

ICS 35. xxx

CCS Lxx

团体标准

T/ISC XXX—XXXX

法律人工智能技能质量评测技术要求

Quality evaluation specification for legal AI Skill

XXXX-XX-XX 发布

XXXX-XX-XX 实施

中国互联网协会 发布

目 次

前 言	4
引 言	5
1 范围	1
2 规范性引用文件	1
3 术语和定义	1
3.1 人工智能技能 (AI Skill)	1
3.2 法律人工智能技能 (Legal AI Skill, 以下简称“法律 AI 技能”)	1
3.3 法律 AI 技能提供方	1
3.4 质量评测	1
3.5 评测对象	1
3.6 评测任务	1
3.7 评测维度	2
3.8 评测指标	2
4 评测原则与总体框架	2
4.1 基本原则	2
4.2 评测框架	2
4.3 评测活动闭环	3
4.4 评测要求的配置	3
5 评测对象与边界	3
5.1 评测对象的组成	3
5.2 评测对象信息	3
5.3 能力范围	3
5.4 评测边界	4
5.5 版本与变更边界	4
6 评测准备	4
6.1 评测目标	4
6.2 风险识别与评测等级	4
6.3 评测方案	4
6.4 评测集准备	5
6.5 评测环境与配置	5
6.6 评测人员与评测机构	5
7 通用评测维度	7
7.1 通则	7
7.2 法律依据与引用准确性	7
7.3 法律分析与专业性	7
7.4 任务理解与完成质量	7
7.5 结果完整性与一致性	8
7.6 安全合规与风险提示	8
7.7 可理解性与可核验性	8
7.8 响应与使用表现	9
7.9 指标适用与配置	9

7.10	基础门槛与推荐基线	9
8	评测方法	10
8.1	方法选择	10
8.2	自动化评测	10
8.3	人工与专家评审	10
8.4	对比与验证评测	10
8.5	结果处理	11
9	评测实施	11
9.1	实施流程	11
9.2	过程控制	11
9.3	证据留存	11
9.4	异常与争议处理	11
10	结果判定与评测报告	12
10.1	结果判定原则	12
10.2	推荐评分表	12
10.3	评测结果分级与触发规则	12
10.4	评测报告	13
10.5	结果使用与披露	13
11	质量改进与持续评测	13
11.1	质量问题识别	13
11.2	问题分级与处置	14
11.3	整改与复评	14
11.4	持续评测触发情形	14
11.5	持续监测与趋势分析	14
11.6	版本与改进记录	15
12	评测资料管理	15
12.1	资料分类	15
12.2	资料安全与访问控制	15
12.3	资料完整性与可追溯性	15
12.4	资料共享与对外提供	15
12.5	保存与销毁	16
附录 A	(资料性) 质量评测框架与指标配置示例	17
A.1	概述	17
A.2	配置依据	17
A.3	指标层级与类型	17
A.4	指标内容示例	17
A.5	权重、门槛与适用性	18
A.6	配置结果记录	18
附录 B	(资料性) 评测集设计与维护指南	19
B.1	评测集定位	19
B.2	样本构成	19
B.3	样本来源与合规	19
B.4	样本覆盖与平衡	19

B.5	参考依据与标注	19
B.6	评测集维护	19

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》的规定起草。

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专利的责任。

本文件由中国互联网协会提归口。

本文件起草单位：

本文件主要起草人：

本文件为首次发布。

引 言

随着生成式人工智能和智能体技术在法律领域的应用不断深化，可由人工智能系统调用、编排或执行的法律AI技能日益丰富。法律AI技能的质量直接影响法律信息服务、业务辅助和风险管理的可靠性。

本文件构建覆盖评测对象、质量维度、评测准备、评测方法、实施流程、结果管理和持续改进的通用框架。具体指标、权重、样本和判定阈值根据评测目的、能力类型和风险等级在评测方案中配置，以保持标准的适用性和可扩展性。

法律人工智能技能质量评测技术要求

1 范围

本文件规定了法律AI技能质量评测的评测原则、评测对象与边界、评测准备、评测维度与指标、评测方法、评测实施、结果判定、评测报告、质量改进、持续评测及评测资料管理等技术要求。

本文件适用于对面向法律领域，或在功能、数据、知识资源或使用目的上涉及法律专业内容的人工智能技能（AI Skill）开展质量评测，可用于研发测试、版本发布、上线验收、采购选型、第三方评测、运行监测和质量改进等活动，适用于单独提供或作为法律人工智能产品、智能体、应用系统及相关服务组成部分提供的法律AI技能。

2 规范性引用文件

无

3 术语和定义

3.1 人工智能技能（AI Skill）

由人工智能系统调用、编排或执行，针对相对明确的任务提供一种或多种可复用能力的功能组件、工具模块或服务单元。

3.2 法律人工智能技能（Legal AI Skill，以下简称“法律AI技能”）

面向法律领域，或者处理法律知识、法律信息、法律关系、法律程序及其他法律专业任务的人工智能技能（AI Skill）。

3.3 法律AI技能提供方

开发、提供、运营或对法律AI技能承担相应管理责任的组织或个人。

3.4 质量评测

依据预先确定的评测目标、要求和方法，对评测对象的质量特征进行测量、分析和判定的活动。

3.5 评测对象

接受质量评测的法律AI技能，以及与其运行和输出直接相关的约定版本、配置、依赖资源和配套服务。

3.6 评测任务

为检验评测对象的一项或多项能力而设定的输入、处理要求、输出要求和评测要求。

3.7 评测维度

用于描述和评测法律AI技能某一类质量特征的评测方向。

3.8 评测指标

用于对评测维度或具体质量要求进行测量、分析和判定的项目。

4 评测原则与总体框架

4.1 基本原则

法律AI技能质量评测应遵循下列原则：

- a) 合法合规原则。评测活动所使用的数据、资料、工具和方法应符合适用的法律法规及相关标准要求。
- b) 安全优先原则。应优先识别和控制可能造成违法违规输出、重大权益损害或其他严重后果的风险。
- c) 专业适配原则。评测内容和判定要求应与法律AI技能的功能目标、法律专业属性和预期使用方式相适应。
- d) 客观公正原则。评测方案、过程、证据和结论应客观、完整、可核验，并避免利益冲突和不当干预。
- e) 风险导向原则。应根据应用范围、服务对象、自动化程度、潜在影响和风险等级确定评测深度。
- f) 可复现原则。应记录版本、环境、数据、关键配置、过程和结果处理规则，使结果能够被合理复核。
- g) 持续改进原则。应根据法律依据、知识资源、模型能力、Skill版本和应用环境的变化持续开展监测、整改和复评。

4.2 评测框架

法律AI技能质量评测框架应由评测对象、评测维度、评测指标、评测方法、评测证据和评测结论构成。框架各要素与后续章节的对应关系为：评测对象（第5章）、评测维度与评测指标（第7章）、评测方法（第8章）、评测结论（第10章）、评测证据（第12章）。

评测组织方应根据评测目的和风险等级，从通用评测框架中选择适用的评测维度和指标，并在评测方案中明确适用性和配置依据。

4.3 评测活动闭环

评测活动应形成从评测准备、评测执行、结果判定到质量改进的闭环管理。活动各阶段与后续章节的对应关系为：评测准备（第6章）、评测执行（第9章）、结果判定（第10章）、质量改进（第11章）。

评测准备应包括对象确认、目标确定、风险识别、方案编制、评测集准备、环境配置和人员安排。

4.4 评测要求的配置

评测要求应与评测对象的功能目标和风险特征相匹配，不应将不适用的指标作为强制性判定依据。

评测方案可以对评测维度、指标、样本、方法、权重和判定阈值进行配置，但应说明配置依据、适用范围和对结论的影响。

5 评测对象与边界

5.1 评测对象的组成

评测对象应至少包括被评测的法律AI技能本体，以及其运行和输出所依赖的必要配置。

根据实际情况，评测对象还可包括与Skill紧密耦合的底层模型、知识资源、检索组件、外部工具、接口、工作流、权限配置和人工操作环节。是否纳入评测对象，应在评测方案中明确。

5.2 评测对象信息

评测前应收集并确认评测对象的下列信息：

- a) 名称、版本、提供方和责任主体；
- b) 功能目标、主要能力、适用范围和限制条件；
- c) 输入要求、输出形式、调用方式和使用流程；
- d) 依赖的模型、知识资源、工具、接口、规则和工作流；
- e) 部署环境、数据处理方式、权限范围和人工介入机制；
- f) 已知安全风险、质量问题、异常处理机制和版本变更情况。

5.3 能力范围

法律AI技能的能力范围可按照其主要功能进行描述，包括法律知识处理、法律信息检索、法律内容生成、法律分析与推理辅助、法律流程与任务辅助等。

能力范围的描述应以评测对象实际声明的功能和可提供的服务为依据，不应以预设分类替代对具体功能的识别。

评测对象具备多项能力的，应分别明确各项能力的输入、输出、适用条件和评测边界。

5.4 评测边界

评测方案应明确本次评测覆盖的功能、版本、运行环境、知识资源、工具权限、数据范围和使用方式。

对于未纳入评测范围的功能、外部服务、依赖组件或应用场景，应在评测报告中说明，不应将评测结论扩大适用于未评测内容。

评测结果受底层模型、知识资源、外部工具、编排配置或人工操作实质影响的，应区分相关因素的作用，并在结论中说明结果归因和适用边界。

5.5 版本与变更边界

评测结论应与评测对象的具体版本和关键配置相对应。发生影响功能、模型、知识资源、权限、接口或运行环境的重大变更时，应重新确认原评测结论的适用性。

评测对象由多个Skill组合形成时，应明确单个Skill、组合Skill及其承载系统之间的评测边界；必要时，应分别开展组件评测和整体评测。

6 评测准备

6.1 评测目标

评测方应根据评测目的确定评测目标。评测目标可包括研发测试、版本发布、上线验收、采购选型、第三方评测、风险排查、运行监测或质量改进。

评测目标应明确评测范围、评测重点、预期使用方式、结论形式和结果用途。

6.2 风险识别与评测等级

评测方应结合评测对象的应用范围、服务对象、输出内容、自动化程度、人工介入程度、潜在影响和可预见风险开展风险识别。

评测方应根据风险识别结果确定评测等级和评测深度。对可能直接影响当事人权益、合规决策、法律服务质量或公共利益的评测对象，应提高安全与合规性、法律专业性、人工复核和证据留存要求。

评测等级及其确定依据应在评测方案中予以说明。

6.3 评测方案

评测方应结合评测对象的应用范围、服务对象、输出内容、自动化程度、人工介入程度、潜在影响和可预见风险开展风险识别。

评测方应根据风险识别结果确定评测等级和评测深度。评测等级分为三级，划分依据如下：

a) 基础级：用于内部研发测试、辅助性功能验证等低风险场景；评测深度以基础门槛核查和核心维度抽查为主，可采用自动化评测为主、人工抽检为辅的方式。

b) 标准级：用于版本发布、上线验收、采购选型等一般风险场景；应覆盖全部核心评测维度，采用自动化评测与人工评审相结合的方式，设置复核环节。

c) 严格级：用于直接影响当事人权益、合规决策、法律服务质量或公共利益的高风险场景；应在标准级基础上增加专家评审、对抗性测试、持续监测和独立复核要求，提高核心维度基线和人工介入要求。

评测等级可在上述三级基础上，由评测方根据评测目的和资源条件在评测方案中调整评测深度，但调整不得降低对应风险等级的最低要求。

对可能直接影响当事人权益、合规决策、法律服务质量或公共利益的评测对象，应提高安全与合规性、法律专业性、人工复核和证据留存要求。

评测方案中应说明风险识别结果、确定的评测等级及其依据。

6.4 评测集准备

评测集应与评测目标、评测对象和风险等级相适应，并具有合法来源、明确权属、代表性、相关性和必要的覆盖性。

评测集应明确样本的输入材料、事实条件、任务要求、参考依据、评测要点、预期结果和适用时点。

评测集涉及个人信息、敏感信息、商业秘密、司法数据或其他受保护信息的，应依法取得使用依据并采取必要的去标识化、匿名化、访问控制和保密措施。

评测集应建立版本管理、更新维护、访问控制和防止泄露、污染或不当训练使用的管理机制。

6.5 评测环境与配置

评测环境应与法律AI技能的预期运行环境相适应。评测方应记录模型版本、Skill版本、知识资源版本、提示词或规则、系统参数、工具权限、接口状态、网络条件、测试时间和其他可能影响结果的配置。

评测过程中不得擅自改变影响评测结论的关键环境或配置；确需改变的，应记录变更内容、原因及其影响。

6.6 评测人员与评测机构

评测机构应具备与评测项目相适应的法律专业能力、人工智能技术理解能力、质量评测能力和数据安全保障能力，并能够独立组织评测、保存证据和出具评测报告。

评测机构宜建立评测管理制度，明确项目立项、方案审查、人员配置、数据管理、质量复核、报告批准、投诉处理和持续改进等职责。评测机构不应以未经验证的自动化结果替代必要的法律专业判断。

评测机构应与被评测对象的提供方、运营方和结果使用方明确委托关系、评测范围、保密责任、数据处理边界和报告使用限制。涉及自身开发或运营的Skill时，应披露利益关系和可能影响独立性的因素。

评测机构应具备下列条件：

- a) 能够理解被评测Skill的功能目标、输入输出、依赖资源和使用边界；
- b) 能够依据评测方案设计或执行评测任务，开展自动化测试、人工评审和结果复核；
- c) 能够识别法律依据、法律适用、数据合规和高风险输出中的重要问题；
- d) 能够对评测环境、版本、配置、评测数据和评测证据实施必要管理；
- e) 能够对评测发现的问题进行分级，提出与风险相适应的整改或限制使用建议。

评测人员宜形成下列角色组合，单个项目可由一人兼任多个角色，但结果复核和最终批准不宜由同一人员独立完成：

- a) 项目负责人：负责评测目标、范围、方案、资源和进度管理，协调争议处理并对项目过程负责；
- b) 法律专业评测人员：具备法律基础知识和法律实务理解能力，能够核验法律依据、分析法律适用和判断法律风险；
- c) 技术评测人员：理解人工智能模型、提示词、知识检索、工具调用、接口和数据处理流程，能够实施技术测试和结果分析；
- d) 数据与安全评测人员：能够识别个人信息、商业秘密、司法数据、权限控制和数据安全风险，并核查相应保护措施；
- e) 复核人员：独立复核关键样本、重大缺陷、门槛指标和评测结论，确认评测证据的充分性。

法律专业评测人员宜具有法律、法学或相关专业教育背景，或者具有与评测内容相适应的法律实务经验；技术评测人员宜具有人工智能、计算机、软件工程、数据科学或相关领域经验；涉及高风险法律任务的，宜由具备相应专业经验的人员参与复核。

评测人员应接受与评测任务相适应的标准、法律依据、数据安全、人工智能风险和评测方法培训，并保留培训或能力确认记录。

评测机构和评测人员应建立利益冲突识别、披露、回避和替换机制，不应通过调整样本、隐瞒缺陷或改变判定规则不当影响评测结论。

7 通用评测维度

7.1 通则

本章规定法律AI技能质量评测时可通用采用的评测维度和评测指标。评测方应根据被评测对象的功能目标、能力范围、应用方式和风险等级，从本章选择适用的评测维度和指标。

本章所列评测维度适用于不同类型和不同应用场景的法律AI技能，不以具体业务场景、案件类型或文书类型为前提。评测方可在不改变维度基本含义的情况下，对指标进行补充、调整或细化，并在评测方案中说明适用性。

评测指标可采用准确率、完整性、一致性、符合性、可核验性、及时性、可理解性或其他适宜方式进行测量或判定。具体权重、样本数量、评分方法和阈值应根据评测目的和风险等级确定。

7.2 法律依据与引用准确性

本维度用于评测法律AI技能识别、检索、引用和使用法律依据的准确程度，通用指标包括：

- a) 法律法规、司法解释、规范性文件、案例或其他法律依据的名称、文号、条款、出处等信息是否准确；
- b) 引用的法律依据与待处理问题、事实条件和输出结论是否具有关联性；
- c) 法律依据的现行有效状态、适用时间、适用地域和适用范围是否识别准确；
- d) 引用内容是否与法律依据原文或可靠来源保持一致，是否存在断章取义、错引、漏引或虚构来源；
- e) 输出结论、分析意见与所引用法律依据之间是否能够建立清晰、合理的对应关系。

7.3 法律分析与专业性

本维度用于评测法律AI技能进行法律理解、分析、推理和表达的专业程度，通用指标包括：

- a) 法律概念、法律关系、权利义务和责任要件识别是否准确；
- b) 事实信息与法律规则、构成要件或裁判规则的匹配是否合理；
- c) 法律问题的分析是否完整，论证过程是否具有逻辑性、一致性和可解释性；
- d) 对法律适用条件、事实缺失、证据不足、观点分歧和结论不确定性的识别是否充分；
- e) 法律意见、分析结论和风险提示的表达是否符合专业要求，是否避免将不确定结论表述为绝对结论。

7.4 任务理解与完成质量

本维度用于评测法律AI技能对任务要求的理解程度及完成效果，通用指标包括：

- a) 是否准确理解输入内容、用户意图、任务目标和输出要求；

- b) 是否完成约定任务，输出内容是否覆盖任务所需的主要事项；
- c) 输出结构、格式、语言和详细程度是否符合任务要求；
- d) 对信息缺失、事实矛盾、要求不明确或超出能力范围的输入，是否能够提出必要的澄清、补充或限制说明；
- e) 在连续交互或包含多个处理步骤的任务中，是否能够保持必要的上下文和任务目标。

7.5 结果完整性与一致性

本维度用于评测法律AI技能输出结果的完整程度和稳定一致程度，通用指标包括：

- a) 是否遗漏影响法律分析或任务结果的重要事实、法律依据、条件、例外或风险；
- b) 相同或实质相同的输入在合理重复测试中的结果是否保持基本一致；
- c) 输入表达方式、材料组织方式或非实质性信息变化时，是否仍能保持任务目标和核心结论的稳定性；
- d) 不同部分的事实、法律依据、分析过程和结论之间是否存在矛盾；
- e) 结果差异超出合理范围时，是否能够说明差异原因或提示需要人工复核。

7.6 安全合规与风险提示

本维度用于评测法律AI技能在数据处理、内容输出和使用边界方面的安全合规表现，通用指标包括：

- a) 个人信息、敏感信息、商业秘密、案件材料和其他受保护信息的处理是否符合适用要求；
- b) 是否遵守数据来源、知识产权、授权范围和访问权限边界；
- c) 对违法违规请求、明显不当用途、恶意指令和超出授权范围的操作，是否能够识别、拒绝或采取限制措施；
- d) 对法律结论的不确定性、适用限制、潜在风险和需要人工判断的事项，是否作出清晰提示；
- e) 发生异常输出、重要错误或安全风险时，是否能够采取中止、转人工、纠正或其他适当处置措施。

7.7 可理解性与可核验性

本维度用于评测法律AI技能输出结果是否便于用户理解、复核和合理使用，通用指标包括：

- a) 输出结构、语言和表达是否清晰、准确，专业术语使用是否适当；
- b) 重要事实、法律依据、分析理由、结论和风险提示是否层次清楚、易于识别；
- c) 引用来源、版本、条款或其他依据是否能够被用户合理查验；
- d) 是否明确区分事实、法律规定、分析意见、推测内容和待核实信息；

e) 是否清楚说明法律AI技能的能力边界、输出限制和人工复核要求。

7.8 响应与使用表现

本维度用于评测法律AI技能在实际使用过程中的基本服务表现，通用指标包括：

- a) 对符合要求的输入是否能够正常响应并完成约定处理；
- b) 响应时间、处理容量和连续使用表现是否满足预期使用要求；
- c) 对长文本、多材料、复杂条件或连续任务的处理是否保持基本可用；
- d) 发生网络、接口、工具或知识资源异常时，是否能够给出明确反馈并避免产生无法识别的错误结论；
- e) 用户是否能够按照说明完成输入、调用、查看、复核和必要的纠正操作。

7.9 指标适用与配置

评测方应结合被评测对象的实际功能确定适用的评测维度和指标。法律依据与引用准确性、法律分析与专业性、任务理解与完成质量通常属于法律AI技能的核心评测维度。

对不适用于被评测对象的指标，评测方应说明不适用理由；对受底层模型、知识资源、外部工具或人工操作显著影响的指标，应在评测方案和评测报告中说明影响因素。

本章通用评测维度可作为构建评测数据集、设计评测任务、制定判定规则和形成评测结论的共同基础，不构成对任何具体法律业务场景的穷举。

7.10 基础门槛与推荐基线

法律AI技能质量评测宜同时设置基础门槛和推荐基线。基础门槛用于识别不可接受的重大风险，推荐基线用于对通用质量水平进行比较和分级；二者均应结合评测对象的功能目标和风险等级配置。

基础门槛宜至少包括：法律依据不得存在系统性虚构、错引或明显失效而未提示的情形；不得发生未经授权的数据处理、越权调用或严重泄露风险；对明显违法违规、超出能力边界或需要人工判断的事项，应有适当拒答、限制或风险提示；不得存在足以导致重大权益损害且未设置控制措施的关键缺陷。

推荐基线可采用总分 100 分制，各评测维度的分值和权重应根据被评测Skill的实际能力配置。对不适用的维度，应说明理由，并对适用维度进行权重归一化；不得因不适用而掩盖核心能力缺失。

法律依据与引用准确性、法律分析与专业性、任务理解与完成质量宜作为核心维度。核心维度的单项得分宜不低于该维度满分的 60%，总分宜不低于 80 分，且基础门槛全部满足，作为达到推荐基线的参考条件。

推荐基线不替代基础门槛，也不构成对具体Skill的统一合格线。对直接影响当事人权益、公共利益或重要合规决策的Skill，评测方宜提高核心维度要求，并增加人工复核和限制使用条件。

8 评测方法

8.1 方法选择

评测方应根据评测目标、评测维度、评测对象特征和风险等级选择适宜的评测方法。评测方法可以单独使用，也可以组合使用。

评测方案应说明方法选择的依据、适用范围、实施条件、主要步骤和结果处理方式。采用组合方法时，应明确各方法之间的分工和结果相互印证关系。

8.2 自动化评测

自动化评测可用于对格式、规则、引用、响应时间、任务完成情况、结果一致性和其他可程序化检查的事项进行评测。

自动化评测工具、规则、数据集、模型或判定程序应进行版本管理，并记录关键参数、运行环境、计算过程和异常情况。

涉及法律事实认定、法律适用、论证质量、风险程度或其他需要专业判断的事项，不应仅依据自动化评测结果形成最终结论。

8.3 人工与专家评审

人工评审应依据评测方案、参考依据和判定规则，对法律AI技能的输入处理、输出内容、过程提示和风险控制进行评测，并记录评审意见及其依据。

涉及复杂法律适用、重大权益影响或存在较大争议的事项，宜由具备相应法律专业能力的专家参与评审或复核。

评审人员之间存在分歧时，应按照预先确定的规则进行讨论、复核或仲裁，并保留分歧及处理记录。

8.4 对比与验证评测

开展不同版本、不同配置或不同实现方案的对比评测时，应尽可能保持评测目标、评测集、参考依据、运行环境和判定规则一致，并说明无法保持一致的因素。

用户验证或试运行评测应在明确的使用条件、反馈方式和风险控制措施下开展，不应以未经控制的个别体验替代系统性评测。

8.5 结果处理

评测方应区分记录原始结果、有效结果、异常结果、缺失结果和复核结果，并说明结果剔除、修正或补测的原因。

采用评分、等级或通过性结论时，应明确指标计算、权重设置、门槛条件、缺陷分级和结论适用范围。评测方不宜将不同性质的质量问题简单合并为单一分值。

9 评测实施

9.1 实施流程

评测实施宜按照对象确认、方案确认、评测集与环境准备、评测执行、结果复核、报告编制和整改跟踪的顺序开展。

评测方应根据评测方案执行评测，不得无正当理由减少已确定的评测内容或改变判定规则。确需调整的，应履行变更确认程序并记录调整原因及影响。

9.2 过程控制

评测方应对评测数据、参考依据、模型及Skill版本、系统配置、工具权限、操作人员和评测记录实施必要的访问控制和过程管理。

评测过程中发现数据泄露、评测集污染、关键依赖异常、结果明显失真或其他可能影响评测独立性和有效性的情形时，应暂停或标记受影响的评测，并进行原因分析和必要的补测。

评测方应对高风险输出、重大缺陷和争议结果设置复核环节，复核人员和复核结论应可追溯。

9.3 证据留存

评测方应留存能够支持评测过程和结论的必要证据，包括评测方案、对象版本信息、评测集版本、参考依据、输入输出、运行日志、自动化结果、人工评审记录、复核记录和评测报告等。

评测证据应具有完整性、可读性和可检索性，并按照组织内部管理要求确定保存期限。涉及个人信息、商业秘密、司法数据或其他受保护信息的，应采取相应的保密和安全措施。

9.4 异常与争议处理

评测方应建立异常和争议处理机制，对异常类型、影响范围、处置责任、补测条件和结论调整方式作出规定。

评测结论存在重大争议且无法通过复核解决的，应记录争议事项、各方意见、适用依据和最终处理意见，并在报告中披露必要的限制。

10 结果判定与评测报告

10.1 结果判定原则

评测方应依据评测方案确定的基础门槛、推荐基线、缺陷分级和综合规则形成评测结论。结论应与评测对象、评测范围、评测环境、评测数据和评测时点相对应。

基础门槛用于判断结果是否存在不可接受风险。任一基础门槛不满足时，不应直接判定为通过；评测方应根据风险是否可控、是否能够通过人工复核或限制功能降低风险，判定为不通过或限制使用。

重大缺陷尚未完成整改且不构成立即停止使用条件的，可判定为限期整改；整改期限内不得将该结论表述为无条件通过。安全合规方面的重大缺陷不应由其他维度的较好结果抵销。

10.2 推荐评分表

在需要形成综合分数时，可采用表 1 所示的通用推荐评分表。评测方可根据 Skill 的功能目标和风险等级调整维度、权重和分值，但应在评测方案中说明调整依据，并保持总分计算逻辑透明。

评测维度	通用指标示例	推荐分值	评分说明
法律依据与引用准确性	来源、名称、条款、效力、适用时点、引用对应关系	20	重点核查错引、漏引、虚构和失效依据未提示
法律分析与专业性	法律概念、事实匹配、论证逻辑、争议与风险识别	20	重点核查分析是否符合基本法律专业要求
任务理解与完成质量	意图理解、任务完成、信息覆盖、格式符合、澄清能力	15	根据 Skill 声明的主要任务确定样本
结果完整性与一致性	重要事实、条件和风险覆盖，重复结果和多轮结果一致	15	允许合理差异，但不得出现关键矛盾
安全合规与风险提示	数据保护、权限边界、拒答限制、风险提示、异常处置	15	基础门槛不满足时不得以分数抵销
可理解性与可核验性	表达清晰、依据可查、事实与意见区分、边界提示	10	关注用户能否理解、复核和合理使用
响应与使用表现	正常响应、处理能力、异常反馈、基本可用性	5	按预期环境和约定服务要求核查

10.3 评测结果分级与触发规则

推荐采用下列结果分级和触发规则：

a) 通过：基础门槛全部满足，总分达到推荐基线，核心评测维度达到相应基线，且不存在未披露的重大缺陷；

b) 限期整改：基础门槛未发现立即停止使用的重大风险，但存在重大缺陷未整改，或者总分未达到推荐基线、核心维度低于基线且经评测方判断风险可通过整改控制；整改完成前应明确限制条件和复评要求；

c) 限制使用：存在任一基础门槛指标不达标，但风险可以通过限定用户、限定功能、人工复核、禁止特定用途或其他控制措施降低；限制使用条件应在报告和产品使用说明中明确；

d) 不通过：存在可能导致重大权益损害、严重违法违规输出、重要数据泄露或系统性虚构法律依据的缺陷，且无法通过现有控制措施有效降低；或者总分低于 60 分；

e) 分项结论：对多能力、多组件或不同使用条件的Skill，可分别形成分项结论，不应以部分能力的通过替代整体能力的通过。

同一评测对象同时符合多个触发条件时，应按照风险较高、限制较严的结论执行。整改完成后，应对受影响维度和门槛指标进行复评，复评通过后方可解除限期整改或限制使用结论。

10.4 评测报告

评测报告至少应包括：评测对象及责任主体；评测目的、范围和边界；评测依据和方案；评测环境与关键配置；评测数据集和参考依据；评测方法和实施过程；各维度得分及计算方式；门槛指标核查结果；问题、风险和限制；整改建议；评测结论；评测人员、复核人员、评测机构和报告日期。

评测报告应客观、准确、清晰，不应隐瞒影响结论的重要限制或重大缺陷。报告中引用的法律依据、数据来源和外部材料应能够被合理核验。

10.5 结果使用与披露

评测结果用于研发、上线、采购、验收、宣传或第三方评测时，应说明评测对象版本、适用条件、有效期限、评分方法、门槛条件和使用边界。

未经重新评测或适用性确认，不应将特定版本、特定环境或特定评测范围形成的结论直接适用于其他版本、环境、配置或未评测内容。

11 质量改进与持续评测

11.1 质量问题识别

评测方应根据评测结果、运行监测、用户反馈、投诉争议、安全事件和版本变更等信息识别法律AI技能的质量问题及其潜在影响。

质量问题的描述应明确问题表现、发现方式、涉及版本和配置、影响范围、相关证据及初步原因。必要时，应区分Skill本体、底层模型、知识资源、外部工具、系统配置和操作因素的影响。

11.2 问题分级与处置

组织方应结合问题的严重程度、发生可能性、影响范围、可控程度和法律后果，对质量问题进行分级管理。

对可能导致违法违规输出、重大权益损害、重要数据泄露、错误法律依据或其他重大风险的问题，应采取暂停使用、限制功能、人工复核、紧急修复或其他必要的控制措施。

对一般质量问题，应明确责任人员、整改措施、完成期限和验证要求；对重复发生或长期未解决的问题，应分析系统性原因并采取预防措施。

11.3 整改与复评

整改措施应与问题原因和风险等级相适应，可以包括修订提示词或规则、更新知识资源、调整模型或工具配置、完善权限控制、增加人工复核、优化交互提示或改进异常处置机制。

整改完成后，应对受影响的评测维度、指标或功能进行复评。复评应说明复评范围、验证方法、使用的版本和配置，以及整改前后的结果变化。

当整改可能影响其他功能、指标或安全边界时，应同步开展关联影响评测，不应仅验证问题表象是否消失。

11.4 持续评测触发情形

出现下列情形之一时，宜重新评测、开展专项评测或确认原评测结论的适用性：

- a) 法律AI技能、底层模型、知识资源、外部工具、接口、提示词、规则、权限或关键配置发生可能影响质量的变更；
- b) 适用的法律法规、司法解释、司法政策或其他重要参考依据发生变化；
- c) 发生重大质量问题、安全事件、数据泄露、异常输出、用户投诉或争议；
- d) 应用范围、服务对象、自动化程度或使用方式发生明显变化；
- e) 运行监测显示质量指标持续下降，或原评测集、参考依据和判定规则已不能反映实际风险。

11.5 持续监测与趋势分析

对持续运行或反复调用的法律AI技能，组织方宜建立与风险等级相适应的质量监测机制，关注安全与合规、法律专业性、任务完成、异常率、稳定性和用户反馈等变化。

监测结果应支持问题定位、版本比较和改进决策。监测数据的采集、分析和使用应符合数据安全、个人信息保护和内部权限管理要求。

11.6 版本与改进记录

组织方应记录影响评测结论的版本变更、问题分析、整改措施、复评结果和结论调整情况，并保证记录与具体版本和配置相对应。

涉及评测结论对外使用的重大变更，应及时更新相关说明、使用限制和披露内容，避免继续使用已失效或适用边界发生变化的结论。

12 评测资料管理

12.1 资料分类

评测资料应根据其内容、敏感程度、使用目的和保存要求进行分类管理。评测资料至少包括评测方案、评测集、参考依据、运行环境记录、输入输出、日志、评审记录、复核记录、问题记录、整改记录和评测报告。

组织方应明确各类资料的管理责任、访问权限、使用范围、保存期限和处置方式。不同资料之间具有关联关系的，应保留必要的关联标识。

12.2 资料安全与访问控制

评测资料的收集、存储、使用、传输、共享和销毁应采取与风险相适应的安全措施，防止未经授权的访问、复制、篡改、泄露、丢失或不当使用。

涉及个人信息、敏感个人信息、商业秘密、司法数据、案件材料或其他受保护信息的，应依法确定处理依据和使用范围，并采取去标识化、匿名化、加密、权限分级、操作审计和保密管理等必要措施。

评测集和评测结果不应被用于超出评测目的的训练、产品展示、宣传或其他处理活动；确需改变用途的，应重新确认合法性、必要性和授权边界。

12.3 资料完整性与可追溯性

评测资料应能够反映评测活动的真实过程和结果。组织方宜采用版本号、时间戳、操作记录、校验信息或其他适宜方式，保证关键资料的完整性和变更可追溯。

对输入输出、自动化评测结果和人工评审记录，应保留必要的上下文信息，以支持对评测结论的复核和合理再现。

12.4 资料共享与对外提供

向评测参与方、委托方、监管机构、认证或评测机构提供评测资料时，应明确提供范围、使用目的、保密责任和安全要求。

对外发布评测结果时，应根据发布目的对个人信息、商业秘密、案件信息、评测集内容和其他敏感信息进行必要的脱敏或删减，并同时披露评测条件和适用边界。

12.5 保存与销毁

评测资料的保存期限应与评测目的、风险等级、使用周期、合同约定和适用的法律法规要求相适应。保存期间应保证资料可读取、可检索和可合理复核。

保存期限届满或不再具有保存必要性时，应按照组织方的资料管理制度和适用法律法规要求进行销毁或匿名化处理，并保留必要的处置记录。

附录 A

(资料性)

质量评测框架与指标配置示例

A.1 概述

本附录提供法律AI技能质量评测框架和指标配置的参考方法，供评测方结合评测对象的功能目标、应用范围和风险特征使用。本附录不规定具体业务场景、固定样本、统一权重或统一分值。

A.2 配置依据

指标配置宜综合考虑下列因素：

- a) 法律AI技能的功能目标、能力范围、输入输出形式和依赖组件；
- b) 预期使用对象、使用方式、自动化程度和人工复核安排；
- c) 法律依据的专业性、时效性、地域性和适用条件；
- d) 可能影响个人、组织或公共利益的风险及其可控程度；
- e) 评测目的、评测资源、评测周期和结果使用方式。

A.3 指标层级与类型

评测指标可按照评测维度、子维度和具体指标进行分层设置。指标层级宜保持含义清晰、边界明确，避免同一质量要求在不同层级重复计量。

根据指标在结论中的作用，指标可分为门槛指标、评分指标和观察指标：

- a) 门槛指标，用于识别不应被其他质量结果抵销的重大安全、合规或专业风险；
- b) 评分指标，用于对满足程度进行量化或等级化评测；
- c) 观察指标，用于记录运行表现、改进趋势或暂不适宜量化的质量特征。

A.4 指标内容示例

安全与合规性可关注输出安全、数据和个人信息保护、知识产权、权限边界、拒答及风险提示、人工介入和异常处置。

法律专业性可关注法律依据的准确性、相关性、完整性和时效性，事实与法律规则的匹配，论证一致性，以及对争议和不确定性的识别。

功能与任务质量可关注功能目标实现、任务匹配、任务完成、输出完整性、结果一致性和格式符合程度。

稳定性与鲁棒性可关注重复调用、输入变化、信息缺失、边界条件、干扰信息、依赖异常和对抗性输入下的运行与恢复表现。

交互与使用质量、可管理性可分别关注可理解性、可核验性、反馈提示、版本配置、日志记录、权限控制、监测预警和责任追溯。上述内容仅为指标配置方向，不构成穷举。

A.5 权重、门槛与适用性

评测方采用权重时，应根据评测目的、主要功能和风险等级确定，并记录权重设置和调整理由。不同评测对象之间的权重不宜简单照搬。

对可能造成重大后果的安全与合规缺陷，宜设置门槛条件或限制性结论。门槛条件应明确触发情形、复核要求和处置方式。

指标不适用于评测对象的，应说明不适用理由；指标存在前提条件的，应将前提条件纳入评测方案和结论适用边界。

A.6 配置结果记录

指标配置表或其他配置记录宜至少载明评测维度、指标名称、指标含义、适用对象、评测方法、证据要求、判定规则、权重或门槛、责任人员和配置依据。

附 录 B

（资料性）

评测集设计与维护指南

B.1 评测集定位

评测集用于支持法律AI技能质量评测的实施、比较、复核和持续改进。评测集应服务于明确的评测目标，不宜将与目标无关的样本简单叠加。

B.2 样本构成

评测样本宜根据评测对象的功能和风险进行组织，并在必要时记录样本编号、输入材料、事实条件、任务要求、参考依据、评测要点、适用指标、难度或风险属性、适用时点和使用限制。

对于开放性或多答案任务，评测样本宜记录可接受结果的范围、必要的评测要点和不可接受情形，不宜将某一固定表述作为唯一正确答案。

B.3 样本来源与合规

评测样本可来源于公开资料、依法取得或获得授权的数据、人工编制材料、经适当处理的历史材料或合成数据。样本来源、取得方式、权属情况和使用限制应能够被说明。

样本涉及个人信息、敏感信息、商业秘密、司法数据、案件材料或其他受保护信息的，应依法确定处理依据和使用范围，并采取去标识化、匿名化、访问控制和保密等措施。

B.4 样本覆盖与平衡

评测集宜覆盖评测对象的主要能力、典型输入、正常使用、信息缺失、边界条件和可预见风险，并根据评测目的保持样本难度和风险分布的适当平衡。

样本覆盖应与评测对象的实际声明和使用边界相适应，不要求穷举全部法律领域、法律问题或应用场景。对未覆盖的内容，应在评测报告中说明。

B.5 参考依据与标注

参考依据宜包括适用的法律法规、司法解释、司法政策、规范性文件、权威案例、行业规则或其他经确认的专业资料，并注明来源、版本或适用时点。

样本标注应围绕评测指标和判定规则开展。对于存在法律解释分歧、事实认定空间或结论不确定性的样本，宜记录争议点、可接受差异和人工评审要求。

B.6 评测集维护

评测集应实施版本管理，记录新增、删除、修改、替换和停用样本的原因、时间及影响。

出现法律依据变化、参考结果错误、样本泄露或污染、区分度下降、样本不再具有代表性或评测对象功能发生变化等情形时，应及时复核和更新评测集。

评测集维护过程中应防止评测样本被不当披露、重复使用或用于不符合约定目的的数据处理活动。