基于大语言模型的高考卷评测应用及其对教育公平的影响研究

王一来　朱华中[1]

（1.浙江工业大学　310014）

【摘　要】随着人工智能技术的迅猛发展，大语言模型在教育领域的应用逐渐增多，尤其是在大语言模型自动化评测和教育公平方面的探索。本文利用 Open Compass 评估平台，评估了多个大语言模型在中国高考全卷中的具体表现，并深入分析了其对教育公平的影响。通过全面评估模型在不同题型中的表现，本文揭示了大语言模型在教育领域应用中面临的主要挑战。这些挑战不仅包括推理能力的提升、对领域知识的掌握不够全面以及训练数据集的优化空间，还涉及教育公平性的问题。随着模型能力的不断提升和优化，未来的研究将更加关注如何通过改进推理能力、丰富领域知识以及优化训练数据集来应对这些挑战。尤其是在高考这种全国性考试中，未来的模型可能会面临更多的公平性问题。因此，确保教育公平性将成为未来研究的重要方向,这需要我们不断探索和引入更加先进的技术提升大语言模型的综合能力,从而实现更公正的评估结果。

【关键词】大语言模型；人工智能；高考卷评测；教育公平

1 引言

在全球范围内，教育一直被视为推动社会进步的重要力量，也是实现个人发展和社会公平的关键途径。教育公平，不仅仅意味着为所有学生提供平等的学习机会，还意味着确保每个学生，无论其背景如何，都能够获得相应的支持，最大程度地发挥其潜力。随着人工智能技术的发展，特别是大语言模型的广泛应用，教育领域迎来了前所未有的变革机会，这种变革不仅体现在教学方式的创新，更在于教育公平的进一步实现。

近年来，伴随着人工智能技术（Artificial Intelligence，AI）的迅猛发展，各个领域都迎来了新一轮的变革浪潮[1]，自然语言处理（Natural Language Processing，NLP）作为计算机科学和人工智能的重要研究方向，大语言模型成为其中的杰出成果之一。在 Transformer 架构[2]提出之前，自然语言处理主要依赖于循环神经网络（Recurrent Neural Networks，RNN）[3]及其变体。这些传统方法由于序列依赖性问题，常常面临梯度消失或梯度爆炸等挑战，并且因顺序处理每个时间步，使得模型训练速度极慢。虽然 LSTM[4]和 GRU[5]通过特殊的门控机制缓解了部分问题，但根本性的问题仍未解决。Transformer 架构通过其独特的注意力机制和并行计算能力革新了 NLP 领域。它摒弃了传统 RNN 的序列依赖性，能够并行处理序列的所有位置，不仅解决了梯度消失和爆炸问题，还显著提升了训练效率，使得更大规模的数据可以参与训练。这一革新性的架构极大地推动了大语言模型的发展。基于 Transformer 的 Bert 和 GPT 模型通过大规模预训练，ChatGPT（Chat Generative Pre-trained Transformer）[6]的发布标志着大语言模型的飞跃，展示了出色的自然语言理解和生成能力，引起了广泛关注和讨论。全球科技巨头纷纷推出自己的大型语言模型，以期在这一新兴技术领域占据领先地位。在这场科技风潮中，如何客观评估这些大语言模型的能力及其未来对教育公平的影响，成为亟需解决的问题。

高考作为我国最重要的人才选拔机制之一，是整个教育体系的核心组成部分[7]。自恢复以来，高考以其公平性和权威性，成为衡量学生学术水平和选拔优秀人才的重要手段。在生成式人工智能时代，我们开始探讨能否利用高考的评估机制来测试人工智能的认知和表达能力。通过将人类的认知测试方法应用于人工智能，客观题由于其结构化和答案明确

的特点，能够有效减少主观因素的影响，公正地评估人工智能在特定领域的知识掌握情况和逻辑推理能力[8]。相对地，主观题因其开放性和复杂性，是衡量人工智能逻辑思维和表达能力的理想方式。通过标准化评分，主观题可以揭示人工智能在不同学科背景下的表达上限和局限性[9]。

在此背景下，本文的研究不仅仅关注人工智能在自然语言处理和知识测试方面的技术能力，还试图通过将人工智能与高考机制相结合，探索如何利用这一新兴技术推动教育公平的实现。通过 Gaokao-bench 测评框架[10]，对不同人工智能模型在高考试题上的表现进行评估，不仅可以为模型优化提供参考，还能够从另一个角度审视当前教育体系的公平性。通过调用文心一言（ERNIE-BOT）、Gemini、Chat GPT 等十三个主流大型语言模型的 API，使这些模型对部分年份高考客观题部分进行分别作答，收集各模型面对不同题型的结果并对这些结果进行分类汇总。基于测评结果我们深入分析人工智能在解答高考全套试卷方面的能力和局限性，为今后大语言模型的优化方向给出一定的参考，并进一步探讨了人工智能对于教育领域以及更广泛的人类社会自身可能产生的影响。在快速发展的技术环境下，如何更好地将教育公平作为人工智能应用的重要目标之一，从而推动整个社会向更加公正和包容的方向发展。

2 大语言模型概述

2.1 大语言模型

大语言模型（Large Language Model，LLM）是一种基于深度学习的自然语言处理技术，通过大规模数据集的自我学习，具备理解和生成自然语言的能力。这些模型的参数规模通常从数百万到数百亿不等，使其能够捕捉语言中的复杂模式，并在多种任务上表现出色。

Transformer 架构是大语言模型的核心组件，是由谷歌研究团队于 2017 年提出如图 1 所示。它通过注意力机制有效捕捉输入序列中的长距离依赖关系，使模型在处理每个输入位置时能关注到其他位置的信息，并动态调整这些信息的重要性。Transformer 由编码器和解码器两部分组成，编码器通过多层自注意力机制和前馈神经网络提取输入序列的特征，生成包含丰富语义信息的隐藏表示，解码器利用这些隐藏表示生成目标序列，例如翻译后的句子或预测的下一个词。

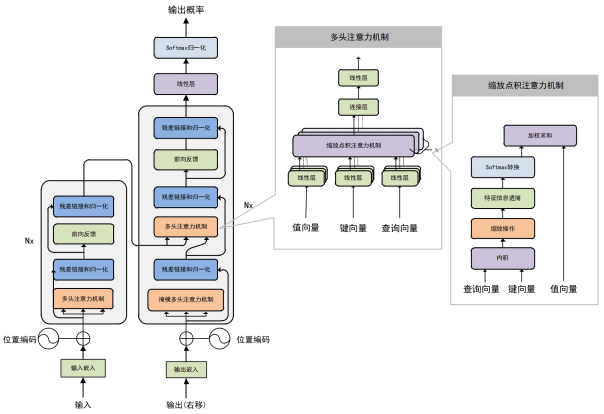

图 1　transformer 架构示意图

为了让模型理解输入序列的顺序，Transformer 引入了位置编码机制，通过将位置信息编码为向量并与词向量相加，使模型能够理解词汇的相对位置。位置编码通常采用正弦和余弦函数，保证了位置信息的周期性和相邻位置之间的平滑变化，增强了模型对序列结构的理解。在 Transformer 架构的基础上，大语言模型通过预训练模型机制进行大规模数据上的无监督或弱监督学习，获取丰富的语言知识。这种预训练奠定了语言理解的基础，使得在后续微调阶段能够以较少的数据和时间达到良好效果，从而提高了模型的泛化性能并加快了开发周期。预训练模型因其强大的语言理解能力和庞大的参数规模，被视为“超级模型”。

2.2 参与测试的大语言模型选取

GAOKAO-Bench 测试评估了 13 个主流大语言模型，包括 GPT-4 系列、Gemini-Pro[11]、ERNIE-Bot 系列[12]、Baichuan 系列[13]、ChatGLM 系列[14]、LLaMA[15]和 Vicuna[16]系列。这些模型在不同任务中的表现各有特点，适用于不同的应用场景。

GPT-4 系列在推理和复杂问题处理方面表现优异，尤其擅长多轮推理和高质量文本生成。它支持多语言任务，适合长文本生成、技术文档撰写等高精度应用。然而，GPT-4 资源需求高，部署成本较高，更适合企业级应用和研究。Gemini-Pro 和 ERNIE-Bot 在中文生成和理解方面具有优势，尤其适用于处理中国文化、历史和法律相关任务。

Gemini-Pro 通过预训练和知识蒸馏提升中文语境理解，ERNIE-Bot 则利用知识图谱增强语义理解。但它们的跨语言处理能力有限，推理速度较慢，不适合实时性要求高的场景。Baichuan 和 ChatGLM 系列在中文对话生成领域表现出色，适合中小型企业应用。Baichuan 注重快速响应和低资源需求，ChatGLM 则优化了模型的轻量化，适用于移动端。虽然它们在一般对话生成中表现良好，但在复杂推理任务中能力有限。LLaMA 和 Vicuna 系列以轻量化和灵活性著称，适合学术研究和快速迭代的开发环境，擅长多任务和小样本学习。然而，它们的文本生成质量和推理能力较弱，在复杂对话生成和深度语义分析上不及 GPT-4 和其他大型模型。

综上所述，各模型在任务复杂性、资源需求和语言支持方面各有优劣，选择时需综合考虑这些因素。

3 高考测试情境下的大语言模型表现

3.1 客观题结果分析

表 1　模型客观题性能对比表

模型	总体	语文	英语	数学理科	数学文科	物理	化学	生物	政治	历史	地理
GPT-4-0314	72.2%	53.9%	93.1%	53.7%	63.3%	55.5%	44.4%	80.7%	75.9%	75.6%	80.0%
GPT-4-0613	71.6%	52.1%	93.2%	54.5%	64.0%	50.8%	43.6%	83.0%	72.5%	74.2%	81.1%
Gemini-Pro	57.9%	46.7%	69.9%	40.7%	47.7%	32.0%	40.3%	70.7%	64.7%	64.5%	68.4%
ERNIE-Bot-0615	56.6%	46.7%	31.0%	38.3%	49.1%	35.9%	66.1%	79.3%	86.9%	79.1%	68.4%
GPT-3.5-turbo-0301	53.2%	34.7%	76.6%	38.8%	47.8%	41.1%	38.7%	56.9%	45.3%	53.9%	54.0%
ERNIE-Bot-turbo-0725	45.6%	35.3%	26.6%	34.1%	36.2%	32.0%	51.6%	64.0%	72.2%	63.4%	44.2%
Baichuan2-13b-Chat	43.9%	26.9%	34.7%	23.8%	31.7%	25.0%	40.3%	53.3%	75.3%	59.9%	61.1%
ChatGLM2-6b	42.7%	31.1%	30.6%	29.0%	35.8%	24.2%	46.0%	71.3%	55.0%	59.2%	41.1%
Baichuan2-7b-Chat	40.5%	31.7%	33.0%	26.6%	28.4%	18.0%	26.6%	48.0%	69.7%	57.8%	49.5%
ChatGLM-6b	30.8%	18.6%	17.0%	25.2%	25.7%	12.5%	30.6%	24.7%	54.1%	59.9%	25.3%
Baichuan2-7b-Base	27.2%	16.2%	21.2%	24.8%	24.8%	0.0%	23.4%	24.0%	55.3%	32.1%	24.2%
LLaMA-7b	21.1%	16.2%	20.5%	24.3%	26.1%	0.0%	22.6%	22.7%	22.2%	19.2%	24.2%
Vicuna-7b	21.0%	12.0%	19.6%	23.8%	23.4%	7.0%	27.4%	20.0%	20.9%	23.0%	23.2%

在高考测试情境下，各大语言模型在智能评分中的表现差异显著如表 1 和图 2 所示，显示出不同模型在各学科任务中的优势与不足。例如：GPT-4 系列在英语科目中的得分超过 93%，展现了强大的语言理解和生成能力，但在中文科目中仅得分 53%-54%，反映出其适应不同语言的差异。相比之下，ERNIE-Bot 系列在中文科目中的得分超过 80%，但在英语科目中的得分仅为 31%，显露出其跨语言处理能力的局限。

从文理科的角度上来讲，在理科科目中，GPT-4 在物理（44%-45%）和化学（44%-45%）中表现不佳，表明其在逻辑推理和精确计算方面存在挑战。然而，ERNIE-Bot 在化学科目中得分高达 66.1%。其他模型如 ChatGLM 和 Baichuan 在物理和数学科目中的得分普遍低于 40%，表现较弱。在文科科目中，GPT-4 在历史和地理科目中的得分高达 80%-81%，显示对广泛知识的处理能力。但在政治科目中，ERNIE-Bot-0615 以 86.9%超越了 GPT-4。LLaMA 和 Vicuna 在这些科目中的得分较低，特别是在历史和地理科目中得分低于 60%。

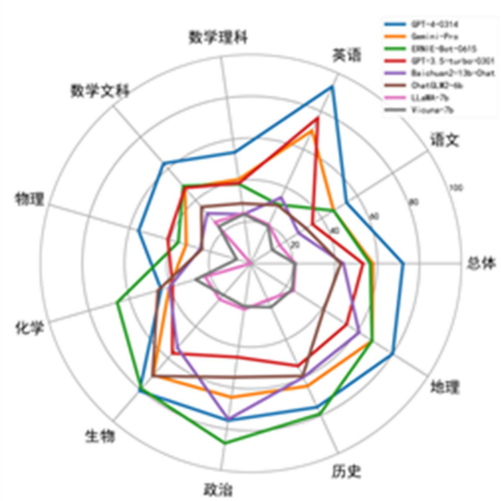

图 2　各科目得分率雷达图

从文理科的角度上来讲，在理科科目中，GPT-4 在物理（44%-45%）和化学（44%-45%）中表现不佳，表明其在逻辑推理和精确计算方面存在挑战。然而，ERNIE-Bot 在化学科目中得分高达 66.1%。其他模型如 ChatGLM 和 Baichuan 在物理和数学科目中的得分普遍低于 40%，表现较弱。在文科科目中，GPT-4 在历史和地理科目中的得分高达 80%-81%，显示对广泛知识的处理能力。但在政治科目中，ERNIE-Bot -0615 以 86.9%超越了 GPT-4。LLaMA 和 Vicuna 在这些科目中的得分较低，特别是在历史和地理科目中得分低于 60%。

综合来看，这些数据揭示了各模型在智能评分中的具体表现差异，反映了训练数据多样性、模型架构和算法设计对学科知识处理能力的影响。GPT-4 在语言类任务中表现出色，但需提升逻辑推理和科学知识处理能力。ERNIE-Bot 在中文和化学科目中表现突出，但在跨语言和逻辑推理任务上仍有不足。这些差异不仅提供了当前大型语言模型智能评分性能的详细画像，还为未来模型改进提供了明确的方向。随着技术的不断进步和训练方法的优化，特别是在跨语言理解、逻辑推理和科学知识处理方面的改进，这些模型的智能评分性能有望在未来得到显著提升。

3.2 主观题结果分析

表 2　模型主观题性能对比表

模型	总体	语文	英语	数学理科	数学文科	物理	化学	生物	政治	历史	地理
GPT-4-0314	51.9%	51.5%	88.3%	24.1%	27.9%	56.7%	35.0%	85.6%	50.0%	63.1%	70.0%
GPT-4-0613	50.8%	50.3%	87.6%	24.6%	27.5%	47.1%	28.5%	85.6%	49.9%	59.9%	71.5%
ERNIE-Bot-0615	48.4%	57.1%	45.0%	17.0%	25.6%	33.5%	30.8%	84.9%	53.0%	60.0%	72.7%
ERNIE-Bot-turbo-0725	39.2%	42.5%	28.8%	14.6%	15.6%	23.2%	25.0%	85.1%	45.3%	47.0%	61.8%
GPT-3.5-turbo-0301	35.8%	33.9%	75.4%	15.2%	15.9%	16.9%	21.4%	36.3%	42.3%	58.4%	62.1%

在高考测试情境下，对多个大型语言模型在主观题智能评分中的表现进行对比，如表 2 所示，清晰地表示了它们之间的性能差异。

GPT-4 系列模型在综合评估中表现突出，尤其在英语语言理解与生成上展现出强大实力，但理科科目如科学数学和物理的得分相对较低，显示出逻辑推理和精确计算的不足。相比之下，ERNIE-Bot 系列在中文和化学科目中表现优异，但在英语和理科科目上则明显逊色，反映出跨语言处理和理科能力的局限。GPT-3.5-turbo-0301 模型整体表现较为平庸，特别是在理科领域挑战较大。这些差异凸显了不同模型在处理知识和逻辑任务时的独特优劣势，根源在于其训练数据、架构设计以及处理复杂任务能力的差异。GPT-4 擅长语言，ERNIE-Bot 在特定领域（中文、化学）表现出色，而 GPT-3.5-turbo 则显得较为均衡但无显著亮点。

为提升这些模型在智能评分及更广泛领域的应用表现，需针对各自不足进行针对性改进。加强理科数据的训练，优化跨语言处理能力，以及提升处理复杂逻辑和科学知识的能力，将是关键的发展方向。通过不断迭代与优化，这些大语言模型有望在教育公平方面发挥更加重要的作用。

为了更深刻地阐释大语言模型在主观题解答中的实际应用效能，我们特别选取了 2024 年高考语文全国甲卷中的第 21 题语言文字运用 II 部分作为分析样本。通过对两个最

新版本的大语言模型在此主观题评测中的答案进行分析。

表 3　简答题示例

题目	参考答案
下面的文字是一位老奶奶在医院看病时的自述，不够简明扼要，不利于和医生高效沟通。请对这段自述进行缩写，要求：保留必要信息，不超过 80 个字 大夫好!今天看病的人太多了，我排了好长时间队才看上。我是你们医院的老病号了，这么多年我的高血压和糖尿病一直是在你们医院看的，好多年前有一次扭伤了脚踝，也是在你们这儿看好的，您可得给我好好看看。是这么回事儿。昨天晚上我老因女来家里，我们一起吃的晚饭。吃过饭看着电视，我就开始头疼，先是头顶一圈疼，一跳一跳的，后来整个头都疼。我试了很多办法，一会儿躺着，一会儿坐着，大口喘气，戴上帽子捂着，都没有用。闺女要带我来医院，我说天太冷了，明天可能就好了，明天再说吧，然后就睡觉了。今天早上醒了还疼，头也不敢动，一晃就更疼了，就赶紧来医院了。	大夫好!我是你们医院的老病号，一直在这儿看高血压和糖尿病。昨天晚上吃完饭后开始头疼先是头顶一圈疼，后来整个头都疼。今天早上醒来仍然头疼，头一动就更疼，所以赶紧来医院了。

GPT-4o　　ERNIE-Bot 4.0

图 3　GPT-4o 和 ERNIE-Bot 4.0 的对比

在分析两个大型语言模型的回答前，首先注意到它们在字数限制方面的表现，如图 3 中显示，GPT 模型回答了 85 字，文心一言则为 75 字，后者在字数控制上更符合要求。接下来，我们分析这两个模型的具体表现，包括删除冗余、信息提炼和内容串联三个方面。首先，在删除冗余信息时，两个模型删除了部分不必要的内容，如排队时间长和寒暄，但都保留了一些无关信息，表明它们在识别冗余信息方面仍有不足。在信息提炼方面，两个模型展现出强大的能力，能从冗余文本中提炼出关键信息，如患者身份、既往病史、症状描述和就诊原因。在语言组织上，两个模型均能按时间顺序和逻辑关系整理信息，使得表述简洁明了。特别是文心一言在字数控制上表现突出，能在保持语义完整的同时符合字数要求。

总体来看，两个模型在信息提炼和组织上表现优异，但在去除冗余信息和个性化表达方面仍有提升空间。文心一言在字数控制上的表现尤为出色。

4 结果分析与展望

在当前教育资源分配不均的背景下，大语言模型的应用具有显著的潜力来缓解这一问题。现阶段，教育资源的不均衡表现为优质教育资源的集中和贫乏地区教育支持的不足，这直接导致了不同区域和群体在教育机会上的不平等。高考作为一种重要的选拔机制，其评卷过程的公平性对教育公平有着至关重要的影响。然而，人工评卷受限于评卷人员的主观因素和工作疲劳，可能造成评分的偏差，从而进一步加剧了教育资源分配的不公平。

目前大语言模型在文科科目中表现较优，而在理科科目如数学和化学上存在局限，暴露出模型在处理逻辑推理和公式应用上的不足。这种差距使得不同学科的评估标准难以统一，也使得模型在高考评测中的应用受到制约。然而，随着技术的发展和模型的不断优化，特别是在理科评测能力的增强方面，大语言模型有望在未来更好地解决这一问题。

在未来大语言模型的成熟应用将能够提供一个标准化的评测过程，通过统一的评测标准和自动化评分，减少因人类评卷员主观判断引发的评分差异。这不仅提高了评卷效率，还确保了所有考生无论地域、学校背景，甚至社会经济地位，都能在评测过程中享有平等的机会，从而减轻了由于教育资源分配不均带来的不公平影响。

随着大语言模型的能力提升，它将能够更准确地评估各类学科的试题，包括目前表现较弱的理科题目。这种能力的扩展将进一步平衡各学科之间的评测标准，减少由于学科差异导致的评分偏差，从而构建一个更为公正和均衡的教育评

估体系。通过优化大语言模型的算法和训练数据集，进一步提高其在高考评测中的应用效果，有助于减少教育资源分配不均对考试公平的影响。最终，这不仅能够提升整体教育质量，还将促进教育机会的公平分配。

5 结束语

本研究通过高考全卷测试，对多个大型语言模型的表现进行了全面分析。尽管当前的大语言模型在语言理解和生成任务中已经展现出优异的表现，但在面对复杂推理任务和科学知识的应用时仍存在不足。这些不足限制了其在高考评测中的全面应用，特别是在需要深度理解和多步骤推理的学科领域。未来研究应提升大语言模型的推理能力和科学知识的整合应用，以应对高考中各类复杂题型的评测需求。此外，引入更丰富和多样化的领域知识，优化训练数据集，将进一步增强模型的准确性和公平性。通过减少人工评卷过程中可能出现的主观偏差，成熟的大语言模型有望在高考评测中实现更为客观、公正的评分，推动教育公平的实现。

随着大语言模型的不断优化和应用扩展，它将为智能教育的发展奠定坚实的技术基础，助力实现更广泛的教育公平和社会资源的优化配置。这不仅将提升整体教育质量，也将为社会福祉的提升做出贡献。

参考文献

[1]郑世林，姚守宇，王春峰. ChatGPT 新一代人工智能技术发展的经济和社会影响[J]. 产业经济评论，2023（3）：5-21.

[2]Vaswani A，Shazeer N，Parmar N，et al. Attention is all you need[C]. Proceedings of the 31st International Conference on Neural Information Processing Systems. New York：Curran Associates Inc.，2017：6000-6010.

[3]Elman J L. Finding structure in time[J]. Cognitive Science，1990，14（2）：179-211.

[4]Hochreiter S，Schmidhuber J.Long short-term memory[J].Neural Computation，1997，9（8）：1735-1780.

[5]Cho K，Merrienboer B，Gulcehre C，et al.Learning phrase representations using RNN encoder-decoder for statistical machine translation[C].Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing.Doha：ACL，2014：1724-1734.

[6]Brown T B，Mann B，Ryder N，et al. Language models are few-shot learners[J]. Arxiv Preprint Arxiv：2005.14165，2020：1-75.

[7]王烽. 面向教育现代化的高考改革设计：不忘初心完善格局[J]. 中小学管理，2017（7）：25-28.

[8]游洋. 几道高考客观题难题的解法[J]. 考试(高考数学版)，2011（5）：10-12.

[9]林文真. 新高考“信息性阅读”主观题教学方法研究[J]. 课堂内外（高中版），2023（7）：4-6.

[10]Zhang X，Li C，Zong Y，et al.Evaluating the performance of large language models on GAOKAO benchmark[J].Arxiv Preprint Arxiv：2305.12474，2023：1-4.

[11]Team G，Anil R，Borgeaud S，et al. Gemini：A family of highly capable multimodal models[J]. Arxiv Preprint Arxiv：2312.11805，2023：1-62.

[12]Sun Y，Wang S，Li Y，et al. Ernie：Enhanced representation through knowledge integration[J]. Arxiv Preprint Arxiv：1904.09223，2019：1-8.

[13]Yang A，Xiao B，Wang B，et al. Baichuan 2：Open large-scale language models[J]. Arxiv Preprint Arxiv：2309.10305，2023：1-28.

[14]Team G L M，Zeng A，Xu B，et al. ChatGLM：A family of large language models from GLM-130B to GLM-4 all tools[J]. Arxiv Preprint Arxiv：2406.12793，2024：1-19.

[15]Touvron H，Lavril T，Izacard G，et al. LLaMA:Open and efficient foundation language models[J]. Arxiv Preprint Arxiv：2302.13971，2023：1-27.

[16]Chiang W L，Li Z，Lin Z，et al. Vicuna：An open-source chatbot impressing GPT-4 with 90%* ChatGPT quality[EB/OL]. https：//lmsys.org/blog/2023-03-30-vicuna/，2023-03-30/2024-08-30.