智能系统学报
CAAI TRANSACTIONS ON INTELLIGENT SYSTEMS

智慧教育中的大语言模型综述

肖建力, 黄星宇, 姜飞

引用本文:
肖建力, 黄星宇, 姜飞. 智慧教育中的大语言模型综述[J]. 智能系统学报, 2025, 20(5): 1054-1070.
XIAO Jianli, HUANG Xingyu, JIANG Fei. A survey of large language models in smart education[J]. *CAAI Transactions on Intelligent Systems*, 2025, 20(5): 1054-1070.

在线阅读 View online: https://dx.doi.org/10.11992/tis.202406040

您可能感兴趣的其他文章

面向智能教育的自适应学习关键技术与应用
Key techniques and application of intelligent education oriented adaptive learning
智能系统学报. 2021, 16(5): 886-898 https://dx.doi.org/10.11992/tis.202105036

人工智能范式的革命与通用智能理论的创生
Paradigm revolution in artificial intelligence and the birth of general theory of intelligence
智能系统学报. 2021, 16(4): 792-800 https://dx.doi.org/10.11992/tis.202103042

多智能体分层强化学习综述
A survey on multi-agent hierarchical reinforcement learning
智能系统学报. 2020, 15(4): 646-655 https://dx.doi.org/10.11992/tis.201909027

人工智能中的封闭性和强封闭性——现有成果的能力边界、应用条件和伦理风险
Criteria of closeness and strong closeness in artificial intelligence——limits, application conditions and ethical risks of existing technologies
智能系统学报. 2020, 15(1): 114-120 https://dx.doi.org/10.11992/tis.202001001

图像情境下的数字序列逻辑学习
Number sequence logic learning in image context
智能系统学报. 2019, 14(6): 1189-1198 https://dx.doi.org/10.11992/tis.201905044

集对分析在人工智能中的应用与进展
Application and development of set pair analysis in artificial intelligence: a survey
智能系统学报. 2019, 14(1): 28-43 https://dx.doi.org/10.11992/tis.201803030

DOI: 10.11992/tis.202406040
网络出版地址: https://link.cnki.net/urlid/23.1538.tp.20250205.1354.002

智慧教育中的大语言模型综述

肖建力[1], 黄星宇[1], 姜飞[2]
(1. 上海理工大学 光电信息与计算机工程学院, 上海 200093; 2. 重庆市科学技术研究院, 重庆 401123)

摘　要: 近年来, 人工智能技术在教育领域的广泛应用正逐步革新现代教育的模式, 教育面临新的机遇和挑战。特别是随着大语言模型的兴起, 人工智能有望融入到教与学的过程中, 教育模式由传统的师−生二元模式正转变为师−生−机三元模式。文章以教育领域内应用的大语言模型为研究焦点, 介绍了大语言模型在教育中的特点。以当前主流的几种大语言模型为例, 详细阐述这些模型在教育中的实际应用情况, 总结了目前教育大模型的共性以及差异性特点。还探讨了如何开发和训练满足教育需求的定制化大语言模型, 这一过程对实际应用至关重要。基于训练完成的教育大模型, 进一步阐释了其存在的局限性, 并展望了未来教育领域可能出现的新型大模型及其发展趋势。

关键词: 人工智能; 智慧教育; 大模型; 教育技术; 自然语言处理; 教育应用; 多模态学习; 学习分析
中图分类号: TP18　　**文献标志码**: A　　**文章编号**: 1673−4785(2025)05−1054−17

中文引用格式: 肖建力, 黄星宇, 姜飞. 智慧教育中的大语言模型综述 [J]. 智能系统学报, 2025, 20(5): 1054–1070.
英文引用格式: XIAO Jianli, HUANG Xingyu, JIANG Fei. A survey of large language models in smart education[J]. CAAI transactions on intelligent systems, 2025, 20(5): 1054–1070.

A survey of large language models in smart education

XIAO Jianli[1], HUANG Xingyu[1], JIANG Fei[2]
(1. School of Optical-Electrical and Computer Engineering, University of Shanghai for Science and Technology, Shanghai 200093, China; 2. Chongqing Academy of Science and Technology, Chongqing 401123, China)

Abstract: In recent years, the application of artificial intelligence (AI) technology in education has gradually advanced the modern educational model. Education currently faces new opportunities and challenges. In particular, with the emergence of large language models (LLMs), AI is expected to be integrated into the teaching and learning processes. The traditional teacher-student binary model of education is transforming into a teacher–student–machine tripartite model. This study aims to focus on LLMs applied in the field of education and introduce the characteristics of them. It takes the current mainstream LLMs as examples and elaborates on their actual applications in education in detail. It summarizes the common and distinctive features of educational large models (EduLLMs). In addition, this study also discusses how to develop and train customized LLMs to meet the needs of education. This process is very important for practical applications. Based on the trained EduLLMs, this study further explains its limitations and explores the possibility and its development trend in new EduLLMs.

Keywords: artificial intelligence; smart education; large language models; educational technology; natural language processing; educational applications; multimodal learning; learning analytics

随着科技的迅猛发展和互联网技术的普及, 传统的教育模式正面临着深刻的冲击。传统的线下教学往往伴随着一系列问题, 例如以教师为中心、学生参与度不高、创新能力培养不足、标准化教学无法满足个体差异等。正是在这种背景下, “互联网+教育”[1]模式应运而生。在早期的“互联网+教育”中, 学生可以获取大量且高质量的教育资源, 摆脱了时间和地点的限制, 实现随时随地地学习。同时, 教师也可以通过线上线下相结合的教学方式, 来提升课堂的趣味性和课后的拓展性。然而, 这种教育模式虽然解决了部分传统教

收稿日期: 2024−06−24.　**网络出版日期**: 2025−02−05.
基金项目: 国家自然科学基金项目 (61603257).
通信作者: 肖建力. E-mail: audyxiao@sjtu.edu.cn.

©《智能系统学报》编辑部版权所有

育的问题, 但仍然存在学生个体差异和课堂参与度低等挑战。

随着大数据时代的到来, 海量的学习数据涌入。然而, 传统的机器学习模型难以应对如此庞大的数据量, 学习效率低、计算资源需求高成为了制约因素[2]。为了应对这些挑战, 近两年来, 大语言模型 (large language models, LLMs) 相继出现, 成为人工智能 (artificial intelligence, AI) 领域最热门的研究方向之一。这类模型能够更好地处理大规模数据, 提高学习的速度和精度, 为教育领域的发展提供新的可能性。

大语言模型是一种由数百亿甚至数千亿参数构建而成的神经网络模型, 能够有效地实现问答、翻译、聊天等自然语言处理任务[3]。由于其庞大的参数量, 大语言模型能够处理大规模数据和复杂任务。在教育领域中, 数据往往以文本形式存在, 包括各种问答、题目、教材等, 因此大语言模型与教育领域具有天然的契合性[4]。智慧教育则是将人工智能技术应用于传统教育上, 对其进行改良优化, 来消除一些存在的弊端[5]。教育大模型 (educational large models, EduLLMs) 是将大语言模型与教育相结合的产物。模型通过大规模数据集的训练, 能够提供个性化定制教学、辅助学习、教育管理以及教育评估等功能[6-7]。

由于教育大模型采用大量的学习数据, 因此能够准确地了解学生的学习特点和方式。教育大模型能捕捉到学生学习的薄弱部分, 通过个性化定制的方式, 对这方面进行加强。与早期的人工智能产品相比, 教育大模型具有上下文学习优势, 在问答环节中能生成准确的内容, 同时语句也更加流畅, 能够帮助学生解答各种问题。此外, 对于教师而言, 利用教育大模型能够准确地了解每位学生在学习上的优势与不足, 帮助教师评估每位学生[8], 调整和优化接下来的课堂教学, 从而更好地开展教育工作。

最近推出的教育产品多数采用大语言模型作为主要卖点来进行宣传。这些产品制造商深知未来教育的趋势是朝着多元化、个性化和精准化的方向发展[9]。然而, 教育大模型也存在一些限制。在学生层面, 尽管正常使用教育大模型可以解答疑惑、理解不懂的知识, 但是如果过度依赖, 学生会失去主动思考问题的能力[10]。因此, 开发者需要制定相应的限制措施。另外, 在教师层面, 如果教师之前没有接触过或使用过教育大模型, 他们可能无法充分发挥模型的功能。为此, 有必要为教师提供专业的培训, 使他们能够迅速上手并高效地使用教育大模型[8]。

本文首先对大语言模型、现代教育和教育大模型的特点进行总结, 其次列举了教育大模型的应用, 介绍了目前教育大模型的共性以及差异性特点, 再对教育大模型的训练步骤进行说明, 最后指出了教育大模型的局限性, 对未来教育大模型进行展望及总结。

1　大语言模型在教育中的特点

本章将探讨大语言模型的特点、现代教育的特点以及融合而成的教育大模型的特点, 如图 1 所示。

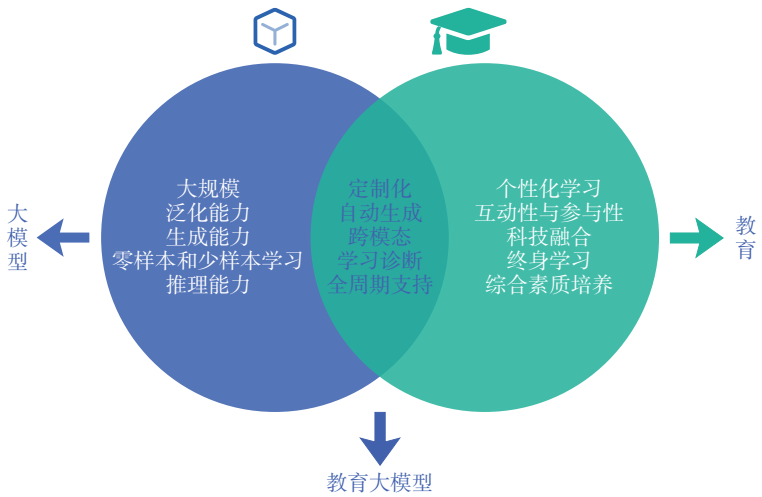

图 1　教育大模型的特点

Fig. 1　Characteristics of EduLLMs

1.1　大语言模型的特点

大规模　大规模是大语言模型的核心特点之一, 模型通常具有庞大的参数量和训练数据量。2017 年, Vaswani 等[11] 提出了基于自注意力机制

的 Transformer 架构后，各种大语言模型问世，从一开始十几亿、上百亿参数量的 GPT-2(generative pre-trained Transformer 2)[12] 模型和 T5(text-to-text transfer Transformer)[13] 模型，到后来有 1 750 亿参数的 GPT-3[14]，甚至万亿参数量的 PanGu-Σ[15]、GPT-4[16] 等。模型参数量越多，它在语义理解、信息生成等任务上会表现得更出色。同时在训练过程中，模型会使用海量的文本数据集，来捕捉语言的多样性和复杂性，提高模型在各种自然语言处理任务中的性能。

泛化能力　大语言模型的泛化能力[17]，指的是模型能够在未见过的数据上保持良好的性能，得益于在庞大的、多样化的数据集上学习到的语言结构和语义信息。利用 Transformer 提取特征，在少样本甚至零样本下，也能较好地完成任务。例如，GPT-3 能够灵活地适应多种语言任务，如写作、编程和翻译，展现出强大的泛化能力，成为自然语言处理领域的关键工具。

上下文理解和生成能力　大语言模型具有强大的上下文理解和生成能力，这种能力源于模型对大量文本数据的训练，使其能够理解语言中的细微差别、上下文关系和语义层次[18]。模型能够根据前文信息生成与主题相关的段落，甚至在长篇文章中保持一致的风格。这种上下文理解能力使得模型在问答、对话生成和故事创作等任务中都有高质量的输出，展现出强大的语言能力。

零样本和少样本学习　零样本学习指的是模型能够在未见过的任务上直接进行推理和生成，而少样本学习是允许模型在提供少量示例的情况下进行推理输出。大语言模型在零样本和少样本学习能力上表现出色[19-20]，用户向模型提供简单的指令或例子，模型可以自动推断出任务要求，生成符合预期的文本，使得大语言模型在实际应用中具有更广泛的适用性和更高的效率。

推理能力　大语言模型具有强大的推理能力，模型能够在文本生成和回答问题时进行逻辑推理和判断。这种能力使模型不仅能够识别和生成语言，还能理解其中的因果关系和逻辑结构[21]。在训练过程中，模型学习到文本中大量的隐含信息和推理模式，在面对复杂问题时，能够结合上下文推导出合理的结论。这种推理能力增强了模型在对话系统、知识问答和文本生成等任务中的表现，使其能够提供更精准、更符合逻辑的回答，甚至解决数学问题或进行情景分析。

1.2　现代教育的特点

个性化学习[22]　个性化学习是现代教育的一个重要发展方向，它根据每个学生的学习风格、兴趣、能力和进度来定制教学内容和方式。与传统的一刀切教育模式不同，个性化学习尊重学生的个体差异，为他们提供量身定制的学习资源和路径。这种方法不仅提升了学生的学习效果，还增强了他们的学习兴趣和参与度。通过大数据分析和人工智能技术，教师和教育机构也可以更准确地了解学生的需求，实时调整教学策略。

互动性与参与性[23]　互动性与参与性是现代教育强调的核心特征之一。教育不仅仅是知识的传递，更是一个师生、同学之间相互交流和启发的过程。通过小组讨论、项目学习和情景模拟等互动方式，学生可以更深入地参与到学习活动中来，增强对知识的理解和应用能力。相较于被动接受知识的学生，主动参与的学生对知识的掌握往往更为牢固。如今，在线教育平台和数字化工具的普及，为互动和参与提供了更多的可能性。例如，在线讨论、实时问答和虚拟课堂等应用，即使学生处于远程学习环境中，学习参与度也能保持高水平。这样的互动学习不仅提高了学生的解决问题的能力，还增强了团队合作和沟通技能。

科技融合[24]　随着科技的发展，教育与技术的融合成为必然趋势。人工智能、大数据、虚拟现实 (virtual reality, VR) 等技术的应用，为教育带来了革命性的变化。人工智能技术能够分析大量的学习数据，为学生提供准确的学习建议；大数据可以为教育决策提供科学依据，帮助学校制定更加合理的教学策略；而虚拟现实技术让学生能够体验到沉浸式的学习场景，让抽象的概念变得直观生动。这些技术的应用不仅提升了教学效率，还打破了传统教育对于时间和空间限制，让学生可以随时随地进行学习。例如，MOOC(massive open online course) 等教育应用让知识传播更加广泛，使更多人能够接触到高质量的教育资源，推动了教育的公平性。

终身学习[25]　现代教育越来越强调终身学习的理念，认为学习不应该局限于学校，而应该贯穿人的一生。随着科技和社会的快速发展，知识更新的速度加快，传统的学历教育不能够满足人们对新知识和新技能的需求。终身学习不仅是一种学习模式，更是一种适应现代社会的生存策略。通过鼓励终身学习，教育能帮助个体在职业生涯中不断更新自我、提升自己，以适应变化多端的职场需求和社会环境。这种终身学习理念为教育注入了持续创新的动力，帮助人们在各个阶段都能获取知识和技能。

综合素质培养　当今社会对于人才的需求已经从单纯的学术能力转向综合素质的发展。现代教育在重视知识传授的同时, 也越来越关注学生的情感智力、社交技能、创新能力和道德素养的培养。这种综合素质教育理念不仅是为了学生的个人成长, 也是为了帮助他们更好地适应未来社会的多元化挑战。例如, 许多学校设置了创新课程、心理健康咨询和社区服务项目, 来培养学生的全面素养。此外, 综合素质的培养还有助于学生在复杂的社会环境中找到平衡, 提升他们的适应力和领导力。通过综合素质的培养, 教育目标不再仅是知识的获取, 而是帮助学生成为具备多元能力的全面发展人才。

1.3　教育大模型的特点

定制化学习路径[26]　教育大模型通过大数据分析, 能够为每位学生量身定制独特的学习路径。这种路径不仅依据学生的知识水平, 还会参考他们的学习风格和兴趣点, 确保学习过程更加高效且符合个人需求。例如, 对于学习速度较慢的学生, 模型提供循序渐进的课程, 而对于学习能力较强的学生, 则推荐更具挑战性的内容, 以激发他们的潜力。

自动生成学习资源　教育大模型的生成能力让它成为学习资源的强大供应源。无论是试题、习题解析, 还是课程内容, 模型都可以自动生成[27], 甚至能根据课程进度动态调整内容。这意味着教师可以将更多时间投入到课堂教学和个性化指导中, 而学生也能获得丰富的学习材料。在日常学习中, 学生可以根据自己的需求获取实时解答和学习建议, 大大提升了学习的便捷性。

跨模态交互能力[28]　部分教育大模型支持多模态交互, 能够处理文字、图像、语音等多种输入形式, 极大丰富了学习体验。例如, 在物理实验课程中, 学生可以通过图像形式输入实验数据, 模型来生成相应的解释, 帮助学生理解实验过程。此外, 模型的语音识别和生成能力允许学生通过语音进行提问, 同时生成语音回答, 模拟真实课堂中的互动, 尤其适合低龄儿童和有特殊需求的学生。

精准学习诊断[29]　教育大模型通过对学生的学习数据进行分析, 可以提供精准的学习诊断。模型能够识别学生对知识的掌握程度, 生成详细的分析报告, 包括学生在哪些知识点上需要改进, 或者在哪些方面具备优势。教师可以基于这些数据进行针对性的教学安排, 而学生也能够根据报告及时调整学习计划, 弥补自身不足。精准的学习诊断帮助教师和学生都能高效地利用教学资源, 避免无效的重复学习。

全周期学习支持　教育大模型不仅服务于在校学生, 还为学习者提供终身支持。模型可以根据职业发展需求或个人兴趣提供个性化的学习资源, 帮助学习者不断更新知识和提升技能。对于职场人士, 教育大模型可以生成符合行业需求的培训课程和相关资料, 帮助他们提升竞争力。这样的支持体系确保了学习的连续性和灵活性, 让学习者在不同的阶段都能获得高效的知识支持。

2　教育大模型的应用

近年来, 大语言模型如谷歌的 PaLM[30]、T5, Meta 公司的 LLaMA[31] 还有 OpenAI 推出的 ChatGPT[32]、GPT-4 等, 它们的发展极大地推动了自然语言处理领域的进步, 它们能够做到文本生成、辅助翻译、问题回答、资料查找并总结等, 能够为学术研究和商业应用开辟广阔的道路。随着对自然语言处理特别是大语言模型的不断探索和改进, 大语言模型的能力进一步增强。得益于其出色的文本处理和生成能力, 大语言模型非常契合教育的应用。这些模型能够处理和分析大量的教育数据, 从而提供个性化的学习体验, 评估学生学习情况, 预测学习成果, 以及辅助教育决策等。

当前, 许多教育技术产品的核心在于通过人工智能和数据分析技术来提升教学体验。这些产品通常聚焦于特定的应用场景, 如个性化学习、辅助写作和研究[33]、评估与测试等。通过分析学生的学习行为、考试成绩以及其他相关数据, 为每位学生提供定制化的学习方案, 同时帮助教师更高效地管理教学过程和评估学生的学习成果。

在教育技术领域, 除了这些专注于特定任务的模型, 近年来还出现了一些更具代表性的教育大模型。这些大模型不仅是工具或辅助系统, 它们拥有强大的数据处理和分析能力, 能够处理和分析海量的教育数据, 包括学生的互动记录、作业提交、考试成绩以及教学内容等。通过挖掘大量数据中的模式和趋势, 这些教育大模型不仅为教育决策提供支持, 还可以直接介入教学过程, 创造自适应的学习环境, 并提供个性化的教学内容。

例如, 这些教育大模型在开放式问答、知识检索、作文批改、情感支持[34]、问题答疑等方面展现出强大的应用潜力。它们不仅能够动态调整学习内容以适应学生的需求, 还能为教师提供关于教学效果的实时反馈。教育大模型的广泛应用为教育学习带来了更多可能性, 使得学习过程更加高效、互动性更强。表 1 介绍了教育大模型在教育上的应用。

表 1　教育大模型的应用

Table 1　Application of EduLLMs

名称	研发团队	使用大模型	是否开源	特点和优势	适用范围
星火语伴	科大讯飞	讯飞星火认知大模型	是	支持多模态交互，具备中英文口语评测、语法纠错及文本问答功能，适合中文学习和口语训练场景	语言学习、口语练习、文本问答、考试模拟
EduChat[35]	华东师范大学	LLaMA和 Baichuan	是	智能问答、作文批改、启发式教学和情感支持，具备检索增强技术，实时更新知识库确保内容的时效性	K-12及高校教育、心理支持、作文评估
智海−三乐	阿里云	通义千问(7B)[36]	是	提供多学科支持，包含搜索、计算引擎和知识库功能，辅助高校课程和AI助教	高等教育课程辅助、AI助教
子曰	网易有道	自研大模型	否	整合多模态知识，提供个性化学习建议，模拟教师引导学生自我探索，满足不同的学习需求	K-12及高校教育、翻译、作文指导
MathGPT	好未来	自研大模型	否	专注于数学领域，具有准确的解题步骤分析，能详细讲解题目，帮助学生在数学学习中构建清晰思维	数学教学、解题演示、题目解析
智适应教育大模型	松鼠Ai	自研大模型	否	聚焦个性化学习，自适应调整教学内容，同时提供情绪支持，帮助学生在良好心态下学习	K-12教育、自适应学习、心理支持
汇雅大模型	超星集团	自研大模型	否	支持在线学习资源管理，整合数字图书馆内容，便于高校师生查阅和管理，适用于教育资源丰富的环境	高等教育、在线学习、教育资源管理
看云大模型	猿辅导	自研大模型	否	涵盖K-12阶段多学科辅导，具备实时答疑功能，帮助学生在课堂外及时解决疑问和巩固知识	K-12教育、题目解析、多学科辅导
Duolingo Max	多邻国	GPT-4	否	增强个性化互动体验，支持多语言学习，提供实时反馈，适合语言学习者提高听、说、读、写各方面技能	语言学习(多语言)、个性化练习
Khanmigo	可汗学院	GPT-4	否	适用于科学和数学教育，提供个性化学习建议和详细解答，帮助学生加深对学科概念的理解和应用	科学和数学教育、个性化辅导

2.1　垂直领域教育大模型

2023 年 6 月 9 日，科大讯飞发布了星火语伴 APP，搭载了讯飞星火认知大模型。该大模型于 2023 年 5 月 6 日发布，原本为通用领域大模型，经过优化适配后，成为了垂直领域教育大模型。该大模型在中文通用大模型评测基准 SuperCLUE (super Chinese language understanding evaluation)[37] 上位列中国第一，全球第三。星火语伴精通中英文，能够随时练习口语；也能作为文本问答工具进行对话；还能够对英语口语发音进行评测并纠错；可以现场模拟口语考试，帮助口语薄弱者提高口语水平。

2023 年 8 月 5 日，华东师范大学计算机科学与技术学院 EduNLP 团队推出了教育大模型 EduChat。这是一种基于大语言模型的智能教育聊天系统，主要应用于智慧教育领域。模型提出并解决了两个教育大模型存在的问题：第一个是大语言模型与教育专家之间的知识差距，不能与现实保持一致；第二个是大语言模型不能实时跟进教育领域的最新知识，还会产生幻觉问题。为了应对这些问题，EduChat 首先在教育书籍和数百万个定制指令上进行了预训练，以获得教育基础知识，并在心理学专家和教师的反馈指导下进行微调，让模型获取教育的特定功能。此外，它引入了检索增强技术，允许模型自动判断检索信息的有效性，并基于相关信息和模型内的知识生成回答，以确保回答的准确性和可信度。EduChat 也可以联网来获取最新的教育资源，确保问题回答的时效性。EduChat 的核心功能有 4 点：开放式问答是利用互联网上实时更新的语料库，EduChat 采用检索增强方法，使其能够自主评估并检索信息的相关性，显著提高了准确性；作文批改是让 EduChat 提供全面的作文评估，包括综合评分、方面级别评分和详细评论，以满足学生个性化指导的需求；启发式教学通过苏格拉底法[38] 的对话方式和多步骤的问答互动，目的培养学生的认知技能和自主学习能力，来提高批判性思维和创新能力；情感支持基于情绪心理学框架，提供个性化诊断和情感支持，分析用户的情感问题并深入了解用户的情感状态，提供准确和专业的

帮助。团队还开发了一个演示系统, 用户可以通过直观的界面选择多种功能, 如开放问答和情感支持等。用户可以轻松地与 EduChat 进行互动对话, 以帮助学生、教师和家长。该系统还具备自适应功能, 不断从用户交互中学习, 获得提高, 以便提供更加个性化和有效的帮助。未来, EduChat 计划扩展更多功能, 如职业规划、课程指导和问题生成等。

2023 年 9 月 18 日, 浙江大学与高等教育出版社、阿里云等单位共同研制教育领域垂直大模型“智海–三乐”, 此模型基于阿里云的通义千问 (7B, “B”代表“十亿”) 开源大模型进行开发, 将教材和论文这些高质量语料作为预训练数据, 用专业指令数据进行模型的微调, 同时拥有搜索引擎、计算引擎和本地知识库等功能。“智海–三乐”与《人工智能引论》[39] 相融合, 可作为学生学习计算机核心课程的重要工具之一, 也可作为 AI 助教、学习助手。模型的发布推动了产教融合, 使得教学迈向数字化和智能化。此外, 也有许多高校 (如清华大学、北京大学、南京大学等) 开始启用 AI 助教, 这些 AI 助教深度介入课程中, 能够提供 24 小时的个性化学习支持、评估和反馈, 同时帮助学生更好地思考, 获得创造性的灵感。在未来, AI 助教将在更多的课程中得到应用, 加速人工智能与教育相融合。

由于教育大模型刚处于起步阶段, 所以开源的教育大模型数量较少, 但是有不少企业正进军此领域, 积极地将教育大模型落地, 开发大模型的应用。由此, 2023 年 7 月 26 日网易有道发布了国内首个教育领域垂直大模型“子曰”。首先, 它能够根据学生的需求提供定制化的分析和建议。其次, 模型通过模拟教师的引导方式, 提出一些问题, 激发学生通过自主探索找到问题的答案。最后, 由于模型能够接入多模态知识库和整合跨学科的知识内容, 因此可以灵活应对学生的学习需求, 帮助学生培养综合能力。并且网易有道还发布了基于“子曰”大模型开发的一些应用, 如翻译、口语教练、作文指导、语法精讲、AI Box 及文档问答, 充分展现了大语言模型在教育领域的广泛应用前景和发展方向。在 2024 年 1 月 4 日, 网易有道发布教育大模型“子曰”2.0 版本, 在 1.0 版本基础上, 对模型的口语对话能力、知识问答能力和文字处理能力进行了升级。作为行业内首个教育类的垂直大模型, 其标志着教育大模型开始进入应用阶段, 推动了行业新发展。

2023 年 8 月, 好未来发布的学而思九章大模型 (MathGPT) 是在数学领域中应用的教育大模型, 主要是以解题和讲题为核心开发的。MathGPT 并没有基于现有的大语言模型进行开发, 而是自主研发, 这样虽然技术难度高, 但是能够打造出自主稳定的模型。模型能够做到解题准确率高、解题步骤稳定且清晰、解题过程有趣且富有个性化。在 CEval-Math(Chinese evaluation math)、AGIEval-Math(artificial general intelligence evaluation math)、APE5K、CMMLU-Math(Chinese multimodal learning for understanding math)、GAOKAO-Math、Math401 这 6 个数学测试集中, MathGPT 取得了多项测试的最高分, 并且在 C-Eval[40] 中学的测试中, 相比于 ChatGLM2(chat general language model) 和 ChatGPT 都有更好的表现。

2024 年 1 月 5 日松鼠 Ai 发布国内首个智适应教育大模型, 将多模态技术融合到教育大模型中。模型由数据层、模型层和应用层组成, 通过大规模的学习资料、效果、行为等数据输入到模型层中, 再对应用层面进行开发。该模型能够快速描绘出学生学习行为和习惯, 为学生提供个性化服务, 从而进一步提高学生的学习效率。此外, 人们往往只关注学习而忽略了情绪这一点, 情绪也会影响学习进度和学习效率, 所以模型同时会关注学生在学习中的反应, 给学生正面的反馈来改善学生的情绪, 解决许多潜在的心理问题, 来获得更好的学习状态。

2024 年 4 月 16 日超星集团发布超星汇雅大模型, 模型使用 30 年积累的海量图书、期刊、报纸等资源进行训练, 模型的参数量达到 340 亿。汇雅大模型具有文本生成、语言理解、知识问答和逻辑推理等核心功能, 支持多样化的教育场景应用。基于实际教学需求, 超星集团开发了一系列 AI 工具, 如机器阅读、视频理解、内容安全检测、相似度分析和 AI 助教等, 为科研和教育提供智能支持, 助力智慧校园建设。超星 AI 助教与泛雅网络平台深度融合, 为学生在课前、课中、课后全程提供个性化辅导。模型还融合图像、语音技术, 支持虚拟形象互动, 推动资源数字化, 促进跨学科知识网络的构建。汇雅大模型同时具备查重和学术管理功能, 有效提升学术不端识别及文献管理效率, 为教育与科研场景提供了全面、智能的支持。

2024 年 5 月 15 日猿辅导开发的看云大模型正式通过大模型备案, 之后在旗下的海豚 AI 学、飞象星球、斑马 APP、猿辅导素养课、小猿学练机等产品上进行落地测试。通过嵌入看云大模型, 这些产品在个性化学习、实时反馈和互动体验上有明显提升, 为学生提供了更灵活、适应性更强

的学习支持。看云大模型不仅能够满足学生在不同学科中的知识答疑需求，还可以动态调整反馈内容以适应不同学习阶段，为学生提供个性化、智能化的辅导，提高学习效率。

对于国外来说，多邻国 (Duolingo) 与可汗学院 (Khan Academy) 在 GPT-4 发布之后第一时间就分别推出了 Duolingo Max 平台和 AI 智能工具 Khanmigo，二者都是与 OpenAI 进行合作的成果，在学生个性化、反馈方面提供帮助。大语言模型融入教育领域，会对智能化、数字化教育进行重塑，很大程度上改变了以往的学习方式，使得学习变得更加轻松、高效。

2.2　通用领域教育大模型

ChatGPT 和 GPT-4 作为通用的大语言模型，应用领域非常广泛，涵盖金融、法律、医疗、交通等许多方面，在教育领域中也不断地进行拓展和优化。ChatGPT 在医学问题的回答任务上取得了显著的进展[41]，该模型的回答水平相当于 3 年医学生的及格线。另外研究人员使用 ChatGPT 对美国医学执照考试 (United States medical licensing examination, USMLE) 的性能进行了评估[42]，ChatGPT 的表现达到了 60% 的及格线，ChatGPT 还是在人类没有进行专门输入的情况下取得的这一成绩。Rizzo 等[43] 的研究表明，GPT-4 的水平相当于一个二到三年级的住院医生，而 GPT-3.5 Turbo 则相当于一个一年级住院医生。ChatGPT 在提供答案时通常逻辑性很强，并且回答的答案还具有相关的背景信息。这些研究表明，像 ChatGPT 这样的大语言模型有望在医学教育中帮助学生，来促进医学学习的进步，未来在医学教育使用大语言模型的人数会持续增加。

在高等教育上，对于高校学生来说，ChatGPT 的用途多种多样，包括考试准备、翻译文献和创建需要的代码等。它甚至可以处理更复杂的科学写作，比如总结文献和改写文本，GPT-4 更能识别流程图、对结构图图像进行解释[44]。另外，Dempere 等[45] 提到，ChatGPT 简化了招生流程，提升了学生学习质量，增强了教学效果，为科研提供了更多的支持。因此，将 ChatGPT 整合到高等教育中是必要的。然而，同样还有一些问题有待解决，包括隐私泄露、AI 滥用、偏见问题、模型幻觉、减少人际互动以及可访问性问题等。

在具体学科中，以 ChatGPT 为主要模型，大语言模型在物理学[46-47] 和化学工程教育[48] 中发挥作用。ChatGPT 等大语言模型有能力解决物理学领域中的定量推理任务，并且能模拟不同学生群体对物理概念的回答。同样地，ChatGPT 在化学工程领域也有着不错的应用，并且作为解决实际问题的工具，有助于增强学生的批判性思维和问题解决能力。适当的设计和提示，提高了这些模型在教育工具中的有效性和实用性。

在学生使用方面，学生可以将大语言模型如 ChatGPT 作为撰写论文的辅助工具，它能够收集资料、生成大纲、提供语言表达建议、自动整理引用格式，以及检查内容一致性等。此外，ChatGPT 还能模拟不同观点来丰富论文内容。然而，学生应该批判性地使用 ChatGPT，在教师的指导下适当地使用此类工具，以确保论文的原创性和学术诚信，充分发挥出 ChatGPT 在学术写作中的辅助作用，同时保持独立思考和创造性。

在教师使用方面，教师可以使用 ChatGPT 对学生的学习成果进行评估，根据得到的评估结果来改善学生学习进度。ChatGPT 可以作为教育自动反馈系统 (automated feedback systems, AFS) 为教师提供反馈，与人类教师相比，模型能够生成更详细、更流畅的反馈意见，同时评价与人类教师也高度一致[49]。ChatGPT 还在小学数学课堂评估对话、提供教学策略和建议，经过专家评估，ChatGPT 生成的反馈与教学改善相关，但是缺乏创新性[50]。所以 ChatGPT 作为评估系统，在优化教师教学方法、提升教学质量方面展现出巨大的潜力，能够快速且准确的判断学生水平并提供反馈，但在评价创新性方面还有不小的挑战需要去攻克。

2.3　教育大模型的技术维度比较分析

在教育大模型的快速发展中，模型在多模态融合、个性化学习、情感支持、智能评估和知识支持等技术维度上展现了丰富的应用潜力。本节对主要的教育大模型进行了深入比较分析，并从中提炼出它们的共性特点及各自的独特优势，来指导不同教育需求的模型应用选择。

2.3.1　共性特点

多模态融合　许多教育大模型支持文本、语音、视频等多模态输入形式，以适应多元化的教育需求，如超星汇雅大模型、智适应教育大模型、星火语伴、看云大模型和 ChatGPT。这种多模态特点为课堂教学、课后辅导等场景提供了灵活的交互体验，增强了模型在复杂教育场景中的适应性。

个性化学习支持　这些模型普遍强调个性化学习路径的构建，通过分析学生的学习行为、答疑需求和知识掌握情况，实现动态调整。EduChat、智适应教育大模型、看云大模型和 GPT-4 在个性化学习支持方面表现优异，为学生提供了更精

准的学习辅导。

情感支持　许多教育大模型不仅关注学生的学业需求, 还能为学生提供情感支持, 帮助他们改善学习过程中产生的负面情绪。EduChat 和智适应教育大模型在情绪分析技术的应用上表现出色, GPT-4 的情感支持功能也逐步应用于教育场景, 为学生提供心理支持与疏导, 并提升学习的动力。

智能评估与反馈　大部分模型具备自动评估和即时反馈功能, 如作文批改和口语评测, 它们为学生提供实时的学习改进建议。超星汇雅、星火语伴、EduChat 和 ChatGPT 都具有作文评估、语言纠错等功能, 让学生在学习过程中获得个性化的反馈, 从而提高学习效率。

大规模知识支持　这些模型基于大规模文本数据进行训练, 涵盖广泛的学科知识。超星汇雅依托海量图书和期刊资源, "智海−三乐"采用高质量教育语料库, EduChat 集成了教育书籍和指令库, ChatGPT 在广泛的领域中均有应用, 为教育场景提供了丰富的知识支持。

2.3.2　差异性特点

星火语伴专注于中英文口语评测和实时语法纠错, 特别适合口语考试和日常对话练习。模型在 SuperCLUE 中文通用评测中表现出色, 尤其在中文处理方面具备明显优势。

EduChat 解决了大语言模型在教育中的"知识差距"与"幻觉"问题。通过检索增强技术和专家指导下的微调, EduChat 实现了知识的实时更新, 适用于需要动态知识的教育场景。其启发式教学和情感支持功能进一步提高了学生的认知和心理能力。

"智海−三乐"基于阿里云的通义千问模型, 重点关注计算机课程辅导和学术资源管理, 广泛应用于高校课程。它与多所高校合作, 开发了 24 小时 AI 助教系统, 是高校教育中的创新工具。

"子曰"模型支持自适应学习和知识扩展, 特别适合多学科的辅助教学。该模型通过模拟教师引导, 激发学生进行自主学习, 并能整合多模态知识库, 广泛应用于难题解析、语法精讲等多种场景。

MathGPT 作为数学领域专用的大语言模型, 解题准确率高并且解题步骤清晰。它在多个数学测试中表现优异, 为数学教育提供了精准的解题与知识拓展支持。

超星汇雅大模型结合实际教学需求, 提供了相似度分析、内容安全检测、AI 助教等工具, 与泛雅网络平台深度融合, 帮助构建跨学科知识网络, 是智慧校园建设的有力助手。

看云大模型在猿辅导旗下多款产品中应用, 主要面向个性化学习和实时互动, 通过动态知识答疑和个性化反馈提供全面支持, 适用于 K-12 教育和多学科辅导。

智适应教育大模型关注学生情绪因素和个性化学习, 通过多模态技术提高学习效率, 适合需要高度个性化支持的教育场景。

Duolingo Max 和 Khanmigo 基于 GPT-4, 专注于语言学习和科学教育, 支持个性化互动和实时反馈, 尤其适合国际化语言学习和科学理论学习需求。

ChatGPT 和 GPT-4 作为通用大模型的代表, 在教育领域的应用逐步深化, 适用于英语、科学、数学等学科, 能够提供详细的问答、语言分析和文本生成支持。模型具有广泛的知识储备和灵活的交互功能, 在全球教育场景中具有强大的应用潜力和适应性。

3　如何训练教育大模型

目前以 ChatGPT 为首的大语言模型的构建主要包括 4 个阶段: 数据预处理、无监督预训练、有监督微调和模型评估与测试[3]。同样教育大模型的构建方法类似, 主要是在微调部分进行变动, 如图 2 所示。

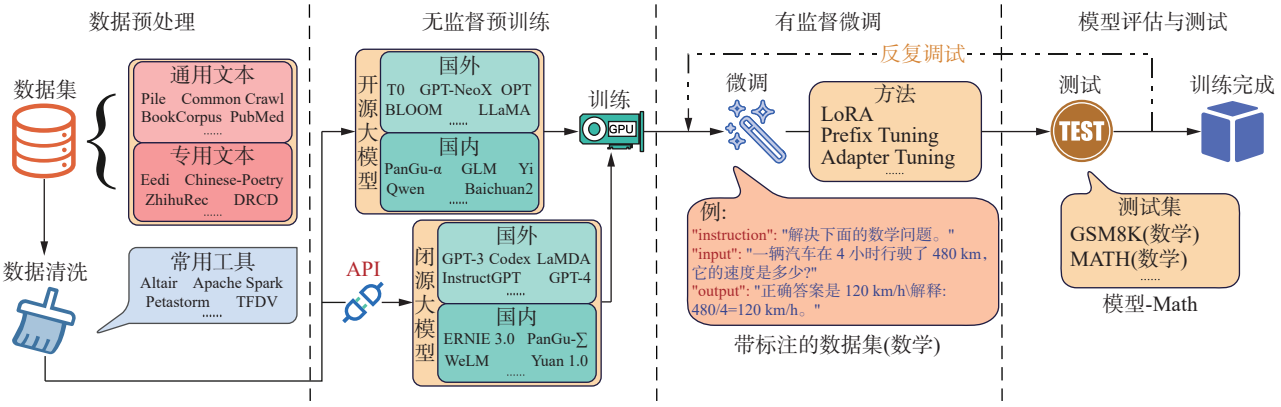

图 2　教育大模型构建流程

Fig. 2　Process of building EduLLMs

3.1　数据预处理

数据预处理是大语言模型的关键步骤，因为模型的性能很大一部分取决于输入数据的质量。首先是寻找大语言模型关于教育类的数据集，这类数据集大多以文本的形式出现，涵盖学习教材、历年的考试题目、发表的论文等，帮助模型提升在个性化教学、问题解答和作业评分等任务中的能力。数据集主要分为 2 种，即通用文本和专用文本。通用文本主要作用是为模型提供广泛的知识基础和增强模型的泛化能力。专用文本则更多地关注特定的领域或主题，提高特定领域的专业性。当收集完数据集之后，需要对数据进行预处理。因为教育数据来源广泛，质量参差不齐，有些还包含学生的个人信息和学习记录，容易引发隐私与安全问题，这些都会影响教育大模型训练的性能，所以拥有一个相对干净的数据集是非常重要的。因此，对数据进行预处理可以提升数据质量，使模型发挥出更准确可靠的性能。可以使用他人已经预处理好的数据集，也可以自己使用相关预处理工具来对数据进行处理。表 2 简要列举了训练教育大模型常见的数据集。

表 2　教育大模型数据集

Table 2　EduLLMs datasets

数据来源	名称	简要介绍	链接
通用文本	Pile[51]	Pile是一个多样化的开源语言建模数据集，由22个较小且高质量的数据集组合在一起	https://pile.eleuther.ai
	OpenWebText2	OpenWebText2是一个大型的过滤数据集，大约1 710万份文档和65.86 GB的未压缩文本	https://openwebtext2.readthedocs.io/en/latest
	Common Crawl[52]	Common Crawl是一个免费且开放的Web爬虫数据存储库，超过2 500亿页。自2007年免费开放语料库以来，其以在超过1万篇研究论文中被引用	https://commoncrawl.org
	ROOTS	ROOTS是一个跨越59种语言(46种自然语言和13种编程语言)的1.6 TB数据集，用于训练拥有1 760亿参数的BLOOM(big science large open-science open-access multilingual language model)模型	https://huggingface.co/bigscience-data
	Wikipedia	Wikipedia数据集是基于维基百科的数据转储构建的，每种语言分为一个子集。每个样本包含一篇完整的维基百科文章内容	https://en.wikipedia.org/wiki/Wikipedia:Database_download
	BookCorpus[53]	BookCorpus数据集是由大量免费小说书籍构成，其中包含16种不同子流派的1万多本书	https://huggingface.co/datasets/bookcorpus
	arXiv	arXiv是一个包含170万篇arXiv文章的数据集，适用于趋势分析、论文推荐引擎、分类预测、知识图谱构建和语义搜索界面等应用	https://www.kaggle.com/datasets/Cornell-University/arxiv
专用文本	PubMed	PubMed数据集主要来自生物医学和健康领域，以及相关的生命科学、行为科学、化学科学等学科	https://pubmed.ncbi.nlm.nih.gov
	SQuAD2.0[54]	斯坦福问答数据集(SQuAD)是一个阅读理解数据集，由维基百科文章上提出的问题组成，其中答案是来自相应阅读段落的文本	https://rajpurkar.github.io/SQuAD-explorer
	SNLI[55]	斯坦福自然语言推理数据集(SNLI)是由57万条人工编写的英语句子构成，手动标记矛盾和中性标签	https://nlp.stanford.edu/projects/snli
	Eedi	Eedi数据集包含两个学年学生的答案，这些数据的答案记录总数超过1 700万，使其成为目前为止最大的教育数据集之一	https://eedi.com/projects/neurips-education-challenge
	ZhihuRec[56]	ZhihuRec数据集来源于知乎平台，由10天内收集的约1亿条交互信息组成，包含问题、答案、话题及用户查询日志	https://github.com/THUIR/ZhihuRec-Dataset
	CCPM[57]	中国古典诗歌匹配数据集(CCPM)，每条数据包含诗歌对应的描述、4个候选诗句和正确诗句的答案编号	https://github.com/THUNLP-AIPoet/CCPM
	ChID[58]	中文成语数据集(ChID)是一个用于完形填空测试的大规模中文成语数据集。ChID包含58万个段落，涵盖多个领域	https://github.com/chujiezheng/ChID-Dataset
	DRCD[59]	Delta阅读理解数据集(DRCD)是一个繁体中文机器阅读理解数据集。该数据集包含来自2 108篇维基百科文章的1万个段落和3万多个问题	https://github.com/DRCKnowledgeTeam/DRCD
	Chinese-Poetry	中华古典文集数据集(Chinese-Poetry)包含5.5万首唐诗、26万首宋诗和其他古典文集	https://github.com/chinese-poetry/chinese-poetry
	APPS[60]	APPS数据集包含1万道题目，这些题目既包含了基础的单行代码练习，也涵盖了复杂的算法挑战	https://github.com/hendrycks/apps

3.2　无监督预训练

当数据集完成预处理之后, 就可以开始进行预训练。大语言模型当前正处于快速发展阶段, 由于模型非常复杂, 个人很难去进行大语言模型的开发, 通常都是由团队进行。所以个人想要去训练教育大模型, 一般会使用开源的模型, 这类模型在通用文本上已经训练相当长一段时间, 对通用知识掌握全面, 后续只需要在此基础上使用专用文本进行训练, 来保证模型对教育领域的指向性。表 3 列举了一些开源项目。

表 3　开源大语言模型

Table 3　Open source large language models

地区	模型	发布时间	大小/B
国外	T0[61]	2021年10月	11
	GPT-NeoX[62]	2022年4月	20
	OPT[63]	2022年5月	175
	BLOOM[64]	2022年11月	176
	LLaMA[31]	2023年2月	7/13/33/65
	LLaMA 2[65]	2023年7月	7/13/70
	Gemma[66]	2024年2月	2/7
	LLaMA 3[67]	2024年4月	8/70
	Gemma 2[68]	2024年6月	9/27
国内	PanGu-α[69]	2021年4月	13
	GLM[70]	2022年10月	130
	Qwen[36]	2023年9月	7/14/72
	Baichuan 2[71]	2023年9月	7/13
	Yi[72]	2023年11月	6/9/34
	YUAN 2.0[73]	2023年11月	2/51/102
	Qwen1.5[74]	2024年2月	0.5/1.8/4/7/14/72
	Qwen2[75]	2024年6月	2/7/72

3.3　有监督微调

当预训练完成之后, 模型掌握了丰富的语言知识、语言识别能力和上下文处理能力, 从而具备了对语言的泛化理解。然而, 为了使模型能够在教育领域中获得更好的性能, 能够更好更精准地回答教育领域的问题, 需要对模型使用标注数据 (即每个样本都有正确答案或标签的数据) 来优化模型在特定任务上的表现。这个过程调整并优化了模型的权重, 使模型能够更好地处理与训练数据。但是由于大语言模型的参数量巨大, 如果在微调过程中对所有参数进行更新, 会导致巨大的存储和计算成本。所以使用高效的微调方法, 改变模型中相对较少的参数, 保留大部分从预训练学到的知识, 同时也能获得良好的性能, 这是微调阶段常见的手段之一。表 4 列举了常见的微调方法。

表 4　高效微调方法

Table 4　Efficient fine-tuning methods

方法	特点	时间
全参数微调	在目标任务上对所有模型参数进行更新	2012年
冻结部分参数	只更新部分层的参数, 减少计算成本	2015年
Adapter Tuning[76]	插入小的神经网络模块, 仅微调部分参数	2019年2月
Prefix Tuning[77]	添加前缀到输入序列, 实现任务调整	2021年1月
P-Tuning[78]	使用长短期记忆网络生成虚拟标签嵌入	2021年3月
Prompt Tuning[79]	对输入前缀进行微调, 提高任务适应性	2021年4月
BitFit[80]	只微调模型的偏置项, 减少参数量	2021年6月
LoRA[81]	低秩矩阵到原权重矩阵, 减少计算复杂度	2021年6月
P-Tuning v2[82]	适应复杂的自然语言理解任务, 更加高效	2021年10月
AdaLoRA[83]	自适应调整低秩矩阵, 提高模型的微调效率	2023年3月
QLoRA[84]	通过量化将模型调优到4 bit, 降低显存占用	2023年5月
DoRA[85]	动态调整微调参数, 提高模型的适应性	2024年2月

训练选用的数据集应该是少量、高质量且包含正确标注的, 同时需要根据自身想要实现的教育功能去寻找或者自己创建数据集。比如在 EduChat 中, 研发人员想要模型提供更适合中文环境的情感支持, 他们通过将广泛使用的英文情感支持数据集 ESConv(emotion support conversation dataset)[86] 翻译成中文并进行人工审查和清理, 创建了 ESConv-zh 数据集。此外, 他们为了提高学生写作技能, 创建了一个作文评估数据集, 其中包括由 ChatGPT 评估的论文和教学专家手工整理的评论, 目的是为学生提供及时且细致的反馈。

3.4　模型评估与测试

构建教育大模型的最后一个步骤是大语言模型的评估和测试, 它也是整个模型开发过程中关键步骤, 目的是确保模型的性能满足预定的标准和应用需求。这不仅有助于提高模型的性能, 还确保模型的使用符合伦理和合规性要求。使用教育类测试集来进行测试和评估, 例如数学类的 GSM8K(grade school math) 和 MATH, 利用测试结果寻找大语言模型的缺陷, 评估的结果可以指导后续的模型调整和改进。表 5 列举了常见的教育测试数据集。

表 5　教育测试数据集
Table 5　Educational testing datasets

类型	名称	特点	链接
英文	MMLU[87]	MMLU是一个包含了57个子任务的英文测试数据集，难度覆盖高中水平到专家水平	https://huggingface.co/datasets/cais/mmlu
	WinoGrande[88]	WinoGrande收集了4.4万个问题，提高了数据集特定偏差的规模和鲁棒性	https://github.com/allenai/winogrande
中文	C-Eval[40]	C-Eval包含1.3万个多项选择题，涵盖了52个学科和4个难度级别	https://cevalbenchmark.com
	AGIEval[89]	AGIEval是一个用于评估基础模型在标准化考试中表现的数据集，涵盖了高考、公务员考试和数学竞赛	https://github.com/ruixiangcui/AGIEval
	CMMLU[90]	CMMLU是一个包含了67个主题的中文测试数据集，涉及自然科学、社会科学、人文以及常识等	https://github.com/haonan-li/CMMLU
	SuperCLUE[37]	SuperCLUE是一个综合性大模型测试数据集，主要关注语言理解与生成、专业技能与知识、Agent智能体和安全性等12项基础能力	https://github.com/CLUEbenchmark/SuperCLUE
	GAOKAO-Bench[91]	GAOKAO-Bench包含由中国高考题目组成的数据集，收集了2010—2022年全国高考卷的题目，其中包括1 781道客观题和1 030道主观题	https://github.com/OpenLMLab/GAOKAO-Bench
多语言	M3Exam[92]	M3Exam包含1.2万个问题，涵盖了从高资源语种到低资源语种共9种语言	https://github.com/DAMO-NLP-SG/M3Exam
数学	GSM8K[93]	GSM8K是一个包含8 000道小学数学问题及答案的数据集，用于评估大语言模型解决基础数学推理问题的能力	https://huggingface.co/datasets/openai/gsm8k
	MATH[94]	MATH数据集是一个评估模型数学推理能力的数据集，包含高中数学竞赛水平问题，涵盖代数、几何、数论等多个领域	https://github.com/hendrycks/math
通用	XTREME[95]	XTREME涵盖了40种类型的语言，并包括9项任务，这些任务需要对不同层次的语法或语义进行推理	https://sites.research.google/xtreme
	GLUE[96]	GLUE涵盖了例如情感分析、问答配对、文本蕴含和语义相似性判断等任务	https://gluebenchmark.com

教育大模型的训练是一个复杂的过程。首先，需要收集大量的文本数据，并进行大量的预处理，包括清洗、标记化以及其他必要的格式调整。随后进行预训练，这一阶段模型学习语言的基本特征和结构。之后，模型使用特定教育任务的数据集进行微调，来适应教育相关的应用场景。在整个过程中，模型需要在不同的数据集上反复评估和调整。最终，经过训练和优化的模型能够理解和生成语言，以解决各种复杂的自然语言处理任务。

4　教育大模型的机会与挑战

4.1　局限性

教育大模型，如 ChatGPT、EduChat 和其他类似的模型，在教育领域应用非常广泛，但同时也普遍存在一些局限性。

理解与推理的局限性[97]　虽然教育大模型在语言处理和生成上表现出色，但是其理解能力依然是基于统计模式，而非真正的语义理解。这意味着在处理复杂逻辑推理或创造性思维任务时，模型的表现往往不够理想。例如，GPT-4 虽然在语言流畅性和回答准确性上有所提升，但面对多步骤推理问题 (如高阶数学证明或科学研究设计) 时，模型会生成表面上合理但推理不完整的答案。在解答涉及多步逻辑的几何问题时，模型会跳过关键步骤，导致回答不准确。

知识更新问题[98]　教育大模型的训练数据通常截止于特定时间，因此无法获取训练结束时间点以后的最新知识。例如，学生若向教育大模型咨询关于最新的量子计算技术或新发布的教育政策，模型可能会提供过时的解答。例如 ChatGPT，当学生提出与最新研究相关的问题时，模型的回答会遗漏近几个月的发现，导致学生无法获取最新且准确的学术信息。

不准确信息[99]　教育大模型在生成内容时有时会出现“幻觉”现象，生成看似合理却不正确的内容。在缺乏人类监督的情况下，模型可能会输出误导性甚至完全错误的信息。例如，在历史学或医学问题中，模型会生成并不存在的事件或医学事实，学生可能因为这些误导信息而对所学知识产生误解。这类不准确信息尤其在专业领域 (如法律或科学研究) 中会带来较大风险。

内容偏见问题[100-101]　教育大模型的偏见问题也是一个重要挑战，模型可能无意中反映训练

数据中的性别、种族、文化或意识形态偏见。例如, 在批改学生作文或解答文化相关问题时, 模型可能会优先推荐某一特定文化内容, 导致学生接触到片面的观点。此外, 模型在缺乏透明度的情况下难以识别和纠正这些偏见。比如在语言学习中, 模型会偏向推荐某一种语言或文化内容, 这对教育的公平性提出了挑战。

数据隐私和安全[102-103]　教育大模型通常需要处理大量的敏感信息, 包括学生的个人信息、学习记录和行为数据, 这在安全方面存在风险。如果这些信息未被妥善管理和保护, 可能导致学生隐私泄露或数据被不当使用。例如, 在在线课堂和个性化学习推荐中, 如果学生的行为数据和成绩被第三方访问或滥用, 可能会对学生造成负面影响。此外, 教育数据的透明性和安全性也对教育机构提出了更高的管理要求。

可接入性和资源限制[104]　教育大模型的应用往往需要高计算资源、存储空间和能耗, 这对资源有限的教育机构来说是一大挑战。模型的实时运行需要高配置的硬件和持续的技术维护, 对于小型教育机构来说, 获得和维持这种技术条件存在困难。此外, 操作和维护这些高级模型需要专业技术人员, 这也使得一些教育机构难以推广和普及这些技术。

4.2　展望

教育大模型预计将会迎来一系列革命性的变化, 将共同推动和创建一个更加智能、能够互动和个性化的学习环境。随着自然语言处理能力的进一步提升, 这些模型将能更深入地理解学生的需求, 提供更准确的反馈和指导, 使学习体验贴合个人的学习风格和节奏。

未来, 多模态技术将会普遍地运用到教育大模型中。多模态方法不仅能处理文本信息, 还能分析图片、声音、视频等不同形式的数据, 从而为学生提供一个更为丰富和综合的学习环境。通过整合视觉和听觉材料, 多模态学习可以增加学生的参与度和兴趣。例如, 互动式视频和虚拟现实可以使学生沉浸在模拟的某个历史场景或科学实验中, 提供身临其境的学习体验。这种学习方式有望在教育领域发挥巨大的作用, 为学生的学习和发展带来更多的可能性。

在隐私安全方面, 随着对数据保护意识的增强, 未来的教育大模型将采用更先进的技术来确保学生信息的安全。包括使用端到端加密、联邦学习[105] 等方法来处理敏感数据, 确保学生的隐私不受侵犯, 同时还保证了教育内容的质量和可靠性。

可访问性和资源效率也是未来发展的重点。通过优化模型结构和计算过程, 以及利用云计算和边缘计算等技术, 教育大模型将能够为更多的学生提供服务, 这不仅能够提升教育资源的覆盖面, 也降低了教育的整体成本。

最重要的是, 大模型将会更加强调与教师的协作。通过辅助教师进行日常的教学和评估工作, 教师能够更专注于课程内容的创新和教学方法的改进, 从而提高教学的整体质量。同时, 通过技术的公平性, 大模型将在消除教育不平等、促进知识的普及和提高社会整体的教育水平方面起到关键作用。

5　结束语

本篇文章将大语言模型在教育领域的应用作为研究核心, 首先对大语言模型、现代教育和教育大模型的特点进行总结, 找出教育与大模型融合的可取之处。文章强调了教育大模型的应用场景, 介绍了目前教育大模型的共性以及差异性特点。文章也详细介绍了如何训练出符合教育需求的大语言模型, 这是实现有效应用的关键环节。在详细分析训练完备的教育大模型的基础上, 文章进一步探讨了这些模型在实际应用中可能遇到的局限性, 并说明了未来教育领域可能出现的新型大语言模型及其潜在的发展趋势, 特别是在个性化学习、多模态交互和数据安全性方面。在教育大模型的支持下, 这种学习方式有望在教育领域发挥更大的作用, 为学生的学习和发展带来更多的可能性。通过这些分析, 文章旨在为教育领域的专业人士提供对于当前技术现状的深刻见解, 并展望未来技术发展可能带来的新机遇。

参考文献:

[1]　张岩. “互联网+教育” 理念及模式探析[J]. 中国高教研究, 2016(2): 70−73.
ZHANG Yan. On the concept and mode of “Internet plus education”[J]. China higher education research, 2016(2): 70−73.

[2]　马世龙, 乌尼日其其格, 李小平. 大数据与深度学习综述[J]. 智能系统学报, 2016, 11(6): 728−742.
MA Shilong, WUNIRI Q Q G, LI Xiaoping. Deep learning with big data: state of the art and development[J]. CAAI transactions on intelligent systems, 2016, 11(6): 728−742.

[3]　ZHAO W X, ZHOU Kun, LI Junyi, et al. A survey of large language models[EB/OL]. (2023−11−24)[2024−06−20]. https://arxiv.org/abs/2303.18223.

[4] 余胜泉, 熊莎莎. 基于大模型增强的通用人工智能教师架构[J]. 开放教育研究, 2024, 30(1): 33−43.
YU Shengquan, XIONG Shasha. General artificial intelligence teacher architecture based on enhanced pre-trained large models[J]. Open education research, 2024, 30(1): 33−43.

[5] 曹培杰. 智慧教育: 人工智能时代的教育变革[J]. 教育研究, 2018, 39(8): 121−128.
CAO Peijie. Smart education: the educational reform at the age of artificial intelligence[J]. Educational research, 2018, 39(8): 121−128.

[6] KASNECI E, SESSLER K, KÜCHEMANN S, et al. ChatGPT for good? On opportunities and challenges of large language models for education[J]. Learning and individual differences, 2023, 103: 102274.

[7] 曹培杰, 谢阳斌, 武卉紫, 等. 教育大模型的发展现状、创新架构及应用展望[J]. 现代教育技术, 2024, 34(2): 5−12.
CAO Peijie, XIE Yangbin, WU Huizi, et al. The development status, innovation architecture and application prospects of educational big models[J]. Modern educational technology, 2024, 34(2): 5−12.

[8] GAN Wensheng, QI Zhenlian, WU Jiayang, et al. Large language models in education: vision and opportunities[C]// 2023 IEEE International Conference on Big Data. Sorrento: IEEE, 2023: 4776−4785.

[9] 祝智庭, 卢琳萌, 王馨怡, 等. 智慧教育理论与实践在中国的发展: 十年回顾与近未来展望[J]. 中国远程教育, 2023(12): 21−33.
ZHU Zhiting, LU Linmeng, WANG Xinyi, et al. The development of smart education theory and practice in China: a ten-year review and near-future prospects[J]. Chinese journal of distance education, 2023(12): 21−33.

[10] RUDOLPH J, TAN S, TAN S. ChatGPT: bullshit spewer or the end of traditional assessments in higher education?[J]. Journal of applied learning and teaching, 2023, 6(1): 342−363.

[11] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[J]. Advances in neural information processing systems, 2017, 30: 5998−6008.

[12] RADFORD A, WU J, CHILD R, et al. Language models are unsupervised multitask learners[J]. OpenAI blog, 2019, 1(8): 9.

[13] RAFFEL C, SHAZEER N, ROBERTS A, et al. Exploring the limits of transfer learning with a unified text-to-text Transformer[J]. Journal of machine learning research, 2020, 21(1): 5485−5551.

[14] BROWN T B, MANN B, RYDER N, et al. Language models are few-shot learners[C]//Proceedings of the 34th International Conference on Neural Information Processing Systems. Vancouver: ACM, 2020: 1877−1901.

[15] REN Xiaozhe, ZHOU Pingyi, MENG Xinfan, et al. Pan-Gu-Σ: towards trillion parameter language model with sparse heterogeneous computing[EB/OL]. (2023−05−20)[2024−06−20]. https://arxiv.org/abs/2303.10845.

[16] OpenAI. GPT-4[EB/OL]. (2024−03−08)[2024−06−20]. https://openai.com/gpt-4.

[17] BOMMASANI R, HUDSON D A, ADELI E, et al. On the opportunities and risks of foundation models[EB/OL]. (2022−07−12)[2024−06−20]. https://arxiv.org/abs/2108.07258.

[18] WANG Shen, XU Tianlong, LI Hang, et al. Large language models for education: a survey and outlook[EB/OL]. (2024−04−01)[2024−06−20]. https://arxiv.org/abs/2403.18105.

[19] GAO Tianyu, FISCH A, CHEN Danqi. Making pre-trained language models better few-shot learners[C]//Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing. [S.l.]: ACL, 2021: 3816−3830.

[20] SCHICK T, SCHÜTZE H. Exploiting cloze questions for few shot text classification and natural language inference[EB/OL]. (2021−01−25)[2024−06−20]. https://arxiv.org/abs/2001.07676.

[21] WEI J, WANG Xuezhi, SCHUURMANS D, et al. Chain-of-thought prompting elicits reasoning in large language models[C]//Proceedings of the 36th International Conference on Neural Information Processing Systems. New Orleans: ACM, 2022: 24824−24837.

[22] 唐雯谦, 覃成海, 向艳, 等. 智慧教育与个性化学习理论与实践研究[J]. 中国电化教育, 2021(5): 124−137.
TANG Wenqian, QIN Chenghai, XIANG Yan, et al. Research on theory and practice of intelligent education and personalized learning[J]. China educational technology, 2021(5): 124−137.

[23] ZUBIRI-ESNAOLA H, VIDU A, RIOS-GONZALEZ O, et al. Inclusivity, participation and collaboration: learning in interactive groups[J]. Educational research, 2020, 62(2): 162−180.

[24] 安涛, 赵可云. 大数据时代的教育技术发展取向[J]. 现代教育技术, 2016, 26(2): 27−32.
AN Tao, ZHAO Keyun. The developmental orientation of educational technology in the big data era[J]. Modern educational technology, 2016, 26(2): 27−32.

[25] RAWAS S. ChatGPT: Empowering lifelong learning in the digital age of higher education[J]. Education and in-

formation technologies, 2024, 29(6): 6895−6908.

[26] 刘凤娟, 赵蔚, 姜强, 等. 基于知识图谱的个性化学习模型与支持机制研究[J]. 中国电化教育, 2022(5): 75−81,90.
LIU Fengjuan, ZHAO Wei, JIANG Qiang, et al. Research on personalized learning model and support mechanism based on knowledge graph[J]. China educational technology, 2022(5): 75−81, 90.

[27] LEE U, JUNG H, JEON Y, et al. Few-shot is enough: exploring ChatGPT prompt engineering method for automatic question generation in English education[J]. Education and information technologies, 2024, 29(9): 11483−11515.

[28] HU Wenbo, XU Yifan, LI Yi, et al. BLIVA: a simple multimodal LLM for better handling of text-rich visual questions[J]. Proceedings of the AAAI conference on artificial intelligence, 2024, 38(3): 2256−2264.

[29] ZHAN Peida, HE Keren. A longitudinal diagnostic model with hierarchical learning trajectories[J]. Educational measurement: issues and practice, 2021, 40(3): 18−30.

[30] CHOWDHERY A, NARANG S, DEVLIN J, et al. PaLM: scaling language modeling with pathways[J]. Journal of machine learning research, 2023, 24(240): 1−113.

[31] TOUVRON H, LAVRIL T, IZACARD G, et al. LLaMA: open and efficient foundation language models[EB/OL]. (2023−02−27)[2024−06−20]. https://arxiv.org/abs/2302.13971.

[32] OpenAI. Introducing ChatGPT[EB/OL]. (2024−03−11)[2024−06−20]. https://openai.com/blog/chatgpt.

[33] 吴娟, 周建蓉, 卢仪珂, 等. 基于复杂学习设计的在线写作模型构建与应用[J]. 电化教育研究, 2024, 45(1): 108−113,121.
WU Juan, ZHOU Jianrong, LU Yike, et al. Construction and application of an online writing model based on complex learning design[J]. E-education research, 2024, 45(1): 108−113,121.

[34] 王洪鑫, 闫志明, 陈效玉, 等. 面向 MOOC 课程评论的主题挖掘与情感分析研究[J]. 开放学习研究, 2021, 26(4): 16−23.
WANG Hongxin, YAN Zhiming, CHEN Xiaoyu, et al. Research on topic mining and emotion analysis for MOOCs course review[J]. Journal of open learning, 2021, 26(4): 16−23.

[35] DAN Yuhao, LEI Zhikai, GU Yiyang, et al. EduChat: a large-scale language model-based chatbot system for intelligent education[EB/OL]. (2023−08−05)[2024−06−20]. https://arxiv.org/abs/2308.02773.

[36] BAI Jinze, BAI Shuai, CHU Yunfei, et al. Qwen technical report[EB/OL]. (2023−09−28)[2024−06−20]. https://arxiv.org/abs/2309.16609.

[37] XU Liang, LI Anqi, ZHU Lei, et al. SuperCLUE: a comprehensive Chinese large language model benchmark [EB/OL]. (2023−07−27)[2024−06−20]. https://arxiv.org/abs/2307.15020.

[38] 刘莉, 刘铁芳. 重审苏格拉底的 “产婆术”[J]. 全球教育展望, 2021, 50(9): 46−62.
LIU Li, LIU Tiefang. A reexamination of Socrates' “midwifery”[J]. Global education, 2021, 50(9): 46−62.

[39] 陈静远, 吴韬, 吴飞. 课程、教材、平台三位一体的“人工智能引论”育人基座能力建设[J]. 计算机教育, 2023(11): 34−37.
CHEN Jingyuan, WU Tao, WU Fei. Construction of the educational foundation for "Introduction to Artificial Intelligence" with integration of courses, textbooks, and platforms[J]. Computer education, 2023(11): 34−37.

[40] HUANG Yuzhen, BAI Yuzhuo, ZHU Zhihao, et al. C-Eval: a multi-level multi-discipline Chinese evaluation suite for foundation models[J]. Advances in neural information processing systems, 2023, 36: 62991−63010.

[41] GILSON A, SAFRANEK C W, HUANG T, et al. How does ChatGPT perform on the United States medical licensing examination (USMLE)? the implications of large language models for medical education and knowledge assessment[J]. JMIR medical education, 2023, 9: e45312.

[42] KUNG T H, CHEATHAM M, MEDENILLA A, et al. Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models [J]. PLoS digital health, 2023, 2(2): e0000198.

[43] RIZZO M G, CAI N, CONSTANTINESCU D. The performance of ChatGPT on orthopaedic in-service training exams: a comparative study of the GPT-3.5 turbo and GPT-4 models in orthopaedic education[J]. Journal of orthopaedics, 2024, 50: 70−75.

[44] NEUMANN M, RAUSCHENBERGER M, SCHÖN E M. “We need to talk about ChatGPT”: the future of AI and higher education[C]//2023 IEEE/ACM 5th International Workshop on Software Engineering Education for the Next Generation (SEENG). Melbourne: IEEE, 2023: 29−32.

[45] DEMPERE J, MODUGU K, HESHAM A, et al. The impact of ChatGPT on higher education[J]. Frontiers in education, 2023, 8: 1206936.

[46] POLVERINI G, GREGORCIC B. How understanding large language models can inform the use of ChatGPT in physics education[J]. European journal of physics, 2024, 45(2): 025701.

[47] KIESER F, WULFF P, KUHN J, et al. Educational data augmentation in physics education research using ChatGPT[J]. Physical review physics education research, 2023, 19(2): 020150.

[48] TSAI M L, ONG C W, CHEN Chengliang. Exploring the use of large language models (LLMs) in chemical engineering education: building core course problem models with Chat-GPT[J]. Education for chemical engineers, 2023, 44: 71−95.

[49] DAI Wei, LIN Jionghao, JIN Hua, et al. Can large language models provide feedback to students? A case study on ChatGPT[C]//2023 IEEE International Conference on Advanced Learning Technologies. Orem: IEEE, 2023: 323−325.

[50] WANG R E, DEMSZKY D. Is ChatGPT a good teacher coach? Measuring zero-shot performance for scoring and providing actionable insights on classroom instruction [EB/OL]. (2023−06−05)[2024−06−20]. https://arxiv.org/abs/2306.03090.

[51] GAO L, BIDERMAN S, BLACK S, et al. The Pile: an 800GB dataset of diverse text for language modeling [EB/OL]. (2020−12−31)[2024−06−20]. https://arxiv.org/abs/2101.00027.

[52] PATEL J M. Introduction to common crawl datasets [M]//Getting Structured Data from the Internet. Berkeley: Apress, 2020: 277−324.

[53] BANDY J, VINCENT N. Addressing “documentation debt” in machine learning research: a retrospective datasheet for BookCorpus[EB/OL]. (2021−05−11)[2024−06−20]. https://arxiv.org/abs/2105.05241.

[54] RAJPURKAR P, JIA R, LIANG P. Know what you don’t know: unanswerable questions for SQuAD[EB/OL]. (2018−06−11)[2024−06−20]. https://arxiv.org/abs/1806.03822.

[55] BOWMAN S R, ANGELI G, POTTS C, et al. A large annotated corpus for learning natural language inference[EB/OL]. (2015−08−21)[2024−06−20]. https://arxiv.org/abs/1508.05326.

[56] HAO Bin, ZHANG Min, MA Weizhi, et al. A large-scale rich context query and recommendation dataset in online knowledge-sharing[EB/OL]. (2021−06−11)[2024−06−20]. https://arxiv.org/abs/2106.06467.

[57] LI Wenhao, QI Fanchao, SUN Maosong, et al. CCPM: a Chinese classical poetry matching dataset[EB/OL]. (2021−06−03)[2024−06−20]. https://arxiv.org/abs/2106.01979.

[58] ZHENG Chujie, HUANG Minlie, SUN Aixin. ChID: a large-scale Chinese IDiom dataset for cloze test[C]//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Florence: ACL, 2019: 778−787.

[59] SHAO C C, LIU T, LAI Yuting, et al. DRCD: a Chinese machine reading comprehension dataset[EB/OL]. (2019−05−29)[2024−06−20]. https://arxiv.org/abs/1806.00920.

[60] HENDRYCKS D, BASART S, KADAVATH S, et al. Measuring coding challenge competence with APPS [EB/OL]. (2021−11−08)[2024−06−20]. https://arxiv.org/abs/2105.09938.

[61] SANH V, WEBSON A, RAFFEL C, et al. Multitask prompted training enables zero-shot task generalization [EB/OL]. (2021−10−15)[2024−06−20]. https://arxiv.org/abs/2110.08207.

[62] BLACK S, BIDERMAN S, HALLAHAN E, et al. GPT-NeoX-20B: an open-source autoregressive language model[EB/OL]. (2022−04−14)[2024−06−20]. https://arxiv.org/abs/2204.06745.

[63] ZHANG Susan, ROLLER S, GOYAL N, et al. OPT: open pre-trained transformer language models[EB/OL]. (2022−06−21) [2024−06−20]. https://arxiv.org/abs/2205.01068.

[64] LE SCAO T, FAN A, AKIKI C, et al. BLOOM: a 176B-parameter open-access multilingual language model [EB/OL]. (2022−11−09) [2024−06−20]. https://arxiv.org/abs/2211.05100v4.

[65] TOUVRON H, MARTIN L, STONE K, et al. LLaMA 2: open foundation and fine-tuned chat models[EB/OL]. (2023−07−19) [2024−06−20]. https://arxiv.org/abs/2307.09288.

[66] TEAM G, MESNARD T, HARDIN C, et al. Gemma: open models based on Gemini research and technology [EB/OL]. (2024−04−16) [2024−06−20]. https://arxiv.org/abs/2403.08295.

[67] DUBEY A, JAUHRI A, PANDEY A, et al. The LLaMA 3 herd of models[EB/OL]. (2024−08−15) [2024−11−11]. https://arxiv.org/abs/2407.21783.

[68] TEAM G, RIVIERE M, PATHAK S, et al. Gemma 2: improving open language models at a practical size [EB/OL]. (2024−10−02) [2024−11−11]. https://arxiv.org/abs/2408.00118.

[69] ZENG Wei, REN Xiaozhe, SU Teng, et al. PanGu-α: large-scale autoregressive pretrained Chinese language models with auto-parallel computation[EB/OL]. (2021−04−26)[2024−06−20]. https://arxiv.org/abs/2104.12369.

[70] ZENG Aohan, XU Bin, WANG Bowen, et al. ChatGLM: a family of large language models from GLM-130B to GLM-4 all tools[EB/OL]. (2024−07−30)[2024−11−11]. https://arxiv.org/abs/2406.12793.

[71] YANG Aiyuan, XIAO Bin, WANG Bingning, et al. Baichuan 2: open large-scale language models[EB/OL]. (2023−09−20)[2024−06−20]. https://arxiv.org/abs/2309.10305.

[72] YOUNG A, CHEN Bei, LI Chao, et al. Yi: open foundation models by 01.AI[EB/OL]. (2024−03−07)[2024−06−20]. https://arxiv.org/abs/2403.04652.

[73] WU Shaohua, ZHAO Xudong, WANG Shenling, et al. YUAN 2.0: a large language model with localized filtering-based attention[EB/OL]. (2023−12−18)[2024−06−20]. https://arxiv.org/abs/2311.15786.

[74] Qwen Team. Introducing Qwen1.5[EB/OL]. (2024−02−04)[2024−06−20]. https://qwenlm.github.io/blog/qwen1.5/.

[75] YANG An, YANG Baosong, HUI Binyuan, et al. Qwen2 technical report[EB/OL]. (2024−09−10)[2024−11−11]. https://arxiv.org/abs/2407.10671.

[76] HOULSBY N, GIURGIU A, JASTRZEBSKI S, et al. Parameter-efficient transfer learning for NLP[C]//International conference on machine learning. Los Angeles: PMLR, 2019: 2790−2799.

[77] LI X L, LIANG P. Prefix-tuning: optimizing continuous prompts for generation[EB/OL]. (2021−01−01)[2024−06−20]. https://arxiv.org/abs/2101.00190.

[78] LIU Xiao, ZHENG Yanan, DU Zhengxiao, et al. GPT understands, too[J]. AI open, 2024, 5: 208−215.

[79] LESTER B, AL-RFOU R, CONSTANT N, et al. The power of scale for parameter-efficient prompt tuning [EB/OL]. (2021−09−02) [2024−06−20]. https://arxiv.org/abs/2104.08691.

[80] BEN ZAKEN E, RAVFOGEL S, GOLDBERG Y. BitFit: simple parameter-efficient fine-tuning for Transformer-based masked language-models[EB/OL]. (2022−09−05) [2024−06−20]. https://arxiv.org/abs/2106.10199.

[81] HU J E, SHEN Yelong, WALLIS P, et al. LoRA: low-rank adaptation of large language models[EB/OL]. (2021−10−16) [2024−06−20]. https://arxiv.org/abs/2106.09685.

[82] LIU Xiao, JI Kaixuan, FU Yicheng, et al. P-Tuning v2: prompt tuning can be comparable to fine-tuning universally across scales and tasks[EB/OL]. (2021−10−18) [2024−06−20]. https://arxiv.org/abs/2110.07602.

[83] ZHANG Qingru, CHEN Minshuo, BUKHARIN A, et al. AdaLoRA: adaptive budget allocation for parameter-efficient fine-tuning[EB/OL]. (2024−07−09) [2024−11−11]. https://arxiv.org/abs/2303.10512.

[84] DETTMERS T, PAGNONI A, HOLTZMAN A, et al. QLORA: efficient finetuning of quantized LLMs[C]//Proceedings of the 37th International Conference on Neural Information Processing Systems. New Orleans: Curran Associates Inc., 2023: 10088−10115.

[85] LIU S Y, WANG C Y, YIN Hongxu, et al. DoRA: weight-decomposed low-rank adaptation[EB/OL]. (2024−07−09) [2024−11−11]. https://arxiv.org/abs/2402.09353.

[86] LIU Siyang, ZHENG Chujie, DEMASI O, et al. Towards emotional support dialog systems[EB/OL]. (2021−06−02)[2024−06−20]. https://arxiv.org/abs/2106.01144.

[87] HENDRYCKS D, BURNS C, BASART S, et al. Measuring massive multitask language understanding[EB/OL]. (2020−09−21) [2024−06−20]. https://arxiv.org/abs/2009.03300.

[88] SAKAGUCHI K, LE BRAS R, BHAGAVATULA C, et al. WinoGrande: an adversarial winograd schema challenge at scale[J]. Communications of the ACM, 2021, 64(9): 99−106.

[89] ZHONG Wanjun, CUI Ruixiang, GUO Yiduo, et al. AGIEval: a human-centric benchmark for evaluating foundation models[EB/OL]. (2023−09−18) [2024−06−20]. https://arxiv.org/abs/2304.06364.

[90] LI Haonan, ZHANG Yixuan, KOTO F, et al. CMMLU: measuring massive multitask language understanding in Chinese[EB/OL]. (2024−01−17) [2024−06−20]. https://arxiv.org/abs/2306.09212.

[91] ZHANG Xiaotian, LI Chunyang, ZONG Yi, et al. Evaluating the performance of large language models on GAOKAO benchmark[EB/OL]. (2024−02−24) [2024−06−20]. https://arxiv.org/abs/2305.12474.

[92] ZHANG W, ALJUNIED M, GAO C, et al. M3Exam: a multilingual, multimodal, multilevel benchmark for examining large language models[J]. Advances in neural information processing systems, 2023, 36: 5484−5505.

[93] COBBE K, KOSARAJU V, BAVARIAN M, et al. Training verifiers to solve math word problems[EB/OL]. (2021−11−18) [2024−06−20]. https://arxiv.org/abs/2110.14168.

[94] HENDRYCKS D, BURNS C, KADAVATH S, et al. Measuring mathematical problem solving with the MATH dataset[EB/OL]. (2021−11−08) [2024−06−20]. https://arxiv.org/abs/2103.03874.

[95] HU J, RUDER S, SIDDHANT A, et al. XTREME: a massively multilingual multi-task benchmark for evaluating cross-lingual generalisation[C]//International Conference on Machine Learning. Virtual Event: PMLR, 2020: 4411−4421.

[96] WANG A, SINGH A, MICHAEL J, et al. GLUE: a multi-task benchmark and analysis platform for natural language understanding[EB/OL]. (2018−09−18) [2024−

06−20]. https://arxiv.org/abs/1804.07461.

[97] MARCUS G. The next decade in AI: four steps towards robust artificial intelligence[EB/OL]. (2020−02−19)[2024−06−20]. https://arxiv.org/abs/2002.06177.

[98] BENDER E M, KOLLER A. Climbing towards NLU: on meaning, form, and understanding in the age of data[C]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Online: ACL, 2020: 5185−5198.

[99] ZELLERS R, HOLTZMAN A, RASHKIN H, et al. Defending against neural fake news[EB/OL]. (2019−05−29)[2024−06−20]. https://arxiv.org/abs/1905.12616v3.

[100] SCHRAMOWSKI P, TURAN C, ANDERSEN N, et al. Large pre-trained language models contain human-like biases of what is right and wrong to do[J]. Nature machine intelligence, 2022, 4: 258−268.

[101] BENDER E M, GEBRU T, MCMILLAN-MAJOR A, et al. On the dangers of stochastic parrots[C]//Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency. Virtual Event: ACM, 2021: 610−623.

[102] RIEKE N, HANCOX J, LI Wenqi, et al. The future of digital health with federated learning[J]. NPJ digital medicine, 2020, 3: 119.

[103] CHEN Yao, GAN Wensheng, WU Yongdong, et al. Privacy-preserving federated mining of frequent itemsets [J]. Information sciences, 2023, 625: 504−520.

[104] STRUBELL E, GANESH A, MCCALLUM A. Energy and policy considerations for modern deep learning research[J]. Proceedings of the AAAI conference on artificial intelligence, 2020, 34(9): 13693−13696.

[105] ZHANG Chen, XIE Yu, BAI Hang, et al. A survey on federated learning[J]. Knowledge-based systems, 2021, 216: 106775.

作者简介：

肖建力，副教授，吴文俊人工智能科学技术奖获得者，中国计算机学会杰出会员，中国自动化学会高级会员，中国人工智能学会会员，电气电子工程师学会 (IEEE) 高级会员，美国计算机协会 (ACM) 高级会员，主要研究方向为人工智能与大数据，著有图书《人工智能怎么学》。E-mail：audyxiao@sjtu.edu.cn。

黄星宇，硕士研究生，主要研究方向为智慧教育。E-mail：233350741@st.usst.edu.cn。

姜飞，副研究员，主要研究方向为智能教学。E-mail：fjiang@sjtu.edu.cn。