国家标准《网络安全技术 生成式人工智能服务安全基本要求》（征求意见稿）编制说明

一、工作简况

1.1 任务来源

为加强网络安全国家标准在国家网络安全保障工作中的基础性、规范性、引领性作用，全国网络安全标准化技术委员会秘书处调研国家网络安全重点工作和技术产业发展需求，研究形成了2023年度第二批网络安全国家标准需求清单，其中包含本标准，中国电子技术标准化研究院联合数十家相关单位参与申报本标准，已完成在网络安全标准化技术委员会的立项工作。

本标准支撑《生成式人工智能服务管理暂行办法》（以下简称《办法》），给出了生成式人工智能服务在安全方面的基本要求，包括训练数据安全、模型安全、安全措施等，并给出了安全评估参考要点。

1.2 制定背景

生成式人工智能已成为继移动互联网技术之后最大的一波技术浪潮，在全球范围内引发科技巨头争相布局、掀起创业热潮的链式反应，同时也带来了新的安全风险和挑战。2023年7月，国家网信办等七部门发布了《办法》，从政策法规层面为我国生成式人工智能健康发展保驾护航，为有序开展相关管理工作明确了方向。为响应中央重视生成式人工智能发展，营造创新生态，重视防范风险的要求，在生成人工智能的训练数据、数据标注、生成内容、用户权益等方面开展安全保护，全方位支撑《办法》，推动生成式人工智能的安全健康发展，《网络安全技术 生成式人工智能服务安全基本要求》被列入国家标准需求。

本标准对《办法》中的安全要求进行细化，规定了生成式人工智能服务在安全方面的基本要求，针对当前生成式人工智能服务研发过程中的网络安全、数据安全、个人信息保护，以及面向服务过程中的应用场景安全风险、软硬件环境安全风险、生成内容安全风险、权益保障安全风险等方面，提出细化安全要求，帮助明确服务网络安全基线、提高企业服务安全水平。充分发挥标准对生成式人工

智能管理工作的支撑作用，对防范生成式人工智能重大安全风险、提升生成式人工智能安全水平起到规范引导作用，促进生成式人工智能行业安全发展。

1.3 主要工作过程

1、2023 年 6 月，成立标准编制组，开展标准制定前期的技术积累，调研国内外生成式人工智能的技术发展、产业化应用情况、安全应用需求等，并启动标准编制工作。

2、2023 年 6 月至 7 月，编制组针对重点问题进行调查研究，分析国内主要生成式人工智能企业情况，调研生成式人工智能常用训练数据集安全性，经多次内部讨论、广泛征求意见，初步形成标准草案。

3、2023 年 8 月，编制组经多次内部讨论、广泛征求意见，结合国内外生成式人工智能发展情况，组织 30 余家相关单位多次展开调研、召开研讨会，多次组织专家会，先后形成 10 余版标准草案。

4、2023 年 8 月 30 日，编制组在大数据安全标准特别工作组立项研讨会上，向工作组汇报了本标准的编制情况，工作组成员单位对本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进行修改，完善标准草案。

5、2023 年 9 月，编制组多次召开内部研讨会，修改完善标准草案。

6、2023 年 9 月 16 日，编制组在网安标委组织的 2023 年第二批网络安全国家标准立项专家评审会上，向专家组汇报本标准编制情况，专家组同意通过对本标准的审查。编制组根据专家意见进行修改，完善标准草案。

7、2023 年 9 月 28 日、10 月 8 日，编制组组织召开专家研讨会，根据专家意见进行修改，完善标准草案。

8、2023 年 11 月 3 日，编制组在网安标委 2023 年第二次标准周活动的大数据安全标准特别工作组会议上，向工作组汇报本标准的编制情况，工作组成员单位对本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。编制组根据意见进行修改，完善标准草案。

9、2023 年 11 月 8 日，编制组组织召开专家研讨会，根据专家意见对标准内容进行修改，完善标准草案。

10、2023 年 11 月-2024 年 3 月，编制组多次召开内部研讨会，密集推动标准草案的完善工作，先后形成了 10 余版征求意见稿。

11、2024 年 4 月 11 日，编制组在网安标委组织的征求意见稿专家审查会上，向专家组汇报标准主要技术内容及意见处理情况，专家组同意通过对本标准的审查。编制组根据专家意见进行修改，完善征求意见稿。

12、2024 年 4 月 12 日、4 月 18 日，编制组多次组织召开专家研讨会，对征求意见稿进行讨论，会后根据专家意见进行修改，完善征求意见稿。

13、2024 年 5 月 6 日，编制组召开内部研讨会，修改完善征求意见稿。

14、2024 年 5 月 11 日，编制组向新技术安全标准特别工作组汇报标准编制情况，工作组对标准征求意见稿进行讨论，同意本标准面向社会征求意见。

15、2024 年 5 月 14 日，编制组召开内部研讨会，根据工作组专家意见进行修改，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：面向生成式人工智能服务的共性安全要求编制本文件，帮助相关单位提高安全水平，为评估工作提供依据。

2）实用性：根据我国生成式人工智能技术的发展情况以及生成式人工智能服务的实际应用场景编制本标准，使其在指导生成式人工智能服务方面具有很强的实用性。

3）符合性：符合国家有关法律法规和已有标准规范的相关要求。

2.2 主要内容及其确定依据

本标准针对生成式人工智能服务存在的安全风险，对现有生成式人工智能技术及其面向公众提供的服务进行分析与调研，对《办法》中的安全要求进行细化，规定了生成式人工智能服务在安全方面的基本要求，包括训练数据安全、模型安全、安全措施等，并给出了安全评估参考要点。

在标准编制过程中，本标准充分吸收了数十家单位的头部企业、研究机构的研究成果和应用实践，具备较好的产业基础，以及技术先进性、创新性。在后续编制过程中，本标准也会根据生成式人工智能发展与安全情况持续完善。标准工作基于 TC260-003《生成式人工智能服务安全基本要求》，该文件已经在各管理部门、各企业有较好共识，并已形成相关安全实践，已经在各企业获得了普遍实

践，标准内容的产业化基础比较充分。

2.3 修订前后技术内容的对比[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

生成式人工智能服务快速发展的当下，在训练数据、数据标注、生成内容安全、生成内容标识、使用者权益保护以及反歧视、透明性等方向均迫切需要有指导性要求为提供者指引，本标准通过基本要求的纲领性条款为服务提供者提供指导性的合规和安全建设的方向，标准范围覆盖生成式人工智能的各个维度。在标准试点执行层面，将选取典型的服务提供者、基于典型的业务场景和典型的安全条款开展标准的应用推广试点工作。

本标准的应用推广牵头单位为北京百度网讯科技有限公司，标准的应用实施试点工作单位拟选取若干已经对公众提供服务的生成式人工智能服务提供者，或者有标准条款中的全部或者部分实施经验的企业。各试点单位结合试点实施报告，梳理各自试点情况，研提对标准的修改意见、标准推广的思路等，形成试点工作总结报告。组织方在各试点单位总结的基础上，完成试点工作整体情况的总结，并召开试点总结会议，对试点工作进行总结，对相关成果进行验收。

3.2 预期的经济效益、社会效益和生态效益

通过试点工作，对标准条款的准确性和适用性进行实操演练，吸纳行业内的最佳实践，解决标准中的难点，为标准的进一步完善和标准推广打好基础。试点中将总结过程中的优秀案例，通过典型企业的典型实践，为行业内树立标杆，促进企业安全与合规能力的建设，快速推进生成式人工智能服务的规范化、合规化。试点也为标准后续实施奠定基础，通过试点工作了解行业现状、企业难点，通过标准编制组、技术支撑单位、参与企业的通力合作，为行业推广时识别风险，扫清障碍，促进标准的落地实施。

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《网络安全技术 生成式人工智能服务安全基本要求》与在研国家标准《网络安全技术 生成式人工智能预训练和优化训练数据安全规范》《网络安全技术 生成式人工智能数据标注安全规范》均为《办法》的配套支撑文件。

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

本标准规定了生成式人工智能服务在安全方面的基本要求，包括训练数据安全、模型安全、安全措施等，并给出了安全评估参考要点。

本标准适用于以交互界面、可编程接口等形式向公众提供生成式人工智能服务的组织或个人。可用于指导生成式人工智能服务提供者开展安全评估，也可为相关主管部门提供参考。

十、其他应当说明的事项

无。

标准编制组

2024 年 5 月 15 日