ICS 35.240
CCS L 70

团　体　标　准

T/TAF 309—2025

生成式人工智能产品和服务风险分类分级指南

Guidelines for classification and grading of generative artificial intelligence products and services based on risks

2025-08-11 发布　　2025-08-11 实施

电信终端产业协会　发布

版权声明

邮箱：tafrb@taf.org.cn

电话：010-82052809

目　　次

前　言

本文件按照GB/T 1.1—2020《标准化工作导则　第1部分：标准化文件的结构和起草规则》的规定起草。

请注意本文件的某些内容可能涉及专利。本文件的发布机构不承担识别专利的责任。

本文件由电信终端产业协会（TAF）提出并归口。

本文件起草单位：中国信息通信研究院、荣耀终端股份有限公司、OPPO广东移动通信有限公司、北京快手科技有限公司、北京抖音信息服务有限公司、维沃移动通信有限公司、华为终端有限公司、高通无线通信技术（中国）有限公司、小米通讯技术有限公司、北京三星通信技术研究有限公司、蚂蚁科技集团股份有限公司、百度在线网络技术（北京）有限公司、上海合合信息科技股份有限公司、北京三快在线科技有限公司、中国联合网络通信有限公司、联想（北京）有限公司、厦门美柚股份有限公司、北京微梦创科网络技术有限公司、南德认证检测（中国）有限公司、北京卡路里科技有限公司、上海得物信息集团有限公司、广州视源电子科技股份有限公司、深圳依时货拉拉科技有限公司、新华三技术有限公司。

本文件主要起草人：王淞鹤、王艳红、李辰淑、武林娜、邓佑军、杜云、陈鑫爱、赵晓娜、王海锋、李根、李腾、落红卫、谷晨、杜蕾、邹恬圆、姚一楠、赵盈洁、衣强、王江胜、刘海涛、吴越、王彬、林冠辰、徐艺溦、宋宏宇、魏亚楠、吴斌、刘觅、刘俊、黄鹏华、任资政、康宇、周世乐、曹昉赫、杨欢、尹志超、白雷、张洪龙、李培。

引　　言

随着生成式人工智能技术在很多领域开始广泛应用，生成内容也从原来相对单一的文本类型扩展至图像、音频、视频，甚至3D模型等，给人们的工作和生活带来了极大便利。但与此同时，由于模型训练的数据规模更大，模型可解释性、生成内容可控性以及模型使用规范性等面临更多挑战，更容易存在侵犯隐私权、肖像权、名誉权等风险。

2023 年国内七部门发布《生成式人工智能服务管理暂行办法》，其中第三条提出对生成式人工智能服务实施包容审慎和分类分级监管的原则，第十六条提出针对生成式人工智能技术特点及其在有关行业和领域的服务应用制定相应的分类分级监管规则或者指引的要求。但是，当前政策法规对于生成式人工智能服务的分类分级方法、依据及相应的管理规定仍不够细致，同时国内相关学者专家对于分类分级的理解与建议也稍存差异，这些都对未来分类分级治理和监管带来挑战。

因此，亟需研究制定基于风险的生成式人工智能产品和服务分类分级标准，为企业或监管部门对相关产品和服务开展分类分级治理工作提供基础和参考。

生成式人工智能产品和服务风险分类分级指南

1 范围

本文件提供了生成式人工智能产品和服务的分类分级方法，包括分类分级原则、框架、方法等方面的指导和建议。

本文件适用于指导组织生成式人工智能产品和服务分类分级工作。

2 规范性引用文件

本文件没有规范性引用文件。

3 术语和定义

下列术语和定义适用于本文件。

3.1

生成式人工智能产品和服务 generative artificial intelligence products and services

利用生成式人工智能技术，向公众提供生成文本、图片、音频、视频等内容的产品和服务。

注：以下简称“产品和服务”。

4 分类分级原则

首先对产品和服务在应用及服务时可能面临的风险进行分类，然后基于该风险分类对产品和服务进行分级。生成式人工智能产品和服务的分类分级宜遵循以下基本原则。

a) 科学实用原则：从便于产品和服务管理和使用的角度，科学选择常见、稳定的属性或特征作为分类分级的依据，并结合实际需要对产品和服务进行分类分级。

b) 边界清晰原则：产品和服务分级的各级别应边界清晰，对不同级别的产品和服务采取相应的保护措施。

c) 点面结合原则：分级既要考虑单一影响对象的影响程度，也要充分考虑多个影响对象汇聚融合后的影响程度，综合确定风险级别。

d) 动态更新原则：根据模型或技术的发展、应用场景的变化、产业发展需求更新等进行调整，并对分类分级结果定期审核。

5 风险分类

产品和服务按照在应用及服务时可能面临的风险进行分类，包括输出危害意识形态、国家安全、社会稳定的违法违规信息，输出威胁现实世界和网络世界安全的信息，生成侵犯组织合法权益的内容，输出内容被滥用，信息泄露风险、生成歧视性内容，形成信息茧房，输出内容混淆事实、误导用户等，具体如下：

a) 输出危害意识形态、国家安全、社会稳定的违法违规信息：AIGC 应用及服务被用于生成虚假文本、图像、音频、视频等，用于传播含有煽动颠覆国家政权、损害国家形象、破坏国家统一、宣扬恐怖主义、挑动社会分裂、宣扬暴力色情等违法违规信息，操纵舆论、影响大众认知；
b) 输出威胁现实世界和网络世界安全的信息：恐怖组织、犯罪分子等通过越狱攻击绕过 AIGC 设置的安全护栏，让大模型输出包含用于设计、制造、使用生化核导、网络攻击武器等信息，或者其它开展违法犯罪活动的信息，严重危害公共安全；
c) 生成侵犯组织合法权益的内容：由于大模型的训练数据集中包含未获得有效的授权、许可的商业秘密或有知识产权的内容，在用户提示词引导下，造成其它组织的商业秘密或知识产权泄露；
d) 输出内容被滥用：受经济、社会等利益驱使，别有用心的组织或个人改变 AIGC 生成内容的正当用途，用于创意剽窃、内容抄袭、学术造假、网络诈骗等违法不当目的；
e) 信息泄露风险：使用服务时，输入个人敏感信息或组织内部敏感数据，被 AIGC 服务提供者保存及记录，如果缺乏有效安全保护或合法利用，可能导致个人隐私、组织商业秘密的泄露；
f) 生成歧视性内容：模型设计及训练过程中，个人偏见被有意、无意引入，或者因训练数据集质量问题，导致输出结果存在民族、宗教、国别、地域等歧视性内容，可能引发社会分裂、人群对立等问题，破坏和谐、稳定的社会环境；
g) 形成信息茧房：根据用户历史交互数据或用户画像持续输出用户偏好的同质化内容，导致其认知局限、思想固化、过度关注某类内容；
h) 输出内容混淆事实、误导用户：由于模型训练质量不高，导致输出的内容不准确，严重不符合科学常识或主流认知；或者内容不可靠，虽然不包含严重错误的内容，但无法对使用者形成帮助。

6 分级规则

6.1 分级方法

根据生成式人工智能产品和服务在应用领域的重要程度，以及一旦产生安全问题，对国家安全、社会秩序、组织权益、用户权益造成的危害程度，将生成式人工智能产品和服务分为高风险、中风险、低风险三个级别。

针对生成式人工智能产品和服务，具体可参考以下步骤进行分级。

a) 确定分级对象：确定待分级的生成式人工智能产品和服务。
b) 风险要素识别：结合本产品和服务特点，按照 6.2 识别涉及的分级风险要素情况。
c) 影响分析：结合分级风险要素识别情况，分析产品和服务的潜在风险、可能的影响对象和影响程度。
d) 确定风险级别：根据风险分析情况，确定风险级别。

6.2 风险要素

产品和服务在应用及服务层的分级风险要素，包括但不限于应用领域及场景、生成内容质量、产品和服务鲁棒性、用户规模、使用群体、业务舆论属性、数据敏感度。

a) 应用领域及场景：产品和服务所属的业务范畴及服务场景，以及对用户生命、财产、健康等产生的影响。
b) 生成内容质量：生成内容的真实性、准确性。
c) 产品和服务鲁棒性：模型抵御对抗攻击、提示词攻击等的安全能力。
d) 用户规模：产品和服务服务的用户量。

e） 使用群体：产品和服务面向的群体，例如专门面向未成年人、老年人、残疾人或其他弱势群体。
f） 业务舆论属性：产品服务的舆论属性或社会动员能力。
g） 数据敏感度：在训练过程中使用到核心数据、重要数据及个人信息等敏感数据。

6.3 影响分析

6.3.1 影响对象

影响对象是指面临安全风险时，可能影响的对象。其中，安全风险主要考虑生成式人工智能产品和服务遭到攻击、篡改、损毁、恶意使用或发生异常时引发的安全问题。影响对象包括国家安全、社会秩序、组织权益、用户权益。

国家安全：影响国家政权稳固、国家统一、民族团结等国家利益安全。

社会秩序：影响社会治安和公共安全、社会日常生活秩序、民生福祉、法治和伦理道德等社会秩序。

组织权益：影响组织自身或其他组织的生产运营、声誉形象、公信力、知识产权等组织权益。

用户权益：影响用户的基本权利，包含但不限于健康权、生命权、肖像权、名誉权、荣誉权、个人信息保护（知情权、决定权、查阅权、复制权、更正权、删除权等）、财产权、公平无歧视等用户权利。

6.3.2 影响程度

影响程度是指生成式人工智能产品和服务一旦遭到攻击、篡改、损毁、恶意使用或发生异常时，对影响对象可能造成的影响程度。影响程度从高到低可分为严重危害、较大危害、一般危害。对不同影响对象进行影响程度判断时，采取的基准不同。如果影响对象是国家安全或社会秩序，则以国家、社会或行业领域的整体利益作为判断影响程度的基准。如果影响对象仅是组织或用户权益，则以组织或用户的权益作为判断影响程度的基准。开展数据影响分析时，应按照以下规则确定影响程度：

a） 当影响对象是国家安全时：
   1） 如果对国家政权稳固、国家统一、民族团结及其他国家利益安全产生严重影响的，将影响程度确定为严重危害；
   2） 如果对国家政权稳固、国家统一、民族团结及其他国家根本利益产生较大影响的，将影响程度确定为较大危害；
   3） 如果对国家政权稳固、国家统一、民族团结及其他国家根本利益产生一般影响的，将影响程度确定为一般危害。
b） 当影响对象是社会秩序时：
   1） 如果对社会治安和公共安全等产生严重影响的，将影响程度确定为严重危害；
   2） 如果对社会治安和公共安全等产生较大影响的，将影响程度确定为较大危害；
   3） 如果对社会治安和公共安全等产生一般影响的，将影响程度确定为一般危害。
c） 当影响对象是组织权益时：
   1） 如果对组织自身或其他组织的生产运营等产生严重影响的，将影响程度确定为严重危害；
   2） 如果对组织自身或其他组织的生产运营等产生较大影响的，将影响程度确定为较大危害；
   3） 如果对组织自身或其他组织的生产运营等产生一般影响的，将影响程度确定为一般危害。
d） 当影响对象是用户权益时：
   1） 如果对用户的基本权利产生严重影响的，将影响程度确定为严重危害；
   2） 如果对用户的基本权利产生较大影响的，将影响程度确定为较大危害；
   3） 如果对用户的基本权利产生一般影响的，将影响程度确定为一般危害。

6.4 级别确定规则

结合分级风险要素识别情况，根据产品和服务对影响对象可能带来风险的影响程度，确定风险级别。风险级别与影响对象、影响程度的对应关系见表1。

表1　风险分级规则确定表

影响对象	影响程度		
	严重危害	较大危害	一般危害
国家安全	高风险	高风险	中风险
社会秩序	高风险	中风险	低风险
组织权益	中风险	低风险	低风险
用户权益	中风险	低风险	低风险
注：如待分级的产品和服务涉及多个影响对象或影响程度，应按照就高从严原则确定级别。			

可从如下步骤开展风险定级：

a） 满足以下任一条件的产品和服务，识别为高风险：

1） 产品和服务一旦产生安全问题，直接对国家安全造成严重或较大危害；

2） 产品和服务一旦产生安全问题，直接对社会秩序造成严重危害；

3） 产品和服务一旦产生安全问题，直接影响用户的生命、健康、财产等权益；

4） 经主管监管部门评估确定的高风险。

b） 未识别为高风险的产品和服务，满足以下任一条件的数据，识别为中风险：

1） 产品和服务一旦产生安全问题，直接对国家安全造成一般危害；

2） 产品和服务一旦产生安全问题，直接对社会秩序造成较大危害；

3） 产品和服务一旦产生安全问题，直接对组织/用户权益造成严重危害；

4） 经主管监管部门评估确定的中风险。

c） 未识别为高风险和中风险的产品和服务，识别为低风险。

参 考 文 献

[1] 《中华人民共和国民法典》，2020 年 6 月 1 日
[2] 《中华人民共和国个人信息保护法》，2021 年 8 月 20 日
[3] 《生成式人工智能服务管理暂行办法》，2023 年 07 月 10 日
[4] GB/T 43697—2024 数据安全技术 数据分类分级规则

电信终端产业协会团体标准

生成式人工智能产品和服务风险分类分级指南

T/TAF 309—2025

*

电信终端产业协会发布
地址：北京市西城区新街口外大街 28 号
电话：010-82052809
电子版发行网址：www.taf.org.cn