ICS 35.240
CCS L70

T/SAIAS

上海市人工智能行业协会团体标准

T/SAIAS XXX—2026

人工智能 智能体能力分级与评测方法

Artificial intelligence—Agent capability grading and evaluation method

（征求意见稿）

XXXX－XX－XX 发布　　XXXX－XX－XX 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。

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：

人工智能 智能体能力分级与评测方法

1 范围

本文件给出了智能体能力评测等级模型和评测方法。

本文件适用于智能体的需求方、开发方以及第三方评测机构等相关组织开展智能体业务能力水平测试评估。

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。

ISO/IEC 22989:2022 信息技术 人工智能 人工智能概念和术语(Information technology — Artificial intelligence — Artificial intelligence concepts and terminology)

3 术语和定义

下列术语和定义适用于本文件。

3.1

智能体 agent

能够感知和响应所处环境并能执行操作以完成目标的自动化实体。

注:本文件涉及的智能体仅指运行在设备上的软件实体。

[来源:ISO/IEC 22989:2022,3.1.1,有修改]

4 基本原则

4.1 价值导向原则

以评估智能体支撑业务的实际能力为评测核心,锚定业务价值对应的能力水平开展评测,避免陷入组件级技术指标细节。

4.2 聚焦业务原则

以真实业务场景需求为锚点,遵循“一智能体一评测”原则开展评测,支持根据业务要求自主调整评测基线及指标权重。

4.3 客观公正原则

确保评测结果客观公正,遵循三重保障机制:

a) 以场景真实数据构建测评数据,规避预设场景片面性;

b) 引入竞对智能体与人工执行数据作参照;

c) 通过盲测、多专家交叉验证及消偏机制,减少主观偏见,确保结果可复现、可追溯。

4.4 独立可控原则

通过搭建独立评测环境等方式,确保评测过程可控,以保障评测结果不受外界因素所干预或影响。

5 智能体能力等级模型

5.1 能力等级

智能体能力等级（简称“能力等级”）规定了智能体对于专业业务的支撑度水平，能力等级自低向高分为L1级基础级（L1）、辅助级（L2）、自主级（L3）、协同级（L4）四个等级。较高的能力等级要求涵盖了低能力等级的要求。

基础级（L1）：智能体需经人类唤醒启动，仅能被动响应外部指令，需严格遵循人类下达的单次指令或预设工作流逐步推进任务，全程需人类管控流程，无任何自主决策与处置权限，是最基础的执行单元。

辅助级（L2）：智能体可在预设场景自动感知，理解特定场景内的常规指令与简单非结构化需求，在既定流程和工具范围内自主完成任务，无需人类逐步骤指导，但无法脱离预设流程，仅能为人类提供流程化辅助支撑，场景变化时需人类重新配置。

自主级（L3）：智能体可主动感知全域信息，精准理解模糊化、高难度甚至创新性需求，可自主完成任务拆解、路径规划与落地执行，独立交付结果，实现从需求到结果的全流程自主闭环。

协同级（L4）：智能体具备跨场景、跨主体的协同感知能力，能理解系统性、全局性需求，可自主确定任务目标、整合全域资源、联动多智能体或人类协同完成工作，同时具备自我迭代进化能力，可通过复盘反思主动纠错迭代，持续优化知识经验与记忆体系。

5.2 能力维度

5.2.1 能力维度框架

智能体的能力维度由能力域和能力子域构成，如图1所示。

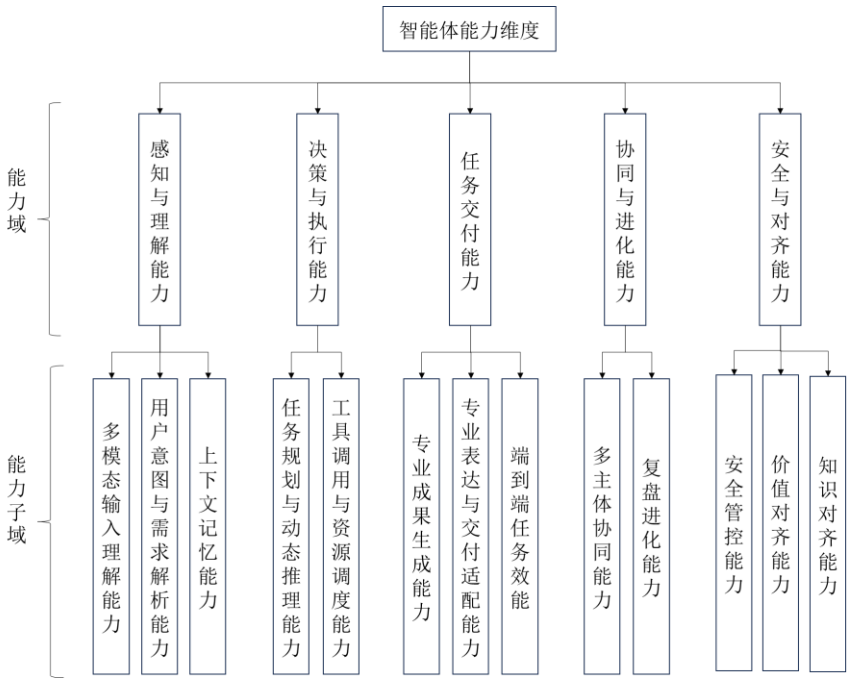

图1 智能体能力维度

5.2.2 能力域

能力域给出了能力等级评测的关键方面，包括感知与理解能力、决策与执行能力、任务交付能力、协同与进化能力、安全与对齐能力五个维度。

——感知与理解能力是指智能体感知外部环境（指令、场景、数据等）、理解信息含义的能力，决定智能体对任务与场景的适配度。

——决策与执行能力是指智能体基于任务目标进行逻辑推理、任务拆解、工具选择与资源调度的能力，是智能体实现自主闭环的核心能力。

——任务交付能力是指智能体端到端交付专业、规范、可用的结果，满足业务对输出内容质量、

格式、时效等要求的能力，是衡量智能体业务价值的关键结果指标。

——协同与进化能力是指智能体联动多主体（人类/其他智能体）协同工作、通过复盘迭代实现自我进化的能力，是高等级智能体的核心特征。

——安全与对齐能力是指智能体在任务全流程中保障运行安全、规避安全风险，同时对齐伦理规范、法律法规、用户核心价值、专业领域业务知识（领域知识、用户需求、场景规则等）的能力，保障智能体的可靠运行和专业性。

5.3 能力等级要求

5.3.1 感知与理解能力

感知与理解能力包括多模态输入理解能力、用户意图与需求解析能力、上下文记忆能力3个能力子域。感知与理解能力按能力等级可划分为不同等级要求，见表1。

表1 感知与理解能力的能力等级要求

能力子域	L1级要求	L2级要求	L3级要求	L4级要求
多模态输入理解能力	a) 能处理纯文本输入,无法解析图像、语音、表格等非文本信息。适用于简单问答或命令式交互。	a) 可理解结构化数据，并能将文本与结构化内容关联，但无法处理原始图像或音频。	a) 支持基础多模态输入，如图像OCR、语音ASR,并能进行图文/文音语义对齐。	a) 具备跨模态融合理解能力，可联合分析视频帧、传感器时序数据、语音情感与文本语境，完成复杂推理。
用户意图与需求解析能力	a) 能响应明确、单轮、无歧义的指令，无法处理模糊、省略或多目标请求。	a) 可识别预设场景内常规意图（含简单非结构化），并基于预设规则完成基础推理匹配。	a) 能识别模糊化、多维度关联复杂意图； b) 可独立提炼核心诉求,无需外部补充。	a) 可识别系统性、全局性协同意图； b) 能兼顾多主体诉求形成统一认知； c) 可联动多主体校验意图识别准确性。
上下文记忆能力	/	a) 具备基础工作记忆能力，可保留最近3轮原始对话,支持简单指代消解。但无法处理长文本输入，一旦上下文超限即截断，导致状态断裂。	a) 能对长程交互进行摘要压缩，动态维护一个精简但语义完整的记忆状态。支持超过10轮的多轮次状态一致性，并能根据新信息更新或遗忘旧记忆。	a) 构建结构化的长期记忆库，支持跨任务、跨会话、跨主体的记忆共享。可形成用户专属的“数字记忆画像”，并在授权下用于个性化服务。

5.3.2 决策与执行能力

决策与执行能力包括任务规划与动态推理能力、工具调用与资源调度能力2个能力子域。决策与执行能力按能力等级可划分为不同等级要求，见表2。

表2 推理与执行能力的能力等级要求

能力子域	L1级要求	L2级要求	L3级要求	L4级要求
任务规划与动态推理能力	a) 无需复杂任务规划，仅支持预设或线性顺序的原子操作，无自主拆解能力。	b) 具备基础推理与任务规划能力，能将明确目标转化为具体的、可执行的子任务列表。	a) 支持动态重新规划，当执行受阻或环境变化时，能基于新信息重新推理并调整后续步骤。	a) 能协同多主体统筹拆解全局任务，将宏观战略目标自主分解为可量化、可执行的额阶段性战术目标与项目蓝图，并持续对齐整体意图。

表2　推理与执行能力的能力等级要求（续）

能力子域	L1级要求	L2级要求	L3级要求	L4级要求
工具调用与资源调度能力	a) 可调用预设范围内的固定工具，所有参数由系统预填充，无法根据任务上下文调整，调用失败即终止流程。	b) 能根据任务规划中的子任务，匹配并调用对应的工具，支持基础参数映射，但无异常处理机制。	a) 能动态适配工具调用：根据执行反馈自动重试、切换备用工具、降级策略或修正参数，确保任务持续推进； b) 支持多工具组合调用（≥5个），具备基础资源冲突检测。	a) 能自主发现、验证并集成新工具，跨系统协调多主体资源，形成弹性执行网络，支撑复杂任务的高可靠落地。

5.3.3　任务交付能力

任务交付能力包括专业成果生成能力、专业表达与交付适配能力、端到端任务效能3个能力子域。任务交付能力按能力等级可划分为不同等级要求，见表3。

表3　任务交付能力的能力等级要求

能力子域	L1级要求	L2级要求	L3级要求	L4级要求
专业成果生成能力	a) 存在明显事实错误、逻辑矛盾或知识缺失；结论不可信。	b) 内容基本正确，逻辑连贯，覆盖常规知识点，但缺乏深度或证据支持。	a) 可生成适配复杂场景的内容；内容达到专业水平：结论有依据、推理完整、符合行业认知；可直接用于业务决策或交付	a) 能识别隐含风险、提出创新方案或跨领域关联。
专业表达与交付适配能力	a) 输出无固定结构，格式混乱，风格不当，详略失衡，形式单一。	a) 能按明确指令调整风格，输出基础格式	a) 自动适配行业或企业标准格式，输出符合规范的形式	a) 能主动选择最优表达策略与交付模态，动态优化沟通效能。
端到端任务效能	a) 需外部全程校验，才能完成任务结果交付。 b) 任务成功率<70%，常遗漏关键要素或输出无效内容。	a) 可交付预设场景下的基础任务结果； b) 结果交付后需外部抽样校验。 c) 成功率≥85%，覆盖标准子流程，成果基本完整	a) 可独立交付复杂任务的精准结果； b) 能自主校验结果准确性并修正偏差。 c) 成功率≥95%，覆盖主干任务	a) 可同步向多主体交付标准化结果； b) 能协同多主体校验结果一致性。 c) 成功率≥99%，主导复杂跨域任务，成果带来可量化业务增益

5.3.4 协同与进化能力

协同与进化能力包括多主体协同能力、复盘进化能力2个能力子域。协同与进化能力按能力等级可划分为不同等级要求，见表4。

表4 协同与进化能力的能力等级要求

能力子域	L1级要求	L2级要求	L3级要求	L4级要求
多主体协同能力	/	a) 可接收多主体发送的基础信息； b) 能响应简单的跨主体协作指令。	a) 可与多主体开展基础协作，同步任务相关信息； b) 能配合其他主体完成分工任务。	a) 可主动联动多主体搭建协同体系； b) 能统筹协调多主体协作节奏与分工； c) 可同步协同进度与结果，保障全域一致性。
复盘进化能力	/	a) 可记录任务执行过程中的基础信息与问题。	a) 可自主复盘单一任务的执行效果； b) 能提炼基础优化点，修正简单执行偏差。	a) 可协同多主体开展全域任务复盘； b) 能提炼系统性优化策略，迭代核心能力； c) 可将优化成果同步至多主体，实现整体进化。

5.3.5 安全与对齐能力

安全与对齐能力包括安全管控能力、价值对齐能力、知识对齐能力3个能力子域。安全与对齐能力按能力等级可划分为不同等级要求，见表5。

表5 安全与对齐能力的能力等级要求

能力子域	L1级要求	L2级要求	L3级要求	L4级要求
安全管控能力	a) 需外部全程管控，保障基础安全。	a) 可识别预设场景内的简单安全风险； b) 能执行基础安全防护指令。	a) 可自主识别复杂场景下的安全风险； b) 能自主启动基础安全防护措施； c) 可记录安全风险与处置信息。	a) 可协同多主体构建全域安全防护体系； b) 能联动处置跨主体安全风险。
价值对齐能力	a) 需外部引导，才能对齐基础价值导向。	a) 可在预设场景内自主对齐基础价值； b) 能规避简单价值偏差行为。	a) 可在复杂场景下精准对齐核心价值； b) 能自主识别并修正价值偏差。	a) 可协同多主体制定统一价值对齐标准； b) 能联动校准多主体价值偏差； c) 可同步价值标准至各关联主体。
知识对齐能力	/	a) 可在外部引导下对齐基础知识。	a) 可自主对齐核心知识，识别知识偏差； b) 能修正基础知识偏差内容。	a) 可协同多主体维护统一知识体系； b) 能联动校准全域知识偏差，保障知识一致性。

注：“/”表示该能力等级对当前能力维度不做要求。

6 评测方法

6.1 评测流程

智能体能力等级的核心评测流程包括：

a) 评测集构建：依据智能体各能力等级要求及实际业务场景，搭建覆盖对应能力维度与要求的标准化评测集；
b) 评测执行：在独立评测环境中开展智能体评测集全流程测试，同步记录完整评测过程和数据；
c) 结果判定：依据评测综合得分判定智能体的能力等级；
d) 结果呈现：梳理评测过程、核心数据及等级判定结论，形成结构化、标准化的测试报告。

6.2 评测集构建

6.2.1 评测集构建原则

6.2.1.1 评测数据应与能力要求对应，每条评测数据的设置需明确是为验证哪个具体能力要求，应确保评测集设置覆盖需要评测的能力等级的所有能力要求。

6.2.1.2 每道评测数据应包含输入信息（即考试题目）、标准参考答案、分值区间三个核心要素：

a) 试题表述应清晰、场景明确，适配评测实际场景；
b) 标准参考答案应界定符合要求的判定边界；
c) 分值区间应依据试题难度、对应能力要求的重要性科学设定，支撑量化评测与结果判定。

6.2.1.3 评测集应基于具体业务场景设计，紧贴实际业务需求，避免脱离业务的抽象化设计，确保评测结果真实反映智能体在实际业务中的能力表现。

6.2.1.4 评测数据的数量应满足全面性要求，每条能力要求的评测数据数量宜不少于 10 条，应根据要求的重要性、业务覆盖需求确定评测集中评测数据的数量分配，核心能力要求可适当增加试题数量，避免因试题不足导致评测偏差，保障评测结果的客观性与可信度。

6.2.2 评测集构建方法

6.2.2.1 评测任务分类

评测任务分为已明确目标等级的评测任务和未明确目标等级的评测任务两类，需根据评测任务分类进行评测集构建。

6.2.2.2 已明确目标等级的评测任务

当业务需求已明确智能体的评测目标等级时，按以下流程构建评测集：

a) 依据目标等级，梳理该等级下所有能力域、能力子域的具体能力要求，形成目标等级能力要求清单；
b) 按照 6.2.1 的原则针对目标等级能力要求清单构建评测集，评测集应覆盖目标等级全部能力要求，对应于每个能力要求的评测数据数量可自行设定（宜大于 10 条），应依据各能力要求对于业务场景的重要性确定，越重要的能力要求评测数据数量占比越大。
c) 对评测数据进行整理形成评测集，包括标记评测集与能力域和能力子域的映射关系、检查评测集的覆盖率、业务贴合度及数量合理性等。

6.2.2.3 未明确目标等级的评测任务

未明确目标等级的评测任务，需逐级判定智能体的能力等级，应逐级构建评测集，直至不通过的等级。各级的评测集构建可按照6.2.2.2的方法进行评测集构建。

6.3 评测执行

6.3.1 运行智能体

以评测集中的每一条评测数据的输入信息作为输入，运行被测智能体，输出过程日志和运行结果并记录。

宜采用沙箱进行评测，评测时智能体可在模拟环境中运行，以保证评测过程的独立性。

6.3.2 运行结果评分

a) 每条评测数据的得分计算：通过自动化工具或人工方式，将智能体输出记录与标准参考答案进行对比，依据答案准确性在分值区间内进行打分。

b) 综合评分按公式（1）计算。

$$T = \sum_{i=1}^{n} t_i / \sum_{i=1}^{n} tmax_i \times 100 \quad (1)$$

其中：

T——在评测集上的综合评分（百分制）；

t_i——第i评测数据的得分；

$tmax_i$——第i评测数据分值区间的上限值；

n——评测集中评测数据的总数量。

6.3.3 评分实施方式

评分的实施方式包括自动化工具、人工方式以及工具与人工协同的方式，宜优先选择协同方式，以兼容效率与质量的平衡。

应设置负面操作的触发条件和惩罚机制，若智能体在任务过程中触发了负面操作（如恶意删除数据、泄露隐私等）应被罚分，如涉及不合法不合规的情形可设置一票否决项，等级判定不通过。

评分实施方式主要包括：

a) 工具主导、人工抽检模式：使用工具进行全量测试数据评分后，自动标记一致性低的考题，再随机抽取不低于10%的常规考题，由人工完成抽检与复核，最终输出得分。该模式适用于普通场景或大规模初筛场景；

b) 人工主导模式：人工完成全量测试数据评分。该模式适用于对质量和可溯性要求较高的重点场景或高风险决策场景；

c) 交叉验证模式：工具与人工独立完成全量测试数据评分，分别输出得分，对比两者一致性，一致项直接采用得分，差异项由专家仲裁小组进行最终判定，形成最终结果。该模式适用于需确保可靠、偏差小、高优先级的场景，如关键里程碑评测任务或对外发布场景。

6.3.4 评测工作的组织方式

评测工作应明确组织主体与责任分工，可采用独立评测、联合评测或委托第三方评测等方式实施。

实施评测的单位及人员应具备相应资质与能力，严格遵守评测规则，保障评测工作的客观性、公正性与规范性。

评审人员需具备丰富的业务经验与专业知识，负责关键维度的复核、权威知识标注与差异项仲裁，保障评测深度。

6.4 结果判定

6.4.1 评测结果表达

评测结果表达包括通过和不通过两种情况：

a) 通过，表示智能体达到当前评测的能力等级；

b) 不通过，表示智能体未达到当前评测的能力等级。

6.4.2 等级判定

6.4.2.1 基准线设置

测试前应对能力等级判定的基准线进行设置，即当前评测等级的综合评分大于该基准线即判定为通过，反之则判定为不通过。

基准线应不低于80分，对于安全性、专业性要求高的智能体，其评测基准线设置应提高。

6.4.2.2 等级判定方法

a) 已明确目标等级的评测任务，目标等级评测的综合评分大于基准线即判定为通过。对于不通过的情形，可进行降级评测或改进后再进行同能力等级评测。

b) 未明确目标等级的评测任务，从低至高逐级进行评测，当前等级通过后进行下一力等级评测，直至不通过，该智能体的能力等级为通过的最高能力等级。

6.5 结果呈现

评测结果宜通过结构化的评测报告呈现，评测报告宜包括如下内容：

a) 评测结果：包括评分结果、判定的能力等级等；
b) 评测方案：包括评测环境描述、评测集描述、能力要求清单、基准线设置及其理由、评测工具情况、评测工作组织方式等；
c) 评测过程记录：包括输入与输出记录、工具评分结果详细记录、专家打分结果详细记录、综合评分计算过程、等级判定过程等；
d) 结果分析：清晰呈现智能体的能力表现、水平等；
e) 优化建议：针对被测智能体的能力短板，结合评测过程中定位的核心原因，形成优化建议，包括能力短板、待优化的能力方向、优化优先级（如高/中/低）等。

参考文献

[1] GB/T 45288.2—2025 人工智能 大模型 第2部分：评测指标与方法
[2] GB/T 45288.3-2025 人工智能 大模型 第3部分：服务能力成熟度评估
[3] AIIA/PG 0152—2024 智能体技术要求与评估方法