文章编号：1003-0077(2024)01-0001-23

大语言模型评测综述

罗 文，王厚峰

（北京大学 计算机学院，北京 100871）

摘 要：大语言模型(Large Language Models，LLMs)在多种自然语言处理(Natural Language Processing，NLP)任务中展现出了卓越性能，并为实现通用语言智能提供了可能。然而随着其应用范围的扩大，如何准确、全面地评估大语言模型已经成为了一个亟待解决的问题。现有评测基准和方法仍存在许多不足，如评测任务不合理和评测结果不可解释等。同时，随着模型鲁棒性和公平性等其它能力或属性的关注度提升，对更全面、更具解释性的评估方法的需求日益凸显。该文深入分析了大语言模型评测的现状和挑战，总结了现有评测范式，分析了现有评测的不足，介绍了大语言模型相关的评测指标和评测方法，并探讨了大语言模型评测的一些新方向。

关键词：自然语言处理；大语言模型；模型评测

中图分类号：TP391 **文献标识码**：A

Evaluating Large Language Models: A Survey of Research Progress

LUO Wen, WANG Houfeng

(School of Computer Science, Peking University, Beijing 100871, China)

Abstract: Large Language Models (LLMs) have demonstrated exceptional performance in various Natural Language Processing (NLP) tasks, providing a potential for achieving general language intelligence. However, their expanding application necessitates more accurate and comprehensive evaluations. Existing evaluation benchmarks and methods still have many short-comings, such as unreasonable evaluation tasks and uninterpretable evaluation results. With increasing attention to robustness, fairness and so on, the demand for holistic, interpretable evaluations is impressing. This paper delves into the current landscape and challenges of LLM evaluation, summarizes existing evaluation paradigms, analyzes limitations, introduces pertinent evaluation metrics and methodologies for LLMs and discusses the ongoing advancements and future directions in the evaluation of LLMs.

Keywords: natural language processing; large language models; model evaluation

0 引言

自 2017 年 Google 提出 Transformer 以来，自然语言处理的研究已逐步统一到这种具有灵活堆叠扩展能力的编解码框架下。特别是，人们可以基于 Transformer 的编码端和解码端，通过无监督的方式，使用大规模数据预训练具有通用语言能力的基础模型，如基于编码端的 BERT[1]、基于解码端的 GPT[2]，以及融入编码和解码结构的 BART[3]、T5[4]等。当这些预训练的基础模型与下游任务适配后，不断地刷新最优结果。为了评估模型的能力，研究人员提出了许多针对这些模型在下游任务上性能表现的评测基准。

预训练语言模型的规模越来越大，参数量从开始的亿级，发展到目前的千亿级甚至万亿级。随着规模的扩大，模型在无须对具体任务适配的情况下，解决下游任务的能力也迅速提升。但与此同时，模型自身的各项能力和属性、应用的局限性、潜在风险及其可控性等仍未得到全面评测和深入研究。由于大语言模型的迅速发展和巨大影响，以及通用性的日益增强，传统基于单一任务的单一评价方法已经

收稿日期：2023-07-19 **定稿日期**：2023-10-30

基金项目：新一代人工智能国家科技重大专项(2022ZD0116308)

无法适应新的评测需求。首先，缺乏广度和深度。面对许多出色的大语言模型，仅在几个已有的基准数据集上往往难以区分它们的优劣。其次，存在数据偏差的问题。许多用于评测的数据集都是从特定的领域或人群中收集，这可能导致模型在基准数据上的表现难以准确反映其在真实应用场景中的性能。再者，忽视模型其他方面的能力或属性评估。先前的评测方法往往只关注模型的性能表现，忽视了对模型其他方面的能力或属性评估。例如，对模型逻辑推理能力的评估、对模型鲁棒性的评估和对模型生成有害内容可能性的评估等。因此，在大语言模型不断发展的同时，模型评估方法也需要进一步研究。

本文首先回顾了自然语言处理中有代表性的评测基准与评估指标，针对大语言模型的评估对评测范式进行了分类，将其分为经典评测范式和新型评测范式，分析了现有评测的不足；再介绍了全面的大语言模型评测思想，以及相关的评测指标和评测方法；最后对目前广受关注的大语言模型评测的一些新方向做了总结。需要说明的是，本文所指的大语言模型并没有严格规定模型规模的大小，凡以预训练为基础具有“通用”能力的语言模型都属于本文所指的大模型。

1 自然语言处理的评测范式

自然语言处理的发展受益于自然语言处理评测。评测通常依赖于一系列的评测基准（Benchmark），模型在这些基准数据集上运行并产生输出结果，评测系统据此返回一个代表模型能力的值。最简单的评测基准由单一任务上的单一数据集构成，这也是常见的自然语言处理基本评测模式。为了全面评估大语言模型，可以将多个数据集聚合和重新组织，形成一个更通用的评测基准。本章针对大语言模型的评估对评测范式进行了分类，将其分为经典评测范式和新型评测范式。表 1 列出了一些典型的评测基准。下面将分别介绍经典评测范式，以及面向多种能力的新型评测范式与现有评测的不足。

表 1 一些典型的评测基准

基准名称	发布时间	发布人/机构	规模	评价方面	语种数
DecaNLP	2018 年	McCann 等	10 个任务	NLU	1
GLUE	2018 年	纽约大学等	9 个任务	NLU	1
SuperGLUE	2019 年	纽约大学等	8 个任务	NLU	1
XTREME	2020 年	卡耐基梅隆大学等	9 个任务	NLU	40
XGLUE	2020 年	微软	11 个任务	NLU	19
CLUE	2020 年	CLUE 团队	9 个任务	NLU	1
GLGE	2020 年	四川大学等	4 个任务	NLG	1
GEM	2021 年	卡耐基梅隆大学等	11 个数据集	NLG	18
CUGE	2021 年	清华大学等	21 个数据集	NLU、NLG	1
EleutherAI LM Harness	2021 年	EleutherAI	200 余个数据集	NLU、NLG	—
BIG-bench	2022 年	谷歌	200 余个数据集	逻辑能力、数学能力、代码理解能力、伦理道德	—
MT-Bench	2023 年	伯克利大学等	80 个样例	与人类偏好对齐程度	1
Chatbot Arena	2023 年	伯克利大学等	—	与人类偏好对齐程度	—
SuperCLUE-Open	2023 年	CLUE 团队	1.2k 个样例	与人类偏好对齐程度	1
C-EVAL	2023 年	上海交通大学等	13948 个样例	知识运用与推理能力	1
SAFETYPROMPTS	2023 年	清华大学等	100k 个样例	安全性	1
LLMEVAL	2023 年	复旦大学等	933 个样例	正确性、流畅性、信息量、逻辑性、无害性等	1
TriviaQA	2017 年	华盛顿大学等	95k 个样例	知识运用能力	1

续表

基准名称	发布时间	发布人/机构	规模	评价方面	语种数
OpenBookQA	2018 年	艾伦人工智能研究所等	约 6k 个样例	知识运用能力	1
GSM8k	2021 年	OpenAI 等	8.5k 个样例	数学推理能力	1
HaluEval	2023 年	人民大学等	35k 个样例	检测幻觉能力	1
MMLU	2021 年	伯克利大学等	57 个数据集	知识运用与问题解决能力	1
HellaSwag	2019 年	华盛顿大学	70k 个样例	常识推理能力	1
HumanEval	2021 年	OpenAI 等	164 个样例	代码生成能力	—
DROP	2019 年	加利福尼亚大学等	96k 个样例	阅读理解和数值推理能力	1

1.1　经典的自然语言处理评测

自然语言处理分为自然语言理解(Natural Language Understanding，NLU)和自然语言生成(Natural Language Generation，NLG)两个大类。但在经典评测范式下都主要关注模型最终输出结果与参考答案的匹配程度。经典评测的结构如图 1 所示。

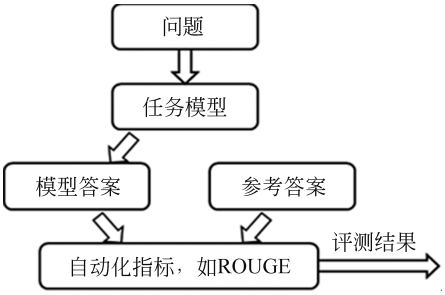

图 1　经典评测的结构

1.1.1　自然语言理解能力评测

常见的自然语言理解任务有情感分析(Sentiment Analysis)、文本匹配(Text Matching)、文本分类(Text Classification)和阅读理解(Reading Comprehension)等。针对具体的任务已有大量的相关评测基准。2018 年，McCann 等人[5]提出了 DecaNLP，试图以统一的问答形式评测 NLU 能力。该基准涉及 10 个任务，与这些任务相关的数据集均以三元组形式表示，如(问题，上下文，答案)。在评测时，给模型输入(问题，上下文)，模型输出“答案”，然后再计算模型答案与参考答案的匹配程度。

纽约大学和华盛顿大学等机构的研究人员提出了评测数据集 GLUE[6]，由 9 个自然语言理解任务组成，包括情感分析、文本蕴含、句子相似性等。随着模型的进一步发展，GLUE 进一步升级为 SuperGLUE[7]。SuperGLUE 在 GLUE 的基础上增加了五个难度更高的评测任务。

上述基准仅限于英语。为了填补跨语言的模型评测空白，卡耐基梅隆大学和谷歌等单位的研究人员提出了 XTREME[8]。XTREME 是一个大规模、多任务、多语言的模型评测基准，涉及 40 种不同的语言，共 9 个任务。几乎与 XTREME 同时出现的 XGLUE[9]也是一个跨语言的模型性能评测基准，由 11 个任务组成，涵盖 19 种语言。

在中文信息处理方面，第一个大规模的中文理解评测基准 CLUE[10]于 2020 年提出，CLUE 中的任务集涵盖了文本分类、阅读理解、自然语言推理等多个中文自然语言理解任务和一个诊断评估数据集，具体包含：长文本分类 IFLYTEK[10]、语义相似度计算 AFQMC[10]、中文命名实体识别 CLUENER[11]、中文自然语言推理 OCNLI[12]、成语完形填空 ChID[10]、小样本(few-shot)测评 FewCLUE[13]和零样本(zerow-shot)测评 ZeroCLUE[13]等。CLUE 提供了一种标准化的评估方式来测评模型的中文理解能力。

1.1.2　自然语言生成能力评测

自然语言生成的典型任务是机器翻译(Machine Translation)、生成式文本摘要(Generative Text Summarization)、自动对话(Dialogue)等。BLEU[14]是评测机器翻译任务中译文质量的一个重要指标。在机器翻译的评测中，每段原文都有一组高质量的参考译文(Reference)，模型生成的译文被称为 Candidate。BLEU 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 的评测得分是一个 0～1 之间的数值，表示生成译文与参考译文的相似程度。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,用于机器翻译生成译文的评估指标还有 METEOR[15]等。

ROUGE[16]是生成式文本摘要任务常见的评测指标,ROUGE 和 BLEU 在计算上非常相似,区别在于 BLEU 更关注精确度,而 ROUGE 更关注召回率。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。

在国内,四川大学和微软的研究人员于 2020 年提出了用于评测生成能力的 GLGE[17]。该基准涵盖了生成式文本摘要、问题生成(Question Generation,QG)、生成式问答(Generative Question Answering,QA)和对话 4 个领域,并且根据难易程度分为三个级别:GLGE-easy、GLGE-medium 和 GLGE-difficult。

SemEval 是一个语义处理国际评测研讨会,目标是推进语义分析技术进步,并帮助创建高质量的标注数据集以应对自然语言语义领域越来越具挑战性的问题。每年的研讨会都包括一系列的共享任务,不同团队设计的计算语义分析系统在这些任务中进行展示和比较。以 SemEval-2022 的任务 9[18]为例,该任务要求模型从英语烹饪食谱和相关视频中回答问题,以此评估模型在表达和推理时具有的语言能力和认知能力。

GEM[19]是一个活跃的自然语言生成评测基准,侧重于通过人类注释和自动化度量对模型的 NLG 能力进行评估。GEM 旨在衡量多语言下各种 NLG 任务的进步,并以数据卡片和模型卡片的方式展示相关的数据集信息和模型评测结果。此外,它还致力于结合自动化度量和人类度量方法制定生成文本评估标准。具体来说,GEM 囊括了 11 个数据集(包括 CommonGEN[20]、Czech Restaurant[21]、DART[22]、E2E clean[23-24]、WebNLG[25]、WikiLingua[26]等),涉及英语、西班牙语、土耳其语、俄语、越南语等 18 种语言。

1.1.3 同时考虑理解和生成的能力评测

随着大语言模型的迅速发展及其在下游任务上的广泛应用,仅仅局限于评估模型某一种能力的评测基准逐渐无法满足评测需求。在这种背景下,许多新的更为全面的评测基准不断推出。这些评测基准的一个重要特点就是它们通常会聚合多个数据集、多个任务以及多个评测指标来对模型进行更全面的能力评测。

为了更好地对通用语言能力进行基准测试,北京大学、清华大学以及北京智源人工智能研究院等研究机构联合提出了一个评估汉语理解和生成能力的评测基准 CUGE[27]。CUGE 在语言能力-任务-数据集层次框架中选择和组织数据集,涵盖了 7 种重要的语言功能,包括:字句级别的语言理解能力(Language Understanding: Word-sentence Level)、语篇级别的语言理解能力(Language Understanding: Discourse Level)、信息获取和问答能力(Information Acquisition and Question Answering)、语言生成能力(Language Generation)、对话式交互能力(Conversational Interaction)、多语言能力(Multilingualism)和数学推理能力(Mathematical Reasoning);在这 7 种语言功能下进一步细分到 18 个主流 NLP 任务,包括:命名实体识别(Named Entity Recognition)、实体关系抽取(Entity Relation Extraction)、语法纠错(Grammatical Error Correction)、阅读理解(Reading Comprehension)、开放领域式问答(Open-domain Question Answering)和机器翻译(Machine Translation)等;再根据相应的 NLP 任务挑选出 21 个数据集,例如,在命名实体识别任务中使用了 CMeEE 数据集[28],在语法纠错任务中使用了 YACLC 数据集[29]。该框架是根据人类语言考试大纲和目前的自然语言处理研究现状精心设计的。

在国外,为了解决大语言模型的量化评估及评估结果复现问题,EleutherAI 提出了 EleutherAI LM Harness[30],这是一个针对自回归大语言模型(Autoregressive Large Language Models)的统一基准测试框架。它涵盖 200 多个数据集,支持包括但不限于 GPT-3[31]、GPT-NeoX[32]和 GPT-J[33]等自回归大语言模型。同时,为了保证评测结果可复现,基于 EleutherAI LM Harness 的评测提供统一的评测接口和对应评测任务的版本控制。

1.2 面向多种能力的新型评测范式

与经典评测范式不同,新型评测范式不仅关注大型语言模型在理解和生成方面的能力,同时也关注模型本身所表现出的更多重要属性。例如,模型生成的内容是否符合社会道德准则。新型评测范式使得研究者能够从更多维度和更深层次去理解和评估自然语言处理模型的性能,从而推动自然语言处理技术的进一步发展和完善。

1.2.1 多种属性的能力评测

为了追踪大语言模型的规模对模型表现的影

响，探究大语言模型本身是否存在基础性能力和属性上的缺陷，Google 聚集 442 名研究人员耗时两年，于 2022 年发布了评测基准 BIG-bench[34]。该基准涵盖 200 多个数据集，分为 9 个主要方向，分别为：传统自然语言处理任务（Traditional NLP Tasks，包括自然语言理解任务和自然语言生成任务）、逻辑、数学和代码（Logic，Math，Code）理解、对世界的理解（Understanding the World）、对人类的理解（Understanding Humans）、对科学技术的理解（Scientific and Technical Understanding）、与模型的交互机制（Mechanics of Interaction with Model）、针对通用语言模型的能力短板（Targeting Common Language Model Technical Limitations）、行为是否符合既定社会道德准则（Pro-social Behavior）以及其他（Other）。对于尚未包括在评测基准里的任务和数据集，BIG-bench 支持研究者提交和更新，这使得评测基准能够随着大语言模型的发展而同步发展，为更加全面地评测大语言模型提供了更多可能。

除了评估大语言模型核心的基础能力外，还存在一些衡量这些模型与人类偏好的对齐程度的评测基准。其中，MT-Bench[35] 和 Chatbot Arena[35] 是两个常用的评测基准。MT-Bench 是一个包含 80 个手工编写的高质量开放式多轮问题的评测基准，其目标是评估大型语言模型在多轮对话和指令遵循方面的能力。该基准涵盖了 8 个常见的人机交互场景，包括：写作、角色扮演、信息提取、推理、数学、编程、自然科学知识和人文社会科学知识。针对每个场景，研究人员精心编写了 10 个多轮问题，用以评估大语言模型在面对这些问题时与人类偏好的一致性。与 MT-Bench 不同，Chatbot Arena 是一个众包匿名基准测试平台。在该平台上，用户可以同时与两个匿名的大语言模型进行交互。用户可以自由地向这两个模型提出相同的问题，并根据个人偏好评价它们的回答。在 Chatbot Arena 开始运作的一个月内，研究者们便收集到了约 30 000 条评测数据，这种众包方式为大量用户动态参与提供了可能，增强了评测结果的广泛覆盖性和多样性。

在中文方面，2023 年 5 月 9 日，SuperCLUE-Open 评测基准正式发布，这是一个评估大语言模型中文对话能力和遵循指令能力的评测基准，包含 1 200 道中文的高质量多轮问题。该基准不仅包括一些普通的常规使用场景，还设计了一些具有挑战性的指令以增加不同模型的区分度。它考察了模型的十大能力，包括：语义理解与抽取、闲聊、上下文对话、角色扮演、知识与百科、生成与创作、代码、逻辑与推理、计算、代码和安全。每个子能力有 60 个题目，每个题目包括两轮问题，从中文语境下与人类偏好的对齐程度方面对大语言模型进行了评估。

C-EVAL[36] 是一个综合的中文评测基准，旨在评估中文语境下大语言模型的知识运用与推理能力，为研究者理解和评估中文语境下的大语言模型能力提供了重要的工具和资源。该评测基准总共包含 13 948 个多项选择题，涵盖了中学、高中、大学和专家四个难度级别。这些题目来自 52 个不同的学科领域，包括人文科学（例如，中国语言文学、艺术学、历史学等）、理工科（例如，高等数学、大学化学、计算机组成、注册电气工程师等）和社会科学（例如，政治学、教育学、工商管理学等）等。此外，研究者还基于 C-EVAL 构建了一个难度更高的评测基准子集 C-EVAL HARD。C-EVAL HARD 中的题目对知识运用与推理能力的要求更高，例如，高等数学题、大学物理考试题等。值得指出的是，为了确保评测数据不被污染，C-EVAL 的题目并非直接从官方的国家考试中选取，而是主要采集自模拟考试和小规模的地方考试。研究者可以通过 C-EVAL 评估中文语境下的大语言模型在各个学科领域和不同难度级别下的表现。

SAFETYPROMPTS[37] 是一个中文大语言模型安全评测基准。该基准从 8 种典型的安全场景和 6 种对抗性的指令攻击场景综合探索了大语言模型应用中的安全性问题。其中，安全场景分别为：侮辱（Insult）、不公平和歧视（Unfairness and Discrimination）、犯罪和非法活动（Crimes and Illegal Activities）、敏感话题（Sensitive Topics）、身体伤害（Physical Harm）、心理健康（Mental Health）、隐私和财产权（Privacy and Property）、伦理和道德（Ethics and Morality）；指令攻击场景分别为：目标劫持（Goal Hijacking）、提示泄漏（Prompt Leaking）、角色扮演（Role Play Instruction）、不安全的话题引导（Unsafe Instruction Topic）、不安全的观点询问（Inquiry with Unsafe Opinion）和逆向曝光（Reverse Exposure）。为了构建 SAFETYPROMPTS 评测基准，研究者们首先根据这 14 个场景人工编写了一个测试数据集，再利用 ChatGPT 对测试数据集进行增广，形成更多的基准数据，最终形成了 10 000 个评测数据。通过 SAFETYPROMPTS 评测基准，研究人员可以较为全面地了解大语言模型在典型安全场景和对抗性指令攻

击下的表现，提升大语言模型的安全性能，减少大语言模型的潜在安全风险。

复旦大学的研究人员提出了一个名为LLMEVAL[①]的中文评测系列，以回答关于大语言模型评估方面、评估方法和排序比较方法的问题。目前已经公开的评测基准包括LLMEVAL-1和LLMEVAL-2。LLMEVAL-1从认知心理学的角度出发，以人类信息处理、思考和问题解决能力为基准，从正确性、流畅性、信息量、逻辑性和无害性五个评估方面构建了一个包含17个大类、453个问题的评测问题集，涵盖了事实性问答、阅读理解、框架生成、段落重写、摘要、数学解题、推理、诗歌生成和编程等多个领域。LLMEVAL-2则以一般用户的日常使用场景为背景，从12个学科（包括，生命科学、化学、汉语言文学、数学、经济学、法学等）出发构建了一个包含480个问题的评测问题集，重点评估了大语言模型在各学科本科生和研究生希望在日常学习和生活中得到帮助的任务上的表现。

除了上述评测基准以外，还存在许多其他用于评估大语言模型的多种能力属性的评测基准。例如，考察大语言模型的知识运用能力的TriviaQA[38]和OpenBookQA[39]、考察大语言模型数学推理能力的GSM8k[40]、评估大语言模型检测幻觉能力的HaluEval[41]等。

1.2.2 模型评测实例——GPT-4的评测

为了凸显GPT-4的总体表现，OpenAI在一系列评测基准上对GPT-4进行了评估[42]。这些评测基准既包含最初为人类设计的模拟考试，也包含在传统自然语言处理任务上用来评估语言模型的评测基准。为人类设计的模拟考试包括：SAT Math、Leetcode等。其中，SAT Math的考察内容是学生在大学和未来生涯期间可能会遇到的数学问题，主要包括：代数核心（Heart of Algebra）、问题求解与数据分析（Problem Solving and Data Analysis）和高等数学基础（Passport to Advanced Math）；Leetcode则主要考察待测者的综合代码能力。上述模拟考试题由多项选择题（Multiple Choice Question）和主观题（Free-Response Question）两种模式组成。传统自然语言处理任务上的评测基准包含MMLU[43]、HellaSwag[44]、HumanEval[45]、DROP[46]等。其中，MMLU是一个涵盖STEM、人文科学、社会科学等57个学科领域（例如，数学、法律、伦理等）的评测基准，旨在考察大语言模型将知识运用于问题解决的能力；HellaSwag关注模型在围绕日常事件的常识性推理方面的能力，包含70k个问题；HumanEval主要考察大语言模型的代码生成能力；DROP则是一个阅读理解与数值推理基准测试数据集，包含96k个问题，用于评测模型在离散推理任务上的表现。这些评测基准不仅关注GPT-4作为一个大语言模型在传统自然语言处理任务上的表现，更关注GPT-4在更高层次问题求解上的能力（例如，推理、知识、语言与理解能力）。评测结果表明，在大多数专业类考试和学术类考试中GPT-4具有与人类相当的表现；而在多个传统的自然语言处理评测基准上GPT-4已经达到了最先进的效果。此外，GPT-4还在评测中展现出了其他方面的能力，例如，处理低资源语言（low-resource language）的能力[42]。研究人员通过Azure Translate将MMLU中的数据翻译成其他多种语言，之后将各个语言版本的MMLU用于评测GPT-4，结果表明GPT-4具有较强的处理其他语言的能力，包括拉脱维亚语（Latvian）、威尔士语（Welsh）和斯瓦希里语（Swahili）等小语种。

以体现人类级别的认知能力与强调和现实世界的紧密联系为原则，微软的研究人员提出了一个以人为中心的评测基准AGIEval[47]，并在其上评测了GPT-4和ChatGPT等大语言模型的表现。与传统评测数据不同，AGIEval中的评测数据来自高标准化、官方的人类考试题，其中包括：研究生入学考试（Graduate Record Examinations，GRE）、学术评估测试（Scholastic Assessment Test，SAT）、中国高考（China College Entrance Exam，Gaokao）、法学院入学考试（Law School Admission Test，LSAT）、美国数学竞赛（American Mathematics Competitions，AMC）、中国公务员考试（Chinese Civil Service Examination）等。与文献[35]不同，为了更加标准和自动地评测大语言模型，AGIEval在题型上删除了所有的主观题，只保留了客观题（包括多项选择和填空）。在AGIEval评测中共有四种设置，即，零样本学习（Zero-shot learning）、小样本学习（Few-shot learning）、零样本思维链（Zero-shot chain-of-thought prompting）和小样本思维链（Few-shot chain-of-thought prompting）。评测结果表明：①GPT-4在LSAT、SAT和数学竞赛中超越了人类的平均表现，在SAT数学考试中达到了95%的准确率，展示了出色性能。②当前的大语言模型（如

① https://github.com/llmeval

GPT-4)在面对需要复杂推理(如 LSAT 分析推理和物理学)或特定领域知识(如法律和化学)的任务时仍然表现不佳。③与先前 GPT-3 系列模型的小样本表现显著优于零样本表现不同,GPT-4 等当前的大语言模型的零样本学习能力开始逐渐接近它们的小样本学习能力。

1.3　现有评测的不足

随着不同通用大语言模型的推出,现有评测及基准的不足开始显现。这使得在应用上如何选择模型以及在开发上如何改进模型都面临极大的挑战。下面简要分析现有评测的不足。

1.3.1　新生任务缺乏相应的评测基准

随着通用大语言模型的迅速发展,需要在更多的应用场景和任务上评测模型的效果。但是,一些新生任务缺乏相应的评测基准。这样,研究者难以了解大语言模型在这些任务上的表现能力,从而制约在该领域的进一步发展。利用评测基准进行评估是衡量模型性能和比较不同模型的重要途径。缺乏评测基准会导致研究人员无法准确评估模型的性能,也难以使许多新生的算法和模型被有效地评估和比较。此外,缺乏评测基准还会影响研究人员对新生任务的理解和定义。因此,建立相应的评测基准对于模型在新生任务上的应用研究至关重要,这也有助于研究者更好地理解大语言模型在新生任务中的应用潜力。

1.3.2　评测任务缺乏区分度

随着大语言模型的发展和规模的不断扩大,其能力也越来越强,以至于它在一些评测任务上的表现已经与人类相当[42],甚至评测结果可以超越人类。在这种情况下,许多原来以较小规模模型为评测目标的评估任务已经逐渐失去了挑战性和区分度,难以为研究者提供有价值的信息。缺乏区分度这一问题不仅是评测基准本身的问题,也反映出了大语言模型发展的一个重要趋势,即现有的大语言模型的发展已经开始超出原有的评估任务的评测范围。因此,需要更加注重评测任务的区分度和难度,以确保评测结果具有实际可参考的意义。

1.3.3　评估方式不公平

在大语言模型的评估中,评估方式的公平性至关重要。然而,目前常用的评估指标和数据集选择存在许多不公平的问题,使得评估结果的准确性和客观性受质疑。例如,当前同一任务下的评测数据集通常有很多,很有可能会产生模型 A 在某个评测数据集上优于模型 B,但是在另一个评测数据集上又劣于模型 B 的矛盾情况。这种情况下,研究者可能只选取有利于自己的结果公布[48]。此外,人为因素也可能导致评估结果的不公平。例如,在人工评测中,评测人员的背景、观点和经验可能影响他们对模型的判断,从而在评测结果中引入人为的偏差;同时,在不同的人工评测过程中,评估标准化程度也可能存在差异,从而进一步削弱了不同模型间的可比性和公平性。

1.3.4　评估不全面

目前,对模型单项能力的评测往往被简化成针对单个任务上的单数据集单指标的评测,无法准确可靠地反映模型在待评测能力方面的强弱[48]。例如,针对自然语言生成能力的评测,需要考察生成文本的连贯性、多样性、幻觉程度和有趣程度等多个方面,但不同方面往往适用不同的评测指标。而且,不同的任务和数据集会涉及不同的语言现象和应用场景,这是单个任务上的单数据集单指标评测有失考量的内容。此外,对模型综合能力的评测大多是单个评测基准的简单聚合,缺乏系统性的交互,也无法全面评估模型的综合能力和多种属性。

1.3.5　评测基准的污染问题

所谓评测基准的污染问题,是指用于评测的数据出现在了模型的训练数据中。为了确保大语言模型评估的公正性和可信度,以及评测基准能够展现的具有一般性的评测结果,评测基准中的测试数据不应当被包含在大语言模型的训练数据中。由于目前的大语言模型是在多个来源的庞大数据集上训练的,研究者很难确定当前使用的评测基准是否泄漏到了模型的训练数据中。这种污染会对评测基准的公正性和可信度产生一定程度的影响。因此,评测基准的构建者需要谨慎考虑以确保评测基准的独立性和代表性;评测基准的使用者也需要注意这一问题。当然,未来大语言模型的研发者应尽可能明确模型在训练时可能存在的污染问题以及污染程度[42]。

1.3.6　评估结果缺乏可解释性

在大语言模型评测中,评测结果的可解释性常常被忽视。现有评测基准通常依赖某个数字指标来概括模型的表现,缺乏对评估过程的解释和分析。这种评估方式虽然可以快速了解不同模型的表现,却难以解释模型表现好坏的原因,也就难以对模型进行有效诊断,进而难以有针对性地对现有模型进行改进和优化。可解释性的缺失主要表现在以下两个方面。第一,评估结果的数字化方式使得研究人

员难以全面了解模型在评测任务中的行为，也就无法直接对模型的优劣进行深入的分析与解释。第二，现有的评测基准往往是针对特定的应用场景和任务设计的，限制了评测结果的可迁移性和可解释性，难以被推广到其他应用场景和任务中。

2 全面的大语言模型评测

随着大语言模型的影响越来越广泛，如何更好地评测模型已经成为研究界关注的热点问题。一项代表性的工作就是 Liang 等人[49]提出的语言模型的全面评估(Holistic Evaluation of Language Models，HELM)方法。

HELM 的出发点是在多个场景、任务和评估指标下评估大语言模型的能力。HELM 首先对自然语言处理涉及的众多场景和任务进行了分类和筛选，并以应用性的任务作为评测重点，基于可行性和全面性从当前主要的评测数据中选择了一部分用于大语言模型的评测。其次，明确了大语言模型评估里需要考虑的 7 个评测指标(如准确率)，同时又设计了 7 个更具针对性的评估维度(如语言能力、推理能力等)。最后，HELM 对 30 个大语言模型(包括 BLOOM[50]、GPT-3、GPT-NeoX、GPT-J、GLM[51]等)在 42 个场景和上述评测指标下进行了评测，并公开了评测结果。HELM 也指出了其评测中存在的遗漏和不足，例如部分场景和任务的缺失、部分评估方法的不足、部分模型和适配策略的遗漏等。

由于不少大模型不再开源(如 ChatGPT)，全面评测大模型存在一定困难。HELM 为了模拟现实中人们通过 API 访问大语言模型的情形[14]，在评估中将大语言模型视为黑盒，这也是上述提及此次评估中的遗漏和不足之一。

下面结合 HELM 用到的评测属性对其分别进行介绍，包括：准确率(Accuracy)、校准度(Calibration)、泛化(Generalization)能力、适配(Adaptation)能力、鲁棒性(Robustness)、效率(Efficiency)、偏见和刻板印象(Bias and Stereotypes)、公平性(Fairness)和有害性(toxicity)。

2.1 准确率

准确率是指模型预测或生成结果的正确比例。一个准确率高的大语言模型能够更好地处理自然语言的相关任务，并提供更准确的预测和生成结果。大语言模型的准确率对于其在具体任务中的应用至关重要。

准确率的评估方法因场景和任务而异。常见的指标有：判别类问题的评测指标，如 F_1(包括 Micro F_1 和 Macro F_1)值和 Accuracy 值；生成类问题的评测指标 BLEU(主要用于机器翻译结果评测)和 ROUGE(主要用于文本摘要结果评测)；判别类问题和生成类问题都用到的精确匹配(Exact Match，EM)；检索类问题常用的 Reciprocal Rank[52] 和 Normalized Discounted Cumulative Gain [53]等。

准确率指标在自然语言处理的评测中广为使用，在很长一段时间里几乎成为模型评测的唯一指标。在今后仍将是重要的指标。

2.2 校准度

准确率衡量的是模型输出结果的正确性，而校准度[54-56]则是衡量模型对输出结果赋予的概率的准确性，也就是模型在预测时给出的置信度(confidence)对真实概率分布进行估计的准确性。

大语言模型的校准度评估是十分有意义的。首先，有助于提高模型的可靠性。在一定程度上，校准度越高，模型的预测结果就越可靠。如果一个大语言模型的校准度低，它的预测结果就更有可能会导致误解和错误的决策。其次，有助于改善置信度估计。在实际的应用场景里，大语言模型的使用通常会伴随着对预测结果的置信度估计。如果模型的校准度很高，置信度估计一般也会更加准确。这样，校准度可以更好地帮助使用者理解模型的预测结果并在必要的时候(例如当模型对预测结果的置信度很低时)进行人工介入。

下面介绍一种常见的校准度评估方法，即期望校准误差(Expected Calibration Error，ECE)[57-58]。ECE 表示模型认为输出正确的概率与模型输出实际上正确的概率之差的绝对值期望。这里介绍一种有限数据情况下的 ECE 计算方法。

首先需要估计期望准确率。将[0，1]概率区间均分成 M 个小区间，每个区间称为一个桶，每个桶的长度均为$\frac{1}{M}$。之后将所有的数据样例按照模型预测的概率放入对应的桶中。若记 B_m 是第 m 个桶中所有样例的集合，因此 B_m 的准确率如式(1)所示。

$$\mathrm{acc}(B_m)=\frac{1}{|B_m|}\sum_{i\in B_m}1(\hat{y}_i=y_i) \tag{1}$$

其中，$\hat{y}_i$ 和 y_i 是样例 l 的预测标签和真实标签。在此基础上，再定义 B_m 的平均置信度(Average Con-

fidence)如式(2)所示。

$$\mathrm{conf}(B_m)=\frac{1}{|B_m|}\sum_{i\in B_m}\hat{p}_i \tag{2}$$

其中，$\hat{p}_i$ 是模型对样例 i 的置信度，即模型赋予样例 i 的预测概率。$\mathrm{acc}(B_m)$和 $\mathrm{conf}(B_m)$之差即估计了桶 B_m 的校准差距(calibration gap)。可以按式(3)计算每个桶对应的校准差距的期望平均，即 ECE。

$$\mathrm{ECE}=\sum_{m=1}^{M}\frac{|B_m|}{n}|\mathrm{acc}(B_m)-\mathrm{conf}(B_m)| \tag{3}$$

其中，n 表示样例的总数目。

2.3　泛化能力

模型泛化能力的评估重点集中于模型在与训练集数据分布不同的域外数据集上的表现。一般来说，泛化能力的评估是在小样本(few-shot)或零样本(zero-shot)设置下进行的[31]。

小样本是指模型在预测时，仅给模型少量的样例作为模型推理的参考。在这个过程中，模型的参数通常不作更新。典型的小样本法是给出 k 个由问题、文本和对应的答案作为相关任务的实例，然后再给出一个真正需要解答的问题和文本，希望模型参照所给的样例输出合理的答案。当前广受关注的上下文学习(In-Context Learning)就属于这种情况。

零样本与小样本类似，不同在于零样本不提供参考样例，只给模型提供需要解答的问题和对应的文本，由模型直接推理出答案。这种方法在应用场景下变得更加简单，但同时也存在一些问题，比如，有时零样本设置可能会导致不清晰的任务定义，从而影响模型的表现[31]。

由于小样本和零样本通常在下游任务上不对模型参数进行更新，所以这种评估方式能够较好地体现模型的泛化能力。泛化能力也在一定程度上预示着模型应用于下游任务时的效果。因此，泛化能力的评估是评估大语言模型能否广泛应用于诸多实际下游应用场景的关键之一，也将成为未来大型模型评估的一个重要组成部分。

2.4　适配能力

目前的大模型普遍强调通用性。虽然通过小样本或零样本可以增强通用模型在具体任务上的能力，但比起在特定任务上经过训练的模型，在该任务上不一定具有优势。因此，需要考虑大模型在下游具体任务上的适配(adaptation)问题。适配是指将原始模型转换成一个适用于下游具体任务的过程；模型的适配能力则是指面对不同的适配策略，模型在具体任务上的性能优劣。适配策略分为三种类型：不更新原模型参数的适配[59]、增加适配层并调整适配层参数的适配[60]，以及对原模型做全参数更新的适配。

在不更新模型参数的适配中，最典型的方法就是通过设计提示(Prompt)和上下文例子(In-Context example)使模型在下游任务上获得更好的效果。提示的作用是提醒模型补充“答案”，这种方式类似于预训练模型时对掩码(Mask)部分的预测或后续内容的生成。以这种方式进行推理与模型预训练的方式一致，减少了推理和训练时形式上的鸿沟(Gap)。但如何选择合适的提示形式非常重要。大量的研究表明，提示形式的轻微变化会导致模型输出结果的明显不同。

增加适配层并调整适配层参数的适配是一类高效率、低损耗的适配方法。这类方法的目标是在保证模型性能的情况下，尽量减少优化迭代的次数，甚至不更新原模型的参数。例如，Houlsby 等人[61]在原有的模型架构上添加只含有少量参数的适配层，即在适配下游任务时，固定原模型本身的参数，而只基于梯度更新适配层的参数，从而缩小更新参数的规模，这也使得原始模型的参数在不同任务中可以共享而不发生变化。

一种极端的适配方式是更新模型的全部参数。具体而言，就是利用下游任务中的数据对模型进行再训练，从而迭代更新整个模型的参数。这种调优方法在之前的模型(如 BERT)规模不足够大时经常使用。但随着模型规模越来越大，重新迭代更新模型所有参数的成本也越来越高，这种方法的实用性也逐渐降低。

需要说明的是，模型对不同适配策略的适配程度与模型的结构设计、预训练方式等因素有关。同一个模型在不同的适配策略下的表现也可能十分不同。从这个角度看，评估模型的适配能力的主要任务之一是在特定类别的任务下研究最适合该模型的适配策略，并探索模型在不同适配策略下产生性能差异的原因。

2.5　鲁棒性

虽然大语言模型在很多任务上的性能越来越出色，甚至在一些数据集上超越了人类的表现，但如果

数据受到轻微的扰动，仍有可能会导致模型性能的大幅下降。特别是，当现实世界比较复杂时，模型的表现可能并不突出[62-65]，这便是模型的鲁棒性不强。鲁棒性用于衡量模型对于输入数据中的扰动或者噪声的抵抗能力。目前，模型鲁棒性的评估方法之一是对文本输入进行扰动，然后观察模型输出的变化。这些扰动大致可以分为两类：对抗扰动（Adversarial Perturbations）[66-69]和非对抗扰动（Non-adversarial Perturbations）[70]。

对抗扰动是指为了误导模型做出错误的预测而故意对输入内容进行修改。尽管这些扰动不会引起人的判断变化，但它们对模型的预测结果会产生明显影响。相比之下，非对抗扰动则是对输入内容更自然和随机的改动。这类扰动并不是刻意用来使模型出错的，而是用于模拟现实世界中输入的复杂情况。

对抗扰动可以用来评估模型对恶意输入的处理能力，而非对抗扰动，可用于衡量模型在现实世界中面对有自然误差的输入时的表现。在评估大语言模型时，需要综合考虑这两种扰动类型的影响，以更全面地评估模型的鲁棒性。

2.6 效率

对于大语言模型而言，效率是一个重要的维度。效率可以分为训练效率和推理效率两个方面。训练效率指模型在训练时的复杂程度，而推理效率则是指模型在不更新参数的情况下的推理复杂度。

针对模型效率的评估指标有多种，如训练时的能量消耗和二氧化碳排放量[71-72]、参数个数[73-74]、FLOPS（运行给定实例模型所需的操作数）[74-77]、实际推理时间[78-79]、执行层数（模型实际推理时输入经过的总层数）[80-81]等。对这些指标的评估可以帮助研究人员选择最合适的模型来满足具体的应用需求。

2.7 偏见和刻板印象

大语言模型通常会应用于多种不同的下游任务，而其中潜在的偏见和刻板印象可能会使它在下游任务中表现出歧视行为[72]，从而限制其在一些领域的应用。

与代表型损害（Representational Harm）[82]对应，本文中的大语言模型偏见和刻板印象指的是针对某个群体和某类属性标签产生的过于笼统且不合事实的概括性观点[83-84]，例如，认为男性天生更擅长数学。目前，评估模型中的偏见和刻板印象的方法主要分为两类：基于表示端的评估方法和基于生成端的评估方法。

基于表示端的评估方法主要利用词向量在语义向量空间中的几何关系表征词汇间的关联程度，从而反映语言模型中的偏见和刻板印象[82,85-89]。其中，上下文嵌入关联测试（Contextualized Embedding Association Test，CEAT）[89]通过待测群体词向量与两组属性标签词向量间的相似度差距来表征待测群体偏向某类属性标签的程度，即刻板印象的程度。以种族偏见为例，两组属性标签分别为“友好、勤劳、有才华”和“冷漠、懒惰、无能”。CEAT首先计算待测群体词向量与两类属性标签词向量的余弦相似度，然后计算这两组相似度的差值，之后再通过统计方法计算效应量（Effect Size）来量化上述差值。效应量的符号代表了偏见的方向（正向偏见或负向偏见），而效应量的绝对值表示偏见程度的大小。然而，由于基于词向量，这类评估方法通常并不能很好地适用于闭源大语言模型。

基于生成端的评估方法侧重于利用模型的生成来衡量其偏见程度[49,84,90-95]。常见做法包括：①利用模型生成内容的统计信息。例如，计算生成内容中不同群体和属性标签的共现频率来反映不同群体与该属性标签的关联程度[49,96]。②利用模型生成过程中给出的概率分数进行估计[84,91,98]。例如，自诊断方法（self-diagnosis）[91]通过设计模板来询问模型生成内容中是否包含偏见成分，并利用模型输出补全时的概率分数估计偏见程度。

上述评测方法通常需要依赖人工筛选的词表集合来代表某个待测群体或某类属性标签。但是研究表明，这些由人工筛选的词表本身可能会引入筛选者的固有偏见[98]；此外，词表中的词汇组成也会对评测结果产生较大的影响[99]。目前，NLP社区对于偏见的评估仍然存在一些问题，例如偏见的界定标准模糊不清[100-101]，某些评估方式与模型在下游应用上表现的相关性并不明确[90,102-104]，除性别、种族外对其他形式的偏见（如宗教、国家等）研究较少，非英语语境下的偏见评估尚缺乏相关研究等。未来，大语言模型研发者需明确模型的预期使用场景，最小化模型在不适合的场景中的应用，并提高模型透明度[105]以减轻偏见在大语言模型实际使用时可能造成的社会危害。

2.8 公平性

随着大语言模型在下游任务中的准确率不断提

高，模型的公平性问题也逐渐受到关注。与分配型损害(allocational harm)[82]对应，公平性更多关注模型在特定下游任务中针对不同特征群体的性能差距[82,102,106-108]。相对而言，偏见和刻板印象是指大语言模型内部的某种固有属性(intrinsic biases[72]，内在偏见)；而公平性则关注实际任务中模型在特征群体间的表现差距(extrinsic harms[72]，外在伤害，通常反映为不同群体间准确率的差距)。例如，机器翻译中某些语言的翻译质量明显低于其他语言；语音识别系统在识别非洲裔美国方言时可能会有更低的准确率[108]。目前，模型公平性评估可以分为三类：预测公平性(Predictive Parity)[109]、机会平等性(Equality of Opportunity)[110]和反事实公平性(Counterfactual Fairness)[111]。

预测公平性和机会平等性评估需先将数据集按群体特征(如性别、语言等)划分为细粒度子群体，然后计算模型在各子群体上的统计量，再将统计量通过聚合函数汇总为最终评估分数。具体来说，预测公平性计算子群体数据上的精确度，机会平等性则计算召回率。聚集函数的选择是多样化的，其本质是反映各个子群体数据集上对应统计量之间的差距。例如，记 S_i 表示第 i 个子群体对应的统计量(如精确度或召回率)，$\overline{S}$ 表示所有子群体统计量的均值，n 是子群体总数，Shen 等人[112]以绝对值差距来反映模型的公平性，如式(4)所示。

$$f=\frac{1}{n}\sum_{i}\mid S_i-\overline{S}\mid \tag{4}$$

而 Lum 等人[113]则计算样本方差来估计模型的公平性，如式(5)所示。

$$f=\frac{1}{n-1}\sum_{i}\mid S_i-\overline{S}\mid^2 \tag{5}$$

上述评估方法通常依赖数据集对子群体信息的预先标注，因此在无预先标注的数据集上通常难以发挥很大的作用[49]。

反事实公平性评估通过对测试样例进行扰动生成反事实数据[114]，然后评估模型基于反事实数据的性能[115-116]。与鲁棒性评估类似，其难度主要在于选择扰动时机和扰动位置[49,117-118]。

随着大语言模型的不断发展，其能力范围和应用形式可能从单语言、单模态逐渐转向多语言、多模态。因此，现有的基于单语言(主要为英语)、单模态、数据标注依赖的公平性评测范式需要进一步迭代，以适应未来更广泛的群体特征及更复杂交融的语言背景[107]。

2.9　有害性

大语言模型的有害性是指模型产生有害言论的能力。当大语言模型部署于社交媒体或互联网时，这种模型产生的有害言论很容易造成不良的社会影响。目前，对大语言模型的有害性评估方法之一是使用有害性检测系统检测文本中可能含有的有害成分(包括大语言模型生成内容中的有害成分)。具有代表性的系统包括 HateBERT[119]和 Perspective API[120]等。

当前，有害言论的定义并没有统一标准，不同群体可能会有不同的理解。因此，开发有害性检测系统时，研发者需要谨慎地考虑多方面的问题，包括系统设计的合理性、数据集标注的准确性和是否存在偏见等。同时，研发有害性检测系统的一个主要挑战是在准确率和公平性之间取得平衡，避免对某些群体的过度惩罚或忽视对他们的有害言论。在这个意义上，研发者应提高系统及其数据的开源性和透明度，以便对系统进行全面评估。这种对检测系统本身的全面评估将有助于提高系统的可信度和有效性，进一步增强有害性评测的准确度和公平性。

3　大语言模型评测的一些新方向

自 ChatGPT 推出以来，生成式大语言模型影响越来越大，与此同时，传统的生成式评测方法又面临巨大的挑战。研究者们开始探索新的评测模式。在这一过程中，涌现出了一些有影响的研究，例如基于模型的评测、幻觉问题的评测和元评测(对评测指标本身进行评估)。这些研究进一步弥补了传统评测的不足，并为评价模型性能(尤其是模型在自然语言生成任务上的性能)提供了更加精准、稳定和可靠的评估结果。下面介绍这三个研究方向以及相应的研究进展。

3.1　基于模型的评测

为了讨论方便，本文将任务中的原文(Source)称为原文本，将任务模型的输出(Hypothesis)称为待测文本，将参考答案文本(Reference)称为参考文本。在自然语言生成领域，早期的自动化评测方法如 BLEU 和 ROUGE 主要基于“形式匹配”。这些方法虽然在某种程度上取得了一定的效果，但同样也存在以下不足：①对语义的忽视。在许多情况下，生成文本可能使用不同的词汇或短语来表达相

同的语义。但是这些方法主要关注词汇表层的形式匹配，容易忽略语义的重要性，导致评测结果不能完全真实地反映模型性能。②对参考文本的依赖。由于需要参考文本作为对照，这些评测方法的评测结果往往受参考文本质量的影响。此外，这些评测指标通常假设存在一个或几个“最优”的参考文本，这在许多NLG任务中并不成立。例如，在开放式对话等任务中，可能存在多种合理但完全不同的生成结果。这种假设限制了这些评测指标在评估生成多样性和创新性方面的能力。③难以抓住不同任务间的细微差别及各个任务上的评测需求。例如，摘要和对话生成这两种任务在语义连贯性、文本多样性和创新性等方面的评测需求可能大相径庭，但是这些差异往往很难被这些只关注表层的精确匹配的自动化评测方法捕捉。

上述局限性使得先前的自动评估指标通常难以准确地评估大语言模型的性能和表现。为了克服这些局限性，研究者开始探索基于模型的评测方法，尤其是基于大语言模型的评测方法。这类方法使用预先构建的评估模型对任务模型进行评测。相比早期的传统评测方法，这些评测模型具有更加强大的表示学习能力和语义理解能力，其中的一些方法也不需要依赖参考文本，并能更好地捕捉到不同生成任务之间的细微差别，与人类评测之间也往往有更好的相关性，为评估大语言模型在自然语言生成任务中的表现提供了更为准确和全面的评价标准。基于模型的评测方法有很多，例如，BERTr[122]、BERTScore[123]、MoverScore[124]、BERT for MTE[125]、COMET[126]、BLEURT[127]、RoBERTa-eval[128]、BARTScore[129]、MAUVE[130]、DiscoScore[131]和基于大语言模型的评测[132-135]等。下面将重点介绍几种有代表性的基于模型的评测方法，分别是依赖参考文本，基于BERT的BERTScore、BERT for MTE与不依赖参考文本，基于大语言模型的GPTScore[132]、Kocmi & Federmann[133]以及PandaLM[135]。

3.1.1 BERTScore

BERTScore是一种基于BERT的评测方法，计算结构如图2所示。其核心思想是利用BERT的词嵌入来计算待测文本中的每个token与参考文本中的每个token的余弦相似度。它首先利用BERT的词嵌入来得到参考文本和待测文本的编码向量，分别记为$(r_1, r_2, \cdots, r_n)$和$(p_1, p_2, \cdots, p_m)$，然后通过待测文本中每个token与参考文本中每个token的余弦相似度来计算精确度与召回率，具体如下：

$$P_{\text{BERT}} = \frac{1}{m}\sum_{j=1}^{m}\max_i r_i^T p_j \tag{6}$$

$$R_{\text{BERT}} = \frac{1}{n}\sum_{i=1}^{n}\max_j r_i^T p_j \tag{7}$$

然后根据精确度和召回率来计算F_1值，如式(8)所示。

$$F_{\text{BERT}} = \frac{2}{\frac{1}{P_{\text{BERT}}} + \frac{1}{R_{\text{BERT}}}} \tag{8}$$

BERTScore即取式(8)计算得到的F_1值。

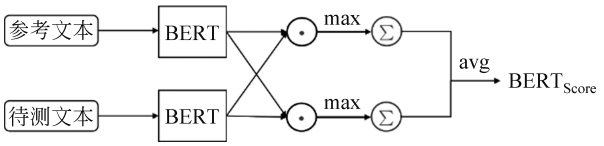

图2 BERT$_{\text{Score}}$的计算结构

3.1.2 BERT for MTE

另一种基于BERT的评测方法是BERT for MTE，该方法通过句子对编码的方式同时编码待测文本和参考文本，并使用基于MLP的回归模型得到最后的指标分数。记参考文本和待测文本的单词序列分别为r和p，BERT for MTE首先利用BERT进行句子对编码，如式(9)所示。

$$v = \text{BERT}([\text{CLS}]; p; [\text{SEP}]; r; [\text{SEP}]) \tag{9}$$

之后再将句子对的嵌入表示送入多层感知机(Multilayer Perceptron, MLP)回归模型中得到最后的指标分数，如式(10)所示。

$$\text{Score} = \text{MLP}(v_{\text{CLS}}) \tag{10}$$

式(10)计算得到的分数即为最终指标值。

3.1.3 GPTScore

GPTScore是一种基于大语言模型的评测方法。其核心在于给定指令和原文本后，经过预训练的大语言模型会对更高质量的生成内容赋予更大的生成概率。具体来说，给定一个生成任务指令d(如“请为以下文本生成一个摘要”)，该任务关注的评估角度a(如流畅度)以及上下文信息S(可以是原文本或参考文本)，GPTScore首先将三者通过提示模板的方式组织成输入文本，然后将GPTScore定义为大语言模型生成待测文本p的加权对数概率和，如式(11)所示。

$$\text{GPTScore} = \sum_t w_t \log P(p_t \mid p_{<t}, T(d, a, S)) \tag{11}$$

其中，$T(\cdot)$是提示模板，用于组织评估的实例，它通常任务相关，并通过提示工程人工构造。

3.1.4　Kocmi & Federmann

与 GPTScore 类似，Kocmi & Federmann 尝试利用大语言模型来对其他的模型进行评估。与 GPTScore 依靠大语言模型给出的概率计算得分不同，Kocmi & Federmann 尝试以一种更加拟人化的形式利用大语言模型进行生成任务上的评估。具体来说，Kocmi & Federmann 利用提示工程将指令 d（如“请评估下面句子的翻译流畅度”）、上下文信息 S（可以是原文本或参考文本，如，需要翻译的原文本）和待测文本（如某个任务模型输出的翻译文本）组织成与人类评估相近的模板形式作为预训练大语言模型的输入，然后让大语言模型直接输出对应的评分，并将这个评分作为该任务的指标分数。

3.1.5　PandaLM

与 GPTScore 和 Kocmi & Federmann 对单个模型的生成内容给出一个绝对的评价不同，PandaLM 是一种基于比较的评测模型。PandaLM 由 LLaMA-7B[136] 调优得到，专注于在指令调优的语境下根据生成内容在各种候选任务模型中选出最优秀的模型。如图 3 所示，PandaLM 接收一个任务的描述，包括指令和与任务相关的输入，再同时接收两个任务模型在这个任务描述下的生成内容，最后给出对哪个任务模型的生成内容更好的评判，并给出评判的原因。

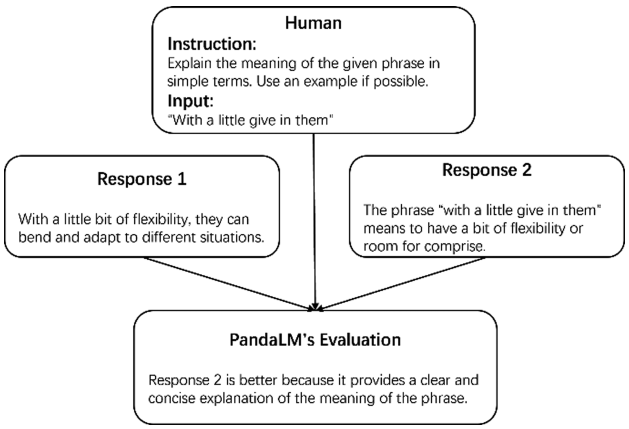

图 3　PandaLM 的评测结构图

由于 PandaLM 评测方法基于不同模型之间的比较，在一定程度上摆脱了对参考文本的依赖。同时，利用大语言模型的泛化能力，PandaLM 超越了传统评估方法主要针对客观正确性的限制，能够通过对指令的设计更好地抓住不同生成任务上对评测需求的微妙差异，如简洁性、清晰度、全面性、正式性等。此外，PandaLM 还可以同时识别和纠正任务模型生成内容中可能存在的逻辑谬误、冗余、语法不准确和上下文不相关等问题，具有较好的鲁棒性。相较于先前的传统自动化评测方法，基于模型的评测方法，特别是基于大语言模型的评测方法，在无参考文本的自然语言生成任务的评估上具有巨大潜力[121]。

下面列出了一些未来可能的基于模型评测的研究方向：

（1）更具鲁棒性的指标。随着现有模型鲁棒性的不断提高，研究者可以开发更具鲁棒性的基于模型的评测指标，以降低噪声对评测结果的影响，从而提高评测结果的稳定性和可靠性。

（2）更可靠的评测方法。虽然大语言模型广泛用于评估生成文本的质量，并展现出了较好的效果[132-135]，但研究表明，基于大语言模型的评测方法同样存在不公平、不可靠的问题[35,121]，例如顺序偏见（大语言模型对不同的位置有特定的偏好）和冗长偏见（大语言模型倾向于偏爱更加冗长的回答，即使这些回答不如更短的回答清晰或准确）等。因此，未来的研究可以进一步发展更加可靠的基于模型的评测方法，增强评测结果的可信度。

（3）知识增强的评测方法。大语言模型在一般场景下可以保持较好的泛化性，但在需要特定知识的专业领域可能表现不佳。基于大语言模型的评测方法也类似：尽管大语言模型在广泛的训练数据上进行了训练，但由于缺乏某些专业知识，它可能仍然无法在专业性较强的领域做出合理准确的评价。然而，如何构建知识增强的大语言模型仍然是一个开放的研究问题[121]：一种方法是将特定领域的数据纳入大语言模型的训练语料中，以便它能够更好地理解和应用该领域的知识；另一种方法是结合外部知识库或专家系统，将其与大语言模型联合使用，以获取该领域专业性的评估能力。未来的研究可以探索将特定知识注入到大语言模型中的方法，从而提高基于大语言模型的评测方法在某些专业领域的表现。

（4）细粒度评估与可解释性增强。过去的许多基于模型的评测方法通常关注生成文本的整体质量，较少关注生成内容中更细维度的质量水平[137]，例如充分性、冗余度、忠实度和趣味性等。由于缺少各个细粒度方面的评价分析，导致在一定程度上缺乏可解释性。未来基于模型的评测研究可以关注评测模型在生成内容的更细粒度划分上的评估方式及可解释性。

（5）摆脱对参考文本的依赖。自然语言生成任务的评测方法通常可以分为两类：需要参考文本的

评测方法和不需要参考文本的评测方法。由于大多数生成式任务具有不确定性和开放性,任务答案往往多样且难以枚举,参考文本通常有限,这就导致需要参考文本的评测方法难以捕捉生成内容的多样性,影响评测结果的准确性。相比之下,无参考文本的评测方法无须枚举可能的答案,在实现对生成内容的多角度、多方面及定制化的评估上有着巨大潜力。未来研究可进一步探索如何利用大语言模型的零样本或小样本泛化能力来摆脱生成式任务评测中对参考文本的依赖,从而获得更易泛化和迁移的评测方法、评测指标和更准确的评测结果。

(6) 人机协作评测。在自然语言生成评测中,人类评测通常被认为是最重要、最准确的评测方法之一。但由于人类评测的时间和资源消耗较大,在模型研发阶段,研究者往往难以利用人类评测实时监测任务模型的能力变化。利用基于模型的评测作为辅助,尤其是基于大语言模型的评测,可以在一定程度上缓解纯人类评测中存在的上述问题。未来研究可尝试提出结合基于模型评测和人类评测的有效方式,从而提高人类评测的可用性和基于模型评测的准确性。

3.2 幻觉问题的评测

随着生成式大语言模型的发展和应用日益广泛,其产生的文本在质量和流畅性上已经达到了十分可观的水平。但模型在生成内容时也可能会产生一种被称为"幻觉"的现象,即生成的文本包含不准确或无根据的信息。这种现象会对模型的实用性和可靠性产生较大的负面影响。因此,越来越多的研究开始集中于幻觉评测。

幻觉是指自然语言生成模型产生的内容不忠实于原文本或不符合现实世界的现象。根据能否通过原文本直接进行验证,幻觉可以分为两类[129]:内在幻觉(Intrinsic Hallucinations)和外在幻觉(Extrinsic Hallucinations)。内在幻觉是指能够直接通过原文本证伪的幻觉现象。以文本摘要任务为例,原文本中包含"苹果公司今天发布了新的iPhone,具有更强大的处理器和摄像头",而待测文本中包含"苹果公司今天发布了新的iPad,具有更强大的处理能力和改进的摄像头",这就是一个内在幻觉的例子。因为待测文本与原文本中的信息直接相矛盾(一者是iPhone,一者是iPad)。外在幻觉是指不能够直接通过原文本得到验证的幻觉现象。同样考虑上述的摘要任务,如果待测文本包含"苹果公司今天发布了新的iPhone,它将在全球范围内同步推出",这就是一个外在幻觉的例子。因为待测文本中存在无法从原文本直接得到验证的内容(iPhone将在全球范围内同步推出)。在原文本中并没有提到产品的发布范围,因此待测文本中的这部分内容既不能由原文本直接支撑,也不能被原文本直接证伪。

为了评估幻觉现象,研究者们提出了多种方法,总体上可以分为非大语言模型的方法[138-139]与基于大语言模型的方法两类。非大语言模型的方法包括基于统计的方法、基于信息抽取的方法、基于生成式问答的方法和基于句子级别分类的方法等。下面重点介绍基于大语言模型的方法。

基于大语言模型方法的核心思想是利用大语言模型的理解和生成能力来评估待测文本的幻觉度。其方法可以分为直接评测方法和间接评测方法。直接评测方法通常将大语言模型作为人的代理,通过模板设计,使其完成一般人类评测员需要完成的工作,即直接评价或直接判断。例如,Sun等人[140]采用自验证的策略,将任务描述、原文本与大语言模型生成的待测文本再次输入大语言模型本身,让其自身对生成的待测文本进行幻觉的检测与幻觉的消除;Mündler等人[141]通过设计模板,使大语言模型能够在给定原文本的情况下,像人类一样直接判断两个和原文本有关的陈述是否互相矛盾。HaluEval[41]结合大语言模型生成和人工标注,创建了一个包含大规模幻觉样例的评测基准以衡量大语言模型检测幻觉和归因幻觉类型的能力。这种评测方法的优势在于能够直接利用大语言模型的泛化能力进行幻觉评测,无须其他额外的计算过程。间接评测方法则是借助大语言模型的生成能力,并结合其他现有的评测指标和方法综合得到最后的幻觉评测结果。例如,给定任务描述、原文本和待测文本,SelfCheckGPT[142]首先将相同的任务描述和原文本输入到一个大语言模型中,并多次随机采样这个大语言模型的输出,得到一组生成文本。如果待测文本中不存在幻觉,那么这组生成文本的内容应当相似,并与待测文本的内容较为一致;反之,这组文本的内容则很可能会发散并与待测文本的内容相互矛盾。因此,给定待测文本和一组生成文本时,可以利用现有的相关指标和方法来表征待测文本和这组生成文本之间的一致性,并将这些指标值综合起来以衡量待测文本的幻觉程度。具体而言,SelfCheckGPT使用了BERTScore、生成式问答与n-gram模型的预测概率三种指标或

方法来衡量待测文本和生成文本集合之间的一致性，并通过加和的方式得到最终衡量幻觉度的指标值。这种间接评测方法的主要优势在于其能够结合大语言模型的生成能力与现有的评测指标与评测方法的优点，得到一个较为综合的度量指标。在幻觉评测中充分利用大语言模型的理解和生成能力，能够在一定程度上帮助处理较为复杂的语义关系，从而评测较为复杂的幻觉现象，如逻辑错误、事实错误及多种错误的耦合等。同时，这种方法一般无须大量的人工标注数据，并可以提供有关幻觉现象的更详细的信息（例如程度信息）。然而，这种方法的局限性在于用于评测的大语言模型本身也同样可能产生幻觉现象。如何控制用于评测的大语言模型本身可能产生的幻觉，将是一项新的挑战性问题。

幻觉评测在未来可能的研究方向有：

（1）更有效的幻觉检测方法。当前的幻觉检测方法在处理较为复杂和模糊的语义时可能会遇到困难。未来的研究可以探索更复杂的模型设计和检测算法以提高幻觉检测的准确性和效率，也可以探索如何利用无标签数据或弱标签数据来提高幻觉评测的性能。

（2）幻觉生成机制的研究。幻觉的全面评测能够帮助研究者进行更深入的有关幻觉生成机制的研究，幻觉生成机制的研究反过来也有助于发展更为全面、更具针对性的幻觉评测方法。若要理解模型为何会产生幻觉，需要深入研究模型的内部工作机制。这可能涉及研究模型的语言理解和生成过程。例如，模型是如何理解并处理语义和语法的，以及这个过程中哪些因素可能会导致幻觉的产生。此外，也可能涉及研究模型的训练过程。例如，模型是如何从训练数据中学习的，训练过程中哪些因素可能导致模型学习到错误的或误导性的信息，从而导致幻觉的产生。

（3）通用的幻觉评测方法设计。在自然语言生成中，不同任务的输入输出形式多样，设计一个与任务无关的通用幻觉评测方法非常重要。这需要深入理解幻觉的本质，以及不同任务中幻觉的共性和特性。同时，不同任务对幻觉的容忍度也不同。在数据到文本生成的任务中，忠实于原文本与事实性正确是两个十分重要的评价方面，对幻觉的容忍度非常低；而故事生成任务对幻觉的容忍度就相对较高，因为在故事生成中往往更加关注例如有趣程度等其他方面。如何设计一个能够捕捉不同任务之间的细微差别，并在各个任务下的评测结果都与人类判断相关性较强的幻觉评测指标，也是目前幻觉评测中的一个挑战。

3.3　元评测

在大语言模型的评测中，元评测是一个不可或缺的部分。元评测是一种衡量评测指标本身有效性和可靠性的过程，也就是对评测的再评测。其核心目标是判断评测方法与人类的评测的相关程度，这对于确保评测质量、减少误差以及提升评测结果可信度具有重要意义。随着大型语言模型在各领域的应用日益广泛，评测大语言模型的方法本身的准确性和可信度也逐渐成为关注焦点。通过对比不同的评测方法，研究者能够发现各种方法的优势和局限性，这将有助于研究者选择更适用于特定任务和场景的评测方法，从而更准确地衡量模型的性能。在下面的讨论中，本文将某个评测指标对模型的 n 个生成内容给出的分数，分别记为 $x_1,\cdots,x_n$，并将人类评测对这 n 个生成内容赋予的分数分别记为 $y_1,\cdots,y_n$。下面将介绍几种元评测中常见的相关性计算方法。

3.3.1　皮尔逊相关系数

皮尔逊相关系数（Pearson Correlation Coefficient）是衡量两个变量之间线性关系强度的指标。给定模型 n 个生成内容上的评测指标分数与人类评测分数的数据点对 $(x_1,y_1),\cdots,(x_n,y_n)$，皮尔逊相关系数的计算，如式（12）所示。

$$\rho=\frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i-\bar{y})^2}} \tag{12}$$

值得指出的是，皮尔逊相关系数衡量的是两个变量之间的线性关系的强弱。其在两个变量之间存在比较强的线性相关时能够表现出较好的性能。同时，它对非线性关系的敏感度较低，并且受异常值的影响较大，数据分布的偏态可能导致相关系数的失真。因此，皮尔逊相关系数不适用于变量之间存在复杂的非线性关系或数据中存在严重异常值或偏态的情况。

3.3.2　斯皮尔曼相关系数

斯皮尔曼相关系数（Spearman's Correlation Coefficient）用于衡量两个变量之间的单调关系，它是基于变量的秩次（相对大小关系）计算得出的。给定模型 n 个生成内容上的评测指标分数与人类评测分数的数据点对 $(x_1,y_1),\cdots,(x_n,y_n)$ 以及它们

对应的秩次$(r_{x_1},r_{y_1}),\cdots,(r_{x_n},r_{y_n})$，斯皮尔曼相关系数的计算如式(13)所示。

$$r=\frac{\sum_{i=1}^{n}(r_{x_i}-\bar{r}_x)(r_{y_i}-\bar{r}_y)}{\sqrt{\sum_{i=1}^{n}(r_{x_i}-\bar{r}_x)^2}\sqrt{\sum_{i=1}^{n}(r_{y_i}-\bar{r}_y)^2}} \tag{13}$$

斯皮尔曼相关系数基于数据的秩次计算，从而对异常值和偏态数据较为鲁棒，并且可以在一定程度上捕捉非线性的关系。但是其只能反映两个变量间的单调关系，当变量之间存在多种依赖关系时，只靠斯皮尔曼相关系数可能难以区分。

3.3.3 肯德尔 τ 系数

肯德尔 τ 系数(Kendall's τ Coefficient)是另一种基于数据秩次的系数，用于衡量两个变量之间的共同趋势。给定模型 n 个生成内容上的评测指标分数与人类评测分数的数据点对$(x_1,y_1),\cdots,(x_n,y_n)$，肯德尔 τ 系数的计算方法如下：①计算配对。对于每一对分数对(x_i,y_i)和(x_j,y_j)，计算它们的差值 x_i-x_j 和 y_i-y_j。②计算一致对(concordant pair)的数目和不一致对(discordant pair)的数目，分别记为 C 和 D。具体来说，若$(x_i-x_j)(y_i-y_j)>0$，则记为一个一致对，若$(x_i-x_j)(y_i-y_j)<0$，则记为一个不一致对。③计算相关系数。肯德尔 τ 系数的计算如式(14)所示。

$$\tau=\frac{C-D}{\binom{n}{2}} \tag{14}$$

与斯皮尔曼相关系数类似，肯德尔 τ 系数是基于数据的秩次，因此对异常值和偏态数据较为鲁棒。但是肯德尔 τ 系数的计算需要枚举每一对数据点对，因此在小样本数据中表现较好，面对大样本数据时计算效率较低。

元评测的实例众多，例如，Sai 等人[143]在摘要、对话、问题生成等多个任务上对包括正确性、流利度、相关性、有趣程度在内的多个评估维度彼此之间的相关程度进行了评估，结果表明即使在同一个任务上，人类在不同评估维度上的评分的相关性往往也并不显著。因此，在这种情况下，仅由自动化评估指标对生成内容赋予一个单一的总分很难全面地评估生成内容在各个细粒度评估维度上的质量。同时，他们还基于扰动方法评估了包括 BLEU、METEOR、BERTScore、BLEURT、MoverScore 在内的多个评测指标的鲁棒性。具体而言，他们通过计算扰动前后评测指标给出的分数差异与人类判断给出的分数差异是否一致来衡量评测指标的鲁棒性。结果显示，相比早期的自动化评测指标，虽然基于模型的评测指标(例如，BERTScore，BLEURT 和 MoverScore 等)在与人类判断的相关程度上表现较好，但是它们面对非常简单的扰动时也无法保持较强的鲁棒性。此外，结果还显示，现有的评测指标往往难以捕捉特定任务上的特殊评测需求。例如，在对话任务中，许多任务模型倾向于生成通用且缺乏针对性的回复，导致与用户的互动效果不佳。然而，在实验中没有一个评测指标对产生诸如“好的”或“你能再重复一遍吗?”等通用回复的扰动具有敏感性。

未来，元评测的研究方向可能包括：

(1) 更细粒度的元评测。不同的自然语言生成任务通常有各自特定的评测需求，即使在同一任务下，也存在多种不同的评估维度，例如连贯性、正确性和相关度等。因此，未来的元评测需要在更细粒度上进行，以评估各评测指标在这些细粒度评估维度上的评测结果与人类判断的相关性，揭示评测指标在捕捉不同生成任务上的微妙差异的能力，为评估方法本身的改进提供指导。

(2) 针对评测指标公平性评估的元评测。现有的评测指标和评测方法通常涉及人类评测与基于模型的评测。其中，人类评测可能受到评测员的专业背景、文化差异等因素的影响；而由于数据的稀缺性，基于模型的评测方法可能面临着在低资源语言上表现更差的问题。元评测需要探究这些因素对评测指标性能的影响，研究评测指标捕捉模型对不同群体或语言的偏见和歧视的能力。这将有助于提高评测方法的公平性，推动更公平、包容的自然语言处理技术的发展。

(3) 针对评测指标鲁棒性评估的元评测。通过基于扰动的方法研究评测指标的鲁棒性，可以揭示其在面对数据噪声、变化或对抗性样本时的稳定性。这种鲁棒性元评测有助于提高评测方法的可靠性，为自然语言处理研究和实践提供更稳健的评估手段。

4 结论

大语言模型评测对大语言模型的应用以及后续发展有非常重要的作用。大语言模型的评测范式分为经典评测范式和新型评测范式。经典评测范式中的传统自然语言处理任务按照内含任务的特点划分

为自然语言理解任务和自然语言生成任务，本文分别介绍了这些任务当前所流行的经典评测基准以及一些新型评测范式下代表性的评测基准和大语言模型评测方面的实例；总结了现有评测中的一些不足之处；然后介绍了全面的大语言模型评测思想以及相关的评测指标和评测方法；最后总结了大语言模型评测的一些新的研究问题、挑战以及未来的研究方向。

参考文献

[1] 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.

[2] RADFORD A, NARASIMHAN K, SALIMANS T, et al. Improving language understanding by generative pre-training [EB/OL]. https://cdn. openai. com/research-covers/language-unsupervised/language_understanding_paper.pdf[2023-05-24].

[3] 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, 2020: 7871-7880.

[4] RAFFEL C, SHAZEER N, ROBERTS A, et al. Exploring the limits of transfer learning with a unified text-to-text transformer[J]. The Journal of Machine Learning Research, 2020, 21(1): 5485-5551.

[5] MCCANN B, KESKAR N S, XIONG C, et al. The natural language decathlon: Multitask learning as question answering[J]. arXiv preprint arXiv: 1806.08730, 2018.

[6] WANG A, SINGH A, MICHAEL J, et al. GLUE: A multi-task benchmark and analysis platform for natural language understanding[C]//Proceedings of the International Conference on Learning Representations, 2018.

[7] WANG A, PRUKSACHATKUN Y, NANGIA N, et al. SuperGLUE: A stickier benchmark for general-purpose language understanding systems [C]//Proceedings of the Advances in Neural Information Processing Systems, 2019.

[8] HU J, RUDER S, SIDDHANT A, et al. Xtreme: A massively multilingual multi-task benchmark for evaluating cross-lingual generalisation[C]//Proceedings of the International Conference on Machine Learning. PMLR, 2020: 4411-4421.

[9] LIANG Y, DUAN N, GONG Y, et al. XGLUE: A new benchmark dataset for cross-lingual pre-training, understanding and generation[C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2020: 6008-6018.

[10] XU L, HU H, ZHANG X, et al. CLUE: A Chinese language understanding evaluation benchmark[C]//Proceedings of the 28th International Conference on Computational Linguistics, 2020: 4762-4772.

[11] XU L, DONG Q, LIAO Y, et al. CLUENER: Fine-grained named entity recognition dataset and benchmark for Chinese[J]. arXiv preprint arXiv: 2001, 04351, 2020.

[12] HU H, RICHARDSON K, XU L, et al. OCNLI: Original chinese natural language inference[G]//Proceedings of the Association for Computational Linguistics: EMNLP,2020: 3512-3526.

[13] XU L, LU X, YUAN C, et al. Fewclue: A Chinese few-shot learning evaluation benchmark[J]. arXiv preprint arXiv: 2107.07498, 2021.

[14] PAPINENI K, ROUKOS S, WARD T, et al. BLEU: A method for automatic evaluation of machine translation[C]//Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, 2002: 311-318.

[15] BANERJEE S, LAVIE A. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments[C]//Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization, 2005: 65-72.

[16] LIN C Y, HOVY E. Automatic evaluation of summaries using n-gram co-occurrence statistics[C]//Proceedings of the Human Language Technology Conference of the North American chapter of the association for Computational Linguistics, 2003: 150-157.

[17] LIU D, YAN Y, GONG Y, et al. GLGE: A new general language generation evaluation benchmark [G]//Proceedings of the Association for Computational Linguistics: ACL-IJCNLP, 2021: 408-420.

[18] TU J, HOLDERNESS E, MARU M, et al. SemEval-2022 task 9: R2VQ-Competence-based multimodal question answering[C]//Proceedings of the 16th International Workshop on Semantic Evaluation, 2022: 1244-1255.

[19] GEHRMANN S, ADEWUMI T, ZHOU J. The GEM benchmark: Natural language generation, its e-

valuation and metrics[C]//Proceedings of the 1st Workshop on Natural Language Generation, Evaluation, and Metrics, Bangkok, Thailand, 2021.

[20] LIN B Y, ZHOU W, SHEN M, et al. CommonGen: A constrained text generation challenge for generative commonsense reasoning[G]//Proceedings of the Association for Computational Linguistics, 2020: 1823-1840.

[21] DUŠEK O, JURCICEK F. Neural generation for czech: Data and baselines[C]//Proceedings of the 12th International Conference on Natural Language Generation, 2019: 563-574.

[22] NAN L, RADEV D, ZHANG R, et al. DART: Open-domain structured data record to text generation[C]//Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, 2021: 432-447.

[23] NOVIKOVA J, DUšEK O, RIESER V. The E2E Dataset: New challenges for end-to-end generation [C]//Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue. Association for Computational Linguistics, 2017.

[24] DUŠEK O, HOWCROFT D M, Rieser V. Semantic noise matters for neural natural language generation [C]//Proceedings of the 12th International Conference on Natural Language Generation, 2019: 421-426.

[25] GARDENT C, SHIMORINA A, NARAYAN S, et al. The WebNLG challenge: Generating text from RDF data[C]//Proceedings of the 10th International Conference on Natural Language Generation, 2017: 124-133.

[26] LADHAK F, DURMUS E, CARDIE C, et al. WikiLingua: A new benchmark dataset for cross-lingual abstractive summarization[G]//Proceedings of the Association for Computational Linguistics, 2020: 4034-4048.

[27] YAO Y, DONG Q, GUAN J, et al. CUGE: A Chinese language understanding and generation evaluation benchmark[J]. arXiv preprint arXiv: 2112.13610, 2021.

[28] HONGYING Z, WENXIN L, KUNLI Z, et al. Building a pediatric medical corpus: Word segmentation and named entity annotation[C]//Proceedings of the Chinese Lexical Semantics: 21st Workshop, Hong Kong, China, 2020.

[29] WANG Y, KONG C, YANG L, et al. YACLC: A Chinese learner corpus with multidimensional annotation[J]. arXiv preprint arXiv: 2112.15043, 2021.

[30] GAO L, TOW J, BIDERMAN S, et al. A framework for few-shot language model evaluation[CP/OL]. https://zenodo.org/records/5371629[2023-05-24].

[31] BROWN T, MANN B, RYDER N, et al. Language models are few-shot learners[C]//Proceedings of the 33rd International Conference on Neural Information Processing Systems, 2020, 33: 1877-1901.

[32] BLACK S, BIDERMAN S, HALLAHAN E, et al. GPT-NeoX-20B: An open-source autoregressive language model[C]//Proceedings of Big Science Episode #5--Workshop on Challenges & Perspectives in Creating Large Language Models, 2022: 95-136.

[33] WANG B, KOMATSUZAKI A. GPT-J-6B: A 6 billion parameter autoregressive language model[CP/OL]. https://docs.adapterhub.ml/classes/models/gptj.html[2023-05-24].

[34] SRIVASTAVA A, RASTOGI A, RAO A, et al. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models[J]. arXiv preprint arXiv: 2206.0461503, 2023.

[35] ZHENG L, CHIANG W L, SHENG Y, et al. Judging LLM-as-a-judge with MT-bench and chatbot arena[J]. arXiv preprint arXiv: 2306.05685, 2023.

[36] HUANG Y, BAI Y, ZHU Z, et al. C-eval: A multi-level multi-discipline Chinese evaluation suite for foundation models[J]. arXiv preprint arXiv: 2305.08322, 2023.

[37] SUN H, ZHANG Z, DENG J, et al. Safety assessment of Chinese large language models[J]. arXiv preprint arXiv: 2304.10436, 2023.

[38] 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, 2017: 1601-1611.

[39] MIHAYLOV T, CLARK P, KHOT T, et al. Can a suit of armor conduct electricity?: A new dataset for open book question answering[C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2018: 2381-2391.

[40] COBBE K, KOSARAJU V, BAVARIAN M, et al. Training verifiers to solve math word problems[J]. arXiv preprint arXiv: 2110.14168, 2021.

[41] LI J, CHENG X, ZHAO W X, et al. HaluEval: A large-scale hallucination evaluation benchmark for large language models[J]. arXiv prints: arXiv: 2305.

11747，2023.
[42] OpenAI. GPT-4 Technical Report[J]. arXiv preprint arXiv：2303.08774，2023.
[43] HENDRYCKS D，BURNS C，BASART S，et al. Measuring massive multitask language understanding [C]//Proceedings of the International Conference on Learning Representations，2020.
[44] 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，2019：4791-4800.
[45] CHEN M，TWOREK J，JUN H，et al. Evaluating large language models trained on code[J]. arXiv preprint arXiv：2107.03374，2021.
[46] DUA D，WANG Y，DASIGI P，et al. DROP：A reading comprehension benchmark requiring discrete reasoning over paragraphs [C]//Proceedings of NAACL-HLT，2019：2368-2378.
[47] ZHONG W，CUI R，GUO Y，et al. A gieval：A human-centric benchmark for evaluating foundation models[J]. arXiv preprint arXiv：2304.06364，2023.
[48] 董青秀，穗志方，詹卫东，等. 自然语言处理评测中的问题与对策[J]. 中文信息学报，2021，35(6)：1-15.
[49] LIANG P，BOMMASANI R，LEE T，et al. Holistic evaluation of language models[J]. arXiv preprint arXiv：2211.09110，2022.
[50] SCAO T L，FAN A，AKIKI C，et al. Bloom：A 176b-parameter open-access multilingual language model[J]. arXiv preprint arXiv：2211.05100，2022.
[51] ZENG A，LIU X，DU Z，et al. GLM-130B：An open bilingual pre-trained model[C]//Proceedings of the 11th International Conference on Learning Representations，2022.
[52] RADEV D，QI H，WU H，et al. Evaluating web-based question answering systems[C]//Proceedings of the 3rd International Conference on Language Resources and Evaluation，2002.
[53] JÄRVELIN K，KEKÄLÄINEN J. Cumulated gain-based evaluation of IR techniques[J]. ACM Transactions on Information Systems，2002，20(4)：422-446.
[54] MURPHY A H. A new vector partition of the probability score[J]. Journal of Applied Meteorology and Climatology，1973，12(4)：595-600.
[55] MURPHY A H，WINKLER R L. Reliability of subjective probability forecasts of precipitation and temperature[J]. Journal of the Royal Statistical Society Series C：Applied Statistics，1977，26(1)：41-47.
[56] DEGROOT M H，FIENBERG S E. The comparison and evaluation of forecasters[J]. Journal of the Royal Statistical Society：Series D，1983，32(1-2)：12-22.
[57] NAEINI M P，COOPER G，HAUSKRECHT M. Obtaining well calibrated probabilities using bayesian binning[C]//Proceedings of the AAAI Conference on Artificial Intelligence，2015.
[58] GUO C，PLEISS G，SUN Y，et al. On calibration of modern neural networks[C]//Proceedings of the International Conference on Machine Learning. PMLR，2017：1321-1330.
[59] LIU P，YUAN W，FU J，et al. Pretrain，prompt，and predict：A systematic survey of prompting methods in natural language processing[J]. ACM Computing Surveys，2023，55(9)：1-35.
[60] HE J，ZHOU C，MA X，et al. Towards a unified view of parameter-efficient transfer learning[C]//Proceedings of the International Conference on Learning Representations，2021.
[61] HOULSBY N，GIURGIU A，JASTRZEBSKI S，et al. Parameter-efficient transfer learning for NLP [C]//Proceedings of the International Conference on Machine Learning. PMLR，2019：2790-2799.
[62] RIBEIRO M T，WU T，GUESTRIN C，et al. Beyond accuracy：Behavioral testing of NLP models with checklist[C]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics，2020：4902-4912.
[63] SAP M，CARD D，GABRIEL S，et al. The risk of racial bias in hate speech detection[C]//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics，2019：1668-1678.
[64] TSIPRAS D. Learning through the lens of robustness [D]. Massachusetts Institute of Technology，2021.
[65] YANG C，BROWER-SINNING R，LEWIS G A，et al. Capabilities for better ml engineering[J]. arXiv preprint arXiv：2211.06409，2022.
[66] ZHANG W E，SHENG Q Z，ALHAZMI A，et al. Adversarial attacks on deep-learning models in natural language processing：A survey[J]. ACM Transactions on Intelligent Systems and Technology，2020，11(3)：1-41.
[67] REN S，DENG Y，HE K，et al. Generating natural language adversarial examples through probability weighted word saliency[C]//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics，2019：1085-1097.
[68] ALSHEMALI B，KALITA J. Improving the reliabil-

ity of deep neural networks in NLP: A review[J]. Knowledge-based Systems, 2020, 191: 105210.

[69] WANG B, XU C, WANG S, et al. Adversarial GLUE: A multi-task benchmark for robustness evaluation of language models[C]//Proceedings of the 35th Conference on Neural Information Processing Systems Datasets and Benchmarks Track, 2021.

[70] MORADI M, SAMWALD M. Evaluating the robustness of neural language models to input perturbations [C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2021: 1558-1570.

[71] BENDER E M, GEBRU T, MCMILLAN-MAJOR A, et al. On the dangers of stochastic parrots: Can language models be too big? [C]//Proceedings of the ACM Conference on Fairness, Accountability, and Transparency, 2021: 610-623.

[72] BOMMASANI R, HUDSON D A, ADELI E, et al. On the opportunities and risks of foundation models [J]. arXiv preprint arXiv: 2108.07258, 2021.

[73] LAN Z, CHEN M, GOODMAN S, et al. ALBERT: A lite BERT for self-supervised learning of language representations[C]//Proceedings of the International Conference on Learning Representations, 2019.

[74] JIAO X, YIN Y, SHANG L, et al. TinyBERT: Distilling BERT for natural language understanding [G]//Proceedings of the Association for Computational Linguistics, 2020: 4163-4174.

[75] LIU W, ZHOU P, WANG Z, et al. FastBERT: A self-distilling BERT with adaptive inference time [C]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020: 6035-6044.

[76] LI X, SHAO Y, SUN T, et al. Accelerating BERT inference for sequence labeling via early-exit[C]// Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, 2021: 189-199.

[77] LIU X, SUN T, HE J, et al. Towards efficient NLP: A standard evaluation and a strong baseline [C]//Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2022: 3288-3303.

[78] SANH V, DEBUT L, CHAUMOND J, et al. DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter[J]. arXiv preprint arXiv: 1910.01108, 2019.

[79] 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,2020: 6640-6651.

[80] ZHOU W, XU C, GE T, et al. Bert loses patience: Fast and robust inference with early exit[C]//Proceedings of the 33rd International Conference on Neural Information Processing Systems, 2020, 33: 18330-18341.

[81] SUN T, ZHOU Y, LIU X, et al. Early exiting with ensemble internal classifiers[J]. arXiv preprint arXiv: 2105.13792, 2021.

[82] RAUH M, MELLOR J, UESATO J, et al. Characteristics of harmful text: Towards rigorous benchmarking of language models[C]//Proceedings of the 35rd International Conference on Neural Information Processing Systems, 2022, 35: 24720-24739.

[83] CALISKAN A, BRYSON J J, NARAYANAN A. Semantics derived automatically from language corpora contain human-like biases[J]. Science, 2017, 356 (6334): 183-186.

[84] 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, 2021: 5356-5371.

[85] MANZINI T, LIM Y C, TSVETKOV Y, et al. Black is to criminal as caucasian is to police: detecting and removing multiclass bias in word embeddings [C]//Proceedings of NAACL-HLT, 2019: 615-621.

[86] LAUSCHER A, GLAVAŠ G. Are we consistently biased? Multidimensional analysis of Biases in distributional word vectors[C]//Proceedings of the 8th Joint Conference on Lexical and Computational Semantics, 2019: 85-91.

[87] MAY C, WANG A, BORDIA S, et al. On measuring social biases in sentence encoders[C]//Proceedings of NAACL-HLT, 2019: 622-628.

[88] KURITA K, VYAS N, PAREEK A, et al. Quantifying social biases in contextual word representations [C]//Proceedings of the 1st ACL Workshop on Gender Bias for Natural Language Processing, 2019.

[89] GUO W, CALISKAN A. Detecting emergent intersectional biases: Contextualized word embeddings contain a distribution of human-like biases[C]//Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society, 2021: 122-133.

[90] 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, 2023.

[91] SCHICK T, UDUPA S, SCHÜTZE H. Self-diagnosis and self-debiasing: A proposal for reducing corpus-based bias in nlp[J]. Transactions of the Association for Computational Linguistics, 2021, 9: 1408-1424.

[92] WEBSTER K, WANG X, TENNEY I, et al. Measuring and reducing gendered correlations in pretrained models[J]. arXiv preprint arXiv: 2010.06032, 2020.

[93] BARTL M, NISSIM M, GATT A. Unmasking contextual stereotypes: Measuring and mitigating BERT's gender bias[C]//Proceedings of the 2nd Workshop on Gender Bias in Natural Language Processing. Association for Computational Linguistics, 2020: 1-16.

[94] 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 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2020: 1953-1967.

[95] KANEKO M, BOLLEGALA D. Unmasking the mask-evaluating social biases in masked language models[C]//Proceedings of the AAAI Conference on Artificial Intelligence, 2022, 36(11): 11954-11962.

[96] BORDIA S, BOWMAN S. Identifying and reducing gender bias in word-level language models[C]//Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop, 2019: 7-15.

[97] KURITA K, VYAS N, PAREEK A, et al. Measuring bias in contextualized word representations[C]//Proceedings of the 1st Workshop on Gender Bias in Natural Language Processing, 2019: 166-172.

[98] ANTONIAK M, MIMNO D. Bad seeds: Evaluating lexical methods for bias measurement[C]//Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, 2021: 1889-1904.

[99] ETHAYARAJH K, DUVENAUD D, HIRST G. Understanding undesirable word embedding associations[C]//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 2019: 1696-1705.

[100] BLODGETT S L, BAROCAS S, DAUMÉ III H, et al. Language (technology) is power: A critical survey of "bias" in NLP[C]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020: 5454-5476.

[101] BLODGETT S L, LOPEZ G, OLTEANU A, et al. Stereotyping norwegian salmon: An inventory of pitfalls in fairness benchmark datasets[C]//Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, 2021: 1004-1015.

[102] GOLDFARB-TARRANT S, MARCHANT R, SÁNCHEZ R M, et al. Intrinsic bias metrics do not correlate with application bias[C]//Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, 2021: 1926-1940.

[103] STEED R, PANDA S, KOBREN A, et al. Upstream mitigation is not all you need: Testing the bias transfer hypothesis in pre-trained language models[C]//Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics, 2022: 3524-3542.

[104] DELOBELLE P, TOKPO E K, CALDERS T, et al. Measuring fairness with biased rulers: A survey on quantifying biases in pretrained language models[J]. arXiv preprint arXiv: 2112.07447, 2021.

[105] MITCHELL M, WU S, ZALDIVAR A, et al. Model cards for model reporting[C]//Proceedings of the Conference on Fairness, Accountability, and Transparency, 2019: 220-229.

[106] HAN X, SHEN A, LI Y, et al. Fairlib: A unified framework for assessing and improving fairness[C]//Proceedings of the the Conference on Empirical Methods in Natural Language Processing: System Demonstrations, 2022: 60-71.

[107] RAMESH K, SITARAM S, CHOUDHURY M. Fairness in language models beyond English: Gaps and challenges[G]//Proceedings of the Association for Computational Linguistics, 2023: 2061-2074.

[108] TATMAN R. Gender and dialect bias in YouTube's automatic captions[C]//Proceedings of the 1st ACL Workshop on Ethics in Natural Language Processing, 2017: 53-59.

[109] HUTCHINSON B, MITCHELL M. 50 years of test (un) fairness: Lessons for machine learning[C]//Proceedings of the Conference on Fairness,

Accountability, and Transparency, 2019: 49-58.
[110] HARDT M, PRICE E, SREBRO N. Equality of opportunity in supervised learning[C]//Proceedings of the 30th International Conference on Neural Information Processing Systems, 2016, 29.
[111] DWORK C, HARDT M, PITASSI T, et al. Fairness through awareness[C]//Proceedings of the 3rd Innovations in Theoretical Computer Science Conference, 2012: 214-226.
[112] SHEN A, HAN X, COHN T, et al. Optimising equal opportunity fairness in model training[C]//Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2022: 4073-4084.
[113] LUM K, ZHANG Y, BOWER A. De-biasing "bias" measurement[C]//Proceedings of the ACM Conference on Fairness, Accountability, and Transparency, 2022: 379-389.
[114] MEADE N, POOLE DAYAN E, REDDY S. An empirical survey of the effectiveness of debiasing techniques for pre-trained language models[C]//Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics, 2022: 1878-1898.
[115] MA Z, ETHAYARAJH K, THRUSH T, et al. Dynaboard: An evaluation-as-a-service platform for holistic next-generation benchmarking[C]//Proceedings of the 34th International Conference on Neural Information Processing Systems, 2021, 34: 10351-10367.
[116] QIAN R, ROSS C, FERNANDES J, et al. Perturbation augmentation for fairer NLP[C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2022: 9496-9521.
[117] DHOLE K, GANGAL V, GEHRMANN S, et al. NL-augmenter: A framework for task-sensitive natural language augmentation[J]. Northern European Journal of Language Technology, 2023, 9(1): 1-41.
[118] ZIEMS C, CHEN J, HARRIS C, et al. VALUE: Understanding dialect disparity in NLU[C]//Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics, 2022: 3701-3720.
[119] CASELLI T, BASILE V, MITROVIĆ J, et al. HateBERT: Retraining BERT for abusive language detection in English[C]//Proceedings of the 5th Workshop on Online Abuse and Harms, 2021: 17-25.
[120] LEES A, TRAN V Q, tay Y, et al. A new generation of perspective api: Efficient multilingual character-level transformers[C]//Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, 2022: 3197-3207.
[121] ZHAO W X, ZHOU K, LI J, et al. A survey of large language models[J]. arXiv preprint arXiv: 2303.18223, 2023.
[122] MATHUR N, BALDWIN T, COHN T. Putting evaluation in context: Contextual embeddings improve machine translation evaluation[C]//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 2019: 2799-2808.
[123] ZHANG T, KISHORE V, WU F, et al. BERTScore: Evaluating text generation with BERT[C]//Proceedings of the International Conference on Learning Representations, 2019.
[124] ZHAO W, PEYRARD M, LIU F, et al. MoverScore: Text generation evaluating with contextualized embeddings and earth mover distance[C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2019.
[125] SHIMANAKA H, KAJIWARA T, KOMACHI M. Machine translation evaluation with bert regressor[J]. arXiv preprint arXiv: 1907.12679, 2019.
[126] REI R, STEWART C, FARINHA A C, et al. COMET: A neural framework for MT evaluation[C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2020: 2685-2702.
[127] SELLAM T, DAS D, PARIKH A. BLEURT: Learning robust metrics for text generation[C]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020: 7881-7892.
[128] ZHAO T, LALA D, KAWAHARA T. Designing precise and robust dialogue response evaluators[C]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020: 26-33.
[129] YUAN W, NEUBIG G, LIU P. Bartscore: Evaluating generated text as text generation[C]//Proceedings of the 34th International Conference on Neural Information Processing Systems, 2021, 34: 27263-27277.
[130] PILLUTLA K, LIU L, THICKSTUN J, et al. MAUVE scores for generative models: theory and practice[J]. arXiv preprint arXiv: 2212.14578, 2022.
[131] ZHAO W, STRUBE M, EGER S. DiscoScore: E-

valuating text generation with BERT and discourse coherence[C]//Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics, 2023: 3847-3865.

[132] FU J, NG S K, JIANG Z, et al. Gptscore: Evaluate as you desire[J]. arXiv preprint arXiv: 2302.04166, 2023.

[133] KOCMI T, FEDERMANN C. Large language models are state-of-the-art evaluators of translation quality[J]. arXiv preprint arXiv: 2302.14520, 2023.

[134] WANG J, LIANG Y, MENG F, et al. Is chatgpt a good NLG evaluator? A preliminary study[J]. arXiv preprint arXiv: 2303.04048, 2023.

[135] WANG Y, YU Z, ZENG Z, et al. PandaLM: An automatic evaluation benchmark for LLM instruction tuning optimization[J]. arXiv preprint arXiv: 2306.05087, 2023.

[136] TOUVRON H, LAVRIL T, IZACARD G, et al. Llama: Open and efficient foundation language models[J]. arXiv preprint arXiv: 2302.13971, 2023.

[137] MEHRI S, ESKENAZI M. USR: An unsupervised and reference free evaluation metric for dialog generation[C]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020: 681-707.

[138] JI Z, LEE N, FRIESKE R, et al. Survey of hallucination in natural language generation[J]. ACM Computing Surveys, 2023, 55(12): 1-38.

[139] HUANG Y, FENG X, FENG X, et al. The factual inconsistency problem in abstractive text summarization: A survey[J]. arXiv preprint arXiv: 2104.14839, 2021.

[140] SUN X, DONG L, LI X, et al. Pushing the limits of ChatGPT on NLP tasks[J]. arXiv preprint arXiv: 2306.09719, 2023.

[141] MÜNDLER N, HE J, JENKO S, et al. Self-contradictory hallucinations of large language models: Evaluation, detection and mitigation[J]. arXiv preprint arXiv: 2305.15852, 2023.

[142] MANAKUL P, LIUSIE A, GALES M J F. Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models[J]. arXiv preprint arXiv: 2303.08896, 2023.

[143] SAI A B, DIXIT T, Sheth D Y, et al. Perturbation checklists for evaluating nlg evaluation metrics[C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2021: 7219-7234.

罗文(2002—),本科生,主要研究领域为自然语言处理与深度学习。
E-mail: llvvvv@stu.pku.edu.cn

王厚峰(1965—),博士,教授,主要研究领域为自然语言处理。
E-mail: wanghf@pku.edu.cn