



计算机科学与探索

Journal of Frontiers of Computer Science and Technology

ISSN 1673-9418, CN 11-5602/TP

## 《计算机科学与探索》网络首发论文

题目: AIGC 大模型测评综述: 使能技术, 安全隐患和应对  
作者: 许志伟, 李海龙, 李博, 李涛, 王嘉泰, 谢学说, 董泽辉  
网络首发日期: 2024-05-27  
引用格式: 许志伟, 李海龙, 李博, 李涛, 王嘉泰, 谢学说, 董泽辉. AIGC 大模型测评综述: 使能技术, 安全隐患和应对[J/OL]. 计算机科学与探索.  
<https://link.cnki.net/urlid/11.5602.tp.20240523.1947.002>



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

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

# AIGC 大模型测评综述：使能技术，安全隐患和应对

许志伟<sup>1,2</sup>, 李海龙<sup>2,3</sup>, 李博<sup>2,3</sup>, 李涛<sup>1,4+</sup>, 王嘉泰<sup>5</sup>, 谢学说<sup>1</sup>, 董泽辉<sup>1</sup>

1. 信创海河实验室, 天津 300450

2. 中国科学院 计算技术研究所, 北京 100080

3. 内蒙古工业大学 数据科学与应用学院, 呼和浩特 010080

4. 南开大学 计算机学院, 天津 300350

5. OPPO 研究院, 北京 100026

+ 通信作者 E-mail: litao@nankai.edu.cn

**摘要：**人工智能生成内容（AIGC）模型因其出色的内容生成能力在全球范围内引起了学术和工业界的广泛关注与应用。特别是以 ChatGPT 为代表的一批多模态 AIGC 大模型的出现，为人类社会生产生活带来了前所未有的智能体验和高阶应用。然而，AIGC 大模型的快速发展也带来了一系列隐患，例如模型生成结果的可解释性、公平性和安全隐私等问题。在 AIGC 大模型融入各行各业的过程中，为了降低不可知风险及其危害，需要事先对 AIGC 大模型进行全面的测评。具体来说，包括以下三个方面：如何根据不同类任务准备数据，选择合适的测评基准/指标，怎样设计测评过程。学术界已经开启了 AIGC 大模型测评相关的研究，旨在有效应对相关挑战，避免潜在的风险。通过对这些模型测评研究的回顾，按照上述思路进行了综述和分析，涵盖了 AIGC 大模型测评过程，当前 AIGC 大模型在不同领域应用过程中面临的新挑战，以及这些挑战对应的应对措施。最后，探讨了 AIGC 大模型测评未来面临的挑战，并展望了其未来发展方向。

**关键词：**AIGC 大模型；大模型测评；可解释性；公平性；鲁棒性；安全与隐私保护

**文献标志码：**A      **中图分类号：**TP18;TP311.31

## A Survey of AIGC Model Evaluation: Enabling Technologies, Vulnerabilities and Mitigation

XU Zhiwei<sup>1,2</sup>, LI Hailong<sup>2,3</sup>, LI Bo<sup>2,3</sup>, LI Tao<sup>1,4+</sup>, WANG Jiatai<sup>5</sup>, XIE Xueshuo<sup>1</sup>, DONG Zehui<sup>1</sup>

1. Haihe Laboratory of Information Technology Application Innovation, Tianjin 300450, China

2. Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100080, China

3. College of Data Science and Application, Inner Mongolia University of Technology, Hohhot 010080, China

4. College of Computer Science, Nankai University, Tianjin 300350, China

5. OPPO Research, Beijing 100026, China

**Abstract:** Artificial Intelligence-Generated Content (AIGC) models have attracted a wide range of attention and applied in academia and industry extensively with their outstanding content generation capabilities. In particular, the emergence of ChatGPT and other multi-modal AIGC models introduced unprecedented intelligent experiences and innovative applications in our daily work and life. Meanwhile, the rapid development of AIGC large models also raised a series of new vulnerabilities, such as concerns about interpretability, fairness, security, and privacy preservation of model-generated content. To address the potential risks and their damages during integrating AIGC large models into practical applications, a comprehensive evaluation method for AIGC large models is highly desired. AIGC large model evaluation should encompass the following three aspects: how to collect data for evaluation; how to select appropriate evaluation benchmarks or metrics; and how to design the evaluation process in terms

**基金项目：**国家自然科学基金地区项目（61962045，62272248）；内蒙古青年科技英才支持项目（NJYT23104）。

This work was supported by the National Science Foundation of China (61962045, 62272248) and Inner Mongolia Support Plan for Young Science and Technology Talents(NJYT23104).

of different types of tasks. Advanced approaches to AIGC large model evaluation aim to effectively conquer the aforementioned challenges and avoid potential problems. By reviewing this research, we analyze the state of the art, including evaluation methods for AIGC large models, the application challenges of the AIGC large models in various fields, and the corresponding evaluation methods. Finally, we discuss the future of AIGC large model evaluation.

**Key words:** AIGC large model ; Large model evaluation; Interpretability; Fairness; Robustness; Security and privacy protection

人工智能生成内容 (Artificial Intelligence Generated Content, AIGC) 是人工智能领域发展的最新成果之一, 其主要是通过学习数据模式和规律, 生成原创的内容, 包括文本、图像、音频等多媒体内容<sup>[1-2]</sup>。2017 年 OpenAI 开发了 GPT 系列模型<sup>[3]</sup>, 经过多年发展, 培育出了全新的人工智能现象级工具 ChatGPT<sup>[4]</sup>, 已经成为人工智能领域全新的发展标杆, 推动了人工智能相关研究向着通用智能方向走出了关键一步。ChatGPT 一经推出就备受瞩目, 其在水机交互、对话、文档生成等任务中展现出了接近人类的能力。衍生出的其他 AIGC 大模型应用已经融入各行各业, 吸引了数以亿万的用户, 相关研究在学术界和工业界得到了广泛的关注<sup>[5-12]</sup>。除了国外其它组织提出的 BRET<sup>[13]</sup>、PaLM<sup>[14]</sup>、OPT<sup>[15]</sup>、BLOOM<sup>[16]</sup>、LLaMA<sup>[17]</sup>、LLaMA2<sup>[18]</sup>、Stable Diffusion<sup>[19]</sup>、Midjourney<sup>[20]</sup>和 GPT4<sup>[21]</sup>等大模型, 我国在 AIGC 大模型领域也取得了显著的进展, 提出了包括清华大学的 ChatGLM<sup>[22]</sup>、ChatGLM2<sup>[23]</sup>和 WEBCPM<sup>[24]</sup>、百度的文心一言<sup>[25]</sup>、阿里的通义千问<sup>[26]</sup>、复旦大学的 MOSS<sup>[27]</sup>、哈尔滨工业大学赛尔实验室的“活字 1.0”和“活字 2.0”两个版本的大语言模型、中国科学院的紫东太初<sup>[28]</sup>和华为的盘古<sup>[29]</sup>等大模型。这些模型的不断进步推动了人工智能技术的前进, 为更加智能化的人机交互体验开辟了新的可能性。

同其它人工智能模型类似, AIGC 生成式大模型同样面临着数据和模型的限制。首先需要收集和预处理数据, 然后训练模型, 并根据实际应用场景进行模型微调, 最后完成模型部署和发布。这一过程的局限具体表现在以下三个方面: (1) 在数据收集阶段, 若使用未经授权数据对模型进行训练, 可能导致数据泄露的问题。(2) 在模型训练过程中, 由于标注策略的局限, 可能导致数据特征分布不均匀。(3) 在内容生成阶段, 不均匀的数据分布将导致生成具有偏见的内容。AIGC 在数据隐私和生成内容公平性方面引发了广泛关注, 另外生成内容导致的虚假信息传播和生成内容版权等大量问题都是急需研究的现实问题。为了防范和杜绝上述

问题及其他潜在隐患, 在模型部署应用之前需要对模型进行测评。AIGC 大模型属于大型深度神经网络模型。长期以来, 由于深度神经网络模型的不可解释性, 我们始终无法完全获知其模型运转机理, 这就造成对深度神经网络模型的测评一直是一个棘手的问题。同时, 考虑到未来 AIGC 大模型的规模和复杂性将快速增长, 如何对其进行安全、隐私性和公平性等方面的测评是一个亟待解决的现实问题。

大模型测评领域的大部分工作都集中在针对新问题创新相应的测评方法方面, 快速地、低成本地测评新模型性能及数据对模型的影响<sup>[10,12,30-36]</sup>。例如, Rao 等人<sup>[10]</sup>设计提示激励客观答案生成, 然后通过替换问题的主语, 最后评估模型生成结果的正确性。Zuccon 等人<sup>[12]</sup>在模型知识和结合提示知识情况下对比模型生成答案的差异。Amos Azaria 等人<sup>[30]</sup>从大模型内部状态的角度来测评生成内容的真实性。Gao 等人<sup>[32]</sup>则采用不同的提示对模型进行测评。Liu 等人<sup>[35]</sup>为更好地测评内容生成质量使用形式填充范式测评模型。Wang 等人<sup>[36]</sup>将 ChatGPT 模型视作人类评价者, 并给出特定的任务和指令, 提示 ChatGPT 对 AIGC 大模型生成结果进行测评。

相应的测评基准也开始涌现。例如, MMLU<sup>[37]</sup>、GAOKAO<sup>[38]</sup>、C-EVAL<sup>[39]</sup>、AGIEval<sup>[40]</sup>、CMMLU<sup>[41]</sup>、M3Exam<sup>[42]</sup>、BIG-bench<sup>[43]</sup>和 HELM<sup>[44]</sup>等基准试图聚合广泛的 NLP 任务, 以进行整体测评。以上测评的工作不仅涉及测评过程的设计, 更是引出了大模型测评的一些新的关注点, 包括生成结果的公平性、Prompt 数据的安全隐私问题等。Chang 等人近期公布了大语言模型测评综述的预印版初稿<sup>[45]</sup>, 从大模型应用着手展开讨论。其当前版本还缺少对大模型测评过程系统介绍, 也没有对大模型应对挑战的解决方案的综述, 只是给出了一些初步的构想。针对这些不足, 本文对最新成果进行更为全面的综述, 同时力争更好地覆盖中文大模型测评问题。

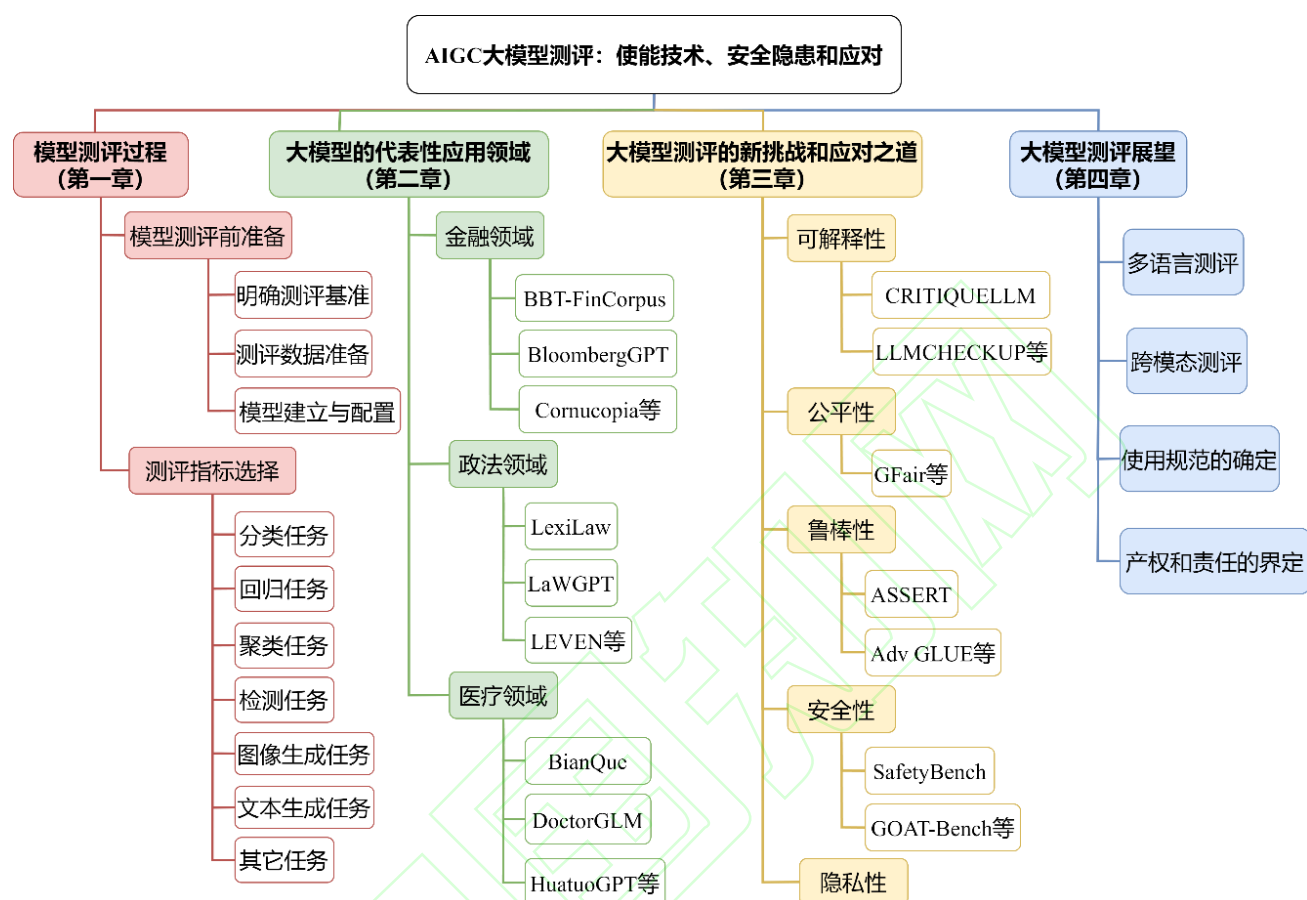


图1 文章结构示意图

Fig.1 Schematic diagram of the article's structure

模型测评发展的方向是获取自动化且可重复的测评结果，以此对不同的模型进行更准确高效的评价<sup>[46-67]</sup>。为了实现对 AIGC 大模型在不同任务和数据上生成的结果的精确量化，需要同时考虑这些任务的特点，有针对性地设计模型测评过程。与其他深度神经网络模型不同，AIGC 大模型大多采用 Prompt 提示方式进行交互，这也为传统模型测评提出了挑战。因此，我们需要重点探索和关注生成结果的公平性、Prompt 数据的安全隐私等新问题，揭示隐藏的风险，形成全面有效的测评结果。主要涉及以下 4 个方面的问题：（1）现有的模型测评方法需要根据 AIGC 大模型的特性进行调整，例如 Prompt 提示设计等，以适应不同领域大模型测评的需求。（2）AIGC 大模型在决策时对不同群体或受数据偏见表现出的差异化行为可能导致对特定群体产生不公平的结果。公平性测评直接影响用户对 AIGC 大模型的信任度问题。因此，如何在确保模型性能的情况下兼顾公平性的约束，是一个需要深

入研究的问题。（3）鲁棒性决定了 AIGC 大模型在面对不同类型攻击和干扰时是否能达到一定的性能水平。提高模型稳定性和优化模型结构能更好应对复杂的实际应用场景。目前的研究采用大量的鲁棒性测评方法对模型进行测试。（4）安全性反映了 AIGC 大模型受到恶意攻击时及时采取措施的能力。而隐私意味着是否有效保护数据，包括 Prompt 数据中的敏感信息。通过对模型进行安全和隐私性测评，提高模型的保护能力和适应能力。由于各个国家地区的法律定义差异，安全和隐私性测评在 AIGC 大模型领域仍处于不断发展和探索阶段。针对上述问题展开系统性的测评，可以在以下几个方面起到关键作用：（1）大模型测评可以帮助评估模型在各种任务和领域中的性能和准确性。这对于了解模型的实际能力、优势和局限性至关重要，以便在实际应用中做出准确的决策。（2）大模型测评可以揭示模型可能存在的问题、偏见、错误或不当行为。通过检测和报告这些问题，可以促使研究人员和开发者



对模型进行改进，提高其鲁棒性和可靠性。（3）大模型测评的结果可以为研究人员提供反馈和启示，指导他们改进模型的训练方法和架构设计。这有助于推动模型研究的进展，提高模型在各种任务和领域中的性能。（4）大模型测评可以确保模型的可靠性和可信度。通过构建透明、可重复的测评过程，可以帮助用户建立对模型性能的信任，并促进大模型的实际部署和应用。

本文主要综述 AIGC 大模型测评的相关内容，包括现有模型测评过程、大模型在特定领域的应用及挑战，以及如何针对这些挑战展开测评以杜绝隐忧等。模型测评过程主要包括确定测评目标、数据准备和预处理、特征工程、测评指标选择和测评方法等方面。然后，对大模型在特定领域的应用及相关的隐忧进行了分析，着重综述了大模型在可解释性、公平性、鲁棒性、安全和隐私性等方面的挑战。最后，讨论了 AIGC 大模型测评目前已取得的进展，并展望了未来大模型测评的研究方向。

文章结构如图 1 所示，其中第 1 节介绍 AIGC 大模型测评的背景知识。第 2 节介绍现有模型测评过程所设计的使能技术。第 3 节讨论了 AIGC 大模型在不同领域应用的过程中暴露出来的问题。第 4 节针对新的问题，从可解释性、公平性、鲁棒性、安全和隐私性等方面详细梳理总结了大模型测评相关研究。第 5 节展望了未来 AIGC 大模型测评的

研究方向，包括多语言测评、跨模态测评、模型测评规范的确定以及测评中产权和责任的界定等内容。第 6 节总结了本文综述内容。

1 模型测评过程

最近，AIGC 大模型已经前所未有的改变了人类生活和社会的各个方面。评估对于了解大模型的真实能力、降低潜在风险并最终进一步造福社会至关重要。相对于传统测评，AIGC 大模型测评更为复杂<sup>[48]</sup>。具体来说，在训练过程中，AIGC 大模型会记忆训练数据的敏感信息，因此存在隐私泄露的风险<sup>[49-54]</sup>。而在部署过程中，特别是在高风险领域，如医疗诊断、司法决策和其他安全关键领域，还需要测试其可信度。

最近的工作从多个方面开展了对大模型的测评。例如，Wang 等人<sup>[55]</sup>发现当前大模型存在对提示信息敏感。Zhu 等人<sup>[56]</sup>发现对抗性即时攻击也是面临的一个挑战。并且，数据污染还会带来严重的安全和隐私问题<sup>[57-59]</sup>。因此，Zhu 等人<sup>[60]</sup>引入了鲁棒性基 PromptBench，旨在衡量大模型对对抗性提示的弹性。我们以测评过程的角度，对测评材料准备和相关任务指标进行回顾。模型测评过程如图 2 所示，大模型测评相关的改进需要在蓝色底纹框标出的部分完成。

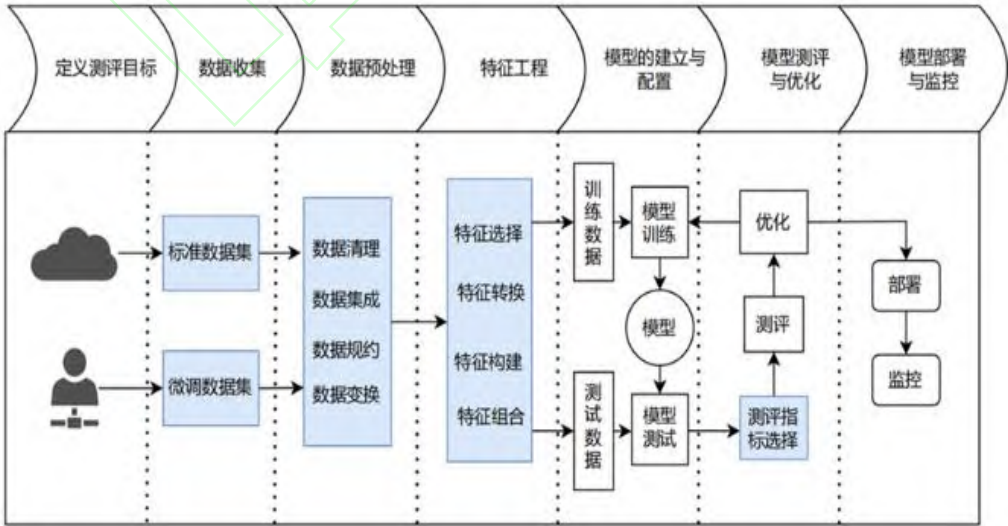


图 2 测评过程  
Fig.2 Evaluation process

## 1.1 模型测评前准备

模型测评之前，需要明确给定数据群体中的随机样本上训练的模型的测评，即数据集的选择和构建。另外，也必须明确定义要测评的任务，例如模型的生成能力、文本理解、机器翻译等。然后确定测评基准收集和优化测评数据，确保测评质量。

### (1) 定义测评目的

让我们考虑一个显而易见的问题：“我们如何确定一个模型的性能？”。通常情况下，我们可能会想到以下步骤：第一步，将训练数据提供给模型进行学习；第二步，预测测试集中样本的标签；第三步，计算样本错误预测的数量。然而，我们更关心的是模型在未见过的数据上的表现如何，即对未知数据进行预测。我们不仅要测试在给定情况下模型的最佳解决方案，还希望比较不同模型在预测中的性能。模型测评目标制定的要点可以总结如下：

1) 我们想要预测模型对未来数据的预测性能。2)

我们希望通过模型的调优，从给定的假设空间中选择最佳的调优参数来提高预测性能。3) 我们想要确定最适合当前任务描述的模型。因此需要比较不同的模型，并从给定的假设空间中选择性能最好的模型。在本文中，我们将讨论处理以上任务的一些方法。在模型在测评前、测评过程中或在处理测评结果时，由于数据问题或模型无意识偏差将可能导致模型测评不准确，这可能是模型测评中最具挑战性的任务之一。

### (2) 明确测评基准

在常规模型测评中，只需将数据集标注信息作为测评基准，然后结合不同指标即可衡量模型的相关性能。然而，对于 AIGC 大模型测评而言，由于使用 Prompt 提示过程来引导模型给出结果，因此需要设计一系列精准问题来构建更为系统的测评基准，以便全面获取模型不同方面的测评结果。我们在表 1 总结了现有的评测基准。

表 1 评测基准

Table 1 Evaluation benchmarks

测评基准	发布机构	简介	链接
MMLU <sup>[37]</sup>	LMSYS Org（大型模型系统组织）	涵盖 STEM、人文科学、社会科学等 57 个科目	<a href="https://github.com/NLP-Core-Team/mmlu_ru">https://github.com/NLP-Core-Team/mmlu_ru</a>
GaoKao <sup>[38]</sup>	复旦大学	1781 道客观题和 1030 道主观题	<a href="https://github.com/OpenMLLab/GAOKAO-Bench">https://github.com/OpenMLLab/GAOKAO-Bench</a>
C-Eval <sup>[39]</sup>	清华大学、上海交通大学和爱丁堡大学合作	52 个不同学科的 13948 个多项选择题	<a href="https://github.com/hkust-nlp/ceval">https://github.com/hkust-nlp/ceval</a>
AGIEval <sup>[40]</sup>	微软亚洲研究院	涵盖全球 20 种面向普通人类考生的官方、公共和高标准录取和资格考试，包含中英文数据。	<a href="https://github.com/ruixiangcui/AGIEval">https://github.com/ruixiangcui/AGIEval</a>
CMMLU <sup>[41]</sup>	MBZUAI、上海交通大学、微软亚洲研究院	67 个学科，11582 道选择题	<a href="https://github.com/haonan-li/CMMLU">https://github.com/haonan-li/CMMLU</a>
M3Exam <sup>[42]</sup>	阿里巴巴达摩院	首个多语言多模态测试基准 M3Exam，涵盖 12317 道题目。	<a href="https://github.com/DAMO-NLP-SG/M3Exam">https://github.com/DAMO-NLP-SG/M3Exam</a>
Big-Bench <sup>[43]</sup>	Google	204 个任务，涉及语言学、儿童发展、数学、常识推理、生物学、物理学、社会偏见、软件开发等领域的问题	<a href="https://github.com/google/BIG-bench">https://github.com/google/BIG-bench</a>
HELM <sup>[44]</sup>	斯坦福大学	在推理场景、含虚假信息场景等多个场景中评估	<a href="https://github.com/stanford-crfm/helm">https://github.com/stanford-crfm/helm</a>
PromptBench <sup>[56]</sup>	微软亚洲研究院	首个大语言模型提示鲁棒性的评测基准	<a href="https://github.com/microsoft/promptbench">https://github.com/microsoft/promptbench</a>
Arena <sup>[61]</sup>	LMSYS Org（大型模型系统组织）	4.2 万匿名用户参与投票	<a href="https://github.com/id-Software/Quake-III-Arena">https://github.com/id-Software/Quake-III-Arena</a>
MT-Bench <sup>[62]</sup>	LMSYS Org（大型模型系统组织）	80 个高质量的多轮对话问题	<a href="https://github.com/SiddhantHegde/Dravidian-MTL-Benchmarking">https://github.com/SiddhantHegde/Dravidian-MTL-Benchmarking</a>
MME <sup>[63]</sup>	腾讯优图实验室	多模态大语言模型的综合评价标准	<a href="https://github.com/open-mmlab/mengine">https://github.com/open-mmlab/mengine</a>
L-Eval <sup>[64]</sup>	复旦大学	涵盖了 17 个大类、453 个问题，包括事实性问答、阅读理解、框架生成、段落重写、摘要、数学解题、推理、诗歌生成、编程等各个领	<a href="https://github.com/OpenMLLab/LEval">https://github.com/OpenMLLab/LEval</a>

		域。	
KoLA <sup>[65]</sup>	清华大学	基于 19 个关注实体、概念和事件的任务。参考了 Bloom 认知体系，从知识的记忆、理解、应用和创造 4 个层级，从深度而非广度去衡量大语言模型处理世界知识的能力	<a href="https://github.com/THU-KEG/KoLA">https://github.com/THU-KEG/KoLA</a>
DynaBench <sup>[66]</sup>	Facebook	第一个用于人工智能领域的动态数据收集和基准测试平台,使用人类测试和模型一起循环迭代,目的是为了创造具有挑战性的新数据并且更优化的人工智能模型	<a href="https://github.com/facebookresearch/dynalab">https://github.com/facebookresearch/dynalab</a>
GLUE-X <sup>[67]</sup>	微软亚洲研究院	基于分布外泛化的自然语言理解模型测试	<a href="https://github.com/YangLinyi/GLUE-X">https://github.com/YangLinyi/GLUE-X</a>
AlpacaEval <sup>[68]</sup>	斯坦福大学	用于对抗生成式预训练 (GPT) 模型的弱点,并更全面地反映模型的真实能力	<a href="https://github.com/ashmita-5/NLP-Assignment-8---Instruction-Tuning-with-hf-AlpacaEval">https://github.com/ashmita-5/NLP-Assignment-8---Instruction-Tuning-with-hf-AlpacaEval</a>
PandaLM <sup>[69]</sup>	微软亚洲研究院	第一个评测大模型的大模型,进行保护隐私、可靠、可复现及廉价的大模型评估	<a href="https://github.com/WeOpenML/PandaLM">https://github.com/WeOpenML/PandaLM</a>
SOCKET <sup>[70]</sup>	微软亚洲研究院	包含 58 个 NLP 任务测试社交知识,分为五个类别:幽默和讽刺、攻击性、情感和可靠性	<a href="https://github.com/facebookincubator/SocketRocket">https://github.com/facebookincubator/SocketRocket</a>
MATH <sup>[71]</sup>	加州大学伯克利分校	测评大模型在数学领域的推理和解决问题的能力	<a href="https://github.com/ossu/math">https://github.com/ossu/math</a>
APPS <sup>[72]</sup>	加州大学伯克利分校	测评大模型代码生成能力	<a href="https://github.com/serhii-londar/open-source-mac-os-apps">https://github.com/serhii-londar/open-source-mac-os-apps</a>
OpenLLM <sup>[73]</sup>	HuggingFace	由 HuggingFace 组织的一个大模型评测榜单	<a href="https://github.com/traceloop/openllmtry">https://github.com/traceloop/openllmtry</a>
OpenCompass	上海 AI 实验室	70+数据集, 40 万道题目	<a href="https://github.com/open-compass/opencompass">https://github.com/open-compass/opencompass</a>
Xiezhi (懈豸)	复旦大学	516 个具体学科, 249587 道题目	<a href="https://github.com/MikeGu721/XiezhiBenchmark">https://github.com/MikeGu721/XiezhiBenchmark</a>
FlagEval	智源	22 个评测数据集, 84,433 道题目	<a href="https://github.com/FlagOpen/FlagEval">https://github.com/FlagOpen/FlagEval</a>
BBT CFLEB <sup>[74]</sup>	超对称(北京)科技有限公司	涵盖 8 个标准语言任务, 衡量不同的模型的多维能力	<a href="https://github.com/ssymmetry/BBT-FinCUGE-Applications">https://github.com/ssymmetry/BBT-FinCUGE-Applications</a>
FinEval <sup>[75]</sup>	上海财经大学	包括 4,661 个问题, 涵盖 34 个不同的学科	<a href="https://github.com/SUFE-AIFLM-Lab/FinEval">https://github.com/SUFE-AIFLM-Lab/FinEval</a>
FinanceIQ		涵盖 10 个金融大类及 36 个金融小类, 总计 7173 个单项选择题	<a href="https://github.com/manan3101/FinanceIQ-Assistant">https://github.com/manan3101/FinanceIQ-Assistant</a>
MultiMedQA <sup>[76]</sup>	Google	涵盖了医学考试、医学研究等领域的问题和回答, 能够更全面地测试大模型的能力	<a href="https://github.com/monk1337/MultiMedQA">https://github.com/monk1337/MultiMedQA</a>
CRITIQUELLM <sup>[77]</sup>	智普 AI	针对各类指令遵循任务上大模型的生成结果提供高质量的评价分数和评价解释	<a href="https://github.com/thu-coai/CritiqueLLM">https://github.com/thu-coai/CritiqueLLM</a> (即将发布)
MentaLLaMA <sup>[78]</sup>	Meta AI	Meta Llama 大模型注重可解释性和公平性, 通过设计透明的网络结构和引入公正性评估指标, 致力于减少人工智能系统中的偏见和不公平性。	<a href="https://github.com/SteveKGYang/MentalLLaMA">https://github.com/SteveKGYang/MentalLLaMA</a>
Adversarial GLUE <sup>[79]</sup>	美国伊利诺伊大学	评估现代大规模语言模型在各种类型的对抗性攻击下的漏洞	<a href="https://github.com/AI-secure/adversarial-glue">https://github.com/AI-secure/adversarial-glue</a>
ASSERT <sup>[80]</sup>	加利福尼亚大学	自动生成涵盖不同类型鲁棒性的提示的方法, 探索了人工智能安全关键领域中大型语言模型鲁棒性的自动评估 CVALUES 手动收集了 10 个场景的对抗性安全提示, 并由专业专家诱导了 8 个领域的责任提示。	<a href="https://github.com/webmozarts/assert">https://github.com/webmozarts/assert</a>

CVALUES <sup>[81]</sup>	阿里巴巴达摩院	手动收集了 10 个场景的对抗性安全提示，并由专业专家诱导了 8 个领域的责任提示。	<a href="https://github.com/X-PLUG/CValues">https://github.com/X-PLUG/CValues</a>
GOAT-Bench <sup>[82]</sup>	香港浸会大学	包含超过 6K 个不同的模因，涵盖一系列主题，由包括五个相互交织的仇恨、厌女、冒犯、讽刺和有害性等方面的开放数据组织构建。	<a href="https://github.com/Ram81/goat-bench">https://github.com/Ram81/goat-bench</a>
SC-Safety <sup>[83]</sup>		针对中文 AIGC 大模型的多轮开放式问题对抗安全基准。包含 4912 个开放式问题，涵盖 20 多个安全子维度，系统地评估 AIGC 大模型的安全性。	<a href="https://github.com/CLUEbenchmark/SuperCLUE-Safety">https://github.com/CLUEbenchmark/SuperCLUE-Safety</a>
SafetyBench <sup>[84]</sup>		评估 AIGC 大模型安全性的综合基准，其中包含 11,435 个不同的多项选择题，涵盖 7 个不同的安全问题类别。	<a href="https://github.com/thu-coai/SafetyBench">https://github.com/thu-coai/SafetyBench</a>
SAFETEXT <sup>[85]</sup>	清华大学	则探索语言模型中的物理安全性的基准。SAFETEXT 是一个包含常识性的物理安全数据集，其中包含人工编写的现实生活场景以及每个场景的安全/不安全建议对。	<a href="https://github.com/sharonlevy/SafeText">https://github.com/sharonlevy/SafeText</a>

单纯针对语言能力的测评基准，大模型更多地关注了对话交互过程。其中，聊天机器人 Arena<sup>[61]</sup>通过用户参与和投票评估模型在现实场景中的性能。类似的，MT-Bench<sup>[62]</sup>通过设计模拟真实世界背景的问题来测评多轮对话的大语言模型。相比于特定任务上的大模型测评，HELM<sup>[44]</sup>从内容生成、内容理解和上下文连贯性等方面，全面评估模型在不同任务和领域的性能。另外，Big-Bench<sup>[43]</sup>通过引入 132 个机构的 450 位作者分享的 204 项任务，涵盖数学、语言学、生物学、物理学等。评估多模态大模型的 MME 基准<sup>[63]</sup>采用指令-答案对的形式评估模型感知和认知能力。L-Eval 基准<sup>[64]</sup>用于评估长上下文语言模型，采用多种指令风格和测评方法，使得针对长上下文语言模型的测评更加可靠。KoLA<sup>[65]</sup>是专门评估大模型语言理解和推理能力的评估基准，该基准主要强调运用知识和理解推理的能力，在语言理解层面有重要意义。在交互式环境中，DynaBench 基准<sup>[66]</sup>研究了针对对抗攻击的鲁棒性评估，推动了动态数据集收集工作。GLUE-X<sup>[67]</sup>也强调了测评鲁棒性的重要性。AlpacaEval<sup>[68]</sup>提供了一系列指标和衡量标准，自动评估大模型在不同语言处理任务中的性能。PandaLM<sup>[69]</sup>作为一种有判别性的大模型，着重关注公平性评估。

参考以人为中心的标准化考试，最早的综合测评基准 MMLU<sup>[37]</sup>已经被广泛用于测评大模型在多个学科上的表现。该测评涵盖 57 个任务，包括基础数学、美国历史、计算机科学、法律等，用于分析

跨领域的模型任务并找出模型的重要缺点。AGIEval<sup>[40]</sup>选取了 20 种面向普通考生的官方、公开、高标准的资格考试题目，包括普通大学入学考试（如中国的高考和美国的 SAT）、司法考试、数学竞赛等来测评大语言模型的表现。为了系统测评大模型性能，GAOKAO<sup>[38]</sup>采用中国高考的问题作为测试样本测评大模型。结果显示，虽然 ChatGPT 在解决客观问题方面表现出色，但在某些逻辑推理和数学问题以及中文较长文本阅读理解方面表现不佳。清华大学和上海交通大学提出的 C-EVAL<sup>[39]</sup>用于测评中文文化背景下基础模型的高级知识推理能力。该测评涵盖了初中、高中、大学和专业四个难度级别的多项选择题，涵盖了 52 个不同的学科。在最新的测评结果中，仅有 GPT-4 的平均准确率可以达到 60% 以上，测评区分度较好。此外，针对中文大模型测评，CMMLU<sup>[41]</sup>涵盖了 67 个主题，涉及自然科学、社会科学、工程、人文及相关常识，可以全面测评大模型在中文知识储备和语言理解能力。为了完成多语言、多模态测评任务，M3Exam<sup>[42]</sup>利用人类考题构建了多语言、多模态、多级别的测试，共涵盖 12317 道题目，以此推动多语言、多模态测评的发展。这些测评基准的建立和应用为评估大模型在不同领域和任务上的性能提供了有价值的参考和指导。SOCKET<sup>[70]</sup>由多项任务和案例研究构成，用来测评大模型在学习和识别社会知识过程中的表现。MATH<sup>[71]</sup>用于测评大模型在数学领域的推理和解决问题的能力。APPS<sup>[72]</sup>衡量大模型根据自然语言规



范生成 Python 代码能力。PromptBench<sup>[56]</sup>重点关注大模型的微调, 提出一个标准化框架进行测评, 实现大模型微调的增强和优化。OpenLLM<sup>[73]</sup>提供了一个开放的平台, 以鼓励研究人员提交他们的模型并在不同的任务上进行测评, 从而推动大模型领域的进步。另外, 上海人工智能实验室开源的大模型评测平台 Open Compass, 涵盖学科、语言、知识、理解、推理等五大评测维度, 全面评估大模型能力。值得一提的是, 以自测的方式针对大模型的综合评估基准 Xiezhi (獬豸) 包括 249587 道选择题, 跨越 516 个不同的学科, 包括金融、医学、心理学、工程学、历史等, 共有四个难度等级。FlagEval (天秤) 大模型评测体系及开放平台全方位评估基础模型及训练算法的性能, 同时探索利用 AI 方法实现对主观评测的辅助, 大幅提升评测的效率和客观性。此外, 金融领域也出现了准 BBT CFLEB、FinEval、FinanceIQ 测评基准。其中, BBT CFLEB 是中文领域金融大模型的专业的评测数据集, 涵盖 8 个标准语言任务, 用以衡量不同的模型的多维能力, 并促进金融大模型研发。FinEval 是一个中文的包含高质量多项选择题的集合, 涵盖金融、经济、会计和证书等领域。它包括 4,661 个问题, 涵盖了 34 个不同的学术科目。FinanceIQ 是一个专注于金融领域的中文评估数据集, 重点评估大语言模型在金融场景下的知识和推理能力。FinanceIQ 涵盖了 10 个金融大类及 36 个金融小类, 总计 7173 个单项选择题。

### (3) 数据准备

数据集是模型测评的关键, 我们跟踪现有数据集, 并将其分为两类: 标准数据集和微调数据集, 并同时融入了大模型特有的 Prompt 提示相关内容, 以下将依次说明。

#### 1) 标准数据集

标准数据集在大模型测评中是高质量数据的重要来源。许多研究工作都是建立在现有的标准数据集的应用之上<sup>[86-90]</sup>。在早期的模型测评阶段, 研究人员主要关注数据集的真实性和可靠性, 包括数据集获取渠道, 数据集的完整性和代表性。为了保障模型性能, 研究人员会对数据进行特征选择、异常处理和去噪等操作以提高数据集质量。然而, AIGC 大模型通常需要更大规模的数据集、更广泛的数据预处理方式以及更强大的计算资源和存储资源来进行训练和测评。以 VQA 数据集为例, 原始样本是一个输入输出对, 其中输入包括图像和自

然语言问题, 输出是基于图像问题的文本答案。这些数据集输入输出对的输入输出是由 GPT 辅助的半自动生成。具体来说, 一些研究工作手动设计 Prompt 描述, 并使用其提示 GPT 生成更多的样本<sup>[91-92]</sup>。但是, 现有 VQA 数据集和字幕数据集的答案通常都很简洁, 因此直接使用这些数据集进行 Prompt 描述可能会存在限制。为了解决这一问题, ChatBridge<sup>[90]</sup>明确给出了 Prompt 数据的简要描述, 来规避上述限制。同样, M<sup>3</sup>IT<sup>[93]</sup>采取的措施是延长现有答案的长度, 通过使用原始问题、答案和上下文提示 ChatGPT 来重新表述原始答案。WebCPM<sup>[94]</sup>可以调整预训练的语言模型来模仿人类的网络搜索并生成基于搜索结果的答案。UltraChat<sup>[95]</sup>是一种多轮教学对话数据集, 包含大量人机交互, 使跨各种主题和指令的对话更加仿真。通过这些改进和调整, AIGC 大模型测评中使用的标准数据集已经能够很好地适应大模型测评的需求, 为测评过程提供更多丰富的信息和更全面的评估。

#### 2) 微调数据集

虽然现有的标准数据集为大模型测评提供了丰富的数据源, 但通常情况下不能很好地满足现实场景下的需求, 尤其是在涉及到多轮对话等复杂任务时。为了解决这个问题, 研究人员手工制作一些符合任务描述的样本作为原始样本, 然后提示 ChatGPT/GPT-4 以这些原始样本为指导来生成更多的样本。大规模的中文新闻故事情节数据集 CStory<sup>[96]</sup>, 人工注释的可解释的因果推理数据集 e-CARE<sup>[97]</sup>, 都可以用于大模型的测评。Allen AI 发布了史上最大文本数据集 Dolma<sup>[98]</sup>, 包含了来自网络、学术出版物、代码书籍和维基百科材料的 3 万亿个数据。LLaVA<sup>[99]</sup>将图像转换为标题, 并提示 GPT-4 在原始样本示例的上下文中生成新数据, 从而构建一个多模态 M-IT 数据集, 也称为 LLaVA-Instruct-150k。此外, 诸如 MiniGPT-4<sup>[100]</sup>、GPT4Tools<sup>[101]</sup>和 DetGPT<sup>[102]</sup>等工作开发了不同的多模态 M-IT 数据集来满足不同的任务需求。

与传统的机器学习模型测评不同, AIGC 大模型测评引入了提示(Prompt)数据作为交互过程中的主要数据输入方式。传统的机器学习模型直接输入数据进行推断或预测, 而大模型通过提示/提问来完成数据输入, 这使得大模型更具灵活性和适应性。在大模型测评过程中, 除了选择合适的测试数据集来评估模型在不同场景下的表现, 还需要合理地收

集 Prompt 数据用于测试模型性能,以确保模型在真实世界的应用中表现良好。

综上所述,通过微调数据集可以使大模型能够更好地适应现实世界的需求,并在更广泛的应用场景中发挥优势。

#### (4) 数据预处理

数据预处理是模型测评前的关键步骤,它对于提高模型的预测和泛化能力至关重要。在真实的数据中,常常包含了大量的缺失值和噪声数据,这些因素都会影响模型的训练效果。因此,使用数据前需要进行数据清理、数据集成、数据规约和数据转换这四个方面,旨在优化原始数据以适应后续的模型训练和测评。数据清理是通过填补缺失值、平滑噪声数据以及平滑或删除离群点等方法来解决数据的不一致性。数据集成则是将多个数据源中的数据汇聚为一个一致的数据存储。数据规约是在保持原数据完整性的前提下,得到数据集的规约表示,以便在规约后的数据集上能进行更好的训练。数据变换则是对数据进行规范化、离散化、稀疏化处理,以达到训练目的。通过数据预处理,可以提高模型的训练和测评效果,使模型更好地适应实际数据和任务需求。对于大模型的 Prompt 提示数据,无论是少量样本提示、基于知识生成提示,还是思维链(COT)提示,其中涉及的数据同样需要进行预处理<sup>[103]</sup>。

#### (5) 特征工程

特征工程在模型测评中扮演着至关重要的角色,其目的是对原始数据进行处理和转换,以提取和构造出更有意义和有用的特征,从而改善模型的性能和泛化能力。特征工程涉及多个方面的操作,包括特征选择、特征转换、特征构建和特征组合等。特征选择旨在减少冗余和噪声,根据任务需求从原始数据中选择与目标变量相关性较高或重要性较大的特征,以优化测评过程。特征转换对原始特征进行用编码、标准化或归一化等操作,以使其适应模型算法的假设和要求,确保了特征的可比性和一致性,从而提高模型的稳定性。特征构建则是根据领域知识或经验,从原始数据中创建新的特征,以丰富特征空间和捕捉更多的信息。特征组合则是将不同特征进行组合,形成更高级和复杂的特征表示,以提取更深层次的特征关系。

大模型测评中的特征工程进一步保证了模型的表达能力,性能和适应性。同时,特征工程需要结合领域知识、数据理解和模型需求进行,是模型测评前不可或缺的重要环节。对于模型的 Prompt 提

示数据,其中包含的特征直接影响能否准确地驱动大模型完成相应的生成任务。因此,特征工程对于优化 Prompt 提示质量具有不可替代的作用。除了完成传统的特征工程的相关任务外,还需要基于特征及其背景知识的逻辑关联,依托思维链这些高级 Prompt 提示形式规划特征的表示和输入,以确保模型在测评过程中具有更好的适应性和表达能力。

#### (6) 模型的建立与配置

模型的建立和配置过程包括确定模型的结构、激活函数、优化算法、损失函数,以及调整模型的超参数。合理的模型配置可以直接影响后续模型的训练和测评过程,从而影响模型的性能和精确度。在建立模型时,根据问题的特点和数据的情况来选择适合的模型结构。

在配置模型时,调整模型的超参数也是必不可少的。超参数是在模型训练过程中需要手动设置的参数,如学习率、批量大小、正则化系数等。模型的建立和配置是一个迭代的过程,需要根据模型的训练和测评结果不断调试和优化,以获得最佳的性能。在此过程中,需要结合数据的情况进行实验并根据实验结果进行调整。Li 等人<sup>[104]</sup>提出的 MoT,无需标注数据集和参数更新,可以通过预先思考和回忆进行自我改进。

### 1.2 测评指标选择

许多日常工作在没有正式测评指标的情况下就可以清楚地判断是否已经达到了预期目标,这种情况被称为"I know it when I see it"<sup>[105]</sup>。然而,对于复杂的任务,特别是像模型测评这样的任务,在测评基准基础上,研究人员需要有针对性地选择测评指标<sup>[106-107]</sup>。

目前,基于任务或模型本身的特性已经开发了各种定量指标来客观测评模型质量<sup>[108]</sup>。回顾之前模型测评的工作,尽管已经提出了许多测评方法,通常具有类似的指标,但由于无法明确模型内部真正运作原理,还是难以抉择哪一种指标是首选。因此,以测评任务角度来看 AIGC 大模型的测评,AIGC 大模型(如 GPT 系列)往往在大量的数据上进行预训练,然后在特定任务上进行微调。这种微调可以包括分类、回归、聚类、检测和生成等任务,以使模型在特定领域问题上表现良好。如表 2 所示,本文根据模型完成的任务的不同,对测评指标进行了介绍。

#### (1) 分类任务

分类任务包括各类检测工作的测评通常都是依赖于常见的准确率（Accuracy）、精确率（Precision）、召回率（Recall）和 F1 值来度量测评结果，这里就不再赘述。准确率关注整体预测准确性，而精确率和召回率仅关注预测结果为正样本的情况，F1 值综合考虑了精确率和召回率的平衡。另外 ROC（Receiver Operating Characteristic）反映了召回率和假阳性的关系，也在测评结果可视化过程中得到了广泛的应用。ROC 曲线及其 AUC（Area

Under Curve）值可以通过更直观且全面方式测评模型的分类准确性，在机器学习领域得到广泛应用。其中，ROC 曲线以召回率为纵轴，假阳性率为横轴，用于衡量模型在不同阈值下的表现，AUC 值是 ROC 曲线下方的面积，作为一个单一的测评指标可以在多类别问题上提供更全面的比较。同时，ROC 曲线通过不断调整模型的预测阈值，处理数据集中类别不平衡和选择最佳阈值问题。

表 2 AIGC 测评任务和相应的代表性测评指标

Table 2 AIGC evaluation tasks and corresponding representative evaluation metrics

任务类型	测评指标
分类	Accuracy: 准确率; Precision: 精确率; Recall: 召回率; F1-score: F1 值;
回归	MAE: 平均绝对误差; MSE: 均方误差; RMSE: 均方根误差; RME: 相对平均误差; R-squared: 决定系数;
聚类	Silhouette Coefficient: 轮廓系数; Calinski-Harabasz: CH 指数, 也称方差比准则; Davies-Bouldin: DB 指数; ARI: 调整兰德指数; NMI: 归一化互信息; PUR: 纯度; PRI: 成对排序指数;
检测	IoU: 交并比; AP: 平均精确度; mAP: 平均精确率均值;
图像生成	IS: 初始分数; FID: 距离得分; SSIM: 结构相似性; PSNR: 峰值信噪比; CLIP score: 是一种用于评估文生图或者图生图, 模型生成的图像与原文本或者原图关联度大小的指标。
文本生成	BLEU: 生成文本与参考文本之间的相似性; ROUGE: 生成文本与参考文本之间的最长公共子序列的重叠系数; METEOR: 生成文本与参考文本之间的语义和词汇匹配; CIDEr: 生成文本与参考文本之间的质量; Perplexity: 困惑度, 生成文本和参考文本之间的质量; ChrF: 生成文本与参考文本之间的相似性和质量; Elo Rating: 评估参与对弈游戏的选手之间相对实力, 根据比赛结果来估计每位选手的等级或能力, 并预测其在未来比赛中的胜率, 应用在文本生成任务中, 用于比较生成文本和参考文本之间的质量; COMET: 除生成文本和参考文本之间的相似性外, 还包括生成文本的流利性、直接性; BLEURT: 在涉及更复杂语言结构和上下文情况下, 评估生成文本和参考文本之间的相似性;
其他	针对不同的任务, 研究人员开发了专门的评估指标体系, 以更有针对性地评估模型的性能。举例来说, 情感分析任务, 其中模型需要判断一段文本的情感倾向; Yelp 评论数据集下的 Cohen's Kappa 系数, 以考虑模型和人类标注者之间的一致性, 从而更好地评估情感分析模型的表现。

(2) 回归任务

基于回归模型的测评主要用于测评模型在预测连续目标变量时的性能表现。其中，一个简单方

法是计算预测值与真实值之间绝对差的平均值。这里平均绝对误差（Mean Absolute Error, MAE）、均方误差（Mean Squared Error, MSE）、均方根误差



(Root Mean Squared Error, RMSE)、相对平均误差 (Relative Mean Error, RME) 从不同角度展现了预测误差。

除此之外, 决定系数 (Coefficient of Determination, R-squared) 用于衡量模型对观测值变异性的解释程度的测评指标, 取值范围在 0 到 1 之间, 越接近 1 表示模型对观测值的变异性解释程度越高, 模型的拟合效果越好。然而, 对于复杂模型可能出现过度拟合的情况, 导致决定系数接近于 1, 但模型的泛化能力较差。因此, 在使用进行模型测评时, 还需要结合其他指标和领域知识综合考量。公式如下:

$$R^2 = 1 - \frac{SSR}{SST} \quad (1)$$

其中,  $SSR$  表示回归平方和, 表示模型预测值与观测值之间的差异的平方和。 $SST$  表示总平方和, 表示观测值与观测值均值之间的差异的平方和。

### (3) 聚类任务

#### ● 轮廓系数 (Silhouette Coefficient)

轮廓系数是一种常用的测评聚类质量的指标, 用于衡量聚类结果的紧密度和分离度。通过结合样本与其所属簇的距离和与其他簇的平均距离, 为每个样本分配一个取值范围在  $[-1, 1]$  之间的值, 数值越接近 1 表示聚类结果越好。轮廓系数也可以用于比较不同聚类算法或不同聚类结果的质量, 较高的轮廓系数表示更好的聚类效果。然而, 轮廓系数在聚类结果的类别数量较少或数据分布不均匀的情况下可能无法准确反映聚类质量。公式如下:

$$S(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}} \quad (2)$$

其中,  $i$  表示样本的索引,  $a(i)$  表示每个样本与同簇其他样本的距离,  $b(i)$  表示每个样本与其他簇中样本的最小距离。

#### ● Calinski-Harabasz 指数

CH 指数也是一种用于测评聚类结果的紧密度和分离度的一种指标。主要通过计算簇内的平方和与簇间的平方和的比值来衡量聚类结果的有效性。CH 指数越大, 表示聚类结果的紧密度越高且簇之间的分离度越好, 即聚类效果更好。与轮廓系数不同, CH 指数是基于簇内离散度和簇间离散度之间的比值进行计算的。但在簇的数量过多或数据分布不均匀的情况下也可能存在一些限制。公式如下:

$$CH = \frac{SB / (k - 1)}{SW / (n - k)} \quad (3)$$

其中,  $i$  表示样本的索引,  $SB = \sum n_i \times (x_i - c_i)^2$  表示簇内离散度,  $SW = \sum n_i \times (x_i - c)^2$  表示簇间的离散度,  $c_i$  表示簇内样本的平均值,  $c$  表示所有样本的全局平均值,  $x_i$  表示属于簇的样本,  $n$  表示样本总数,  $k$  表示簇的数量。

#### ● Davies-Bouldin 指数

同样的, DB 指数也是一种用于测评聚类结果的紧密度和分离度的一种指标。它通过计算每对簇之间的平均相似度和簇内样本的紧密度来度量聚类结果的优劣, 数值越小表示簇之间的分离度高、簇内的紧密度高, 表示聚类结果越好。然而, 该指数对聚类簇的形状和大小敏感, 以及在高维数据和不平衡数据集上的表现可能较差。公式如下:

$$DB = \frac{1}{N \times \sum \max((S_i + S_j) / d(c_i, c_j))} \quad (4)$$

其中,  $N$  表示聚类簇数,  $S_i$  表示簇  $i$  中所有样本与簇中心的距离的平均值,  $C_i$  表示簇  $i$  的中心点,  $d(c_i, c_j)$  表示簇  $i$  和簇  $j$  中心点之间的距离。

#### ● 调整兰德指数 (Adjusted Rand Index, ARI)

ARI 用于衡量聚类结果与真实标签之间的相似度。ARI 是在 RI 的基础上进行调整, 用于解决 RI 对于随机因素引起的聚类结果的不确定性的不足。在聚类任务中存在不平衡的簇大小或者存在噪声样本的情况下, ARI 可以提供更准确的结果。ARI 的取值范围在  $[-1, 1]$  之间, 值越接近 1 表示聚类结果与真实标签越相似, 值越接近 -1 表示聚类结果与真实标签越不相似, 值为 0 表示聚类结果与真实标签随机无关。公式如下:

$$ARI = \frac{RI - \text{Expected}(RI)}{\max(RI) - \text{Expected}(RI)} \quad (5)$$

其中, RI 表示在所有样本对中, 聚类结果和真实标签都相同或都不相同的样本对的比例, 即聚类结果与真实标签之间的一致性度量。 $\max(RI)$  表示在所有可能的聚类结果中, RI 的最大值。 $\text{Expected}(RI)$  表示如果聚类结果和真实标签之间没有关系, 根据随机分配的期望 RI 值。

$$RI = \frac{a + d}{a + b + c + d} \quad (6)$$

其中,  $a$  表示聚类结果中同一簇内的样本对数量,  $b$  表示聚类结果中不同簇之间的样本对数量,  $c$  表示真实标签中同一类别内的样本对数量,  $d$  表示



真实标签中不同类别之间的样本对数量。

● 归一化互信息 (Normalized Mutual Information, NMI)

NMI 用于度量聚类结果与真实类别标签之间的相似度。基于信息论的概念, 通过计算聚类结果与真实标签之间的互信息来度量它们之间的相关性。互信息反映了两个随机变量之间的相关程度, 包括它们共享的信息量和彼此独立的信息量。取值范围在 0 到 1 之间, 其中 1 表示完全一致, 0 表示完全不一致。公式如下:

$$NMI = \frac{2 \times I(C, T)}{H(C) + H(T)} \quad (7)$$

$$I(C, T) = \sum P(i, j) \cdot \log\left(\frac{P(i, j)}{P(i) \cdot P(j)} + \varepsilon\right) \quad (8)$$

其中,  $I(C, T)$  表示聚类结果和真实类别标签之间的互信息,  $H(C) = -\sum P(i) \cdot \log(P(i) + \varepsilon)$  表示聚类结果的熵,  $H(T) = -\sum P(j) \cdot \log(P(j) + \varepsilon)$  表示真实类别标签的熵,  $P(i, j)$  表示聚类结果中簇  $i$  和真实类别标签中类别  $j$  的样本占比。  $P(i)$  表示聚类结果中簇  $i$  的样本占比。  $P(j)$  表示真实类别标签中类别  $j$  的样本占比。  $\varepsilon$  表示为避免对数计算时出现无穷大或非数值的微小值 (通常取一个很小的正数, 如  $1e-10$ )。

● 纯度 (Purity, PUR)

PUR 是衡量聚类结果中同一类别的样本占比。在给定的聚类结果下, 将每个样本的真实标签与其他样本的标签进行比较所产生的不确定性。公式如下:

$$PUR = \frac{1}{N \times \sum \max(|C_i \cap G_j|)} \quad (9)$$

其中,  $N$  表示样本总数,  $C_i$  表示第  $i$  个聚类簇中的样本集合,  $G_j$  表示第  $j$  个类别中的样本集合,  $|C_i \cap G_j|$  表示聚类簇  $C_i$  与类别  $G_j$  的交集的样本数。

● 成对排序指数 (Pairwise Rand Index, PRI)

PRI 用于衡量聚类结果与真实标签之间的相似度。适用于评估具有不同簇大小的聚类结果。取值范围在  $[0, 1]$  之间, 值越接近 1 表示聚类结果与真实标签之间的一致性越高。公式如下:

$$PRI = \frac{a+b}{C} \quad (10)$$

其中,  $a$  表示聚类结果中同一类别的样本对数,  $b$  表示聚类结果中不同类别的样本对数,  $C$  表示样本对的总数。

(4) 检测任务

虽然 AIGC 图像大模型无法统一所有的视觉任务, 常见的 Stable Diffusion<sup>[19]</sup>、DALL·E 2<sup>[109]</sup>等生成类大模型主要聚焦于文生图和图生图, 同样涵盖了目标检测任务。

● 交并比 (Intersection over Union, IoU)

IoU 用于衡量目标检测任务中预测框和真实框之间的重叠程度。通常将 IoU 与设定的阈值进行比较, 以确定预测框是否与真实框匹配。一般来说, 当 IoU 大于等于设定阈值时, 认为预测框与真实框匹配, 否则认为不匹配。公式如下:

$$IoU = \frac{S_{com}}{S_{pre} + S_{real} - S_{com}} \quad (11)$$

其中,  $S_{com}$  表示预测框和真实框的重叠部分的面积,  $S_{pre}$  表示预测框的总面积,  $S_{real}$  表示真实框的总面积。

● 平均精确度 (Average Precision, AP)

AP 是用于衡量模型在单个类别上的准确率和召回率的平衡性。AP 的取值范围在 0 到 1 之间, 数值越高表示模型在该类别上的性能越好。公式如下:

$$AP = \frac{1}{n} \sum_{k=1}^n P(k) \cdot R(k) \quad (12)$$

其中,  $n$  表示检索出的结果数目,  $P(k)$  表示在前  $k$  个结果中的精确率,  $R(k)$  表示第  $k$  个结果是否为相关结果 (相关为 1, 不相关为 0)。

● 平均精确率均值 (mean Average Precision, mAP)

mAP 是用于衡量模型在不同类别上的平均精度。对于目标检测任务中存在多个类别的情况, 可以计算每个类别的 AP, 并取它们的平均值作为 mAP, 也可以单独报告每个类别的 AP 值。对于目标检测任务中存在多个类别的情况, 可以计算每个类别的 AP, 并取它们的平均值作为 mAP, 也可以单独报告每个类别的 AP 值。

(5) 图像生成任务

● 初始分数 (Inception score, IS)

IS 是测评模型生成图像质量和多样性的测评指标。IS 综合考虑了真实度和多样性, 它使用 Inception-V3 模型来计算生成图像的条件概率分布和互信息。具体的, IS 对生成图像进行分类预测, 计算每个生成图像的预测分布的熵, 表示真实度。

IS 计算生成图像集合的条件概率分布的 KL 散度，表示多样性。IS 的值越高，表示生成图像具有更好的真实度和多样性。然而，需要注意的是，IS 仅通过使用 Inception-V3 模型的分类能力来评估生成图像的质量，它并不考虑图像的细节和语义一致性。公式如下：

$$IS(G) = \exp(E_{x \sim p} D_{KL}(p(y|x) \| p(y))) \quad (13)$$

其中， $p(y|x)$  表示模型生成图像  $x$  的分类预测分布， $p(y)$  表示真实数据集的分类分布， $D_{KL}$  表示 KL 散度。

● 距离得分 (Frechet Inception Distance score, FID)

FID 是用于测评生成图像和真实图像之间的相似性。FID 越低，表示生成图像和真实图像之间的相似性越高，生成图像的质量越好。公式如下：

$$FID_{g,r} = \left\| \mu_g - \mu_r \right\|_2^2 + \text{Tr} \left( \sum g + \sum r - 2 \left( \sum g \sum r \right)^{\frac{1}{2}} \right) \quad (14)$$

其中， $\left\| \mu_g - \mu_r \right\|_2^2$  表示生成图像特征向量均值与真实图像特征向量均值之间的欧氏距离的平方。 $\text{Tr}(\cdot)$  表示生成图像特征向量协方差矩阵与真实图像特征向量协方差矩阵之间的迹的值。

● 结构相似性 (Structural Similarity Index, SSIM)

SSIM 用来衡量图像质量，包括考虑了生成图像与参考图像之间的亮度、对比度和结构信息。通过计算 SSIM 值，可以得到生成图像和参考图像之间的结构相似性的度量，数值越接近 1，表示两个图像之间的相似性越高。SSIM 能够提供更全面的图像质量评估，尤其在对比度和结构信息方面具有较高的灵敏度。公式如下：

$$SSIM(x, y) = [l(x, y) \cdot c(x, y) \cdot s(x, y)]^\beta \quad (15)$$

其中， $l(x, y) = \frac{2\mu_x\mu_y + C_1}{\mu_x^2 + \mu_y^2 + C_1}$  表示衡量图像亮度信

息的相似程度。 $c(x, y) = \frac{2\sigma_x\sigma_y + C_2}{\sigma_x^2 + \sigma_y^2 + C_2}$  表示衡量图像

对比度信息的相似程度。 $\sigma_x$  和  $\sigma_y$  分别表示生成图像

$x$  和参考图像  $y$  的对比度标准差。 $s(x, y) = \frac{\sigma_{xy} + C_3}{\sigma_x\sigma_y + C_3}$

表示衡量图像结构信息的相似程度。 $\sigma_{xy}$  表示生成图像  $x$  和参考图像  $y$  像素间的协方差。

● 峰值信噪比 (Peak Signal-to-Noise Ratio, PSNR)

PSNR 用于测评图像或视频重建质量，主要是用于比较原始图像（或视频）与经过压缩或其他处理后的重建图像（或视频）之间的相似程度。PSNR 值越大，表示重建图像与原始图像之间的差异越小，即图像质量越高，重建图像与原始图像越相似。但 PSNR 仅考虑了像素级的差异，而忽略了人眼对于图像细节和结构的感知。公式如下：

$$PSNR = 10 \cdot \log_{10} \left( \frac{MAX_I^2}{MSE} \right) \quad (16)$$

其中， $MAX$  值的最大可能取值， $MSE$  误差，表示原始图像与重建图像之间对应像素值之间的平方差的均值。

(6) 文本生成任务

● Bilingual Evaluation Understudy (BLEU)

BLEU 用于衡量机器翻译系统生成文本与参考文本之间的相似性。BLEU 通过比较生成文本中的  $n$ -gram（连续的  $n$  个词）与参考句子中的  $n$ -gram 的重叠度来计算得分，从而测评翻译质量。BLEU 的取值范围为 0 到 1，其中 1 表示完全匹配，更接近 1 的分数表示生成句子与参考句子之间的重叠度更高，即更好的翻译质量。通常使用 BLEU-1 到 BLEU-4 来评估不同长度的  $n$ -gram 重叠度。公式如下：

$$BLEU = BP \times \exp \left( \sum_{n=1}^N W_n \log P_n \right) \quad (17)$$

其中， $BP$  表示用于惩罚生成句子长度短于参考句子的情况，如果生成句子长度大于等于参考句子

长度， $BP = e^{1 - \frac{l_s}{l_c}}$ ； $l_s$  表示参考文本总数， $l_c$  表示生成文本总数，如果生成句子长度小于参考句子长度， $BP=1$ ； $W_n$  表示每个  $n$ -gram 的权重， $P_n$  表示生成文本在参考文本中所占的比例。

● Recall-Oriented Understudy for Gisting Evaluation (ROUGE)

ROUGE 用于衡量生成摘要和参考摘要之间的相似程度。ROUGE 主要关注召回率，即生成摘要是否能够涵盖参考摘要中的关键信息。ROUGE 包含多个变体，例如 ROUGE-N 用于衡量  $N$ -gram 的匹配情况、ROUGE-L 基于最长公共子序列的匹配度量，考虑了生成摘要与参考摘要之间的长度序列匹配情况。ROUGE-N 公式如下：

$$ROUGE-N = \frac{\sum Count_{match}(gram_n)}{\sum Count_{ref}(gram_n)} \quad (18)$$

其中,  $Count_{match}(\cdot)$  表示生成文本和参考文本中匹配的  $n$ -gram 序列的数量,  $Count_{ref}(\cdot)$  表示参考文本中的总  $n$ -gram 序列数量。

ROUGE-L 公式如下:

$$ROUGE-L = \frac{LCS}{REF} \quad (19)$$

其中,  $LCS$  表示生成文本和参考文本之间的最长公共子序列的长度,  $REF$  表示参考文本总长度。

● Metric for Evaluation of Translation with Explicit ORdering(METEOR)

METEOR 用于衡量生成文本与参考文本之间的相似度, METEOR 结合了多个词级和句级的比对方法, 包括精确匹配、词干匹配和同义词匹配。但对文本长度比较敏感。公式如下:

$$METEOR = (1 - Penalty) \frac{(\alpha^2 + 1)P}{R + \alpha P} \quad (20)$$

其中,  $Penalty = \gamma \left( \frac{chunks}{unigrams\_matched} \right)^\theta$  表示调整因子, 用于惩罚候选翻译中匹配词的不连续性。  $chunks$  表示生成文本与参考文本中匹配的短语数量,  $unigrams\_matched$  表示生成文本和参考文本中匹配的单词数量,  $\gamma$  是权重参数, 用于平衡短语匹配和单词匹配的重要性。  $P$  表示生成文本与参考文本的精确匹配得分。  $R$  表示生成文本与参考文本的召回率得分。  $\alpha$  是权重参数, 用于平衡精确匹配和不完全匹配的重要性。

● Consensus-based Image Description Evaluation (CIDEr)

CIDEr 用于衡量生成文本与参考文本之间的相似度。CIDEr 通过综合考虑多个参考文本之间的一致性来测评生成文本质量。相比于 BLUE 和 METEOR, CIDEr 更加注重多个参考文本中共同包含的特征信息。公式如下:

$$CIDEr_n(c, S) = \frac{1}{M} \sum_{i=1}^M \frac{g^n(c) \cdot g^n(S_i)}{\|g^n(c)\| \times \|g^n(S_i)\|} \quad (21)$$

其中,  $i$  表示参考文本的索引,  $M$  表示参考文本的数量,  $c$  表示生成文本,  $S$  表示生成文本集合,  $n$  表示测评的是  $n$ -gram,  $g^n$  表示基于  $n$ -gram 的 TF-IDF 向量。

● 困惑度(Perplexity)

困惑度用于衡量模型对给定数据的拟合程度和预测能力, 较低的困惑度表示模型更好地适应了数据, 并且能够更准确地预测词语序列, 而较高的困惑度表示模型的预测性能较差。困惑度本质上是计算词语序列的概率分布, 公式如下:

$$PPL(W) = P(w_1, w_2, \dots, w_N)^{\frac{1}{N}} \quad (22)$$

其中,  $w_1, w_2, \dots, w_N$  表示词语序列,  $N$  表示词语序列的长度,  $P(w_1, w_2, \dots, w_N)$  表示词语序列的概率分布。

● Character n-gram F-score(ChrF)

ChrF 是一种用于评估文本生成任务的评价指标之一。与其他文本生成评价指标类似, ChrF 旨在衡量生成文本与参考文本之间的相似性和质量。ChrF 通过计算生成文本和参考文本之间的字符级  $n$ -gram (连续的  $n$  个字符) 重叠来度量相似性。公式如下:

$$ChrF = \frac{(1 + \beta^2)(precision \times recall)}{(\beta^2 \times precision + recall)} \quad (23)$$

其中,  $precision$  是生成文本中的  $n$ -gram 与参考文本中的  $n$ -gram 之间的重叠比率,  $recall$  是参考文本中的  $n$ -gram 与生成文本中的  $n$ -gram 之间的重叠比率, 而  $\beta$  是一个权重参数, 用于平衡  $precision$  和  $recall$  的影响。

● Elo Rating

Elo Rating 是由物理学家阿帕德·埃洛创建的一个评价方法, 最开始用于评估国际象棋中棋手的水平, 后被广泛应用于围棋、足球和篮球等其他竞技项目中, 用于衡量竞技对弈选手之间的相对实力。将 Elo Rating 应用在文本生成任务中, 用于衡量不同模型之间的文本生成能力。通过为每个模型初始化一个分数, 以表示其初始的实力估计, 然后使用预定义的指标如 BLEU、ROUGE 等来评估每个模型生成文本的质量, 最后计算每个模型的预期胜率。通过预测竞技结果和每个模型生成文本质量以及之前的分数, 使用 Elo Rating 机制更新每个模型的分数, 模型打败比其能力更弱模型, 分数增加较少, 反之, 获得更多的分数, 重复进行分数计算和更新的步骤, 直至模型的分数趋于稳定。模型的分数越高, 则模型的生成能力越强。通过 Elo 机制, 可以得到表现最好的模型。公式如下:

$$R' = R + K \times (S - E) \quad (24)$$

其中,  $R$  是模型的初始分数,  $R'$  是模型更新之



后的分数,  $K$  用于调整分数变化的幅度,  $S$  是实际比赛结果,  $E$  是根据模型之间的分数差异计算的预期胜率。预期胜率  $E$  公式如下:

$$E = \frac{1}{1 + e^{-(R_a - R_b)}} \quad (25)$$

其中,  $R_a$  是模型  $a$  的分数,  $R_b$  是模型  $b$  的分数。

#### ● COMET

COMET<sup>[110]</sup>是一种用于训练多语言机器翻译评估模型的神经框架。该框架通过利用跨语言预训练模型产生多语言且适应性强的机器翻译来评估模型。具体而言, COMET 支持两种不同的架构: 估算器模型和翻译排名模型。估算器模型在给定源、假设和参考文本的嵌入向量, 采用 RUSE<sup>[111]</sup>提取源特征、参考特征、源差异和参考差异, 通过组合以上特征突出语义特征空间中嵌入之间的差异。翻译排名模型主要通过最小化源文本和参考文本在语义特征空间的距离, 最后将所得距离转换为介于 0 和 1 之间的相似度得分。综上, 前者用于计算质量分数, 后者用于比较生成文本和参考文本之间的相似度。

#### ● BLEURT

BLEURT<sup>[112]</sup>是一种基于 BERT 的评估指标, 通过用数千个可能有偏差的训练示例来模拟人类判断。相比与 BLEU, BLEURT 通过使用 BERT 捕获更丰富的语义信息, 解决了某些情况下对翻译质量的限制。其主要分为得分计算和参数优化两个部分。首先, 得分计算通过 BERT 计算生成文本和参考文本之间的相似度。然后通过优化参数, 使正确翻译在排序中的得分高于错误翻译。

#### (7) 其他

#### ● Cohen's Kappa

Cohen's Kappa (科恩的 Kappa 系数) 是一种用于衡量两位标注者之间一致性的统计指标, 通常用于评估分类问题的一致性。它考虑了实际观察到的一致性与预期一致性之间的差异。Kappa 系数的值范围在 -1 到 1 之间, 其中负数表示观察到的一致性低于随机一致性, 0 表示观察到的一致性与随机一致性相等, 1 表示观察到的一致性与随机一致性完全一致。Kappa 系数计算公式如下:

$$Kappa = \frac{P_o - P_e}{1 - P_e} \quad (26)$$

其中,  $P_o$  是实际观察到的分类一致性的比例,  $P_e$  是预期分类一致性的比例。  $P_e$  的计算公式如下:

$$P_e = \frac{\sum_{i=1}^k (n_{i+} \times n_{+i})}{n^2} \quad (27)$$

其中,  $n_{i+}$  表示标注者 1 在第  $i$  类上的标注数目,  $n_{+i}$  表示标注者 2 在第  $i$  类上的标注数目,  $i$  表示总的标注数目。

## 2 大模型的代表性应用领域

由于 AIGC 大模型在不同实际领域中的广泛应用, 相应问题逐渐显现。本章我们总结了金融、政法和医疗领域的大模型的发展以及暴露出来的问题, 并在表 3 总结了这些挑战。

### (1) 金融领域

AIGC 大模型的广泛应用为金融行业提供更准确的风险识别、市场趋势预测和交易决策支持。金融领域第一个大语言模型 BloombergGPT<sup>[113]</sup>正式发布。针对金融领域的大模型及其测评工作也已经涌现。例如, XuanYuan (轩辕)<sup>[114]</sup>在 BLOOM-176B 的基础上针对中文通用领域和金融领域进行了针对性的预训练与微调。Cornucopia (聚宝盆)<sup>[115]</sup>通过中文金融公开数据对 LLaMA 进行指令微调。此外, BBT-FinCUGE-Applications<sup>[116]</sup>开源了中文金融领域开源语料库 BBT-FinCorpus, 中文金融领域知识增强型预训练语言模型 BBT-FinT5 及中文金融领域自然语言处理评测基准 CFLEB。由于金融数据涉及大量敏感信息, 包括个人信息和财务状况, 模型测评中需要考虑数据安全和隐私。另外, 在实际应用中, 大模型会面临时刻变化的金融数据的不确定性问题, 因此需要持续监测和更新模型, 以保持其准确性和鲁棒性。对于模型的决策依据和推理过程, 使用者还需要考虑其可信性, 需要可解释性分析。

有关金融大模型的测评, 一种直接的方法是通过金融数据集 (由领域专家注释的高质量数据集、由 ChatGPT 或 GPT-4 构建的数据集和涵盖各种类型如新闻、报告和时间序列数据等的开源数据集) 测评其在下流任务上的表现, 比如使用分类指标 (召回率、精确率和 F1-score 等) 预测股票走势在投资策略中的潜在价值; 使用回归指标 (MAE、MSE 和 RME 等) 分析金融文本中的价格变动。此外, 直接评估大模型在特定任务上的性能的指标, 比如高级金融分析通过多轮问答对话给出受益报告; 金融情绪分析评测财经新闻的积极、消极或中性的情绪标准。除了特定任务评估外, 采用综合评估系统评测大模型整体的质量可以作为金融领域评估大模



型的指南以确保做出有意义的决策,包括涵盖不同场景的任务,包含不同方面的指标,包括公平性和鲁棒性等。

## (2) 政法领域

大模型在法律领域具有广阔的前景,不仅可以提高法律实践的效率,还能为法律从业者提供更好的决策支持,并为非法律人员提供法律服务。有一些具体应用已经涌现出来,展现了大模型在法律领域的应用潜力。例如, LexiLaw<sup>[117]</sup>是一个经过微调的中文法律大模型,通过在法律领域的数据集上进行微调,使其在法律咨询和支持方面具备更高的性能和专业性。LaWGPT<sup>[118]</sup>在通用中文基座模型的基础上扩充法律领域专有词表,并进行大规模中文法律预料预训练,从而增强了大模型在法律领域的基础语义理解能力。通过构建法律领域对话问答数据集和中国司法考试数据集,该模型进一步提升了对法律内容的理解和执行能力。Lawyer LLaMA<sup>[119]</sup>在大规模法律语料上进行预训练,并结合 ChatGPT 进行指令微调,从而在具体应用场景中增强了模型的应用能力,其中包括中国国家统一法律职业资格考试客观题的分析和对法律咨询的回答。还有一部分工作甚至促进其前沿的发展,例如, Yao 等人<sup>[120]</sup>提出的大规模法律事件数据集 LEVEN 也为政法领域的大模型提供了支持。因此,在政法领域应用大模型的过程中,需要保证模型推理的可靠性,在保证提示模型鲁棒性同时避免安全隐私问题。

随着大模型的不断进步,评测其在政法领域理解和分析政策和法律能力是重中之重。一种可行的方法是利用大模型内的监管推理和法律推理,通过民主程序和立法实现与所确定的社会价值观相一致的“法律知情的人工智能”。这种“法律告知守则”依赖于通过反复辩论和诉讼创建的适应性政策和法律标准,采用问题生成、政法讨论等形式实现民主程序的既定有效性。比如评测大模型在早期预测信托义务何时被违反的能力。此外,鉴于政法工作敏感性和专业性,除了部署大模型除了采取严格的数据隐私、减少偏见、做出决策保持问责制等,评估大模型从政策和法律分析,落地和预测等任务,采用检索增强等方式,评估大模型仅从内部知识库回答问题的能力,降低难以驾驭政策和法律体系的人适用性的难度是评测的另一面。

## (3) 医疗领域

在医疗领域,信任问题<sup>[130]</sup>是一个非常重要且不

容忽视的因素,它极大地制约了医疗 AIGC 大模型的应用<sup>[121-124]</sup>。近期出现的医疗领域大模型,如 DoctorGLM<sup>[125]</sup>、BenTsao<sup>[126]</sup>、BianQue<sup>[127]</sup>、Huatu-oGPT<sup>[128]</sup>等的广泛应用,使得对医疗大模型的测评变得尤为重要。谷歌医疗领域模型 Med-PaLM2<sup>[129]</sup>可以回答患者的问题,并从海量的医学语料中总结出专业的医疗知识。该模型是目前第一个在美国医疗执照考试上达到“专家”水平的大语言模型。此外,谷歌还考虑加入多模态功能(例如输入 X 光片)以便模型能整合更多维度的医疗信息。这些模型在医疗问答效果、病例自动生成和医学影像分析等方面发挥着重要作用。由 7 个医学 QA 相关的数据集组成的医学评估基准 MultiMedQA<sup>[130]</sup>主要测评大模型在临床的表现。

评估医学大模型通过采用各种指标评估大模型处理临床数据或回答标准化测试问题的能力。大模型的新兴能力的评估已显然扩展了其潜力,超越了主要围绕文本处理和问题回答的传统标准化医学语言处理任务使用 ROUGE、METEOR 和 BLEU 进行的评估。然而,由于缺乏医学证据总结的错误类型的标准化术语,生成式医学人工智能将各种类型的医疗数据(电子病历、医学图像)映射到所需的输出不受现有知识、参考或数据的支持,存在“幻觉”,评估医学大模型受到了限制。考虑到医学的高风险性,对事实错误的容忍度应该降低,评估大模型生成输出的事实一致依赖于各种生物医学数据的可靠性。

因此,医疗 AIGC 大模型离不开可靠的医疗数据,大模型的测评也面临着数据质量等方面的挑战。在现实世界中,医疗机构并非出于模型研究目的而收集医疗数据,因此收集到的数据可能包含偏见、错误或不完整的信息。保证数据质量和选择数据收集方式对于医疗 AIGC 大模型的应用和测评至关重要。在测评医疗大模型时,需要综合考虑数据质量、数据真实性以及模型鲁棒性和安全性等因素。增强医疗大模型的信任度,从而促进其在医疗领域的应用和推广。

## 3 大模型测评的新挑战和应对之道

在 AIGC 大模型实际应用之前,需要对其进行系统性的测评,以确保模型的性能和准确性,发现潜在问题,推动模型改进,保证模型的顺利部署和应用。考虑到 AIGC 大模型的复杂结构和黑盒特性,

在落实上述测评需求过程中，不可避免地需要应对全新挑战。因此，本章将从可解释性入手，对大模型测评需要关注的新问题进行解构，以期获得全面的研究分析结论，为未来 AIGC 大模型测评奠定基础得出。

### 3.1 可解释性测评

AIGC 大模型属于深度神经网络模型。与其它

机器学习模型不同，AIGC 大模型作为一种复杂的黑盒模型，目前仍无法为深度神经网络构建一套数学模型，即无法基于输入数据推导模型输出<sup>[107]</sup>。对生成结果可靠性有要求的领域例如自动驾驶、医疗诊断和财务决策，这种不透明可能会造成严重后果。对可解释性的测评因任务和模型而不同，目前无法使用普适的方法完成可解释性分析<sup>[131]</sup>。因此，对于 AIGC 大模型，对可解释性的测评尤为重要<sup>[132-133]</sup>。

表 3 AIGC 测评任务和相应的代表性测评指标

Table 3 AIGC evaluation tasks and corresponding representative evaluation metrics

文献	代表性领域	新挑战				
		可解释性	公平性	鲁棒性	安全性	隐私性
BloombergGPT <sup>[113]</sup>	金融				√	√
XuanYuan <sup>[114]</sup>	金融			√		√
Cornucopia <sup>[115]</sup>	金融			√	√	√
BBT-FinCorpus <sup>[116]</sup>	金融			√		√
LexiLaw <sup>[117]</sup>	政法		√	√	√	√
LaWGPT <sup>[118]</sup>	政法		√		√	√
Lawyer LLaMA <sup>[119]</sup>	政法		√		√	√
LEVEN <sup>[120]</sup>	政法		√		√	√
ChatGPT 医疗评估 <sup>[121]</sup>	医疗	√	√	√	√	√
ChatGPT 生育咨询 <sup>[122]</sup>	医疗	√		√	√	√
ChatGPT 遗传性分析 <sup>[123]</sup>	医疗	√	√	√		
ChatGPT 行医执照考试 <sup>[124]</sup>	医疗	√	√		√	√
DoctorGLM <sup>[125]</sup>	医疗	√		√	√	√
BenTsao <sup>[126]</sup>	医疗	√		√	√	√
BianQue <sup>[127]</sup>	医疗	√		√	√	√
HuatuogPT <sup>[128]</sup>	医疗	√				√
Med-PaLM2 <sup>[129]</sup>	医疗	√		√		√
MultiMedQA <sup>[130]</sup>	医疗				√	√

可解释性测评是为决策过程提供解释或解释能力的过程。它关注模型如何解释其预测结果、决策或推荐，并提供了对模型内部逻辑和决策依据的可解释性。可解释性测评可以增强模型的可信度和可接受性<sup>[62,134]</sup>。传统模型测评对模型可靠性进行更细粒度的评估,从而消除模型在实际部署应用中的潜在风险。同时，一些方法采用可视化模型的中间激活状态、内部特征表示方式以及可视化模型中的数据流程图来支撑模型可解释性<sup>[135]</sup>。关于可解释性的研究，一方面从模型本身入手，调整内部参数，对系统得到的结果进行分析，测评系统中不同变量的

重要性，推测系统作出决策的依据。另一方面，直接构建本身具有可解释性质的模型，去探索更结构化的可解释性<sup>[136]</sup>。当模型能够提供清晰、透明的解释时，用户可以更好地理解模型如何得出预测结果或决策，从而更容易信任和接受模型的输出。此外，可解释性测评还有助于揭示模型的内部逻辑和决策过程<sup>[137-138]</sup>。

自然语言处理社区开始制作大型语言模型充当批评家来评估生成文本的质量，并针对模型的可解释性进行测评。CRITIQUELLM<sup>[77]</sup>考虑评估大模型的关键因素，基于对话的提示方法获得的高质量

参考或无参考评估数据进行训练。该模型可用于评估预训练大模型的预测性能和可解释性的基准,并提供英文和中文评估集。该模型的扩展特性以及生成的批评作对 AIGC 大模型生成质量产生了积极影响。虽然预训练语言模型为许多 NLP 任务带来了巨大改进,但人们越来越关注探索模型的功能并解释其预测。然而,现有的工作通常只关注某些能力和一些下游任务。缺乏直接评估掩码词预测性能和预训练语言模型可解释性的数据集。Shen 等人<sup>[139]</sup>提出了一种新颖的评估基准,提供英文和中文注释数中尤其如此,可解释性研究的重点是仅适用于合成数据集的“解开”措施,而不是基于人为因素。ROSS 等人<sup>[140]</sup>引入了一项任务来量化 AIGC 大模型表示的人类可解释性,其中用户交互地修改表示以重建目标实例。该任务区分了广泛相信但从未被证明可以产生或多或少可解释模型的表示学习方法。因此,性能与人类模型理解有意义地相关,改善合成数据集解纠缠的方法也可以提升 AIGC 大模型的可解释性。

以对话形式提供解释的可解释性工具已证明其在增强用户理解方面的功效。然而当前基于对话的解释的解决方案需要许多依赖性,并且不容易转移到它们不是设计用于的任务。LLMCHECKUP<sup>[141]</sup>提供了一个易于访问的工具,允许用户与任何最先 AIGC 大模型讨论其行为。LLMCHECKUP 使 AIGC 大模型能够自行生成所有解释,并通过将其与广泛的可解释人工智能工具连接起来,无需进行微调即可处理意图识别。AIGC 大模型自学解释以互动对话的形式呈现,支持后续问题并生成建议。LLMCHECKUP 提供系统中可用操作的教程,适合具有不同 XAI 专业知识水平的人,并支持多种输入模式。

Lei 等人<sup>[142]</sup>提出了一种新的推荐系统模型解释方法,研究了使用 AIGC 大模型作为替代模型来增强推荐系统的可解释性的潜力。该方法通过使用 AIGC 大模型作为代理模型,并学习模仿和理解目标推荐模型。具体来说,包括:行为对齐、意图对齐和混合对齐。行为对齐在语言空间中进行,将用户偏好和物品信息表示为文本,以学习推荐模型的行为;意图对齐在推荐模型的潜在空间中工作,使用用户和项目表示来理解模型的行为;混合对齐结合了语言和潜在空间来进行对齐训练。该方法有效地使 AIGC 大模型能够理解推荐模型的模式并生成

据。该评估基准涵盖了常见的评估维度,即语法、语义、知识、推理和计算。此外,还创建扰动实例并通过扰动下基本原理的一致性来评估忠实度。它包含每个原始实例的扰动实例,以便使用扰动下的基本原理一致性作为忠实度的度量,这是可解释性的一个角度。通过这样多角度的测评,可以有效帮助改进 AIGC 大模型。

然而,尽管人们对可解释性的兴趣日益浓厚,但对于如何衡量它仍然没有达成共识。在表示学习

高度可信的推荐解释。

由于传统的判别方法泛化能力差、可解释性低,最近 AIGC 大模型已被探索用于社交媒体上的可解释心理健康分析,旨在提供详细的解释以及零样本或少样本设置中的预测。MentaLLaMA<sup>[75]</sup>将可解释的心理健康分析正式建模为文本生成任务,并构建了第一个具有 105K 数据样本的多任务、多源可解释心理健康指令(IMHI)数据集,以支持 AIGC 大模型指令调整和评估。此外,为了确保解释的可靠性,MentaLLaMA 对生成数据的正确性、一致性和质量进行严格的自动和人工评估。最新的 AIGC 大模型在自动化心理健康分析方面展现出强大的能力。然而,现有的相关研究存在一些局限性,包括评估不充分、缺乏激励策略以及忽视探索 AIGC 大模型的可解释性。为此,Yang 等人<sup>[143]</sup>在 5 个任务的 11 个数据集上全面评估了 AIGC 大模型的心理健康分析和情绪推理能力。该方法通过 CoT 提示解释大模型的预测,探讨了 AIGC 大模型在可解释的心理健康分析中的潜力。与此同时,开发了一种可靠的注释协议,用于对 AIGC 大模型生成的解释进行人工评估,并对人工注释的现有自动评估指标进行基准测试。此外,利用无监督和远程监督的情感信息探索不同的提示策略的效果。根据这些提示,通过指导 AIGC 大模型为他们的每个决定生成解释,探索可解释的心理健康分析。

此外,可解释性还用于测评代码大模型。代码大模型中的可解释性测评是指理解和阐明这些模型在处理和生成代码时如何做出预测或决策的能力,涵盖对代码大模型的开发和部署至关重要的多个方面。例如,它可以指模型决策过程的透明度,了解代码的哪一部分使模型认为代码片段容易受到攻击。Ma 等人<sup>[144]</sup>研究了 CodeBERT 和 GraphCodeBERT 在理解代码语法和语义方面的可解释性。Li 等人<sup>[145]</sup>



通过向代码大模型提供不同的输入（包括不同的掩蔽率和足够的输入子集方法）来评估这些模型，表明这些模型在很大程度上对特定输入不敏感。用于解释 AIGC 大模型的其他技术包括反事实分析<sup>[146]</sup>、因果推理<sup>[147]</sup>和探测等。此外，许多研究中使用 LIME、BreakDown 和 SHAP 等与模型无关的解释技术来解释代码大模型的预测。然而，研究表明，不同方法产生的解释结果可能会相互冲突<sup>[148]</sup>。

研究人员还提出了增强代码大模型可解释性的方法，并展示了可解释性如何使其受益。Ji 等人<sup>[149]</sup>提出了一种基于因果关系分析的方法，可以提供对 AIGC 大模型有效性的见解并帮助最终用户理解预测。Palacio 等人<sup>[150]</sup>提出 ASTxplainer，它提供 AIGC 大模型预测的可视化，帮助最终用户理解模型预测。Zhang 等人<sup>[151]</sup>提出可解释的程序合成，旨在为用户提供对合成过程的洞察和控制。

总而言之，AIGC 大模型可解释性测评是一个多层次、多方位的过程，提高 AIGC 大模型的可解释性至关重要。上述测评方法目的在于提高对 AIGC 大模型的可解释性，同时推动可解释性研究的进一步发展，以满足不同层面和角度的评测需求。对于一般最终用户来说，可解释性通过以可理解的方式阐明模型预测背后的推理机制来建立适当的信任，而无需技术专业知识。这样，最终用户就能够了解 AIGC 大模型的功能、局限性和潜在缺陷。其次，对于研究人员和开发人员来说，解释 AIGC 大模型行为可以提供洞察力来识别意外偏差、风险和性能改进领域。换句话说，可解释性充当调试辅助工具，可以快速提高下游任务的模型性能。然而，由于 AIGC 大模型存在局限性，当前对大模型可解释性的测评工作较为缺乏<sup>[152]</sup>。因此，未来研究可以致力于制定全方位、多层次的测评基准，以提高 AIGC 大模型可解释性方法的测评质量和可信度，进而支撑公平性、鲁棒性、安全性和隐私性等方面的测评工作。

### 3.2 公平性测评

由于 AIGC 大模型依赖于训练数据，在训练过程中模型可能会有意或无意的偏向某个群体或个人，从而导致模型中存在偏见<sup>[153]</sup>。哈尔滨工业大学秦兵教授在“大语言模型之人类价值观对齐”报告中指出如何将大语言“型的价值观与人类对齐是当前待解决的重要问题。

大模型出现前，公平性主要用在人为生成内容

的评价。随着 AIGC 大模型的生成能力的提升，其生成内容的公平性也得到了广泛的重视。公平性测评的目标是确保 AIGC 大模型对不同群体和具有敏感特征的群体，在处理数据和做出决策时，避免社会中存在各种敏感属性，例如，性别、种族和年龄等。通常其相关个体数量会有较为显著的差异，反映在这些属性上，收集到的数据也并不均匀。如果 AIGC 大模型在处理数据和做出决策时存在偏见，和属性上可能会加剧社会的不平通过公平性测评，可以识别和纠正模型的不公平行为，确保模型在不同群体之间提供公正的服务，尤其在金融、政法和医疗等关键领域更为重要。公平性测评需要建立完善的测评指标和测评基准，才能够确保部署到实际应用中的 AIGC 大模型具有公平性，从而保证大模型的社会可接受性，减少公众的担忧和不信任<sup>[154]</sup>。

公平性测评的目的是确定 AIGC 大模型处理是否公平。识别和解决模型中存在的潜在偏见和歧视问题，确保模型的生成结果对所有个体和群体是公平的<sup>[155]</sup>。公平性可以涵盖多个维度，在不同的维度，准确定义公平性对于后续的测评和改进至关重要。其中几个常见的维度是个体公平、群体公平和机会均等<sup>[156-157]</sup>。个体公平关注模型对个体的公平对待，群体公平关注模型在不同群体之间的差异待遇，而机会均等则强调不同群体之间应享有平等的机会。通过不同的公平维度，可以量化和比较模型在不同方面的公平性表现，从而帮助改善模型的设计和决策过程的优化。

评估大模型的潜在偏见和公平性变得至关重要，因为现有方法依赖于仅关注少数群体的有限提示，缺乏全面的分类视角。Bi 等人<sup>[158]</sup>提出了一种表征不同社会群体的新颖的层次结构，从群体公平的角度评估大模型的偏见。具体来说，该方法构建了一个数据集 GFair，封装跨多个维度的目标属性组合。此外引入了语句组织，以揭示 AIGC 大模型中复杂的偏见。对热门 AIGC 大模型的广泛评估揭示了固有的安全问题。为了从群体公平的角度减轻 AIGC 大模型的偏见，该方法还提出了一种新颖的思想链方法 GF-Think，以从群体公平的角度减轻 AIGC 大模型的偏见。此外，由于 AIGC 大模型基于人类语言，潜在有害的偏见可能会扩散到 AIGC 系统中并产生不公平的结果、歧视少数群体或产生法律问题。因此，Freiberger 等人<sup>[159]</sup>系统地设计了 AIGC 大模型的 6 个公平标准，并可进一步细化为 18 个子类



别。该标准为操作和测试流程提供了基础,以从审核员和被审核组织的角度证明公平性。该方案有文献和专家访谈的支持,研究结果允许为广泛的AIGC应用程序开发公平性认证,包括大型语言模型和其他文本生成人工智能方法。

Huang 等人<sup>[160]</sup>着重于解决在量化和减少语言模型表现出的特定类型的偏差:生成文本的情感偏差。给定条件上下文和语言模型,分析生成文本的情绪是否受到敏感属性值变化的影响,使用反事实评估的形式调节上下文。该方法通过采用公平机器学习文献中的个人和群体公平指标来量化情绪偏差。然后,提出对语言模型的潜在表示进行嵌入和情感预测衍生的正则化。正则化提高了公平性指标,同时保留了可比水平的困惑度和语义相似性。此外,该方法使用自动指标和人类对情感和语义相关性的评估来评估所提出的方法,并发现自动指标和人类评估之间存在很强的相关性。

Zhuo 等人<sup>[161]</sup>使用传统的测试集和指标对 ChatGPT 的偏见、可靠性、稳健性和毒性进行了系统评估。该方法对 ChatGPT 的评估是在零样本设置下进行的,这更准确地反映了未提供上下文示例的典型人机交互场景。实验结果表明,在执行问题回答或文本生成任务时,与当前的 AIGC 大模型相比,ChatGPT 表现出较低水平的偏差。此外,ChatGPT 展示了无需进行上下文中的小样本学习即可完成任务的能力,这表明对有限上下文的全面理解。此外,这也意味着现有的 AIGC 大模型可能比 ChatGPT 存在更严重的偏见。Ferrara<sup>[162]</sup>则探讨了与大规模语言模型中的公平性相关的独特挑战和风险。这些偏差源于训练数据、模型规范、算法约束、产品设计和政策决策等因素。此外,该文分析了减轻偏见的复杂性,承认某些偏见不可避免地持续存在,并考虑在不同应用程序中部署这些模型的后果。

除了社会偏见之外,AIGC 大模型在信息搜索和自动决策辅助方面的爆炸性采用强调了了解其局限性和偏见的重要性。Hartmann 等人<sup>[163]</sup>重点关注民主社会最重要的决策过程之一:政治选举。通过来自两个领先投票建议应用程序的 630 条政治声明以及三个预先注册的实验中与国家无关的政治指南针测试来提示 ChatGPT,揭示了 ChatGPT 的亲环境、左翼自由主义意识形态。与此同时,研究人员<sup>[164]</sup>试图通过让 ChatGPT 回答有关常用政治偏见衡量标准的问题来研究 AIGC 大模型的自我认知和政治偏见。

结果显示显示出进步观点的倾向。此外,使用 OCEAN 测试来测试 ChatGPT 的大五人格特征,并使用迈尔斯-布里格斯类型指标(MBTI)测试来查询其人格类型。结果显示,ChatGPT 认为自己高度开放且令人愉快,具有 ENFJ 人格类型,并且是具有最不明显的黑暗特征的测试者之一。

多位研究者已经使用基于反事实公平性的方法对 AIGC 大模型进行测评。这种方法通过改变提示中的各种关键词来测试模型输出的一致性,任何输出的变化都可能指示偏差的存在。HELM<sup>[44]</sup>测评框架通过替换输入提示中的名词、术语以及性别和名字标识,并对比标准英语和非裔美国英语方言中的使用情况,检查 AIGC 大模型在处理与性别和种族相关的问题时的公平性水平。进一步地,Li 等人<sup>[164]</sup>针对 ChatGPT 在教育、刑事司法、金融和医疗保健这些高灵敏度领域的性能进行了公平性评估。这项评估使用了反事实公平的方法来考察 ChatGPT 在处理有偏见和无偏见提示时是否表现出公平性。他们设计了包含任务描述、案例上下文、数据特征和问题的四部分提示,使用了来自上述领域的具体数据集。研究结果指出,即使 ChatGPT 的表现超过了较小模型,但它在消除偏见方面仍然有待提高。

为了评价 AIGC 大模型行为中可能存在的偏差,研究者们常常观察 AIGC 大模型在一些任务上对不同人群的处理是否存在性能上的不一致,尤其是在问答任务中这一点格外明显。具体而言,BBQ<sup>[165]</sup>测评集就是专门开发来测评问答任务在处理九种社会偏见方面的性能,这需要考虑含糊与具体、肯定与否定问题形式以及多选答案。测评侧重于比较模型在面对各种问题时选择答案的差异,以此来判定模型响应是否存在偏差。UnifiedQA<sup>[166]</sup>上的实验结果表明,AIGC 大模型在缺乏充分上下文信息时,会不同程度地显示出依赖社会偏见的预测倾向;但随着上下文清晰度的增加,这种偏差倾向有所减少。利用 BBQ 的评估基准,研究工作 HELM<sup>[44]</sup>进一步对 30 种领先的 AIGC 大型模型所含偏见和成见进行了系统评估。

AIGC 大模型在各种高影响力的应用中取得了显著的成功,改变了我们与技术互动的方式。然而,如果没有适当的公平保障措施,AIGC 大模型就有可能做出可能导致歧视的决定,从而引发严重的道德问题和日益增加的社会关注。因此,公平性测评方法在 AIGC 大模型的测评中具有重要性,它们帮

助我们识别模型中的公平性问题，揭示潜在的平等，推动大模型朝着更加公正和包容的方向发展。这些方法的综合应用可以为我们提供全面的公平性测评，推动大模型的公平性发展，并为决策制定者和从业者提供更可靠和可信的模型结果。尽管 AIGC 大模型公平性得到了比较广泛的研究，并且在之前的一些工作中已经进行了讨论，但我们发现这些研究仍然有限，应该进行更多的探索。与此同时，AIGC 大模型仍处于开发更全面、对社会无害的制度阶段，其公平性是社会关注的焦点。在未来的发展中，我们可以进一步探索和改进公平性测评方法，以确保大模型的应用更加公正和包容。

### 3.3 鲁棒性测评

鲁棒性测评是评估 AIGC 大模型在面对各种不确定建议、对抗性攻击和领域转移等挑战时的性能和稳定性的过程<sup>[167]</sup>。它的目的是指 AIGC 大模型在真实世界环境中的行为表现，而不仅仅是在理想化的条件下评。鲁棒方法通过模拟实际应用场景中的各种变化和干扰，评估模型对这些变化的适其力和表现质量<sup>[168-169]</sup>测评。鲁棒性测评可以使用多种评估指标、实验设置和测试场景来量化和比较 AIGC 大模型的性能。通过鲁棒性测评，可以发现 AIGC 大模型的弱点、漏洞和局限性，并提供指导来改进 AIGC 大模型的鲁棒性和健壮性。这有助于确保 AIGC 大模型在复杂、不确定和恶劣条件下的可靠性和稳定测评提高 AIGC 大模型在实际应用中的效果和用户体验大。

鲁棒性是指当输入受到扰动或轻微变化时，AIGC 大模型能够一致且正确地执行操作。不稳健的 AIGC 大模型可能会造成负面影响，甚至灾难性的后果。鲁棒性测评可以分为两种类型：白盒测评和黑盒测评。白盒测评假设测试人员完全了解目标模型，包括其架构、参数（权重和偏差）、训练数据和训练算法。测试人员可以访问模型的内部状态和梯度来制作测试输入。在黑盒测评中，测试人员可能只能在给定特定输入的情况下查询模型的输出。这是一个更现实的场景，特别是对于模型远程托管且用户可以发送输入以获取输出的服务攻击者可以多次查询模型来制作对抗性示例。

Adversarial GLUE<sup>[79]</sup>是一种用于语言模型鲁棒性评估的多任务基准。该基准从不同的角度和层次考虑文本对抗性攻击，包括词级转换、句子级操作和人类编写的对抗性示例，以便 AdvGLUE 能够覆

盖尽可能多的对抗性语言现象。AdvGLUE 采用众包的方式来识别高质量的对抗数据以进行可靠的评估。为了全面了解语言模型在不同 NLU 任务中的鲁棒性，AdvGLUE 涵盖了广泛使用的 GLUE 任务，并创建了 GLUE 基准的对抗版本来评估语言模型的鲁棒性。此外，AdvGLUE 具有很高的对抗性可迁移性，可以有效地攻击各种最先进的模型。然而，然而，我们尚不了解 AIGC 大模型不同层的贡献所实现的固有鲁棒性。因此，Kashyap 等人<sup>[170]</sup>提出从两个角度系统地逐层分析 AIGC 大模型的鲁棒性。该方法通过使用诊断探针分析编码的句法和语义信息来研究表示的鲁棒性。研究发现，对于来自看不见的域的数据，相似的层具有相似数量的语言信息。

基于对抗性示例的可转移性，许多研究人员提出可转移性的方法：为一个模型生成的对抗性示例也可能对另一种模型有效。Liu 等人<sup>[171]</sup>首先训练替代模型来模仿受害者模型的行为。他们为替代模型生成对抗性示例，并用它们来攻击原始受害者模型。Zhang 等人<sup>[172]</sup>发现使用基于梯度的方法为较小的 AIGC 大模型生成的对抗性示例可以成功地转移到多个较大的 AIGC 大模型。

随着 AIGC 大模型融入社会，对一系列提示的鲁棒性对于在高方差环境中保持可靠性变得越来越重要。ASSERT<sup>[80]</sup>是一组自动生成涵盖不同类型鲁棒性的提示的方法，探索了人工智能安全关键领域中大型语言模型鲁棒性的自动评估。该方法使用新颖的语义对齐增强、有针对性的引导和对抗性知识注入方法以探索语言模型中的稳健性概念。ASSERT 语义对齐增强方法生成语义等效的提示，并且有针对性的引导创建具有相关但语义不等效的场景的样本。同时，对抗性知识注入会生成对抗性样本，旨在与不可信的知识结合时反转真实标签。然而，这种功能带来了即时注入攻击的风险，攻击者将指令注入 AIGC 大模型的输入中以引发不良行为或内容。为此，Li 等人<sup>[173]</sup>建立了一个基准来评估指令跟踪 AIGC 大模型针对即时注入攻击的鲁棒性。该方法的目标是确定 AIGC 大模型受注入指令影响的程度以及它们区分这些注入指令和原始目标指令的能力。通过对 AIGC 大模型进行广泛的实验，发现了它们对此类攻击的鲁棒性的重大漏洞。实验结果表明，某些模型过度调整以遵循提示中的任何嵌入指令，过度关注提示的后半部分，而没有完全掌握整个上下文。相比之下，更好地掌握上下文和指



令跟踪能力的模型可能更容易受到注入指令的影响。此外还进行了各个方面的深入分析,包括攻击和防御机制的影响、注入指令的类型以及它们在提示中的注入位置。通过这样的鲁棒性测评可以揭示这些漏洞,提供有价值的见解,指导未来工作中更强大的解决方案的开发。

PromptBench<sup>[56]</sup>是一个稳健性基准,旨在衡量 AIGC 大模型对对抗性提示的适应能力。这项研究使用了大量的对抗性文本攻击,针对多个级别的提示:字符、单词、句子和语义。对抗性提示旨在模仿拼写错误或同义词等看似合理的用户错误,旨在评估轻微偏差如何影响 AIGC 大模型结果,同时保持语义完整性。然后,这些提示可用于各种任务,包括情感分析、自然语言推理、阅读理解、机器翻译和数学问题解决。此外,通过 PromptBench 对对抗性提示的鲁棒性进行全面评估,并进行广泛的分析,包括对观察到的漏洞的可视化解释、对抗性提示的可迁移性分析、词频分析,为下游用户提供实用指导,并提示工程师制作更多强大的提示。该研究还发现当代 AIGC 大模型对于对抗性提示并不鲁棒,并使用注意力可视化进一步分析了其背后的原因。但是,现有方法面临着数据污染、对提示的敏感性以及基准创建成本高昂等挑战。为了解决这个问题,Li 等人<sup>[174]</sup>提出了一种基于无损数据压缩的评估方法,该方法测试模型的预测能力在训练截止后如何泛化。具体来说,训练和测试期间的性能差距作为鲁棒性的衡量标准。该方法通过跨时间的无损数据压缩来评估大型语言模型的泛化性和鲁棒性,避免了现有基于基准的评估中的数据污染和不同提示的潜在干扰。

最近 GPT-4 等 AIGC 大模型在文本生成和视觉输入方面取得了前所未有的性能。为了确保 AIGC 大模型符合人类价值观并生成安全文本,Qiu 等人<sup>[175]</sup>提出了一个评估 AIGC 大模型安全性和鲁棒性的基准,强调平衡方法的必要性。为了全面研究文本安全性和输出鲁棒性,该方法引入了一个潜在的越狱提示数据集,每个数据集都涉及恶意指令嵌入。此外,为了进一步分析安全性和鲁棒性,该方法设计了一个分层注释框架,对 AIGC 大模型的安全性和鲁棒性进行了系统分析,涉及显式正常指令的位置、单词替换和指令替换指示。而针对视觉大模型,Zhao 等人<sup>[176]</sup>根据经验评估了最先进的 AIGC 大模型的对抗鲁棒性,特别是针对那些接受

视觉输入的 AIGC 大模型(例如,基于图像的文本生成或联合生成)。该方法研究了最现实和高风险的场景,其中对手只有黑盒系统访问权限,并试图欺骗模型返回目标响应。研究结果表明,AIGC 大模型的鲁棒性高度依赖于其最脆弱的输入模式。

AIGC 大模型的主要应用之一将是这些行业的企业和用户在工业生产中的实际部署。然而,AIGC 大模型在工业场景中的准确性和鲁棒性尚未得到很好的研究。Li 等人<sup>[177]</sup>对中国工业生产领域的 AIGC 大模型的准确性和鲁棒性进行了全面的实证研究。该方法设计了一个变质测试框架,包含四个特定行业的稳定性类别和八种能力,总共 13,631 个问题及其变体,用于评估 AIGC 大模型的鲁棒性。Li 等人指出,不同行业的鲁棒性得分有所不同,本地 AIGC 大模型总体表现低于全球 AIGC 大模型。此外,AIGC 大模型的鲁棒性因能力而异。全球 AIGC 大模型在逻辑相关变体下更加鲁棒,而高级本地 AIGC 大模型在与理解中国工业术语相关的问题上表现更好。

鉴于 AIGC 大模型广泛融入日常生活,AIGC 大模型必须保持针对各种输入的鲁棒性,以便为最终用户提供最佳表现。鲁棒性测评是评估 AIGC 大模型在面对各种不确定性、噪声、对抗性攻击和领域转移等挑战时的性能和稳定性的过程。它通过模拟实际应用场景中的变化和干扰,评估模型对这些变化的适应能力和表现质量。例如,相同的提示但具有不同的语法和表达式可能会导致 ChatGPT 和其他 LLM 生成不同的结果,这表明当前的 LLM 对输入并不稳健。虽然之前有一些关于鲁棒性评估的工作,但还有很大的进步空间,例如包括更多样化的评估集、检查更多的评估方面以及开发更有效的评估来生成鲁棒性任务。同时,鲁棒性的概念和定义也在不断发展。总之,鲁棒性测评方法之间的区别在于评估角度、扰动来源和技术手段,但它们共同的目标是提高大 AIGC 大模型在真实世界环境中的可靠性和稳定性,并为改进模型设计和训练策略提供指导。

### 3.4 安全性和隐私性测评

在 ChatGPT 等生成式大模型备受追捧的同时,我们必须认识到其面临 AI 自身数据和模型方面的安全隐患<sup>[95]</sup>。尽管生成式大模型带来了各种革命性的技术进步,但其自身也带来的一系列安全与隐私

问题也值得我们注意，例如可能引发数据泄漏和助长虚假信息传播等<sup>[125]</sup>。因此，在 AIGC 大模型测评中，安全性和隐私性测评起着至关重要的作用。

安全性测评旨在通过评估模型的安全性能、鲁棒性和抗攻击能力来检测。这涉及对模型的代码、架构和环境进行设计，攻击和对抗性样本的测试，以验证模型在不同场景下的表现，可靠性和抵御恶意攻击的能力<sup>[45]</sup>。

隐私性测评主要关注模型对个人数据的隐私保护和合规性，包括对数据传输、存储和处理过程中的隐私保护措施的评估，检查是否存在数据泄露、滥用或未经授权的访问风险。隐私性测评的目标是确保模型在处理个人信息时符合相关法规和隐私标准，保护用户数据的隐私权益。

安全性和隐私性测评在 AIGC 大模型的评估中具有重要意义。它们保护用户数据和隐私，增强模型的可信度和透明度，遵守法规要求，保护商业利益和声誉。通过全面评估和提高模型的安全性和隐私性，能够构建更可靠的大模型，为用户和社会带来更大的价值和信任。

### 3.4.1 安全性测评

安全性测评是大对 AIGC 大模型在面对安全威胁和攻击时的评估和测试过程。随着 AIGC 大模型的广泛应用，对其安全性的关注也越来越重要。安全性测评旨在评估模型在面对各种潜在威胁和攻击时的强度和鲁棒性，以识别潜在的漏洞和弱点，并提供相应的保护和防御策略<sup>[178]</sup>。

数据中毒是 AIGC 大模型安全性领域中一个重要问题，它涉及到攻击者通过修改或注入恶意例子到模型的训练数据集中来进行攻击。这种攻击形式在如今变的尤为关注，因为开源和重用已成为开发文化的核心。攻击者可以利用这种方法，通过公开的方式收集并发布含有恶意代码的数据集，或者直接在开源平台上操纵数据。在开放式的协作环境中，如 GitHub，研究发现存在用于操纵或破坏开源仓库的虚假账户。攻击者可能会利用这些账户创建含有易受攻击代码的仓库，并通过各种手段提升这些仓库的受欢迎度，使得在收集训练 AIGC 大模型的数据时，这些恶意的仓库数据也被包括进去。因此，研究人员测评了 AIGC 大模型针对不同任务的数据中毒带来的风险，包括 API 推荐、代码搜索、代码表示和代码生成。

CVALUES<sup>[81]</sup>是具有对抗性和诱导性提示的人

类价值评估基准，用于衡量 AIGC 大模型在安全责任和标准方面的一致性能力。CVALUES 手动收集了 10 个场景的对抗性安全提示，并由专业专家诱导了 8 个领域的责任提示。此外，该方法不仅进行人工评估以进行可靠比较，还构建了多项选择提示以进行自动评估。研究结果表明，虽然大多数中国 AIGC 大模型在安全方面表现良好，但在责任方面还有很大的改进空间。GOAT-Bench<sup>[82]</sup>是综合性模因基准，其中包含超过 6K 种不同的模因，涵盖隐性仇恨言论、性别歧视和网络欺凌等主题。利用 GOAT-Bench，深入研究 AIGC 大模型准确评估仇恨、厌女症、攻击性、讽刺和有害内容。例如，在关注性别歧视（例如厌女症）的任务中，通过 GOAT-Bench 测评基准可以观察到 AIGC 大模型之间存在显著差异；在诸如危害性等更广泛的任务中，需要先进的背景知识和推理，AIGC 大模型的表现往往更加集中，而且通常比较温和。综上所述，通过对于 AIGC 大模型安全性的测评基准，可以进一步让研究人员能够多维地了解他们的模型在解决基于模因的社会虐待方面的能力，并为 AIGC 大模型安全见解的进步做出贡献。

然而，缺乏全面的安全评估基准对有效评估和提高大模型的安全性构成了重大障碍。SC-Safety<sup>[83]</sup>是针对中文 AIGC 大模型的多轮开放式问题对抗安全基准。SC-Safety 包含 4912 个开放式问题，涵盖 20 多个安全子维度，系统地评估 AIGC 大模型的安全性。通过 SC-Safety 对中文 AIGC 大模型安全性测评发现：1) 闭源模型在安全性方面优于开源模型；2) 中国发布的模型表现出与 GPT-3.5turbo 等 AIGC 大模型相当的安全水平；3) 一些具有 6B-13B 参数的较小型号可以在安全性方面有效竞争。SafetyBench<sup>[84]</sup>是一个评估 AIGC 大模型安全性的综合基准，其中包含 11,435 个不同的多项选择题，涵盖 7 个不同的安全问题类别。此外，SafetyBench 还纳入了中英文数据，方便双语评估。具体来说，SafetyBench 的试题类型丰富，涵盖对话场景、现实场景、安全对比、安全知识查询等。这种多样化的阵列确保 AIGC 大模型在各种与安全相关的环境和场景中经过严格的测试。与上述方法不同的是，SAFETEXT<sup>[85]</sup>则探索语言模型中的物理安全性的基准。SAFETEXT 是一个包含常识性的物理安全数据集，其中包含人工编写的现实生活场景以及每个场景的安全/不安全建议对。该方法使用数据集根据



经验量化 AIGC 大模型 中的常识性物理安全性。通过 SAFETEXT 测试结果表明, AIGC 大模型能够生成不安全的文本, 并且不能轻易拒绝不安全的建议。总体而言, 我们还需要创建更多的 AIGC 大模型安全性测评基本, 全面的测评 AIGC 大模型, 从而可以有效促进协作, 以创建更安全、更值得信赖的 AIGC 大模型。

安全性测评可以帮助识别和评估 AIGC 大模型所面临的安全威胁和风险, 从而提供及时的安全防护措施和应对策略。通过对 AIGC 大模型的安全性进行评估, 可以减少未经授权访问、数据泄露、攻击和滥用等安全风险的发生。然而, 安全性测评也存在一些缺点和挑战。由于大模型的复杂性和规模, 安全性测评可能面临时间和计算资源的限制。评估 AIGC 大模型的安全性可能需要大量的时间和计算资源, 增加评估的复杂性和成本。此外, 安全性测评方法和技术仍处于不断发展和完善的阶段。当前的安全性测评方法无法覆盖所有的安全威胁和攻击方式。因此, 需要进一步研究和改进安全性测评方法提高评估的准确性和全面性<sup>[179]</sup>。未来的研究方向应该包括将隐私保护技术、公平性评估考虑纳入安全性测评中, 以实现更全面的 AIGC 大模型测评。对于新兴的安全威胁和攻击方式, 需要持续跟踪和研究, 不断改进安全性测评方法, 以提高模型的安全性和可信度。

### 3.4.2 隐私性测评

在 AIGC 大模型的实际开发和应用中, 隐私泄露<sup>[180]</sup>是一个重要的问题, 尤其涉及大量敏感数据的处理。这些数据可能包含个人身份信息、健康记录、财务数据等敏感信息。AIGC 大模型的主要目的是生成和人类语言风格相近的语言, 训练数据在不经意间会被转换成包含敏感和隐私的个人信息, 如账号密码, 病例信息等生成内容, 从而导致隐私泄露的风险。同时大模型可以推断出潜在的敏感信息如用户的偏好、兴趣和行为等, 给用户带来诱导性的虚假信息, 来操控用户观点和行为。

隐私性测评是评估和分析 AIGC 大模型在处理敏感信息时的隐私保护程度和潜在风险的过程。通过隐私性测评, 可以及时识别和评估大模型面临的隐私泄露风险<sup>[181]</sup>, 从而提供有针对性的隐私保护措施和控制策略。隐私性测评可以帮助发现模型中存在的安全漏洞和弱点, 从而增强模型的安全性和鲁棒性。此外, 隐私性测评还有助于验证模型的合规

性, 确保模型在处理个人敏感信息时符合适用的隐私法规和规范要求<sup>[139]</sup>。

隐私概念包括 AIGC 大模型训练数据中包含的个人隐私或与隐私相关的个人的其他权利。AIGC 的大部分训练数据都是从偶尔隐藏软件机密的开源存储库中提取的。研究工作<sup>[182-183]</sup>已披露, 这些存储库可能包含大量敏感数据元素, 例如 API 密钥、密码和个人身份信息 (包括电子邮件地址等)。模型可能会无意中学习和复制这些敏感数据, 从而导致此类数据的无意泄露。另一种情况是“未经授权的训练”, 即模型是否在未经数据集所有者许可的情况下在数据集上进行训练。例如, 最近的新闻报道称, GitHub 因未经许可使用开源存储库中的代码而被起诉。Niu 等人<sup>[184]</sup>设计了一组可能从 GitHub Copilot 引发隐私信息的提示, 发现大约 8% 的提示会导致隐私泄露。因此, 随着 AIGC 大模型的迅速发展, 对其隐私性测评也变得尤为重要。

由于 AIGC 大模型将信息暴露存在额外风险, 这可能会引发隐私和安全问题。最近一些研究已经探索了用通用标记替换文本数据中的识别信息。Vats 等人<sup>[185]</sup>利用 AIGC 大模型来建议屏蔽标记的替代品, 并在下游语言建模任务上评估其有效性。提出了多种预训练和微调的基于 AIGC 大模型的方法, 并该方法在各种 NLP 数据集上进行实证实验, 以相应地比较适应的模型。实验结果表明, 在混淆语料库上训练的模型与在没有隐私保护令牌屏蔽的原始数据上训练的模型具有相当的性能。这种混淆技术有助于保护用户数据不被暴露给对手。通过使用 AIGC 大模型生成屏蔽令牌的替代品, 模型可以在不损害原始信息的隐私和安全的情况下对混淆数据进行训练。

随着 AIGC 大模型的迅速发展, AIGC 大模型也逐渐应用于医疗领域。由于隐私限制, 它们与医疗保健数据并不兼容。目前的 AIGC 大模型是专有模型, 要求用户将数据发送到专有源进行处理 (例如 OpenAI), 这就需要去识别患者数据, 但是删除所有患者健康信息是一项劳动密集型工作, 而且对于大型报告集来说也是不可行的。因此隐私性测评对于 AIGC 大模型来说也是至关重要的。研究人员<sup>[186]</sup>探索使用 ChatGPT 实施加密, 最终保护数据机密性。尽管缺乏广泛的编码技能或编程知识, 作者仍然能够通过 ChatGPT 成功实现加密算法。这凸显了个人利用 ChatGPT 执行加密任务的潜力。

为确保 AIGC 大模型的隐私性, 研究人员采取了多种措施处理训练数据, 减少隐私泄露的风险。例如, InCoder<sup>[187]</sup>的开发者们利用正则表达式识别出训练数据中所有的电子邮件地址, 并使用占位符来替换它们, 以避免在模型训练过程中泄露这些个人信息。Allal 等人<sup>[188]</sup>采取了使用自动化的检测工具, 这些工具激活了所有默认的插件, 以识别可能包含敏感信息的内容, 并采取正则表达式来特别检测出电子邮件地址、IPv4 地址和 IPv6 地址, 提高了隐私保护措施的综合性和有效性。Li 等人<sup>[189]</sup>训练一个秘密检测模型 “starpaii”, 它可以以更高的准确度和精度识别代码中的秘密, 并使用该工具删除敏感信息。

上述隐私性测评方法表明, 在大模型测评中, 隐私性测评具有重要的意义。隐私性测评通过评估模型对个人隐私的保护能力和隐私风险, 有助于确保个人数据的安全和隐私权益。然而, 现有隐私性测评也存在限制。首先, 隐私性测评涉及复杂的技术和专业知识, 对于一般开发者和组织来说难以理解 and 应用。此外, 隐私性测评方法的标准化和一致性仍然需要进一步研究和发展。缺乏一致性的测评指标和基准, 导致不同测评结果间存在差异<sup>[190]</sup>。未来的研究需要进一步推动隐私性测评方法的标准化和规范化, 以确保测评结果的一致性和可比性。随着隐私保护技术的不断发展, 研究人员可以探索新的隐私性测评方法和指标, 以适应不断演变的隐私保护需求。加强隐私法规和合规要求的研究, 将隐私性测评与法律法规的要求相结合, 确保隐私性测评的合规性和有效性。

## 4 AIGC 大模型测评展望

### 4.1 多语言测评

多语言测评是对涉及多种语言的模型或系统进行测评的过程。在文本生成任务中, 如自然语言处理、机器翻译、语音识别等领域, 由于不同语言之间存在差异, 比如语言特征和文化差异等方面, 构建适合的多语言测评的数据集变得至关重要, 因此, 大量研究人员正在努力收集和标注多语言数据集, 以便在不同语言环境下进行测评对比。例如, 在中文语境下, 构建涵盖多个任务领域的大规模数据集为中文模型测评提供了坚实的基础。目前, 针对教育和语言的大语言模型测评方面具有极大的潜力, 尤其是在中文的大语言模型测评方面仍有改进的空

间, 未来的工作可以集中于开发方法来提高模型在该方面的性能。

### 4.2 跨模态测评

AIGC 大模型的发展正在迅速从单纯的语言模型转向多模态大模型。尽管 GPT-4 作为一个多模态大模型取得了突破性进展, 但其受限于 GPU 资源。多模态大语言模型依赖于大语言模型丰富的知识储备以及强大的推理和泛化能力, 以此来解决多模态问题。目前多模态大模型已经涌现出一些令人惊叹的能力, 如看图写作和看图写代码。例如 Zhang 等人<sup>[191]</sup>提出了一个具有内在特征的大型语言模型 SpeechGPT, 具有跨模式会话能力, 能够感知和生成多个模型内容, 赋予大模型多模态会话能力。然而, 仅凭这些样例很难充分反映多模态大语言模型的性能, 当前仍然缺乏对多模态大语言模型的全面测评。清华大学 NLP 小组开源的 VisCPM 是基于 CPM-Bee-10B 进行多模态扩展得到的。VisCPM 系列包括多模态对话和文本生成图片。从发布的测评数据看, VisCPM 模型对图片理解较好。Fu 等人<sup>[63]</sup>提出的测评基准 MME 首次对现有的 10 种开源多模态大语言模型进行了全面定量测评并公布了 16 个排行榜, 包含感知和认知两个榜以及 14 个子榜。然而, 目前现有的多模态大语言模型定量测评方法主要以下三个缺点: 一是在传统的公开数据集上进行测评, 例如图像描述和视觉问答数据集, 但这些数据集难以准确反映模型能力, 并且无法保证这些数据集是否被其他模型用于训练; 二是在新数据集进行开放式测评, 但这种方法存在着数据集不公开和数量有限的问题; 三是仅聚焦于多模态大语言模型的某个特定方面进行测评, 如仅测评鲁棒性, 难以做到全面测评。四是评测大模型是否包含特定的知识, 训练数据不包含的知识能否通过一些方法推理得到。Wu 等人<sup>[192]</sup>对预训练语言模型是否能够在预训练过程中对本体知识进行有效编码以及是否能够深入理解语义内容进行了全面的探讨。Yin 等人<sup>[193]</sup>通过评估大模型识别无法回答的问题来调查模型的自我认知。实验结果表明, 虽然这些模型具有一定程度的自我认知, 但与人类的自我认知相比仍然存在明显的差距。因此, 需要进一步探索更全面、准确和公平的多模态大语言模型测评方法, 以确保对模型性能的全面理解和比较。这些测评将使 AIGC 大模型做出更准确、更可靠的反应, 将对他们在不同领域的应用产生积极影响。



### 4.3 使用规范的确定

测评 AIGC 大模型的性能可能受到各种因素的影响,这可能导致其评估结果的不稳定和不可靠。例如,恶意用户可能使用大模型编写诈骗短信、钓鱼邮件,甚至开发恶意软件和勒索软件等<sup>[194-195]</sup>。大模型本身也可能被恶意分子用来违法侵犯他人肖像权、隐私权、名誉权。例如,他人可能用 AIGC 技术生成虚假肖像来进行诈骗<sup>[196]</sup>。如果测评过于依赖固定和人工编写的标准,可能无法全面、公正地评估大语言模型的性能。一个模型在解决一个基准问题时表现出色,但在稍作修改(甚至只是修改提示)后可能得到完全相反的结果。测评通常依赖于人工编写的“ground truth”文本,但在需要专业知识的领域,这样的文本往往稀缺。随着模型在某些领域超越人类在基准测试中的表现,我们无法获得与“人类水平”性能的比较,并且难以直观地判断一个语言模型是否具备解决其他相关问题的能力。这导致了对 AIGC 大模型综合测评的困难,因为需要严格的基准来确定各种输入的弱点。测评过程容易出现脆弱问题,稍微修改基准提示或评估协议可能会导致完全不同的结果。因此,AIGC 大模型的测评需要根据使用者的国籍和现有法律来进行自适应挑战。最重要的是,需要一个完善的当地监管系统来检测其使用是否符合当地法律法规。

### 4.4 产权和责任的界定

AIGC 大模型由于其出色的语言处理能力,在各类语言工作、学习研究、新闻媒体编辑等场合被广泛应用。虽然大模型生成的内容符合知识产权的全部形式要求,但由于它们无法自主创作和拥有权利,仅能作为用户的辅助工具,因此无法成为著作权的主体,不能直接行使诸多权利。尽管如此,我们也应该探索新的机制来定性并保护 AIGC 生成产品,使其与传统意义上的著作权作品有所区别。这样的探索可以合理的给予 AIGC 生成产品一定的保护,进而对人工智能社区做出进一步完善。

AIGC 大模型以问答的形式存在,其回复存在不可信问题,或者无法判断其正确,存在似是而非的错误回复。使用者在使用大模型时应具备自己的判断和思考能力,避免盲目接受或传播错误的信息。因为大模型的回复可能是虚构的,模型无法提供合理的证据进行可信性的验证<sup>[197]</sup>。因此,在使用 AIGC 大模型生成的内容时,用户应该保持批判性思维,谨慎对待模型的回答,并在需要时进行有针对性的

测评。在模型等到有效测评的前提下,对于重要的信息和决策,最好依然依赖于可靠的来源、专业意见,以确保获得准确和可信的信息。

### 4.5 与人类对齐

尽管大模型拥有出强大的能力,但产生了看似合理,实际上并不正确的回答。大模型出现这种“幻觉”的原因之一在于未经人类对齐时发生的“过泛化”情况。比如,让大模型描述街景图片时,无论画面中是否有行人出现,模型都会因为自身过度的泛化问题,输入对行人的描述。这种现象在大模型中普遍存在。如何避免“幻觉”的出现?将人类纳入训练循环的人类反馈强化学习广泛应用。然而,人类反馈强化学习严重依赖专业的高质量人类反馈数据,在实践中难以正确实施。因此大模型对齐的发展仍处于初级阶段,还有很大的改进空间。比如细粒度指令数据对齐,通过提出不同的训练指令使不同的方法的评估具备公平性;人类-大模型联合评估,现有大模型评估框架要么利用大模型评估,要么利用人类评估。开发人类-大模型联合评估框架,根据各自优势为人类和大模型分配不同的评估任务,以保持大模型对齐评估过程中的效率和质量。

### 4.6 智能体评测

长期以来,自主代理一直被认为是实现通用人工智能的一种有前途的方法,智能体通过自我指导和规划来完成任务。这与人类的思维过程存在很大的不同,人类可以从更广泛的环境中学习,所以智能体通常无法复制人类水平的决策过程,尤其是在不受约束的开放域环境中。基于大模型的智能体关键在于模拟各种现实世界的人类行为。具体来说,理想的模拟应该准确地复制人类知识。在这种情况下,大模型可能会表现出压倒性的能力,接受海量网络知识库的培训,这些知识远远超出了普通人的知识范围。大模型的巨大能力可以显著影响模拟的有效性。例如,当尝试模拟用户对各种电影的选择行为时,确保大模型假设对这些电影没有先验知识是至关重要的。然而,大模型有可能已经获得了有关这些电影的信息。如果不实施适当的策略,大模型可能会根据其广泛的知识做出决策,即使现实世界的用户无法事先访问这些电影的内容。基于此,可以得到当前的重要问题是如何通过测评约束智能体模拟人类行为的重要能力。



## 5 结论

随着 ChatGPT 等大型模型驱动的 AIGC 服务进入人们的日常生活, 其给用户导致的各种问题逐渐凸显出来, 成为 AIGC 时代亟待解决的问题之一。因此, 对 AIGC 大模型进行充分多方面的测评变得尤为重要。系统化的测评可以确保模型的性能、准确性和可靠性, 发现潜在问题并推动模型改进和研究进展。这对于 AIGC 大模型的应用和社会的可持续发展具有重要意义。本文综述了 AIGC 大模型测评的指标和方法。首先, 介绍了模型测评的前期准备, 包括数据集的准备和模型的建立与选择, 并根据不同的学习方式和任务类型分别描述了模型测评指标。然后, 从多个角度描述了 AIGC 大模型在不同实际领域的新挑战和应对方法, 包括可解释性、公平性、鲁棒性、安全性和隐私性测评。最后, 在文章末尾讨论了 AIGC 大模型在当前所面临的挑战和趋势, 指出了未来可能的研究方向。我们希望这篇综述能够概述 AIGC 模型测评的问题, 并为学术界和工业界如何更好的进行 AIGC 大模型测评提供新的思路。通过持续地推动 AIGC 大模型的测评研究, 我们可以更好地应对其应用所带来的挑战, 并不断完善和发展这一领域。

**作者贡献声明:** 作者一、作者二和作者三完成了综述框架初步规划、相关文献收集整理、综述撰写等工作, 作者四优化了综述框架, 审阅和优化了综述内容, 作者五负责调研 AIGC 大模型的行业应用现状, 作者六和作者七提出具体修改意见并进行了投稿前修改。

## 参考文献:

- [1]. HO J, JAIN A, ABBEEL P. Denoising diffusion probabilistic models[J]. *Advances in Neural Information Processing Systems*, 2020, 33: 6840-6851.
- [2]. KARRAS T, LAINE S, AILA T. A style-based generator architecture for generative adversarial networks[C]//*Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2019: 4401-4410.
- [3]. ZHANG C, ZHANG C, ZHENG S, et al. A complete survey on generative AI (AIGC): Is ChatGPT from GPT-4 to GPT-5 all you need?[J]. *arXiv preprint arXiv:2303.11717*, 2023.
- [4]. OPENAI. ChatGPT: Optimizing language models for dialogue[EB/OL]. (2022-11-30) [2024-05-17]. <https://openai.com/blog/chatgpt>
- [5]. BANG Y, CAHYAWIJAYA S, LEE N, et al. A multitask, multilingual, multimodal evaluation of Huang F, Kwak H, An J. Is ChatGPT better than human annotators? Potential and limitations of ChatGPT in explaining implicit hate speech[J]. *arXiv preprint arXiv:2302.07736*, 2023.
- [6]. LI J, TANG T, ZHAO W X, et al. Pretrained language models for text generation: A survey[J]. *arXiv preprint arXiv:2201.05273*, 2022.
- [7]. FLORIDI L, CHIRIATTI M. GPT-3: Its nature, scope, limits, and consequences[J]. *Minds and Machines*, 2020, 30: 681-694.
- [8]. 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.
- [9]. QIN C, ZHANG A, ZHANG Z, et al. Is ChatGPT a general-purpose natural language processing task solver?[J]. *arXiv preprint arXiv:2302.06476*, 2023.
- [10]. RAO H, LEUNG C, MIAO C. Can ChatGPT assess human personalities? A general evaluation framework[J]. *arXiv preprint arXiv:2303.01248*, 2023.
- [11]. YANG X, LI Y, ZHANG X, et al. Exploring the limits of ChatGPT for query or aspect-based text summarization[J]. *arXiv pre-print arXiv:2302.08081*, 2023.
- [12]. ZUCCON G, KOOPMAN B. Dr ChatGPT, tell me what I want to hear: How prompt knowledge impacts health answer correctness[J]. *arXiv preprint arXiv:2302.13793*, 2023.
- [13]. DEVLIN J, CHANG M W, LEE K, et al. Bert: Pre-training of deep bidirectional transformers for language understanding[J]. *arXiv preprint arXiv:1810.04805*, 2018.
- [14]. CHOWDHERY A, NARANG S, DEVLIN J, et al. Palm: Scaling language modeling with pathways[J]. *arXiv preprint arXiv:2204.02311*, 2022.
- [15]. ZHANG S, ROLLER S, GOYAL N, et al. Opt: Open pre-trained transformer language models[J]. *arXiv pre print arXiv:2205.01068*, 2022.
- [16]. 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.
- [17]. META AI. Introducing LLaMA: A foundational, 65-billion-parameter large language model[EB/OL]. (2023-01-15) [2024-05-17]. <https://ai.facebook.com/blog/large-language-model-llama-meta-ai>
- [18]. META AI. LLaMA2[EB/OL]. (2023-07-19) [2024-05-17]. <https://ai.meta.com/research/publications/llama-2-open-foundation-and-fine-tuned-chat-models/>
- [19]. ROMBACH R, BLATTMANN A, LORENZ D, et al. High-resolution image synthesis with latent diffusion models[C]//*Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2022: 10684-10695.
- [20]. RUSKOV M. Grimm in Wonderland: Prompt Engineering with Midjourney to Illustrate Fairytales[J]. *arXiv preprint arXiv:2302.08961*, 2023.
- [21]. OPENAI R. Gpt-4 technical report. *arXiv 2303.08774*[J]. *View in Article*, 2023, 2: 13.
- [22]. DU Z, QIAN Y, LIU X, et al. Glm: General language model pretraining with autoregressive blank infilling[J]. *arXiv pre-print arXiv:2103.10360*, 2021.
- [23]. QINGHUA DA XUE. ChatGLM2[EB/OL]. (2023-03-14)

- [2024-05-17]. <https://hub.misakamoe.com/topics/chatglm2>
- [24]. QIAN C, HAN C, FUNG Y R, et al. CREATOR: Disentangling Abstract and Concrete Reasonings of Large Language Models through Tool Creation[J]. arXiv preprint arXiv:2305.14318, 2023.
- [25]. BAIDU. 文心一言[EB/OL]. (2023-03-16) [2024-05-17]. <https://yiyan.baidu.com/>
- [26]. ALIYUN. 通义大模型[EB/OL]. (2023-09-16) [2024-05-17]. <https://tongyi.aliyun.com/>
- [27]. FU DAN DA XUE. MOSS[EB/OL]. (2023-04-21) [2024-05-17]. <https://moss.fudan.edu.cn/>
- [28]. ZHONG GUO KE XUE YUAN. SHENG SI[EB/OL]. (2023-06-16) [2024-05-17]. <https://www.mindspore.cn/largeModel>
- [29]. HUA WEI YUN. PAN GU DA MO XING, [EB/OL]. (2023-07-16) [2024-05-17]. <https://www.huaweicloud.com/pan-gu-large-model>
- [30]. AZARIA A, MITCHELL T. The internal state of an llm knows when its lying[J]. arXiv preprint arXiv:2304.13734, 2023.
- [31]. CHIANG C H, LEE H. Can Large Language Models Be an Alternative to Human Evaluations?[J]. arXiv preprint arXiv:2305.01937, 2023.
- [32]. GAO M, RUAN J, SUN R, et al. Human-like summarization evaluation with ChatGPT[J]. arXiv preprint arXiv:2304.02554, 2023.
- [33]. LIN Y T, CHEN Y N. LLM-Eval: Unified Multi-Dimensional Automatic Evaluation for Open-Domain Conversations with Large Language Models[J]. arXiv preprint arXiv:2305.13711, 2023.
- [34]. LIU J, XIA C S, WANG Y, et al. Is your code generated by ChatGPT really correct? rigorous evaluation of large language models for code generation[J]. arXiv preprint arXiv:2305.01210, 2023.
- [35]. LIU Y, ITER D, XU Y, et al. Gpteval: Nlg evaluation using GPT-4 with better human alignment[J]. arXiv preprint arXiv:2303.16634, 2023.
- [36]. WANG J, LIANG Y, MENG F, et al. Is ChatGPT a good NLG evaluator? a preliminary study[J]. arXiv preprint arXiv:2303.04048, 2023.
- [37]. HENDRYCKS D, BURNS C, BASART S, et al. Measuring Massive Multitask Language Understanding[C]//International Conference on Learning Representations. 2020.
- [38]. ZHANG X, LI C, ZONG Y, et al. Evaluating the Performance of Large Language Models on GAOKAO Benchmark[J]. arXiv preprint arXiv:2305.12474, 2023.
- [39]. 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.
- [40]. ZHONG W, CUI R, GUO Y, et al. Agieval: A human-centric benchmark for evaluating foundation models[J]. arXiv preprint arXiv:2304.06364, 2023.
- [41]. LI H, ZHANG Y, KOTO F, et al. CMMLU: Measuring massive multitask language understanding in Chinese[J]. arXiv preprint arXiv:2306.09212, 2023.
- [42]. ZHANG W, ALJUNIED S M, GAO C, et al. M3Exam: A Multilingual, Multimodal, Multilevel Benchmark for Examining Large Language Models[J]. arXiv preprint arXiv:2306.05179, 2023.
- [43]. ZHOU J, GANDOMI A H, CHEN F, et al. EvSrivastava A, Rastogi A, Rao A, et al. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models[J]. arXiv preprint arXiv:2206.04615, 2022.
- [44]. LIANG P, BOMMASANI R, LEE T, et al. Holistic evaluation of language models[J]. arXiv preprint arXiv:2211.09110, 2022.
- [45]. CHANG Y, WANG X, WANG J, et al. A Survey on Evaluation of Large Language Models[J]. arXiv preprint arXiv:2307.03109, 2023.
- [46]. CARVALHO D V, PEREIRA E M, CARDOSO J S. Machine learning interpretability: A survey on methods and metrics[J]. Electronics, 2019, 8(8): 832.
- [47]. RÜPING S. Learning interpretable models (Ph. D. thesis)[J]. University Dortmund.[Google Scholar], 2006.
- [48]. ZHOU J, KHAWAJA M A, LI Z, et al. Making machine learning useable by revealing internal states update-a transparent approach[J]. International Journal of Computational Science and Engineering, 2016, 13(4): 378-389.
- [49]. CARLINI N, LIU C, ERLINGSSON Ú, et al. The secret sharer: Evaluating and testing unintended memorization in neural networks[C]//28th USENIX Security Symposium (USENIX Security 19). 2019: 267-284.
- [50]. FREDRIKSON M, JHA S, RISTENPART T. Model inversion attacks that exploit confidence information and basic countermeasures[C]//Proceedings of the 22nd ACM SIGSAC conference on computer and communications security. 2015: 1322-1333.
- [51]. GANJU K, WANG Q, YANG W, et al. Property inference attacks on fully connected neural networks using permutation invariant representations[C]//Proceedings of the 2018 ACM SIGSAC conference on computer and communications security. 2018: 619-633.
- [52]. SALEM A, BHATTACHARYA A, BACKES M, et al. {Updates-Leak}: Data set inference and reconstruction attacks in online learning[C]//29th USENIX security symposium (USENIX Security 20). 2020: 1291-1308.
- [53]. SHOKRI R, STRONATI M, SONG C, et al. Membership inference attacks against machine learning models[C]//2017 IEEE symposium on security and privacy (SP). IEEE, 2017: 3-18.
- [54]. SONG C, RISTENPART T, SHMATIKOV V. Machine learning models that remember too much[C]//Proceedings of the 2017 ACM SIGSAC Conference on computer and communications security. 2017: 587-601.
- [55]. WANG J, HU X, HOU W, et al. On the robustness of ChatGPT: An adversarial and out-of-distribution perspective[J]. arXiv preprint arXiv:2302.12095, 2023.
- [56]. ZHU K, WANG J, ZHOU J, et al. PromptBench: Towards Evaluating the Robustness of Large Language Models on Adversarial Prompts[J]. arXiv preprint arXiv:2306.04528, 2023.
- [57]. WILLIG M, ZECEVIC M, DHAMI D S, et al. Causal parrots: Large language models may talk causality but are not

- causal[J]. preprint, 2023, 8.
- [58]. ZHOU K, ZHU Y, CHEN Z, et al. Don't Make Your LLM an Evaluation Benchmark Cheater[J]. arXiv preprint arXiv:2311.01964, 2023.
- [59]. ZHU K, CHEN J, WANG J, et al. Dyval: Graph-informed dynamic evaluation of large language models[J]. arXiv preprint arXiv:2309.17167, 2023.
- [60]. ZHU K, ZHAO Q, CHEN H, et al. Promptbench: A unified library for evaluation of large language models[J]. arXiv preprint arXiv:2312.07910, 2023.
- [61]. LMSYS. CHATBOT ARENA: Benchmarking LLMs in the wild with ELO ratings, [EB/OL]. (2023-05-01) [2024-05-17]. <https://lmsys.org>
- [62]. 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.
- [63]. FU C, CHEN P, SHEN Y, et al. MME: A Comprehensive Evaluation Benchmark for Multimodal Large Language Models[J]. arXiv preprint arXiv:2306.13394, 2023.
- [64]. AN C, GONG S, ZHONG M, et al. L-Eval: Instituting Standardized Evaluation for Long Context Language Models[J]. arXiv preprint arXiv:2307.11088, 2023.
- [65]. YU J, WANG X, TU S, et al. KoLA: Carefully Benchmarking World Knowledge of Large Language Models[J]. arXiv preprint arXiv:2306.09296, 2023.
- [66]. KIELA D, BARTOLO M, NIE Y, et al. Dynabench: Rethinking benchmarking in NLP[J]. arXiv preprint arXiv:2104.14337, 2021.
- [67]. ZHOU Y, MURESANU A I, HAN Z, et al. Large language models are human-level prompt engineers[J]. arXiv preprint arXiv:2211.01910, 2022.
- [68]. DUBOIS Y, GALAMBOSI B, LIANG P, et al. Length-controlled alpacaEval: A simple way to debias automatic evaluators[J]. arXiv preprint arXiv:2404.04475, 2024.
- [69]. 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.
- [70]. CHOI M, PEI J, KUMAR S, et al. Do LLMs Understand Social Knowledge? Evaluating the Sociability of Large Language Models with SockET Benchmark[J]. arXiv preprint arXiv:2305.14938, 2023.
- [71]. HENDRYCKS D, BURNS C, KADAVATH S, et al. Measuring mathematical problem solving with the math dataset[J]. arXiv preprint arXiv:2103.03874, 2021.
- [72]. HENDRYCKS D, BASART S, KADAVATH S, et al. Measuring coding challenge competence with apps[J]. arXiv preprint arXiv:2105.09938, 2021.
- [73]. HUGGINGFACE. Open-source large language models leaderboard[EB/OL]. (2023-01-01) [2024-05-17]. <https://huggingface.co/spaces/HuggingFaceH4/open-llm-leaderboard>
- [74]. 超对称(北京)科技有限公司. BBT CFLEB, [EB/OL]. (2023-07-28) [2024-05-17]. <https://bbt.ssymmetry.com/evaluation.html>
- [75]. ZHANG L, CAI W, LIU Z, et al. Fineval: A Chinese financial domain knowledge evaluation benchmark for large language models[J]. arXiv preprint arXiv:2308.09975, 2023.
- [76]. SINGHAL K, AZIZI S, TU T, et al. Large language models encode clinical knowledge[J]. Nature, 2023: 1-9.
- [77]. KE P, WEN B, FENG Z, et al. Critiquellm: Scaling llm-as-critic for effective and explainable evaluation of large language model generation[J]. arXiv preprint arXiv:2311.18702, 2023.
- [78]. YANG K, ZHANG T, KUANG Z, et al. Mentalllama: Interpretable mental health analysis on social media with large language models[J]. arXiv preprint arXiv:2309.13567, 2023.
- [79]. WANG B, XU C, WANG S, et al. Adversarial glue: A multi-task benchmark for robustness evaluation of language models[J]. arXiv preprint arXiv:2111.02840, 2021.
- [80]. MEI A, LEVY S, WANG W Y. ASSERT: Automated Safety Scenario Red Teaming for Evaluating the Robustness of Large Language Models[J]. arXiv preprint arXiv:2310.09624, 2023.
- [81]. XU G, LIU J, YAN M, et al. Cvalues: Measuring the values of Chinese large language models from safety to responsibility[J]. arXiv preprint arXiv:2307.09705, 2023.
- [82]. LIN H, LUO Z, WANG B, et al. GOAT-Bench: Safety Insights to Large Multimodal Models through Meme- Based Social Abuse[J]. arXiv preprint arXiv:2401.01523, 2024.
- [83]. XU L, ZHAO K, ZHU L, et al. Sc-safety: A multi-round open-ended question adversarial safety benchmark for large language models in Chinese[J]. arXiv preprint arXiv:2310.05818, 2023.
- [84]. ZHANG Z, LEI L, WU L, et al. Safetybench: Evaluating the safety of large language models with multiple choice questions[J]. arXiv preprint arXiv:2309.07045, 2023.
- [85]. LEVY S, ALLAWAY E, SUBBIAH M, et al. SafeText: A benchmark for exploring physical safety in language models[J]. arXiv preprint arXiv:2210.10045, 2022.
- [86]. CHEN F, HAN M, ZHAO H, et al. X-LLM: Bootstrapping advanced large language models by treating multi-modalities as foreign languages[J]. arXiv preprint arXiv:2305.04160, 2023.
- [87]. GAO P, HAN J, ZHANG R, et al. LLAMA-adapter v2: Parameter-efficient visual instruction model[J]. arXiv preprint arXiv:2304.15010, 2023.
- [88]. XU Z, SHEN Y, HUANG L. Multiinstruct: Improving multi-modal zero-shot learning via instruction tuning[J]. arXiv preprint arXiv:2212.10773, 2022.
- [89]. ZHANG R, HAN J, ZHOU A, et al. LLAMA-adapter: Efficient fine-tuning of language models with zero-init attention[J]. arXiv preprint arXiv:2303.16199, 2023.
- [90]. ZHAO Z, GUO L, YUE T, et al. Chatbridge: Bridging modalities with large language model as a language catalyst[J]. arXiv preprint arXiv:2305.16103, 2023.
- [91]. GONG T, LYU C, ZHANG S, et al. Multimodal-GPT: A vision and language model for dialogue with humans[J]. arXiv preprint arXiv:2305.04790, 2023.
- [92]. LI K C, HE Y, WANG Y, et al. Videochat: Chat-centric video understanding[J]. arXiv preprint arXiv:2305.06355, 2023.
- [93]. LI L, YIN Y, LI S, et al. M<sup>3</sup> IT: A Large-Scale Dataset towards Multi-Modal Multilingual Instruction Tuning[J].



- arXiv preprint arXiv:2306.04387, 2023.
- [94]. QIN Y, CAI Z, JIN D, et al. WebCPM: Interactive Web Search for Chinese Long-form Question Answering[J]. arXiv preprint arXiv:2305.06849, 2023.
- [95]. DING N, CHEN Y, XU B, et al. Enhancing Chat Language Models by Scaling High-quality Instructional Conversations[J]. arXiv preprint arXiv:2305.14233, 2023.
- [96]. SHI K, WANG X, YU J, et al. CStory: A Chinese Large-scale News Storyline Dataset[C]//Proceedings of the 31st ACM International Conference on Information & Knowledge Management. 2022: 4475-4479.
- [97]. DU L, DING X, XIONG K, et al. e-CARE: a new dataset for exploring explainable causal reasoning[J]. arXiv preprint arXiv:2205.05849, 2022.
- [98]. ALLEN INSTITUTE FOR AI. Dolma[EB/OL]. (2023- 5-07) [2024-05-17]. <https://blog.allenai.org/dolma-3-trillion-tokens-open-llm-corpus-9a0ff4b8da64>
- [99]. LIU H, LI C, WU Q, et al. Visual instruction tuning[J]. Advances in neural information processing systems, 2024, 36.
- [100]. ZHU D, CHEN J, SHEN X, et al. Minigt-4: Enhancing vision-language understanding with advanced large language models[J]. arXiv preprint arXiv:2304.10592, 2023.
- [101]. YANG R, SONG L, LI Y, et al. GPT4Tools: Teaching large language model to use tools via self-instruction[J]. arXiv preprint arXiv:2305.18752, 2023.
- [102]. PI R, GAO J, DIAO S, et al. DetGPT: Detect What You Need via Reasoning[J]. arXiv preprint arXiv:2305.14167, 2023.
- [103]. LIU P, YUAN W, FU J, et al. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing[J]. ACM Computing Surveys, 2023, 55(9): 1-35.
- [104]. LI X, QIU X. MoT: Pre-thinking and Recalling Enable ChatGPT to Self-Improve with Memory-of-Thoughts[J]. arXiv preprint arXiv:2305.05181, 2023.
- [105]. DOSHI-VELEZ F, KIM B. Towards a rigorous science of interpretable machine learning[J]. arXiv preprint arXiv:1702.08608, 2017.
- [106]. BAYAT V, PHELPS S, RYONO R, et al. A severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) prediction model from standard laboratory tests[J]. Clinical Infectious Diseases, 2021, 73(9): e2901-e2907.
- [107]. GUIDOTTI R, MONREALE A, RUGGIERI S, et al. A survey of methods for explaining black box models[J]. ACM Computing Surveys (CSUR), 2018, 51(5): 1-42.
- [108]. EXPLAINABLE AI: Interpreting, explaining and visualizing deep learning[M]. Springer Nature, 2019.
- [109]. Ramesh, Aditya, et al. Hierarchical text-conditional image generation with CLIP latents[EB/OL]. (2022-04-13) [2024-05-17]. <https://arxiv.org/abs/2204.06125>
- [110]. REI R, STEWART C, FARINHA A C, et al. COMET: A neural framework for MT evaluation[J]. arXiv preprint arXiv:2009.09025, 2020.
- [111]. PIRES T, SCHLINGER E, GARRETTE D. How multilingual is multilingual BERT?[J]. arXiv preprint arXiv:1906.01502, 2019.
- [112]. SELLAM T, DAS D, PARIKH A P. BLEURT: Learning robust metrics for text generation[J]. arXiv preprint arXiv:2004.04696, 2020.
- [113]. WU S, IRSOY O, LU S, et al. BloombergGPT: A large language model for finance[J]. arXiv preprint arXiv:2303.17564, 2023.
- [114]. ZHANG X, YANG Q, XU D. XuanYuan 2.0: A Large Chinese Financial Chat Model with Hundreds of Billions Parameters[J]. arXiv preprint arXiv:2305.12002, 2023.
- [115]. YANGMU YU, WENHUAN HONG, et al. Cornucopia[EB/OL]. (2023-03-21) [2024-05-17]. <https://github.com/jerry1993-tech/Cornucopia-LLaMA-Fin-Chinese>
- [116]. LU D, WU H, LIANG J, et al. BBT-Fin: Comprehensive construction of Chinese financial domain pre-trained language model, corpus and benchmark[J]. arXiv preprint arXiv:2302.09432, 2023.
- [117]. LEXILAW. LexiLaw[EB/OL]. (2023-07-18) [2024-05-17]. <https://github.com/CSHaitao/LexiLaw>
- [118]. NGUYEN H T. A brief report on LawGPT 1.0: A virtual legal assistant based on GPT-3[J]. arXiv preprint arXiv:2302.05729, 2023.
- [119]. HUANG Q, TAO M, AN Z, et al. Lawyer LLaMA Technical Report[J]. arXiv preprint arXiv:2305.15062, 2023.
- [120]. YAO F, XIAO C, WANG X, et al. Leven: A large-scale Chinese legal event detection dataset[J]. arXiv preprint arXiv:2203.08556, 2022.
- [121]. CASCELLA M, MONTOMOLI J, BELLINI V, et al. Evaluating the feasibility of ChatGPT in healthcare: an analysis of multiple clinical and research scenarios[J]. Journal of Medical Systems, 2023, 47(1): 33.
- [122]. CHERVENAK J, LIEMAN H, BLANCO-BREINDEL M, et al. The promise and peril of using a large language model to obtain clinical information: ChatGPT performs strongly as a fertility counseling tool with limitations[J]. Fertility and Sterility, 2023.
- [123]. DUONG D, SOLOMON B D. Analysis of large- language model versus human performance for genetics questions[J]. European Journal of Human Genetics, 2023: 1-3.
- [124]. GILSON A, SAFRANEK C W, HUANG T, et al. How does ChatGPT perform on the United States medical licensing examination? The implications of large language models for medical education and knowledge assessment[J]. JMIR Medical Education, 2023, 9(1): e45312.
- [125]. XIONG H, WANG S, ZHU Y, et al. Doctorglm: Fine-tuning your chinese doctor is not a herculean task[J]. arXiv preprint arXiv:2304.01097, 2023.
- [126]. LI S T. Ben-Tsao Gong-Mu (Chinese Botanical Encyclopedia)[J]. Taipei, Taiwan: Great Taipei Publishing, 1990.
- [127]. LIANG Y, HUANG Y. Bian Que, the founder of diagnostics of traditional Chinese medicine[J]. Journal of Traditional Chinese Medical Sciences, 2022, 9(2): 93-94.
- [128]. ZHANG H, CHEN J, JIANG F, et al. HuatuoGPT, towards Taming Language Model to Be a Doctor[J]. arXiv preprint arXiv:2305.15075, 2023.
- [129]. SINGHAL K, TU T, GOTTWEIS J, et al. Towards expert-level medical question answering with large language models[J]. arXiv preprint arXiv:2305.09617, 2023.
- [130]. HALL P, GILL N, SCHMIDT N. Proposed guidelines for

- the responsible use of explainable machine learning[J]. arXiv preprint arXiv:1906.03533, 2019.
- [131]. 陈珂锐, 孟小峰. 机器学习的可解释性[J]. 计算机研究与发展, 2020, 57(9): 1971-1986.
- CHEN K Y, MENG X F. Interpretability of machine learning[J]. Journal of Computer Research and Development, 2020, 57(9): 1971-1986.
- [132]. 梁峥,王宏志,戴加佳等.预训练语言模型实体匹配的可解释性[J].软件学报,2023,34(03):1087-1108.
- LIANG Z, WANG H Z, DAI J J, et al. Interpretability of Entity Matching Based on Pre-trained Language Model [J]. Journal of Software, 2023, 34(03): 1087-1108.
- [133]. 王冬丽,杨珊,欧阳万里等.人工智能可解释性:发展与应用[J].计算机科学,2023,50(S1):19-25.
- WANG D L, YANG S, OUYANG W L, et al. Explainability of Artificial Intelligence: Development and Application [J]. Computer Science, 2023, 50(S1): 19-25.
- [134]. MARKUS A F, KORS J A, RIJNBEEK P R. The role of explainability in creating trustworthy artificial intelligence for health care: a comprehensive survey of the terminology, design choices, and evaluation strategies[J]. Journal of Biomedical Informatics, 2021, 113: 103655.
- [135]. 纪守领, 李进锋, 杜天宇, 李博. 机器学习模型可解释性方法、应用与安全研究综述[J]. 计算机研究与发展, 2019, 56(10): 2071-2096.
- JI S L, LI J F, DU T Y, LI B. Survey on Techniques, Applications and Security of Machine Learning Interpretability [J]. Journal of Computer Research and Development, 2019, 56(10): 2071-2096.
- [136]. 成科扬, 王宁, 师文喜, 詹永照. 深度学习可解释性研究进展[J]. 计算机研究与发展, 2020, 57(6): 1208-1217.
- CHENG K Y, WANG N, SHI W X, ZHAN Y Z. Research Advances in the Interpretability of Deep Learning[J]. Journal of Computer Research and Development, 2020, 57(6): 1208-1217.
- [137]. DOSHI-VELEZ F, KIM B. Considerations for evaluation and generalization in interpretable machine learning[J]. Explainable and interpretable models in computer vision and machine learning, 2018: 3-17.
- [138]. LIPTON Z C. In machine learning, the concept of interpretability is both important and slippery[J]. Queue, 2018, 16: 28.
- [139]. SHEN Y, WANG L, CHEN Y, et al. An Interpretability Evaluation Benchmark for Pre-trained Language Models[J]. arXiv preprint arXiv:2207.13948, 2022.
- [140]. ROSS A, CHEN N, HANG E Z, et al. Evaluating the interpretability of generative models by interactive reconstruction[C]//Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems. 2021: 1-15.
- [141]. WANG Q, ANIKINA T, FELDHUS N, et al. LLMCheckup: Conversational Examination of Large Language Models via Interpretability Tools[J]. arXiv preprint arXiv:2401.12576, 2024.
- [142]. LEI Y, LIAN J, YAO J, et al. RecExplainer: Aligning Large Language Models for Recommendation Model Interpretability[J]. arXiv preprint arXiv:2311.10947, 2023.
- [143]. YANG K, JI S, ZHANG T, et al. Towards interpretable mental health analysis with large language models[C]//Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. 2023: 6056-6077.
- [144]. MA W, ZHAO M, XIE X, et al. Are Code Pre-trained Models Powerful to Learn Code Syntax and Semantics?[J]. arXiv preprint arXiv:2212.10017, 2022.
- [145]. LI Y, ZHANG T, LUO X, et al. Do pre-trained language models indeed understand software engineering tasks?[J]. IEEE Transactions on Software Engineering, 2023.
- [146]. HOODA A, CHRISTODORESCU M, ALLAMANIS M, et al. Do Large Code Models Understand Programming Concepts? A Black-box Approach[J]. arXiv preprint arXiv:2402.05980, 2024.
- [147]. RODRIGUEZ-CARDENAS D, PALACIO D N, KHATI D, et al. Benchmarking Causal Study to Interpret Large Language Models for Source Code[C]//2023 IEEE International Conference on Software Maintenance and Evolution (ICSME). IEEE, 2023: 329-334.
- [148]. ROY S, LABERGE G, ROY B, et al. Why don't XAI techniques agree? characterizing the disagreements between post-hoc explanations of defect predictions[C]//2022 IEEE International Conference on Software Maintenance and Evolution (ICSME). IEEE, 2022: 444-448.
- [149]. JI Z, MA P, LI Z, et al. Benchmarking and Explaining Large Language Model-based Code Generation: A Causality-Centric Approach[J]. arXiv e-prints, 2023: arXiv:2310.06680.
- [150]. PALACIO D N, VELASCO A, RODRIGUEZ-CARDENAS D, et al. Evaluating and explaining large language models for code using syntactic structures[J]. arXiv preprint arXiv:2308.03873, 2023.
- [151]. ZHANG T, CHEN Z, ZHU Y, et al. Interpretable program synthesis[C]//Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems. 2021: 1-16.
- [152]. 杨朋波, 桑基韬, 张彪等. 面向图像分类的深度模型可解释性研究综述[J]. 软件学报, 2023, 34(01): 230-254.
- YANG P B, SANG J T, ZHANG B, et al. Survey on Interpretability of Deep Models for Image Classification [J]. Journal of Software, 2023, 34(01): 230-254.
- [153]. CHEN H, JI Y. Adversarial training for improving model robustness? Look at both prediction and interpretation[C]//Proceedings of the AAAI Conference on Artificial Intelligence. 2022, 36(10): 10463-10472.
- [154]. 王昱颖, 张敏, 杨晶然, 等. 深度学习模型中的公平性研究[J]. 软件学报, 2023, 34(09): 4037-4055.
- WANG Y Y, ZHANG M, Y J R, et al. Research on fairness in deep learning models[J]. Journal of Software, 2023, 34(09): 4037-4055.
- [155]. 刘文炎, 沈楚云, 王祥丰, 等. 可信机器学习的公平性综述[J]. 软件学报, 2021, 32(05): 1404-1426.

- LIU Wenyan, SHEN Chuyun, WANG Xiangfeng, et al. Survey on Fairness in Trustworthy Machine Learning [J]. *Journal of Software*, 2021, 32(05): 1404-1426.
- [156]. ZHOU Y, MURESANU A I, HAN Z, et al. Large language models are human-level prompt engineers[J]. *arXiv preprint arXiv:2211.01910*, 2022.
- [157]. SAH C K, XIAOLI D L, ISLAM M M. Unveiling Bias in Fairness Evaluations of Large Language Models: A Critical Literature Review of Music and Movie Recommendation Systems[J]. *arXiv preprint arXiv:2401.04057*, 2024.
- [158]. BI G, SHEN L, XIE Y, et al. A Group Fairness Lens for Large Language Models[J]. *arXiv preprint arXiv:2312.15478*, 2023.
- [159]. FREIBERGER V, BUCHMANN E. Fairness Certification for Natural Language Processing and Large Language Models[J]. *arXiv preprint arXiv:2401.01262*, 2024.
- [160]. HUANG P S, ZHANG H, JIANG R, et al. Reducing sentiment bias in language models via counterfactual evaluation[J]. *arXiv preprint arXiv:1911.03064*, 2019.
- [161]. ZHUO T Y, HUANG Y, CHEN C, et al. Exploring ai ethics of chatgpt: A diagnostic analysis[J]. *arXiv preprint arXiv:2301.12867*, 2023.
- [162]. FERRARA E. Should chatgpt be biased? challenges and risks of bias in large language models[J]. *arXiv preprint arXiv:2304.03738*, 2023.
- [163]. HARTMANN J, SCHWENZOW J, WITTE M. The political ideology of conversational AI: Converging evidence on ChatGPT's pro-environmental, left-libertarian orientation[J]. *arXiv preprint arXiv:2301.01768*, 2023.
- [164]. LI Y, ZHANG Y. Fairness of chatgpt[J]. *arXiv preprint arXiv:2305.18569*, 2023.
- [165]. PARRISH A, CHEN A, NANGIA N, et al. BBQ: A hand-built bias benchmark for question answering[C]// *Findings of the Association for Computational Linguistics: ACL 2022*. 2022: 2086-2105.
- [166]. KHASHABI D, MIN S, KHOT T, et al. Unifiedqa: Crossing format boundaries with a single qa system[J]. *arXiv preprint arXiv:2005.00700*, 2020.
- [167]. RUTINOWSKI J, FRANKE S, ENDENDYK J, et al. The self-perception and political biases of chatgpt[J]. *Human Behavior and Emerging Technologies*, 2023, 2024.
- [168]. FERREIRA S L C, CAIRES A O, BORGES T S, et al. Robustness evaluation in analytical methods optimized using experimental designs[J]. *Microchemical Journal*, 2017, 131: 163-169.
- [169]. BRENDL W, RAUBER J, KÜMMERER M, et al. Accurate, reliable and fast robustness evaluation[J]. *Advances in neural information processing systems*, 2019, 32.
- [170]. KASHYAP A R, MEHNAZ L, MALIK B, et al. Analyzing the domain robustness of pretrained language models, layer by layer[C]// *Proceedings of the Second Workshop on Domain Adaptation for NLP*. 2021: 222-244.
- [171]. LIU Q, JI S, LIU C, et al. A practical black-box attack on source code authorship identification classifiers[J]. *IEEE Transactions on Information Forensics and Security*, 2021, 16: 3620-3633.
- [172]. ZHANG C, WANG Z, MANGAL R, et al. Transfer Attacks and Defenses for Large Language Models on Coding Tasks[J]. *arXiv preprint arXiv:2311.13445*, 2023.
- [173]. LI Z, PENG B, HE P, et al. Evaluating the instruction-following robustness of large language models to prompt injection[J]. *arXiv preprint arXiv:2308.10819*, 2023.
- [174]. LI Y, GUO Y, GUERIN F, et al. Evaluating Large Language Models for Generalization and Robustness via Data Compression[J]. *arXiv preprint arXiv:2402.00861*, 2024.
- [175]. QIU H, ZHANG S, LI A, et al. Latent jailbreak: A benchmark for evaluating text safety and output robustness of large language models[J]. *arXiv preprint arXiv:2307.08487*, 2023.
- [176]. ZHAO Y, PANG T, DU C, et al. On evaluating adversarial robustness of large vision-language models[J]. *arXiv preprint arXiv:2305.16934*, 2023.
- [177]. LI Z, QIU W, MA P, et al. An Empirical Study on Large Language Models in Accuracy and Robustness under Chinese Industrial Scenarios[J]. *arXiv preprint arXiv:2402.01723*, 2024.
- [178]. NEUMANN P G. Computer system-Security evaluation[C]// *Managing requirements knowledge, international workshop on*. IEEE Computer Society, 1978: 1087-1087.
- [179]. TOUBIANA V, NARAYANAN A, BONEH D, et al. Ad-nostic: Privacy preserving targeted advertising[C]// *Proceedings Network and Distributed System Symposium*. 2010.
- [180]. Plant R, Giuffrida V, Gkatzia D. You are what you write: Preserving privacy in the era of large language models[J]. *arXiv preprint arXiv:2204.09391*, 2022.
- [181]. HARDT M, TALWAR K. On the geometry of differential privacy[C]// *Proceedings of the forty-second ACM symposium on Theory of computing*. 2010: 705-714.
- [182]. ZHANG C, WANG Z, MANGAL R, et al. Transfer Attacks and Defenses for Large Language Models on Coding Tasks[J]. *arXiv preprint arXiv:2311.13445*, 2023.
- [183]. FENG R, YAN Z, PENG S, et al. Automated detection of password leakage from public github repositories[C]// *Proceedings of the 44th International Conference on Software Engineering*. 2022: 175-186.
- [184]. JUNGWIRTH G, SAHA A, SCHRÖDER M, et al. Connecting the dotfiles: Checked-In Secret Exposure with Extra (Lateral Movement) Steps[C]// *2023 IEEE/ACM 20th International Conference on Mining Software Repositories (MSR)*. IEEE, 2023: 322-333.
- [185]. VATS A, LIU Z, SU P, et al. Recovering from privacy-preserving masking with large language models[J]. *arXiv preprint arXiv:2309.08628*, 2023.
- [186]. ABBASIAN M, AZIMI I, RAHMANI A M, et al. Conversational health agents: A personalized llm-powered agent framework[J]. *arXiv preprint arXiv:2310.02374*, 2023.
- [187]. FRIED D, AGHAJANYAN A, LIN J, et al. Incoder: A generative model for code infilling and synthesis[J]. *arXiv preprint arXiv:2204.05999*, 2022.
- [188]. ALLAL L B, LI R, KOCETKOV D, et al. SantaCoder: don't reach for the stars![J]. *arXiv preprint arXiv:2301.*



- 03988, 2023.
- [189]. LI R, ALLAL L B, ZI Y, et al. Starcoder: may the source be with you![J]. arXiv preprint arXiv:2305.06161, 2023.
- [190]. LYU C, XU J, WANG L. New trends in machine translation using large language models: Case examples with chatgpt[J]. arXiv preprint arXiv:2305.01181, 2023.
- [191]. ZHANG D, LI S, ZHANG X, et al. Speechgpt: Empowering large language models with intrinsic cross-modal conversational abilities[J]. arXiv preprint arXiv:2305.11000, 2023.
- [192]. WU W, JIANG C, JIANG Y, et al. Do PLMs Know and Understand Ontological Knowledge?[C]//Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2023: 3080-3101.
- [193]. YIN Z, SUN Q, GUO Q, et al. Do Large Language Models Know What They Don't Know?[J]. arXiv preprint arXiv:2305.18153, 2023.
- [194]. CHARAN P V, CHUNDURI H, ANAND P M, et al. From Text to MITRE Techniques: Exploring the Malicious Use of Large Language Models for Generating Cyber Attack Payloads[J]. arXiv preprint arXiv:2305.15336, 2023.
- [195]. DERNER E, BATISTIČ K. Beyond the Safeguards: Exploring the Security Risks of ChatGPT[J]. arXiv preprint arXiv:2305.08005, 2023.
- [196]. DASH B, SHARMA P. Are ChatGPT and deepfake algorithms endangering the cybersecurity industry? A review[J]. International Journal of Engineering and Applied Sciences, 2023, 10(1).
- [197]. TSIGARIS P, TEIXEIRA DA SILVA J A. Can ChatGPT be trusted to provide reliable estimates?[J]. Accountability in Research, 2023: 1-3.



**许志伟**(1979—), 信创海河实验室, 博士, 研究员, 硕士生导师, CCF 会员。中国科学院计算技术研究所, 客座研究员。主要研究方向为面向边缘智能的可信计算与隐私计算、AIGC 大模型测评。

**XU Zhiwei**, born in 1979. Haihe Laboratory of Information Technology Application Innovation, PhD, Research Professor, Master supervisor, member of CCF, while working as an Adjunct Research Professor of Institute of Computing Technology, Chinese Academy of Sciences, His main research interests include trusted computing and privacy computing for edge intelligence, AIGC model evaluation.



**李海龙**(2000—), 内蒙古工业大学, 硕士。主要研究方向为 AIGC 大模型测评。

**LI Hailong**, Inner Mongolia University of Technology, born in 2000. Master. His main research interests include AIGC model evaluation.



**李博**(1998—), 内蒙古工业大学, 硕士。主要研究方向为 AIGC 大模型测评。

**LI Bo**, born in 1998. Inner Mongolia University of Technology, Master. His main research interests include AIGC model evaluation.



**李涛**(1977—), 南开大学教授、博导, CCF 杰出会员、理事。主要研究方向为异构计算、智能物联网和区块链系统。

**LI Tao**, born in 1977. Nankai University, Professor, PhD. Distinguished member of CCF. His main research interests include heterogeneous computing, machine learning and IoT.



**王嘉泰**(1998—), OPPO 研究院, 硕士, 算法工程师。主要研究方向为多模态表示学习、大语言模型。

**WANG Jiatai**, born in 1998, OPPO Research, Master, Algorithm Engineer. His main research interests include multi-modal representation learning, large language models.



**谢学说**, 信创海河实验室, 博士, 副研究员。主要研究方向为物联网云边端智能协同调度、安全智能物联网、物联网与人工智能及区块链等技术融合。

**XIE Xuesuo**, PhD., Haihe Laboratory of Information Technology Application Innovation, Associate Research Professor. His main research interests include cloud-edge-device intelligent collaborative scheduling of the Internet of Things, secure intelligent Internet of Things, Internet of Things, artificial intelligence and blockchain technology integration.



**董泽辉**(1997—), 信创海河实验室, 硕士, 助理工程师。主要研究方向为面向边缘智能的可信计算与隐私计算、AIGC 大模型测评。

**DONG Zehui**, born in 1997, Haihe Laboratory of Information Technology Application Innovation, Master, Assistant Engineer. His main research interests include trusted computing and privacy computing for edge intelligence, AIGC model evaluation.