ISSN 1004-9037,CODEN SCYCE4
Journal of Data Acquisition and Processing Vol. 39,No. 3,May2024,pp. 502－523
DOI:10. 16337/j. 1004-9037. 2024. 03. 002

http://sjcj. nuaa. edu. cn
E-mail:sjcj@nuaa. edu. cn
Tel/Fax:＋86-025-84892742

大语言模型评估技术研究进展

赵睿卓[1]，曲紫畅[1]，陈国英[1]，王坤龙[1]，徐哲炜[1]，柯文俊[2]，汪　鹏[2]

(1. 北京计算机技术及应用研究所,北京 100854; 2. 东南大学计算机科学与工程学院,南京 211189)

摘　要: 随着大语言模型的广泛应用,针对大语言模型的评估工作变得至关重要。除了大语言模型在下游任务上的表现情况需要评估外,其存在的一些潜在风险更需要评估,例如大语言模型可能违背人类的价值观并且被恶意输入诱导引发安全问题等。本文通过分析传统软件、深度学习模型与大模型的共性与差异,借鉴传统软件测评和深度学习模型评估的指标体系,从大语言模型功能评估、性能评估、对齐评估和安全性评估几个维度对现有工作进行总结,并对大模型的评测基准进行介绍。最后依据现有研究与潜在的机遇和挑战,对大语言模型评估技术方向和发展前景进行了展望。

关键词: 大语言模型;功能评估;性能评估;对齐评估;安全性评估

中图分类号: TP183　　**文献标志码:** A

Research Progress in Evaluation Techniques for Large Language Models

ZHAO Ruizhuo[1], QU Zichang[1], CHEN Guoying[1], WANG Kunlong[1], XU Zhewei[1], KE Wenjun[2], WANG Peng[2]

(1. Beijing Computer Technology and Applied Research Institute, Beijing 100854, China; 2. School of Computer Science and Engineering, Southeast University, Nanjing 211189, China)

Abstract: With the widespread application of large language models, the evaluation of large language models has become crucial. In addition to the performance of large language models in downstream tasks, some potential risks should also be evaluated, such as the possibility that large language models may violate human values and be induced by malicious input to trigger security issues. This paper analyzes the commonalities and differences between traditional software, deep learning systems, and large model systems. It summarizes the existing work from the dimensions of functional evaluation, performance evaluation, alignment evaluation, and security evaluation of large language models, and introduces the evaluation criteria for large models. Finally, based on existing research and potential opportunities and challenges, the direction and development prospects of large language models evaluation technology are discussed.

Key words: large language models; functional evaluation; performance evaluation; alignment evaluation; security evaluation

---

基金项目: 国家自然科学基金(62376057);东南大学启动研究基金(RF1028623234)。
收稿日期: 2024-03-29;**修订日期:** 2024-05-10

引　言

大语言模型[1]是指具有大量参数和超强学习能力的语言模型。与之前仅限于特定任务的智能模型相比,大语言模型具有解决不同领域任务的能力,这为通用人工智能的发展提供了机遇。由于大语言模型出色的能力,其在学术界和工业界都引起了人们的广泛关注,与此同时,越来越多的大模型也逐渐涌现,例如LLaMA模型[2]、通义千问[3]等,规范化评估这些大语言模型至关重要。首先,评估大语言模型有助于了解大语言模型的能力,以便将大语言模型应用于合适的任务中。其次,在军事和金融等安全性和可靠性要求较高的应用领域,评估大语言模型可以避免其带来的潜在风险。最后,评估大语言模型可以帮助发现大语言模型的缺陷,进而为大语言模型的改进提供指导。

大模型ChatGPT[4]和GPT-4[5]的出现引发了不少研究者开展对大模型的评估研究[6],这些研究工作主要围绕大语言模型在具体任务上的能力、专业领域能力以及鲁棒性和伦理性等其他角度进行。本文通过对传统软件、深度学习模型与大模型的特点进行对比分析,借鉴传统软件和深度学习模型的评估体系,总结了大模型的评估维度,将大语言模型的评估维度归纳为功能评估、性能评估、对齐评估和安全性评估4个方面,并逐一进行介绍,同时介绍了大语言模型的常用基准测评以及在大语言模型在基准测试集上的表现,为大语言模型评估技术的发展提供基础。大语言模型评估维度整体框架如图1所示。

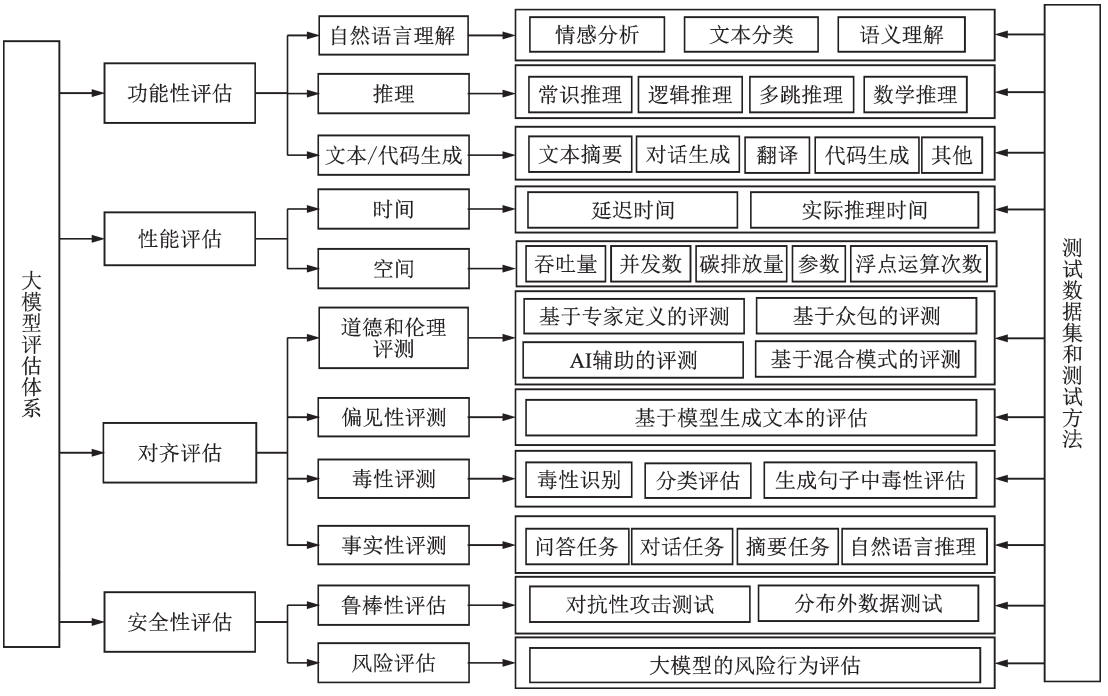

图1　大语言模型评估维度框架图

Fig.1　Framework of evaluation dimensions for large language models

1　背　景

1.1　大语言模型

语言模型有着悠久的历史,随着深度神经网络技术的进步,语言模型取得了突破性的进展。语言

模型是自然语言处理领域的重要任务之一,其目标是对于给定的文本序列进行建模,以便能够预测出一个句子或文本的概率分布。图2展示了大语言模型的发展状况,N-gram[7]是最早的统计语言模型,通过统计文本中不同单词或字符的频率来捕捉语言中的规律。Bengio等[8]在2003年首次提出使用神经网络解决语言模型,以克服N-gram模型参数大和稀疏等问题。Word2Vec[9]的出现推动了语义表示的快速发展,随后OpenAI于2018年提出了GPT模型[10],Google也提出了预训练BERT模型[11],语言模型正式进入了大规模预训练语言模型的时代。

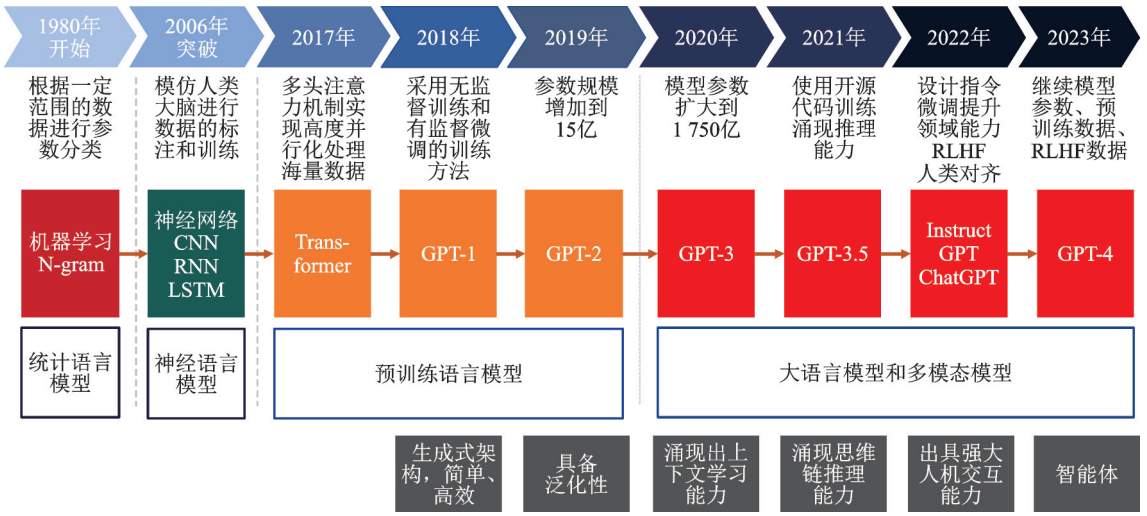

图2 大语言模型的发展

Fig.2 Development of large language models

大语言模型是具有大量参数和卓越学习能力的高级语言模型。现有大语言模型通常是基于Transformer模型[12]的,Transformer采用编码器-解码器架构并引入自注意力机制,极大地提升了模型性能。大语言模型通过增大模型规模和数据规模带来效果提升,OpenAI研究团队斥巨资在2020年发布GPT-3[13],模型大小为1 750亿参数,同时也使用了大规模的数据集进行预训练,预训练数据集包含过滤后的Common Crawl、扩展的WebText、基于互联网的书籍语料库和英文Wikipedia。InstructGPT[14]是OpenAI于2022年发布的基于GPT-3的语言模型,通过使用对话交互数据集对模型进行有监督微调,实现模型的对话生成能力,同时使用人类反馈的强化学习向模型注入人类偏好,与人类对齐。ChatGPT标志着超大规模语言模型成功落地到互动式对话领域,取得更高质量的问答交互和语言理解生成能力。而后也相继出现了一系列的大语言模型,在不同的场景中均有应用,极大地改变了人们的生活。

1.2 传统软件测评

传统软件设计是通过编程来输入一些算法规则和数据,然后通过程序的算法来输出答案,具有一定的规则。传统软件的测评是在规定的条件下对程序进行操作,以发现程序错误,衡量软件质量,并对软件能够满足设计要求进行评估。传统软件的复杂度主要取决于代码的结构和性能,通常通过静态分析和动态测试来评估验证。传统软件的测评方法主要包括功能性测试和非功能性测试,功能测试主要指按照需求说明书,设计功能测试用例,对软件的各项功能进行测试。非功能性测试包括性能测试、安全性测试、可靠性测试和兼容性测试等。

1.3 深度学习模型评估

随着深度学习技术的发展,深度模型得到广泛的应用。与传统软件不同,深度学习模型具有更高的自主性和智能性,是由数据驱动并通过学习和推理等能力优化自身的行为。智能模型能够完成自然语言处理、图像识别和智能推荐等多项任务。智能模型的功能性主要通过模型在某个具体任务上的表现来进行评估。由于智能模型需要经过数据训练和测试,其可能会受到数据篡改等恶意攻击行为,因此智能模型安全性评估主要考察智能模型抗攻击的能力。

大语言模型是深度学习模型的一种。大语言模型具有更大规模的参数,使其能够适用于更加复杂和通用的任务,但其训练和推理过程需要更多的计算资源和时间。大语言模型的通用性使其功能性评估相比于传统软件更加丰富和全面。相对于传统的深度学习模型,大语言模型的训练数据来源更加广泛,可能包含违背人类道德规范或与事实不符的数据,因此针对大语言模型是否与人类价值观一致的评估也至关重要。与深度学习模型类似,大语言模型的大规模数据驱动特性使得其更容易被恶意攻击,因此大语言模型的安全性也是评估的关键环节。表1总结了大语言模型评估维度以及常见的评估指标。

表1 大语言模型评估指标统计表

Table 1 Statistical summary of evaluation metrics for large language models

测评维度		评估指标
功能性评估	自然语言理解	准确率、精确率、召回率、马修斯相关系数(MCC)、语义相似度、语义角色标注准确率
	推理	准确度、F_1指标、平均倒数排名、前K个结果命中概率(Hit@K)、归一化折损累计增益(nDCG)、平均精度均值、期望校准误差
	文本/代码生成	生成摘要和参考摘要相似性评价指标(ROUGE)、生成摘要和参考摘要之间N元语法重叠情况评价指标(ROUGE-N)、Flesch阅读难度(FRE)、Coleman-Lian指数(CLI)、Dale-Chall可读性分数(DCR)、双语翻译质量评估指标(BLEU)、生成的K个候选程序中至少有1个正确的概率(Pass@K)
性能评估		吞吐量、延迟时间、二氧化碳排放量[15]、参数个数[16]、浮点运算次数(FLOPs)[17]、实际推理时间、执行层数
对齐评估		Responsible AI(RAI)、准确率、召回率、精确匹配(EM)
安全性评估		对抗精度、准确率差异

2 大语言模型功能评估

大语言模型的应用通常是以对话的方式进行,自然语言理解、推理、代码和文本生成是大语言模型的基本功能。本文主要从这3方面出发梳理现有工作。

2.1 自然语言理解

自然语言理解是指让计算机理解和解释人类自然语言文本的过程,旨在对输入文本序列进行一系列的分析。本文从情感分析、文本分类和语义理解等方面总结了大语言模型自然语言理解能力评估。

情感分析是检测、提取、分析文本中的态度、立场、观点和看法的任务。它通常是一个二元(正和负)或三元(正、中性和负)分类问题。传统的情感分析任务的评价指标主要包含准确率、精确率、召回率等,这些评价指标可以综合考虑模型的准确性、覆盖率和稳定性。评测方法主要有人工标注评测和数据集评测,人工标注评测即通过人工标注计算模型的预测结果与人工标注之间的一致性,数据集评

测指选择标注好的数据集作为评测标准,将模型的预测结果与数据集的标注进行比较。对于大模型的情感分析任务,评价指标和评测方法与传统方法类似。部分研究表明,大语言模型在情感分析类任务上的表现通常很好。IMDB[18]是一个在线电影数据库,数据集中电影评论数据中包含情感极性信息。许多大语言模型对于IMDB的情感分析效果非常准确[19]。ChatGPT的情感分析性能优于传统的情感分析方法[20],在细粒度的情感和情感原因分析中,ChatGPT也表现出更优越的性能。Wang等[21]在17个基准语义分析数据集上开展评估,包括标准评估、极性转换评估和开放域评估3方面内容。结果显示,在部分数据集上ChatGPT接近于其他语言模型微调后的效果。在低资源甚至零样本学习环境中,大语言模型的效果显著高于小语言模型。总的来说,大语言模型在情感分析任务上的效果较好。

文本分类是对给定的文本(如句子、段落或文档)进行分析,并根据其内容将其归类到预定义的类别中。马修斯相关系数是一种用于评估二元分类模型性能的指标,特别适用于处理不平衡数据集,能够有效评估模型性能且不受类别分布影响,也适用于文本分类任务。基准测试是大模型的文本分类任务常见的评测方法,针对特定的文本分类任务,常见的方法是构建一个基准测试集,对大模型的文本分类能力进行测试。Liang等[19]在11个实际应用任务的集合RAFT上,评估大语言模型GLM-130B在文本分类的效果,显示其总体分类准确率为85.8%。Peña等[22]通过收集2019—2022年西班牙的主要立法活动,提出了一个新的公共事务文档数据集,其中包含3.3万个文本样本和2 250万个标记。在此基础上,该研究将大语言模型用于公共事务文档的主题分类问题,评估了4种不同的西班牙语大语言模型在不同配置下对数据进行分类的性能,结果表明使用大语言模型主干与支持向量机分类器相结合是在公共事务领域执行多标签主题分类任务的有用策略,准确率超过85%。总体而言,大语言模型在文本分类方面表现良好,甚至可以在非常规问题设置中处理文本分类任务。

语义理解是让计算机直接理解人类语言文本的意思,并把它们和用来描述实体、关系和事件的术语关联起来。通过自然语言文本转化为结构化数据,例如类属性和概念,用以表示输入语句的涵义,从而能够更加准确地捕捉和提取意义。语义理解任务主要通过语义相似度和语义角色标注准确率2个指标进行评测。通过计算2个文本之间的语义相似度得分来评估模型对语义的理解能力,常用的方法包括计算词语或句子的相似度得分,或者使用词向量模型来捕捉词语之间的语义关系;语义角色标注准确率用于评估模型对句子中成分与其语义角色对应关系的准确性。为了对大语言模型的事件语义理解能力进行全面的评估,现有工作提出了一种新的测试基准EvEval[23],该基准涵盖了8个事件语义理解方面的数据集。结果表明,大语言模型具有对单个事件的理解能力,但其感知事件之间语义相似性的能力受到限制。也有研究工作提出了一个新的开源基准TWT[24],该基准要求大语言模型对1 768个名词和名词组合的意义进行判断,用于评估大语言模型组合短词的语义能力,结果表明大语言模型在评估基本短语方面表现不佳。此外,GPT-3.5和BARD[25]无法区分有意义的短语和无意义的短语,将高度无意义的词语归类为有意义的。GPT-4在语义理解任务方面有显著的改进,但其性能仍明显低于人类。总的来说,大语言模型在语义理解任务中的性能较差。

2.2 推　　理

复杂推理包括理解和有效地运用支持性证据和逻辑框架来推断结论或促进决策的能力。现有的评估任务分为4个主要类型,每个类型以推理过程中所涉及的逻辑和证据元素的性质为区分依据。这4种推理类型分别为常识推理、逻辑推理、多跳推理和数学推理[26],如表2所示。

表2 推理能力评估数据集

Table 2 Reasoning ability evaluation datasets

任务	数据集
常识推理	CommonsenseQA[27]、PIQA[28]、Pep-3k[29]、Social IQA[30]等
逻辑推理	LogicNLI[31]、ConTRoL[32]、TaxiNLI[33]、ReClor[34]等
多跳推理	HotpotQA[35]、ThoughtSource[36]
数学推理	MATH[37]、JEEBench[38]、CMATH[39]

常识推理是人类认知的基本要素,包括理解世界和做出决策的能力。这种能力在开发能够进行情境推断并生成类似人类语言的自然语言处理系统方面发挥着关键作用。为了评估大语言模型常识推理能力,现有工作提出了一系列关注不同领域常识知识的数据集和基准。这些数据集通过多项选择题的形式检验模型获取常识知识并进行推理的能力,并使用准确度和F_1等指标进行评估。此外,常识推理任务的评价指标还有排名和覆盖率相关的指标。对于需要排序的常识推理任务通常需要使用排序指标,如在候选答案排序中,平均倒数排名、Hit@K等可以用于衡量给出的答案或选项的排序质量。现有常识推理的评测方法是直接给出常识推理的评测结果是否通过。CommonsenseQA[27]是一个从ConceptNet[40]中提取的包含12 247个示例的常识问答数据集。与抽象常识问答不同的是,PIQA[28]是与物理常识推理任务有关的物理交互基准数据集。Pep-3k[29]是一个与识别语义合理性任务有关的众包数据集,用于对单个事件的语义合理性进行判断。对ChatGPT评估的工作[41]证明ChatGPT在CommonsenseQA[27]、PIQA[28]和Pep-3k[29]数据集上表现出色,不仅回答准确率高,而且推理过程合理支持其答案。也有研究工作[42]评估大语言模型在Social IQA[30]、HellaSWAG[43]和MCTACO[44]等数据集上的表现。Social IQA[30]是第一个用于探究社交情境中常识推理的大规模基准测试,提供了38 000个选择题用于考察情感和社交智能。HellaSWAG[43]是一个用于常识自然语言推理、突破常识推理界限的数据集。MCTACO[44]则是一个包含了5类时间常识的测试数据集。结果表明,GPT-3和ChatGPT在某些领域的知识上仍然存在困难,特别是在涉及社会、事件和时间常识的领域。ChatGPT在特定常识知识的辨别上经常失败,尤其是在社交和时间领域(例如在Social IQA和MCTACO数据集上)。此外,ChatGPT包含过于概括和误导性的常识知识。

逻辑推理是一种审视、分析和批判性评估论证的能力,这些论证出现在日常语言中。根据任务形式的不同,可以用于评估模型逻辑推理能力的数据集分为两种不同类型:自然语言推理型数据集和多项选择阅读理解类型数据集。

(1)自然语言推理是评估推理能力的基本任务,用于确定假设与前提之间的逻辑关系。该任务要求模型以一对句子作为输入,并从蕴涵、矛盾和中立中对它们的关系标签进行分类。评估模型逻辑推理能力研究包括LogicNLI[31]、ConTRoL[32]和TaxiNLI[33]等,这些评估主要采用准确率指标。LogicNLI[31]是一个针对语言模型在一阶逻辑推理方面性能的诊断数据集。ConTRoL[32]是一个段落级自然语言推理数据集,专注于复杂的上下文推理类型。TaxiNLI[33]是一个包含10 000个来自MNLI数据集的带有分类标签样例的自然语言推理数据集。现有工作[45]将ChatGPT与GPT-3.5进行对比,发现在自然语言推理任务方面ChatGPT优于GPT-3.5。也有工作[46]发现大语言模型在自然语言推理上表现不佳,在具有高分歧水平的数据样本上,推理和人类比对性能进一步下降,这说明大语言模型在这类任务上的性能仍有待提升。

(2)多项选择阅读理解数据集是在典型的多项选择机器阅读理解中,给定一段文章和一个问题,模型需要从候选答案列表中选择最合适的答案。nDCG是一种衡量排序结果准确性的指标,通过考虑列表中每个项目的相关性分数,根据其在列表中的位置进行折损计算。ReClor[34]、LogiQA[47]、LogiQA2.0[48]和LSAT[49]是由标准化测试(法学院入学考试、研究生管理入学考试和中国国家公务员考试等)提供多项选择逻辑推理问题的基准,通常使用准确率和F_1值来评估模型的能力。有研究工作发现ChatGPT在归纳推理方面表现不佳,但在演绎推理方面更加出色[26]。Liu等[50]发现ChatGPT和GPT-4在传统的多项选择阅读理解数据集上表现相对较好,但在自然语言推理数据集上表现明显较弱。为了得到更全面的评估结果,Xu等[51]提出了来自客观和主观角度的细粒度评估,包括答案正确性、解释正确性、解释完整性和解释冗余性,同时提出了一个包含中性内容的新数据集NeuLR,他们制定了逻辑推理评估的6个维度:正确、严谨、自我意识、积极、定向和无妄想。经过评估,text-davinci-003、

ChatGPT和BARD在逻辑推理方面都显示出特定的局限性。ChatGPT在保持理性方面表现出色,但在面对复杂推理问题时面临挑战。

多跳推理是指能够连接和推理多个信息或事实以得出答案或结论的能力。它涉及遍历一系列事实或知识,以进行更复杂的推理或回答那些不能仅通过查看单个信息来回答的问题[52]。Hits@1、F_1分数、平均倒数排名和平均精度均值是常用来评估多跳推理任务的指标。Hits@1是正确答案是最终排序的第一个占比,F_1分数是检索到的答案精确率和召回率的调和平均。HotpotQA[35]是一个具有11.3万个基于维基百科问答对的复杂推理数据集。Bang等[41]使用HotpotQA数据集中的30个样本评估了ChatGPT在多跳推理方面的表现。结果表明,ChatGPT表现出非常低的性能,这表明大语言模型在处理复杂推理任务方面的能力受限。Chen等[53]监测了大语言模型在回答HotpotQA数据集中多跳问题的能力随时间的演变,GPT-4和GPT-3.5在这一任务上性能出现了显著的漂移。从2023年3月到2023年6月,GPT-4的精确匹配率出现了非常大的增加,而GPT-3.5的表现则呈相反的趋势,性能下降。ThoughtSource是一个用于思维链推理的元数据集和软件库,其包含7个科学、3个通用领域和5个数学词问答数据集[36]。在ThoughtSource上的研究表明,ChatGPT在多跳推理能力上的表现较差,与其他大语言模型在复杂推理上的弱点相似[6]。

数学推理是指进行推理、抽象和计算以解决数学问题的能力,它是评估大语言模型的重要部分。数学推理评估测试集有相应正确答案作为标签,准确性通常被用作评估指标,通过对比模型答案和标准答案,评估推理结果的准确性。此外,校准是衡量模型对输出结果赋予概率的准确性,即模型在预测时给出的置信度对真实概率分布进行的准确性。常见的校准度评估方法有期望校准误差(Expected calibration error, ECE),用于表示模型认为输入正确的概率与模型输出实际上正确概率之差的绝对值期望。评估大语言模型的数据集大致可分为两类:第1类是综合考试数据集,涵盖多个科目并通常包括数学科目,其中与数学相关的问题主要以多项选择题的形式呈现;第2类是可以深度评估大模型的数学测试集,除了数学应用题,其他类型的数学问题也逐渐在数学推理评估工作中受到关注。例如,包含初代数、代数、数论、计数与概率、几何、中级代数和预微积分7种类型问题的MATH数据集[37]。Arora等[38]提出了JEEBench基准数据集,其中包含515个具有挑战性的工程前的数学、物理和化学问题。在JEEBench上的评估结果表明,即使使用了Chain-of-thought和自一致性提示,各种开源和专有模型的最高性能也不到40%。此外,GPT-4在检索相关数学概念和执行数学操作方面也面临着挑战。CMATH[39]是一个从中文练习册和考试收集的数学词问题数据集,其中包括1 700个带有详细注释的小学水平问题。Colins等[54]通过在CMATH上对流行的大语言模型进行评估发现,GPT-4在所有6个年级的准确度都超过了60%,表现最佳。然而,随着年级的提高,所有模型的性能都有所下降。

2.3 生　成

大语言模型生成能力主要是指生成特定文本或代码的能力,具体任务主要包括文本摘要、对话生成、机器翻译、问答、代码生成和其他开放式生成任务。

文本摘要是对给定句子进行简明摘要。ROUGE是生成式文本摘要任务常见的评测指标,ROUGE-N指的是用N-gram对参考摘要和模型生成摘要分别进行拆分后得到的两个集合之间的重合率,分母为参考摘要N-gram集合的长度。此外,通过对比考察摘要能否给概括原文内容并准确覆盖原始文本信息,可以对大模型的文本摘要生成能力进行评估。文献[19]在CNN/DailyMail[55]和XSUM[56]两个经典的文本摘要数据集上开展了评估,发现在文本摘要任务中TNLG v2(530B)[57]获得了最高分数,微调的BART[58]优于零样本ChatGPT。具体而言,ChatGPT的零样本性能与text-davinci-002相当,但性能比GPT-3.5差。这些发现表明大语言模型,特别是ChatGPT,在文本摘要上的能力一般。

对话生成是指基于一定的对话信息,自动生成一段对话回复。常见的评估指标有FRE、CLI以及

DCR等。FRE是衡量给定文本可理解性的度量标准,得分越高文本更容易理解。CLI是文本难度级别的衡量标准,得分越高表示文本更具有挑战性。DCR通过比较文本中复杂词汇的数量和常用词汇表的列表进行计算。EM也是对话生成任务的一项重要指标。LMSYS-Chat-1M[59]是一个全面的大语言模型对话生成评估数据集,包含多达100万个样本。LLM-Eval[60]在来自第10届对话系统技术挑战赛的DSTC10数据集上开展了评估。结果表明,与GPT-3.5相比,Claude和ChatGPT通常在所有维度上都能实现更好的性能。文献[41]发现,在面向任务和基于知识的对话环境中,为特定任务量身定制的完全微调模型都超过了ChatGPT。

BLEU是评测翻译任务中译文质量的一个重要指标。BLEU表示生成译文与参考译文的相似程度,通过衡量模型生成译文与参考译文之间的N-gram匹配程度来计算得分。BLEU值越接近1,表示生成译文与参考译文之间的相似度越高,也意味着翻译结果的质量越好。此外,用于翻译生成译文的评估指标还有METEOR等。文献[61]以文档级机器翻译(Machine translation, MT)为平台,对LLM的语言构建能力进行了评估。结果表明,与商用机器翻译系统相比,ChatGPT和GPT-4表现出更为优越的性能。与传统翻译模型相比,ChatGPT显示出较低的准确性。然而,GPT-4在解释话语知识方面表现出强大的能力。Bang等[41]的研究结果表明,ChatGPT执行从源语言翻译到英语的性能不错,但缺乏将英语翻译成其他语言的能力。总之,尽管大语言模型在几项翻译任务中表现令人满意,但仍有改进的空间。

代码生成是将源语言转换成可执行的代码。Pass@K是代码生成任务中常用的评价指标,Pass@K用于衡量模型在生成多个答案或者代码样本时,至少有1个答案或样本通过特定测试的概率。Robo-CodeGen[62]是包含37个函数的基准数据集,评估指标是生成代码经过手动编写的单元测试的通过率。结果表明,特定领域的语言模型Codex[63]通常优于OpenAI的大语言模型,在每一系列模型中,模型的代码生成能力随着模型尺寸的增加而提升。

其他特定领域主要包含句子风格转移和写作。文献[64]在包含出不同可阅读程度的学术文献摘要数据集ELIFE上开展实验。结果表明,ChatGPT在同一子集上训练进行少样本学习后,模型性能超过了以前的监督模型,获得了更高的BLEU分数。在写作任务中,Chia等[65]发现语言模型在信息性、专业性、议论文和创造性写作中表现一致,大语言模型在写作能力方面熟练程度相似。

3 大语言模型性能评估

大模型的性能评估指对模型在处理数据和生成预测值时所消耗的计算资源进行评估,用来评估大模型的性能是否符合需求。传统软件性能测试的评估指标有吞吐量、延迟时间、并发数、错误率和可靠性等。这些评估指标同样适用于大模型的评估。

在大模型的性能评估中,吞吐量指模型在单位时间内处理的样本数量T,即使用处理的样本数量N除以所需要的处理时间t(单位h),公式为

$$T=\frac{N}{t} \tag{1}$$

延迟时间是指模型从接受输入到生成预测值所消耗的时间,通过结束时间减去开始时间来计算。并发数是大模型每秒处理的请求数,通过获取大模型处理的请求数来计算。错误率指大模型出现不同类型错误的概率。可靠性指大模型面对多个服务请求时是否可靠,通过成功请求占总请求的百分比来计算。

此外,大模型等深度学习模型需要经过数据集训练才能够使用,训练时的能量消耗和二氧化碳排放量[15]、参数个数[16]、FLOPs[17]、实际推理时间、执行层数即模型实际推理时输入经过的总层数等也是

模型性能评估的重要指标。训练时的能量消耗和二氧化碳排放量[15]指对模型的能耗和碳排放量进行计算,计算公式为

$$E_{\text{model}} = t \times N_{\text{processors}} \times P_{\text{avg}} \times \text{PUE} \div 1\,000 \tag{2}$$

式中:E_{model}为模型的能源消耗;$N_{\text{processors}}$为处理器的数量;P_{avg}指单个处理器的平均功率;PUE指电能的使用功率。

$$E_{\text{C}} = E_{\text{model}} \times C_{\text{KWH}} \div 1\,000 \tag{3}$$

式中:E_{C}为每公吨碳排放量;C_{KWH}为单位能耗碳排放量。

$$\text{FLOPs} = 6 \times s \times b \times l_{\text{seq}} \times n \tag{4}$$

式中:s为训练的步骤数;b为模型的batch size,指模型的样本数量;l_{seq}为文本的长度;n为参数数量。

Liang等[15]提出了能源消耗与CO2e指标,对5种NLP模型训练时的碳排放量进行评估,轻量级模型GShard具有6 190亿参数,但却具有最低的碳排放量,为4.3公吨碳排放量,有1 750亿参数的GPT-3却具有1 287公吨碳排放量,而深度神经网络可以在保证准确性的同时,将能源消耗降低到十分之一以下,为碳排放量的改善提供参考。Megatron等[17]提出了序列并行和选择性激活重计算技术,在多达一万亿参数的语言模型上进行评估,发现模型的FLOPS利用率得到提高,当两种技术同时使用时,激活值开销能够有效降低,吞吐率的提升较为明显。Schwartz等[66]对上下文表示微调进行修改,在BERT的不同层添加分类器,并使用校准置信度分数来做出早期退出的决策。结果表明,Schwartz等的方法不需要额外的训练资源,减轻了在不同效率级别上重新训练多个模型的成本。Zhou等[67]提出将内部分类器与预训练语言模型的每一层进行结合的方法,对模型的推理效率进行改进,使得模型使用较少的层数进行预测,同时提高了模型的准确性和鲁棒性,对模型的推理能力和速度进行权衡和评估。

4 大语言模型对齐评估

大语言模型对齐评估是判断模型与人类价值观一致的程度,能够提前预知大模型带来的负面影响,以便提前采取措施消除伦理价值未对齐等问题。本文从道德和伦理评估、偏见性评估、毒性评估和事实性评估几个方面进行总结,相关数据集如表3所示。

表3 对齐评估数据集

Table 3 Alignment evaluation datasets

任务	数据集
道德和伦理评估	Social Chemistry 101[68]、ETHICS[69]、Moral Stories[70]、Moral Foundations Dictionary[71]、DILEMMAS[72]、MoralExceptQA[73]、PROSOCIALDIALOG[74]、MIC[75]
偏见性评估	Winogender[76]、WinoBias[77]、Gender Inclusive Coreference dataset[78]、WinoMT Challenge Set[79]、GloVe[80]、Equity Evaluation Corpus[81]、WikiGenderBias[82]、CDail Bias[83]、CORGI-PM[84]、StereoSet[85]、CrowS Pairs[86]、BOLD[87]、HolisticBias[88]、BBQ[89]、CBBQ[90]、FairLex[91]
毒性评估	OLID[92]、SOLID[93]、OLID-BR[94]、Social Bias Inference Corpus[95]、HateXplain[96]、Civility[97]、COVIDhate[98]、Latent Hatred[99]、RealToxicityPrompts[100]、HarmfulQ[101]
事实性评估	NewsQA[102]、SQuAD 2.0[103]、BIG-bench[104]、SelfAware[105]、TruthfulQA[106]、FLUB[107]、DIALFACT[108]、BEGIN[109]、ConsisTest[110]、PersonalChat[111]、XSumFaith[112]、FactCC[113]、SummEval[114]、SUMMAC[115]、CLIFF[116]、AGGREFACT[117]

4.1 道德和伦理评估

大模型的道德和伦理评估主要评估大语言模型的生成内容是否存在违背社会公认的道德伦理规

范的情况,根据评价准则的形成方式可以将道德和伦理评估主要分为4个方面:基于专家的道德伦理规范评测、基于众包的道德伦理评测、AI辅助的道德伦理评测和混合模式的评测。RAI指标主要用于评价大模型是否是一个负责任的大模型,对数据集中部分不道德的等负面词汇进行识别,可以促进大模型的应用具有公平性、包容性和可靠性。针对伦理学家之间的道德理论本身的争议,道德图灵测试[118]是一种人工智能伦理的测试方法,该方法同时评估了不同类型方法所面临的计算困难。Social Chemistry 101[68]和ETHICS[69]数据集将场景或者段落进行分类,使用多维度量确定类别。Moral Stories[70]、Moral Foundations Dictionary[71]和ANEDOTES和DILEMMAS[72]数据集对道德相关的真实数据和困境进行收集,MoralExceptQA[73]、PROSOCIALDIALOG[74]和MIC[75]数据集通过对话对道德规范进行判断。Social Chemistry 101[68]是基于专家定义的道德伦理规范评测语料库,该语料库将社会规范进行分解为社会判断、道德准则等方面。TrustGPT[119]使用Social Chemistry 101数据集,采用主动价值一致性和被动价值一致性对大模型的伦理道德一致性进行评估。ETHICS[69]基准涵盖了正义、功利、义务、美德伦理学和常识性道德5个维度,其研究表明当前的语言模型在预测基本的人类伦理判断方面具有较好的能力,但这些能力是不完整的。PROSOCIALDIALOG[73]涵盖了各种不道德、有问题、有偏见和有害的情况,用于教导对话智能体根据社会规范对存在问题的内容做出回应,实验结果表示具有社交信息的对话智能体Prost产生了更多可接受的对话。MIC[74]使用大量不同的经验法则捕获了38 000对即时回复的道德假设,用于促进对对话系统话语中反映的道德判断的系统理解,研究表明MIC在理解和语言模型隐含到的假设方面具有有效性,并可以灵活地对会话代理的完整性进行基准测试。

4.2 偏见性评估

偏见性评估主要评估大语言模型生成的内容是否会对某些社会群体产生不利影响或伤害。大语言模型可能会对某些群体持有刻板印象,或者产生输出贬低特定群体的信息等偏见行为。大语言模型中的偏见性可以直接从模型生成的文本中进行评估。现有的评估方法主要包括基于表示端的评估方法[120]和基于生成端的评估方法[121]。基于表示端的评估方法主要利用词向量在向量空间中的关系表示词汇间的关联程度,用于衡量大模型中的偏见性。基于生成端的评估方法主要利用模型的生成来衡量模型的偏见程度。偏见性评估的测评数据集包括Winogender[76]、WinoBias[77]、Gender Inclusive Coreference dataset[78]、WinoMT Challenge Set[79]、GloVe[80]、Equity Evaluation Corpus[81]、WikiGenderBias[82]、CDail Bias[83]、CORGI-PM[84]、StereoSet[85]、CrowS Pairs[86]、BOLD[87]、HolisticBias[88]、BBQ[89]、CBBQ[90]和FairLex[91]。Cao等[77]创建了Gender Inclusive Coreference dataset跨性别个体数据集,在该数据集上目前最好的模型效果F_1值仅为34%。CDail Bias[83]数据集引入了第1个带注释的中国社会偏见检测对话数据集,涵盖种族、性别、地区和职业类别。BOLD[87]使用从维基百科上收集的句子,从性别极性、尊重、情感和毒性4个方面对大语言模型进行评估。HolisticBias[88]根据人类偏好等标准对GPT-2、DialoGPT等模型的输出进行评估。BBQ[89]将对比组答案和正确答案相结合,提出统计偏差得分指标来评估多个问答模型,研究表明当正确答案与社会偏见一致时,模型平均准确率提高3.4%,而当模型在以性别为目标的例子中进行测试时,差异会扩大到5%以上。CBBQ[90]依据中国的社会文化因素对数据集进行更改,发现中国的大语言模型偏差更高。

4.3 毒性评估

毒性评估主要评估大语言模型生成的内容中是否含有仇恨、侮辱和淫秽等有害信息。大模型毒性评估分为毒性识别和分类评估以及生成句子中毒性评估。现有的毒性评估方法主要是使用毒性检测系统检测文本中可能包含的毒性语句等。在毒性识别和分类评估任务中,OLID[92]、SOLID[93]和OLID-BR[94]数据集主要用于对有毒数据的识别和分类,Social Bias Inference Corpus[95]、HateXplain[96]、Civili-

ty[97]、COVID-hate[98]和Latent Hatred[99]数据集用于对大模型的毒性鉴定和分类能力进行评估。在生成句子中毒性评估中,测评数据集主要包括RealToxicityPrompts[100]、HarmfulQ[101]。OLID是从Twitter上抓取的攻击性语言数据集,SOLID基于该数据集,使用半监督学习方法对数据集进行标注。Wang等[122]对大模型进行零样本即时毒性检测研究,Zhu等[123]对ChatGPT生成情绪分析和仇恨言论标签的能力进行分析,Huang等[124]研究了ChatGPT识别和分类隐性仇恨言论的能力。RealToxicityPrompts[100]是由10万个自然发生的提示组成的数据集,与毒性分类器的毒性评分配对。Deshpande等[125]使用该数据集,给ChatGPT分配特定的角色,通过角色扮演发现其毒性可以增加到6倍。HarmfulQ[101]是包含多个明显有毒问题的数据集,用于评估大语言模型生成答案的毒性。

4.4 事实性评估

事实性评估主要评估模型生成的内容是否真实、准确,以及是否符合事实。大语言模型事实性评估主要基于问答任务、对话任务、摘要任务和自然语言推理任务来实现。事实性评估的评价指标有准确率、召回率、EM等[126]。常用的评测方法可以分为直接评测方法和间接评测方法,直接评测方法主要利用大模型,使大模型可以完成一般人类完成的直接评价等,间接评测方法指利用大模型和现有的评测指标和基准等对模型的事实性进行评估。问答任务的数据集主要包括NewsQA[102]、SQuAD 2.0[103]、BIG-bench[104]、SelfAware[105]、TruthfulQA[106]、FLUB[107]。对话任务的数据集主要包括DIAL-FACT[108]、BEGIN[109]、ConsisTest[110]、PersonalChat[111]。文本摘要任务的数据集主要包括XSum-Faith[112]、FactCC[113]、SummEval[114]、SUMMAC[115]、CLIFF[116]、AGGREFACT[117]。SelfAware用于评估大模型在缺乏足够的信息来提供问题的明确答案时如何识别知识的边界,结果表明GPT-3等大模型具有内在的自我认识能力,但与人类在认识知识局限性方面的熟练程度具有差距。TruthfulQA[106]包括38个不同类别的817个问题,用来衡量语言模型在生成问题答案时是否真实,Lin等[106]在GPT-3、GPT-Neo/J等大模型上使用该数据集进行测试,结果表明大语言模型的最高答对率为58%。在对话任务中,大语言模型的事实性评估主要包括事实核查和事实一致性评估两类数据集。Gupta等[107]使用DIAL-FACT基准来进行对话中的事实核查任务。Dziri等[109]提出BEGIN基准用于评估对话中的事实一致性。文本摘要任务中,Tam等[127]提出了事实不一致基准来衡量大语言模型的事实一致性,研究表明部分大语言模型通常给事实一致的摘要比事实不一致的摘要分配更高的分数。FLUB[107]设计了3个难度递增的任务,主要包含从真实网络环境中收集到的狡猾、幽默和误导的问题,在几个大语言模型上的评估结果表明只有性能最好的模型在该数据集上的效果超过了及格线。

5 大语言模型安全性评估

大语言模型的安全性评估是指对其进行系统性分析和测试,以降低它们在使用过程中的安全风险。安全性评估包括鲁棒性评估和风险评估,具体数据集如表4所示。

表4 风险性评估数据集

Table 4 Risk evaluation datasets

任务	数据集
鲁棒性评估	PromptBench[128]、Justice[129]、AdvGLUE[130]、ANLI[131]、DDXPlus[132]、WMT[133]、RobuT[134]、SynText-Bench[135]、JGLUE[136]、ReCode[137]
风险评估	AgentBench[138]、WebArena[139]

5.1 鲁棒性评估

大语言模型的鲁棒性评估是指其面对未知情况和攻击时的稳定性和可靠性进行评估。这种评估旨在确认大语言模型在实际应用中是否能够保持良好的性能,并且不容易受到干扰或攻击的影响。使用攻击方法对模型进行攻击,随后计算对抗精度是鲁棒性评估的一种重要方法。此外,对样本进行扰动,观察模型输出的变化,评估模型在干净样本和扰动样本的准确率差异也是评估模型鲁棒性的方法。对抗性攻击测试和分布外数据测试是鲁棒性评估的两个重要方向。对抗性攻击测试指大语言模型可能会受到对抗性攻击,即有意设计的输入样本,误导大语言模型或导致错误的输出。评估大语言模型对不同类型对抗性攻击的鲁棒性,如对抗性样本、对抗性噪声等,有助于了解其在面对这些攻击时的表现。分布外数据测试指大语言模型在实际应用中可能会遇到不同于训练数据分布的数据。评估模型在处理分布外数据时的表现,以及其对于未知数据的适应能力,是评估其鲁棒性的重要部分。

鲁棒性评估的数据集主要包括PromptBench[128]、Justice[129]、AdvGLUE[130]、ANLI[131]、DDXPlus[132]、WMT[133]、RobuT[134]、SynTextBench[135]、JGLUE[136]和ReCode[137]。Wang等[140]使用AdvGLUE[120]、ANLI[121]和DDXPlus[122]数据集,从对抗性的角度评估了ChatGPT和其他大语言模型。Zhuo等[141]提出了基于提示的大型代码语言模型对抗性鲁棒性研究,结果表明,代码语言模型容易受到精心制作的对抗性示例的攻击,进而提出了在不需要大量标记数据或大量计算资源的情况下提高大语言模型鲁棒性的方法。Yang等[142]通过扩展GLUE[143]数据集评估分布外数据鲁棒性,结果表明在操纵输入情况下大语言模型存在潜在安全风险。Li等[144]总体分析了大语言模型的对抗性鲁棒性、领域泛化和数据集偏差,同时强调了当前的挑战和未来的研究前景。Zhu等[128]提出从字符、单词、句子和语义层面评估大语言模型面对对抗性文本攻击的鲁棒性,结果表明,大语言模型容易受到对抗性提示的影响。

5.2 风险评估

随着技术的发展,大语言模型的能力迅速接近或达到人类水平,这带来了极大的安全性风险。传统软件的风险评估主要包括安全需求分析、安全架构设计审查、渗透测试、持续监控和程序数据扫描等评估方法,用来查找软件中存在的风险隐患。大语言模型的风险评估除了要涵盖以上几个方面,在评估模型设计中的风险、保证模型自身具有安全性方面,还应该包括:大语言模型是否追求权利和财富、大语言模型是否追求短期利益、大语言模型是否有自我意识、大语言模型是否具有协作能力以及大语言模型与环境的交互能力。Perez等[145]通过自动构建数据集来研究大语言模型的风险行为,他们发现大语言模型不仅表现出取悦人类的行为,还表现出对权力和资源的渴望。Chan等[146]通过评估大语言模型在与其他智能体在高风险交互中的行为来分析大语言模型的合作能力。他们使用众包和语言模型生成具有特定博弈论结构的场景,并根据生成的数据形成评估数据集。GPT-3在该数据集上的评估结果表明,基于指令微调的大语言模型倾向于以合作的方式来行动。Liu等[138]提出了基准测试AgentBench用于评估大语言模型在8种不同环境中的推理和决策能力,结果表明大模型在复杂环境中具有强大的智能能力。Zhou等[139]创建了一个逼真的网站环境,提出了一组基准测试评估大语言模型在复杂环境下完成任务的正确性。结果表明,基于GPT-4的智能体只实现了14.41%的端到端任务成功率,大语言模型在解决复杂任务方面仍面临着挑战。

6 评测基准

大模型的评测基准可以对大模型的能力进行全面的评估,发现模型存在的漏洞和不足,为模型的改进提供参考。表5列出了一些常用的大模型评测基准,并对评测基准的所在领域以及实验结果进行介绍。

表5 大模型评测基准

Table 5 Large language models evaluation benchmark

基准	所在领域	模型表现
APPS[147]	代码生成能力	可以通过约20%的入门问题的测试用例
ARC[148]	知识获取能力	DGEM的得分仅有27.11
C-Eval[149]	推理能力	仅有GPT-4可以达到60%以上的平均准确率
Chatbot Arena[150]	问答能力	GPT-4能够很好匹配人类偏好
DROP[151]	推理能力	最好的模型BERT只达到了32.7%的F_1
GSM8K[152]	推理能力	GPT-4的准确率已经超过90%
HELM[19]	多场景	在HELM之前模型平均仅在17.9%的场景上评估
HumanEval[63]	代码生成能力	Codex能够解决28.8%的问题
LAMBADA[153]	文本理解能力	Chinchilla模型的准确率仅有77.4%
MMLU[154]	文本模型的多任务	多数模型的准确率接近25%,GPT-3为43.9%
NQ[155]	开放问答能力	BERT的F_1分数只有52%,低于人类
PromptBench[128]	鲁棒性	UL2表现最好,性能下降8%
QMSum[156]	文本摘要能力	HMNet表现最后,ROUGE得分为32.29
SuperGLUE[157]	自然语言理解能力	BERT得分为71.5,低于人类
TriviaQA[158]	文本理解能力	模型表现仅有40%

C-Eval[149]是中文大模型的评估基准,旨在评估基础模型在中文背景下的高级推理能力,C-Eval Hard是从C-Eval中选择出来的部分具有挑战性的科目,对大模型的高级推理能力进行评估。结果表明,在C-Eval中,仅有GPT-4可以达到60%以上的平均准确率,大模型仍具有很大的改进空间。DROP[151]是对文本内容进行推理的基准,含有9万个问题,要求模型对问题进行解析,并对问题进行离散操作,对段落内容进行推理,需要模型对段落内容具有全面的了解。结果表明,最好的模型只达到了32.7%的F_1。SuperGLUE[157]保留了GLUE[143]中最难的两个任务,其余任务通过公开征集确定,具有更加多样化的任务格式和更具挑战性的NLU任务,结果表明,BERT的平均得分仍然低于人类,WSC[159]与人类的差距最大,模型中最小的差距也有10分。GSM8K[152]包含8 500个高质量、语言多样的小学数学文字问题的数据集,结果表明,即使是最大的Transformer模型也不能在测试中取得高性能,但通过训练验证器来判断模型完成度的正确性,能够进一步提高性能。

MMLU[154]是用于衡量文本模型的多任务准确率的一个基准,涵盖57个任务,包括基础数学、计算机科学等多个领域,结果表明,大多数最新模型的准确率接近随机水平,最大的GPT-3模型的准确率比随机水平高,但仍需要改进。LAMBADA[153]通过单词预测任务来评估模型的文本理解能力,包含多篇叙事性文章,实验表明,Chinchilla模型的准确率仅有77.4%。TriviaQA[158]是阅读理解基准,包含超过65万个问题-答案-证据的三元组,具有相对复杂的问题,并需要有较多的跨句推理寻找答案,结果表明,常用的神经网络在该基准上的表现仅有40%,低于人类。Chatbot Arena[150]用于展示大模型与真实用户对话的性能和评分,结果表明,GPT-4能够很好地匹配人类偏好,达到了超过80%的一致性,与人类的一致性水平相同。Natural Questions[155]是评估开放问答领域的基准,包括30万个问题以及人工解释,实验表明,BERT的F_1分数只有52%,而人工的最优成绩远高于BERT。

APPS[147]是用于代码生成的基准,用于衡量模型根据任意自然语言规范生成Python代码的能力,APPS包括10 000个问题,从简单的解决方案到复杂的算法,结果表明,最新的模型可以通过约20%的

入门问题的测试用例。HumanEval[63]用于衡量从文档生成程序的功能正确性,包含了164个手写编程问题,每个问题有明确的函数名和文档字符串、完整的函数体以及针对该函数的单元测试,结果表明,Codex解决了28.8%的问题,GPT-J解决了14%的问题。

HELM[19]涵盖了16个场景和7个类别的度量,覆盖范围广泛,且具有多个评价指标,全面评估大语言模型,结果表明,模型的生成结果受到场景的影响。ARC[148]对模型的知识和推理能力进行评估,其中包含了7 787个问题,比之前的问答数据集具有更加全面的推理问题。结果表明,在先前的问答数据集中表现较好的模型在该基准上表现一般。QMSum[156]包含了来自多个领域的232次会议的1 808个查询-摘要对,用于对多个领域的会议做摘要,模型根据查询选择和总结会议的相关部分,实验表明,模型对长时间的会议进行摘要仍面临重大挑战。PromptBench[128]旨在衡量大模型对对抗性提示的鲁棒性基准,通过使用大量的对抗性文本攻击,评估保持语义完整性的同时,也包括轻微偏差对大模型结果的影响。结果表明,当代大模型对对抗性提示并不鲁棒。

7 结束语

随着大语言模型的广泛应用和不断发展,面向大语言模型的功能、性能、对齐和安全等方面的评估已取得一定进展,但目前仍存在评估指标不足、测试环境与应用现实存在差距等问题。在大语言模型功能评估方面,评估的实时性和动态性需要进一步完善,未来的评估应更加注重模型在实时场景下的表现,以及在不同时间段内的变化趋势。在性能评估方面,需进一步评估大模型系统在多用户同时发送请求时的响应时长,以判断大语言模型在真实应用场景下的并行处理能力是否能满足用户需求。在对其评估方面,现有的对齐评估方法可能无法完全适应大语言模型的需求,未来可以研究跨领域的对齐评估,以确保大语言模型在多领域的输出内容都满足人类价值观。在安全性评估方面,现有方法缺乏对与大模型的可解释性和透明度的评估,为了增强人们对大语言模型的信任度,可以增加对于大模型可解释性和透明度的评估。随着技术的不断进步和应用场景的不断拓展,大语言模型的评估将继续成为研究和实践的重要议题,需要通过先进的评估方法和手段提升大语言模型的质量,促进其在各个领域的可持续发展。

参考文献:

[1] 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.

[2] TOUVRON H, LAVRIL T, IZACARD G, et al. LLaMA: Open and efficient foundation language models[EB/OL]. (2023-11-06). https://arxiv.org/abs/ 2302.13971.

[3] NEWSWIRE P. Alibaba cloud unveils new AI model to support enterprises' intelligence transformation[EB/OL]. (2023-04-10)[2023-08-11]. https://finance.yahoo.com/news/alibaba-cloud-unveils-ai-model-031300094.html.

[4] OpenAI. ChatGPT[EB/OL]. (2023). https://openai.com/chatgpt.

[5] ACHIAM J, ADLER S, AGARWAL S, et al. GPT-4 technical report[EB/OL]. (2023-03-15). https://arxiv. org/abs/2303.08774.

[6] CHANG Y, WANG X, WANG J, et al. A survey on evaluation of large language models[EB/OL]. (2023-07-09). https://arxiv.org/abs/2307.03109v2.

[7] BROWN P F. Class-based N-gram models of natural language[J]. Computational Linguistics, 1990, 18: 18.

[8] BENGIO Y, DUCHARME R, VINCENT P. A neural probabilistic language model[J]. Advances in Neural Information Processing Systems, 2000. DOI:10.1162/153244303322533223.

[9] MIKOLOV T, CHEN K, CORRADO G, et al. Efficient estimation of word representations in vector space[EB/OL]. (2013-09-07). https://arxiv.org/abs/1301.3781v1.

[10] RADFORD A, NARASIMHAN K, SALIMANS T, et al. Improving language understanding by generative pre-training[EB/OL]. (2018-01-01)[2024-01-30]. https://www.docin.com/p-2176538517.html.

[11] DEVLIN J, CHANG M W, LEE K, et al. BERT: Pre-training of deep bidirectional transformers for language understanding [C]//Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. [S.l.]: ACL, 2019: 4171-4186.

[12] 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. Long Beach, CA, USA: [s.n.], 2017: 6000-6010.

[13] BROWN T, MANN B, RYDER N, et al. Language models are few-shot learners[J]. Advances in Neural Information Processing Systems, 2020, 33: 1877-1901.

[14] 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.

[15] PATTERSON D, GONZALEZ J, LE Q, et al. Carbon emissions and large neural network training[EB/OL]. (2021-04-30). https://arxiv.org/abs/2104.10350.

[16] JIAO X, YIN Y, SHANG L, et al. TinyBERT: Distilling BERT for natural language understanding[C]// Proceedings of Findings of the Association for Computational Linguistics. [S.l.]: ACL, 2020: 4163-4174.

[17] KORTHIKANTI V A, CASPER J, LYM S, et al. Reducing activation recomputation in large transformer models[J]. Proceedings of Machine Learning and Systems, 2023. DOI:10.48550/arXiv.2205.05198.

[18] SHAMSEERA S P, SREEKANTH E S. Learning word vectors to dual sentiment analysis[EB/OL]. [2024-03-27]. http://www.ijcter.com/papers/volume-2/issue-5/word-vectors-in-sentiment-analysis.pdf.

[19] LIANG P, BOMMASANI R, LEE T, et al. Holistic evaluation of language models[J]. Transactions on Machine Learning Research, 2023. DOI:10.1111/nyas.15007.

[20] ZHANG W, DENG Y, LIU B, et al. Sentiment analysis in the era of large language models: A reality check[EB/OL]. (2023-05-24). https://arxiv.org/abs/2305.15005.

[21] WANG Z, XIE Q, DING Z, et al. Is ChatGPT a good sentiment analyzer? A preliminary study[EB/OL]. (2023-04-10). https://arxiv.org/abs/2304.04339.

[22] PEÑA A, MORALES A, FIERREZ J, et al. Leveraging large language models for topic classification in the domain of public affairs[C]//Proceedings of International Conference on Document Analysis and Recognition. Cham: Springer Nature Switzerland, 2023: 20-33.

[23] TAO Z, JIN Z, BAI X, et al. Eveval: A comprehensive evaluation of event semantics for large language models[EB/OL]. (2023-05-24). https://arxiv.org/abs/2305.15268.

[24] RICCARDI N, DESAI R H. The two word test: A semantic benchmark for large language models[EB/OL]. (2023-06-10). https://arxiv.org/abs/2306.04610.

[25] MANYIKA J, HSIAO S. An overview of BARD: An early experiment with generative AI[EB/OL]. (2023)[2024-03-20]. https://ai. google/static/documents/google-about-bard. pdf.

[26] GUO Z, JIN R, LIU C, et al. Evaluating large language models: A comprehensive survey[EB/OL]. (2023-11-09). https://arxiv.org/abs/2310.19736.

[27] TALMOR A, HERZIG J, LOURIE N, et al. COMMONSENSEQA: A question answering challenge targeting commonsense knowledge[C]//Proceedings of NAACL-HLT. Minneapolis, MN, USA: Association for Computational Linguistics, 2019: 4149-4158.

[28] BISK Y, ZELLERS R, GAO J, et al. PIQA: Reasoning about physical commonsense in natural language[C]//Proceedings of the AAAI Conference on Artificial Intelligence. [S.l.]: AAAI, 2020: 7432-7439.

[29] WANG S, DURRETT G, ERK K. Modeling semantic plausibility by injecting world knowledge[C]//Proceedings of NAACL-HLT. New Orleans, Louisiana, USA: Association for Computational Linguistics, 2018: 303-308.

[30] SAP M, RASHKIN H, CHEN D, et al. Social IQA: Commonsense reasoning about social interactions[C]//Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). Hong Kong, China: ACL, 2019: 4463-4473.

[31] TIAN J, LI Y, CHEN W, et al. Diagnosing the first-order logical reasoning ability through LogicNLI[C]//Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. Punta Cana, Dominican Republic: ACL, 2021: 3738-3747.

[32] LIU H, CUI L, LIU J, et al. Natural language inference in context-investigating contextual reasoning over long texts[C]//Proceedings of the AAAI Conference on Artificial Intelligence. [S.l.]: AAAI, 2021: 13388-13396.

[33] JOSHI P, ADITYA S, SATHE A, et al. TaxiNLI: Taking a ride up the NLU Hill[C]//Proceedings of the 24th Conference on Computational Natural Language Learning. Punta Cana, Dominican Republic: [s.n.], 2020: 41-55.

[34] YU W, JIANG Z, DONG Y, et al. ReClor: A reading comprehension dataset requiring logical reasoning[C]//Proceedings of International Conference on Learning Representations. New Orleans, USA: [s.n.], 2019.

[35] YANG Z, QI P, ZHANG S, et al. HotpotQA: A dataset for diverse, explainable multi-hop question answering[C]//Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. [S.l.]: ACL, 2018: 2369-2380.

[36] OTT S, HEBENSTREIT K, LIÉVIN V, et al. ThoughtSource: A central hub for large language model reasoning data[J]. Scientific Data, 2023, 10(1): 528.

[37] HENDRYCKS D, BURNS C, KADAVATH S, et al. Measuring mathematical problem solving with the MATH dataset [C]//Proceedings of Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2). [S.l.]: SAIL, 2021.

[38] ARORA D, SINGH H G. Have LLMS advanced enough? A challenging problem solving benchmark for large language models [EB/OL]. (2024-01-20). https://arxiv.org/abs/2305.15074.

[39] WEI T, LUAN J, LIU W, et al. CMATH: Can your language model pass Chinese elementary school math test? [EB/OL]. (2023-07-12). https://arxiv.org/abs/2306.16636.

[40] SPEER R, CHIN J, HAVASI C. Conceptnet 5.5: An open multilingual graph of general knowledge[C]//Proceedings of the AAAI Conference on Artificial Intelligence. [S.l.]: AAAI, 2017.

[41] BANG Y, CAHYAWIJAYA S, LEE N, et al. A multitask, multilingual, multimodal evaluation of ChatGPT on reasoning, hallucination, and interactivity[EB/OL]. (2023-04-20). https://arxiv.org/abs/2302.04023.

[42] BIAN N, HAN X, SUN L, et al. ChatGPT is a knowledgeable but inexperienced solver: An investigation of commonsense problem in large language models[EB/OL]. (2023-06-18). https://arxiv.org/abs/2303.16421.

[43] ZELLERS R, HOLTZMAN A, BISK Y, et al. HellaSwag: Can a machine really finish your sentence?[C]//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Florence, Italy: ACL, 2019: 4791-4800.

[44] ZHOU B, KHASHABI D, NING Q, et al. Study of temporal commonsense understanding[C]//Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). Hong Kong, China: ACL, 2019: 3363-3369.

[45] QIN C, ZHANG A, ZHANG Z, et al. Is ChatGPT a general-purpose natural language processing task solver? [C]//Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. Singapore: ACL, 2023: 1339-1384.

[46] LEE N, AN N M, THORNE J. Can large language models capture dissenting human voices? [C]//Proceedings of 2023 Conference on Empirical Methods in Natural Language Processing. Singapore: ACL, 2023.

[47] LIU J, CUI L, LIU H, et al. LogiQA: A challenge dataset for machine reading comprehension with logical reasoning[C]//Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence. Yokohama, Japan: [s.n.], 2021: 3622-3628.

[48] LIU H, LIU J, CUI L, et al. LogiQA 2.0 An improved dataset for logical reasoning in natural language understanding[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2023. DOI: 10.1109/TASLP.2023.3293046.

[49] WANG S, LIU Z, ZHONG W, et al. From LSAT: The progress and challenges of complex reasoning[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2022, 30: 2201-2216.

[50] LIU H, NING R, TENG Z, et al. Evaluating the logical reasoning ability of ChatGPT and GPT-4[EB/OL]. (2023-04-07). https://arxiv.org/abs/2304.03439.

[51] XU F, LIN Q, HAN J, et al. Are large language models really good logical reasoners? A comprehensive evaluation from deductive, inductive and abductive views[EB/OL]. (2023-10-02). https://arxiv.org/abs/2306.09841.

[52] TANG Y, NG H T, TUNG A. Do multi-hop question answering systems know how to answer the single-hop sub-questions? [C]//Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics. [S.l.]: ACL, 2021: 3244-3249.

[53] CHEN L, ZAHARIA M, ZOU J. How is ChatGPT's behavior changing over time? [EB/OL]. (2023-07-19). https://arxiv.org/abs/2307.09009.

[54] COLLINS K M, JIANG A Q, FRIEDER S, et al. Evaluating language models for mathematics through interactions[EB/OL]. (2023-06-06). https://arxiv.org/abs/2306.01694.

[55] HERMANN K M, KOCISKY T, GREFENSTETTE E, et al. Teaching machines to read and comprehend[J]. Advances in Neural Information Processing Systems, 2015. DOI: 1048550/arxiv.1506.03340.

[56] NARAYAN S, COHEN S B, LAPATA M. Don't give me the details, just the summary! Topic-aware convolutional neural networks for extreme summarization[EB/OL]. (2018-08-27). https://arxiv.org/abs/1808.08745.

[57] SMITH S, PATWARY M, NORICK B, et al. Using deep speed and megatron to train megatron-turing NLG 530B, a large-scale generative language mode[EB/OL]. (2022-02-04). https://arxiv.org/abs/2201.11990.

[58] LEWIS M, LIU Y, GOYAL N, et al. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension[C]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Seattle, Washington, USA: ACL, 2020: 7871-7880.

[59] ZHENG L, CHIANG W L, SHENG Y, et al. LMSYS-Chat-1M: A large-scale real-world LLM conversation dataset[EB/OL]. (2023-09-21). https://arxiv.org/abs/2309.11998.

[60] LIN Y T, CHEN Y N. LLM-EVAL: Unified multi-dimensional automatic evaluation for open-domain conversations with large language models[C]//Proceedings of the 5th Workshop on NLP for Conversational AI (NLP4ConvAI 2023). Toronto, Canada: ACL, 2023: 47-58.

[61] WANG L, LYU C, JI T, et al. Document-level machine translation with large language models[C]//Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. Singapore: ACL, 2023: 16646-16661.

[62] LIANG J, HUANG W, XIA F, et al. Code as policies: Language model programs for embodied control[C]//Proceedings of 2023 IEEE International Conference on Robotics and Automation (ICRA). [S.l.]: IEEE, 2023: 9493-9500.

[63] CHEN M, TWOREK J, JUN H, et al. Evaluating large language models trained on code[EB/OL]. (2021-07-14). https://arxiv.org/abs/2107.03374.

[64] PU D, DEMBERG V. ChatGPT vs human-authored text: Insights into controllable text summarization and sentence style transfer[C]//Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics. Toronto, Canada: ACL, 2023: 1-18.

[65] CHIA Y K, HONG P, BING L, et al. Instructeval: Towards holistic evaluation of instruction-tuned large language models [EB/OL]. (2023-06-18). https://arxiv.org/abs/2306.04757.

[66] SCHWARTZ R, STANOVSKY G, SWAYAMDIPTA S, et al. The right tool for the job: Matching model and instance complexities[C]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Seattle, Washington, USA: ACL, 2020: 6640-6651.

[67] ZHOU W, XU C, GE T, et al. BERT loses patience: Fast and robust inference with early exit[J]. Advances in Neural Information Processing Systems, 2020, 33: 18330-18341.

[68] FORBES M, HWANG J D, SHWARTZ V, et al. Social Chemistry 101: Learning to reason about social and moral norms [C]//Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). [S.l.]: ACL, 2020: 653-670.

[69] HENDRYCKS D, BURNS C, BASART S, et al. Aligning AI with shared human values[C]//Proceedings of International Conference on Learning Representations.[S.l.]: [s.n.], 2021.

[70] EMELIN D, LE BRAS R, HWANG J D, et al. Moral stories: Situated reasoning about norms, intents, actions, and their consequences[C]//Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. Punta Cana, Dominican Republic: ACL, 2021: 698-718.

[71] REZAPOUR R, SHAH S H, DIESNER J. Enhancing the measurement of social effects by capturing morality[C]//

Proceedings of the Tenth Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis. Minneapolis, USA: [s.n.], 2019: 35-45.

[72] LOURIE N, LE BRAS R, CHOI Y. Scruples: A corpus of community ethical judgments on 32,000 real-life anecdotes[C]// Proceedings of the AAAI Conference on Artificial Intelligence. [S.l.]: AAAI, 2021: 13470-13479.

[73] JIN Z, LEVINE S, GONZALEZ ADAUTO F, et al. When to make exceptions: Exploring language models as accounts of human moral judgment[J]. Advances in Neural Information Processing Systems, 2022, 35: 28458-28473.

[74] KIM H, YU Y, JIANG L, et al. ProsocialDialog: A prosocial backbone for conversational agents[C]//Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing. [S.l.]: ACL, 2022: 4005-4029.

[75] ZIEMS C, YU J, WANG Y C, et al. The moral integrity corpus: A benchmark for ethical dialogue systems[C]//Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics. Dublin, Germany: ACL, 2022: 3755-3773.

[76] RUDINGER R, NARADOWSKY J, LEONARD B, et al. Gender bias in coreference resolution[C]//Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. New Orleans, Louisiana, USA: ACL, 2018: 8-14.

[77] ZHAO J, WANG T, YATSKAR M, et al. Gender bias in coreference resolution: Evaluation and debiasing methods[EB/OL]. (2018-04-18). https://arxiv.org/abs/1804.06876.

[78] CAO Y T, DAUMÉ III H. Toward gender-inclusive coreference resolution[C]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Seattle, Washington, USA: ACL, 2020: 4568-4595.

[79] STANOVSKY G, SMITH N A, ZETTLEMOYER L. Evaluating gender bias in machine translation[EB/OL]. (2019-06-03). https://arxiv.org/abs/1906.00591.

[80] PENNINGTON J, SOCHER R, MANNING C D. Glove: Global vectors for word representation[C]//Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). Doha, Qatar: [s.n.], 2014: 1532-1543.

[81] KIRITCHENKO S, MOHAMMAD S. Examining gender and race bias in two hundred sentiment analysis systems[C]// Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics. New Orleans, Louisiana, USA: ACL, 2018: 43-53.

[82] GAUT A, SUN T. Towards understanding gender bias in relation extraction[EB/OL]. (2020-08-08). https://arxiv.org/abs/1911.03642.

[83] ZHOU J, DENG J, MI F, et al. Towards identifying social bias in dialog systems: Frame, datasets, and benchmarks[EB/OL]. (2022-10-28). https://arxiv.org/abs/2202.08011.

[84] ZHANG G, LI Y, WU Y, et al. CORGI-PM: A Chinese corpus for gender bias probing and mitigation[EB/OL]. (2023-01-01). https://arxiv.org/abs/2301.00395.

[85] NADEEM M, BETHKE A, REDDY S. StereoSet: Measuring stereotypical bias in pretrained language models[C]// Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). [S.l.]: ACL, 2021: 5356-5371.

[86] NANGIA N, VANIA C, BHALERAO R, et al. CrowS-Pairs: A challenge dataset for measuring social biases in masked language models[C]//Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). [S.l.]: ACL, 2020: 1953-1967.

[87] DHAMALA J, SUN T, KUMAR V, et al. Bold: Dataset and metrics for measuring biases in open-ended language generation [C]//Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency. [S.l.]: ACM, 2021: 862-872.

[88] SMITH E M, HALL M, KAMBADUR M, et al. Mtrics for measnew biases in language models with a holistic descriptor dataset[C]//Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing. [S.l.]: ACL, 2022: 9180-9211.

[89] PARRISH A, CHEN A, NANGIA N, et al. BBQ: A hand-built bias benchmark for question answering[C]//Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics. Dublin, Germany: ACL, 2022: 2086-2105.

[90] HUANG Y, XIONG D. CBBQ: A Chinese bias benchmark dataset curated with human-AI collaboration for large language models[EB/OL]. (2023-06-28). https://arxiv.org/abs/2306.16244.

[91] CHALKIDIS I, PASINI T, ZHANG S, et al. Fairlex: A multilingual benchmark for evaluating fairness in legal text processing

[EB/OL]. (2023-03-14). https://arxiv.org/abs/2203.07228.

[92] ZAMPIERI M, MALMASI S, NAKOV P, et al. Predicting the type and target of offensive posts in social media[C]// Proceedings of NAACL-HLT. Minneapolis, MN, USA: ACL, 2019: 1415-1420.

[93] ROSENTHAL S, ATANASOVA P, KARADZHOV G, et al. SOLID: A large-scale semi-supervised dataset for offensive language identification[C]//Proceedings of Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021. [S.l.]: ACL, 2021: 915-928.

[94] TRAJANO D, BORDINI R H, VIEIRA R. OLID-BR: Offensive language identification dataset for Brazilian Portuguese[J]. Language Resources and Evaluation, 2023. DOI:10.1007/s10579-023-09657-0.

[95] SAP M, GABRIEL S, QIN L, et al. Social bias frames: Reasoning about social and power implications of language[C]// Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Seattle, Washington, USA: ACL, 2020: 5477-5490.

[96] MATHEW B, SAHA P, YIMAM S M, et al. Hatexplain: A benchmark dataset for explainable hate speech detection[C]// Proceedings of the AAAI Conference on Artificial Intelligence. [S.l.]: AAAI, 2021: 14867-14875.

[97] ZAMPIERI M, MALMASI S, NAKOV P, et al. SemEval-2019 Task 6: Identifying and categorizing offensive language in social media (OffensEval) [C]//Proceedings of the 13th International Workshop on Semantic Evaluation. Minneapolis, Minnesota, USA: [s.n.], 2019: 75-86.

[98] HE B, ZIEMS C, SONI S, et al. Racism is a virus: Anti-Asian hate and counterspeech in social media during the COVID-19 crisis[C]//Proceedings of the 2021 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining. [S.l.]: IEEE, 2021: 90-94.

[99] ELSHERIEF M, ZIEMS C, MUCHLINSKI D, et al. Latent Hatred: A benchmark for understanding implicit hate speech [C]//Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. Punta Cana, Dominican Republic: ACL, 2021: 345-363.

[100] GEHMAN S, GURURANGAN S, SAP M, et al. Realtoxicityprompts: Evaluating neural toxic degeneration in language models[EB/OL]. (2020-09-25). https://arxiv.org/abs/2009.11462.

[101] SHAIKH O, ZHANG H, HELD W, et al. On second thought, let's not think step by step! Bias and toxicity in zero-shot reasoning[EB/OL]. (2023-06-04). https://arxiv.org/abs/2212.08061.

[102] TRISCHLER A, WANG T, YUAN X, et al. NewsQA: A machine comprehension dataset[C]//Proceedings of the 2nd Workshop on Representation Learning for NLP. Vancouver: ACL, 2017: 191-200.

[103] RAJPURKAR P, JIA R, LIANG P. Know what you don't know: Unanswerable questions for SQuAD[C]//Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics. Melbourne, Australia: ACL, 2018: 784-789.

[104] SRIVASTAVA A, RASTOGI A, RAO A, et al. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models[J]. Transactions on Machine Learning Research, 2023. DOI: 10.48850/arxiv.2206.04615.

[105] YIN Z, SUN Q, GUO Q, et al. Do large language models know what they don't know?[C]//Proceedings of Findings of the Association for Computational Linguistics. [S.l.]: ACL, 2023: 8653-8665.

[106] LIN S, HILTON J, EVANS O. TruthfulQA: Measuring how models mimic human falsehoods[C]//Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics. Dublin, Germany: ACL, 2022: 3214-3252.

[107] LI Y, ZHOU Q, LUO Y, et al. When LLMs meet cunning questions: A fallacy understanding benchmark for large language models[EB/OL]. (2024-02-16). https://arxiv.org/abs/2402.11100.

[108] GUPTA P, WU C S, LIU W, et al. DialFact: A benchmark for fact-checking in dialogue[C]//Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics. Dublin, Germany: ACL, 2022: 3785-3801.

[109] DZIRI N, RASHKIN H, LINZEN T, et al. Evaluating attribution in dialogue systems: The BEGIN benchmark[J]. Transactions of the Association for Computational Linguistics, 2022, 10: 1066-1083.

[110] LOTFI E, DE BRUYN M, BUHMANN J, et al. What was your name again? Interrogating generative conversational models for factual consistency evaluation[C]//Proceedings of the 2nd Workshop on Natural Language Generation, Evaluation, and Metrics (GEM). Abu Dhabi, United Arab Emirates: [s.n.], 2022: 509-519.

[111] ZHANG S, DINAN E, URBANEK J, et al. Personalizing dialogue agents: I have a dog, do you have pets too?[C]//

Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics. Melbourne, Australia: ACL, 2018: 2204-2213.

[112] MAYNEZ J, NARAYAN S, BOHNET B, et al. On faithfulness and factuality in abstractive summarization[C]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Seattle, Washington, USA: ACL, 2020: 1906-1919.

[113] KRYŚCIŃSKI W, MCCANN B, XIONG C, et al. Evaluating the factual consistency of abstractive text summarization[C]//Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). [S.l.]: ACL, 2020: 9332-9346.

[114] FABBRI A R, KRYŚCIŃSKI W, MCCANN B, et al. Summeval: Re-evaluating summarization evaluation[J]. Transactions of the Association for Computational Linguistics, 2021, 9: 391-409.

[115] LABAN P, SCHNABEL T, BENNETT P N, et al. SummaC: Re-visiting NLI-based models for inconsistency detection in summarization[J]. Transactions of the Association for Computational Linguistics, 2022, 10: 163-177.

[116] CAO S, WANG L. CLIFF: Contrastive learning for improving faithfulness and factuality in abstractive summarization[C]//Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. Punta Cana, Dominican Republic: ACL, 2021: 6633-6649.

[117] TANG L, GOYAL T, FABBRI A R, et al. Understanding factual errors in summarization: Errors, summarizers, datasets, error detectors[C]//Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics. Toronto, Canada: ACL, 2023: 11626-11644.

[118] ALLEN C, VARNER G, ZINSER J. Prolegomena to any future artificial moral agent[J]. Journal of Experimental & Theoretical Artificial Intelligence, 2000, 12(3): 251-261.

[119] HUANG Y, ZHANG Q, SUN L. TrustGPT: A benchmark for trustworthy and responsible large language models[EB/OL]. (2023-06-30). https://arXiv preprint arXiv:2306.11507.

[120] RAUH M, MELLOR J, UESATO J, et al. Characteristics of harmful text: Towards rigorous benchmarking of language models[J]. Advances in Neural Information Processing Systems, 2022, 35: 24720-24739.

[121] TOKPO E K, DELOBELLE P, BERENDT B, et al. How far can it go? On intrinsic gender bias mitigation for text classification[C]//Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics. Dubrovnik, Croatia: [s.n.], 2023: 3418-3433.

[122] WANG Y S, CHANG Y. Toxicity detection with generative prompt-based inference[EB/OL]. (2022-05-24). https://arxiv.org/abs/2205.12390.

[123] ZHU Y, ZHANG P, HAQ E U, et al. Can ChatGPT reproduce human-generated labels? A study of social computing tasks [EB/OL]. (2023-04-24). https://arxiv.org/abs/2304.10145.

[124] HUANG F, KWAK H, AN J. Is ChatGPT better than human annotators? Potential and limitations of ChatGPT in explaining implicit hate speech[C]//Proceedings of Companion proceedings of the ACM Web Conference 2023. Austin, TX, USA: [s.n.], 2023: 294-297.

[125] DESHPANDE A, MURAHARI V, RAJPUROHIT T, et al. Toxicity in ChatGPT: Analyzing persona-assigned language models[C]//Proceedings of 2023 Findings of the Association for Computational Linguistics: EMNLP 2023. [S.l.]: ACL, 2023: 1236-1270.

[126] WANG C, LIU X, YUE Y, et al. Survey on factuality in large language models: Knowledge, retrieval and domain-specificity [EB/OL]. (2023-10-08). https://arxiv.org/abs/2310.07521.

[127] TAM D, MASCARENHAS A, ZHANG S, et al. Evaluating the factual consistency of large language models through news summarization[C]//Proceedings of Findings of the Association for Computational Linguistics. [S.l.]: ACL, 2023: 5220-5255.

[128] ZHU K, WANG J, ZHOU J, et al. Promptbench: Towards evaluating the robustness of large language models on adversarial prompts[EB/OL]. (2023-06-07). https://arxiv.org/abs/2306.04528.

[129] LIU Y, YAO Y, TON J F, et al. Trustworthy LLMs: A survey and guideline for evaluating large language models' alignment [EB/OL]. (2023-08-15). https://arxiv.org/abs/2308.05374.

[130] WANG B, XU C, WANG S, et al. Adversarial GLUE: A multi-task benchmark for robustness evaluation of language models

[C]// Proceedings of the Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track. [S.l.]: SAIL, 2021.

[131] NIE Y, WILLIAMS A, DINAN E, et al. Adversarial NLI: A new benchmark for natural language understanding[C]// Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Seattle, Washington, USA: ACL, 2020: 4885-4901.

[132] TCHANGO A F, GOEL R, WEN Z, et al. DDXPlus: A new dataset for automatic medical diagnosis[J]. Advances in Neural Information Processing Systems, 2022, 35: 31306-31318.

[133] JIAO W, WANG W, HUANG J, et al. Is ChatGPT a good translator? Yes with GPT-4 as the engine[EB/OL]. (2023-08-15). https://arxiv.org/abs/2301.08745.

[134] ZHAO Y, ZHAO C, NAN L, et al. RobuT: A systematic study of table qa robustness against human-annotated adversarial perturbations[C]//Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics.Toronto, Canada: ACL, 2023: 6064-6081.

[135] KO C Y, CHEN P Y, DAS P, et al. On robustness-accuracy characterization of large language models using synthetic datasets [C]// Proceedings of Workshop on Efficient Systems for Foundation Models. [S.l.]: ICML, 2023.

[136] KURIHARA K, KAWAHARA D, SHIBATA T. JGLUE: Japanese general language understanding evaluation[C]// Proceedings of the Thirteenth Language Resources and Evaluation Conference. Marseille, France: European Language Resources Association, 2022: 2957-2966.

[137] WANG S, LI Z, QIAN H, et al. ReCode: Robustness evaluation of code generation models[C]//Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics. Toronto, Canada: ACL, 2023: 13818-13843.

[138] LIU X, YU H, ZHANG H, et al. Agentbench: Evaluating LLMS as agents[EB/OL]. (2023-08-07). https://arxiv.org/abs/2308.03688.

[139] ZHOU S, XU F F, ZHU H, et al. Webarena: A realistic web environment for building autonomous agents[EB/OL]. (2023-07-15). https://arxiv.org/abs/2307.13854.

[140] WANG J, HU X, HOU W, et al. On the robustness of ChatGPT: An adversarial and out-of-distribution perspective[C]// Proceedings of ICLR 2023 Workshop on Trustworthy and Reliable Large-Scale Machine Learning Models. [S.l.]: ICLR, 2023.

[141] ZHUO T Y, LI Z, HUANG Y, et al. On robustness of prompt-based semantic parsing with large pre-trained language model: An empirical study on codex[C]//Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics. Dubrovnik, Croatia: ACL, 2023: 1090-1102.

[142] YANG L, ZHANG S, QIN L, et al. Glue-X: Evaluating natural language understanding models from an out-of-distribution generalization perspective[EB/OL]. (2023-05-12). https://arxiv.org/abs/2211.08073.

[143] WANG A, SINGH A, MICHAEL J, et al. Glue: A multi-task benchmark and analysis platform for natural language understanding[C]// Proceedings of the 7th International Conference on Learning Representations. New Orleans, LA, USA: ICLR, 2019.

[144] LI X, LIU M, GAO S, et al. A survey on out-of-distribution evaluation of neural NLP models[C]//Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence. Barcelona, Catalonia, Spain: [s.n.], 2023: 6683-6691.

[145] PEREZ E, RINGER S, LUKOŠIŪTĖ K, et al. Discovering language model behaviors with model-written evaluations[C]// Proceedings of Findings of the Association for Computational Linguistics. [S.l.]: ACL, 2023: 13387-13434.

[146] CHAN A, RICHÉ M, CLIFTON J. Towards the scalable evaluation of cooperativeness in language models[EB/OL]. (2023-03-16). https://arxiv.org/abs/2303.13360.

[147] HENDRYCKS D, BASART S, KADAVATH S, et al. Measuring coding challenge competence with APPS[C]// Proceedings of the Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2). [S.l.]: SAIL, 2021.

[148] CLARK P, COWHEY I, ETZIONI O, et al. Think you have solved question answering? tryARC, theAI2 reasoning challenge[EB/OL]. (2018-03-17). https://arxiv.org/abs/1803.05457.

[149] HUANG Y, BAI Y, ZHU Z, et al. C-EVAL: A multi-level multi-discipline Chinese evaluation suite for foundation models[J]. Advances in Neural Information Processing Systems, 2024. DOI: 10.48850/arxiv.2305.08322.

[150] ZHENG L, CHIANG W L, SHENG Y, et al. Judging LLM-as-a-judge with MT-bench and chatbot arena[J]. Advances in Neural Information Processing Systems, 2024, 36. DOI: 10.48850/arxiv.2306.05685.
[151] DUA D, WANG Y, DASIGI P, et al. DROP: A reading comprehension benchmark requiring discrete reasoning over paragraphs[C]//Proceedings of NAACL-HLT. Minneapolis, MN, USA: ACL, 2019: 2368-2378.
[152] COBBE K, KOSARAJU V, BAVARIAN M, et al. Training verifiers to solve math word problems[EB/OL]. (2021-11-18). https://arxiv.org/abs/2110.14168.
[153] PAPERNO D, KRUSZEWSKI MARTEL G D, LAZARIDOU A, et al. The LAMBADA dataset: Word prediction requiring a broad discourse context[C]// Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics Proceedings of the Conference. [S.l.]: ACL, 2016, 3: 1525-1534.
[154] HENDRYCKS D, BURNS C, BASART S, et al. Measuring massive multitask language understanding[C]// Proceedings of International Conference on Learning Representations. [S.l.]: ICLR, 2020.
[155] KWIATKOWSKI T, PALOMAKI J, REDFIELD O, et al. Natural questions: A benchmark for question answering research [J]. Transactions of the Association for Computational Linguistics, 2019, 7: 453-466.
[156] ZHONG M, YIN D, YU T, et al. QMSum: A new benchmark for query-based multi-domain meeting summarization[EB/OL]. (2021-04-13). https://arxiv.org/abs/2104.05938.
[157] WANG A, PRUKSACHATKUN Y, NANGIA N, et al. SupergLUE: A stickier benchmark for general-purpose language understanding systems[J]. Advances in Neural Information Processing Systems, 2019. DOI: 10.48850/arxiv.1905.00537.
[158] JOSHI M, CHOI E, WELD D S, et al. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension[C]//Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics. Vancouver, Canada: ACL, 2017: 1601-1611.
[159] LEVESQUE H J, DAVIS E, MORGENSTERN L. The Winograd schema challenge[C]//Proceedings of the Thirteenth International Conference on Principles of Knowledge Representation and Reasoning. [S.l.]: AAAI, 2012: 552-561.

作者简介:

赵睿卓(1995-),女,工程师,研究方向:智能软件测试,E-mail:ruizhuozhao@163.com。

曲紫畅(1996-),男,工程师,研究方向:智能软件测试。

陈国英(2002-),女,硕士研究生,研究方向:智能软件测试。

王坤龙(1990-),男,高级工程师,研究方向:软件工程。

徐哲炜(1989-),**通信作者**,男,高级工程师,研究方向:深度学习、智能试验评估,E-mail:xuzhewei@intellisoftx.com。

柯文俊(1990-),男,副教授,研究方向:人工智能、自然语言处理。

汪鹏(1977-),男,教授,研究方向:人工智能、知识图谱。

(编辑:张黄群)