airiti

Enhancing the Accuracy of LLM-Based Automatic Scoring: A Study of Zero-Shot, Few-Shot, and Self-Consistency Strategies in Socio-Scientific Issues

Chi-Jung Sui

OVERVIEW

LLMs combined with few-shot prompting and self-consistency achieve high accuracy inautomatically scoring student argumentation on socio-scientific issues.

KEY POINTS

- Few-shot outperforms zero-shot across models.
- Self-consistency benefits vary by model architecture.
- Evidence scoring less sensitive to prompting.
- High accuracy suitable for formative assessment.
- Strategic prompting transforms LLMs into specialized tools.

INTRODUCTION

Argumentation regarding socio-scientific issues (SSIs) plays a critical role in science education by connecting scientific knowledge to broader ethical, social, and political considerations (Sadler, 2004). As complex problems like climate change dominate public discourse, students must develop abilities to examine multiple viewpoints, assess evidence, and construct well-informed arguments. Traditional assessment of such complex argumentation has been labor-intensive, but large language models (LLMs) now offer affordable, adaptable alternatives for nuanced evaluation (Farrokhnia et al., 2024). Through prompt engineering techniques like zero-shot, few-shot, and self-consistency methods, LLMs can perform sophisticated assessments without extensive fine-tuning (Lee et al., 2024). This study examines how different prompting strategies affect LLM performance in scoring SSI arguments across multiple state-of-the-art models.

FINDINGS

The study evaluated 238 eighth-grade student arguments on energy-related SSIs using GPT-4.1, Gemini-2.5, and O4-Mini.

As shown in Figure 1, few-shot prompting consistently outperformed zero-shot across all models. For claim evaluation, improvements ranged from 5.9% to 8.1%, with Gemini-2.5 achieving over 90% accuracy. Justification scoring improved by 3.1% to 5.6%, while evidence evaluation showed minimal differences, suggesting objective factual assessments require less exemplar guidance. Notably, O4-Mini achieved the highest justification accuracy at 87.8%, demonstrating that model size doesn't necessarily correlate with specialized task performance.

airiti

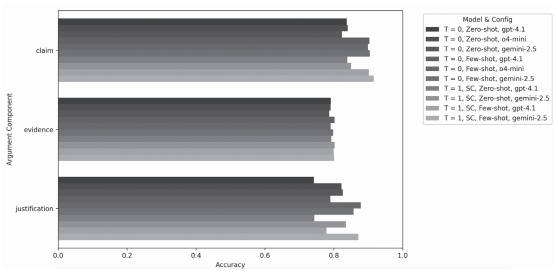

Figure 1. Accuracy by model and configuration

The self-consistency method revealed model-dependent variations. Gemini-2.5 with self-consistency achieved the highest accuracy across components, particularly for claims (91.6%) and justification (87.1%), while GPT-4.1 showed negligible improvements. Figure 2 shows Cohen's Kappa coefficients for inter-rater reliability. Evidence evaluation demonstrated the strongest agreement, while claim evaluation showed the most variability (0.227–0.457). Combined few-shot and self-consistency approaches achieved accuracy exceeding 80% across most conditions.

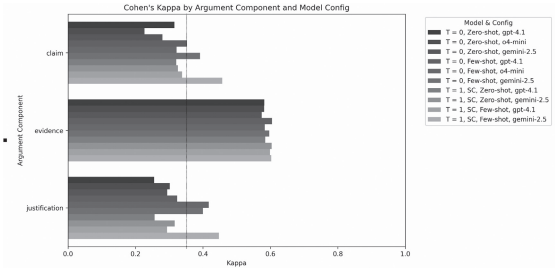

Figure 2. Cohen's Kappa by model and configuration

TAKEAWAYS

This research demonstrates that strategic prompt engineering can transform general-purpose LLMs into effective tools for educational assessment without costly fine-tuning (Latif & Zhai, 2024). The findings have significant implications for formative assessment, where immediate feedback serves as scaffolding for student development (Wilson et al., 2024). The component-specific analysis enables targeted interventions—identifying when evidence lacks relevance or when justifications merely repeat without logical bridges. While current accuracy may not meet high-stakes assessment requirements, it offers opportunities for personalized learning pathways (Yang et al., 2025). The cost-effectiveness of this approach democratizes access to AI-supported assessment for resource-limited institutions. Future research should explore transferability across domains, grade levels, and cultural contexts.

airiti

科學教育學刊
2025, 第三十三卷第三期, 265-286
DOI:10.6173/CJSE.202509_33(3).0003

Contemporary Journal of Science Education
2025, 33(3), 265-286

Enhancing the Accuracy of LLM-Based Automatic Scoring: A Study of Zero-Shot, Few-Shot, and Self-Consistency Strategies in Socio-Scientific Issues

Chi-Jung Sui

Department of Learning and Materials Design, University of Taipei

Abstract

This study investigates the efficacy of large language models (LLMs) for the automated scoring of students' argumentation on socio-scientific issues (SSIs), focusing on the distinct components of claim, evidence, and justification. We compared the performance of zero-shot, few-shot, and self-consistency prompting strategies across three models: GPT-4.1, Gemini-2.5, and O4-Mini, using a dataset of 238 student arguments on energy-related topics. The findings reveal that few-shot prompting significantly outperforms zero-shot prompting, consistently improving scoring accuracy for the more nuanced components of claims and justifications across all models. This underscores the importance of providing concrete examples to align the models' evaluative reasoning with complex scoring criteria. Conversely, the self-consistency method demonstrated model-dependent effectiveness; it reliably enhanced the accuracy of Gemini-2.5 but provided negligible benefit for GPT-4.1, suggesting that a model's architecture influences its susceptibility to such deliberative prompting. From a practical standpoint, the combination of few-shot prompting and self-consistency achieved high accuracy levels, highlighting the potential of LLMs as powerful tools for formative assessment. These methods offer a low-cost, scalable alternative to manual grading, capable of providing students with immediate feedback in adaptive learning environments. However, the observed error rates warrant caution for high-stakes summative applications. In conclusion, this study demonstrates through strategic prompt design that LLMs can be transformed into effective and professional tools for assessing argumentation on socio-scientific issues, and offers concrete approaches for real-time diagnosis and individualized scaffolding in SSI instruction.

Key words: Argumentation, Automatic Scoring, Large Language Model, Socio-Scientific Issue

* Corresponding author: Chi-Jung Sui, chrosui@gmail.com; ORCID: 0000-0003-4559-0641
Received: 2025/6/22, Revised: 2025/7/22, Accepted: 2025/7/29, Available Online: 2026/1/26

airiti

Introduction

Argumentation regarding socio-scientific issues (SSIs) occupies a critical place in science education, connecting scientific knowledge to broader ethical, social, and political considerations. In today's world, where complex problems like climate change and energy policy dominate public discourse, it's increasingly important for students not only to grasp scientific concepts but also to engage thoughtfully in debates and critical reasoning. By integrating SSIs into science curricula, educators encourage students to examine multiple viewpoints, assess varying evidence, and develop well-informed, rational arguments (Sadler, 2004).

Educational environments are currently experiencing transformative changes due to rapid advances in artificial intelligence (AI). Traditionally, AI applications in education required extensive data collection, meticulous manual labeling, and the creation of specialized, task-specific models, making them both expensive and cumbersome (Akhmadieva et al., 2023; Guo et al., 2024). However, the emergence of large language models (LLMs) has radically simplified this process, offering educators and researchers affordable, adaptable tools. These models facilitate nuanced evaluations of students' argumentative abilities regarding SSIs, providing assessments that were previously labor-intensive or impractical (Farrokhnia et al., 2024; van den Berg & du Plessis, 2023).

Automated assessment systems represent a key aspect of modern educational technology, essential for personalized and adaptive learning environments (Jia et al., 2024). Such systems offer students instant and individualized feedback, helping them swiftly identify and rectify misunderstandings, ultimately enhancing their analytical and problem-solving skills (Boonstra, 2025). Unlike early automated scoring tools, which were rigid and limited in functionality, contemporary LLMs allow for more sophisticated evaluation methods through careful prompt engineering. Techniques such as zero-shot and few-shot prompting enable LLMs to perform complex tasks by following straightforward instructions or minimal examples (Boonstra, 2025; U. Lee et al., 2024). Additionally, methods like self-consistency—derived from the chain of thought (CoT) approach—further boost reliability by combining multiple logical reasoning pathways and selecting the most frequent or robust outcome (G.-G. Lee et al., 2024).

Despite these advances, a significant gap remains concerning automated assessments of argumentation on SSIs. This form of argumentation involves not only making claims but also using concrete evidence and engaging in sophisticated dialectical reasoning. Students must apply interdisciplinary knowledge, drawing on natural sciences, environmental studies, and social sciences, demonstrating their ability to navigate and address complex real-world issues beyond rote memorization. However, existing research suggests that while LLMs perform well on simpler tasks, their effectiveness can decline significantly with more intricate and varied problems (Chang et al., 2024).

This study aims to bridge this gap by examining the effectiveness of LLMs and prompt engineering strategies in assessing student arguments about SSIs. Specifically, this study evaluates and compares the accuracy of zero-shot and few-shot prompting strategies across three distinct LLMs: GPT-4.1, Gemini-2.5, and O4-Mini. Furthermore, it examines whether applying the self-consistency method can further enhance the reliability and

airiti

accuracy of automated scoring for each of these models. Thus, the research questions guiding this inquiry are:

(1) How do the accuracies of zero-shot and few-shot prompting strategies compare across different LLM models (GPT-4.1, Gemini-2.5, and O4-Mini) for automatically scoring students' socio-scientific arguments?

(2) To what extent does applying a self-consistency method improve the scoring accuracy and reliability of these models when used with zero-shot and few-shot prompting?

Literature Review

Argumentation in SSIs

Argumentation, fundamentally, is an interactive and reflective process through which individuals or groups develop claims, support these claims with relevant evidence, and articulate justifications that link evidence coherently to the claims made (Erduran et al., 2004). Rather than being merely persuasive, argumentation serves as an intellectual activity emphasizing logical coherence, evidence-based reasoning, and dialectical engagement. In educational contexts, it fosters critical thinking, deepens conceptual understanding, and promotes thoughtful deliberation (Driver et al., 2000).

Within the educational landscape, SSIs have increasingly gained attention as valuable contexts for cultivating learners' argumentation skills (Chien & Chang, 2015; Sadler et al., 2007; Wahono et al., 2021; Zeidler & Nichols, 2009). SSI are complex, controversial problems that intersect scientific understanding and social implications, such as climate change, genetic engineering, and public health policies. These issues inherently require participants to utilize scientific knowledge alongside moral, ethical, social, and political considerations, thus demanding reasoning skills that integrate multiple disciplinary perspectives (Fang et al., 2019; Wu & Tsai, 2007).

To operate effectively within SSI contexts, this research delineates argumentation into three distinct yet interconnected elements: claim, evidence, and justification. This conceptual division, rooted in the argumentation framework by Toulmin (1958), facilitates clarity in analyzing and fostering argumentative skills. The claim denotes a specific, contestable statement or conclusion drawn by the arguer. Evidence encompasses relevant data, facts, or observations used to substantiate the claim. Justification refers to the explicit reasoning that logically and coherently connects the evidence to the claim (Zohar & Nemet, 2002), highlighting the rationale underlying the argument. By distinctly highlighting these components, educators and researchers can better assess learners' argumentation skills, ensuring that students explicitly recognize the necessity of coherent evidence-supported claims and rigorous logical reasoning (Capkinoglu et al., 2020).

Existing research emphasizes that explicitly structuring SSI argumentation around these three distinct components, such as claim, evidence, and justification, facilitates clearer thinking, more rigorous analysis, and deeper conceptual understanding among learners. McNeill and Krajcik's (2012) work, for instance, provided empirical evidence that students who receive structured instruction explicitly emphasizing claims, evidence, and justifications develop more sophisticated arguments and demonstrate enhanced comprehension of complex

airiti

scientific topics. Moreover, Sadler and Dawson (2012) pointed out that argumentation within SSI contexts uniquely requires integrating scientific evidence with ethical, social, and moral justifications. Their research demonstrates that when learners explicitly recognize the distinct roles of evidence (grounded in scientific data), claims (assertions that demand scrutiny), and justifications (reasoning that incorporates ethical and socio-political perspectives), they are better able to manage the complexity and multidimensionality inherent in SSI.

The tripartite framework also provides valuable support for educators and researchers in systematically assessing the quality and coherence of learners' arguments. By clearly distinguishing among claim formulation, evidence quality, and justification coherence, educators can more precisely pinpoint learners' strengths and weaknesses. Venville and Dawson's (2010) findings indicated that explicit framework-based interventions enable teachers to effectively scaffold student argumentation, leading to measurable improvements in both argumentation quality and conceptual understanding of SSI. Given the demonstrated value of explicitly structured argumentation frameworks, automated scoring technologies grounded in the claim-evidence-justification framework have emerged as particularly promising. These automated tools could have potential at systematically assessing learners' arguments, offering consistent, detailed, and immediate feedback on each argumentative component (Akhmadieva et al., 2023). Traditional manual scoring methods, by contrast, often struggle to provide timely due to practical constraints. Automated scoring approaches mitigate these challenges, thereby significantly enhancing the practicality and effectiveness of SSI argumentation instruction and assessment. Thus, grounding automated scoring technologies in the explicit tripartite framework is not only theoretically sound but also practically advantageous, underscoring its relevance and necessity in contemporary SSI education.

LLMs for automatic scoring

LLMs represent a class of deep learning-based natural language processing models, typically characterized by billions to trillions of parameters (Brown et al., 2020). These models undergo extensive pre-training on massive text datasets (Devlin et al., 2019), enabling them to perform diverse linguistic tasks such as text generation, translation, summarization, and question-answering. Central to their architecture is the Transformer model (Vaswani et al., 2017), which employs a self-attention mechanism that allows deep contextual understanding and semantic modeling. Consequently, LLMs have demonstrated exceptional performance in both academic research and practical applications.

Due to their expansive knowledge base, strong contextual awareness, and flexibility, LLMs have increasingly garnered attention in the domain of automated scoring (e.g., Latif & Zhai, 2024; Y. Wei et al., 2025; Yang et al., 2025). Specifically, the emergence of LLMs has significantly shifted research in automated scoring away from labor-intensive feature engineering towards the more streamlined processes of prompt design and domain adaptation. Prompt engineering, a technique aimed at optimizing LLM performance, involves supplying the model with tailored contextual information or task-specific instructions without altering its foundational parameters (Liu et al., 2023).

airiti

Compared to fine-tuning, prompt engineering offers a cost-effective and straightforward implementation (Lester et al., 2021).

Currently, several types of prompts exist, particularly beneficial when extensive training data is unavailable for novel tasks. Zero-shot prompting provides only instructions to guide the model's response (Kojima et al., 2022), thus requiring the model to infer task-specific requirements without examples. Few-shot prompting, in contrast, enhances the model's learning by including representative examples, helping the model generalize from these demonstrations (U. Lee et al., 2024).

Among various prompting strategies, CoT prompting has become widely recognized. CoT prompts guide models to generate a sequence of logically coherent steps, substantially improving reasoning capabilities (J. Wei et al., 2022). Variations of CoT include tree of thought (ToT) and self-consistency. The ToT method encourages models to evaluate multiple reasoning paths, selecting the most robust pathway through backtracking and forward-looking evaluations (Yao et al., 2023). Conversely, self-consistency enhances model performance by generating multiple reasoning paths and identifying the most consistent answer, thereby improving accuracy and stability (Wang et al., 2022).

Prompt engineering has been notably effective in enhancing LLM performance for automated scoring tasks. For instance, G.-G. Lee et al. (2024) demonstrated that combining GPT-4 with carefully designed CoT prompts surpasses zero-shot baselines and offers balanced predictions across different proficiency levels. Extending this line of research, Y. Wei et al. (2025) explored complex physics problems requiring multi-step numerical solutions. Comparing one-shot, few-shot, CoT, and ToT prompts, they found ToT prompts significantly narrowed performance gaps among models.

However, ToT's complexity, constructing and exploring an extensive reasoning tree, demands careful state management and incurs high computational costs, complicating practical implementation. In contrast, the self-consistency method simply involves multiple generations followed by majority voting, aligning neatly with the need for high-quality, stable scoring outcomes in automated assessment contexts. Nevertheless, limited research has explored the utility of self-consistency specifically within automated scoring contexts.

This study addresses this gap by examining the performance of zero-shot and few-shot prompting strategies, coupled with a systematic evaluation of self-consistency. We specifically focus on evaluating student arguments concerning SSIs, dissecting the components of claims, evidence, and justifications. Given the inherent complexity of assessing these components, where claims are straightforward, but evidence requires judgment on relevance and adequacy, and justifications necessitate sophisticated logical reasoning, this research offers timely insights into enhancing the reliability and precision of automated scoring using LLMs.

Methods

Dataset

The dataset for this study was collected from eighth-grade junior high school students in Northern Taiwan. These students participated in lessons focused on SSIs related to energy, including topics such as natural gas, nuclear power, and wind energy. The dataset comprises 238 student arguments

airiti

for each of the three topics, with each argument structured into three components: claim, evidence, and justification.

A detailed coding scheme was developed to assess the quality of students' argumentation. A score of 1 point was awarded for a claim if the student presented a reasonable stance on the issue, with incomplete sentences or phrases being acceptable. For evidence, a score of 1 point was given for providing valid empirical or theoretical support for the claim; however, evidence was only scored if the accompanying claim was deemed valid. Similarly, the justification was awarded 1 point if it provided a logical bridge explaining how the evidence supported the claim, substantiated by relevant scientific, technological, or social principles. The evaluation of the justification was only performed if both the claim and the evidence had been awarded points.

To ensure the reliability of the human scoring, the author first scored a subset of the responses and discussed the application of the coding scheme with two science education experts until all discrepancies were resolved. Following this, a second independent rater, a science researcher, was trained to use the scheme and coded 25% of the responses for each topic (60 responses per topic). As shown in Table 1, inter-rater agreement was excellent across all coding components. The final human scores for the entire dataset used in this study were then completed by the author.

Table 1. Inter-rater agreement of human scoring

Argument's component	Agreement (%)	Cohen's Kappa
Claim	98.8	.952
Evidence	95.1	.912
Justification	98.1	.910

Prompt engineering

Prompt engineering involves designing inputs that effectively elicit accurate and relevant outputs from LLMs through carefully structured instructions, examples, and context (Boonstra, 2025). To enhance the reasoning capability of the LLM, we utilized the CoT method by explicitly instructing the model to "think step-by-step" when justifying its scores. This approach encourages more interpretable and logically grounded outputs. Two main prompting strategies, zero-shot prompting, and few-shot prompting, were employed in conjunction with CoT to evaluate students' SSI arguments.

Zero-shot prompting provided the LLM with task-specific instructions without including illustrative examples. Following Boonstra (2025), this method is appropriate when simplicity and scalability are priorities. The zero-shot prompt explicitly defined the evaluation task, detailed the scoring rubric (Claim, Evidence, Justification), and instructed structured outputs with brief justifications. Few-shot prompting, by contrast, provided the model with several annotated examples to guide its understanding of the expected output format and reasoning processes. According to Boonstra, few-shot prompting generally enhances model performance by demonstrating typical task-specific response patterns. For this study, each few-shot prompt included five carefully selected annotated examples from authentic student responses across diverse SSI contexts (wind power, fossil fuels, and nuclear power). These examples represented a range of scoring scenarios to guide the model's judgment effectively.

The structured few-shot prompting examples (Table 2) were strategically chosen to highlight clearly correct reasoning, quantitative support,

airiti

insufficient logical development, incomplete connections, and repetitive justification, enabling the model to generalize more effectively across varied student responses. Both zero-shot and few-shot prompting employed identical rubrics, response formats, and role/context descriptions (Table 3). Prompts consistently specified evaluation criteria for Claim, Evidence, and Justification components,

Table 2. Few-shot prompting examples of student responses and corresponding scores

Type	Argument	Score	Explanation
High-quality ideal example	Claim: Wind power is more environmentally friendly than fossil fuels. Evidence: In 2015, traditional fossil fuels accounted for 30% of Taiwan's 85.33 million tons of CO_2 emissions, while wind power produces no greenhouse gases due to non-combustion operation. Justification: Wind turbines generate energy by harnessing wind to drive motors.	1/1/1	All components are logically connected.
Quantitative evidence but weak justification	Claim: Natural gas is more eco-friendly than coal. Evidence: Natural gas emits only 60% as much as coal. Justification: $CH_4 + 2O_2 \rightarrow CO_2 + 2H_2O$.	1/1/0	Justification does not clearly connect to environment.
Weak claim and insufficient support	Claim: Use wind power. Evidence: Newspaper. Justification: None.	0/0/0	Lacks argumentative structure and supporting details.
Incomplete logical connection	Claim: Nuclear power is more powerful. Evidence: Unlike coal and natural gas, nuclear power does not emit carbon. Justification: Nuclear fission releases large amounts of energy, thus it is more eco-friendly.	1/0/0	Evidence is unrelated to claim; logic is incomplete.
Justification repeats claim and evidence	Claim: Nuclear energy is cheaper. Evidence: In 2023, Taiwan's nuclear energy cost NT$1.5 per kWh, cheaper than wind/solar (NT$5) and natural gas (NT$3.5). Justification: Compared to other sources, nuclear is cheaper.	1/1/0	Justification simply repeats claim, no explanation.

Table 3. The structures of zero-shot and few-shot prompt

Component	Zero-shot Prompt	Few-shot Prompt
Role prompting	You are an expert evaluator of social-scientific arguments.	Same as zero-shot.
Contextual prompting	Includes a power issue and a task.	Same as zero-shot.
Rubric description	Presents the scoring rubric with criteria for Claim, Evidence, and Justification, each worth 1 point, along with conditions for evaluation.	Same as zero-shot.
Examples included	None.	Includes 5 annotated examples of student responses and corresponding scores to illustrate expected patterns.
Input format	Claim: [Student's Claim]. Evidence: [Student's Evidence]. Justification: [Student's Justification].	Same as zero-shot.
Output format	Claim Score: [0 or 1]. Evidence Score: [0 or 1]. Justification Score: [0 or 1].	Same as zero-shot.

airiti

each worth 1 point, clearly outlining conditions for scoring.

Self-consistency approach

To enhance the reliability of the automated scoring, this study employed a self-consistency approach, which extends the CoT method to improve accuracy through majority voting. This technique was applied to both the zero-shot and few-shot prompting strategies. For each student argument, the scoring prompt was executed five times to generate distinct reasoning paths, with the final score determined by the most frequent outcome. For instance, if the five individual evaluations yielded scores of [1, 1, 1, 0, 0], the final assigned score would be 1. The process involved the following steps:

- Generating diverse reasoning paths: The same prompt was sent to the LLM multiple times. To encourage the generation of varied reasoning paths and perspectives, the model's temperature parameter was set to 1.0. This strategy was therefore limited to models that support temperature adjustment (GPT-4.1 and Gemini-2.5 Flash).
- Extracting the answer: The numerical score (0 or 1) was extracted from each of the five generated responses.
- Choosing the most common answer: The final score was determined by selecting the most common answer from the five generated paths.

Data analysis

The data analysis for this study was conducted within a specific technical environment to ensure reproducibility. The experiments were run on an Ubuntu 24.04.2 LTS operating system using Python version 3.12.3. The LLMs were accessed via their respective Python libraries: "google-generativeai" version 0.8.5 for the "gemini-2.5-flash" model (Gemini-2.5, released June 17, 2025), and "openai" version 1.88.0 for the "gpt-4.1" model (GPT-4.1, released April 14, 2025) and the "o4-mini" model (O4-mini, released April 16, 2025).

The three large language models utilized in this study each have distinct characteristics. GPT-4.1 is an API (application programming interface) -focused model with significant improvements in coding, instruction following, and long-context capabilities (OpenAI, 2025a). OpenAI o4-mini is a model optimized for fast and cost-effective reasoning, excelling in areas like math and programming (OpenAI, 2025b). Gemini-2.5 Flash is a powerful and efficient model from Google, designed for speed and cost-effectiveness (Google, n.d.). These models were evaluated on the GPQA (Graduate-Level Google-Proof Q&A) benchmark (Rein et al., 2023), which consists of graduate-level science questions designed to be "Google-proof", meaning they are difficult for skilled non-experts to answer even with unrestricted web access. On this challenging test, Gemini-2.5 Flash achieved the highest accuracy at 82.8%, followed by O4-mini at 81.4%, and GPT-4.1 at 66.3%, according to performance data released by their respective companies (Google, n.d.; OpenAI, 2025a, 2025b).

To evaluate the performance of the automated scoring, we adopted standard classification metrics derived from a confusion matrix, which compares the model's predictions against the human-coded ground truth. In this context: True Positive (TP): The model correctly assigns a score of "1" when the human score is "1". True Negative (TN): The model correctly assigns a score of '0' when the human score is '0'. False Positive (FP): The model

airiti

incorrectly assigns a score of '1' when the human score is '0'. False Negative (FN): The model incorrectly assigns a score of '0' when the human score is '1'. Using these components, we assessed model performance with two key metrics: accuracy and Cohen's Kappa (Latif & Zhai, 2024).

Accuracy is the most intuitive metric, representing the proportion of all predictions that the model got correct. It is suitable when the classes (scores of '1' and '0') are relatively balanced. The formula is:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

In addition to accuracy, Cohen's Kappa was used to measure the agreement between the model's predictions and the human scores. Kappa is a robust statistic that corrects for the probability of agreement occurring by chance, providing a more reliable measure of consistency than simple accuracy, especially in cases with imbalanced class distributions. The kappa coefficient ranges from -1 to 1, where 1 indicates perfect agreement, 0 indicates agreement equivalent to chance, and values less than 0 indicate that the agreement is worse than chance.

Results

Accuracy of LLM-based automatic scoring across models and prompting strategies

The analysis examined automated scoring performance across three energy-related SSIs: fossil fuel, nuclear power, and wind power. Each topic comprised 238 student responses, resulting in a total dataset of 714 responses. For each response, three argument components (claim, evidence, and justification) were evaluated. The accuracy and Cohen's Kappa coefficients were calculated by comparing LLM-generated scores against human expert ratings across all energy topics and argument components (Table 4).

GPT-4.1 demonstrated variable performance across the three argument components when using different prompting strategies. For claim evaluation, GPT-4.1 achieved 90.3% accuracy with few-shot prompting and 83.8% accuracy with zero-shot prompting, representing a 6.5 percentage point improvement with few-shot prompting. Evidence evaluation showed more modest differences, with 80.2% accuracy for few-shot prompting compared to 79.1% for zero-shot prompting, a difference of only 1.1%. Justification evaluation revealed the largest performance gap, with few-shot prompting achieving 79.0% accuracy compared to 74.2% for zero-shot prompting, representing a 4.8 percentage point improvement.

The effectiveness of zero-shot and few-shot prompting strategies varied considerably across models and argument components. For claim evaluation, all three models showed substantial improvements with few-shot prompting. Gemini-2.5 demonstrated the largest improvement (8.1%), followed by GPT-4.1 (6.5%) and O4-Mini (5.9%). Evidence evaluation presented a different pattern, with minimal differences between prompting strategies across all models. Gemini-2.5 showed a slight improvement of 1.1%, while GPT-4.1 and O4-Mini showed improvements of 1.1% and 0.0%, respectively. This suggests that evidence evaluation may be less sensitive to the presence of examples in the prompt.

Cohen's Kappa coefficients revealed significant variations in inter-rater agreement across ar-

airiti

Table 4. Accuracy and Cohen's Kappa for LLM-based automatic scoring by model, component, and prompting strategy

Component	Model	Strategy	Accuracy	Kappa
Claim	Gemini-2.5	Few-shot	.905	.391
		Zero-shot	.824	.280
	GPT-4.1	Few-shot	.903	.353
		Zero-shot	.838	.315
	O4-Mini	Few-shot	.899	.322
		Zero-shot	.840	.227
Evidence	Gemini-2.5	Few-shot	.798	.596
		Zero-shot	.787	.574
	GPT-4.1	Few-shot	.802	.605
		Zero-shot	.791	.582
	O4-Mini	Few-shot	.791	.583
		Zero-shot	.791	.582
Justification	Gemini-2.5	Few-shot	.857	.400
		Zero-shot	.826	.294
	GPT-4.1	Few-shot	.790	.323
		Zero-shot	.742	.255
	O4-Mini	Few-shot	.878	.417
		Zero-shot	.822	.302

gument components and prompting strategies. Evidence evaluation demonstrated the strongest and most consistent inter-rater agreement, with Kappa values ranging from .574 to .605, all falling within the moderate to substantial agreement range. This superior performance suggests that evidence assessment may involve more objective criteria that align well with human expert judgment. In contrast, claim evaluation showed the most variable agreement, with Kappa values ranging from .227 to .391, spanning from fair to moderate agreement levels. Justification evaluation fell between these extremes, with Kappa values ranging from .255 to .417, indicating fair to moderate agreement depending on the model and prompting strategy employed.

The comparison between prompting strategies revealed a consistent pattern favoring few-shot prompting across all models and components. Few-shot prompting consistently produced higher Kappa values than zero-shot prompting across different conditions. Notably, all models achieved moderate inter-rater agreement (> .40) for evidence evaluation when using few-shot prompting, with GPT-4.1 reaching .605, Gemini-2.5 achieving .596, and O4-Mini attaining .583. This convergence toward moderate agreement levels with few-shot prompting suggests that providing examples significantly enhances the models' ability to align with human expert scoring standards, particularly for evidence assessment where objective criteria can be more clearly demonstrated through exemplars.

Few-shot prompting proved consistently more effective than zero-shot prompting across

airiti

all models and components, with the magnitude of improvement varying by task complexity. The strategy was most beneficial for claim evaluation, moderately beneficial for justification evaluation, and least beneficial for evidence evaluation. The results suggest that providing examples helps models better understand the scoring criteria, particularly for tasks requiring subjective judgment such as claim validity and logical reasoning assessment.

Justification evaluation revealed the most complex pattern of results. O4-Mini achieved the highest accuracy overall (87.8%) with few-shot prompting and showed the largest improvement (5.6%) compared to zero-shot conditions. Gemini-2.5 achieved 85.7% accuracy with few-shot prompting, representing a 3.1 percentage point improvement. GPT-4.1, while showing improvement with few-shot prompting, demonstrated the lowest overall performance for justification evaluation among the three models.

Impact of self-consistency method on automated scoring performance

To address the second research question, which examines whether applying a self-consistency method improves the accuracy of automated scores, we compared the performance of both zero-shot and few-shot prompting strategies with and without self-consistency enhancement across all three argument components. The implementation of self-consistency methods demonstrated variable improvements across different models, prompting strategies, and argument components. Table 5 presents the comprehensive comparison of accuracy and Cohen's Kappa coefficients before and after applying self-consistency methods.

The analysis revealed variable improvements across components and prompting strategies. Specifically, Gemini-2.5 exhibited consistent accuracy and Kappa enhancements when employing self-consistency, particularly in claim and justification evaluations (Figure 1 and Figure 2). Notably,

Table 5. Performance comparison of standard prompting and self-consistency enhanced prompting

			Self-consistency		Gain	
Component	Model	Strategy	Accuracy	Kappa	Δ Accuracy	Δ Kappa
Claim	Gemini-2.5	Zero-shot	.850	.326	+0.026	+0.046
		Few-shot	.916	.457	+0.011	+0.066
	GPT-4.1	Zero-shot	.839	.321	+0.001	+0.006
		Few-shot	.902	.339	-0.001	-0.014
Evidence	Gemini-2.5	Zero-shot	.802	.604	+0.015	+0.030
		Few-shot	.801	.626	+0.003	+0.030
	GPT-4.1	Zero-shot	.792	.585	+0.001	+0.003
		Few-shot	.799	.599	-0.003	-0.006
Justification	Gemini-2.5	Zero-shot	.835	.316	+0.009	+0.022
		Few-shot	.871	.447	+0.014	+0.047
	GPT-4.1	Zero-shot	.744	.257	+0.002	+0.002
		Few-shot	.779	.294	-0.011	-0.029

airiti

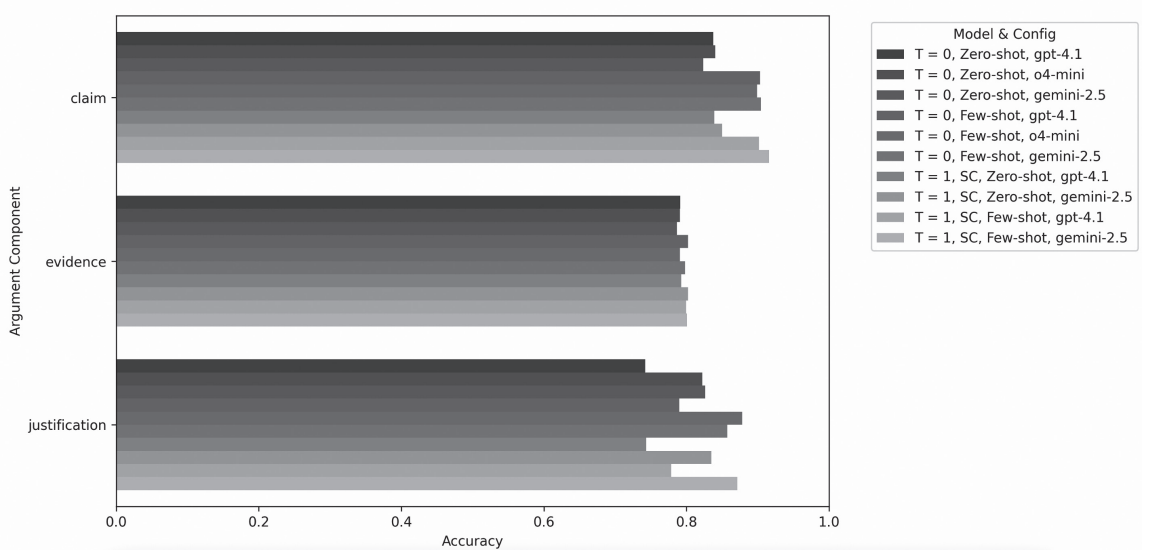

Figure 1. Accuracy by model and configuration

Note: SC, self-consistency.

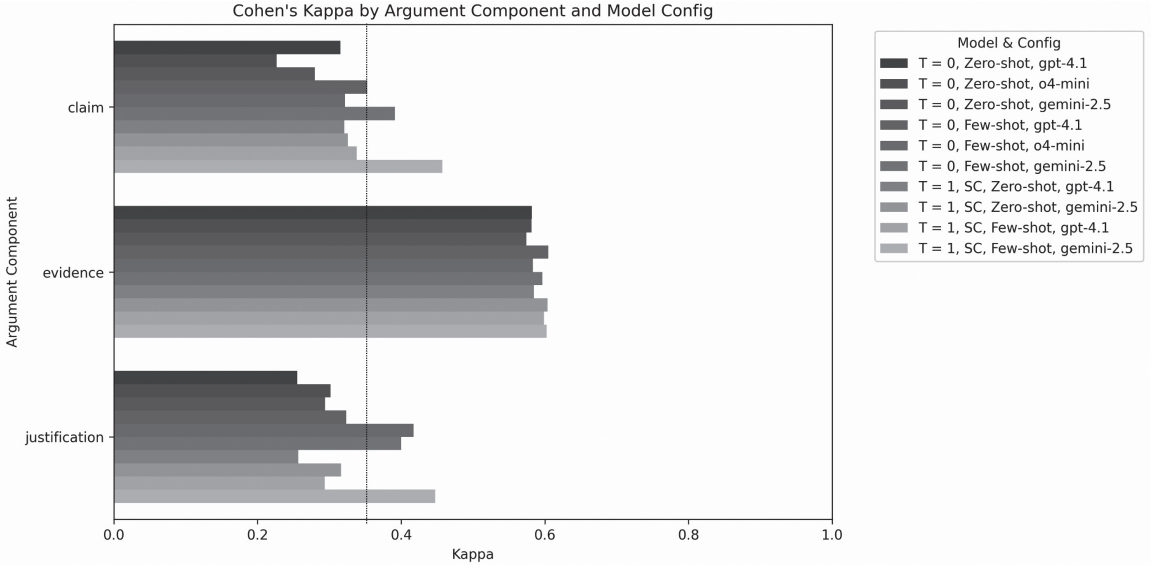

Figure 2. Cohen's Kappa by model and configuration

for the justification component using few-shot prompting, Gemini-2.5 achieved an increase in accuracy by .014 and Kappa by .047. In contrast, GPT-4.1 demonstrated mixed results. While zero-shot prompting generally benefited from the self-consistency method, improvements were modest. Interestingly, few-shot prompting occasionally showed slight declines in performance for GPT-4.1, suggesting diminishing returns or potential overfitting to provided examples. Overall, these

airiti

findings highlight that the self-consistency method can enhance scoring reliability, particularly in contexts requiring subjective judgment such as claims and justification, but the degree of improvement varies by model and prompting strategy.

Discussion

Effectiveness of few-shot prompting across models

The results consistently demonstrate that few-shot prompting serves as an effective strategy for enhancing LLM-based automatic scoring across multiple models and argument components. This finding aligns with previous research suggesting that providing examples helps models better understand task-specific requirements and scoring criteria (Boonstra, 2025; U. Lee et al., 2024). The cross-model effectiveness observed in our study—with Gemini-2.5, GPT-4.1, and O4-Mini all showing improvements with few-shot prompting—indicates that this strategy represents a generalizable approach rather than a model-specific phenomenon.

Notably, the magnitude of improvement varies by argument component, with claim evaluation showing the most substantial gains (5.9% ~ 8.1%), followed by justification evaluation (3.1% ~ 5.6%), and evidence evaluation showing minimal improvements (0.0% ~ 1.1%). This pattern suggests that few-shot prompting is particularly beneficial for tasks requiring subjective judgment and interpretation, where concrete examples can effectively demonstrate the application of scoring criteria. Evidence evaluation, being more objective and factual in nature, appears less dependent on exemplar guidance.

Among the three models tested, O4-Mini demonstrated the strongest overall performance when combined with few-shot prompting, particularly excelling in justification evaluation (87.8% accuracy). This finding is particularly noteworthy given that O4-Mini is positioned as a more efficient model compared to GPT-4.1, suggesting that model size and computational complexity do not necessarily correlate with superior performance in specialized tasks like argumentation scoring.

Model-dependent effectiveness of self-consistency

We has so far highlighted the general superiority of few-shot prompting for enhancing the performance of LLMs in argumentation analysis. It was noted that O4-Mini, despite being a smaller model, achieved the highest accuracy in the evaluation of the justification component. This finding aligns with research suggesting that smaller, domain-specific models can sometimes outperform larger, more generalized models on niche tasks (Al Zubaer et al., 2023). However, a more comprehensive analysis requires a discussion of the other core components of argumentation: the claim and the evidence. The tasks of claim and evidence identification are fundamentally about information retrieval and classification, where the broader knowledge base and more advanced natural language processing capabilities of models like GPT-4.1 and Gemini-2.5 would likely provide a competitive edge. However, this study employed a self-consistency mechanism to enhance the overall reliability of the argumentation analysis. By generating multiple reasoning paths and selecting the most consistent answer, this approach leverages the chosen model's strengths while mitigating potential weaknesses in

airiti

any single component evaluation. This methodology demonstrates that even if a model does not lead in every individual metric, its overall performance can be significantly bolstered to achieve robust and dependable results across the entire argumentation structure.

The self-consistency approach revealed striking model-dependent variations in effectiveness, with Gemini-2.5 showing consistent improvements while GPT-4.1 demonstrated mixed or negligible benefits. This differential response suggests that the underlying architecture and training approaches of different models may influence their susceptibility to self-consistency enhancement. The findings indicate that the self-consistency method works by mitigating the effects of model stochasticity through generating multiple reasoning paths and selecting the most common output, effectively averaging out random errors and reinforcing more stable reasoning processes. The consistent improvements observed with Gemini-2.5, particularly in justification evaluation (Kappa improvement of .047 with few-shot prompting), suggest that its reasoning process benefits from this "deliberative" approach, leading to outputs that better align with human scoring rubrics. Conversely, the negligible or negative impact on GPT-4.1, particularly in few-shot conditions, may indicate that its baseline predictions are already highly consistent and that introducing multiple reasoning paths does not add value. In some cases, this approach may even introduce noise that deviates from the strong anchor provided by few-shot examples. This finding suggests that self-consistency may be most beneficial for models with higher inherent variability in their reasoning processes.

High accuracy with divergent Kappa values

Although the automated scores for all three argumentation components reached high and broadly similar overall accuracy, their Cohen's Kappa coefficients diverged sharply. To better understand this discrepancy, we adopt additional reliability indices: the Prevalence Index, which quantifies the degree of class imbalance, and the Bias Index, which captures systematic tendencies for both raters to favor one response category over another (Byrt et al., 1993). These two indices allow us to disentangle statistical artifacts from substantive disagreement in inter-rater reliability.

Taking GPT-4.1 with the zero-shot method as an example (see Table 6), we observe that Cohen's Kappa corrects observed agreement by subtracting the agreement expected by chance. When one category predominates—as is the case for negatives ("0") in Claim and Justification—the expected-chance term increases steeply, forcing Kappa downward even when the observed agreement remains high. The Prevalence Index in Table 6 captures this skew, with values of .725 for Claim and .611 for Justification. This pattern exemplifies

Table 6. Accuracy, Cohen's Kappa, Prevalence Index, and Bias Index of GPT-4.1 with zero-shot method

Component	Accuracy	Kappa	Prevalence Index	Bias Index
Claim	.838	.315	.725	.053
Evidence	.791	.582	.090	.080
Justification	.742	.255	.611	.232

airiti

the prevalence paradox identified by Feinstein and Cicchetti (1990): the rarer the positive cases become, the lower Kappa appears, despite unchanged accuracy.

Secondly, beyond prevalence, bias further influences Cohen's Kappa. The Bias Index for Justification is .232, indicating that both human raters and the language model tend to favor the same decision direction, jointly classifying borderline responses as negative. Evidence, by contrast, shows a much more balanced class distribution (Prevalence Index = .090) and only minimal shared bias (Bias Index = .080), allowing its Cohen's Kappa to better reflect genuine agreement. Several practical measures can make Kappa less vulnerable to prevalence-bias effects. For instance, collecting or oversampling additional positive instances of high-quality Claims and Justifications would reduce prevalence imbalance at its source. Augmenting prompts with carefully chosen positive exemplars should temper the model's default-negative inclination and thereby lower shared bias.

Implications

The accuracy levels achieved using few-shot prompting combined with self-consistency, which consistently exceeded 80% across most conditions, have significant implications for educational practice. However, the application context must be carefully considered when interpreting these results. For high-stakes summative assessments, the current accuracy levels may raise fairness concerns, as even small error rates can have substantial consequences for individual students. The reliability requirements for such assessments typically demand higher accuracy thresholds and more extensive validation procedures. Therefore, while promising, the current performance levels may not yet meet the stringent requirements for consequential educational decisions.

However, these findings present compelling opportunities for AI-enhanced adaptive learning systems, where the focus shifts from high-stakes evaluation to formative assessment and learning support. In such contexts, the combination of few-shot prompting and self-consistency offers several advantages. First, it provides a simple, low-cost implementation approach that does not require extensive fine-tuning procedures (Latif & Zhai, 2024). Second, it enables the generation of real-time, structured feedback that can function as a powerful scaffolding tool for students. In an SSI instructional setting, a student can submit an argument and instantly receive not just a score, but a component-specific analysis derived from the model's step-by-step reasoning process. For example, the feedback could pinpoint that the provided "evidence" is irrelevant to the "claim" or that the "justification" merely repeats earlier points without providing a true logical bridge. This iterative cycle of receiving feedback and revising their work allows students to deconstruct the complex skill of argumentation and progressively internalize the standards for high-quality reasoning (Winne & Hadwin, 1998). Third, the automated assessment system with these methods provides a data-driven foundation for creating personalized learning pathways (Guo et al., 2024). By tracking a student's performance over time, an adaptive learning system can identify persistent areas of difficulty. For instance, if a student consistently receives low scores on the "justification" component, the system could automatically deliver targeted interventions, such as showing examples of strong justifications, providing a short tutorial on logical fallacies, or

airiti

offering guiding questions. This personalization allows students to focus on their specific weaknesses, leading to more efficient and effective skill development (Wilson et al., 2024; Yang et al., 2025; Zhai et al., 2020). Last, it can support personalized learning pathways based on continuous assessment of argumentation skills. The cost-effectiveness of this approach is particularly noteworthy, as it eliminates the need for expensive model fine-tuning while maintaining reasonable accuracy levels. This accessibility makes it feasible for educational institutions with limited resources to implement AI-supported assessment systems, potentially democratizing access to advanced educational technologies.

Limitations

Several limitations should be acknowledged in interpreting these findings. First, the study focused exclusively on eighth-grade students in Northern Taiwan, which may limit the generalizability of results to other educational contexts, grade levels, or cultural settings. The argumentation patterns and linguistic expressions used by students in different contexts may require different prompting strategies or model adaptations. Second, the evaluation was conducted on a specific set of energy-related SSIs. The effectiveness of the proposed approaches may vary across different scientific domains or argument structures. Future research should examine the transferability of these findings to other subject areas and argument types. Third, the study employed a relatively simple scoring rubric with binary (0/1) scoring for each component. More nuanced scoring schemes that capture gradations of argument quality might reveal different patterns of model performance and require different optimization strategies. Fourth, the self-consistency approach was limited to models that allow temperature adjustment, which excluded some specialized reasoning models from this analysis. This limitation may have affected the comprehensiveness of the model comparison. Finally, the human scoring, while achieving high inter-rater reliability (98%), was primarily conducted by a single expert with a secondary rater scoring only 25% of responses. A more robust human scoring protocol with multiple expert raters throughout the dataset might provide more reliable ground truth for model evaluation.

Conclusion

This study investigated the efficacy of LLMs for the automated scoring of students' argumentation on SSIs, focusing on the distinct components of claim, evidence, and justification. By comparing zero-shot, few-shot, and self-consistency prompting strategies across multiple LLMs, we sought to identify reliable methods for enhancing scoring accuracy. Our findings clearly demonstrate that few-shot prompting is a significantly more effective strategy than zero-shot prompting for this task. Across all tested models—GPT-4.1, Gemini-2.5, and O4-Mini—the inclusion of annotated examples consistently improved performance, particularly for the more nuanced components of claim and justification. This suggests that providing models with concrete examples is crucial for aligning their evaluative reasoning with the complex, often subjective criteria of argumentation quality. In contrast, the scoring of evidence, a more objective task, showed minimal benefit from few-shot examples, indicating that direct instructions suffice for fact-based evaluations. Furthermore, the study revealed that the self-consistency method, while

airiti

promising, exhibits model-dependent efficacy. It consistently enhanced the accuracy and reliability of Gemini-2.5, suggesting that this model benefits from a "deliberative" process of generating and comparing multiple reasoning paths. However, for GPT-4.1, the gains were negligible and, in some cases, slightly negative, implying that its baseline reasoning may already be sufficiently stable, with additional processing offering diminishing returns.

From a practical standpoint, the high accuracy levels achieved, often exceeding 80% and approaching 90% with the combination of few-shot prompting and self-consistency, underscore the potential of LLMs as powerful tools for formative assessment. These methods offer a low-cost, scalable alternative to manual grading, capable of providing students with immediate, structured feedback to improve their argumentation skills within adaptive learning environments. In conclusion, this study confirms that strategically engineered prompts can transform general-purpose LLMs into effective, specialized tools for educational assessment. The combination of few-shot examples and, for certain models, self-consistency methods, presents a robust framework for the automated scoring of complex scientific argumentation, paving the way for more responsive and scalable educational technologies.

Acknowledgements

This work was financially supported by the National Science and Technology Council of Taiwan under contracts the NSTC 114-2410-H-656-013-MY3 (科學探究評量與自我評量的交會：探討國中學生自我調整科學學習的縱貫式研究).

Reference

Akhmadieva, R. S., Udina, N. N., Kosheleva, Y. P., Zhdanov, S. P., Timofeeva, M. O., & Budkevich, R. L. (2023). Artificial intelligence in science education: A bibliometric review. *Contemporary Educational Technology*, *15*(4), Article ep460. https://doi.org/10.30935/cedtech/13587

Al Zubaer, A., Granitzer, M., & Mitrović, J. (2023). Performance analysis of large language models in the domain of legal argument mining. *Frontiers in Artificial Intelligence*, *6*, Article 1278796. https://doi.org/10.3389/frai.2023.1278796

Boonstra, L. (2025). *Prompt engineering*. https://reurl.cc/GG4o6Z

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., ..., Amodei, D. (2020). Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan & H. Lin (Eds.), *Advances in neural information processing systems 33* (pp. 1877-1901). Curran Associates.

Byrt, T., Bishop, J., & Carlin, J. B. (1993). Bias, prevalence and kappa. *Journal of Clinical Epidemiology*, *46*(5), 423-429. https://doi.org/10.1016/0895-4356(93)90018-V

Capkinoglu, E., Yilmaz, S., & Leblebicioglu, G. (2020). Quality of argumentation by seventh-graders in local socioscientific issues. *Journal of Research in Science Teaching*, *57*(6), 827-855. https://doi.org/10.1002/tea.21609

airiti

Chang, Y., Wang, X., Wang, J., Wu, Y., Yang, L., Zhu, K., Chen, H., Yi, X., Wang, C., Wang, Y., Ye, W., Zhang, Y., Chang, Y., Yu, P. S., Yang, Q., & Xie, X. (2024). A survey on evaluation of large language models. *ACM Transactions on Intelligent Systems and Technology*, *15*(3), Article 39. https://doi.org/10.1145/3641289

Chien, Y.-T., & Chang, C.-Y. (2015). Supporting socio-scientific argumentation in the classroom through automatic group formation based on students' real-time responses. In M. S. Khine (Ed.), *Science education in East Asia: Pedagogical innovations and research-informed practices* (pp. 549-563). Springer. https://doi.org/10.1007/978-3-319-16390-1_22

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. In J. Burstein, C. Doran, & T. Solorio (Eds.), *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: Human language technologies*, Vol 1 (long and short papers) (pp. 4171-4186). Association for Computational Linguistics. https://doi.org/10.18653/v1/N19-1423

Driver, R., Newton, P., & Osborne, J. (2000). Establishing the norms of scientific argumentation in classrooms. *Science Education*, *84*(3), 287-312. https://doi.org/10.1002/(SICI)1098-237X(200005)84:3<287::AID-SCE1>3.0.CO;2-A

Erduran, S., Simon, S., & Osborne, J. (2004). TAPping into argumentation: Developments in the application of Toulmin's Argument Pattern for studying science discourse. *Science Education*, *88*(6), 915-933. https://doi.org/10.1002/sce.20012

Fang, S.-C., Hsu, Y.-S., & Lin, S.-S. (2019). Conceptualizing socioscientific decision making from a review of research in science education. *International Journal of Science and Mathematics Education*, *17*(3), 427-448. https://doi.org/10.1007/s10763-018-9890-2

Farrokhnia, M., Banihashem, S. K., Noroozi, O., & Wals, A. (2024). A SWOT analysis of ChatGPT: Implications for educational practice and research. *Innovations in Education and Teaching International*, *61*(3), 460-474. https://doi.org/10.1080/14703297.2023.2195846

Feinstein, A. R., & Cicchetti, D. V. (1990). High agreement but low Kappa: I. the problems of two paradoxes. *Journal of Clinical Epidemiology*, *43*(6), 543-549. https://doi.org/10.1016/0895-4356(90)90158-L

Google. (n.d.). Gemini 2.5 Flash.

Guo, S., Zheng, Y., & Zhai, X. (2024). Artificial intelligence in education research during 2013–2023: A review based on bibliometric analysis. *Education and Information Technologies*, *29*(13), 16387-16409. https://doi.org/10.1007/s10639-024-12491-8

Jia, F., Sun, D., & Looi, C.-k. (2024). Artificial intelligence in science education (2013–2023): Research trends in ten years. *Journal of Science Education and Technology*, *33*(1), 94-117. https://doi.org/10.1007/s10956-023-10077-6

Kojima, T., Gu, S. S., Reid, M., Matsuo, Y., & Iwasawa, Y. (2022). Large language models are zero-shot reasoners. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, & A. Oh (Eds.), *Advances in*

airiti

neural information processing systems 35 (pp. 22199-22213). Curran Associates.

Latif, E., & Zhai, X. (2024). Fine-tuning ChatGPT for automatic scoring. *Computers and Education: Artificial Intelligence*, *6*, Article 100210. https://doi.org/10.1016/j.caeai.2024.100210

Lee, G.-G., Latif, E., Wu, X., Liu, N., & Zhai, X. (2024). Applying large language models and chain-of-thought for automatic scoring. *Computers and Education: Artificial Intelligence*, *6*, Article 100213. https://doi.org/10.1016/j.caeai.2024.100213

Lee, U., Jung, H., Jeon, Y., Sohn, Y., Hwang, W., Moon, J., & Kim, H. (2024). Few-shot is enough: Exploring ChatGPT prompt engineering method for automatic question generation in English education. *Education and Information Technologies*, *29*(9), 11483-11515. https://doi.org/10.1007/s10639-023-12249-8

Lester, B., Al-Rfou, R., & Constant, N. (2021). The power of scale for parameter-efficient prompt tuning. *arXiv*. https://doi.org/10.48550/arXiv.2104.08691

Liu, P., Yuan, W., Fu, J., Jiang, Z., Hayashi, H., & Neubig, G. (2023). Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, *55*(9), Article 195. https://doi.org/10.1145/3560815

McNeill, K. L., & Krajcik, J. S. (2012). *Supporting grade 5-8 students in constructing explanations in science: The claim, evidence, and reasoning framework for talk and writing*. Pearson.

OpenAI. (2025a). *Introducing GPT-4.1 in the API*. https://reurl.cc/8b05lX

OpenAI. (2025b). *Introducing OpenAI o3 and o4-mini*. https://reurl.cc/5b4Moq

Rein, D., Hou, B. L., Stickland, A. C., Petty, J., Pang, R. Y., Dirani, J., Michael, J., & Bowman, S. R. (2023). *GPQA: A graduate-level google-proof Q&A benchmark*. https://doi.org/10.48550/arXiv.2311.12022

Sadler, T. D. (2004). Informal reasoning regarding socioscientific issues: A critical review of research. *Journal of Research in Science Teaching*, *41*(5), 513-536. https://doi.org/10.1002/tea.20009

Sadler, T. D., Barab, S. A., & Scott, B. (2007). What do students gain by engaging in socioscientific inquiry? *Research in Science Education*, *37*(4), 371-391. https://doi.org/10.1007/s11165-006-9030-9

Sadler, T. D., & Dawson, V. (2012). Socio-scientific issues in science education: Contexts for the promotion of key learning outcomes. In B. J. Fraser, K. Tobin, & C. J. McRobbie (Eds.), *Second international handbook of science education* (pp. 799-809). Springer. https://doi.org/10.1007/978-1-4020-9041-7_53

Toulmin, S. E. (1958). *The uses of argument*. Cambridge University Press.

van den Berg, G., & du Plessis, E. (2023). ChatGPT and generative AI: Possibilities for its contribution to lesson planning, critical thinking and openness in teacher education. *Education Sciences*, *13*(10), Article 998. https://doi.org/10.3390/educsci13100998

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S.

airiti

Vishwanathan & R. Garnett (Eds.), *Advances in Neural Information Processing Systems 30* Curran Associates.

Venville, G. J., & Dawson, V. M. (2010). The impact of a classroom intervention on grade 10 students' argumentation skills, informal reasoning, and conceptual understanding of science. *Journal of Research in Science Teaching*, *47*(8), 952-977. https://doi.org/10.1002/tea.20358

Wahono, B., Chang, C.-Y., & Khuyen, N. T. T. (2021). Teaching socio-scientific issues through integrated STEM education: An effective practical averment from Indonesian science lessons. *International Journal of Science Education*, *43*(16), 2663-2683. https://doi.org/10.1080/09500693.2021.1983226

Wang, X., Wei, J., Schuurmans, D., Le, Q., Chi, E., Narang, S., Chowdhery, A., & Zhou, D. (2022). *Self-consistency improves chain of thought reasoning in language models*. https://doi.org/10.48550/arXiv.2203.11171

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E., Le, Q., & Zhou, D. (2022). *Chain-of-thought prompting elicits reasoning in large language models*. https://doi.org/10.48550/arXiv.2201.11903

Wei, Y., Zhang, R., Zhang, J., Qi, D., & Cui, W. (2025). Research on intelligent grading of physics problems based on large language models. *Education Sciences*, *15*(2), Article 116. https://doi.org/10.3390/educsci15020116

Wilson, C. D., Haudek, K. C., Osborne, J. F., Buck Bracey, Z. E., Cheuk, T., Donovan, B. M., Stuhlsatz, M. A. M., Santiago, M. M., & Zhai, X. (2024). Using automated analysis to assess middle school students' competence with scientific argumentation. *Journal of Research in Science Teaching*, *61*(1), 38-69. https://doi.org/10.1002/tea.21864

Winne, P. H., & Hadwin, A. F. (1998). Studying as self-regulated learning. In D. J. Hacker, J. Dunlosky, & A. C. Graesser (Eds.), *Metacognition in educational theory and practice* (pp. 277-304). Routledge.

Wu, Y. T., & Tsai, C. C. (2007). High school students' informal reasoning on a socio-scientific issue: Qualitative and quantitative analyses. *International Journal of Science Education*, *29*(9), 1163-1187. https://doi.org/10.1080/09500690601083375

Yang, J., Latif, E., He, Y., & Zhai, X. (2025). Fine-tuning ChatGPT for automatic scoring of written scientific explanations in Chinese. *Journal of Science Education and Technology*, *34*(4), 719-736. https://doi.org/10.1007/s10956-025-10199-z

Yao, S., Yu, D., Zhao, J., Shafran, I., Griffiths, T. L., Cao, Y., & Narasimhan, K. (2023). *Tree of thoughts: Deliberate problem solving with large language models*. https://doi.org/10.48550/arXiv.2305.10601

Zeidler, D. L., & Nichols, B. H. (2009). Socioscientific issues: Theory and practice. *Journal of Elementary Science Education*, *21*(2), 49-58. https://doi.org/10.1007/BF03173684

Zhai, X., Yin, Y., Pellegrino, J. W., Haudek, K. C., & Shi, L. (2020). Applying machine learning in science assessment: A systematic review. *Studies in Science Education*, *56*(1), 111-151. https://doi.org/10.1080/03057267.2020.1735757

airiti

Zohar, A., & Nemet, F. (2002). Fostering students' knowledge and argumentation skills through dilemmas in human genetics. *Journal of Research in Science Teaching*, *39*(1), 35-62. https://doi.org/10.1002/tea.10008

airiti

科學教育學刊
2025, 第三十三卷第三期, 265-286
DOI:10.6173/CJSE.202509_33(3).0003

Contemporary Journal of Science Education
2025, 33(3), 265-286

提升大型語言模型自動評分的準確性：以零範例、少量範例與自我一致性策略分析社會科學議題

隋奇融*
臺北市立大學學習與媒材設計學系

摘要

本研究探討大型語言模型於社會科學議題 (socio-scientific issues, SSIs) 中，針對學生論證文本進行自動評分的效能，並聚焦於主張、證據與辯證三個論述的元素。我們比較三種提示策略零範例 (zero-shot)、少量範例 (few-shot) 與自我一致性 (self-consistency) 於三種模型 (GPT-4.1、Gemini-2.5 及 O4-Mini) 中的表現，並使用一組包含 238 筆學生關於能源議題的論證資料進行分析。研究結果顯示在所有模型中，少量範例提示策略在主張與辯證等較具複雜性的成分評分上明顯優於零範例，突顯提供具體範例能協助模型對齊其評分邏輯與複雜的評分標準。相較之下，自我一致性策略的效能則依模型而異：對 Gemini-2.5 有穩定提升準確度的效果，對 GPT-4.1 則幾乎無效，顯示模型架構可能影響其對深度推理提示策略的敏感性。從實務面來看，少量範例結合自我一致性策略可達到高準確率，展現大型語言模型作為形成性評量工具的潛力，具備低成本、高擴展性的優勢，可即時回饋學生於自適應學習環境中。然而，考量其仍有一定錯誤率，高風險的總結性評量仍需審慎應用。總結而言，本研究透過策略性提示設計，證實大型語言模型可轉化為有效、專業的社會性科學議題的論證評量工具，並對 SSI 教學中的即時診斷與個別化鷹架提供具體操作途徑。

關鍵詞：論證、自動評分、大型語言模型、社會科學議題

*通訊作者：隋奇融，chrosui@gmail.com；ORCID：0000-0003-4559-0641
投稿：2025/6/22，修訂：2025/7/22，接受：2025/7/29，線上出版：2026/1/26