以增强多模态大型语言模型（MLLMs）的能力。例如，Sachin等[65]使用语义分割和实例分割等视觉模型来提升MLLMs在目标计数任务中的表现。

5. 模型联合进化

模型联合进化是指在面对复杂且异质性较高的任务场景时，多个模型通过动态协作与共同演化的方式，相互交流和共享知识的过程。在这一过程中，各模型不仅依托各自独立的学习路径进行自适应优化与性能提升，同时还能够影响其他模型的演化方向，从而实现跨任务的泛化能力、参数共享与知识迁移。对于架构差异、任务需求或数据分布存在显著差异的应用场景，联合进化能够有效平衡模型的专用性与泛化性，共同应对异质性带来的挑战。通过这种联合进化，模型的鲁棒性与效率将得到提升，使其能够在更广泛的任务条件下保持较好表现，适用于在资源受限与隐私保护等通常出现在分布式和联邦环境中的约束性场景下。

本节将从模型、任务和数据三个层面的异质性出发，探讨模型联合进化的策略。第5.1节将重点讨论模型异构条件下的联合进化，介绍包括参数共享、双向知识蒸馏以及基于超网络的参数投影等关键技术。第5.2节将关注任务异构环境下的双向学习、对抗式学习和模型合并等方法。第5.3节则聚焦于数据异质性，深入探讨联邦学习与分布外知识蒸馏等策略。每节内容将围绕如何优化多样化环境中的模型协作与效率展开详细讨论。

5.1. 模型异质性下的联合进化

5.1.1. 子模型同构条件下的参数共享

在子模型结构同构的前提下实现参数共享，近年来已取得显著进展。此方法不仅在平衡模型专用化学习与共享参数效率方面起到了重要作用，而且推动了多任务环境中模型的协同进化。Haller等[66]提出了“稀疏共享”（sparse sharing）的概念，该方法通过在较大模型中利用相互重叠的子网络，并结合基于Lottery Ticket假设的迭代式幅度剪枝，有效提高了参数的利用效率。随后，Ding等[67]将该思想扩展为多级别的稀疏共享模型，通过在不同层次上实施更细粒度的共享与专用特征控制，实现了更高效的参数配置。Wang等[68]提出了多任务提示微调，通过将共享知识蒸馏至可转移的提示中，帮助大型语言模型在不同任务间实现高效适应。Zhang等[69]在视觉质量评估任务中采用共享编码器，并通过对比学习动态调整共享参数，从而提升了模型性能。Chen等[70]提出了针对目标检测Transformer的组检测Transformer，通过在目标查询中引入分组的参数共享机制，显著提高了检测效率。Ghosh等[71]在联邦学习场景中提出了迭代式联邦聚类算法，通过在不同用户集群间共享表示层，进一步保留了集群特定的学习特点。Ye等[72]提出了OpenFedLLM框架，结合联邦指令微调和价值对齐等技术，在去中心化系统中实现了大型语言模型的协同训练，避免了原始数据的暴露。这些研究成果共同表明，参数共享不仅能提升模型的效率和减少冗余，还能在多样化任务需求的场景中提供稳健的性能。

5.1.2. 双向知识蒸馏

在模型联合进化的研究范式中，双向知识蒸馏（dual knowledge distillation, dual KD）逐渐成为应对模型异质性的重要策略。该方法使得模型同时担任教师和学生的双重角色，在不同结构的模型之间实现双向的知识传递，从而促进互补学习。与传统的单向蒸馏方法不同，双向蒸馏更强调模型间的相互学习。例如，互斥式对比学习[73]、自适应跨结构互学蒸馏[74]以及多合一知识蒸馏[75]等框架均基于这一思想。AIO-KD通过在多个学生模型间进行同步优化，采用动态梯度分离和互学策略，提升了知识交换效率，同时保持了教师模型的高性能。在半监督学习的应用中，多阶段协作知识蒸馏[76]通过多次细化伪标签，减

少了过拟合现象并增强了序列生成任务的泛化能力。双向蒸馏方法在文本到图像合成等任务中同样至关重要，例如，某些自适应教师-学生协作机制[77]通过“oracle”方式迭代指导学生模型的输出。此外，知识过滤同样关键，诸如Selective-FD [78]等方法通过过滤具有歧义或分布外问题的预测，确保在联邦学习环境中的高效且准确的知识共享。总体而言，双向知识蒸馏通过迭代协同学习机制，不仅有效应对了不同结构和领域之间的差异，还能够显著提升模型的性能和泛化能力。

5.1.3. 基于超网络的参数投影

在联合进化体系中，基于超网络的参数投影方法逐渐成为解决模型异质性，特别是大规模预训练模型异质性的重要手段。超网络最初是为了生成目标网络的权重而提出的，在跨模型知识转移的场景中，超网络通过学习从共享潜在空间到各异质模型参数空间的映射，协调了多模型间的信息传递。对于已经在不同任务或领域上完成微调的模型，超网络提供了一种统一的机制来融合各个模型的表示。利用超网络，模型能够在无需大规模再训练的情况下，动态生成适应目标模型的专用参数，便于模型根据输入或任务需求进行灵活切换。在知识融合方面，超网络可以无缝整合来自异质模型的输出，例如，在“用于大型语言模型的知识融合”[79]和“混合适应”[80]等方法中，均强调了不同模型间的词元化或适应模块对齐。该方法可以与模型集成[81]及集成策略[82]等其他模型聚合技术结合使用，提供更加灵活的参数空间探索方式，同时保持各模型的独特特征。此外，回归平均[83]和基于排序的融合[84]等方法，通过无需下游数据的方式进行模型参数合并，进一步展示了超网络在参数融合中的灵活性。通过对参数投影空间的有效探索，超网络能够在异质环境中构建更具一致性和高效性的联合进化流程。

5.2. 任务异质性下的联合进化

5.2.1. 双向学习

双向学习是近年来应对任务异质性挑战的关键范式之一，其通过利用成对任务之间的内在对偶性来提高学习效率并提升跨领域性能。在无偏学习排序领域，Yu等[85]提出了一种结合列表输入和点输入的上下文双向学习算法。该方法通过自注意力机制捕捉列表输入模型的局部上下文信息，同时，点输入模型则用于蒸馏相关性判断，从而有效缓解了位置和上下文的偏差，并在Baidu-ULTR数据集上取得了优于现有方法的效果。在约束优化场景中，Park与Van Hentenryck引入了自监督原-对偶学习[86]，通过模拟增广拉格朗日法同时训练原模型与对偶模型，实现了对可行性与最优性之间的平衡，显著降低了约束违反程度和最优解偏差。此外，Fei等[87]通过对任务间结构信息的对齐，提出了文本到文本生成中的句法结构共同回响与交叉重构策略，并在机器翻译和图像描述等任务中取得了显著性能提升。针对视频描述任务，Ji等[88]提出了基于注意力机制的双向学习方法，通过多头注意力机制建立视频与描述之间的双向交互，关注有效信息，生成更加准确和连贯的描述。Li等[89]提出的多次双向学习框架，用于风格化对话生成，模型通过映射不同上下文与风格回复之间的关系，并引入判别器确保风格一致性，从而获得了最先进的结果。此外，双向学习在模型联合进化中的潜力还体现在自反射翻译[90]与动态知识蒸馏框架[91]等工作中，这些方法通过双向学习反馈机制或对比语言-图像预训练模型，成功应对了任务异质性的挑战。

5.2.2. 对抗学习

对抗学习在任务异质性下的模型联合进化中具有重要作用，其核心思想是通过模型或模块间的对抗博弈提升不同任务间的鲁棒性、对齐性和整体性能。在词汇简化任务中，某些LLM增强对抗编辑系统采用了混淆损失与不变损失，以预测词汇编辑并确保语义一致性，同时区分复杂词和简单词[92]。另外，潜在对抗训练方法[93]通过生成针对性对抗样本来触发并缓解有害输出，从而去除语言模型中的不良行为。在AI文本检测领域，研究者通过对抗学习在改写模型与检测器之间建立对抗关系，增强了检测器对改写攻击的鲁棒性[94]。在最差类对抗训练[95]中，研究者集中关注最劣势类别，通过无悔动力学改善类别不平衡所带来的对抗鲁棒性问题。弱监督语义分割领域的研究[96]则利用分类器与重构器之间的对抗学习来引导分类器生成更加精确的类别激活图，进而提升分割精度。这些方法展示了对抗机制在应对任务异质性挑战中的巨大潜力。

5.2.3. 模型合并

模型合并是解决任务异质性问题的有效手段，其目标是在多个异构任务之间构建一个统一的模型，同时尽量减少相互干扰。最基础的合并方法是参数平均[97]，尽管该方法简洁，但由于任务冲突的存在，往往导致性能不理想。为了解决这一问题，研究者提出了基于加权的合并方式，如球面线性插值[98]，通过衡量各模型或任务向量的重要性来优化合并系数。此外，逐层或逐参数加权的合并方法也得到了发展，如逐层自适应模型合并[99]与基于

Fisher信息加权的合并方式[100]，这些方法通过自适应调整不同层或参数的权重，进一步提升了合并效果。另一类基于子空间的方法，包括剪枝与重标[101]及去除和重缩放[102]，主要通过剪除不重要的参数，利用神经网络的过参数化特性合并稀疏子空间，从而减少任务间的干扰。在推理阶段，路由策略能够动态调整合并过程，以适应不同输入的差异。例如，双重合并[103]与基于权重集成的专家网络[104]通过路由网络引导合并过程，优化了合并效果。最后，后校准技术，如表示修正[105]，通过对合并模型与独立模型表示进行对齐，缓解了表示偏差，从而进一步提升了合并后的模型性能。

5.3. 数据异质性下的联合进化

5.3.1. 联邦学习

OpenFedLLM [72]构建了一整套训练流程，包括联邦指令微调和联邦价值对齐（FedVA），增强了模型执行指令的能力和对齐度，在不同数据方之间有效保障隐私安全。Zhang等[106]提出了多模态大语言模型辅助的联邦学习，通过全局多模态预训练、联邦微调和全局对齐三大步骤，显著提升了在异质且长尾数据分布下的联邦学习性能，同时有效减轻了客户端的隐私风险和计算负担。Bai等[107]提出了针对LLM的联邦学习方案，该方案能够根据客户端的资源情况动态调整LoRA的秩值，从而提升模型在任务和资源异质性环境下的泛化能力。为了进一步改善协同模型的表现，FedMKT [108]采用了基于最小编辑距离的联邦选择性互惠知识转移与token对齐机制，从而实现LLM与小模型在协作过程中的性能互补与共同提升。针对如何有效增强小模型能力，Li等[109]提出了从LLM生成并提炼通用及领域专有知识的方案，并在保护隐私的基础上，将这些知识传递给本地小模型。Fan等[110]提出利用逐步蒸馏方法，将LLM的能力分阶段传递给小模型，并通过先进的提示与推理编码技术，确保在知识扰动及随后的蒸馏过程中保持信息的完整性。

5.3.2. 分布外知识蒸馏

知识蒸馏通常将功能强大的LLM视为教师模型，通过训练资源友好且高效的小模型来传授其能力，从而在保持模型性能的同时降低计算资源的占用。然而，传统的合成或无数据蒸馏方法在处理超出分布场景时，常常会出现性能下降的问题。为应对这一挑战，Gholami等[111]提出了一种与具体任务无关的分布外知识蒸馏框架，通过反复利用LLM的反馈迭代优化小模型模型，显著提升其在分布外环境下的泛化能力。Li等[112]针对视觉-语言模型在分布外蒸馏场景中的问题，通过改进教师模型的提示一致性和语言表达方式，使其能够更好地对齐教师与学生在视觉-语言任务上的表现。Agarwal等[113]提出了泛化知识蒸馏方法，利用基于强化学习的微调技术对训练与推理过程进行对齐，并根据教师模型对学生模型输出结果的反馈进行调整，从而提高模型的泛化能力。Chen等[114]采用了一种扰动蒸馏策略，从分数、类别和实例三个层面对模型进行修正，帮助小模型更好地学习教师模型的知识，从而有效应对领域泛化的挑战。

6. 面向科学、工程与社会的人工智能

后大模型时代的到来标志着人工智能在多个领域（特别是在科学、工程和社会领域）的角色发生了重大转变。这些领域既有共同的挑战，又有各自独特的特征。因此，这就需要面向各领域对人工智能方法进行量身定制。如图3所示，本节将详细综述面向科学、工程和社会领域人工智能开发流程的各个环节，包括假设提出、问题定义、问题解决以及方法可解释性强化，探讨了知识、协作和共同演化如何支撑这些开发环节。

6.1. 假设提出

假设的提出是科学、工程和社会领域共同面临的基础性挑战。根据领域的不同，假设的形式可能各不相同。在科学研究中，假设通常是旨在解释自然现象的理论命题，并且通常需要通过实证检验[29]。例如，科学研究中的假设可能会预测某个特定变量对生物过程的影响，或者预测在特定条件下化学反应的结果。在工程研究中，假设往往表现为实现特定目标或满足运行约束而设定的任务目标[115]。例如，电力系统、空间站或自动驾驶汽车等复杂系统的运行，通常需要做出关乎效率和安全的假设。这些假设更具实用性，可以作为系统设计和决策的基础，帮助工程师在给定条件下实现最优的运行和控制。在社会领域，假设通常与行为或政策结果相关[22]。例如，人工智能模型可能假设通过某些干预措施（如公共意识提升活动或基础设施调整）能够改善医疗普惠性或交通管理水平。这些假设通常会在模拟或试点项目中进行验证，之后再进行更广泛的实施。尽管不同领域假设的类型各异，但它们之间存在一些共通点。例如，虽然场景不同，但都是系统行为的预测，或仿真模型的不同设置在投入实际应用前的基础性预测。这些共通的和独特的假设为各领域后续的问题定义和问题解决过程提供了指导。

在后大模型时代，知识赋能的人工智能模型在假设的

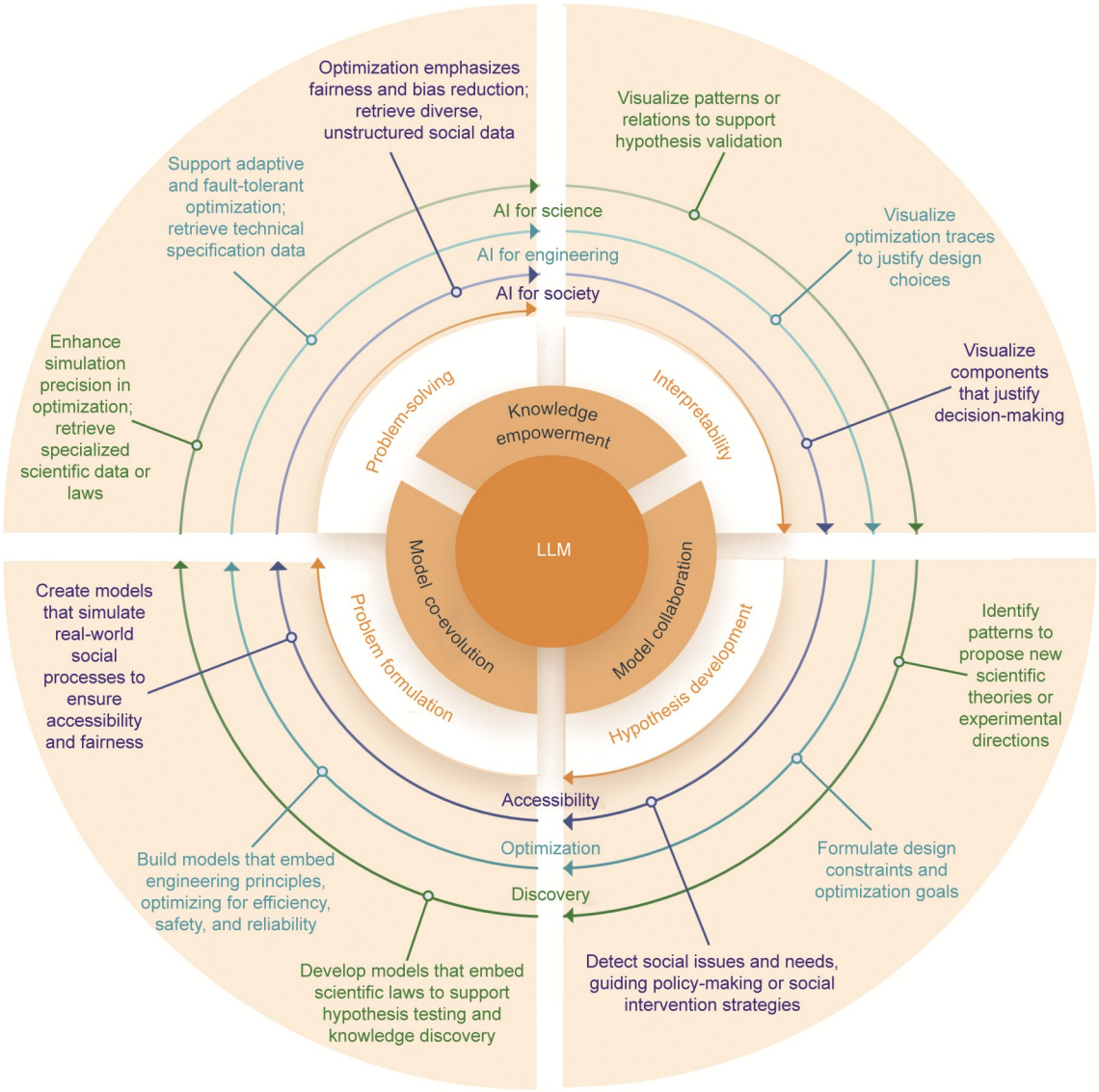

图3. 人工智能赋能科学、工程和社会的架构图。

构建中将发挥重要作用，因其融合了领域特定的专业知识，从而提高了假设的准确性和可靠性。例如，像“盘古”（Pangu）[116]、“风乌”（FengWu）[117]和“伏羲”（FuXi）[118]等先进的气象人工智能模型，可以与领域知识相结合，提升可再生能源（如风能和太阳能）预测的准确性，这对于将可再生能源源源不断地接入电力系统至关重要。多个小模型的协作通过多视角交叉验证的方式可以增强假设的鲁棒性，也将在假设验证中起到关键作用。这种协作方法有助于减少偏差并提供更全面的问题理解。此外，共同演化也将促进假设的迭代优化。通过从成功和失败中不断学习，模型可以发展出更准确有效的假设，继而对假设提出的转变具有深远影响，使得更深层次的理论推理和更广泛的数据驱动探索成为可能，推动科学、工程和社会领域的应用进步。共同演化的迭代过程使得假设能够更好地适应变化的环境，更加契合领域特定的挑战，并最终推动各自领域的发展和进步。

6.2. 问题定义

将大模型应用于对现实世界的建模，目前是科学、工程和社会研究的一个公共焦点。研究方向涉及三种类型的建模：客观实体的建模、客观环境的建模以及客观规律的建模。

对于实体建模，可通过引入多智能体系统来有效模拟教育场景中的个性化角色，如学生和教师[119]。在基于大型模型智能体的实体建模中，对不同个性化角色模拟规则的有效整合可用于揭示教育教学的规律，从而推动教育领域的发展。

对于环境建模而言，其关键挑战在于如何实现多个智

能体之间的有组织互动。对该过程的一种实践是利用多智能体创建虚拟课堂平台。该虚拟平台应用大模型智能体来模拟多名学生，并探索如何培养他们的学术能力。Yue等[119]将教学过程涉及的领域知识整合到课堂模拟中。利用精心设计的角色模拟器，通过有效的角色互动协调，来实现对课堂教学过程进行的细致探索。

探索物理世界的客观规律是人工智能发展的主要目标之一[120]。面向该目标提出了物理信息神经网络（PINNs）[11]，它利用物理定律来提高模型对物理世界预测的精度和泛化能力。与传统神经网络相比，PINNs能够利用较少的训练数据，获取产生符合物理定律预测的能力。此外，该模型对噪声和其他干扰的鲁棒性也更强。PINNs已在物理研究的许多领域得到广泛应用，如流体力学和热传导研究。在热传导研究中，它们可以帮助分析物理世界中的客观现象，如热扩散[121]。尽管PINNs的研究取得了巨大进展，但PINNs训练速度慢和收敛困难等问题仍然存在。此外，PINNs在处理高维数据和求解高维方程时表现不佳。

6.3. 问题解决

人工智能在科学、工程和社会领域已得到广泛应用[122]。在符号主义主导的人工智能发展时期，一些研究[123]设计了各种逻辑自动推理器，它们运用一阶逻辑和高阶逻辑进行科学研究，如自动数学证明器[124]和自动物理推理器[125]。然而，这些手动设计的推理器中存储的知识（即逻辑规则）数量通常有限，在解决更复杂的科学问题时表现可能不够理想。随着深度学习的兴起，一些研究者[123]开始将目光转向具有更强知识存储和利用能力的大型神经网络。此类研究依据神经网络的功能大致可被分为两类：一种方式是将深度学习模型设计为检索器[122]。深度学习模型负责从知识数据库中检索每个推理步骤所需的知识，从而协助逐步解决科学问题。另一种方式是将深度学习模型视为纯粹的记忆器[126]。在训练过程中，深度学习模型充分记忆知识。在后续的推理过程中，深度学习模型直接生成充分且完整的解决方案，无需检索知识数据库。

用于解决现实世界问题的大语言模型（LLMs），如数学模型DeepSeek_prover_v1.5 [126]，将神经网络归纳和存储领域知识的能力发挥到了极致。随着大型模型的发展，多个智能体的合作也逐渐应用于科学研究领域。例如，通过模仿现实世界中计算与验证的分离，联合构建数学问题解决智能体和数学结论验证智能体[127]。这两种类型智能体的有效合作，实现了对数学问题更准确的解决。

6.4. 方法可解释性强化

在人工智能驱动的研究中，除了得出结论，解释方法的推理过程也是一个重要问题。为此，一些研究者提出了元编程框架MetaGPT [128]，它将标准操作程序（SOPs）整合到多智能体系统的流程中。该框架旨在增强任务分解和协调能力，这种能力对于管理复杂软件工程项目至关重要。通过将SOPs编码成提示序列，MetaGPT使智能体能够利用类似人类世界的专业知识进行操作，验证中间结果并减少预测错误。通过模仿人类专家的行为，这种整合SOPs的方法提高了模型操作的可解释性。提高模型的人机交互能力是提高可解释性的另一种方式。基于这一理念，Qian等[129]引入了一个新的框架，用于由多个基于大语言模型的智能体驱动的软件开发。这些智能体通过自然语言和编程语言进行交流以完成合作。该研究通过聊天链和去幻觉机制来提高所开发软件的完整性、可执行性和一致性。

支撑假设是实现模型可解释性的另一个重要目标。Fang等[130]提出了KANO，这是一种基于知识图谱增强的分子对比学习方法，它整合了化学领域的知识，提供可解释的分子表征和优越的预测性能。KANO通过生成功能提示来唤醒与下游任务相关的知识，从而增强模型预测结果的可解释性。Li等[131]介绍了modSAR，这是一种基于优化的定量构效关系（QSAR）建模技术，通过确定关键的断点特征和构建分段线性回归方程，提供透明且可解释的预测。该模型生成清晰规则并为分子片段分配SHAP（全称为SHapley Additive exPlanations）值的能力强化了对其预测合理性的解释。这种能力使modSAR成为药物发现领域中具备现实价值的工具。

7. 未来方向与新兴应用

7.1. 未来研究方向

除了上述讨论的主题外，还有若干重要且相关的领域值得进一步探索。

（1）**具身人工智能（embodied AI）**。具身人工智能是一个有前景的后大语言模型的发展方向。收集高质量的机器人数据集需要大量劳动力，并且过度依赖仿真数据加剧了仿真与现实之间的差距，这要求进行协同数据集创建并改进仿真器。有效地整合人类示范数据并推动复杂环境中的认知进展，对于构建自适应模型至关重要。此外，实现因果推理、持续学习和统一的评估基准，对于构建稳健、可扩展和具有普适性的具身人工智能系统也至关重要。

（2）**类脑人工智能（brain-like AI）**。受人类大脑结

构和功能启发的人工智能系统和算法，旨在模拟大脑的并行处理、适应性和高效性，从而增强计算模型。与神经科学的跨学科整合，通过采纳大脑学习、记忆和决策过程的洞察，可能产生能够紧密模拟人类认知功能的人工智能模型。此外，神经科学的进展也有望激发稳健的类脑人工智能模型，这些模型能够作出自然的情感和情境响应，提升人机共情和适应性的潜力。此外，开发可扩展、高效且负责任的AI框架，能够在现实应用中可靠运行，尤其是在资源受限或敏感领域，将是一个重要的机会。通过整合神经回路结构和适应性的洞察，研究人员可以提高类脑人工智能模型的韧性、高效性和透明度，最终朝着负责任、适应性强、能响应复杂人类需求的AI迈进。

（3）**非transformer基础模型（non-transformer foundation models）**。尽管transformer架构在大基础模型中占据主导地位，但一些替代架构展现出作为潜在替代方案的前景。例如，Hyena [132]通过将数据控制的门控与隐式参数化的长卷积结合，提出了一种高效结构，为大规模序列处理提供了次二次时间复杂度的解决方案。其他模型则利用状态空间模型（SSMs）[133]来实现线性扩展，并提高相对于传统transformer的效率。RetNet [134]通过用多尺度保持机制替代多头注意力，能够有效捕捉序列信息，同时减少内存使用并显著加速训练。因此，这些模型可视为有效的transformer替代方案。

（4）**LLM参与的模型生成（LLM-involved model generation）**。利用大语言模型（LLM），通过总结用户需求和一些领域内数据为潜在变量，进而解码生成小型、任务特定的模型，这些模型可直接用于预测[135]，这一方向可能是后大语言模型阶段的一个有前景的发展方向。

7.2. 新兴应用

在后大语言模型时代，下一代人工智能将以知识赋能、模型协作和共同演化为特征，必将重新定义人工智能的能力，并重塑我们对这些新型AI系统的认知。其持续发展的特性将为我们的现实社会带来新的可能，满足日益复杂、专业化、适应性强且更符合人类需求的应用。

知识赋能的特性意味着，后大语言模型的AI系统将更加注重融合更多专业、事实性和结构化的信息，显著提高它们在特定领域的专长，具备精确性和逻辑推理能力，并最终超越今天的通用型人工智能模型。特别是，随着来自科学、工程和人类社会丰富知识来源的整合，下一代AI预计将深入探索科学规律，生成新的科研假设和发现，并预测事件的演变轨迹。例如，在数学领域，AI的赋能作用将更加普及，大规模神经网络将被用于存储数学知识并进行推理，同时通过多代理协作提高问题解决的准确性。这也将惠及其他新兴的跨学科AI领域，如在线教育、物理学等。例如，AI的个性化应用将出现在教育领域，通过模拟和整合学生与教师之间的互动洞察，丰富教学互动和体验。在物理学领域，诸如PINNs等技术将利用物理法则提升模型的预测准确性和泛化能力。

后大语言模型时代的模型协作将涉及更加深入的异构数据与异构模型的合作。通过融合来自多种来源（如文本、图像、音频和传感信号）的数据，全模态AI系统将能够获得对物理世界的更加全面的理解，这在自动驾驶、跨媒体内容生成和数字孪生等领域将特别有用。大模型（通用型）与小模型（专业型）之间的协作也是协作AI中的另一个新兴趋势。大模型在生成、推理和知识整合方面表现出强大的能力，而小模型则具有效率高、延迟低、安全性强、隐私保护好等优点。实现大模型和小模型之间的更深层次协作是未来的发展趋势，这不仅涉及有效的数据交换，还包括知识共享和任务分解，以应对复杂的任务场景，特别是在具身智能等领域。随着应用场景的拓展，结合大模型和小模型的个性化和自适应协作系统将成为智能助手和服务机器人等领域的重要发展方向。

受生物生态系统启发的模型联合进化是建立下一代人工智能的另一个关键要素，在这种模式下，人工智能模型通过相互依赖的过程共同进化、学习和适应。群体智能模型之间的这种动态和持续演化关系预计将显著提高AI系统的智能水平和适应能力，增强其应对动态和未知物理世界变化的鲁棒性。合并多样化功能模型可能是联合进化AI系统的潜在方法，因为它可以将多个模型的信息综合成一个统一的框架。然而，仍有若干关键挑战亟待解决，包括对合并机制的理论理解不足（尤其是涉及在不同数据集或任务上训练的模型时）以及合并方案的高计算成本和内存成本等问题。联合进化的人工智能模型具有广泛的应用潜力，如自动驾驶、矿业机器人和工业制造等。

知识赋能、协作和联合进化的AI预计将使AI系统达到一个新的层次，具有更高的智能、韧性和自主性，拓展AI处理复杂现实应用的能力，如科学发现、工程设计、个性化教育、制造等。同时，人工智能模型的自主性和互联性不断增强，也可能带来安全性和社会影响方面的挑战，需要建立机制对AI系统进行监控和控制，以防止其出现不可预见的行为。

8. 结论

本文围绕后大模型时代的发展蓝图，重点探讨了知识引导、模型协同及联合进化等新兴研究范式。大型语言模型（LLM）在垂域应用依然面临诸多固有局限，本文围绕构建更加鲁棒、可扩展且具备高度适应性的人工智能模型，探讨了多种技术路径和应用场景，为后大模型时代研究与实践提供了可行的技术路线参考。随着人工智能技术的持续进步，亟需跨越领域、学科的壁垒，构建既具备强大的理解推理能力，又能够灵活适应多场景易变需求，与人类价值观始终对齐一致的人工智能系统。

致谢

本研究部分由国家自然科学基金委员会（62441605）资助。

Compliance with ethics guidelines

Fei Wu, Tao Shen, Thomas Bäck, Jingyuan Chen, Gang Huang, Yaochu Jin, Kun Kuang, Mengze Li, Cewu Lu, Jiaxu Miao, Yongwei Wang, Ying Wei, Fan Wu, Junchi Yan, Hongxia Yang, Yi Yang, Shengyu Zhang, Zhou Zhao, Yueting Zhuang, and Yunhe Pan declare that they have no conflict of interest or financial conflicts to disclose.

References

[1] Pan Y. 2018 special issue on artificial intelligence 2.0: theories and applications. Front Inform Technol Electron Eng 2018;19(1):1–2.
[2] Lyu YG. Artificial intelligence: enabling technology to empower society. Engineering 2020;6(3):205–6.
[3] Lyu YG, Wu F. Toward a more general empowering artificial intelligence. Engineering 2023;25:1–2.
[4] Lyu YG, Wu F. Further empowering humans in specific fields and rethinking AGI testing. Engineering 2024;34:1–2.
[5] Li DF, Xu F. Synergizing knowledge graphs with large language models: a comprehensive review and future prospects. 2024. arXiv:2407.18470.
[6] Pan S, Luo L, Wang Y, Chen C, Wang J, Wu X. Unifying large language models and knowledge graphs: a roadmap. IEEE Trans Knowl Data Eng 2024;36(7):3580–99.
[7] Kau A, He X, Nambissan A, Astudillo A, Yin H, Aryani A. Combining knowledge graphs and large language models. 2024. arXiv:2407.06564.
[8] Yuan B, Chen Y, Zhang Y, Jiang W. Hide and seek in noise labels: noise-robust collaborative active learning with LLMs-powered assistance. In: Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics; 2024 Aug 11–16; Bangkok, Thailand. Stroudsburg: Association for Computational Linguistics (ACL); 2024. p. 10977–1011.
[9] Hao Z, Jiang H, Jiang S, Ren J, Cao T. Hybrid SLM and LLM for edge-cloud collaborative inference. In: Proceedings of the Workshop on Edge and Mobile Foundation Models; 2024 Jun 3–7; Tokyo, Japan. New York City: Association for Computing Machinery (ACM); 2024. p. 36–41.
[10] Zhang K, Wang J, Ding N, Qi B, Hua E, Lv X, et al. Fast and slow generating: an empirical study on large and small language models collaborative decoding. 2024. arXiv:2406.12295.
[11] McClenny LD, Braga-Neto UM. Self-adaptive physics-informed neural networks. J Comput Phys 2023;474:111722.
[12] Sharma P, Chung WT, Akoush B, Ihme M. A review of physics-informed machine learning in fluid mechanics. Energies 2023;16(5):2343.
[13] Zhou C, Liu P, Xu P, Iyer S, Sun J, Mao Y, et al. LIMA: less is more for alignment. In: Proceedings of the Advances in Neural Information Processing Systems 36 (NeurIPS 2023); 2023 Dec 10–16; New Orleans, LA, USA. Trier: NeurIPS Proceedings; 2024.
[14] Akyürek E, Bolukbasi T, Liu F, Xiong B, Tenney I, Andreas J, et al. Towards tracing factual knowledge in language models back to the training data. 2022. arXiv:2205.11482.
[15] Shen T, Mao Y, He P, Long G, Trischler A, Chen W. Exploiting structured knowledge in text via graph-guided representation learning. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP); 2020 Nov 16–20; online. Stroudsburg: Association for Computational Linguistics (ACL); 2020. p. 8980–94.
[16] Zhang D, Yuan Z, Liu Y, Zhuang F, Chen H, Xiong H. E-BERT: a phrase and product knowledge enhanced language model for E-commerce. 2020. arXiv:2009.02835.
[17] Tian H, Gao C, Xiao X, Liu H, He B, Wu H, et al. SKEP: sentiment knowledge enhanced pre-training for sentiment analysis. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics; 2020 Jul 5–10; online. Stroudsburg: Association for Computational Linguistics (ACL); 2020. p. 4067–76.
[18] Gao T. Knowledge authoring and question answering with KALM. 2019. arXiv:1905.00840.
[19] Wang X, Gao T, Zhu Z, Zhang Z, Liu Z, Li J, et al. KEPLER: a unified model for knowledge embedding and pre-trained language representation. Trans Assoc Comput Linguist 2021;9:176–94.
[20] Li S, Li X, Shang L, Sun C, Liu B, Ji Z, et al. Pre-training language models with deterministic factual knowledge. In: Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing; 2022 Dec 7–11; Abu Dhabi, UAE. Stroudsburg: Association for Computational Linguistics (ACL); 2022. p. 11118–31.
[21] Xiong W, Du J, Wang WY, Stoyanov V. Pretrained encyclopedia: weakly supervised knowledge-pretrained language model. 2019. arXiv:1912.09637.
[22] Ji J, Wang K, Qiu T, Chen B, Zhou J, Li C, et al. Language models resist alignment. 2024. arXiv:2406.06144.
[23] Zhang S, Dong L, Li X, Sen Zhang XS, Wang S, Li J, et al. Instruction tuning for large language models: a survey. 2023. arXiv:2308.10792.
[24] Gekhman Z, Yona G, Aharoni R, Eyal M, Feder A, Reichart R, et al. Does fine-tuning LLMs on new knowledge encourage hallucinations? 2024. arXiv:2405.05904.
[25] Wang J, Huang W, Qiu M, Shi Q, Wang H, Li X, et al. Knowledge prompting in pre-trained language model for natural language understanding. 2022. arXiv:2210.08536.
[26] Ye H, Zhang N, Deng S, Chen X, Xiong F, Chen X, et al. Ontology-enhanced prompt-tuning for few-shot learning. In: Proceedings of the ACM Web Conference 2022; 2022 Apr 25–29; online. New York City: Association for Computing Machinery (ACM); 2022. p. 778–87.
[27] Luo H, Tang Z, Peng S, Guo Y, Zhang W, Ma C, et al. ChatKBQA: a generate-then-retrieve framework for knowledge base question answering with fine-tuned large language models. 2023. arXiv:2310.08975.
[28] Luo L, Li YF, Haffari G, Pan S. Reasoning on graphs: faithful and interpretable large language model reasoning. 2023. arxiv:2310.01061.
[29] Ovadia O, Brief M, Mishaeli M, Elisha O. Fine-tuning or retrieval? Comparing knowledge injection in LLMs. 2023. arXiv:2312.05934.
[30] Yang D, Rao J, Chen K, Guo X, Zhang Y, Yang J, et al. IM-RAG: multi-round retrieval-augmented generation through learning inner monologues. In: Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval; 2024 Jul 14–18; Washington, DC, USA. New York City: Association for Computing Machinery (ACM); 2024. p. 730–40.
[31] Mussmann S, Ermon S. Learning and inference via maximum inner product search. In: Proceedings of the International Conference on Machine Learning PMLR; 2016 Jun 20–22; New York City, NY, USA. Birmingham: Proceedings of Machine Learning Research; 2016. p. 2587–96.
[32] Lewis P, Perez E, Piktus A, Petroni F, Karpukhin V, Goyal N, et al. Retrieval-augmented generation for knowledge-intensive NLP tasks. In: Proceedings of the 34th International Conference on Neural Information Processing Systems;

2020 Dec 6 – 12; Vancouver, BC, Canada. New York City: Association for Computing Machinery (ACM); 2020. p. 9459–74.
[33] Wu Y, Zhao Y, Hu B, Minervini P, Stenetorp P, Riedel S. An efficient memory-augmented transformer for knowledge-intensive NLP tasks. 2022. arXiv: 2210.16773.
[34] Guu K, Lee K, Tung Z, Pasupat P, Chang MW. REALM: retrieval augmented language model pre-training. In: Proceedings of the International Conference on Machine Learning; 2020 Jul 13 – 18; online. Birmingham: Proceedings of Machine Learning Research; 2020. p. 3929–38.
[35] Logan R, Liu NF, Peters ME, Gardner M, Singh S. Barack's wife Hillary: using knowledge graphs for fact-aware language modeling. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics; 2019 Jul 28 – Aug 2; Florence, Italy. Stroudsburg: Association for Computational Linguistics (ACL); 2019. p. 5962–71.
[36] Zhang Y, Li H, Zhang S, Wang R, He B, Dou H, et al. LLMCO4MR: LLMs-aided neural combinatorial optimization for ancient manuscript restoration from fragments with case studies on Dunhuang. In: Leonardis A, Ricci E, Roth S, Russakovsky O, Sattler T, Varol G, editors. Computer vision—ECCV 2024; 2024 Sep 29–Oct 4; Milan, Italy. Cham: Springer; 2024.
[37] Sun Y, Wang S, Feng S, Ding S, Pang C, Shang J, et al. ERNIE 3.0: large-scale knowledge enhanced pre-training for language understanding and generation. 2021. arXiv:2107.02137.
[38] Sun T, Shao Y, Qiu X, Guo Q, Hu Y, Huang X, et al. CoLAKE: contextualized language and knowledge embedding. In: Proceedings of the 28th International Conference on Computational Linguistics; 2023 Dec 8 – 13; Barcelona, Spain. Stroudsburg: Association for Computational Linguistics (ACL); 2020. p. 3660–70.
[39] Zhang T, Wang C, Hu N, Qiu M, Tang C, He X, et al. DKPLM: decomposable knowledge-enhanced pre-trained language model for natural language understanding. Proc Conf AAAI Artif Intell 2022;36(10):11703–11.
[40] Yu W, Zhu C, Fang Y, Yu D, Wang S, Xu Y, et al. Dict-bert: enhancing language model pre-training with dictionary. 2021. arXiv:2110.06490.
[41] Li S, Gao Y, Jiang H, Yin Q, Li Z, Yan X, et al. Graph reasoning for question answering with triplet retrieval. 2023. arXiv:2305.18742.
[42] Luo L, Ju J, Xiong B, Li YF, Haffari G, Pan S. ChatRule: mining logical rules with large language models for knowledge graph reasoning. 2023. arXiv: 2309.01538.
[43] Wang J, Sun Q, Chen N, Li X, Gao M. Boosting language models reasoning with chain-of-knowledge prompting. 2023. arXiv:2306.06427.
[44] Shazeer N, Mirhoseini A, Maziarz K, Davis A, Le Q, Hinton G, et al. Outrageously large neural networks: the sparsely-gated mixture-of-experts layer. 2017. arXiv:1701.06538.
[45] Wei J, Wang X, Schuurmans D, Bosma M, Ichter B, Xia F, et al. Chain-of-thought prompting elicits reasoning in large language models. In: Proceedings of the 36th International Conference on Neural Information Processing Systems; 2022 Nov 28 – Dec 9; New Orleans, LA, USA. Red Hook: Curran Associates Inc.; 2022.
[46] Kraaijenbrink J, Wijnhoven F. Managing heterogeneous knowledge: a theory of external knowledge integration. Knowl Manag Res Pract 2008;6 (4):274–86.
[47] Dogan A, Birant D. A weighted majority voting ensemble approach for classification. In: Proceedings of the 2019 4th International Conference on Computer Science and Engineering (UBMK); 2019 4th International Conference on Computer Science and Engineering (UBMK 2019); 2019 Sep 11–15; Samsun, Turkey. New York City: IEEE; 2019. p. 1–6.
[48] Kwon H, Park J, Lee Y. Stacking ensemble technique for classifying breast cancer. Healthc Inform Res 2019;25(4):283–8.
[49] Du N, Huang Y, Dai AM, Tong S, Lepikhin D, Xu Y, et al. Glam: Efficient scaling of language models with mixture-of-experts. In International Conference on Machine Learning; 2022.
[50] Wang K, Xu Y, Wu Z, Luo S. LLM as prompter: low-resource inductive reasoning on arbitrary knowledge graphs. 2024. arXiv:2402.11804.
[51] Touvron H, Martin L, Stone K, Albert P, Almahairi A, Babaei Y, et al. Llama 2: open foundation and fine-tuned chat models. 2023. arXiv:2307.09288.
[52] Nayak A, Timmapathini HP. LLM2KB: constructing knowledge bases using instruction tuned context aware large language models. 2023. arXiv: 2308.13207.
[53] Wang H, Li R, Jiang H, Tian J, Wang Z, Luo C, et al. BlendFilter: advancing retrieval-augmented large language models via query generation blending and knowledge filtering. 2024. arXiv:2402.11129.
[54] Parisi A, Zhao Y, Fiedel N. TALM: Tool augmented language models. 2022. arXiv preprint arXiv:2205.12255.
[55] Schick T, Dwivedi-Yu J, Dessì R, Raileanu R, Lomeli M, Hambro E, et al. Toolformer: language models can teach themselves to use tools. In: Proceedings of the Thirty-Seventh Conference on Neural Information Processing Systems NeurIPS; 2023 Dec 10; New Orleans, LU, USA. New York City: Association for Computing Machinery (ACM); 2023.
[56] Shen Y, Song K, Tan X, Li D, Lu W, Zhuang Y. HuggingGPT: solving AI tasks with ChatGPT and its friends in hugging face. 2023. arXiv:2303.17580v4.
[57] Yao S, Yu D, Zhao J, Shafran I, Griffiths T, Cao Y, et al. Tree of Thoughts: deliberate problem solving with large language models. In: Proceedings of the Thirty-Seventh Conference on Neural Information Processing Systems NeurIPS; 2023 Dec 10; New Orleans, LU, USA. New York City: Association for Computing Machinery (ACM); 2023.
[58] Besta M, Blach N, Kubicek A, Gerstenberger R, Podstawski M, Gianinazzi L, et al. Graph of Thoughts: solving elaborate problems with large language models. 2024. arXiv:2308.09687v4.
[59] Rabby G, Auer S, D'Souza J, Oelen A. Fine-tuning and prompt engineering with cognitive knowledge graphs for scholarly knowledge organization. 2024. arXiv:2409.06433.
[60] Ein-Dor L, Toledo-Ronen O, Spector A, Greta S, Dankin L, Halfon A, et al. Conversational prompt engineering. 2024. arXiv:2408.04560.
[61] Yu Z, Ouyang X, Shao Z, Wang M, Yu J. Prophet: prompting large language models with complementary answer heuristics for knowledge-based visual question answering. 2023. arXiv:2303.01903.
[62] Lu X, Liao Y, Liu C, Lio P, Hui P. Heterogeneous model fusion federated learning mechanism based on model mapping. IEEE Internet Things J 2022; 9 (8):6058–68.
[63] Wu TH, Lian L, Gonzalez JE, Li B, Darrell T. Self-correcting LLM-controlled diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2024 Jun 16–22; Seattle, WA, USA. New York City: IEEE; 2024. p. 6327–36.
[64] Wang Y, Zhu S, Fu F, Miao X, Zhang J, Zhu J, et al. Efficient multi-task large model training via data heterogeneity-aware model management. 2024. arXiv: 2409.03365.
[65] Sachin DN, Annappa B, Hegde S, Abhijit CS, Ambesange S. FedCure: a heterogeneity-aware personalized federated learning framework for intelligent healthcare applications in IoMT environments. IEEE Access 2024;12:15867–83.
[66] Haller M, Lenz C, Nachtigall R, Awayshehl FM, Alawadi S. Handling non-IID data in federated learning: an experimental evaluation towards unified metrics. In: Proceedings of the 2023 IEEE International Conference on Dependable, Autonomic and Secure Computing (DASC); 2023 Nov 14 – 17; Abu Dhabi, UAE. New York City: IEEE; 2023. p. 0762–70.
[67] Ding K, Dong X, He Y, Cheng L, Fu C, Huan Z, et al. MSSM: a multiple-level sparse sharing model for efficient multi-task learning. In: Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval; 2021 Jul 11 – 15; online. New York City: Association for Computing Machinery (ACM); 2021. p. 2237–41.
[68] Wang Z, Panda R, Karlinsky L, Feris R, Sun H, Kim Y. Multitask prompt tuning enables parameter-efficient transfer learning. 2023. arXiv:2303.02861.
[69] Zhang W, Zhai G, Wei Y, Yang X, Ma K. Blind image quality assessment via vision-language correspondence: a multitask learning perspective. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2023 Jun 18–22; Vancouver, BC, Canada. New York City: IEEE; 2023. p. 14071–81.
[70] Chen Q, Chen X, Wang J, Zhang S, Yao K, Feng H, et al. Group DETR: fast DETR training with group-wise one-to-many assignment. In: Proceedings of the IEEE/CVF International Conference on Computer Vision; 2023 Oct 2 – 6; Paris, France. New York City: IEEE; 2023. p. 6633–42.
[71] Ghosh A, Chung J, Yin D, Ramchandran K. An efficient framework for clustered federated learning. In: Proceedings of the 34th International Conference on Neural Information Processing Systems; 2020 Dec 6 – 10; Vancouver, BC, Canada. Red Hook: Curran Associates Inc.; 2020.
[72] Ye R, Wang W, Chai J, Li D, Li Z, Xu Y, et al. OpenFedLLM: training large language models on decentralized private data via federated learning. In: Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining; 2024 Aug 25 – 29; Barcelona, Spain. New York City: Association for Computing Machinery (ACM); 2024. p. 6137–47.
[73] Yang C, An Z, Zhou H, Zhuang F, Xu Y, Zhang Q. Online knowledge distillation via mutual contrastive learning for visual recognition. IEEE Trans Pattern Anal Mach Intell 2023;45(8):10212–27.
[74] Ni J, Tang H, Shang Y, Duan B, Yan Y. Adaptive cross-architecture mutual knowledge distillation. In: Proceedings of the 2024 IEEE 18th International Conference on Automatic Face and Gesture Recognition (FG); 2024 May 27–31; Istanbul, Turkiye. New York City: IEEE; 2024. p. 1–5.
[75] Miao Z, Zhang W, Su J, Li X, Luan J, Chen Y, et al. Exploring all-in-one

knowledge distillation framework for neural machine translation. In: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing; 2023 Dec 6–10; Singapore. Stroudsburg: Association for Computational Linguistics (ACL); 2023. p. 2929–40.
[76] Zhao J, Zhao W, Drozdov A, Rozonoyer B, Sultan MA, Lee JY, et al. Multistage collaborative knowledge distillation from a large language model for semi-supervised sequence generation. In: Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics; 2024 Aug 11 – 16; Bangkok, Thailand. Stroudsburg: Association for Computational Linguistics (ACL); 2024. p. 14201–14.
[77] Starodubcev N, Fedorov A, Babenko A, Baranchuk D. Your student is better than expected: adaptive teacher-student collaboration for text-conditional diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2024 Jun 16–22; Seattle, WA, USA. New York City: IEEE; 2024. p. 9275–85.
[78] Shao J, Wu F, Zhang J. Selective knowledge sharing for privacy-preserving federated distillation without a good teacher. Nat Commun 2024;15:349.
[79] Wan F, Huang X, Cai D, Quan X, Bi W, Shi S. Knowledge fusion of large language models. 2024. arXiv:2401.10491.
[80] Wang Y, Agarwal S, Mukherjee S, Liu X, Gao J, Awadallah AH, et al. AdaMix: mixture-of-adaptations for parameter-efficient model tuning. 2022. arXiv: 2205.12410.
[81] Wortsman M, Ilharco G, Gadre SY, Roelofs R, Gontijo-Lopes R, Morcos AS, et al. Model soups: averaging weights of multiple fine-tuned models improves accuracy without increasing inference time. In: Proceedings of the International Conference on Machine Learning PMLR; 2022 Jul 17 – 23; Baltimore, MD, USA. Seattle: Proceedings of Machine Learning Research; 2022. p. 23965–98.
[82] Arpit D, Wang H, Zhou Y, Xiong C. Ensemble of averages: improving model selection and boosting performance in domain generalization. 2022. arXiv: 2110.10832.
[83] Jin X, Ren X, Preotiuc-Pietro D, Cheng P. Dataless knowledge fusion by merging weights of language models. 2022. arXiv:2212.09849.
[84] Perin G, Chen X, Liu S, Kailkhura B, Wang Z, Gallagher B. RankMean: module- level importance score for merging fine-tuned LLM models. In: Proceedings of the Findings of the Association for Computational Linguistics: ACL 2024; 2024 Aug 11–16; Bangkok, Thailand. Stroudsburg: Association for Computational Linguistics (ACL); 2024. p. 1776–82.
[85] Yu L, Bi K, Ni S, Guo J. Contextual dual learning algorithm with listwise distillation for unbiased learning to rank. 2024. arXiv:2408.09817.
[86] Park S, Van Hentenryck P. Self-supervised primal-dual learning for constrained optimization. Proc Conf AAAI Artif Intell 2023;37(4):4052–60.
[87] Fei H, Wu S, Ren Y, Zhang M. Matching structure for dual learning. In: Proceedings of the International Conference on Machine Learning PMLR; 2022 Jul 17 – 23; Baltimore, MD, USA. Seattle: Proceedings of Machine Learning Research; 2022. p. 6373–91.
[88] Ji W, Wang R, Tian Y, Wang X. An attention based dual learning approach for video captioning. Appl Soft Comput 2022;117:108332.
[89] Li J, Xia Y, Yan R, Sun H, Zhao D, Liu T, et al. Stylized dialogue generation with multi-pass dual learning. In: Proceedings of the 35th International Conference on Neural Information Processing Systems; 2021 Sep 6–14; online. Red Hook: Curran Associates Inc.; 2021.
[90] Chen A, Lou L, Chen K, Bai X, Xiang Y, Yang M, et al. DUAL-REFLECT: enhancing large language models for reflective translation through dual learning feedback mechanisms. 2024. arXiv:2406.07232.
[91] Dong J, Zhang M, Zhang Z, Chen X, Liu D, Qu X, et al. Dual learning with dynamic knowledge distillation for partially relevant video retrieval. In: Proceedings of the IEEE/CVF International Conference on Computer Vision; 2023 Oct 4–6; Paris, France. New York City: IEEE; 2023. p. 11302–12.
[92] Wang Y, Sun T, Li S, Yuan X, Ni W, Hossain E, et al. Adversarial attacks and defenses in machine learning-empowered communication systems and networks: a contemporary survey. IEEE Comm Surv and Tutor 2023;25(4):2245–98.
[93] Cheng P, Yang Y, Li J, Dai Y, Hu T, Cao P, et al. Adversarial preference optimization: enhancing your alignment via RM-LLM game. In: Proceedings of the Findings of the Association for Computational Linguistics: ACL 2024; 2024 Aug 11 – 16; Bangkok, Thailand. Stroudsburg: Association for Computational Linguistics (ACL); 2024. p. 3705–16.
[94] Tan K, Luo K, Lan Y, Yuan Z, Shu J. An LLM-enhanced adversarial editing system for lexical simplification. 2024. arXiv:2402.14704.
[95] Sheshadri A, Ewart A, Guo P, Lynch A, Wu C, Hebbar V, et al. Targeted latent adversarial training improves robustness to persistent harmful behaviors in LLMs. 2024. arXiv:2407.15549.
[96] Hu X, Chen PY, Ho TY. RADAR: robust AI-text detection via adversarial learning. In: Proceedings of the 37th International Conference on Neural Information Processing Systems; 2023 Dec 10–16; OrleansNew, LA, USA. Red Hook: Curran Associates Inc.; 2023. p. 15077–95.
[97] Thota S, Vangoor VKR, Reddy AK, Ravi CS. Federated learning: privacy-preserving collaborative machine learning. DLBSAR 2019;5:168–90.
[98] Goddard C, Siriwardhana S, Ehghaghi M, Meyers L, Karpukhin V, Benedict B, et al. Arcee's mergekit: a toolkit for merging large language models. 2024. arXiv:2403.13257.
[99] Yang E, Wang Z, Shen L, Liu S, Guo G, Wang X, et al. AdaMerging: adaptive model merging for multi-task learning. 2024. arXiv:2310.02575.
[100] Matena M, Raffel C. Merging models with fisher-weighted averaging. In: Proceedings of the 36th International Conference on Neural Information Processing Systems; 2022 Nov 28–Dec 9; New Orleans, LA, USA. Red Hook: Curran Associates Inc.; 2022. p. 17703–16.
[101] Yadav P, Tam D, Choshen L, Raffel C, Bansal M. TIES-MERGING: resolving interference when merging models. In: Proceedings of the 37th International Conference on Neural Information Processing Systems; 2023 Dec 10–16; New Orleans, LA, USA. Red Hook: Curran Associates Inc.; 2023.
[102] Le Yu BY, Yu H, Huang F, Li Y. Language models are super Mario: absorbing abilities from homologous models as a free lunch. 2023. arXiv:2311.03099.
[103] Lu Z, Fan C, Wei W, Qu X, Chen D, Cheng Y. Twin-merging: dynamic integration of modular expertise in model merging. 2024. arXiv:2406.15479.
[104] Tang A, Shen L, Luo Y, Yin N, Zhang L, Tao D. Merging multi-task models via weight-ensembling mixture of experts. 2024. arXiv:2402.00433.
[105] Yang E, Shen L, Wang Z, Guo G, Chen X, Wang X, et al. Representation surgery for multi-task model merging. In: Proceedings of the 41st International Conference on Machine Learning; 2024 Jul 21 – 27; Vienna, Austria. Seattle: Proceedings of Machine Learning Research; 2024.
[106] Zhang J, Yang HF, Li A, Guo X, Wang P, Wang H, et al. MLLM-FL: multimodal large language model assisted federated learning on heterogeneous and long- tailed data. 2024. arXiv:2409.06067.
[107] Bai J, Chen D, Qian B, Yao L, Li J. Federated fine-tuning of large language models under heterogeneous language tasks and client resources. 2024. arXiv: 2402.11505.
[108] Fan T, Ma G, Kang Y, Gu H, Song Y, Fan L, et al. FedMKT: federated mutual knowledge transfer for large and small language models. 2024. arXiv: 2406.02224.
[109] Li H, Zhao X, Guo D, Gu H, Zeng Z, Han Y, et al. Federated domain-specific knowledge transfer on large language models using synthetic data. 2024. arXiv: 2405.14212.
[110] Fan T, Kang Y, Chen W, Gu H, Song Y, Fan L, et al. PDSS: a privacy-preserving framework for step-by-step distillation of large language models. 2024. arXiv:2406.12403.
[111] Gholami M, Akbari M, Hu C, Masrani V, Wang J, Zhang Y. GOLD: generalized knowledge distillation via out-of-distribution-guided language data generation. 2024. arXiv:2403.19754.
[112] Li X, Fang Y, Liu M, Ling Z, Tu Z, Su H. Distilling large vision-language model with out-of-distribution generalizability. In: Proceedings of the IEEE/CVF International Conference on Computer Vision; 2023 Oct 4 – 6; Paris, France. New York City: IEEE; 2023. p. 2492–503.
[113] Agarwal R, Vieillard N, Zhou Y, Stanczyk P, Ramos S, Geist M, et al. On-policy distillation of language models: learning from self-generated mistakes. In: Proceedings of the Twelfth International Conference on Learning Representations; 2024 May 7–11; Vienna, Austria. London: ICLR; 2024.
[114] Chen Z, Wang W, Zhao Z, Su F, Men A, Meng H. PracticalDG: perturbation distillation on vision-language models for hybrid domain generalization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2024 Jul 16–22; Seattle, WA, USA. New York City: IEEE; 2024. p. 23501–11.
[115] Feng S, Sun H, Yan X, Zhu H, Zou Z, Shen S, et al. Dense reinforcement learning for safety validation of autonomous vehicles. Nature 2023;615(7953): 620–7.
[116] Bi K, Xie L, Zhang H, Chen X, Gu X, Tian Q. Accurate medium-range global weather forecasting with 3D neural networks. Nature 2023;619 (7970):533–8.
[117] Chen K, Han T, Gong J, Bai L, Ling F, Luo JJ, et al. FengWu: pushing the skillful global medium-range weather forecast beyond 10 days lead. 2023. arXiv:2304.02948.
[118] Zhong X, Chen L, Liu J, Lin C, Qi Y, Li H. FuXi-extreme: improving extreme rainfall and wind forecasts with diffusion model. 2023. arXiv:2310.19822.
[119] Yue M, Mifdal W, Zhang Y, Suh J, Yao Z. MathVC: an LLM-simulated multi-character virtual classroom for mathematics education. 2024. arXiv:2404.06711.
[120] Müller J, Zeinhofer M. Achieving high accuracy with PINNs via energy natural

gradient descent. In: Proceedings of the International Conference on Machine Learning; 2023 Jul 23–29; Honolulu, HI, USA. New York City: IEEE; 2023.

[121] Aymerich E, Pisano F, Cannas B, Sias G, Fanni A, Gao Y, et al. Physics informed neural networks towards the real-time calculation of heat fluxes at W7-X. Nucl Mater Energy 2023;34:101401.

[122] Yang K, Swope A, Gu A, Chalamala R, Song P, Yu S, et al. LeanDojo: theorem proving with retrieval-augmented language models. In: Proceedings of the 37th International Conference on Neural Information Processing Systems; 2023 Dec 10–16; New Orleans, LA, USA. Red Hook: Curran Associates Inc; 2024.

[123] Zhan B. AUTO2, a saturation-based heuristic prover for higher-order logic. In: Proceedings of Interactive Theorem Proving; 2016 Aug 22–25; Nancy, France; 2016.

[124] Steen A, Sutcliffe G, Scholl T, Benzmüller C. Solving modal logic problems by translation to higher-order logic. In: Proceedings of the International Conference on Logic and Argumentation; 2023 Sep 10 – 12; Hangzhou, China. Cham: Springer Nature Switzerland; 2023. p. 25–43.

[125] Foulis DJ, Randall CH. The empirical logic approach to the physical sciences. In: Hartkämper A, Neumann H, editors. Foundations of quantum mechanics and ordered linear spaces. Marburg: Advanced Study Institute; 1973. p. 230–49.

[126] Xin H, Ren ZZ, Song J, Shao Z, Zhao W, Wang H, et al. DeepSeek-prover-V1. 5: harnessing proof assistant feedback for reinforcement learning and Monte-Carlo tree search. 2024. arXiv:2408.08152.

[127] Zhou JP, Staats C, Li W, Szegedy C, Weinberger KQ, Wu Y. Don't trust: verify-grounding LLM quantitative reasoning with autoformalization. 2024. arXiv: 2403.18120.

[128] Hong S, Zheng X, Chen J, Cheng Y, Wang J, Zhang C, et al. MetaGPT: meta programming for multi-agent collaborative framework. 2023. arXiv: 2308.00352.

[129] Qian C, Liu W, Liu H, Chen N, Dang Y, Li J, et al. ChatDev: communicative agents for software development. In: Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistic; 2024 Aug 11 – 16; Bangkok, Thailand. Stroudsburg: Association for Computational Linguistics (ACL); 2024. p. 15174–86.

[130] Fang Y, Zhang Q, Zhang N, Chen Z, Zhuang X, Shao X, et al. Knowledge graph- enhanced molecular contrastive learning with functional prompt. Nat Mach Intell 2023;5(5):542–53.

[131] Li Y, Cardoso-Silva J, Kelly JM, Delves MJ, Furnham N, Papageorgiou LG, et al. Optimisation-based modelling for explainable lead discovery in malaria. Artif Intell Med 2024;147:102700.

[132] Poli M, Massaroli S, Nguyen E, Fu DY, Dao T, Baccus S, et al. Hyena hierarchy: towards larger convolutional language models. In: Proceedings of the International Conference on Machine Learning PMLR; 2023 Jul 23 – 29; Honolulu, HI, USA. Seattle: Proceedings of Machine Learning Research; 2023. p. 28043–78.

[133] Gu A, Dao T. Mamba: linear-time sequence modeling with selective state spaces. 2023. arXiv:2312.00752.

[134] Sun Y, Dong L, Huang S, Ma S, Xia Y, Xue J, et al. Retentive network: a successor to transformer for large language models. 2023. arXiv:2307.08621.

[135] Tang Z, Lv Z, Zhang S, Wu F, Kuang K. ModelGPT: unleashing LLM's capabilities for tailored model generation. 2024. arXiv:2402.12408.