Teach Me How to Argue:
A Survey on NLP Feedback Systems in Argumentation

Camélia Guerraoui[1,2,3] **Paul Reisert**[4] **Naoya Inoue**[5,2] **Farjana Sultana Mim**[6]
Keshav Singh[7] **Jungmin Choi**[1,2] **Irfan Robbani**[5]
Shoichi Naito[1,2,8] **Wenzhi Wang**[1,2] **Kentaro Inui**[9,1,2]
[1]Tohoku University [2]RIKEN [3]INSA Lyon [4]Beyond Reason
[5]JAIST [6]Tufts University [7]CTW Inc. [8]Ricoh Company, Ltd. [9]MBZUAI
{guerraoui.camelia.kenza.q4, naito.shoichi.t1, wang.wenzhi.r7}@dc.tohoku.ac.jp, beyond.reason.sp@gmail.com, naoya-i@jaist.ac.jp
farjana.mim@tufts.edu, keshav.singh29@gmail.com, jungmin.choi@riken.jp, robbaniirfan@jaist.ac.jp, kentaro.inui@tohoku.ac.jp

Abstract

The use of argumentation in education has shown improvement in students' critical thinking skills, and computational models for argumentation have been developed to further assist this process. Although these models are useful for evaluating the quality of an argument, they often cannot explain why a particular argument score was predicted, i.e., why the argument is good or bad, which makes it difficult to provide constructive feedback to users, e.g., students, so that they can strengthen their critical thinking skills. In this survey, we explore current NLP feedback systems by categorizing each into four important dimensions of feedback (Richness, Visualization, Interactivity and Personalization). We discuss limitations for each dimension and provide suggestions to enhance the power of feedback and explanations to ultimately improve user critical thinking skills.

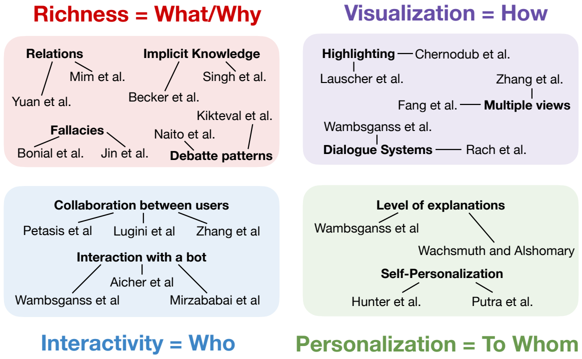


Figure 1: Overview of some NLP feedback systems categorized into our feedback dimensions.

1 Introduction

Argumentation is the field of elaborating and presenting arguments to engage in debate, convince others, and eventually reach agreements. In this context, *an argument* is made of a conclusion (i.e., a claim) supported by reasons (i.e., premises) (Toulmin, 1958). *Computational argumentation* emerged as a way to support argumentation. It is a subfield of natural language processing (NLP) that deals with the automated representation, evaluation, and generation of arguments. It includes tasks such as mining arguments (Al-Khatib et al., 2016), assessing arguments' quality (El Baff et al., 2018), reconstructing implicit assumptions in arguments (Habernal et al., 2018) or even providing constructive feedback for improving arguments (Naito et al., 2022), to name a few.

In education, learning how to argue (e.g., writing argumentative essays, debates, etc.) has been shown to improve students' critical thinking skills (Pithers and Soden, 2000; Behar-Horenstein and Niu, 2011). To further improve critical thinking skills, several researchers have been working on computational argumentation and specifically argumentative feedback systems to provide support and to assist learners in improving the quality of their arguments (Habernal et al., 2017; Wachsmuth et al., 2017; Lauscher et al., 2022).

Although argumentative feedback systems are proven to assist students' learning and reduce teachers' workload (Twardy, 2004; Wambsganß et al., 2021), such systems still lack the ability to *deeply explain* how an argument can be improved; i.e., not only providing a holistic label or score, but explaining particularly *why* this result was given by automatic evaluation rubrics. Such explanations as feedback can ultimately *explain and visualize the results comprehensively* for the users so that users can understand and improve their argumentation skills. The lack of ability in current systems to provide deep explanations as feedback motivated our interest in investigating the current state of argumentative feedback generation.

In this survey, we focus on different kinds of feedback given to learn how to argue. Inspired by the sections *Tutorial Feedback* and *Architecture and Technology* mentioned in Scheuer et al. (2010), we combine features of feedback systems,

Proceedings of the 10th Workshop on Argument Mining, pages 19–34
December 7, 2023. ©2023 Association for Computational Linguistics

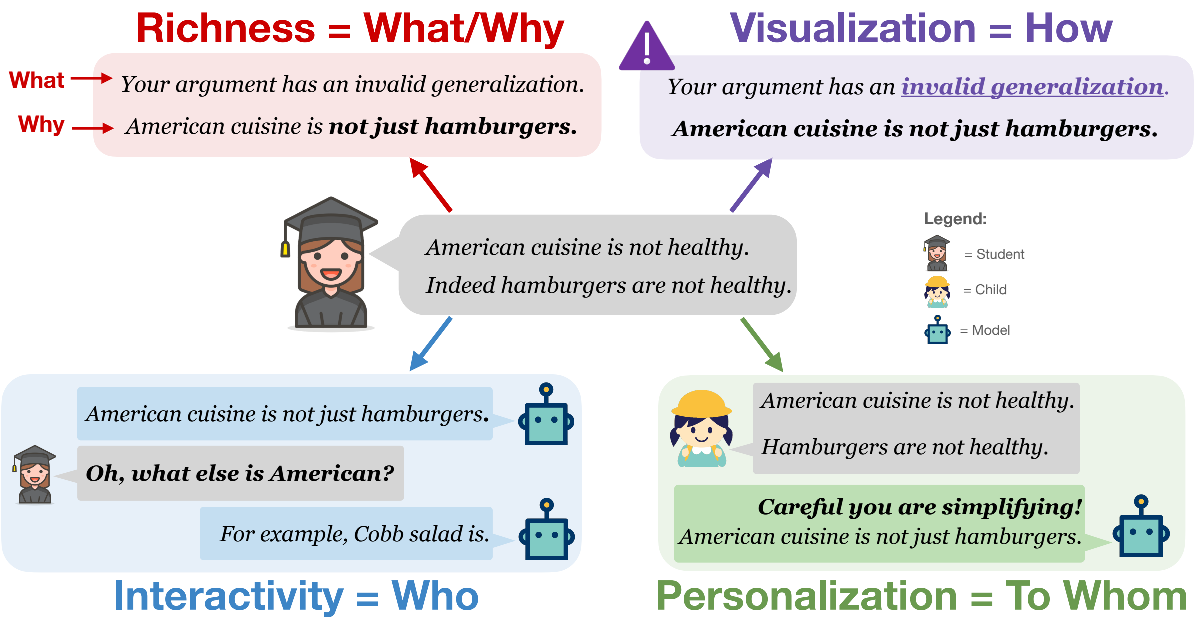

Richness = What/Why
What
Why
Your argument has an invalid generalization.
American cuisine is not just hamburgers.
Visualization = How
Your argument has an invalid generalization.
American cuisine is not just hamburgers.
American cuisine is not healthy.
Indeed hamburgers are not healthy.
Legend:
= Student
= Child
= Model
American cuisine is not just hamburgers.
Oh, what else is American?
For example, Cobb salad is.
American cuisine is not healthy.
Hamburgers are not healthy.
Careful you are simplifying!
American cuisine is not just hamburgers.
Interactivity = Who
Personalization = To Whom

Figure 2: Example of four feedback for each dimension (*Richness*, *Visualization*, *Interactivity* and *Personalization*).

formulate four distinct dimensions and categorize existing papers into these dimensions (Figure 1):

- *Richness*: Level of feedback details given by a model, i.e., *what* is the error identified by the model and *why* it is an error.
- *Visualization*: Model's ability to present feedback, i.e., *how* the feedback is shown to the end user.
- *Interactivity*: Model's ability to allow the user to communicate with other users or the model itself, i.e., with *whom* the user is talking.
- *Personalization*: Model's ability to adapt the feedback to the users' background, i.e., *to whom* the feedback is given.

Figure 2 shows four different dimensions of feedback (*Richness*, *Visualization*, *Interactivity*, and *Personalization*), for a given argument consisting of two claims and one premise. In this example, in the *Richness* dimension, a faulty generalization in the argument is identified (cf. What) and explained (cf. Why). *Visualization* would add symbols and highlight important feedback elements to make it more understandable. *Interactivity* would allow the user to ask for more explanations to the model. *Personalization* would consider that the user is a child and provide appropriate feedback on that basis.

Towards better argumentative feedback, this survey aims to give an overview of argumentative feedback systems. We explore work that provide feedback answering one or multiple questions among the types: *What/Why* (§4), *How* (§5), *Who* (§6), and *To Whom* (§7). Finally, we discuss remaining challenges and potential ways to overcome them (§8) in order to develop systems that provide feedback or detailed explanations in a way so that learners can improve their critical thinking skills. We believe this survey can aid researchers in understanding current explanations in argumentation and broaden their horizon on argumentative feedback.[1]

2 Related Work

Several surveys have been done in the field of argumentation (Ke and Ng, 2019; Habernal and Gurevych, 2016; Lawrence and Reed, 2020; Wang et al., 2022) and explainability (Danilevsky et al., 2020; Islam et al., 2021; Hartmann and Sonntag, 2022). As we would like to focus on how well a model can explain its results as a type of feedback for learners, we present here recent surveys related to feedback or explainability in argumentation.

Beigman Klebanov and Madnani (2020) present the progress in automated writing evaluation, using Page (1966) to frame the presentation. In this survey, the succinct feedback section enumerates different systems for writing assistant and highlights the inconclusiveness of research on effectiveness

[1] For more details, papers mentioned in this survey are categorized at https://kmilia.github.io/teach_me_how_to_argue/.

of automated writing evaluation.

Vassiliades et al. (2021) highlights the potential of argumentation in explainable AI systems. They provide an exhaustive overview of argumentation systems by grouping them based on domain, such as law. For each domain, papers are compared by tasks (e.g., argument classification). Despite the extensiveness of their survey, some topics to improve explanations in argumentative systems received little attention. For example, frameworks that include arguments with commonsense knowledge and diverse attack relations between them have rarely been discussed, even though they can enhance the model's explainability (Saha et al., 2021).

Čyras et al. (2021) focus on the different frameworks, types, and forms of explanations. They distinguish intrinsic approaches (i.e., models using argumentative methods) from post-hoc approaches (i.e., non-argumentative models that provide complete or partial explanations). They discuss multiple forms of argumentation, such as dialogue. Their final roadmap covers the need to focus more on properties and computational aspects of argumentation-based explanations. Whereas they focus on how argumentation can be used to enhance the explainability of models, our work discusses what kind of feedback (i.e., explanations) on argumentation models can provide.

Moreover, our work distinguishes itself from the surveys previously mentioned by giving an overview of automatized feedback on argumentation from the angle of *rich* (§4), *visual* (§5), *interactive* (§6), and *personalized* (§7) explanations inspired by Scheuer et al. (2010).

3 Pedagogy

Before discussing the four dimensions mentioned a priori, it is essential to know the pedagogy used to teach argumentation and adopted by computational models. This section presents some standard pedagogical methods used in teaching how to argue.

Toulmin model The Toulmin model (Toulmin, 1958), often seen as the foundation of teaching argumentation, is a popular framework for constructing, analyzing and evaluating arguments, and can contribute to the improvement of students' argumentative writing (Rex et al., 2010; Yeh, 1998) as well as critical thinking skills (Giri and Paily, 2020). This approach deconstructs an argument into six elements (Appendix, Figure 4), and students are taught to identify each element within an argument. By identifying elements from the Toulmin model, models can provide users with *rich* feedback.

Rhetorical structure theory Based on Mann and Thompson (1988), the rhetorical structure theory was originally developed in the context of computer-based text generation in order to attribute a formal structure to a text (Hou et al., 2020). This theory employs graphical representations, such as mind maps or graphs, to illustrate the relationships between different components of the text's architecture. This visual approach can help students visualize the connections between different concepts and enhance their understanding of complex topics (Matsumura and Sakamoto, 2021). The advent of tools like Tiara (Putra et al., 2020) has given rise to the deployment of the rhetorical structure theory, i.e. the generation of *visual* feedback.

Collaborative argumentation In collaborative argumentation-based learning, also described as CABLE by Baker et al. (2019), individuals work together to construct, refine, and evaluate arguments on a particular topic or issue. The main goal of collaborative argumentation is to foster constructive dialogue, critical thinking, and the exploration of different perspectives. Weinberger and Fischer (2006) differentiate four dimensions of CABLE:

- *Participation*: Do learners participate at all? Do they participate on an equal basis?
- *Epistemic*: Are learners engaging in activities to solve the task (on-task discourse) or rather concerned with off-task aspect?
- *Argumentative*: Are learners following the structural composition of arguments and their sequences?
- *Social*: To what extent do learners refer to the contributions of their learning partners? Are they gaining knowledge by asking questions?

Veerman et al. (2002); Baker et al. (2019) show CABLE's positive effects on students' argumentation development. Nevertheless, they also highlight the challenges of this method, as not every dialogue can be predicted. By using CABLE, models can generate *interactive* feedback.

Socratic questioning The Socratic questioning is a common teaching strategy, described in Schauer (2012); Abrams (2015). With this method, the student is guided through reflexive questions towards solving a problem on their own, instead of receiving directly a solution. The user receives feedback which is tailored to their background, i.e., *personalized* feedback.

Recently, this method has been integrated into large language models (LLMs) to more effectively adhere to user-provided queries (Ang et al., 2023; Pagnoni et al., 2023), to enhance the ability of such models in generating sequential questions (Shridhar et al., 2022), and also to enhance the explainability of these models (Al-Hossami et al., 2023).

Nevertheless, the Socratic questioning is now raising debates among researchers focusing on pedagogy in argumentation. Indeed, Kerr (1999) and Christie (2010) pointed out its inefficiency and abusiveness as students are forced to give imperfect answers in a hurry and endure criticism.

4 Richness - What is an Error and Why?

To improve students' critical thinking skills, we first need to evaluate their argumentative texts, i.e., identify argumentative errors. In this section, we focus on models providing shallow explanations, i.e., models that identify *what* should be corrected in the arguments. We discuss relevant works that identify properties such as the structure of arguments which is helpful in this process.

Components Identifying argumentative components is one of the fundamental tasks in argumentation (Teufel, 1999; Stab and Gurevych, 2014; Jo et al., 2020). Such works primarily focus on identifying components such as *claims* and *premises*. More recently, the usefulness of identifying such components can be seen in tasks such as counter-argument generation. For example, in Alshomary et al. (2021), weak premises are identified and ranked to generate counter-arguments.

Relations After identifying the different components of an argumentative text, it is necessary to distinguish the multiple relations between them, ultimately to assert the arguments' quality. Indeed, supporting or refuting a claim is made of complex logical moves, such as promoting, contradicting, or acknowledging a fact. To identify the different relations patterns, Yuan et al. (2021) focus on finding interactive argument pairs, whereas Mim et al. (2022) enables annotating complex attack relations.

Schemes In addition to components and relations, Walton et al. (2008) proposed a set of roughly 80 logical argumentation schemes to categorize the underlying logic. Each scheme has a set of critical questions which provide a template to assess the strength of the argument depending upon the associated scheme. Since the first work on automatically detecting argumentation schemes in argumentative texts (Feng and Hirst, 2011), the use of such schemes has been explored in tasks such as essay scoring (Song et al., 2014).

Fallacies Although a good structure with a claim and premises is necessary for a good argument, it is not sufficient. An argument has more complex properties, such as its logical, dialectical, and rhetorical aspects. A fallacy is a logical error or deceptive argument that undermines the validity of a conclusion or reasoning, which poses a substantial issue due to its propensity to generate miscommunication. Towards teaching students to avoid making errors in logical reasoning, logical fallacies have received attention (Habernal et al., 2017; Bonial et al., 2022; Zhivar et al., 2023; Nakpih and Santini, 2020). Motivated by the gamification method made by Habernal et al. (2017), Bonial et al. (2022) aimed to capture similar fallacy types for news articles, but the low distribution of fallacy types in the wild makes identification challenging. However, most natural texts do not have recurrent specific patterns, compared to current datasets, like the Logic and LogicClimate datasets (Jin et al., 2022). Moreover, given the large number of logical fallacies that exist (over 100 types), long arguments can be grouped into multiple fallacies, resulting in difficulties in classification (Goffredo et al., 2022).

Debate patterns In a case of a debate, an opponent is willing to give a counter-argument synchronously and interactively. Analyzing and evaluating a debate is a difficult task as we need to retrieve not only the argumentation structure of each opponent but also the relations between them. Bao et al. (2022) focuses on argument pair extraction (APE), which consists of finding two interactive arguments from two argumentative passages of a discussion. Although the APE task gives insights into relations between different argumentative texts, it does not indicate complex relations (i.e., how claims, supports, attacks and the intention of the speakers are interrelated). To palliate this issue, Hautli-Janisz et al. (2022) identified and analyzed the dialogical argumentative structure of debates using Inference Anchoring Theory (IAT) (Budsziyska et al., 2014). Following the same IAT theory, Kikteva et al. (2022) showed that the type of questions (e.g., pure, assertive, and rhetorical questions) leads to different argumentative discourse. Focused more on the opponent's side, Naito et al. (2022) propose diagnostic comments for assessing the quality of counter-arguments by providing expressive, informative and unique templates. The feedback is then written by template selection and slot filling.

In-Depth Explanations Although identifying such argumentative structures (components, relations, and schemes) and properties (fallacies and debate patterns) is important, it has limitations in terms of effective feedback. Identifying a missing claim or a wrong premise is insufficient to understand how to improve the argumentation properly. Thus, we relate the identification of structure and properties to shallow explanations in the sense that users can still benefit from the output of the models.

Shallow explanations can be difficult to understand, especially for beginners, as they tend to be minimalist and lack guidance. To explain more effectively the errors in an argument, a model should go a step further, hence by providing *in-depth* explanations, which attempt to identify the argument's implicit components to explain *why* it is an error in a particular argument. In Figure 2, we implicitly know that hamburgers belong to the American cuisine, as same as the Cobb salad, a healthy garden salad from California. Therefore, if the model is able to reason out this implicit knowledge, it can better explain the invalid generalization in Figure 2.

Implicit Knowledge and Reasoning in Arguments To provide *in-depth* explanations, we need to know how to refine the argument, i.e., how to identify implicit information. Recently, many works have focused their attention on this aim. The main goal of such studies is to make the structure and reasoning of arguments explicit to explain the arguments for humans better. Additionally, this focus can eventually help build robust argumentation machines that can be enriched with language understanding capacity. Following the pioneer works of Razuvayevskaya and Teufel (2017), the ExpLAIN project (Becker et al., 2021) and Jo et al. (2021) are one such example that focuses extensively on reconstructing implicit knowledge in arguments by relying on knowledge graphs among others. Taking a step further in this direction, Heinisch et al. (2022) and Saadat-Yazdi et al. (2023) proposed to utilize such implicit information to bridge the implicit reasoning gap in arguments to help students explain their arguments better.

Large annotated corpora are required to improve implicit reasoning detection for models. To address this need, various studies have proposed methods for annotating implicit knowledge, leading to the development of multiple datasets (Becker et al., 2020; Singh et al., 2021, 2022). In Singh et al. (2021), semi-structured warrants, i.e. links between a claim and evidence (c.f. Appendix Figure 4), were annotated via crowdsourcing, whereas Becker et al. (2020) focus on reconstructing omitted information, semantic clause types, and commonsense knowledge relations through expert annotation. Corpora can be dedicated to a specific domain or sentence patterns. For example, (Singh et al., 2022) focused on domain-specific knowledge using six topics. However, implicit knowledge may take various forms, such as warrants, causal relations, facts, beliefs, or assumed-known arguments. Thus, revealing implicit knowledge in an unknown text through annotated datasets can be challenging.

In recent years, LLMs have made significant progress in exhibiting reasoning abilities. A comprehensive overview of the current state of reasoning abilities in LLMs is provided in the survey Huang and Chang (2023). The increasing interest in LLMs and implicit reasoning prompted the first ever workshop on natural language reasoning and structured explanations in 2023 (Dalvi Mishra et al., 2023). This workshop discussed that while LLMs have demonstrated good capabilities to find implicit components within an argument, they often cannot correctly explain the logical reasons behind their responses. To bridge this gap, a novel category of explanation techniques has arisen, playing a vital role in shaping the logical reasoning of models. One such example is the chain-of-thought prompting (Wei et al., 2022; Wang et al., 2023a), which employs explanations as a means for LLMs to emulate human reasoning procedures. While the references Huang and Chang (2023) and Dalvi Mishra et al. (2023) do not primarily focus on argumentative tasks, they can be a valuable source of inspiration in argumentation.

5 Visualization - How to Show the Error?

The effectiveness of any argument does not solely rely on its content but also on its presentation. This is where visualization of argumentative feedback emerges as a crucial factor. Visualizing feedback empowers individuals to perceive the intricacies of an argument in a more comprehensive manner. By using visual aids like graphs, feedback becomes more accessible and engaging, fostering constructive discussions. In this section, we discuss how visualization impacts argumentative feedback.

Highlights A simple approach to visualization is highlighting, i.e., application of visual emphasis on a specific pattern with the intention of drawing the viewer's attention to this specific pattern. For example, Lauscher et al. (2018) identify the argument component (Claim, background, data) and visualizes them by highlighting the text in different colors. Similarly, Chernodub et al. (2019) allow the user to choose the model to use and the components to highlight. Wambsganss et al. (2022b) take a step further by highlighting and presenting scores that give a quick overview of users' skills.

Highlighting serves as an essential key step in the cognitive input process, enabling viewers to quickly identify crucial argumentative structures. However, its use should be complemented with other visualization techniques to ensure a more profound exploration and comprehension of complex explanations. Studies conducted by Lauscher et al. (2018); Chernodub et al. (2019); Wambsganss et al. (2022b) shed light on the potentials and limitations of highlighting, paving the way for future advancements in data visualization methodologies.

Multiple views To overcome the shallowness of highlighting, several researchers add to their system other views, such as diagrams showing the argumentative structure. For example, to compare two drafts of an essay, Zhang et al. (2016); Afrin et al. (2021) use a revision map made of color-coded tiles, whereas Putra et al. (2021) rely on a tree to reorder arguments.

Based on the work of Wambsganß et al. (2020), Xia et al. (2022) and Wambsganss et al. (2022a) use a text editor which highlights components, a graph view which shows the argumentative structure, and a score view showing the user's performance. Based on the classroom-setting evaluation, students using such systems wrote texts with a better formal quality of argumentation compared to the ones using the traditional approach.

Nevertheless, the current accuracy of such systems' feedback still leaves a large improvement space in order for users to be motivated to use them. More recent work such as Zhang et al. (2023) incorporate feedback generated by state-of-the-art LLMs in their graphical systems. Nonetheless, factual inaccuracies, as well as inconsistent or contradictory statements, are still generated, exposing the user to confusion and leaving room for improvement.

Dialogue Systems In the realm of visualization, a novel approach gaining attraction is the integration of dialogue systems to enhance the interaction between users and visual representations. Dialogue systems, commonly known as chatbots like ChatGPT, have been increasingly explored for their potential to facilitate information comprehension (Rach et al., 2020; Wambsganß et al., 2021).

This kind of representation is challenging in terms of user-friendliness. Particularly, in a pedagogical context, users may have difficulties visualizing their previous feedback and progress. Indeed, users may be lost in the discussion flow and struggle to keep track of the ongoing discussions, lessons, or feedback because the representation does not provide clear signposts or structure. Students may forget a specific lesson and want to verify some information, or they simply need to reread their lessons and exercises. However, finding specific information in a chat discussion may take much effort. Thus, it is important (i) to have a chat session per lesson, exercise or test and (ii) to keep structured notes of the issues users face and how these issues can be solved. Eventually, a personal dashboard showing a user's progress through time could be beneficial not only for students but also for teachers. Indeed, with a dashboard, teachers can see if a specific student needs more attention. Moreover, teachers sometimes need to compare students among them, specifically during a test. Therefore, we believe that to improve the user-friendliness of pedagogical dialogue systems, other visual elements should be used.

Despite the growing popularity of both graphs and chatbots in data visualization, limited work has directly compared their effectiveness in improving critical thinking skills. Further research is needed to provide more nuanced insights on the comparison on one hand between both approaches and on the other between works among the same approach.

The importance of visualization lies in its ability to enhance the understanding of complex ideas. In this section, we highlighted the potential of the visualization of argumentative feedback and how it can improve students' learning process.

6 Interactivity - Who Talks to the User?

Teaching how to argue is a multifaceted task that demands more than the dissemination of theoretical knowledge; it requires fostering interactive learning environments that facilitate active engagement and practice. The traditional approach to teaching argumentation often centers on lecturing and one-way communication, where instructors impart information to students. While didactic methods have their place in education, a more interactive pedagogical approach, one that encourages learners to actively participate, can be used. In this section, we will see in which ways current argumentative computational models enable a form of interaction.

Interaction between different users NLP systems mostly allow communication between a user and a conversational agent. Nonetheless, some works chose to apply the CABLE pedagogy (§3) by allowing a user to dialog with *other users*. Following the footsteps of Petasis (2014), Lugini et al. (2020) track real-time class discussions and help teachers annotate and analyze them. Recent works such as Zhang et al. (2023) plan to add a collaborative setting in their future work.

The collaboration between multiple users within NLP systems is promising. Nevertheless, only a few works focus on the CABLE pedagogy. It is essential to acknowledge that some challenges and barriers have hindered its use in NLP, possibly due to the difficulty of designing and evaluating such tools, as human resources in a real-class setting (e.g., students, teachers) are required.

Interaction with a conversational agent As seen in §5, several research papers have showcased the feasibility of employing current conversational agents for educational purposes (Lee et al., 2022; Macina et al., 2023; Wang et al., 2023b). Often based on state-of-the-art language models, these agents have shown great capabilities in understanding and generating human-like responses. They can engage in dynamic and contextually relevant conversations, making them potentially valuable tools for educational purposes.

The use of conversational agents as dialog tutors has been explored outside of argumentation (Wambsganß et al., 2021; Mirzababaei and Pammer-Schindler, 2022; Aicher et al., 2022). For instance, in Mirzababaei and Pammer-Schindler (2022), an agent examines arguments to determine a claim, a warrant, and evidence, identifies any missing elements, and then assists in completing the argument accordingly. Wambsganß et al. (2021) create an interactive educational system that uses interactive dialogues to teach students about the argumentative structure of a text. The system not only provides feedback on the user's texts but also learning sessions with different exercises.

Research on chatbots in education is still preliminary due to the limited number of studies exploring the application of effective learning strategies using chatbots. This indicates a significant opportunity for further research to facilitate innovative teaching methods using conversational agents (Hwang and Chang, 2021). However, extraction and classification of useful data remain challenging, as the data collected are noisy and much effort still has to be made to make it trainable (Lin et al., 2023). Researchers must also continue to account for ethical considerations, including biased representations and data privacy safeguards, to ensure that their chatbots positively impact users (Kooli, 2023).

Overall, integrating interaction in teaching how to argue is not merely a pedagogical choice but an essential requirement to cultivate adept arguers who can navigate the intricacies of argumentation. Therefore, we encourage researchers to consider this dimension in their future pedagogical systems.

7 Personalization - To Whom is it For?

Even if the feedback mentioned in §4 are a step towards good guidance, they are static, which can be problematic. Beginners and professionals in argumentation do not need the same amount of feedback. A child and an adult have different levels of understanding and knowledge. Therefore, it is essential that a model knows *to whom* it should explain the errors and hence how to adapt its output by providing *personalized* explanations.

Levels of explanations A first approach to personalization is to discretize different users' proficiency levels in argumentation into a small number of categories. For instance, with the system described in Wambsganß et al. (2020) and Wambsganss et al. (2022a), users can select their own level among the following categories: Novice, Advanced, Competent, Proficient, Expert.

Although Wambsganß et al. (2020) and Wambsganss et al. (2022a) propose different granularity levels of explanations, their study is restrained to students from their university. Having end-users from different backgrounds may imply the need for new levels of explanations. Wachsmuth and Alshomary (2022) show that the explainee's age affects the way an explainer explains the topic at hand. Thus, we consider that information such as the learner's age should be considered in future interactive argumentative feedback systems, where terminology such as *fallacy* and their existence would require different explanation approaches for younger students (i.e., elementary) compared to older students.

Self-personalization For more personalized feedback, systems such as Hunter et al. (2019) and Putra et al. (2020) rely on user's inputs. They allow users to make their custom tags or to choose their preferences among a set of rubrics. Nevertheless, manually personalizing the system can be overwhelming and time-consuming for users.

Next directions Hunter et al. (2019) argue that the next direction for personalized argumentative feedback would be to develop argumentation chatbots for persuasion and infer the user's stance based on the discussion. Chatbots' personalization capabilities enable them to tailor their responses to individual learners' needs and learning styles, potentially enhancing the effectiveness of the tutoring process (Lin et al., 2023). However, bridging the gap among personalized chatbots (Qian et al., 2021; Ma et al., 2021), personalized educational methods (González-González et al., 2023; Ismail et al., 2023; Liu et al., 2020) and argumentation has remained unexplored. Thus, we think researchers should focus in the future on providing more *personalized* explanations (i.e., precisely adjusted by considering the learner's background) to improve the users' critical thinking skills efficiently.

8 Discussions

Teaching how to argue through NLP systems holds significant promise for enhancing students' learning process. However, existing research in this area presents various open issues. In this section, we explore some difficulties in designing and evaluating computational models for argumentation and discuss some methods for mitigating them.

Evaluating different systems The evaluation of NLP systems often relies on human assessment, which is insightful. However, this reliance makes it hard to reproduce the evaluation and to compare different systems. To the best of our knowledge, no research has focused on comprehensive comparative studies of different end-to-end systems. The lack of direct comparisons between similar systems hampers the understanding of their relative advantages and limitations. As researchers and educators, it becomes overwhelming to discern which system best fits specific pedagogical objectives. A possible reason for this issue resides in the restricted access to various tools. Indeed, many systems may not be accessible, limiting researchers to test them. Additionally, the lack of guidelines to evaluate systems for learning argumentation exacerbates the difficulty in evaluating these systems in a systematic manner. Current systems' performance is evaluated with metrics such as coherence. Nevertheless, new evaluation methods such as the ones described in Heuer and Buschek (2021) should be explored. Therefore, we should promote open-source projects and the research of standard guidelines.

Domain Adaptation Towards effectively explaining output to improve critical thinking skills of users, future systems must be capable of understanding the topic of discussion in a way that argumentation errors (e.g., fallacies) can be identified. In a pedagogical setting, teachers have the ability to choose new topics of discussion annually; hence, systems must also be capable of adapting to various domains. Recent works have focused on domain adaptation for tasks such as short answer scoring (Funayama et al., 2023), which focus on training models for several tasks to learn common properties helpful in evaluating unseen topics. We must also adopt such strategies for computational argumentation to ensure the most reliable feedback is given to the user.

Collaboration NLP researchers and pedagogical researchers generally conduct their research independently, thus creating a gap. We suggest that researchers from both fields must come together to ensure that appropriate and sufficient explanations are provided to learners. Ideally, a system for linking various educational schools and providers with artificial intelligence researchers could significantly help assist with ensuring systems can be properly evaluated.

Ethics Tailoring a constructive feedback system to each user's background and current worldview would benefit the user significantly. Nevertheless, the creation of such a system presents significant challenges in navigating ethical issues (Hovy et al.,

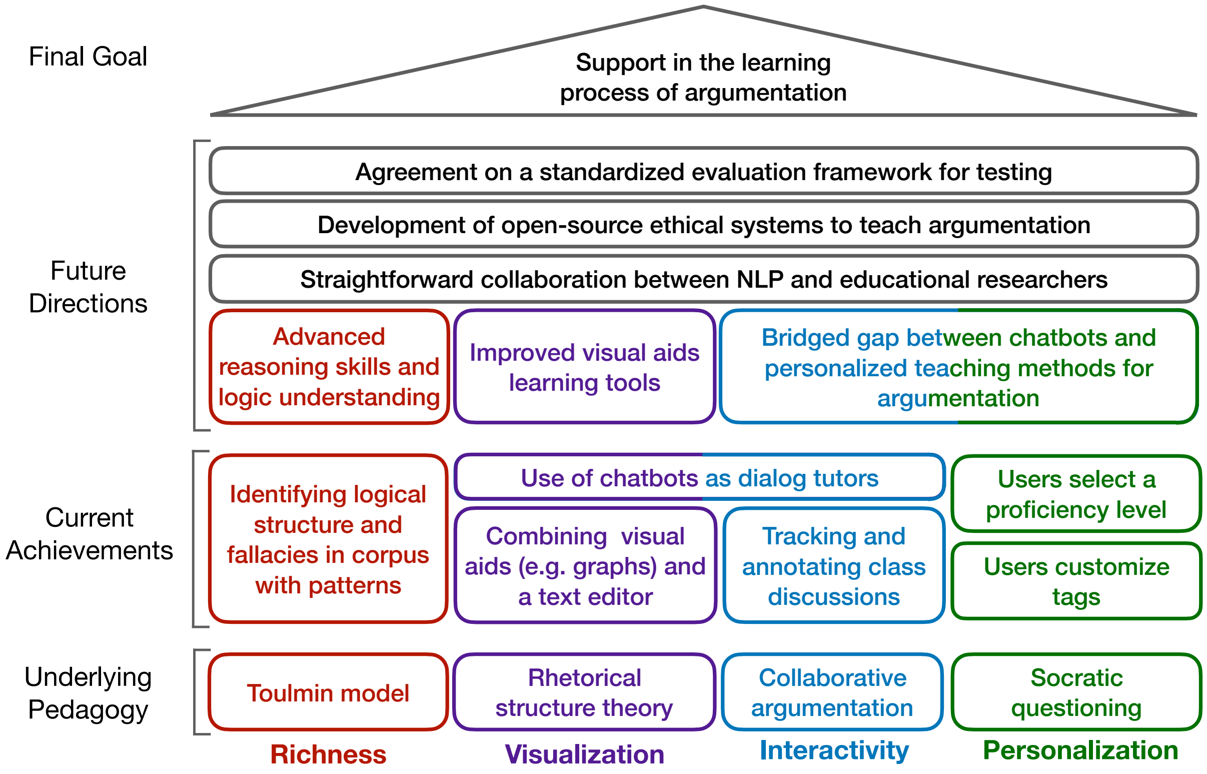

Final Goal
Support in the learning process of argumentation
Future Directions
Agreement on a standardized evaluation framework for testing
Development of open-source ethical systems to teach argumentation
Straightforward collaboration between NLP and educational researchers
Advanced reasoning skills and logic understanding
Improved visual aids learning tools
Bridged gap between chatbots and personalized teaching methods for argumentation
Current Achievements
Identifying logical structure and fallacies in corpus with patterns
Use of chatbots as dialog tutors
Combining visual aids (e.g. graphs) and a text editor
Tracking and annotating class discussions
Users select a proficiency level
Users customize tags
Underlying Pedagogy
Toulmin model
Rhetorical structure theory
Collaborative argumentation
Socratic questioning
Richness
Visualization
Interactivity
Personalization

Figure 3: Current and future directions of teaching argumentation with NLP systems. Boxes with a specific color correspond to a specific dimension, whereas the ones in black are general directions.

2017; Trust et al., 2023). Hence, conceiving novel systems with an *ethics by design* approach remains important (Leidner and Plachouras, 2017). *Ethics by design* is a concept that emphasizes the integration of ethical considerations and principles into the design and development of products, systems, technologies, and processes from the very beginning. It promotes the idea that ethical considerations should be a fundamental part of the design process rather than added as an afterthought or compliance requirement. This approach aims to prevent and mitigate potential ethical issues, such as privacy violations, bias, discrimination, and lack of transparency, by building ethical principles and values into the core of a project. In order to add this principles in a project, Leidner and Plachouras (2017) suggest an Ethics Review Board (ERB) for companies and research institutions, as well as a list of remedies that researchers can consider when facing ethical dilemmas.

9 Conclusion

In our survey, we explored several works providing feedback in argumentation, following various dimensions: *Richness* (§4), *Visualization* (§5), *Interactivity* (§6), and *Personalization* (§7). Figure 3 summarizes the pedagogy, current achievements and potential future directions of each dimension.

As potential areas for improvement to enhance the quality of educational argumentative systems, we highlighted the following points: (1) generate accurate, constructive feedback for a real-life input(§4-5), (2) tailor the output based on the user's background (§6-7), (3) evaluate and compare end-to-end systems more deeply(§8), (4) improve models' abilities to adapt to unknown topics(§8), (5) collaborate with pedagogical teams and actual students(§8), and finally (6) take into consideration ethical issues(§8). For instance, in challenge (2), the use of conversational agents becomes increasingly frequent. However, such systems still leave room for improvement, particularly their ability to tailor discussions based on the user's background.

We hope our survey contributes to enriching the research community focused on argumentation with a comprehensive understanding of current perspectives in NLP systems for teaching how to argue. In our future work, we will focus further on real-life and end-to-end systems (Challenges (1) and (3)). We plan to prototype a system to measure the effects of different feedback on users and evaluate it in actual classrooms (Appendix, Figure 5).

Limitations

This survey offers an overview of NLP feedback systems in argumentation. Despite our best efforts, some limitations may still exist in this research.

Paper selection Our survey primarily focuses on argumentative feedback systems in the context of NLP and human-machine interaction, but there may be valuable insights from other feedback systems that could be applied to argumentation. For instance, feedback systems for grammatical errors, such as (Liang et al., 2023), could inspire new argumentative feedback systems. Moreover, we excluded non-English articles in our survey and prioritized works dedicated to students rather than teachers (e.g., Datta et al., 2021).

Categorization Based on our understanding and subjective opinions, we have categorized the works into four dimensions. It could be relevant to have external opinions on this categorization.

Descriptions The descriptions provided in this survey are generally concise to ensure comprehensive coverage within the constraints of page limits. We hope this survey can be a reference, directing readers to more detailed information in the respective works.

Experiments It is important to note that this survey is purely informational and lacks experimental data or empirical results. Conducting comparative experiments with different feedback systems could offer more substantial guidance. However, this aspect is left for future research.

Acknowledgements

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

References

Jamie Abrams. 2015. Reframing the socratic method. *Journal of Legal Education*, 64(4):562–585.

Tazin Afrin, Omid Kashefi, Christopher Olshefski, Diane Litman, Rebecca Hwa, and Amanda Godley. 2021. Effective interfaces for student-driven revision sessions for argumentative writing. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (CHI 2021)*. ACM.

Annalena Aicher, Nadine Gerstenlauer, Isabel Feustel, Wolfgang Minker, and Stefan Ultes. 2022. Towards building a spoken dialogue system for argument exploration. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 1234–1241, Marseille, France. European Language Resources Association.

Erfan Al-Hossami, Razvan Bunescu, Ryan Teehan, Laurel Powell, Khyati Mahajan, and Mohsen Dorodchi. 2023. Socratic questioning of novice debuggers: A benchmark dataset and preliminary evaluations. In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*, pages 709–726, Toronto, Canada. Association for Computational Linguistics.

Khalid Al-Khatib, Henning Wachsmuth, Matthias Hagen, Jonas Köhler, and Benno Stein. 2016. Cross-domain mining of argumentative text through distant supervision. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1395–1404, San Diego, California. Association for Computational Linguistics.

Milad Alshomary, Shahbaz Syed, Arkajit Dhar, Martin Potthast, and Henning Wachsmuth. 2021. Counter-argument generation by attacking weak premises. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1816–1827, Online. Association for Computational Linguistics.

Beng Heng Ang, Sujatha Das Gollapalli, and See-Kiong Ng. 2023. Socratic question generation: A novel dataset, models, and evaluation. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 147–165, Dubrovnik, Croatia. Association for Computational Linguistics.

Michael Baker, Jerry Andriessen, and Baruch Schwarz. 2019. *Collaborative Argumentation-Based Learning*, pages pp. 76–88. Routledge.

Jianzhu Bao, Jingyi Sun, Qinglin Zhu, and Ruifeng Xu. 2022. Have my arguments been replied to? argument pair extraction as machine reading comprehension. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 29–35, Dublin, Ireland. Association for Computational Linguistics.

Maria Becker, Katharina Korfhage, and Anette Frank. 2020. Implicit knowledge in argumentative texts: An annotated corpus. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 2316–2324, Marseille, France. European Language Resources Association.

Maria Becker, Siting Liang, and Anette Frank. 2021. Reconstructing implicit knowledge with language models. In *Proceedings of Deep Learning Inside Out (DeeLIO): The 2nd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures*, pages 11–24, Online. Association for Computational Linguistics.

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

Beata Beigman Klebanov and Nitin Madnani. 2020. Automated evaluation of writing – 50 years and counting. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7796–7810, Online. Association for Computational Linguistics.

Claire Bonial, Austin Blodgett, Taylor Hudson, Stephanie M. Lukin, Jeffrey Micher, Douglas Summers-Stay, Peter Sutor, and Clare Voss. 2022. The search for agreement on logical fallacy annotation of an infodemic. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4430–4438, Marseille, France. European Language Resources Association.

Kasia Budsziyska, Mathilde Janier, Chris Reed, Patrick Saint-Dizier, Manfred Stede, and Olena Yakorska. 2014. A model for processing illocutionary structures and argumentation in debates. In *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC 2014)*, volume 14, pages electronic–medium. European Language Resources Association (ELRA).

Artem Chernodub, Oleksiy Oliynyk, Philipp Heidenreich, Alexander Bondarenko, Matthias Hagen, Chris Biemann, and Alexander Panchenko. 2019. TARGER: Neural argument mining at your fingertips. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 195–200, Florence, Italy. Association for Computational Linguistics.

Christie A. Linskens Christie. 2010. What critique have been made of the socratic method in legal education. *European Journal of Law Reform*, 12.

Bhavana Dalvi Mishra, Greg Durrett, Peter Jansen, Danilo Neves Ribeiro, and Jason Wei, editors. 2023. *Proceedings of the 1st Workshop on Natural Language Reasoning and Structured Explanations (NLRSE 2023)*. Association for Computational Linguistics, Toronto, Canada.

Marina Danilevsky, Kun Qian, Ranit Aharonov, Yannis Katsis, Ban Kawas, and Prithviraj Sen. 2020. A survey of the state of explainable ai for natural language processing.

Debajyoti Datta, Maria Phillips, James P. Bywater, Jennifer Chiu, Ginger S. Watson, Laura Barnes, and Donald Brown. 2021. Virtual pre-service teacher assessment and feedback via conversational agents. In *Proceedings of the 16th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 185–198, Online. Association for Computational Linguistics.

Roxanne El Baff, Henning Wachsmuth, Khalid Al-Khatib, and Benno Stein. 2018. Challenge or empower: Revisiting argumentation quality in a news editorial corpus. In *Proceedings of the 22nd Conference on Computational Natural Language Learning*, pages 454–464, Brussels, Belgium. Association for Computational Linguistics.

Vanessa Wei Feng and Graeme Hirst. 2011. Classifying arguments by scheme. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 987–996, Portland, Oregon, USA. Association for Computational Linguistics.

Hiroaki Funayama, Yuya Asazuma, Yuichiroh Matsubayashi, Tomoya Mizumoto, and Kentaro Inui. 2023. Reducing the cost: Cross-prompt pre-finetuning for short answer scoring. In *Lecture Notes in Computer Science*, page 78–89, Berlin, Heidelberg. Springer-Verlag.

Vetti Giri and M. U. Paily. 2020. Effect of scientific argumentation on the development of critical thinking. *Science & Education*, 29.

Pierpaolo Goffredo, Shohreh Haddadan, Vorakit Vorakitphan, Elena Cabrio, and Serena Villata. 2022. Fallacious argument classification in political debates. In *Proceedings of the 31st International Joint Conference on Artificial Intelligence IJCAI 2022*, pages 4143–4149. International Joint Conferences on Artificial Intelligence Organization. Main Track.

Carina González-González, Vanesa Muñoz-Cruz, Pedro Antonio Toledo-Delgado, and Eduardo Nacimiento-García. 2023. Personalized gamification for learning: A reactive chatbot architecture proposal. *Sensors*, 23(1).

Ivan Habernal and Iryna Gurevych. 2016. Argumentation mining in user-generated web discourse. *CoRR*, abs/1601.02403.

Ivan Habernal, Raffael Hannemann, Christian Pollak, Christopher Klamm, Patrick Pauli, and Iryna Gurevych. 2017. Argotario: Computational argumentation meets serious games. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 7–12, Copenhagen, Denmark. Association for Computational Linguistics.

Ivan Habernal, Henning Wachsmuth, Iryna Gurevych, and Benno Stein. 2018. The argument reasoning comprehension task: Identification and reconstruction of implicit warrants. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1930–1940, New Orleans, Louisiana. Association for Computational Linguistics.

Mareike Hartmann and Daniel Sonntag. 2022. A survey on improving nlp models with human explanations.

Annette Hautli-Janisz, Zlata Kikteva, Wassiliki Siskou, Kamila Gorska, Ray Becker, and Chris Reed. 2022. QT30: A corpus of argument and conflict in broadcast debate. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 3291–3300, Marseille, France. European Language Resources Association.

Philipp Heinisch, Anette Frank, Juri Opitz, and Philipp Cimiano. 2022. Strategies for framing argumentative conclusion generation. In *Proceedings of the 15th International Conference on Natural Language Generation*, pages 246–259, Waterville, Maine, USA and virtual meeting. Association for Computational Linguistics.

Hendrik Heuer and Daniel Buschek. 2021. Methods for the design and evaluation of HCI+NLP systems. In *Proceedings of the First Workshop on Bridging Human–Computer Interaction and Natural Language Processing*, pages 28–33, Online. Association for Computational Linguistics.

Shengluan Hou, Shuhan Zhang, and Chaoqun Fei. 2020. Rhetorical structure theory: A comprehensive review of theory, parsing methods and applications. *Expert Systems with Applications*, 157:113421.

Dirk Hovy, Shannon Spruit, Margaret Mitchell, Emily M. Bender, Michael Strube, and Hanna Wallach, editors. 2017. *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*. Association for Computational Linguistics, Valencia, Spain.

Jie Huang and Kevin Chen-Chuan Chang. 2023. Towards reasoning in large language models: A survey. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 1049–1065, Toronto, Canada. Association for Computational Linguistics.

Anthony Hunter, Lisa Chalaguine, Tomasz Czernuszenko, Emmanuel Hadoux, and Sylwia Polberg. 2019. Towards computational persuasion via natural language argumentation dialogues. In *KI 2019: Advances in Artificial Intelligence*, pages 18–33, Cham. Springer International Publishing.

Gwo-Jen Hwang and Ching-Yi Chang. 2021. A review of opportunities and challenges of chatbots in education. *Interactive Learning Environments*, 0(0):1–14.

Sheikh Rabiul Islam, William Eberle, Sheikh Khaled Ghafoor, and Mohiuddin Ahmed. 2021. Explainable artificial intelligence approaches: A survey.

Heba Ismail, Nada Hussein, Saad Harous, and Ashraf Khalil. 2023. Survey of personalized learning software systems: A taxonomy of environments, learning content, and user models. *Education Sciences*, 13(7).

Zhijing Jin, Abhinav Lalwani, Tejas Vaidhya, Xiaoyu Shen, Yiwen Ding, Zhiheng Lyu, Mrinmaya Sachan, Rada Mihalcea, and Bernhard Schoelkopf. 2022. Logical fallacy detection. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 7180–7198, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Yohan Jo, Jacky Visser, Chris Reed, and Eduard Hovy. 2020. Extracting implicitly asserted propositions in argumentation.

Yohan Jo, Haneul Yoo, JinYeong Bak, Alice Oh, Chris Reed, and Eduard Hovy. 2021. Knowledge-enhanced evidence retrieval for counterargument generation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, Online. Association for Computational Linguistics.

Zixuan Ke and Vincent Ng. 2019. Automated essay scoring: A survey of the state of the art. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence IJCAI 2019*, pages 6300–6308. International Joint Conferences on Artificial Intelligence Organization.

Orin Kerr. 1999. The decline of the socratic method at harvard. *Nebraska law review*, 78:113.

Zlata Kikteva, Kamila Gorska, Wassiliki Siskou, Annette Hautli-Janisz, and Chris Reed. 2022. The keystone role played by questions in debate. In *Proceedings of the 3rd Workshop on Computational Approaches to Discourse*, pages 54–63, Gyeongju, Republic of Korea and Online. International Conference on Computational Linguistics.

Chokri Kooli. 2023. Chatbots in education and research: A critical examination of ethical implications and solutions. *Sustainability*, 15:5614.

Anne Lauscher, Goran Glavaš, and Kai Eckert. 2018. ArguminSci: A tool for analyzing argumentation and rhetorical aspects in scientific writing. In *Proceedings of the 5th Workshop on Argument Mining*, pages 22–28, Brussels, Belgium. Association for Computational Linguistics.

Anne Lauscher, Henning Wachsmuth, Iryna Gurevych, and Goran Glavaš. 2022. Scientia potentia Est—On the role of knowledge in computational argumentation. *Transactions of the Association for Computational Linguistics*, 10:1392–1422.

John Lawrence and Chris Reed. 2020. Argument Mining: A Survey. *Computational Linguistics*, 45(4):765–818.

Mina Lee, Percy Liang, and Qian Yang. 2022. CoAuthor: Designing a human-AI collaborative writing dataset for exploring language model capabilities. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems (CHI 2022)*. ACM.

Jochen L. Leidner and Vassilis Plachouras. 2017. Ethical by design: Ethics best practices for natural language processing. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, pages 30–40, Valencia, Spain. Association for Computational Linguistics.

Kai-Hui Liang, Sam Davidson, Xun Yuan, Shehan Panditharatne, Chun-Yen Chen, Ryan Shea, Derek Pham, Yinghua Tan, Erik Voss, and Luke Fryer. 2023. ChatBack: Investigating methods of providing grammatical error feedback in a GUI-based language learning chatbot. In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*, pages 83–99, Toronto, Canada. Association for Computational Linguistics.

Chien-Chang Lin, Anna Huang, and Stephen Yang. 2023. A review of ai-driven conversational chatbots implementation methodologies and challenges (1999–2022). *Sustainability*, 15:4012.

Haochen Liu, Zitao Liu, Zhongqin Wu, and Jiliang Tang. 2020. Personalized multimodal feedback generation in education. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 1826–1840, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Luca Lugini, Christopher Olshefski, Ravneet Singh, Diane Litman, and Amanda Godley. 2020. Discussion tracker: Supporting teacher learning about students' collaborative argumentation in high school classrooms. In *Proceedings of the 28th International Conference on Computational Linguistics: System Demonstrations*, pages 53–58, Barcelona, Spain (Online). International Committee on Computational Linguistics (ICCL).

Zhengyi Ma, Zhicheng Dou, Yutao Zhu, Hanxun Zhong, and Ji-Rong Wen. 2021. One chatbot per person: Creating personalized chatbots based on implicit user profiles. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2021)*. ACM.

Jakub Macina, Nico Daheim, Lingzhi Wang, Tanmay Sinha, Manu Kapur, Iryna Gurevych, and Mrinmaya Sachan. 2023. Opportunities and challenges in neural dialog tutoring.

William Mann and Sandra Thompson. 1988. Rhetorical structure theory: Toward a functional theory of text organization. *Text - Interdisciplinary Journal for the Study of Discourse*, 8(3):243–281.

Kana Matsumura and Teruyo Sakamoto. 2021. A structure analysis of japenese efl students' argumentative paragraph writings with a tool for annotating discourse relations. *Bulletin of the JACET Kansai Branch Writing Guidance Study Group*, 14:pp. 31–50.

Farjana Sultana Mim, Naoya Inoue, Shoichi Naito, Keshav Singh, and Kentaro Inui. 2022. LPAttack: A feasible annotation scheme for capturing logic pattern of attacks in arguments. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2446–2459, Marseille, France. European Language Resources Association.

Behzad Mirzababaei and Viktoria Pammer-Schindler. 2022. Learning to give a complete argument with a conversational agent: An experimental study in two domains of argumentation. In *Educating for a New Future: Making Sense of Technology-Enhanced Learning Adoption*, pages 215–228, Cham. Springer International Publishing.

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

Callistus Ireneous Nakpih and Simone Santini. 2020. Automated discovery of logical fallacies in legal argumentation. *International Journal of Artificial Intelligence & Applications*.

Ellis Page. 1966. The imminence of... grading essays by computer. *The Phi Delta Kappan*, 47(5):238–243.

Artidoro Pagnoni, Alex Fabbri, Wojciech Kryscinski, and Chien-Sheng Wu. 2023. Socratic pretraining: Question-driven pretraining for controllable summarization. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL 2023)*, pages 12737–12755, Toronto, Canada. Association for Computational Linguistics.

Georgios Petasis. 2014. Annotating arguments: The nomad collaborative annotation tool. In *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC 2014)*, Reykjavik, Iceland. European Language Resources Association (ELRA).

R.T. Pithers and Rebecca Soden. 2000. Critical thinking in education: a review. *Educational Research*, 42(3):237–249.

Jan Wira Gotama Putra, Kana Matsumura, Simone Teufel, and Takenobu Tokunaga. 2021. Tiara 2.0: an interactive tool for annotating discourse structure and text improvement. *Language Resources and Evaluation*, 57:5 – 29.

Jan Wira Gotama Putra, Simone Teufel, Kana Matsumura, and Takenobu Tokunaga. 2020. TIARA: A tool for annotating discourse relations and sentence reordering. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 6912–6920, Marseille, France. European Language Resources Association.

Hongjin Qian, Xiaohe Li, Hanxun Zhong, Yu Guo, Yueyuan Ma, Yutao Zhu, Zhanliang Liu, Zhicheng Dou, and Ji-Rong Wen. 2021. Pchatbot: A large-scale dataset for personalized chatbot.

Niklas Rach, Yuki Matsuda, Johannes Daxenberger, Stefan Ultes, Keiichi Yasumoto, and Wolfgang Minker. 2020. Evaluation of argument search approaches in the context of argumentative dialogue systems. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 513–522, Marseille, France. European Language Resources Association.

Olesya Razuvayevskaya and Simone Teufel. 2017. Finding enthymemes in real-world texts: A feasibility study. *Argument Computation*.

Lesley Rex, Ebony Thomas, and Steven Engel. 2010. Applying toulmin: Teaching logical reasoning and argumentative writing. *The English Journal*, 99.

Ameer Saadat-Yazdi, Jeff Z. Pan, and Nadin Kokciyan. 2023. Uncovering implicit inferences for improved relational argument mining. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2484–2495, Dubrovnik, Croatia. Association for Computational Linguistics.

Swarnadeep Saha, Prateek Yadav, Lisa Bauer, and Mohit Bansal. 2021. ExplaGraphs: An explanation graph generation task for structured commonsense reasoning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7716–7740, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Frederick Schauer. 2012. *Thinking like a Lawyer*. Harvard University Press.

Oliver Scheuer, Frank Loll, Niels Pinkwart, and Bruce Mclaren. 2010. Computer-supported argumentation: A review of the state of the art. *I. J. Computer-Supported Collaborative Learning*, 5:43–102.

Kumar Shridhar, Jakub Macina, Mennatallah El-Assady, Tanmay Sinha, Manu Kapur, and Mrinmaya Sachan. 2022. Automatic generation of socratic subquestions for teaching math word problems. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 4136–4149, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Keshav Singh, Naoya Inoue, Farjana Sultana Mim, Shoichi Naito, and Kentaro Inui. 2022. IRAC: A domain-specific annotated corpus of implicit reasoning in arguments. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4674–4683, Marseille, France. European Language Resources Association.

Keshav Singh, Farjana Sultana Mim, Naoya Inoue, Shoichi Naito, and Kentaro Inui. 2021. Exploring methodologies for collecting high-quality implicit reasoning in arguments. In *Proceedings of the 8th Workshop on Argument Mining*, pages 57–66, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Yi Song, Michael Heilman, Beata Beigman Klebanov, and Paul Deane. 2014. Applying argumentation schemes for essay scoring. In *Proceedings of the First Workshop on Argumentation Mining*, pages 69–78, Baltimore, Maryland. Association for Computational Linguistics.

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

Simone Teufel. 1999. *Argumentative Zoning: Information Extraction from Scientific Text*. Ph.D. thesis, University of Edinburgh.

Stephen Toulmin. 1958. *The Uses of Arguments*, 1 edition. Cambridge University Press.

Torrey Trust, Jeromie Whalen, and Chrystalla Mouza. 2023. Editorial: Chatgpt: Challenges, opportunities, and implications for teacher education. *Contemporary Issues in Technology and Teacher Education*, 23(1):1–23.

Charles Twardy. 2004. Argument maps improve critical thinking. *Teaching Philosophy*, 27.

Alexandros Vassiliades, Nick Bassiliades, and Theodore Patkos. 2021. Argumentation and explainable artificial intelligence: a survey. *The Knowledge Engineering Review*, 36:e5.

Arja Veerman, Jerry Andriessen, and Gellof Kanselaar. 2002. Collaborative argumentation in academic education. *Instructional Science*, 40(3).

Henning Wachsmuth and Milad Alshomary. 2022. "mama always had a way of explaining things so I could understand": A dialogue corpus for learning to construct explanations. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 344–354, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Henning Wachsmuth, Nona Naderi, Yufang Hou, Yonatan Bilu, Vinodkumar Prabhakaran, Tim Alberdingk Thijm, Graeme Hirst, and Benno Stein. 2017. Computational argumentation quality assessment in natural language. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 176–187, Valencia, Spain. Association for Computational Linguistics.

Douglas Walton, Christopher Reed, and Fabrizio Macagno. 2008. *Argumentation schemes*. Cambridge University Press.

Thiemo Wambsganss, Andrew Caines, and Paula Buttery. 2022a. ALEN app: Argumentative writing support to foster English language learning. In *Proceedings of the 17th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2022)*, pages 134–140, Seattle, Washington. Association for Computational Linguistics.

Thiemo Wambsganss, Andreas Janson, Tanja Käser, and Jan Marco Leimeister. 2022b. Improving students argumentation learning with adaptive self-evaluation nudging. *Proceedings of the ACM on Human-Computer Interaction (PACMHCI 2022)*, 6(520):1–31.

Thiemo Wambsganß, Tobias Kueng, Matthias Söllner, 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 2021)*, pages 1–13.

Thiemo Wambsganß, Christina Niklaus, Matthias Cetto, Matthias Söllner, Siegfried Handschuh, and Jan Marco Leimeister. 2020. Al: An adaptive learning support system for argumentation skills. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (CHI 2020)*.

Boshi Wang, Sewon Min, Xiang Deng, Jiaming Shen, You Wu, Luke Zettlemoyer, and Huan Sun. 2023a. Towards understanding chain-of-thought prompting: An empirical study of what matters. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL 2023)*, pages 2717–2739, Toronto, Canada. Association for Computational Linguistics.

Lingzhi Wang, Mrinmaya Sachan, Xingshan Zeng, and Kam-Fai Wong. 2023b. Strategize before teaching: A conversational tutoring system with pedagogy self-distillation.

Xinyu Wang, Yohan Lee, and Juneyoung Park. 2022. Automated evaluation for student argumentative writing: A survey.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, pages 24824–24837. Curran Associates, Inc.

Armin Weinberger and Frank Fischer. 2006. A framework to analyze argumentative knowledge construction in computer-supported collaborative learning. *Computers & Education*, 46(1):71–95. Methodological Issues in Researching CSCL.

Meng Xia, Qian Zhu, Xingbo Wang, Fei Nie, Huamin Qu, and Xiaojuan Ma. 2022. Persua: A visual interactive system to enhance the persuasiveness of arguments in online discussion. *Proceedings of the 25th ACM Conference On Computer-Supported Cooperative Work And Social Computing (CSCW 2022)*, 6(CSCW2):1–30.

Stuart Yeh. 1998. Empowering education: Teaching argumentative writing to cultural minority middle-school students. research in the teaching of english. *Research in the Teaching of English*, 33(1):49–83.

Jian Yuan, Zhongyu Wei, Donghua Zhao, Qi Zhang, and Changjian Jiang. 2021. Leveraging argumentation knowledge graph for interactive argument pair identification. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2310–2319, Online. Association for Computational Linguistics.

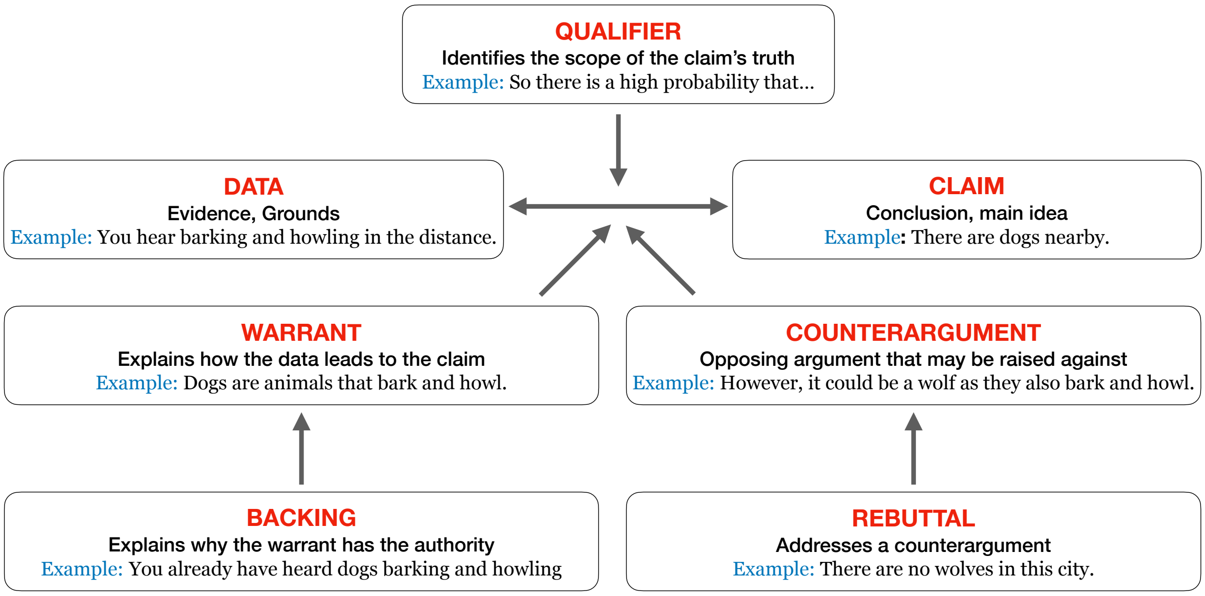
Fan Zhang, Rebecca Hwa, Diane Litman, and Homa B. Hashemi. 2016. ArgRewrite: A web-based revision assistant for argumentative writings. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, pages 37–41, San Diego, California. Association for Computational Linguistics.

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.

Sourati Zhivar, Ilievski Filip, Sandlin Hông-Ân, and Mermoud Alain. 2023. Case-based reasoning with language models for classification of logical fallacies.

Kristijonas Čyras, Antonio Rago, Emanuele Albini, Pietro Baroni, and Francesca Toni. 2021. Argumentative xai: A survey.

A Appendix



QUALIFIER
Identifies the scope of the claim's truth
Example: So there is a high probability that...

DATA
Evidence, Grounds
Example: You hear barking and howling in the distance.

CLAIM
Conclusion, main idea
Example: There are dogs nearby.

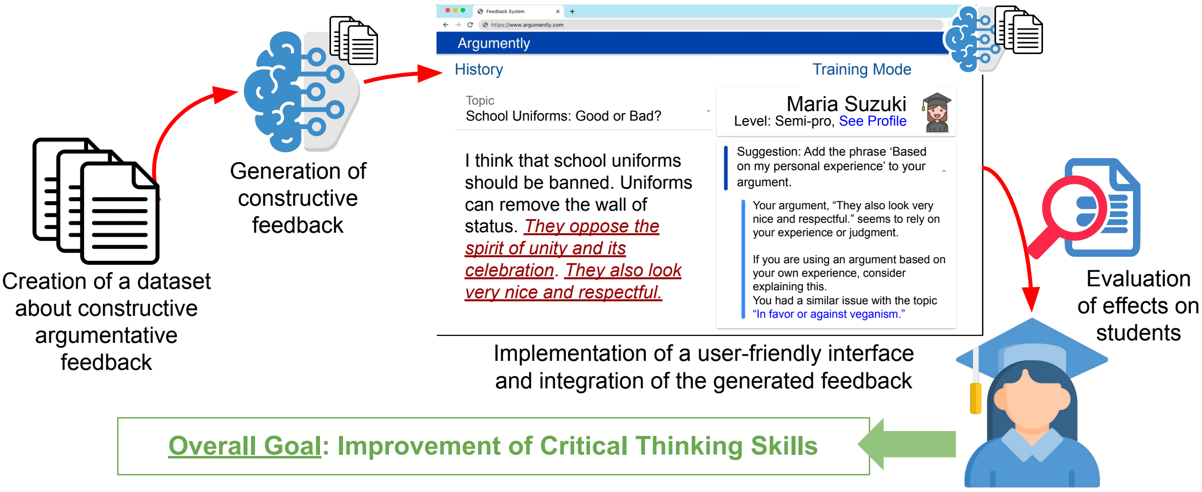
WARRANT
Explains how the data leads to the claim
Example: Dogs are animals that bark and howl.

COUNTERARGUMENT
Opposing argument that may be raised against
Example: However, it could be a wolf as they also bark and howl.

BACKING
Explains why the warrant has the authority
Example: You already have heard dogs barking and howling

REBUTTAL
Addresses a counterargument
Example: There are no wolves in this city.

Figure 4: Six elements of the Toulmin's model.



Creation of a dataset about constructive argumentative feedback
Generation of constructive feedback
Feedback System
https://www.argumently.com
Argumently
History
Training Mode

Topic
School Uniforms: Good or Bad?

Maria Suzuki
Level: Semi-pro, See Profile

I think that school uniforms should be banned. Uniforms can remove the wall of status. They oppose the spirit of unity and its celebration. They also look very nice and respectful.

Suggestion: Add the phrase 'Based on my personal experience' to your argument.

Your argument, "They also look very nice and respectful." seems to rely on your experience or judgment.

If you are using an argument based on your own experience, consider explaining this.
You had a similar issue with the topic "In favor or against veganism."

Implementation of a user-friendly interface and integration of the generated feedback
Evaluation of effects on students
Overall Goal: Improvement of Critical Thinking Skills

Figure 5: Preliminary sketch design of an end-to-end system to learn argumentation.