奕衡
YI HENG

『奕衡』通用大模型评测体系白皮书

中国移动研究院
中国移动技术能力评测中心

目 录

前 言

人工智能大模型（以下简称大模型）是实现生成式人工智能服务（AIGC）的重要技术，ChatGPT上线两个月活跃用户（MAU）突破 1 亿，激发了大模型技术的爆发式发展，全球科技公司开启大模型“科技竞赛”。国外科技巨头微软、谷歌、META等，加快大模型研发，并迅速应用到搜索、办公、音乐、视频等领域。我国头部企业积极开展自主可控的大模型研发，百度、腾讯、华为、阿里、中科院自动化所、智谱AI、科大讯飞等公司的大模型也相继推向市场。各家公司也加快大模型的迭代升级速度，OpenAI、谷歌、百度已经在短短几个月内多次升级大模型版本，能力提升明显，大模型行业竞争激烈。

与此同时，随着大模型评测需求逐渐增加,相关研究也进一步深入。大模型相比传统模型，泛化能力更强、灵活性更高、适应性更广，多任务、多场景，评测维度、评测指标和数据集更复杂，面向大模型的评估方法、评测基准、测试集成为新的研究课题。业界头部公司、主流科研机构和重点高校等权威组织，如OpenAI、微软、斯坦福大学、信通院，在评测框架、评测指标、数据构建方法等方面发表了一些论文和研究报告，从准确性、鲁棒性、毒害性、公平性等评测维度对相关大模型进行了评测，为用户和行业充分掌握大模型能力发挥了积极作用。

目前业界多家机构发布了大模型的评测榜单，但是评测维度及侧重点各有不同。从推动AI大模型成熟应用、促进生态繁荣、指引产业优化方向的角度，有必要从用户视角，构建一套客观全面、公平公正的大模型评测体系。

中国移动技术能力评测中心作为中国移动的专业评测机构，也在关注和跟进大模型评测技术发展。自 2019 年起陆续开展了专业公司 31+N考核对标评测、技术中台能力准入等工作，涵盖人工智能、互联网、物联网、大数据、大视频等 20 余个领域 1000 余项产品和能力，积累了丰富的产品技术能力评测经验和数据。基于前期积累，对业界各类大模型评测技术进行了充分调研和评测验证，构建了“弈衡”通用大模型“2-4-6”评测体系，并基于该体系对已发布的大模型进行了广泛的评测。

随着大模型技术的不断发展，“弈衡”通用大模型评测体系也将持续迭代完善，希望通过发布《“弈衡”通用大模型评测体系白皮书》，与产业界相关企业和研究机构一道，加强交流合作，逐步完善测试指标、测试方法、测试数据、测试自动工具，共同建立评测产业标准化生态，为业界大模型评测提供参考依据，促进大模型技术的产业成熟和应用落地。

01 大模型评测背景

1.1 大模型发展现状

随着大模型技术的快速发展，其巨大的参数量、计算量以及模型复杂度，在解决复杂任务方面具有很大的优势，主要体现在强大的理解和生成能力、高度的泛化能力、优秀的可迁移学习特性及端到端训练优势。大模型技术受到各类行业的广泛关注，通过将大模型与实际业务相结合，可为用户提供更加个性化、更符合用户需求的服务。大模型在多个领域的应用示例如下：

行业	领域	应用
通用能力	搜索领域	用于实现更智能、更准确的信息检索和推荐。
	语音识别与合成领域	识别并合成语音，实现更智能、更自然的语音助手。
垂直行业	内容创作与审核领域	用于自动撰写文章、新闻、绘画、音乐等任务。
	教育科技领域	为教育领域提供智能化支持。
	金融科技领域	帮助金融机构提高决策效率和质量。
	医疗健康领域	协助医生和研究人员提高工作效率，提高医疗水平。
	智能制造领域	助力工厂实现智能化生产、降本增效。
	软件开发领域	提高开发人员的工作效率，降低人力成本。
	法律领域	用于文书的撰写、法律咨询等任务，降低法律服务成本。
	人力资源领域	帮助企业优化人力资源管理。
	媒体与娱乐领域	为创作者提供创意灵感，提高创作效率。
	语言学习领域	辅助语言教师授课，帮助学习者提高语言能力。
	旅游领域	提供个性化的旅行建议和服务。
	公共服务领域	提高政府服务效率，优化公共资源配置。
	客服领域	应用于智能客服助手等任务，提高客服效率，降低成本。
	市场分析领域	帮助企业洞察市场动态，优化产品、提供更加安全的服务。

随着大模型的发展，模型能力还将不断扩展，通过文本、图像和语音等多种形式。与更多新兴的应用场景相结合，赋能千行百业。

1.2 评测需求

由于大模型高度复杂的结构，如何对其进行全面、客观的评测成为了一个亟待解决的问题。与传统AI模型单一的应用领域相比，大模型在多任务和多领域方面展现出卓越的性能和泛化能力。因此，针对大模型产品的评估通常需结合多种不同任务，从多个维度展开综合评价。在现阶段的研究与实践中，大模型评测的主要需求包括但不限于以下几类：

文本类大模型：此类模型需要能够依据提示创作符合需求的文本内容，并依赖知识和文本逻辑，推理并回答用户问题。在文本生成任务中，主要考察模型生成内容是否满足使用者要求，并具备正确性、流畅性、规范性和逻辑性；在知识应用任务中，则需要模型覆盖尽可能多的领域，并具备一定深度，同时还应具备对知识的理解与运用能力；在推理任务中，还需对模型生成内容是否符合人类思维的判断、推理过程质量、推理过程与答案一致性、数值计算正确性等指标进行评估。

图像类大模型：此类模型需要识别并定位图像中的各种物体，对其进行分类，并将不同对象或区域分割开来，在此基础上，通常还要求模型根据给定的描述生成新的图像。在图像分类任务中，核心指标包括分类的准确性、鲁棒性及对新类别的泛化能力；物体检测任务更关注对复杂场景的处理能力和检测的准确率、覆盖率；图像分割任务更能体现模型对细节的处理能力；对于图像生成任务，对于图像质量和创新型的评测需要更综合的评测方法。

语音类大模型：此类模型需要能够识别多种人类语音，实现文本和语音的双向转化。在语音识别任务中，需要评估模型是否能够准确、高效地将人类语音转化为文字表达，关注模型括识别准确率、噪声抑制效果、多语种处理能力等；在语音合成任务中更关注合成语音的括语音质量、语音流畅度、音韵准确性等。

除上述几类模型中的评测需求之外，针对模型及产品的各项能力，还需探究大模型生成结果的置信度、训练数据与生成结果的一致性、对生成内容的规划能力、噪声和扰动下的稳定性、对于提示词的敏感性等传统NLP、CV及语音任务涉及较少的评测指标，形成更为标准化和通用的解决办法。

大模型评测对于推动人工智能技术的发展具有重要的意义。一方面，通过对大模型性能的评测，可以为模型优化和改进提供有力依据，从而提高其应用效果和商业价值。另一方面，大模型评测可以了解大模型在不同行业的性能和适用性，促进人工智能技术在各行业的发展和应用。此外，大模型评测还可以促进不同领域研究者的技术交流和合作，推动人工智能技术的共同发展。

1.3 评测问题与挑战

技术发展日新月异，大模型评测需要与时俱进。随着人工智能领域的飞速发展，评测难度也在逐渐增加。为了保障评测针对性和有效性，需要不断更新评测标准和方法。

首先，大模型复杂性对评测提出挑战。

随着人工智能不断发展，大模型复杂性不断增长，评测需求多样性更加显著。大模型涉及到文本生成、问答系统、知识图谱、图像创作、语音生成等多个任务领域。如文章写作任务中，模型的生成质量是重要指标之一，需要考虑到文本是否自然、流畅，是否符合语言规范，是否有语法错误等。而图片创作任务中，图片的视觉效果、清晰度、色彩鲜艳度等是评估模型性能的重要指标。面对以上问题，需要制定一套更为全面的评测体系，以全面评价模型能力。

其次，大模型泛化性对评测提出更高要求。

大模型在很多任务上已经达到或超过了人类的水平，但在某些特定领域中，它们的性能仍然有待提高。对于低资源任务，评测者需要关注模型在使用少量语料时的表现，需要考虑到语言之间的差异性和复杂性，以便更好地评估模型在不同场景下的泛化能力。对于专业领域任务，需要关注模型对领域特定术语、概念和规则的理解和应用，使用更广泛的数据集和跨领域的评测任务，以确保评测结果具有泛化性和可靠性。

再者，大模型安全性也需要重点考虑。

数字化时代，攻击者可能会利用特定数据来攻击模型，或者破坏模型的性能。对抗性攻击是一种常见的攻击类型，通过向模型输入有意制造的数据或恶意样本来欺骗模型或破坏模型的性能。对抗性样本可以模拟现实世界中的攻击。如图像分类任务，针对正确分类的样本，可以通过添加一些扰动来生成对抗性样本，导致模型对其错误分类。面对以上问题，需要考虑各种攻击模型，并设计相应任务来评估模型安全性。

总之，随着大模型的不断发展和应用，评测工作所面临的挑战逐渐增加。需要重点考虑多样性、普适性、客观性和公正性等评测需求，充分评估大模型的性能和潜力，为大模型技术的进一步发展提供支持。