《人工智能 智能体能力分级与评测方法》（征求意见稿）编制说明

一、标准编制背景

随着人工智能赋能行业应用落地深化，单纯聚焦基础认知的大模型已难以满足自主解决实际问题的需求，智能体应运而生。区别于大模型评测，智能体评测需依托真实业务场景，核心验证其能否精准达成业务目标、合规完成任务流程。为提供一套智能体能力等级评测基准框架，统一行业评测维度和评测流程，2026年上海市人工智能行业协会下达的年度计划，《人工智能 智能体能力分级与评测方法》正式立项。由上海市人工智能行业协会提出并归口，由上海库帕思科技有限公司牵头并联合上海脉博深处科技有限公司、上海华东电信研究院、工业互联网创新中心（上海）有限公司、上海市人工智能行业协会等单位共同起草。

二、 编制过程

本标准的修订主要包括以下几个阶段：

（一）标准起草和立项申请阶段

牵头起草单位组建由科研机构、骨干企业及标准化技术组织构成的起草工作组，开展资料收集与产业调研，梳理智能体领域国内外标准，如国标、团标、国际标准，以及核心学术研究成果，摸排产业应用实践案例与实际需求，确定了标准内容框架和标准的核心技术要素，并依据GB/T 1.1—2020起草规范编写标准草案。

项目组完成建议书和标准草案等立项材料的编写，并根据上

海市人工智能行业协会要求完成立项答辩并正式立项，并由协会进行公示。

（三）征求意见阶段

组建标准工作组，并基于立项答辩中专家提出的意见或建议，表准工作组召开了多轮内部讨论会，并以专家学者、业务骨干座谈会以及企业调研等形式进行深入调研和讨论，对标准文本进行修改完善，形成《人工智能 智能体能力分级与评测方法》（征求意见稿）及编制说明。

三、 编制原则

本标准编制符合以下原则：

（一）规范性原则

本文件符合国家和本市现行法律、法规和规范性文件；并符合GB/T 1.1-2020 的起草要求。

（二）前瞻性原则

标准编制立足智能体产业当前发展实际，兼顾技术与应用的未来演进趋势，预判行业发展方向与潜在需求。衔接国际相关标准与前沿技术成果，在能力分级、评测方法等核心内容上预留拓展空间，避免标准滞后于产业发展，确保标准具备长期指导价值与适用性。

（三）科学性原则

评测体系设置严格遵循客观规律，围绕智能体能力分级与评测核心，构建逻辑清晰、层次分明的体系框架。能力分级维度、评测指标、实施流程等均基于实际产业调研、实证分析与学术研究，确保体系科学合理、可验证、可追溯。

（四）适用性原则

标准编制聚焦各行业智能体实际业务需求，立足产业应用痛点，避免脱离实际的抽象化设定。核心内容贴合智能体业务应用场景，评测方法、分级要求具备较强可操作性，可适配不同领域、不同类型智能体的评测需求，服务于产业实践。

四、 主要内容

（一）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。

（二）规范性引用文件

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）

（三）术语和定义

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。[来源：IS0/IEC 22989:2022，3.1.1，有修改]。

注:本文件涉及的智能体仅指运行在设备上的软件实体。

（四）基本原则

包括价值导向原则、聚焦业务原则、客观公正原则和独立可控原则。

（五）智能体能力等级模型

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，能力等级分为四个等级，自低向高分别为L1级基础级（L1）、辅助级（L2）、自主级（L3）、协同级（L4）。具体如下：

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。

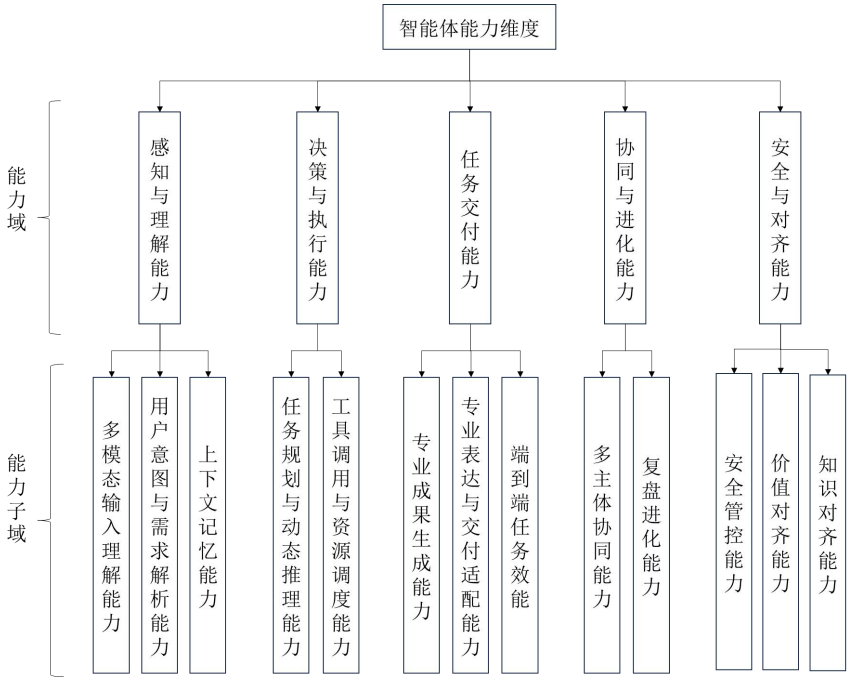
辅助级（L2）：智能体可在预设场景自动感知，理解特定场景内的常规指令与简单非结构化需求，在既定流程和工具范围内自主完成任务，无需人类逐步骤指导，但无法脱离预设流程，仅能为人类提供流程化辅助支撑，场景变化时需人类重新配置。

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。

2.能力维度

智能体的能力维度由能力域和能力子域构成，如下图所示。



3．能力等级要求

列出了各个能力等级对应的能力子域的要求。

（六）评测方法

1．评测流程

智能体能力等级的核心评测流程包括：评测集构建、评测执行、结果判定、结果呈现。

2．评测集构建

给出了评测集构建的原则：

a）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。

b）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：

c）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。

d）评测数据的数量应满足评测全面性要求，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。

给出了评测集构建方法：

根据评测场景差异，评测场景分为已明确目标等级和未明确目标等级两类，需根据评测场景分类进行智能体能力等级的评测集构建。

3．评测执行

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。并通过自动化工具或人工方式，进行评分。

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。

a）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。

b）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。

4．结果判定

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。

根据评分进行等级判定。

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：评测结果、评测方案、评测过程记录、结果分析、优化建议等。

五、 重大分歧意见的处理结果及理由

本标准在修订过程中无重大分歧意见。

六、 标准作为强制性或推荐性标准发布的意见

推荐性。

七、 推动标准实施的措施建议

本标准发布后将尽快组织宣贯，加大贯彻实施力度。第一，在适用主体中推广应用该标准，形成经验；第二，根据试点经验，复制推广试点经验和标准适用主体范围扩大；第三，广泛收集意见和建议，及时归纳和总结，并不断完善标准，必要时提出标准修订。

八、 其他应予以说明的事项

无。

《人工智能 智能体能力分级与评测方法》标准编制组

2026年2月6日