ICS 35.xxx
CCS Lxx

团　　体　　标　　准

T/ISC XXX—XXXX

可信互联网智能体能力评估规范

Capability Assessment Specification for Trusted Internet Agents

（征求意见稿）

XXXX－XX－XX 发布　　XXXX－XX－XX 实施

中 国 互 联 网 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》的规则起草。

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/本部分由中国互联网协会提出并归口。

本文件起草单位：中国信息通信研究院。

本文件主要起草人：郭雪、卫斌、马铭洋、景韩愈。

引　言

随着生成式人工智能、智能决策与自动执行技术持续演进，互联网智能体正从单一问答工具加速迈向具备联网检索、任务规划、工具调用、数据处理和协同执行能力的综合服务形态，广泛应用于办公协作、代码辅助、生活服务、系统运维、内容生产等多类场景，逐步成为智能互联网的重要服务载体。

与此同时，互联网智能体在快速发展过程中，也面临真实性、可信性与安全性方面的突出挑战：一是企业主体信息、智能体基本信息披露不充分，服务来源、责任主体和能力边界不清晰；二是功能说明、性能指标、服务可用性承诺与实际运行效果之间可能存在偏差，影响用户判断与服务信任；三是智能体在内容生成、权限申请、数据访问、任务执行、系统操作等环节存在越权调用、错误执行、敏感信息泄露和行为失控风险；四是联网访问、外部交互、插件组件调用、多智能体协同及动态策略变更等复杂运行过程缺乏统一安全约束与审计机制，难以满足高敏感场景下的治理要求。

为规范互联网智能体的部署运营与能力评估，提升互联网智能体在信息披露、能力验证、行为约束和运行防护等方面的可信水平，本文件立足互联网智能体全生命周期治理需求，围绕“信息真实性、能力真实性、行为可信性、运行安全性”四个维度，系统构建覆盖文档审查、功能演示、安全测试、权限校验、边界管控、过程审计的技术要求框架，明确不同场景、不同权限等级下的能力要求和安全要求。

本文件旨在为互联网智能体开发者、服务提供方、平台运营方、评估机构及相关管理单位提供统一、可操作的技术参考和评估依据，推动互联网智能体形成信息真实透明、能力描述准确、执行过程可信、运行全程安全的治理体系，促进互联网智能体产业规范有序发展。

可信互联网智能体能力评估规范

1 范围

本文件定义了可信互联网智能体的能力评估指标，包括四方面：一是企业信息真实性披露包括企业信息和互联网智能体信息；二是互联网智能体能力真实性，具体指智能体功能有效性、性能可靠性和服务可用性；三是互联网智能体行为可信性，按操作权限划分为用户级操作可信与系统级操作可信，结合不同技术类型与应用场景实施分级管控；四是互联网智能体运行安全性，包含基础安全、交互安全、安全审计三类管控方向，覆盖智能体全生命周期运行安全。

本文件适用于互联网智能体的设计开发、部署运营、合规测评与监督治理，适用对象包括互联网智能体开发者、服务提供商、平台运营方、第三方评估机构，以及金融、医疗、工业、政务等各行业应用单位与相关管理部门。本文件根据互联网智能体的权限等级、服务场景与业务类型差异，设置差异化的技术指标要求。

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

3 术语和定义

GB/T 41867-2022界定的以及下列术语和定义适用于本文件。

3.1 智能体 agent

能感知环境，并主动采取行动以实现特定目标的实体。

注：一般为软件系统。

3.2 互联网智能体 Internet agent

互联网智能体是对内提供联网搜索功能或对外提供互联网服务，能够自主执行任务，与其他应用、工具或智能体交互协作的软件系统。

4 缩略语

下列缩略语适用于本文件：

API：Application Programming Interface，应用程序编程接口

ISO：International Organization for Standardization，国际标准化组织

SLA：Service Level Agreement，服务等级协议

5 企业基本信息和互联网智能体基本信息披露

企业应如实披露企业基本信息和互联网智能体基本信息。其中，企业基本信息包括企事业法人证书、规模、人员等；互联网智能体基本信息包括名称、版本号、起始时间等。

真实性披露细则要求如下，详见表1：

表 1 企业基本信息和互联网智能体信息披露表

项目	是否必选	需披露的材料
企业基本信息		
企事业单位法人证书	必选	企事业单位法人证书
规模	必选	公司整体社保缴纳信息证明
组织结构	必选	组织机构证书
连续几年纳税证明	可选	纳税证明
互联网智能体信息		
互联网智能体名称	必选	服务基本信息
互联网智能体版本号	必选	正式对外运营时间
功能说明	必选	服务产品手册
起始时间	必选	线上或线下

6 互联网智能体能力真实性

6.1 能力要求方法

本维度考察互联网智能体服务协议及配套文档中承诺的功能、性能、可用性等指标的完备性、规范性与真实性。

能力要求细则：

a) 采用文档审查方式，审核企业提供的功能说明书、性能测试报告、服务协议（含SLA）及其他公开文档，确认相关指标描述的完整性、规范性和一致性；

b) 采用功能演示方式，按照互联网智能体业务功能手册逐一操作，验证所有承诺功能可正常使用且与实际描述相符；对响应时效性、请求吞吐量、系统稳定性等性能指标进行实测，验证测试结果与文档描述的一致性。

6.2 功能有效性

本维度考察互联网智能体服务协议及配套文档中产品功能约定内容的完备性与规范性，指标指向产品手册载明的全部基础与可选业务功能。

能力要求细则：

a) 应按照互联网智能体业务功能手册进行操作，所有承诺功能均可正常使用；

b) 功能演示应与功能说明书描述一致。

6.3 性能可靠性

本维度考察互联网智能体性能相关指标描述的完备性与规范性，覆盖响应时效、并发吞吐、长期运行稳定性相关约定内容。

能力要求细则：

a) 应展示互联网智能体性能测试流程及报告，测试指标包括响应时效性、请求吞吐量、系统稳定性等；

b) 性能测试结果应与互联网智能体说明文档或服务清单中的描述相符。

6.4 服务可用性

本维度考察互联网智能体服务可用性条款及服务效果指标内容的完备性与规范程度，服务提供方应结合自身产品明确相关定义并形成书面约定。

能力要求细则：

a) 应展示互联网智能体服务可用性承诺，在相关协议中明确定义可用性，明确可用性概率数值、计算方式、技术指标含义、统计周期、赔偿方案等；

b) 应展示互联网智能体服务效果与准确性指标，包括答案正确性、任务完成率、模型幻觉率等，并与实际服务场景相匹配。

7 互联网智能体行为可信性

7.1 能力要求方法

本维度考察互联网智能体在完成任务过程中的行为可信，依据互联网智能体的类型差异，本文件针对不同场景其技术指标要求略有不同。

能力要求细则：

1)技术维度：依据智能体的交互模式与操作权限，分为：

a) 信息问答/内容生成类互联网智能体，仅执行文本问答交互，不涉及文件读写、应用调用、系统设置等能力。应满足本维度所规定的指标能力要求；

b) 任务执行类互联网智能体，一类是具备用户级操作权限，能够读取用户个人文件、执行用户指定的简单任务（如日程提醒、信息查询等），无法访问系统核心资源或其他用户数据，应满足7.2用户级操作可信所规定的指标能力要求；

c) 任务执行类互联网智能体，另一类是具备系统级操作权限，访问系统核心资源、操作系统接口、硬件设备等，能执行系统管理、运维、监控等任务，应满足7.2用户级操作可信和7.3系统级操作可信所规定的指标能力要求。

2）场景维度：依据智能体面向的产品功能与服务场景，分为：

a) 文档处理类：面向办公、法务、金融、学术等专业文档场景，以长文本理解、结构化生成、格式规范输出为核心。该类智能体以文本问答、内容生成为核心，应满足本维度所规定的指标能力要求；

b) 代码辅助类：面向软件开发全生命周期，覆盖前后端、算法、运维等技术栈，以代码理解、生成、调试、优化为核心。该类智能体为用户提供代码文本建议，应满足本维度所规定的指标能力要求；

c) 生活服务类：面向日常生活服务场景，如餐饮订购、商品购买、出行规划等，以自然语言交互、服务调用、订单处理为核心。若仅提供商品信息推荐、服务介绍及使用咨询，应满足本维度可信要求；若涉及用户账户数据访问、订单指令执行、外部商家或配送平台调用及支付链路操作，应满足7.2用户级操作可信所规定的指标能力要求；

d) 系统应用类：面向系统运维、服务器监控、安全审计、资源调度等技术基础设施场景，以系统级操作、自动化运维、异常检测为核心。若仅提供运维知识问答、故障排查建议及操作指南咨询，应满足本维度可信要求；若涉及业务系统日志读取、监控数据查看、非核心配置文件修改及常规运维工具调用，应满足7.2用户级操作可信的要求；若涉及系统核心资源访问、操作系统底层接口调用、内核配置修改、核心进程管理、硬件设备控制或安全审计权限操作，应满足7.2用户级操作可信和7.3系统级操作可信所规定的指标能力要求。

7.2 用户级操作可信

7.2.1 数据访问边界

本维度考察互联网智能体是否仅能访问已被授权的用户数据，是否实现不同用户间的数据隔离，以及是否具备敏感数据识别能力。

能力要求细则：

a) 应验证智能体仅能读取或操作当前用户已授权的数据，无法跨用户访问其他用户的数据；

b) 应验证不同用户会话之间的数据存储和访问相互隔离，无数据混叠或泄露风险；

c) 应具备识别敏感数据类型（如身份证号、银行卡号、生物特征等）的能力，对导出或共享敏感数据的操作实施额外授权校验。

7.2.2 行为可干预

本维度考察用户能否在互联网智能体执行任务过程中随时暂停、回退或终止其操作。

能力要求细则：

a) 应提供明确的操作控制界面（如暂停、停止、回退按钮），用户可随时中断智能体正在执行的任务；

b) 应支持用户回退至操作前的状态（如提供撤销功能或备份恢复机制）；

c) 应提供“一键关停”机制，管理员可立即中止智能体所有正在执行的高敏感任务。

7.3 系统级操作可信

7.3.1 权限申请与开放

本维度考察互联网智能体在执行高敏感系统操作、管理关键系统服务、访问底层资源与接口时，是否能够向用户充分说明操作目的、影响范围和调用方式，并在相关操作过程中体现授权透明、使用规范和结果可核验。

能力要求细则：

a) 应测试智能体在发起高敏感系统操作（如静默安装应用、读取其他应用的私有数据、修改系统安全设置）前，必须弹出明确的授权请求，说明操作内容和风险，并获得用户显式点击“允许”后方可执行；

b) 应对系统关键进程（如init、systemd、安全监控进程）及核心服务实施访问控制，仅执行经明确授权的进程管理操作（如启动、停止、重启）；

c) 应展示智能体访问系统底层资源（如内核接口、硬件抽象层）所使用的API或通道，并说明其安全合规性；

d) 禁止调用未公开、未备案的私有系统接口，对未知接口调用请求直接拦截。

7.3.2 高风险知情确认

本维度考察互联网智能体针对等高风险操作是否设置多层防御机制，是否具备操作风险识别能力，以及是否提供回滚或恢复机制。

能力要求细则：

a) 应对“删除系统文件”“系统配置修改”“恢复出厂设置”等高风险操作设置至少两层防御措施，例如用户语音二次确认、动态验证码、生物识别等，防止单点失误或恶意利用；

b) 应具备操作风险识别能力，在影响系统稳定性或安全性的操作执行前主动评估风险并提示用户确认；

c) 应对系统配置变更类操作提供完整的回滚或恢复机制，确保系统可恢复至操作前状态，并定期演练回滚流程；

d) 应完整记录所有系统资源访问操作，包含时间、进程名、操作类型、参数等信息，支持安全审计人员查阅。

8 互联网智能体运行安全性

8.1 能力要求方法

本维度考察互联网智能体在运行、联网、调用工具、处理数据、协同执行和动态变更过程中的安全控制能力。依据互联网智能体的服务场景、权限范围、部署方式和业务类型差异，本文件针对不同场景的技术指标要求略有不同。

能力要求细则：

a) 采用文档审查方式，审核企业提供的安全策略、权限配置、组件清单、日志记录、应急预案及其他相关材料，确认安全要求描述的完整性、规范性和一致性；

b) 采用功能演示和安全测试方式，对互联网智能体的身份认证、访问控制、工具调用、联网访问、数据处理、异常拦截和应急处置等能力进行验证，确认各项安全措施可有效执行且与实际描述相符；

c) 对政府、金融、医疗、工业等高敏感场景，应结合业务影响、数据敏感程度和操作权限等因素开展重点核验，验证差异化安全措施与实际运行情况的一致性。

8.2 基础安全

8.2.1 环境设施防护

本维度考察互联网智能体运行环境、部署基础设施和相关接口的基础安全防护能力。

能力要求细则：

a) 应对主机、容器、云环境、接口服务、开放端口和访问路径进行安全配置核查，避免未授权暴露、默认配置运行或高风险接口开放；

b) 应控制互联网智能体访问网络、远程管理通道和内部资源的范围，限制与业务无关的网络可达性和资源访问；

c) 应对关键配置、认证凭据和运行依赖实施访问控制，不得以明文方式在代码、配置文件、公开目录或日志中暴露。

8.2.2 身份访问管控

本维度考察互联网智能体在身份认证、分级授权、权限申领与使用全流程管控，以及权限风险防控方面的安全能力。

能力要求细则：

a) 应建立统一的身份认证与分级授权机制，严格遵循最小权限原则，仅为不同用户、业务角色、智能体实例、子智能体分配完成业务所必需的对应操作权限，禁止提前申请未来业务所需权限，实现业务层面与系统层面的权限隔离；

b) 应保证智能体公开声明的权限列表与实际运行时申请、使用的权限保持一致，无未申报的敏感权限，不存在超出业务功能需要的权限请求，如读取无关的个人文件或联系人信息；

c) 执行业务敏感操作或系统高风险操作前，应核验操作主体的身份与对应授权凭证，拦截未授权、超出权限范围的操作请求；

d) 应支持用户随时查看已向智能体授予的权限列表，并可自主撤销或临时授予某项权限，智能体首次访问用户数据（如通讯录、相册、位置）或执行敏感操作前，应弹出明确的授权请求，说明操作目的和范围，并提供 “允许”“拒绝” 选项供用户自主选择；

e) 对敏感数据访问、外部发送、系统配置修改等高风险操作，应在执行前实施严格的访问控制校验，设置多层确认机制，必要时获取用户显式确认；

f) 应对跨系统、跨工具、跨智能体访问建立独立的权限控制关系，防止权限扩散、权限叠加或权限失控。

8.2.3 组件供应管理

本维度考察互联网智能体对工具、技能、插件、连接器、协议服务（如MCP服务）及第三方依赖的安全管理能力。

能力要求细则：

a) 应建立已接入工具、技能、插件、连接器、协议服务（如MCP服务）和第三方组件清单，明确其功能用途、调用权限、数据访问范围和风险等级；

b) 对具备写入、执行、联网、外发或高权限访问能力的第三方组件，应实施独立授权、风险审查和最小权限控制；

c) 应建立供应链变更管理机制，对新增、升级、替换或下线的模型、插件、依赖库和外部服务进行安全核验，防范恶意组件、失效组件或异常调用风险。

8.3 交互安全

8.3.1 指令与内容安全

本维度考察互联网智能体对注入攻击、恶意诱导、违规内容及多模态风险输入的识别与防护能力。

能力要求细则：

a) 应建立全类型输入校验机制，覆盖文本指令、图像、音频、视频、文档附件等多模态输入，可识别提示注入、越狱攻击、编码混淆、变形诱导、角色伪装等可疑输入，按风险等级执行拦截、风险提示或限制处理，不得因多轮铺垫、上下文干扰绕过安全约束；

b) 应建立任务意图审查机制，对套取内部信息、越权访问、伪装授权、规避安全规则等恶意请求进行风险识别和处置；

c) 应建立输出内容安全管控机制，明确违规违法、虚假有害、歧视仇恨等不当内容的判定规则与处置流程，对敏感信息输出具备专项识别、脱敏或屏蔽能力，防止敏感信息通过生成内容泄露；

d) 具备联网检索或知识库引用能力的智能体，应对引用内容的真实性、安全性、相关性进行校验，保证引用来源真实可信，避免引入虚假信息、违规内容或无关信息。

8.3.2 边界与隐私防护

本维度考察互联网智能体对高风险操作的边界约束与敏感数据的全生命周期防护能力。

能力要求细则：

a) 应明确高风险能力的启用条件、使用范围和禁止事项，不得超出预设安全策略实施相关操作；

b) 对系统修改、批量文件处理、个人数据批量导出、资金处理、权限变更、数据外发等高风险操作，应设置明确的触发条件、多层确认机制和人工复核措施；针对涉及用户个人文件批量删除、大量个人数据导出等直接影响用户数据安全的操作，须设置二次确认弹窗，明确告知操作后果并获得用户再次确认；同时建立异常中止机制，在出现越权执行、违规联网或数据泄露风险时及时停止操作；

c) 应分类识别全场景敏感数据，涵盖个人信息、商业秘密等隐私类数据，以及系统令牌、密钥、核心配置等技术类数据，遵循最小必要原则管控数据访问、读取、导出等操作，执行授权校验；

d) 应对批量采集敏感数据等高风险行为开展专项管控，明确敏感数据留存周期、删除与脱敏规则，防止敏感信息在日志、缓存、调试信息中泄露。

8.3.3外部与端侧管控

本维度考察互联网智能体对外部交互行为与端侧运行场景的安全管控能力。

能力要求细则：

a) 应对外部接口调用、资源访问与信息外发建立范围管控机制，优先访问可信来源，限制高风险、低可信及与任务无关的外部资源访问；

b) 应完整记录外部交互的目标对象、调用结果与引用来源，保证交互过程可追溯；对邮件发送、代码下载、外部脚本执行、访问外部网络资源等高风险外部交互行为，应向用户清晰告知目标地址、操作类型、风险及潜在影响，并获得用户明确确认后方可执行；

c) 端侧部署或本地模型运行场景下，应严格管控模型文件、缓存数据与本地接口的访问权限，防止因端侧存储或接口开放导致信息泄露；

d) 调用摄像头、麦克风、位置、通讯录等终端采集能力时，应向用户明示采集目的与范围，并获取用户明确授权。

8.4 安全审计

本维度考察互联网智能体在协同运行、配置变更、运行监测及安全事件场景下的审计留痕、追溯核查与风险告警能力。

能力要求细则：

a) 应对多智能体协同、任务转派、权限调用、数据共享等关键环节留痕，保障执行链路可追溯，支持异常与越权行为核查；

b) 应对提示词、安全规则、模型配置、工具编排等变更建立审计机制，记录变更主体、内容、生效时间与回退情况；

c) 应建立运行安全监测机制，对异常调用、越权访问、违规输出等高风险行为实时告警，留存完整可追溯的关键活动审计日志；

d) 应建立安全事件留痕与复盘机制，记录告警、处置、恢复全流程信息，定期开展风险测试与应急演练。