司法领域智能体技术应用指南

V1.0（2026 年 5 月）

深圳市中级人民法院、深圳市大数据研究院、清华大学互联网司法研究院、
香港大学法律学院
联合课题组

孟天一、刘庄执笔

引言

本指南旨在为法官、司法工作人员及司法人工智能技术研发者，在履行司法职能及开展相关技术活动时，提供智能体技术应用与系统研发的通用准则与指导原则。指南强调智能体等前沿人工智能技术和产品方法应用的辅助性、规范性、安全性及权责一致性，致力于推动人工智能技术与司法工作的深度融合，在提升审判质效的同时，严守司法伦理、数据安全与公平正义的底线。

一、 智能体技术带来的机遇与风险

智能体技术是指能够感知环境、进行决策并执行行动以实现特定目标的计算机程序或系统。在司法语境下，特指为完成特定审判辅助任务而设计的智能化模块或系统。当前的智能体技术，已经能

够基于大语言模型等先进人工智能技术，根据目标自主进行任务规划、分步推理、工具调用与环境交互。其“自主性”体现在为完成任务而动态制定策略与步骤的能力。

2026年以来，以智能体为代表的人工智能技术取得显著进展，其强大的任务规划与执行灵活性为司法领域带来新的赋能机遇，体现在：显著提升效率——可自动、批量处理文书送达、信息录入、简单查询等重复性事务；灵活响应需求——能更低成本、更灵活地响应个性化的案件管理或当事人服务需求；降低研发门槛——通过模块化、可编排的智能体，降低复杂司法人工智能应用的开发与部署难度。

智能体技术在带来便利的同时，也因其高度的自主性与交互性引入了新的、不容忽视的风险，若管控不当，将对司法程序、责任体系与司法公信力造成严重冲击，主要表现在：

（一）权责僭越风险：司法核心职权面临“空心化”

存在核心司法职权被不当转移的潜在可能。法官可能因对智能体能力的错误认知或便利性依赖，过度授权，例如：

结果导向的过度委托：仅下达最终任务指令（如法官直接指令“生成判决书”），忽视中间的事实认定与法律推理步骤，导致决策过程“黑箱化”。

权责不清的全面外包：未加区分地将证据审查、事实判断等本应亲历性的司法核心工作交由智能体处理，形成事实上的“自动化裁判链”，架空司法责任制。

（二）指令污染风险：任务目标在“自主”中被恶意扭曲

任务指令在复杂执行过程中存在被曲解、篡改的风险。智能体实现“自主性”的关键在于中间步骤的“信息自主”，使得智能体收到的任务指令、读取的上下文、调用工具的返回值均可能在传递过程中被污染或篡改，进而让真实任务目标被恶意改写。例如：

使用者越界攻击：使用者（包括诉讼当事人借助代理工具或文本材料）可能尝试绕过智能体内置的安全约束，通过角色扮演、编码绕过、多轮逐步诱导、系统提示泄露等方法诱使智能体做出本应拒绝的行为，即业界所称的“越狱”类攻击。司法场景下需特别防范使用者通过递交材料、诱导提问、设置个性化参数等方式进行深层攻击。

第三方指令注入：第三方内容可能被植入恶意指令，诱导智能体读取并执行；第三方内容可能包括以图像、音频、扫描件等非文本形式承载的指令隐写，试图绕过普通文本过滤进入智能体任务。

知识与记忆污染：诉讼材料中包含由人工智能生成的虚假法条、虚构案例等不实信息，若不加甄别地将此类“污染知识”纳入任务上下文，将导致基于错误前提的推理与输出；具备长期记忆能力的

智能体，一次任务中被植入的错误或恶意内容可能跨会话持续影响后续所有任务，形成“慢性污染”。

（三） 程序失范风险：动态执行链偏离规范轨道

智能体为完成任务而动态调用的工具（技能）及自生成的执行步骤，可能偏离法定程序或内部规范。例如：

工序不可验证：智能体自主调用外部工具或者自行制造工具，导致生成中间步骤的过程可能缺乏透明度和可解释性，难以审计和验证其合规性。

技能违规引入：智能体调用或自主创设的某个外部工具或“技能”本身可能包含错误逻辑、恶意代码或不符合司法程序的规定，导致任务执行在表面上完成，实则违反了办案规范，带来程序性风险。这一风险在能力组件的供应链上尤其突出：打包分发的扩展、第三方提供的工具接口、智能体通过代码创建的组件、动态派生的子智能体等均可能成为污染链路的入口；恶意能力组件可能完全绕过运行时安全策略，在策略引擎的管辖范围之外运作，传统安全审查难以覆盖。

（四） 系统失控风险：权限滥用与级联故障放大危害

智能体获得系统操作权限后，可能引发越权操作、数据泄露等安全隐患。例如：

高权限误操作：智能体在权限边界不清、异常情形识别不足的情况下，可能错误调用高权限接口或触发关键业务流程，对案件审理造成难以挽回的干扰或损害。

数据边界渗透：在跨系统或多数据源场景下，智能体可能为完成任务而尝试访问其权限范围之外的敏感数据，造成数据越界访问与泄露风险。

多智能体级联错误：在多智能体协同场景下，单个智能体的异常操作可能通过任务分派与跨组件调用链路引发连锁错误，进一步放大影响半径。

二、紧抓机遇：坚持人工智能技术司法应用的“正向研发原则”

智能体技术是大语言模型等人工智能技术发展的组合创新产物，即使其展现出了更强的自主能力，其应用必须继续遵循前期人工智能辅助审判探索经验中沉淀的四项研发原则，即，一方面鼓励以“大胆探索、审慎应用”的态度开展创新实践，引导法院系统积极稳妥推进人工智能司法应用，另一方面保持清醒的技术定位认知，坚决避免出现“机器代替法官”的认识误区，确保技术发展始终服务于审判能力现代化这一重要目标。

（一）锚定辅助工具定位：智能体等人工智能技术的司法应用应当严格划定人机协作关系的基础伦理界限，明确智能体等人

工智能技术的辅助工具属性。无论技术如何发展，始终坚持尊重法官自主决策权，全过程充分贯彻“亲历性原则”，不允许人工智能以任何方式替代法官作出裁判，在辅助工具研发中确保“法官决策在环路”。

（二）深度融入审判环节：智能体等人工智能技术的司法应用应当深度契合审判工作规律与法官办案习惯。通过“通专结合多模型引擎＋司法业务系统工程+权威法律知识库”的工程化应用模式，以司法业务思维方法牵引人工智能技术组合方式，精准识别各类案件核心办理环节的难点任务，围绕各类案件的业务标准框架进行设计，实现技术与业务的深度融合。

（三）同步落实司法责任：智能体等人工智能技术的司法应用应当始终坚持责任归属的明确到人，确保权责清晰。辅助工具必须设置必要的法官审核、确认与决策环节，确保案件的判断决策由裁判者做出，司法责任最终由裁判者承担，系统全程提供决策辅助，知识推送，同步记录决策过程，促进法官审慎司法，确保“让审理者裁判、由裁判者负责”原则在人工智能应用中贯彻

（四）全程确保数据安全：智能体等人工智能技术的司法应用应当贯彻数据全生命周期安全管理，严格保障司法数据安全。在系统设计、技术方案和管理制度上，辅助工具应始终将数据安全视为研发工作红线，全面采取专网部署、专人管理、专门安全架构等方

式，确保人工智能技术的训练、微调、推理、应用所涉及司法数据、交互信息及生成内容均在封闭的内网环境中安全运行、闭环处理，从根本上杜绝敏感数据外泄风险。

三、 严防风险：智能体技术司法应用的“负面清单”

为平衡智能体技术的应用潜力与安全风险，在遵循前述四项一般原则基础上，提出以下四项“负面清单”，以明确禁止性行为框定安全底线。清单旨在为智能体技术的司法应用划定明确的禁区，在坚守安全、伦理与合规红线的前提下，为有益的技术创新与探索保留合理空间。

（一） 禁止转授决策性职权

任何情况下，不得将裁判权、审批权、决定权等核心司法决策职权交由智能体行使。必须始终坚持“法官决策、智能体执行”的根本模式，所有实质性的证据采信、事实认定、法律适用及裁判结论必须由法官明确作出。为此，需通过技术手段严格禁止智能体以“裁判建议”、“一键生成结论”等形式提供替代性决策方案，系统应对用户的此类指令进行有效引导与拦截。在涉及多个步骤的“工作流智能体”设计中，必须保留并强制经过所有关键决策节点的人工确认环节，坚决杜绝形成跳过人工判断的自动化闭环，避免任何形式的黑箱式、误导性决策。

（二） 禁止注入未治理数据

不得使用未经清洗、脱敏、合规与安全审查的原始数据直接训练智能体或作为其执行任务的上下文信息。所有输入智能体的提示词、案件材料、知识信息等，必须经过严格的来源审核、内容清洗、知识核验、敏感信息识别与处理等全流程治理。必须建立有效机制防范“提示词注入攻击”，原则上禁止将未经结构化提取的原始诉讼材料直接作为智能体的任务输入，从源头切断恶意指令的注入途径。同时，应确保智能体进行法律推理所依据的知识必须直接来源于经过认证的权威知识库，并对诉讼材料中可能存在的虚假法律信息进行识别与警示，严防“知识污染”。此外，对智能体调用外部工具所获取内容，包括检索所得网页文本等第三方信息，须实施指令过滤与格式隔离。对使用者的直接输入实施安全过滤与越狱攻击检测，拒绝明显的越狱式诱导指令。对智能体的长期记忆、跨会话状态、决策缓存等实施定期清洗、按案件隔离、按任务清零，防止一次攻击跨会话扩散形成慢性污染。

（三） 禁止应用未认证技能

不得允许智能体调用未经安全评估、功能认证与合规审核的外部工具、插件或能力组件。系统应对所有可调用技能实行严格的白名单管理制度，严禁用户自行安装或配置未经验证的第三方技能。需通过技术手段限制智能体自行创建或从外部获取新技能的能力，

将其行动范围牢固约束在预先审核通过的工具集内。智能体及能力组件须经校验后方可加载；对被调用能力组件及被激活子智能体作完整性验证。必须建立统一的技能管理机制，由技术管理部门与审判管理部门共同负责，对拟上线的技能进行前置安全认证、功能合规性审查与上线后的定期复评，同步建立变更审批与回滚机制。所有技能的调用必须全程留痕，形成完整、可审计的追踪链条。此外，建立智能体运行的实时监控与异常检测机制，一旦发现能力组件调用偏离预期轨迹或触发安全阈值，系统应自动中断当前任务链路并回滚至安全状态，同时通知审判主体介入处置。

（四） 禁止逾越安全性护栏

不得绕过或破坏系统为智能体设置的网络隔离、权限控制、数据访问边界、操作审计及风险熔断等安全防护机制。智能体必须在严格的网络隔离环境下运行，严禁擅自访问外部网络或进行跨安全域通信，严守**网络边界**。必须实施严格的数据分级分类与最小权限访问控制，智能体只能接触完成当前任务所必需且经过脱敏处理的、在用户权限范围内的数据，严守**数据边界**。智能体的操作权限必须与用户身份紧密绑定，并遵循“随案授权”、“按岗授权”原则，为其执行的每项任务创建独立的“安全沙箱”环境，防止权限滥用与横向移动，严守**权限边界**。在多智能体协同场景中，智能体之间的消息传递必须通过经认证的加密通道进行；禁止智能体之间使用

未经认证的通信方式或绕过通信中间件直接交互；对智能体间通信实施协议版本管控，禁止降级至不安全的通信协议，严守**通信边界**。针对多智能体任务链路中的级联失败风险，建立任务级和系统级熔断机制，当单个智能体出现异常行为（如输出格式异常、置信度低于阈值、响应超时）时，自动暂停该智能体及其下游的整条任务链路，隔离故障节点并通知人员介入，严守**链路边界**。此外，必须对智能体的所有操作进行全量日志记录与实时行为监控，并设立异常操作风险预警与自动阻断机制，实施严格的**行为监控**。

附：《指南》实践基础与形成过程

自 2023 年起，深圳法院率先探索大语言模型等人工智能技术在司法领域的稳慎融合，研发上线人工智能辅助审判系统并持续迭代升级。2026 年，实现“多智能体安全协同”的深度应用架构。深圳市大数据研究院、清华大学互联网司法研究院、香港大学法律学院在深入调研深圳法院实践经验的基础上，结合人工智能技术最新发展态势与司法应用特殊要求，与深圳市中级人民法院联合编制本《指南》，力求为司法领域规范、安全、有效地推进人工智能及智能体技术应用提供参考与指引。