一种针对垂类模型的综合成效评测框架*

宋　元[1]，张　衍[1,2]，任熠辉[1]，黄晓鹏[1]

（1. 苏州市人工智能有限公司，江苏　苏州　215100；2. 苏州国际发展集团有限公司，江苏　苏州　215007）

摘　要：针对垂类模型在评测实践中存在的评价维度单一、缺乏领域适配性以及方法碎片化等问题，提出了一套综合成效评测框架。该研究旨在通过标准化方案解决技术研发与产业应用之间的“评价断层”，为垂类模型的开发、部署和监管提供科学依据。研究方法包括构建以安全合规、技术性能和应用价值为核心的多维指标体系，并配套设计评测数据集构建策略与混合评测方法，后者融合了自动化测试、人工评估和大模型作为裁判的评估手段。研究结果形成了一套结构化的评测体系，涵盖评价对象分类、指标定义和方法实施，能够实现对不同类型垂类模型的全面、可比较评估。结论表明，该框架有助于提升评测的客观性和可操作性，推动垂类模型在关键领域的可信赖应用，未来需通过实践验证和动态优化以适应技术发展。

关键词：人工智能；垂类模型；模型评测

中图分类号：TP391. 1　　**文献标识码：**A　　**DOI：**10. 19358/j. issn. 2097 - 1788. 2025. 11. 004

引用格式：宋元，张衍，任熠辉，等. 一种针对垂类模型的综合成效评测框架［J］. 网络安全与数据治理，2025，44（11）：18 - 23，29.

A comprehensive effectiveness evaluation framework for domain-specific models

Song Yuan[1]，Zhang Kan[1,2]，Ren Yihui[1]，Huang Xiaopeng[1]

（1. Suzhou Artificial Intelligence Co. , Ltd. , Suzhou 215000，China；

2. Suzhou International Development Group Co. , Ltd. , Suzhou 215007，China）

Abstract：This paper addresses the issues of single evaluation dimensions，lack of domain adaptability，and fragmented methods in the evaluation practice of domain-specific models，and proposes a comprehensive effectiveness evaluation framework. This study aims to address the "evaluation gap" between technology research and development and industrial application through standardized solutions，providing a scientific basis for the development，deployment，and supervision of domain-specific models. The research method includes constructing a multidimensional indicator system centered on security compliance，technical performance，and application value，and designing a supporting evaluation dataset construction strategy and a hybrid evaluation method. The latter integrates automated testing，manual evaluation，and large models as evaluation means. The research results form a structured evaluation system that covers the classification of evaluation objects，indicator definition，and method implementation，which can achieve a comprehensive and comparable evaluation of different types of domain-specific models. The conclusion shows that the framework helps to improve the objectivity and operability of the evaluation and promote the trustworthy application of domain-specific models in key areas. In the future，it will need to be verified in practice and dynamically optimized to adapt to technological development.

Key words：artificial intelligence；domain-specific model；model evaluation

0　引言

以大模型为核心的人工智能技术正加速重构全球产业格局，成为驱动新质生产力发展、推动经济社会高质量转型的关键引擎。相较于通用性基础大模型，面向特定行业、领域或场景的垂类模型正凭借其对专业需求的深度适配性，在制造、医疗、金融、政务、农业等关键领域实现落地。例如，工业垂类模型可优化生产流程的故障诊断效率[1]，医疗垂类模型能辅助临床影像的精准

*基金项目：中国博士后科学基金（2024M762322）

识别[2]，政务智能体系统可提升公共服务的响应速度[3]。然而，随着垂类模型应用场景的多元化与技术架构的复杂化，行业内对其成效的评价仍缺乏统一、系统的标准体系，导致技术研发与产业应用之间存在“评价断层”。

当前针对模型评价实践中，存在三方面核心问题。其一，评价维度单一化，多数研究仅聚焦技术性能，如响应速度、准确率，忽视了安全合规的前置性要求与实际应用场景中的价值转化能力，难以全面反映模型的综合成效[4]；其二，评价对象同质化，未针对各领域间的差异化特征设计适配的评价指标，导致评价结果对不同类型模型的指导性不足；其三，评价方法碎片化，部分评价依赖主观经验判断，缺乏标准化的数据集构建规范与量化计算逻辑，难以保证评价结果的客观性与可复现性[5]。这些问题不仅制约了垂类模型技术迭代的方向，也为产业界选择适配模型，政府部门开展监管、引导与奖励带来了困难。

本文提出了一套垂类模型综合成效评价框架，首先明确评价对象的分类标准与准入条件，随后构建以安全合规、技术性能、应用价值为基础的三大维度评价指标体系。同时，框架配套设计了标准化的评价方法，实现对不同类型垂类模型成效的精准、可比评价。

1　评测框架现状

近年来，随着大模型在通用及垂直场景中的广泛落地，评测体系的关注点已从单一任务精度逐步转向“多维度—多场景”的综合评估范式。以 HELM[6] 为代表的研究提出将模型能力、可信赖性、鲁棒性、公平性、安全性以及效率等多类指标纳入统一评估框架，强调应在具体应用情境下对模型进行密集化、多角度的系统对比，以全面揭示模型在不同指标间的权衡关系与能力盲区。HELM 所倡导的“情境化多指标评估”理念，已成为评估语言模型透明性与可靠性的重要参考。

在理论框架持续完善的同时，工程化的评测工具链也日趋成熟，各类模型评测框架更是加速涌现。EleutherAI 的 lm-evaluation-harness[7]、HuggingFace 的 LightEval[8] 及 Open LLM Leaderboard[9] 等工具，为研究与实践提供了可复现、自动化的评测流水线支持，使同一测试集能在不同模型与后处理流程下进行一致性比较。在垂直领域方面，社区已积累了一批高质量的专业基准。例如，在医疗领域，MultiMedQA[10] 等综合问答集被广泛用于评估模型在医学知识理解、专科问答和患者导诊等任务上的表现；在金融领域，FinQA[11] 等数据集则聚焦财报解析与复杂数值推理。它们通常由领域专家参与构建，并融合可解释的推理链或金标准参考步骤。这类垂类基准的出现，使得比较通用大模型与领域微调模型在专业任务上的性能差异成为可能。

尽管取得上述进展，当前垂类模型评测体系仍存在若干明显短板。其一，跨领域一致性指标体系尚未建立，不同行业基准所侧重的能力维度差异显著，导致模型结果难以在同一尺度上进行跨领域横向对比[12]；其二，自动化评测指标尚无法完全替代人工评估[13]，在医疗诊断、法律咨询、投资建议等高风险场景中，输出内容的正确性高度依赖细粒度事实核验、推理逻辑严谨性与合规审查，而自动评分方法易忽略语义细节或受对抗性提示干扰；其三，数据污染、测试泄露与基准覆盖偏差等问题依然普遍，部分模型在公开排行榜上的高分可能无法反映其真实泛化能力；其四，针对模型置信度、幻觉现象及可解释性等安全相关属性的量化标准仍不统一，现有方法多依赖启发式检测与人工复审，缺乏系统性的度量支持。

综上所述，仅依靠现有通用评估框架难以充分满足垂直领域对可靠性、合规性与专业性的高要求。在此背景下，构建一套面向垂直场景，统筹准确性、可解释性、安全性、价值成效与工程可行性的系统化评测框架，已成为推动垂类模型切实落地的关键环节。基于这一现实需求，本文在既有研究基础上提出一套垂类模型综合成效评测框架，以期为该领域提供更加完整、可操作的解决方案。

2　概念与特征

当前，产业界正积极推动人工智能垂类模型的研发与设计，致力于采用最新技术手段提升模型性能，并探索其在各细分领域中的应用潜力。研究表明，面向特定领域的预训练语言模型在专业任务上的表现显著优于通用模型[14]。然而，目前关于人工智能垂类模型的内涵界定与能力特征，仍有待进一步明确。

2.1　垂类模型的概念内涵

当前学界与业界对于垂类模型的边界尚未形成统一共识。大部分研究将其笼统视为面向特定领域的模型，而另一些研究则更强调其在任务层面上的适配能力。总体而言，垂类模型的分类仍存在一定模糊性与交叉性，尤其是在行业层面与业务层面的划分上，常常存在重叠。为便于研究与讨论，本文尝试在现有基础上，对垂类模型进行系统梳理，并提出三类较为典型的形态：行业大模型、垂域大模型与智能体系统。

行业大模型依托于通用基础大模型，通过注入特定行业数据进行微调，从而形成面向制造、医疗、金融、

交通、文旅、建筑、政务服务、教育、农业、能源与消费等多个典型行业的专用模型。其显著特征在于对行业核心场景的广泛覆盖以及对共性需求的适配能力。领域自适应预训练可以显著提升模型在特定领域的表现[15]，在实际应用中，行业大模型在特定领域的基准任务上表现出优于通用基础模型的性能，成为支撑行业智能化升级的重要支点。

在行业大模型的进一步细化过程中，垂域大模型逐渐凸显出差异化优势。垂域大模型通常针对具体的下游业务场景开展深度优化，其训练过程不仅依赖于行业大模型的通用能力，还整合了领域规则、知识图谱等结构化知识资源，以提升对细分场景的适配性。与行业大模型相比，垂域大模型的数据规模相对有限，但其突出的特征在于对特定业务需求的高度专注性与精细化表达能力。这类模型能够在细分任务中展现出较高的性能稳定性和针对性，是推动行业智能化由通用领域向特定领域发展的关键。

基于大模型及规则驱动的智能体系统，为垂类模型的应用拓展提供了新的实现路径。智能体系统不仅具备语言与知识的表达能力，更融合了环境感知、自主决策与行动执行等多维能力。通过多步骤的环境交互，如 API 调用与工具操作，智能体能够完成复杂任务，从而突破传统模型在静态推理与单一任务上的局限[16]。

2.2 垂类模型的能力特征

垂类大模型在通用大模型的基础能力上，针对特定行业与场景进行了深度优化，它不仅提升了模型在专业领域的应用价值，也推动了人工智能在产业链中的深层次融合[17]。与通用大模型相比，垂类模型更强调专业化、合规性、更新性与集成性，能够为行业提供更具针对性的智能化支持。具体来看，面向高质量应用要求的垂类大模型通常需具备以下能力特征：

（1）场景应用的专业性

垂类大模型需根据行业及细分业务场景的特征进行优化，确保模型能够识别并适配复杂的场景需求。例如，在医疗领域，模型应突出诊疗数据解读与临床知识推理；在金融领域，则应着重于风险评估、市场预测与合规审查。通过这种专业化适配，垂类模型能够在特定场景中展现出更高的精准度和可靠性。

（2）内容输出的合规性

垂类模型的应用往往涉及敏感信息和严格的监管要求。因此，其输出必须符合行业规范与伦理标准，确保生成结果的安全性与可控性。例如，在金融与政务等场景中，模型应具备对违规请求的过滤能力，并通过审查机制保证内容的可解释性与可追溯性。只有在合规性得到保障的前提下，垂类模型才能在实际业务中被广泛采纳。

（3）知识更新的持续性

由于行业知识迭代速度较快，垂类模型需具备动态更新能力，不断学习新数据与新规则，以保证知识库的有效性与时效性。例如，医疗模型需快速吸收最新临床指南，交通与能源模型则需同步最新的政策与技术标准。持续的知识更新不仅能降低因知识陈旧带来的风险，也能使模型在快速变化的行业环境中保持竞争力。

（4）技术应用的集成性

垂类模型需具备良好的开放性与扩展性，能够与外部工具、知识库及行业平台进行深度集成。例如，智能体系统可通过 API 调用完成跨平台任务，行业大模型可与知识图谱结合提升决策准确度。通过提供统一接口与插件化支持，垂类模型能够实现多源数据与多种技术的互联互通，从而催生更丰富的行业应用生态。

3 体系框架

如图 1 所示，本文所提框架围绕安全合规、技术性能、应用能力与价值表现四个核心维度构建评测体系。这四个模块共同反映模型在行业场景中的综合表现，为模型的开发、部署与优化提供科学依据。在设计中，每一模块独立衡量其核心目标，同时通过标准化指标体系实现模块间的协同评价，从而确保评估结果能够全面反映模型的功能性能与实际应用价值。

安全合规模块聚焦模型输出内容的合法性、可靠性与可控性，核心任务在于评估模型输出是否存在危害国家安全、宣扬暴力恐怖、传播低俗信息或其他违法违规内容，以及其是否符合国家法律法规与行业标准[18]。该模块不仅关注模型单次输出的合规性，还涵盖对连续交互中潜在风险的识别能力、内容过滤机制的有效性，以及对恶意或不当指令的处置能力。安全合规是模型在行业环境中可靠部署的基础保障，也是其在复杂任务与动态场景下安全应用的重要前提。

技术性能模块用于衡量模型在运行环境中的响应效率与资源消耗，反映其在实际业务场景中高效稳定运行的能力。该模块包括响应效能与资源效率两个方面，响应效能主要考察模型对用户请求的即时响应能力，具体包括平均响应时长、首词元响应时间及高并发负载下的系统稳定性；资源效率则关注模型运行对硬件资源的占用情况与经济成本控制，典型指标包括显存利用率、单位请求成本等。通过对这两方面的综合考量，本模块可有效评估模型在不同负载条件下的可用性、经济性与可扩展性，为模型部署策略与性能优化提供依据。

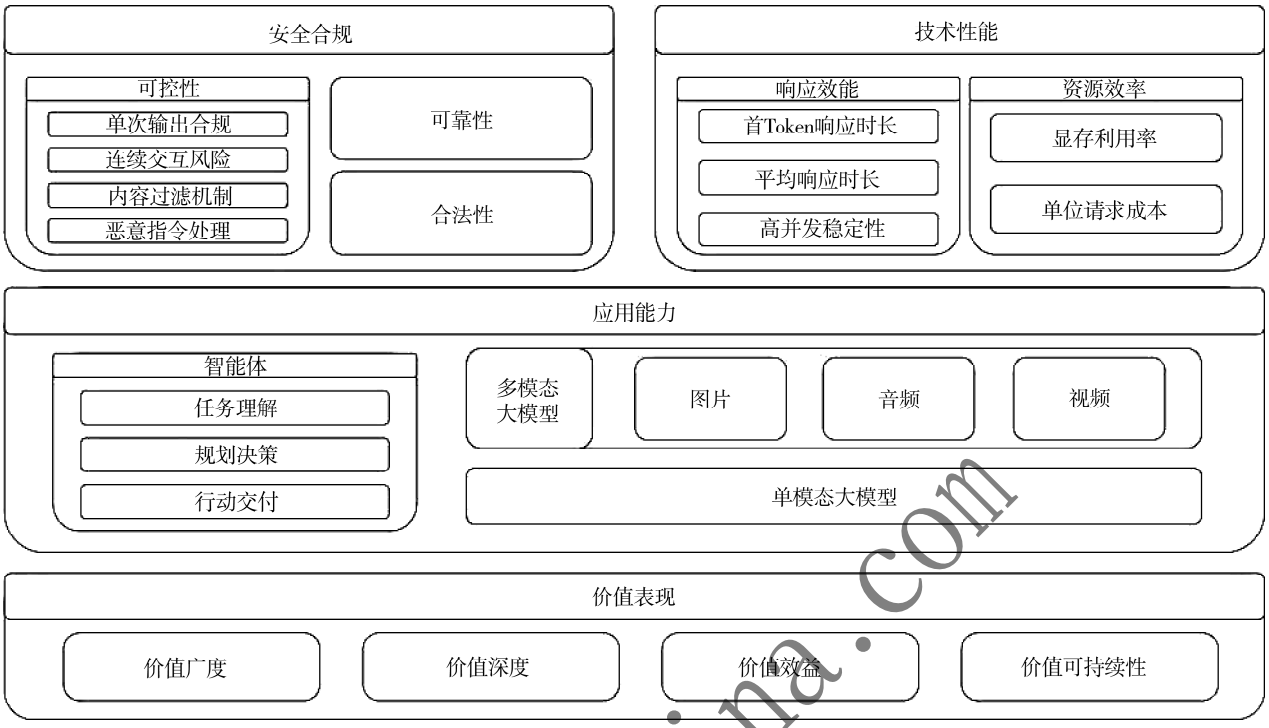


图 1　垂类模型评测框架图

应用能力模块分为大模型能力评测与智能体系统核心执行能力评测两大板块。针对大模型，该模块评估其对不同信息载体的处理能力，单模态能力涵盖文本信息的理解与生成、文本推理与问答等[18]；多模态能力则评估其对图文、文音、图音及图文音等跨模态信息的融合处理能力，包括图文检索与推理、音视频协同的异常检测、有声视频问答等方向[19]。针对智能体系统，该模块以“任务执行全流程”为主线，围绕任务理解能力、规划决策能力与行动交付能力三大维度进行评测，任务理解能力侧重对用户指令的解析，包括关键信息识别、上下文关联、歧义化解与深层需求挖掘；规划决策能力关注将需求转化为可行方案的能力，涉及目标拆解、路径规划、资源调度与应对意外调整等；行动交付能力则评估任务完成与结果呈现能力，涵盖具体操作执行、任务链条推进、结果匹配度与执行过程解释等方面。

价值表现模块与安全合规、技术性能、应用能力模块形成联动——安全合规是价值实现的前提，技术性能与应用能力是价值转化的基础，三者共同支撑价值表现的达成。该模块构建以价值广度、价值深度、价值效益与价值可持续性为核心的评估框架，形成从短期价值贡献到长期发展潜力的完整评估链条。价值广度关注模型覆盖范围与行业影响力，通过行业公开案例数量、服务用户总量、政策契合度等指标衡量其行业辐射能力；价值深度侧重评估技术渗透与创新壁垒，以行业标准引用情况、人工替代能力、专利与论文数量等判断其在业务与技术层面的核心竞争力；价值效益从经济与社会双重视角出发，既核算模型或智能体的直接营收，也评估其对产业链的带动作用[20]，以及在公共服务、文化传播、生态保护等领域的综合社会效益；价值可持续性围绕应用成熟度、方案扩展性与团队发展能力展开，通过数据安全机制、架构灵活性、研发投入强度等维度，判断评估对象在长期稳定运行与迭代优化方面的潜力。

4　关键策略

4.1　评测数据集构建

构建信度与效度俱佳的高质量评测数据集，是人工智能模型评估工作得以科学开展的基石。为规避数据偏见、确保评测结论的普适性与可靠性，本研究在数据集的构建与管理中确立了以下策略。

数据安全与隐私保护是首要原则。所有涉及个人可识别信息或敏感商业机密的数据，在纳入数据集前均须经由严苛的匿名化与脱敏处理，确保过程不可逆，从根本上杜绝隐私泄露风险[21]。在此基础上，本研究建立了一套涵盖数据分级分类、基于角色的权限控制及全链路操作审计的完整数据治理体系。该框架确保了数据从采

集、存储、处理到销毁的全生命周期均处于安全可控状态，为评测工作的合法合规性提供制度性保障。

为全面评估模型在复杂现实场景中的泛化能力，本研究致力于构建一个在应用形态、任务模态及情境维度上均具有充分多样性的数据集合。本研究采用分层抽样与情境构造相结合的方法，针对每一项核心评测指标，系统性地构建足量、具代表性的数据样本，确保其能够覆盖主流应用场景、长尾分布用例及潜在的对抗性输入。

鉴于人工智能技术的快速演进，本研究引入了数据集的动态维护范式。该机制要求对数据集进行定期的版本发布，其核心活动包括：纳入源自新兴应用场域的数据、修正经核验的标注谬误，并依据技术前沿动态，前瞻性地扩充新的评测维度与挑战集。这一持续性更新策略旨在维系数据集的“活性”，确保其评测能力与技术发展保持同步，从而准确反映迭代模型的真实性能水平。

标注质量是决定数据集信度的关键。框架通过制定精确的标注规范、实施标注员分级培训与资格认证、引入多轮交叉校验与专家仲裁机制，构建了一套完整的标注质量管理体系，旨在最大化标注结果的客观性、一致性与可追溯性，从源头上保障最终评测数据的科学性与权威性[21]。

4.2 评测方法

为确保对人工智能模型进行全面、公正且高效的评估，构建了一套融合自动化测试、人工测试与大模型裁判的混合评测方法。该方法通过三种方法的优势互补，旨在从效率、主观判断多个维度对模型性能进行立体化评测。

自动化测试可以保障评测效率与客观性。在本策略中，强调测试流程的标准化与透明化。具体而言，在测试执行前，必须清晰定义所有评估任务所需的预期输出标准、参考答案或关键性能指标的阈值。在此基础上，自动化测试脚本将内置明确的评价指标计算方法、评分规则与判定逻辑，确保每一次测试执行均可重复、可验证，从而为模型性能的迭代优化提供稳定、量化的数据支持。

针对当前大语言模型的快速发展，本研究引入大模型作为裁判进行辅助测试。在此策略中，将重点关注其应用的稳健性与可靠性。首要步骤是审慎选择与特定评价任务高度相关的大模型作为裁判，并采用多个模型进行交叉验证，以提升测试结论的稳定性。本研究通过精心设计的提示工程，将人工定义的评价标准与评分规则，高效地转化为能够激发大模型最佳判断性能的输入提示词，确保大模型能够严格按照既定标准执行评价。然而，大模型并非完美，因此本研究在测试流程中设置了强制性的人工审核机制，由人工专家对模型的评判结果进行抽样审核与校准，及时发现问题并调整评价策略，从而保障最终评价的准确与公正。

人工测试为评测引入了不可或缺的领域专业知识与深层语义理解。为实现高质量的人工评估，本研究制定了多维度的策略：第一，必须编制详尽、无歧义的评价标准指南，并对所有评价人员进行统一培训，确保其对标准有一致的理解和执行尺度；第二，在评价过程中，需持续监控评价结果的分布与评分者间的一致性，通过统计方法及时识别潜在的个体偏差或系统性误差；此外，评价人员的选择至关重要，应优先考虑具备相关领域背景的专家，以保证评价结果的准确性与专业性[22]。为辅助其工作，为评价人员提供了专用的评价工具与平台。最后，本研究建立了长效的优化机制，包括定期对评价人员进行复训以更新其知识与技能，尤其是在评价标准发生调整时；同时，定期收集评价人员的反馈，用以持续优化整个评价流程与标准体系。

通过上述自动化、人工与大模型三位一体的关键策略，本文构建了一个兼具效率、深度与智能的综合性评测方法，为客观衡量模型能力提供了坚实的方法论基础。

5 结论

本文提出了一套用于评估垂类模型综合成效的评测框架。针对当前评测实践中存在的评价维度单一、缺乏领域适配以及方法碎片化等显著短板，本框架通过明确行业大模型、垂域大模型与智能体系统的分类标准，构建了以安全合规、技术性能、应用能力与价值表现为核心的多维度指标体系，形成了结构化的评测方案，其中安全合规模块保障模型部署的合法性与可控性，技术性能模块衡量模型运行的效率与资源经济性，应用能力模块覆盖各种模态处理及智能体任务执行全流程，价值表现模块衔接短期贡献与长期发展潜力，最终实现了不同类型垂类模型的一致化、可比较评估，为模型的开发优化、产业选型与监管决策提供了科学依据。

与之配套的关于数据集构建和混合评测方法的关键策略，构成了本框架的可操作性支柱。强调数据安全、生态效度、动态演化和标注规范的数据集构建原则，确保了评测基础既稳健又贴合实际。同时，协同利用自动化测试的效率、人类专家的知识以及大模型裁判能力的三位一体评测策略，提供了一种平衡且可扩展的方法，以同时捕捉定量指标和细微的定性性能。

本框架在模型评测方面目前仍存在两方面明显局限。

其一，价值表现维度的社会效益、技术创新壁垒等核心指标仍较多依赖定性描述与间接量化手段，不同行业的价值内涵与表现形式差异显著，比如医疗领域的社会效益侧重诊疗公平性与资源节约，金融领域则聚焦风险防控成效与市场稳定贡献，而现有框架缺乏针对各行业特性的标准化量化模型，导致跨行业间的价值对比缺乏统一标尺，精准度与说服力不足；其二，混合评测方法中自动化测试、人工评估与大模型裁判的权重分配，尚未脱离行业经验与专家共识的依赖，未建立基于任务风险等级的动态计算模型，像医疗诊断、金融风控这类高风险场景与文旅咨询、信息查询这类低风险场景的评测需求差异明显，固定化的权重配比难以适配不同场景的核心诉求，尤其在高风险场景中可能因人工评估参与度不足或大模型裁判复核机制缺失，影响评测结果的客观性与可靠性。

对于垂类模型评测，未来仍有多个具体方向需深入探索。本框架需在制造、医疗、金融、政务等更多关键领域开展大规模实践验证，通过收集不同场景下的实测数据与应用反馈，针对性优化指标定义的精准度与评测方法的可操作性。随着 AI 技术的快速迭代，指标体系与数据集需建立常态化动态更新机制，需新增模型幻觉抑制效果、复杂场景可解释性等关键指标，同时针对不同行业设计差异化的长期社会经济指标。此外，需进一步研究自动化测试、人工评估与大模型裁判三者的协同交互机制，基于任务的专业化程度与安全关键等级建立动态权重分配模型，在医疗诊断、金融风控等高度敏感场景中强化人工评估的深度参与和大模型裁判结果的复核流程，在低风险场景中提升自动化测试的效率占比，通过这些针对性优化，让框架持续适配技术发展与产业需求，为垂类模型评测标准的成熟完善、现实场景的规模化可靠部署提供更坚实的支撑。

参考文献

[1] LEE J, DAVARI H, SINGH J, et al. Industrial Artificial Intelligence for industry 4.0-based manufacturing systems [J]. Manufacturing Letters, 2018, 18: 20 – 23.

[2] LITJENS G, KOOI T, BEJNORDI B E, et al. A survey on deep learning in medical image analysis [J]. Medical Image Analysis, 2017, 42: 60 – 88.

[3] SUN T Q, MEDAGLIA R. Mapping the challenges of artificial intelligence in the public sector: evidence from public healthcare [J]. Government Information Quarterly, 2019, 36 (2): 368 – 383.

[4] YANG Q, STEINFELD A, ZIMMERMAN J. Unremarkable AI: fitting intelligent decision support into critical, clinical decision-making processes [C]//Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, 2019: 1 – 11.

[5] DODGE J, GURURANGAN S, CARD D, et al. Show your work: improved reporting of experimental results [J]. arXiv preprint arXiv: 1909.03004, 2019.

[6] LIANG P, BOMMASANI R, LEE T, et al. Holistic evaluation of language models [J]. arXiv preprint arXiv: 2211.09110, 2022.

[7] GAO L, TOW J, ABBASI B, et al. The language model evaluation harness [EB/OL]. [2025 – 10 – 17]. https://github.com/EleutherAI/lm-evaluation-harness.

[8] FOURRIER C, The Hugging Face Community. LLM evaluation guidebook [EB/OL]. [2025 – 10 – 17]. https://github.com/huggingface/evaluation-guidebook.

[9] MYRZAKHAN A, BSHARAT S M, SHEN Z. Open-LLM-Leaderboard: from multi-choice to open-style questions for LLMs evaluation, benchmark, and arena [J]. arXiv preprint arXiv: 2406.07545, 2024.

[10] SINGHAL K, AZIZI S, TU T, et al. Large language models encode clinical knowledge [J]. Nature, 2023, 620 (7972): 172 – 180.

[11] CHEN Z, CHEN W, SMILEY C, et al. FinQA: a dataset of numerical reasoning over financial data [J]. arXiv preprint arXiv: 2109.00122, 2021.

[12] ZHANG Z, ZHAO X, FANG X, et al. Redundancy principles for MLLMs benchmarks [J]. arXiv preprint arXiv: 2501.13953, 2025.

[13] LIU C W, LOWE R, SERBAN I V, et al. How not to evaluate your dialogue system: an empirical study of unsupervised evaluation metrics for dialogue response generation [J]. arXiv preprint arXiv: 1603.08023, 2016.

[14] GURURANGAN S, MARASOVI C A, SWAYAMDIPTA S, et al. Don′t stop pretraining: adapt language models to domains and tasks [J]. arXiv preprint arXiv: 2004.10964, 2020.

[15] HAN X, ZHANG Z, DING N, et al. Pre-trained models: past, present and future [J]. AI Open, 2021, 2: 225 – 250.

[16] SCHICK T, DWIVEDI-YU J, DESSÌ R, et al. Toolformer: language models can teach themselves to use tools [J]. Advances in Neural Information Processing Systems, 2023, 36: 68539 – 68551.

[17] 陈浩泷，陈罕之，韩凯峰，等. 垂直领域大模型的定制化：理论基础与关键技术 [J]. 数据采集与处理，2024，39 (3)：524 – 546.

[18] 苏艳芳，袁静，薛俊民. 大模型安全评估体系框架研究 [C]//第 39 次全国计算机安全学术交流会论文集，2024：107 – 111.

[19] 国家市场监督管理总局，国家标准化管理委员会. 人工智能 大模型 第 2 部分：评测指标与方法 (GB/T 45288.2 – 2025) [S]. 2025.

(下转第 29 页)

[4] 高松. 生成式人工智能的数据安全风险与应对措施 [J]. 中国信息安全，2024（6）：32－37.
[5] GOLDA A，MEKONEN K，PANDEY A，et al. Privacy and security concerns in generative AI：a comprehensive survey [J]. IEEE Access，2024：1248126－1248144.
[6] TAKALE D G，MAHALLE P N，SULE B. Cyber security challenges in generative AI technology [J]. Journal of Network Security Computer Networks，2024，10（1）：28－34.
[7] 张欣. 生成式人工智能的数据风险与治理路径 [J]. 法律科学（西北政法大学学报），2023，41（5）：42－54.
[8] 斜晓东. 论生成式人工智能的数据安全风险及回应型治理 [J]. 东方法学，2023（5）：106－116.
[9] 赵梓羽. 生成式人工智能数据安全风险及其应对 [J]. 情报资料工作，2024，45（2）：30－37.
[10] 虎嵩林. 大模型的安全风险及应对建议 [J]. 中国信息安全，2024（6）：14－17.
[11] GUPTA M，AKIRI C K，ARYAL K，et al. From ChatGPT to ThreatGPT：impact of generative AI in cybersecurity and privacy [J]. IEEE Access，2023，11：80218－80245.

（收稿日期：2025－06－23）

作者简介：

林杰（1967－），男，博士，教授，主要研究方向：人工智能、机器学习、决策支持系统、供应链管理。

王家盛（2000－），男，硕士，主要研究方向：数据安全治理。

（上接第 23 页）

[20] PATWARDHAN T，DIAS R，PROEHL E，et al. GDPval：evaluating AI model performance on real-world economically valuable tasks [J]. arXiv preprint arXiv：2510.04374，2025.
[21] 中国信息通信研究院人工智能研究所，清华大学计算社会科学与国家治理实验室，中国人工智能产业发展联盟数据委员会. 人工智能高质量数据集建设指南 [Z]. 2025.
[22] 王浩，孟祥峰，王权，等. 人工智能医疗器械用数据集管理与评价方法研究 [J]. 中国医疗设备，2018，33（12）：1－5.

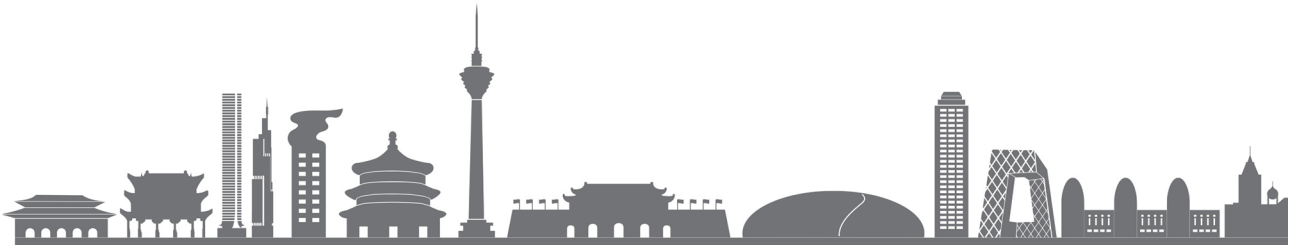
（收稿日期：2025－10－27）

作者简介：

宋元（1987－），男，硕士，主要研究方向：模型评测、大语言模型。

张衍（1987－），男，博士，主要研究方向：大语言模型、RAG、模型评测。

任熠辉（2001－），通信作者，男，硕士，主要研究方向：模型评测、智能体。E-mail：renyihui2001@sina.com。



版权声明

凡《网络安全与数据治理》录用的文章，如作者没有关于汇编权、翻译权、印刷权及电子版的复制权、信息网络传播权与发行权等版权的特殊声明，即视作该文章署名作者同意将该文章的汇编权、翻译权、印刷权及电子版的复制权、信息网络传播权与发行权授予本刊，本刊有权授权本刊合作数据库、合作媒体等合作伙伴使用。同时，本刊支付的稿酬已包含上述使用的费用，特此声明。

《网络安全与数据治理》编辑部