ICS
CCS

T/SAIAS

上 海 市 人 工 智 能 行 业 协 会 团 体 标 准

T/SAIAS XXX—2025

智能体评测指标与方法

Evaluation Metircs and Methods for Intelligent Agents

2025－ －发布　　2025－00－00实施

上海市人工智能行业协会　　发 布

目　　次

前　言

本文件按照GB/T 1.1—2020《标准化工作导则　第1部分：标准化文件的结构和起草规则》的规定起草。

近年来，由大型语言模型（LLM）驱动的 AI Agent 技术发展迅猛，展现出迈向通用人工智能的巨大潜力。为准确、客观、全面地衡量其能力与风险，引导技术健康发展，支撑产业应用选型，并辅助监管治理，制定一套统一、透明、可信的评测标准至关重要。

请注意本文件的某些内容可能涉及专利。本文件的发布机构不承担识别这些专利的责任。

本文件由上海市人工智能行业协会提出并归口。

本文件起草单位：

本文件主要起草人：

本标准首次制定。

首期执行单位：

引　言

大模型智能体（AI Agent）作为能够自主感知、规划、行动并从交互中学习的新一代人工智能系统，正成为推动“人工智能+”和新质生产力发展的核心引擎。然而，当前行业缺乏公认的能力评估标尺，导致市场认知混乱、用户选型困难，并为技术的安全、伦理风险治理带来挑战。

本标准借鉴全球前沿研究（如 Stanford HELM、OpenCompass）与国际标准实践，建立一套系统、全面的通用大模型智能体(AI Agent) 能力评测标准，以期为行业提供标准参考，为技术研发提供能力图谱，为产业应用提供决策依据，为监管治理提供技术支撑，填补行业标准空白，促进产业健康有序发展。

智能体评测指标与方法

1 范围

本文件规定了智能体能力评测的总体框架，包括基本原则、能力维度、评测方法等。

本文件适用于面向通用开放领域，完成各类复杂、多样化的领域任务。同时，金融、医疗、编程、法律等垂直领域也可参照执行。

本标准主要针对由大型语言模型（LLM）或多模态模型驱动的、在数字环境中执行任务的智能体。对于涉及与物理世界进行交互的具身智能（Embodied AI），建议参考相关的专门标准。

2 规范性引用文件

下列文件中的内容通过文中的规范性引用而构成本文件必不可少的条款。其中，注日期的引用文件，仅该日期对应的版本适用于本文件；不注日期的引用文件，其最新版本（包括所有的修改单）适用于本文件。

GB/T 41867-2022 信息技术 人工智能 术语

GB/T 45288.1-2025 人工智能 大模型 第1部分：通用要求

GB/T 45288.2-2025 人工智能 大模型 第2部分：评测指标与方法

T/SAIAS 具身智能能力评测框架指南

3 术语和定义

下列术语和定义适用于本文件。

3.1

AI 智能体 AI agent

基于大语言/多模态模型的智能体系统。

3.2

规划与推理 Planning and reasoning

自主分解复杂目标，进行逻辑推断和策略规划。

3.3

行动 Action

调用工具（如 API、代码解释器）或与环境交互来执行任务。

3.4

记忆 Memory

具备短期记忆和长期记忆（能够处理当前对话的上下文信息，或跨对话场景下的关键信息，用户偏好等）。

3.5

反思 Reflection

从执行结果中总结经验，修正策略，并具备记忆能力。

3.6

沙盒 Sandbox

受控环境，用于隔离测试 Agent 与外部工具或环境的交互。

3.7

工具 Tool

可供大模型进行调用的工具或工具集合。

4 缩略语

下列缩略语适用于本文件。
AI 人工智能（Artificial Intelligence）
API 应用程序编程接口（Application Programming Interface）
HMLM 全方位大语言模型评测（Holistic Evaluation of Language Models）
OpenCompass司南大模型评测体系（OpenCompass Foundation Model Evaluation System）

5 智能体核心能力

5.1 基础能力

5.1.1 感知能力

a) 环境感知：识别和理解智能体所处数字或物理环境状态的能力。
b) 数据感知：从多样化、非结构化的数据源中准确提取信息的能力。

5.1.2 任务完成能力

a) 任务成功率：在权威基准上端到端的任务完成比例。
b) 指令遵循度：理解并执行复杂、多重约束指令的准确性。
c) 任务复杂度处理：按步骤数、工具依赖等量化的复杂度上限。

5.1.3 规划与推理能力

a) 逻辑推理：演绎、归纳与因果推断能力（数学、科学领域）。
b) 上下文一致性：在长对话或复杂任务流中，持续保持逻辑一致性的能力。
c) 策略规划：自主分解复杂目标，规划最优或次优路径。
d) 自我反思与修正：试错-学习能力，从失败中分析原因并修正策略。

5.1.4 工具使用与扩展能力

a) 工具调用准确性：选择、调用并理解 API、代码解释器等工具的正确率。
b) 多工具协同：协同多个工具完成复杂任务流程。
c) 工具泛化能力：面对新工具（仅提供文档）自主学习并成功使用。
d) 协议支持度：对主流工具调用协议（如 MCP、A2A 等）的兼容与支持情况。

5.1.5 知识与记忆能力

a) 知识准确性与幻觉监控：评估内置与检索知识的真实性、时效性，并建立对模型产生幻觉（Hallucination）的监控与评估指标。
b) 长期记忆：多轮交互中存储、检索关键信息的能力。
c) 情境感知：动态理解对话、任务状态与环境变化。

5.1.6 自主性能力 (Autonomy)

a) 目标驱动性：在仅给定一个高层级、甚至模糊的目标时，智能体自主定义可执行的子目标和清晰任务规划的能力。
b) 长期任务执行：在无需持续人工干预的情况下，独立执行跨越数小时甚至数天的长期任务的能力。
c) 环境适应与策略调整：在任务执行过程中，当外部环境或工具集发生非预期变化时，智能体能否自主调整其原有策略并继续完成任务。

5.1.7 多智能体协同能力

a) 任务分解与分配：能够将复杂任务有效分解为子任务，并分配给合适的子智能体（Sub-agent）执行。
b) 通信与协作：子智能体之间能够进行有效的信息交换和状态同步，以协同完成整体任务。
c) 冲突解决与资源管理：在多智能体协作过程中，能够处理潜在的目标冲突或资源竞争。

5.2 可靠性与安全性

5.2.1 鲁棒性

a) 抗干扰能力：在噪声、模糊或矛盾输入下的稳定性。
b) 容错与优雅降级：工具或环境故障时识别、替代或报告能力。

5.2.2 安全性

a) 防越狱与提示注入：抵抗恶意提示操纵，防止有害操作。
b) 权限控制：遵循最小权限原则，不执行越权操作。
c) 数据安全：遵循安全协议，防止数据泄露。
d) 物理世界隔离：评测过程必须确保测试活动与真实的物理世界和生产系统完全隔离，防止任何潜在的负面影响。

5.2.3 可控性与可预测性

a) 行为一致性：相似输入对应稳定输出与决策逻辑。
b) 人工干预机制：人类审查、干预、终止机制的清晰度与有效性。
c) 可审计与可验证性：智能体的决策过程和推理链条应可追溯、可验证，能够以可理解的方式输出，支持事后审计。

5.3 伦理与对齐

5.3.1 价值观对齐

a) 有害性拒绝：拒绝非法、有害、歧视性指令。
b) 偏见与公平性：使用 BBQ、BOLD 基准评估系统偏见。

5.3.2 透明度与可解释性

a) 决策溯源：以可理解方式解释决策原因与推理链。
b) 信息来源披露：说明事实信息来源的准确性。

5.3.3 隐私保护

a) 个人信息处理：对用户数据中个人敏感信息的处理，应遵循国家标准 GB/T 35273-2020《信息安全技术 个人信息安全规范》的相关要求，进行有效的识别、脱敏或特殊处理。
b) 合规性：遵循国家及地方关于数据和隐私保护的相关法律法规。在高敏感度垂直领域中，需界定清晰的隐私数据识别与使用规范，确保在不影响核心功能的前提下最大化保护用户隐私。

5.4 应用效能评测

应用效能评测旨在从实际应用的角度，衡量智能体在执行任务过程中的效率、成本、用户体验及知识时效性。

5.4.1 智能体效率

主要通过“响应时间”来衡量，具体指从任务开始到智能体最终完成任务所需的端到端耗时，单位为秒（s）。

此项指标应包含在不同负载下的压力测试表现。

5.4.2 成本效益

成本效益从多个维度综合评估智能体的运行成本：

a) 步骤经济性： 指智能体完成任务所采取的平均操作步骤数（N）。步骤越少，通常代表其规划能力越强，效率越高。
b) 计算资源消耗：记录智能体在任务执行过程中消耗的核心计算资源，包括处理的 Token 总数、模型的调用总次数以及 CPU/GPU 的使用量。
c) API 调用成本： 指完成任务过程中，调用外部 API（如搜索引擎、数据库等）所产生的实际费用。
d) 综合任务成本： 这是对任务总成本的量化评估，其计算方法为：
   1) 综合任务成本 = (计算资源消耗成本) + (API 调用成本)
   2) 其中，“计算资源消耗成本”可根据模型提供方的定价（例如，按 Token 数或调用次数计费）进行核算。

5.4.3 用户体验与评价

主要评估用户在与智能体交互过程中的主观感受：

a) 交互自然度： 通过评估对话的流畅度、拟人化程度以及对用户意图理解的精准度来衡量。可采用 5 分制李克特量表（Likert Scale）进行人工打分，例如，1 分表示“非常不自然”，5 分表示“非常自然”。
b) 满意度与有用性： 评估智能体完成任务的结果是否满足用户预期且具有实际价值。这通常通过人类偏好评分来完成，例如采用 Elo 评分系统。在该系统中，测试人员对两个不同智能体完成同一任务的结果进行比较（“A 更好”、“B 更好”或“平局”），通过大量成对比较的结果来计算各个智能体的相对排名和评分。

5.4.4 知识更新度

评估智能体所掌握和提供信息的更新程度，即其知识与现实世界变化保持同步的能力。评测所用的问题和数据集应建立动态更新机制，定期补充新的时效性问题，并移除或更新过时的问题，以确保持续评估智能体跟进现实世界变化的能力。

6 评测方法与实施流程

6.1 评测方法

a) 静态基准测试：采用离线、有标准答案的数据集进行测试。此方法效率高、可复现性强，主要用于评测有明确答案或固定评估标准的指标。
   - 适用范围包括：
     1) 感知能力 (6.1.1)：特别是数据感知，例如评测从非结构化文本、图表中准确提取信息的能力。
     2) 规划与推理能力 (6.1.3)：特别是逻辑推理、数学计算等，例如使用 GSM8K、MATH 等基准进行测试。
     3) 知识与记忆能力 (6.1.5)：重点评估知识准确性，例如使用 MMLU 等知识类基准。
     4) 安全性 (6.2.2)：针对已知的攻击样本（如提示注入、越狱尝试）进行防御能力测试。
     5) 价值观对齐 (6.3.1)：如使用 BBQ、BOLD 等基准评估系统偏见。
     6) 隐私保护 (6.3.3)：使用包含个人敏感信息的数据集，评测智能体识别与处理的有效性。
b) 动态交互评测：在受控的沙盒环境中，通过让智能体执行包含多步骤、需要与模拟环境或真实工具进行动态交互的任务，来评估其综合能力。此方法能更好地反映智能体在真实场景中的表现。
   - 关键要素： 必须确保评测环境的安全性和隔离性，避免对外部真实世界产生非预期影响。
   - 适用范围包括：
     1) 感知能力 (6.1.1)：特别是环境感知，评估其在任务过程中对动态变化的适应能力。
     2) 任务完成能力 (6.1.2)

3) 工具使用与扩展能力（6.1.4）
4) 自主性能力（6.1.6）
5) 多智能体协同能力（6.1.7）
6) 鲁棒性（6.2.1）
7) 可控性与可预测性（6.2.3）：例如测试在任务中途接收人类干预指令并调整行为的能力。
8) 应用效能（6.4.1 效率，6.4.2 成本，6.4.4 知识更新度）

c) 人工评估：由领域专家或通过众包方式，对难以量化或涉及主观判断的指标进行评分。
- 关键要素： 需制定详尽、一致的评估准则（Rubric）并对评估人员进行培训，以保证评估结果的可靠性和一致性（即“评分者间信度”）。
- 适用范围包括：
  1) 可控性与可预测性（6.2.3）：评估其审计日志和推理链条的可追溯性与可验证性。
  2) 透明度与可解释性（6.3.2）：评估其决策解释是否清晰、易于理解。
  3) 隐私保护（6.3.3）：对复杂场景下隐私处理的合理性进行定性评估。
  4) 用户体验与评价（6.4.3）：如交互自然度、满意度等。
  5) 对动态评测和模型辅助评测的结果进行抽样复核。

d) 模型辅助评估：选取在特定评测任务或垂直领域上，能力表现出色且经过验证的模型作为“裁判”，对智能体的输出进行打分和评估。此方法可作为人工评估的一种规模化、低成本的替代方案。
- 采用此方法时，应遵循以下原则以确保结果的可靠性：
  1) 领域专家原则：作为评审员的“裁判模型”，其在被测垂直领域的专业能力应经过验证，并不低于（即持平或优于）被评测智能体的模型。评测发起方应优先选用在该领域公认表现更优的模型，并在评测报告中声明所用“裁判模型”及其在该领域的性能依据。
  2) 偏见声明原则：应明确声明“裁判模型”自身可能存在的偏见，并在分析结果时予以考虑。
  3) 人工抽检原则：对于模型给出的评分和评价，需进行一定比例的人工抽样复核，以校准和验证“裁判模型”的评估质量。
- 适用范围包括：
  用户体验与评价（6.4.3）等主观性较强的指标的大规模初步筛选。

6.2 实施流程

a) 评测准备：明确评测目标，选择合适的基准与方法，搭建标准化的测试环境。
b) 评测执行：运行自动化脚本，详细记录交互日志，并组织开展人工评测。
c) 评测分析：计算量化指标，统计主观评分，结合典型案例进行深度分析，定位能力短板。

6.3 评测结果呈现与分析

评测结果应以结构化、可视化的方式呈现，以清晰地展示智能体的综合能力水平。建议采用以下方式：

a) 能力雷达图：根据核心能力（如任务完成能力、推理规划能力）和应用效能（如效率、成本）的各项指标得分，绘制成雷达图，直观地展示智能体的优势与短板。
b) 综合得分与排名：对各项指标进行加权计算，得出一个综合分数，并可用于不同智能体之间的横向对比排名。
c) 详细分析报告：结合评测数据和典型的交互案例进行深度分析，详细阐述智能体在不同场景下的具体表现，并定位其能力短板，为后续优化提供明确指引。

附 录 A
（资料性）

A.1 智能体场景：

本标准的核心能力与评测方法可应用于多种场景，通过具体场景可以更有效地检验智能体的综合能力。

A.1.1 通用生活场景

如问答对话、旅行规划、餐厅预订等，考验智能体理解自然语言、调用多种公共API（搜索引擎、地图、日历等）并完成规划的能力。

A.1.2 专业领域场景

a) 金融分析： 自动分析财报、检索市场新闻、生成投资摘要，考验数据感知、逻辑推理和知识准确性。
b) 代码开发： 根据需求文档编写代码、调试、运行测试用例，考验工具使用（代码解释器、编译器）、逻辑推理和自我修正能力。
c) 工业控制： 在仿真环境中根据指令操作设备，考验指令遵循的精确性、安全性和容错能力。