CAICT 中国信通院

蓝皮报告

大模型基准测试体系研究报告（2024 年）

中国信息通信研究院人工智能研究所

人工智能关键技术和应用评测工业和信息化部重点实验室

2024年6月

版权声明

前　言

近几年，大模型推动人工智能技术迅猛发展，极大地拓展了机器智能的边界，展现出通用人工智能的“曙光”。如何准确、客观、全面衡量当前大模型能力，成为产学研用各界关注的重要问题。设计合理的任务、数据集和指标，对大模型进行基准测试，是定量评价大模型技术水平的主要方式。大模型基准测试不仅可以评估当前技术水平，指引未来学术研究，牵引产品研发、支撑行业应用，还可以辅助监管治理，也有利于增进社会公众对人工智能的正确认知，是促进人工智能技术产业发展的重要抓手。全球主要学术机构和头部企业都十分重视大模型基准测试，陆续发布了一系列评测数据集、框架和结果榜单，对于推动大模型技术发展产生了积极作用。然而，随着大模型能力不断增强和行业赋能逐渐深入，大模型基准测试体系还需要与时俱进，不断完善。

本研究报告首先回顾了大模型基准测试的发展现状，对已发布的主要大模型评测数据集、体系和方法进行了梳理，分析了当前基准测试存在的问题和挑战，提出了一套系统化构建大模型基准测试的框架——“方升”大模型基准测试体系，介绍了基于“方升”体系初步开展的大模型评测情况，并对未来大模型基准测试的发展趋势进行展望。面向未来，大模型基准测试仍存在诸多开放性的问题，还需要产学研各界紧密合作，共同建设大模型基准测试标准，为大模型行业健康有序发展提供有力支撑（联系人：韩旭，hanxu5@caict.ac.cn）。

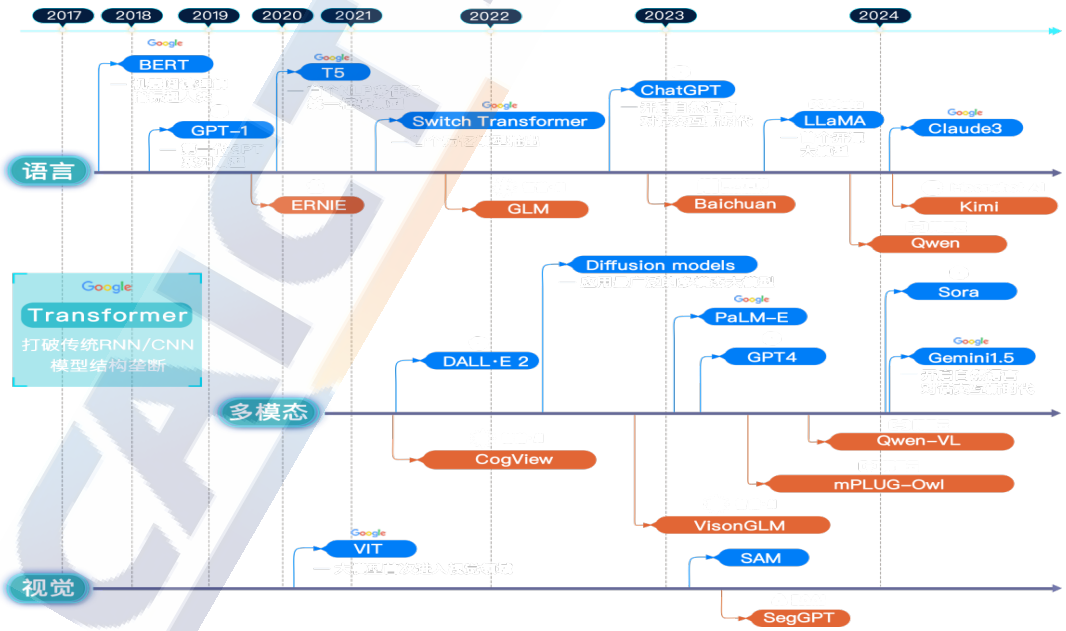
目 录

图 目 录

表 目 录

一、大模型基准测试发展概述

近几年，大模型推动人工智能技术迅猛发展，极大地拓展了机器智能的边界，展现出通用人工智能的“曙光”，全球各大科技巨头和创新型企业纷纷围绕大模型加强布局。如图 1 所示，2018 年，谷歌公司提出基于 Transformer 实现的预训练模型 BERT，在机器阅读理解水平测试 SQuAD 中刷新记录。同年，OpenAI 公司发布了第一代生成式预训练模型 GPT-1，擅长文本内容生成任务。随后几年，OpenAI 相继推出了 GPT-2 和 GPT-3，在技术架构、模型能力等方面进行持续创新。2022 年 11 月，OpenAI 发布的 ChatGPT 在智能问答领域上的表现引起产业界轰动。除了大语言模型，2023 年，OpenAI 还发布了多模态大模型 GPT-4。同期国内大模型的发展也呈现不断加速态势，已经发布了华为“盘古”、百度“文心一言”、阿里“通义千问”、腾讯“混元”和智谱“清言”等 200 多个通用和行业大模型产品。



来源：中国信息通信研究院

图 1 大模型发展历程

随着大模型产品的不断推出，对大模型的能力进行评测逐渐成为产业界关注的重点。1950 年代提出的图灵测试（Turing Testing）作为一种经典的人工智能测试方法，一直被认为是衡量机器智能水平的“试金石”。 2023 年 7 月《自然（Nature）》发表文章《ChatGPT broke the Turing test — the race is on for new ways to assess AI》，指出图灵测试已经无法满足大模型的评测要求，应该探索新方法来评估人工智能水平。

大模型基准测试（Benchmark）的目标是通过设计合理的测试任务和数据集来对模型的能力进行全面、量化的评估。大模型基准测试体系涵盖了大模型的测评指标、方法、数据集等多项关键要素，是指导大模型基准测试落地实践的规范。

（一）大模型基准测试的重要意义

当前，基准测试已赋能大模型“建用管”全生命周期的多个阶段，在大模型研发、应用和管理中扮演重要角色，主要表现在：

一是指引学术研究。过去一年，在 ChatGPT 的引领下，国内外的大模型企业也从最初摸索和尝试，逐渐步入研发和应用深水区。大模型研发迭代周期正在缩短，OpenAI 在一年时间内先后发布 ChatGPT、GPT4、GPT-4V 等多款大模型，Meta 的 LLaMA 大模型一经发布便迅速带动了 Alpaca、Vicuna 等几十个开源大模型，形成“羊驼”开源大模型生态圈。在如此高的迭代频率下，大模型基准测试可以验证模型研发效果，快速挖掘大模型当前的不足与痛点问题，推动大模型能力持续提升。并且，大模型评测不应该是开发流程的终点，

而应该作为起点驱动模型开发。构建以能力提升为目标的评估（Enhancement-Oriented Evaluation）策略对大模型发展十分重要，建立“开发-部署-应用-测试”的闭环流程将缩短产品迭代周期。

二是指导产品选型。近期，商业公司和研究机构等纷纷推出大模型榜单来对大模型的能力进行排序，大模型“打榜”逐渐成为各界关注的话题。国外大模型榜单 Open LLM Leaderboard 使用 4 个公开数据集对大模型进行综合测评。加州大学伯克利分校借鉴 Elo 评分系统推出了 Chatbot Arena，采用众包方式对大模型进行匿名、随机化的对战，得到模型的能力分级。斯坦福大学的 AlpacaEval 使用强大的语言模型（如 GPT-4）对大模型进行评估，提升评测效率。国内的 OpenCompass、FlagEval、SuperCLUE、SuperBench 等分别发布大模型评测榜单，对中文大模型进行重点评测。大模型能力“榜单”确实能够在一定程度上反映出大模型能力，对于大模型的科学研究和能力提升提供正向借鉴意义。此外，在大模型的实际应用中，大模型的使用方需要综合考虑业务需求、花费成本、系统架构、安全要求等因素进行大模型的产品选型（POC）测试。大模型基准测试利用客观数据集对模型能力进行全面、客观的验证，这已经成为 POC 测试的主要落地方式，在大模型行业和应用落地中扮演重要角色。

三是支撑行业应用。近期，“人工智能+”行动的开展驱动了大模型在各应用场景中落地。大模型已经在金融、医疗、软件工程、教育、法律、科研、政务、电信、能源、工业、汽车、机器人等行业领域中取得一定的应用成果。同时，面向行业的大模型基准测试也取得

显著进展，目前已推出多种面向行业应用的评测数据集，例如金融领域的 FinEval，医疗领域的 PubMedQA，软件领域的 MBPP、HumanEval 等。用户在进行大模型行业应用时，无论通过外部采购或自主研发的方式构建大模型能力，都需要利用基准测试对备选大模型进行量化评估，才能保障大模型的行业应用效果。

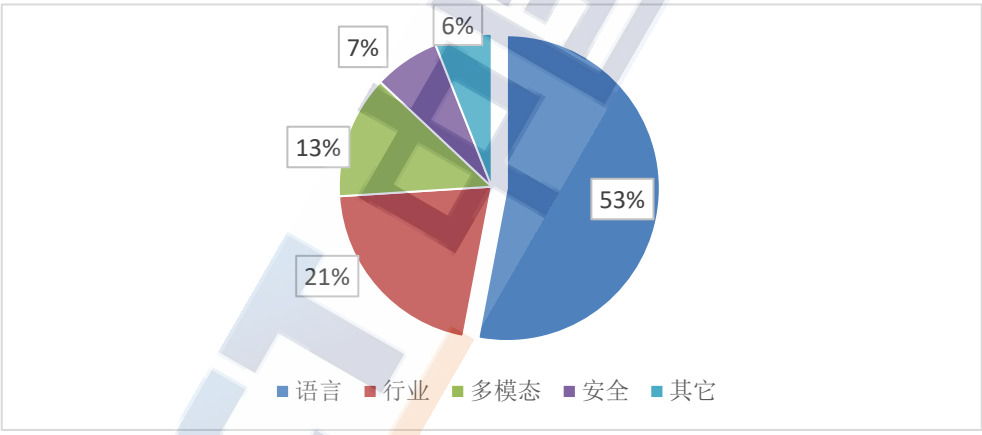
四是辅助监管治理。随着大模型性能的不断提升，安全隐患和威胁的阴影始终如达摩克里斯之剑悬在人类头顶。近期，人工智能专家 Geoffrey Hinton 在接受《60 分钟》公开采访中表示了对人工智能存在的安全隐患的担忧，并担心人类将会被其接管。目前随着 TOXIGEN、CVALUES 等数据集推出，对大模型的内容合规评测等已经取得一定进展，但在大模型的诚实性、自主意识和隐私保护等方面仍缺乏高质量基准。大模型基准测试对保障模型内容安全和能力监控发挥重要作用，可以引导其朝着更健康、更安全的方向发展，让大模型的成果惠及全人类。

（二）蓬勃发展的大模型基准测试

据中国信息通信研究院（以下简称“中国信通院”）统计，截止到 2023 年底，产学研各界已经报道 325 个大模型基准测试的相关数据集、方法和榜单等研究成果。其中，使用频次较高的评测数据集包括加州大学伯克利分校的 MMLU、Open AI 的 GSM8K、上海交通大学的 C-Eval 等；大模型基准测试体系和工具包括美国斯坦福大学的 HELM 和 HEIM、上海 AI 实验室的 OpenCompass、北京智源研究院的 FlagEval、ChineseGLUE 的 SuperCLUE、清华大学的 SuperBench

等；大模型评测榜单包括 Hugging Face 推出的 Open LLM Leaderboard、加州大学伯克利分校的 Chatbot Arena、斯坦福大学的 AlpacaEval 等。通过对现有成果进行梳理，观察到如下现象。

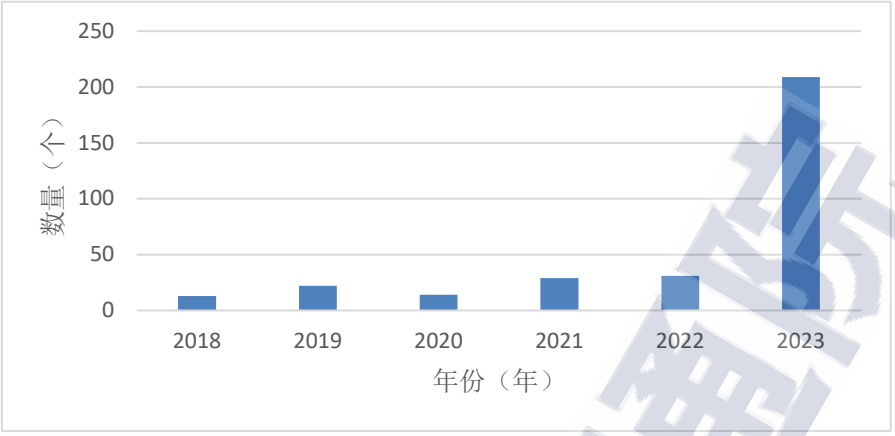
一是从测试领域占比来看，如图 2 所示，由于大语言模型是当前产业应用的主流，因此针对大模型的通用语言类评测数据集最多，占比超过 50%，多模态大模型评测数据集数量仅占 13%。面向行业类的评测数据集 2023 年也迎来爆发式发展，其中 80%也针对语言类任务构建。而对于模型安全、可靠性和鲁棒性评测的数据集较少，需要持续投入。此外，当前对大模型产业应用效果的评测数据集和方法论相对缺乏，亟需产学研各界重点关注。



来源：中国信息通信研究院

图 2 大模型基准测试领域占比分布

二是从发布时间来看，2023 年不但是大模型的涌现年，也是大模型基准测试的爆发年。如图 3 所示，仅 2023 年一年出现的大模型基准测试数据集的数量远远超过之前 5 年，达到 209 个。预计在 2024 年，大模型基准测试数据集的数量仍会持续攀升。



来源：中国信息通信研究院

图 3 大模型基准测试数据集发布时间

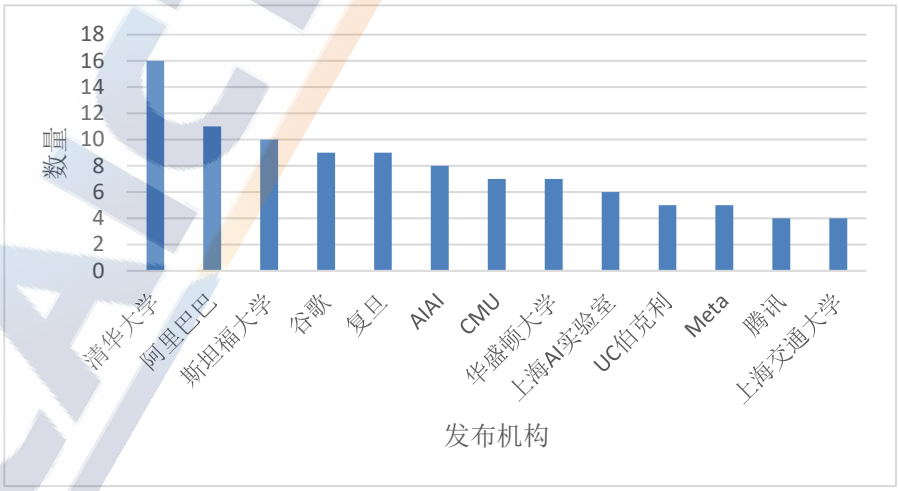
三是从大模型基准测试数据集的使用频次来看，如表 1 所示，通过对 GPT-4、LLaMA 2、LLaMA 3、Gemini、Claudes 3、Mixtral 8x7B、GLM4 等大模型官方发布结果中使用的评测数据集进行统计，MMLU、GSM8K、ARC、HumanEval、Math、BBH、WinoGrande、HellaSwag 等基准的使用频次较高，其中大部分为传统的自然语言处理评测数据集，并主要针对大模型的英文能力进行测试。对于多模态大模型，LLaVA-Bench、VisIT-Bench、MMBench 等使用较为广泛。

表 1 代表性大模型官方发布结果中使用的评测数据集

□□□□□□□	GPT-4	LlaMA2	LlaMA3	Gemini	Claude3	Mixtral8x7B	GLM4
MMLU	√	√	√	√	√	√	√
GSM8K	√	√		√	√	√	√
ARC	√	√	√		√	√	√
HumanEval	√	√		√	√	√	√
Math		√		√	√	√	√
BBH		√	√	√	√		√
WinoGrande	√	√	√		√	√	
HellaSwag	√	√		√		√	

MBPP		√			√	√	
DROP	√			√	√		
TriviaQA		√	√			√	
GPQA			√	√	√		
AGIEval		√	√				
PIQA		√				√	
MGSM				√	√		
NQ		√				√	
MGSM				√	√		
SQuAD		√	√				
BoolQ		√	√				

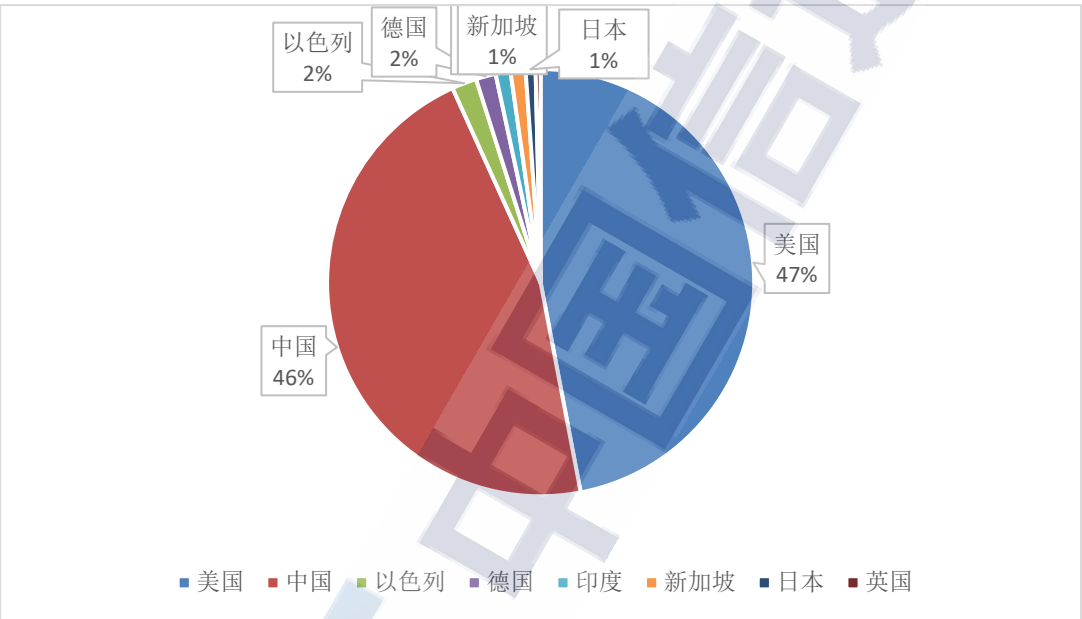
四是从发布机构上来看，学术机构在此领域的研究中扮演了重要角色。如图 4 所示，清华大学和斯坦福大学位于发布评测数据集数量的第一名和第三名，其中清华大学的成果大多集中在 2023 年。美国艾伦人工智能研究所（AI2）由于在传统自然语言处理数据集上的贡献，仍然位居前列。谷歌、阿里巴巴、Meta 和腾讯成为上榜的四家企业。



来源：中国信息通信研究院

图 4 大模型基准测试数据集发布机构排名

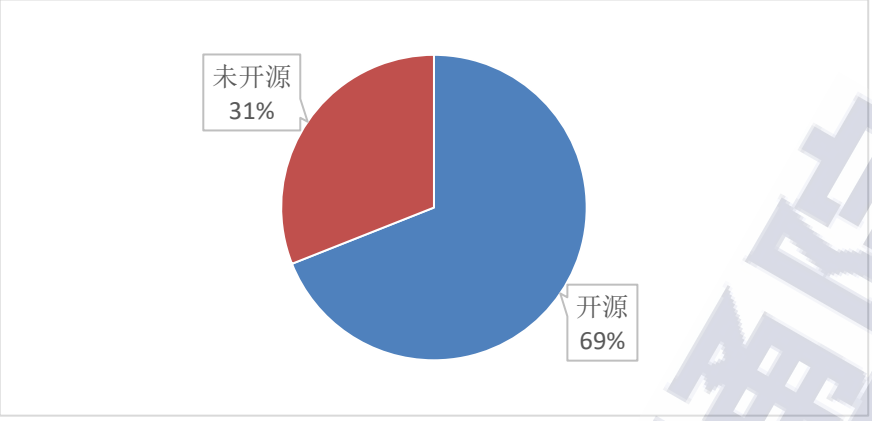
五是从发布国家来看，如图 5 所示，中美发布数量旗鼓相当，占比均为 47%左右。2023 年，国内大模型基准测试数据集“井喷式”发展，推出包括 C-Eval、CMMLU 等评测数据集 100 多个，在中文评测领域具有显著的影响力。虽然国内提出的基准数据集在数量上有明显提升，但与美国提出的基准测试数据集相比，在国际上的影响力仍然差距明显。



来源：中国信息通信研究院

图 5 大模型基准测试数据集发布国家分布

六是从测试数据集开源状况来看，如图 6 所示，开源测试数据集更多，占比达到 69%，而闭源数据集仅占 31%。评测数据集开源对其推广影响很大，产学研各界只有充分获取数据才可以高效进行测试。但同时数据的开源会容易导致模型“作弊”的现象发生。因此，如何在保证数据充分开放的前提下，对模型的数据污染状况进行检测成为当前研究的热点。



来源：中国信息通信研究院

图 6 大模型基准测试数据集开源分布

（三）大模型评测发展共性与差异

当前人工智能测试已经由机器学习、深度学习测试时期进入大模型测试时期，未来还将迈向通用人工智能（AGI）测试时期。产学研各界推出的大模型基准测试数据集众多，这些数据集的构成和测试重点各不相同，但表现出一些共性：

一是通用能力测试为主。目前产学研各界所发布的大模型基准测试数据集大都侧重于模型的通用能力，包括大模型理解、生成、推理、知识能力等，MMLU 和 GSM8K 等成为当前大模型最常用的评测基准，而近期面向行业和应用的评测数据集已得到产业界广泛关注。

二是通过考试方式执行。虽然 Chatbot Arena 等采用“模型对战”的方式完成评测，但当前大模型基准测试主要仍以考试方式为主，通过在考题上的表现来衡量大模型能力。AGIEval、KoLA 等利用客观选择题评测大模型知识能力，PubMedQA 等通过问答题评测生成能力。

三是测试数据构成类似。大模型基准测试的输入通常为测试数据，常见的测试数据类型包括单选、多选、问答等。为提升自主测试效率，

数据集还会提供标准答案、Prompt 样例和测试脚本等。同时，评测数据集中通常还会包含一定量的模型微调数据来提升大模型表现。

四是测试结果仍需主观评估。当测试题目为客观选择题，测试结果评估可以通过脚本高效执行。当测试题目为主观题或开放问答时，仍然需要人工主观评估。虽然大模型已经作为“裁判”参与结果评估，但据论文《Large Language Models are not Fair Evaluators》研究表明，使用 GPT-4 进行结果评估容易受到“答案顺序”等因素影响。

除了上述共性外，大模型基准测试数据集也表现出一定差异性，主要为：**一是评测数据数量上的差异，**知识类考察数据集的题目数量通常会超过 1 万，例如 MMLU 和 C-Eval 的题目数量分别为 15858 和 13948，而代码类评测数据集中题目数量较少，如 MBPP 和 HumanEval 的题目数量仅为 974 和 164。**二是评测环境上的差异，**对语言大模型的评测通常以考试的方式进行，而对于 AI 智能体（AGENT）或具身智能系统的评测通常需要搭建仿真环境。**三是评测目标上的差异，**大模型的训练可分为预训练、监督式微调、强化学习训练等几个阶段，不同的评测数据集所针对的目标模型不相同。**四是评测方法上不统一，**根据提示工程中提供样例多少，大模型可通过 zero-shot、few-shot 等方式进行评测，但各大模型在评测方式上并不统一。

二、大模型基准测试现状分析

2023 年，大模型基准测试迎来飞速发展的一年，大模型的评测体系、数据集、方法、工具如雨后春笋般出现。本章对已发布的大模型基准测试成果进行简要介绍，主要分为评测体系、数据集和方法等，以梳理大模型基准测试的整体发展趋势，并探寻未来发展方向。

（一）大模型基准测试体系总体介绍

与传统认为 Benchmark 仅包含评测数据集不同，大模型基准测试体系包括关键四要素：测试指标体系、测试数据集、测试方法和测试工具。指标体系定义了“测什么？”，测试方法决定“如何测？”，测试数据集确定“用什么测？”，测试工具决定“如何执行？”。

图 7 大模型基准测试体系构成

1. 测试指标体系

在进行大模型基准测试时，首先需要确定测试的指标体系，明确评测的维度和对应指标。大模型评测的指标体系可以按照<场景-能力-任务-指标>四层结构进行构建。测试场景定义了待测试模型的外在

环境条件的组合，如通用场景、专业场景、安全场景等。测试能力决定了模型的测试维度，如理解能力、生成能力、推理能力、长文本处理能力等。针对待测试的能力，可以通过多种任务完成测试。如语言大模型的理解能力可以重点考察在文本分类、情感分析、阅读理解、自然语言推理、语义歧义消解等任务中的表现。对于不同的测试任务，需要与不同的指标进行关联。如文本分类可以计算准确率、召回率等指标，而阅读理解可以利用准确率、F1 Scores、BLUE、ROUGE 等进行考察。

2. 测试数据集

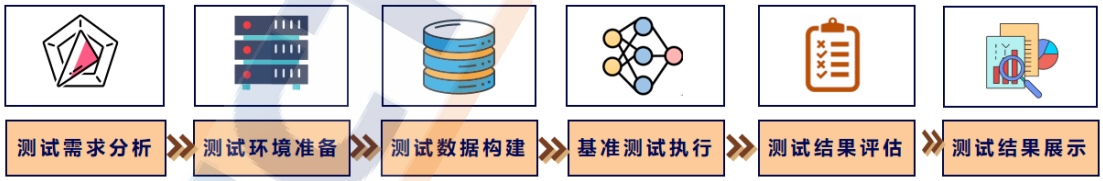
按照大模型可处理的信息模态，可以将大模型分为语言大模型、多模态模型、语音大模型等。其中，语言大模型和多模态大模型的研究和应用最为广泛。语言大模型的输入和输出均为自然语言，多模态大模型的输入和输出为不同模态的数据。下面对语言大模型和多模态模型评测中常用数据集进行梳理和介绍。

对语言大模型的通用能力进行评测需要考察理解能力、生成能力、推理能力、知识能力、学科能力、多语言能力、长文本能力、思维链能力、角色扮演能力、工具使用能力、可靠性、鲁棒性等。代表性的评测数据集如附录表 1 所示，包括 MMLU、BBH、GSM8K 等。对语言大模型的行业能力进行评测需要考察行业通用能力、行业知识能力、行业场景能力、行业安全能力等。代表性的评测数据集如附录表 2 所示，包括 FinEval、PubMedQA、JEC-QA 等。对语言大模型的应用能力进行评测需要考察大模型在智能客服、知识管理、数据分析、办公

助手、内容创作、网页助手、代码助手、任务规划、智能代理、具身智能等应用中的效果。代表性的评测数据集如附录表 3 所示，包括 GAIA、APPS、AgentBench 等。对语言大模型的安全能力进行评测需要考察大模型内容安全、伦理安全、隐私安全、模型安全等，代表性的评测数据集如附录表 4 所示，包括 SafetyBench、TOXIGEN、JADE 等。当前对多模态大模型的评测主要集中在通用能力，主要包括视觉问答、视觉推理、视觉处理、视觉描述、视觉生成、可靠性等。代表性评测数据集如附录表 5 所示，包括 MMBench、LLaVA-Bench、POPE、OCRBench 等。

3. 测试方法

大模型基准测试方法的研究主要集中在大模型的整体评测流程或评测方式的创新。如图 8 所示，大模型的评测流程包括测试需求分析、测试环境准备、测试数据构建、基准测试执行、测试结果评估和测试结果展示等。本报告对每个环节涉及的内容进行介绍。



来源：中国信息通信研究院

图 8 大模型基准测试流程

测试需求分析通常是大模型测试过程中的第一步，通过对测试需求进行全面和准确的覆盖，有助于确保测试活动的有效性和高效性。大模型测试需求分析需要完成以下任务：确定评测目的，预评估待测模型，测试体系设计，测试方案设计，测试输入（输出）分析，测试

可实施性分析等。

测试环境准备是大模型测试的基础，需要搭建配套的软硬件平台保证测试顺利执行。首先，根据被测模型的实际性能要求需要搭建测试软硬件环境。其次，对于单一模型的少样本测试，可利用脚本完成测试，而对于多个模型的大数据量测试，需要使用测试框架，可将其部署在单一服务器或集群中。再者，对私有化部署大模型，需要将其部署在环境中。最后，可使用少量测试数据对测试环境功能进行验证。

大模型评测数据可以通过人工构建、题目自动化扩充和智能算法生成三种方式进行定期补充或更新。人工构建方式主要是通过人工采集、标注的方式构建测试数据。面向大模型的测试数据的构建流程一般包括方案设计、数据采集、数据标注、数据清洗、数据增强、数据规范化和数据存储等环节。在实际大模型评测中，应针对模型的薄弱点定期进行评测数据集的更新工作，以保证评测数据的有效性。

题目自动扩充主要利用“模板”化信息提取算法或对抗样本生成对题库中题目的可变量进行“替换”，从而“衍生”生成相似题目。其在一定程度上防止大模型通过“刷题”和“记题”方式获取更高的分数，并验证大模型的鲁棒性。微软提出动态测试框架 DyVal，利用有向无环图动态生成测试数据，减少测试数据被大模型记忆的可能。PromptBench 对大模型的提示工程词进行字符级别、单词级别、句子级别和语义级别的黑盒攻击，来对语言大模型的鲁棒性进行评测。智能算法生成主要是利用一些先进的人工智能技术（如大模型）自动化生成一定量的新题目。目前基于大模型的智能出题已有实际的应用范

例，例如考试星推出的智能考试命题服务中，使用大模型对一段长文本进行自动化出题，涉及题型包括单选、多选、问答等。香港中文大学推出了数学推理问题的合成数据方法 MathGenie，通过训练一个反向翻译模型对种子试题集的增广解决方案进行反演，从而得到更多的数学题目。目前智能算法生成的题目质量很难保证，需要人工进行核验，以确保测试题目质量。

为了保证测试结果的公正性，大模型评测数据集应该提供统一、标准的提示工程（Prompt）范例，支持 Zero-shot、Few-Shot 等多种评测模式。通过优化提示工程词内容可以提升大模型的表现，但为了保证结果的可比性，推荐使用评测数据集所提供的提示工程样例，并且所有的大模型所使用的评测提示工程词应该保持一致。

测试执行阶段需要将测试数据输入被测模型，并观察被测模型的输出结果。从执行方式上，根据实际需求（测试数据量、测试成本等）可使用单点和分布式两种方式执行。单点执行在单台服务器上将测试数据依次输入大模型，并收集大模型的输出结果。分布式执行通过中心节点对测试任务和数据集进行切分，再分发至单点服务器上分布式执行，最后通过中心节点对大模型输出结果进行汇总并统计，测试成本相对较高。

对于大模型生成的结果需要使用合理的评估指标进行衡量，以确保生成内容的正确性和准确性。**大模型生成内容的评估方式可以分为自动化评估和人工评估。**传统自动化评估通过计算特定指标完成模型生成内容和标准答案的对比。对客观类评测题目（如选择题）的结果

评估相对简单，若模型的回答不满足提示工程词要求，会采取特定的策略(如正则匹配)完成答案的对比。由于大模型生成内容较为灵活，对主观类题目（如问答题）进行自动化评估难度较高。若生成内容较为规范，如机器翻译和文本摘要等，可以计算 BLEU、ROUGE 等指标。但对于较复杂或专业的生成内容，需要专家对结果的正确性和准确性进行人工评判，其对评估人员资质和具体评测方式等有一定要求，如评估人员需要具有专业化背景、评估人员数量要充足等。

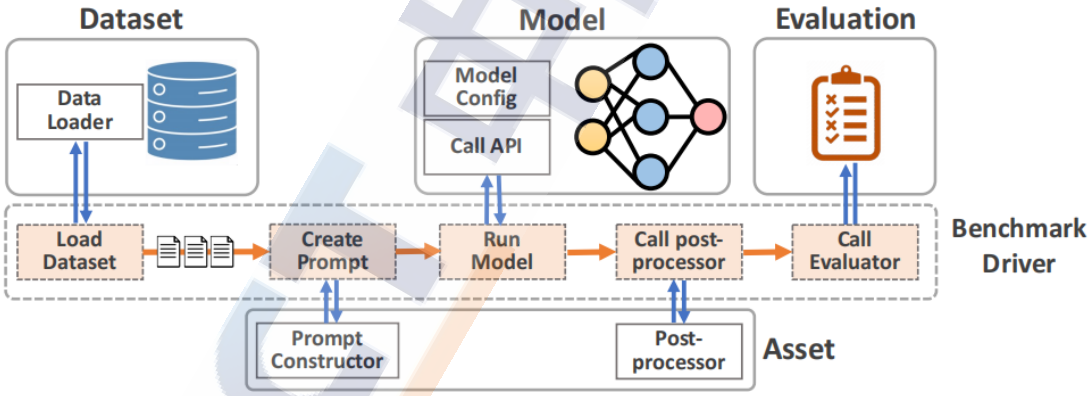
现有研究尝试将大模型作为自动化结果评估工具来对其它模型的生成内容质量进行评估，例如 AlpacaEval 等采用 GPT-4 对其它模型的生成结果进行质量分级。根据《Benchmarking Foundation Models with Language-Model-as-an-Examiner》等论文结果，这种评估方式有望成为人工评估的有效替代。其按照技术原理可分为基于提示工程词和模型微调两种方式。前者一般会设计高质量的提示词，利用大模型来对生成内容进行打分。该方式可通过优化提示词内容或构建大模型裁判网络来提升评估效果。中科院在论文《Wider and deeper llm networks are fairer llm evaluators》中以大模型作为神经元搭建“裁判”网络 WideDeep，人机评估一致率达到 93%。基于模型微调的方式主要利用相关数据对大模型进行训练以提升评判的准确率。代表性的成果包括清华大学的 CRITIQUELLM 和北京智源研究院的 JudgeLM 等。

大模型基准测试结果可以通过测试报告、模型榜单、雷达图、柱状图等多种形式进行展示。大模型测试报告中需要包含评测目标、数据集描述、测试任务描述、测试环境描述、评估指标、量化结果、可

视化结果、对比分析、评测结论、建议提升方向、错误样例等内容。

4. 测试工具

测试工具是测试方法的落地实践方式，是提升大模型评测效率的重要手段。大模型基准测试工具通常需要支持数据集管理、模型库管理、API 管理、测试任务分发、测试指标计算、测试结果分析、测试结果展示等多种基础功能。图 9 展示了由卡塔尔计算研究所所提出的开源大模型基准测试评测工具 LLMeBench 原理图。从图中可以发现，其包含数据加载模块、提示工程词模块、模型执行模块、后处理模块和结果评估模块，与大模型的基准测试流程基本一致。当前大模型的基准测试工具在测试数据集构建和测试结果评估阶段仍然需要人工参与，全自动化的基准测试工具仍是产业界的迫切需求。



来源：《LLMeBench: A Flexible Framework for Accelerating LLMs Benchmarking》

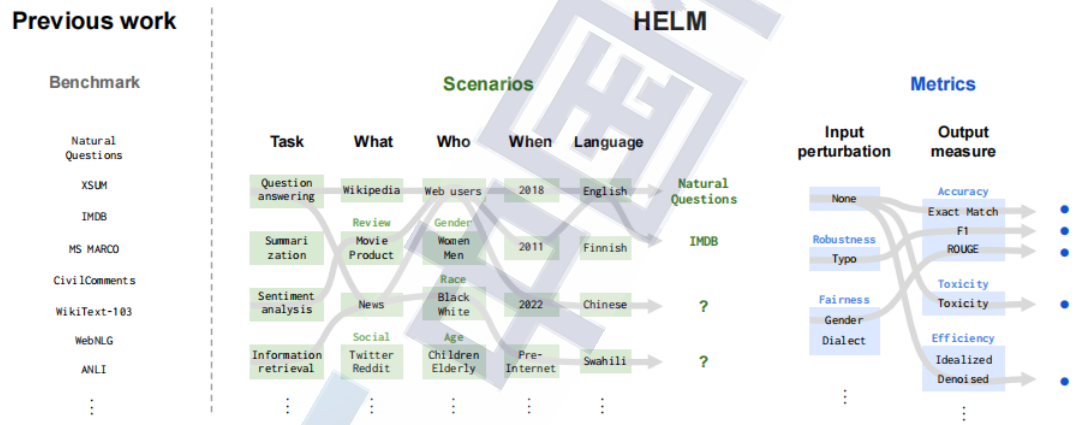
图 9 大模型基准测试工具 LLMeBench 框架图

（二）代表性的大模型基准测试体系

当前已发布的评测榜单背后均有相应的评测体系和方法，国内外知名度较高的大模型基准测试体系包括：

1. HELM

HELM（Holistic Evaluation of Language Models）是由斯坦福大学在2022年推出的大模型评测体系。该体系主要包括了场景（Scenarios）、适配（Adaptation）和指标（Metrics）三个核心模块，每次评测都需要“自顶而下”指定一个场景、一个适配模型的提示工程词和一个或多个指标来进行。如图 10 所示，HELM 使用了几十个场景和多个指标的核心集完成大模型评测，场景涉及问答、信息检索、摘要、毒性检测等多种典型评测任务，指标包括准确性、校准、鲁棒性、公平性、偏差、毒性、效率等。



来源：《Holistic Evaluation of Language Models》

图 10 大模型评测基准 HELM 原理图

2. HEIM

HEIM（Holistic Evaluation of Text-to-Image Models）是由斯坦福大学在 2023 年推出的多模态大模型评测体系。与之前文本生成图像的评测主要关注文本图像对齐和图像质量不同，HEIM 定义包括文本图像对齐、图像质量、美学、原创性、推理、知识、偏见、毒性、公平性、鲁棒性、多语言性和效率在内的 12 个维度。HEIM 确定包含

这些维度的 62 个场景，并在这个场景上评测了 26 个最先进的文本到图像的生成模型。

3. HRS-Bench

HRS-Bench（Holistic Reliable Scalable Bench）是由沙特的 KAUST 在 2023 年推出的全面、可靠、可扩展的多模态大模型评测体系。与之前文本生成图像仅考察有限维度不同，HRS-Bench 重点评测大模型的 13 种技能，可分为准确率、鲁棒性、泛化性、公平性和偏见 5 个类别，覆盖了包括动物、交通、食物、时尚等 50 多个场景。

4. OpenCompass

OpenCompass（司南）是由上海 AI 实验室推出的开源、高效、全面的评测大模型体系及开放平台，其包括评测工具 CompassKit、数据集社区 CompassHub 和评测榜单 CompassRank。在已发布的评测榜单中，对语言大模型主要考察语言、知识、推理、数学、代码和智能体方面的表现。对多模态大模型主要评测在 MMBench、MME 等数据集上的指标。OpenCompass 提供了开源大模型基准测试工具，已集成大量的开源大模型和闭源商业化 API，在产业界影响力较大。

5. FlagEval

FlagEval （天秤）是由北京智源研究院推出的大模型评测体系及开放平台，其旨在建立科学、公正、开放的评测基准、方法、工具集，协助研究人员全方位评估基础模型性能，同时探索提升评测的效率和客观性的新方法。FlagEval 通过构建“能力-任务-指标”三维评测框架，细粒度刻画基础模型的认知能力边界，包含 6 大评测任务，近 30

个评测数据集和超 10 万道评测题目。在 FlagEval 已发布的榜单中，其主要通过中、英文的主、客观题目对大模型进行评测，具体任务包括选择问答和文本分类等。

6. SuperCLUE

SuperCLUE 是由 ChineseCLUE 团队提出的一个针对中文大模型的通用、综合性测评基准。其评测范围包括模型的基础能力、专业能力和中文特性，基础能力包括语言理解与抽取、闲聊、上下文对话、生成与创作、知识与百科、代码、逻辑与推理、计算、角色扮演和安全。目前提供的基准榜单包括 OPEN 多轮开放式问题评测、OPT 三大能力客观题评测、琅琊榜匿名对战基准、Agent 智能体能力评估、Safety 多轮对抗安全评估等。除此之外，还针对长文本、角色扮演、搜索增强、工业领域、视频质量、代码生成、数学推理、汽车等领域单独发布大模型能力榜单。

（三）问题与挑战

虽然当前大模型基准测试发展迅速，涉及内容范围广泛，但仍存在一些挑战性问题:

1.建立规范化的评测体系

业界对于大模型应测哪些内容、如何测、使用哪些评测集并没有统一的规范，这容易导致大模型评测榜单结果存在差异，很难精确对比大模型能力。例如，在 2023 年底，谷歌发布 Gemini 大模型，并表示在 MMLU 上的得分率高于 GPT-4。但通过分析谷歌发布的技术报告《Gemini: A Family of Highly Capable Multimodal Models》，Gemini

Ultra 采用“CoT@32”（使用了思维链提示技巧，尝试 32 次并从中选择最好结果）的测试方法，这与 GPT-4 采用的“Few-Shots”明显不同，因此评测结果的公正性受到质疑。

2.构建面向产业应用的基准

由于行业需求经常高度定制和专业化，仅测试大模型的通用能力无法充分评估模型在特定行业中的应用效果。当前一些行业仍然缺乏公开的高质量评测数据集，这加大了对大模型在实际场景中进行全面评测的难度。例如在 Meta 发表的论文《GAIA:A Benchmark for General AI Assistants》中，在 AI 助手的评测基准 GAIA 上，人类回答问题的准确率为 92%，而配备了插件的 GPT-4 只有 15%，这说明大模型在实际应用场景上仍然有较大的提升空间。

3.模型安全能力评估

当前大型模型在常见问题上的回答稳定性较好，但在特定敏感问题或某些“边缘场景”下可能会存在风险，目前国内外针对模型风险的评测基准数量仍然较少。例如，近期大连理工大学联合多家机构发表论文《Automated Evaluation of Large Vision-Language Models on Self-driving Corner Cases》，重点对多模态模型在自动驾驶“边缘场景”中的表现进行评估。

4. 评测结果与用户体验的差异

当前大模型的评测大多对通用知识能力进行考察，对用户的实际使用体验关注度不够，这容易导致用户实际反馈与模型测试排名并不一致，如 Open LLM Leaderboard 和 Chatbot Arena 的评测结果在大模

型的排名上有明显的差异。在清华大学发表的论文《Understanding User Experience in Large Language Model Interactions》中，作者表示当前缺少面向用户体验评估的 Benchmark。

5. 测试数据集的“污染”问题

据美国佐治亚理工大学的论文《Investigating Data Contamination in Modern Benchmarks for Large Language Models》，当前大模型的测试数据容易被包含在训练数据中进行训练，造成数据“污染”问题。产学研各界需要研究数据“污染”的检测手段，降低大模型“刷榜”对评测结果的公正性和可信度产生的影响。

6. 评测数据集的“饱和”使用问题

目前 MMLU、GSM8K 等高质量评测数据已经被大模型评测多次，准确率已经达到一定水平，产学研各界应对评测数据集的选择和构建形成更加科学的方法论。

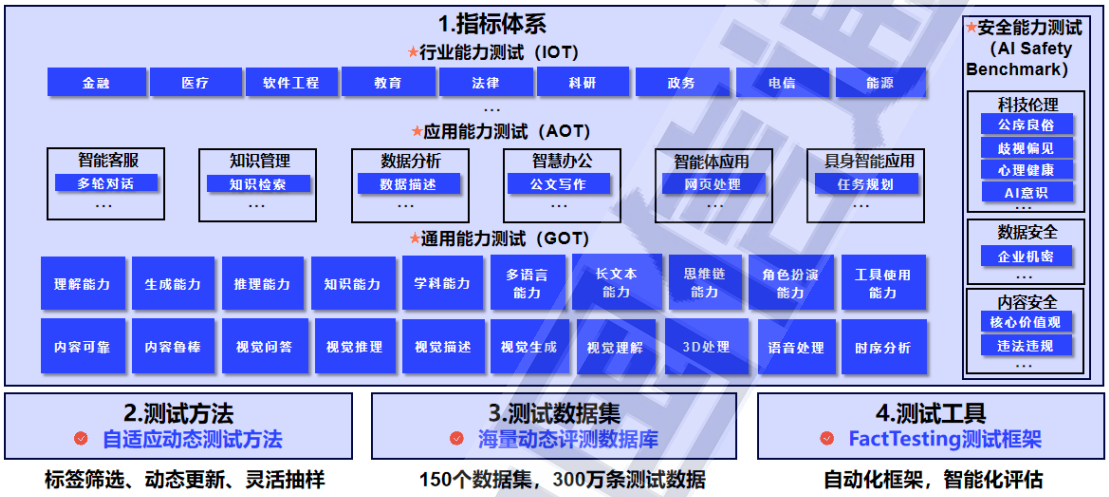
三、大模型基准测试体系框架

大模型基准测试体系涵盖大模型的测评指标、方法、数据集等多项关键要素，是指导大模型基准测试落地实践的规范。大模型基准测试体系的建设和完善，旨在形成一个全面、客观、规范的大模型基准测试的方法论，从而保障大模型评测结果的公正性和客观性。当前大模型的基准测试偏重模型的通用能力，产业界也亟需面向具体场景和实际落地效果的模型评测能力。针对上述问题，中国信通院从指标体系、测试方法、测试数据集和测试工具四个维度出发，构建“方升”大模型基准测试体系，重点面向产业应用效果进行评估，并且推出自适应动态测试方法，努力保证评测结果的公正性和科学性。

（一）“方升”大模型基准测试体系

为提供大模型基准测试体系的规范化建设思路，2023 年底，中国信通院发布“方升”大模型基准测试体系。“方升”体系的发布，由北京智源研究院、认知智能全国重点实验室、天津大学和中国信通院共同见证。此外，国网智能电网研究院、首都之窗、天津大学、中国电信研究院、中国联通软件研究院、华为、甲骨易、海天瑞声、东方财富 9 家单位成为“方升”大模型基准测试首批合作伙伴。如图 11 显示，“方升”测试体系涵盖基准测试的四个关键要素，即指标体系、测试方法、测试数据集和测试工具。其中测试能力主要规定了测试维度与指标，其由“三横一纵”的框架构成，“三横”自顶至下依次为大模型的行业能力测试（Industry-Oriented Testing，IOT）、应用能力测试（Application- Oriented Testing，AOT）和通用能力测试（General-

Oriented Testing，GOT），而“一纵”为大模型的安全能力测试，其在行业、应用和通用能力中都会涉及。显然，“方升”测试体系将从行业、应用、通用和安全能力四个维度全面评估大模型的表现，特别其将重点评估行业和应用能力这两个维度，这对大模型的产业落地具有重要参考价值。



来源：中国信息通信研究院

图 1 “方升”大模型基准测试体系

构建一个高质量的评测基准，不能仅考虑数据集和指标，“方升”测试体系除了对大模型的指标体系进行科学化设计，还对测试方法、测试数据集和测试工具提供规范化的建设思路。在指标体系中，“方升”测试体系除了关注通用能力和安全能力，还重点考察大模型在行业和实际应用中的表现。为保证测试结果的科学性和客观性，大模型的评测需要保证环境和输入的一致性，并在测试方法上进行精细化设计，以满足高效、精准的评测目标。在评测数据集方面，应该满足一定质量要求，如充分性、多样性、新颖性、区分度、合理性、可追溯性等，才能从源头上保证测试真实有效。在评测工具方面，应该进行合理的模块化设计，满足功能和性能要求，支持端到端自动化执行测

试。通过对测试数据自动构建、测试结果自动评估、测试分布式执行等关键环节进行探索，提升大模型基准测试的效率。

面向大模型的通用能力测试（GOT）在产学研各界已得到蓬勃的发展，**“方升”测试体系将全面吸收产学研各界的优秀成果，并在评测大模型的生成能力和内容可靠性等方面进行重点探索，打造全面和坚实的通用能力测试底座。**“方升”测试体系已具备针对大语言模型的理解能力、生成能力、推理能力、知识能力、学科能力、多语言能力、长文本能力、思维链能力、角色扮演能力、工具使用等方面的评测，针对多模态大模型，联合产学研机构建立视觉问答、视觉推理、视觉生成等能力的评测数据集。在“方升”测试体系中，已对大模型的通用能力测试进行全面梳理，形成<领域-能力-任务-指标>的关联关系，从而构建全面且体系化的通用能力评测基础底座。

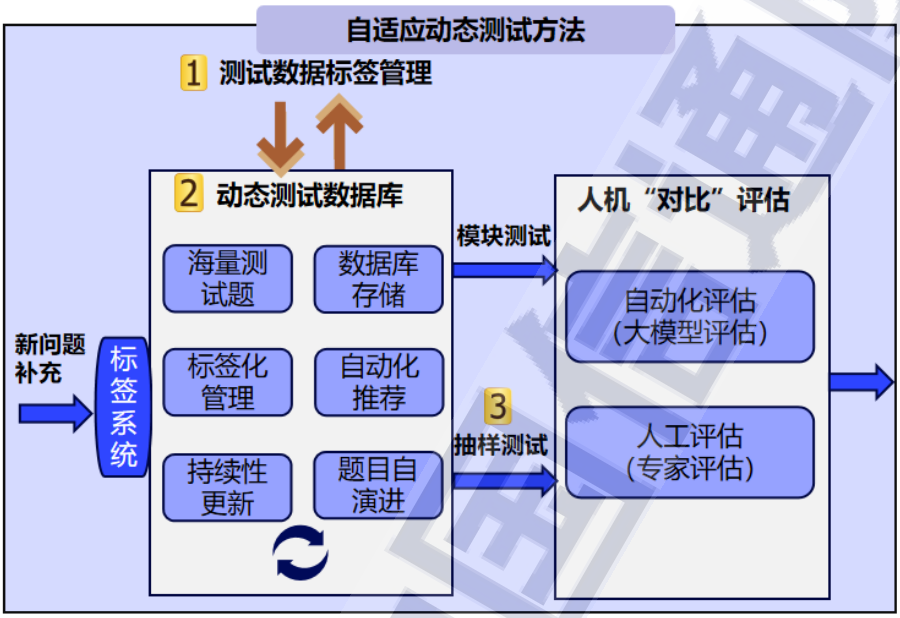
“方升”测试体系在大模型行业测试（IOT）领域进行重点布局，助力大模型赋能千行百业。在大模型实际落地过程中，通常会用行业数据对基础大模型进行微调得到行业大模型，而后将行业大模型应用在实际业务中。然而，由于每个行业的需求和应用场景不同，因此评测方案和数据集也不相同，评测难度明显提升。在“方升”测试体系中，已针对多个重点行业中的典型应用场景进行梳理，形成 “通用-知识-场景-安全” 的多维度评测方案，并在政务、电信等行业进行验证。当前“方升”测试体系中涵盖包括金融、医疗、工程、教育、法律、科研、设计、汽车、机器人等多个行业的评测数据集，并在政务、电信、能源等领域与产学研机构共建评测数据集，助力大模型行业能力评估。

大模型的实际应用通常限定在具体场景和特定任务，**“方升”测试体系面向大模型的应用测试（AOT）进行重点探索，解决大模型业务落地的“最后一公里”问题。**当前大模型常见的落地场景包括智能客服、知识管理、数据分析、办公助手、内容创作、代码生成等。在上述领域中，为了保证大模型生成结果的准确性，通常会利用外挂知识库的方式来进行技术落地。随着大模型能力的提升，可利用外部工具完成更为复杂的任务，例如网络购物、数据库操作等，这需要大模型智能体（AGENT）技术的支撑。“方升”测试体系将针对智能客服、知识管理、RAG、数据分析、代码助手、办公助手、AGENT、具身智能等多个重点应用领域的测试方法进行研究，并通过设计合理的评测指标对实际任务的落地效果进行评估，为大模型应用效果评估遇到的评测数据缺乏问题提供解决方案，全面衡量大模型在实际业务落地中发挥的作用。

安全能力是保障大模型实际落地应用的重要基石，已经成为人工智能领域的核心议题。AI Safety Benchmark **着力打造公平公正、面向产业应用的大模型安全能力测试体系，为大模型产业安全健康发展保驾护航。一是数据集层面，**构建完备的安全测评数据集，涵盖 40 余万条数据，26 个细粒度安全类别和 4 种数据模态。从内容安全、数据安全、科技伦理等方面综合评估大模型安全能力。其中，内容安全涉及价值观、违法违规等；数据安全包括个人隐私、企业机密等；科技伦理包括歧视偏见、心理健康、AI 意识等。**二是评测指标层面，**设置科学的测评指标，从安全性和负责任性两个角度分别衡量大模型的性

能。其中，安全性分数主要关注模型输出的绝对安全性，负责任性分数更加关注模型回答的正向积极性和与人类价值对齐的情况。

（二）“方升”自适应动态测试方法



来源：中国信息通信研究院

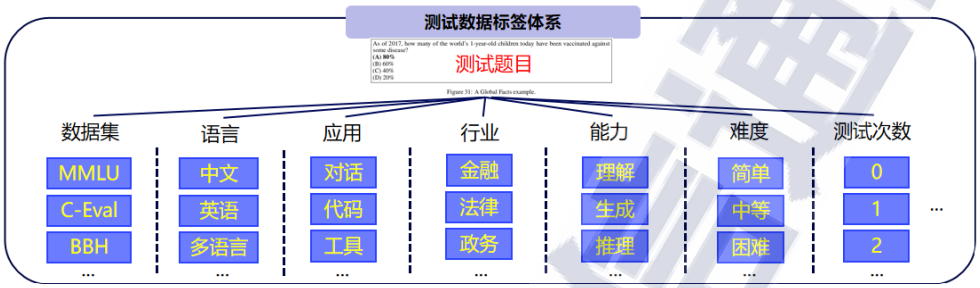
图 2 自适应动态测试方法原理图

“方升”测试体系中的指标部分通过对大模型测试领域和指标的全领域梳理，从方法论上指导用户更加精准且科学的完成测试。除此之外，为了解决测试数据集管理难、大模型测试“刷榜”等问题，“方升”测试体系提出自适应动态测试（Adaptive Dynamic Testing，ADT）方法对大模型进行评测，以保证大模型基准测试能高质、高准、高效的完成。如图 12 所示，自适应动态测试方法包含三个关键部分，即测试数据标签化管理、动态测试数据库和高质量测试数据抽样算法。其中测试数据标签化管理重点解决测试数据集格式繁多、难管理问题，动态测试数据库主要解决大模型测试“刷榜”和评测数据“静态化”问

题，高质量测试数据抽样算法主要解决大模型的精准缺陷挖掘困难高、测试效率较低等问题。

自适应动态测试方法的关键特性包括以下几方面：

1.全量筛选，测试标签匹配化



来源：中国信息通信研究院

图 3 大模型基准测试标签体系

在大模型的实际测试过程中，测试人员很难直接获取相关测试数据和指标，需要花费大量的人力去搜集和整理数据，测试门槛高。并且当测试题库中的测试题目达到百万量级，该如何管理这些数据是一个难点，直接影响测试数据的价值。**“方升”测试体系对全量测试数据进行“标签化”处理，完成测试数据精准“画像”**。如图 13 所示，“方升”测试体系中的测试数据会赋予特定的“测试标签”，例如所属数据集、测试行业、测试领域、测试任务、测试能力、题目难度等。通过多维度的数据标签刻画，充分提升测试数据的利用效率。“方升”测试体系希望通过多层次的梳理和筛选，在构建全面、统一的测试基准同时，可以自动化推荐基准测试的“数据”和“指标”，从而降低大模型基准测试的“门槛”。测试人员在实际测试时，可以参照“方升”测试体系“自顶至下”依次在“行业”、“应用”和“通用”中选择需要的测试维度，

“方升”测试体系可以根据用户的选择自动化推荐测试所需的“数据”和“指标”。

2.动态更新，测试题库实时化

为了防止大模型测试的“刷榜”问题，“方升”测试体系的底层测试数据库采用动态方式构建，保证每次参与测试题目的都不相同，以解决存在题目封闭、考题过时、模型作弊等问题。动态测试数据库中的题目会通过人工补充、题目自生成和智能算法生成三种方式定期进行扩充，从而保证每次测试时都有一定占比的题目从未用于大模型测试。这些数据在一定时间内不会进行公开，后续会根据产业需求进行开放。测试过程中，会参考已有测试结果，通过人工方式定期补充测试数据，对已发现的大模型薄弱能力进行反复测试。题目自生成方式主要针对题库中已有题目“生成”出一些评测题目，从而防止大模型通过“刷题”和“记题”等方式提升模型表现。智能算法生成常利用高质量提示工程驱动大模型自动化生成一定量的测试题目，但这些题目的质量很难保证，需要人工对大模型生成题目进行核验。

3.灵活抽样，测试方案定制化

为了避免大模型测试据集质量不高的问题，“方升”测试体系采用全量测试、模块测试和抽样测试三种不同模式对大模型进行评测。全量测试针对动态测试数据库中所有题目进行遍历测试，其测试覆盖领域全面，但测试的成本高、周期长。并且如果大模型已经存在“刷题”等问题，部分测试题目已经失效，重复测试意义降低。模块测试即从动态测试数据库中按照特定维度选择特定测试题目进行评测，其针对

大模型的特定能力进行评估，测试方式较为灵活，定制化较强，但无法表征大模型的全面能力，在特定的业务需求下可以执行。

抽样测试即从题库中动态选择题目进行测试，该方法仅用少量有效数据即完成大模型的测试，避免很多无意义的测试过程，测试成本低、效率高、综合性强。如何从海量数据集中选择高质量测试数据是一个难点，产业界缺少成熟的方案。高质量评测数据集需要标准且量化的定义，如满足充分性、多样性、新颖性、区分度、合理性、有效性、追溯性等多项要求。通过定义函数的目标函数，量化制定每一个质量维度的权重，最终使用智能算法完成高质量测试数据集抽取。抽样算法可选择随机抽样或定向抽样，也可将数据集质量作为优化目标，使用演进类等优化算法反复迭代计算得到高质量测试数据集。由于每次使用的测试题目均不相同，使用抽样数据进行测试可以在一定程度上避免大模型“刷题”对测试结果的影响。

（三）“方升”大模型测试体系实践

为全面和深入认知大语言模型能力及其缺陷，跟踪国内外大语言模型发展态势，并验证“方升”大模型基准测试体系的有效性，中国信通院于 2024 年初启动“方升”首轮试评测，实际测试执行时间为 2 月 19 日至 2 月 29 日。本次评测基于“方升”测试体系，针对大模型的通用、行业、应用和安全能力进行全方位评测。被测对象为 30 多家国内外主流的闭源（商业）大模型和开源大模型，如 GPT-4、Qwen-72B- Chat、LLaMA2 等。本次评测向参测方提供大模型评测报告及提升建议，并展示少量的错误样例，以推动大语言模型健康发展。

来源：中国信息通信研究院

图 4 “方升”大模型首轮试评测模式

如图 14 所示，“方升”大模型首轮试评测提供综合评测和专项评测两种测试模式，其中综合评测是必测项目，专项评测是选测项目。综合测试主要针对大模型的通用、行业、应用和安全能力进行全面评估以衡量大模型的综合能力，专项评测则对于大模型的指定能力进行测试，例如面向特定行业或场景的定向评估。为了提升大模型测试效率，本次综合测试的评测数据集是通过动态测试数据库中定向抽取所构建，这种动态抽取题库的方式保证每次评测题目的新颖性，一定程度上可规避模型“刷榜”的问题。后续测试会增加抽样的次数以保证结果的客观性。专项评测主要对于参侧方所选择的评测维度进行全量精细化评测，以全面衡量大模型在该领域内的客观表现。

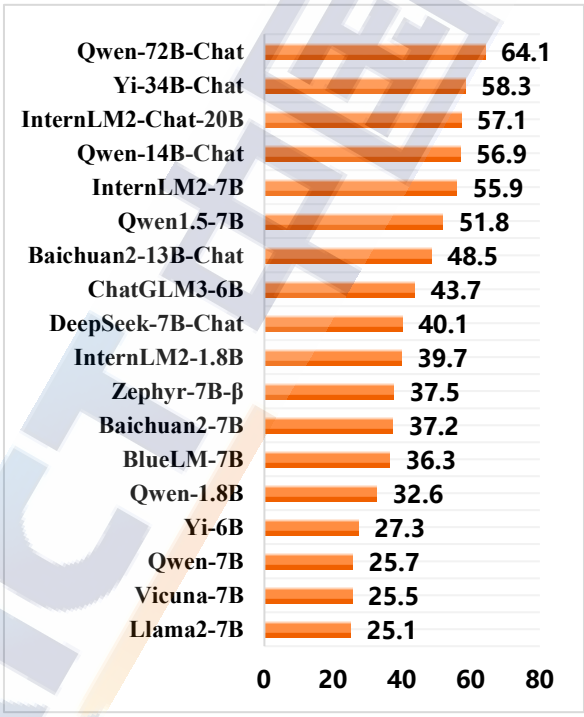
“方升”大模型首轮试评测中的综合评测题目从包含 203 万的评测数据库中定向抽取构建，全面覆盖通用、行业、应用、安全 4 个一级测试维度，通用、行业、应用的评测题目数量占比大约为 40%、40%

和 20%，安全采用 AI Safety Benchmark 专项测试。测试题型包含单选、多选、填空、判断、问答等多种形式，其中客观题占比大约为 90%。本次评测的高、中和低频测试题目的数量占比约为 10%、40%和 50%，其中低频测试题目为新构建的测试题目，从未用于大模型的评测，而中频测试题目为测试次数较少的题目。本次评测中文测试题目数大约为 70%，英文测试题目为 30%，即保证了国内的主要应用市场，也衡量大模型的多语言处理能力。从题目的难易度上来说，难题、中等题、简单题数目的比例为 30%、40%和 30%，题目难易的分级为利用 GPT4-Turbo 进行自动化识别得到。

本次评测被测对象为 30 家国内外主流的闭源（商业）大模型和开源大模型，其中闭源（商业）大模型 12 个，开源大模型 18 个。在 12 个闭源大模型中，除 GPT-4-Trubo 和 GPT-3.5-Turbo 外，10 个为国内商业大模型。所评测开源大模型中既包含国外大模型 LLaMA2、Vicuna、Zephyr，也包括国内的大模型 Qwen、ChatGLM、Baichuan、Yi、InternLM、DeepSeek、BlueLM 等。闭源商业大模型主要是通过 API 的模式参与测试。在选择开源大模型时，考虑了不同的参数量与版本发布时间。需要注意，本报告仅提供开源大模型的评测结果，且本次评测结果只能从特定维度来表征大模型的能力，不代表各大模型产品的全面能力，仅供研究分析使用。

图 15 展示了开源大模型的评测结果，从其可以发现开源大模型的表现除了依赖参数量，还与模型版本迭代时间相关。从排名来看，Qwen-72B-Chat、Yi-34B-Chat、InternLM2-Chat-20B、Qwen-14B -Chat

分别占据了前几名，余下的开源大模型参数量基本都小于 10B，这说明大模型的参数量在一定程度上影响大模型的表现。但大模型的能力不仅仅依赖于模型的参数量，还与训练技术和数据质量密切相关。大模型技术迭代速度快，往往两三个月内即有新版本出现，在模型参数量基本不变的情况下，新版本的大模型能力对比上一版本往往明显增强，有的甚至发生跨越式的提升，例如 Qwen-7B 在本次评测中仅为 25.8 分，但 Qwen1.5-7B 的分数却显著提升至 51.8 分。且发布时间较晚的 InternLM2-1.8B 和 Qwen-1.8B 在参数量大大减少的情况下，整体表现甚至优于部分参数量为 6B 或 7B 的大模型。



来源：中国信息通信研究院

图 15　开源大模型评测榜单结果

需要注意的是，本次评测结果仅从特定维度对大模型能力进行考察。在实际的商业应用选型中，并不能仅考虑能力这一个因素，还需要结合应用场景、部署成本、推理时延、自主可控、用户体验等其他

因素，通过综合决策选择最适配的大模型。例如，开源大模型 Qwen-72B-Chat 与 Qwen-1.8B 相比，Qwen-72B-Chat 在能力上的优势十分明显，但在部署成本上，Qwen-1.8B 的部署难度更低。除此之外，在选择商业大模型时，还要重点考虑价格以及是否支持私有化部署等因素。

四、总结与展望

伴随着大模型基准测试的蓬勃发展，针对大模型各个维度的测试方法如雨后春笋般出现。大模型基准测试不应该仅仅作为大模型研发的终点，以发布测试榜单为目的，更重要的是切实发现大模型问题，驱动大模型能力的提升，指导大模型的研究方向和应用路线。因此，产学研各界应该在探索新的测试方法、构建自动化测试平台以及共享高质量评测数据集等方面协同发力。未来，对 AGI 进行全方位、科学化的评估，将成为人工智能领域亟待解决的重要问题。

（一）形成面向产业应用的大模型评测体系

随着人工智能技术的不断发展，大模型的应用日益广泛，为各行各业带来了巨大的变革和可能性。在金融、医疗、法律、交通、教育等各个领域，大模型展现出了巨大的应用潜力，有望提升工作效率，优化应用效果。此外，基于大模型的 AI 原生应用也逐渐进入人们的视野，大模型不仅能完成智能客服、知识管理、数据分析等简单任务，还可借助外部工具助力人类进行网络购物、旅行规划、餐馆预定等复杂活动。然而，由于当前产业应用数据大多在行业用户的手中。因此，虽然行业测评基准已初步建立，但面向大模型应用评测的评测数据集仍较为缺乏。

随着“人工智能+”行动的开展，各行业将以大模型实际落地的效果为评估目标，形成不同行业和应用效果评估的体系和方法论，积极建立面向产业场景化应用的评测数据集，探索面向行业和场景化应用的新型评测方法，切实推动大模型基准测试在行业场景中进行落地，

全面正向驱动大模型的发展与应用。

（二）构建超自动化的大模型基准测试平台

大模型基准测试的流程包括测试需求分析、测试环境准备、测试数据构建、基准测试执行、测试结果评估、测试结果展示等。其中，测试数据准备和测试结果评估这两步均需要投入大量人力，工作繁琐。并且，大基准测试执行可通过单点、分布式等方式进行，不同的硬件环境将直接影响模型的评测效率。由于评测结果会直接指引下一步研发方向，因此基准测试的自动化、工程化和批量化处理非常关键，可直接决定大模型整体的迭代效率。如何全自动化地完成大模型的测试、快速挖掘大模型缺陷、降低测试人力的投入是该领域值得深入研究的问题。

基准测试不应该仅作为 AI 应用开发的终点，而是要成为一个新起点，驱动大模型的能力持续提升。未来将会出现企业级的自动化大模型基准测试平台，保证从测试需求分析到测试结果统计的全流程质量把控。其不仅需要具备测试任务高效分发、分布式批量执行、测试结果自动统计等基础功能，还应该支持流程中的测试数据构建和测试结果评估等工作。例如，当前自动生成的测试数据质量很难保证，需要人工进行复核，上述操作可以在平台页面上完成。测试平台中可以集成已训练好的“裁判”大模型，助力大模型生成内容的正确性评估，降低评估的人力成本。

（三）探索 AGI 等先进人工智能的评测技术

人工智能技术发展迅速，大模型、RAG、AGENT、具身智能、AGI 等新概念和新技术层出不穷。**大模型基准测试作为研究较为深入的领域，将带动其他新技术的研究。**当前虽然 AGI 仍未有明确的定义，但针对 AGI 的探索性评测研究已有初步成果。例如微软发布论文《通用人工智能的火花：GPT-4 的早期实验》，通过数学、编程、视觉、医学、法律、心理学等复杂度较高的任务证明 GPT-4 已经进入 AGI 的早期阶段。北京通用人工智能研究院发布《通智测试：通用人工智能具身物理与社会测试评级系统》，提出一种基于能力和价值维度的 AGI 的评测方法。中国科学院和美国俄亥俄州立大学等先后推出 AGIBench 和 MMMU 评测数据集，从多模态、多学科、多粒度等维度衡量大模型距离 AGI 的差距。虽然当前 AGI 的发展仍然处于初期阶段，但通过基准测试的研究，可以为未来 AGI 的发展方向提供思路，并对 AGI 的能力进行监控以指引其正向发展。

附录

附表 1 语言大模型通用能力的代表性评测数据集

基准名称	评测目标	国家	时间	题目类型
MMLU	理解、知识	美国	2021	客观
C-Eval	理解、知识	中国	2023	客观
CMMLU	理解、知识	中国	2023	客观
MT-Bench	生成（对话）	美国	2022	主观
MT-Bench-101	生成（对话）	中国	2024	主观
AlpacaEval	生成（对话）	美国	2023	主观
Lmsys-chat-1m	生成（对话）	美国	2023	主观
DialogSum	生成（摘要）	中国	2021	主观
LCSTS	生成（摘要）	中国	2015	主观
StoryCloze	推理能力	美国	2016	客观
BBH	推理能力	美国	2022	客观
GSM8K	推理能力	美国	2021	客观
CMATH	推理能力	中国	2023	客观
MATHVISTA	推理能力	中国	2023	客观
AGIEval	知识能力	美国	2023	客观
KoLA	知识能力	中国	2023	主观
SOCKET	知识能力	美国	2023	主观
GAOKAO	学科能力	中国	2023	主观/客观
M3Exam	学科能力	中国	2024	主观/客观
M3KE	学科能力	中国	2023	客观
XTREME	多语言	美国	2020	主观
MEGA	多语言	美国	2023	主观
L-EVAL	长文本	中国	2023	主观
LongBench	长文本	中国	2023	主观
CharacterEval	角色扮演	中国	2023	主观/客观

ToolQA	工具使用	美国	2023	主观
TruthfulQA	可靠性	英国	2022	主观/客观
UHGEval	可靠性	中国	2023	主观/客观
PromptBench	鲁棒性	中国	2023	主观

附表 2 语言大模型行业能力的代表性评测数据集

基准名称	行业	国家	时间	题目类型
PIXIU	金融	中国	2023	主观/客观
FinEval	金融	中国	2023	主观/客观
FINANCEBENCH	金融	美国	2023	主观
PubMedQA	医疗	美国	2019	主观/客观
MedQA	医疗	美国	2021	主观
CMExam	医疗	中国	2023	主观/客观
JEC-QA	法律	中国	2020	主观/客观
CUAD	法律	美国	2021	主观
LAiW	法律	中国	2023	主观
LegalBench	法律	美国	2023	主观/客观
DevOps-Eval	软件	中国	2023	客观
LogBench	软件	中国	2023	客观
OpsEval	软件	中国	2023	主观/客观
SciEval	科研	中国	2024	主观/客观
SCIBENCH	科研	美国	2023	客观
SciQA	科研	德国	2023	客观
ChemLLMBench	科研	美国	2023	客观
NetEval	通信	中国	2023	客观
TeleQnA	通信	中国	2023	客观
CGAEval	政务	中国	2023	主观/客观
NuclearQA	能源	美国	2023	主观
CloudEval-YAML	互联网	中国	2023	主观

MSQA	互联网	美国	2023	主观
battery-device-data-qa	工业	英国	2023	主观
GameEval	游戏	中国	2023	主观
AvalonBench	游戏	美国	2023	主观

附表 3 语言大模型应用能力的代表性评测数据集

基准名称	应用场景	国家	时间	题目类型
GAIA	智能助手	美国	2023	主观/客观
CFBenchmark	智能助手	中国	2023	主观/客观
RGB	知识管理	中国	2023	主观
CRUD-RAG	知识管理	中国	2024	主观
MMC-Benchmark	数据分析	美国	2023	客观
QTSUMM	数据分析	美国	2023	主观/客观
TableQAEval	数据分析	中国	2023	主观
MBPP	代码助手	美国	2021	主观
APPS	代码助手	美国	2021	主观
HumanEval	代码助手	美国	2021	主观
WikiSQL	代码助手	美国	2017	主观
VGEN	代码助手	美国	2023	主观
VerilogEval	代码助手	美国	2023	主观
AgentBench	AI 智能体	中国	2023	主观
AgentSims	AI 智能体	中国	2023	主观/客观
BOLAA	AI 智能体	中国	2023	主观
TELeR	AI 智能体	美国	2023	主观
SQA3D	具身智能	中国	2022	主观
BEHAVIOR-1K	具身智能	美国	2023	主观
ALFRED	具身智能	美国	2023	主观

附表 4 语言大模型安全能力的代表性评测数据集

基准名称	评测目标	国家	时间	题目类型
DECODINGTRUST	综合安全	美国	2023	主观
Safety-Prompts	综合安全	中国	2023	主观
TRUSTGPT	综合安全	中国	2023	主观
SafetyBench	综合安全	中国	2023	主观
TOXIGEN	内容安全	美国	2022	主观
CPAD	内容安全	中国	2023	主观
JADE	内容安全	中国	2023	主观
Do-Not-Answer	内容安全	阿联酋	2023	主观
CVALUES	伦理安全	中国	2023	主观
ETHICS	伦理安全	美国	2020	主观
BBQ	伦理安全	美国	2021	主观
DialogueSafety	伦理安全	中国	2023	主观
CONFAIDE	隐私安全	美国	2023	客观
R-Judge	模型安全	中国	2024	主观

附表 5 多模态大模型通用能力的代表性评测数据集

基准名称	评测目标	国家	时间	题目类型
MME	综合能力	中国	2023	主观/客观
MMBench	综合能力	中国	2023	主观/客观
SEED-Bench	综合能力	中国	2023	主观/客观
LVLM-eHub	综合能力	中国	2023	主观/客观
OwlEval	综合能力	中国	2023	主观/客观
MM-Vet	综合能力	新加坡	2023	主观/客观
TouchStone	综合能力	中国	2023	主观/客观
LLaVA-Bench	综合能力	美国	2023	主观/客观
VQA	视觉问答	美国	2015	主观/客观

OK-VQA	视觉问答	美国	2019	主观/客观
SCIGRAPHQA	视觉问答	美国	2023	主观/客观
CORE-MM	视觉推理	中国	2023	主观/客观
CONTEXTUAL	视觉推理	美国	2024	主观/客观
Mementos	视觉推理	美国	2024	主观/客观
OCRBench	视觉处理	中国	2023	主观/客观
Q-Bench	视觉处理	新加坡	2023	主观/客观
T2I-CompBench	图像生成	香港	2023	主观/客观
HRS-Bench	图像生成	沙特	2023	主观/客观
POPE	可靠性	中国	2023	主观
AMBER	可靠性	中国	2023	主观

缩略语

AI	Artificial Intelligence	人工智能
AGI	Artificial General Intelligence	通用人工智能
GPU	Graphics Processing Unit	图形处理器
API	Application Programming Interface	应用程序编程接口
GPT	Generative Pre-trained Transformer	生成式预训练变换器
NLP	Natural Language Processing	自然语言处理
HELM	Holistic Evaluation of Language Models	语言模型整体评估
SOTA	State-Of-The-Art	领域最佳性能
RAG	Retrieval Augmented Generation	检索增强生成

参考文献

1.WX Zhao, K Zhou, J Li, T Tang, X Wang, Y Hou, et al. A survey of large language models. arXiv:2303.18223, 2023.

2.张奇，桂韬，郑锐，黄萱菁. 大规模语言模型从理论到实践. 中国工信出版集团, 2023.

3.Y Chang, et al. A survey on evaluation of large language models [J]. ACM Transactions on Intelligent Systems and Technology, 2023, 15(3):1-45.

4.Z Guo, R Jin, C Liu, Y Huang, D Shi, L Yu, Y Liu, J Li, B Xiong, D Xiong. Evaluating large language models: A comprehensive survey. arXiv:2310.19736, 2023.

5.罗文，王厚峰. 大语言模型评测综述 [J]. 中文信息学报, 2024, 38(1):1-23.

6.D Hendrycks, C Burns, S Basart, A Zou, M Mazeika, D Song, J Steinhardt. Measuring Massive Multitask Language Understanding [C]. International Conference on Learning Representations (ICLR), 2020.

7.Y Huang, Y Bai, et al. C-eval: A multi-level multi-discipline chinese evaluation suite for foundation models [C]. Advances in Neural Information Processing Systems (NeurIPS), 2024.

8.T Zhang, F Ladhak, E Durmus, P Liang, et al. Benchmarking large language models for news summarization [J]. Transactions of the Association for Computational Linguistics, 2024, 12:39-57.

9.K Zhu, J Chen, J Wang, NZ Gong, D Yang, X Xie. Dyval: Graph-informed dynamic evaluation of large language models. arXiv:2309.17167, 2023.

10.CH Chiang, H Lee. Can large language models be an alternative to human evaluations? arXiv:2305.01937, 2023.

11.C Li, Z Gan, Z Yang, J Yang, L Li, L Wang, J Gao. Multimodal Foundation Models: From Specialists to General-Purpose Assistants. arXiv:2309.10020, 2023.

12.T Lee, M Yasunaga, C Meng, Y Mai, JS Park, A Gupta, Y Zhang, D Narayanan, H Teufel. Holistic Evaluation of Text-to-Image Models [C]. Advances in Neural Information Processing Systems (NeurIPS), 2024.

13.Y Liu, H Duan, Y Zhang, B Li, S Zhang, W Zhao, et al. Mmbench: Is your multi-modal model an all-around player? . arxiv:2307.06281, 2023.

14.X Liu, H Yu, H Zhang, et al. Agentbench: Evaluating llms as agents. arxiv:2308.03688, 2023.

15.Q Xie, W Han, X Zhang, Y Lai, M Peng, A Lopez-Lira, J Huang. Pixiu: A large language model, instruction data and evaluation benchmark for finance. arxiv:2306.05443, 2023.

16.L Zhang, W Cai, Z Liu, et al. Fineval: A chinese financial domain knowledge evaluation benchmark for large language models. arxiv:2308.09975, 2023.

17.Z Fei, X Shen, D Zhu, F Zhou, Z Han, S Zhang, K Chen, Z Shen, J Ge. Lawbench: Benchmarking legal knowledge of large language models .arxiv:2309.16289, 2023.

18.J Chen, H Lin, X Han, L Sun. Benchmarking large language models in retrieval-augmented generation [C]. Proceedings of the AAAI Conference on Artificial Intelligence (AAAI), 2024.

19.X Liu, X Lei, S Wang, et al. Alignbench: Benchmarking chinese alignment of large language models. arXiv:2311.18743, 2023.

20.Y Zhuang, Q Liu, Y Ning, et al. Efficiently measuring the cognitive ability of llms: An adaptive testing perspective. arxiv:2306.10512, 2023.

21.X Zhang, B Yu, H Yu, Y Lv, T Liu, F Huang, H Xu, Y Li. Wider and deeper llm networks are fairer llm evaluators. arxiv:2308.01862, 2023.

编制说明

本研究报告自 2023 年 12 月启动编制，分为前期研究、框架设计、文稿起草、征求意见和修改完善五个阶段，面向大模型基准测试的技术供应方和服务应用方开展了深度的调研等工作。本报告由中国信息通信研究院人工智能研究所撰写，撰写过程中得到了人工智能关键技术和应用评测工业和信息化部重点实验室的大力支持。

参编单位：中国科学院大学、中国科学院软件研究所、北京智源人工智能研究院、天津大学、北京邮电大学、北京交通大学、中国移动通信集团有限公司、中国电信集团有限公司、中国联合网络通信集团有限公司、广州数据集团有限公司、航天信息股份有限公司、煤炭科学研究总院、华为云计算技术有限公司、百度云计算技术有限公司、腾讯计算机系统有限公司、阿里云计算有限公司、科大讯飞股份有限公司、浪潮通信信息系统有限公司、荣耀终端有限公司、蚂蚁科技集团股份有限公司、北京海天瑞声科技股份有限公司、东方财富信息股份有限公司、甲骨易语言科技股份有限公司、远光软件股份有限公司、南京新一代人工智能研究院。

中国信息通信研究院 人工智能研究所

地址：北京市海淀区花园北路 52 号

邮编：100191

电话：010-62301618

传真：010-62301618

网址：www.caict.ac.cn