

TC260-PG-20269A

网络安全标准实践指南

——人工智能应用安全指引 教育

(v1.0-202609)



全国网络安全标准化技术委员会秘书处

2026 年 09 月

本文档可从以下网址获得：

www.tc260.org.cn/



全国网络安全标准化技术委员会
National Technical Committee 260 on Cybersecurity of SAC



前 言

《网络安全标准实践指南》（以下简称《实践指南》）是全国网络安全标准化技术委员会（以下简称“网安标委”）秘书处组织制定和发布的标准相关技术文件，旨在围绕网络安全法律法规政策、标准、网络安全热点和事件等主题，宣传网络安全相关标准及知识，提供标准化实践指引。

本文件起草单位：教育部科学技术与信息化司、教育部教育管理信息中心、中国电子技术标准化研究院、北京师范大学、北京邮电大学、中国科学院自动化研究所、中国劳动关系学院、清华大学、北京交通大学、华东师范大学。

本文件主要起草人：任昌山、陈冰雅、刘臻、杨伟平、孙强、吴子坚、张然、高峰、张妍婷、吕飞霄、童莉莉、祁彬斌、曾佳、李青、李兵、刘雨帆、文煜鑫、张小平、郎丛妍、周爱民、谈飞。



声 明

本《实践指南》版权属于网安标委秘书处，未经秘书处书面授权，不得以任何方式抄袭、翻译《实践指南》的任何部分。凡转载或引用本《实践指南》的观点、数据，请注明“来源：全国网络安全标准化技术委员会秘书处”。





摘 要

为全面提升各行业人工智能应用安全水平，以高水平安全保障人工智能高质量发展，依据《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》，以及《生成式人工智能服务管理暂行办法》《人工智能生成合成内容标识办法》《互联网信息服务算法推荐管理规定》《互联网信息服务深度合成管理规定》等政策法规，制定人工智能应用安全指引系列文件。

人工智能应用安全指引系列文件包含通用文件以及行业领域文件两类。通用文件提供通用性实践安全指导，适用于各行业领域开展人工智能应用活动。行业领域文件给出适用于特定行业领域的实践指引。

本《实践指南》属于行业领域文件，用于指导各级各类教育机构开展人工智能应用活动、提高安全能力，也用于指导教师、教育机构安全使用人工智能。本文件与《网络安全标准实践指南——人工智能应用安全指引 总则》等通用文件配合使用。



目 录

1 范围	1
2 规范性引用文件.....	1
3 术语定义.....	1
4 缩略语.....	2
5 教育通用安全指引.....	3
6 教育场景安全指引.....	4
6.1 智能助教.....	4
6.2 智能助学.....	6
6.3 智能助研.....	6
6.4 智能助管.....	7
6.5 智能助评.....	8
附 录 A （资料性） 教育人工智能应用场景.....	11





1 范围

本文件给出了教育通用安全指引、教育场景安全指引，为各级各类教育机构开展人工智能应用活动提供安全指导。

本文件适用于各级各类教育机构开展人工智能应用活动，也可供有关主管监管部门及第三方评估机构参考。

2 规范性引用文件

下列文件中的内容通过文中的规范性引用而构成本文件必不可少的条款。其中，注日期的引用文件，仅该日期对应的版本适用于本文件；不注日期的引用文件，其最新版本(包括所有的修改单)适用于本文件。

GB/T 25069-2022 信息安全技术 术语

GB/T 41867-2022 信息技术 人工智能 术语

GB 45438-2025 网络安全技术 人工智能生成合成内容标识方法

GB/T 45654-2025 网络安全技术 生成式人工智能服务安全基本要求

TC260-PG-20268A 网络安全标准实践指南——人工智能应用安全指引 总则

3 术语定义



TC260-PG-20268A《网络安全标准实践指南——人工智能应用安全指引 总则》附录 A 中界定的以及下列术语和定义适用于本文件。

3.1 教育机构 educational institutions

教育机构包括各级教育行政部门及其直属单位、各级各类学校。

3.2 教育数据 educational data

教育数据是教育机构在开展教育教学、管理和服务等相关活动中所产生的以电子方式对信息的记录。

3.3 智能助教 AI-assisted teaching

智能助教是利用人工智能技术为教师提供智能助手的应用场景。

3.4 智能助学 AI-assisted learning

智能助学是利用人工智能技术为学生提供智能助手的应用场景。

3.5 智能助研 AI-assisted research

智能助研是利用人工智能技术加速科研创新与技术突破的应用场景。

3.6 智能助管 AI-assisted management

智能助管是利用人工智能技术提升教育机构管理服务水平的应用场景。

3.7 智能助评 AI-assisted assessment

智能助评是利用人工智能技术提升教育评价水平的应用场景。

4 缩略语



TC260-PG-20268A《网络安全标准实践指南——人工智能应用安全指引 总则》附录 B 中的缩略语适用于本文件。

5 教育通用安全指引

教育机构在开展人工智能应用活动过程中，教育通用安全指引包括：

- a) 参考 TC260-PG-20268A《网络安全标准实践指南——人工智能应用安全指引 总则》，提升人工智能应用安全水平，防范人工智能应用安全风险，引导使用者在使用人工智能过程中规范使用行为、提升安全意识、保护自身的安全。
- b) 建立教育人工智能应用场景管理机制。根据使用者角色、年龄、学段、学科、任务类型明确使用边界、使用方式、成果要求与控制权限，实行白名单准入、算法备案、安全评估与数据分类分级管理，并对规则执行情况进行监督。
- c) 建立教育人工智能应用伦理安全审查机制。组建跨部门伦理安全审查小组并制定专项审查清单，审查内容覆盖人类福祉、公平公正、可控可信、透明可解释、责任可追溯、隐私保护等方面，并明确跟踪审查频次，确保教育人工智能应用始终服务于教育教学目标，坚持育人导向，尊重学生成长规律，保障学生身心健康，保护未成年人数据与隐私。



- d) 建立教育人工智能应用个人权益保障机制。人工智能应用坚持辅助角色定位，不能作为教学、评价、决策、研究的主体。人工智能生成的建议、判断或评价仅作参考，应经人工审核确认，不应直接替代教育机构作最终决定，并向相关人提供结果告知、异议提出、人工复核、结果更正与申诉处理渠道。
- e) 建立教育人工智能应用未成年人保护机制。面向未成年人提供人工智能服务时，应采用与其年龄和认知能力相适应的告知方式和默认保护设置，并依法取得监护人授权同意。未成年人使用人工智能辅助学习应全程在其监护人监护或教师监督下使用，并提供必要的使用时长、功能限制、监护设置和异常提醒措施。

6 教育场景安全指引

在第 5 章教育通用安全指引外，本章针对智能助教、智能助学、智能助研、智能助管和智能助评具体教育人工智能应用场景给出安全指引。同时涉及多个场景的，应结合实际功能同时符合相应场景要求。具体场景示例见附录 A，教育场景安全指引包括：

6.1 智能助教



智能助教场景安全指引如下：

- a) 教学活动中坚持以人为主导，人工智能作为辅助，教师把握人工智能介入教育教学的时机与频次，尽到监督指导职责，推荐可靠工具、设定使用边界，并培养学生甄别、批判与规范使用人工智能的能力，未经权威资料核验不应盲信。不应干扰、替代真实师生互动，不应直接以人工智能回答学生问题或提供咨询作为最终答案。
- b) 应提供访问控制和应急接管功能，支持按使用范围、管理层级采取实时干预和接管措施，同时支持学生主动终止个人会话。课堂教学、考试测评等连续性要求较高的教学环节发生异常时，应优先保障正常教学秩序，及时向教师或者教育机构管理员提示，请求人工接管或者转换为非人工智能教学方式。
- c) 应提供答案输出控制和人工复核功能，支持根据学段、学科、任务类型和学习目标设置答案输出层级，优先提供解题思路、方法提示和分步引导，避免在不适宜的场景中直接输出完整答案。使用人工智能辅助作业批改时，对于主观题、开放性任务应由人工复核确认。
- d) 应提供人工智能生成内容的解释说明功能，提供必要的引用来源、推荐理由和不确定性提示等，提升人工智能服务的透明度。进入课堂教学环节的人工智能生成内容应由人工进行审核，发



现连续错误输出、异常内容或者越权调用时，应暂停相关功能，经整改复核后方可重新启用。

6.2 智能助学

智能助学场景安全指引如下：

- a) 应按照学段采取人工智能应用的指导措施，小学阶段学生不应独自使用开放式内容生成功能，应在教师或监护人在场指导下使用；初中阶段可适度开展逻辑探究，高中阶段可结合原理探究使用；高等教育阶段应在任课教师规定范围内自主使用，研究生科研场景按 6.3 智能助研要求执行。
- b) 应提供使用时长、使用频率和使用时段管理功能，支持教师或监护人根据学生年龄、学段及教学需要设置连续使用时长、休息间隔和夜间使用限制，加强个性化推荐、拟人化互动和付费功能限制，并向教师或监护人提供使用情况统计、规则设置和异常提醒等监督工具。
- c) 应对学生学段、年龄、认知水平和特殊教育需求进行适配，人工智能服务生成的学习资源、学习建议和交互内容等应符合课程标准或者人才培养目标，仅限用于实现教育教学目的，无法确定内容适配性的，应提示教师或监护人进行复核。

6.3 智能助研



智能助研场景安全指引如下：

- a) 应限制人工智能辅助科研创新活动边界，仅可用于文献检索、资料整理、文字润色等常规辅助工作，研究选题、核心设计、数据解读、论点撰写等关键环节必须由研究者本人完成，并标注人工智能使用环节、提交过程性创作材料。严禁使用人工智能直接生成项目申报材料、论文或研究报告。
- b) 应对人工智能生成的实验模拟数据、实验操作建议、科研分析结果、模型验证结果，明确标注其数据属性，并与真实观测数据分开存储，经合理性、一致性和可重复性验证后方可使用，禁止利用人工智能伪造实验数据、捏造访谈、合成不实证据，或直接使用未经核实的人工智能生成内容。涉及民族、宗教、意识形态等敏感内容的研究，应按有关制度审查，避免人工智能的意识形态偏差与歧视、偏见。
- c) 应根据人工智能处理研究数据的来源、敏感程度和使用主体采取分级保护措施。涉及关键核心技术、未公开重大科研成果或者其他敏感科研数据的，优先采用安全可控的本地化或者私有化部署方式，并对接口调用、跨平台共享和数据出境风险进行评估。未成年人教学研究和学生课题数据应根据学生年龄、数据类型和研究目的采取特殊保护。

6.4 智能助管



智能助管场景安全指引如下：

- a) 应明确人工智能辅助教育管理的使用范围，教育机构使用人工智能仅限用于学情分析、教情管理、心理健康促进、危机预警与干预、学业困难帮扶、校园安全以及其他经明确论证的教育管理活动。学生管理、学籍管理、人事管理、资源分配及学位授予等重大管理处置，不应直接执行人工智能结论，需经教育机构人工决议。
- b) 应对人工智能服务监测预警、数字画像、异常识别、行为轨迹类功能进行必要性审查，预警输出应附依据与置信度，不应作负面定性，并应标注人工智能参与，由教育机构人工审核。数字画像仅作为辅助用途，不应影响相关人利益。被预警、被处置对象应有人工复核与纠错撤回机制，防止标签长期伴随。
- c) 应对人工智能服务输入输出内容进行分析，在检测到输入输出内容涉及人身伤害、违法违规、行为诱导、心理操纵、极端意识形态、网络欺凌、人身攻击或者其他可能危及学生身心健康的高风险行为和倾向时，人工智能应用应根据风险程度采取限制回答、暂停交互、提供安全引导等应急处置措施，并向教师、监护人或者教育机构发出风险提示。涉及生命财产安全风险的，应及时联动具有相应职能部门进行处置。

6.5 智能助评



智能助评场景安全指引如下：

- a) 应根据评测、评价事项开展算法安全评估，重点检测算法滥用、算法漏洞、算法恶意利用及其他风险，在不同群体之间是否存在缺乏合理依据的系统性差异，并评估结果的稳定性、准确性和可解释性。禁止用于对个人进行负面评价，如固化标签、歧视性对待，或者作为处分、淘汰、升学限制、资格取消以及其他惩罚性、排斥性决定。发现可能造成不公平影响的，应采取干预策略、暂停使用或者其他纠正措施。
- b) 应保障被评测、评价对象的知情权和选择权，确保人工智能服务的透明度，明确告知数据来源、算法、标准、结果适用范围及其局限性等，取得被评测、评价对象的授权同意，并保留可追溯依据。如被评测、评价对象拒绝人工智能服务，或对结果质疑，应提供申诉渠道，采用相同方法进行人工复核，并以人工复核结果为准。
- c) 应对人工智能服务做出的主观成果评价进行人工复核，如作文、艺术作品、开放性作业、课程论文、报告等主观类成果的自动批改结果不应直接作为最终评价，应由人工结合真实观察修订，评价结果不应套用模板或空洞格式化。
- d) 应对人工智能生成的评价结果按比例进行抽样审核，如升学推荐、奖学金评定、学位授予、学生分类分班、绩效考核、职称评审以及其他对个人权益有重大影响的评价，应由审核人员对



全国网络安全标准化技术委员会
National Technical Committee 260 on Cybersecurity of SAC

人工智能评价结果进行批准、否决或调整，并作出相应解释说明。



全国网络安全标准化技术委员会
National Technical Committee 260 on Cybersecurity of SAC



附录 A

(资料性)

教育人工智能应用场景

本附录给出教育人工智能应用的典型场景、主要功能、主要风险和对应管控条款，包括智能助教、智能助学、智能助研、智能助管和智能助评。同一人工智能应用可能同时涉及多个场景，结合实际功能同时参照相应条款。

A.1 智能助教

人工智能应用可作为教学辅助工具参与备课、教学资源生成、课堂互动、答疑、作业批改和教学分析等活动，支持教师减少重复性工作、改进教学组织。智能助教不替代教师开展教学设计、课堂教学和教学决策。典型场景如下：

- a) 虚拟助教与答疑：为师生提供知识点讲解、问题解答和学习提示，支持教师减少重复答疑工作。可支持提高常见问题反馈效率。主要风险包括事实性错误、内容不适配、直接输出完整答案和学生过度依赖。
- b) 作业批改辅助：对客观题提供自动批改，对主观题、开放性任务提供初步分析、反馈建议或者评分参考，最终结果由教师根据任务性质复核确认。可支持减少机械性批改工作。主要风险包括评分偏差、反馈失当、评价依据不透明和生成内容错误。
- c) 个性化学习路径规划：根据学生知识掌握情况、学习进度和学习目标，为教师提供资源组合、教学顺序和分层指导建议。可



支持发现学习薄弱环节和优化教学安排。主要风险包括学生画像固化、推荐偏差、过度收集学习数据和错误路径引导。

- d) 教学行为分析：分析课堂互动、学习参与和作业完成等信息，为教师提供教学观察和改进建议。可支持发现课堂组织中的异常情况。主要风险包括过度监测、行为标签化、将短期数据用于不当评价以及侵犯学生隐私。
- e) 教案、课件与教学资源生成：辅助教师生成教案初稿、课件素材、试题、案例和课堂活动方案。可支持提高教学准备效率。主要风险包括知识错误、版权或者引用问题、不适宜内容进入课堂和教师过度依赖。
- f) 虚拟课堂与课堂互动辅助：在实体课堂或者线上课堂中提供实时问答、课堂互动、字幕翻译、内容总结和学习反馈。可支持扩大课堂参与和远程教学覆盖。主要风险包括服务中断、异常内容实时传播、学生身份与互动数据泄露以及人工智能替代教师主导。

A.2 智能助学

人工智能应用可通过学习内容交互、学习数据分析和资源推荐等方式，为学生提供自主学习、差异化学习和终身学习支持。相关应用应根据学生年龄、学段、认知水平和特殊教育需要设置差异化保护。典型场景如下：

- a) 学习辅助与自主学习平台：为学生提供习题讲解、模拟练习、自主答题反馈和学习过程跟踪。包含拟人化角色或者持续性情感互动功能时，同时属于拟人化互动场景。可支持学生开展自



主学习。主要风险包括不适宜内容、错误答案、过度使用、情感依赖和个人信息过度收集。

- b) 个性化学习资源推荐：根据学生的知识掌握情况、学习进度、能力水平、学习目标和个人兴趣推荐课程、练习和学习材料。可支持资源与学习需求匹配。主要风险包括推荐偏差、信息范围收窄、标签固化和不透明画像。
- c) 语言学习与交流辅助：通过口语练习、写作反馈、多语种翻译和情境对话等方式支持语言学习。可支持增加练习机会和即时反馈。主要风险包括错误翻译、不适宜对话内容、文化偏差、声音等个人信息处理风险。
- d) 虚拟实验与仿真实训：构建模拟实验、虚拟实训和操作演练环境，支持科学实践和职业技能培养。可支持降低部分真实操作的训练成本和初次操作风险。主要风险包括错误实验建议、仿真结果被误认为真实数据、危险操作迁移到现实环境和设备权限失控。
- e) 学习发展档案：持续记录和分析学习经历、学习成果和能力发展信息，为学习指导和成果认证提供参考。可支持形成连续性学习记录。主要风险包括长期画像、数据用途扩张、评价标签固化、信息泄露和难以更正。
- f) 在线培训与职业技能提升：结合课程目标和岗位能力要求，推荐或者生成培训内容，支持职业技能提升和继续教育。可支持提高培训内容匹配度。主要风险包括错误能力判断、岗位推荐偏差、培训内容质量不稳定和个人职业数据泄露。



- g) 社会化知识服务：通过智能问答、知识共享和社群学习等方式提供面向公众的学习支持。可支持扩大教育资源覆盖。主要风险包括虚假信息传播、不良内容扩散、用户身份和互动数据泄露以及线上平台责任边界不清。
- h) 生涯规划与职业发展辅助：根据学习经历、能力特征和职业信息，为学生提供专业选择、升学路径和职业发展建议。可支持学生了解多样化发展路径。主要风险包括推荐偏差、依据不透明、对学生形成固化判断以及将建议误作为确定性结论。

A.3 智能助研

人工智能应用可辅助科研人员和学生开展文献检索、数据分析、研究设计、成果表达和科研诚信风险识别。相关应用不替代研究者的专业判断和科研责任。典型场景如下：

- a) 文献检索与知识关联：检索、梳理和关联学科知识、文献资料与前沿成果，辅助开展课题探索。可支持提高资料发现和整理效率。主要风险包括虚假文献、引用错误、来源不可核验和未公开研究信息泄露。
- b) 研究数据分析与报告辅助：辅助开展数据清理、统计分析、建模、结果可视化和科研报告初步撰写。可支持提高数据处理和成果表达效率。主要风险包括分析方法错误、数据泄露、伪造数据、结论夸大和人工智能使用不透明。
- c) 科研方案与实验设计辅助：辅助形成研究问题、实验方案、技术路线和操作建议。可支持研究人员比较不同方案。主要风险



包括危险实验建议、忽略伦理和审批要求、智能体越权操作以及错误建议进入真实实验。

- d) 学术诚信检测与风险提示：识别重复内容、异常数据、引用缺失和其他可能的学术不端风险，为使用者提供核查线索。可支持提高问题发现效率。主要风险包括误报、漏报、检测规则不透明和将检测结果直接作为处分依据。
- e) 科研表达与线上答辩辅助：辅助生成汇报提纲、演示材料、答辩问题和模拟反馈，支持线上或者线下科研交流。可支持研究者完善表达和准备答辩。主要风险包括代写代答、泄露未公开成果、生成错误论证和隐藏人工智能实质性使用。
- f) 区域发展与社会服务研究辅助：辅助开展地方产业调研、区域经济数据分析、乡村振兴和教育公共服务研究。可支持高校服务地方经济社会发展。主要风险包括政务或者产业敏感数据泄露、分析偏差、跨部门数据越权使用和研究结论被不当用于决策。

A.4 智能助管

人工智能应用可辅助教育机构开展教学科研管理、校务服务、校园安全、学生支持、设施运行和跨机构协同。涉及个人重大权益、敏感个人信息或者公共安全的管理活动，属于重点风险关注场景。典型场景如下：

- a) 排课与资源调度：辅助生成课表，配置教室、设备、师资和其他教育资源。可支持减少排课冲突和提高资源利用效率。主要



风险包括规则设置错误、资源分配不公平、系统故障影响教学和自动结果缺少人工复核。

- b) 校园安全风险感知与应急辅助：分析门禁、视频、设备和事件信息，为校园安全预警和应急响应提供辅助。可支持提高异常发现和汇总效率。主要风险包括人脸等生物识别信息滥用、误识别、过度监控、权限越级和错误预警。
- c) 学生成长支持与风险预警：综合学习、行为等信息识别可能需要学业帮扶或者成长支持的学生，为教师和教育机构提供参考。可支持较早发现需要帮助的情况。主要风险包括标签化、歧视性处理、数据用途扩张、错误预警和将帮扶结果转用于惩罚。
- d) 心理健康教育辅助与危机转介：提供心理健康知识、一般性情绪支持和风险线索识别，并在发现可能危及人身安全的情况时转介具有相应职责和专业能力的人员。人工智能应用不替代专业人员开展诊断、治疗或者危机处置。可支持提高一般性知识服务和风险线索发现效率。主要风险包括错误判断、延误专业干预、敏感心理数据泄露和学生产生情感依赖。
- e) 校务公开与智慧服务：辅助开展招生咨询、就业服务、校友联络、财务报销、办事问答和信息公开等校务服务。可支持提高服务响应效率。主要风险包括个人信息泄露、回答错误、权限越级、服务歧视和自动建议被误作正式决定。
- f) 校园能源与设施运行管理：辅助分析水、电、暖、设备和建筑运行数据，支持节能调度、故障预警和绿色校园建设。可支持



减少能源浪费和提高设施维护效率。主要风险包括控制指令错误、设备权限过大、运行数据泄露和系统故障影响校园运行。

- g) 跨校协同与资源共享：支持学校之间开展课程、师资、科研资源、平台能力和教育数据的协同共享。可支持扩大优质资源覆盖。主要风险包括共享范围不清、访问权限失控、数据跨机构流转风险和责任边界不明确。

A.5 智能助评

人工智能应用可辅助开展学生、教师、课程教材、学校办学和区域教育质量测评，为评估主体提供数据分析和反馈支持。人工智能生成的测评结果不直接替代具有相应职责和权限的主体作出最终测评决定。典型场景如下：

- a) 学生综合素质测评辅助：分析学业表现、学习行为和活动参与等信息，为教师和教育机构提供综合素质评价参考。可支持汇总多来源测评证据。主要风险包括固化标签、群体歧视、唯数据、评价黑箱和对学生重大权益产生不当影响。
- b) 教学质量测评辅助：分析课堂互动、教学设计、学生反馈和教学成效等信息，为教师改进教学和专业发展提供参考。可支持发现教学改进线索。主要风险包括将单一数据用于教师排名或者绩效决定、算法偏差、评价依据不透明和监测范围过度。
- c) 课程与教材效果评估：分析课程使用、学习成效和使用反馈等信息，为课程内容和教材改进提供参考。可支持持续发现内容和学习需求之间的差异。主要风险包括数据代表性不足、将相关性误作因果关系、评价指标单一和生成结论错误。



- d) 学校办学质量测评辅助：分析人才培养、教育教学、资源配置和发展成效等信息，为学校自评、外部测评和改进提供参考。可支持整合多维度办学信息。主要风险包括唯指标化、学校标签化、数据口径不一致、评价模型偏差和评价结果被不当用于资源配置。
- e) 区域教育质量监测辅助：分析多校、多区域教育数据，为教育行政部门开展质量监测、精准帮扶和政策研究提供参考，可支持发现区域性差异和资源需求。主要风险包括数据跨区域共享风险、群体标签化、算法偏差、指标口径不一致和将监测结果直接用于惩罚性排名。