团体标准

《生成式人工智能多模态合成数据鉴别的安全技术规范》

编制说明

标准编制小组

2026年3月

《生成式人工智能多模态合成数据鉴别的安全技术规范》编制说明

一、标准制定的必要性

随着 ChatGPT、DeepSeek 等生成式人工智能和多模态大模型技术的快速发展，人工智能在文本、图像、音频、视频等多种模态内容生成方面的能力显著提升，合成数据在信息传播、内容生产和智能应用中的占比持续上升。在推动数字经济和智能社会发展的同时，多模态合成数据被滥用所引发的安全风险、信任风险和治理风险日益凸显，对国家安全、社会稳定和公共治理体系构成了新的挑战。

一方面，生成式人工智能模型能够高质量生成与真实数据高度相似的多模态内容，使得虚假信息、深度伪造、身份冒用、舆论操纵等风险呈现出规模化、自动化和跨模态扩散的特征。合成数据与真实数据之间的边界日益模糊，传统基于人工审核或单一模态特征的鉴别手段难以有效应对，亟须系统化、技术化的鉴别方法和统一的技术规范予以支撑。

另一方面，多模态大模型在训练数据、推理过程和输出结果等环节，可能引入模型幻觉、统计偏差、隐含指纹、异常分布等可用于鉴别的重要特征，但目前相关技术研究和工程实践仍处于分散状态。不同机构和产品在合成数据鉴别的技术路线、特征选取、评估指标和测试方法上差异较大，缺乏统一的术语体系、技术框架和评价标准，导致鉴别结果难以横向对比，检测能力难以客观评估，也不利于相关技术的规模化应用和监管落地。 从行业发展现状看，我国人工智能综合实力已进入全球第一方阵，形成了覆盖算力基础设施、算法模型、数据资源和应用场景的完整产业体系。截至2024年6月，我国人工智能核心产业规模已接近6000亿元，产业链上下游企业超过4700家，生成式人工智能应用正加速向政务、金融、媒体、教育等重点领域渗透。

在此背景下，合成数据的可信性与可鉴别性已成为人工智能产业健康发展的基础性问题，迫切需要通过标准化手段加以规范和引导。 因此，制定《生成式人工智能多模态合成数据鉴别的安全技术规范》，系统规范多模态合成数据鉴别的技术原则、特征分析方法、检测流程、测试要求和评估指标，明确不同模态及跨模态场景下的鉴别技术要求，有助于为行业提供统一、可复现、可验证的技术

依据，提升合成数据治理和安全监管能力，促进生成式人工智能技术在安全、可信、可控的轨道上持续发展。

二、标准编制原则及依据

1. 按照TC260-003《生成式人工智能服务安全基本要求》进行编写。

2. 本标准参考GB/T 45288.2-2025《人工智能大模型第2部分：评测指标与方法》、GB 45438-2025《网络安全技术 人工智能生成合成内容标识方法》等相关标准起草。

3. 参照《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》《中华人民共和国科学技术进步法》《生成式人工智能服务管理暂行办法》等相关法律法规和规定。

4. 本文件的制定工作遵循“统一性、协调性、适用性、一致性、规范性”的原则，本着先进性、科学性、合理性和可操作性的原则，按照GB/T 1.1—2020《标准化工作导则 第一部分：标准化文件的结构和起草规则》给出的规则编写。

三、项目背景及工作情况

（一）任务来源

根据《中国高技术产业发展促进会标准化工作委员会团体标准管理办法》的有关规定，经中国高技术产业发展促进会标准化工作委员会及相关专家技术审核，批准《生成式人工智能多模态合成数据鉴别的安全技术规范》团体标准制定计划，计划编号为：CHI2026033。本标准由内蒙古科技大学提出，由中国高技术产业发展促进会归口。

根据计划要求，本标准完成时限为10个月。

（二）标准起草单位

本标准的主要起草单位是内蒙古科技大学，负责标准文档起草及相关文件的编制等。太原科技大学、新疆大学、重庆理工大学、桂林电子科技大学、中原工学院、重庆文理学院、集美大学、河南科技大学等单位参与起草，负责标准中重要技术点的研究和建议，并参与标准内容的讨论。

（三）标准研制过程及相关工作计划

1.前期准备工作

项目立项前，标准编制小组查阅、研读相关生成式人工智能合成数据相关的鉴别检测文献，广泛搜集与多模态大模型合成数据鉴别相关材料。同时，标准编

制小组安排相关人员，多次与生成式人工智能数据中心的专家进行深入调研和交流，广泛征求多模态大模型合成数据鉴别及安全检测方面的意见和建议。

2. 标准起草过程

2026 年 1 月 26 日，由中国高技术产业发展促进会向国家标准委全国标准服务平台提交立项，立项编号为：CHI2026033，并向全社会公示了十五日。

2026年1月28日，编制组召开第一次起草会议。会议围绕标准适用范围、术语定义及章节内容进行充分讨论，明确分工和阶段性任务及时间节点。

2026年2月10日，召开第二次起草会议，对标准内容进行细化，在标准起草期间，编制小组主编单位及参编单位组织了数次内部研讨会，经过多次修改，于2026年3月2日完成了标准初稿及编制说明的撰写工作。

2026年3月10日，将标准征求意见稿及编制说明提交中国高技术产业发展促进会标准化工作委员会审核通过，并于3月16日在全国标准信息公共服务平台公开征求意见30日。

（四）主要试验（或验证）情况分析

围绕《生成式人工智能多模态合成数据鉴别的安全技术规范》中提出的技术原则、鉴别方法及评估要求，研制单位开展了基于实际系统平台的验证性试验工作，对标准中涉及的多模态合成数据鉴别能力和测试流程的可行性与有效性进行了验证。

前期研究依托自研的《生成式AI合成数据鉴别检测平台》，构建了覆盖文本、图像和视频等多模态数据的测试环境。相关验证工作重点围绕多模态合成内容的检测能力、分析方法及评估结果展示方式展开，为本标准技术内容的制定提供了实践依据。

该合成数据检测平台作为本标准验证工作的技术载体之一，面向多模态大模型应用场景，集成了模型内生安全检测与内容安全检测两类功能模块，可支持多模态合成数据鉴别技术的验证与评估。其中，模型内生安全检测模块主要用于验证大模型在面对合成输入和异常提示条件下的检测表现，支持对抗攻击检测和越狱攻击检测等能力；内容安全检测模块重点面向多模态合成内容，支持对文本、图像和视频等数据的深度伪造特征分析与检测。平台能够对不同模态输入输出的检测结果进行统一展示，并生成相应的分析与评估结果图（如图1所示），用于验证多模态合成数据在不同场景下的鉴别效果。

图1 合成数据鉴别测试结果

基于上述平台，研制单位对标准中提出的多模态合成数据鉴别流程、特征分析方法及评估指标进行了验证性模块开发，运用智能体开发技术，可对相关测试结果进行智能体统计分析，并可视化呈现（如图2所示）。平台涵盖不同模态数据的检测结果、判别特征分布情况以及综合评估分析结果，验证了本标准所规定技术方法在工程实现和应用场景中的可操作性和有效性。

图2 智能体合成数据特征分析结果

四、标准制定的基本原则

标准编制过程中，遵循了以下基本原则：

1） 标准需要具有行业特点，指标及其对应的分析方法要积极参照采用国家标准和行业标准。

2） 标准能够体现出《生成式人工智能多模态合成数据鉴别的安全技术规范》的技术要素。

3） 标准能够为多模态大模型数据中心的安全建设与应用发挥较大的指导性作用。

4） 标准需要具有科学性、先进性和可操作性。

5） 要能够结合行业实际情况和多模态大模型合成数据安全检测平台风险特点。

6） 与相关标准法规协调一致。

7） 促进行业健康发展与技术进步。

五、标准主要内容

本标准聚焦生成式人工智能多模态合成数据鉴别安全，明确了文本、图像、音频、视频及跨模态合成数据的鉴别技术要求、流程与方法，为多模态合成数据安全鉴别提供全维度技术规范，适用于内容审核、身份认证、信息溯源等各类应用场景，核心内容如下：

1.基础定义与范围：界定多模态合成数据、合成数据鉴别、鉴别置信度等核心术语，明确标准覆盖单模态（文本、图像、音频、视频）及跨模态组合合成数据的鉴别，不涉及合成数据生成方法与模型实现，适用于鉴别系统的设计、开发、测试与应用。

2.合成数据分类与安全风险：将多模态合成数据分为文本、语音 / 音频、图像、视频及跨模态合成五类，梳理出身份伪造与欺骗、内容篡改与虚假信息传播、数据滥用与隐私侵害三类核心安全风险，为鉴别技术设计明确目标导向。

3.单模态合成数据鉴别技术：针对四类单模态数据，分别制定特征分析、鉴别方法及流程判定体系。文本聚焦语言流畅性、语义一致性等特征，结合统计分析与深度学习方法鉴别；图像从视觉统计、纹理频域、光照几何特征入手，融合传统统计/频域分析与 CNN、Transformer 等深度学习模型；音频分析频谱声纹、时序韵律等特征，采用信号处理与端到端深度学习模型检测；视频兼顾单帧视觉、跨帧时序、音视频同步特征，结合帧级检测、时序建模与多模态融合技术鉴别。

4.多模态综合鉴别方法：构建跨模态特征一致性检测、多模态联合判定、

不确定性与置信度评估三大核心方法，通过校验不同模态语义、时序逻辑一致性，采用证据理论等融合模型综合判定结果，同时对鉴别结论进行可信度量化，为风险分级和人工复核提供依据。

5.鉴别工具与评估体系：明确鉴别工具需具备多模态处理、自动化分析、结果可解释及批量检测能力，支持主流格式解析与实时流检测；构建包含文本、图像、音频、视频的多类型评测集，从准确率、召回率、F1-Score 等指标评估鉴别性能，同时通过对抗攻击、环境扰动测试验证方法的鲁棒性与泛化能力。

六、与有关法律法规和强制性标准的关系

1.本标准符合《中华人民共和国标准化法》及现行法律法规的规定，本标准与其他相关标准没有矛盾之处。

2.本标准针对生成式人工智能多模态合成数据鉴别安全进行了技术细节的补充，以满足市场和创新的需要。

3.参照相关法律法规和规定，在编制过程中着重考虑了科学性、适用性和可操作性。

4.本标准的技术要求严格遵循《强制性国家标准管理办法》的规定，确保不低于强制性标准的要求，以保障人身健康和生命财产安全。

七、重大意见分歧的处理依据和结果

本标准起草过程中没有重大分歧意见。

八、采标程度，国内外同类标准水平的对比情况

国内外尚没有《生成式人工智能多模态合成数据鉴别的安全技术规范》这方面的标准。

九、涉及专利的有关说明

无

十、后续贯彻措施

做好宣传培训，建议由各行业主管部门组织、主要起草单位配合开展标准宣贯培训工作，使相关检测人员了解标准、熟悉标准，掌握标准的各项技术要求，强化示范效应，让标准在行业内得到广泛推广和应用，使标准的应用落到实处。

对《生成式人工智能多模态合成数据鉴别的安全技术规范》团体标准执行情况进行跟踪调查，及时发现标准中执行的问题，不断修改完善，提高标准水平，提高标准的科学性、合理性、协调性和可操作性。

建议本标准发布之日起半年内实施。

标准编制小组

2026年3月