多模态大模型生成内容标识合规模型清单发布

给AI生成内容造假者戴上“紧箍咒”

► 本报记者 李洋

近日，中国软件评测中心选取一批国内市场上活跃的知名多模态大模型产品，开展显式标识检测和隐式标识检测工作。检测结果显示，被检测样本文生图场景完全合规率为81.8%，文生视频场景完全合规率为90.9%。与此同时，检测发现某些大模型存在生成的图片或视频未发现任何显式标识和隐式标识等情况，推测其生成功能的合规改造可能尚未完成或存在技术疏漏，不符合强制性标准要求。

据悉，此次选取的国内市场上活跃的知名多模态大模型产品，检测时间为2025年9月1-2日，覆盖文生图和文生视频两大主流应用场景。

技术路线已逐步成熟

今年3月，国家互联网信息办公室联合多个部门发布《人工智能生成合成内容标识办法》（以下简称《标识办法》）。当月，国家市场监督管理总局发布了强制性国家标准《网络安全技术 人工智能生成合成内容标识方法》，已于2025年9月1日起正式施行。上述政策规定及标准均明确规定，人工智能生成内容（AIGC）服务提供者需对生成内容进行显著标识，且标识应包含显式标识与隐式标识两部分。

中国软件评测中心（工业和信息化部软件与集成电路促进中心）人工智能研究测评事业部副总经理孙佰鑫在接受本报记者采访时表示，AIGC内容的标识要求是确保技术透明、可信的重要基石。

孙佰鑫介绍说，显式标识的核心要求在于“清晰可见、不易去除、易于理解”，具体是指必须在生成内容的显著位置（如图片/视频的角落、音频的开始或结束、文本的标题或末尾）以明确的方式标注。例如，使用“AI生成”或“由人工智能创作”等字样，或使用统一的公众易于识别的徽标或水印，该标识应贯穿内容传播的全周期。而隐式标识的关键信息可以理解为，内容的“数字身份证”或“基因密码”，通常以元数据的形式嵌入在文本内部，需要特定工具才能读取，其关键信息应至少包含生成模型标识、生成主体信息、生成时间戳、内容摘要/指纹。隐式标识的核心目的是为了溯源、取证和验证，为监管和责任认定提供技术依据。

此次检测结果显示，各大厂商在显示标识方面普遍采用水印策略。在隐式标识方面，各主流厂商采用在元数据区中添加特定对象或键值的方式实现标准要求。同时，得益于主流视频容器格式对元数据的良好支持以及标准化处理工具链的普及，视频领域通过文本元数据添加标识的技术路径已经逐步成熟。

孙佰鑫解释说，被检测样本文生视频场景完全合规率为超九成，略高于文生图场景完全合规率，造成这一结果的原因是部分厂商仅支持文生图功能，不支持文生视频，导致参与测试的文生图厂商及样本数量更多；且部分文生图大模型厂商未能及时嵌入AIGC标识，拉低了整体合规率。

然而，孙佰鑫也表示，文生视频（Text-to-Video）是当前全球人工智能（AI）领域最前沿、技术难度最高的挑战之一，涉及复杂的时空一致性、物理逻辑等问题，目前厂商的精力仍集中在攻克基础生成质量、延长视频时长等核心技术方面。

应实行分级分类治理

专家表示，强制标识制度能够有效提升公众对AIGC内容的辨别能力，同时也能倒逼厂商在技术研发之初就将合规性、可控性、可追溯性纳入核心设计，而不是仅仅追求生成效果。

“这将促使厂商在识别精度、防篡改技术、溯源能力等方面展开良性竞争，从而优化整个行业的技术生态和竞争格局，淘汰那些只追求短期利益、忽视社会责任的参与者。”孙佰鑫说。

孙佰鑫认为，随着AIGC技术向教育、医疗、媒体等更多领域渗透，标识制度应进行动态化和精细化调整，实行“一类一策”甚至“一场景一策”的分级分类治理模式。

“不能用一把‘尺子’衡量所有行业。”孙佰鑫表示，首先，风险等级存在差异，不同行业的风险容忍度完全不同。在医疗、金融、司法等领域，AIGC内容的错误可能直接危及生命财产安全或社会公平正义，因此，需要最高级别的标识要求，如强制、不可关闭的显式标识和更详尽的隐式溯源信息。在教育和媒体等领域，内容需要兼顾可信度和体验，标识方式可以更灵活，但标准则需要严格。在娱乐、艺术创作领域，在确保基本标识的前提下，可以允许更艺术化、更不干扰体验的标识方式。

其次，不同行业需要强调的标识信息侧重点不同。如，医疗健康内容标识必须强烈警示“此内容不可作为医疗诊断依据”，并需包含生成所依据的数据来源和模型局限性说明；新闻媒体需明确标注AI生成的是文字、图片还是视频，并可能要求注明信息源和自动化生成的程度；教育需要标识内容符合教学大纲版本、适合学龄段以及是否经过教育专家审核等额外信息。

第三，针对实时交互的AI医生、沉浸式VR（虚拟现实）教育等不同形态的内容，还需要研发新的标识技术，如语音播报标识、虚拟环境中的标识等。