ICS 35. xxx
CCS Lxx

团　　体　　标　　准

T/ISC XXX—XXXX

数字人产品功能质量分级评估规范

Grading and Classification of Functional Quality for Digital Human Products

（征求意见稿）

XXXX－XX－XX 发布　　XXXX－XX－XX 实施

中国互联网协会　　发布

目 次

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出并归口。

本文件起草单位：

本文件主要起草人：

引 言

随着生成式人工智能和虚拟交互技术的快速发展，数字人已在政务服务、传媒传播、电商直播、教育培训等领域得到广泛应用，成为数字经济的重要组成部分。鉴于数字人产品融合硬件设备、软件系统、算法模型、内容资源等多要素的复杂构成特性，传统技术评估体系难以全面、精准地反映其真实质量水平。

本文件构建 TCEMS 评估模型，涵盖技术实力（T）、内容表现（C）、用户体验与运维（E）、商业可用性（M）、安全合规（S）五大维度，为数字人产品功能质量评估、等级判定和风险控制提供体系化依据。

本文件可作为政企单位、金融机构、电商企业、文旅教育等行业用户在采购、选型和验收数字人系统或服务时的评测依据，也可作为数字人技术提供商开展产品自评、能力对标、等级认证以及行业协会、研究机构制定相关标准或发布评测报告的技术参考依据。

数字人产品功能质量分级评估规范

1 范围

本文件规定了数字人产品（包括 2D 数字人、3D 数字人）功能质量评估的评估框架、核心维度、评分规则、动态权重、总分计算方法及等级划分；主要评价数字人产品的功能质量，不评价模型训练能力、算法先进性、企业经营能力、市场规模和品牌影响力，也不替代网络安全、数据安全、个人信息保护等专项合规要求。

本文件适用于面向政务、企业及行业应用场景的各类数字人相关产品和解决方案的功能质量评测与分级。

2 规范性引用文件

本文件没有规范性引用文件。

3 术语和定义

下列术语和定义适用于本文件。

3.1 数字人产品 digital human product

利用建模、驱动、语音合成、自然语言处理、渲染等技术形成的，可在特定场景中呈现拟人化形象并提供内容表达或交互服务的软件、硬件或服务组合。

3.2 TCEMS 模型 TCEMS model

由技术实力、内容表现、用户体验与运维、商业可用性、安全合规五个核心维度构成的五维评估模型。技术实力衡量数字人的底层技术性能与稳定性。内容表现衡量数字人在实际内容输出中的可用性与可控性。用户体验与运维衡量使用方与最终用户的体验质量。商业可用性关注产品是否具备可规模化落地的能力。安全合规对安全性、合规性、可追溯性进行评估。

3.3 基础权重 base weight

在不区分具体应用场景时，用于计算数字人产品功能质量总分的各评估维度默认权重。基础权重应体现通用场景下各维度对功能质量的相对重要程度。

3.4 场景修正系数 scenario correction coefficient

根据数字人产品应用场景差异，对基础权重进行调整的系数。场景修正系数用于体现特定场景对某一评估维度的强化或弱化需求。

3.5 场景权重 scenario weight

基础权重经场景修正系数调整并归一化后形成的权重，用于特定场景下功能质量总分计算。

4 符号和缩略语

缩略语	中文含义	英文名称
2D	二维	2-Dimensional
3D	三维	3-Dimensional
TCEMS	技术实力-内容表现-用户体验与运维-商业可用性-安全合规	Technology-Content-Experience-Market Usability-Security
AI	人工智能	Artificial Intelligence
API	应用程序接口	Application Programming Interface
RAG	检索增强生成	Retrieval-Augmented Generation

5 评估框架

数字人产品功能质量评估框架（TCEMS 模型）是一个多维度、动态的系统性框架。该框架涵盖技术实力（T）、内容表现（C）、用户体验与运维（E）、商业可用性（M）、安全合规（S）五大维度。

各维度之间的关系为：指标项评分形成维度得分，维度得分经基础权重或场景权重汇总形成产品功能质量总分，总分结合安全合规约束确定产品等级。

6 评估方法

6.1 概述

数字人产品功能质量分级评估应按照统一流程实施。评估活动宜采用材料审查、功能演示、场景测试、数据验证和专家复核相结合的方式开展。评分应以可验证证据为依据，并保持评分口径一致。

6.2 评估流程

图 1 数字人产品功能质量分级评估流程

1. 申请受理：由数字人产品提供方或授权单位提出评估申请，提交产品基本信息、功能说明、部署形态、应用场景、测试账号、接口说明、合规证明等材料，评估机构对材料完整性和符合性进行审查。
2. 评测准备：评估机构确认评测范围、评测环境、统一测试数据集、测试样本、人工标注规则、应用场景、评分表适用性和场景修正系数，形成评测计划。

3. 测试评估：依据本文件规定的评价指标体系，对数字人产品开展功能演示验证、场景测试和必要的数据抽样验证，形成测试记录，并同步留存原始数据、日志、截图、视频和评分依据。
4. 评分汇总：根据测试结果对各指标项进行评分，计算各维度得分，并按基础权重或场景权重计算功能质量总分。
5. 等级判定：依据功能质量总分、分等分级规则和安全合规约束，确定数字人产品的功能质量等级。
6. **结果发布：形成评估结果文件，明确产品名称、版本、评估范围、评分依据、功能质量总分、等级和必要的风险提示。**

6.3 基础权重、动态权重与总分计算

基础权重是各评估维度在通用应用场景下参与综合评分的默认权重，用于体现各维度在产品功能质量评价中的相对重要性。各维度基础权重参考值见表 1。

除评估方案另有规定外，同一维度内各指标项采用等权方式计算维度得分；当针对特定应用场景开展评估时，可依据场景特点对各维度基础权重进行动态调整，并重新归一化后参与综合评分。

表 1 TCEMS 维度基础权重与评分表对应关系

维度	维度名称	基础权重	指标项	评分表
T	技术实力	25%	多模态生成、渲染与建模、动作驱动、智能交互、模型自适应	第 8.4.1～8.4.5 条
C	内容表现	20%	形象逼真度、语音质量、思维能力、动作表现、输出质量	第 8.4.6～8.4.10 条
E	用户体验与运维	20%	平台易用性、响应速度、系统稳定性、客户支持、集成兼容	第 8.4.11～8.4.15 条
M	商业可用性	15%	成本结构、扩展性、交付效率、落地案例、生态兼容	第 8.4.16～8.4.20 条
S	安全合规	20%	数据安全、内容合规、知识产权、算法透明度、人机伦理	第 8.4.21～8.4.25 条

- **指标项得分**

每个指标项依据第 8.4 条对应评分表进行评价，得到指标项得分，记为 Sij。指标项得分范围为 0～100 分。

*说明：Sij 表示第 i 个维度中第 j 个指标项的评分结果。

- **维度得分**

各维度得分由该维度内所有指标项得分按权重计算得到。当未设置专项权重时，各指标项采用等权计算。

$Di = \Sigma(Sij \times Wij) / \Sigma Wij$。

其中：

- **Di** —— 第 *i* 个维度得分；
- **Sij** —— 第 *i* 个维度中第 *j* 个指标项得分；
- **Wij** —— 第 *i* 个维度中第 *j* 个指标项权重；未设置专项权重时，各指标项权重相同。

*说明：维度得分可理解为该维度内所有指标项的加权平均分。

- **场景动态权重**

各维度基础权重可根据具体应用场景进行动态修正。

$W_i' = B_i \times K_i$

其中：

- **Bi** —— 第 *i* 个维度基础权重；
- **Ki** —— 场景修正系数，建议取值范围为 **0.5～1.5**；
- **Wi′** —— 修正后的维度权重。

*说明：场景修正仅用于体现不同应用场景下各维度重要性的差异，不得改变 TCEMS 五维评价体系总体结构，不得合并、拆分、取消或替代任何维度，也不得使任何维度失去评价意义。例如，在电商直播场景中，可适当提高"商业可用性"维度权重；在政务服务场景中，可适当提高"安全合规"维度权重。

- **权重归一化**

为保证所有维度权重之和始终为 100%，应对修正后的权重进行归一化处理。

$W_i = W_i' / \Sigma W_i' \times 100\%$

其中：

- **Wi** —— 归一化后的维度权重。

安全合规维度归一化后权重不得低于 20%；若低于 20%，应将其调整至 20%，其余维度按比例缩减。

*说明：归一化处理仅用于保证各维度权重总和保持 100%，不会改变各维度的重要性排序；场景修正不应使任何维度失去评价意义。

- **综合得分**

产品功能质量综合得分按照各维度得分及对应权重计算。

$Q = \Sigma(Di \times Wi) / 100$。

其中：

- **Q** —— 产品功能质量综合得分；
- **Di** —— 各维度得分；
- **Wi** —— 各维度归一化后的权重。

*说明：综合得分为各维度得分按权重加权求和，满分为 100 分。

6.4 评分档位适用规则

第 8.4 条各评分表应依据评分项证据形态采用区间评分（代表分法）、离散评分、布尔评分三类评分方式。区间评分（代表分法）和离散评分涉及多项关键条件的，应按照“最低满足项”原则判定；布尔评分涉及多个独立判定条件的，应拆分为多个布尔子指标分别评分并按子项权重汇总。

区间评分（代表分法）：适用于准确率、时延、成本等连续性能指标，根据指标实际测量值所在区间赋予对应代表分，不采用线性插值计算。区间统一采用左闭右开表达，最高档测量值区间包含上限端点 100，即 [0,20)、[20,40)、[40,60)、[60,80)、[80,100]；五档代表分统一为 10、30、50、70、90。

离散评分：适用于模态数量、平台数量、API 数量、案例数量、语言数量等能力数量指标，根据能力档位赋予对应代表分。五档代表分统一为 10、30、50、70、90，采用“**最低满足项原则**”：即所评测项必须同时达到对应档位，任一条件未达到时，按较低档计分；对于 API 数量、动作数量、模板数量、案例数量、语言数量等数量型指标，最高档除数量要求外还应同时满足覆盖范围、代表性、可用性和可验证性要求，且覆盖全部规定测试项、证据完整、无重大异常时，方可按满分触发条件取 100 分。

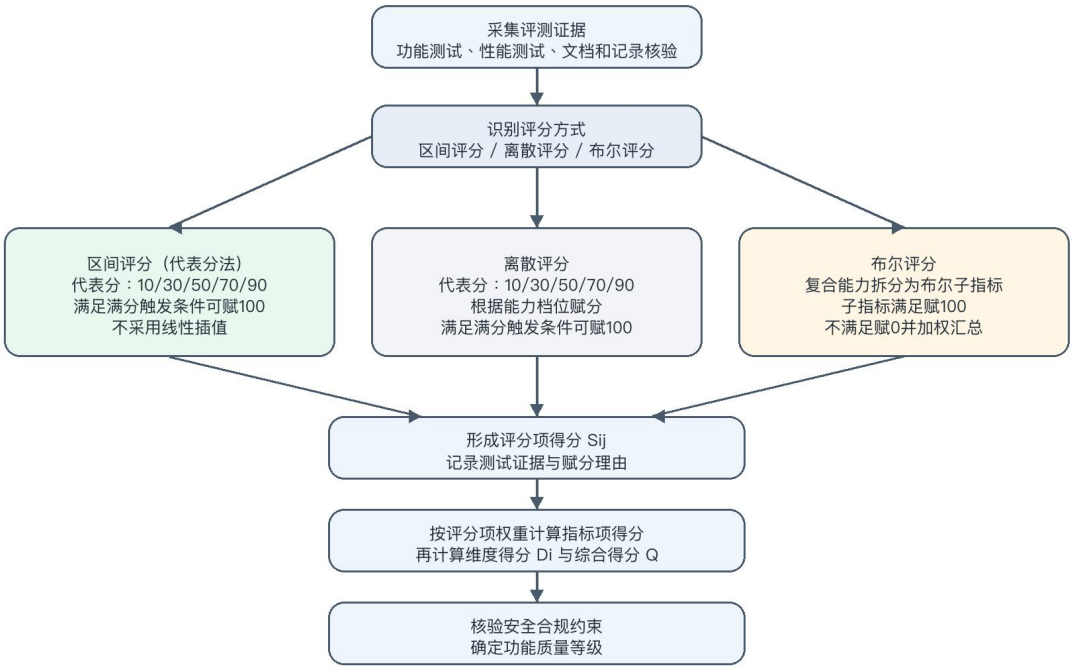
布尔评分：适用于是否支持、是否具备、是否提供、是否建立等功能存在性或合规存在性指标。布尔子指标满足赋 100 分，不满足赋 0 分；包含多个独立判定条件的功能能力，应拆分为多个布尔子指标分别评分，并按子项权重汇总，避免部分满足导致评分歧义。涉及安全合规底线的布尔评分项不满足时，应同时依据第 8.3 条进行等级限制或直接判定不通过。

评分原则：同类指标应采用同类评分方式，区间评分（代表分法）、离散评分和布尔评分不得在同一性质指标中混用；评估记录应载明评分方式、测试证据、档位或取值判定依据和赋分理由，保证评价过程统一、评价结果可重复、可复现。

评分方式可按图 2 执行。

图 2 数字人产品功能质量评分标准流程图

数字人产品功能质量评分标准流程



6.5 统一测试数据集与测量规范

统一测试数据集应由公开基础集、场景专项集、边界异常集和复核留存集组成；同一评分项应使用同一产品版本、同一测试环境、同一脚本、同一输入模板和同一统计口径。

准确率类指标的有效样本数原则上不少于 100 个；当实际可测对象少于 100 个时应全量测试。每个主要类别宜至少覆盖 10 个样本，并覆盖典型、边界和异常场景。

一致率、偏差率类指标宜采用同一输入重复测试不少于 3 次，或不少于 30 组可比较样本；覆盖率类指标应对全部强制对象全量测试，对非强制对象按分层抽样测试，抽样比例不低于 30% 且不少于 20 项。

人工标注应至少由 2 名标注人员独立完成，分歧由第 3 名复核人员裁决；标注规范、金标准和争议处理结果应一并归档。

6.6 量化指标计算口径与定义

涉及准确率、偏差率、一致率、覆盖率等量化指标的，应明确评价对象、计算公式、单位和统计方式。

- 准确率 = 正确样本数 / 总样本数 × 100%；
- 偏差率 = |实测值 − 基准值| / max(|基准值|, ε) × 100%，基准值为 0 时采用绝对偏差；
- 一致率 = 一致结果数 / 比较样本数 × 100%；
- 覆盖率 = 已覆盖项数 / 应覆盖项数 × 100%。
- 时延、响应时间和成本等连续指标应在同一环境、同一版本、同一采样口径下统计，主统计量宜采用中位数或均值，必要时同时给出 P95 或单位成本。

6.7 评分档位设置依据

评分档位设置：综合行业调研、典型产品能力分布、专家咨询和历史测试数据确定。10、30、50、70、90 分分别对应明显不足、基础可用、行业常见、明显优于平均和接近顶级的能力状态；100 分仅作为满分触发结果，不作为常规代表分。

评分档位依据现状、典型产品能力水平、公开技术资料、行业应用实践以及标准编制组专家咨询和产品调研结果确定，遵循可区分、可测量、可实施的原则设置，用于反映当前阶段数字人产品功能质量能力水平。随着技术发展和产业成熟度提升，相关评价阈值可依据标准修订程序进行动态更新。

6.8 数量型指标评价规则

对于 API 数量、动作数量、模板数量、案例数量、语言数量等数量型指标，评分不得仅按数量堆积判定；最高档除数量条件外还应同时满足覆盖范围、代表性、可用性和可验证性要求。评分表应在评价要求栏中明确核心覆盖对象，必要时将数量与质量、覆盖范围或可复测性组合评价。

6.9 动态权重修正规则

场景修正仅用于体现不同应用场景下各维度重要性的差异，不得改变 TCEMS 五维评价体系总体结构，不得合并、拆分、取消或替代任何维度，也不得使任何维度失去评价意义。场景修正后的权重应重新归一化，且安全合规维度仍应保留最低约束。

6.10 评估证据管理要求

评估证据应包括原始测试记录、测试数据、日志、视频、截图、脚本、版本信息、人工标注记录、专家评审记录、异议处理记录和最终评分结果。评估机构应为每项证据赋予唯一编号，注明时间、来源、版本和责任人；证据保存期限不少于 3 年，涉及争议、投诉或复评的，应保存至争议解决后再不少于 3 年。

6.11 术语和表述统一要求

本标准主要评价数字人产品功能质量，不评价模型训练能力、算法先进性、企业经营能力、品牌影响力和市场规模，也不替代网络安全、数据安全、个人信息保护等专项监管要求。全文应统一使用指标项、评分项、测试样本、测试方法、评分规则、维度得分和综合得分等术语，同一概念不得混用不同表述。

7 评估维度与指标要求

7.1 技术实力（T）

衡量数字人产品在核心技术层面的成熟度与先进性，重点评估其在多模态生成、驱动控制、智能交互及模型能力演进等方面的技术实现水平，是数字人产品功能实现的基础能力。

指标项	考察要点	关键指标参考
多模态生成	数字人对文本、语音、图像等多模态信息的理解、融合与生成能力及其一致性表现。	模态覆盖范围、跨模态转换精度、多模态协同能力。
渲染与建模	数字人模型精度、视觉细节表现及在不同终端环境下的实时渲染稳定性。	渲染技术类型、模型细节精度、动态光影效果、实时渲染性能。
动作驱动	数字人面部表情与肢体动作的自然程度及其与语	骨骼动画精度、语义动作生成逻

指标项	考察要点	关键指标参考
	音、语义的同步协调能力。	辑、口型同步延迟、物理引擎适配能力。
智能交互	数字人对用户输入的理解准确性、多轮交互能力及复杂场景下的应答表现。	自然语言理解深度、意图识别准确率、多轮对话上下文保持能力、RAG 能力、幻觉率控制能力。
模型自适应	数字人模型在不同场景和需求变化下的自我调整与持续优化能力。	模型微调效率、数据兼容性、迁移学习能力、个性化训练界面质量、国产化软硬件生态适配能力。

7.2 内容表现（C）

评估数字人在视觉、听觉及行为层面的内容输出质量，重点关注其表现是否自然、生动、一致，是否具备良好的内容呈现效果和表达能力。

指标项	考察要点	关键指标参考
形象逼真度	数字人外观形象在真实感、细节表现和整体观感上的自然程度。	面部微表情丰富度、肢体动作自然度、皮肤和毛发动态效果。
语音质量	数字人语音输出在清晰度、自然度和稳定性方面的综合表现。	音色库多样性、情感语调匹配度、抗噪声能力、长语音连贯性。
思维能力	数字人在多内容类型及多交互场景下，情绪表达、人格设定与语义逻辑的一致性与稳定性。	脚本质量、情感与人格一致性、情感衰减逻辑、跨场景情感迁移能力。
动作表现	数字人在内容表达过程中，通过表情、肢体动作、节奏控制及镜头配合增强信息传达效果与用户吸引力的能力。	场景动作库覆盖度、镜头语言适配能力、互动反馈能力。
输出质量	数字人生成内容在准确性、完整性和可用性方面的整体水平。	视频分辨率/码率、直播卡顿率、动画渲染一致性。

7.3 用户体验与运维（E）

衡量数字人产品在实际使用和运维过程中的易用性、稳定性和服务支持能力，反映产品在真实应用环境中的可持续运行水平。

指标项	考察要点	关键指标参考
平台易用性	数字人平台在配置、管理和使用过程中的便捷性与友好程度。	界面交互逻辑流畅度、模板分类粒度、说明文档易用性、批量操作支持度。
响应速度	数字人在交互和服务过程中的响应时延及其在高负载条件下的表现。	冷启动时延、生成速度、复杂任务耗时。
系统稳定性	数字人在长时间运行和复杂场景下的稳定性与故障恢复能力。	崩溃率、资源占用率、并发承载能力、灾备机制健全度。
客户支持	产品提供方在技术支持、问题响应和售后服务方面的保障能力。	响应渠道多样性、问题解决时效、领域知识库质量。

指标项	考察要点	关键指标参考
集成兼容	数字人与现有业务系统、平台及接口的对接和兼容能力。	API 接口丰富度、第三方系统适配案例、低代码平台对接能力。

7.4 商业可用性（M）

评估数字人产品在实际商业应用中的落地能力和经济可行性，重点关注其成本、交付、扩展和生态适配水平。

指标项	考察要点	关键指标参考
成本结构	数字人产品在采购、部署和运行过程中的成本合理性与透明度。	计费颗粒度、按量/包年定价对比、定制化项目报价透明度。
扩展性	数字人产品在用户规模、功能能力和应用场景扩展方面的支持能力。	多角色管理容量、语言覆盖范围、终端适配类型。
交付效率	数字人产品从实施到上线交付的周期和执行效率。	标准化项目周期、定制化需求响应速度、验收标准明确性。
落地案例	数字人产品在实际行业应用中的落地数量、稳定性及可复制性。	行业案例数量、案例效果数据、客户评价。
生态兼容	数字人产品与上下游生态系统及第三方工具的协同适配能力。	硬件适配清单、云服务集成度、大模型协同能力、部署模式支持能力。

7.5 安全合规（S）

评估数字人产品在数据、内容、算法和伦理等方面的合规性与风险控制能力，是数字人产品规模化应用和长期运行的重要保障。

指标项	考察要点	关键指标参考
数据安全	数字人产品在数据采集、存储、传输和访问控制方面的安全防护能力。	加密标准、数据生命周期管理、隐私保护措施。
内容合规	数字人生成内容在法律法规和行业监管要求下的合规控制能力。	违禁内容过滤规则、AI 生成标识规范、人工审核机制、违规内容拦截与留痕能力。
权利合规	数字人产品在模型、数据、形象、声音和内容使用中的权利及授权合规情况。	素材版权来源、生成内容归属权、商用授权范围、肖像与声音授权。。
算法透明度	数字人产品在算法原理、使用边界及风险可追溯性方面的透明程度。	决策可追溯性、模型日志留存、审计接口开放。
人机伦理	数字人产品在身份标识、使用边界、人格设定、虚假代言、诱导性互动等伦理风险方面的感知、预警、处置和追溯能力。	身份冒用防护、显著身份标识、虚假代言检测、伦理风险识别准确率、实时预警与阻断能力、人工复核与追溯记录。

8 分级规范

8.1 分级原则

- 分级结果应以功能质量总分为基础，并结合安全合规约束进行判定。
- 评分过程应保证测试样本、测试场景、数据来源、专家复核和评分记录可追溯。
- 同一产品存在多个部署形态或应用版本时，应分别明确评估范围，不宜将不同版本测试结果混用。

8.2 产品功能质量等级划分

依据功能质量总分 Q，将数字人产品功能质量划分为五个等级：卓越级（L5）、先进级（L4）、成熟级（L3）、发展级（L2）、起步级（L1）。等级划分见表 2。

表 2 数字人产品功能质量等级划分

等级（标识）	等级名称	最终得分区间	等级说明
L5（2D/3D）	卓越级	[90,100]	实现逼真情绪表达与深度交互，技术、内容、体验、商业和安全能力均衡突出，可胜任媒体主持、复杂智能体等高要求任务。
L4（2D/3D）	先进级	[75,90)	具备高拟真与场景理解能力，核心功能稳定，可在多行业实现自主适配与扩展。
L3（2D/3D）	成熟级	[60,75)	多模态表现较一致，能够满足教育讲解、品牌内容、政务服务等规模化专业应用需求。
L2（2D/3D）	发展级	[40,60)	能稳定生成基本可控内容，满足客服、播报等常规场景的基础应用需求，部分能力仍需完善。
L1（2D/3D）	起步级	[0,40)	仅具备基础语音、形象或动作驱动能力，适用于简单播报和固定流程场景，功能质量存在明显提升空间。

8.3 安全合规约束

- 安全合规维度得分低于 60 分时，产品功能质量等级不应高于 L2。
- 存在违法违规、重大数据安全或个人信息保护风险、严重内容合规问题、身份冒用、虚假代言、知识产权侵权、缺失必要安全能力或其他严重伦理风险且未完成整改的，评估结果应直接判定为不通过；存在可整改问题的，可暂缓发布，整改并复测通过后方可重新评估。
- 安全合规问题整改后，可依据整改证据和复测结果重新计算安全合规维度得分并进行等级判定。

8.4 数字人产品功能质量评分表

本条给出 TCEMS 模型各指标项的评分方法。每个指标项均拆分为若干评分项，各评分项满分为 100 分，按权重加权形成该指标项得分。各评分项应依据第 6.4 条划分为区间评分（代表分法）、离散评分或布尔评分，并在评分项总表中列明评分方式、测试方法、评分规则、样本规模、统计方法、人工标注要求和证据留存要求。区间评分（代表分法）和离散评分均采用 10、30、50、70、90 五档代表分；满分 100 仅在满分触发条件满足时使用，不作为常规代表分；布尔评分采用 0 或 100 二值取分。

评分时应按照“最低满足项”原则判定：区间评分（代表分法）和离散评分同一档位涉及多项关键条件的，所有关键条件均满足时方可进入该档；若某一关键条件未满足，应按未满足条件对应

的较低档位赋予代表分。布尔评分涉及多个独立判定条件的，应拆分为布尔子指标分别评分并按子项权重汇总；单一布尔子指标满足赋 100 分，不满足赋 0 分。

表 3 评分方式说明表

评分方式	适用对象	评分形式	典型指标
区间评分（代表分法）	准确率、时延、成本等连续性能指标。	根据指标实际测量值所在区间赋予对应代表分；五档代表分为 10、30、50、70、90；100 分仅在覆盖全部规定测试项、最高档指标均满足且无重大异常时按满分触发条件使用。	跨模态转换准确率、口型同步与时延、意图识别准确率、成本水平、客户满意度等。
离散评分	模态数量、平台数量、API 数量、案例数量、语言数量等能力数量指标。	根据能力档位赋予对应代表分；五档代表分为 10、30、50、70、90；数量型指标应同时满足覆盖范围、代表性和可用性要求，100 分仅在满分触发条件满足时使用。	模态覆盖能力、API 覆盖能力、支持渠道、案例数量、语言覆盖范围、部署模式支持等。
布尔评分	是否支持、是否具备、是否提供、是否建立等功能存在性或合规存在性指标。	布尔子指标满足赋 100 分，不满足赋 0 分；包含多个独立判定条件的功能能力，应拆分为多个布尔子指标分别评分，并按子项权重汇总。	灾备恢复能力、验收标准明确性、数据生命周期管理、隐私保护能力、审核机制、审计接口等。

8.4.1 多模态生成评分表

多模态生成得分 = 模态覆盖能力得分 × 30% + 跨模态转换准确率得分 × 40% + 实时协同能力得分 × 30%

表 4 多模态生成评分项、评分方式及权重

二级指标	评分方式	权重	测试方法	评分规则
模态覆盖能力	离散评分	30%	功能测试、模态清单核验、样例验证	根据能力档位赋予对应代表分。
跨模态转换准确率	区间评分（代表分法）	40%	样本集转换测试、人工复核、准确率统计	根据指标实际测量值所在区间赋予对应代表分。
实时协同能力	区间评分（代表分法）	30%	场景联动测试、延迟测试、稳定性统计	根据指标实际测量值所在区间赋予对应代表分。

表 4-1 模态覆盖能力评分表

代表分	评价要求
10	仅支持 1 种模态，或模态能力不可验证。
30	支持 2 种模态，但仅能独立输入或输出，协同能力弱。
50	支持 2 种及以上模态，并能完成基础跨模态输入或输出。
70	支持 3 种及以上模态，具备稳定的跨模态协同生成能力。
90	支持 4 种及以上模态，可在同一任务中实时协同生成并保持一致性。

表 4-2 跨模态转换准确率评分表

得分区间	代表分	评价要求
[0,20)	10	准确率 <60%，转换结果不可稳定使用。
[20,40)	30	准确率 ≥60% 且 <70%，可完成简单转换。
[40,60)	50	准确率 ≥70% 且 <85%，满足常规场景基础要求。
[60,80)	70	准确率 ≥85% 且 <95%，复杂样本下仍较稳定。
[80,100]	90	准确率 ≥95%，多场景测试结果稳定。

表 4-3 实时协同能力评分表

得分区间	代表分	评价要求
[0,20)	10	平均延迟 >5s，或多模态协同频繁失败。
[20,40)	30	平均延迟 ≤5s，协同稳定性一般。
[40,60)	50	平均延迟 ≤3s，基础协同可用。
[60,80)	70	平均延迟 ≤2s，协同输出较稳定。
[80,100]	90	平均延迟 ≤1s，实时协同稳定，输出一致性高。

8.4.2 渲染与建模评分表

渲染与建模得分 = 渲染性能得分 × 35% + 建模细节质量得分 × 35% + 动态光影与稳定性得分 × 30%

表 5 渲染与建模评分项、评分方式及权重

二级指标	评分方式	权重	测试方法	评分规则
渲染性能	区间评分（代表分法）	35%	性能测试、分辨率/帧率/卡顿率检测	根据指标实际测量值所在区间赋予对应代表分。
建模细节质量	离散评分	35%	模型检查、视觉质量评审、细节样张核验	根据能力档位赋予对应代表分。
动态光影与稳定性	离散评分	30%	功能演示、材质光影检查、终端适配验证	根据能力档位赋予对应代表分。

表 5-1 渲染性能评分表

得分区间	代表分	评价要求
[0,20)	10	分辨率＜480p，或帧率＜15fps，或卡顿率＞5%，或测试数据无法验证。
[20,40)	30	分辨率 480p～<720p，帧率 15～<24fps，卡顿率≤5%。
[40,60)	50	分辨率 720p～<1080p，帧率 24～<30fps，卡顿率≤3%。
[60,80)	70	分辨率 1080p～<4K，帧率 30～<60fps，卡顿率≤1%，运行稳定。
[80,100]	90	分辨率≥4K，帧率≥60fps，卡顿率≤0.5%，并通过多终端渲染稳定性测试。

表 5-2 建模细节质量评分表

代表分	评价要求
10	模型轮廓粗糙，面部或肢体细节明显缺失。

代表分	评价要求
30	具备基础模型结构，但细节较少。
50	建模细节较完整，可满足常规展示。
70	建模细节完善，近景展示无明显瑕疵。
90	高精度建模，细节丰富且跨角度表现一致。

表 5-3 动态光影与稳定性评分表

代表分	评价要求
10	不支持动态光影或材质表现明显异常。
30	支持简单光影，环境变化下效果不稳定。
50	支持基础动态光影，材质表现基本自然。
70	支持动态光影和较自然材质，跨终端表现稳定。
90	支持实时全局光照或等效高阶效果，复杂环境下稳定。

8.4.3 动作驱动评分表

动作驱动得分 = 动作资源与驱动能力得分 × 30% + 口型同步与时延得分 × 40% + 语义动作匹配得分 × 30%。

表 6 动作驱动评分项、评分方式及权重

二级指标	评分方式	权重	测试方法	评分规则
动作资源与驱动能力	离散评分	30%	动作库清单核验、驱动方式测试、动作样例验证	根据能力档位赋予对应代表分；涉及数量条件时应同时核验覆盖范围、可用性和证据完整性。
口型同步与时延	区间评分（代表分法）	40%	音视频同步测试、延迟测量	根据指标实际测量值所在区间赋予对应代表分。
语义动作匹配	区间评分（代表分法）	30%	语义场景样本测试、匹配准确率统计	根据指标实际测量值所在区间赋予对应代表分。

表 6-1 动作资源与驱动能力评分表

代表分	评价要求
10	不支持语音驱动或动作库 <50 个。
30	动作库 ≥50 个，主要依赖固定动作。
50	动作库 ≥100 个，支持基础语音或语义驱动。
70	动作库 ≥300 个，驱动能力覆盖常用场景。
90	动作库 ≥500 个，支持复杂场景组合驱动。

表 6-2 口型同步与时延评分表

得分区间	代表分	评价要求

得分区间	代表分	评价要求
[0,20)	10	口型明显不同步，或延迟 >500ms。
[20,40)	30	口型同步延迟 >300ms 且 ≤500ms。
[40,60)	50	口型同步延迟 ≤300ms，基础场景可接受。
[60,80)	70	口型同步延迟 ≤150ms，同步效果较自然。
[80,100]	90	口型同步延迟 ≤80ms，长语音场景稳定自然。

表 6-3 语义动作匹配评分表

得分区间	代表分	评价要求
[0,20)	10	不支持语义动作匹配。
[20,40)	30	支持简单语义触发，准确率 <75%。
[40,60)	50	语义匹配准确率 ≥75%，可满足基础表达。
[60,80)	70	语义匹配准确率 ≥85%，动作与语义较协调。
[80,100]	90	语义匹配准确率 ≥95%，支持物理引擎或等效动作约束。

8.4.4 智能交互评分表

智能交互得分 = 意图识别准确率得分 × 35% + 多轮上下文保持得分 × 30% + 响应速度与知识增强得分 × 35%。

表 7 智能交互评分项、评分方式及权重

二级指标	评分方式	权重	测试方法	评分规则
意图识别准确率	区间评分（代表分法）	35%	问答样本测试、意图识别准确率统计	根据指标实际测量值所在区间赋予对应代表分。
多轮上下文保持	离散评分	30%	多轮对话场景测试、上下文保持轮次核验	根据能力档位赋予对应代表分。
响应速度与知识增强	区间评分（代表分法）	35%	响应时延测试、知识增强样本测试、幻觉率统计	根据指标实际测量值所在区间赋予对应代表分。

表 7-1 意图识别准确率评分表

得分区间	代表分	评价要求
[0,20)	10	准确率 <60%，无法支撑有效交互。
[20,40)	30	准确率 ≥60% 且 <75%，可识别简单意图。
[40,60)	50	准确率 ≥75% 且 <85%，满足常规交互。
[60,80)	70	准确率 ≥85% 且 <95%，复杂表达下较稳定。
[80,100]	90	准确率 ≥95%，多领域样本下表现稳定。

表 7-2 多轮上下文保持评分表

代表分	评价要求

代表分	评价要求
10	不支持多轮上下文。
30	支持不超过 2 轮上下文。
50	支持 3 轮及以上上下文，偶有遗忘。
70	支持 5 轮及以上上下文，一致性较好。
90	支持复杂上下文推理，长轮次对话仍保持一致。

表 7-3 响应速度与知识增强评分表

得分区间	代表分	评价要求
[0,20)	10	响应时间 >5s，或无知识增强与幻觉控制。
[20,40)	30	响应时间 ≤5s，知识增强能力有限。
[40,60)	50	响应时间 ≤3s，具备基础知识检索能力。
[60,80)	70	响应时间 ≤2s，知识增强结果较稳定。
[80,100]	90	响应时间 ≤1s，知识增强有效且幻觉率控制充分。

8.4.5 模型自适应评分表

模型自适应得分 = 微调效率得分 × 35% + 场景迁移能力得分 × 30% + 数据兼容与国产适配得分 × 35%。

表 8 模型自适应评分项、评分方式及权重

二级指标	评分方式	权重	测试方法	评分规则
微调效率	区间评分（代表分法）	35%	微调任务测试、训练界面核验、训练/上线周期统计	根据指标实际测量值所在区间赋予对应代表分。
场景迁移能力	离散评分	30%	迁移场景测试、场景数量核验、迁移后效果验证	根据能力档位赋予对应代表分。
数据兼容与国产适配	离散评分	35%	格式兼容测试、国产软硬件环境适配验证	根据能力档位赋予对应代表分。

表 8-1 微调效率评分表

得分区间	代表分	评价要求
[0,20)	10	不支持微调。
[20,40)	30	微调周期 >7 天。
[40,60)	50	微调周期 ≤7 天。
[60,80)	70	微调周期 ≤3 天。
[80,100]	90	微调周期 ≤1 天，并支持自动化训练。

表 8-2 场景迁移能力评分表

代表分	评价要求

代表分	评价要求
10	不支持场景迁移，或迁移功能不可用、不可验证。
30	支持 **1 个场景**迁移，仅能完成基础配置或固定脚本复用；迁移结果可验证。
50	支持 2～3 个场景迁移，核心功能和内容能够复用，迁移后效果达到基础测试要求。
70	支持 **4～5 个场景**迁移，覆盖至少 2 类业务场景，迁移流程可重复，迁移后效果稳定。
90	支持 **≥6 个行业或业务场景**快速迁移，覆盖至少 3 类典型行业或场景；迁移后功能、内容和交互效果稳定，且通过规定样本测试。

表 8-3 数据兼容与国产适配评分表

代表分	评价要求
10	数据格式单一，不能适配国产环境。
30	仅兼容少量数据格式，国产适配不足。
50	兼容 2 类及以上数据格式，具备基础适配能力。
70	兼容 3 类及以上数据格式，适配主流国产环境。
90	兼容 5 类及以上数据格式，支持国产软硬件环境部署和运行。

8.4.6 形象逼真度评分表

形象逼真度得分 = 微表情丰富度得分 × 30% + 动作自然度得分 × 35% + 皮肤与毛发动态效果得分 × 35%。

表 9 形象逼真度评分项、评分方式及权重

二级指标	评分方式	权重	测试方法	评分规则
微表情丰富度	离散评分	30%	微表情样例核验、类别数量统计、一致性抽检	根据能力档位赋予对应代表分。
动作自然度	区间评分（代表分法）	35%	动作轨迹测试、偏差统计、专家复核	根据指标实际测量值所在区间赋予对应代表分。
皮肤与毛发动态效果	离散评分	35%	渲染样张核验、材质和毛发模拟检查	根据能力档位赋予对应代表分。

表 9-1 微表情丰富度评分表

代表分	评价要求
10	无可识别微表情，或微表情功能不可用、不可验证；有效微表情种类为 0 种
30	有效微表情 **1～5 种**，一致率 **≥75%**，每种微表情均可稳定触发和复现
50	有效微表情 **6～9 种**，一致率 **≥85%**，覆盖至少两类基础情绪，测试结果可复核
70	有效微表情 **10～14 种**，一致率 **≥90%**，覆盖主要基础情绪和常见交互场景，表现自然稳定
90	有效微表情 **≥15 种**，一致率 **≥95%**，覆盖主要基础情绪、复合情绪和细微情绪变化，并通过规定样本测试

表 9-2 动作自然度评分表

得分区间	代表分	评价要求
[0,20)	10	动作僵硬，轨迹偏差 >20%。
[20,40)	30	动作轨迹偏差 ≤20%，自然度较弱。
[40,60)	50	动作轨迹偏差 ≤12%，整体可接受。
[60,80)	70	动作轨迹偏差 ≤8%，动作较自然。
[80,100]	90	动作轨迹偏差 ≤5%，近似真实动作捕捉效果。

表 9-3 皮肤与毛发动态效果评分表

代表分	评价要求
10	无动态渲染，材质明显失真。
30	渲染简单，缺乏动态细节。
50	具备基础皮肤或材质渲染。
70	支持基础皮肤动态，毛发效果较自然。
90	支持实时皮肤光照和毛发动态模拟，帧间一致性 ≥98%。

8.4.7 语音质量评分表

语音质量得分 = 音色多样性得分 × 25% + 情感匹配与自然度得分 × 35% + 抗噪与长语音连贯性得分 × 40%。

表 10 语音质量评分项、评分方式及权重

二级指标	评分方式	权重	测试方法	评分规则
音色多样性	离散评分	25%	音色清单核验、试听样本测试	根据能力档位赋予对应代表分。
情感匹配与自然度	区间评分（代表分法）	35%	情感语音样本测试、一致率统计、人工复核	根据指标实际测量值所在区间赋予对应代表分。
抗噪与长语音连贯性	区间评分（代表分法）	40%	噪声环境测试、长语音连续性测试、错误率统计	根据指标实际测量值所在区间赋予对应代表分。

表 10-1 音色多样性评分表

代表分	评价要求
10	仅支持单一音色，或音色不可区分、不可稳定调用。
30	支持 2～10 种音色，且每种音色均可正常调用。
50	支持 11～29 种音色，覆盖至少两类主要音色，且样本调用测试通过。
70	支持 30～49 种音色，覆盖性别、年龄等主要类型，且音色切换和输出稳定。
90	支持 ≥50 种音色，覆盖性别、年龄、方言等类型，完成规定多音色样本测试，音色质量和调用稳定性均达到要求。

表 10-2 情感匹配与自然度评分表

得分区间	代表分	评价要求
[0,20)	10	情感失配明显，语音机械感强。
[20,40)	30	情感匹配度 ≥75%，自然度一般。
[40,60)	50	情感匹配度 ≥85%，语音较自然。
[60,80)	70	情感匹配度 ≥90%，语调节奏较稳定。
[80,100]	90	情感识别与输出一致率 ≥95%，长文本表达自然。

表 10-3 抗噪与长语音连贯性评分表

得分区间	代表分	评价要求
[0,20)	10	抗噪能力差，长语音不连贯。
[20,40)	30	抗噪较弱，连续语音错误率 ≤10%。
[40,60)	50	信噪比 ≥20dB 时准确率 ≥85%，错误率 ≤5%。
[60,80)	70	信噪比 ≥15dB 时准确率 ≥90%，错误率 ≤3%。
[80,100]	90	信噪比 ≥10dB 时准确率 ≥95%，错误率 ≤1%。

8.4.8 思维能力评分表

思维能力得分 = 语义逻辑准确性得分 × 35% + 人格与情感一致性得分 × 35% + 跨场景迁移能力得分 × 30%

表 11 思维能力评分项、评分方式及权重

二级指标	评分方式	权重	测试方法	评分规则
语义逻辑准确性	区间评分（代表分法）	35%	脚本和专业问答样本测试、逻辑错误率和事实准确率统计	根据指标实际测量值所在区间赋予对应代表分。
人格与情感一致性	区间评分（代表分法）	35%	角色设定样本测试、一致率统计、情感变化和衰减逻辑复核	根据指标实际测量值所在区间赋予对应代表分。
跨场景迁移能力	离散评分	30%	多场景脚本测试、迁移场景数量核验	根据能力档位赋予对应代表分。

表 11-1 语义逻辑准确性评分表

得分区间	代表分	评价要求
[0,20)	10	逻辑混乱，专业内容不可用。
[20,40)	30	逻辑错误率 ≤20%，仅适合简单内容。
[40,60)	50	逻辑错误率 ≤10%，常规内容可用。
[60,80)	70	逻辑错误率 ≤5%，专业内容准确率 ≥95%。
[80,100]	90	逻辑错误率 ≤2%，专业内容准确率 ≥98%。

表 11-2 人格与情感一致性评分表

得分区间	代表分	评价要求
[0,20)	10	人格不一致，无稳定情感逻辑。
[20,40)	30	人格一致率 ≥75%，情感逻辑不稳定。
[40,60)	50	人格一致率 ≥85%，情感偶有突变。
[60,80)	70	人格一致率 ≥90%，情感衰减基本合理。
[80,100]	90	多轮对话一致率 ≥95%，情绪变化符合时间逻辑。

表 11-3 跨场景迁移能力评分表

代表分	评价要求
10	无法跨场景表达。
30	跨场景迁移能力弱，仅可复用固定脚本。
50	支持 3 类及以上场景。
70	支持 4 类及以上场景迁移。
90	支持 5 类及以上场景迁移，情感一致率 ≥90%。

8.4.9 动作表现评分表

动作表现得分 = 场景动作覆盖得分 × 35% + 镜头语言适配得分 × 30% + 互动反馈能力得分 × 35%。

表 12 动作表现评分项、评分方式及权重

二级指标	评分方式	权重	测试方法	评分规则
场景动作覆盖	离散评分	35%	动作类别清单核验、场景演示测试	根据能力档位赋予对应代表分。
镜头语言适配	离散评分	30%	镜头脚本测试、镜头类型清单核验、匹配抽检	根据能力档位赋予对应代表分。
互动反馈能力	区间评分（代表分法）	35%	互动场景测试、反馈延迟和准确率统计	根据指标实际测量值所在区间赋予对应代表分。

表 12-1 场景动作覆盖评分表

代表分	评价要求
10	无可用场景动作库，或动作不可调用、不可验证。
30	支持 1～10 类场景动作，至少覆盖基础表达动作，动作可正常调用。
50	支持 11～19 类场景动作，覆盖基础高频场景，动作调用测试通过。
70	支持 20～49 类场景动作，覆盖主要业务场景和常见交互动作，动作匹配准确、运行稳定。
90	支持 ≥50 类场景动作，覆盖高频、边界和异常场景，并通过规定样例和稳定性测试。

表 12-2 镜头语言适配评分表

代表分	评价要求

代表分	评价要求
10	无镜头语言适配能力，或镜头语言不可配置、不可调用、不可验证。
30	支持 **1 种**镜头语言，能够完成基础镜头调用。
50	支持 **2～4 种**镜头语言，镜头切换和基础匹配功能可用，匹配结果可复核。
70	支持 **5～9 种**镜头语言，覆盖主要景别、机位和运镜方式，匹配准确率 **≥90%。**
90	支持 **≥10 种**镜头语言，覆盖主要景别、机位、运镜和转场场景，匹配准确率 **≥95%**，并通过多场景样本测试。

表 12-3 互动反馈能力评分表

得分区间	代表分	评价要求
[0,20)	10	无互动反馈。
[20,40)	30	互动反馈延迟 >2s。
[40,60)	50	互动反馈延迟 ≤2s。
[60,80)	70	互动反馈延迟 ≤1s，动作匹配准确率 ≥90%。
[80,100]	90	实时响应延迟 ≤500ms，反馈准确率 ≥95%。

8.4.10 输出质量评分表

输出质量得分 = 分辨率与码率得分 × 35% + 直播稳定性得分 × 35% + 动画一致性得分 × 30%。

表 13 输出质量评分项、评分方式及权重

二级指标	评分方式	权重	测试方法	评分规则
分辨率与码率	区间评分（代表分法）	35%	输出检测、码率监测、端到端播放测试	根据指标实际测量值所在区间赋予对应代表分。
直播稳定性	区间评分（代表分法）	35%	直播压测、卡顿率和中断率统计	根据指标实际测量值所在区间赋予对应代表分。
动画一致性	区间评分（代表分法）	30%	帧一致性检测、跳帧和闪烁统计	根据指标实际测量值所在区间赋予对应代表分。

表 13-1 分辨率与码率评分表

得分区间	代表分	评价要求
[0,20)	10	分辨率＜720p，或平均码率＜3 Mbps，或码率波动率＞20%，或无法提供可验证的输出测试数据。
[20,40)	30	分辨率为 720p～<1080p，且平均码率为 3～<5 Mbps，码率波动率≤20%。
[40,60)	50	分辨率为 1080p～<4K，且平均码率为 5～<8 Mbps，码率波动率≤15%。
[60,80)	70	分辨率为 1080p～<4K，且平均码率为 8～<15 Mbps，码率波动率≤10%，输出过程稳定。
[80,100]	90	分辨率≥4K，且平均码率≥15 Mbps，码率波动率≤5%，并通过多终端播放和持续输出稳定性测试。

*分辨率、平均码率和码率波动率应分别确定对应档位，表 13-1 最终代表分取三项中的最低分。任一指标未达到对应档位时，不得进入该档，按最低满足项确定最终得分。

表 13-2 直播稳定性评分表

得分区间	代表分	评价要求
[0,20)	10	卡顿率 >5%。
[20,40)	30	卡顿率 ≤5%，但波动明显。
[40,60)	50	卡顿率 ≤3%。
[60,80)	70	卡顿率 ≤1%。
[80,100]	90	直播卡顿率 ≤0.5%（P95）。

表 13-3 动画一致性评分表

得分区间	代表分	评价要求
[0,20)	10	渲染不连续，跳帧或闪烁明显。
[20,40)	30	存在明显渲染不稳定。
[40,60)	50	动画一致性 ≥90%。
[60,80)	70	动画一致性 ≥95%。
[80,100]	90	动画帧一致性 ≥98%，无明显跳帧或闪烁。

8.4.11 平台易用性评分表

平台易用性得分 = 操作路径与成功率得分 × 35% + 模板检索与文档得分 × 35% + 批量操作能力得分 × 30%。

表 14 平台易用性评分项、评分方式及权重

二级指标	评分方式	权重	测试方法	评分规则
操作路径与成功率	区间评分（代表分法）	35%	典型任务可用性测试、操作步数和成功率统计	根据指标实际测量值所在区间赋予对应代表分。
模板检索与文档	离散评分	35%	模板库核验、检索测试、说明文档核验	根据能力档位赋予对应代表分；涉及数量条件时应同时核验覆盖范围、可用性和证据完整性。
批量操作能力	离散评分	30%	批量任务功能测试、操作类型清单核验	根据能力档位赋予对应代表分。

表 14-1 操作路径与成功率评分表

得分区间	代表分	评价要求
[0,20)	10	关键任务成功率〈80%，操作复杂。
[20,40)	30	关键操作路径 >5 步，成功率 ≥80%。
[40,60)	50	关键操作路径 ≤5 步，成功率 ≥90%。

得分区间	代表分	评价要求
[60,80)	70	关键操作路径 ≤4 步，成功率 ≥95%。
[80,100]	90	关键操作路径 ≤3 步，成功率 ≥98%。

表 14-2 模板检索与文档评分表

代表分	评价要求
10	无清晰模板体系，或模板不可检索、不可使用，说明文档缺失或不可验证。
30	模板分类 **1～10 类**，文档覆盖率 **＜80%**，关键任务自助解决率 **＜60%**，支持基本检索。
50	模板分类 **11～19 类**，文档覆盖率 **80%～＜90%**，关键任务自助解决率 **60%～＜80%**，检索结果可复核。
70	模板分类 **20～29 类**，文档覆盖率 **90%～＜100%**，关键任务自助解决率 **80%～＜90%**，支持分类、标签检索和版本更新。
90	模板分类 **≥30 类**，覆盖主要业务场景，文档覆盖率 **100%**，关键任务自助解决率 **≥90%**，支持分类、标签检索、版本更新和权限管理。

表 14-3 批量操作能力评分表

代表分	评价要求
10	不支持批量操作，或批量操作功能不可用、不可验证。。
30	支持 **1～3 类**批量操作，覆盖基础操作类型；批量任务成功率 **＜90%**，或无法提供可复核的测试结果。
50	支持 **4～5 类**批量操作，批量任务成功率 **≥90%且＜95%**，测试结果可复核。
70	支持 **6～9 类**批量操作，批量任务成功率 **≥95%且＜99%**，操作流程稳定，错误可追溯。
90	支持 **≥10 类**批量操作，覆盖主要业务操作类型，批量任务成功率 **≥99%**，并通过连续批量压力测试，无重大异常。

8.4.12 响应速度评分表

响应速度得分 = 冷启动时延得分 × 30% + 生成速度得分 × 35% + 复杂任务耗时得分 × 35%。

表 15 响应速度评分项、评分方式及权重

二级指标	评分方式	权重	测试方法	评分规则
冷启动时延	区间评分（代表分法）	30%	冷启动性能测试、平均时延统计	根据指标实际测量值所在区间赋予对应代表分。
生成速度	区间评分（代表分法）	35%	文本/音频/视频生成性能测试、吞吐统计	根据指标实际测量值所在区间赋予对应代表分。
复杂任务耗时	区间评分（代表分法）	35%	复杂任务脚本测试、完成时间统计	根据指标实际测量值所在区间赋予对应代表分。

表 15-1 冷启动时延评分表

得分区间	代表分	评价要求
[0,20)	10	冷启动时延 >5s。

得分区间	代表分	评价要求
[20,40)	30	冷启动时延 >3s 且 ≤5s。
[40,60)	50	冷启动时延 ≤3s。
[60,80)	70	冷启动时延 ≤2s。
[80,100]	90	冷启动时延 ≤1s。

表 15-2 生成速度评分表

得分区间	代表分	评价要求
[0,20)	10	生成缓慢或不稳定。
[20,40)	30	文本生成速度 ≥30 字/s。
[40,60)	50	文本生成速度 ≥50 字/s。
[60,80)	70	文本生成速度 ≥70 字/s。
[80,100]	90	文本生成速度 ≥100 字/s，或视频生成 ≤5s/段。

表 15-3 复杂任务耗时评分表

得分区间	代表分	评价要求
[0,20)	10	复杂任务不可完成或耗时 >60s。
[20,40)	30	复杂任务耗时 >30s 且 ≤60s。
[40,60)	50	复杂任务耗时 ≤30s。
[60,80)	70	复杂任务耗时 ≤20s。
[80,100]	90	复杂任务耗时 ≤10s。

8.4.13 系统稳定性评分表

系统稳定性得分 = 崩溃率与资源占用得分 × 40% + 并发承载能力得分 × 30% + 灾备恢复能力得分 × 30%。

表 16 系统稳定性评分项、评分方式及权重

二级指标	评分方式	权重	测试方法	评分规则
崩溃率与资源占用	区间评分（代表分法）	40%	稳定性测试、崩溃率、CPU和内存占用统计	根据指标实际测量值所在区间赋予对应代表分。
并发承载能力	区间评分（代表分法）	30%	并发压测、吞吐量和错误率统计	根据指标实际测量值所在区间赋予对应代表分。
灾备恢复能力	布尔评分	30%	灾备方案核验、恢复演练或测试记录核验	拆分为布尔子指标分别评分；各子指标满足赋 100 分，不满足赋 0 分，并按子项权重汇总。

表 16-1 崩溃率与资源占用评分表

得分区间	代表分	评价要求
[0, 20)	10	崩溃率＞3%，或 CPU 占用率＞95%，或内存占用率＞95%；或者任一指标无法提供可验证的测试结果
[20, 40)	30	崩溃率为 1%＜崩溃率≤3%，CPU 占用率为 85%＜CPU≤95%，内存占用率为 85%＜内存≤95%
[40, 60)	50	崩溃率为 0.5%＜崩溃率≤1%，CPU 占用率为 80%＜CPU≤85%，内存占用率为 80%＜内存≤85%
[60, 80)	70	崩溃率为 0.1%＜崩溃率≤0.5%，CPU 占用率为 70%＜CPU≤80%，内存占用率为 75%＜内存≤80%。
[80, 100]	90	崩溃率≤0.1%，CPU 占用率≤70%，内存占用率≤75%，且完成规定压力测试，无重大异常。

*崩溃率、CPU 占用率和内存占用率分别按照上述区间确定单项代表分，最终得分取三项单项代表分中的最低分。三项指标必须均有同一测试环境、同一版本和统一统计口径下的可验证数据；任一指标未达到对应档位时，按该指标对应的较低档位确定最终代表分。

表 16-2 并发承载能力评分表

得分区间	代表分	评价要求
[0, 20)	10	并发能力＜100，或无法提供可验证的并发压测结果；系统在测试中出现明显不可用、严重错误或频繁中断。
[20, 40)	30	并发能力为 **100～199**，且系统基本可用，错误率符合测试规范。
[40, 60)	50	并发能力为 **200～499**，且吞吐量、响应时间和错误率均符合测试规范。
[60, 80)	70	并发能力为 **500～999**，且压力测试稳定，无重大异常，资源使用处于可控范围
[80, 100]	90	并发能力 **≥1000**，且完成规定压力测试；系统持续运行稳定，无重大异常，吞吐量、响应时间、错误率和资源使用均达到最高档要求，并具备资源监测和容量预警能力。

*并发能力以统一测试环境下可同时维持的有效会话数或有效任务数计，不得将瞬时峰值、理论上限或未完成业务请求计入。并发数量、错误率、响应时间、吞吐量和资源使用条件须同时满足对应档位；任一条件未满足时，按最低满足项确定代表分。

表 16-3 灾备恢复能力评分表

布尔子指标	子项权重	满足条件	取值规则
备份机制	35%	具备可验证的数据、配置或服务备份机制。	满足赋 100 分；不满足赋 0 分。
故障切换机制	35%	具备可验证的主备切换、容灾切换或等效故障切换机制。	满足赋 100 分；不满足赋 0 分。
恢复演练或测试记录	30%	提供恢复演练、恢复测试或故障恢复记录。	满足赋 100 分；不满足赋 0 分。

8.4.14 客户支持评分表

客户支持得分 = 支持渠道得分 × 25% + 响应与解决时效得分 × 40% + 知识库与自助能力得分 × 35%。

表 17 客户支持评分项、评分方式及权重

二级指标	评分方式	权重	测试方法	评分规则
支持渠道	离散评分	25%	服务渠道清单核验、可用性抽检	根据能力档位赋予对应代表分。
响应与解决时效	区间评分（代表分法）	40%	工单记录核验、响应和解决时间统计	根据指标实际测量值所在区间赋予对应代表分。
知识库与自助能力	离散评分	35%	知识库内容核验、自助问答测试、覆盖抽检	根据能力档位赋予对应代表分。

表 17-1 支持渠道评分表

代表分	评价要求
10	无稳定支持渠道。
30	支持渠道 ≤2 种。
50	支持渠道 ≥3 种。
70	支持渠道 ≥4 种。
90	支持渠道 ≥5 种，覆盖在线、电话、工单等方式。

表 17-2 响应与解决时效评分表

得分区间	代表分	评价要求
[0,20)	10	关键问题响应不可控。
[20,40)	30	关键问题响应 >30 分钟。
[40,60)	50	响应 ≤30 分钟，解决 ≤1 天。
[60,80)	70	响应 ≤10 分钟，解决 ≤4 小时。
[80,100]	90	响应 ≤5 分钟，解决 ≤2 小时。

表 17-3 知识库与自助能力评分表

代表分	评价要求
10	无知识库，或知识库不可访问、不可检索、不可验证；关键问题自助解决率＜30%
30	知识库覆盖率＜80%，且关键问题自助解决率为 30%～<60%；支持基本检索功能。
50	知识库覆盖率为 80%～<90%，且关键问题自助解决率为 60%～<80%；知识条目基本可用，检索结果可复核。
70	知识库覆盖率为 90%～<95%，且关键问题自助解决率为 80%～<90%；支持分类、检索和内容更新，主要问题可通过知识库解决。
90	知识库覆盖率≥95%，且关键问题自助解决率≥90%；支持分类、标签、检索、版本更新和权限管理，并通过规定的自助问答测试。

*知识库覆盖率＝已建立有效知识条目的问题数量÷评估范围内应覆盖问题总数×100%。关键问题自助解决率＝用户通过知识库或自助问答功能独立解决的问题数量÷关键问题总数×100%。有效知识条目应内容准确、可访问、可检索，并与当前产品版本一致。

8.4.15 集成兼容评分表

集成兼容得分 = API 覆盖能力得分 × 35% + 第三方适配案例得分 × 35% + 低代码接入得分 × 30%。

表 18 集成兼容评分项、评分方式及权重

二级指标	评分方式	权重	测试方法	评分规则
API 覆盖能力	离散评分	35%	API 文档核验、接口调用测试、核心功能覆盖核验	根据能力档位赋予对应代表分；涉及数量条件时应同时核验覆盖范围、可用性和证据完整性。
第三方适配案例	离散评分	35%	集成案例材料核验、接口联调记录抽检	根据能力档位赋予对应代表分。
低代码接入	离散评分	30%	低代码平台接入测试、组件和模板清单核验	根据能力档位赋予对应代表分；涉及数量条件时应同时核验覆盖范围、可用性和证据完整性。

表 18-1 API 覆盖能力评分表

代表分	评价要求
10	无标准 API 或接口不可调用、无文档、无法提供验证证据。
30	API 数量 1～10 个，具备基本接口文档，基础接口可调用。
50	API 数量 11～19 个，覆盖部分核心功能，文档基本完整，抽检接口调用通过。
70	API 数量 20～49 个，覆盖主要核心功能，接口文档完整，核心接口调用验证通过。
90	API 数量 ≥50 个，覆盖生成、管理、调度等核心功能，接口文档完整，核心接口调用全部验证通过，并具备版本管理和权限控制能力。

表 18-2 第三方适配案例评分表

代表分	评价要求
10	无第三方适配案例，或案例无法提供合同、上线证明、联调记录等可核验证据。
30	适配案例 1～3 个，完成基本联调，至少有一项可核验证据。
50	适配案例 4～9 个，覆盖至少 2 类业务系统，案例材料基本完整，联调结果可复核。
70	适配案例 10～19 个，覆盖主要业务系统类别，具备标准化适配流程，案例运行稳定。
90	适配案例 ≥20 个，覆盖主流业务系统和不少于 3 类典型行业或场景，提供完整上线证明、联调记录和运行效果数据。

*案例数量按已完成部署或上线、能够提供可核验证据且通过抽检的独立适配项目计数。同一客户、同一系统的重复部署不得重复计为多个案例。

表 18-3 低代码接入评分表

代表分	评价要求
10	不支持低代码平台接入，或无法完成基本接入验证。
30	支持 1 类低代码平台，但需要较多定制开发；接入成功率 ＜80%，或只能完成展示级接入。
50	支持 2 类低代码平台，可完成标准组件接入；接入成功率 ≥80%且＜90%，仍需少量开发配置。
70	支持 3～4 类主流低代码平台，可完成主要业务流程接入；接入成功率 ≥90%且＜95%，配置流程标准化。
90	支持 ≥5 类主流低代码平台，可完成完整业务流程接入；接入成功率 ≥95%，主要场景无需开发即可接入，并通过多平台兼容性测试。

8.4.16 成本结构评分表

成本结构得分 = 报价透明度得分 × 40% + 成本水平得分 × 35% + 计费颗粒度得分 × 25%。

表 19 成本结构评分项、评分方式及权重

二级指标	评分方式	权重	测试方法	评分规则
报价透明度	布尔评分	40%	报价文件、合同条款和费用说明核验	拆分为布尔子指标分别评分；各子指标满足赋 100 分，不满足赋 0 分，并按子项权重汇总。
成本水平	区间评分（代表分法）	35%	成本测算、行业均值对比、报价样本核验	根据指标实际测量值所在区间赋予对应代表分。
计费颗粒度	离散评分	25%	计费规则核验、账单样例核验	根据能力档位赋予对应代表分。

表 19-1 报价透明度评分表

布尔子指标	子项权重	满足条件	取值规则
报价组成完整	40%	提供采购、部署、运行、增值服务等费用组成说明。	满足赋 100 分；不满足赋 0 分。
计费规则明确	35%	提供计费口径、计费周期、计费单位和触发条件说明。	满足赋 100 分；不满足赋 0 分。
可变费用可预测	25%	说明可变费用、超额费用或价格调整条件，支持费用预测。	满足赋 100 分；不满足赋 0 分。

表 19-2 成本水平评分表

得分区间	代表分	评价要求
[0,20)	10	成本明显高于行业水平且原因不清。
[20,40)	30	费用波动 >20%。
[40,60)	50	成本略高，费用波动 ≤20%。
[60,80)	70	成本接近行业均值（±10%）。
[80,100]	90	单位成本低于行业均值 ≥20%。

表 19-3 计费颗粒度评分表

代表分	评价要求
10	无明确计费规则。
30	仅支持单一粗颗粒计费。
50	支持单一计费模式，颗粒度基本明确。
70	支持按量和包年等计费模式。
90	支持按调用、时长、并发等细颗粒度计费。

8.4.17 扩展性评分表

扩展性得分 = 角色管理容量得分 × 35% + 语言覆盖范围得分 × 30% + 终端适配类型得分 × 35%。

表 20 扩展性评分项、评分方式及权重

二级指标	评分方式	权重	测试方法	评分规则
角色管理容量	离散评分	35%	角色管理功能测试、容量上限核验	根据能力档位赋予对应代表分。
语言覆盖范围	离散评分	30%	语言清单核验、多语言样本测试	根据能力档位赋予对应代表分；涉及数量条件时应同时核验覆盖范围、可用性和证据完整性。
终端适配类型	离散评分	35%	终端兼容测试、适配清单核验	根据能力档位赋予对应代表分。

表 20-1 角色管理容量评分表

代表分	评价要求
10	多角色管理容量〈50。
30	50≤ 多角色管理容量 ≤100。
50	101≤多角色管理容量 ≤200。
70	201≤ 多角色管理容量 ≤500
90	501≤ 多角色管理容量

表 20-2 语言覆盖范围评分表

代表分	评价要求
10	仅支持 1 种语言，或语言功能不可稳定使用。
30	支持 **2～5 种语言**，每种语言均能完成规定的识别、生成和交互基础测试。
50	支持 **6～10 种语言**，覆盖至少 2 类主要语系或目标市场语言，且样本测试通过。
70	支持 **11～19 种语言**，覆盖主要目标市场语言，识别、生成和交互质量达到测试要求。
90	支持 **≥20 种主流语言**，覆盖评估范围内全部目标市场主要语种，完成规定多语言样本测试，且识别、生成、交互和切换功能稳定。

*语言数量按可独立完成规定识别、生成和交互测试的语言种类统计。同一语言的方言、音色或语音包不单独计为一种语言；仅支持单向识别或单向生成的，不计入完整语言覆盖数量。各档位须同时满足语言数量、目标语种覆盖和测试稳定性要求，未满足任一条件时按较低档计分。

表 20-3 终端适配类型评分表

代表分	评价要求
10	不支持终端适配，或终端功能不可用、不可验证。
30	支持 1～2 类终端，基本功能可用；

代表分	评价要求
50	支持 3 类终端，兼容性测试通过；
70	支持 4 类终端，兼容性和稳定性满足测试要求；
90	支持 ≥5 类终端，覆盖主要使用终端，完成兼容性和稳定性验证，并支持规模化场景扩展。

8.4.18 交付效率评分表

交付效率得分 = 标准项目周期得分 × 35% + 定制需求响应得分 × 30% + 验收标准明确性得分 × 35%。

表 21 交付效率评分项、评分方式及权重

二级指标	评分方式	权重	测试方法	评分规则
标准项目周期	区间评分（代表分法）	35%	项目记录核验、上线周期统计	根据指标实际测量值所在区间赋予对应代表分。
定制需求响应	区间评分（代表分法）	30%	需求响应记录核验、响应时间统计	根据指标实际测量值所在区间赋予对应代表分。
验收标准明确性	布尔评分	35%	验收方案、验收记录模板和合同条款核验	拆分为布尔子指标分别评分；各子指标满足赋 100 分，不满足赋 0 分，并按子项权重汇总。

表 21-1 标准项目周期评分表

得分区间	代表分	评价要求
[0,20)	10	无明确交付周期。
[20,40)	30	项目周期 >1 个月。
[40,60)	50	标准项目周期 ≤1 个月。
[60,80)	70	标准项目周期 ≤2 周。
[80,100]	90	标准项目周期 ≤1 周。

表 21-2 定制需求响应评分表

得分区间	代表分	评价要求
[0,20)	10	响应不稳定或不可控。
[20,40)	30	定制需求响应 >5 天。
[40,60)	50	定制需求响应 ≤5 天。
[60,80)	70	定制需求响应 ≤3 天。
[80,100]	90	定制需求响应 ≤2 天。

表 21-3 验收标准明确性评分表

布尔子指标	子项权重	满足条件	取值规则
验收标准明确	40%	形成明确的功能、性能、合规或交付验收标准。	满足赋 100 分；不满足赋 0 分。
验收标准可执行、可	35%	验收标准具有可执行步骤、测试口径和复测条件。	满足赋 100 分；不满

布尔子指标	子项权重	满足条件	取值规则
复测			足赋 0 分。
验收记录模板完整	25%	提供验收记录模板或等效验收证据留存机制。	满足赋 100 分；不满足赋 0 分。

8.4.19 落地案例评分表

落地案例得分 = 案例数量得分 × 30% + 效果数据完备性得分 × 40% + 客户满意度得分 × 30%

表 22 落地案例评分项、评分方式及权重

二级指标	评分方式	权重	测试方法	评分规则
案例数量	离散评分	30%	案例材料核验、合同/上线证明抽检	根据能力档位赋予对应代表分；涉及数量条件时应同时核验覆盖范围、可用性和证据完整性。
效果数据完备性	离散评分	40%	效果数据材料核验、指标覆盖清单检查	根据能力档位赋予对应代表分。
客户满意度	区间评分（代表分法）	30%	客户访谈、问卷或满意度数据核验	根据指标实际测量值所在区间赋予对应代表分。

表 22-1 案例数量评分表

代表分	评价要求
10	无可核验的实际落地案例，或案例无法提供合同、上线证明、验收记录等有效证据。
30	可核验实际案例 **1～3 个**，至少完成部署或上线，具备基本案例材料。
50	可核验实际案例 **4～9 个**，覆盖至少 2 类业务系统或应用场景，案例材料基本完整。
70	可核验实际案例 **10～19 个**，覆盖主要业务系统类别或行业场景，具备标准化交付和适配记录，运行情况可复核。
90	可核验实际案例 **≥20 个**，覆盖不少于 3 类典型行业或场景，且每个案例均有上线证明、验收记录或可核验效果数据。

表 22-2 效果数据完备性评分表

代表分	评价要求
10	无效果数据。
30	效果数据缺失或不可核验。
50	有基础效果数据。
70	80% 及以上案例有量化效果数据。
90	效果数据完备，覆盖 ROI、转化率或效率提升等指标。

表 22-3 客户满意度评分表

得分区间	代表分	评价要求
[0,20)	10	无可核验客户评价，或有效评价样本不足、评价来源不明、数据无法复核。
[20,40)	30	客户满意度为 **70%～<85%**，有效评价样本不少于规定数量，评价结果可核验。

得分区间	代表分	评价要求
[40,60)	50	客户满意度为 **85%～<90%**，有效评价样本和评价材料完整，覆盖主要使用方。
[60,80)	70	客户满意度为 **90%～<95%**，评价结果稳定，覆盖不同类型使用方，且投诉或重大负面反馈处于可接受范围。
[80,100]	90	客户满意度 **≥95%**，有效评价样本充足，覆盖主要使用方和典型应用场景，并有可核验的复购、续约或持续使用证据。

*客户满意度＝评价为“满意”或以上的有效评价数量÷有效评价总数量×100%。有效评价应来自实际使用产品的客户或最终用户，并能够核验评价主体、评价时间、使用场景和评价内容。无效、重复、明显异常或无法确认来源的评价不得计入。

*有效评价样本原则上不少于 30 份；实际客户数量少于 30 个时，应进行全量调查。调查结果应保留原始问卷、访谈记录、统计过程和客户确认材料。

8.4.20 生态兼容评分表

生态兼容得分 = 部署模式支持得分 × 35% + 云服务与模型协同得分 × 30% + 国产软硬件适配得分 × 35%

表 23 生态兼容评分项、评分方式及权重

二级指标	评分方式	权重	测试方法	评分规则
部署模式支持	离散评分	35%	部署模式清单核验、部署演示或记录核验	根据能力档位赋予对应代表分；涉及数量条件时应同时核验覆盖范围、可用性和证据完整性。
云服务与模型协同	离散评分	30%	云服务集成测试、模型协同场景验证	根据能力档位赋予对应代表分。
国产软硬件适配	离散评分	35%	国产 CPU/GPU、操作系统和中间件适配验证	根据能力档位赋予对应代表分。

表 23-1 部署模式支持评分表

代表分	评价要求
10	仅支持固定部署方式。
30	部署复杂，周期 >2 周。
50	支持单一部署模式，部署周期 ≤2 周。
70	支持 2 种及以上部署模式，私有化部署 ≤1 周。
90	同时支持私有化、公有云 SaaS、混合云，轻量部署 ≤1 天，并提供对应部署记录或演示验证。

表 23-2 云服务与模型协同评分表

代表分	评价要求
10	无云服务集成，不支持模型协同。
30	云集成或模型协同能力有限。
50	支持基础云服务集成或单模型协同。
70	支持多云或多模型协同，接口稳定。

代表分	评价要求
90	支持主流云服务和多模型协同，调度策略可配置。

表 23-3 国产软硬件适配评分表

代表分	评价要求
10	不支持国产软硬件环境。
30	仅少量国产环境可运行，适配不完整。
50	兼容部分主流国产环境。
70	兼容主流国产环境，适配清单较完整。
90	兼容 3 类及以上国产 GPU/CPU 平台，适配清单完整。

8.4.21 数据安全评分表

数据安全得分 = 加密覆盖能力得分 × 35% + 数据生命周期管理得分 × 35% + 隐私保护能力得分 × 30%

表 24 数据安全评分项、评分方式及权重

二级指标	评分方式	权重	测试方法	评分规则
加密覆盖能力	区间评分（代表分法）	35%	传输和存储加密配置核验、覆盖率统计	根据指标实际测量值所在区间赋予对应代表分。
数据生命周期管理	布尔评分	35%	制度文件、系统配置、日志和流程记录核验	拆分为布尔子指标分别评分；各子指标满足赋 100 分，不满足赋 0 分，并按子项权重汇总。
隐私保护能力	布尔评分	30%	脱敏、匿名化、敏感数据识别机制核验和功能测试	拆分为布尔子指标分别评分；各子指标满足赋 100 分，不满足赋 0 分，并按子项权重汇总。

表 24-1 加密覆盖能力评分表

得分区间	代表分	评价要求
[0,20)	10	加密覆盖率 <70% 或无统一加密策略。
[20,40)	30	加密覆盖率 ≥70%。
[40,60)	50	加密覆盖率 ≥85%。
[60,80)	70	加密覆盖率 ≥95%。
[80,100]	90	采用国密或同等强度算法，加密覆盖率 100%。

表 24-2 数据生命周期管理评分表

布尔子指标	子项权重	满足条件	取值规则
采集与存储管理	25%	建立数据采集和存储环节的管理规则或系统控制。	满足赋 100 分；不满足赋 0 分。
使用与访问管理	25%	建立数据使用、访问控制和权限管理机制。	满足赋 100 分；不满

布尔子指标	子项权重	满足条件	取值规则
			足赋 0 分。
归档与销毁管理	25%	建立数据归档、留存期限和销毁机制。	满足赋 100 分；不满足赋 0 分。
流程证据留存	25%	能够提供制度、系统配置、日志或流程记录作为证据。	满足赋 100 分；不满足赋 0 分。

表 24-3 隐私保护能力评分表

布尔子指标	子项权重	满足条件	取值规则
数据脱敏机制	35%	具备可验证的数据脱敏机制。	满足赋 100 分；不满足赋 0 分。
匿名化机制	30%	具备可验证的数据匿名化或去标识化机制。	满足赋 100 分；不满足赋 0 分。
敏感数据识别机制	35%	具备敏感数据识别、提示或拦截机制。	满足赋 100 分；不满足赋 0 分。

8.4.22 内容合规评分表

内容合规得分 = 风险识别覆盖得分 × 35% + 标识与拦截得分 × 35% + 审核机制得分 × 30%

表 25 内容合规评分项、评分方式及权重

二级指标	评分方式	权重	测试方法	评分规则
风险识别覆盖	区间评分（代表分法）	35%	风险样本集测试、风险类型覆盖和识别准确率统计	根据指标实际测量值所在区间赋予对应代表分。
标识与拦截	区间评分（代表分法）	35%	AI 标识覆盖测试、实时拦截率统计	根据指标实际测量值所在区间赋予对应代表分。
审核机制	布尔评分	30%	审核流程、人工介入规则、审核记录核验	拆分为布尔子指标分别评分；各子指标满足赋 100 分，不满足赋 0 分，并按子项权重汇总。

表 25-1 风险识别覆盖评分表

得分区间	代表分	评价要求
[0,20)	10	无有效风险识别或过滤机制，或测试数据无法验证风险识别能力。
[20,40)	30	风险识别覆盖 **1～3 类**，识别准确率 **≥80%**，且测试结果可复核。
[40,60)	50	风险识别覆盖 **4～5 类**，识别准确率 **≥90%**，覆盖至少 2 类主要风险，测试结果可复核。
[60,80)	70	风险识别覆盖 **6～9 类**，识别准确率 **≥95%**，覆盖主要风险类别，误报和漏报处于可接受范围。
[80,100]	90	风险识别覆盖 **≥10 类**，识别准确率 **≥98%**，覆盖评估范围内全部规定风险类型，且具备风险分级、拦截或处置联动能力。

表 25-2 标识与拦截评分表

得分区间	代表分	评价要求
[0,20)	10	无 AI 生成内容标识或拦截机制。
[20,40)	30	标识不完整，拦截能力不稳定。
[40,60)	50	标识覆盖率 ≥85%，具备基础拦截。
[60,80)	70	标识覆盖率 ≥95%，拦截率 ≥97%。
[80,100]	90	标识覆盖率 100%，实时拦截率 ≥99%。

表 25-3 审核机制评分表

布尔子指标	子项权重	满足条件	取值规则
审核流程建立	30%	建立人工审核或人机协同审核流程。	满足赋 100 分；不满足赋 0 分。
触发条件与责任规则明确	25%	明确审核触发条件、责任角色和处置边界。	满足赋 100 分；不满足赋 0 分。
响应时限明确	20%	明确审核响应时限或处理时限。	满足赋 100 分；不满足赋 0 分。
审核记录留存	25%	审核过程、结果和处置记录可留存、可追溯。	满足赋 100 分；不满足赋 0 分。

8.4.23 知识产权评分表

知识产权得分 = 素材数据可追溯得分 × 35% + 权属规则明确性得分 × 35% + 商用授权覆盖得分 × 30%

表 26 知识产权评分项、评分方式及权重

二级指标	评分方式	权重	测试方法	评分规则
素材数据可追溯	区间评分（代表分法）	35%	训练数据、素材、形象来源清单和追溯记录核验	根据指标实际测量值所在区间赋予对应代表分。
权属规则明确性	布尔评分	35%	合同条款、权属规则和系统配置核验	拆分为布尔子指标分别评分；各子指标满足赋 100 分，不满足赋 0 分，并按子项权重汇总。
商用授权覆盖	区间评分（代表分法）	30%	授权文件、合同和应用范围核验	根据指标实际测量值所在区间赋予对应代表分。

表 26-1 素材数据可追溯评分表

得分区间	代表分	评价要求
[0,20)	10	来源不可追溯，存在明显侵权风险。
[20,40)	30	可追溯率 <85%。
[40,60)	50	可追溯率 ≥85%。
[60,80)	70	可追溯率 ≥95%。
[80,100]	90	训练数据、素材和形象来源可追溯率 100%，无侵权记录。

表 26-2 权属规则明确性评分表

布尔子指标	子项权重	满足条件	取值规则
生成内容归属明确	40%	明确数字人生成内容的归属规则。	满足赋 100 分；不满足赋 0 分。
素材、形象和声音权利边界明确	35%	明确训练素材、数字人形象、声音等权利边界。	满足赋 100 分；不满足赋 0 分。
合同或系统配置可核验	25%	通过合同条款、授权文件或系统配置提供可核验证据。	满足赋 100 分；不满足赋 0 分。

表 26-3 商用授权覆盖评分表

得分区间	代表分	评价要求
[0,20)	10	未取得必要商用授权，或者存在授权缺失、过期、主体不一致等问题
[20,40)	30	商用授权覆盖率 <80%，且缺失项不涉及安全合规底线；授权范围、期限或使用渠道存在明显限制
[40,60)	50	商用授权覆盖率 80%～<90%；主要授权文件可核验，但部分使用场景、媒介、地域或期限未覆盖<90%，主要授权文件可核验，但部分使用场景、媒介、地域或期限未覆盖
[60,80)	70	商用授权覆盖率 90%～<100%；核心对象和主要商业场景均有有效授权，仅存在非核心缺口
[80,100]	90	商用授权覆盖率 =100%；授权主体、对象、用途、期限、地域、媒介和渠道均覆盖本次评估范围，授权文件完整、有效且可追溯，不存在超范围使用。

8.4.24 算法透明度评分表

算法透明度得分 = 可解释能力得分 × 35% + 日志留存得分 × 35% + 审计接口得分 × 30%

表 27 算法透明度评分项、评分方式及权重

二级指标	评分方式	权重	测试方法	评分规则
可解释能力	区间评分（代表分法）	35%	决策路径样本核验、解释结果抽检	根据指标实际测量值所在区间赋予对应代表分。
日志留存	离散评分	35%	日志范围、留存时长、查询能力和导出记录核验	根据能力档位赋予对应代表分。
审计接口	布尔评分	30%	审计接口调用测试、审计数据导出核验	拆分为布尔子指标分别评分；各子指标满足赋 100 分，不满足赋 0 分，并按子项权重汇总。

表 27-1 可解释能力评分表

得分区间	代表分	评价要求
[0,20)	10	关键算法不可解释。
[20,40)	30	可解释率 <80%。
[40,60)	50	可解释率 ≥80%。
[60,80)	70	可解释率 ≥90%。

得分区间	代表分	评价要求
[80,100]	90	关键决策路径可解释率 ≥95%。

表 27-2 日志留存评分表

代表分	评价要求
10	无日志留存，或日志无法访问、无法验证，或无法确认日志内容和留存周期。
30	日志留存时间 **＜30 天**，支持基本日志查看；不要求导出和按请求追溯。
50	日志留存时间 **30～＜90 天**，支持日志查询和导出，日志内容基本完整。
70	日志留存时间 **90～＜180 天**，支持按条件查询和导出，覆盖主要业务操作，并可关联用户、时间、操作和结果信息。
90	全量日志留存时间 **≥180 天**，支持按请求追溯、条件查询和导出，日志字段完整，能够关联请求主体、时间、操作对象、操作内容、处理结果和异常信息，并通过审计复核测试。

*日志留存时长按评估范围内全量日志中实际可查询日志的连续留存时间计算，不按单个日志文件或部分日志的最长保存时间计算。全量日志是指评估范围内规定业务操作、接口调用、权限变更、异常事件和安全事件等日志均纳入留存范围。

表 27-3 审计接口评分表

布尔子指标	子项权重	满足条件	取值规则
审计接口或等效导出能力	60%	提供可用审计接口，或可导出等效审计数据。	满足赋 100 分；不满足赋 0 分。
审计数据字段完整	20%	审计数据覆盖时间、主体、操作、对象、结果等必要字段。	满足赋 100 分；不满足赋 0 分。
审计可验证	20%	支持内部或第三方审计核验。	满足赋 100 分；不满足赋 0 分。

8.4.25 人机伦理评分表

人机伦理得分 = 身份防护与标识得分 × 35% + 虚假代言检测得分 × 35% + 伦理风险预警闭环得分 × 30%。

表 28 人机伦理评分项、评分方式及权重

二级指标	评分方式	权重	测试方法	评分规则
身份防护与标识	布尔评分	35%	身份标识展示、冒用防护机制和高风险提示测试	拆分为布尔子指标分别评分；各子指标满足赋 100 分，不满足赋 0 分，并按子项权重汇总。
虚假代言检测	区间评分（代表分法）	35%	公众人物、品牌和代言样本检测测试、覆盖率统计	根据指标实际测量值所在区间赋予对应代表分。
伦理风险预警闭环	布尔评分	30%	预警、阻断、人工复核、追溯记录和处置闭环核验	拆分为布尔子指标分别评分；各子指标满足赋 100 分，不满足赋 0 分，并按子项权重汇总。

表 28-1 身份防护与标识评分表

布尔子指标	子项权重	满足条件	取值规则
显著身份标识	35%	在必要场景中提供清晰、显著的数字人身份标识。	满足赋 100 分；不满足赋 0 分。
身份冒用防护	35%	具备身份冒用检测、防护或拦截机制。	满足赋 100 分；不满足赋 0 分。
高风险场景提示	30%	在金融、医疗、政务、代言等高风险场景中提供必要提示。	满足赋 100 分；不满足赋 0 分。

表 28-2 虚假代言检测评分表

得分区间	代表分	评价要求
[0,20)	10	无虚假代言检测能力，或无法提供可验证的检测结果。
[20,40)	30	检测覆盖率 **＜80%**，或检测准确率 **＜80%**；仅能识别少量固定样本。
[40,60)	50	检测覆盖率 **80%～＜90%**，且检测准确率 **≥85%**；覆盖公众人物、品牌或代言素材中的至少一类。
[60,80)	70	检测覆盖率 **90%～＜95%**，且检测准确率 **≥90%**、漏报率 **≤10%**；覆盖公众人物、品牌和主要敏感场景。
[80,100]	90	检测覆盖率 **≥95%**，且检测准确率 **≥95%**、漏报率 **≤5%**、误报率 **≤5%**；覆盖评估范围内全部规定的公众人物、品牌及敏感场景，并具备告警、拦截或人工复核联动能力。

表 28-3 伦理风险预警闭环评分表

布尔子指标	子项权重	满足条件	取值规则
风险感知与预警	30%	具备伦理风险感知、识别和预警能力。	满足赋 100 分；不满足赋 0 分。
阻断与处置	25%	具备高风险内容或行为的阻断、暂停或处置机制。	满足赋 100 分；不满足赋 0 分。
人工复核	20%	具备人工复核入口、流程或责任机制。	满足赋 100 分；不满足赋 0 分。
追溯闭环与记录	25%	伦理风险处置过程可追溯，并形成闭环记录。	满足赋 100 分；不满足赋 0 分。

8.5 场景权重调整与总分计算示例

8.5.1 示例说明

本条给出场景权重调整和总分计算示例，用于帮助标准使用者理解各评分表、维度得分与总体评价之间的关系。

8.5.2 场景权重调整示例

以政务服务场景为例，假设对用户体验与安全合规要求较高，可设置场景修正系数如表 29。

表 29 政务服务场景权重调整示例

维度	维度名称	基础权重	场景修正系数	归一化后场景权重
T	技术实力	25%	1.0	25.25%
C	内容表现	20%	0.9	18.18%
E	用户体验与运维	20%	1.1	22.22%
M	商业可用性	15%	0.8	12.12%
S	安全合规	20%	1.1	22.22%

8.5.3 总分计算示例

表 30 总分计算示例

维度	维度得分 Di	场景权重 Wi	加权得分 Di×Wi
T	82	25.25%	20.71
C	78	18.18%	14.18
E	84	22.22%	18.66
M	72	12.12%	8.73
S	88	22.22%	19.55
合计	-	99.99%（四舍五入）	81.83

根据表 30，功能质量总分 Q = 81.83，且安全合规维度得分为 88 分，未触发等级限制，因此该产品在政务服务场景下判定为先进级（L4）。

参考文献

[1] GB/T 46483—2025 信息技术 客服型虚拟数字人通用技术要求

[2] ITU-T F.748.15 Framework and Metrics for Digital Human Application Systems

附录 A（资料性）

数字人产品功能质量评估记录表（示例）

本附录仅用于评估实施参考，不影响正文评价规则和等级判定逻辑。

表 A.1 数字人产品功能质量评估记录表（示例）

项目	填写内容	填写说明
评估机构		填写承担评估工作的机构全称。
产品名称		填写数字人产品全称。
产品版本		填写版本号、发布日期或构建号。
评估范围		填写纳入评估的产品形态、功能模块和应用场景。
测试环境		填写软硬件环境、网络环境、接口账号和版本信息。
测试样本信息		填写样本来源、数量、类别分布和标注方式。
评估人员		填写测试人员、评分人员和专家复核人员。
测试时间		填写起止时间和关键测试日期。
评分记录		填写各指标项得分、评分方式和主要依据。
证据编号		填写原始记录、日志、截图、视频等证据编号。
专家复核意见		填写专家复核结论、修改意见和确认结果。
最终等级判定		填写最终等级或不通过、暂缓发布结论。
签字确认		填写评估、复核及审批签字。