



小型微型计算机系统

Journal of Chinese Computer Systems

ISSN 1000-1220, CN 21-1106/TP

《小型微型计算机系统》网络首发论文

题目: AI 预训练大模型发展综述
作者: 蔡睿, 葛军, 孙哲, 胡冰, 徐玉华, 孙知信
收稿日期: 2023-11-29
网络首发日期: 2024-05-11
引用格式: 蔡睿, 葛军, 孙哲, 胡冰, 徐玉华, 孙知信. AI 预训练大模型发展综述[J/OL]. 小型微型计算机系统. <https://link.cnki.net/urlid/21.1106.tp.20240510.1900.010>



网络首发: 在编辑部工作流程中, 稿件从录用到出版要经历录用定稿、排版定稿、整期汇编定稿等阶段。录用定稿指内容已经确定, 且通过同行评议、主编终审同意刊用的稿件。排版定稿指录用定稿按照期刊特定版式(包括网络呈现版式)排版后的稿件, 可暂不确定出版年、卷、期和页码。整期汇编定稿指出版年、卷、期、页码均已确定的印刷或数字出版的整期汇编稿件。录用定稿网络首发稿件内容必须符合《出版管理条例》和《期刊出版管理规定》的有关规定; 学术研究成果具有创新性、科学性和先进性, 符合编辑部对刊文的录用要求, 不存在学术不端行为及其他侵权行为; 稿件内容应基本符合国家有关书刊编辑、出版的技术标准, 正确使用和统一规范语言文字、符号、数字、外文字母、法定计量单位及地图标注等。为确保录用定稿网络首发的严肃性, 录用定稿一经发布, 不得修改论文题目、作者、机构名称和学术内容, 只可基于编辑规范进行少量文字的修改。

出版确认: 纸质期刊编辑部通过与《中国学术期刊(光盘版)》电子杂志社有限公司签约, 在《中国学术期刊(网络版)》出版传播平台上创办与纸质期刊内容一致的网络版, 以单篇或整期出版形式, 在印刷出版之前刊发论文的录用定稿、排版定稿、整期汇编定稿。因为《中国学术期刊(网络版)》是国家新闻出版广电总局批准的网络连续型出版物(ISSN 2096-4188, CN 11-6037/Z), 所以签约期刊的网络版上网络首发论文视为正式出版。

AI 预训练大模型发展综述

蔡 睿, 葛 军, 孙 哲, 胡 冰, 徐玉华, 孙知信

(南京邮电大学 江苏省邮政大数据技术与应用工程研究中心, 南京 210003)

(南京邮电大学 国家邮政局邮政行业技术研发中心(物联网技术), 南京 210003)

(南京邮电大学 宽带无线通信与传感网技术教育部重点实验室, 南京 210003)

E-mail: sunzx@njupt.edu.cn

摘 要: 本文首先介绍了 AI 预训练大模型相关的部分核心技术, 其中包括 Transformer 架构和人类反馈强化学习技术以及近端策略优化技术; 研究了通用大模型的发展, 重点关注了基于 Transformer-Decoder 架构的 GPT 系列、LLaMA 系列模型与基于 Transformer-Encoder 架构的 BERT、ALBERT、DeBERTa 与 RoBERTa 模型, 深入研究了它们的架构和训练方法, 总结了它们的特点, 探讨了其在不同领域中的应用; 关注了垂直领域的大模型发展, 如金融、医学、法学、自然科学和代码编程等领域。在金融领域, 研究了 BloombergGPT、GPT-InvestAR 和 TradingGPT 模型; 在医学领域, 探讨了 Med-PaLM 和 PMC-LLaMA 等模型; 在法学领域, 分析了 Lawformer 和 Chatlaw 模型; 在自然科学领域, 介绍了华为盘古气象大模型和 FLUID-GPT 模型; 在代码编程领域, 研究了 CodeGeex 和 PanGu-Coder2 模型。最后, 对当前 AI 预训练大模型在知识产权、歧视、成本等方面的局限性与未来发展进行了讨论。

关键词: 人工智能; AI 大模型; 通用大模型; 垂直大模型

Overview of the Development of AI Pre-trained Large Models

CAI Rui, GE Jun, SUN Zhe, HU Bing, XU Yu-hua, SUN Zhi-xin

(Post Big Data Technology and Application Engineering Research Center of Jiangsu Province, Nanjing University of Posts and Telecommunications, Nanjing 210003, China)

(Post Industry Technology Research and Development Center of the State Posts Bureau (Internet of Things Technology), Nanjing University of Posts and Telecommunications, Nanjing 210003, China)

(Key Lab of Broadband Wireless Communication and Sensor Network Technology, Ministry of Education, Nanjing University of Posts and Telecommunications, Nanjing 210003, China)

Abstract: The paper firstly introduces the core technologies related to AI pre-training large models, including Transformer architecture, human feedback reinforcement learning technology, and proximal policy optimization technology. Then, it studies the development of general large models, focusing on the GPT series, LLaMA series models based on the Transformer-Decoder architecture, and the BERT, ALBERT, DeBERTa, and RoBERTa models based on the Transformer-Encoder architecture. The paper deeply analyzes their architectures and training methods, summarizes their characteristics, and discusses their applications in different fields. It also pays attention to the development of large models in vertical fields such as finance, medicine, law, natural science, and code programming. In the finance field, the paper studies the models such as BloombergGPT, GPT-InvestAR, and TradingGPT; in the medical field, it explores the models like Med-PaLM and PMC-LLaMA; in the legal field, it analyzes the models like Lawformer and Chatlaw; in the natural science field, it introduces the Pangu-Weather and the FLUID-GPT; in the code programming field, it studies the models like CodeGeex and PanGu-Coder2. Finally, the paper discusses the limitations and future development of current AI pre-training large models in terms of intellectual property, discrimination, and cost.

Key words: Artificial Intelligence; AI Large language models; General-purpose large models; Vertical large models

1. 引 言

随着科技的飞速发展, 人工智能已经成为当今社会和科

学研究领域的焦点之一。特别是近年来, ChatGPT 的崭露头角引领了人工智能领域的一场革命。AI 大模型, 以其惊人的计算能力和广泛的应用领域, 已经成为了解决各种复杂问题的有力工具。

收稿日期: 2023-11-29 收修改稿日期: 2024-04-01 基金项目: 国家自然科学基金项目(62272239, 61972208)资助; 国家自然科学基金(青年项目)(62302237)资助。作者简介: 蔡 睿, 男, 2001 年生, 硕士研究生, 研究方向为人工智能; 葛 军, 男, 1981 年生, 博士, 讲师, 研究方向为图像处理与三维可视化; 孙 哲, 男, 1982 年生, 博士, 副教授, 研究方向为智能优化算法、欠驱动系统控制与优化; 胡 冰, 女, 1989 年生, 博士, 讲师, 研究方向为无线传感器网络时间同步、移动边缘计算; 徐玉华, 女, 1989 年生, 博士, 讲师, 研究方向为网络异常检测、网络数据隐私安全; 孙知信(通信作者), 男, 1964 年生, 博士, 教授, 研究方向为信息安全、人工智能。

当前 AI 大模型主要分为 2 类：通用大模型，垂直大模型。通用大模型是一种跨领域通用的大型人工智能模型，训练成本十分高昂，具备高度的特征提取和规律发现能力。这些模型在大规模无标注数据上进行训练，以寻找特征并发现规律，从而具备“举一反三”的强大泛化能力。因此，通用大模型可以在不进行微调或仅进行少量微调的情况下，完成多种场景下的任务。这种能力相当于 AI 接受了“通识教育”。垂直大模型是指在特定行业领域中训练和优化的大规模深度学习模型。这些模型通过学习大量的行业相关数据和专业知识，以及针对特定行业任务的训练目标和优化策略，旨在提供更准确、更专业的解决方案和预测结果,相当于成为某些领域的“行业专家”。

AI 大模型也面临着一系列挑战和问题，包括计算资源的需求、数据隐私和安全性、模型的鲁棒性等。本综述将探讨这些问题，同时还将关注 AI 大模型对未来技术发展的潜在影响。

本文将探讨 AI 大模型的主要技术、主流模型和应用案例。同时还将讨论当前 AI 大模型领域的前沿研究和挑战，以及未来可能的发展趋势。希望本综述论文能够为研究人员、决策者和社会大众提供有关 AI 大模型的全面了解，以促进这一领域的可持续发展和创新应用。

2. AI 大模型主要技术

2.1 Transformer 模型

Transformer 模型^[1]是 Google 在 2017 年提出的重要模型，它完全基于注意力机制，摒弃了传统的循环神经网络和卷积神经网络。首先，与循环神经网络和卷积神经网络相比，Transformer 模型的训练时间显著缩短。其次，它能够在不同的表示子空间中捕捉长距离依赖关系，表达能力得到提高。并且，它具有更高的并行性，它的多个头中可以同时关注不同信息，计算效率得到提高。它的适应性还十分广泛，不仅在自然语言处理领域取得优异成果，还具有广泛的应用潜力。Transformer 模型的架构如图 1 所示。

2.1.1 多头自注意力机制

多头自注意力机制^[1]（Multi-Head Self-Attention）是 Transformer 模型中的一个关键部分，它允许模型在不同的表示子空间中联合关注输入序列中的信息。与单一注意力头相比，多头注意力头可以使模型在不同的位置同时关注不同的信息，从而提高模型的表达能力。

在多头自注意力机制中，输入序列被分成多个头，每个头有自己的查询、键和值矩阵。查询矩阵和键矩阵通过点积计算注意力权重，然后将注意力权重与值矩阵相乘以获取关注的信息。多个头的结果被拼接在一起，再经过一个线性变换，得到最终的注意力权重。最后，将注意力权重与值矩阵相乘，得到关注后的信息。

多头自注意力机制的优势在于它可以使模型在不同的表示子空间中捕捉长距离依赖关系，提高模型的并行计算能力。同时，通过使用多个头，模型可以更好地学习输入序列中的不同任务，使注意力分布更加明确。在 Transformer 模型中，多头自注意力机制与位置编码相结合，有效地解决了长距离依赖问题，提高了模型的性能。

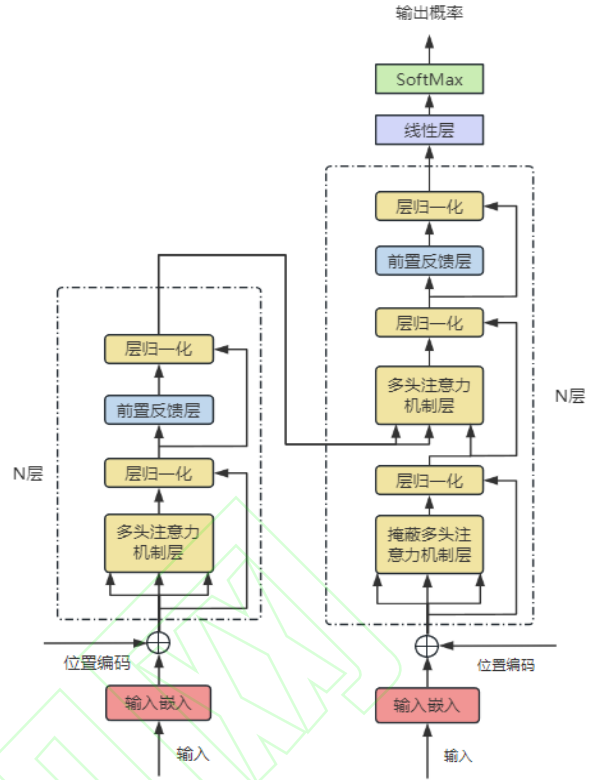


图 1 Transformer 模型架构图

Fig.1 Transformer Architecture Diagram

2.1.2 编码器

编码器是 Transformer 的主要组成部分，用于将输入序列转换为连续的向量表示。编码器通过多层的多头自注意力机制和前馈神经网络^[2]（Feedforward Neural Networks，简称 FNN）进行特征提取和非线性变换，从而逐步将输入序列转换为更高维度的向量表示。

具体来说，编码器中的自注意力机制用于对输入序列中的每个元素进行交互，并根据元素的重要性分配不同的权重，学习输入字符串内部关系。而前馈神经网络用于对输入序列进行非线性变换，并引入新的特征表示。

在编码器中，自注意力机制子层和前馈神经网络子层是交替出现的。通过不断叠加层数，编码器可以逐渐提取出序列中更抽象、更高层次的特征表示。每个子层都采用了残差连接^[3]（Residual Connection），紧接着进行应用层规范化^[4]（Layer Normalization）。这里的编码过程是并行计算的，相比于原来的循环神经网络和卷积神经网络模型，极大地提高了效率。

2.1.3 解码器

Transformer 中的解码器用于将编码器输出的向量表示转换为目标序列，例如生成文本或回答问题等。解码器与编码器类似，也由多层的自注意力机制和前馈神经网络组成，但其输出序列与输入序列相反，即从高维度向低维度转换。

在解码器中，自注意力机制的查询向量、键向量和值向量分别与编码器输出的向量表示进行交互，从而实现对编码器输出的理解和解析。通过不断叠加层数，解码器可以逐渐将高维度的向量表示转换为目标序列中的具体元素，例如单词或字符。

2.2 近端策略优化技术

近端策略优化^[5] (Proximal Policy Optimization, 简称 PPO) 是一种强化学习算法, 用于训练智能代理程序以执行任务和决策, 以最大化累积奖励。它的核心思想是通过在每次训练迭代中保持策略更新的幅度受限, 以确保稳定性。PPO 的训练过程为:

1. 收集数据: 代理程序与环境互动, 收集观测值、采取的动作和获得的奖励。
2. 估计优势函数: 计算每个状态-动作对的优势函数, 以衡量选择每个动作相对于平均水平的好坏程度。
3. 更新策略: 使用 PPO 的目标函数来更新策略, 以增加优势函数。
4. 控制策略更新幅度: PPO 通过引入剪切优化来控制策略更新的幅度, 以确保策略改进不会过于剧烈。

文献[6]中指出, 虽然 PPO 算法在某些基准测试中表现良好, 但是当超越这些基准测试的范围时, 其性能可能会下降。于是, 研究者们提出了一些替代的代理目标和政策参数化方法, 这些方法可以在面对这些失败模式时提供解决方案。文章强调了 PPO 算法的一些实践和失败模式, 并提醒我们注意强化学习算法的设计选择与特定环境的紧密关系。因此, 虽然 PPO 在某些基准测试中表现良好, 但也需要关注其失败模式, 并谨慎选择对 PPO 算法的设计, 以使其适用于特定环境。

2.3 人类反馈强化学习技术

人类反馈强化学习^[7] (Reinforcement Learning from Human Feedback, 简称 RLHF) 是一种先进的 AI 系统训练方法, 它将强化学习与人类反馈相结合。强化学习的基本思想是通过与环境的交互来学习一个最优策略。模型通过与环境互动并接受奖励或惩罚, 尝试找出在给定状态下采取哪种行为是最好的。RLHF 是一种通过将人类训练师的智慧和经验纳入模型训练过程中, 以创建更健壮的学习过程的方法。该技术涉及使用人类反馈创建奖励信号, 然后通过强化学习来改善模型的行为。RLHF 通过用人类生成的反馈替换或补充预先定义的奖励函数, 使得模型能够更好地捕捉复杂的人类偏好和理解。RLHF 的过程为:

1. 识别问题: 需要认识到项目中存在的问题或者需要改进的地方。这个问题可以来自于用户反馈、专家评估或其他途径。
2. 列出可能原因: 在确定问题后, 需要分析可能导致该问题的各种原因。这些原因可以是潜在的缺陷、不足的设计或其他因素。
3. 假设原因: 针对列出的可能原因, 提出一个或多个假设来解释问题的根本原因。这些假设可以是关于问题产生的机制、影响因素等。
4. 测试假设: 通过实验或数据分析来验证提出的假设。如果假设被证实, 那么可以采取相应的措施来解决问题; 如果假设被否定, 那么需要重新回到第一步, 继续识别问题并迭代上述过程。

在 RLHF 技术推出之后, RLHF 的相关变体也竞相推出。

为了解决传统 RLHF 技术中的近端策略优化需要超过监督微调 3 倍的内存, 导致其使用成本大大提高的问题。

Santacrocce 等^[8]提出了一个 Hydra-RLHF 技术。Hydra-RLHF 方法首先将监督微调和奖励模型进行整合, 以监督微调模型作为主体模型, 奖励模型作为辅助模型, 通过对监督微调的训练来优化 PPO 算法。在训练过程中, Hydra-RLHF 方法会动态关闭低秩适应^[9] (Low-Rank Adaptation, 简称 LoRA), 即当训练过程中出现较大损失时, 才会启动 LoRA 模型进行辅助训练, 从而减少内存的使用。实验结果表明, 使用 Hydra-RLHF 方法可以显著减少内存的使用, 同时提高模型与人类偏好的一致性。

Lee 等^[10]提出了一种名为 AI 反馈强化学习 (Reinforcement Learning from AI Feedback, 简称 RLAIFF) 的强化学习技术。它与 RLHF 的最主要区别在于两者的训练数据的来源不同, RLHF 技术利用人类反馈来训练语言模型, 而 RLAIFF 则利用大型语言模型或其他 AI 模型生成的反馈来训练语言模型。在 RLAIFF 中, 先使用 AI 模型生成反馈, 然后将这些反馈作为奖励信号来微调预训练的 AI 模型, 使其生成的反馈更加符合人类偏好。在评估方面, RLAIFF 和 RLHF 表现相似, 都优于监督微调模型。

3. AI 大模型发展状况

3.1 通用大模型

3.1.1 基于 Transformer-Decoder 架构的自回归模型



图2 通用大模型分类情况

Fig.2 Classification of general large models

Transformer-Decoder 模型, 顾名思义为仅使用了 Transformer 架构中的解码器模块, 被称作为自回归模型。自回归模型广泛应用于生成式任务, 例如文本生成。但自回归模型的实现难度较高, 由于其仅仅为单向编码, 而下游语言理解任务往往需要双向的上下文信息, 无法完全捕捉上下文内在联系, 因此其在建模上下文时效果不佳。但自回归模型也有其优点, 它擅长进行生成式自然语言处理任务, 因此更适用于文本生成。

自然语言处理 (Natural Language Processing, 简称 NLP) 领域的快速发展在过去的几年中引领了人工智能研究的前沿。2018 年 6 月, OpenAI 发布了其第一代预训练大模型——GPT-1^[11], GPT-1 采用 Transformer 来代替长短期记忆网络^[12] (Long Short-Term Memory, 简称 LSTM) 作为特征提取器, 且仅采用了 Transformer 的解码器部分, 训练 GPT-1 时, 作者 Alec Radford 等采用了先基于海量无标注语料进行无监督通用的生成式预训练, 然后针对下游任务使用有标注的数据进行微调的训练思路。这种训练方式使得在训练时避免了大量耗时且昂贵的人工标注数据, 提高了模型训练的效率。相较于传统的 ELMO 模型^[13] (Embeddings from Language Models) 使用两对双层双向 LSTM 分别提取上文信息和下

文信息，然后再将提取的信息进行拼接使用的方式。GPT-1 使用的多层单向 Transformer 座位特征提取器的比传统 ELMO 模型在信息提取方面更加简便。但 GPT-1 存在着许多问题，例如：

- 由于其自回归模型的特性导致其缺乏对上下文的理解。

- 训练所用的参数量太少，导致生成内容的可靠性低。

2019 年，延续着 GPT-1 模型相似的架构，Alec Radford 等^[14]推出了 GPT-2。针对 GPT-1 的一些缺陷，研究者们做出许多改进：去掉了微调层，使模型更加适用于通用任务，通用大模型也因此萌芽；增加了数据集，GPT-2 使用了更加广泛、语料更加丰富的数据集。该数据集包含 800 万个网页，大小为 40G，是经过过滤筛选过后的质量文本，使得模型生成的内容真实性得到提高；增加了 Transformer 网络层数，将 Transformer 的层数增加至 48 层，隐层的维度为 1600，参数量达到 15 亿，大大提升了模型生成内容质量。GPT-2 通过零样本学习^[15]（Zero-Shot Learning），使得模型在进行迁移任务时不需要额外的标注数据，大大降低了训练成本。参数量的提升，使得 GPT-2 的生成文本能力相比 GPT-1 更加出色，能够产生更加连贯、语义更加准确的文本。同时也证明了大规模语义模型在合适的数据集上的预训练可以实现对 NLP 任务的零样本学习和适配。但大模型中还有许多问题尚未解决，例如产生幻觉、参数量不够多以及存在歧视等。

Vetagiri 等^[16]提出通过 GPT-2 进行自动分类在线性别歧视内容。在 EXIST 2023 社会网络性别歧视识别共享任务的背景下使用的数据集在自动分类的训练和评估模型中效果十分良好，因此研究者选择它作为 GPT-2 微调的数据集。歧视内容分类由三部分任务组成：任务 1 将任务分为性别歧视或者是非性别歧视，在分类完后，模型会输出一个二进制标签“YES”或者“NO”，指示文本是性别歧视或者是非性别歧视。若任务 1 输出显示为歧视内容，任务 2 将继续对歧视内容进行进一步分类，对歧视内容的原意图进行分类，分别为：直接的、间接的和批判性的。任务 3 则是用来处理复杂歧视问题，可以对例如推特推文中的歧视问题进行多层次多标签分类。经过实验，模型在识别“NO”标签方面表现较好，但在其他标签上的分类准确性较低。总体而言，模型的准确率为 51%。研究内容为社交媒体平台等在线平台提供了一种自动分类工具，可以帮助监测和管理在线性别歧视内容，从而促进网络环境的健康发展。

2020 年，OpenAI 发布了 GPT-3^[17]，这是一款参数量达到 1750 亿的语言模型。相较于之前的 GPT 系列模型，GPT-3 的参数量提升了 2 个量级。GPT-3 通过使用少量或零样本的方式来训练，使其能够以自然语言文本的形式理解任务。借助上下文学习技术^[18]（In-Context Learning，简称 ICL），在训练模型时只需要给预训练模型展示一些输入-输出示例，就能学会做一些全新的事情。虽然 GPT-3 的论文没有明确讨论大型模型所具备的能力，但在文献^[19]中，Jared Kaplan 等观察到了其超越基本标度定律的显著性能提升。尽管 GPT-3 的能力非常出众，但其不可避免地会受到暴力、色情、恶意冒犯和错误语料的影响，导致生成一些错误文本。为了解决这个问题，OpenAI 引入了 RLHF 技术，将其融入到大型模型的训练中，以生成真实、有用且无害的内容。

2022 年 1 月，Long Ouyang 等^[20]提出了 InstructGPT，这是一种基于人类反馈的三阶段强化学习模型，用于改进 GPT-3 模型。RLHF 算法在减轻大型模型生成有害或有毒内容的问题上非常有效，这对于大型模型在实际部署中的安全性至关重要。虽然 GPT-3 性能十分强大，并且在 RLHF 算法的帮助下能生成符合人类偏好且无害的内容，但其仍存在一些缺陷：

- 不支持多模态，而同时期百度的文心一言^[21]大模型已经支持多模态。

- 对歧视等问题消除不够彻底，仍然会生成有害的内容。

Zong 等^[22]提出使用 GPT-3 解关于方程组的数学应用题，研究者们主要探究了三个问题：GPT-3 在将问题划分为不同主题方面的能力如何？GPT-3 在从问题描述中直接提取线性方程组的能力如何？GPT-3 在生成有效问题的创新能力如何？为了探究这些问题，研究者们创建了几个基于从网络上抓取的 200 个问题的数据集，并对数据集中的问题进行手动分类，分为五个不同的类别：求和与差、计算物品价值、计算矩形周长、求解相对运动距离以及计算液体浓度。经过测试，GPT-3 在除了“计算物品价值”类别以外的其他类别的分类准确率都超过了 80%。其中，“计算液体浓度”和“计算矩形周长”组别的识别准确率达到 100%。研究者指出：GPT-3 在分类数学文字问题方面具有高准确性，但在提取方程方面表现较差。但是随着提供给模型的问题数量的增加，提取方程的准确性也会提高。因此，GPT-3 在数学教学中可以作为一个智能辅助工具，帮助学生解决数学词问题，并提供步骤和解释，从而提高学生的学习效果。

Brand 等^[23]认为利用 GPT-3.5-turbo 进行市场研究能快速准确地获取消费者反馈和偏好，提高市场营销效率。研究者首先将其视为消费者模拟器，假设它能模拟人类消费者的喜好和购买决策，因为 GPT-3.5-turbo 在训练过程中吸收了大量人类生成的文本数据。通过引导它进行产品对比，研究者认为 GPT-3.5-turbo 的回应反映了其训练数据中消费者回答的分布。为探究商品属性变化对选择概率和市场份额的影响，研究者多次询问 GPT-3.5-turbo。为获得回答的分布情况，研究者将 GPT-3.5-turbo 的随机性设为最大值，以获取多个回答进行市场研究分析。研究发现，向它提供明确的指导和限制会影响它的回答。比如，询问消费者最高支付价格时，要求它给出具体价格而非长篇回答，能获得更简洁回应。此外，研究者将 GPT-3.5-turbo 的回答与传统市场研究工具进行对比，发现它的回答与现有研究结果一致，并能生成接近实际消费者选择数据的结果。这为市场研究人员和实践者提供了新的工具和方法。

2023 年 3 月，OpenAI 发布了 GPT-4^[24]。GPT-4 相比 GPT-3 最大的提升在于其是首款支持多模态能力的模型，例如图片与文字的混合输入、图标与文字的混合输入。它可以处理多种媒体数据并将其整理到统一的语义空间中。但目前该项功能处于研究阶段，暂时未对外开放。但笔者认为，GPT-4 中的多模态能力的加入并非是重新训练的多模态模型，而是在自然语言模型的基础之上再加入了图像理解、图像生成、图像识别能力。例如外挂一个类似于 Midjourney^[25]之类的绘画模型。

为了研究 GPT-4 中的产生歧视的问题, Zhao 等^[26]提出了一个两阶段的方法。在初始阶段, 大模型的任务是自动完成陈述, 其中可能包含隐性社会偏见。然而, 在接下来的阶段, 同一个大模型重新判断自己产生的偏见陈述, 并反驳它。研究者们认为这种重新判断的不一致性可能类似于人类无意识的内隐社会偏见和有意识的外显社会偏见之间的一致性。Zhao 等^[27]提出了一种名为计数器属性数据增强 (Counter Attribute Data Augmentation, 简称 CADA) 的降歧方法, 该方法通过在预定义属性对的基础上替换属性项来构造相反的数据集, 从而增强训练数据, 减少歧视。

P versus NP 问题^[28]是一个在计算机科学领域备受关注的难题, 被收录在千禧年大奖难题中。该问题最早于 1971 年由 Stephen A. Cook 和 Leonid Levin 提出。多年来, 无数人投身于该问题的研究中, 但至今仍未有人能够给出确切的答案。Dong 等^[29]提出使用 GPT-4 来增强和加速对 P versus NP 问题的研究。研究人员提出了一个名为苏格拉底推理的通用问题解决框架, 灵感来自于古希腊哲学家苏格拉底的方法。该框架通过一系列问题激发批判性思维, 并引导 GPT-4 模型建立高层次的推理路径, 以解决高度复杂的任务。苏格拉底推理通过人与 GPT-4 之间的一系列对话回合进行, 作为一种递归机制, 用于解决 GPT-4 面临的复杂挑战。实验结果表明, 在第 97 轮对话回合中, 他们与 GPT-4 进行了严格的推理, 最终得出了「 $P \neq NP$ 」的结论。该研究证明了利用大型语言模型作为合作伙伴, 对于增强和加速不同领域的科学研究进程是很有必要的。

Westermann 等^[30]提出通过 GPT-4 辅助解决在线争议问题, LLMediator 是一个实验性平台, 用于探索在在线纠纷解决环境中使用大型语言模型的可能性。该平台利用 GPT-4 模型来改写具有情绪性的消息、为用户提供改写建议、在谈判中直接发送改写后的消息。研究者利用 GPT-4 对参与纠纷的各方的信息进行重述, 使其更加中立和客观。通过改变语气和措辞, 模型能够保留信息的重要元素, 同时减少产生歧视的可能性。该项研究通过借助大型语言模型的力量, 为在线纠纷解决拓展了新的可能性, 提升他们的工作效率和准确性, 并为缺乏充分培训的调解人员或推广人员所在的地区提供解决方案, 从而为在线纠纷解决平台的发展和改进提供了宝贵的启示, 引领了新的发展趋势。

表 1 自回归模型对照表

Table 1 Comparison table of general large models

模型名称	发布时间	参数规模	开源情况
GPT-1	2018-6-11	0.117B	✓
GPT-2	2019-2-14	1.5B	✓
GPT-3	2020-5-28	175B	×
GPT-4	2023-3-14	175B	×
LLaMA	2023-2-24	65B	✓
LLaMA2	2023-7-18	70B	✓

2023 年 2 月, Meta 团队发布了其第一代开源大模型——LLaMA^[31], 是 Meta 公司基于 Transformer 架构打造的自回归模型。其拥有 70 亿、130 亿、330 亿、650 亿参数四个

版本。研究者在原本的 Transformer 架构上做出许多改进。

- 使用了由 Zhang 和 Sennrich 引入的均方根层归一化函数^[32] (Root Mean Square Normalization, 简称 RMSNorm), 将每个 Transformer 子层的输入进行归一化, 而不是对输出进行归一化, 提高了训练的稳定性。

- 使用 Shazeer 引入的简化门线性单元^[33] (Simplified Gate Linear Units, 简称 SwiGLU) 取代了修正线性单元^[34] (Rectified Linear Unit, 简称 ReLU) 以提高性能。

2023 年 7 月, Meta 发布了其新一代开源大模型——LLaMA2^[35]。在预训练方面, LLaMA2 分别有 70 亿、130 亿、340 亿、700 亿参数的版本, 数据集上采用了 Meta 公司产品与服务以外的公开来源的新数据组合, Meta 团队经过性能与成本评估后, 最终选择 LLaMA2 在基于 2 万亿 token 的真实数据上进行训练, 使得 LLaMA2 在保持良好性能的基础上, 生成内容的真实性、知识面的广阔性以及模型幻觉的抑制也得到保证。LLaMA2 与 LLaMA 主要的区别在于:

- 文本长度由上一代的 2k 增加到 4k。
- Token 数量由上一代的 1 亿增加到 2 亿。
- 分组查询注意力机制^[36] (Grouped Query Attention, 简称 GQA) 在 LLaMA 2 的两个较大参数的版本得到支持, 该机制可以提高大模型的推理可扩展性。

LLaMA 2 使用与 LLaMA 相同的分词器: 它采用了字节对编码算法^[37] (Byte Pair Encoding, 简称 BPE), 使用 SentencePiece 算法^[38]进行实现。与 LLaMA 一样, 将所有数字拆分为单独的数字, 并使用字节来分解未知的 UTF-8 字符。

经过测试, LLaMA2 的性能虽然在部分项目中与当前主流大模型持平, 但由于其参数量较少, 它的总体水平还是不如那些拥有海量参数的大模型。

Zhang 等^[39]进行了 LLaMA 模型对于写作辅助方面的研究, 研究人员选择了七个具有代表性的写作任务, 并选择了十个数据集来创建一个全面的评估基准。这些任务涵盖了语法、流畅度、清晰度、连贯性、简化、中立化和改写等不同方面的写作要求。除此之外还收集了公开可用的约 60,000 个训练数据, 并将其转化为指令格式。然后使用这些数据对 LLaMA 模型实施多任务指导调优, 显著提升了其在各项写作任务中的表现。在七个写作任务中, LLaMA-7B 模型的平均得分从 30.64 提高到 43.07, 尤其在语法、连贯性、简明和改写任务中, 指导调优对性能的提升尤其显著。本研究的创新之处在于, 研究人员采用多任务指令对 LLaMA 模型进行调整, 以适应各种特定的写作任务。通过融入通用指令和特定场景指令, 使模型更精准地理解任务要求, 并生成更为准确、流畅的文本。

3.1.2 基于 Transformer-Encoder 架构的自编码模型

Transformer-Encoder 模型顾名思义为仅采用了 Transformer 中的 Encoder 模块, 其被称作自编码模型。其原理类似于在给定上下文的情况下, 重新生成当中被抹去的内容。其中 BERT^[40] (Bidirectional Encoder Representations from Transformers) 为典型的自编码模型。

2018 年 10 月, 谷歌团队推出了基于 Transformer-Encoder 架构的 BERT 自回归模型。BERT 包含

了 BERT-Base 与 BERT-Large 两个版本。BERT-Base 通常包含 12 个编码层。这意味着文本输入将经过 12 个编码器层，每个层都包含自注意力机制和前馈神经网络。BERT-Base 的参数较少，适用于一般的自然语言处理任务。BERT-Large 是一个更大的变体，通常包含 24 个编码层。相对于 BERT-Base，它拥有更多的参数，因此在某些复杂的任务上可能表现更好，但也需要更多的计算资源进行训练和推理。BERT 通过使用双向编码器，允许模型全面理解文本中的语境，使其能够同时考虑输入文本中的左侧和右侧上下文信息，从而更好地捕获语义信息。

Liu 等^[41]推出了 BERT 的优化版本——RoBERTa (Robustly Optimized BERT Pretraining Approach)。与 BERT 等先前的方法相比，RoBERTa 在许多 NLP 任务上取得了更好的性能。它相较于 BERT 的区别如下：

- RoBERTa 采用了更大数据，更大的批大小，优化了模型的推理速度与表现。
- 使用了动态掩码方式，对于每一个输入样本序列，都会复制 10 条，然后复制的每一个都会重新随机掩码，其中每个句子被掩码的 token 不同，即拥有不同的掩码 tokens。
- 通过一系列实验证明了下一句预测技术^[42] (Next Sentence Prediction, 简称 NSP) 的不足并将其在 RoBERTa 中去除。

Lan 等^[43]提出了 BERT 的精简版——ALBERT (A Lite BERT)。ALBERT 采用了跨层参数共享技术^[44] (Cross-layer Parameter Sharing, 简称 CPS) 以及句子顺序预测技术^[45] (Sentence-order prediction, 简称 SOP)来取代 NSP 来对 BERT 进行优化。

- CPS 技术将 ALBERT 层与层之间的变化变得平滑，提高了模型的稳定性。
- SOP 技术使得模型学习到的句子之间更加连贯，解决了原版 BERT 中 NSP 损失低效的问题。

He 等^[46]提出了 DeBERTa(Decodingenhanced BERT with disentangled attention)，DeBERTa 相较于 BERT 做出了两项改进：

- 相较于在 BERT 中每个 token 仅用一个向量表示，DeBERTa 中的每个 token 都是用两个向量表示的，分别对内容和位置进行编码，根据 token 的内容和相对位置，利用分散矩阵计算 token 之间的注意权值。这意味着 token 对的注意权重不仅取决于它们的内容，还取决于它们的相对位置。
- DeBERTa 在预训练时采用了增强型掩码解码器，增强了 BERT 的输出层。大大缓解训练前和微调之间的不匹配问题。

表 2 自编码模型对照表

Table 2 Comparison table of autoencoder models

模型名称	主要特点
BERT	由多层 Transformer 的解码器组成
ALBERT	精简版版本 BERT，易于模型扩展
RoBERTa	优化后的 BERT 版本，模型更为精细，参数量增加

DeBERTa	基于注意力解耦机制的解码增强型 BERT
---------	----------------------

Nguyen 等^[47]提出了一种基于 BERT 的面部微表情识别模型，名为 Micron-BERT。微表情是指人脸上极其微小的表情变化，持续时间短暂，通常仅有 0.25 至 0.5 秒，很难被人眼捕捉。过去的研究主要依赖高速摄像机来捕捉微表情，然而现有的方法在准确性和鲁棒性方面仍存在局限性。为此，研究者们采用了对角线微注意力机制 (Diagonal Micro Attention, 简称 DMA)，以精准地检测两个连续视频帧之间的细微差异。此外，他们还引入了一种创新的兴趣区域模块 (Patch of Interest, 简称 PoI)，以无监督的方式进行训练，有效定位包含微表情的面部区域，而无需使用任何面部标签，例如面部边界框或特征点。通过这些组件整合到一个端到端的深度网络中，Micron-BERT 在各种微表情任务中的表现显著优于前人方法。该模型能够大规模地在无标签数据集上进行训练，并在新的未见过的面部微表情数据集上实现高准确率。

Salogni 等^[48]提出了一种利用 BERT 进行地理定位的方法。地理定位是指根据文本内容确定其地理位置的任务。研究者们通过对 BERT 模型进行微调，实现了对展示非标准语言变体的序列进行地理定位。首先，他们选用了包含 15000 个带有地理标记的推特推文数据集。随后将数据集划分为训练、验证和测试集，并对目标和输出坐标数据进行归一化处理。接着，利用 BERT 模型进行预训练，并在任务特定数据集上进行微调。实验结果显示，运用 BERT 进行微调在地理定位任务上能取得优异的性能。然而，研究者们也指出了一些局限性和需进一步研究的方向，例如如何更有效地提取地区和方言模式等问题。

随着当前主流模型向着更大参数、更大训练数据集、更加全能化的方向发展，BERT 类模型仍有其一席之地。在处理自然语言理解任务^[49] (Natural Language Understanding, 简称 NLU) 或者特定领域任务，例如实体识别、信息抽取、文本分类时，BERT 模型能处理的很好。并且由于其体量小的特性，它的部署成本很低，单卡 GPU 即可部署，并且处理速度也十分理想。因此 BERT 类模型仍然可以在其特定领域中发光发热。

3.2 垂直大模型

现阶段的通用大模型虽然涵盖了许多领域，但针对特定领域的垂直大模型仍具有不可替代的优势。本章节主要讲述应用于金融、医学、法学、自然科学、代码编程五个领域中的垂直大模型。

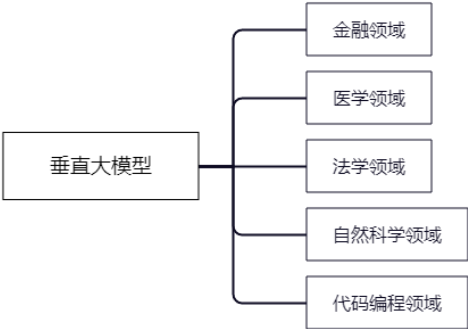


图3 垂直大模型分类情况

Fig.3 Classification of vertical large models

3.2.1 金融领域

Wu 等^[50]针对金融领域提出了 BloombergGPT 金融大模型。BloombergGPT 的训练数据库由一系列英文金融信息组成,包括新闻、文件、新闻稿、网络爬取的金融文件以及提取到的社交媒体消息。为了提高数据质量,研究者增加了来自维基百科等通用数据集的 3450 亿 token 来提升其通用性能。每个数据集都经过过去重处理,保证了数据质量。经测试, BloombergGPT 在财务特定任务上远超过现有模型,同时在通用场景上的表现与现有模型也能一比高下。

Gupta 等^[51]提出了 GPT-InvestAR, 用于分析公司年报以增强股票投资策略, 研究人员获取了美国市值前 1500 家公司涵盖了 2002 年至 2023 年的报告, 共计 24200 份文档。利用 GPT-3.5-turbo 模型分析公司年度报告, 提取关键信息, 并结合历史股价数据进行模型训练。这种方法能够更准确地预测股票的表现, 提供更好的投资策略, 预测未来一年内表现最佳的股票。但此项研究也有其不足之处: 由于训练的数据仅是从 2002 年至 2023 年, 对于 2023 年以后的数据预测可能不准确, 因此可能无法准确反映当前的市场情况。需要准确预测市场情况就需要经常对模型进行及时的训练更新, 因此训练的成本也较大。

Li 等^[52]提出了 TradingGPT, 将其用于在金融市场的股票交易中。TradingGPT 也基于来自 OpenAI 公司的 GPT-3.5-turbo, 使用了来源于金融数据库的自 2020 年 8 月 15 日到 2023 年 8 月 15 日的大量多模式财务数据用作数据集, 例如 Databento 股票价格数据库、Alpaca 新闻 API、方舟公开的每日持股历史记录等。由于大型语言模型的记忆系统并不像人类一样分为长、中、短期层, 因此这导致大型语言模型并不能像人类一样优先处理即时与关键任务, 在金融市场中也不能及时处理突发的情况。为了解决这个问题, 研究人员将大模型的记忆分为三个不同的层次, 并且在每个层次中, 通过三个关键指标——新近度、相关性和重要性, 在记忆内建立事件的分层情况。除此之外, 每个层次都由一个自定义的衰减机制控制来模拟人类的记忆衰减情况。该系统旨在通过模拟人类交易者的认知行为, 确保在不断变化的市场场景中具备良好的响应能力, 以实现优于其他交易代理系统的交易性能。

3.2.2 医学领域

谷歌团队的 Singhal 等^[53]提出了 Med-PaLM, 该模型的前身为 Flan-PaLM^[54], 它们都基于谷歌早先提出的 PaLM^[55]模型, Med-PaLM 的训练过程采用了指令调优的方法, 通过在多个数据集上进行微调来扩大任务数量和模型规模, 并使用思维链^[56]数据作为指令。这种方法的有效性在于其能够提高模型的性能, 尤其是在需要推理的任务中表现十分出色。具体来说, Med-PaLM 的训练过程包括在多个数据集上进行微调, 每个数据集的例子都以一些指令或少量示例为前缀。这种方法能够帮助模型更好地理解任务的背景和要求, 从而提高其性能。Med-PaLM 模型在 BIG-bench^[57] (Beyond the Imitation Game Benchmark) 和 MMLU^[58] (Massive Multitask Language Understanding) 等基准测试上取得了最先进的性能。特别是在使用思维链数据进行微调后, Med-PaLM 模型表现

出了出色的推理能力, 能够更好地应对复杂的任务。Med-PaLM 在经过精心调优后, 在医学领域的多个评价指标上都有卓越表现。具体而言, Med-PaLM 在解答医学问题时, 其答案的完整性与准确性均有显著提升。在一次由临床医生参与的评估活动中, Med-PaLM 的答案得分率达到了 95.4%, 与临床医生的答案得分率 (97.8%) 相差无几。相较之下, 未经指令提示调优的 Flan-PaLM 在同一评估中的得分率仅为 76.3%。此外, Med-PaLM 在解答消费者医学问题方面的表现同样得到了优化。在一个名为 Health-SearchQA 的评估数据集上, Med-PaLM 的答案获得了 61.9% 的认同率, 相较于 Flan-PaLM 的 51.4% 认同率更胜一筹。综合来看, 经过指令提示调优的 Med-PaLM 模型在医学领域表现卓越, 不仅在医学知识检索方面缩小了与临床医生的差距, 而且在解答消费者医学问题方面也展现出了明显优势。

Wu 等^[59]提出了 PMC-LLaMA, 该模型基于 Meta 公司发布的开源模型 LLaMA, 其拥有 130 亿参数。在知识注入阶段, 研究团队收集了大量的医学相关文献和教科书作为数据源。其中, 医学文献主要来自生物医学论文和公共医学中心数据库 (PubMed Central, 简称 PMC), 共计约 4.8 百万篇论文。教科书则来自多个渠道, 包括开放图书馆、大学图书馆和知名出版商, 共计约 3 万本。在医学特定指令的调整阶段, 研究团队主要考虑了医学对话、医学合理性问答和知识图谱三个方面。其中, 医学对话和医学合理性问答的结合, 能够显著提升模型的性能。同时, 引入知识图谱的指令调整, 进一步强调了关键的医学概念。通过这两个关键步骤, 研究团队成功构建了一个医学领域的开源语言模型, 并将其命名为 PMC-LLaMA。经过测试, 该模型在参数量小于其他模型的情况下, 在多个医学问答基准测试中表现出色, 超过了其他主流模型。

3.2.3 法学领域

Xiao 等^[60]提出了针对中文法律长文档调优的预训练语言模型——Lawformer。在数据处理阶段, 研究者收集了数千万份中国政府发布的案件文件, 并从中筛选出刑事案件和民事案件。将每个文档分为四个部分: 当事人信息、事实描述、法院观点和判决结果。只保留事实描述超过 50 个标记的文档进行后续处理。在预训练阶段, 研究者对 Lawformer 的学习率设置为 5×10^{-5} , 序列长度设置为 4096, 批大小设置为 32。由于法律文件的长度通常小于 4,096, 于是研究者将不同文档连接在一起以充分利用输入长度。预训练时, Lawformer 进行了 200,000 个步骤, 前 3,000 个步骤用于热身。通过与其他模型例如 BERT、RoBERTa 进行比较, 在刑事和民事案件的判决预测任务中, Lawformer 在准确率、召回率等指标上均取得了优于以上模型的结果。在法律领域中, 长文档的处理是一项具有挑战性的任务, Lawformer 提供了一种有效的解决方案, 为法律文本分析提供了新的思路和方法。并且在多个法律任务上取得了优异的性能, 可以为法律从业者提供更准确、高效的法律信息检索和分析工具。

Cui 等^[61]提出了一个开源的法律大型语言模型——Chatlaw, 同时集成外部知识库, 以解决法律领域的研究问题。构建数据集阶段, 研究者通过多种途径构建了全面且多样化的数据集, 其中包括收集大量的原始法律数据、整理真

实的法律咨询数据以及创建用于法律考试的多项选择题数据集等。训练模型时,研究者以 LLaMA-13B 模型为基础,通过应用 LoRA 方法进行微调,并引入自监督角色以减轻模型产生幻觉的问题。此外,研究者利用多个 GPU 进行训练,并结合 DeepSpeed 技术^[62]以进一步降低训练成本,从而提高模型的性能和效率。为了增强模型的训练效果,研究者在推理过程中引入了“咨询”、“参考”、“自我建议”和“回应”四个模块,将垂直模型与知识库相结合。将特定领域的知识融入模型,并从知识库中获取准确信息,旨在减少幻觉现象的发生,从而提高模型的准确性和可靠性。尽管模型在微调以及挂载外部知识库后性能得到很大提升,但是模型在处理复杂的法律问题时可能存在信息不全或错误理解的情况,并且对于一些特定领域的法律问题,模型的表现可能不如专业领域的法律专家。

3.2.4 自然科学领域

Bi 等^[63]在提出了华为盘古气象大模型。为了解决此前 AI 气象预报模型的精度不足的问题,研究团队首先将高度信息整合到一个新的维度中,使得深度神经网络的输入和输出可以在三维空间中概念化。他们设计了一个三维地球特定 Transformer 模型 (Three-dimensional Earth-specific Transformer, 简称 3D-EST), 将地球特定的先验知识注入到深层网络中。与二维模型相比,这个三维模型通过将高度表述为单个维度,能够捕捉不同压力水平下大气状态之间的关系,显著提高了预测的准确性。由于气象数据分辨率很大,因而相比于常见的视觉 Transformer 方法^[64] (Vision Transformer, 简称 ViT), 研究人员将网络的编码器和解码器减少到 2 级,同时采用滑窗注意力机制^[65],以减少网络的计算量。此外,研究团队还应用了一种层次化时域聚合策略,该算法训练了几种不同时间间隔的模型并通过贪心算法来调用这些模型使得模型调用的次数最少。因此,在应对不同中期预报的情况时,该算法都能找到最少调用模型的次数,从而使中期天气预报所需的迭代次数大大减少,减轻了预报误差。在训练效率方面,盘古气象大模型在一张 V100 显卡上只需要 1.4 秒就能完成 24 小时的全球气象预报,相比传统数值预报提速 10000 倍以上。而传统数值预报对算力的消耗非常大,需在超过 3000 个节点的超级计算机上花费数小时进行仿真。

Yang 等^[66]提出 FLUID-GPT 混合机器学习模型,用于预测颗粒轨迹和表面侵蚀速率。首先,研究者利用 GPT-2 模型对时间序列数据进行训练,以便识别序列数据中的模式和相互关系。接着,研究者采用卷积神经网络模型^[67] (Convolutional Neural Network, 简称 CNN), 基于 GPT-2 模型生成的轨迹数据进行侵蚀预测。研究者采用了自定义的 GPT-2 架构,并借助 PyTorch 框架实现了 FLUID-GPT 模型。为提高模型学习能力,研究者对输入数据进行了增强,将计算中使用的五个初始参数(颗粒大小、主进口速度、主进口压力、次进口速度和次进口压力)纳入到输入数据中。在模型训练过程中,研究者还优化了学习率、窗口大小和步长等超参数,以提高模型的训练效率和准确性。研究结果表明,较小的步长可以提高预测准确性,但会增加训练时间。研究者还探讨了解码器层数和注意力头数对模型性能的影响,并找到了最佳配置参数。研究者比较分析了

FLUID-GPT 模型、LSTM 模型和双向长短期记忆网络模型^[68] (Bi-directional Long Short-Term Memory, 简称 BiLSTM) 的训练效率和准确性,发现 FLUID-GPT 模型在分析序列数据方面具有高效和准确的优势,为预测复杂的颗粒动力学迈出了重要一步。

3.2.5 代码编程领域

Zheng 等^[69]提出了用于代码生成和多语言代码生成的预训练模型 CodeGeeX, 参数量为 130 亿。研究者使用包含 23 种不同编程语言的大规模代码数据集对 CodeGeeX 进行了训练,使得 CodeGeeX 能够生成高质量的代码,并具备多语言代码生成能力以及代码翻译能力。为了评估 CodeGeeX 的性能,研究者提出了一个名为 HumanEval-X 的多语言基准测试,该基准测试包含了 820 个问题—解决的方案对,涵盖了 Python、C++、Java、JavaScript 和 Go 这五种目前市面上的主流编程语言,并且每个问题都经过手工重写,以确保符合相应语言的编程风格和语法规范。经测试,在代码生成方面,CodeGeeX 在不同语言上的生成能力存在差异,其中 Python 的生成能力最强,其他语言相对较弱。此外,CodeGeeX 生成的代码在语法错误和语义错误方面的表现良好。在代码翻译方面,CodeGeeX 在不同语言之间的翻译能力存在不对称性,即 A 到 B 的翻译性能与 B 到 A 的翻译性能通常呈负相关。这表明多语言代码生成模型在代码翻译过程中对源语言和目标语言的关注程度存在不平衡。CodeGeeX 为开发者提供了强大的代码生成和翻译功能,可以提高开发效率和代码质量,减少开发成本和工作量。

Shen 等^[70]提出了一种根据反馈对回答进行排序框架 (Rank Responses to align Test&Teacher Feedback, 简称 RRTF), 用于提高代码生成大型语言模型的性能。该框架通过使用测试信号和人类偏好作为反馈,有效地提高了预训练的代码生成大型语言模型的性能。并通过这种方法,训练出了代码生成大型语言模型 PanGu-Coder2。RRTF 框架的主要组成部分如下:

- 测试信号: 这些信号用于评估模型在特定任务上的性能。通过对比模型生成的代码与预期输出,可以确定模型在哪些方面需要改进。

- 教师反馈: 教师反馈是一种指导模型如何改进的方法。在 RRTF 框架中,作者使用了人类偏好作为一种教师反馈,以引导模型生成更符合人类编程习惯的代码。

- 回答排名: 作者通过对模型生成的代码进行排序,根据人类偏好和测试信号来调整模型。这种方法有助于在训练过程中让模型关注更有价值的代码片段。

- 模型训练与优化: 在 RRTF 框架下,作者采用了多种训练和优化技术,如监督微调、指令调优和强化学习等。这些方法有助于提高模型在代码生成任务上的性能。

通过 RRTF 框架,研究者成功地提高了预训练的代码生成大型语言模型的性能。在框架下开发的 PanGu-Coder2 模型在 HumanEval 基准测试中取得了 62.20% 的通过率,同时在 CoderEval 和 LeetCode 基准测试中表现优于之前发布的所有代码生成大型语言模型。这表明 RRTF 框架在提高代码生成模型性能方面具有实际价值。

3.2.6 小结

总的来说,在更强的通用大模型推出之前,垂直大模

型相较于通用大模型在特定领域有其独特的优势：

- 垂直大模型可以在特定领域进行深入训练，从而具备更多的专业知识和领域特定的语义理解，这使得它在该领域内的问题回答和任务执行方面更加准确和可靠。

- 由于垂直大模型专注于特定领域，因此它可以更高效地处理该领域的任务和问题，这意味着它在执行特定任务时可能比通用大模型更快速和高效。

- 垂直大模型可以根据特定领域的需求进行调整和优化，以提供更好的适应性和定制性。这使得它可以更好地满足特定领域的需求，并针对该领域的特定问题提供更准确的解决方案。

4. AI 大模型的局限与未来发展

4.1 AI 大模型的局限

当前的 AI 大模型虽然能力已经十分强大，但其发展仍然面临种种挑战。

1. AI 大模型会对知识产权保护产生侵害。在训练的过程中，AI 大模型学习了大量来自互联网的作品，某些作者可能享有这些作品的著作权。举个例子，Midjourney 可以上传图片，Midjourney 可以对上传的图片进行二次创作，二次创作可能会对原作者著作权的产生侵犯。但由于我国当前并未就 AI 大模型及相关行为进行明确立法，因此这些作者的著作权与知识产权也就得不到相应的保护。

2. AI 大模型仍会产生歧视问题。由于目前社会环境中存在大量歧视问题，人们在平时的作品中会产生歧视的元素，因此模型也避免不了学习到歧视的元素^[71]。著名的图像生成 AI 在测试中也显示出了明显的种族偏见问题。当要求生成“医生”、“高管”等高薪职业的人物图像时，几乎所有的输出都呈现了白人形象。

3. AI 大模型部署与训练的算力与电力成本巨大。部署 Meta 公司开源的 650 亿参数的 LLaMA 模型需要使用到 130GB 显存，意味着在消费领域需要 8 张 4090 显卡才能部署该模型。LLaMA 模型在训练中消耗的总电量，如表 3 所示。

表 3 训练 LLaMA 模型所需的耗电量

Table 3 The electricity consumption required for training the LLaMA model.

	GPU	功耗	计算时间	总耗电量
LLaMA-7B	A100-80GB	400W	82,432	36 兆瓦时
LLaMA-13B	A100-80GB	400W	135,168	59 兆瓦时
LLaMA-33B	A100-80GB	400W	530,432	233 兆瓦时
LLaMA-65B	A100-80GB	400W	1,022,362	449 兆瓦时

训练 LLaMA-65B 的电费，以国内电价 0.6 元/度估算，需要大约 40 万元人民币。一张 A100-80GB 显卡的零售价为 9.1999 万人民币，训练 LLaMA 使用了 2048 块 A100-80GB，可以计算出仅 GPU 的成本为 1.88 亿元。这使得大型科技公司在计算资源上形成壁垒，让小型科技公司与普通消费者望而却步。

4. AI 大模型的训练与部署对环境有影响。由于 AI 大模型需要海量的 GPU 资源进行训练与微调，因此 AI 大模型训练过程中的碳排放量十分巨大。据统计，Meta 公司使用

自家的研究超级集群以及内部生产集群（都使用英伟达 A100GPU），在 33 万小时的 LLaMA2 模型训练的过程中，共计产生了 539 吨二氧化碳。

4.2 AI 大模型的未来发展

未来 AI 大模型的应用领域将会更加广泛。在智能客服领域，AI 大模型将能够更好地理解客户需求并提供更准确的解答，从而提高客服效率，降低企业运营成本，并提升客户满意度。在广告推荐领域，AI 大模型将能够更准确地挖掘用户兴趣和行为，从而提供更精准的广告推荐，为企业带来更多的收益。在舆情监测领域，AI 大模型将能够更准确地挖掘网络上的舆情信息，并提供更全面的舆情分析报告，为政府和企业提供更好的决策支持。

目前已经有一些成功的 AI 大模型应用案例。例如，在自然语言处理领域，GPT 系列模型已经能够生成高质量的自然语言文本，AI 生成内容（AI Generated Content，简称 AIGC）将会成为一种全新的内容生产模式。相较于专业生成内容（Professional Generated Content，简称 PGC）和用户生产内容（User Generated Content，简称 UGC），AIGC 不仅具有更高的产出效率、更为稳定的内容质量、更低的产出成本，其内容的可拓展性也将更强^[72]。在计算机视觉领域，ResNet 系列模型已经在图像分类、目标检测等领域取得了显著成果。

对于未来 AI 大模型的期望，首先希望能够解决模型可解释性的问题。目前，许多 AI 大模型的决策过程缺乏透明度，使得人们难以理解其决策依据。因此，未来的 AI 大模型需要能够提供更清晰的解释和说明。

其次，期望能够加强隐私保护和数据安全。随着 AI 大模型的广泛应用，个人数据的安全和隐私保护问题日益突出。未来的 AI 大模型需要采用更加严格的隐私保护和数据安全措施，确保个人数据的安全性和保密性。

最后，期望能够优化模型的算法和结构，提高模型的性能和准确性。未来的 AI 大模型需要不断优化和改进，以提高其处理复杂任务的能力和效率。

相信随着技术的不断进步和优化，未来的 AI 大模型将会为我们带来更多的惊喜和改变。

References:

- [1] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[J].Proceedings of the 31st International Conference on Neural Information Processing Systems,2017:5998-6008.
- [2] Ketkar N, Moolayil J, Ketkar N, et al. Feed-forward neural networks[J].Deep Learning with Python: Learn Best Practices of Deep Learning Models with PyTorch,2021:93-131.
- [3] He K, Zhang X, Ren S, et al. Deep residual learning for image recognition[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016:770-778.
- [4] Ba J L, Kiros J R, Hinton G E. Layer normalization[J]. arXiv preprint arXiv:1607.06450,2016.

- [5] Wang Y, He H, Tan X. Truly proximal policy optimization[C]//Uncertainty in Artificial Intelligence, PMLR,2020:113-122.
- [6] Queeney J, Paschalidis Y, Cassandras C G .Generalized proximal policy optimization with sample reuse[J].Advances in Neural Information Processing Systems,2021, 34:11909-11919.
- [7] Sumers T R, Ho M K, Hawkins R D, et al. Learning rewards from linguistic feedback[C]//Proceedings of the AAAI Conference on Artificial Intelligence,2021:6002-6010.
- [8] Santacrose M, Lu Y, Yu H, et al. Efficient RLHF: reducing the memory usage of PPO[J].arXiv preprint arXiv:2309.00754,2023.
- [9] Hu E J, Wallis P, Allen Zhu Z, et al. LoRA: low-rank adaptation of large language models[C]//International Conference on Learning Representations,2021.
- [10] Lee H, Phatale S, Mansoor H, et al. Rlaif: scaling reinforcement learning from human feedback with ai feedback[J].arXiv preprint arXiv:2309.00267,2023.
- [11] Radford A, Narasimhan K, Salimans T, et al. Improving language understanding by generative pretraining[EB/OL].https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf,2023.
- [12] Greff K, Srivastava R K, Koutnik J, et al. LSTM: a search space odyssey[J].IEEE Transactions on Neural Networks and Learning Systems,2016,28(10):2222-2232.
- [13] Peters M E, Neumann M, Iyyer M, et al. Deep contextualized word representations[C]//Proceedings of NAACL HLT,2018:2227-2237.
- [14] Radford A, Wu J, Child R, et al. Language models are unsupervised multitask learners[J].OpenAI Blog,2019, 1(8): 9.
- [15] Guo Y, Ding G, Han J, et al. Zero-shot learning with transferred samples[J].IEEE Transactions on Image Processing,2017,26(7):3277-3290.
- [16] Vetagiri A, Adhikary P K, Pakray P, et al. Leveraging GPT-2 for automated classification of online sexist content[J]. Working Notes of CLEF,2023.
- [17] Brown T, Mann B, Ryder N, et al. Language models are fewshot learners[J].Advances in Neural Information Processing Systems,2020,33:1877-1901.
- [18] Xun G, Jia X, Gopalakrishnan V, et al. A survey on context learning[J].IEEE Transactions on Knowledge and Data Engineering,2016,29(1):38-56.
- [19] Kaplan J, McCandlish S, Henighan T, et al. Scaling laws for neural language models[J].2020,doi:10.48550
- [20] Ouyang L, Wu J, Jiang X, et al. Training language models to follow instructions with human feedback[J].Advances in Neural Information Processing Systems,2022,35: 27730-27744.
- [21] Sun Y, Wang S, Feng S, et al. ERNIE 3.0: large-scale knowledge enhanced pre-training for language understanding and generation[J].2021,doi:10.48550.
- [22] Zong M, Krishnamachari B. Solving math word problems concerning systems of equations with gpt-3[C]//Proceedings of the AAAI Conference on Artificial Intelligence,2023:15972-15979.
- [23] Rivas P, Zhao L. Marketing with chatgpt: navigating the ethical terrain of gpt-based chatbot technology[J].AI, 2023, 4(2):375-384.
- [24] OpenAI.GPT-4 technical report[EB/OL].<https://cdn.openai.com/papers/gpt-4.pdf>,2023.
- [25] Chiu T K F. The impact of generative AI (GenAI) on practices, policies and research direction in education: a case of ChatGPT and Midjourney[J].Interactive Learning Environments,2023: 1-17.
- [26] Zhao Y, Wang B, Zhao D, et al. Mind vs. mouth: on measuring rejudge inconsistency of social bias in large language models[J].arXiv preprint arXiv:2308.12578, 2023.
- [27] Zhao J, Fang M, Shi Z, et al. CHBias: bias evaluation and mitigation of Chinese conversational language models[J]. arXiv preprint arXiv:2305.11262,2023.
- [28] Fortnow L. The status of the P versus NP problem[J]. Communications of the ACM,2009,52(9):78-86.
- [29] Dong Q, Dong L, Xu K, et al. Large language model for science: a study on P vs. NP[J].arXiv preprint arXiv:2309.05689,2023.
- [30] Westermann H, Savelka J, Benyekhlef K. Llmmediator: Gpt-4 assisted online dispute resolution[J].arXiv preprint arXiv:2307.16732,2023.
- [31] Touvron H, Lavril T, Izacard G, et al. Llama: open and efficient foundation language models[J].arXiv preprint arXiv:2302.13971,2023.
- [32] Zhang B, Sennrich R. Root mean square layer normalization[C]//Proceedings of the 33rd International Conference on Neural Information Processing Systems, 2019:12381-12392.
- [33] Shazeer N. Glue variants improve transformer[J].arXiv preprint arXiv:2002.05202,2020.
- [34] He J, Li L, Xu J, et al. Relu deep neural networks and linear finite elements [J].Journal of Computational Mathematics, 2020,38(3):502-527.
- [35] Touvron H, Martin L, Stone K, et al. Llama 2: open foundation and fine-tuned chat models[J].arXiv preprint arXiv:2307.09288,2023.
- [36] Hudson D A, Manning C D. Gqa: a new dataset for real-world visual reasoning and compositional question answering[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition,2019: 6700-6709.

- [37] Sennrich R, Haddow B, Birch A. Neural machine translation of rare words with subword units[C]//54th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics (ACL),2016: 1715-1725.
- [38] Kudo T, Richardson J. Sentence Piece: a simple and language independent subword tokenizer and detokenizer for Neural Text Processing[C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing: System Demonstrations, Association for Computational Linguistics,2018.
- [39] Zhang Y, Cui L, Cai D, et al. Multi-task instruction tuning of LLaMa for specific scenarios: a preliminary study on writing assistance[J].arXiv preprint arXiv:2305.13225, 2023.
- [40] Kenton J D M W C, Toutanova L K. BERT: pre-training of deep bidirectional transformers for language understanding[C]//Proceedings of NAACL-HLT,2019: 4171-4186.
- [41] Liu Y, Ott M, Goyal N, et al. RoBERTa: a robustly optimized BERT pretraining approach[J]. 2019,doi:10.48550.
- [42] Shi W, Demberg V. Next sentence prediction helps implicit discourse relation classification within and across domains[C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing,2019:5790-5796.
- [43] Lan Z, Chen M, Goodman S, et al. ALBERT: a lite BERT for self-supervised learning of language representations[C]//International Conference on Learning Representations,2019.
- [44] Hanif M A, Shafique M. Cross-layer optimizations for efficient deep learning inference at the edge[M]//Embedded Machine Learning for Cyber-Physical, IoT, and Edge Computing: Software Optimizations and Hardware/Software Codesign, Cham: Springer Nature Switzerland,2023:225-248.
- [45] Cui B, Li Y, Chen M, et al. Deep attentive sentence ordering network[C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing,2018: 4340-4349.
- [46] He P, Liu X, Gao J, et al. DeBERTa: decoding enhanced bert with disentangled attention[C]//International Conference on Learning Representations,2020.
- [47] Nguyen X B, Duong C N, Li X, et al. Micron-BERT: BERT-based facial micro-expression recognition[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition,2023: 1482-1492.
- [48] Salogni I. Salogni at GeoLingIt: geolocalization by fine-tuning BERT[C]//Proceedings of the 8th Evaluation Campaign of Natural Language Processing and Speech Tools for Italian, Final Workshop (EVALITA),CEUR.org, Parma,Italy,2023.
- [49] Bates M. Models of natural language understanding[J]. Proceedings of the National Academy of Sciences,1995, 92(22): 9977-9982.
- [50] Wu S, Irsoy O, Lu S, et al. Bloomberg GPT: a large language model for finance[J].arXiv preprint arXiv:2303.17564, 2023.
- [51] Gupta U. GPT-InvestAR: enhancing stock investment strategies through annual report analysis with large language models[J].arXiv preprint arXiv:2309.03079, 2023.
- [52] Li Y, Yu Y, Li H, et al. Trading GPT: multi-agent system with layered memory and distinct characters for enhanced financial trading performance[J].arXiv preprint arXiv:2309.03736,2023.
- [53] Singhal K, Azizi S, Tu T, et al. Large language models encode clinical knowledge[J].Nature,2023,620(7972): 172-180.
- [54] Wei J, Bosma M, Zhao V, et al. Finetuned language models are zero-shot learners[C]//International Conference on Learning Representations,2021.
- [55] Chowdhery A, Narang S, Devlin J, et al. PaLM: scaling language modeling with pathways[J].2022,doi:10.48550.
- [56] Wei J, Wang X, Schuurmans D, et al. Chain-of-thought prompting elicits reasoning in large language models[J]. Advances in Neural Information Processing Systems, 2022,35:24824-24837.
- [57] Ghazal A, Rabl T, Hu M, et al. Bigbench: towards an industry standard benchmark for big data analytics[C]//Proceedings of the ACM SIGMOD International Conference on Management of Data,2013: 1197-1208.
- [58] Hendrycks D, Burns C, Basart S, et al. Measuring massive multitask language understanding[C]//International Conference on Learning Representations,2020.
- [59] Wu C, Zhang X, Zhang Y, et al. Pmc-llama: further finetuning llama on medical papers[J].arXiv preprint arXiv:2304.14454,2023.
- [60] Xiao C, Hu X, Liu Z, et al. Lawformer: a pre-trained language model for Chinese legal long documents[J].AI Open,2021:79-84.
- [61] Cui J, Li Z, Yan Y, et al. Chatlaw: open-source legal large language model with integrated external knowledge bases[J].arXiv preprint arXiv:2306.16092,2023.
- [62] Rasley J, Rajbhandari S, Ruwase O, et al. Deepspeed: System optimizations enable training deep learning models with over 100 billion parameters[C]//Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining,2020:3505-3506.

- [63] Bi K, Xie L, Zhang H, et al. Accurate medium-range global weather forecasting with 3D neural networks[J].Nature, 2023,619(7970): 533-538.
- [64] Han K, Wang Y, Chen H, et al. A survey on vision transformer[J].IEEE Transactions on Pattern Analysis and Machine Intelligence,2022,45(1):87-110.
- [65] Liu Z, Lin Y, Cao Y, et al. Swin transformer: Hierarchical vision transformer using shifted windows[C]// Proceedings of the IEEE/CVF International Conference on Computer Vision,2021:10012-10022.
- [66] Yang S D, Ali Z A, Wong B M.FLUID-GPT fast learning to understand and investigate dynamics with a generative pretrained transformer: efficient predictions of particle trajectories and erosion[J].Industrial & Engineering Chemistry Research,2023.
- [67] LeCun Y, Bottou L, Bengio Y, et al. Gradient-based learning applied to document recognition[J].Proceedings of the IEEE,1998,86(11): 2278-2324.
- [68] Zhou P, Shi W, Tian J, et al. Attention-based bidirectional long short-term memory networks for relation classification[C]//Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, 2016:207-212.
- [69] Zheng Q, Xia X, Zou X, et al. Codegeex: a pre-trained model for code generation with multilingual benchmarking on humaneval-x[C]//Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, 2023:5673-5684.
- [70] Shen B, Zhang J, Chen T, et al.Pangu-coder2:boosting large language models for code with ranking feedback[J].arXiv preprint arXiv:2307.14936,2023.
- [71] Carlini N, Hayes J,Nasr M, et al. Extracting training data from diffusion models[C]//32nd USENIX Security Symposium,2023:5253-5270.
- [72] CHEN Y W. Beyond ChatGPT: opportunities, risks and challenges of generative AI[J].Journal of Shandong University(Philosophy and Social Sciences Edition),2023, (3):127-143.

附中文参考文献:

- [72] 陈永伟. 超越 ChatGPT: 生成式 AI 的机遇、风险与挑战[J]. 山东大学学报 (哲学社会科学版), 2023, (3): 127-143.