Evaluating the Reliability of Human–AI Collaborative Scoring of Written Arguments Using Rational Force Model

Noriko Takahashi
Montclair State University
takahashin1@montclair.edu

Abraham Onuorah
Montclair State University
onuoraha1@montclair.edu

Alina Reznitskaya
Montclair State University
reznitskayaa@montclair.edu

Evgeny Chukharev
Iowa State University
evgeny@iastate.edu

Ariel Sykes
Montclair State University
Sykesa@montclair.edu

Michele Flammia
Independent Researcher
micheledapila@gmail.com

Joe Oyler
Maynooth University
joe.oyler@mu.ie

Abstract

This study aims to improve the reliability of a new AI collaborative scoring system used to assess the quality of students' written arguments. The system draws on the Rational Force Model and focuses on classifying the functional relation of each proposition in terms of support, opposition, acceptability, and relevance.

We evaluated GPT-4o under zero-shot and few-shot prompting. Results show that few-shot prompting improved classification accuracy: Acceptability Support (AS) reached an F1 score of 0.95, Relevance Support (RS) rose from 0.08 to 0.72, and Acceptability Objection (AO) increased from 0.42 to 0.74. Relevance Objection (RO) was rare but false positives decreased. Error analysis revealed that misclassifications often stemmed from overreliance on lexical cues rather than contextual nuance. For instance, GPT-4o tended to treat extreme words like never or any as objections, even when the context indicated support. These findings highlight the potential of RFM-guided prompts to enhance automated essay scoring and provide more reliable, reasoning-focused feedback.

1 Introduction

Research on automated essay scoring (AES) for argumentative writing has advanced significantly over the past decade. Foundational studies established methods for identifying core argumentative elements such as claims, reasons, and evidence (Stab and Gurevych, 2014; Persing and Ng, 2015). Building on this foundation, more recent systems increasingly employ transformer-based large language models (LLMs), including BERT, GPT, and LLaMA, to improve scoring accuracy and robustness. For example, Carlile et al. (2018) created a dataset of student essays labeled for persuasiveness and related qualities, offering early resources for argumentative writing research, while Toledo et al. (2019) leveraged BERT-based architectures to rank arguments. Hicke et al. (2023) introduced a transformer-based method for labeling persuasive segments as "effective" or "ineffective," reaching near-human performance. Similarly, Sun and Wang (2024) developed a multi-dimensional model that assesses vocabulary, grammar, and coherence with high predictive accuracy.

Despite these gains, most AES systems still operate at the level of isolated features or segments and therefore struggle to capture how propositions interconnect to form a coherent line of reasoning. Argumentative writing unfolds through chains of interdependent propositions: some supply direct evidence, others provide conceptual linkage, and still others contest earlier claims. Treating these components independently obscures the discourse-level relationships that determine overall logical quality and persuasiveness. Modeling these relationships remains a central challenge.

This limitation is especially consequential in educational settings. Scholars have argued that emphasizing the mere presence of claims, evidence,

Proceedings of the Artificial Intelligence in Measurement and Education Conference (AIME-Con) – Volume 2: Works in Progress, pages 135–140

and counterarguments can divert attention from the coherence and quality of reasoning (Chinn et al., 2016; Newell et al., 2011; Rapanta et al., 2013). Backman et al. (2023) further contend that such structural checklists can impede both teachers and students from developing a deeper understanding of what distinguishes strong from weak arguments. Accordingly, automated scoring should be aligned with educational perspectives that prioritize the quality of reasoning, not just its components.

To address the limitations of the current approaches, we adopt the Rational Force Model (Naess, 1959; Backman et al., 2012, 2023), a framework that evaluates important but largely overlooked dimensions of argument quality, specifically focusing on the relational role each proposition plays in connection to another, as well as its acceptability and relevance. We discuss this framework next.

2 Rational Force Model (RFM)

The Rational Force Model (RFM), developed by Naess (1959) and extended by other researchers (e.g., Backman et al., 2012, 2023; Björnsson et al., 1994), provides a fine-grained framework for evaluating argumentative quality. Rather than focusing on the mere presence of certain argument elements (claims, reasons), RFM centers on both the proposition's function and its epistemic strength, thus examining how a proposition supports or opposes another proposition within the overall line of reasoning.

According to a more recent version of the RFM, discussed by Backman and colleagues (2023), RFM proceeds in two phases. In the descriptive (reconstruction) phase, a text is segmented into discrete idea units (propositions). Each proposition is mapped to a target (the main claim or another proposition) and classified by intended function: Acceptability Support (AS): A proposition intended to increase another proposition's acceptability, Acceptability Objection (AO): A proposition intended to decrease another proposition's acceptability, Relevance Support (RS): A proposition intended to increase another proposition's relevance, or Relevance Objection (RO): A proposition intended to decrease another proposition's relevance, as shown in Table 1. This reconstruction yields a directed structure of support and opposition.

In the evaluative (scoring) phase, each proposition receives two scores: Acceptability (A), the degree to which there is reason to believe the proposition is true; and Relevance (R), the degree to which, if true, the proposition advances resolution of the issue or supports its target. The proposition's rational force is the product of these values: $\mathrm{RF}_i = A_i R_i$.

In sum, RFM highlights not just the presence of argumentative components, but their functional roles, accuracy, and logical strength. As such, RFM provides a principled basis for analyzing important, but largely overlooked, dimensions of written arguments, thus generating valuable diagnostic information to support meaningful feedback.

	Support	**Objection**
Acceptability	Acceptability Support(AS)	Acceptability Objection(AO)
Relevance	Relevance Support (RS)	Relevance Objection (RO)

Table 1: Four types of propositions in an RFM analysis.

Note. Adapted from Backman et al. (2012, 2023).

3 Aims

The present study is part of a larger project (Reznitskaya et al., 2025) aimed at developing a related AES system. Here, we focus on one key component of that effort: evaluating the ability of AES systems to assess not just individual propositions, but the relationships between them.

Specifically, we frame each proposition in terms of its function—Acceptability Support (AS), Relevance Support (RS), Acceptability Objection (AO), or Relevance Objection (RO)—within the structure of reasoning. By treating function identification as a classification task, we examine the extent to which AI systems can recover the relational architecture of arguments.

Our research questions (RQ) are:

RQ1: Can GPT-4o reliably classify the functional relation between two propositions as AS, RS, AO, or RO, compared to a human label?

RQ2: Does few-shot prompting improve GPT-4o ability to distinguish these roles?

4 Sample

Our study draws on a corpus of 504 argumentative essays written by Grade 5 students (10–11 years old) in public schools at two research sites in the United States (New Jersey and Ohio). The essays

were collected as part of a quasi-experimental study aimed at improving students' argumentation skills (Wilkinson et al., 2023; Reznitskaya and Wilkinson, 2020).

In New Jersey (n = 239), students were primarily White (60.7

The writing task was based on a short story, The Pinewood Derby (776 words), in which a boy named Jack faces a moral dilemma of whether to report his classmate Thomas, who cheated by not building his model car himself. After hearing and reading along with the story, students were asked to write a letter to their teacher explaining whether Jack should tell on Thomas, supporting their opinion with reasons and evidence, addressing possible counterarguments, and concluding their response. Students were given 25 minutes to complete the task, which pilot studies confirmed was sufficient time.

From this larger dataset of 504 essays, we randomly selected 25 essays for detailed manual annotation in the current study.

5 Method

Each essay was segmented into idea units. An idea unit "expresses one action or event or state, and generally corresponds to a single verb clause" (Mayer, 1985, p. 71). This segmentation step ensures that long or complex student sentences are broken down into smaller, analytically meaningful parts, each representing a distinct claim or piece of reasoning. For example, the sentence "Thomas should tell on Jack because he cheated" would be divided into two idea units: one expressing the main claim ("Thomas should tell on Jack"), and another the supporting reason ("he cheated").

Within the RFM framework, each idea unit, we called a source, is aimed at one other idea unit, we called a target. Trained annotators labeled each source idea unit specifying its relation to the target, selecting from the four RFM categories: Acceptability Support (AS), Relevance Support (RS), Acceptability Objection (AO), or Relevance Objection (RO) (see Table 1).

To improve clarity and reduce ambiguity, annotators also created a reconstructed, or standardized idea unit, which paraphrased the student's statement into its core meaning. These reconstructions, shown in brackets [ ], helped resolve cases where children's writing was unclear, incomplete, or colloquial. For example, if a student wrote "There is no reason to feel sorry for Thomas," the reconstructed idea unit might be [Mean people don't deserve empathy], ensuring that the intended meaning was explicit. This process was essential for maintaining consistency in annotation and for allowing both human raters and the AI system to work with clearly defined propositions (see Table 2).

We analyzed 200 pairs of annotated idea units and examined the reliability of GPT classification compared to a human label. We used GPT-4o with prompt strategies. The target–source pairs were provided, but without human labels. GPT was prompted to assign an RFM label to the target idea unit. We had two experimental conditions: a zero-shot prompt and a few-shot prompt. In the zero-shot condition, we provided general instructions about RFM labels without examples. In the few-shot condition, the labels were explained in more detail with examples and exceptions to provide clearer classification guidelines.

No.	Target Idea	Source Idea	Label
1	because Thomas didn't make the model car by himself. [Thomas didn't build the car on his own]	I think so because in the story it said "No my brother did it." [Thomas said "No, my brother made it"]	AS
2	This shows that he is very mean. [Thomas was mean]	I think Thomas should be nicer to other people. [Thomas should be nicer to people]	RS
3	because Thomas cheated and he won. [Thomas won by cheating]	And that's not fair to everyone else [Winning by cheating is unfair]	AS
4	Some people might say "no" because he was unliked by many kids [Students didn't like Thomas]	There is no reason to feel sorry for Thomas [Mean people don't deserve empathy]	AO

Table 2: Sample of dataset with annotated labels and reconstructed idea units.

Category	Zero-shot			Few-shot		
	Precision	Recall	F1 Score	Precision	Recall	F1 Score
AS	0.83	0.70	0.76	0.96	0.94	0.95
RS	0.06	0.14	0.08	0.81	0.64	0.72
AO	0.89	0.42	0.42	0.63	0.90	0.74
RO	0	0	0	0	0	0

Table 3: F1 scores for each category.

6 Results

In comparing the zero-shot and few-shot prompt results (see Table 3), F1 scores improved overall in the few-shot condition. The AS category showed an increase in F1 score from 0.76 to 0.95. The most substantial change occurred in RS, which increased from 0.08 to 0.72. AO also showed improvement, with F1 rising from 0.42 to 0.74. RO, which rarely appeared in the essays and was not part of human labeling, remained at 0 for both prompts. However, the False Positives for RO decreased from 8 in the zero-shot prompt to 1 in the few-shot prompt, indicating an improvement.

The few-shot prompt (see Table 4) contributed substantially to the improvement in RS and AO by providing clearer definitions for these categories. The model struggled to distinguish between AS and RS, as well as AS and AO in the zero-shot prompt, so differences between them were added in the few-shot prompt. The few-shot prompt also encouraged considering the context of the Pinewood Derby story and the student's likely intent.

Based on these results, the answers to the research questions are:

RQ1: GPT-4o reliably classified the functional relation for AS with a high F1 score. Other categories were also reliably classified with the few-shot prompt.

RQ2: Few-shot prompting improved GPT-4o's performance, particularly for RS and AO. However, RS and AO still show variability in Precision and Recall, indicating areas that require further refinement.

7 Discussion

The results demonstrate that the few-shot prompt improved GPT-4o's ability to classify functional relations, particularly RS and AO. However, further refinement is needed. For example, for the target idea "Thomas never did any of the hard work," the source "Thomas painted and decorated his car" was labeled AS by human annotators because it provides evidence for the claim 'did not do hard work,' as painting and decorating are considered easy tasks in the context of the story. In contrast, GPT-4o labeled it as AO, since the source describes Thomas doing some work, even though it's considered less difficult. This discrepancy likely arises because GPT-4o tends to focus on extreme expressions like 'never' or 'any,' which it interprets as strong markers of absolute negation. As a result, GPT struggles to account for the nuanced difference between what is considered 'hard work' versus 'easy work' in the context of the story. This highlights the need for further improvement in the few-shot prompt.

This study highlights the potential to improve the Human–AI Collaborative Scoring system using the RFM framework. It supports the development of scoring procedures that (1) target theoretically and pedagogically important aspects of argument quality and (2) can be applied reliably to naturally occurring student arguments. Despite some inconsistencies in the scoring system due to the variety of propositions in student essays, the results suggest a path toward refining rules to handle exceptional cases.

Limitations

This study has several limitations. First, the distribution of categories was uneven, with Relevance Objections (RO) almost absent in the student essays. As a result, the model's performance on this category could not be meaningfully evaluated. Second, the study focused on a single model (GPT-4o) under two prompting conditions (zero-shot and few-shot), which limits the scope of the findings. Additional experiments with other models, prompting strategies, and fine-tuning approaches are needed to test the robustness of the results. Finally, annotation according to the Rational Force Model (RFM) involves nuanced judgments of acceptability and relevance, which can be open to interpretation. Disagreements among annotators may influ-

	Zero-shot	Few-shot (Added)
AS	The *source* strengthens the truthfulness or plausibility of the *target*	(+) AS if the *source* answers "Why believe?"
RS	The *source* strengthens the relevance or usefulness of the *target*	(+) RS if the *source* answers "Why care?" or adds moral/social importance (+) RS if the *source* explains general moral or social norms, not AS
AO	The *source* challenges truthfulness or plausibility of the *target*	(+) AO if the *source* disagrees with truthfulness or plausibility of the *target*, or pushes back against it (rebuttal)
RO	The *source* challenges the relevance or usefulness of the *target*	(+) RO if the *source* disagrees with relevance or usefulness of the *target*, or pushes back against it (rebuttal)
Others	For each pair, classify the *source* label in relation to the *target*	(+) Use only AS or AO when the *target* is a main claim (+) Reference the story "Pinewood Derby" (+) If a sentence is unclear, use the [bracketed] reconstructed idea unit to understand each idea clearly

Table 4: Zero-shot vs. Few-shot prompting comparison.

ence the gold-standard labels and, in turn, affect the evaluation of model accuracy. Strengthening inter-annotator reliability therefore remains an important direction for future studies.

Acknowledgments

This material is based upon work supported by the National Science Foundation under Grant Nos. 2506473 and 2506474.

References

Ylva Backman, Viktor Gardelli, Tobias Gardelli, and Anders Persson. 2012. *Scientific Thinking Tools: A Base for Academic Studies*. Studentlitteratur.

Ylva Backman, Alina Reznitskaya, Viktor Gardelli, and Ian A. G. Wilkinson. 2023. Beyond structure: Using the rational force model to assess argumentative writing. *Written Communication*, 40(2):555–585.

Gunnar Björnsson, Ulrik Kilhbom, Folke Tersman, and Anders Ullholm. 1994. *Argumentationsanalys*. Natur och Kultur.

Winston Carlile, Nishant Gurrapadi, Zixuan Ke, and Vincent Ng. 2018. Give me more feedback: Annotating argument persuasiveness and related attributes in student essays. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 621–631, Melbourne, Australia. Association for Computational Linguistics.

Clark A. Chinn, Richard G. Duncan, Lung-Chi Hung, and Robert W. Rinehart. 2016. Epistemic criteria and reliable processes as indicators of argument quality in science students' argumentation. In *Proceedings of the Annual Meeting of the American Educational Research Association (AERA 2016)*, Washington, DC, USA.

Yann Hicke, Tonghua Tian, Karan Jha, and Choong Hee Kim. 2023. Automated essay scoring in argumentative writing: Deberteachingassistant. arXiv preprint arXiv:2307.04276.

Richard E. Mayer. 1985. Structural analysis of science prose: Can we increase problem solving performance? In *Understanding of Expository Text*, pages 65–87. Erlbaum, Hillsdale, NJ.

Arne Naess. 1959. *Communication and Argument: Elements of Applied Semantics*. Allen & Unwin.

George E. Newell, Richard Beach, Jennifer Smith, Jennifer VanDerHeide, Deanna Kuhn, and Jeroen Andriessen. 2011. Teaching and learning argumentative reading and writing: A review of research. *Reading Research Quarterly*, 46(3):273–304.

Isaac Persing and Vincent Ng. 2015. Modeling argument strength in student essays. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 543–552, Beijing, China. Association for Computational Linguistics.

Chrysi Rapanta, Maria Garcia-Mila, and Sergi Gilabert. 2013. What is meant by argumentative competence?

an integrative review of methods of analysis and assessment in education. *Review of Educational Research*, 83(4):483–520.

Alina Reznitskaya, Michele Flammia, Noriko Takahashi, Abraham Onuorah, Ariel Sykes, Joe Oyler, and Evgeny Chukharev. 2025. Enhancing diagnostic and instructional value of assessments designed to evaluate written arguments. In *EARLI Conference: Realising Potentials through Education: Shaping the Minds and Brains for the Future*, Graz, Austria.

Alina Reznitskaya and Ian A. G. Wilkinson. 2020. Measuring production and comprehension of written arguments in upper-elementary grades. *Studia Paedagogica*, 24(Special Issue on Argumentation):63–84.

Christian Stab and Iryna Gurevych. 2014. Identifying argumentative discourse structures in persuasive essays. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 46–56, Doha, Qatar. Association for Computational Linguistics.

Kun Sun and Rong Wang. 2024. Automatic essay multidimensional scoring with fine-tuning and multiple regression. arXiv preprint arXiv:2406.01198.

Assaf Toledo, Shai Gretz, Edo Cohen-Karlik, Roni Friedman, Elad Venezian, Dan Lahav, Michal Jacovi, Ranit Aharonov, and Noam Slonim. 2019. Automatic argument quality assessment – new datasets and methods. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5625–5635, Hong Kong, China. Association for Computational Linguistics.

Ian A. G. Wilkinson, Alina Reznitskaya, and Joseph V. D'Agostino. 2023. Professional development in classroom discussion to improve argumentation: Teacher and student outcomes. *Learning and Instruction*, 85:101732.