ArgVantage: the New Pedagogical System to Learn Argumentation

Camélia Guerraoui[1,2,3] Paul Reisert[4] Naoya Inoue[5,2] Wenzhi Wang[1,2]
Shoichi Naito[1,2,6] Jungmin Choi[2] Irfan Robbani[5] Kentaro Inui[7,1,2]
[1]Tohoku University [2]RIKEN [3]INSA Lyon [4]Beyond Reason
[5]JAIST [6]Ricoh Company, Ltd. [7]MBZUAI

{guerraoui.camelia.kenza.q4, wang.wenzhi.r7, naito.shoichi.t1}@dc.tohoku.ac.jp, beyond.reason.sp@gmail.com

naoya-i@jaist.ac.jp, jungmin.choi@riken.jp, robbaniirfan@jaist.ac.jp, kentaro.inui@mbzuai.ac.ae

Abstract

Argumentation in education can help students develop their critical thinking, and computational models for argumentation have been developed to further assist this process. While existing argumentative systems have proven beneficial, existing works have emphasized a need for more Detailed, Visual, Interactive, and Personalized (DVIP) feedback. In response, we introduce ArgVantage, an open-source web application designed to enhance users' argumentation skills. By integrating the DVIP dimensions, ArgVantage aims to provide a comprehensive and interactive learning experience. This paper navigates through the design and implementation of ArgVantage, highlighting the rationale behind its integration of the DVIP principle and discussing feasible evaluation methods.

1 Introduction

Critical thinking is the ability to objectively analyze, evaluate, and make reasoned judgments about information, situations, or problems. It involves being receptive to different viewpoints, questioning assumptions, and making well-informed decisions based on evidence and logical reasoning. Learning argumentation has been shown to improve students' critical thinking skills[1; 2]. To assist students' learning and reduce teachers' workload, several researchers have been working on computational argumentation and specifically argumentative feedback systems[3].

Guerraoui et al., 2023[4] presents a survey on existing educational feedback systems that teach argumentation. Their survey demonstrates that although such systems are proven to assist students' learning and reduce teachers' workload[5], they still lack the ability to deeply explain how an argument can be improved. Overall, they argue that a good argumentative feedback system should be (i) Detailed, (ii) Visual, (iii) Interactive, and (iv) Personalized (DVIP), while current systems only implement some of these four dimensions.

Although this survey gives an informative overview, it lacks experimental data or empirical results, leading us to question whether these four dimensions (DVIP) are necessary and/or sufficient for an argumentative feedback system. Towards answering this question, we explore the DVIP principle by designing and implementing an end-to-end system integrating all four dimensions. In this context, we develop ArgVantage, a comprehensive system designed to enhance users' argumentation skills. ArgVantage is an open-source web application, prompting users to write counter-arguments in response to a given initial argument.

After presenting the related work(§2), we discuss the design of ArgVantage (§3), and implementation (§4) while explaining the reasoning behind its integration of the DVIP principle. While the comprehensive evaluation of this system in a classroom setting is beyond the scope of this paper, potential challenges and approaches for assessment are discussed in the last section (§5). This study seeks to offer valuable insights into computational argumentation, laying a robust foundation for implementing the DVIP principle.

2 Related Work

Argumentative Feedback Recent advancements in Large Language Models (LLMs) have led to the exploration of automated systems for generating argumentative feedback. Notably, Wang et al., 2023[6] demonstrate the

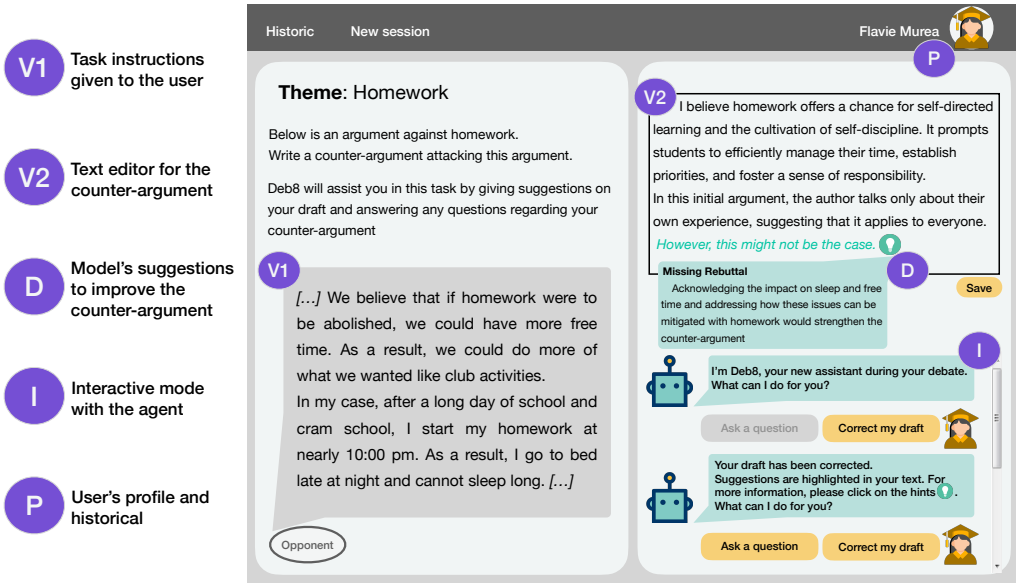


Figure 1: ArgVantage design: a **D**etailed, **V**isual, **I**nteractive and **P**ersonalized system. In this scenario, feedback has been generated and displayed after the user pressed the *"Correct my draft"* button.

feasibility of utilizing ChatGPT, a language model, to provide argumentative feedback. Similarly, Su et al., 2023[7] explore the possibilities of ChatGPT in assisting students with diverse argumentative tasks such as content revision while highlighting some limitations (e.g. ethical issues).

Focusing on writing counter-arguments, Wang et al, 2024[8] and Naito et al., 2024[9] rely on templates to generate feedback. These templates serve as a scaffold, guiding the models in finding patterns and formulating responses that adhere to logical argumentative structures. Both approaches leverage state-of-the-art models to enhance the quality of generated feedback in argumentation

End-to-end Tools Several end-to-end tools have been developed to assist in learning argumentation. Among these, Argumate[10] offers a platform that combines a mind map and an agent to assist users in crafting arguments. Similarly, Persua[11] is an interactive visual system, which provides guidance on strategies to enhance the persuasiveness of arguments. Additionally, ArgueTutor[12] focuses on guiding users through the process of constructing effective arguments by offering tailored interactive support.

While these tools contribute valuable insights, this paper distinguishes itself by emphasizing the importance of Detailed, Visual, Interactive, and Personalized feedback. Our work investigates the impact of these four dimensions on improving argumentation skills and more generally critical thinking skills, thereby extending the existing landscape of end-to-end tools in the field.

3 Design

3.1 Methodology

Research on comparative studies between existing automated and teacher-provided feedback ([13; 14]), has shown that feedback provided by models tends to be verbose, lacking specificity, and occasionally contained inaccuracies. In contrast, feedback from teachers tends to be more brief and centered on interpersonal communication.

In Guerraoui et al., 2023[4], areas for improvement to enhance the quality of educational argumentative systems are highlighted, such as: (1) how to generate more accurate, constructive feedback for real-life input, (2) how to tailor the output based on the user's profile, (3) how to evaluate and compare end-to-end systems more deeply, and (4) how to improve models' abilities to adapt to unknown topics.

Informed by these findings, our system design aims to address some of these identified limitations. The following subsection outlines the specific issues we have chosen, providing precise definitions and delving into their scope.

3.2 Goals/Functionalities

In this section, we define several Design Goals (DG) we plan to achieve by describing ArgVantage's functionalities.

DG1. Foster user's ability to counterattack an argument Following the steps of Wang et al, 2024[8] and Naito et al., 2024[9], ArgVantage aims to help users in crafting effective counter-arguments. We believe this approach offers benefits to users, considering the greater complexity of constructing counter-arguments compared to arguments. Crafting a counter-argument requires not only raising valid points but also addressing the logic of the initial argument. Moreover, we think this task is a good scenario for real-life inputs such as everyday debates.

DG2. Help user noticing logical errors in their argumentation ArgVantage is designed to provide users with valuable insights into the logical structure of their arguments, facilitating the identification and correction of potential errors, ultimately leading to the refinement of argumentative skills. Its feedback should be balanced between the verbosity of automated models and the brevity of teacher feedback to prioritize the user experience.

DG3. Guarantee interaction between the user and the agent Following the guideline of Guerraoui et al. 2023[4], promoting an interactive user-agent dynamic is fundamental for effective learning. We aim to encourage a two-way interaction, ensuring that users actively engage with our system, receive timely feedback, and have the opportunity to seek clarification or further guidance.

DG4. Minimize visual distractions Even if having graphical elements showing the overall structure of an argument is beneficial for the user[4], we decided, as a first step, to minimize the visual components of ArgVantage's interface, to optimize the user experience. The interface is designed to be clean and intuitive, focusing on the core elements necessary for effective learning without overwhelming users with unnecessary visual stimuli.

DG5. Minimal personalization It has been demonstrated that personalized feedback is essential to learning argumentation[4]. However, ensuring the system generates effective feedback takes precedence before incorporating personalization. As a result, we have temporarily set aside the personalization of feedback. Nonetheless, users have the option to review their history, made of prior counter-arguments and the corresponding feedback received. This feature enables users to track progress and gain a retrospective view of their interactions with the system, setting the stage for potential future personalization enhancements.

4 Implementation

This section outlines the implementation of ArgVantage, illustrating how key aspects align with the design goals (DG) previously presented(§3). It starts with a use case, followed by a concise explanation of the architecture.

4.1 Use Case

Upon account creation and login, a user must select a topic (e.g., *homework* or *the death penalty*), choose a stance (either *in favor* or *against* the chosen topic), and then write a counter-argument (**DG1**).

To achieve this objective, a user is guided by the Deb8 agent, with two interaction options presented as buttons: (1) *"Correct my draft"* and (2)*"Ask a question"* (**DG3**). Opting for *"Correct my draft"* allows the user to receive suggestions for improving their counter-argument. Detected logical errors in the user's counter-argument are highlighted in green (**DG2**). Hovering the mouse cursor over a highlight triggers an informative alert with detailed explanations (Figure 1). To dismiss the alert and revert the text to its original state, the user can click on the alert's close icon (**DG4**). In case the user finds Deb8's suggestions unclear, they can request further clarification by choosing the *"Ask a question"* option. The agent will then respond within a chat box, with irrelevant questions disregarded.

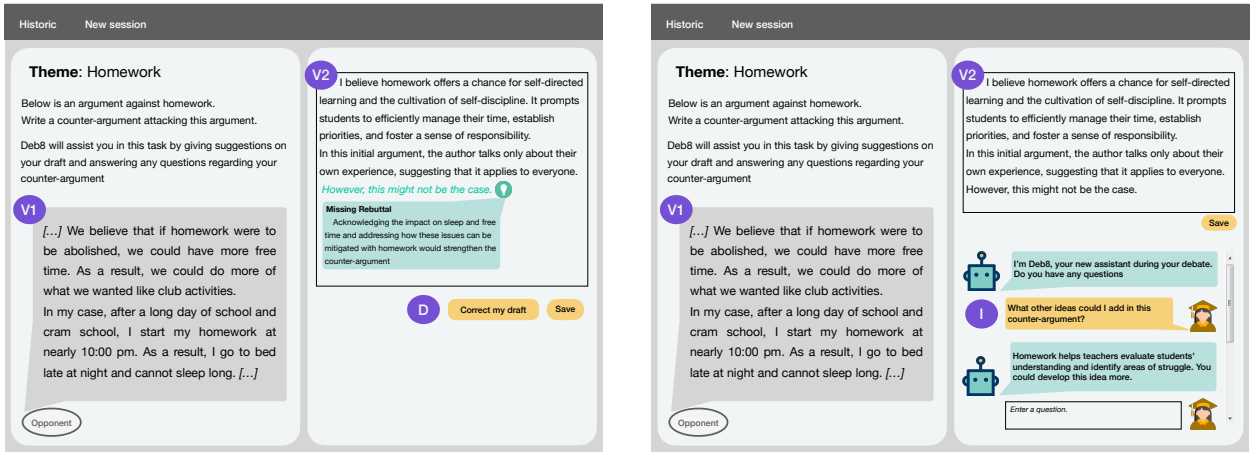
Each time the user asks a question or requests suggestions, a version of their current counter-argument is saved. Additionally, users can manually save their work by clicking the *"Save"* button. To access this historical (**DG5**), the user should access their profile by clicking on their name.

4.2 Architecture

To ensure straightforward upkeep and visibility, we opted to deploy our system as much as possible within a Free/Libre and Open Source Software environment.

Front-end We employed VueJS along with Bootstrap, creating an intuitive and responsive user interface, accessible on computers and tablets, both of which are devices that can be used in school.

Back-end Our server runs on NodeJS with Express JS, following a RESTful architecture. Secure connections are facilitated through JSON Web Tokens, and our database is powered by PostgreSQL. These choices guarantee an easily accessible and secure web system.



(a) **V**isual and **D**etailed System

(b) **V**isual and **I**nteractive System

Figure 2: Comparison between a **D**etailed and an **I**nteractive system

Deb8's Feedback The server triggers a Python script to generate feedback, utilizing the GPT4 model via its API. Although GPT4 is neither open nor free, its API facilitates easy deployment with good accuracy for feedback generation. One challenge in feedback generation is the model's need to fully comprehend the nuances of the provided argument. To address this issue, we opted for a pragmatic approach, leveraging current state-of-the-art models as a short-term solution. This enables us to promptly provide empirical solutions and explore the DVIP principle without waiting for further model development. Ultimately, we aim to integrate a study that investigates the effectiveness of various models. Actual inputs and outputs of GPT4, along with the prompt used, are detailed in Appendix 3.

Topic The initial arguments come from the TYPIC corpus[15], which contains two topics: homework and death penalty. Chosen for its openness, this resource includes human-written arguments, counter-arguments, and two types of feedback: one provided by an annotator and another generated with a template. These instances provide a valuable basis for comparison with our system. As a first step, the interface is designed within a closed domain to consider the evaluation. Nevertheless, the model is implemented in an open domain. The model does not have predefined limitations on the topics it can handle, making it versatile and more powerful in real-life settings.

Deployment The entire system is structured using a Dockerized approach, consisting of three distinct images: one for the database, another for the client, and the third for the server. This modular architecture enhances scalability and facilitates efficient management of each component.

5 Discussion and Future Work

In this section, we discuss several challenges and methods to evaluate our system, which we leave for future work.

Variables and Codependency ArgVantage involves two variables: the model-generated explanation (Textual feedback) and its interface (Visual feedback). Assessing one variable at a time is crucial for meaningful results.

Metrics Quantitative metrics such as feedback accuracy, as well as qualitative metrics, should be used to measure the system performance. Qualitative metrics involve user surveys such as the Questionnaire for User Interface Satisfaction[16], giving insights into the user experience.

Protocol As depicted in Figure 2, we intend to compare the effectiveness of a Detailed system and an Interactive system. Figure 2a represents a system that only shows textual feedback, while Figure 2b illustrates a system that interacts with the user only via questions. This experiment aims to identify the different impacts of detailed and interactive elements on critical thinking skills development.

These requirements will help measure ArgVantage's effects on learning argumentation and refining its system to ensure optimal functionality in real-class settings.

6 Conclusion

This paper navigates through the design (§3), implementation (§4), and potential evaluation(§5) of ArgVantage, elucidating the logic behind its assimilation of the DVIP principle. This study provides valuable perspectives in computational argumentation, establishing a strong foundation of the DVIP principle.

Acknowledgements

This work was supported by JSPS KAKENHI Grant Number 22H00524 and by JST CREST Grant Number JPMJCR20D2, Japan.

References

[1]R.T. Pithers and Rebecca Soden. Critical thinking in education: a review. **Educational Research**, Vol. 42, No. 3, pp. 237–249, 2000.

[2]Linda Behar-Horenstein and Lian Niu. Teaching critical thinking skills in higher education: A review of the literature. **Journal of College Teaching and Learning**, Vol. 8, , 02 2011.

[3]John Lawrence and Chris Reed. Argument mining: A survey. **Computational Linguistics**, Vol. 45, No. 4, pp. 765–818, December 2019.

[4]Camelia Guerraoui, Paul Reisert, Naoya Inoue, Farjana Sultana Mim, Keshav Singh, Jungmin Choi, Irfan Robbani, Shoichi Naito, Wenzhi Wang, and Kentaro Inui. Teach me how to argue: A survey on NLP feedback systems in argumentation. In **Proceedings of the 10th Workshop on Argument Mining**, pp. 19–34, Singapore, December 2023. Association for Computational Linguistics.

[5]Charles Twardy. Argument maps improve critical thinking. **Teaching Philosophy**, Vol. 27, , 06 2004.

[6]Li Wang, Xinya Chen, Chung Wang, Lingna Xu, Rustam Shadiev, and Yan Li. Chatgpt's capabilities in providing feedback on undergraduate students' argumentation: A case study. **Thinking Skills and Creativity**, Vol. 51, p. 101440, 2024.

[7]Yanfang Su, Yun Lin, and Chun Lai. Collaborating with chatgpt in argumentative writing classrooms. **Assessing Writing**, Vol. 57, p. 100752, 2023.

[8]Wenzhi Wang, Shoichi Naito, Paul Reisert, Naoya Inoue, Camélia Guerraoui, Jungmin Choi, Irfan Robbani, and Kentaro Inui. Exploring task decomposition for assisting large language models in counter-argument logical structure analysis. **30th Annual Meeting of the Natural Language Processing (NLP2024)**, 2024.

[9]内藤昭一, 王文質, Paul Reisert, 井之上直也, Camélia Guerraoui, 山口健史, Jungmin Choi, Irfan Robbani, 乾健太郎. 反論の論理パターン解析: データセット構築と実現性検証. **30th Annual Meeting of the Natural Language Processing (NLP2024)**, 2024.

[10]Kai Guo, Jian Wang, and Samuel Kai Wah Chu. Using chatbots to scaffold efl students' argumentative writing. **Assessing Writing**, Vol. 54, p. 100666, 2022.

[11]Meng Xia, Qian Zhu, Xingbo Wang, Fei Nie, Huamin Qu, and Xiaojuan Ma. Persua: A visual interactive system to enhance the persuasiveness of arguments in online discussion. **Proceedings of the 2022 ACM Human-Computer Interaction**, Vol. 6, No. Computer-Supported Cooperative Work and Social Computing, 2022.

[12]Thiemo Wambsganss, Tobias Kueng, Matthias Soellner, and Jan Marco Leimeister. Arguetutor: An adaptive dialog-based learning system for argumentation skills. In **Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems**, CHI '21, New York, NY, USA, 2021. Association for Computing Machinery.

[13]Joshua Wilson and Amanda Czik. Automated essay evaluation software in english language arts classrooms: Effects on teacher feedback, student motivation, and writing quality. **Computers & Education**, Vol. 100, pp. 94–109, 2016.

[14]Mohaddeseh Mehrzad Stephanie Link and Mohammad Rahimi. Impact of automated writing evaluation on teacher feedback, student revision, and writing improvement. **Computer Assisted Language Learning**, Vol. 35, No. 4, pp. 605–634, 2022.

[15]Shoichi Naito, Shintaro Sawada, Chihiro Nakagawa, Naoya Inoue, Kenshi Yamaguchi, Iori Shimizu, Farjana Sultana Mim, Keshav Singh, and Kentaro Inui. TYPIC: A corpus of template-based diagnostic comments on argumentation. In **Proceedings of the Thirteenth Language Resources and Evaluation Conference**, pp. 5916–5928, Marseille, France, June 2022. European Language Resources Association.

[16]John P. Chin, Virginia A. Diehl, and Kent L. Norman. Development of an instrument measuring user satisfaction of the human-computer interface. In **Proceedings of the SIGCHI Conference on Human Factors in Computing Systems**, CHI '88, p. 213–218, New York, NY, USA, 1988. Association for Computing Machinery.

A Appendix

Prompt instructed to GPT4

Case of "Correct my draft"

You are a personal tutor in argumentation. You will be provided an initial argument and a counter-argument attacking the initial argument.

Your task is to give suggestions to improve the counter-argument and to cite the passage(s) of the counter-argument that needs improvement. Give as many suggestions as you can to improve the counter-argument.

A suggestion should be one sentence long. Use the following format to give one suggestion:

```
{
    "suggestion": "INSERT SUGGESTION HERE",
    "citation": "INSERT CITATION HERE"
}
```

If the counter-argument is not linked to the initial argument then simply write: "The counter-argument is off-topic.".
Do not make any suggestions regarding the grammatical errors in the counter-argument.

Case of "Ask a question"

You are a personal tutor in argumentation.

You will be provided an initial argument, a counter-argument attacking the initial argument, and a question linked to the counter-argument.

You are instructed to answer the question concisely.

If the provided question is linked neither to argumentation, the initial argument's logic nor the counter-argument's logic, don't answer.

Example of an initial argument given to GPT4 (Input)

Hello everyone. Today's topic is "Homework should be abolished". We have two points: The first point is "free time" and the second point is "decrease burden on teachers". I will explain the first point of "free time". We believe that if homework were to be abolished, we could have more free time. As a result, we could do more of what we really wanted like club activities, hobbies, or playing with friends. In my case, I go to tennis club after class until 5:00 pm and then I go to cram school until 8:00 pm. After this full day, I arrive at my home around 8:40 pm to eat dinner and take a shower. At nearly 10:00 pm I start my homework. I have a lot of homework. As a result, I go to bed late at night at nearly 1:00 am in the morning and I don't have the opportunity to sleep for a long period of time. It is not healthy. Therefore, homework should be abolished. Thank you.

Example of a counter-argument given to GPT4 (Input) and its feedback in green generated by GPT4 (Output)

They said that if homework were to be abolished, we could have more free time. In other words, we can not do more of what we really wanted like club activities, hobbies, or playing with friends because of a lot of homework. *However, I think that we have enough time for doing more of what we really like.*

Provide empirical evidence or studies demonstrating that current time management allows sufficient time for hobbies and rest.

We can play with friends and devote oneself to club activities and hobbies on weekend like Saturday and Sunday.

Clarify the relevance of weekend time availability in relation to the argument about homework on weekdays.

If we don't have enough time on weekend, we should consider our ways to spend time.

In this case, we don't have enough time on weekday.

Address the health concerns associated with lack of sleep as mentioned in the initial argument to strengthen the counter-argument's position.

Offer a structured plan or provide examples of how to spend time more effectively that reconciles both educational responsibilities and personal time.

So, we have to consider our ways to spend time more effectively.

Counter the initial argument directly by explaining how homework contributes positively to student life rather than suggesting the abolition of homework might not be the solution.

Figure 3: Prompt, data inputs and outputs of ArgVantage's model