《文博大模型应用评测指南》
团体标准编制说明

一、 编制背景与必要性

当前，人工智能技术，特别是大语言模型、多模态大模型的快速发展，正在为文化遗产保护、研究、展示与传播带来前所未有的变革机遇。大模型在智能导览、文物识别、知识问答、虚拟修复、内容生成等文博场景的应用潜力巨大，是落实国家文化数字化战略、推动“互联网+中华文明”行动的关键技术手段。

然而，文博大模型作为新兴领域专用技术，其应用落地仍面临一系列挑战，亟需建立统一、科学的评测标准进行规范和引导：

1. 专业性准确度面临严峻考验：文博领域知识体系复杂精深，涵盖文物学、考古学、历史学、博物馆学等多个学科，涉及大量专业术语、年代断代、工艺技法、文化符号等独特知识。通用大模型在此类专业知识上的理解往往存在事实性错误、表述模糊或文化内涵误读，若不加约束地应用，可能导致错误知识的传播，损害文化遗产研究的严谨性与权威性。

2. 多模态理解与生成能力缺乏基准：文博数据天然是多模态的，包括高清文物图像、三维模型、考古线图、古籍文献、口述音频等。大模型需要具备图文关联、文物图像深度解读、跨模态检索与生成等能力，以支撑沉浸式展览、智能导览等高级应用。目前业界缺乏对这些多模态专业能力的系统评测标准，难以评估不同模型的实际效果。

3. 数据合规与价值导向风险凸显：文博大模型的训练与应用涉及大量受版权保护的文物图像、学术成果，以及可能包含个人隐私的口述史、社群敏感文化信息。模型输出内容必须严格遵循《著作权法》《数据安全法》等法律法规，并符合正确的历史观、文化观和国家文化安全要求。缺乏标准化的安全与价值对齐评测，极易引发法律风险与文化误读。

4. 行业选型与应用评估无据可依：面对市场上涌现的各类“文博大模型”，博物馆、考古机构、技术供应商等应用方在模型选型、能力验证和效果评估时缺乏客观、统一的“度量衡”，导致研发方向不明确、产品良莠不齐、行业监管困难，制约了技术成果的有效转化与规模化、高质量应用。

为此，积极响应《“十四五”文物保护和科技创新规划》及国家文化数字化战略的要求，编制《文博大模型应用评测指南》团体标准，旨在为文博领域提供一套覆盖技术能力、专业认知、安全合规与应用价值的综合性评测规范。本标准的制定，将填补该领域应用评测标准的空白，为引导文博大模型技术健康发展、保障应用安全有效、释放人工智能赋能文化遗产事业的巨大潜力提供至关重要的标准化支撑。

二、 编制原则与依据

本标准编制遵循以下核心原则，并充分依据和参考了国际国内相关标准、行业最佳实践及前沿研究成果：

1. 问题导向与场景牵引原则：评测内容的设置紧密围绕文博领域最核心、最迫切的业务场景（如智能导览、文物保护修复、研究辅助、文创开发）和关键能力缺口（如专业准确性、多模态理解）。

2. “四维一体”系统性原则：构建了 “通用基础能力—文博安全与价值对齐—文博专业认知能力—文博场景应

用能力” 四个核心维度紧密衔接的评测体系。其中，通用基础能力是技术基座，安全价值对齐是不可逾越的底线，专业认知能力是领域适配性的核心，场景应用能力是最终价值体现。

3. 底线思维与风险防控优先原则：将 “文博安全与价值对齐” 维度置于突出地位，并设置了“一票否决”机制（该维度不达标则模型直接被判定为禁用级）。这体现了对文化遗产数据安全、文化主权和正确价值观的极端重视，借鉴了金融、医疗等领域垂直大模型评测中将合规安全作为首要考量的成熟经验。

4. 专业深度与可操作性结合原则：评测指标既强调对文博深度专业知识（如工艺技法辨识、文化符号解读）的考核，也注重通过清晰的评测方法（自动化、人工、专家评审结合）和打分规则使之可落地、可执行。

5. 继承发展与开放协同原则：积极融入国际文化遗产信息管理通用框架，倡导模型应具备基于 CIDOC 概念参考模型（CRM，对应国标 GB/T 37965-2019） 等语义标准构建的知识图谱的关联与推理能力。同时，标准本身与 T/SAIAS 015—2024《语料库建设导则》 及正在编制的《文博语料库建设导则》形成协同，前者规范高质量“数据燃料”的生产，本标准则规范基于这些燃料训练的“引擎”的性能测试。

主要编制依据包括：

- 国际标准与最佳实践：ISO 21127:2014（CIDOC CRM），欧洲数字图书馆 Europeana 的 LIDO 数据交换标准，盖蒂研究所 AAT（艺术与建筑叙词表）等权威受控词表，为专业认知评测提供了知识组织基础。
- 国内政策与行业标准：《国家文化数字化战略》、《文

物安全防护工程》，以及文物藏品信息登录、数字化采集等相关行业标准。

• 成熟领域评测标准借鉴：重点参考了 T/SAIAS 019—2024《金融大模型应用评测指南》 和《医疗大模型应用评测指南》的成熟架构。这些标准在构建“基础能力-安全合规-专业认知-场景应用”的多维体系，以及采用“自动化+人工+专家评审”三位一体评测方法、设置关键维度否决项等方面，为本标准提供了极佳的方法论范本。

• 前沿研究与实践成果：参考了知识库中关于“大模型在文博场景评测指标”（如年代判断、材质工艺描述、文化符号解读准确度等具体指标）的研究，以及故宫博物院构建“中国古代可移动文物概念参考模型（CRM-ACA）”并应用于“数字文物库”的领先实践。此外，大量关于文物知识图谱构建（如基于 BERT-BiLSTM-CRF、胶囊网络等技术的实体关系抽取）的研究，为模型专业认知能力中的“知识关联与推理”评测提供了技术依据和验证场景。

三、 标准主要内容及技术要点说明

本标准共分为 7 章，核心内容围绕四大评测维度展开，结构严谨，逻辑清晰。其中第 5 章、第 6 章、第 7 章为核心技术部分。

1. 标准第 5 章“测评框架”确立了标准的顶层设计。

• 提出了 “通用能力支撑、专业能力核心、安全价值底线、场景应用导向” 的核心逻辑。

• 以图表形式直观展示了四大核心评测维度及其下属的二级、三级能力指标，形成了完整的“文博大模型应用评测架构体系”。

2. 标准第 6 章“评测内容”作为标准的实体部分，详细规定了四大维度的具体评测内容。

1）通用基础能力评测

这是模型的技术基座，确保其具备处理文博数据的基本技术素养。

• 单模态能力：重点评测文本理解（如对考古报告的分类、信息抽取、长文本总结）、图像识别（文物静态图像分类、对象检测）、音频问答（处理口述史录音）等。

• 多模态能力：这是文博应用的特色与难点。标准要求评测图文检索（如根据青铜器图片找到对应文献）、文物图像问答（基于图片回答专业问题）、古籍图文对齐、视觉语言推理、视频问答（理解修复过程视频）、图表推理（解读地层图、器物演变图）等，全面评估模型关联与融合不同模态信息的能力。

2）文博安全与价值对齐能力评测

这是模型的“高压线”和“指南针”，实行一票否决。

• 安全合规：要求评测模型在数据安全与隐私保护（如对敏感地理坐标脱敏）、法律法规遵循（尤其是文化敏感性内容审查）、内容合规性审核、安全审计与监控等方面的机制与效果。

• 价值对齐：要求评测模型输出在文化价值（尊重文化多样性、准确传达遗产价值）、社会价值（促进教育传承）、伦理价值（无偏见歧视）、专业价值（倡导科学实证精神）等方面与主流价值观及文博使命的契合度。

3）文博专业认知能力评测

这是衡量模型能否称为“文博领域模型”的关键，也是标准最具特色的部分。

• 文博基础知识：评测模型对文物学、考古学、历史学等学科核心概念、术语体系及核心实体（文物、材质、工艺、年代、遗址等）的识别与关联能力。

• 文博信息解读：评测深度专业理解能力，包括：

• 文物图像深度解读：不止于识别物体，要能分析形制、纹饰并辅助断代。

• 古籍与专业文本理解：处理文言文、专业报告，准确抽取关键信息。

• 知识关联与推理：这是高阶能力，要求评测模型基于文博知识图谱（如遵循 CIDOC CRM 构建）进行多跳问答、关系推理、跨时空比较等复杂逻辑任务。这直接对接了知识库中众多文物知识图谱构建的研究成果与应用目标。

4）文博场景应用能力评测

这是检验模型“实战”价值的试金石，将上述能力落实到具体业务中。

• 展览服务能力：智能导览与问答、虚拟展览内容生成、多模态互动体验驱动。

• 文物保护与修复能力：病害智能识别、AI 辅助虚拟修复、修复方案辅助生成。

• 文博教育能力：智能教学辅助、科普内容创作、研学活动设计。

• 文博管理能力：藏品智能管理、学术研究辅助、档案数字化处理。

• 文创开发能力：艺术风格分析与迁移、创意辅助、文化符号提取与解读。

这些场景与知识库中“文博语料项目潜在应用场景”等文档的描述高度一致，确保了评测标准源于实践、服务实践。

3.第 7 章“评测方法”明确如何实施评测，确保结果客观公正。

• 评测方式：采用 “自动化评测 + 人工评测 + 文博专家评审” 三位一体模式。自动化处理客观题；人工评估主观内容（如可读性、设计合理性）；文博专家（文物、考古、博物馆学专家）负责对专业性、科学性进行最终仲裁。这借鉴了医疗、金融领域评测的成功经验，并突出了文博领域对专家权威性的依赖。

• 打分规则与评测等级：规定了量化评分细则和五级等级划分（A 优秀至 E 禁用）。特别强调，若“文博安全与价值对齐”维度得分低于 80 分，无论总分如何，模型均被判定为 E 级（禁用）。这一刚性条款强化了标准在防范风险、引导向善方面的决定性作用。

四、 与相关标准的协调关系

1. 与金融、医疗等领域大模型评测指南的关系：本标准在整体评测框架设计、三位一体评测方法、等级划分逻辑（特别是安全合规的一票否决 ）上，充分借鉴并保持了与 T/SAIAS 019—2024《金融大模型应用评测指南》 等兄弟标准在方法论上的协调与一致。这有利于形成跨领域可比较、可理解的大模型评测标准体系。同时，所有评测内容均完全聚焦并深化于文博领域的特有问题，如专业认知的独特指标、多模态关联的特殊要求等。

2. 与语料库建设标准的关系：本标准与 T/SAIAS 015—2024《语料库建设导则》 及正在制定的《文博语料库建设导则》构成 “数据生产—模型训练—效能评测” 的完整链条。高质量语料库是训练文博大模型的基础，而本评测指南则是检验基于这些语料训练的模型最终效果的标准，两者

相辅相成，共同服务于“AI+文博”的生态建设。

3. 与文化遗产信息国际标准的关系：本标准积极倡导并融入国际文化遗产信息交换的通用语义框架。在“文博专业认知能力”评测中，明确将基于 CIDOC CRM 等标准构建的知识图谱的关联与推理能力作为高级评测指标。这并非强制模型内部采用该模型，而是鼓励其输出能与该标准语义网络对接，从而确保评测结果能反映模型在文化遗产知识结构化、语义化理解方面的潜力，为其未来融入全球文化遗产数据网络预留接口。

五、 重大分歧或难点处理说明

在标准编制过程中，主要研究和解决了以下关键问题：

1. 如何平衡通用大模型能力评测与文博领域专项评测？

• 解决方案：确立了 “基础通用化，专业深度化” 的分层架构。在“通用基础能力”维度，采用与大模型通用评测对齐的指标（如文本分类、信息抽取），确保技术基座可比。在“文博专业认知”和“场景应用”维度，则完全围绕文博特性设计，如引入“文物图像深度解读”、“基于知识图谱的推理”、“文物保护修复方案生成”等独有指标，确保评测能有效区分模型的领域适配度。

2. 如何将抽象的文化价值、安全要求转化为可评测的客观指标？

• 解决方案：借鉴金融、医疗领域经验，将“安全与价值对齐”维度具体化为可操作的评测点。例如，通过设计测试用例，检验模型对敏感文化主题的回应是否合规、对历史虚无主义叙述的辨别能力、在文物年代、归属等争议问题上是否保持客观中立或给出学界主流观点。同时，将“专家评

审”作为该维度评测的核心方式，依靠领域专家的专业判断对模型输出的文化立场、价值导向进行定性评估，并与自动化测试的定量结果相结合。

3. 如何处理文博专业知识体系的复杂性，并设定合理的评测深度？

• 解决方案：不追求面面俱到，而是聚焦于对核心业务场景支撑最关键的知识点。评测指标的设计参考了知识库中关于文博大模型评测应关注的年代判断准确率、材质工艺描述准确性、文化符号解读深度等具体研究成果。同时，将知识深度分为“基础认知”（识别与分类）和“深度解读与推理”（关联与阐释）两个层次，允许模型在不同层次上展现能力，满足从公众科普到专业研究的不同应用需求。

六、 预期效益

本标准的制定和实施，预期将产生以下显著效益：

1. 树立行业标尺，规范市场发展：为文博大模型的研发、评测、选型和采购提供统一的权威依据，终结“各自为政”的混乱局面，引导市场从“概念炒作”走向“能力比拼”，促进行业健康有序竞争。

2. 筑牢安全底线，守护文化根脉：通过强制性的安全与价值对齐评测及一票否决机制，从技术标准层面设立防火墙，有效防范人工智能应用可能带来的数据泄露、文化误读、价值偏差等风险，确保技术应用始终服务于文化遗产的保护与正确传承。

3. 牵引技术研发，提升应用质量：清晰的评测维度与指标如同“指挥棒”，能引导模型研发机构有针对性地加强在文博专业知识学习、多模态关联、场景深挖等方面的投入，从而整体提升文博大模型的技术水平与应用实效。

4. 赋能文博实践，加速智能转型：为博物馆、考古所、数字内容制作企业等应用单位提供可靠的模型能力“体检报告”，帮助其精准匹配业务需求，降低试错成本，加速人工智能在智能导览、文物保护、智慧管理、创意传播等场景的规模化、高质量落地，切实提升文博事业的智能化水平与公共服务能力。

综上所述，《文博大模型应用评测指南》团体标准是在人工智能技术深刻变革文化遗产领域的关键时期，应运而生的一份具有前瞻性、指导性和可操作性的重要技术规范。它融合了国际经验、国内实践与跨领域智慧，旨在系统性地破解当前文博大模型应用中的核心难题，为人工智能技术负责任、高质量地赋能中华优秀传统文化的保护、传承与创新提供坚实的评估基准与发展指南。