国家标准《网络安全技术 人工智能应用安全分类分级方法》（征求意见稿）编制说明

一、工作简况

1.1 任务来源

根据国家标准化管理委员会2026年下达的国家标准制修订计划，《网络安全技术 人工智能应用安全分类分级方法》由中国电子技术标准化研究院负责承办，计划号：20260951-T-469。本标准由全国网络安全标准化技术委员会归口管理。

1.2 制定背景

2023年10月18日，习近平主席在第三届“一带一路”国际合作高峰论坛开幕式主旨演讲中宣布中方提出《全球人工智能治理倡议》，倡议中明确指出对人工智能实施分类分级管理。同时，我国不同管理文件中均提出人工智能从安全角度进行分类分级管理的要求，如《生成式人工智能服务管理暂行办法》《人工智能拟人化互动服务管理暂行办法》等，急需研制人工智能应用安全分类分级方法标准，作为我国人工智能安全治理机制的关键支撑，从顶层规划人工智能在安全角度分类分级的框架思路。

本标准的编制基于全国网络安全标准化技术委员会《人工智能安全治理框架》相关研究基础，以及编制组围绕人工智能应用安全分类分级国际国内相关做法的动态研究，充分参考借鉴了国际国内的现行管理机制与做法。标准编制过程中与各有关单位进行了充分的沟通，充分论证其技术可行性，包括多次组织人工智能、网络安全、行业领域相关单位研讨。特别是组织教育、广电、卫健、金融、应急等行业领域研究单位参与研讨，共同形成共识。

1.3 起草单位

本标准由中国电子技术标准化研究院牵头，参与单位包括浦江国家实验室、北京中关村实验室、国家计算机网络应急技术处理协调中心、浙江大学、中国政法大学、清华大学、北京航空航天大学、复旦大学、北京理工大学、中央网信办（国家网信办）数据与技术保障中心、公安部第三研究所、中国信息安全测评中心、中国信息通信研究院、中国移动通信集团有限公司、华为技术有限公司、北京小米移动软件有限公司、中国电信集团有限公司、科大讯飞股份有限公司、北

京快手科技有限公司、北京火山引擎科技有限公司、中国科学院信息工程研究所、中国科学院软件研究所、OPPO 广东移动通信有限公司、浙江工业大学、广西壮族自治区标准技术研究院、哈尔滨工业大学、中国网络空间研究院、国家广播电视总局广播电视科学研究院、国家卫生健康委统计信息中心、教育部教育管理信息中心、应急管理部大数据中心、自然资源部地图技术审查中心、中国工商银行股份有限公司、煤炭科学研究总院有限公司、广西电网有限责任公司、行吟信息科技（上海）有限公司、北京深度求索人工智能基础技术研究有限公司、河南科技大学、西安交通大学、工业和信息化部电子第五研究所、上海赛西科技发展有限责任公司、上海燧原科技股份有限公司、北京天融信网络安全技术有限公司、北京神州绿盟科技有限公司、北京奇虎科技有限公司。

本文件主要起草人：姚相振、郝春亮、张妍婷、胡侠、周睿康、刘勇、任奎、王迎春、张震、孟令宇、赵鞾、王志伟、盛小宝、秦湛、徐阳、骆意、武姗姗、张凌寒、苏紫敏、方强、史倩君、刘建伟、杨珉、宣琦、吕飞霄、李子威、徐葳、贺敏、马梦娜、柳嘉琪、落红卫、谷晨、郭建领、张志勇、关振宇、姜伟、洪延青、徐甲、王姣、王秉政、范博、董汀、王嘉凯、崔栋、赵宇航、石霖、乔玉平、王蕊、王俊杰、唐迪、吴少卿、刘莹、李根、黄龙、刘北水、李寒雨、梅敬青、刘帅、张琳琳、郭晓霞、叶红、宋宇宸、吴子坚、王普、王龑、张睿、邹权臣、林秋诗、费凡芮、宋昊、张立尧、黄晴、李芒。

其中：姚相振、郝春亮、张妍婷、胡侠、周睿康负责标准整体推进、设计标准框架，以及组织调查研究、技术研讨、文本撰写等工作。吕飞霄、李子威、孟令宇、徐阳、方强、史倩君、柳嘉琪、落红卫、谷晨、郭建领、王嘉凯、崔栋、赵宇航、费凡芮、宋昊、张立尧、林秋诗、王普负责对标准内容进行研讨、技术论证并撰写相应条款。刘勇、任奎、王迎春、刘建伟、杨珉、宣琦、徐葳、王志伟、贺敏、张震、盛小宝、秦湛、张凌寒、关振宇、姜伟负责对标准技术内容进行深度研讨并做技术把关。武姗姗、苏紫敏、张志勇、洪延青、乔玉平、王蕊、王俊杰、唐迪、吴少卿、梅敬青、王姣、王秉政、刘帅负责国内外情况调查研究。马梦娜、徐甲、范博、李根、刘北水、董汀、李寒雨、黄晴、李芒负责对标准条款进行优化完善。赵鞾、骆意、黄龙、石霖、张琳琳、刘莹、郭晓霞、叶红、宋宇宸、吴子坚、王龑、张睿、邹权臣负责研判标准内容与行业领域实际情况的适

配性。

1.4 起草过程

1、2024年12月，广泛调研国内外人工智能应用情况、安全需求、分类分级工作等。

2、2025年1月，针对人工智能应用情况以及主要安全风险开展调研，收到浦江国家实验室、中关村实验室、快手、抖音等17家单位反馈。

3、2025年2月，经多次内部讨论，结合我国人工智能安全治理需求、国内外人工智能应用及安全治理工作情况，形成人工智能应用安全分类分级初步方案。

4、2025年3月13日，组织研讨会，围绕不同风险应用场景展开讨论，浦江国家实验室、CNCERT、中关村实验室、中国政法大学、北航、华为等27家单位参与研讨。

5、2025年4月17日，在网安标委2025年第一次标准周活动的新技术安全标准特别工作组会议上，编制组汇报标准编制情况，工作组成员单位对标准草案进行讨论并投票，最终表决通过。编制组根据各单位意见修改完善标准草案。

6、2025年6月23日，组织浦江国家实验室、中关村实验室、抖音、华为、快手等14家单位研讨，修改完善标准草案。

7、2025年8月6日，组织交通、金融、广电、教育、应急等行业领域单位研讨，修改完善标准草案。

8、2025年9月8日，在新技术安全标准特别工作组组织的专家评审会上，编制组汇报本标准编制情况，听取专家意见，并按专家意见进行修改完善。

9、2025年9月16日，在网安标委2025年第二次标准周活动的新技术安全标准特别工作组会议上，编制组汇报标准编制情况，工作组成员单位对标准草案进行讨论并投票，最终表决通过，推荐转为征求意见稿。编制组根据各单位意见修改完善标准草案。

10、2025年10月22日，组织标准启动会，会上介绍标准研制情况以及后续工作规划，进行广泛研讨，并邀请广电、交通等行业领域单位分享行业实践。

11、2025年11月-2026年1月，编制组经多次内部讨论、广泛征求意见，完善标准草案。

12、2026年1月19日，秘书处组织专家评审会，编制组汇报标准编制情况，根据专家意见完善标准草案。

13、2026年4月3日，在网安标委2026年第一次标准周活动的人工智能安全标准工作组会议上，编制组汇报标准编制情况，工作组成员单位对标准草案进行讨论，编制组根据各单位意见修改完善标准草案。

14、2026年5月27日，秘书处组织征求意见稿专家审查会，对标准征求意见稿及相关材料进行了审查和质询，与会专家一致同意通过审查，建议编制工作组修改完善后发起公开征求意见。编制组根据专家意见修改完善征求意见稿。

二、标准编制原则、主要内容及其确定依据

2.1 标准编制原则

本标准的编制原则是：

1）通用性：面向有关行业主管（监管）部门、地方主管部门以及人工智能应用开发者、运营者等编制本文件，提供统一的人工智能应用安全分类分级方法，帮助各方在不同地区、不同行业开展分类分级工作，并为制定行业标准规范、实施细则及配套管理措施提供依据。

2）实用性：结合我国人工智能安全治理的管理需求与行业实践基础编制本标准，明确分类维度、分级等级、分级要素及实施步骤，便于按流程操作、按要素判定、按结果复核，确保在实际工作中可直接应用、可推广复用。

3）符合性：本标准以国家人工智能安全治理总体部署为牵引，标准内容与我国网络安全、数据安全、个人信息保护等相关法律法规要求保持一致，并与现行政策文件及已有标准规范相衔接，确保标准体系内在一致、适用边界清晰。

2.2 主要内容及其确定依据

本标准按照《人工智能安全治理框架》明确的总体思路，进一步规定了人工智能应用安全分类分级的方法与实施流程，主要包括基本原则、分类方法、分级方法等内容，适用于适用于人工智能应用开发者、运营者等开展应用安全分类分级活动。本标准依据《全球人工智能治理倡议》提出的分类分级管理要求，以及《生成式人工智能服务管理暂行办法》《人工智能拟人化互动服务管理暂行办法（征求意见稿）》等政策文件对分类分级监管的工作部署，参考国内外相关做法，结合教育、广电、医疗、金融、应急等行业实践，通过多轮征求意见、专题研讨

和专家论证，形成可供各行业、各地区共同遵循的统一技术依据，支撑相关制度要求落地实施。

2.3 修订前后技术内容的对比[仅适用于国家标准修订项目]

不涉及。

三、试验验证的分析、综述报告，技术经济论证，预期的经济效益、社会效益和生态效益

3.1 试验验证的分析、综述报告

人工智能应用在不同行业快速落地，应用形态、部署方式、服务对象与风险外溢程度差异显著，行业主管（监管）部门与实施主体迫切需要一套可统一执行、可复核、可动态调整的安全分类分级方法，为监管要求落地、行业细则制定和企业自评自查提供依据。为验证本标准方法的可操作性与适用性，编制组结合教育、广电、医疗、金融、应急等典型行业场景，组织相关研究单位、企业及技术支撑机构开展多轮研讨。

3.2 技术经济论证

本标准的分类分级方法以已有研究基础和行业实践为支撑，实施流程清晰，所涉及的分级要素在各行业应用中具备可获取性和可操作性，技术可行性良好。在经济性方面，本标准方法依托现有安全管理体系即可开展实施，不依赖额外的专用技术工具或平台，实施增量成本可控；同时，统一分类分级口径有助于减少各行业、各实施主体间重复评估与制度对接带来的资源消耗，降低合规成本，整体投入产出合理，具备推广实施的经济可行性。

3.3 预期的经济效益、社会效益和生态效益

在经济效益方面，本标准可为行业提供可对照、可复用的参考路径，帮助应用开发者、运营者更高效开展分类分级与差异化安全投入，降低重复评估和合规沟通成本，提高资源配置效率，带来直接的经济性收益。在社会效益方面，标准推广将促进各行业形成统一的安全分类分级认知，提升人工智能应用安全治理的可执行性，增强公众对人工智能应用的信任，支撑相关制度要求更好落地。在生态效益方面，统一的分类分级规则有助于行业主管（监管）部门在同一基准上制定配套规范与实施细则，带动测评、审查、咨询、工具与管理流程等配套服务协同发展，形成良好闭环的治理生态。

四、与国际、国外同类标准技术内容的对比情况，或者与测试的国外样品、样机的有关数据对比情况

不涉及。

五、以国际标准为基础的起草情况，以及是否合规引用或者采用国际国外标准，并说明未采用国际标准的原因

不涉及。

六、与有关法律、行政法规及相关标准的关系

本标准是我国构建人工智能安全治理机制的关键标准支撑。法律法规层面，与《网络安全法》《数据安全法》《个人信息保护法》等保持一致；政策制度层面，落实《全球人工智能治理倡议》提出的“分类分级管理”以及《生成式人工智能服务管理暂行办法》提出的“分类分级监管”，为行业主管监管部门提供顶层标准依据。标准层面，本标准《网络安全技术 人工智能应用安全分类分级方法》与在研国家标准《网络安全技术 人工智能安全能力成熟度评估方法》形成配套关系，本标准提供分类分级的基础方法，成熟度标准进一步开展安全能力水平评估相关工作，两者共同支撑人工智能安全治理工作的规范化推进。

七、重大分歧意见的处理经过和依据

不涉及。

八、涉及专利的有关说明

不涉及。

九、实施国家标准的要求，以及组织措施、技术措施、过渡期和实施日期的建议等措施建议

本标准规定了人工智能应用安全分类分级方法，适用于行业领域主管（监管）部门参考制定本行业本领域的人工智能应用安全分类分级标准规范，也适用于各地区、各部门开展人工智能应用安全分类分级工作，同时为人工智能应用开发者、运营者进行应用安全分类分级提供参考。

十、其他应当说明的事项

我单位已按照公平竞争审查表中各项内容进行了逐项审查，从市场准入与退出、商品要素自由流动、生产经营成本、生产经营行为等方面逐项开展审查，未发现有内容存在排除、限制市场竞争的情形，符合公平竞争要求。

本标准不存在版权问题。

《网络安全技术 人工智能应用安全分类分级方法》标准编制组

2026 年 7 月 10 日