# 人工智能语境下未成年人模式的困境与出路

**陈惠琳[1]，张　浩[1]，龙玉竹[1]**

(1.西华大学，四川 成都 610039)

**摘　要：**在生成式人工智能深度嵌入未成年人日常生活的背景下，未成年人模式的实际保护效能面临考验。本文通过机制对比与 369 份用户反馈考察发现当前模式存在分龄管理缺位、外部监管乏力等多重困境，保护实效未达预期。2026 年 7 月施行的《人工智能拟人化互动服务管理暂行办法》将规制触角延伸至拟人化互动服务，但重心限于特定业态，通用型 AIGC 工具的未成年人模式仍缺乏细化标准。据此，应突破限制型保护逻辑，在完善法律体系、技术生态重构等方面协同发力，推动构建与 AIGC 特性相适配的未成年人模式。

**关键词：**未成年人模式；未成年人网络保护；生成式人工智能

**基金项目：**四川省省级大学生 2025 年创新创业训练计划 创新训练项目“生成式人工智能语境下未成年人网络保护机制研究——基于青少年模式法律规制的实证考察”（项目编号：S202510623085）

DOI：doi.org/10.70693/202608292723

## 一、引言

生成式人工智能技术正深刻重塑未成年人在互联网的学习、社交与娱乐图景。据第 57 次《中国互联网络发展状况统计报告》，截至 2025 年 12 月，生成式人工智能产品用户规模已达 2.49 亿，19 岁及以下用户占比 26.4%，在各年龄段居首[1]。同年中国青少年研究中心调查揭示，超七成未成年人用 AI 辅助完成作业，逾三成与 AI 聊天或用于娱乐[2]。然而未成年人身心发育尚未成熟，辨别能力较弱。生成式 AI 在带来便利的同时，也潜藏不良信息接触、诱发心理依赖等新型风险。国外已发生几起 AI 聊天机器人被指控教唆青少年自残等恶性事件，为我国敲响了警钟。

国家网信办自 2019 年起推行青少年模式，近年向未成年人模式转型，现成为 AIGC 领域等各类平台的标配功能；但其实效未达预期：用户信任度低、家长斥其形同虚设等现象在传统领域屡见不鲜[3]。2026 年 7 月施行的《人工智能拟人化互动服务管理暂行办法》（以下简称《办法》）明确禁止向未成年人提供虚拟亲密关系服务，并要求建立未成年人模式等机制。施行前夕，多家平台下线智能体功能，印证了未成年人模式在 AIGC 场景下存在实效困境。《办法》主要针对拟人化互动业态，通用型 AIGC 工具的未成年人模式仍缺乏细化的执行标准。

现有研究从多个维度进行了探讨，但多从局部环节切入：规制探讨偏重理念阐述，缺乏对现行模式实效的检验；实证研究聚焦传统平台，对 AIGC 业态的覆盖不够充分；对《办法》这一最新监管动态尚未形成系统回应。鉴于此，本文通过考察未成年人模式在 AIGC 语境下的运行实效并提出相应优化路径，以期为完善 AIGC 时代的未成年人网络保护机制提供数据参照与学理思考。

## 二、AIGC 语境下“未成年人模式”实效分析

### （一）主流 AIGC 应用未成年人模式的实施现状

作者参考 QuestMobile《2025 年春季中国移动互联网报告》、七麦数据 2025 年 6 月应用下载量排行及《中国未成年人互联网运用报告（2025）》所涉 AI 应用等数据，兼顾市场占有率、用户活跃度与类型范畴，并纳入曾

**作者简介：**陈惠琳（2005—），女，法学院本科在读，研究方向为民法学；
张　浩（2005—），男，法学院本科在读，研究方向为行政法学；
龙玉竹（2005—），女，法学院本科在读，研究方向为民法学。

因内容合规问题被监管部门约谈的应用以增强风险代表性[4]，选取了 10 款应用作为分析样本①，依托《中华人民共和国未成年人保护法》等规定，从静态设置与动态运行两个维度进行对比以考察未成年人模式在生成式人工智能领域的运行实效。

表 1 10 款应用在未成年人模式方面的设计差异

| 名称 | 模式触发方式 | 防护措施 | 设计逻辑 |
| --- | --- | --- | --- |
| 元宝、DeepSeek、天工、千问、文心、Kimi | 无 | 自动过滤有害信息，支持举报；对自伤、违法活动等危险话题引导至专业帮助渠道。 | 对话层实时拒绝，依赖底层模型安全策略。 |
| 豆包 | 需自行在设置中激活并设密码。 | 严格过滤信息；<br>限制核心功能，仅保留拍题答疑等正向服务。 | 降级为答疑助手，功能牺牲换取安全。 |
| 筑梦岛 | 弹窗提示可跳过或在设置页激活；需设独立密码，无实名验证。 | 1. 每日限用 40 分钟，每 15 分钟休息提醒<br>2. 单次充值≤50 元, 月累计≤200 元。<br>3. 按 8 岁以下、9-15 岁、16-18 岁设置时长与内容。 | 防沉迷与分龄管控相结合。 |
| 猫箱 | 首次弹窗确认年龄且不可关闭：须实名认证并设长期密码。 | 屏蔽成人向情节，关闭创作和评论功能，仅保留系统特定科普性角色。 | 入口身份绑定，自动干预。 |
| 星野 | 弹窗提示可跳过，自主申报实名认证进入模式。 | 屏蔽敏感情节；关闭创作等社交功能；部分版本有时间及消费限制。 | 以实名认证为基础，综合内容屏蔽与功能限制。 |

### 1、静态机制

#### (1) 模式设置

《未成年人网络保护条例》第四十三条明确设置独立醒目的未成年人模式是网络服务提供者的法定义务。如表 1 所示，仅四成应用设计独立模式入口，余六成单纯依赖后端模型过滤。该分化表明行业内对未成年人模式应以显性入口还是隐性算法为载体尚未形成共识。

#### (2) 防护措施

《中华人民共和国未成年人保护法》第七十四条规定网络服务提供者应针对未成年人使用其服务设置相应的时间管理、权限管理、消费管理等功能。《移动互联网未成年人模式建设指南》（以下简称《指南》）要求未成年人模式允许用户对每日上网时长进行总量限制并提出分龄推荐标准。各平台防护措施可归纳为三种策略，成本投入与保护效果相应递减。综合管控型以筑梦岛为典型，因强娱乐和付费属性而集成时间、消费及分龄管控等多重手段，合规最为全面。功能简化型以豆包为代表，开启后即关闭 AI 创作等功能，合规彻底但抑制了合理需求。内容拦截型涵盖所有未设独立模式的应用，依托对话层实时过滤，效果受限于算法识别精度，在时间、消费及分龄内容管理上均付之阙如，依赖用户自我约束。

### 2、动态运行

#### (1) 模式开启

① 10 款应用包括：元宝、Deepseek、豆包、筑梦岛、星野、kimi、猫箱、文心、千问、天工，样本选择侧重类型覆盖而非完全随机抽样，结论主要反映上述产品的规律性特征。

未设独立模式者不存在“开启”操作，用户对保护机制几乎无感知。设有独立模式者开启路径也各有障碍：豆包需自行进入设置页；筑梦岛、星野弹窗引导但可跳过；仅猫箱采取强制弹窗并绑定实名认证，管控强度最高。既有研究指出开启环节被动化是传统领域未成年人模式的共性短板[5]。在 AIGC 语境下，这一问题反而因部分平台取消独立模式而加剧。

(2) 模式退出

设有独立模式者多以密码验证作为退出条件，部分平台输入次数未设限，客观上降低了暴力解除的门槛。密码重置流程虽严却可被规避：猫箱、豆包、筑梦岛在用户忘记密码时要求通过申诉重置，一般需家长提供手持身份证照片等材料进行人工审核，验证门槛较高。然而游客模式登录或切换成年人账号亦可绕过限制。

(3) 设计逻辑

上述应用的设计逻辑可归为两种范式。未设置独立模式者属后端内容安全型，将保护窄化为“内容过滤”，法定功能覆盖面缺失；其余归于前端功能限制型，通过显性管控限制特定功能、使用时长等，保护效果依赖监护人的管理能力，且在分龄适配和用户体验上仍有改进空间。

### （二）未成年人用户使用反馈

为弥补供给侧机制分析在需求侧视角的不足，作者对成都市使用过生成式 AI 对话的未成年人进行了一轮小规模交流考察，系小范围现状摸底，目的是发现潜在问题，共获有效反馈记录 369 份，受访者集中于 12 至 18 岁。

#### 1、深度使用与情感索取需求并存

未成年人对生成式 AI 的使用需求呈现工具性与情感性并存的特征，逾七成受访者将 AI 用于完成日常任务，而部分使用者的行为呈现从工具性向关系性延伸的倾向。值得注意的是，部分未成年人向 AI 寻求情感慰藉的动因并非单纯猎奇心理，而是如家庭沟通、学校心理资源等现实支持体系失灵，过半数受访者表示因害怕被评判等顾虑而向 AI 倾诉，折射出其社交与情感需求在现实层面未获满足（见表 2）。

**表 2 未成年人向生成式 AI 倾诉的原因**

| 类型 | 人数 | 占比 |
|---|---|---|
| AI 不会评判用户 | 179 | 48.51% |
| 现实中缺乏合适的倾诉对象 | 198 | 53.66% |
| 好奇心理驱动 | 217 | 58.81% |
| 其他原因（如测试不同 AI 的差别、单纯用于娱乐） | 64 | 17.34% |

#### 2、模式体验呈现限制与需求的错位

关闭诱因的分布揭示了保护机制与用户需求之间的错位，关闭或回避模式的主因是模式过度限制、正常使用需求无法得到满足（见表 3）。就外部监管而言，近四成反馈表明，监护人的监管止于口头提醒等柔性措施，仅三成表示监护人会主动开启模式。同时，过半数受访者表示明确知晓规避限制的方法并会主动使用，印证了现行模式在防绕过方面存在脆弱性。

**表 3 关闭模式的诱因**

| 原因 | 人数 | 占比 |
|---|---|---|
| 模式限制太严格，影响正常使用体验 | 245 | 66.40% |
| 功能单调枯燥，缺乏个性化设计 | 217 | 58.81% |
| 其他原因（如与监护人一同使用无需开启、单纯不感兴趣） | 136 | 36.86% |

#### 3、法律与风险意识滞后

直接询问“是否了解相关法律”易诱发称许性回答，而将其嵌入具体情境更能反映未成年人对行为边界的判

断水平，故作者设置了一道要求受访者识别正确说法的情景化多选题，正确答案为 B 项及 D 项。一定比例受访者未能准确识其中有明确违法性的行为（见表 4）。对 AI 诱导自伤、低俗对话等风险事件的知晓程度亦不足，仅三成受访者了解相关事件。由此可见，当技术尝试触及法律或伦理红线时，部分未成年人缺乏清晰的判断标准。

表 4 法律情景判断情况

| 选项 | 人数 | 占比 |
|---|---|---|
| A：甲未经同意利用 AI 生成他人肖像的搞笑视频，不涉嫌违法 | 169 | 45.80% |
| B：乙经同意获取他人照片，利用 AI 生成涉黄图像进行传播，此举违法。 | 255 | 69.11% |
| C：某平台的 AI 模型与未成年人聊天时教唆自残，该平台无需担责。 | 142 | 38.48% |
| D：某杂志社未经授权采集真人写真，利用 AI 修改成插画风格并刊登，此举违法。 | 185 | 50.14% |

## 三、现行模式的困境

当前模式存在的弊病不能简单归因于个别平台执行偏差，而是既有治理思路与人工智能应用环境适配度低。

### （一）防护机制漏洞与分龄体系建构迟滞

AIGC 语境下的未成年人模式始终遵循限制型保护逻辑，即通过身份识别、权限设置与家长赋能三个环节，将未成年人与网络风险进行物理隔离[6]。此举在传统领域尚可奏效，但移植到内容动态生成、交互高度开放的生成式人工智能环境便面临不适配性。现有模式沿袭了传统领域的固有缺陷：开启环节主要以主动操作为前提，呈现被动化特征；退出机制防绕开能力不足，管控密码机制易被设备借用等方式突破；身份验证依靠用户自主申报或算法画像，精确度有限。上述问题共同催生了规避现状，使模式在实际运行中趋于形式化。同时生成式人工智能基于概率模型实时构造内容，输出有不可预测性，用户可通过精心设计的提示词或迭代对话规避内容过滤，如使用隐喻等手法，印证了单纯依赖后端防护的内在局限。

更为突出的是，当前模式普遍欠缺分龄治理思维，功能供给呈现“重管控、轻匹配”的取向，有助于降低内容风险，却致使模式使用黏性减弱。分龄管理缺位实为诱发规避行为之症结所系，并使保护实践陷入过度干预与保护真空并存的矛盾现状。对心智发育与媒介素养已超越模式预设的未成年人而言，一刀切式管控无异于变相排斥，当使用需求在模式内无法获得适龄化探索余地时，主动绕开限制便成为具备合理性的行为反应。

### （二）行业标准缺失与市场激励错位

各平台防护思路不统一表明行业明显缺乏统一标准，如未能针对 AI 业态类型设定分级要求，致使用户可从限制严格的平台无缝迁移至相对宽松者，保护效果被稀释。此外，商业利益与保护责任的内生矛盾持续存在。平台模型迭代依赖用户活跃度与数据积累，而未成年人模式恰会抑制这些指标。既有研究指出未成年人模式难以嵌入盈利机制，平台因而缺乏持续投入的激励[7]。严格的模式设计会削弱产品吸引力，进而拖累整体数据规模，这驱使平台在设计和执行上留有余地。

### （三）法律规制覆盖缺失与监管失衡

专门性规范存在覆盖空白与归责指引模糊。《办法》主要规制拟人化互动服务，而通用型 AI 工具的模式设计依赖《未成年人网络保护条例》等多为原则性表述的既有规范，未就 AIGC 场景下模式应包含的核心功能给出具体答案。制度供给的模糊使平台在合规选择上缺乏明确指引，进而采取低成本的举措。《民法典》构建了平台与用户的责任结构，而人工智能背景下裂变为模型开发者、技术服务提供者、应用运营方与用户的多主体交织的责任链条。当 AI 生成内容诱导未成年人自伤或实施违法行为时，责任归属于模型训练者、内容审核方还是应用运营方，现有规范未就各主体间责任边界作出清晰划分。归责的模糊状态为各方相互推诿预留了空间，最终使保护义务难以落实。

惩戒力度不足淡化平台改善意愿。既有监管实践对 AIGC 场景下模式失范的处罚仍属少数，并主要依赖专项整治等阶段性手段。模式失范对平台引发的主要后果是舆论风险，而非法律层面的实质性负担，由此形成了低成

本的违规生态。在此制度环境下，平台倾向于以形式合规替代实质保护，在合规边缘维持最低限度的投入，乃至回避未成年人模式的独立建设。

### （四）协同联动失效与风险教育脱节

现行模式在 AIGC 语境下成效有限，还受制于家庭素养、学校教育、社会履责三重短板。

家庭端监管介入不足。既有研究证实，家庭关系越不理想，未成年人在网络交往中越易感到孤独，生成式 AI 凭借算法内嵌的情绪识别系统能在某种程度上填补家庭沟通、朋辈陪伴乃至专业心理辅导的空白，但其程式化交互缺乏真实情感联结，青少年易形成单向情感依赖，长期如此可能引发现实社交退缩与认知偏差[8]。在此情形下，监护人的数字素养若未能并行跟进，便难以有效引导未成年人的 AI 使用行为。

未成年人对 AI 法律边界与新型风险认知模糊，成因或在于学校未将有关 AI 伦理、法律风险与心理健康教育系统嵌入课程体系，社会亦未形成有效的补充供给。更深层的问题在于家庭、学校与平台三方缺乏系统联动。平台端未及时将用户行为反馈至家校，学校缺乏判断机制，家庭端的监管水平参差不齐，多重因素叠加使模式实效式微。

## 四、现行模式的优化进路

### （一）健全法律体系，构建刚性约束框架

提升专门性立法的效力层级。现行《生成式人工智能服务管理暂行办法》对数据来源和基础模型的规定过于原则，未能系统覆盖语料采集、审核及使用全过程；相关安全技术标准多属推荐性国家标准，法律强制力有限；《办法》作为部门规章，裁判适用时易产生效力争议。2026 年政府工作报告明确提出“完善人工智能治理”，作者建议推动制定《人工智能法》以建立统一框架。在此之前，可先行出台行政法规层级的《生成式人工智能服务未成年人模式管理条例》，推动既有框架在 AIGC 场景下实现细化落地，健全行业标准、压实企业自律责任。立法亦应避免简单禁止的倾向，《办法》中诱导沉迷、情感操纵等概念的内涵与外延尚待厘清，难以准确区分正常互动与不当干预，亦或催生平台分体运营规避监管。事实上，智能体下架所引发的用户投诉及虚拟财产诉求验证相关措施未能根本回应陪伴需求，从长远看并不利于未成年人综合发展。

此外，生成式人工智能提供服务者应向监护人展示信息来源、推荐逻辑与潜在风险，定期披露模式运行报告，涵盖功能改进、风险干预触发次数及第三方评估，以可见的治理成效换取用户信任。违反防沉迷或防依赖义务的，网信部门可依据情节给予暂停业务、罚款等处罚；贯彻“首违不罚”原则，辅以约谈、黑名单等柔性措施，健全举报平台[9]。当 AI 生成内容诱导未成年人自伤或实施违法行为时，应在司法解释中明确各环节的责任边界：语料提供者承担源头审查义务，宜采用无过错责任；模型训练者在故意或重大过失情形下担责；服务提供者应履行主动管控义务，重点监控高风险模型，而非仅适用“通知—删除”规则[10]。对于未成年人使用诱导性提示词引发的侵权，应考虑前端算法偏差等因素，适用共同侵权与连带责任制度，由司法裁判合理分配各方责任[11][12]。

### （二）重建技术生态，算法调适与分龄体系精准赋能

平台应依法采集未成年人实名信息，并结合交互内容、使用记录等信息生成用户画像，以适龄设计为导向，依据不同年龄段的心智水平与发展需求，在模式的功能供给、视觉形态、交互风格上进行定向适配[13]。参照《指南》等既有框架可建立三级体系：12 岁以下严格限制功能与时长，提供基础科普内容；12 至 15 岁强调引导，开放启发式交互；16 至 18 岁侧重赋能，增设风险话题即时警示，依托行为轨迹与兴趣数据动态拓展功能。分龄体系的要义在于以差异化的权限梯度替代功能阉割，提升模式设计的吸引力与适配性，从源头消解规避动机。此外，可引入成就徽章等游戏化激励机制，化解模式内容枯燥、功能受限引发的抵触情绪，也有助于平台协调商业利益与社会责任的张力，并建设支持监护人弹性调整权限的智能模块，增强模式的可接受性。

生成式 AI 输出的不可预测性使后端过滤存在天然突破点，单纯提升识别精度难以弥补内生局限。平台应在前端交互层动态监测用户与 AI 的交互模式，具体而言应以对话时长、话题偏移频次、情绪表达强度、深夜使用频率等行为指标为监测维度，当交互模式偏离正常范围时，如捕捉到用户反复倾诉负面情绪、呈现自伤倾向等即

自动触发干预：轻度风险推送心理援助资源，高危信号启动与专业机构的转介通道。

### （三）构建协同教育防线，织密保护网络

强化监护人职责，是确保模式有效触发的首道关口。应促使监护人了解 AI 技术基础逻辑与潜在隐患以建立陪伴式引导的能力基础，而非单纯诉诸技术被动拦截。学校作为连结支点，应将人工智能素养教育嵌入课程体系，借助跨学科实践与 AI 技术革新教学形态，指导学生理解信息甄别、算法逻辑及新型风险。社会主体应整合区域内公共及科研资源，以讲座、工作坊等形式提供补充性支持。为实现多元主体在保护网络的有效联动，未成年人模式应增设功能使用统计与实时监控模块，并建立信息互通机制：平台一旦监测到异常行为，即向家校推送预警，形成保护闭环。

未成年人网络保护作为一项系统工程，需要法律、技术、教育等方面的协同发力。唯有在发展中保护、在保护中发展，才能真正为下一代构筑清朗、安全且富有成长性的数字家园。

**参考文献：**

[1] 中国互联网络信息中心. 第 57 次《中国互联网络发展状况统计报告》[R/OL]. 2026-02-05. https://www.cnnic.cn/n4/2026/0304/c88-11549.html
[2] 孙宏艳. 调查发现超六成受访中小学生用过 AI[N/OL]. 中国青年报, 2026-03-26(08)[2026-07-27]. http://sc.people.com.cn/n2/2026/0326/c345515-41534430.html .
[3] 季为民, 杨子函.推动 AI 精准陪伴赋能, 促进未成年人互联网智能化成长[M].范承志,季为民,沈杰,叶俊,杨斌艳,季琳.中国未成年人互联网运用报告（2025）.北京:社会科学文献出版社.2025 年:1-16.
[4] 孙文轩, 冯仪. AI 聊天软件调查:擦边内容尺度大 青少年模式仍存漏洞[N/OL]. 央广网. (2025-06-13)[2026-07-27]. https://edu.cnr.cn/dj/20250613/t20250613_527208239.shtml
[5] 谭凯月. 网络青少年模式的运行困境与优化路径——基于 20 款 APP 青少年模式实证研究[J].遵义师范学院学报,2024,26(1):51-56.
[6] 刘晓春. 从限制到调控：人工智能背景下未成年人保护模式的转型与重塑[J].财经法学,2025,(5):115-130.
[7] 方增泉,吴京京,杨舒逸.移动互联网未成年人模式发展定位、存在问题与创新路径[J].社会治理,2026,(2):68-83.
[8] 叶俊，时瑞泽，曹雨欣.AI 技术对未成年人信息获取及认知的影响[M].范承志,季为民,沈杰,叶俊,杨斌艳,季琳.中国未成年人互联网运用报告（2025）.北京:社会科学文献出版社.2025 年:107-123.
[9] 李岩,康铭. 生成式人工智能防沉迷立法问题研究[J].东北师大学报(哲学社会科学版),2024,(2):112-126.
[10] 王利明.生成式人工智能侵权的法律应对[J].中国应用法学,2023,(5):27-38.
[11] 李昕梦.生成式人工智能版权注意义务的界定[J].数字法治,2025(5):156-169.
[12] 邓建鹏.生成式人工智能侵权谁负责？[EB/OL].(2025-09-30).https://cssn.cn/skgz/bwyc/202509/t20250930_5917880.shtml
[13] 叶俊，张译文.未成年人使用人工智能设备研究[M].范承志,季为民,沈杰,叶俊,杨斌艳,季琳.中国未成年人互联网运用报告（2025）.北京:社会科学文献出版社.2025 年:17-36.

# Dilemmas and Solutions of Minor Mode in the Context of Artificial Intelligence

**Chen Huilin[1], Zhang Hao[1], Long Yuzhu[1]**

*1.Xihua University, Chengdu, China*

**Abstract:** As generative AI becomes deeply embedded in minors' daily lives, the protective efficacy of existing minor modes is increasingly questioned. Through regulatory comparison and analysis of 369 user feedback entries, this study identifies deficiencies such as absent age-differentiated management and weak external oversight, resulting in suboptimal protection. The July 2026 Interim Measures on AI Anthropomorphic Interactive Services extend oversight to anthropomorphic services but remain confined to specific formats, leaving general-purpose AIGC tools without detailed minor-mode standards. The article argues for transcending restriction-based protection and advancing coordinated efforts—including, but not limited to, legal refinement and technological ecosystem restructuring—to develop a minor mode compatible with generative AI.

**Keywords**: minor mode; minor online protection; generative artificial intelligence