

# CAT: A CounterArgument Tutoring System based on Logical Structure Analysis

Wenzhi Wang<sup>1,2</sup>, Thimo Wambsganss<sup>3</sup>, Chenyang Wang<sup>3</sup>, Paul Reisert<sup>4</sup>,  
Léane Wettstein<sup>3</sup>, Katja Pott<sup>3</sup>, Kenshi Yamaguchi<sup>2</sup>, Machi Shimmei<sup>1,2</sup>,  
Roman Rietsche<sup>3</sup>, Surawat Pothong<sup>5</sup>, Naoya Inoue<sup>5,2</sup>, Kentaro Inui<sup>6,1,2</sup>

<sup>1</sup>Tohoku University, <sup>2</sup>RIKEN, <sup>3</sup>Bern University of Applied Sciences,

<sup>4</sup>Beyond Reason, <sup>5</sup>JAIST, <sup>6</sup>MBZUAI

Correspondence: wang.wenzhi.r7@dc.tohoku.ac.jp

## Abstract

Engaging in counterarguments (CAs) has been shown to be a promising approach to effectively support the development of critical thinking skills in educational settings. Developing an argumentation tutoring system that automatically provides personalized feedback on CAs would benefit both teachers and students by mitigating the constraints imposed by limited instructional resources. However, the development of such systems remains underexplored. To address this gap, we develop **CAT**, a CounterArgument Tutoring system that analyzes the underlying logical structure of CAs and provides tailored feedback accordingly. To evaluate CAT, we conduct a user study with 38 undergraduate students over two weeks in a real classroom, comparing it against an existing argumentation tutoring system as a baseline. The results indicate that CAT supports students in critically evaluating opposing arguments, considering alternative perspectives, and refining their own reasoning. Statistical analysis further shows that CAT elicits more engaged interaction than the baseline system. Moreover, students perceive CAT as easy to use and valuable for improving both their thinking skills and argumentation skills. The source code and demonstration materials are publicly available.<sup>1</sup>

## 1 Introduction

Critical thinking is an essential cognitive ability in higher education (Andreucci-Annunziata et al., 2023) and encompasses many skills, including analysis, explanation, interpretation, evaluation, inference, and self-regulation (Facione, 1990). Policy bodies recognize this importance by incorporating critical thinking skills into competency frameworks and curricular objectives (UNESCO, 2019; OECD, 2024; Cardona et al., 2023; World Economic Forum, 2025). Towards improving one’s critical thinking skills, engaging in argumentation is an effective

way (Kuhn, 2019; Ennis, 2011; Meral et al., 2021), especially in the task of formulating CAs (Kuhn, 1991, 2019; Roy and Macchiette, 2005; Liu and Stapleton, 2014; Johnson and Johnson, 1993; Nussbaum, 2008). In forming CAs, one must critically evaluate an opponent’s argument to increase overall effectiveness. Additionally, personalized feedback in argumentation-based education has been shown to maximize learning gains (Black and Wiliam, 2009; Hattie and Timperley, 2007).

Towards providing personalized feedback, previous work on argumentation tutoring systems has primarily focused on argument mining (Lawrence and Reed, 2019), which teaches students to construct argumentative units (claims/premises) and rhetorical relations (support/attack) within a single argument (Wambsganss et al., 2021, 2020; Zhang et al., 2023; Pan et al., 2025). In the context of CAs, several systems have been developed to help students become aware of different perspectives (Wu et al., 2014; Tanprasert et al., 2024; Pitre and Luther, 2024), or assist them with identifying CAs in an argumentative essay (Wolfe et al., 2009). Recent work has also explored AI agents that argue against users in order to support learning (Wang et al., 2026; Brender et al., 2026a,b). However, to use CAs as a means of practicing critical thinking skills, providing feedback based solely on argument mining or the identification of CAs is insufficient, as critical thinking encompasses more complex cognitive abilities, such as evaluating arguments, considering alternatives, and reflecting on one’s own judgments (Facione, 1990). Furthermore, passively presenting CAs to learners also has limited effectiveness. In contrast, actively engaging students in generating CAs leads to greater learning gains (Iordanou and Rapanta, 2021). Moreover, although engaging students in oppositional peer argumentation provides opportunities to practice generating CAs, it does not explicitly provide feedback on how to formulate a logically stronger CA.

<sup>1</sup>[https://github.com/oubunshitsu/CAT\\_demo](https://github.com/oubunshitsu/CAT_demo)

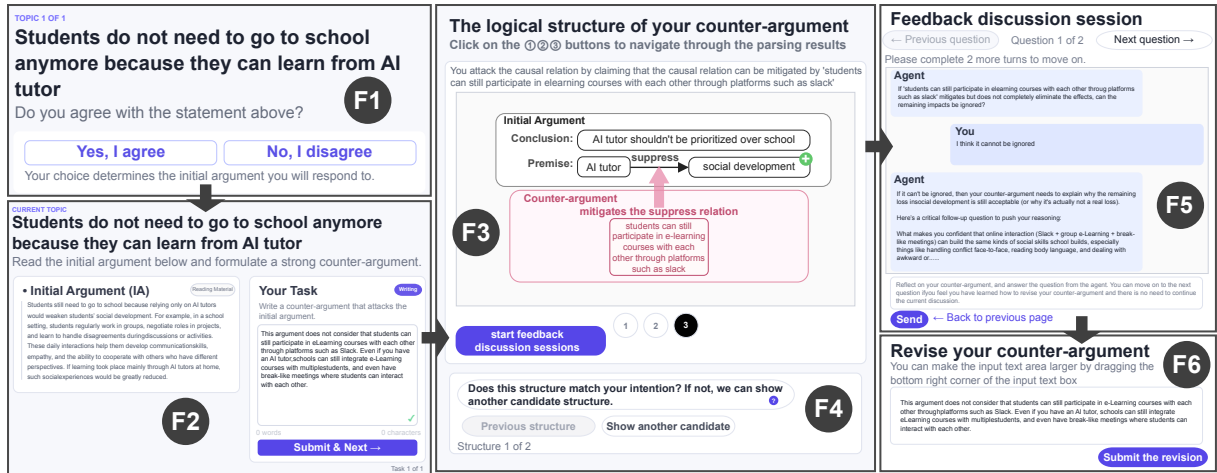


Figure 1: User Interface (UI) of CAT. The figure is based on the actual system UI, with the component sizes enlarged for readability.

Meanwhile, recent advances in NLP have opened up new avenues for in-depth analysis of the logical structure of CAs. Naito et al. (2024) propose a typology for representing the logical structure of CAs, whereas Wang et al. (2025) introduce a new task for identifying multiple possible logical structures of CAs, aiming to address the inherent logical ambiguities in CAs. Despite efforts in both the development of tutoring systems for CAs and advances in NLP, to the best of our knowledge, the existing literature has not yet explored the development of a dedicated feedback system that teaches CA formulation grounded in logical structures to improve students’ critical thinking skills.

Toward addressing this gap, we develop **CounterArgument Tutoring (CAT)**, a novel system designed to foster critical thinking by focusing on CA formulation using logical structures. CAT explicitly analyzes the logical relationship between an initial argument (IA) and its CA based on two recent CA studies in NLP (Naito et al., 2024; Wang et al., 2025), and it teaches students how to rebut the IA effectively by providing feedback on the CA’s underlying logical structure (Figure 1). CAT has two core functionalities: 1) it analyzes the logical structure of a CA and presents users with a dynamic logical structure diagram to facilitate reflection on their own reasoning; 2) based on the analyzed logical structure, it provides critical question as feedback and discusses it with users in order to improve their reasoning.

Figure 2 shows an example of the logical structure analysis of a CA compared to the traditional argument mining based analysis. Here, the IA states

that social media should be abolished for teenagers under 16 since it could cause mental health problems. The CA challenges this by arguing that social media also has benefits, such as opportunities for teenagers to connect with one another. The logical structure of the CA can be represented with the template “Although  $X$  promotes a negative outcome  $Y$ , it also promotes a positive outcome  $Z$ ”. While traditional analysis is limited in providing surface-level feedback, such as “Please add more evidence to support your claim.”, the logical structure analysis enables us to provide feedback based on how the CA attacks the IA. For instance, we can give feedback to prompt students to consider comparing the advantages of social media, as argued by the CA, with its disadvantages, as argued by the IA.

We evaluate our CAT system through a user study conducted in a classroom setting with 38 undergraduate students over a two-week period. We compare CAT against ArgueTutor (Wambsganss et al., 2021), a proven argumentation tutoring system for fostering argumentation skills, which we further adapted for CA tasks. The results from our classroom experiments indicate that CAT contributes to improvements in students critical thinking skills to some degree, particularly by enabling them to evaluate opposing arguments, consider alternative perspectives, and refine their reasoning.

Our work’s main contributions are as follows:

1. We propose the first tutoring system that provides feedback on the logical structure of CAs to foster students’ critical thinking skills.
2. We demonstrate the effectiveness of our system by conducting a user study and comparing



2003) has shown that external representations help learners reason and understand complex concepts. Prior work empirically shows the effectiveness of utilizing graphical representations in building argumentation systems (Nussbaum et al., 2007; Hsu et al., 2015; Pinkwart et al., 2009; Wambsganss et al., 2020). Based on the theoretical and empirical evidence, we use a dynamic logical structure diagram to visually show user’s CA logical structure. Each detected logical structure is rendered with a React-based SVG diagram template. The diagram components are written as JSX and transformed in the browser using Babel. CAT selects the corresponding diagram template based on the resulting logical structure, and then populates it with relevant contextual information, such as slot-fillers. The contextualized diagram is then presented to users to help them understand and reflect on their own reasoning in the CA (F3). To help users understand the logical structure diagram more easily, we also present it progressively in three steps, in which users can control the transition between steps. Furthermore, since an argumentative text can inherently have multiple possible logical interpretations (Iser, 1972; Tierney et al., 1983), CAT identifies multiple possible logical structures in the backend. Users are expected to review different candidate logical structures of their CA and select the one that best matches their intention if the default option does not (F4).

**Critical Question Discussion (F5)** Pedagogical Conversational Agents (PCAs) have been shown to be effective in helping users understand complex concepts (D’mello and Graesser, 2013; Graesser et al., 2005; Wouters et al., 2013). Moreover, engagement with critical questions has been shown to help students reason more deeply (Nussbaum, 2011). Therefore, we design a PCA to facilitate discussion of critical questions concerning logical structure, thereby encouraging users to consider these questions carefully, reflect on their reasoning, and identify their own solutions without being explicitly instructed on how to improve their CAs. The critical questions are designed in the style of those in the Argumentation Scheme (Walton et al., 2008) and are tailored to the logical structure of CAs. Each question is designed as a template and instantiated with the corresponding CA context during runtime. Once users select the logical structure that aligns with their intention, they can proceed to the feedback discussion session (F5). The discus-

sion begins with a critical question, and the subsequent interaction is conducted using GPT-5.2 with all the context. If the logical structure has multiple associated critical questions, users must complete multiple discussion sessions, respectively.

**CA Refinement (F6)** Finally, users are prompted to refine their CA based on feedback from the discussion session with the PCA (F6). The system forces them to modify their CAs in order to submit.

**Beginner Mode** If the composed CA does not have a logical structure, the user will be directed to a discussion session with a PCA, intended to help refine the CA. The user may then resubmit the refined CA, and if the revision has a valid logical structure, the user will continue using the system according to the normal procedures and functionalities described above.

By adopting such a hybrid approach that combines template-based logical structure representations and critical questions with LLM-driven interaction, CAT balances the controllability of structured methods with the flexibility and contextual sensitivity of LLMs.

## 2.2 ArgueTutor System

To evaluate CAT, we compare it with a well-established argumentation tutoring system, ArgueTutor (Wambsganss et al., 2021). The backend algorithm is based on argument mining (Lawrence and Reed, 2019), which classifies the claims and premises in an argumentative essay. ArgueTutor has three main functionalities: 1) It provides rule-based feedback based on the number of claims and premises in the user’s CA. For instance, if the CA contains only a claim without any premises, the user receives feedback indicating they should add more premises to support the claim. 2) It can display the actual claims and premises to the user if they want more details about their CA. 3) It can also display the definitions of a claim and a premise, along with examples, to support beginner learners. We show more details in Appendix B.

## 3 Automatic Evaluation

Before conducting the classroom user study, we evaluate the performance of GPT-5.2 on the task of identifying multiple logical structures. We formulate the task as a multi-class classification problem and use recall as the evaluation metric, since in real-world usage, users select only one structure

from many that matches their intention. We use the F1 score of BERTScore (Zhang et al., 2020) to evaluate slot-fillers when the logical structure template requires them. Specifically, we assess the performance of the model on the task of identifying multiple logical structures using the CALSA+ dataset which contains 134 CAs proposed in Wang et al. (2025), following their experimental setting. The task is to answer a yes/no question to represent the existence of a predicate (i.e., a decomposed fine-grained logic component) and to extract a slot-filler (i.e., a phrase) from the CA if the predicate requires one. After obtaining all predicate-level results, we aggregate them based on deterministic rules to derive the final results of the identified logical structure. We consider a model’s predicted logical structure to be correct for a CA if it matches any of the labels. We calculate the score of each predicted slot-filler against all labeled slot-fillers and take the highest score as the score for that corresponding predicted logical structure. If multiple predicted logical structures match the labels for a CA, we average the highest slot-filler scores across all matched logical structures for that instance. Finally, we average across all CA instances to obtain the final slot-filler quality score. As a result, we obtained a recall score of 0.91 and an average slot-filler quality score of 0.78, confirming that the model is capable of performing this task.

## 4 Human Evaluation

We evaluate CAT through a classroom-based user study, drawing on both students’ perceptions of the system and their performance. We are specifically interested in the following questions:

- Q1: *How do students perceive the interaction with CAT during their argumentation process?*
- Q2: *What are the impacts of a system that provides deep logical level feedback on students’ critical thinking skills compared to a system that provides feedback based on the argument mining?*

Before conducting the classroom user study, we also carried out an internal pilot study to confirm that CAT was functioning properly and that the interface was understandable. Details can be found in Appendix C.

### 4.1 Procedures of the Study

We conducted a crossover user study in which we asked two classes of bachelors students at a Western European university to use both the CAT and the ArgueTutor system. The study spanned four sessions in two different weeks. The study consists of three parts: 1) pre-test survey, 2) interaction with the system, and 3) post-test survey. All participants are required to complete the same pre-test and post-test. The detailed settings can be found in Appendix D and example survey items are shown in Appendix E.

**1) Pre-test** The pre-test consists of 22 questions across 4 constructs. Participants are asked to indicate the extent of their agreement with the items using a 5-point or 7-point Likert scale.

**2) Interaction with the tool** After finishing the pre-test, we ask the participants to use both CA tutoring systems. In total, 32 students use CAT (treatment group), whereas 34 students use ArgueTutor (control group). We use the following two topics for the writing task: “Students do not need to go to school anymore because they can learn from AI tutors” and “Social media should be banned for teenagers below 16 years old”.

**3) Post-test** We ask a total of 17 questions across 6 constructs, measured on a 7-point Likert scale. We also ask 3 qualitative questions about users’ overall experience with the systems.

### 4.2 Q1: Students’ Perception of CAT

We report the resulting scores of the self-evaluated survey items in the post-test and the qualitative user feedback.

**Results of the Post-test** We use a linear mixed-effects model (Seabold et al., 2010) to account for the hierarchical structure of the data when measuring the results. Specifically, observations were nested within students; some students participated in both experiments (the first and second week), while others participated in only one. The model includes the student as a random effect to account for within-subject correlations, and includes group, day, and class as fixed effects to control for potential confounding factors. This approach allows for a more accurate estimation of the treatment effect.

The results reveal significant treatment effects for the following constructs: Perceived Feedback

| Category                             | Instance                                                                                                                                                                                                                                                                       |
|--------------------------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| On the overall impacts of the system | <i>"It didn't hallucinate or always agree with me, as other LLM's currently do. It also analyzed my arguments in more depth and also drew common key messages from them. It was also helpful that it showed me which arguments were better/worse or caused contradictions"</i> |
| On the logical structure diagram     | <i>"I liked after arguing that new design and that pictures were used."</i>                                                                                                                                                                                                    |
| On the thinking practice perspective | <i>"That during the conversation points stand out that I didn't even think about at the beginning"</i>                                                                                                                                                                         |
| Improvements on the interface design | <i>"Button 'To the next question' below, as it was not clearly visible."</i>                                                                                                                                                                                                   |

Table 1: Representative examples of the user feedback.

Accuracy ( $\beta = 0.98$ ,  $SE = 0.27$ ,  $p < 0.001$ ), Perceived Usefulness ( $\beta = 0.75$ ,  $SE = 0.25$ ,  $p = 0.003$ ), and Intention to Use ( $\beta = 1.32$ ,  $SE = 0.27$ ,  $p < 0.001$ ), indicating medium to very large effects. These results suggest that students perceived CAT as providing more accurate feedback, being more useful, and eliciting stronger intention to use compared to ArgueTutor. This is likely attributable to CATs ability to deliver contextualized feedback grounded in the logical structure of students CAs, whereas ArgueTutor provides more generic, template-based feedback based on the number of claims and premises.

**Qualitative User Feedback** Overall, participants expressed a positive attitude toward CAT, noting that the system provided critical analysis of their CAs and that the logical structure diagram was helpful. However, they also offered suggestions for improving the interface design. Examples are shown in Table 1.

### 4.3 Q2: Impacts on Critical Thinking Skills

Improvements in critical thinking skills are difficult to measure objectively, and it is also challenging to observe absolute gains in critical thinking over a short period of time (Snyder and Wiles, 2015). Therefore, we discuss the statistical results and provide a qualitative analysis of the discussion chat history and the CA revisions to report the potential impact of CAT on students' critical thinking.

#### 4.3.1 Statistical Analysis

Table 2 shows the statistics for the number of turns and words in the user messages in chat history for both CAT and ArgueTutor. A significant difference is observed in the number of words in users' messages, suggesting that students may have more engaged discussions with the CAT system than

|                              | CAT      | ArgueTutor |
|------------------------------|----------|------------|
| mean(# turns)                | 3.5      | 3.8        |
| median(# turns)              | 3.0      | 4.0        |
| range(# turns)               | [0, 14]  | [0, 10]    |
| mean(# words in user msgs)   | 113      | 26         |
| median(# words in user msgs) | 87       | 24         |
| range(# words in user msgs)  | [0, 416] | [0, 74]    |

Table 2: Statistics of the discussion chat history, including the number of turns and words in user messages for both systems.

|         | Category                         | #instances |
|---------|----------------------------------|------------|
| VALID   | Add claim                        | 2/12       |
|         | Add attack on IA                 | 4/12       |
|         | Add premise                      | 4/12       |
|         | Paraphrase the same point        | 2/12       |
| INVALID | Fatigued by the 2nd discussion   | 8/21       |
|         | Fatigued by too many questions   | 1/21       |
|         | Defend one's own ideas           | 2/21       |
|         | Struggle with critical questions | 3/21       |
|         | Repeat one's own points          | 3/21       |
|         | Show revision intent only        | 3/21       |
|         | Have balanced stance             | 5/21       |

Table 3: Results of the qualitative analysis for CAT. We categorize revision instances into different categories. Note that a single invalid chat may be assigned to multiple categories; therefore, the categories are not mutually exclusive.

with the ArgueTutor system. This is potentially due to the fact that CAT facilitates deeper engagement with the logical structure of CAs, prompting students to articulate and elaborate their reasoning, whereas ArgueTutor provides feedback primarily based on the number of premises and claims. The differences in interaction design likely also account for the variation in average message length, as richer, dialogic exchanges tend to elicit more elaborated responses, while simpler, rule-based interactions require comparatively fewer words.

#### 4.3.2 Qualitative Analysis

We conducted a qualitative analysis of the chat history and CA revisions for both the CAT and ArgueTutor systems. We discuss the results below.

**CAT** We received a total of 33 chat histories. Of these, 12 resulted in valid CA revisions, whereas the remaining 21 either resulted in invalid revisions or did not have a corresponding revision. We analyzed the original CAs, their corresponding revisions, where applicable, and the associated chat histories to examine how these discussions influenced revision behavior. We show the summary of the results in Table 3. Among all valid revisions,

|                                                                                                                                                                                                                                                                                                                                                                                                                                 |
|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Original CA                                                                                                                                                                                                                                                                                                                                                                                                                     |
| While all of that is true i also think that social media can be important for teenagers to connect with their friends. They want to chat with them and post the cool things they did.....                                                                                                                                                                                                                                       |
| CA Revision                                                                                                                                                                                                                                                                                                                                                                                                                     |
| While all of that is true i also think that social media can be important for teenagers to connect with their friends. They want to chat with them and post the cool things they did.... <b>So thats why i think the solution to this problem is a gernerel regulation on how the algorithm works for everyone. Social media companys should not push harmful content anymore but instead for example informative content..</b> |

Table 4: An example of a CA and its revision. Green text represents the newly added portion during revision.

11 of 12 were motivated by the discussions. 4 of 12 revisions introduced a new point that directly addresses the IAs point, even though the original CAs merely presented new points supporting the opposing stance to the IA without actually challenging the IAs point. An example of a CA and its revision after interaction with CAT is shown in Table 4. The topic under discussion is whether social media should be banned for teenagers under 16. The student's position is that it should not be banned. Accordingly, CAT presents an IA with the opposing stance, arguing that "social media causes many mental health problems." The student's original CA argues that social media offers the benefit of helping teenagers connect with friends. However, this does not address the IAs main point (i.e., "social media causes many mental health problems"); instead, the student's CA merely introduces a new, parallel point. To help the student improve their CA, CAT analyzes the logic of the CA and raises the critical question of why the student believes that the point about connecting with friends is more important than the IAs concern about mental health problems. The multi-turn discussion ultimately leads the student to include a new argument in the revised CA that directly addresses mental health issues by proposing the regulation of social media algorithms. This suggests that CAT is capable of helping students practice critically evaluating opponent's arguments and weighing their own reasoning against them. However, we also observed that some students lost patience when the interaction became overly long and when the system questioned students views in a critical way (Table 3).

**ArgueTutor** Overall, we found that ArgueTutor provides straightforward feedback that directly and effectively guides students to add more claims or premises in their revisions (14/26 valid revisions). However, the system failed to offer concrete guid-

ance on how to develop these claims or premises in ways that engage with the opponent's arguments. As a result, students tend to engage in superficial discussions and incorporate into their revisions only claims or premises that do not directly address the IA (11/14 revisions).

## 5 Discussion

Our results have several implications for designing systems that aim to improve higher-order thinking skills using CAs: 1) Systems involving engagement with complex concepts should encourage users to move beyond their comfort zones and invest cognitive effort during the interaction to achieve optimal learning outcomes. However, a balance between the necessary cognitive effort and overwhelming information is necessary. For instance, as a potential future improvement, the CAT system could allow students to choose whether to participate in additional feedback discussion sessions after completing one, in order to reduce student overload. 2) When providing feedback on CAs, systems need not explicitly instruct users on what actions to take; rather, they should guide users through the reasoning process and help them determine the appropriate revisions on their own. However, without explicit instruction, users may find it difficult to determine which actions to take. Future systems should incorporate mechanisms to help balance users workload and better support them throughout their thinking process.

## 6 Conclusion

In this work, we developed and evaluated CAT, a dedicated CA tutoring system that aims to improve students' critical thinking skills by providing personalized feedback on the logical structures of their CAs. We compared CAT to a well-established system, ArgueTutor, in a classroom experiment. We found that CAT helps students practice critically evaluating opponents arguments and weighing their own reasoning against them, which are important aspects of critical thinking. Furthermore, our findings offer valuable insights for the design of future argumentation systems. Overall, we believe that CAT represents a significant advancement in the continued development of argumentation tutoring systems for higher-order critical thinking skills.

## Limitations

Given the difficulties of conducting a user study in a real classroom setting, our study was comparatively small and conducted over a short period of time, which may not be ideal for measuring significant changes in critical thinking skills. Nevertheless, we believe that the results of our study indicate the potential for substantial outcomes if the user study is scaled into a long-term experiment. Furthermore, in the current study, we implemented the backend model solely using the OpenAI API. However, the design of the CAT system also allows for the use of local small models. While we have not tested the feasibility of using such a local model in a real-world scenario due to GPU constraints, we believe this would be worthwhile to examine, as it might reduce the cost of relying on API calls in a large-scale experiment. Moreover, although our results show that CAT is effective, we did not determine the individual effect of each component. Future work could focus on conducting an ablation study to identify the individual effects of the logical structure diagram and critical question discussion.

## Ethical Considerations

All data collected in the user study were obtained with the participants' consent. We specifically prepared a consent form that explains the purpose of the study and explicitly states that the data will be stored and analyzed anonymously. The anonymized data will be kept strictly confidential and used only for scientific research, such as publications, presentations, and subsequent studies. Only those who explicitly agreed to the form participated in our user study. We also informed participants that their participation in this study was voluntary. They had the right to withdraw from the study at any time without explanation.

## Acknowledgments

This work was supported by JSPS KAKENHI Grant Number JP22H00524. We sincerely thank all the reviewers and meta-reviewers for the time and effort devoted to reviewing our work. The authors prepared the entire manuscript independently but used ChatGPT (OpenAI) for assistance with phrasing. The authors also used Codex for coding support. All generated suggestions were reviewed, modified as necessary, and verified by the authors, who assume full responsibility for the final manuscript and code.

## References

- Ritu Agarwal and Elena Karahanna. 2000. [Time flies when you're having fun: Cognitive absorption and beliefs about information technology usage](#). *MIS Quarterly*, 24(4):665–694.
- Paola Andreucci-Annunziata, Andrea Riedemann, Susana Cortés, Augusto Mellado, María Teresa del Río, and Alejandro Vega-Muñoz. 2023. [Conceptualizations and instructional strategies on critical thinking in higher education: A systematic review of systematic reviews](#). *Frontiers in Education*, Volume 8 - 2023.
- Susan J. Ashford. 1986. [Feedback-seeking in individual adaptation: A resource perspective](#). *The Academy of Management Journal*, 29(3):465–487.
- Paul Black and Dylan Wiliam. 2009. [Developing the theory of formative assessment](#). *Educational Assessment, Evaluation and Accountability (formerly: Journal of personnel evaluation in education)*, 21(1):5–31.
- Jérôme Brender, Aitor Perez, Patrick Jermann, Engin Bumbacher, Francesco Mondada, and Chenyang Wang. 2026a. [More correcting or confounding? when an ai arguing agent informed by common student misconceptions joins case-based peer-argumentation in a real course](#). In *Proceedings of the Impactful and Responsible AI Systems for Education Workshop*, volume 339 of *Proceedings of Machine Learning Research*, pages 80–89. PMLR.
- Jérôme Brender, Aitor Perez, Patrick Jermann, Francesco Mondada, Engin Walter Bumbacher, and Chenyang Wang. 2026b. [Arguing to learn at scale: A semester-long study of an ai arguing agent in a graduate robotics course](#).
- Miguel A. Cardona, Roberto J. Rodríguez, and Kristina Ishmael. 2023. [Artificial intelligence and the future of teaching and learning: Insights and recommendations](#). Technical report, U.S. Department of Education, Office of Educational Technology.
- Sidney D'mello and Art Graesser. 2013. [Autotutor and affective autotutor: Learning by talking with cognitively and emotionally intelligent computers that talk back](#). *ACM Trans. Interact. Intell. Syst.*, 2(4).
- Robert Ennis. 2011. [Critical thinking: Reflection and perspective part ii](#). *Inquiry: Critical thinking across the Disciplines*, 26(2):5–19.
- Peter Facione. 1990. [Critical thinking: A statement of expert consensus for purposes of educational assessment and instruction \(the delphi report\)](#).
- A. C. Graesser, P. Chipman, B. C. Haynes, and A. Olney. 2005. [Autotutor: an intelligent tutoring system with mixed-initiative dialogue](#). *IEEE Trans. on Educ.*, 48(4):612618.

- John Hattie and Helen Timperley. 2007. [The power of feedback](#). *Review of Educational Research*, 77(1):81–112.
- P.-S. Hsu, M. Van Dyke, Y. Chen, and T.J. Smith. 2015. [The effect of a graph-oriented computer-assisted project-based learning environment on argumentation skills](#). *Journal of Computer Assisted Learning*, 31(1):32–58.
- Kalypso Iordanou and Chrysi Rapanta. 2021. [argue with me: A method for developing argument skills](#). *Frontiers in Psychology*, Volume 12 - 2021.
- Wolfgang Iser. 1972. [The reading process: A phenomenological approach](#). *New Literary History*, 3:279.
- David W. Johnson and Roger T. Johnson. 1993. [Creative and critical thinking through academic controversy](#). *American Behavioral Scientist*, 37(1):40–53.
- Deanna Kuhn. 1991. *The Skills of Argument*. Cambridge University Press.
- Deanna Kuhn. 2019. [Critical thinking as discourse](#). *Human Development*, 62(3):146–164.
- John Lawrence and Chris Reed. 2019. [Argument mining: A survey](#). *Computational Linguistics*, 45(4):765–818.
- Fulan Liu and Paul Stapleton. 2014. [Counterargumentation and the cultivation of critical thinking in argumentative writing: Investigating washback from a high-stakes test](#). *System*, 45:117–128.
- Elif Meral, İbrahim Fevzi Şahin, and Yavuz Akbaş. 2021. [The effects of argumentation-based teaching approach on students critical thinking disposition and argumentation skills: population in our country unit](#). *International Journal of Psychology and Educational Studies*, 8(1):5174.
- Shoichi Naito, Wenzhi Wang, Paul Reisert, Naoya Inoue, Camélia Guerraoui, Kenshi Yamaguchi, Jungmin Choi, Irfan Robbani, Surawat Pothong, and Kentaro Inui. 2024. [Designing logic pattern templates for counter-argument logical structure analysis](#). In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 11313–11331, Miami, Florida, USA. Association for Computational Linguistics.
- E Michael Nussbaum. 2008. [Using argumentation vee diagrams \(avds\) for promoting argument-counterargument integration in reflective writing](#). *Journal of Educational Psychology*, 100(3):549.
- E Michael Nussbaum, Denise L Winsor, Yvette M Aquí, and Anne M Poliquin. 2007. [Putting the pieces together: Online argumentation vee diagrams enhance thinking during discussions](#). *International Journal of Computer-Supported Collaborative Learning*, 2(4):479–500.
- Michael Nussbaum. 2011. [Critical questions and argument stratagem: A framework for enhancing and analyzing students’ reasoning practices](#). *Journal of the Learning Sciences*, 20:433–488.
- OECD. 2024. [Critical thinking](#).
- Qiyu Pan, Jianqiao Zeng, Jie Wang, Junyu Liu, Yihan Qiu, Kangyu Yuan, and Zhenhui Peng. 2025. [AMQuestioner: Training critical thinking with question-driven interactive argument maps in on-line discussion](#). *Proc. ACM Hum.-Comput. Interact.*, 9(7).
- Niels Pinkwart, Kevin Ashley, Collin Lynch, and Vincent Aleven. 2009. [Evaluating an intelligent tutoring system for making legal arguments with hypotheticals](#). *Int. J. Artif. Intell. Ed.*, 19(4):401424.
- Priya Pitre and Kurt Luther. 2024. [ArguMentor: Augmenting user experiences with counter-perspectives](#). *Preprint*, arXiv:2406.02795.
- Philip M Podsakoff and Jiing-Lih Farh. 1989. [Effects of feedback sign and credibility on goal setting and task performance](#). *Organizational Behavior and Human Decision Processes*, 44(1):45–67.
- Abhijit Roy and Bart Macchiette. 2005. [Debating the issues: A tool for augmenting critical thinking skills of marketing students](#). *Journal of Marketing Education*, 27(3):264–276.
- Skipper Seabold, Josef Perktold, and 1 others. 2010. [Statsmodels: econometric and statistical modeling with python](#). *scipy*, 7(1):92–96.
- Julia J. Snyder and Jason R. Wiles. 2015. [Peer led team learning in introductory biology: Effects on peer leader critical thinking skills](#). *PLoS ONE*, 10.
- Edward M. Sosu. 2013. [The development and psychometric validation of a critical thinking disposition scale](#). *Thinking Skills and Creativity*, 9:107–119.
- Daniel D. Suthers. 2003. [Representational Guidance for Collaborative Inquiry](#), pages 27–46. Springer Netherlands, Dordrecht.
- Thitaree Tanprasert, Sidney S Fels, Luanne Sinnamon, and Dongwook Yoon. 2024. [Debate chatbots to facilitate critical thinking on youtube: Social identity and conversational style make a difference](#). In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, CHI ’24, New York, NY, USA. Association for Computing Machinery.
- Robert J Tierney and 1 others. 1983. [Author’s intentions and readers’ interpretations](#). technical report no. 276.
- UNESCO. 2019. [Global standards for media and information literacy curricula development guidelines](#).
- Viswanath Venkatesh, Michael G Morris, Gordon B Davis, and Fred D Davis. 2003. [User acceptance of information technology: Toward a unified view1](#). *MIS quarterly*, 27(3):425–478.

- Douglas Walton, Christopher Reed, and Fabrizio Macagno. 2008. *Argumentation Schemes*. Cambridge University Press.
- Thiemo Wambsganss, Tobias Kueng, Matthias Soellner, and Jan Marco Leimeister. 2021. [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. Association for Computing Machinery.
- Thiemo Wambsganss, Christina Niklaus, Matthias Cetto, Matthias Söllner, Siegfried Handschuh, and Jan Marco Leimeister. 2020. [AL: An adaptive learning support system for argumentation skills](#). CHI '20, page 114, New York, NY, USA. Association for Computing Machinery.
- Chenyang Wang, Filippo Forte, Léane Wettstein, Pierre Dillenbourg, and Thiemo Wambsganss. 2026. [Argue-Mate: Designing an arguing agent with maximised disagreement to support student peer-argumentation exercise](#). In *Proceedings of the Extended Abstracts of the 2026 CHI Conference on Human Factors in Computing Systems*, CHI EA '26, New York, NY, USA. Association for Computing Machinery.
- Wenzhi Wang, Paul Reisert, Shoichi Naito, Naoya Inoue, Machi Shimmei, Surawat Pothong, Jungmin Choi, and Kentaro Inui. 2025. [Identification of multiple logical interpretations in counter-arguments](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 6418–6433, Suzhou, China. Association for Computational Linguistics.
- Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, and 3 others. 2020. [Hugging-face's transformers: State-of-the-art natural language processing](#). *Preprint*, arXiv:1910.03771.
- Christopher R Wolfe, M Anne Britt, Melina Petrovic, Michael Albrecht, and Kristopher Kopp. 2009. [The efficacy of a web-based counterargument tutor](#). *Behavior Research Methods*, 41(3):691–698.
- World Economic Forum. 2025. [Future of jobs report 2025](#).
- Pieter Wouters, Christof Van Nimwegen, Herre Van Oostendorp, and Erik D Van Der Spek. 2013. [A meta-analysis of the cognitive and motivational effects of serious games](#). *Journal of educational psychology*, 105(2):249.
- Yu Wu, Patrick C Shih, and John M Carroll. 2014. [Design for supporting dialectical constructivist learning activities](#). In *EDULEARN14 proceedings*, pages 4156–4164. IATED.
- Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. [Bertscore: Evaluating text generation with BERT](#). In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.
- Zheng Zhang, Jie Gao, Ranjodh Singh Dhaliwal, and Toby Jia-Jun Li. 2023. [Visar: A human-ai argumentative writing assistant with visual programming and rapid draft prototyping](#). In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, UIST '23, New York, NY, USA. Association for Computing Machinery.

## A CALSA Logical Structures

The logical structure of CAs depends on the logic of IA. We show the text description of all ten logical structures in CALSA typology in Table 5 for the IA: *x should be abolished, because {x} suppresses a good {y}*. The details of the logical structures can be found in Naito et al. (2024).

## B Details of ArgueTutor

The ArgueTutor is a system designed to support students' persuasive writing. Based on a human-centered design approach, the ArgueTutor tool includes both a free-chat window and a few predefined shortcut buttons for common messages in the user interface to facilitate user interaction. However, since the tool was designed before the current LLM era, the interaction with the PCA is entirely template-based. If users choose to interact with the tool via free-text messages, the intent of those messages will be classified first, and then a template-based response will be returned. The original ArgueTutor tool cannot be used directly in our CA context, as the backend NLP model for classifying claims and premises is trained on a German dataset and not specifically on CAs. Therefore, we re-implement the backend model with a more advanced GPT-5.2 model via the OpenAI API. We keep all other main functionalities the same as the original tool. Users can interact with the tool either through free-text messages or via three predefined shortcut buttons corresponding to the main functionalities described above. The intent of free-text user messages is also determined by GPT-5.2 and mapped to one of the three main functionalities. To control the differences and similarities in the design between CAT and ArgueTutor, we keep the task the same for both tools. Specifically, for ArgueTutor, the general task is also to write a CA based on an IA and a topic, interact with the PCA to improve the CA, and then revise it. The functionalities F1, F2, and F6 in the interfaces shown in Figure 1 are the same. Rather than first receiving the structures (F3, F4) and subsequently discussing the critical questions (F5), users begin by conversing directly with a PCA, which provides feedback based on premises and claims.

## C Pilot Study

Before the classroom experiment, we conducted a pilot study with several PhD students to evaluate the interface design and gathered their perceptions

through brief interviews. Overall, the feedback was positive, particularly regarding the logical structure diagrams and the chat interface. Participants found the diagrams easy to understand and perceived the Critical Question Discussion function as useful, as it helped them identify areas for improving their CAs. Nevertheless, participants also provided feedback on the interface design, such as the color scheme, button placement, and related elements. Several participants also noticed issues with the logical structure diagrams at the beginning, as some visual elements were not drawn in the correct place for certain CAs. We refined the interface based on feedback from the pilot study and used the final version for the classroom experiment.

## D User Study Settings

The experiment was conducted over four sessions spanning two weeks, with each session held within a German-speaking undergraduate course in which students use English as a second language. In the first week, two sessions were conducted. 18 students in class A were assigned to the treatment group and used CAT, while 20 students in class B were assigned to the control group and used ArgueTutor. In the second week, the remaining two sessions were conducted. The groups were swapped. 14/18 students who participated on the first day in class A also participated on the second day. They were assigned to the control group and used the ArgueTutor system. 14/20 students who participated on the first day in class B also participated on the second day. They were assigned to the treatment group and used the CAT. Participants in the treatment group had an average age of 24.75 (SD = 3.75). 7 were female, and 23 were male; 2 were self-identified. In the control group, participants' average age was 24.81 (SD = 3.82). 7 were female, 25 male. Two did not complete the survey.

## E Pre/Post-tests Details

The constructs and the example items are shown in Table 6.

| Structure Name                                        | Description                                                                                                                                                     |
|-------------------------------------------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Mitigation                                            | While it is not denied that {x} suppresses a positive outcome {y}, the causal relationship can be mitigated through the means of z.                             |
| Alternative                                           | While it is not denied that {x} suppresses a positive outcome {y}, another factor {z} also suppresses {y}, and it should be prioritized to reduce {z} over {x}. |
| No evidence                                           | There is no evidence that {x} suppresses {y}.                                                                                                                   |
| Another true cause                                    | The true cause of suppressing {y} is {z}, not {x}; therefore, abolishing {x} will not solve the problem.                                                        |
| Missing mechanism #1                                  | {x} promotes another factor {z}, and {z} promotes {y}. Therefore, {x} does not suppress {y} but rather promotes {y}.                                            |
| Missing mechanism #2                                  | {x} suppresses another factor {z}, and {z} suppresses {y}. Therefore, {x} does not suppress {y} but rather promotes {y}.                                        |
| No need to address                                    | {y} is sufficient as it is, and there is no need to take action for {y}.                                                                                        |
| Negative effect due to y                              | Since {y} leads to a negative outcome {z}, {y} is not a positive outcome.                                                                                       |
| Positive effects of a different perspective from y #1 | Since {x} promotes a positive outcome {z}, which is a different perspective from {y}, {x} should not be abolished.                                              |
| Positive effects of a different perspective from y #2 | Since {x} suppresses a negative outcome {z}, which is a different perspective from {y}, {x} should not be abolished.                                            |

Table 5: All ten logical structure names and their corresponding textual descriptions from the CALSA typology (Naito et al., 2024).

| Survey | Scale | Construct                                                  | Example Item                                                                                     |
|--------|-------|------------------------------------------------------------|--------------------------------------------------------------------------------------------------|
| Pre    | 7     | Attitudes towards Technology (Agarwal and Karahanna, 2000) | I like to experiment with new information technologies                                           |
|        | 5     | Wish for Feedback (Ashford, 1986)                          | It is important for me to receive feedback on my performance                                     |
|        | 7     | Argumentation Skills                                       | How do you rate your ability to argue?                                                           |
|        | 5     | Critical Thinking Dispositions (Sosu, 2013)                | I usually try to think about the bigger picture during a discussion                              |
| Post   | 7     | Perceived Feedback Accuracy (Podsakoff and Farh, 1989)     | The feedback I received reflected my true performance.                                           |
|        | 7     | Intention to Use (Venkatesh et al., 2003)                  | Imagine the argumentation tool would be available in your next course, I would use it.           |
|        | 7     | Perceived Usefulness (Venkatesh et al., 2003)              | Using the argumentation tool helps me complete tasks more quickly.                               |
|        | 7     | Perceived Ease of Use (Venkatesh et al., 2003)             | Learning to operate the system would be easy for me.                                             |
|        | 7     | Argumentation Skills                                       | How do you now rate your ability to argue?                                                       |
|        | 7     | Argumentation Skills Improvement                           | After completing the task, my ability to argue in a structured way in written texts has improved |

Table 6: Constructs used in the pre- and post-tests, along with their examples. For all scales, 1 represents "strongly disagree" and the highest score represents "strongly agree" when the items are presented as statements; 1 represents "extremely poor" and the highest score represents "excellent" when the items are presented as questions.