团体标准《智能体评测指标与方法》（征求意见稿）

编制说明

一、工作简况

（一）任务来源

为积极响应国务院发布的《关于深入实施 “人工智能 +” 行动的意见》政策号召，落实国家对新一代智能终端、智能体等应用普及率发展目标，顺应人工智能技术蓬勃发展态势，特制定本《智能体评测指标与方法》。当下，智能体已从工具属性蜕变为能感知、决策、执行的自主系统，在多领域实现深度落地。但评测规范缺失导致行业发展乱象丛生，产品质量参差不齐、跨领域应用受阻、公众信任不足等问题凸显。本标准聚焦建立统一评测体系，通过主客观结合的评测方法精准度量用户体验，为智能体技术发展优化提供方向指引，同时规范市场秩序、加速技术跨领域落地、提升公众信任度，填补智能体评测标准化领域的关键空白，全面推动技术在各行业的健康、高效应用，赋能数字经济高质量发展。

本标准由上海人工智能创新中心提出，上海市人工智能行业协会批准立项。《智能体评测指标与方法》主要起草单位：上海人工智能创新中心，计划应完成时间为 2026 年 2 月。

（二）主要参与起草单位

本标准由上海人工智能创新中心、上海市人工智能行业协会、上海计算机软件技术开发中心、中国移动通信集团有限公司、人形机器人（上海）有限公司、国家安全防范报警系统产品质量检验检测中心（上海）、公安部第三研究所、上海市药品检验研究院、上海市公安局科技信息化总队、北京大学长沙计算与数字经济研究院等 10 余家单位共同负责起草。

（三）主要工作过程与主要起草人所做工作

本标准编制过程到目前主要经历了标准工作组成立、标准调研、立项阶段和标准研制四个阶段。各阶段主要工作总结如下：

1. 标准工作组成立

2025 年 1 月，为保证标准编写质量，面向行业征集标准起草单位，成立了标准起草工作组，吸纳了国内智能体研发与应用领域的实验室、高校和企业的专家，共同开展标准研制工作。

2. 标准调研

2025 年 2 月 - 2025 年 5 月，标准工作组对智能体评测存在的问题、标准发展情况、建设与应用需求等进行资料收集、查阅、分析和整理。对团体标准政策制度进行梳理，确定标准主要内容和标准框架结构。对照《标准化法》、《团体标准管理规定》和 GB/T 1.1-2020 等标准，编写标准草案。期间为提高标准关于智能体评测的普适性和可操作性，标准工作组通过召开研讨会以及书面反馈意见等多种形式，对标准文本进行修改完善。

3. 立项阶段

2025 年 6 月，标准工作组组织各参编单位代表开展了《智能体评测指标与方法》团体标准立项会，会上标准工作组汇报了标准编制的必要性、目的和意义，以及标准草案的框架和主要内容，与会专家一致同意标准立项。

4. 标准研制

2025 年 8 月-10 月，标准工作组组织上海人工智能实验室及各参编单位代表开展了共 3 轮《智能体评测指标与方法》团体标准的标准研制会，进一步针对标准框架和技术内容展开技术研讨，征求各相关方意见，形成标准征求意见稿及编制说明。

二、标准编制原则和确定主要内容的论据及解决的主要问题

（一）原则

（1）符合性原则，本团体标准编制修订符合国家有关团体标准的法律法规、行政规章的要求；

（2）先进性原则，本团体标准的主要技术指标不低于强制性标准的技术要求，充分考虑在我国医疗大模型应用测评领域的可操作性；

（3）协调性原则，本团标标准文本与现行标准协调一致；

（4）规范性原则，本团体标准编写符合 GB/T 1.1 的要求。

（二）确定主要内容的论据

该标准基于人工智能领域通用的概念和技术逻辑，并参照包括《信息技术 人工智能 术语》（标准号 GB/T 41867-2022）、《人工智能 面向机器学习的数据标注规程》（标准号 GB/T 42755-2023）、《人工智能 大模型 第 1 部分：通用要求》（标准号 GB/T 45288.1-2025）、《人工智能 大模型 第 2 部分：评测指

标与方法》（标准号 GB/T 45288.2-2025）、《网络安全技术 人工智能生成合成内容标识方法》（标准号 GB 45438-2025）中关于人工智能、大模型相关定义界定和划分说明文件，提出了相应的评测维度及其具体能力。标准研制期间，标准工作组邀请人工智能、大模型、计算机视觉、信息安全等不同领域的专家多次讨论，并就“智能体核心能力”维度的评测方面进行征集，进一步丰富了各评测维度的具体内容。

本标准覆盖了智能体的通用性基础能力、行业专业能力以及具体场景的应用能力，同时还考虑了智能体在应用中价值对齐问题和伦理安全问题，基本覆盖了智能体技术应用的各个方面。

（三）解决的主要问题

大模型智能体（AI Agent）作为具备环境自主感知、任务动态规划、目标精准执行及交互持续学习能力的新一代人工智能系统，正成为驱动“人工智能+”战略落地、催生新质生产力的核心引擎。其技术特性已在工业智能调度、服务业个性化交互、科研辅助决策等领域展现应用潜力，逐步重塑产业价值创造模式。当前行业面临关键瓶颈：缺乏公认的统一能力评估体系。这一方面导致市场认知碎片化，企业与用户难以对 AI Agent 的实际效能进行量化核验，显著提升选型成本与决策风险；另一方面使得技术安全边界界定模糊、伦理风险防控缺乏量化依据，为行业监管治理带来现实挑战。

本标准基于全球前沿研究成果，深度借鉴 Stanford HELM、OpenCompass 等评测框架的技术逻辑，并参考国际标准制定的规范化流程，构建覆盖感知精度、规划合理性、执行鲁棒性、学习迭代效率的多维度通用评测体系。该标准将为行业提供统一能力参照基准，为技术研发提供清晰优化图谱，为产业应用提供科学决策支撑，为监管治理提供可落地的技术依据，进而填补行业标准空白，保障 AI Agent 产业高质量有序发展。

二、主要试验情况分析

在《智能体评测指标与方法》编制过程中，工作组采用资料深度挖掘、行业实地调研、跨领域专家研讨等多元方式，紧密围绕《关于深入实施“人工智能 +”行动的意见》等国家政策导向，严格遵循人工智能标准化建设要求与团体标准管理规范，开展多轮次、系统性的实验验证工作。

针对评测体系构建，工作组通过反复研讨、多轮专家论证与广泛行业调研，从基础能力、可靠性与安全性、伦理与对齐三大核心维度出发，构建起全面且具有前瞻性的评测框架。在评测方法设计上，创新提出融合客观指标量化分析与主观体验深度评估的综合评测规则，明确单项能力评分细则及综合得分计算逻辑，确保评测结果兼具科学性与实用性。

此外，严格遵循团体标准制定流程，组织多轮次跨行业意见征集与专业技术审查，充分吸纳产学研各界智慧，对标准内容进行持续优化完善，全方位保障《智能体评测指标与方法》的科学性、行业适配性与实践指导性。

四、知识产权情况说明

本标准不涉及知识产权问题。

五、产业化情况、推广应用论证和预期达到的经济效果

《智能体评测指标与方法》为产业规模化发展提供核心规范支撑，有效破解当前市场标准碎片化、技术参数不统一等突出问题。通过明确统一的量化维度，企业可精准对标行业能力基准，推动智能体技术从实验室研发阶段向工程化应用阶段加速转型。该指标体系通过量化智能体的基础核心能力、任务执行效率及运行可靠性，引导市场主体聚焦核心技术迭代与服务质量提升，避免低水平同质化竞争，助力成熟产品实现规模化落地，构建“硬件终端-软件系统-增值服务”协同发展的产业生态。

由行业协会、标准化组织及企业通过培训、研讨会和试点应用推广，配合政府融入监管框架，保障市场公平。在经济效益方面，标准统一评测要求，降低重复测试成本，缩短研发周期，提升市场竞争力，支撑智能服务市场建设；在社会效益方面，标准规范伦理、安全与合规性，保护用户隐私，提升公众信任，促进大模型技术广泛应用；在环境效益方面，减少重复测试的算力消耗，提升资源利用率，助力绿色可持续发展。

六、转化国际标准和国外先进标准情况

无。

七、与现行相关法律、法规、规章及相关标准的协调性

符合国家有关法律法规、政策制度的要求。

八、重大分歧意见的处理经过和依据

无。

九、标准性质的建议

本标准批准后作为推荐性团体标准使用。

十、贯彻标准的要求和措施建议

建议本标准批准发布3个月后实施。

建议本标准由上海人工智能实验室宣贯实施。

十一、替代或废止现行相关标准的建议

无。

十二、其它应予说明的事项

无。

《智能体评测指标与方法》团体标准编制起草组

2025-11-03