

团 体 标 准

T/DGAG XXXX—XXXX

自然资源领域法律法规智能体应用能力 评估规范

Specification for assessment of intelligent agent application capability in natural
resources laws and regulations

（征求意见稿）

在提交反馈意见时，请将您知道的相关专利连同支持性文件一并附上。

XXXX – XX – XX 发布

XXXX – XX – XX 实施

广东省数字政务协会 发 布

目 次

前言 II

1 范围 1

2 规范性引用文件 1

3 术语和定义 1

4 缩略语 1

5 总体要求 1

 5.1 评估要求 1

 5.2 评估职责 2

 5.3 评估依据 2

 5.4 测试数据集要求 2

6 评估流程 3

 6.1 成立评估小组 3

 6.2 征集参评对象 2

 6.3 开展评估工作 3

 6.4 评估结果 4

7 评估应用 4

附录 A（规范性） 测试指标及计算公式要求 5

 A.1 指标体系 5

 A.2 指标计算公式 6

附录 B（规范性） 指标任务分配规则 8

 B.1 指标匹配规则 8

 B.2 任务目录 8

 B.3 指标任务匹配 8

 B.4 自然资源业务场景应用指标组合示例 10

 B.5 自然资源专属业务条线场景指标组合示例 11

附录 C（规范性） 参评材料格式及填写要求一览表 13

参考文献 14

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。

本文件由广州市基础地理信息中心提出。

本文件由广东省数字政务协会归口。

本文件起草单位：

本文件主要起草人：

自然资源领域法律法规智能体应用能力评估规范

1 范围

本文件规定了自然资源领域法律法规智能体应用能力评估的总体要求、评估流程、评估结果和结果应用。

本文件适用于自然资源领域业务主管部门或第三方评估机构，对应用于土地资源、矿产资源、测绘地理信息、国土空间规划、海洋、林草湿地、不动产登记等自然资源业务场景的法律法规智能体，开展基础能力和场景实际应用价值的测试评估工作。

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 45654—2025 网络安全技术 生成式人工智能服务安全基本要求

3 术语和定义

下列术语和定义适用于本文件。

3.1

自然资源领域法律法规

涉及土地、矿产、测绘地理信息、国土空间规划、森林、草原、海洋等自然资源管理关系的法律、行政法规、部门规章、地方性法规及规范性文件的统称。

3.2

通用问答应用服务

基于基础LLM+意图识别+工具调用+提示词工程+RAG检索增强构建的通用法律咨询问答服务。

3.3

场景应用服务

面向自然资源行政审批、行政公开、行政处罚、复议诉讼等业务场景，提供请求分析、行政行为事实分析、争议焦点分析、历史类案分析、法律适用分析与文书审查等服务的智能体应用服务。

4 缩略语

下列缩略语适用于本文件。

LLM：大语言模型（Large Language Model）

RAG：检索增强生成（Retrieval-Augmented Generation）

JSON：JavaScript对象表示法（JavaScript Object Notation）

QA：问答（Question and Answer）

LCS：最长公共子序列（Longest Common Subsequence）

TP：真正例（True Positive）

FP：假正例（False Positive）

FN：假反例（False Negative）

5 总体要求

5.1 评估要求

评估自然资源领域法律法规智能体的应用能力，包括以下三个：

- a) 基础记忆能力：评估智能体是否具备法条与业务知识的准确召回与复述能力，可在知识检索与输出环节实现零偏差的信息还原；
- b) 理解分析能力：评估智能体是否具备对法律位阶关系的准确辨析能力、对法条引用逻辑的严密把握能力，以及对法规时效状态的独立判定能力；
- c) 应用推理能力：评估智能体在法律适用场景下法条匹配的高度准确性，是否能够显著提升法律分析报告的质量；在请求分析与文书审查两类核心场景中，是否达到可直接落地应用的服务水准。

5.2 评估职责

评估机构的职责包括：

- a) 制定指标：依据本规范附录 A 指标体系，结合参评对象类别与业务场景，制定具体测评方案、指标权重配置与评分细则；
- b) 接受申请：受理自然资源领域业务主管部门、智能体开发或运营单位的评估申请，对参评对象资质、版本与提交材料进行核验；
- c) 组织开展：组建评估小组，部署自动化测试工具，调度教师模型与组织专家，按本规范第 6 章流程开展三类协同测评；
- d) 输出报告：汇总自动化指标、教师模型评分与专家评分结果，按本规范第 6.5 章要求生成评估报告，并按统一格式提交。

5.3 评估依据

评估机构按照指标体系（参见附录A）实施评估。

评估内容围绕自然资源领域法律法规智能体的三维度能力展开，包括：

- a) 基础记忆能力：法条与业务知识的准确召回与复述；
- b) 理解分析能力：法律位阶、引用逻辑与时效性的辨析；
- c) 应用推理能力：法律适用场景下的条文匹配、案件分析与文书/报告生成。

5.4 评估对象

参评对象主要为已部署运行的自然资源法律法规智能体所提供的各类应用服务，包括以下两大类：

- a) 通用问答服务；
- b) 场景应用服务。

5.5 测试数据集要求

测试数据集是评估工作的基础，应满足以下要求：

- a) 数据来源：测试数据应来源于自然资源领域法律法规条文、真实业务案例及专家标注样本，涵盖国家、省、市三级法规文件；
- b) 案例来源：真实业务案例宜取自自然资源行政审批卷宗、行政执法案卷、行政复议与诉讼案例（脱敏后使用）；
- c) 数据规模：通用问答测试集不少于 500 条，场景应用测试集不少于 200 条；
- d) 标注规范：每条测试数据应包含问题文本、参考答案、参考法条及评分要点，标注须由 2 名以上领域专家交叉审核；
- e) 边界覆盖：数据集须覆盖时效性案例（含已废止或修订法规）、争议性案例及边界案例，边界案例占比不低于 15%；
- f) 业务条线覆盖：数据集应覆盖土地、矿产、测绘地理信息、国土空间规划、海洋、林草等业务条线，各条线案例占比宜均衡；
- g) 数据安全：涉及个人信息、国家秘密及商业秘密的数据须脱敏处理后使用，数据存储与传输应符合 GB/T 45654-2025 的要求；
- h) 版本管理：测试数据集应标注法规截止日期，定期更新以反映法规修订动态，每次评估须记录所用数据集版本。

6 评估流程

6.1 成立评估小组

评估机构根据参评对象类别组建评估小组，小组应包括：

- a) 项目负责人：负责评估全过程组织与结果审定；
- b) 自动化测试工程师：负责测试工具开发、数据导入与自动化指标计算；
- c) 教师模型测评工程师：负责教师模型选型、评分细则生成与评分执行；
- d) 领域专家：自然资源领域政策法规专业人员与法律专家，负责场景类任务的主观评分与校准。

6.2 材料收集

参评单位应按照下列要求提交参评材料（详见附录C）：

- a) 提交的材料应真实、准确、完整，不得包含涉密内容及个人敏感信息，涉及第三方知识产权的应附授权使用证明；
- b) 材料以电子版方式提交，文字类材料采用 Word 或 PDF 格式，数据类材料采用 Excel 或 CSV 格式，演示视频采用 MP4 格式，扫描件采用 PDF 格式且图像清晰可辨；
- c) 各项材料的具体内容、格式及填写要求见附表；
- d) 材料经参评单位审核并加盖公章后，按征集通知规定的方式和时限提交；逾期提交或材料不全的，不予受理。

6.3 开展评估工作

6.3.1 评估场景

评估场景分为通用问答与场景应用。

通用问答覆盖法律条文记忆、业务知识记忆、法律体系理解、法律时效性识别等基础能力；

场景应用覆盖自然资源行政审批、行政公开、行政处罚、复议诉讼四类真实业务场景，以请求分析、行政行为事实分析、争议焦点分析、历史类案分析、法律适用分析与文书审查为核心任务。

如需评估其他自然资源业务场景，可结合以上两类场景，可参考附录B.4内容，进行指标组合的设计。

6.3.2 评估任务（示例见附录）

针对不同应用场景设计评估任务，明确指标组合规则与权重配置。任务设计应满足：

- a) 自动化指标与专家评分协同设置；
- b) 教师模型评分须配置领域专家校准；
- c) 专家评分须覆盖全部能力维度；
- d) 数据集须覆盖边界案例与时效性案例；
- e) 测评结果须支持可追溯分析。

6.3.3 任务匹配（匹配机制见附录）

评估任务与参评对象、评估方法之间按以下规则匹配：

- a) 通用问答服务：将从基础记忆能力、理解分析能力两个维度，验证智能体对自然资源领域法律法规与业务知识的掌握程度，验证范围涵盖法律条文记忆、业务知识记忆、法律体系理解、法律时效性识别、案件要素提取、法律风险识别等多项能力；
- b) 场景应用评估：面向场景应用服务，在真实案件场景中，从申请人请求分析、行政行为事实分析、争议焦点分析、历史类案分析、法律适用分析、文书审查等六个维度，验证智能体在自然资源法律法规领域的应用推理能力。测评单位可根据自然资源法律法规智能体实际应用场景进行任务匹配，场景任务匹配示例详见附录 B.3.2。

6.3.4 评估方法

评估可采用多维度协同测评模式（测评指标见附录A.2），具体如下：

- a) 自动化测试：F1 值、ROUGE-N、ROUGE-L、BLEU-4、检索召回率/精确率等客观指标，可通过开发批量测试工具，实现测试数据的导入与计算；
- b) 教师模型测评：准确性、完整性、相关性、有效性、一致性等指标，可选用具备中文法律领域能力的闭源或开源大模型作为教师模型，包括但不限于：DeepSeek-V3、文心一言、Qwen-Max、ChatGLM、豆包等；由评估机构根据测评任务与时效性要求选定 1 至 3 个教师模型，教师模型须与待评模型同属中文语种、版本可追溯。教师模型基于标注参考样本生成测试任务、标准答案与评分细则，并对待评模型/系统服务的输出进行评分；
- c) 专家评估：准确性、完整性、可参考性、高价值性等指标，由自然资源领域政策法规类专业人员与法律专家组成专家组，对场景类任务进行主观评分。

6.4 评估结果

6.4.1 形成结果

汇总自动化指标、教师模型评分与专家评分三类结果，按附录 A.2 公式计算各维度得分，并根据附录B确定各任务的指标权重计算综合得分；按预设得分区间（0~0.2、0.2~0.4、0.4~0.6、0.6~0.8、0.8~1.0）开展分布诊断，识别能力短板。

6.4.2 结果复审

对得分异常或专家分歧大的任务进行复审，复审范围包括：

- a) 异常案例记录与 error analysis；
- b) 教师模型评分与专家评分差异超过预设阈值（默认 15%）的任务；
- c) 边界、争议性与时效性案例。

6.4.3 评估报告

评估报告应至少包含：

- a) 评估概述（目的、范围、依据）；
- b) 被评对象基本信息（版本号、模型基座、知识库日期、参评对象类别）；
- c) 测试数据集说明（来源、规模、标注说明）；
- d) 指标与权重配置；
- e) 各维度得分及等级评定；
- f) 得分分布分析图表；
- g) 能力短板诊断；
- h) 与历史版本的迭代对比（如有）；
- i) 优化建议（数据/模型/提示词/系统架构层面）。

7 评估应用

评估结果可应用于以下方面：

- a) 智能体优化迭代：识别能力短板，指导数据扩充、模型微调、提示词工程与系统架构改进；
- b) 智能体选型与采购：为业务主管部门遴选合规、合用、安全的智能体提供第三方评价依据；
- c) 智能体上线验收：作为项目验收与上线投入运行的客观依据；
- d) 智能体持续监测：纳入政务智能体常态化监测体系，定期复评并形成能力画像；
- e) 行业自律与标准符合性评价：作为行业内智能体应用能力等级评定与互认的依据。

附录 A
(规范性)
测试指标及计算公式要求

A.1 指标体系

本标准指标体系按三维度能力组织，对应评估任务与评估方法如下：

- a) 基础记忆能力：包含法律条文记忆、业务知识记忆两项任务；由自动化指标 + 教师模型评分协同得出；
- b) 理解分析能力：包含法律体系理解、法律时效性识别、案件要素提取、法律风险识别四项任务；由自动化指标 + 教师模型评分协同得出；
- c) 场景应用推理能力：包含申请人请求分析、行政行为事实分析、争议焦点分析、历史类案分析、法律适用分析、文书审查六项任务；由教师模型评分 + 专家评估协同得出。

表 A.1 给出了指标体系明细表。

表 A.1 指标体系明细表

服务场景	能力	任务	指标名称
通用问答服务	基础记忆能力	法律条文记忆	自动化指标（ROUGE-L）、教师模型评分
		业务知识记忆	自动化指标（ROUGE-L）、教师模型评分
	理解分析能力	法律体系理解	自动化指标（ROUGE-L、F1-SCORE）、教师模型评分
		法律时效性识别	自动化指标（ROUGE-L、F1-SCORE）、教师模型评分
		案件要素提取	自动化指标（F1-SCORE）、教师模型评分
		法律风险识别	自动化指标（ROUGE-L）、教师模型评分
场景应用服务	场景应用推理能力	申请人请求分析	自动化指标（ROUGE-L）、教师模型评分、专家评估
		行政行为事实分析	自动化指标（ROUGE-L）、教师模型评分、专家评估
		争议焦点分析	自动化指标（ROUGE-L）、教师模型评分、专家评估
		历史类案分析	教师模型评分 + 专家评估
		法律适用分析	教师模型评分 + 专家评估
		文书审查	自动化指标（F1-SCORE、ROUGE-L）、教师模型评分、专家评估

A.1.1 自动化指标

自动化指标由批量测试工具得出，指标定义见表A.2。

表 A.2 基础性能指标

指标	定义
F1 值	精确率与召回率的调和均值
ROUGE-N	生成文本与参考文本的 N-gram 重叠度
ROUGE-L	基于最长公共子序列的相似度
BLEU-4	4-gram 修正精确率的几何平均

A.1.2 教师模型指标

任务性能指标由教师模型评分得出，采用 5 维度评分体系，各维度取 1~5 整数分，按预设权重计算 Teacher-LLM-Score，指标定义见表A.3。

表 A.3 教师模型评分指标

指标	定义
准确性	生成内容的真实程度及符合场景任务知识体系的一致性

完整性	生成内容是否能够覆盖相关能力任务要求的信息种类
相关性	生成的内容能够与用户输入信息的相关性
一致性	多次生成的内容语义和内涵相同的程度
有效性	生成内容符合问题需求、能够有效解决场景任务中的问题的程度

A.1.3 专家评估指标

专家评估指标由自然资源领域政策法规人员与法律专家按场景任务主观评分，采用 5 维度评分体系，按预设权重计算FeedBackScore。指标与适用任务见表A.4。

表 A.4 专家评估指标

指标	定义
准确性	生成内容符合法律法规与业务实际要求，不存在事实性错误的程度
完整性	生成内容完整覆盖场景任务要求的全部核心要点，无关键信息遗漏的程度
可参考性	生成内容能够为业务办理提供明确指引、具备实际参考应用价值的程度
高价值性	生成内容能够深度匹配自然资源业务的特殊性需求，解决复杂场景下实际业务问题的价值程度

A.2 指标计算公式

A.2.1 教师模型评分计算公式

教师模型评分计算方式见公式（A.1）。

$$\text{Teacher-LLM-Score} = (\sum (\text{维度得分}_i \times \text{维度权重}_i)) \times 2 / 10 \dots\dots\dots (\text{A.1})$$

式中：

维度得分_i —— 教师模型对第 i 个维度给出的评分，取值 1~5 的整数；

维度权重_i —— 第 i 个维度的预设权重，各维度权重之和为 1；

Teacher-LLM-Score —— 标准化后的教师模型评分，取值范围 [0,1]。

A.2.2 F1 值计算公式

F1 值计算方式见公式（A.2）。

$$\text{精确率: } P = TP / (TP + FP)$$

$$\text{召回率: } R = TP / (TP + FN)$$

$$F1 = 2 \times P \times R / (P + R) \dots\dots\dots (\text{A.2})$$

式中：

TP —— 真正例数量；

FP —— 假正例数量；

FN —— 假反例数量。

A.2.3 ROUGE-N 计算公式

ROUGE-N 计算方式见公式（A.3）。

$$ROUGE - N = \frac{\sum_{S \in \{Reference\}} \sum_{gram_n \in S} Count_{match}(gram_n)}{\sum_{S \in \{Reference\}} \sum_{gram_n \in S} Count(gram_n)} \dots\dots\dots (\text{A.3})$$

式中：

gramn——表示一个n-gram，例如，假设我们以2-gram，那么gramn可以是摘要中的“建设用地”“行政复议”等词对；

Countmatch(gramn)——表示生成文本与参考文本中共同出现的n-gram最大数量；

Count(gramn)——表示参考摘要中n-gram的总数，即统计所有出现的n-gram总量。

A.2.4 ROUGE-L 计算公式

ROUGE-L计算方式见公式（A.4）。

$$\text{ROUGE-L} = (1 + \beta^2) \times R_{\text{lcs}} \times P_{\text{lcs}} / (R_{\text{lcs}} + \beta^2 \times P_{\text{lcs}}) \cdots \cdots \cdots (\text{A.4})$$

式中：

β —— 预设参数，通常取 1；

LCS —— 生成文本与参考文本的最长公共子序列长度；

R_{lcs} —— $\text{LCS} / \text{参考文本长度}$ ； $P_{\text{lcs}} = \text{LCS} / \text{生成长度}$ 。

A.2.5 BLEU-4 计算公式

BLEU-4 计算方式见公式 (A.5)。

$$\text{BLEU} = \text{BP} \times \exp(\sum(w_n \times \log p_n)) \cdots \cdots \cdots (\text{A.5})$$

式中：

p_n —— 修正的 $n\text{-gram}$ 精确率 ($n=1,2,3,4$)；

w_n —— 权重，通常各取 25%；

BP —— 简短惩罚因子， $\text{BP} = \min(1, \exp(1 - \text{参考长度} / \text{生成长度}))$ 。

A.2.6 专家评分计算公式

专家评分计算方式见公式 (A.6)。

$$\text{FeedBackScore} = (\sum (\text{维度得分}_i \times \text{维度权重}_i)) \times 2 / 10 \cdots \cdots \cdots (\text{A.6})$$

注：多专家评分时，取所有专家评分的算术平均值。

附录 B
(规范性)
指标任务分配规则

B.1 指标匹配规则

指标匹配规则如下：

- a) 通用问答服务：从基础记忆能力与理解分析能力两个维度配置指标组合，以自动化指标与教师模型评分协同评估，权重配置见 B.3.1；
- b) 场景应用服务：从应用推理能力维度配置指标组合，以教师模型评分与专家评分协同评估为主，辅以自动化指标，权重配置见 B.3.2；
- c) 自然资源其他业务场景应用服务：从场景涉及的能力维度出发，设置混合指标，权重按各子能力维度占比分配，可参考 B.4 示例。

B.2 任务目录

评估任务按基础记忆、理解分析、应用推理三个能力维度组织，覆盖法律条文记忆、法律体系理解、争议焦点分析等 12 项任务，详见表B.1。

表 B.1 评估任务目录

序号	能力维度	任务名称	任务说明	权重
1	基础记忆	法律条文记忆	考察智能体对自然资源领域法律法规条文的准确复述能力	7%
2		业务知识记忆	考察智能体对自然资源业务规范、审批要件、流程等业务知识的掌握	8%
3	理解分析	法律体系理解	考察智能体对法律位阶、新旧法/上下位法关系的辨析能力	9%
4		法律时效性识别	考察智能体对法条生效/废止/过渡条款的独立判定能力	8%
5		案件要素提取	从自然语言案件描述中抽取关键事实与法律要素	8%
6		法律风险识别	识别业务场景下的合规风险与法律风险点	8%
7	应用推理	申请人请求分析	在自然资源行政审批场景下解析申请人诉求与诉求项	8%
8		行政行为事实分析	对行政行为的合法性与事实依据进行分析	9%
9		争议焦点分析	在复议诉讼场景下识别争议焦点并归类	9%
10		历史类案分析	检索匹配类案并提炼裁判规则	9%
11		法律适用分析	在具体场景中匹配可适用的法律规范	9%
12		文书审查	对法律文书/审批文件进行合规性审查	8%

B.3 指标任务匹配

B.3.1 通用问答测评指标组合

表B.2给出了法律条文记忆指标。

表 B.2 法律条文记忆

测评内容	指标名称	指标类型	计算方式	建议权重
法律条文记忆能力	ROUGE-L	自动化	ROUGE-L	50%
	教师模型评分	教师模型	Teacher-LLM-Score=(0.3*准确性+0.2*完整性+0.5*一致性) *2/10	50%

表B.3给出了业务知识记忆指标。

表 B. 3 业务知识记忆

测评内容	指标名称	指标类型	计算方式	建议权重
业务知识记忆能力	ROUGE-L	自动化	ROUGE-L	30%
	教师模型评分	教师模型	Teacher-LLM-Score=(0.3*准确性+0.3*完整性+0.4*相关性)*2/10	70%

表B.4给出了法律体系理解指标。

表 B. 4 法律体系理解

测评内容	指标名称	指标类型	计算方式	建议权重
效力位阶关系	ROUGE-L	自动化	ROUGE-L	20%
从属关系	ROUGE-L	自动化	ROUGE-L	20%
法条引用关系	F1-SCORE	自动化	F1-SCORE	20%
适用规则判断	ROUGE-L、教师模型评分	自动化+教师模型	0.5*ROUGE-L+0.2*(准确性+完整性+相关性+有效性+一致性)*2/10)	40%

表B.5给出了法律时效性识别指标。

表 B. 5 法律时效性识别

测评内容	指标名称	指标类型	计算方式	建议权重
时效状态判断	F1-SCORE、教师模型评分	自动化+教师模型	0.6*F1-SCORE+0.4*((0.4*准确性+0.3*完整性+0.3*相关性)*2/10)	30%
修订内容对比	ROUGE-L、教师模型评分	自动化+教师模型	0.5*ROUGE-L+0.5*((0.4*准确性+0.3*完整性+0.3*相关性)*2/10)	30%
特定时期适用法条	ROUGE-L、教师模型评分	自动化+教师模型	0.5*ROUGE-L+0.5*((0.4*准确性+0.3*完整性+0.3*相关性)*2/10)	40%

表B.6给出了案件要素提取指标。

表 B. 6 案件要素提取

测评内容	指标名称	指标类型	计算方式	建议权重
案件要素提取	F1-SCORE	自动化	F1-SCORE	50%
	教师模型评分	教师模型	(0.4准确性+0.4完整性+0.2*一致性)*2/10	50%

表B.7给出了法律风险识别指标。

表 B. 7 法律风险识别

测评内容	指标名称	指标类型	计算方式	建议权重
法律风险识别	ROUGE-L	自动化	ROUGE-L	50%
	教师模型评分	教师模型	(0.3*准确性+0.3*完整性+0.2*相关性+0.2*有效性)*2/10	50%

B. 3. 2 场景应用指标组合

表B.8给出了申请人请求分析指标。

表 B. 8 申请人请求分析

测评内容	指标名称	指标类型	计算方式	建议权重
申请人请求分析	ROUGE-L	自动化	ROUGE-L	10%
	教师模型评分	教师模型	(0.4*准确性+0.4*完整性+0.2*相关性)*2/10	20%

	专家评分	专家评分	专家评分结果*2/10	70%
--	------	------	-------------	-----

表B.9给出了行政行为事实分析指标。

表 B. 9 行政行为事实分析

测评内容	指标名称	指标类型	计算方式	建议权重
行政行为事实分析	ROUGE-L	自动化	ROUGE-L	10%
	教师模型评分	教师模型	$(0.4*准确性+0.4*完整性+0.2*相关性)*2/10$	20%
	专家评分	专家评分	专家评分结果*2/10	70%

表B.10给出了争议焦点分析指标。

表 B. 10 争议焦点分析

测评内容	指标名称	指标类型	计算方式	建议权重
争议焦点分析	ROUGE-L	自动化	ROUGE-L	10%
	教师模型评分	教师模型	$(0.4*准确性+0.4*完整性+0.2*相关性)*2/10$	20%
	专家评分	专家评分	专家评分结果*2/10	70%

表B.11给出了历史类案分析指标。

表 B. 11 历史类案分析

测评内容	指标名称	指标类型	计算方式	建议权重
评案类检索结果	教师模型评分、专家评分	教师模型+专家评分	$0.3*(0.4*准确性+0.4*完整性+0.2*有效性)*2/10+0.7*专家评分结果*2/10$	30%
评案类洞察分析报告	教师模型评分、专家评分	教师模型+专家评分	$0.4*(0.4*准确性+0.4*完整性+0.2*有效性)*2/10+0.6*专家评分结果*2/10$	70%

表B.12给出了法律适用分析指标。

表 B. 12 法律适用分析

测评内容	指标名称	指标类型	计算方式	建议权重
评案类检索结果	教师模型评分、专家评分	教师模型+专家评分	$0.3*(0.4*准确性+0.4*完整性+0.2*有效性)*2/10+0.7*专家评分结果*2/10$	30%
评案类洞察分析报告	教师模型评分、专家评分	教师模型+专家评分	$0.4*(0.4*准确性+0.4*完整性+0.2*有效性)*2/10+0.6*专家评分结果*2/10$	70%

表B.13给出了文书审查指标。

表 B. 13 文书审查

测评内容	指标名称	指标类型	计算方式	建议权重
文书审查	F1-SCORE	自动化	F1-SCORE	10%
	ROUGE-L	自动化	ROUGE-L	10%
	教师模型评分	教师模型	$(0.4*准确性+0.4*完整性+0.2*相关性)*2/10$	30%
	专家评分	专家评分	专家评分结果*2/10	50%

B. 4 自然资源业务场景应用指标组合示例

表B.14给出了自然资源行政审批场景指标组合。

表 B. 14 自然资源行政审批场景指标组合

指标名称	指标类型	建议权重	建议阈值
F1-Score	自动化	20%	≥ 0.92
ROUGE-N	自动化	20%	≥ 0.85
ROUGE-L	自动化	15%	≥ 0.82

BLEU-4	自动化	15%	≥ 0.75
教师模型评分	教师模型	15%	≥ 0.80
专家评分	专家评分	15%	≥ 0.82

表B.15给出了自然资源行政公开场景指标组合。

表 B. 15 自然资源行政公开场景指标组合

指标名称	指标类型	建议权重	建议阈值
F1-SCORE	自动化	40%	≥ 0.90
ROUGE-L	自动化	15%	≥ 0.82
BLEU-4	自动化	15%	≥ 0.78
教师模型评分	教师模型	15%	≥ 0.82
专家评分	专家评分	15%	≥ 0.82

表B.16给出了自然资源行政处罚场景指标组合。

表 B. 16 自然资源行政处罚场景指标组合

指标名称	指标类型	建议权重	建议阈值
F1-Score	自动化	25%	≥ 0.93
ROUGE-N	自动化	12%	≥ 0.85
ROUGE-L	自动化	18%	≥ 0.80
教师模型评分	教师模型	22%	≥ 0.80
专家评分	专家评分	23%	≥ 0.82

表B.17给出了自然资源复议诉讼场景指标组合。

表 B. 17 自然资源复议诉讼场景指标组合

指标名称	指标类型	建议权重	建议阈值
F1-Score	自动化	25%	≥ 0.85
ROUGE-N	自动化	12%	≥ 0.82
BLEU-4	自动化	10%	≥ 0.72
ROUGE-L	自动化	8%	≥ 0.72
教师模型评分	教师模型	22%	≥ 0.78
专家评分	专家评分	23%	≥ 0.80

B. 5 自然资源专属业务条线场景指标组合示例

面向自然资源调查监测、国土空间规划、用途管制、耕地保护、矿产资源、不动产登记等专属业务条线，可结合本附录 B.4 行政四类场景设计指标组合。以下示例侧重各业务条线的法规适用与事实判定特点，可与行政四类场景组合使用。

表B.18给出了国土空间规划合规性审查场景指标组合。

表 B. 18 国土空间规划合规性审查场景指标组合

指标名称	指标类型	建议权重	建议阈值
F1-Score	自动化	15%	≥ 0.88
ROUGE-L	自动化	15%	≥ 0.80
BLEU-4	自动化	10%	≥ 0.72
教师模型评分	教师模型	25%	≥ 0.82
专家评分	专家评分	35%	≥ 0.82

表B.19给出了土地征收与供应合法性审查场景指标组合。

表 B. 19 土地征收与供应合法性审查场景指标组合

指标名称	指标类型	建议权重	建议阈值
F1-Score	自动化	20%	≥ 0.90
ROUGE-N	自动化	15%	≥ 0.85
ROUGE-L	自动化	15%	≥ 0.82
教师模型评分	教师模型	25%	≥ 0.82
专家评分	专家评分	25%	≥ 0.82

附 录 C
(规范性)
参评材料格式及填写要求一览表

表C.1给出了参评材料格式及填写要求一览表。

表 C. 1 参评材料格式及填写要求一览表

序号	材料名称	格式要求	填写要求
1	参评申报表	Word 或 PDF, A4 纵向, 按征集通知提供的模板填写	逐项如实填写, 单位名称使用全称并加盖单位公章, 由法定代表人或授权代表签字; 联系人及联系方式应保持畅通
2	参评智能体基本情况说明	Word 或 PDF, 篇幅一般不超过 10 页	说明建设背景、总体架构、功能范围(对应通用问答服务或场景应用服务)、部署运行情况、服务对象与规模等, 内容应与申报的参评类别一致
3	系统功能演示材料	MP4 视频(单个不超过 15 min, 分辨率不低于 1080P)或 PPT、PDF 演示文稿	按评估场景完整演示核心功能与业务流程, 画面清晰、操作连贯, 附必要的文字说明或解说; 演示内容应与申报功能一致, 不得剪辑伪造运行效果
4	接口与部署说明	Word 或 PDF, 接口文档建议采用 OpenAPI (Swagger) 规范	说明系统访问地址、测试账号及有效期、接口清单、请求与响应示例、依赖环境及数据更新频率等, 满足评估小组独立开展测评的需要
5	测试数据集及说明	Excel 或 CSV, 配套说明采用 Word 或 PDF	说明数据集来源、规模、字段结构、标注方式及授权情况; 数据集应覆盖所申报的评估任务, 不得包含涉密内容和个人敏感信息
6	自评估报告	Word 或 PDF, 按本文件规定的测试指标体系编写	逐项对照测试指标给出自评结果、计算过程及佐证材料(截图、日志或数据); 结论表述应客观, 不得夸大或虚报
7	资质及合规证明材料	扫描件 PDF, 页面完整、图像清晰	包括软件著作权证书、生成式人工智能服务备案证明、网络安全等级保护测评报告等; 复印件需加盖单位公章
8	参评承诺函	按模板打印后签字盖章, 扫描为 PDF	承诺所提交材料真实有效、数据来源合法、不侵犯第三方合法权益; 由法定代表人或授权代表签字并加盖单位公章

参 考 文 献

- [1] GB/T 17694—2023 地理信息 术语
 - [2] GB/T 39623—2020 基础地理信息数据库系统质量测试与评价
 - [3] GB/T 45288.1—2025 人工智能 大模型 第1部分：通用要求
 - [4] GB/T 45288.2—2025 人工智能 大模型 第2部分：评测指标与方法
 - [5] GB/T 45081—2024 人工智能 管理体系
 - [6] GB/T 33447—2016 地理信息系统软件测试规范
 - [7] 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
 - [8] GB/T 46800—2025 生成式人工智能技术应用社会影响评估指南
 - [9] GB/Z 42759—2023 智慧城市 人工智能 技术应用场景分类指南
 - [10] T/EGAG 019—2023 信息技术应用创新项目验收规范
-