



## TOWARDS EFFECTIVE ARGUMENTATION: DESIGN AND IMPLEMENTATION OF A GENERATIVE AI-BASED EVALUATION AND FEEDBACK SYSTEM

**Hunkoog Jho,  
Minsu Ha**

### Introduction

Science is a logical language that includes mathematics and various types of data. Logical activities such as discussions, verifications, and refutations among scientists have had a profound impact on the development of science. Moreover, argumentation, the ability to understand, judge, and make decisions based on diverse information is considered very important not only in science but also in various fields such as economics, society, and reading (Berland & Hammer, 2012; Duschl et al., 2007; National Research Council, 2011). In science, argumentation is necessary to develop hypotheses or models and evaluate data and evidence obtained from the inquiry. In language education, it plays a significant role in communicating with others. It is significant that the public should be capable of appreciating given information and making appropriate decisions as the informed citizens, and argumentation is closely connected to such a goal of K-12 education (Yi & Guo, 2021; Hodson, 2011; Ratcliffe & Grace, 2003). A number of countries survey students' ability related to argumentation and scientific literacy periodically (Tsai, 2015).

Argumentation refers to the process of constructing and presenting reasoned, logical arguments to persuade or convey a point of view effectively. It involves the use of evidence, reasoning, and persuasive techniques to support one's claims and engage in constructive dialogue with others. Argumentation plays a crucial role in critical thinking, communication skills, and the exchange of ideas, often employed in academic, professional, and everyday contexts to explore, debate, and resolve complex issues and differing viewpoints. In science education, argumentation plays a vital role in fostering critical thinking skills and enhancing the learning process. It encourages students to engage actively with scientific concepts and practices by constructing, evaluating, and communicating reasoned arguments. One of the key roles of argumentation in science learning is that it promotes a deeper understanding of scientific concepts. When students are encouraged to formulate arguments based on evidence and reasoning, they are more likely to develop a robust comprehension of scientific principles and phenomena. Furthermore, argumentation in science education encourages students to question and explore ideas, leading to increased curiosity and

**Abstract.** *This study aimed at examining the performance of generative artificial intelligence to extract argumentation elements from text. Thus, the researchers developed a web-based framework to provide automated assessment and feedback relying on a large language model, ChatGPT. The results produced by ChatGPT were compared to human experts across scientific and non-scientific contexts. The findings revealed marked discrepancies in the performance of AI for extracting argument components, with a significant variance between issues of a scientific nature and those that are not. Higher accuracy was noted in identifying claims, data, and qualifiers, as opposed to rebuttals, backing, and warrants. The study illuminated AI's promise for educational applications but also its shortcomings, such as the increased frequency of erroneous element identification when accuracy was low. This highlights the essential need for more in-depth comparative research on models and the further development of AI to enhance its role in supporting argumentation training.*

**Keywords:** *argumentative writing, artificial intelligence, automated assessment, natural language processing, web architecture*

**Hunkoog Jho**  
Dankook University, Korea  
**Minsu Ha**  
Seoul National University, Korea



motivation to learn. It also helps students develop the ability to evaluate the validity of scientific claims and engage in constructive discussions with peers and educators.

It is expected that argumentation plays a crucial role in achieving the abilities necessary to the future. The rapid development of science and technology brought about a variety of communicating media, such as the Internet, social media, and the Internet of Things, and enabled the public to access enormous amounts of information and knowledge. However, such a change also brought about mis/disinformation and made it worse combining with the uncertainty. In order to make the right decision, a person should be able to make judgment about the reliability of the given information along with its source and establish logical inferences by taking into account various perspectives and con on the ground of abundant data and evidence. A series of thoughts and inferences are often implemented in argumentation. The previous studies turned out that argumentation-related activities are effective to improve such abilities. However, examining the elements and schemes of argumentation is, to some extent, subjective, and there are differences in identifying them among evaluators. Accordingly, argumentative activities have been utilised not through immediate assessment or feedback for a large number of students, but rather through focused one-on-one meetings or discussion formats with a small number of students.

The recent significant progress in technology related to artificial intelligence is beneficial in overcoming these difficulties. Traditionally, automated assessment and feedback have been conducted in the context of multiple choice questions or objective type of questions. Since the 2010s, deep learning technologies in natural language processing have made it possible to quickly and easily understand sentences, analyse them, and give instructions (Wambsganss et al., 2022). In particular, a transformer model consisting of multiple encoders and decoders improves the performance in generating, classifying and summarizing texts (Gillioz et al., 2020; Jho, 2023; Vaswani et al., 2017; Zhang et al., 2022).

In recent years, there has been an upsurge in research focused on the assessment and feedback of students' argumentative essays utilizing artificial intelligence technologies such as machine learning and deep learning. Zhai et al. (2023) employed a cognitive diagnostic model to analyse students' arguments from three dimensions: claim, evidence, and warrant. Human-scored data are automatically evaluated using their developed tool, the Constructed Response Classifier (CRC) to build scoring models. The CRC is an ensemble model that concurrently harnesses eight classification algorithms (Martin et al., 2023). On the other hand, Wilson et al. procured high-quality data based on credible argumentative prompts and a systematic evaluation rubric. They then optimized the automated assessment model by applying an ensemble approach that leverages eight distinct machine learning classification models (Wilson et al., 2023). Additionally, Martin et al. (2023) conducted a study that utilized data-driven clustering analysis and unsupervised learning to categorize and assess argument patterns. Their research encompasses the utilization of pre-trained large-scale language models, various deep learning techniques, and collaboration with humans in analysing students' argumentative essays (Zhai et al., 2023). While these are all recent studies, there has been no research to date that exclusively entrusts the assessment of argumentative essays to generative artificial intelligence.

Recent developments in Natural Language Processing (NLP), including large language models, have made it possible to go beyond merely generating text to also analysing and evaluating reasoning. NLP refers to the field of computer science and artificial intelligence that focuses on the interaction between computers and human language (Rothman & Gulli, 2022). It involves the development of algorithms and models that enable computers to understand, interpret, generate, and respond to human language in a way that is both meaningful and contextually relevant. NLP encompasses a wide range of tasks, including text analysis, speech recognition, language translation, sentiment analysis, and chatbot development, among others, with the goal of enabling machines to communicate with humans in a manner that resembles natural human language understanding and generation. With the release of pre-trained models, it has become easier to take use of generative AI models in a variety of situations. The release of ChatGPT in 2022 had a great impact on societies and it became aware that large language models can be utilised in various situations such as inference, problem solving, and image generation. A Large Language Model (LLM) refers to an artificial intelligence system designed to understand, interpret, generate, and engage with human language at a large scale, often consisting of billions of parameters (Amaratunga, 2023). There are a number of LLMs such as ChatGPT by OpenAI, LLaMA (Large Language Model Meta AI), Claude, PaLM (Pathways Language Model) and Alpaca. The use of LLMs in argumentation become heeded in the community of NLP and education. Many of researchers conduct a study whether the use of AI helps students to evaluate and improve their ability relevant to argumentation based on a variety of schemes (Walton, 2016; Martin et al., 2023; Spector & Mar, 2019). In particular, Toulmin's model is useful to analyse the detailed information about the argumentation and to give a visual interpretation of the logic-based argumentation (Bentahar et al., 2010; Racharak & Tojo, 2021).

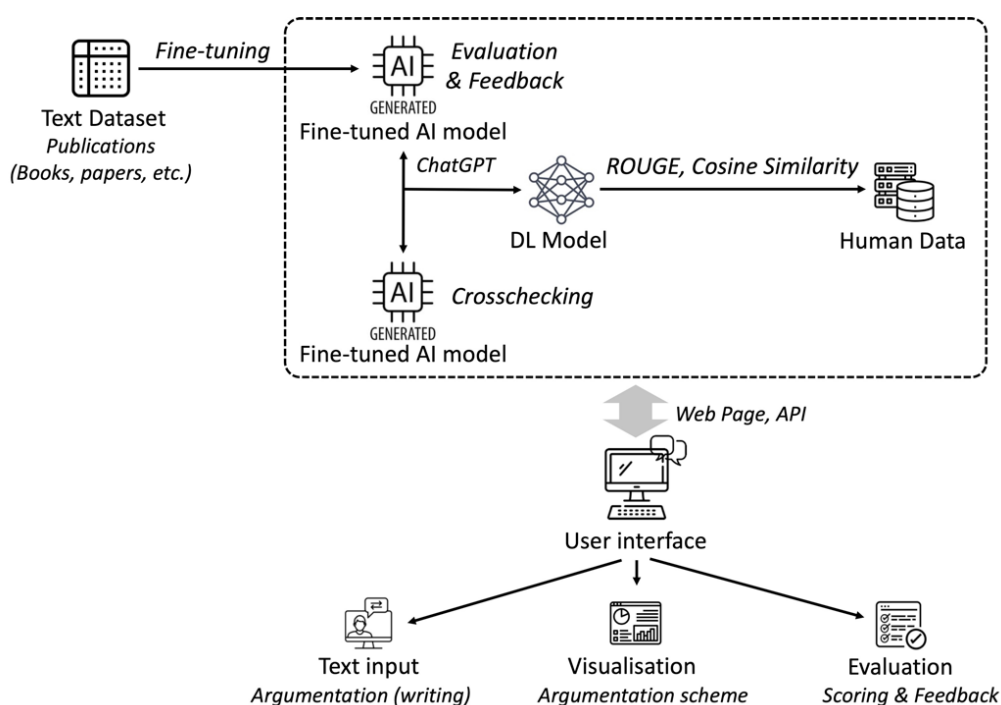
### Research Focus

This research aimed to develop a web-based framework that utilizes large language models to extract, evaluate, and provide feedback on argumentative elements within texts written by learners, and to verify its performance. Utilising ChatGPT, which has recently gained significant attention, this framework is designed to identify six argumentative elements from students' writings, assess the level of argumentation, and provide results accordingly. The performance of the artificial intelligence-generated results is evaluated by comparing them with assessments made by human experts. Through this, the study intends to offer insights into the role artificial intelligence can play, particularly in educational activities related to argumentation.

### Research Methodology


#### General Background

Figure 1 represents how the system you developed analysed students' data and provided students with feedback. While there are various methods for analysing argumentation based on its structure or the credibility and accuracy of its content, this research focuses on the structure of argumentation. Following the widely utilised definition by Stephen Toulmin (2003) in research evaluating the elements and structure of argumentation, the elements of argumentation have been defined as six components: claim, data, warrant, backing, qualifier, and rebuttal. The system handling the input and output of the texts to be analysed is divided into a web-based client page and a server responsible for the actual analysis. When a user submits a text through the client page, the server scores the six argumentation elements and the level of argumentation by interfacing with ChatGPT and sends related feedback to the client, which can then be viewed on the webpage. Moreover, to assess the accuracy of the extracted argumentation elements, the results generated by artificial intelligence were compared with those evaluated and extracted by human experts using various natural language processing techniques such as ROUGE and Cosine Similarity. In particular, the performance was verified by distinguishing whether the given problem situation was scientific or not. Through this, the study sought to provide implications for how large language models can be utilised in evaluating or classifying various students' works and what methods are necessary for their utilisation.

**Figure 1***The Architecture of Automated Assessment and Feedback Using an Generative AI Model*

The system provides several visualised screens once users input their writings on the website. The website explains what elements are included in their writings, evaluates their writing according to the rubric suggested by Osborne and others, gives feedback for better writing, and shows up a diagram of the relationships among the elements. The website was developed using Python, and the client page was run by HTML and JavaScript and communicated using Ajax.

**Figure 2**  
*Screenshots of Web Sites Showing the Main Functions to Extract Key Elements of TAP, to Visualise the Scheme and to Give Scores and Feedback to Users*



## Analysis & evaluation of argumentation using GPT

---

Please select one problem among the following ones. Salt and Metabolism

Based on the following texts (A) and (B), write down your argument with evidence.

(A) Salt is merely a crystalline structure formed by the bonding of a single sodium atom with a chlorine atom, and the amount of salt required by humans is no more than about 3 grams per day. However, once ingested, salt divides into sodium and chloride ions, significantly impacting metabolism. For instance, it becomes a major component of bodily fluids such as blood and gastric juice, aiding in the transport of nutrients throughout the body and facilitating the excretion of various metabolic wastes through sweat and urine.

(B) The sodium and chloride ions, once separated in the body, readily bind with water, effectively drawing it away from cells. Consuming overly salty food can lead to insufficient hydration of body cells, preventing them from functioning correctly. The depletion of cellular water by salt diminishes our metabolic efficiency. Furthermore, excessive salt intake can narrow blood vessels, complicating the transport of vitamins, trace nutrients, enzymes, and proteins to the cells.

Please input your name (max: 20 letters) Joseph

When salt splits into sodium and chloride ions within the body, it readily binds with water, meaning it has the property of drawing water away from cells. Considering this characteristic, consuming excessively salty food can lead to inadequate hydration of body cells, preventing them from functioning properly. Furthermore, the intake of foods high in salt can constrict blood vessels, making it difficult for vitamins, enzymes, and proteins to move to the cells. Despite salt's ability to become a major component of bodily fluids, distributing nutrients throughout our body efficiently, organisations like the WHO and AHA recommend a daily salt intake of less than 5 grams. Therefore, it is imperative that we humans avoid excessive consumption of salt.

Send Abort

---

Your request is complete.

Analysis Visualization Evaluation and Feedback

<b>Claim</b>	Consuming excessively salty food can lead to inadequate hydration of body cells and prevent them from functioning properly.
<b>Rebuttal</b>	Not included.
<b>Data</b>	Salt splits into sodium and chloride ions within the body and readily binds with water, drawing water away from cells.
<b>Warrant</b>	The property of salt drawing water away from cells leads to inadequate hydration and improper functioning of body cells.
<b>Backing</b>	Organisations like the WHO and AHA recommend a daily salt intake of less than 5 grams.
<b>Qualifier</b>	It is imperative that humans avoid excessive consumption of salt.

Analysis Visualization Evaluation and Feedback

**Data**

Salt splits into sodium and chloride ions within the body and readily binds with ...

**Qualifier**

It is imperative that humans avoid excessive consumption of salt.

**Claim**

Consuming excessively salty food can lead to inadequate hydration of ...

→

**Warrant**

The property of salt drawing water away from cells leads to inadequate ...

↑

**Backing**

Organisations like the WHO and AHA recommend a daily salt intake of less ...

↑

Analysis Visualization Evaluation and Feedback

<b>Score</b>	4 Point(s) (Argumentation shows arguments with a claim with a clearly identifiable rebuttal. Such an argument may have several claims and counterclaims as well.)
<b>Feedback</b>	

The researchers first developed writing tasks with different contexts in order to figure out whether the contextual conditions bring about any differences in the user's quality of argumentative reasoning: inference of a person's nationality grounded on the given information and the cost and benefit of salt intake from a metabolic perspective.

The system is composed of three different models for evaluation and feedback; one model (ChatGPT) extracts each element of TAP and provides feedback; the other model (ChatGPT) examines whether the elements are properly included in the writing; and finally, another transformer-based model compares the similarities between automated and human-made results relying on ROUGE score. A website is operated on Linux and developed in Python.

### *Sample*

In this study, essays written by upper secondary school students were evaluated to examine the feasibility of utilising artificial intelligence for extracting argumentative elements in educational contexts. Scientific argumentation is considered a crucial skill not only for students aspiring to enter the fields of science and technology but also for all students from the perspective of scientific literacy that citizens should possess. Therefore, eleventh graders who have completed the common curriculum were selected as the study subjects. Since scientific argumentation is important not only in scientific contexts but also in everyday situations, students were asked to write texts on problems in both scientific and non-scientific scenarios: a scientific question about the impact of salt intake on metabolism and a non-scientific question regarding nationality determination based on nationality law.

Responses were gathered from a total of 180 students across two schools regarding two distinct problems. These responses were organized into a dataframe for analysis. Both ChatGPT and human experts assessed these responses, identifying argumentative elements based on the Toulmin Argumentation Pattern (TAP) framework. The analyses by ChatGPT were performed using the widely adopted ChatGPT-3.5 version, accessed through API calls. The consistency between ChatGPT's categorizations and those made by human experts was then compared to evaluate alignment.

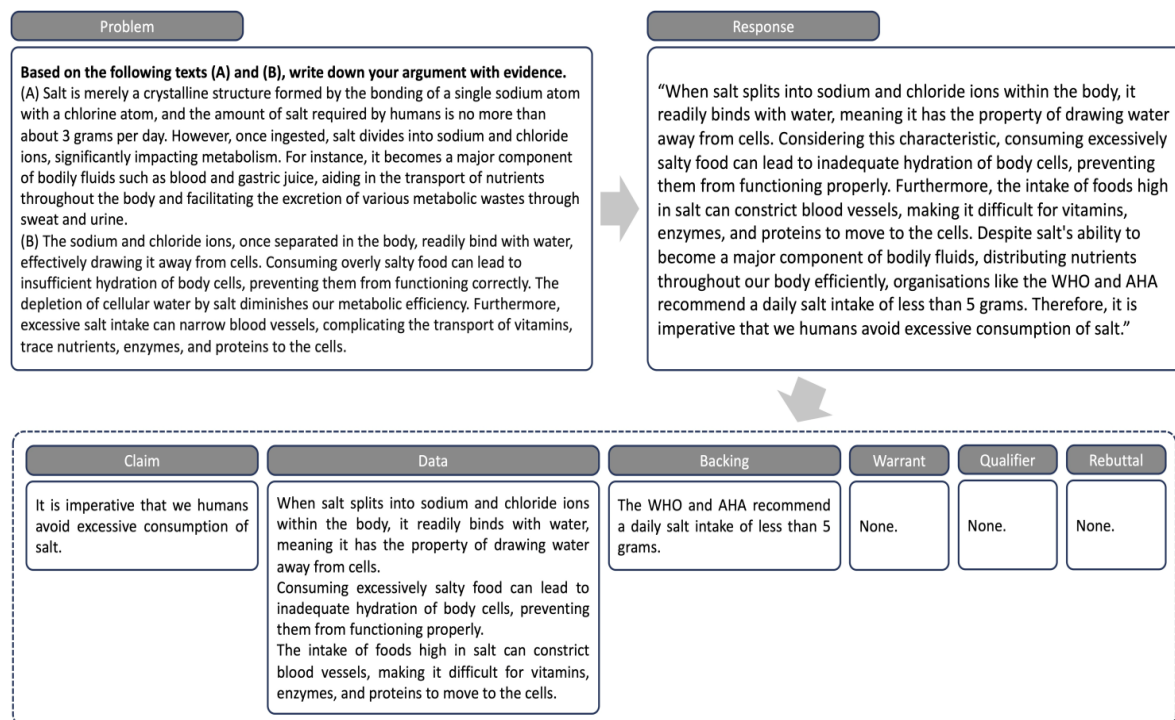
### *Instrument and Procedures*

In this research, the texts written by students were analysed based on the extraction of the six argumentation elements proposed by Toulmin. Building upon previous studies that utilised these elements to assess the level of argumentation, this study aimed to measure the level of argumentation present in the students' texts. One of widely used ones is Toulmin's Argumentation Patterns (TAP), advocated by Toulmin (2003). It consists of six key elements as follows:

1. **Claim:** The central statement or proposition that the argument is trying to establish or prove. It represents the main point or conclusion of the argument.
2. **Grounds (Evidence):** The evidence or data that support the claim. Grounds serve as the foundation for the argument and provide reasons for accepting the claim.
3. **Warrant:** The underlying reasoning or logic that connects the grounds to the claim. It explains why the provided evidence is relevant and sufficient to support the claim.
4. **Backing:** Additional evidence or support that reinforces the warrant. Backing may not always be present in every argument but can provide further assurance of the argument's validity.
5. **Qualifier:** An acknowledgment of the degree of certainty or scope of applicability of the argument. Qualifiers help clarify the strength of the claim, making it clear whether the claim is absolute, probabilistic, or conditional.
6. **Rebuttal (Reservation):** Counterarguments or potential objections to the claim. Acknowledging and addressing potential counterarguments strengthens the overall argument by showing that the author has considered alternative perspectives.

As shown in Figure 3, each of student's text were categorised into six argumentation elements by both AI and human experts.

**Figure 3**  
An Example of How to Analyse Students' Writing Based on the Six Argumentation Elements Given to the Problems



This framework facilitates the evaluation of the structural aspects of argumentation based on whether each of the elements is present. Osborne et al. (2004) proposed a new framework that allows the level of argumentation to be presented in five stages, based on the six elements. This framework has been referenced in various studies, and it was also followed in this research (see Table 1).

**Table 1**  
Analytic Framework Based on Toulmin's Argumentation Patterns (Osborne et al., 2004)

Level	Description
1	Level 1 argumentation consists of arguments that are a simple claim versus a counterclaim or a claim versus a claim.
2	Level 2 argumentation has arguments consisting of claims with either data, warrants, or backings, but do not contain any rebuttals.
3	Level 3 argumentation has arguments with a series of claims or counterclaims with either data, warrants, or backings with the occasional weak rebuttal.
4	Level 4 argumentation shows arguments with a claim with a clearly identifiable rebuttal. Such an argument may have several claims and counterclaims as well, but this is not necessary.
5	Level 5 argumentation displays an extended argument with more than one rebuttal.

*Data Analysis*

The performance of AI-generated assessment was followed by ROUGE scores and cosine similarity. ROUGE refers to an important metric used to evaluate the performance of automatic summarization and text generation models. It is primarily employed in machine learning and natural language processing research for evaluation purposes. Rouge scores measure how well a model's output, such as a summary or translation, matches the reference text. Here, the researchers analysed three different kinds of ROUGE metrics: ROUGE-1, ROUGE-2, and ROUGE-L.

Rouge-N (N-gram overlap) counts the number of overlapping N-grams (consecutive sequences of N words) between the model-generated summary or translation and the reference text. Rouge-L (Longest Common Subsequence): Rouge-L calculates the length of the longest common subsequence (LCS) between the model's summary or translation and the reference text. It considers word order and helps capture the sequence of words (Lin, 2004; Mitrović & Müller, 2015). In this study, the researcher took into account the ROUGE 0.3 score as a benchmark to compare and determine the correspondence between human evaluations and scoring results.

On the one hand, in the field of natural language processing, it is widely common to determine the similarity between two documents or sentences using cosine similarity. In this study, the researchers chose to employ BERT to obtain document-based embedding vectors and calculate the cosine similarity between the two vectors. Finally, this research took into account both metric methods to examine the similarities between human- and computer-rated assessment quantitatively.

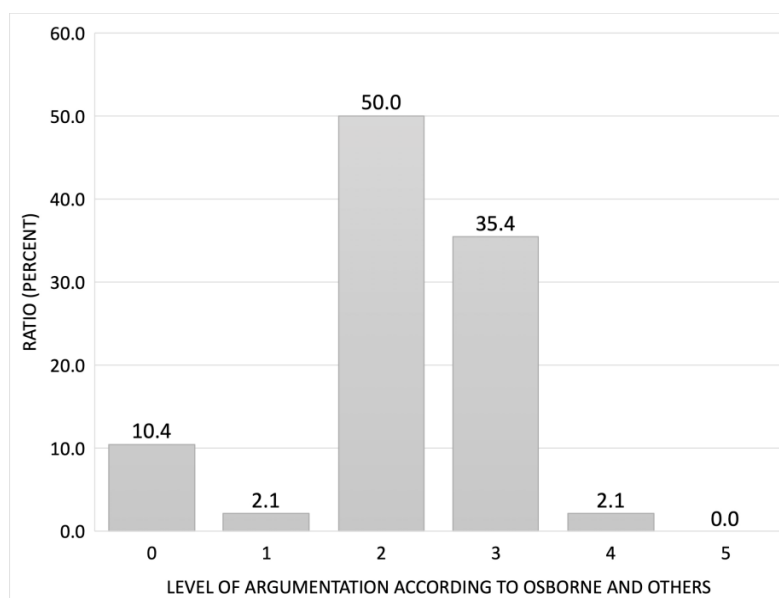
## Research Results

In this study, students were presented with argumentative problems encompassing both scientific and non-scientific contexts, and an analysis was conducted on the level of their arguments as well as the argumentative elements they contained. Accordingly, the aim was to investigate the nature and level of argumentative elements present in students' writings, centred around the contexts provided in the problems.

First of all, many of the students showed lower levels of argumentation skills. According to the rubric suggested by Osborne et al. (2004), students with high levels of argumentative skills tend to use multiple evidence and to consider more elements of argumentation. However, students in this study used single evidence or showed a simple structure of argumentation consisting of a claim solely. As shown in Figure 4, 10.4% of the students (Level 0) did not clearly present their arguments, and simple cases involving only the claim and one other element accounted for 50.0% of the total. Students exhibiting a complex argument structure (Level 4) were only a mere 2.1%. Even no students showed up the highest level.

**Figure 4**

*A Percentage of Students' Argumentative Skills according to the Rubric Suggested by Osborne et al. (2004)*

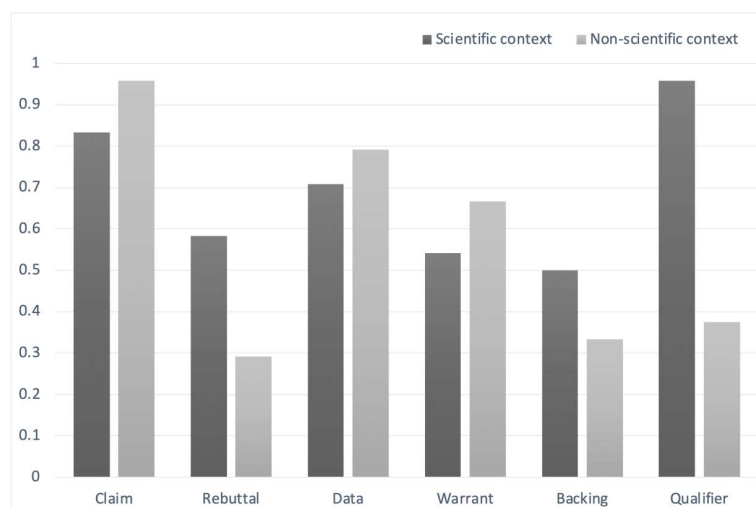


When evaluating the accuracy of argumentative elements extracted from students' writings, it is obvious that the accuracy in scientific contexts is significantly higher compared to non-scientific contexts. In particular, there is a notable difference in the aspects of warrant, backing, and qualifying. This may be attributed to the conjecture that relevant information is more easily accessible depending on the context of the text. However, it is also related

to the fact that students often fail to include these relevant elements in their arguments due to their lower level of argumentation skills.

**Figure 5**

*Comparison of Performance in Extracting Argumentative Elements according to the Different Contexts*



In this study, the researchers aimed to validate evaluation capabilities between humans and artificial intelligence, in addition to assessing the accuracy of argumentative element extraction based on given contexts. Table 2 presents a comparison of the accuracy of argumentative elements extracted by humans and artificial intelligence. Generally, ROUGE scores and Cosine Similarity show similar results, but there are differences in some elements. When compared to the actual accuracy results, ROUGE scores appear to be superior to Cosine Similarity, which considers the angle between embedding vectors.

As Figure 5 illustrates, Table 2 also reflects lower agreement in elements included in highly argumentative texts, possibly related to a lack of data for training and testing, as observed. Furthermore, backing and qualifying pose challenges as they require inferring omitted argument structures, unlike claims or evidence, which is difficult even for experts and spends a significant amount of time.

**Table 2**

*Comparison between AI-rated and Human-rated Assessment in Argumentation Elements*

Context	Argumentation Element	ROUGE	Cosine Similarity
Scientific	Claim	0.496	0.504
	Rebuttal	0.583	0.583
	Data	0.536	0.440
	Warrant	0.141	0.364
	Backing	0.502	0.523
	Qualifier	0.491	0.475
Non-scientific	Claim	0.675	0.587
	Rebuttal	0.203	0.229
	Data	0.434	0.506
	Warrant	0.360	0.395
	Backing	0.250	0.300
	Qualifier	0.303	0.304



The gap in evaluation results between humans and artificial intelligence often arises from cases where the AI fails to find the relevant argumentative elements altogether. Table 3 summarises cases of inconsistency between experts and artificial intelligence, where the AI either falsely identifies elements or fails to detect them when the experts do. According to this table, the discrepancy is more prevalent in elements other than claims and evidence, and as mentioned earlier, it is relatively higher in backing and qualifying compared to the frequency of these elements in the written content. This suggests that while the extraction of explicit content is relatively easier for artificial intelligence, tasks requiring multiple stages of reasoning and inference still present limitations.

**Table 3**

*Assessment without Human-rated or AI-rated Result in Examining the Elements of Argumentation*

Context	Argumentation Element	No human result	No AI result
Scientific	Claim	2	0
	Rebuttal	0	10
	Data	0	5
	Warrant	2	8
	Backing	0	11
	Qualifying	0	12
Non-scientific	Claim	0	0
	Rebuttal	0	16
	Data	3	0
	Warrant	2	4
	Backing	0	16
	Qualifying	1	13

Most of the inaccurate results occur when experts determine the absence of the corresponding elements. This can be attributed to the tendency of generative artificial intelligence to infer these elements even when they are absent in the text. In contrast to training datasets for artificial intelligence, where missing or false information is rare, this study on argument analysis showed that in most cases, there were relatively fewer relevant elements. This characteristic is primarily observed in algorithms like GPT series, as opposed to encoding-centric algorithms such as BERT.

## Discussion

This study's findings highlight the possibilities and limitations of generative artificial intelligence in educational settings, particularly when using Large Language Models (LLMs) like ChatGPT for assessing argumentation elements based on Toulmin's framework. LLMs demonstrate remarkable accuracy for certain elements, yet there is a significant disparity in their capability to extract different elements, as well as a marked difference in performance depending on whether the context is scientific or not. This suggests that the role of LLMs as evaluators in place of teachers or experts appears to be challenging. Despite the advancements in LLM performance over recent years, including their enhanced capacity for rational reasoning, this study reveals inconsistencies in their ability to recognise argumentation elements reliably.

However, the use of LLMs for assessment and feedback can be beneficial for students to self-diagnose their capabilities and gain a variety of ideas. Despite the limitations of artificial intelligence in extracting complex logical structures or argumentative elements, various studies applying AI to argumentative activities have shown that it can have a positive effect on generating creative ideas (Guo et al., 2022; Guo et al., 2023; Heeg & Avraamidou, 2023; Kim et al., 2022; Rapanda & Walton, 2016; Urhan et al., 2024). ChatGPT can inspire creative ideas and stronger arguments among students, but it falls short in supporting the reasoning process effectively.

The challenge of misinformation during the argumentation process has been noted, with initial wide-scale public introductions of LLMs leading to significant media coverage on the issue. Developers have made considerable efforts to mitigate these concerns. The transition to Transformer-based models, including BERT for information extraction and GPT for natural text generation, marked a significant evolution in LLM capabilities. However, the predominance of GPT-based models, despite their success in generating natural texts, underscores a vulnerability in producing reliable outputs. Recent literature advocates for the combination of multiple models to address this issue, a strategy that our study explored but found limited in rectifying inaccuracies in extracted argumentative elements.

The findings underscore the need for further development before generative AI can be effectively applied in educational settings. The generic training of most LLMs does not inherently cater to specific educational needs. For a more effective application, educators might need to finetune these models with datasets tailored to their specific objectives or adjust the models using techniques like knowledge distillation. This study revealed performance disparities between scientific and non-scientific contexts and in identifying different argumentative elements, suggesting that educators should develop targeted tasks to evaluate argumentation, gather student responses, and refine LLM quality accordingly.

Fortunately, the accessibility of web-based repositories offers numerous opportunities for obtaining and improving pre-trained models for text analysis at no cost. Future research should focus on comparing the performance of various LLMs across identical tasks to better understand and mitigate the limitations observed in this study.

## Conclusions and Implications

The purpose of this study was to establish a system for automated extraction and feedback for argumentation using an LLM and to examine its performance while comparing the results generated by human and AI. The results showed significant variance in the extraction performance of argumentation elements based on several factors. The use of LLM showed meaningful results in extracting claims, data, and qualifiers with over 80% accuracy, whereas rebuttals, backing, and warrants exhibited lower accuracies in the range of 50-60%. Furthermore, the rate of hallucination, where LLMs incorrectly identified non-existent elements as present, was higher for elements with lower accuracy. This issue is presumed to be partly due to the fact that texts written by students often lack relevant argumentation elements. In addition, for elements with low accuracy, the content evaluated by humans was also inconsistent or inaccurate, which could be attributed to the lack of specific and clear definitions for backing and warrants in Toulmin's model. The study also found significant differences in accuracy depending on whether the problem was scientific in nature, with tasks related to scientific contexts showing higher accuracy for claims, data, and warrants, while the opposite was true for the other elements. In particular, a very large variance was observed in non-scientific contexts. After all, this study indicates that directly applying LLMs developed for general purposes in educational settings could lead to various accuracy issues. In addition, as performance differences could exist between different models for such purposes, further research comparing them is necessary.

## Declaration of Interest

The authors declare no competing interest.

## Acknowledgement

This work was supported by the Ministry of Education of the Republic of Korea and the National Research Foundation of Korea (NRF-2021S1A5A2A03061991).

## References

- Amaratunga, T. (2023). *Understanding large language models: Learning their underlying concepts and technologies*. Apress.
- Bentahar, J., Moulin, B., & Bélanger, M. (2010). A taxonomy of argumentation models used for knowledge representation. *Artificial Intelligence Review*, 33, 211–259. <https://doi.org/10.1007/s10462-010-9154-1>
- Berland, L. K., & Hammer, D. (2012). Framing for scientific argumentation. *Journal of Research in Science Teaching*, 49(1), 68–94. <https://doi.org/10.1002/tea.20446>
- Duschl, R. A., Schweingruber, H. A., & Shouse, A. W. (2007). *Taking science to school: Learning and teaching science in grade K-8*. National Academies Press.
- Gillioz, A., Casas, J., Mugellini, E., & Khaled, O. A. (2020, December 6-9). Overview of the transformer-based models for NLP tasks. In *Proceedings of the 15th Conference on Computer Science and Information Systems (FedCSIS 2020)*. Sofia, Bulgaria.
- Guo, K., Zhong, Y., Li, D., & Chu, S. K. W. (2023). Investigating students' engagement in chatbot-supported classroom debates, Interactive Learning Environments. <https://doi.org/10.1080/10494820.2023.2207181>
- Heeg, D. M., & Avraamidou, L. (2023). The use of artificial intelligence in school science: A systematic literature review. *Educational Media International*, 60(2), 125–150. <https://doi.org/10.1080/09523987.2023.2264990>
- Hodson, D. (2011). *Looking to the future: Building a curriculum for social activism*. Sense Publishers.
- Jho, H. (2023). Understanding of generative artificial intelligence based on textual data and discussion for its application in science education. *Journal of the Korean Association for Science Education*, 43, 307–319.
- Kim, M., Kim, N., & Heidari, A. (2022). Learner experience in artificial intelligence-scaffolded argumentation. *Assessment & Evaluation in Higher Education*, 47(8), 1301–1316. <https://doi.org/10.1080/02602938.2022.2042792>
- Li, Y., & Guo, M. (2021). Scientific literacy in communicating science and socio-scientific issues: Prospects and challenges. *Frontiers in Psychology*, 12, Article 758000. <https://doi.org/10.3389/fpsyg.2021.758000>
- Lin, C.-Y. (2004, July 21-26). ROUGE: A package for automatic evaluation of summaries. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics* (pp. 74–81). Barcelona, Spain.
- Martin, P. P., Kranz, D., Wulff, P., & Graulich, N. (2023). Exploring new depths: Applying machine learning for the analysis of student argumentation in chemistry. *Journal of Research in Science Teaching*. <https://doi.org/10.1002/tea.21903>
- Mitrović, S., & Müller, H. (2015, September 8-11). Summarizing citation contexts of scientific publications. In *Proceedings of the International Conference of the Cross-Language Evaluation Forum for European Languages*. Toulouse, France.
- National Research Council. (2011). *A framework for K-12 science education: Practices, crosscutting concepts, and core ideas*. National Academies Press.
- Osborne, J., Erduran, S., & Simon, S. (2004). Enhancing the quality of argumentation in school science. *Journal of Research in Science Teaching*, 41(10), 994–1020. <https://doi.org/10.1002/tea.20035>
- Racharak, T., & Tojo, S. (2022). On the relationship with Toulmin method to logic-based argumentation. In A.P. Rocha, L. Steels, & J. van den Herik (Eds.), *Agents and Artificial Intelligence: ICAART 2021* (Lecture Notes in Computer Science, Vol. 13251). Springer, Cham. [https://doi.org/10.1007/978-3-031-10161-8\\_10](https://doi.org/10.1007/978-3-031-10161-8_10)
- Rapanda, C., & Walton, D. (2016). The use of argument maps as an assessment tool in higher education. *International Journal of Educational Research*, 79, 211–221. <http://dx.doi.org/10.1016/j.ijer.2016.03.002>
- Ratcliffe, M., & Grace, M. (2003). *Science education for citizenship: Teaching socio-scientific issues*. Open University Press.
- Rothman, D., & Gulli, A. (2022). *Transformers for neural language processing: Build, train, and fine-tune deep neural network architectures for NLP with Python, Hugging Face, and OpenAI's GPT-3, ChatGPT, and GPT-4*. Packt Publishing.
- Spector, J. M., & Ma, S. (2019). Inquiry and critical thinking skills for the next generation: From artificial intelligence to human intelligence. *Smart Learning Environments*, 6, 8. <https://doi.org/10.1186/s40561-019-0088-z>
- Tsai, C.-Y. (2015). Improving students' PISA scientific competencies through online argumentation. *International Journal of Science Education*, 37, 321–339. <https://doi.org/10.1080/09500693.2014.982229>
- Toulmin, S. E. (2003). *The uses of argument*. Cambridge University Press.
- Urhan, S., Gençaslan, O., & Dost, Ş. (2024). An argumentation experience regarding concepts of calculus with ChatGPT. *Interactive Learning Environments*. <https://doi.org/10.1080/10494820.2024.2308093>
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, A. (2017). Attention is all you need. <https://arxiv.org/abs/1706.03762>
- Walton, Douglas. (2016). Some Artificial Intelligence Tools for Argument Evaluation: An Introduction. *Argumentation*, 30, 317–340. <https://doi.org/10.1007/s10503-015-9387-x>
- Wambsganss, T., Janson, A., & Leimeister, J. M. (2022). Enhancing argumentative writing with automated feedback and social comparison nudging. *Computers & Education*, 191, 104644. <https://doi.org/10.1016/j.compedu.2022.104644>
- Wilson, C. D., Haudek, K. C., Osborne, J. F., 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>

- Zhai, X., Haudek, K. C., & Ma. W. (2023). Assessing argumentation using machine learning and cognitive diagnostic modeling. *Research in Science Education*, 53, 405-424. <https://doi.org/10.1007/s11165-021-09982-2>
- Zhang, F., An, G., & Ruan, Q. (2002, October 21-24). Transformer-based natural language processing and generation. In *Proceedings of the 16th IEEE International Conference on Signal Processing*. Beijing, China.

Received: January 14, 2024

Revised: March 22, 2024

Accepted: April 02, 2024

Cite as: Jho, H., & Ha, M. (2024). Towards effective argumentation: Design and implementation of a generative AI-based evaluation and feedback system. *Journal of Baltic Science Education*, 23(2), 280-291. <https://doi.org/10.33225/jbse/24.23.280>



**Hunkoog Jho**

PhD, Physics Education, Associate Professor, Department of Science Education, Dankook University, 152, Jukjeon-ro, Suji-gu, Yongin-si, Gyeonggi-do, Korea.  
E-mail: [hjho80@dankook.ac.kr](mailto:hjho80@dankook.ac.kr)  
Website: <https://www.dankook.ac.kr>  
ORCID: <https://orcid.org/0000-0002-8740-6550>

**Minsu Ha**  
(Corresponding author)

PhD, Associate Professor, Department of Biology Education, Seoul National University, 1, Gwanak-ro, Gwanak-gu, Seoul, Korea.  
E-mail: [msha101@snu.ac.kr](mailto:msha101@snu.ac.kr)  
Website: <https://biologyedu.snu.ac.kr>  
ORCID: <https://orcid.org/0000-0003-3087-3833>

