Advances in Education 教育进展, 2026, 16(9), 934-941
Published Online September 2026 in Hans. https://www.hanspub.org/journal/ae
https://doi.org/10.12677/ae.2026.1691981

Hans汉斯

AI生成文本检测技术研发赋能学科建设路径研究

智路平，方俪颖*，蔡　妙

上海理工大学管理学院，上海

收稿日期：2026年8月6日；录用日期：2026年9月7日；发布日期：2026年9月11日

摘　要

针对生成式人工智能(AIGC)引发的学术不端风险与考核同质化危机，本研究立足人工智能赋能教育的政策导向，在相关理论与前期技术研究基础上，构建“技术检测、制度规范、教育赋能”三位一体的探索性协同框架。技术上，尝试构建“双轮驱动”大语言模型文本检测体系，并通过公开数据集实验初步验证其检测性能；制度上，从应用边界、评价体系、监管流程、长效保障四个维度开展全链条设计，探索设计“系统检测、专家复核、过程溯源”三级监管机制；教育上，将检测技术转化为认知反馈工具，设计三阶段培养路径与“教师－生成式AI－AI生成文本检测－学生”四元协同教学组织形式。研究为智能时代高校学术诚信治理与人才培养模式转型提供了一个探索性的框架，为后续教育场景实证研究与实践验证提供参考。

关键词

生成式人工智能，AI文本检测，AIGC监管，人才培养模式转型

A Study on the Pathways for Empowering Discipline Construction through the R&D of AI-Generated Text Detection Technology

Luping Zhi, Liying Fang*, Miao Cai

Business School, University of Shanghai for Science and Technology, Shanghai

Received: August 6, 2026; accepted: September 7, 2026; published: September 11, 2026

*通讯作者。

文章引用: 智路平, 方俪颖, 蔡妙. AI 生成文本检测技术研发赋能学科建设路径研究[J]. 教育进展, 2026, 16(9): 934-941. DOI: 10.12677/ae.2026.1691981

Abstract

To address the risks of academic misconduct and the homogenization of assessment practices associated with generative artificial intelligence (AIGC), this study is guided by the policy orientation of AI-enabled education, building on relevant theoretical perspectives and prior research on AI-generated text detection, and develops an exploratory collaborative framework integrating technological detection, institutional regulation, and educational empowerment. Technically, a dual-driven large language model-based text detection system is explored, with its detection performance preliminarily validated through experiments on public dataset. Institutionally, a whole-process framework is designed across four dimensions: application boundaries, assessment systems, regulatory procedures, and long-term safeguards, with an exploratory three-tier regulatory mechanism comprising system-based detection, expert review, and process traceability. Educationally, AI-generated text detection is repositioned as a cognitive feedback tool, and incorporated into a three-stage talent development pathway and a four-party collaborative teaching model involving instructors, generative AI, AI-generated text detection, and students. The study provides an exploratory framework for academic integrity governance and the transformation of talent cultivation models in higher education in the intelligent era, offering a reference for subsequent empirical research and practical validation in educational settings.

Keywords

Generative Artificial Intelligence, AI-Generated Text Detection, AIGC Regulation, Talent Cultivation Model Transformation

1. 引言

随着 AIGC 的爆发式发展，大语言模型已被广泛应用于高等教育与学术研究的各个环节。然而，技术红利相伴而生的学术不端风险与考核同质化危机，正深刻冲击着传统的人才培养与学科建设模式。如何有效界定“技术辅助”与“人类原创”的边界，成为高等教育治理必须回答的时代命题。

2025 年 5 月，教育部发布《中国智慧教育白皮书》[1]，前瞻性提出“未来教师、未来课堂、未来学校、未来学习中心”的“四个未来”构想，全面启动国家教育数字化战略行动 2.0，为智慧教育的深化发展明确了顶层方向；2026 年 4 月，教育部等五部门关于印发《“人工智能+教育”行动计划》的通知[2]明确提出，既要发挥 AI 赋能教育的变革性作用，又要筑牢人工智能教育应用的安全屏障，然而，当前高校在推进“AI+”学科建设时，面临着“唯结果论”传统评价体系失灵、学术诚信监管缺乏高精度底层技术支撑、人才培养路径与智能时代脱节等现实瓶颈。

2. AIGC 教育应用与学术治理研究现状

当前，生成式 AI 的应用带来了不可忽视的学术诚信风险。AIGC 学术治理与规范是教育技术与学术诚信领域的核心议题。在理论与实践层面，祝智庭等明确了“技术承担基础性任务，人类主导创造性活动”的分工原则[3]，并指出数字化转型必须实现技术、制度、实践三者联动[4]。同时，李金正归纳出限制与披露并举的治理共识[5]，国内高校也相继出台了全流程披露等治理规范。这些治理实践的理论基础，

源于 AI 在教育中的科学定位：在人机协同视角下，Salomon 提出 AI 作为高级认知工具能解放人类认知资源[6]，Holmes 等进一步将其定位为“认知伙伴”[7]；同时 Miao 等强调 AI 系统必须服务于学习者发展与伦理安全[8]。在高等教育情境中，人机协同框架强调互补性分工，以避免 AI 过度主导导致的人类能力退化[9]。为此，Bassett 等主张构建“检测–人工复核–过程溯源”的闭环机制[10]。Sun 等指出高精度检测模型能够嵌入学位论文管理全流程实现分级管控[11]，Pudasaini 等则进一步提出将检测结果转化为认知反馈工具，推动教育评价向“重过程、重能力”转型[12]。

综上，现有研究已在学术规范治理、人机协同理论、以及 AI 赋能教育等方面形成了较为丰富的理论基础与实践探索。然而，现有研究多集中于单一维度，现有研究较少从技术检测、制度规范与教育赋能的协同视角，对高校复杂学术场景中的 AIGC 治理进行系统整合，针对高校复杂学术场景的系统性治理仍存在明显缺口。本研究在前期高精度 AI 生成文本检测模型研发的基础上，构建“检测–规范–赋能”三位一体的协同工作机制，为 AI 时代高等教育学科建设提供可操作的理论支撑与实践路径。

3. AI 生成文本检测模型开发

针对 LLM 文本检测存在场景复杂、部署成本高两大难题，本文构建双轮驱动一体化检测体系，两套模型互补协同、构成完整检测整体：模型一为级联检测模型，适配复杂真实学术场景；模型二依托对比学习与 SVM 实现轻量化高效检测。实验表明，模型一可精准识别改写类对抗文本，表现出优于部分主流基线模型的鲁棒性；模型二经对比学习优化特征，兼顾高精度与低算力，适配大批量文本筛查。二者形成“深度精检 + 轻量化快筛”协同架构，为不同检测需求下的文本筛查提供一种技术实现思路。二者的模型框架如图 1 所示：

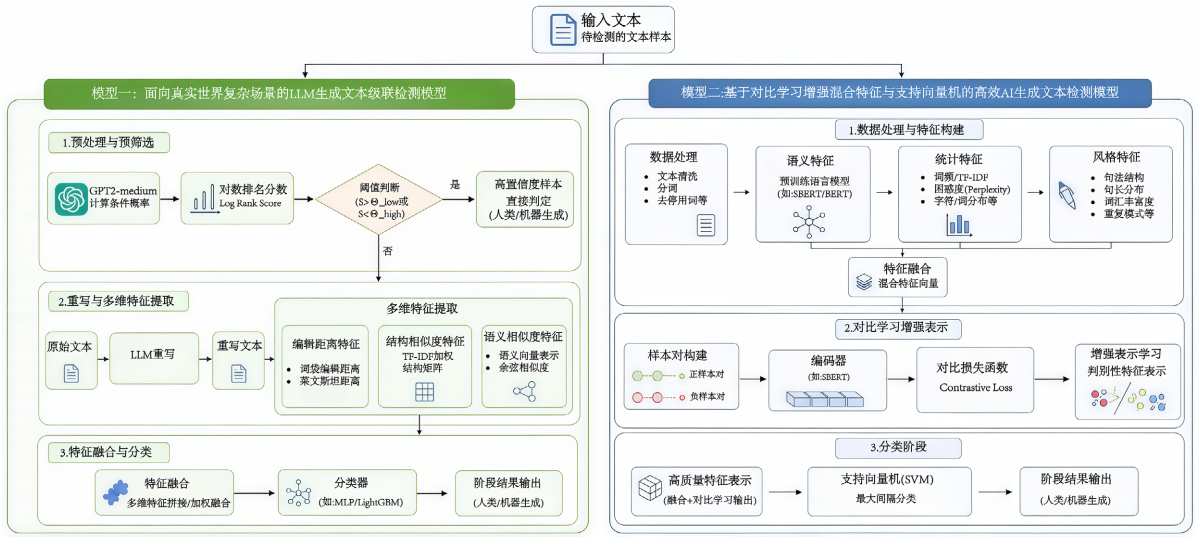

Figure 1. Model framework
图 1. 模型框架图

3.1. 模型原理

模型一采用两阶段级联检测策略。在预处理阶段，GPT2-medium 作为一种标准的自左向右(left-to-right)语言模型被采用。该模型计算每个样本第 m 个词元在给定前序 m − 1 个词元条件下生成的条件概率，据此得到该词元在所有可能生成的词元条件概率中的排名。相较于直接排名，对数排名分数能更有效地量化 AI 生成文本与人类撰写文本在词元选择“可预测性偏置”上的差异，因此被选作预处理层的关

键指标。RAIDAR 利用重写前后文本的修改程度差异性来识别文本来源，该检测算法在字符或词元的微观层面进行编辑距离计算，通过计算词袋编辑距离和基于莱文斯坦分数的编辑距离，来衡量人类和 AI 生成文本重写前后在字符或词元层面的修改量差异[13]。编辑距离仅侧重于“文本字面如何被修改”，模型在此基础上增加结构相似度的计算，它关注文本的词汇分布、重要性及整体组织模式，通过 TF-IDF 加权技术构建结构矩阵来量化词元相对重要性。它反映了“文本骨架”和关键词分布是否变化。由于一部分 LLM 使用者倾向于使用对抗性规避策略，如同义词替换、句式重组等，通过模拟人类文本的修改模式来规避基于字面或结构相似度的检测。模型增加了语义相似度特征，通过评估重写前后文本的深层含义一致性，提供了独特的语义视角。它与编辑距离和结构相似度共同构建了多维度、更全面的特征工程，有助于从编辑、结构和语义多个层面对文本变化进行表征，并为提升模型的判别能力提供支持。

模型二以对比学习为核心，通过构建人类文本与 AI 生成文本的正负样本对，采用对比损失函数学习判别性文本表示，使同类别文本在特征空间内更加紧密聚合、不同类别文本充分分离，从而强化模型对 AI 生成文本细微特征差异的感知能力，提高对不同大语言模型生成文本的泛化性能。在特征提取阶段，模型融合深层语义特征、语言统计特征与文本风格特征构建混合特征表示，其中语义特征由预训练语言模型获取上下文语义信息，统计特征刻画词汇分布、困惑度、词频等生成规律，风格特征则反映句法结构、词汇丰富度、句长变化、重复模式及语言流畅性等写作特征，多维特征互补能够更加全面地揭示 AI 文本与人类文本在语言生成机制上的本质差异。在分类阶段，模型采用支持向量机(Support Vector Machine, SVM)作为轻量化分类器，利用最大间隔原理在高维特征空间中构建最优分类超平面，并结合核函数进一步增强对复杂非线性边界的建模能力，实现对 AI 生成文本与人类文本的高精度分类。相较于直接采用大型神经网络进行端到端分类，该方法充分发挥了对比学习增强特征表示能力与 SVM 小样本、高维数据分类优势，在保证检测精度的同时显著降低模型参数规模和计算成本，实验结果表明，该方法在特定数据集上表现出较好的分类性能，并在模型规模与计算成本方面具有一定的轻量化优势，为后续探索其在教育场景中的应用提供了技术基础。

3.2. 实验及结果

模型一采用公开的全英文数据集 CHEAT，该数据集包含 15,395 篇人类撰写摘要及 35,304 篇 ChatGPT 生成摘要。以 F1 分数、精确率、召回率为评价指标，同 RAIDAR、Fast-detectgpt、RoBERTa 等主流基线模型对比，结果见表 1。

Table 1. Experimental results of Model 1 on the CHEAT dataset
表 1. 模型一在 CHEAT 数据集上的实验结果

model	F1	精确率	召回率
Ours	0.8308	0.8351	0.8313
RAIDAR	0.7895	0.7896	0.7895
Fast-detectgpt	0.7231	0.6258	0.8562
roberta-base-openai-detector	0.6668	0.5002	0.9996
roberta-large-openai-detector	0.6814	0.5595	0.8714
rank_threshold	0.6814	0.5579	0.8752
log_rank_threshold	0.7502	0.7244	0.7780
likelihood_threshold	0.7537	0.7396	0.7682
entropy_threshold	0.6668	0.5001	1.0000

模型一全部指标优于所有对照基线，其中识别人类文本召回率 88.46%、识别 AI 生成文本精确率 87.08%，说明人机文本误判风险显著更低。

模型二以公共野生数据集 SAID 中的 Quora 数据集作为训练和评估载体，包含了来自社交媒体平台 Quora 的真实 AI 生成文本。其中包含 14,648 条人类生成文本和 22,892 条 AI 生成文本。以准确率、F1、加权准确率作为评价指标，与 Fast-DetectGPT、RoBERTa 基线模型对比，结果如表 2 所示：

Table 2. Model evaluation results of Model 2 on the SAID_quora_test dataset
表 2. 模型二在 SAID_quora_test 测试集上的模型评估结果

模型	准确率	F1	加权准确率
Ours	0.89	0.89	0.93
Fast-DetectGPT	0.84	0.83	0.83
RoBERTa	0.88	0.88	0.91

从实验结果可以看出，模型二在这三项指标上均优于 Fast-DetectGPT 与 RoBERTa 基线，均衡识别性能与类别适配能力更强，表明该方法在特定数据集场景下具有潜在的应用价值。

4. AIGC 检测技术赋能学术治理与人才培养的探索性框架

依托本研究研发的高精度 AI 生成文本检测技术底座，为破解智能时代高校学术诚信监管难题、适配 AIGC 与学科教育融合趋势，初步探索技术、制度、教学三位一体的赋能框架。

4.1. AIGC 学术使用边界与透明披露机制

构建 AIGC 学术规范体系，核心是明确 AI 学术使用边界，建立可量化、可追溯的强制性标注与管控机制，防范 AI 滥用冲击学术原创性。结合国内高校管理规范与学科差异，本研究建立差异化分级管控标准，为学术 AIGC 使用的精细化合规管理提供可行的设计思路。

量化管控层面，实行学位论文 AIGC 生成内容分级标准：AI 生成占比 ≤20%为合格，可正常答辩；20% < 占比 ≤30%，由导师审核认定；30% < 占比 ≤40%，经导师审核后提交答辩委员会复核；占比 > 40%判定为不合格，需修改延期复检。同时兼顾学科差异，人文社科类从严管控，理工医科可结合实验、实证原创内容占比适度调整标准。

边界界定层面，明确 AIGC 学术使用的适用场景与禁用红线。允许 AI 辅助文献梳理、数据处理、文本润色、格式优化、思路启发等基础性工作；严禁 AI 生成论文核心内容、参与核心观点提炼与关键逻辑推演。研究方案设计、核心结论提炼等创新环节须由学生独立完成，坚守学术原创性。

溯源披露层面，推行贯穿科研全周期的 AIGC 透明披露机制。开题阶段提交 AI 使用规划，明确应用场景与用途；研究过程建立电子档案，留存提示词、原始生成内容、版本迭代记录等溯源材料；成果提交时，通过脚注、附录标注 AI 生成内容，在致谢中附加标准化披露声明，明确生成内容范围与占比，严格落实学科管控标准。

4.2. 融入 AIGC 检测的过程性学业评价机制

针对传统学业评价重终稿、轻过程、难以甄别 AI 使用合规性的弊端，本研究将 AIGC 检测技术融入评价全流程，以探索技术检测与形成性评价之间的协同可能，为学生学术能力与合规素养的评价提供多维度的参考依据，释放教师精力，聚焦高阶创新性评价工作。

本评价体系采用百分制多维度加权设计，从四项核心维度考核学生综合学术能力，具体指标如下：

(1) 知识掌握(30%)：考察学生学科基础理论、核心概念与前沿知识的应用能力。通过结构化测试、概念绘制等方式评估，依托 AIGC 检测工具筛查文献综述等内容的原创性与准确性，保障学生知识积累的真实性。

(2) 过程参与与迭代能力(40%)：核心权重指标，评估学生研究参与度、问题解决、成果迭代及人机协同合规能力。结合学生留存的提示词日志、文稿迭代记录、过程成果与 AIGC 检测报告量化分析，规避单一终稿评价的弊端。

(3) 创新性与批判性思维(20%)：考察学生原创问题提出、分析框架构建、批判性反思与跨学科研究能力。采用专家质性评审 + 技术量化辅助模式，重点评判论点新颖性、证据严谨性，规避 AI 同质化内容导致的思维惰性。

(4) 学术规范与伦理意识(10%)：考核学生学术诚信与 AIGC 使用伦理素养，涵盖 AI 内容标注完整性、引用规范性、使用比例合规性、过程日志透明度，引导学生树立合规的学术研究理念。

4.3. AIGC 检测监管与复核工作机制

为保障学术规范与过程性评价落地，对标国家 AI 内容管理相关办法，依托自研高精度检测技术，建立“系统检测－专家复核－过程溯源”三级闭环监管机制，通过技术筛查与人工判断相结合，降低单一检测结果直接用于学术判断所可能带来的误判风险。

第一，智能初检：依托微调后的高精度检测模型，对学位论文、课程作业等学术文本全面扫描，统计 AI 生成内容占比，标记超标文本，结合句法、语义、逻辑特征生成异常分析报告，为复核提供量化依据。第二，专家复核：针对初检异常成果，由院系学术伦理委员会或答辩专家团队质性审核，结合学生日常表现、答辩情况与过程材料综合判定成果原创性与 AI 使用合规性，弥补机器在混合文本场景下的识别短板，保障监管公平公正。第三，过程溯源：联合校内信息化部门搭建学术过程记录系统，存档文献查阅、文稿迭代、AI 使用等全过程数据。针对学术争议可完整溯源研究全貌，精准界定自主创作与 AI 辅助边界，为成果认定提供数据支撑。

4.4. 面向 AI + 学科建设的分阶段人才培养路径

基于前述检测技术、规范体系与监管机制，本研究进一步提出一种面向 AI 时代高校人才培养的分阶段设计思路，秉持“技术赋能、人本核心、创新为本”理念，释放 AI 低阶辅助价值、坚守人类高阶创新思维核心，构建分层递进的培养路径，实现技术赋能与人才培养的深度协同。

(1) 基础认知培育阶段：通过 AI 素养通识课程与实训，教授提示词工程基础技巧，结合自研检测工具开展对抗性验证训练，帮助学生识别 AI 文本特征，建立正确的 AI 认知与基础合规辨识能力。

(2) 人机协同进阶阶段：推动 AI 与专业训练深度融合，培养学生“人为主导、AI 辅助”的协同能力。学生依托检测工具实时核查 AI 辅助内容的原创性与准确性，通过“生成－检测－优化”闭环训练，主导研究创新与逻辑重构，旨在推动学生由基础 AI 使用逐步向更加主动、审慎的人机协同能力发展。

(3) 创新范式变革阶段：引导学生主导 AI 赋能新型研究范式，将 AIGC 融入研究全流程，构建“AI 辅助预处理－人类批判性优化”研究路径。通过跨学科人机协同项目，探索 AI 学科应用边界，依托全过程检测溯源体系，保障研究成果的原创性与学术价值。

4.5. 四元协同创新教学组织模式

为适配 AIGC 赋能教学的新需求，打破传统师生双主体教学局限，构建“教师－生成式 AI－AI 生成文本检测－学生”四元协同教学生态，实现技术层、制度层、培养层的闭环融合，如图 2 所示。

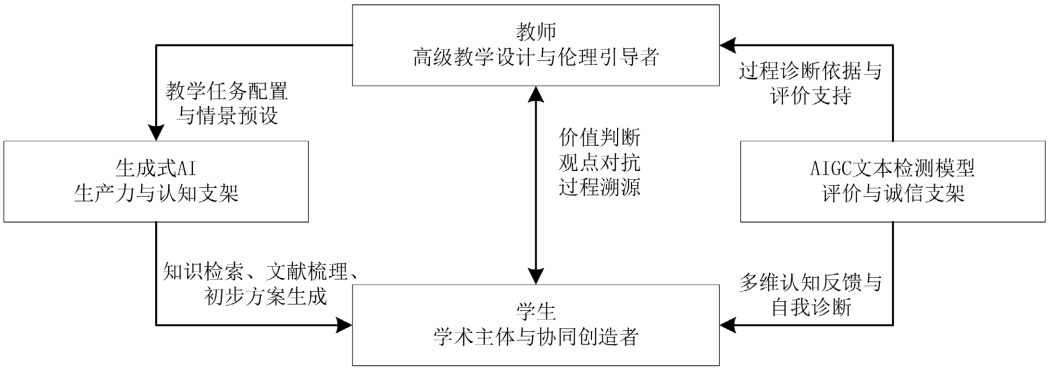

Figure 2. Four-party collaborative teaching model of “teachers - generative AI - AI-generated text detection - students”
图 2. “教师–生成式 AI–AI 生成文本检测–学生”四元协同教学组织形式

生成式 AI 承担认知支架与辅助功能，承接文献梳理、基础答疑、格式校对等重复性工作，降低师生事务性负担，提升教研效率，且所有 AI 辅助内容全程纳入检测监管，杜绝滥用。与此同时，AI 生成文本检测模型则扮演着学术诚信守门人与过程评价枢纽的角色，在制度层将研发的高精度检测技术深度嵌入至学位论文管理与学业形成性评价全流程，并建立三级监管机制，在理论设计层面，该模式尝试将检测结果转化为过程性认知反馈信息，并探索其在形成性评价中的应用可能，从而为减轻教师重复性审核负担、推动学业评价由结果导向向过程与能力导向转变提供一种可供验证的实践思路。

学生作为核心学术主体，在人机交互中持续优化研究成果，锤炼独立思考与创新重构能力，其留存的全过程创作数据，既是过程性评价依据，也是学术诚信治理的基础支撑。

教师角色转型升级为教学设计师、思维引导者与伦理守护者，摆脱基础批改与审核工作，聚焦高阶教学。教师以检测报告为辅助诊断依据，结合学生文稿迭代轨迹研判学习成效，重点开展创新研讨、观点对抗与伦理引导，督促学生证明独立思考能力。

5. 总结与展望

本研究紧扣 AIGC 给高等教育带来的机遇与挑战，以《“人工智能+教育”行动计划》为指引，搭建本研究紧扣 AIGC 给高等教育带来的机遇与挑战，以《“人工智能+教育”行动计划》为指引，在前期 AI 生成文本检测模型研究基础上，探索构建“技术检测、制度规范、教育赋能”三位一体的闭环协同框架，核心内容如下：针对学术文本检测泛化不足、算力成本较高等问题，构建双轮驱动 AI 生成文本检测体系，并通过公开数据集实验对其检测性能进行初步验证。在此基础上，沿着“规范建构–评价转型–闭环监管–伦理引导”思路，提出 AIGC 学术使用分级标准与场景边界、全周期透明披露机制、量化与质性相结合的过程性评价体系以及“系统检测、专家复核、过程溯源”三级监管机制，探索弥补单一技术检测局限的治理路径。同时，尝试转变检测工具单纯用于结果筛查的定位，将其作为能力培养的认知反馈载体，设计三阶段分层培养路径，提出“教师–生成式 AI–AI 生成文本检测–学生”四元协同教学模式。总体而言，本文形成的相关机制与模式主要属于基于技术研究和相关理论的探索性框架，其实际应用效果仍有待进一步验证。

但本研究中构建的教育治理与人才培养框架尚未经过真实课程场景的实证验证，部分评价指标及权重具有探索性；同时，检测模型主要基于公开数据集开展实验，其在不同学科和真实高校环境中的泛化能力仍需进一步检验。

未来可通过课程试点，将本文提出的评价与教学模式应用于具体课程，并结合问卷、访谈、学习过

程记录及学业成果对比等方式开展实证研究，进一步验证框架的有效性与可行性。同时扩大不同学科和文本类型的数据来源，根据实践结果持续优化检测模型、评价指标及治理机制。

参考文献

[1] 《中国智慧教育白皮书》和启动国家教育数字化战略行动 2.0 [EB/OL]. http://www.moe.gov.cn/jyb_xwfb/xw_zt/moe_357/2025/2025_zt06/cgfb/202505/t20250507_1189603.html, 2025-05-16.

[2] 教育部等五部门关于印发《“人工智能+教育”行动计划》的通知[EB/OL]. http://www.moe.gov.cn/srcsite/A16/s3342/202604/t20260410_1433240.html, 2026-04-10.

[3] 祝智庭, 戴岭, 赵晓伟. “近未来”人机协同教育发展新思路[J]. 开放教育研究, 2023, 29(5): 4-13.

[4] 祝智庭, 朱晓悦, 陈怡, 闫寒冰. 教育技术创新的系统图谱: 全景解读与生态系统框架研究——“跟跑·并跑·领跑”战略视域下的中国探索[J]. 电化教育研究, 2025, 46(12): 5-18.

[5] 李金正, 李德. 中西方期刊界对生成式人工智能参与科研论文写作的风险治理比较分析[J]. 中国科技期刊研究, 2025, 36(7): 824-834.

[6] Salomon, G. (1993) No Distribution without Individuals' Cognition: A Dynamic Interactional View. In: Salomon, G., Ed., *Distributed Cognitions*: *Psychological and Educational Considerations*, Cambridge University Press, 111-138.

[7] Holmes, W. (2019) Artificial Intelligence in Education: Promises and Implications for Teaching and Learning. Center for Curriculum Redesign.

[8] Miao, F. and Holmes, W. (2021) AI and Education: A Guidance for Policymakers. UNESCO Publishing.

[9] Atchley, P., Pannell, H., Wofford, K., Hopkins, M. and Atchley, R.A. (2024) Human and AI Collaboration in the Higher Education Environment: Opportunities and Concerns. *Cognitive Research*: *Principles and Implications*, **9**, Article No. 20. https://doi.org/10.1186/s41235-024-00547-9

[10] Bassett, M.A., Bradshaw, W., Bornsztejn, H., Hogg, A., Murdoch, K., Pearce, B., *et al*. (2026) Heads We Win, Tails You Lose: AI Detectors in Education. *Journal of Higher Education Policy and Management*, 1-16. https://doi.org/10.1080/1360080x.2026.2622146

[11] Sun, Y., Liao, Y. and Ma, X. (2026) Trusting AI to Detect AI? A Systematic Evaluation of the Reliability and Robustness of Current AIGC Detection Tools for Student Academic Work. *Computers & Education*, **249**, Article 105616. https://doi.org/10.1016/j.compedu.2026.105616

[12] Pudasaini, S., Miralles-Pechuán, L., Lillis, D. and Llorens Salvador, M. (2024) Survey on AI-Generated Plagiarism Detection: The Impact of Large Language Models on Academic Integrity. *Journal of Academic Ethics*, **23**, 1137-1170. https://doi.org/10.1007/s10805-024-09576-x

[13] Mao, C., Vondrick, C., Wang, H. and Yang, J. (2024) Raidar: Generative AI Detection via Rewriting. https://arxiv.org/abs/2401.12970