TC260-PG-2026NA

网络安全标准实践指南

——智能体交互安全要求

（征求意见稿 v0.23—202607）

全国网络安全标准化技术委员会秘书处

2026 年 7 月

本文档可从以下网址获得：
www.tc260.org.cn/

全国网络安全标准化技术委员会
National Technical Committee 260 on Cybersecurity of SAC

前　言

《网络安全标准实践指南》（以下简称《实践指南》）是全国网络安全标准化技术委员会（以下简称“网安标委”）秘书处组织制定和发布的标准相关技术文件，旨在围绕网络安全法律法规政策、标准、网络安全热点和事件等主题，宣传网络安全相关标准及知识，提供标准化实践指引。

本文件起草单位：中国移动通信集团有限公司、中国电子技术标准化研究院、北京中关村实验室、国家计算机网络应急技术处理协调中心、中兴通讯股份有限公司、北京快手科技有限公司、复旦大学、阿里云计算有限公司、启元实验室战略研究中心、蚂蚁科技集团股份有限公司、哈尔滨工业大学。

本文件起草人：邱勤、粟栗、杨凯、李子威、贺敏、崔勇、陈佳、赵宇航、冉鹏、孙洋、李邦玲、鲁冬杰、徐思嘉、张妍婷、张蕾、晁艺涵、杜晨光、石桂欣、李晔、王锟、王博、周继华、梅傲婷、曹晓琦、邵萌、王键、谷晨、杨珉、洪赓、陈沛、彭骏涛、徐源、李韵佳、林冠辰、杨小芳、彭晋、姜伟、叶麟。

声　明

摘　要

为指导人工智能智能体交互安全，防范智能体交互过程中的身份伪造、越权访问、多智能体级联幻觉扩散等安全风险，依据《中华人民共和国网络安全法》《中华人民共和国数据安全法》及相关国家标准，制定本文件。

为此，本指南规定了智能体交互通用安全要求，以及智能体间交互安全要求、智能体与工具交互安全要求，旨在为智能体服务提供者、工具提供者提供交互过程的安全实践指引，也可供第三方评估机构等组织参考。

目　录

1 范围 5
2 规范性引用文件 5
3 术语定义 6
4 缩略语 8
5 概述 8
6 智能体交互通用安全要求 8
7 智能体间交互安全要求 11
8 智能体与工具交互安全要求 14
附录 A（资料性）智能体交互安全风险清单 16
参考文献 23

1 范围

本文件规定了智能体与智能体、智能体与工具的交互安全要求，包括智能体交互通用安全要求、智能体间交互安全要求、智能体与工具交互安全要求。

本文件适用于指导智能体服务提供者、工具提供者保障智能体交互过程的安全，也适用于第三方评估机构等组织对智能体交互安全评估。

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。

GB/T 25069 信息安全技术 术语

GB 45438—2025 网络安全技术 人工智能生成合成内容标识方法

GB/T 45574—2025 数据安全技术 敏感个人信息处理安全要求

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

GB/Z 185.1—2026 人工智能 智能体互联 第1部分：总体架构

GB/Z 185.3—2026 人工智能 智能体互联 第3部分：身份管理

GB/Z 185.4—2026 人工智能 智能体互联 第4部分：智能体描

述

GB/Z 185.5—2026 人工智能 智能体互联 第5部分：智能体发现

ISO/IEC 22989:2022 Information technology-Artificial intelligence-Artificial intelligence concepts and terminology

3 术语定义

3.1 智能体 AI agent

泛指一种自动化实体，即在指定条件下无需人工干预或受控人工干预即可运行，能够感知并响应其环境，并采取行动以实现其目标的流程或系统。

[来源：ISO/IEC 22989:2022, 3.1.1 有修改]

3.2 智能体服务提供者 AI agent service provider

提供智能体服务的组织或个人。

3.3 智能体描述 AI agent description

用以描述智能体名称、功能等，机器及人类可理解的信息。

[来源：GB/Z 185.4—2026, 3.1]

注：在特定系统中，智能体描述一般遵从特定表达形式。

3.4 智能体凭证 AI agent credential

一个包含智能体身份属性声明、防篡改的数据集合，用于身份鉴别。

[来源：GB/Z 185.1—2026, 3.4]

3.5 智能体发现 AI agent discovery

匹配并获得符合业务要求的 1 个或多个智能体描述的过程。

[来源：GB/Z 185.5—2026, 3.1]

3.6 智能体发现服务 agent discovery service

实现或提供智能体发现功能的过程。

[来源：GB/Z 185.5—2026, 3.2]

3.7 智能体身份注册服务 agent identity registration service

处理智能体身份注册请求、执行智能体身份核验并管理智能体身份账户的服务。

[来源：GB/Z 185.3—2026, 3.4，有修改]

3.8 关键信息基础设施 critical information infrastructure

公共通信和信息服务、能源、交通、水利、金融、公共服务、电子政务、国防科技工业等重要行业和领域，以及其他一旦遭到破坏、丧失功能或者数据泄露，可能严重危害国家安全、国计民生、公共利益的重要网络设施、信息系统等。

[来源：GB/T 39204—2022, 3.1]

3.9 工具 tool

可被智能体应用调用的功能实体。

注：工具的示例如云端服务、应用程序、智能体、外部设备等。

3.10 最小特权 minimum privilege

对某一主体，其访问权限仅限于执行所授权任务所必需的权限。

[来源：GB/T 5271.8—2001, 08.04.15, 有修改]

3.11 标识 identity

可唯一确定一个实体身份的信息。

[来源：GM/T 0090—2020, 3.1]

注：标识由实体无法否认的信息组成，如实体的可识别名称、电子邮箱、身份证号、电话号码、街道地址等。

4 缩略语

下列缩略语适用于本文件。

AI：人工智能（Artificial Intelligence）

TLS：传输层安全性协议 (Transport Layer Security)

IP：互联网协议 (Internet Protocol)

5 概述

智能体交互安全要求包括智能体交互通用安全要求、智能体间交互安全要求、智能体与工具交互安全要求，如图 1 所示：

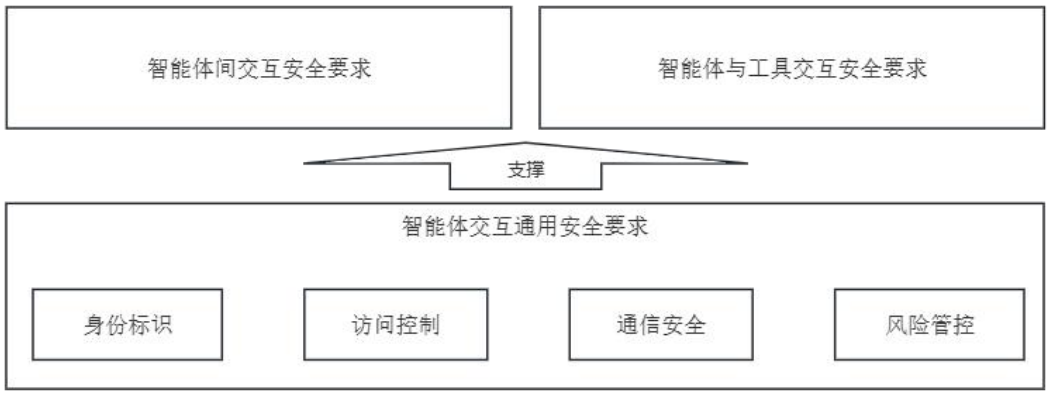

图 1 **智能体交互安全架构**

6 智能体交互通用安全要求

6.1 身份标识

智能体应具有标识，该标识可唯一识别该智能体。智能体应具有与该标识对应的智能体凭证。

6.2 访问控制

智能体交互双方应支持访问控制机制，以决定是否允许访问和中断访问。

6.3 通信安全

智能体交互的通信安全要求如下:

a) 智能体应采用安全的通信信道，该通道支持对访问对象的身份鉴别和提供双方行为的不可否认性。

b) 智能体应采用加密与完整性校验策略，宜采用 TLS 1.2 及以上版本的传输层协议。

c) 处理关键信息基础设施承载的关键业务的智能体，在与交互对象通信过程中，应按照国家密码管理部门与行业有关要求使用密码算法进行身份鉴别、数据加密及完整性校验。

d) 智能体服务提供者应支持流量监控和清洗机制，对异常高频请求进行识别和限流阻断，防止资源耗尽性攻击。

e) 智能体服务提供者宜支持传输路径冗余机制，当主路径受损时自动切换备用路径。

6.4 风险管控

6.4.1 主动防范

智能体交互的主动防范安全要求如下:

a) 智能体服务提供者应为智能体建立交互限流机制，控制单个智

能体在单位时间内的交互对象数量和交互频率。

b) 智能体服务提供者宜支持基于历史交互日志、行为特征库对潜在异常交互行为进行前置研判，对高风险交互请求实施拦截或人工复核。

6.4.2 异常检测

智能体交互的异常检测安全要求如下：

a) 智能体服务提供者应在获得用户授权的条件下，支持智能体交互状态的监控功能，实时跟踪通信时间、交互对象、交互状态、交互行为等信息，及时发现并预警异常行为。

b) 当检测到非收敛交互或通信异常，智能体服务提供者应强制终止该交互，记录异常信息。

注：非收敛交互是指智能体间的对话或操作序列陷入重复、循环或矛盾的状态，无法推进任务，或者由于网络故障、对方智能体崩溃或无响应导致的互联中断。

c) 当智能体服务提供者检测到交互双方的异常行为，应记录异常事件详情。

6.4.3 攻击分析

智能体交互的攻击分析安全要求如下：

a) 智能体服务提供者应针对已发现攻击活动，分析攻击目标、攻击路线。

b) 智能体服务提供者应具备机制，对已识别的攻击行为进行防护。

c) 智能体服务提供者宜建立多智能体异常协同检测机制。

6.4.4 应急响应与处置

智能体交互的应急响应与处置安全要求如下:

a) 智能体服务提供者应制定智能体交互安全事件分级响应预案，明确一般、较大、重大、特别重大事件处置流程。

b) 智能体服务提供者应制定智能体交互应急策略，当智能体交互安全事件发生后，应立即启动应急响应处置措施，结合实际情况快速实施切断通信连接、阻断交互、隔离处置、漏洞修复，对多次发生异常的对象发起二次验证等，形成事件报告与处置记录，留存完整交互日志与证据链，防止安全风险通过交互链条扩散。

c) 较大以上安全事件发生后，智能体服务提供者应及时向主管部分进行报告。

6.4.5 溯源

智能体服务提供者应具备攻击溯源机制，宜根据分析结果实现攻击者画像。

7 智能体间交互安全要求

7.1 智能体注册安全

智能体服务提供者应选择智能体身份注册服务，服务应支持对注

册过程进行如下安全机制，在注册时:

a) 应鉴别智能体身份。

b) 应实现注册监控机制，支持对异常注册的智能体监控、告警、注销。

c) 应对智能体服务提供者实施合规审核，仅允许通过审核的提供者注册其智能体。

d) 应支持智能体身份生命周期管理机制。

e) 应根据智能体描述检测能力真实性和行为合规性，包含但不限于检测能力范围、输入输出类型、提供方和明显的虚假/恶意注入等，推荐引入第三方评估机构实施测试。

7.2 智能体发现安全

智能体交互的发现安全要求如下:

a) 智能体服务提供者应基于规范化的智能体发现机制披露智能体描述。

b) 发现方智能体服务提供者，需要验证发现返回的列表信息的完整性和来源。

c) 智能体服务提供者选择基于服务的智能体发现机制时，应选择支持如下安全机制的智能体发现服务:

1) 应具备防干扰的智能体推荐排序机制;

2) 应支持智能体列表防篡改机制;

3) 应支持恶意查询识别和阻断机制。

7.3 智能体间调用安全

智能体交互的调用安全要求如下:

a) 智能体调用双方应实施双向身份鉴别。

b) 智能体调用双方应通过标准接口或其他形式，交互彼此的智能体描述、安全策略、数据需求及潜在风险，在双方均充分知晓潜在影响的基础上，达成调用约定。

c) 智能体调用双方应遵循最小特权原则，应支持权限协商机制，且权限的有效性可验证。

d) 智能体应确保协作过程中的决策自主性，防止未授权方篡改其决策逻辑或目标。

e) 智能体调用双方在交互过程中，宜支持调用会话的生命周期管理和防止会话劫持机制。

注：会话为每个会话对应一个多智能体交互过程，由请求智能体创建，具有一个编号标识，用于管理该过程中涉及的服务智能体、消息及任务。

f) 智能体调用双方应对输入输出的攻击行为和内容具备安全防护机制，针对关键信息，应具备验证其真实性、完整性的能力。

g) 智能体调用双方应支持识别超出智能体描述范围的异常行为的机制。

h) 交互过程中，被调用方智能体应禁止未经用户授权的启用高危系统权限。

i) 调用方智能体服务提供者应基于用户同意授予的最小特权，授权被调用方智能体应验证该权限是否与用户原始授权一致，并

保存调用日志。

8 智能体与工具交互安全要求

8.1 工具发现安全

智能体交互的工具发现安全要求如下:

a) 工具发现过程中,智能体服务提供者应确保工具列表可验证来源和完整性。

b) 工具提供者应向调用方公开完整且可校验完整性和来源的工具属性信息，至少包含工具标识、工具名称、工具描述、工具输入输出参数等信息。

8.2 工具调用安全

智能体交互的工具调用安全要求如下:

a) 智能体调用工具时应获得用户授权。

b) 智能体调用工具时,应基于双方都支持的身份鉴别模式进行身份鉴别。

注：身份鉴别模式可为如下任意一种：双向身份鉴别，智能体单向鉴别工具，工具单向鉴别智能体，无身份鉴别。

c) 工具提供者应实施最小接口暴露原则,仅向智能体暴露最小必要的功能接口。

d) 智能体调用工具时，应遵循最小特权原则。

e) 智能体基于协议方式调用工具时,应基于双方认可的标准化协议。

f) 智能体应基于明确的业务必要性与风险评估结果，严格限制高危系统权限的申请与使用。

g) 涉及高危系统权限的工具调用时，智能体服务提供者应充分告知用户安全风险，获取用户明示同意。

h) 智能体应支持恶意工具识别机制，识别到被调用工具为恶意工具时不应发起调用行为。

i) 智能体应检测工具反馈信息，识别到恶意内容或恶意行为时停止调用行为。

附　录　A

（资料性）

智能体交互安全风险清单

表 1　智能体交互安全风险清单

编号	风险	风险描述	覆盖章节	解决方案
T01	信息篡改	输入指令篡改、工具列表篡改、调用参数/返回结果篡改	6.3 通信安全 7.3 智能体间调用安全(f) 8.1 工具发现安全（b) 8.2 工具调用安全（i)	1. 采用加密与完整性校验的通信信道（6.3 b）。 2. 智能体间调用时，对关键信息验证其真实性、完整性（7.3 f）。 3. 工具发现时，验证工具列表和工具属性信息的完整性（8.1 a，b）。 4. 工具调用时，检测工具反馈信息，识别恶意内容或行为（8.2 i）。
T02	信息泄露	调用参数/返回结果信息泄露、输出敏感信息、记忆	6.3 通信安全 7.3 智能体间调用安全(b，f)	1. 采用加密通信信道（6.3 b）。 2. 智能体间调用前交互数据需求与安全策略(7.3 b)，对输出内容进行安全防护

		窃取	8.2 工具调用安全 (g)	（7.3 f）。 3. 严格限制高危系统权限的申请与使用，仅在受控环境调用（8.2 g）。 4. 遵循最小特权原则（6.2，7.3 c，8.2 d）。
T03	信息缺失	日志不完整、责任难追溯	6.4.5 溯源 7.3 智能体间调用安全 (i)	1. 智能体服务提供者应具备攻击溯源机制（6.4.5）。 2. 智能体间调用应保存调用日志（7.3 i）。
T04	有害信息	注入有害指令、间接注入、业务逻辑绕过、工具描述投毒、调用参数/返回结果恶意内容、输出有害/恶意内容、传播	6.4.2 异常检测 7.3 智能体间调用安全 (f，g) 8.2 工具调用安全 (h，i)	1. 监控交互状态，检测并终止非收敛或异常交互（6.4.2）。 2. 智能体间调用时，对输入输出进行安全防护，识别异常行为（7.3 f，g）。 3. 工具调用时，支持恶意工具识别机制（8.2 h），检测工具反馈的恶意内容（8.2 i）。

		虚假信息		
T05	身份错误	身份冒用、智能体身份伪造、调用方身份伪造、用户身份混淆、接受方身份不明、会话劫持	6.1 身份标识 7.1 智能体注册安全 (a) 7.3 智能体间调用安全 (a, e) 8.2 工具调用安全 (b)	1. 智能体应具有唯一标识和对应凭证（6.1）。 2. 智能体注册时需鉴别身份（7.1 a）。 3. 智能体间调用必须实施双向身份鉴别（7.3 a）。 4. 智能体调用工具时，应基于双方支持的模式进行身份鉴别（8.2 b）。 5. 支持调用会话的生命周期管理和防止会话劫持机制（7.3(e)）。
T06	权限失控	越权访问、越权执行、环境越狱、运行环境逃逸、越权、系统权	6.2 访问控制 7.3 智能体间调用安全 (c, h, i) 8.2 工具调用安全 (a,	1. 交互双方应支持访问控制机制（6.2）。 2. 智能体间调用遵循最小特权原则，支持权限协商与验证（7.3 c），禁止未经用户授权启用高危权限

		限提升、过度申请权限	d，g)	（7.3 h），验证权限与用户原始授权一致（7.3 i）。 3. 工具调用需获得用户授权（8.2 a），遵循最小特权（8.2 d），严格限制高危权限在受控环境使用（8.2 g）。
T07	资源占用	DoS 攻击、规划混淆、资源超载、系统功能滥用	6.3 通信安全 (d) 6.4.1 主动防范 (a) 6.4.2 异常检测 (a)	1. 支持流量监控和清洗，对异常高频请求进行识别和限流阻断（6.3 d）。 2. 建立交互限流机制，控制单位时间内的交互对象数量和频率（6.4.1 a）。 3. 监控交互状态，及时发现并预警异常行为（6.4.2 a）。
T08	攻击行为	DoS 攻击工具、输出攻击行为、输出恶意代码/脚本、	6.4.3 攻击分析 7.3 智能体间调用安全 (f)	1. 针对已发现的攻击活动进行分析和防护，宜建立多智能体异常协同检测机制（6.4.3）。 2. 智能体间调用时，对输

		多智能体协同攻击、安全事件应急响应与处置机制缺失，导致攻击扩散	8.2 工具调用安全 (i) 6.4.4 应急响应与处置	入输出的攻击行为具备安全防护机制（7.3 f）。 3. 工具调用时，检测工具反馈信息，识别到恶意行为时停止调用（8.2 i）。 4. 应急响应与处置，实施分级预案、切断/隔离/修复、二次验证、留存证据链、上报等处置（6.4.4）。
T09	组件风险	恶意物料、有脆弱性的物料、恶意扩展	7.1 智能体注册安全 (e) （间接通过引用 GB/T 43698, GB/T 44111）	1. 注册时根据智能体描述检测能力真实性和行为合规性，推荐引入第三方评估（7.1 e）。 2. 规范性引用文件中包含软件供应链安全标准（GB/T 43698, GB/T 44111），为防范组件风险提供了标准依据。
T10	意愿劫持	侵害工具选择权、工具排名操	7.2 智能体发现安全 (c, a)	1. 当选择基于服务的发现机制时，应选择具备防干扰的智能体推荐排序机制的

		纵、智能体描述不规范或被篡改导致发现阶段误匹配、调用错对象	8.1 工具发现安全(b)	服务（7.2 c 1））。 2. 基于规范化的智能体发现机制披露智能体描述（7.2(a) ）。 3. 工具属性信息应公开完整且可校验完整性（8.1(b)）。
T11	工具滥用	恶意工具、工具组合攻击、工具降级、非法操作、工具滥用、工具投毒、工具越权执行	7.1 智能体注册安全(e) 8.1 工具发现安全 (a) 8.2 工具调用安全 (g, h, i)	1. 注册时检测智能体能力真实性和行为合规性（7.1 e）。 2. 工具发现时确保工具列表可验证来源和完整性（8.1 a）。 3. 高危工具调用限于受控环境并显式获得用户授权（8.2 g）。 4. 支持恶意工具识别机制（8.2 h），检测到恶意内容或行为时停止调用（8.2 i）。

T14	意图偏离	推理劫持、目标劫持、幻觉放大、意图偏移累积、长上下文攻击、工具选择熵增、任务执行异常	6.4.2 异常检测 7.3 智能体间调用安全(d, g)	1. 监控交互状态，当检测到非收敛交互陷入循环、矛盾）时强制终止并记录（6.4.2）。 2. 智能体应确保协作过程中的决策自主性，防止未授权方篡改其决策逻辑或目标（7.3 d）。 3. 支持识别超出智能体描述范围的异常行为的机制（7.3 g）。

参考文献

[1] GB 45438—2025 网络安全技术 人工智能生成合成内容标识方法

[2] GB/T 34975—2017 信息安全技术 移动智能终端应用软件安全技术要求和测试评价方法

[3] GB/T 35273 信息安全技术 个人信息安全规范

[4] GB/T 39720—2020 信息安全技术 移动智能终端安全技术要求及测试评价方法

[5] GB/T 39276—2020 信息安全技术 网络产品和服务安全通用要求

[6] GB/T 45574—2025 数据安全技术 敏感个人信息处理安全要求

[7] GB/T 45652—2025 网络安全技术 生成式人工智能预训练和优化训练数据安全规范

[8] 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

[9] GB/T 45674—2025 网络安全技术 生成式人工智能数据标注安全规范

[10] 鸿蒙智能体框架白皮书

[11] YD/T 3973—2021 5G 网络切片 端到端总体技术要求

[12] YD/T 6095—2024 基于 SDN/NFV 的新一代网络架构 固移融合专线业务总体技术要求