《智能体全链路动态风险评测方法》

编制说明

标准起草小组

2026 年 10 月

1. 标准范围

本文件规定了智能体全链路动态风险的评测方法，包括评测对象与边界、风险模型、资产发现评测方法、安全扫描执行与判定方法、分类风险评测方法、智能体运行行为风险评测、评测指标与计算方法、评测实施流程与结果报告，以及评测能力与结果互认。

本文件适用于智能体及其 AI 资产的开发运营方、智能体平台与协同平台及智能体网关的供给方、第三方评测机构以及行业应用单位开展的安全评测活动，包括自评测、接入评测与上线后复测、独立评测以及选型与验收评测。本文件所称全链路，指从资产发现、安全扫描、风险研判到报告交付的完整评测过程；所称动态评测，指在受控评测环境中对处于运行状态的 AI 资产施加只读探测与行为观测，并依据确定性规则、探针与指纹命中等证据形成可复现结论的评测方式。本文件不适用于智能体安全功能的设计与实现要求、大模型服务与应用层的安全评测、模型算法内生安全性与模型鲁棒性评测、语料集与训练数据质量评测、评测机构资质认定等五类内容，不适用范围逐条列于第 5.2 条评测边界。

本文件的主要用途包括:

（1）产品自测与整改: 为智能体及其承载资产在开发、接入、上线各环节开展安全自测提供统一的资产识别、

风险分类、判定规则、指标口径与报告范式，支撑上线前整改与上线后复测;

（2）第三方独立评测与结果互认：为第三方评测机构开展智能体安全评测提供可复现的判定依据与统计口径，并通过明确的指标门槛与互认条件，支撑不同机构对同一对象给出可比较的结论;

（3）平台准入与行业选型：为智能体平台、协同平台与网关的供给方开展接入评测，以及行业应用单位在选型、验收环节开展评测提供方法依据;

（4）要求类标准的验证支撑：为智能体相关要求类标准所规定的能力条目提供方法层的验证手段，使"要求"可被检测、"结论"可被核对。

2. 工作简况

本文件由中国互联网协会提出并归口，广州市云山人工智能安全研究院、广东著一智慧科技有限公司等单位共同起草。2026 年，起草组向中国互联网协会提交团体标准项目建议书及标准应用落地能力评估材料，经立项审核专家函审通过后，由中国互联网协会于 2026 年 10 月发布立项公告，本文件正式进入起草阶段。

起草过程分为四个阶段。第一阶段为预研与技术验证阶段：起草组基于智能体安全评测平台的研发与运行实践，完成资产发现、确定性探针、双引擎交叉核验、级联资产测

绘、金标回归与覆盖可见等核心能力的工程化，并在相关智能体上开展实际评测，形成可复用的判定口径与指标口径。第二阶段为框架设计阶段：围绕评测对象、边界、风险模型、判定规则与指标五条主线确定标准骨架，明确以资产为观测单元、以确定性判定为方法主线的技术路线。第三阶段为文本起草阶段：形成标准草案，并同步编写评测用例库要求与评测环境参考两个附录。第四阶段为立项评审与修改阶段：根据立项审核专家意见对草案逐条修改，形成本征求意见稿。

立项评审阶段共收到两位专家的 11 条意见，集中于术语定义表述、缩略语精简、评测对象枚举一致性、章节名称、评测框架示意图、范围条款范式、英文名称与统一用例形式等方面。起草组对全部意见逐条研究处理，其中 10 条完全采纳、1 条部分采纳，处理情况详见《团体标准意见汇总处理表》。

3. 标准编制原则和确定标准主要内容

3.1 标准编制原则

本文件依据 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》的规定编制。编制过程中主要遵循以下原则：

一是科学性原则。以智能体运行期的真实风险为对象，按资产承载层与运行行为两条线索建立风险模型，使风险编

号、评测用例与指标口径三者一一对应，避免风险分类与检测方法脱节。

二是可复现原则。这是本文件区别于既有评测类文件的核心原则：全部风险结论以可复现的规则匹配、探针响应与指纹命中为依据，逐条附判定依据与证据；大模型仅用于研判分级与解读生成，不参与判定。由此使不同机构对同一对象、同一版本的评测结论可以比较与互相核对。

三是覆盖可见原则。评测过程中模块失败或部分覆盖时，应如实标注状态与原因，技术失败单独统计且不计入指标分母，不得将"未测到"呈现为"未发现风险"。该原则针对当前评测实践中普遍存在的结论虚高问题。

四是实用性原则。条款以评测活动可直接执行的动作组织：给出资产发现的三种模式与三层识别管线、扫描深度的三级划分、判定规则的四项约束、指标的六个计算公式，并配套规范性附录规定用例库要求，使标准可直接用于评测方案编制与报告复核。

五是协调性原则。与现行网络安全、人工智能相关国家标准和协会已发布团体标准保持衔接，在范围与边界条款中明确划出与要求类标准、服务与应用层评测标准的分工，不与法律法规和强制性国家标准相冲突。

3.2 标准主要内容

本文件主要内容如下：

（1）范围、规范性引用文件、术语和定义、缩略语

明确文件的适用对象与不适用范围，界定全链路、动态评测、确定性判定、判定依据、密库引用、确定性探针、金标回归、有效判分率、评测环境矩阵等 15 个术语，统一评测活动的基础概念，避免同一表述在不同机构产生不同理解。

（2）评测对象与总体框架

将评测对象确定为十类 AI 资产，按接入形态分为网络面（推理服务、活体智能体网关）、制品面（容器镜像与制品、模型文件、Agent Skills 包、MCP 配置、智能体配置）与运行面（MCP 服务端点、代码与凭据、依赖目录与 SBOM）；规定以“智能体标识 + 版本 + 运行配置”三元组唯一确定评测对象，配置变更后原结论不再适用。同时给出四层十五种风险模型（R1.1～R4.3）、用例层—执行层—判定层—指标层四层评测框架及框架与边界的关系图（图 1）、基础级与增强级两级评测等级及判定基准，以及与要求类标准的映射、报告有效性、口径一致、范围声明四项衔接要求。

（3）资产发现评测方法

规定发现任务与扫描任务分离、全程只读、外网授权门控、发现不出具漏洞结论四项通用要求；给出网段发现、端点发现、主机发现三种模式与存活探测、正向指纹、服务识

别三层识别管线；规定暴露画像分的四项封顶计分规则（0～10 分）及其"仅用于排序、不作风险量化结论"的定位；规定发现台账登记与资产血缘，以及发现报告的要素构成。

（4）安全扫描执行与判定方法

规定单任务队列、四种构建任务方式、计划预览与任务参数落库等执行要求；给出 L1～L3 三级扫描深度与模块调度口径，并明确评测任务应全程采用 L3 深度；规定级联扫描的三类顶层资产与凭证继承、谱系留存、跨层关联、安全纪律四项要求；规定确定性判定、覆盖可见、技术失败单独统计、归并口径四项判定规则；规定证据、判定依据、运行日志、操作日志与产物归档五类记录要求。

（5）分类风险评测方法

按基础设施层、模型层、应用层、数据层四层十五种风险逐一给出建模要点、用例构造与判定规则，覆盖镜像与依赖已知漏洞、容器运行时配置缺陷、组件版本已知漏洞、模型文件完整性破坏、推理服务暴露面、MCP 服务威胁模式、MCP 配置缺陷、MCP 工具冲突、Agent Skills 供应链风险、智能体配置基线缺陷、活体智能体网关行为风险、工具调用链与对话面风险、凭据泄露、敏感数据泄露、知识库投毒与违规内容。

（6）智能体运行行为风险评测

从运行行为视角按端到端运行链路组织评测，覆盖输入与目标控制、规划与决策、知识与记忆、工具与外部依赖、多智能体协作、输出与数据处置六个环节，并规定动态执行五项要求；给出五类 11 项运行行为风险（AIA01～AIA11）及其与资产视角风险编号的映射规则；按领域大类组织沙箱环境场景；规定结果运用与记录要求，明确模型层与应用层风险应分别统计与呈现。

（7）评测指标与计算方法

建立输出层、覆盖层、质量层、画像层四层指标框架，给出检出率、误报率、漏报率、风险覆盖度、资产覆盖率、模块覆盖率、级联链路完整率、证据附着率、判定可复现率、有效判分率等指标的计算公式，并明确计数单位、严重度与置信度分离、指标保留位数、分项报告与综合评分的关系、平均值与极差并报等统计口径。

（8）评测实施流程与结果报告

规定准备、实施、判定、报告、复测五个环节；规定四项否决情形与增强级结论的四项门槛（有效判分率不低于 95%、用例覆盖率不低于 80%、关键模块金标回归一致率 100%、证据附着率不低于 99%）；规定报告要素及"结论必须回溯至风险编号、任务编号与发现编号"的可追溯要求；规定复测触发条件与有效期。

（9）评测能力与结果互认

以推荐性要求规定评测机构的能力条件，以及一份报告被另一互认方采信应同时满足的六项条件（同一版本标准、用例库版本一致或已确认等效、判定规则与阈值一致、环境等效性说明、报告要素完备、抽样复核一致率不低于95%），并规定互认声明的内容构成。

（10）附录

附录 A 为规范性附录，规定评测用例库的覆盖度矩阵、要素完备性、有效性验证、样本管理与更新机制；附录 B 为资料性附录，给出五条评测用例示例；附录 C 为资料性附录，给出判定对象与留痕要素、确定性探针观测类型、隔离沙箱与样本库、评测环境矩阵、评测平台能力与风险编号映射等参考内容。

4. 主要试验（或验证）的分析、综述报告

本文件的方法已在相关智能体、以及若干开源智能体与 MCP 服务样本上开展验证性评测，验证重点与结论如下:

（1）资产识别与级联测绘的有效性。在含网关、容器与 MCP 服务端点的混合环境中，三种发现模式均可完成识别并登记台账，级联扫描可将上层资产缺陷定位到下层具体资产，资产血缘关系可追溯，验证了第 5 章与第 6 章对象界定与级联要求的可执行性。

（2）确定性判定的可复现性。同一对象在同一用例集、同一判定口径下多次评测，判定结果保持一致；金标回

归中缺陷样本命中、加固样本不命中，验证了规则、探针与指纹的区分度，支撑第 7.4 条确定性判定规则与第 10.4 条判定可复现率指标。

（3）覆盖可见与统计口径的可操作性。对目标不可达、凭据失效等情形，平台将其记为技术失败并单独统计，不计入指标分母；有效判分率与用例覆盖率可稳定计算，验证了第 10 章指标与第 11.4 条门槛的工程可行性。

（4）运行行为评测的可执行性。对具备原生工具调用能力的智能体，多轮交互可在受控环境中完整走通输入、规划、记忆读写、工具调用与结果处理链路，工具调用参数与任务轨迹可留痕并回放，验证了第 9 章评测链路与动态执行要求的可执行性。

上述验证属起草组内部验证与单点实践，尚未形成覆盖多机构、多对象的独立第三方比对试验综述报告。征求意见阶段及标准实施后，拟结合智能体平台、行业应用单位与第三方测评机构的实际评测数据，补充开展多机构比对验证，重点验证不同评测环境下判定一致性与指标可比性，完善试验验证材料。

5. 标准在起草过程中遇到的问题及解决办法；重大分歧意见的处理经过和依据；有无重要技术问题需要说明

起草过程中主要面临六个问题。

一是"全链路"的边界如何界定。若按生命周期理解，全链路将覆盖模型训练、数据集构建等环节，与其他标准重叠且不可评测。解决方式是在术语 3.3 中将全链路界定为评测活动自身的链路（资产发现、安全扫描、风险研判、报告交付），并加注明确不包含模型训练与数据集构建；第 9 章另行界定智能体运行链路（输入理解至输出处置）并说明两者关系，使"全链路"在文件中始终有确定所指。

二是判定应否引入大模型。评测实践中普遍以大模型作为裁判，其结论受模型版本、提示与随机性影响，不可复现。经讨论确定以确定性判定为方法主线，大模型仅用于研判分级与解读生成，并在第 5.2 条边界约束、第 7.4 条判定规则与第 9.1 条动态执行要求中三处一致体现，同时在术语 3.6 中给出确定性判定的定义。

三是评测对象枚举前后不一致。审议中发现术语 3.2 与第 5.1 条对资产类数的表述分别为十类与十一类，差异来自 MCP 服务与 MCP 服务端点、依赖目录与 SBOM 的口径不同。解决办法是统一术语与第 5.1 条的枚举口径，经征求意见稿修订校核将资产枚举最终确定为十类，并在术语与第 5.1 条采用完全相同的枚举与排序，避免再出现两处口径不一。

四是章节名称与标准名称重复。第 9 章原名与文件名称高度重合，易使读者误解为标准的总则章。解决办法是将

第 9 章改名为"智能体运行行为风险评测"，并同步修改章首语、表 3 标题与相关引用，使章名准确指向该章的观测视角。

五是资产视角与行为视角两套风险编号的关系。文件同时存在 R 编号（四层十五种）与 AIA 编号（五类 11 项），易被认为重复建模。解决办法是在第 9 章章首语明确其与第 5.3 条资产视角风险模型的衔接，并在第 9.4 条结果运用与记录要求中规定两套编号的映射要求，明确 R 编号用于资产扫描与分类评测的组织、AIA 编号用于运行行为用例的构造与结果统计。

六是评测规则如何统一。若由各评测机构自行构造用例，判定口径必然发散。解决办法是在第 5.4 条规定评测应以统一的用例（测试例）形式组织实施，用例的要素、覆盖度、有效性验证与样本管理在附录 A 中统一规定，第 6 章至第 9 章给出的用例构造与判定规则为对应风险类别用例的设计规格；同时说明本文件不统一用例文本，原因是评测对象与风险面差异较大，统一文本会显著削弱适用性，统一要素与判定口径已足以保证结论可比。

在起草与立项评审过程中未出现重大原则性分歧。立项审核专家提出的 11 条意见已全部处理:

（1）术语定义表述类：将"密库引用"的定义由实现方式描述改为术语本质描述，实现细节不再写入术语，另加注说明引用名书写格式由评测环境约定。

（2）缩略语类：删除正文未使用的 DSP、LLM、SCN、WS 四项，第 4 章缩略语由 24 项精简为 20 项，保留正文实际使用的各项。

（3）对象枚举类：统一术语 3.2 与第 5.1 条对评测对象的枚举口径（含类数与排序），经征求意见稿修订校核将资产枚举确定为十类，并在两处采用完全相同的枚举与排序。

（4）章节名称类：第 9 章改为"智能体运行行为风险评测"。

（5）对象关系类：在第 9 章明确运行行为视角与第 5.3 条资产视角风险模型的关系，并规定 R 编号与 AIA 编号的映射要求。

（6）英文名称类：封面英文名称改为 Full-Chain Dynamic Risk Evaluation Methods for AI Agent。

（7）范围条款范式类：第 1 章改为“本文件规定了……本文件适用于……”的规范表述，原四类适用主体的分列改为一句表述，不适用范围在第 5.2 条评测边界逐条列出。

（8）术语中英对照类：仅"智能体"保留英文对应词并改为 AI agent，其余术语不再逐条给出英文对应词。

（9）确定性探针定义类：删除"向被测对象发送的"表述，改为"以只读方式获取被测对象可复现观测事实"。

（10）示意图类：在第 5.4 条新增图 1"评测框架与评测边界示意"，并在正文增加图引用。

（11）统一测试例类：部分采纳，见前述第六个问题的处理方式。

当前标准不存在尚未解决的重大技术问题。需要说明的一项技术判断是：本文件采用的确定性判定方法以资产与配置事实为主要证据来源，对于完全依赖语义理解的模型内生安全问题（如幻觉、价值观偏差），本文件不作为判定对象，此类问题由模型层评测文件规定，本文件仅在运行行为层面观测其可被判定的外部表现。

6. 与国外标准的关系：包括采用国际标准和国外先进标准的程度，与国外标准主要技术内容的差异

本文件未采用国际标准或国外先进标准文本，不存在采标关系。

国外相关文件主要分布在三个方向：一是治理与风险管理框架，如 ISO/IEC 42001 人工智能管理体系、NIST AI 风险管理框架，给出组织层面的治理要求与风险管理工作流，不提供可执行的评测方法与判定规则；二是模型与应用

的攻击面分类，如 OWASP 大模型应用风险清单、MITRE ATLAS 攻击技术矩阵，给出威胁分类与攻击技术描述，不规定评测对象界定与判定口径；三是测试实践类文件，如世界数字科学院（WDTA）发布的 AI 智能体运行安全测试标准，采用全链路风险分析框架（覆盖输入输出、大模型、知识检索、记忆与工具等链路环节），与本文件在"全链路"的技术取向上接近，但未规定统一的风险编号体系、指标计算口径与结果互认条件。

本文件与上述文件的主要差异是：以资产与运行行为为统一观测对象，给出可复现的确定性判定规则，并配套指标门槛与互认条件，形成"对象—风险—用例—判定—指标—互认"的完整链条；国外文件目前尚未提供同等粒度的可操作方法。因此本文件在起草中参考了上述文件的分类思路，未采用其文本，亦不构成等效或非等效采用关系。

7. 修订标准时，说明与标准前一版本的重大技术变化，并列出所涉及的新、旧版本的有关章条；废止/代替现行有关标准的建议

本文件为首次制定，无前一版本，本条不涉及。本文件不废止、不代替任何现行标准。

7. 说明标准与其他标准或文件的关系，特别是与有关现行法律、法规和强制性国家标准的关系

与国家标准的关系：本文件引用 GB/T 41867—2022《信息技术　人工智能　术语》界定智能体等基础概念，参考 GB/T 45288.2—2025《人工智能　大模型　第 2 部分：评测指标与方法》的指标组织思路，但不重复其面向模型能力的指标；本文件的评测方法可为 GB/T 45654—2025《网络安全技术　生成式人工智能服务安全基本要求》所提要求的落实情况提供验证手段。

与协会团体标准的关系：T/ISC 0146—2026《大模型服务与应用安全评测技术规范》面向大模型服务与应用，以输入—输出为观察单元，处于通用层；本文件面向智能体及其承载资产的运行期全链路动态风险，以资产、配置、运行行为与判定证据为统一观测对象，处于专项层。两者为通用层与专项层的衔接关系，在评测对象、判定依据与指标口径上互补，不构成重复。T/ISC 0123—2026《智能体交互身份与访问控制安全能力要求》、T/ISC 0145—2026《系统智能体权限管理技术要求》等要求类文件规定智能体应具备的安全能力，本文件为其提供方法层验证手段，衔接要求见第 5.6 条；T/ISC 0125—2026《大模型安全护栏能力技术规范》规定护栏应具备的能力，本文件不重复护栏功能要求，仅在推理服务暴露面与工具调用链等条款中将护栏配置状态作为观测事实之一。此外，本文件将 T/SAIAS 059—2026《智能体

评测指标与方法》列入参考文献，作为智能体评测指标与方法类参考文件。

与法律法规的关系：本文件内容符合《网络安全法》《数据安全法》《个人信息保护法》《生成式人工智能服务管理暂行办法》等法律法规要求。本文件在数据层风险评测中要求敏感数据与凭据不得进入评测证据、报告与日志，与上述法律关于个人信息与数据处理的规定一致。

与强制性国家标准的关系：本文件为推荐性团体标准，不与现行强制性国家标准相冲突。GB 45438—2025《网络安全技术　人工智能生成合成内容标识方法》为强制性国家标准，本文件不规定标识的技术方法，仅在运行行为评测的输出与数据处置环节将标识与来源追溯的落实情况作为观测点，并引用该标准作为判定依据，不构成对强制性国家标准的替代或解释。

8. 标准作为强制性标准或推荐性标准的建议

建议本文件作为推荐性团体标准发布实施。

9. 贯彻标准的要求和措施建议；标准发布后，对国内外业界可能产生的影响

贯彻标准建议采取四项措施：一是宣贯培训，围绕评测对象界定、判定规则、指标计算与报告要素组织专题培训，重点讲清"确定性判定"与"覆盖可见"两项方法纪律；二是试

点应用，选择智能体平台、政务与能源等行业应用单位、第三方测评机构各若干家开展试点评测，形成典型案例与实施指南；三是配套建设，同步建设与本文件风险编号对应的公共评测用例库与报告模板，降低实施门槛；四是互认衔接，推动评测机构在协会框架下就判定口径与指标门槛达成一致，逐步形成结果互认机制。

标准发布后可能产生的影响有三方面。对产业侧，为智能体上线前自测与平台准入提供统一方法，减少"自证安全"与"结论不可比"造成的重复评测成本；对评测侧，明确判定依据与统计口径，使不同机构的结论具备横向可比性，抑制以主观研判替代评测的做法；对治理侧，为要求类标准与强制性国家标准所提要求提供方法层验证手段，支撑监管抽查与行业自律。国际上，本文件提出的确定性判定与覆盖可见口径可为智能体评测的国际标准化讨论提供可参照的技术方案，但不构成对国外框架的替代。

10. 标准是否涉及知识产权的情况说明；如标准中含有自主知识产权，说明产品研发程度、产业化基础及进程

目前未发现本文件涉及必须披露的专利等知识产权内容。文件中涉及的资产发现、确定性探针、级联扫描与金标回归等能力，均以通用方法形式表述，不指向特定产品的实现细节，不作为对特定产品的推荐。如后续在标准实施、评

测工具与用例库建设中涉及自主知识产权成果，将按规定另行补充说明。

11. 其他应予说明的事项

无。