

《人工智能应用场景下的法律知识库建设 规范》

编制说明

一、工作简况

（一）任务来源

本项目根据中国互联网协会团体标准立项计划（标准编号：T/ISC XXX—2026），项目名称为“人工智能应用场景下的法律知识库建设规范”的任务而进行制订。本文件由中国互联网协会提出并归口。

（二）起草单位及主要起草人

本文件起草单位：北京新橙科技有限公司、XXX。

本文件主要起草人：张萌、XXX。

（三）标准制定目的和意义

近年来，法律科技与生成式人工智能快速发展，法律数据库、智能检索系统、法律大模型等产品在司法办案、律师执业、企业合规、公共法律服务等场景中广泛应用。法律知识库作为上述产品的核心数据底座，其质量与管理水平直接决定法律信息服务的准确性与公信力。

从行业现状看，法律知识库建设尚缺乏统一标准：数据来源认定、采集加工规则、分类编目方法、更新维护机制等做法不一，部分产品存在法规版本陈旧、失效文件未标注、案例引用不规范等问题。法律知识具有强时效性与强权威性，法规的立、改、废若不能及时同步至知识库，将直接导致产品输出错误结论，引发执业与合规风险。同时，法律大模型训练与检索增强生成（RAG）等应用对高质量结构化语料的需求日益迫切，裁判文书、案例数据的采集使用还涉及个人信息保护、著作权、数据安全等多重合规要求，亟需统一规范。

制定本文件的必要性主要体现在三个方面：一是质量保障需求，通过统一来源核验与更新时限要求，确保法律知识产品输出的准确性与权威性；二是产业发展需求，通过统一分类编目与结构化标准，形成行业可复用的高质量语料供给；三是安全合规需求，通过明确数据分类分级、个人信息保护与授权合规要求，守住数据安全底线。制定目的在于：确立法律知识库全生命周期建设的基本规范，

统一各环节技术要求和管理要求，为法律人工智能的健康发展提供标准支撑。

（四）主要工作过程

2026 年 7 月 16 日，召开了本团体标准立项评估会，确定了标准框架结构，收集了与会各单位对标准草案（初稿）的修改意见，明确了工作分工及会后安排。

2026 年 8 月 24 日，本团体标准由中国互联网协会正式立项，立项名称为：《法律知识库建设规范》。

2026 年 8 月 30 日，召开了本团体标准第一次讨论会，与会专家及代表对标准的技术内容、章节结构、关键指标等进行了深入讨论，编制组认真听取并记录了相关调整建议。

2026 年 9 月 8 日，根据专家意见对本团体标准的立项名称进行了修改，从《法律知识库建设规范》修改为《人工智能应用场景下的法律知识库建设规范》。

2026 年 9 月至 10 月，对《人工智能应用场景下的法律知识库建设规范》（征求意见稿）公开征求意见。

二、标准编制原则和依据

（一）编制原则

标准起草小组在编制标准过程中，以国家、行业现有的标准为制订基础，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》的规定及相关要求编制，遵循以下原则：

（1）协调性。

标准编写符合 GB/T 1.1—2020 的规定，技术内容与现行国家标准、行业标准及法律法规协调一致。

（2）科学性

本文件依据法律知识管理规律、数据治理原理和人工智能技术应用特点设计，围绕法律知识库全生命周期确立各环节技术要求，各项指标均有明确依据。

（3）实用性

本文件立足法律知识库建设运营实际，明确来源管理、权威性核验、时效性监测、版本管理等可操作的实施要求，并以附录形式给出元数据规范、监测要点清单和数据结构示例，为建设主体提供可行依据。

（二）标准主要内容与确定依据

一、适用范围

本文件规定了人工智能应用场景下法律知识库建设的总体要求、数据采集、分类与编目、存储与结构化、更新与维护、数据安全与合规管理以及应用管理要求，适用于法律人工智能产品研发企业、律师事务所、司法行政机关、企业法务及合规部门等组织法律知识库的规划、建设、运营与应用活动。该范围的确定基于当前行业现状和人工智能应用发展趋势，并考虑了标准的普适性和可推广性。

二、技术要求

（1）总体要求

本文件提出合法合规、权威准确、完整系统、动态时效、安全可控、开放兼容六项建设原则，确立数据源层、采集层、加工层、存储层、服务层、应用层的总体架构和建设流程，并规定组织与人员、收录范围管理要求。

（2）数据采集要求

要求建立数据来源清单管理制度，规范性法律文件以发布机关官方网站、官方公报为准，司法案例以司法机关官方发布平台为准；采集活动应遵守著作权法、网络安全法、数据安全法、个人信息保护法等法律法规；入库前应进行权威性核验，重要数据宜交叉校验；时效性要求高的数据宜每日监测，采集日志保存不少于一年。

（3）分类与编目规范

确立法律法规、司法案例、学术与实务观点、合同与文书范本四大基本类目，并建立多维分类标签；为每个知识条目分配稳定、不可复用的唯一标识符；以规范性附录 A 确立核心元数据字段，关键元数据缺失或存疑的条目不得标注为“现行有效”对外提供服务；自动化标注的人工抽检比例不低于 5%，准确率合格标准不低于 95%。

（4）存储与结构化要求

入库数据应经过去重、纠错、格式规范化处理；法律法规类数据应结构化至条文层级，采用 JSON、XML 等通用格式和 UTF-8 编码；面向大模型训练与 RAG 应用的知识内容应合理分块并确保检索结果可溯源；商业化服务的知识库安全保护等级不宜低于 GB/T 22239—2019 规定的第三级。

（5）更新与维护机制

实行分级监测，法律、行政法规、司法解释及核心规章宜每日监测，其他规范性文件监测周期不超过一周；新发布的法律、行政法规、司法解释应自发布之日起 2 个工作日内完成入库，部门规章、地方性法规宜在 5 个工作日内完成入库；规范性文件历次版本应完整保存形成版本链，废止文件应保留可查并标注“已废止”。

（6）数据安全与合规管理

参照 GB/T 43697—2024 对知识库数据分类分级并实行差异化管控；裁判文书、案例数据入库前应完成脱敏处理；收录他人享有著作权的作品应取得合法授权；敏感操作审计日志保存不少于三年；同时对数据出境、安全事件应急、第三方委托处理、系统安全防护作出规定，特别要求防范提示词注入、数据投毒、对抗样本等恶意攻击。

（7）应用管理要求

对外提供知识库服务应提供标准化 API，接口返回的失效或废止数据应带有明确状态标识；训练语料应满足来源合法性、权威性核验与个人信息保护要求且可追溯；RAG 场景应保证效力状态随检索结果一并返回；不得将未经核验的机器生成内容直接作为确定性法律结论对外提供，面向公众提供生成式法律内容服务的还应遵守《生成式人工智能服务管理暂行办法》及 GB 45438—2025 标识要求。

（8）质量评价

建立内容准确率（法律法规类核心数据不低于 99.5%）、收录完整率（核心类目达到 99%）、更新及时率（核心法规不低于 98%）、溯源完备率（不低于 95%）、脱敏合格率（达到 100%）五项指标体系，指标数值综合考虑了法律服务的高准确性要求与行业实际能力水平。质量评价每年不少于一次，不达标应制

定整改计划并落实。

三、国内外情况简要说明

经查，与本标准相关的现行国家标准情况如下：

GB/T 22239—2019《信息安全技术 网络安全等级保护基本要求》，本文件存储系统安全防护等级要求与该标准相衔接；

GB/T 35273—2020《信息安全技术 个人信息安全规范》，本文件裁判文书、案例数据脱敏及个人信息保护要求与该标准保持一致；

GB/T 37988—2019《信息安全技术 数据安全能力成熟度模型》，本文件安全风险评估与数据安全能力提升要求参考了该标准；

GB/T 41867—2022《信息技术 人工智能 术语》，本文件术语和定义与该标准协调一致；

GB/T 43697—2024《数据安全技术 数据分类分级规则》，本文件数据分类分级要求参照该标准制定；

GB 45438—2025《网络安全技术 人工智能生成合成内容标识方法》，本文件生成式法律内容标识要求与该标准相衔接。

经检索，目前国内外尚无专门针对人工智能应用场景下法律知识库建设规范的国家标准、行业标准，本标准为首次制定，填补了该领域标准空白。

四、与有关的现行法律、法规和强制性国家标准的关系

本标准的制定过程、技术要求的选定、质量指标的确定等均符合《中华人民共和国网络安全法》、《中华人民共和国数据安全法》、《中华人民共和国个人信息保护法》、《中华人民共和国著作权法》、《生成式人工智能服务管理暂行办法》、《数据出境安全评估办法》等现行法律法规和强制性国家标准的规定，与 GB 45438—2025《网络安全技术 人工智能生成合成内容标识方法》等强制性国家标准协调配套，无冲突。

五、重大分歧意见的处理经过和依据

本文件在制定过程中未出现重大分歧意见。

六、废止现行有关标准的建议

本标准不涉及对现行标准的废止。

七、其他应予说明的事项

无。

《人工智能应用场景下的法律知识库建设规范》团体标准编制组

2026 年 9 月