# Exploring the Feasibility of AI-Based Analysis of Elementary Chemistry Science Fair Reports in Taiwan

**Chao-Min Hu [1*], Chin-Cheng Chou [1]**

*[1] Department of Science Education, National Taipei University of Education, Taipei 106, Taiwan*

*jimmyhu19908047@gmail.com*

**Abstract**

Science fairs in Taiwan serve as crucial platforms for fostering students' scientific literacy, inquiry competence, and creativity. However, teachers often struggle to guide elementary students in developing feasible yet innovative research projects due to limited time and ambiguous evaluation criteria. This study explores the feasibility of using ChatGPT-4o—a large language model—to analyze and predict award-winning reports in the elementary chemistry division of Taiwan's National Science Fair. A total of 38 reports from 2021 to 2024 were analyzed, with 29 reports (2021–2023) used for model training and 9 reports (2024) for prediction testing. The model successfully identified all three top-award reports from 2024, achieving 100% prediction accuracy. Repeated trials indicated high test–retest consistency under identical prompts. The results suggest that constructing a feature-based evaluation model enhances the reliability and interpretability of AI-assisted assessments. This stu dy highlights the potential of AI tools to complement human judgment in science fair evaluation, offering new directions for integrating AI into inquiry-based science education and assessment practices.

**Keywords:** AI-assisted assessment, ChatGPT, elementary chemistry reports, inquiry-based learning, science fair.

## Introduction

### 1.1 Background of Science Fairs in Taiwan

Since their establishment in 1955, Taiwan's science fairs have become essential mechanisms for promoting students' inquiry and problem-solving abilities [1,2]. The fairs follow a three-tier progression system—school, municipal, and national levels—creating a structured pathway for project advancement [3]. The elementary division currently covers eight disciplinary categories [4], with top-performing projects often linked to academic recognition and secondary school admissions. These competitions therefore serve both educational and selection functions.

Despite their long history and institutional support, disparities in mentoring quality and resource allocation persist. Teachers often face difficulties providing consistent guidance, especially when assisting students with limited disciplinary knowledge but strong creative potential. Consequently, establishing more objective and transparent evaluation mechanisms has become essential to ensure fairness and improve instructional alignment.

### 1.2 Challenges in Guiding Elementary Student Research

Guiding elementary students through independent research projects is both pedagogically demanding and time-intensive [5,6]. Students at this level often exhibit high imagination but limited understanding of experimental control, data interpretation, or logical structure. Teachers must therefore strike a delicate balance between offering necessary scaffolding and maintaining student ownership of the work—a

dilemma reinforced by national regulations requiring students to complete and defend their projects independently [5].

Drawing from Vygotsky's Zone of Proximal Development [7,8], appropriate scaffolding can help students transcend their current competencies. However, interdisciplinary projects—such as those combining chemistry, physics, and environmental science—often exceed a single teacher's expertise, highlighting the need for collaborative or technology-assisted approaches [6].

### 1.3 AI as a Collaborative Assistant in Science Education

Artificial intelligence (AI) has recently emerged as a potential "second mentor" in education [9,10]. Large language models (LLMs) can support teachers by accelerating literature reviews, suggesting research ideas, and simulating rubric-aligned feedback [11]. Prior studies have shown that AI can identify learning gaps, provide formative feedback, and enhance scoring consistency [9,10]. Within science education, AI-assisted evaluation could reduce teachers' workload while maintaining fairness and transparency in assessment.

Recent evidence also supports the concept of human–AI collaboration. According to Shneiderman's Human-Centered AI (HCAI) framework [12], AI should be designed to augment rather than replace human capacities, emphasizing safety, explain ability, and empowerment. In both educational and industrial contexts, AI-augmented teams have demonstrated improved performance, creativity, and collaboration quality [13,14].

The integration of AI into science fairs presents an opportunity to bridge the gap between student creativity and systematic evaluation. By analyzing past award-winning projects, AI can help identify latent patterns and features associated with successful reports—insights that teachers can use to design instructional scaffolds and guide students' project development [9,11].

### 1.4 Research Purpose and Significance

This study investigates whether an AI model, specifically ChatGPT-4o, can accurately identify and predict award-winning reports in Taiwan's National Elementary Science Fair (Chemistry Division). The research aims to:

A. Develop a feature-based AI model derived from historical science fair reports.

B. Evaluate the model's predictive accuracy and stability.

C. Explore the implications of AI-assisted assessment for supporting teacher decision-making in science education.

By examining the feasibility of AI-based report evaluation, this study contributes to emerging discussions on AI-supported inquiry assessment, providing empirical evidence for how AI can function as a formative evaluation assistant in science learning environments.

## 2. Methodology

This study used OpenAI's GPT-4o, released on 13 May 2024, which operates at approximately twice the speed of GPT-4 at roughly 50% lower cost, and remains accessible to free-tier users with limited quotas [15,16]. According to official OpenAI documentation, free users are limited to a small number of messages within a rolling five-hour window [16], Plus subscribers can send approximately 80 messages per three-hour window [17], and Team/Pro workspaces have higher caps [14]. Selecting GPT-4o therefore enables secondary school teachers to replicate the analytic pipeline without subscription fees,

aligning with the practitioner-oriented objectives of this study.

To assess output stability, each prompt was re-run three times between June and July 2025 under identical parameter settings (e.g., temperature, random seed), and the results were compared for consistency. This procedure follows the test–retest reliability framework proposed by Mondal et al., who reported a Pearson correlation coefficient of r = .71 for ChatGPT-3.5 in statistical-test recommendation tasks, and extends this approach to GPT-4o [18]. Upon completion of all report processing, GPT-4o re-examined and summarized the main writing features of award-winning reports based on the entire dataset.

While this study primarily focuses on developing and analyzing an AI-assisted evaluation model, the model itself was intentionally designed to be applicable across different subject domains rather than restricted to a specific discipline. By training the model on datasets containing student work from multiple subjects, we aimed to identify common evaluative features that transcend disciplinary boundaries. Interestingly, the extracted features showed strong correlations with domain-specific patterns, and models trained on diverse data often yielded more consistent and accurate scoring outcomes than those based solely on standardized rubrics. These findings suggest that cross-disciplinary AI models have the potential to provide teachers with data-informed insights for evaluating and guiding student learning across subjects [19,20].

**Step 1: Data Collection and Preparation**

The data for this study were drawn from publicly available chemistry projects in the Taiwan National Science Fair, which annually announces approximately 10–15 reports, all of which are submitted by winners of local competitions in each county and city. In Taiwan's ranking system, awards are granted to the first, second, and third place winners, followed by honorable mentions, referred to locally as Merit Awards. From this pool, we collected 38 chemistry reports from 2021 to 2024, as shown in Table 1. The sample included the top three national award-winning reports for each year (if multiple reports shared the same rank, all were included) as well as non-awarded projects. Merit Awards were excluded because they represent an intermediate ranking between winners and non-winners, which could introduce classification ambiguity in system analysis.

To confirm feasibility, the model was trained on reports from 2021–2023 and tested on those from 2024. Because the ChatGPT-4o version used in this study (June–July 2025) allowed a maximum of ten files to be uploaded per run, the analysis was conducted in batches of up to ten reports. This design consisted of three independent rounds, with awarded and non-awarded works compared within the same year. During data preparation, cover pages and any content revealing award status were removed to ensure complete blinding. In addition, persistent conversation history was disabled during analysis to avoid influence from prior data. This design minimizes evaluation bias and ensures replicability for future research.

**Table 1 2021-2024 National Chemistry Exhibition Reports**

| Number of entries(Number of selected entries) | Top three(First, second, and third place) | honorable mentions | Non-awarded | Total selection |
|---|---|---|---|---|
| 2021 | 3 (randomly selected from 5) | 4 (Not used) | 7 (randomly selected from 11) | 10 |
| 2022 | 3(randomly selected from 5) | 3 (Not used) | 7 (all 7 Selected) | 10 |
| 2023 | 3 (randomly selected from 4) | 3 (Not used) | 6(all 6 Selected) | 9 |
| 2024 | 3(randomly selected from 4) | 3 (Not used) | 6(all 6 Selected) | 9 |

**Step 2: Model Training**

Reports were entered into ChatGPT-4o for analysis of their strengths and weaknesses (e.g., research motivation, methodology, data processing, and writing quality). To improve output quality, the prompts were iteratively refined within the same session.

Because the ChatGPT-4o version used in the experimental tests (June–July 2025) allowed a maximum of ten files to be uploaded per run, the training was conducted in three cumulative rounds:

Round 1: Training with up to 10 reports from 2021 only, analyzed according to the official National Science Fair scoring criteria (research motivation, method design, data processing, and report writing).

Round 2: Training with up to 10 reports from 2021 and 2022, incorporating the 2022 data into the previous round and establishing a revised feature set of award-winning standards.

Round 3: Training with up to 10 reports from 2021, 2022, and 2023, further extending the dataset to refine and validate the feature model.

For each round, the system was asked to:

1. Compare the advantages and disadvantages of award-winning reports and non-award-winning reports from the 2021–2023 in the following aspects, as specified in the official evaluation standards:

A. Research motivation

B. Method design

C. Data processing

D. Report writing

2. Summarize the main writing features that appeared in the award-winning reports.

In Rounds 2 and 3, two additional system prompts were added: (a) repeat the results of the previous round before analyzing the new set of reports, and (b) after completing all provided reports, re-identify and summarize the key writing features of the award-winning reports across the entire dataset.

**Step 3: Building the Award-Winning Report Feature Model**

Based on the results of the three training rounds, key writing features that consistently appeared in award-winning reports were identified. These features included strengths in research motivation, method design, data processing, and report writing.

The recurring characteristics were then synthesized to form a preliminary “award-winning report

feature model." This model served as the reference framework for evaluating new reports in the subsequent prediction and validation stage.

**Step 4: Prediction and Validation**

Using the feature model constructed from the 2021–2023 reports, we conducted a prediction test on the award-winning reports from the 2024 National Chemistry Fair (a total of nine reports).

This step aimed to determine whether the award-winning report feature model could accurately identify winning projects in a 2024 dataset, while also assessing the reliability of the system under repeated testing conditions.

The direct human evaluation of students' science fair reports was conducted strictly in accordance with the official scoring criteria announced by the National Taiwan Science Education Center [5]. In contrast, the AI model was independently developed by refining the official rubric into four analytical dimensions—scientific reasoning, experimental design, data interpretation, and presentation clarity—derived from the core components of the original framework. Because the written reports did not include presentation-related elements, the modeling process focused solely on the content-based aspects of evaluation. This approach enabled the model to preserve alignment with the official standards while enhancing its analytical precision and generalizability.

## 3.Results

### 3.1 Prediction Accuracy

Using science fair reports from 2021–2023 to build a predictive model yielded stronger results for forecasting 2024 award-winning projects. To ensure blinding, the cover pages and acknowledgments of award-winning reports were removed before analysis. Due to system constraints (maximum of 10 reports per run), the training process was conducted year by year. In each round, the model was re-instructed to analyze all available reports according to the four official evaluation criteria: (A) research motivation, (B) method design, (C) data processing, and (D) report writing. The evaluation of students' science fair reports was conducted strictly in accordance with the official scoring criteria announced by the National Taiwan Science Education Center [5]. Building on this framework, the model further refined the criteria into four analytical dimensions to enhance the consistency and interpretability of AI-based evaluation, in line with established principles of performance-based and formative assessment [19,20].

Table 2 summarizes the predictive accuracy under three evaluation conditions, all conducted in June 2025. When using only the official evaluation rubric (Test 1), the model correctly identified 2 out of 3 award-winning reports (success rate: 67%). In contrast, when predictions were based on the round 3 trained model (Test 2), accuracy improved to 100% (3/3). Finally, when the round 3 model was re-established and reapplied (Test 3), the system again achieved 100% success (3/3). Across all tests, the total number of reports considered remained nine.

**Table 2. AI Prediction Accuracy Across Different Evaluation Conditions(June 2025)**

| Evaluation Condition– All conducted in June 2025 | Predicted 2024 award-winning reports (Success Rate) | Total Reports |
|---|---|---|
| Test 1 – Using official rubric only | 2/3 (67%) | 9 |
| Test 2 – Using round 3 trained model | 3/3 (100%) | 9 |
| Test 3 – Re-establishing round 3 model | 3/3 (100%) | 9 |

### 3.2 Recurrent Features of Award-Winning Reports

In addition to higher predictive accuracy, the third round of training revealed recurring features that consistently characterized award-winning reports:

1. Creative questions inspired by daily life – Projects often originated from real-world problems or observations.
2. Well-planned experiments with clear variables and creative tools – Designs included control and test groups, frequently using self-made devices or innovative methods.
3. Strong data analysis with clear explanations and real-life applications – Data were presented transparently, results explained logically, and applications linked to daily life.
4. Clear reports with strong writing and visuals – Reports were well-organized, easy to read, and supported with charts or pictures.
5. Creative integration of ideas from multiple fields – Knowledge from different subjects was combined to generate novel insights or propose improvements.

Overall, these results indicate that while the official evaluation rubric provides a useful baseline for prediction, its accuracy was limited when applied directly. In contrast, the trained AI model was able to capture additional latent features and patterns from prior years' reports that were not explicitly represented in the rubric. The iterative training process allowed the system to refine its recognition of nuanced characteristics—such as innovative approaches, depth of analysis, and clarity of presentation—that often distinguish award-winning projects. This highlights the potential of AI-assisted evaluation to complement traditional rubrics by detecting subtle, multidimensional qualities beyond explicit scoring guidelines.

## 4. Discussion and Conclusion

### 4.1 Discussion

This study demonstrates that constructing a **feature-based evaluation model** enhances AI performance in recognizing and predicting award-winning science fair reports. Compared with direct application of the official rubric, the trained model achieved higher accuracy and reliability. The iterative process allowed the AI to detect nuanced elements—such as coherence, creativity, and contextual reasoning—that are difficult to quantify but crucial in authentic student research.

Repetition tests indicated that GPT-4o maintained a high level of response consistency within three identical prompt iterations, aligning with prior findings on LLM stability [18]. Beyond this threshold, slight variations emerged due to stochastic generation processes. These fluctuations underscore the need for clear and detailed prompts when employing AI for evaluative purposes.

A limitation observed in this study was the AI's occasional overreliance on visual proxies (layout consistency, image quality) rather than genuine visual interpretation. Because GPT-4o primarily processed text, its inference of "clear visuals" likely resulted from statistical associations rather than semantic understanding. Future models integrating vision–language architectures may better capture multimodal aspects of report quality.

### 4.2 Educational Implications

For educators, the findings suggest that AI can serve as a diagnostic and formative assessment assistant rather than a grading authority. By identifying characteristic features of exemplary reports, teachers can design more effective scaffolding for student research projects. The AI-generated feature

summaries may also assist in professional development by illustrating how high-quality reports demonstrate creativity, logical reasoning, and coherence.

### 4.3 Limitations and Future Work

Although the feature model performed consistently within the tested dataset, generalization remains limited by the small sample size and subject specificity. Expanding the corpus to include other disciplines (e.g., biology, physics) and multimodal data (e.g., oral defenses, presentation slides) could improve robustness. Additionally, longitudinal studies could examine how AI-supported feedback influences students' inquiry competence and teachers' assessment literacy.

### 4.4 Conclusion

This research provides empirical evidence that AI-based evaluation, when structured around feature extraction and iterative refinement, can effectively identify the characteristics of award-winning science fair reports. The study confirms the feasibility and educational potential of AI-assisted assessment in elementary science education. While human judgment remains indispensable, integrating AI as a complementary tool can enhance fairness, reduce teacher workload, and promote sustained engagement in inquiry-based learning.

### Reference


[1] Wu, C.-S., J. Natl. Taiwan Normal Univ.: Educ., 55(2), 1–34 (2010).

[2] National Research Council, A Framework for K-12 Science Education: Practices, Crosscutting Concepts, and Core Ideas, National Academies Press, Washington, DC (2012).

[3] Ministry of Education, National Primary and Secondary School Science Fair Implementation Guidelines, Ministry of Education, Taipei, Taiwan (2022).

[4] Ministry of Education (Taiwan), Guidelines for the National Primary and Secondary School Science Fair (revised on Jan 16, 2025), Ministry of Education, Taipei (2025).

[5] Bogden, M.; Wilkerson, N., "A Teacher's Guide to Science Fair," Vivify STEM Blog, Houston, TX (2023).

[6] Edutopia, "Connecting Across Disciplines in Project-Based Learning," Edutopia, San Rafael, CA (2024).

[7] Simply Psychology, "Zone of Proximal Development," Simply Psychology (2025).

[8] Vygotsky, L. S., Mind in Society: The Development of Higher Psychological Processes, Harvard University Press, Cambridge, MA (1978).

[9] Holmes, W.; Bialik, M.; Fadel, C., Artificial Intelligence in Education: Promises and Implications for Teaching and Learning, Center for Curriculum Redesign, Boston, MA (2019).

[10] Luckin, R.; Holmes, W.; Griffiths, M.; Forcier, L. B., Intelligence Unleashed: An Argument for AI in Education, Pearson, London, UK (2016).

[11] Monteith, B., "AI: The Future of Personalized Mentorship in Science Fairs," Medium, San Francisco, CA (2024).

[12] Shneiderman, B., Human-Centered AI, Oxford University Press, New York, NY (2022).

[13] Kong, X.; Fang, H.; Chen, W.; Xiao, J., Humanit. Soc. Sci. Commun., 12, 821 (2025).

[14] OpenAI, "What is the Message Cap on ChatGPT Team?" OpenAI Help Center, San Francisco, CA (2024).

[15] Buchanan, N., "Microsoft-Backed OpenAI Unveils Most Capable AI Model, GPT-4o," Investopedia, New York, NY (2024).

[16] OpenAI, “ChatGPT Free Tier — Usage Limits,” OpenAI Help Center, San Francisco, CA (2025).

[17] OpenAI, “What is ChatGPT Plus? — Usage Limits,” OpenAI Help Center, San Francisco, CA (2025).

[18] Mondal, H.; Mondal, S.; Mittal, P., Perspect. Clin. Res., 15(4), 178–182 (2024).

[19] Brookhart, S. M., How to Create and Use Rubrics for Formative Assessment and Grading, ASCD, Alexandria, VA (2013).

[20] Wiggins, G., Educative Assessment: Designing Assessments to Inform and Improve Student Performance, Jossey-Bass, San Francisco, CA (1998).