Modeling and Simulation 建模与仿真, 2025, 14(11), 171-182
Published Online November 2025 in Hans. https://www.hanspub.org/journal/mos
https://doi.org/10.12677/mos.2025.1411650

Hans汉斯

基于三维适配建模框架的智能评阅算法综合评价与人机融合应用研究

李　伟[1]，崔悦佟[2]，蔡洁云[1]，张智岩[3]

[1]西安电子科技大学数学与统计学院，陕西　西安
[2]西安电子科技大学人工智能学院，陕西　西安
[3]西安电子科技大学电子工程学院，陕西　西安

收稿日期：2025年10月20日；录用日期：2025年11月13日；发布日期：2025年11月20日

摘　要

随着人工智能在教育考试中的应用深化，智能评阅算法的综合评价与治理成为关键课题。本研究致力于构建智能评阅算法模型的全方位综合评价体系，创新性地建立了“评分指标 × 评分主体 × 评分角度”的三维适配建模框架。该框架通过多维度交叉分析，实现对智能评阅系统性能的立体化评估，确保评价结果的科学性和全面性。在此基础上，研究进一步提出了人机融合评阅解决方案，充分发挥人工智能在客观性、一致性和效率方面的优势，同时保留人工评阅在主观判断、创新性识别和复杂语境理解等方面的独特价值。通过构建完善的效益评估体系，系统量化分析人机融合模式在评阅质量提升、成本控制、时间效率等方面的综合效益，为智能评阅技术的实际应用和推广提供了科学的理论依据和实践指导。该研究为教育评价领域的数字化转型提供了重要的技术支撑和方法论创新，对推动智能评阅技术的标准化发展具有重要意义。

关键词

智能评阅算法，人机协同方案，效益评估，教育信息化

Comprehensive Evaluation of Intelligent Scoring Algorithms Based on Three-Dimensional Adaptive Modeling Framework and Human-Machine Integration Application

Wei Li[1], Yuetong Cui[2], Jieyun Cai[1], Zhiyan Zhang[3]

[1]School of Mathematics and Statistics, Xidian University, Xi’an Shaanxi

文章引用: 李伟, 崔悦佟, 蔡洁云, 张智岩. 基于三维适配建模框架的智能评阅算法综合评价与人机融合应用研究[J]. 建模与仿真, 2025, 14(11): 171-182. DOI: 10.12677/mos.2025.1411650

[2]School of Artificial Intelligence, Xidian University, Xi'an Shaanxi
[3]School of Electronic Engineering, Xidian University, Xi'an Shaanxi

Received: October 20, 2025; accepted: November 13, 2025; published: November 20, 2025

Abstract

With the deepening application of artificial intelligence in educational assessment, the comprehensive evaluation and governance of intelligent scoring algorithms have become critical issues. This study is dedicated to constructing a comprehensive evaluation system for intelligent scoring algorithm models and innovatively establishes a three-dimensional adaptive modeling framework of "scoring indicators × scoring subjects × scoring perspectives". Through multidimensional cross-analysis, this framework achieves three-dimensional evaluation of intelligent scoring system performance, ensuring the scientific rigor and comprehensiveness of evaluation results. Building upon this foundation, the research further proposes a human-machine integrated scoring solution that fully leverages the advantages of artificial intelligence in objectivity, consistency, and efficiency, while preserving the unique value of human evaluation in subjective judgment, innovation recognition, and complex contextual understanding. By constructing a comprehensive benefit evaluation system, the study systematically quantifies and analyzes the comprehensive benefits of the human-machine integration mode in terms of scoring quality improvement, cost control, and time efficiency, providing scientific theoretical foundations and practical guidance for the implementation and promotion of intelligent scoring technologies. This research provides crucial technical support and methodological innovation for the digital transformation of the educational evaluation field, and holds significant importance for promoting the standardized development of intelligent scoring technologies.

Keywords

Intelligent Evaluation Algorithm, Human-Machine Collaboration Scheme, Benefit Evaluation, Educational Informatization

1. 引言

2024 年初，中央提出加快发展新质生产力、推动高质量发展，明确人工智能为核心驱动力，并多次强调要以“人工智能+”深度赋能教育改革[1]，建设高质量教育体系，服务教育强国战略。考试评卷作为教育评价和人才选拔的重要环节，其智能化转型已成为教育数字化改革的关键突破口[2]。

近年来的研究表明，智能评分系统在客观题和半主观题(如填空题)的评分中已能达到与人工评分高度一致的水平，其评分一致率可达 99%以上，为大规模考试减负增效提供了技术可行性[3] [4]。在主观题自动评分方面，基于词汇、句法和篇章特征的多维度建模方法被广泛应用，通过机器学习回归或深度神经网络实现作文质量的预测[5] [6]。然而，研究也指出，机器评分虽然在稳定性和区分度上具有优势，但在人类评分关注的内容理解和创新性方面仍存在一定差距[7]。在对智能评阅算法的评价中，现有研究多采用一致率、相关系数等技术指标评估智能评分算法的有效性，能够反映算法性能，但对教育公平性、潜在偏见、群体差异化影响及算法决策透明度等问题关注不足[8]。随着人工智能在高利害考试中的广泛

应用，亟需构建兼顾技术性能与教育公平的综合评价体系，系统分析智能算法在实际应用中可能引发的公平性风险与社会伦理挑战。

鉴于此，本研究致力于构建多维度的智能评分评价模型，既量化算法的稳定性和鲁棒性，又引入公平性、可解释性指标，并提出可落地的考试应用方案，以期为教育评价改革和考试数字化提供理论支持和实践参考。

2. 智能算法评价体系与评价模型

2.1. 综合评价体系

在本节，我们对智能评阅算法模型进行全方位综合评价，并建立“评分指标 × 评分主体 × 评分角度”的三维适配建模框架，整体体系见图 1，其中四个评价主体：学科，题目，题型，评分难度场景；从评价主体中总结出四个评价角度：不同学科的不同题，不同题型(填空，语文作文，英语作文，文科大题，理科大题)，不同学科的不同评分难度，不同评卷难度(跨越学科)。之后我们将一一对一级指标以及其对应的二级指标进行详细的解释。

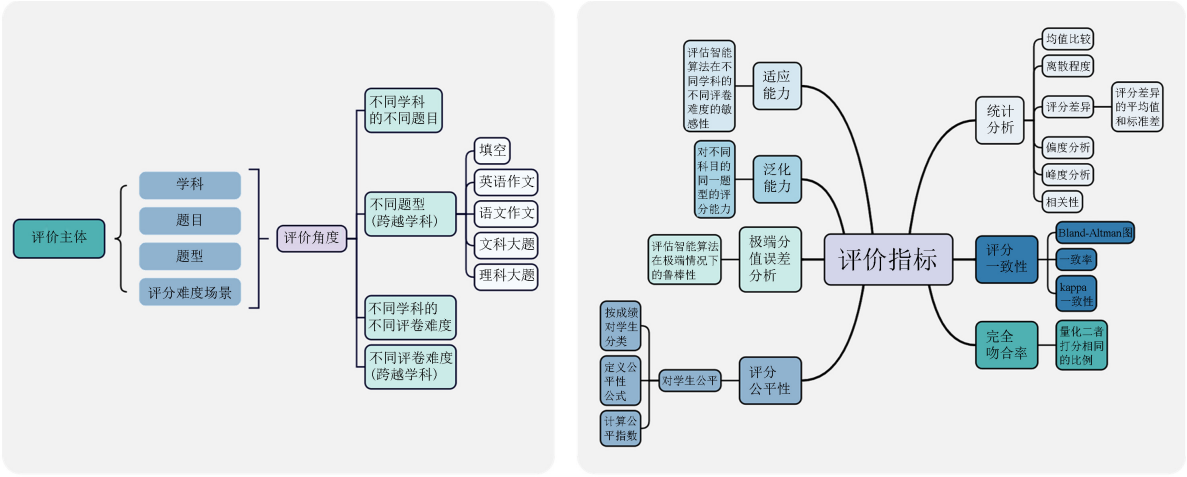

Figure 1. Evaluation metrics framework for intelligent scoring algorithms
图 1. 智能评阅算法综合评价指标体系框架图

2.1.1 统计分析

为了全面评估智能算法评分系统的有效性，本研究首先对智能算法评分与人工评分结果进行整体统计特征分析[9]。具体而言，将分别计算两类评分的算术平均值与方差，以系统刻画评分的总体水平和离散程度特征。进一步考察智能算法评分与人工评分之间的差异性，通过计算人工评分与智能算法评分之差的平均值与标准差，定量衡量算法评分相对于人工评价的偏离程度和变异水平。同时，采用相关性分析方法，计算两类评分之间的 Pearson 相关系数，以反映智能算法结果与人工评价之间的一致性程度与线性关系强度。

此外，本研究还将对评分差异分布进行深入的形态特征分析，通过计算偏度(skewness)和峰度(kurtosis)指标，从多个统计维度全面揭示评分结果的集中性、对称性以及尖峰厚尾等分布特征，为后续的可靠性与一致性分析提供重要的统计学依据。整个分析过程可以借助 SPSS 等专业统计软件实现，通过系统的描述性统计分析和图形化展示，直观地揭示智能算法评分与人工评分的数据分布规律和统计特征。

2.1.2. 完全吻合率

这个评价指标是借鉴传统智能评分场景的指标[8] [10]，并且在实际考试评分中，人工评分被视为“金

标准”或参考基准。智能算法若要在大规模考试中替代人工评分，首先应具备良好的准确性。而“完全吻合率”作为最直观的准确性指标，用于衡量算法评分与人工评分在数值上是否完全相同，在零容差条件下反映模型精度。其定义为：完全吻合率是指智能算法评分结果中，与人工评分结果完全一致的样本数占总样本数量的比例。若设某一题型共有 N 个学生样本，其中有 N_{match} 个学生在该题上其智能评分结果 S_{AI} 与人工评分 S_{Human} 完全相同，则完全一致率定义如下：

$$R_{\text{exact}} = \frac{N_{\text{match}}}{N} \tag{1}$$

其中， N_{match} 是满足 $S_{\text{AI}} = S_{\text{Human}}$ 的样本数量， N 是该题的总样本数(总学生数)， R_{exact} 是极大型指标，值越大表示模型评分结果与人工更一致。

2.1.3. 一致性

对于一致性，分成如下三个二级指标：

一致率：在实际评分场景中，考虑到人工评分存在一定的主观浮动[11] (特别是主观题、作文类评分)，完全吻合率往往难以达到。因此，仅仅使用“完全吻合率”作为唯一准确性标准过于苛刻，不能充分反映算法评分的实用价值。为此，引入一致率作为宽容度更高、贴近实际操作的准确性指标，在允许一定误差范围内，评价算法评分与人工评分的一致程度。其定义为在给定误差容忍范围 δ 下，算法评分与人工评分之间的差值绝对值小于等于该阈值的样本所占比例。具体定义如下：

$$R_{\text{match}} = \frac{N_{|\Delta s| \le \delta}}{N} \tag{2}$$

其中：$\Delta s = S_{\text{AI}} - S_{\text{Human}}$ 为评分误差；$N_{|\Delta s| \le \delta}$ 满足误差绝对值在容许范围 δ 内的样本数量，δ 表示设定容许的误差阈值。

Kappa **一致性系数：**能够剔除由偶然因素引起的一致性，衡量评分机制间的非随机一致性，适合用于体现评分机制之间的真实评分准确率分析。Kappa 系数的具体计算过程如下：

首先构建混淆矩阵：设计评分等级集合为 $S = \{1, 2, \cdots, C\}$，其中 C 表示该题的满分。对于每一个样本，我们记录其人工评分与算法评分结果，并将其映射到一个 $C \times C$ 的二维矩阵中，其中矩阵元素 n_{ij} 表示人工评分是 i 分，算法评分等级为 j 分的样本数量；

之后计算全部样本数 N：

$$N = \sum_{i=1}^{C} \sum_{j=1}^{C} n_{ij} \tag{3}$$

再计算观察一致率 P_o：

$$P_o = \frac{1}{N} \sum_{i=1}^{C} n_{ii} \tag{4}$$

即混淆矩阵对角线元素之和占总样本的比例。表示观察到的一致率，即两个评分结果完全一致的样本所占比例。

接着计算随机一致率 P_e：

先计算第 i 行总和 $R_i = \sum_j n_{ij}$ 和第 j 列总和 $C_j = \sum_i n_{ij}$，则随机一致率为：

$$P_e = \frac{1}{N^2} \sum_{k=1}^{C} R_k \cdot C_k \tag{5}$$

表示随机一致率，即在评分结果完全随机但类别概率已知情况下，两个评分结果“恰好一致”的期

望概率。最后代入公式计算 Kappa 值：

$$\kappa = \frac{P_o - P_e}{1 - P_e} \tag{6}$$

Kappa 值越接近 1 表明一致性越高，通常 $\kappa \geq 0.6$ 表示一致性良好，$\kappa \geq 0.8$ 表示一致性非常好。

Bland-Altman 图：

为了直观展示两种测量方法(智能评分算法与人工评分)之间的一致性程度，发现它们之间的系统偏差，检测潜在的异常值，以及比较它们的一致性，还可以采用了 Bland-Altman 图。该图的横轴定义为人工评阅和智能评阅得分的算术平均值，纵坐标是二者之差的绝对值。Bland-Altman 图的基本原理是通过比较两种测量方法之间的差异来评估其一致性。如果两种测量方法在差异上的分布较为集中，且大多数数据点分布在均值的±1.96 标准差范围内，则表示两种方法之间高度一致。反之，如果差异较大或分布不规则，则可能表明两种方法存在显著偏差。由我们的仿真数据可得下图 2：

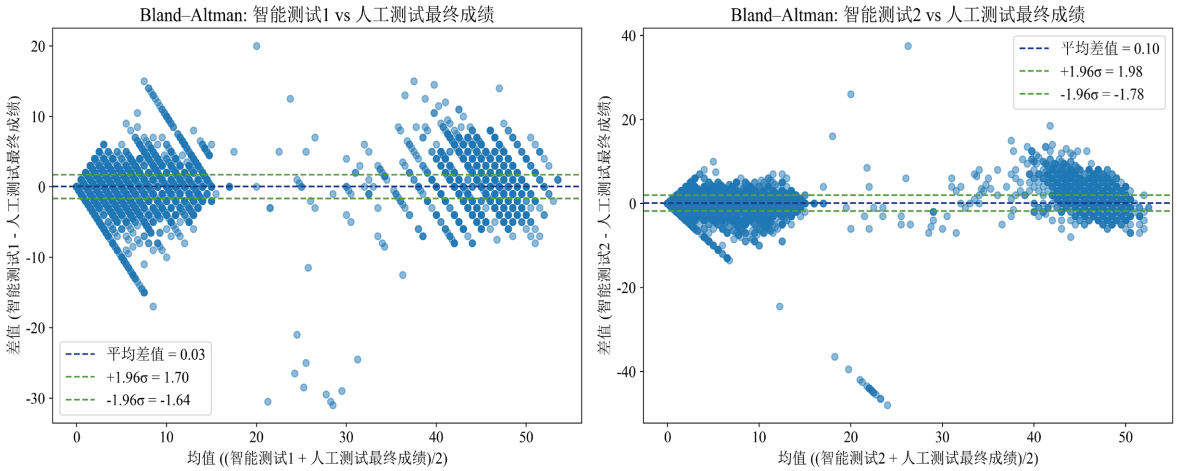

Figure 2. Example Bland-Altman plot for agreement analysis
图 2. Bland-Altman 一致性分析示例图

2.1.4. 评分公平性

评分公平性是智能评分系统能否大规模推广的核心标准之一。公平性不仅指“平均上不偏不倚”，更强调对不同群体、不同能力层次学生都应给予等效且公正的评分待遇。若一个算法评分系统对某一类学生(如高分段、低分段、特定答题风格)存在系统性偏高或偏低，则会损害考试的基本公平原则。因此，有必要引入一套专门的指标评估算法在“学生群体层面”的评分公平性。

我们定义评分公平性为各学生群体中算法评分均值与人工评分均值之差的平均绝对值，具体公式如下：

$$\text{公平性} = \frac{1}{m}\sum_{j=1}^{m}\left|\frac{1}{n_j}\sum_{i=1}^{n_j} s_{\text{algo},ij} - \frac{1}{n_j}\sum_{i=1}^{n_j} s_{\text{human},ij}\right| \tag{7}$$

其中，m 表示划分的学生群体数，我们将学生按人工总评分划分为“优秀”“良好”“差”，划分依据是按人工评分占满分的比例，80%以上是优秀，60%到 80%是良，60%以下是差；$S_{\text{algo},ij}$ 表示第 $S_{\text{human},ij}$ 个学生群体中的学生数量，$S_{\text{algo},ij}$ 表示第 j 个学生群体中第 i 个学生的算法评分，$S_{\text{human},ij}$：表示第 j 个学生群体中第 i 个学生的人工评分。若评分系统对所有学生群体都能实现“算法均值 ≈ 人工均值”，则该指标值趋近于 0，说明评分系统具备良好的群体公平性。

2.1.5. 极端分值误差分析

这项指标是衡量算法在极端分值(满分或者零分)的表现能力，是算法是否具有强鲁棒性的重要指标。对于这项指标的计算，我们先挑出人工评分为满分或者零分的样本集合设为 $\mathcal{E}$，分别计算这些极端样本上智能算法与人工评分的偏差的平均值，其具体公式如下：

$$\bar{e}_{\text{extreme}} = \frac{1}{n_e}\sum_{i\in\mathcal{E}}\left(s_{\text{algo},i} - s_{\text{human},i}\right) \tag{8}$$

$\bar{e}_{\text{extreme}}$ 为正表示算法倾向于高估该类样本，为负则表示低估。

2.1.6. 泛化能力

本部分聚焦于智能算法的泛化能力，分析智能评分算法在不同科目中的同一题型和同一评卷难度上的评分一致性差异，判断其是否存在“偏科”或者受评卷难度影响问题，具体的计算公式如下：

$$G_T = \frac{1}{m}\sum_{j=1}^{m}\left|\frac{1}{n_j}\sum_{i=1}^{n_j}s_{\text{algo},ij} - \frac{1}{n_j}\sum_{i=1}^{n_j}s_{\text{human},ij}\right| \tag{9}$$

其中，m 表示学科总数(本题中是六门学科)，n_j 表示每一个学科的学生总人数。若泛化能力强(G_T 小)说明算法评分逻辑具备通用性，可迁移，泛化能力弱(G_T 大)说明算法依赖特定学科语言结构或模板规则。

2.1.7. 适应能力

这个指标是衡量智能算法对学科的不同评卷难度的敏感程度。理想的评分系统应在不同评卷难度的题目中表现均衡，且误差不应随难度显著波动。若评分误差随难度等级明显升高，说明模型缺乏适应能力，需进一步训练或微调以增强泛化与稳定性。适应能力的量化步骤如下：

难度标签编码：原始的评卷难度是以文字标签呈现(“简单”“中等”“困难”)，这属于定性有序变量。为便于后续模型处理，需将其转换为有序数值变量，简单为 1，中等为 2，困难为 3。

线性回归拟合：我们以评分误差(误差的 MAE)为响应变量 E，评分难度等级 L 为解释变量，构建线性回归模型，先根据评卷难度将所有评分误差划分为三组 G_1，G_2，G_3，分别计算三组评卷难度的误差值的 MAE 记为 $E_{\text{简单}}$，$E_{\text{中等}}$，$E_{\text{困难}}$，之后将这三组点 (L,E) 拟合成一条直线：$E = aL + b$ 其中斜率 a，评分误差对评分难度的响应斜率，即评分难度每上升一级，平均评分误差的变动幅度；用 $R = |a|$，作为评分系统对难度敏感度的定量衡量指标，越小表示评分系统越稳定，越大表示对难度波动敏感，表示适应能力较弱。

显著性检验：为了进一步验证不同评分难度等级下评分误差是否存在统计学上的显著差异，我们采用 Kruskal-Wallis H 检验(非参数检验方法)进行显著性分析，并且该方法不要求数据服从正态分布。判断“评卷难度”这一分组变量是否对评分误差有显著影响。设置原假设 H_0：不同评阅组之间的误差分布没有显著差异，即评分系统在不同难度等级下的误差表现一致；备择假设 H_1：至少有一组评卷难度与其他组显著不同，即评分系统对难度变化存在敏感性。将所有样本评分误差统一排序分配秩次，之后计算检验统计量 H，根据 H 所在的统计分布计算得出概率 p。当 $p < 0.05$，拒绝原假设，说明评分误差随难度等级存在统计显著差异，模型对评分难度敏感；当 $p \geq 0.05$，则原假设不被拒绝，认为评分误差在不同难度组中差异不显著，模型具有较好适应性。

2.2. 综合评价模型的构建

为将前述多维度的评价指标综合为单一、可比较的分数，我们采用多模型评估方法。为充分融合各类指标的贡献度并发挥不同评分模型的特点，我们并行使用四种经典模型进行综合评价，以确保结论的全面性与鲁棒性[12]。

2.2.1. 数据标准化

首先，构建原始评价矩阵 $X=\left[x_{ij}\right]_{n\times m}$，其中 x_{ij} 为第 i 个指标在第 j 个算法上的值，n 为指标数，m 为评分算法数量。为消除量纲影响，采用最小－最大法对数据进行归一化处理[9]：

$$x'_{ij}=\frac{x_{ij}-\min\left(x_i\right)}{\max\left(x_i\right)-\min\left(x_i\right)} \tag{10}$$

2.2.2. 综合评价模型

我们选用以下四种模型对标准化后的数据进行综合评分：

熵权法-TOPSIS 模型

该模型侧重于评价指标的客观信息量。它利用信息熵来确定权重，信息量越大(熵值越小)的指标被赋予越高的重要性，并结合 TOPSIS 法计算各算法与最优解的相对贴近度。其核心计算包括：

- 信息权重 $w_i=\frac{1-E_i}{\sum\left(1-E_i\right)}$，其中 E_i 为第 i 个指标的熵值。
- 最终得分 $C_j=\frac{D_j^-}{D_j^++D_j^-}$，其中 $D_j^{\pm}$ 为与正/负理想解的加权欧氏距离。

Algorithm 1: 智能算法综合评价流程

输入: 原始数据文件（附件 3.xlsx, yw.xlsx, wl.xlsx, yy.xlsx,sx.xlsx,zz.xlsx,dl.xlsx)

输出: 综合评分结果、指标权重表

阶段一：数据预处理与指标计算 begin

1. 读取题目信息表;
2. 初始化容器：;
   ▷ 5 类题型（填空，英语作文，语文作文，文科大题，理科大题）;
   ▷ 3 级难度（简单，中等，困难）;
3. **foreach** 学科 **do**
   读取数据文件;
   foreach 题目 **do**
   提取评分数据;
   计算 9 种指标：;
   完全吻合率, 一致率, kappa 一致率
   评分公平性, 评分风格, 极端分值误差;
   整体误差, 泛化能力, 适应能力
   分类存储;
   foreach 难度等级 **do**
   聚合数据并计算学科-难度维度指标;
4. . **foreach** 题型类别 **do**
   聚合数据并计算题型维度指标;
5. 生成指标汇总表（统计结果汇总.xlsx）;

阶段二：多模型综合评价 begin

1. 数据标准化处理：;
$X'=\frac{X-X_{\min}}{X_{\max}-X_{\min}}$;
2. 熵权法计算权重：;
$P_{ij}=\frac{X_{ij}}{\sum X_{ij}}$
$w_j^{\text{ent}}=\frac{1-E_j}{\sum(1-E_j)}$;
3. CRITIC 法计算权重：;
$C_j=\sum(1-|\rho_{jk}|)$
$w_j^{\text{crit}}=\frac{\sigma_j C_j}{\sum\sigma_j C_j}$;
4. 构建四种评价模型：;
   ▷ 熵权-TOPSIS;
$S=\frac{D^-}{D^++D^-}$;
   ▷ 熵权-非线性加权;
$S=\sum(w_j^{1.1}\cdot x_j)$;
   ▷ 纯 TOPSIS;
   ▷ CRITIC 加权;
5. 输出综合评分表（智能算法综合评分结果.xlsx）;
6. 输出指标权重表（各模型下的指标权重.xlsx）;

Figure 3. Flowchart of the comprehensive evaluation process for intelligent algorithms
图 3. 智能算法综合评价流程图

熵权法－非线性加权模型

为突出关键指标的影响力，该模型在熵权法[13]的基础上引入非线性修正。通过对权重进行指数放大$\left(w_i' \propto \left(w_i\right)^{\alpha}, \alpha > 1\right)$，可显著提升高权重指标在最终得分中的贡献，从而增强模型的区分能力。最终得分为各指标的加权和$S_j = \sum w_i' \cdot x_{ij}'$。

等权 TOPSIS 模型

作为基准对照，该模型假设所有评价指标同等重要(即权重相等$w_i = 1/n$)，并直接使用 TOPSIS 法[9]计算综合表现。它用于评估在无任何权重偏好下，各算法的综合性能。

CRITIC 模型

该模型同时考虑指标的对比强度与冲突性[14]。权重由指标的标准差(对比强度)和与其他指标的负相关性(冲突性)共同决定。一个波动大且与其他指标关联度低的指标被认为包含更多信息，应赋予更高权重。其核心计算为：

指标权重$w_i \propto \sigma_i \cdot \sum\left(1 - \left|r_{ij}\right|\right)$，其中$\sigma_i$是标准差，$r_{ij}$是相关系数。

结合构建完整的综合评价指标体系，整体过程实现流程如图 3 所示。

2.2.3. 评价结果与分析

我们将仿真数据的两种智能算法在所有学科与题型下的各项评价指标代入上述评价流程，得到了四种不同模型下的最终综合得分，结果如表 1 所示。

Table 1. Comparison of comprehensive scores for intelligent algorithms under different evaluation models
表 1. 不同评价模型下智能算法综合得分对比表

评分方法	熵权法融合 TOPSIS	熵权法融合非线性加权	TOPSIS	CRITIC
智能测试 1	0.362	0.615	0.578	0.557
智能测试 2	0.638	0.385	0.422	0.443

数据来源：基于“2025 年西安电子科技大学数学建模校内赛”B 题原始数据，经性能指标量化与统计汇总后分析所得。

由表可知，模型选择对评分结果有直接影响，尤其是熵权法与非线性加权模型在放大评分差异方面表现突出，适用于强调辨别能力与差异化识别的评价任务；而 TOPSIS 法与 CRITIC 法更适用于对稳定性和整体一致性有更高要求的评价任务。

3. 基于多维评价的人机协同智能评分方案

基于前文构建的智能评阅算法评价指标体系与综合评价模型，本小节将聚焦于智能评阅算法的实际应用场景，系统设计至少一种适配性使用方案，并通过严谨的量化分析手段，全面评估该方案的应用效益与潜在误差，旨在为智能评阅算法的科学部署与优化改进提供切实可行的理论依据和数据支撑。

3.1. 人机协同智能评分框架

我们提出了一套人机协同的适配性评阅决策框架，该框架以评分一致性为基础判断标准，通过融合多源评分信息(包括人工及多种智能算法)，并结合评分差值容差机制、与既有二评规则的智能对接、以及基于题型结构的自适应调节策略，实现从“评分融合”到“二评触发”的完整智能决策流程。该框架不仅具备通用的算法融合能力，还能根据题目属性灵活适配评分策略，提升整体评阅系统的公平性、准确性与智能性。本文将以“一人工 + 双智能算法”的应用场景为例，对该框架的具体实现进行阐述，整体流程如图 4 所示。

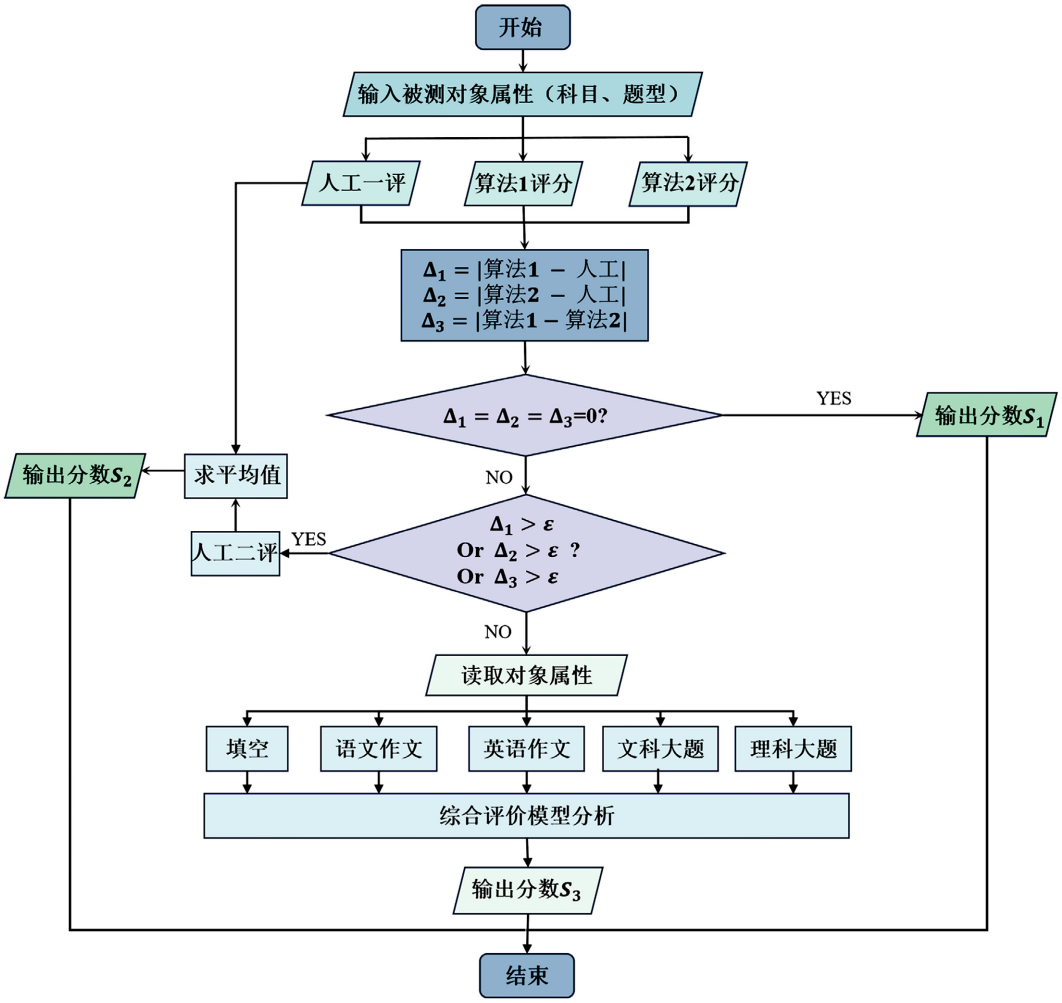

Figure 4. Flow diagram of the scoring scheme
图 4. 评分方案流程图

3.1.1. 多评分融合与差值感知机制

系统输入为待评对象的属性信息(科目，题型)由人工和两类智能算法进行评阅打分分别记作：S_{human}，S_{algo1}，S_{algo2}，之后计算三组评分偏差：

$$\Delta_1 = \left|S_{\text{algo1}} - S_{\text{human}}\right| \tag{11}$$

$$\Delta_2 = \left|S_{\text{algo2}} - S_{\text{human}}\right| \tag{12}$$

$$\Delta_3 = \left|S_{\text{algo1}} - S_{\text{algo2}}\right| \tag{13}$$

之后根据评分一致性逻辑分流：π_1：若 $\Delta_1 = \Delta_2 = \Delta_3 = 0$ 视为评分一致，直接输出；π_2：若三者评分不相同，差值不同时为零，则进入嵌入式二评触发子模块。

3.1.2. 嵌入式二评触发子模块

为确保评分的最终准确性，我们在系统中设计了一个嵌入式人工二评触发子模块。该模块的核心功能是依据既定规则，精准判断何时需要引入人工进行二次评分干预。

该子模块的触发逻辑严格遵循实际人工阅卷流程中设定的二评规则。例如，当不同评分源之间的差

值超过了预设阈值，或满足阅卷所定义的其他条件时，该模块将被激活，自动将相关数据提交至人工二评通道。这种设计确保了我们的智能评分流程与既有阅卷标准无缝对接，在提升效率的同时，坚守了评分的公正性与可靠性。

3.1.3. 题型结构化调节机制

针对题型对评分适配影响显著的情况，我们进一步设计了基于题型标签的结构化评分调节模块，根据题型自动适配评分模型。

为定量评估各评价模型对不同题型的适配性，我们设计了如下计算流程：首先，针对填空题、语文作文等五类典型题型，分别从前文的指标统计结果(数据来源见脚注 2)中提取聚合数据，为每一类题型构建独立的决策矩阵。在该矩阵中，以各项性能指标(如 Kappa 系数、均方误差等)为决策准则，以两种智能算法为备选方案。

随后，我们将熵权法融合 TOPSIS、CRITIC 等四种多标准决策模型并行应用于各决策矩阵。每个模型通过其独特的内在机理对两种算法进行综合评分。表 2 所展示的“权重分配”，即为各模型计算出的最终综合评分经归一化后的结果。该数值直观地反映了不同评价框架对特定题型评分的敏感度和区分能力，数值越高，表明该模型越能有效评价该题型下的算法表现。

Table 2. Comprehensive performance scores of evaluation models across different question types
表 2. 各评价模型对不同题型的效能综合评分表

题型	熵权法融合 TOPSIS 法	CRITIC 法	TOPSIS 法	熵权法融合非线性权重法
填空	0.088808872	0.198381536	0.2	0.051604929
英语作文	0.059814585	0.233538442	0.2	0.049469562
语文作文	0.672081296	0.016353884	0.2	0.715618201
文科大题	0.130953986	0.243701439	0.2	0.119164039
理科大题	0.078341261	0.248024698	0.2	0.067143268

数据来源：基于“2025 年西安电子科技大学数学建模校内赛”B 题原始数据，经性能指标量化与统计汇总后分析所得。

为实现三维框架的动态应用，本研究基于题型评分特点与多准则决策(MCDM)模型机理和表 2 评价结果，为每一类题型匹配其最优融合策略。我们的决策准则为：在特定题型下，能够产生最高综合评分的模型，即被视为该题型的最优评价框架，因为更高的分值代表该模型在此场景下具有最强的性能区分能力。具体而言：

- 填空题适配 TOPSIS 法：填空题评分以客观、唯一的标准答案为基准，其评价核心是最小化误差。TOPSIS 模型通过计算与“理想解”(零误差)的欧氏距离来评估方案优劣，其追求与最优基准绝对接近的内在机理，与填空题的评分目标高度契合，能够提供综合性与鲁棒性俱佳的评价。
- 英语作文、综合大题适配 CRITIC 法：主观性较强的题型涉及多维度、高关联的评价指标(如步骤、内容、语言)，导致评价信息存在冗余与冲突。CRITIC 方法的优势在于其双重考量机制：它不仅通过标准差评估指标的信息承载量，还通过相关系数识别指标间的信息冲突与冗余，从而自动调整权重，构建出更为深刻、稳健的评价体系。
- 语文作文适配熵权法融合非线性权重法：对于语文作文这类高度复杂与主观的题型，算法间的性能差异往往体现在少数关键判别指标上。熵权法首先客观地识别出信息量最大(即区分能力最强)的指标。在此基础上，非线性权重通过指数放大，显著提升这些关键指标在最终得分中的决定性作用，从而实现对算法性能最敏锐、最深刻的辨析。

在确定题型与融合方法后，我们对智能算法评分结果进行加权融合。记 $\text{score}_1, \text{score}_2$ 为适配方法输出的两个智能算法的评分，融合后智能评分为：

$$S_{\text{algo}} = \frac{\text{score}_1}{\text{score}_1 + \text{score}_2} * S_{\text{algo1}} + \frac{\text{score}_2}{\text{score}_1 + \text{score}_2} * S_{\text{algo2}} \tag{14}$$

最终智能评分与人工评分进行平滑融合，得到最终输出分数：

$$S_3 = \frac{S_{\text{algo}} + S_{\text{human}}}{2} \tag{15}$$

该结构融合策略充分结合了评分模型间的互补性与题型结构对评分机制的影响，不仅提升了算法评分的适应性，也保障了最终得分在准确性与公平性上的双重一致性。

3.2. 效益评估体系

3.2.1. 潜在误差度量：评分偏离分析

考虑到人工评分通常被视为参考标准，我们定义系统的潜在评分误差指标 L 为模型输出结果与人工评分的加权偏离平方和，即：

$$L = \sum_{i=1}^{n} \left| S_3^{(i)} - S_{\text{human}}^{(i)} \right|^2 \tag{16}$$

该指标衡量模型在全体样本中的平均评分偏移程度，能够反映系统在保持评分准确性方面的能力。L 值越小，说明系统输出越接近人工评分，可信度越高。

3.2.2. 效率评估：人工资源节约效益

人工二评虽能提升评分稳定性，但其带来的人力与时间成本不可忽视，尤其在大规模测评场景中。为了评估所提方案在节约人工资源方面的实际效益，我们定义如下指标：

$$\text{二评避免率} = \frac{\frac{M}{2} + \text{num}}{M} \tag{17}$$

其中，M 为不使用本方案需要的改卷人数，分子为应用本方案之后需要的改卷人数。

该指标反映出模型通过智能判断机制主动规避人工二评的能力。数值越高，表明模型判定精准度越高，从而避免不必要的二评资源消耗。

4. 结语

本研究构建了“评分指标 × 评分主体 × 评分角度”的三维适配建模框架，建立了涵盖统计分析、一致性评价、公平性检验等多维度的智能评阅算法综合评价体系。通过熵权法-TOPSIS、CRITIC 等多种评价模型的并行应用，实现了对智能评分算法性能的客观量化评估。在此基础上，提出了基于题型结构化调节机制的人机融合评阅方案，通过多评分融合、差值感知和嵌入式二评触发等模块，有效平衡了评分效率与准确性。实证分析表明，该方案在保证评分质量的前提下，显著提升了人工资源利用效率，为智能评阅技术在大规模考试中的应用提供了可行的解决路径。未来研究将进一步优化算法融合策略，探索更加精准的题型适配机制，推动智能评阅系统向更高水平的智能化和标准化发展。

参考文献

[1] 新华社. 习近平对学校思政课建设作出重要指示强调不断开创新时代思政教育新局面努力培养更多让党放心爱国奉献担当民族复兴重任的时代新人[N]. 人民日报, 2024-05-12(01).

[2] 吴岩. 深入推进“人工智能 + 教育”赋能教育评价改革[J]. 中国考试, 2024(1): 1-5.

[3] 何屹松, 孙媛媛, 江光贤, 等. 人工智能评分参与高考网评一评的应用实践[J]. 中国考试, 2021(9): 40-46.

[4] 竺博, 付瑞吉, 盛志超, 等. 人工智能在教育考试评测中的应用探索[J]. 前沿科技, 2019(14): 5-9, 100.

[5] 何屹松, 徐飞, 刘惠, 等. 新一代智能网上评卷系统的技术实现及在高考网评中的应用实例分析[J]. 中国考试, 2019(1): 57-65.

[6] 王建, 王小芳. 词汇量化特征对作文机评分数的预测能力分析[J]. 语言教育, 2020, 8(3): 26-32, 39.

[7] 任瑞娟, 高莉. 人、机英语作文评分比较研究[J]. 黑龙江教育(高教研究与评估), 2018(1): 28-31.

[8] 王伟, 赵英华. 人机协同评分质量控制方法[J]. 教育考试, 2023(4): 97-104.

[9] Hwang, C.L. and Yoon, K. (1981) Multiple Attribute Decision Making: Methods and Applications. Springer-Verlag. https://doi.org/10.1007/978-3-642-48318-9

[10] 胡智丹, 田娜, 王萌. 普通话水平测试“命题说话”项计算机评测质量的考察与评价[J]. 现代语文, 2021(10): 61-67.

[11] 谭雄飞. 基于人工智能技术的智能考试评卷系统设计与应用[J]. 软件, 2024, 45(4): 147-149.

[12] Velasquez, M. and Hester, P.T. (2013) An Analysis of Multi-Criteria Decision Making Methods. *International Journal of Operations Research*, **10**, 56-66.

[13] Shannon, C.E. (1948) A Mathematical Theory of Communication. *The Bell System Technical Journal*, **27**, 379-423. https://doi.org/10.1002/j.1538-7305.1948.tb01338.x

[14] Diakoulaki, D., Mavrotas, G. and Papayannakis, L. (1995) Determining Objective Weights in Multiple Criteria Problems: The Critic Method. *Computers & Operations Research*, **22**, 763-770. https://doi.org/10.1016/0305-0548(94)00059-H