Research on a Multi-dimensional Evaluation Model Based on an Intelligent Scoring Algorithm Using Entropy Weighting

Nuo Chen *

Rocket Force University of Engineering, Xi'an, China

* Corresponding Author Email: 3261545527@qq.com

Abstract. This paper focuses on the research of intelligent scoring algorithms, constructing a multi-dimensional evaluation model and intelligent scoring scheme based on the entropy weight method. The study first analyses the distribution characteristics of evaluation data from human reviewers and two AI algorithms, establishes a statistical model for evaluation data, and reveals the patterns of differences in scoring consistency across different question types. Subsequently, an evaluation indicator system is constructed based on accuracy, stability, and adaptability. Information entropy is used to quantify indicator variability and weights, establishing a comprehensive evaluation model that confirms the performance degradation of AI algorithms in subjective question scenarios. The study was then expanded to six academic disciplines, constructing a discipline-weighted scoring model to further analyse and compare the scoring effectiveness of the two AI algorithms across different disciplines. Finally, an intelligent scoring scheme was designed by combining error thresholds with dynamic weights based on the entropy weight method, validating the model's effectiveness across multiple disciplines and providing a data-driven objective evaluation framework for intelligent educational scoring.

Keywords: Entropy weight method; intelligent scoring algorithms; comprehensive evaluation model; error threshold.

1. Introduction

In the context of the intelligent transformation of educational evaluation, assessing the accuracy and reliability of intelligent scoring algorithms has become a critical issue. This paper focuses on a comparative study between human scoring and two AI algorithms, establishing a multi-dimensional evaluation system centred on the entropy weight method [1-2]. The study first conducts a statistical analysis of the distribution characteristics of human review data and data from the two AI algorithms, establishing a review data model based on basic statistics to reveal the patterns of differences in scoring consistency across different question types. Secondly, an evaluation indicator system is constructed from three dimensions: accuracy, stability, and adaptability. The entropy weight method is used to determine the weights of each indicator, forming a comprehensive evaluation model [3]. Furthermore, the evaluation dimensions are extended to multi-disciplinary scenarios, and a subject-level weight calculation model is constructed based on the entropy weight method, achieving the construction of a multi-level evaluation system from question types to disciplines [4]. Finally, an intelligent scoring scheme was designed in conjunction with error thresholds, dynamically adjusting cross-question type consistency weights using the entropy weight method to form a closed-loop mechanism of 'AI scoring - consistency verification - human intervention' [5], providing a scientifically sound and operationally feasible solution for the practical application of intelligent scoring in the education field.

2. Review Data Statistics Model

2.1. Model Establishment

First, this paper analyses the distribution characteristics of manually reviewed data and two types of AI algorithm-reviewed data, establishes a statistical model for review data based on basic statistics,

and calculates the basic statistics of the three review methods, including mean, variance, standard deviation, median, skewness, and kurtosis.

2.2. Model Solution and Analysis

This paper calculates the basic statistics of three evaluation methods. The results show that the distribution characteristics of the scoring data for the four questions are analysed as follows:

Sub-object 1: Manual reviews are statistically similar to the two AI algorithms, indicating high review consistency; data skewness and kurtosis are close to 0, with a symmetrical and slightly flat distribution. Sub-object 2: Human and AI1 data are similar, with a symmetrical and flat distribution; AI2 has a higher mean and left skewness, indicating more lenient grading. Sub-object 3: AI2 has significantly higher mean and variance than human and AI1, indicating more lenient and dispersed scoring; all three methods have skewness > 0.5 and right skewness, with scores concentrated in the low-score range. Sub-object 4: AI1 is highly similar to human data, with high skewness and kurtosis; AI2 has high variance and a dispersed distribution, with reduced skewness and weakened right skewness, near-zero kurtosis, and a distribution approaching normal distribution.

Sub-object 1 is a fill-in-the-blank question, while sub-objects 2, 3, and 4 are short-answer questions. Through overall comparative analysis, the distribution characteristics of human grading and the two AI algorithms can be summarized: when grading fill-in-the-blank questions, the three scoring methods are highly consistent, with distributions approaching symmetry and flatness; when grading short-answer questions, human grading is closer to AI 1, while AI 2 shows higher scores in the high-score segment and greater dispersion.

3. Comprehensive Evaluation Model

3.1. Model Establishment

This section continues to select different evaluation perspectives to construct an evaluation indicator system for the ‘intelligent grading algorithm’ and comprehensively evaluate the two types of artificial intelligence algorithms using the given data. For the four sub-objects, statistical measures such as root mean square error and mean absolute error were calculated from three evaluation perspectives: accuracy, stability, and adaptability. The weights of each indicator are analyzed using the entropy weight method, and a comprehensive evaluation model for the intelligent grading algorithm is designed to evaluate and compare the artificial intelligence grading algorithms.

3.1.1. Evaluation indicator system

(1) Accuracy is mainly measured by root mean square error (RMSE), mean absolute error (MAE), Pearson correlation coefficient, complete agreement rate, and average deviation.

The average deviation indicates the direction of systematic bias between the AI algorithm's scores and human scores, reflecting the model's overall tendency to overestimate or underestimate.

$$\begin{cases} ME = \dfrac{1}{M}\sum_{i=1}^{p}\sum_{j=1}^{q}(s_{Ai} - s_{Hj}) \cdot \omega_{ij} \\ \omega_{ij} = n_{Ai} \cdot n_{Hj} \\ M = \sum \omega_{ij} \end{cases} \tag{1}$$

Where ME represents the average deviation, p represents the total number of AI scoring samples, q represents the total number of manual scoring samples, S_{Ai} represents the ith score given by the AI algorithm, S_{Hj} represents the j score given by manual scoring, ω_{ij} represents the cross weight of AI score i and manual score j, n_{Ai} represents the number of samples given the AI

score i, n_{Hi} represents the number of samples given the manual score j, and M represents the total weight.

(2) Stability is mainly measured by the standard deviation of errors and the tolerance error consistency rate, where the tolerance error consistency rate refers to the ratio of the number of samples within the allowable error range between the AI review algorithm score and the manual score to the total number of samples, reflecting the acceptability of the algorithm score under the rules. The higher the tolerance error consistency rate, the better the stability of the AI algorithm within the allowable error range.

$$TEAR = \frac{n_{tear}}{n}, \quad n_{tear} = n(|A_{ki} - H_i| < t) \tag{2}$$

Where n_{tear} represents the number of samples where the difference between the AI and manual review results is within the acceptable error range, and t represents the error threshold.

(3) Adaptability

Cross-question type consistency refers to the adaptability of the intelligent review algorithm's scoring performance across different sub-objects, reflecting the algorithm's adaptability to different types of questions.

$$\begin{cases} MAE' = \dfrac{MAE}{FULL} \\ \mu' = \dfrac{\sum_{i=1}^{m} MAE_i'}{m}, \sigma' = \sqrt{\dfrac{\sum_{i=1}^{m} (MAE_i' - \mu')^2}{m-1}} \\ CV = \dfrac{\sigma'}{\mu'} \times 100\% \end{cases} \tag{3}$$

Where MAE represents the average absolute error of each sub-object, $FULL$ represents the full score of each sub-object, MAE' represents the relative average absolute error, m represents the number of sub-objects, μ' represents the mean value of the relative average absolute error, and σ' represents the standard deviation of the relative average absolute error. CV Represents the coefficient of variation, which is used to measure the volatility of the errors of each sub-object. The lower the coefficient of variation, the higher the consistency across question types.

3.1.2. Comprehensive evaluation model

Based on the evaluation indicator system of the 'intelligent grading algorithm,' the entropy weight method is used to determine the weights of each indicator, thereby establishing a comprehensive evaluation model for the intelligent grading algorithm. The entropy weight method is an objective weighting method that determines the weights of each indicator by calculating their information entropy. The smaller the information entropy, the greater the variability of the indicator, the more information it provides, and the higher its weight [6].

(1) Data standardization

The standardisation formulas for negative and positive indicators are shown below.

$$\begin{cases} x'_{negative} = \dfrac{\max(x) - x}{\max(x) - \min(x)} \\ x'_{positive} = \dfrac{x - \min(x)}{\max(x) - \min(x)} \end{cases} \tag{4}$$

Where x represents the evaluation indicator data, x' represents the standardized indicator data, $x'_{negative}$ represents the standardized negative indicator data, and $x'_{positive}$ represents the standardized positive indicator data.

(2) Calculation of information entropy and weights

Information entropy is a fundamental concept in information theory, used to describe the uncertainty of an information source. The more ordered a system is, the lower the entropy; the more disordered, the higher the entropy.

$$\begin{cases} p_j = \dfrac{x'_j}{\sum_{i=1}^{\lambda} x'_{ji}} \\ E_j = -\dfrac{1}{\ln(\lambda)} \sum_{i=1}^{\lambda} p_{ji} \ln(p_{ji}) \\ D_j = 1 - E_j \\ W_j = \dfrac{D_j}{\sum_{i=1}^{\lambda} D_{ji}} \times 100\% \end{cases} \tag{5}$$

Where p_j represents the weight of evaluation indicator j, x'_j represents the standardized data of indicator j, λ represents the total number of evaluation indicators j, E_j represents the information entropy of evaluation indicator j, D_j represents the coefficient of variation of evaluation indicator j, and W_j represents the weight of evaluation indicator j.

(3) Comprehensive evaluation model

$$sc_o = \sum W_j \cdot x', \quad sc = \frac{sc_o}{\max\ (sc_o)} \times 100 \tag{6}$$

Where sc_0 represents the initial score of each sub-object evaluation indicator, and sc represents the score of each sub-object evaluation indicator on a percentage basis.

Since cross-question consistency is a core dimension of algorithm evaluation, it was not considered in the analysis process described above. After obtaining the evaluation scores for each sub-object, the reasonable weight of the coefficient of variation CV was calculated using the principle of information entropy balance.

$$\begin{cases} p = \dfrac{x'}{\sum x'} \\ H_{sc_0} = -\sum (p_{sc_0} \cdot \ln(p_{sc_0})) \\ H_{cv} = -\sum (p_{cv} \cdot \ln(p_{cv})) \\ W_{cv} = \dfrac{H_{cv}}{H_{sc_0} + H_{cv}} \times 100\% \end{cases} \tag{7}$$

Where p_{sc_o} represents the weight of the evaluation index score, p_{cv} represents the weight of the coefficient of variation CV, H_{sc_o} represents the entropy of the evaluation index score of each sub-object, H_{cv} represents the entropy of the coefficient of variation CV, and W_{cv} represents the weight of the coefficient of variation CV.

After obtaining the weight of the coefficient of variation CV, the scores of each sub-object can be updated.

$$sc_0' = sc_0 \cdot (1 - W_{cv}) + CV \cdot W_{cv}, \quad sc' = \frac{sc_0'}{\max(sc_0')} \times 100 \tag{8}$$

Where sc'_0 represents the updated scores for each sub-object.

After obtaining the AI-evaluated scores for each question using the entropy weight method, a hierarchical weighting system for calculating the total subject score must be constructed to integrate the question-level scores into a subject-level comprehensive score. The weighting of each question is determined based on its percentage of the total score for the subject.

$$W_{\mathrm{BQ}i} = \frac{FULL_i}{\sum FULL_i} \tag{9}$$

Where $FULL_i$ represents the total score for question i, and W_{BQi} represents the weighting between questions. This comprehensive evaluation model can be used to calculate the comprehensive scores for AI 1 and AI 2.

$$SCORE = \sum sc' \cdot W_{\mathrm{BQ}i} \tag{10}$$

Where $SCORE$ represents the evaluation score of artificial intelligence algorithm for that course.

3.2. Model Solution and Analysis

To accurately evaluate artificial intelligence algorithms, a comprehensive evaluation model based on an evaluation indicator system was established. After calculating the weight values of each indicator using the entropy weight method, the algorithm scores were calculated on a percentage basis, as shown in Table 1.

Table 1. Preliminary Scores for Two Types of Artificial Intelligence Algorithms.

	Sub-object 1	Sub-object 2	Sub-object 3	Sub-object 4
AI 1	98.3	83	64.3	65.64
AI 2	100	52.44	0.014	0.27

The results show that for sub-object 1, both types of AI grading algorithms scored relatively high, with AI 2 performing slightly better than AI 1. However, for sub-objects 2, 3, and 4, the scores of both algorithms dropped significantly, especially for AI 2, with the lowest score reaching as low as 0.014. This indicates that AI grading algorithms exhibit significantly reduced accuracy and stability when grading subjective questions such as short-answer questions, with AI 2 showing a more pronounced decline in performance. This also reflects the poor consistency of AI algorithms across question types.

By introducing cross-question-type consistency using information entropy principles and assigning a weight of 43% to the coefficient of variation (CV), it is indicated that stability across question types is a core dimension in algorithm evaluation. By updating the scores of each sub-object using the CV weight and constructing a hierarchical weighting system for subject-level total score calculation, a subject-level comprehensive score is derived, as shown in Table 2. For the subject of physics, the evaluation results of AI 1 are better than those of AI 2.

Table 2. Comprehensive Scores for Two Types of Artificial Intelligence Algorithms.

	Sub-object 1	Sub-object 2	Sub-object 3	Sub-object 4	Overall score
AI 1	95.54	84.55	72.69	73.55	78.41
AI 2	100	68	33.68	50.69	55.44

4. Interdisciplinary Expansion and Model Upgrades

This section adds sample data from human and two types of artificial intelligence algorithms for Chinese, mathematics, English, politics, and geography, in addition to physics. Based on the model in the previous section, the evaluation process is elevated to a higher disciplinary dimension. Therefore, this section not only requires calculating the scores of the two types of artificial intelligence algorithms under different question types but also establishing a more complex global evaluation model to conduct comparative evaluations of grading effectiveness at the subject dimension level.

4.1. Discipline-based Weighted Scoring Model Establishment

4.1.1. Evaluation indicator system

This section utilises the evaluation indicator system for 'intelligent grading algorithms' established in the previous section, combined with the provided multi-subject data, to comprehensively evaluate the two types of artificial intelligence algorithms. For each question under each discipline, we calculate statistical metrics such as root mean square error, average absolute error, Pearson correlation coefficient, complete agreement rate, error standard deviation, tolerance error agreement rate, and cross-question type consistency from three evaluation perspectives: accuracy, stability, and adaptability.

4.1.2. Discipline-Layer Weight Calculation Model

Based on the evaluation metric system of the 'intelligent grading algorithm,' the entropy weight method is used to determine the weights of each metric. Additionally, cross-question type consistency is introduced using information entropy principles. Based on the evaluation models for each question in the intelligent grading algorithm, a subject-level hierarchical weight calculation model is constructed to integrate question-level scores into subject-level comprehensive scores, enabling evaluation and comparison of grading effectiveness at the subject dimension. The key formulas are consistent with those in the previous section.

4.2. Model Solution and Analysis

To precisely evaluate AI algorithms at the subject level, a subject-tiered weighting calculation model was designed to integrate question-level scores into subject-level comprehensive scores, analysing the evaluation effectiveness of each subject both horizontally and vertically. Specific score comparisons are shown in Fig. 1.

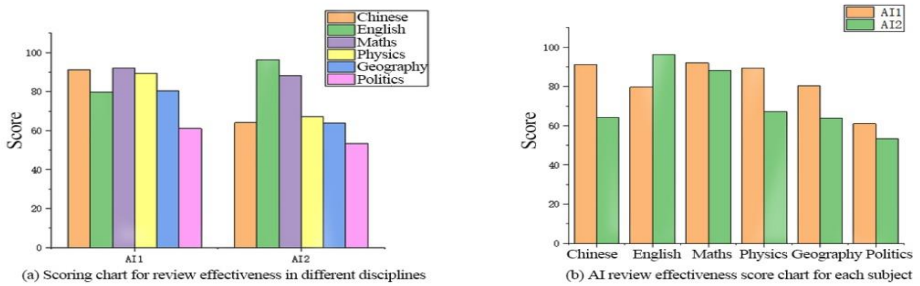

Fig 1. Review results for each subject.

As shown in Fig. 1(a), the overall evaluation results of AI 1 are better than those of AI 2. As shown in Fig. 1(b), in the subjects of Chinese language, mathematics, physics, geography, and politics, AI 1 outperforms AI 2 in terms of grading effectiveness. However, in the English subject, AI 2 demonstrates superior grading effectiveness. As concluded in the previous sections, the grading results of AI 1 are closer to human grading, while AI 2 is more lenient in scoring, resulting in significant differences from human grading. Except for English, which primarily consists of objective questions, the other subjects have a higher proportion of subjective questions, leading to higher scores from AI 1.

5. Intelligent Grading Scheme Model

This section uses two types of artificial intelligence algorithms to grade six subjects. It is necessary to fully consider the grading performance of AI 1 and AI 2 in different subjects, as well as the error thresholds for each question, to establish an intelligent grading scheme model based on error thresholds. A comprehensive, efficient, reliable, and economical intelligent grading scheme is designed, and the designed scheme is analyzed and evaluated.

5.1. Model Establishment

In the data preprocessing stage mentioned earlier, it is assumed that manual re-evaluation is conducted when the difference between the results of the two types of artificial intelligence algorithms is less than the error threshold. Therefore, this question establishes an intelligent grading scheme model based on the error threshold.

(1) Consistent artificial intelligence scoring

$$| A_{k1} - A_{k2} | < t \tag{11}$$

Where A_{k1} and A_{k2} represent the review results of AI 1 and AI 2, respectively, and represents the error threshold for each question in each subject.

At this point, the review results of the two types of AI algorithms are highly consistent. Combining the AI review effectiveness scores for each subject given in the previous section, the AI with the higher score is selected to review the questions.

(2) Inconsistent AI scoring

$$| A_{k1} - A_{k2} | \geq t \tag{12}$$

This situation indicates that the two types of AI algorithms produce inconsistent results, requiring human intervention to ensure scoring accuracy. In summary, the intelligent scoring scheme model based on error thresholds is as follows:

$$P = \begin{cases} A_{kbetter}, & if\ | A_{k1} - A_{k2} | < t \\ H, & if\ | A_{k1} - A_{k2} | \geq t \end{cases} \tag{13}$$

Where P represents the final review result of the question, H represents the manual review result, $A_{kbetter}$ represents the review result of the higher score between the two types of artificial intelligence algorithms, and t represents the error threshold of each question in each subject.

A consistency test is conducted on the two scoring methods. If the difference between the two evaluation results falls within the error threshold range, the final evaluation score is calculated using the subject-based weighting model, and the AI evaluation result with the higher score is output. If the difference exceeds the error threshold, the initial human evaluation result is output.

5.2. Model Solving and Analysis

For the evaluation results of the two types of artificial intelligence across various disciplines, an intelligent evaluation scheme model based on error thresholds was employed to derive the final output scores for each question. These scores were then compared with human evaluation results through error analysis, calculating the complete agreement rate, average deviation, and standard deviation of errors for each discipline. This process was used to validate the reasonableness and reliability of the model. And the results show that this model is highly effective in evaluating mathematics, English, and physics, especially physics. Mathematics and English have a small average deviation, but the error fluctuates greatly.

6. Summary

This study systematically analysed the performance of intelligent scoring algorithms in different question types and subjects by constructing an evaluation system centred on the entropy weight method, producing research results with both theoretical value and practical significance. The study first analysed the distribution of evaluation data from human reviewers and two AI algorithms, established a statistical model, and revealed the patterns of differences in scoring consistency across different question types. Subsequently, an evaluation indicator system was constructed based on accuracy, stability, and adaptability, and an integrated evaluation model was established using the entropy weight method, revealing that AI algorithms exhibit significantly reduced performance in subjective question scenarios. The study was then expanded to six academic disciplines, where a discipline-weighted scoring model was constructed to further analyse and compare the scoring effectiveness of the two AI algorithms across different disciplines. Finally, an intelligent scoring scheme was designed based on error thresholds, and the Entropy Weighting Method was used to balance the consistency weights across question types, validating the model's effectiveness in multi-disciplinary scenarios. This provides a data-driven, objective evaluation framework for intelligent educational scoring.

References

[1] Xiao Guoliang, Ma Lei, Yuan Feng, et al. Evaluation of the Effectiveness of Intelligent Scoring Technology Applications [J]. China Examination, 2023, (10): 17-27.

[2] Wang Xuyong. Research on Intelligent Customer Service Scoring Methods Based on Entropy Weighting [J]. Computer Programming Techniques and Maintenance, 2023, (02): 121-123+134.

[3] Zhou Yi, Song Hongwen, Tian Shaoai, et al. Research on Live Streaming E-commerce Service Quality Evaluation Model and Regulatory Strategy Based on Entropy Weight Method [J]. Business Exhibition Economy, 2022, (14): 74-76.

[4] Xiang Bazhuoma, Wang Zhenzhen, Chang Hongsheng, et al. Research on Intelligent Scoring of Subjective Questions in Pharmacology Examinations Based on Large Language Models [J]. Chinese Medical Education Technology, 2024, 38 (05): 572-579.

[5] Wang Cixiao, Xu Junyan, Guo Liming, et al. Research on an Evaluation Framework for Multi-scenario Human-Computer Collaborative Online Teaching: An Analysis Based on the Analytic Hierarchy Process and Entropy Weight Method [J]. Modern Educational Technology, 2023, 33 (01): 74-82.

[6] Zhang Tian, Yan Hongcan. Optimisation and Application of Multi-Attribute Decision-Making Algorithms Based on Entropy Weighting [J]. Journal of North China University of Science and Technology (Natural Science Edition), 2022, 44 (01): 82-88.