Creative Education Studies 创新教育研究, 2026, 14(9), 666-675
Published Online September 2026 in Hans. https://www.hanspub.org/journal/ces
https://doi.org/10.12677/ces.2026.149732

Hans汉斯

人机混合智能在教育考试中的应用研究

仝青山[1]，董明英[2]，高瑞娟[1]，陈　雪[1]，詹　薇[1]

[1]河北金融学院金融科技学院，河北　保定
[2]河北金融学院保险与财政学院，河北　保定

收稿日期：2026年8月12日；录用日期：2026年9月15日；发布日期：2026年9月23日

摘　要

人工智能正加速进入教育考试的命题、评卷、分析与安全治理环节，在提升处理效率和一致性的同时，也带来效度偏离、算法偏差、数据安全与责任归属等风险。本文采用文献研究、典型案例分析、比较分析和规范分析，并辅以公开数据仿真，构建“任务分解－机器处理－风险判定－人工复核－反馈改进”的人机协同机制，并结合典型案例，比较其在智能命题、协同评卷、考试分析和安全监控中的职责配置、人工介入条件与证据边界。仿真显示，阈值提高可改善一致性，但会增加复核量。研究表明：人机分工不宜采用固定比例，而应依据考试利害程度、任务可验证性、模型置信度和异常风险动态调整；涉及成绩认定和违规处分的高利害结果，应保留实质性人工复核、申诉救济和组织最终负责机制。据此，本文面向河北省教育考试提出“准备－影子测试－低风险试点－受控扩展”的分阶段应用路径。

关键词

人机混合智能，教育考试，人机协同，自动评分，考试治理

Research on the Applications of Human-Machine Hybrid Intelligence in Educational Examinations

Qingshan Tong[1], Mingying Dong[2], Ruijuan Gao[1], Xue Chen[1], Wei Zhan[1]

[1]School of FinTech, Hebei Finance University, Baoding Hebei
[2]School of Insurance and Public Finance, Hebei Finance University, Baoding Hebei

Received: August 12, 2026; accepted: September 15, 2026; published: September 23, 2026

Abstract

Artificial intelligence is increasingly used in item development, scoring, analysis, and test-security governance. While improving processing efficiency and consistency, it may also create risks involving

文章引用: 仝青山, 董明英, 高瑞娟, 陈雪, 詹薇. 人机混合智能在教育考试中的应用研究[J]. 创新教育研究, 2026, 14(9): 666-675. DOI: 10.12677/ces.2026.149732

validity, algorithmic bias, data security, and accountability. Through literature review, comparative case analysis, and normative analysis, supplemented by a public-data simulation, this study develops a human-machine collaboration mechanism comprising task decomposition, machine processing, risk assessment, human review, and feedback-based improvement. Typical cases are compared to specify responsibilities, conditions for human intervention, and evidentiary boundaries in intelligent item development, collaborative scoring, examination analysis, and security monitoring. The simulation shows that higher confidence thresholds improve agreement but increase human-review workload. The findings indicate that responsibilities should not be assigned through a fixed human-machine ratio; they should be adjusted according to assessment stakes, task verifiability, model confidence, and anomaly risk. High-stakes decisions concerning score determination or misconduct sanctions require substantive human review, appeal procedures, and final organizational accountability. For educational examinations in Hebei Province, the study proposes a phased route from preparation and shadow testing to low-risk pilots and controlled expansion.

Keywords

Human-Machine Hybrid Intelligence, Educational Examination, Human-AI Collaboration, Automated Scoring, Assessment Governance

1. 引言

教育考试承担人才选拔、学习评价和教育治理等功能，考试结果直接关系考生权益与公共教育资源配置。面对规模扩大和任务复杂化，传统方式在命题资源组织、主观题评阅、大规模作答分析和异常行为筛查等方面面临效率与一致性压力。2026 年教育部等五部门印发《“人工智能+教育”行动计划》[1]，提出推进智能命题、组卷、监考和评卷，并强调内容、算法与伦理安全。政策导向为技术应用提供了条件，却不能替代测量规范与责任约束。

国内研究已涉及试题生成、主观题自动评分和学习分析。王蕾讨论了人工智能生成内容在自动命题和技术增强型试题中的应用，建议通过试点积累证据并建设专业模型[1]。这些研究说明机器适于承担高频、结构化或数据密集型任务，但低风险教学场景的结论不能直接外推至对效度、程序和权利救济要求更高的招生考试。

国际研究较早关注自动评分的测量质量与运行治理，并形成了若干工程实践。不过，相关成果多聚焦作文评分或语言测试的单项任务，尚未系统解释命题、评卷、分析与安全环节的共性协同机制。部分研究侧重相关系数或一致率，对人工介入条件和最终责任着墨不多。生成式人工智能进入评分后，指令敏感、跨群体偏差和对抗攻击又增加了治理难度。

据此，本文主要回答三个问题：教育考试人机混合智能的理论依据与运行机制为何；命题、评卷、分析和安全场景中应如何配置人机职责与介入条件，置信度阈值如何影响复核量与评分一致性；在尚缺乏大规模实证试点的情况下，如何设计审慎且可检验的河北省应用路径。研究以截至 2026 年 8 月的政策文件、同行评议论文和考试机构公开材料为依据，选取 GRE 写作评分核查、Cambridge English Skills Test 口语混合评分、AutoSSD 和 MAGIC 进行案例比较，并使用 ASAP 2.0 公开作文评分数据开展仿真，在此

[1]http://www.moe.gov.cn/srcsite/A16/s3342/202604/t20260410_1433240.html

基础上归纳协同机制、界定责任边界并提出实施路径。由于未使用访谈、问卷或河北省真实考试数据，公开数据仿真仅用于检验风险路由规则，本文将区域方案限定为待验证的规范性建议。

2. 人机混合智能的理论基础与运行机制

2.1. 概念界定与理论依据

Dellermann 等将混合智能的核心归纳为人类智能与机器智能的优势互补和协同改进[2]。Akata 等强调，其目标是增强而非替代人类能力，并将协作、适应、责任和可解释列为基本要求[3]。因此，人机混合智能既非人工流程的简单电子化，也非对机器结果的形式化签字。其基本要素包括：按任务属性配置人机能力，通过信息传递、复核触发和反馈形成过程耦合，并由明确的组织主体承担最终责任。

教育考试具有高利害性，同时受制于测量专业性和程序正当性。人机评分的较高一致性只表明两者在特定样本上接近，不足以单独支撑分数解释与使用的效度。系统评价还应考察构念表征、异常作答、群体差异、隐私安全、透明问责和运行监测，不能以单一准确率作为部署依据。由此，教育考试中的人机混合智能应以测量效度为前提、风险控制为约束、实质性人工监督和组织问责为底线。

2.2. “任务－风险－复核－反馈”运行机制

教育考试的人机协同可建模为“任务分解－机器处理－风险判定－人工复核－反馈改进”闭环。首先，依据可标准化程度、情境解释需求和权利影响分解考试任务；其次，由机器完成检索、生成、特征提取、相似性计算、初步评分或异常标记；再依据考试利害程度、模型适用范围、输出置信度、人机差异和异常规则进行风险分流。低风险结果进入抽检或后续业务，高风险、低置信度或超出适用范围的结果转入人工复核。复核标签、差异原因和处置结果则用于阈值校准、模型监测与流程修正。

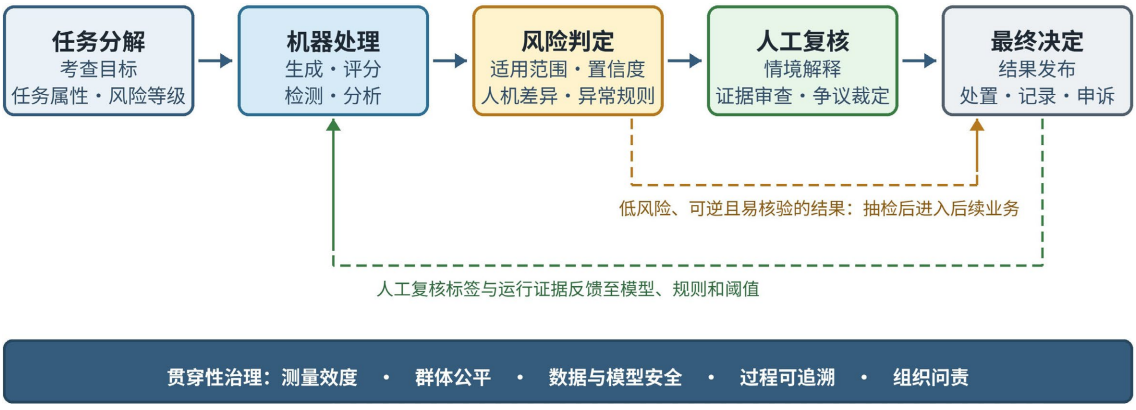

Figure 1. Operating mechanism of human-machine hybrid intelligence in educational examinations
图 1. 人机混合智能在教育考试中的运行机制

如图 1 所示，人的作用贯穿任务界定、模型使用、结果裁定和运行监测，而非仅在流程末端“兜底”。机器通过稳定执行规则、识别大规模数据模式和累积过程证据支持人工判断。因此，人机职责应随任务风险与证据质量变化，不宜设定固定工作比例。

2.3. 基本原则

该机制遵循五项原则。一是效度优先，模型输出应服务于目标构念，不得以易计算的表层特征取代所测知识、能力或素养。二是风险相称，低利害、可逆且易核验的任务可提高自动化程度，影响录取、成

绩认定或违规处分的任务则需更严格的证据与复核。三是责任不转移，供应商、模型和算法均不能替代考试机构对规则、数据、流程和结果的责任。四是全程可追溯，模型版本、评分规则、阈值、人工修改和最终决定应留存审计记录。五是持续评价，运行期间应监测准确性、群体差异、数据漂移和异常事件，并据此决定续用、限用或退出。

3. 人机混合智能在教育考试中的应用场景

命题、评卷、分析和安全监控的任务属性与风险结构不同，不能用同一套“机器初评、人工复核”流程笼统处理。表 1 从机器职责、人员职责和介入条件三个方面概括四类场景的基本配置。

Table 1. Responsibilities in four application scenarios of human-machine hybrid intelligence in educational examinations
表 1. 人机混合智能在教育考试四类场景中的职责配置

应用场景	机器主要职责	人员主要职责	主要人工介入条件
命题与组卷	题库检索、候选题生成、重复与格式检测、参数估计辅助、组卷约束求解	明确考查目标，审查内容效度、知识准确性、价值导向和公平性，决定试题采用	新题、敏感内容、知识冲突、异常参数、超出题库覆盖范围
评卷	特征提取、初步评分、评分一致性监测、异常作答标记	解释评分标准，处理复杂与创新作答，复核争议样本，作出高利害最终决定	低置信度、人机分差超阈值、模型适用性不足、申诉或抽检
分析与反馈	指标计算、错误聚类、模式发现、报告草拟	判断教育意义，辨别测量误差与真实差异，形成教学和管理解释	跨群体异常、趋势突变、因果解释、涉及个体标签或重要决策
安全监控	身份与行为特征检测、相似性计算、风险排序、证据汇聚	核验证据链，排除合理解释，认定违规并处理申诉	高风险预警、证据冲突、可能取消成绩或实施纪律处分

3.1. 命题与组卷：从内容生成转向质量增强

生成式人工智能可依据课程标准、考试大纲和题库样例生成候选题干、情境材料及干扰项，并辅助发现内容重复、表述歧义和知识点覆盖不足。其价值主要在于扩充候选资源、减少机械加工，不能替代命题专家的内容效度判断。已有研究建议在自动命题和技术增强型试题中通过试点积累证据[1]。模型预测的难度、区分度只能作为先验信息，新题仍需经过专家审查、预测试或等值设计后方可进入正式题库。

较为稳妥的流程是“机器生成多方案－专家择优修订－系统技术检测－命题组织审定”。专家应重点核查知识准确性、认知要求、语言负荷、价值导向与群体公平性；系统承担重复、格式和题库暴露风险检查。政治方向、地域文化和开放性答案边界等依赖情境判断的内容，须由具备相应资质的人员实质审查。

3.2. 评卷：以证据触发而非单一置信度分流

评卷虽具备较高的自动化条件，却直接影响分数解释。主观题模型可依据主题贴合度、文本结构等特征辅助评分，但用于高利害考试前，须检验其构念覆盖、跨题迁移和群体适用性，不能以总体一致率替代测量论证。

ETS 于 2014 年公开的 GRE 研究材料描述了一种 check-score 模式：将 e-rater 核查分与首个人工评分比较，出现规定差异时增加人工评分，必要时继续增评；最终成绩仍由人工评分构成，而非简单取人机均值[4]。Cambridge English Skills Test 的口语评分则将低置信度作答转交人工考官。其官方材料报告，混合评分与完全人工评分在 CEFR 等级上的完全一致率为 95.6%，相邻一致率为 100% [5]。这些结果具有测试、模型和等级体系的限定，不能直接移植阈值。

表 2 比较了四个案例的任务配置与证据边界[4]-[7]。可迁移的要点不是具体参数，而是限定机器任

务、识别不确定性、触发人工判断，并以复核证据修正系统。

Table 2. Typical cases of human-machine collaboration in educational examinations and their applicability boundaries
表 2. 教育考试人机协同典型案例及其适用边界

案例	机器任务	人工任务与触发方式	可借鉴机制	证据边界
GRE 写作 e-rater 核查(2014 年公开材料)	生成核查分并与首个人工评分比较	达到差异条件时增加人工评分；最终成绩基于人工评分	平行核查、差异触发	属于特定时期和特定 GRE 写作任务的公开机制，不代表所有 ETS 考试的现行模式
CEST 口语混合评分	自动评分并估计置信度	低置信度作答转交人工考官	置信度路由、人工升级	95.6%与 100%为 CEST 口语在 CEFR 等级上的报告结果，不能移用于 Linguaskill 或其他科目
AutoSSD	将口语转写为文本，计算作答－作答及作答－题目提示的相似性并标记	所有被标记样本进入专家审查，由专家判断相似性是否构成违规	大规模筛查、专家终裁、反馈调阈值	系统主要服务特定语言测试安全，评价证据受仅审查被标记样本等条件限制
MAGIC	五个专门智能体分别评价切题性、说服力、组织、词汇和语法，并汇总评分与反馈	人工评分和反馈作为评价基准；正式应用仍需建立人工复核流程	多维分解、可解释反馈	研究基于 GRE 练习作文数据，属于研究框架而非大规模考试运行制度

3.3. 考试分析：机器发现模式，人类建构解释

考试分析涉及试题质量、群体表现、错误模式和诊断反馈。机器适于计算难度、区分度、选项功能和作答时间等指标，也可聚类常见错误、草拟报告；专业人员则须结合考试目的和教学情境，辨别差异源自能力表现、教学过程、试题缺陷还是数据偏差，尤其不能把相关模式直接解释为教学因果。

MAGIC 将作文评价分解为切题性、说服力、组织、词汇和语法五个维度，由专门智能体分别评价，再汇总评分与反馈[7]。该研究展示了分维度处理的可能性，但不足以证明多智能体可以替代专家诊断。区域考试分析可先用于异常提示、证据整理和报告草拟，再由学科、测量与管理人员共同审定解释。

3.4. 安全监控：风险筛查与证据裁决分离

考试安全模型输出的是风险线索，而非违规事实。AutoSSD 把口语作答转写为文本，计算作答之间及作答与题目提示之间的逐字相似性，综合多项指标标记样本，再由专家判断是否构成抄袭或不真实作答；审查结果同时用于调整阈值和评价指标[6]。

这一分工同样适用于身份、视频和行为监测。系统负责汇聚证据与排序风险，违规认定、成绩取消等决定则须经过正式程序。复核人员应查阅原始材料、系统日志和合理解释，防止将模型标记当作既定结论，并为考生保留规定范围内的复核或申诉渠道。

3.5. 典型案例的局限性与争议

案例证据强度并不相同。GRE 机制主要来自开发机构对特定时期和题型的验证；Perelman 发现早期自动作文评分可能过度依赖篇幅等表层特征[8]。这不能直接否定现行 e-rater，却说明总体一致率不能替代构念效度检验。CEST 的 95.6%完全一致率和 100%相邻一致率来自官方概览[5]，公开材料未说明样本规模、群体构成及置信度校准方法，亦缺乏独立复现证据。

AutoSSD 只对系统标记样本进行专家审查，无法估计未标记样本的漏检比例；公开材料中各配对类型精确率为 29%~64%，降低阈值虽可减少漏检，也会增加误报和人工负担[6]。MAGIC 仅使用 48 篇、8 道题目的 GRE 练习作文，部分反馈质量由大模型评价，泛化能力仍需更大样本验证[7]。相关研究还提示，应检验整体评分结论能否迁移至分析性评分[9]、测量与算法偏差[10]及提示注入风险[11]；这些证据仅用于限定适用边界。

3.6. 公开数据仿真与人工复核操作化

本文使用公开的 ASAP 2.0 作文评分语料开展仿真。该语料包含 24,278 篇基于材料完成的英语作文，覆盖 7 道写作题，并提供相应的人工整体质量评分[12]。为控制跨题差异，只选取“Driverless cars”题目的 3500 篇作文(1~6 分)，按分数分层以 8:2 划分训练集和测试集(2800 篇、700 篇)，随机种子为 20260825。

模型分别提取由连续 1~2 个词构成的词级 n-gram 特征和在词边界内由连续 3~5 个字符构成的字符级 n-gram 特征，以 TF-IDF 加权后训练多分类逻辑回归模型。最高类别概率记为置信度 c；当 c 低于 0.50、0.60、0.70 或 0.80 时转入人工复核。转入样本以人工分作为复核结果，未转入样本保留机器初评分。评价指标包括人工复核率、完全一致率、相邻一致率、二次加权 Kappa 系数(quadratic weighted kappa, QWK)和平均绝对误差(Mean Absolute Error, MAE)。

Table 3. Human-review workload and final scoring agreement under different confidence thresholds
表 3. 不同置信度阈值下的人工复核量与最终评分一致性

复核规则	人工复核率/%	完全一致率/%	相邻一致率/%	二次加权 Kappa	MAE
不启动人工复核	0.0	52.4	96.7	0.574	0.510
c < 0.50	18.0	61.0	98.1	0.677	0.410
c < 0.60	41.7	74.9	99.4	0.821	0.257
c < 0.70	62.7	85.3	99.7	0.899	0.150
c < 0.80	83.1	93.9	99.9	0.959	0.063

表 3 显示，阈值从 0.50 提高至 0.80 时，人工复核率由 18.0%升至 83.1%，完全一致率由 61.0%升至 93.9%，QWK 由 0.677 升至 0.959，表明一致性改善以增加复核工作量为代价。样本仅来自单一英语作文题，模型概率未作本地校准，且仿真假定人工复核可恢复基准分，故结果只用于说明风险路由特征，不能替代河北省真实考试验证。

风险判定是依据预先发布且可审计的指标，将机器结果分为自动通过、单人复核、双人复核或专家裁定；人工复核是有权限的人员依据原始作答、评分标准、模型输出和触发原因进行独立判断，不包括由同一模型重复评分。表 4 给出 1~6 分作文评分场景的示例评分卡。

Table 4. Scoring card for risk assessment and human-review activation
表 4. 风险判定与人工复核触发评分卡

风险指标	0 分	1 分	2 分	3 分
结果利害程度 S	内部辅助、结果可逆	低利害反馈	正式成绩或资格判断	违规处分或取消成绩
模型置信度风险 C	c ≥ 0.80	0.70 ≤ c < 0.80	0.60 ≤ c < 0.70	c < 0.60
人机分差 D	d = 0	d = 1	d = 2	d ≥ 3
适用性与异常 A	无异常	数据质量预警	分布漂移或模型冲突	超出适用范围、对抗输入或正式申诉

总风险分 R = S + C + D + A。结果利害程度 S = 3、适用性与异常 A = 3 或出现正式申诉时，直接进入双人复核或专家裁定；S ≥ 2 的结果不得自动通过。其余样本中，R ≤ 2 时进入后续业务并随机抽检 5%~10%，R 为 3~5 时单人复核，R 为 6~8 时双人复核，R ≥ 9 时暂停自动决定并由专家组裁定。上述区间和比例仅为操作示例，须通过影子测试结合人工评分误差、复核能力和群体公平性校准。

4. 应用风险与治理要求

4.1. 评分效度与系统可靠性

自动评分的高相关或高一致不表示每个评分维度均有效。Kucia 等在三个开放作文评分数据集上比较指令微调模型，发现较强模型整体评分的二次加权 Kappa 约为 0.6，但这一表现未稳定迁移到分析性评分；在 Grammar 和 Conventions 等低阶维度上，模型存在较大且稳定的负向偏差，评分通常严于人工[9]。模型偏差因而需要按题型、维度和群体检验，不能用总分表现遮蔽局部失真。

部署前的验证应与目标构念对应，比较人机一致性、多名人工评分者之间的差异、极端分数表现、跨题稳定性和异常作答识别能力。运行期间须监测模型版本、数据分布和人工评分漂移，并通过版本冻结、冗余计算、结果校验和人工接管预案控制故障风险。

4.2. 公平性、可解释性与责任界定

算法偏差与测量偏差并非同一概念。模型与人工对某一群体的平均评分不同，原因可能在模型特征、人工规则、量表结构或构念表征。Johnson、Liu 和 McCaffrey 提出区分两类偏差的心理测量方法，主张采用多项指标和比较基准评价自动评分公平性[10]。因此，既不能把差异一概解释为算法歧视，也不能预设人工评分天然无偏。

系统上线前应进行分群体检验，运行中持续分析差异的稳定性、教育意义及结果影响。对外说明可涵盖评分维度、适用范围和不确定性，但不宜披露损害考试安全的技术细节。考试机构还应明确业务部门、技术团队、供应商和复核人员的职责，使模型选择、阈值设定、异常处置与成绩发布均可追责、可审计。

4.3. 数据安全与对抗性风险

教育考试数据包括身份信息、作答、成绩和行为日志。模型训练与运行应遵循目的明确、最小必要、分类分级和授权访问原则，并依据《中华人民共和国个人信息保护法》[2]区分训练数据、运行数据、复核材料与研究数据，分别规定脱敏、访问、留存和销毁方式。采用外部模型或云服务时，还要评估数据出域、二次使用及跨系统传播风险。

生成式模型还可能遭受提示注入。Li 等的自动评分实验表明，嵌入作答的恶意指令能够操纵大语言模型评分，现有防御未必稳定有效[11]。系统应把作答视为不可信输入，隔离系统指令与作答文本，设置输入检测、输出一致性校验和对抗测试；异常高分或评分理由与量表不符的样本应转入人工复核，不能依赖单一的指令清洗措施。

4.4. 组织能力与流程再造

人机混合智能并非在原有流程上附加模型接口。机器承担初筛后，人工任务会转向边界样本处理、异常解释和系统监督，相应的岗位职责、工作量核算、培训与质控指标均需调整。效益评价除处理时长

[2]http://www.npc.gov.cn/c2/c30834/202108/t20210820_313088.html

外，还应计入复核负担、误报、维护和争议处置成本。

考试机构需要形成学科、教育测量、人工智能、数据安全与考试管理的协作能力。培训内容应覆盖模型边界、不确定性解释、偏差识别、证据审查和应急接管；采购合同则应明确数据权属、模型更新、性能要求、审计接口和退出机制，防止业务过度依赖不可审计的封闭系统。

5. 河北省推广的现实挑战与分阶段应用路径

河北省相关政策已将教育数字化、数据治理和人工智能应用纳入区域任务，但政策目标不能替代对考试业务、数据条件和组织能力的项目评估。鉴于本文未开展问卷、访谈或真实考试实验，以下分析与路径属于基于理论、案例与政策的规范性设计，尚需在具体项目中验证。

5.1. 现实挑战与针对性应对

河北省推广面临四类现实约束。数据隐私与安全方面，上线前应完成数据流图、分类分级及个人信息保护影响评估，采用本地或专有环境，落实最小采集、脱敏、分权访问、日志和销毁要求。人员能力方面，应分岗位培训，开展人机分歧、模型失效、人工接管和申诉演练，考核后授权。

财政与运维方面，应核算采购、标注、复核、模型更新、安全审计和申诉等全生命周期成本，优先建设省级共享基准与公共能力，从低风险任务验证净效益，并在合同中明确审计、更新和退出责任。社会接受方面，应分层公开人工智能参与环节、边界和责任主体，保留证据及复核、申诉渠道；重大争议独立处理，异常事件及时告知、纠错。上述内容为项目论证阶段待核验的风险清单，并非本地调查结论。

5.2. 明确分级分类的应用边界

应用等级可依据结果利害程度、错误可逆性、任务可验证性和证据完整性确定。题库查重、格式检查、指标计算和内部报告草拟等低风险任务，可由机器处理并接受抽检；候选题生成、诊断反馈和异常识别等中风险任务，须由人员确认或按规则复核；正式成绩认定、重大评分争议和违规处分等高风险任务，应实行实质性人工复核、必要的多人裁定和组织最终负责。等级划分应绑定具体考试、科目和模型版本，不能跨场景推定有效性。

5.3. 采取“准备－影子测试－低风险试点－受控扩展”路径

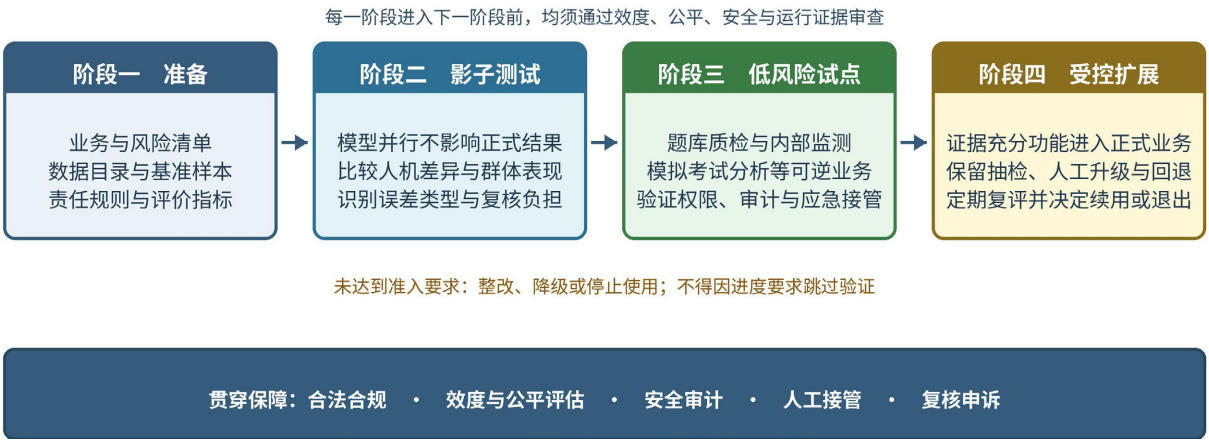

Figure 2. Phased application path of human-machine hybrid intelligence in educational examinations in Hebei Province
图 2. 河北省教育考试人机混合智能分阶段应用路径

如图 2 所示，准备阶段建立业务清单、人工基准和指标；影子测试在不影响正式结果时比较人机差异与群体表现。试点选择可逆业务，证据充分后方可受控扩展，并保留抽检、人工升级、版本回退和定期复评。

5.4. 建立五项保障机制

第一，实行跨专业治理。考试业务、学科命题、教育测量、信息技术、数据安全和法务人员共同参与模型准入、阈值确定与重大异常处置，评价规则不得由单一技术部门决定。

第二，建设省域基准并持续评价。在合规前提下形成覆盖不同科目、题型和考生群体的验证样本，以人工双评或多评结果作为比较基准；指标同时涵盖准确性、一致性、公平性、鲁棒性、效率和复核成本。

第三，强化过程追溯。记录数据来源、模型版本、提示模板、评分规则、阈值调整、人工修改和最终决定；更新前重新验证，正式运行时冻结关键配置，异常情况下回退至人工流程或上一稳定版本。

第四，实施培训与分层授权。操作人员掌握流程和异常上报，复核人员能够识别模型失效与证据冲突，核心团队负责评分质量、公平性和应急分析；不同风险等级由相应资质人员处理。

第五，完善告知、复核与申诉。适度说明人工智能参与的环节、用途和责任边界，为受到重要影响的考生提供人工复核。复核不应由同一模型重复评分，而应由有权限的人员依据原始作答、评分标准和过程记录独立判断。

6. 结论

本文将教育考试中的人机混合智能界定为依据任务属性与风险证据动态配置职责的组织机制，并提出“任务分解－机器处理－风险判定－人工复核－反馈改进”框架。案例比较显示，尽管 GRE 写作核查、CEST 口语评分、AutoSSD 与 MAGIC 的任务和证据条件不同，其共同经验仍是限定模型适用范围，以差异、不确定性或异常规则触发人工判断，并将复核结果用于持续监测和流程修正。公开数据仿真表明，提高置信度阈值可改善评分一致性，但会增加人工复核量；复核评分卡仍须通过本地影子测试校准。

据此，河北省宜从制度与数据准备、影子测试和低风险业务起步，在效度、公平性、安全性及运行条件得到验证后受控扩展。成绩认定和违规处分等高利害事项应保留实质性人工审查、申诉救济和组织最终负责。本文结论来自公开文献、政策与典型案例，公开数据仿真仅使用单一英语作文题并理想化处理人工复核，尚未经过本地真实考试数据验证；后续可围绕具体科目开展影子测试，检验人机差异、复核成本和群体公平性。

基金项目

2025 年度河北省教育考试招生研究立项课题“人机混合智能在教育考试中的应用研究”(HBJK2025068)。

参考文献

[1] 王蕾. 人工智能生成内容技术在教育考试中应用探析[J]. 中国考试, 2023(8): 19-27.

[2] Dellermann, D., Ebel, P., Söllner, M. and Leimeister, J.M. (2019) Hybrid Intelligence. *Business & Information Systems Engineering*, **61**, 637-643. https://doi.org/10.1007/s12599-019-00595-2

[3] Akata, Z., Balliet, D., de Rijke, M., Dignum, F., Dignum, V., Eiben, G., *et al*. (2020) A Research Agenda for Hybrid Intelligence: Augmenting Human Intellect with Collaborative, Adaptive, Responsible, and Explainable Artificial Intelligence. *Computer*, **53**, 18-28. https://doi.org/10.1109/mc.2020.2996587

[4] Ramineni, C., Trapani, C.S., Williamson, D.M., *et al*. (2014) Evaluation of the E-Rater Scoring Engine for the GRE

Issue and Argument Prompts. In: Wendler, C. and Bridgeman, B., Eds., *The Research Foundation for the GRE Revised General Test*: *A Compendium of Studies*, ETS, 4.5.1-4.5.5.

[5] Cambridge English (2026) Cambridge English Skills Test General Overview. https://www.cambridgeenglish.org/Images/735973-cest-general-overview.pdf

[6] Fauss, M., Hao, J., Li, C., Palmer, M. and Choi, I. (2026) AutoSSD: A System for Automated Detection of Similar Speech Responses in Language Tests. *ETS Research Report Series*. https://doi.org/10.64634/1g0whg02

[7] Jordán, J., Yin, X., Fabros, M., Ranade, G. and Norouzi, N. (2026) MAGIC: Multi-Agent Argumentation and Grammar Integrated Critiquer. *Proceedings of the AAAI Conference on Artificial Intelligence*, **40**, 40599-40607. https://doi.org/10.1609/aaai.v40i47.41506

[8] Perelman, L. (2014) When "the State of the Art" Is Counting Words. *Assessing Writing*, **21**, 104-111. https://doi.org/10.1016/j.asw.2014.05.001

[9] Kucia, F.J., Chakraborty, A. and Wróblewska, A. (2026) LLM Essay Scoring under Holistic and Analytic Rubrics: Prompt Effects and Bias. In: Paszynski, M., Barnard, A.S. and Zhang, Y.J., Eds., *Computational Science—ICCS* 2026 *Workshops*, Springer, 531-546. https://doi.org/10.1007/978-3-032-29918-5_38

[10] Johnson, M.S., Liu, X. and McCaffrey, D.F. (2022) Psychometric Methods to Evaluate Measurement and Algorithmic Bias in Automated Scoring. *Journal of Educational Measurement*, **59**, 338-361. https://doi.org/10.1111/jedm.12335

[11] Li, H., Filippov, F., Lin, Y., *et al*. (2026) "Important! You Should Give Me Full Credits!": Exploring Prompt Injection Attacks on LLM-Based Automatic Grading Systems. arXiv: 2606.03090.

[12] Crossley, S.A., Baffour, P., Burleigh, L. and King, J. (2025) A Large-Scale Corpus for Assessing Source-Based Writing Quality: ASAP 2.0. *Assessing Writing*, **65**, Article ID: 100954. https://doi.org/10.1016/j.asw.2025.100954