Please quote as: Neshaei, S. P., Tolzin, A., Berkle, Y., Leuchter, M., Leimeister, J. M., Janson, A. & Wambsganss, T. (2025). Leveraging Learner Errors in Digital Argumentation Learning: How ALure Helps Students Learn from their Mistakes and Write Better Arguments. Proceedings of the ACM on Human-Computer Interaction, Computer-Supported Cooperative Work and Social Computing (CSCW), : Association for Computing Machinery.

Leveraging Learner Errors in Digital Argumentation Learning: How ALure Helps Students Learn from their Mistakes and Write Better Arguments

SEYED PARSA NESHAEI, EPFL, Switzerland
ANTONIA TOLZIN, Research Center for IS Design (ITeG), University of Kassel, Germany
YVONNE BERKLE, RPTU University Kaiserslautern-Landau, Germany
MIRIAM LEUCHTER, RPTU University Kaiserslautern-Landau, Germany
JAN MARCO LEIMEISTER, University of St.Gallen, Switzerland and Research Center for IS Design (ITeG), University of Kassel, Germany
ANDREAS JANSON, University of St.Gallen, Switzerland
THIEMO WAMBSGANSS, Institute for Digital Technology Management, Bern University of Applied Sciences, Switzerland

Providing argumentation feedback is considered helpful for students preparing to work in collaborative environments, helping them with writing higher-quality argumentative texts. Domain-independent natural language processing (NLP) methods, such as generative models, can utilize learner errors and fallacies in argumentation learning to help students write better argumentative texts. To test this, we collect design requirements, and then design and implement two different versions of our system called ALure to improve the students' argumentation skills. We test how ALure helps students learn argumentation in a university lecture with 305 students and compare the learning gains of the two versions of ALure with a control group using video tutoring. We find and discuss the differences of learning gains in argument structure and fallacies in both groups after using ALure, as well as the control group. Our results shed light on the applicability of computer-supported systems using recent advances in NLP to help students in learning argumentation as a necessary skill for collaborative working settings.

CCS Concepts: • **Computing methodologies** → **Natural language processing**; • **Human-centered computing** → **Natural language interfaces**; **Field studies**.

Additional Key Words and Phrases: argumentation learning, writing assistants, learning from errors, natural language processing

ACM Reference Format:
Seyed Parsa Neshaei, Antonia Tolzin, Yvonne Berkle, Miriam Leuchter, Jan Marco Leimeister, Andreas Janson, and Thiemo Wambsganss. 2025. Leveraging Learner Errors in Digital Argumentation Learning: How ALure

Authors' addresses: Seyed Parsa Neshaei, seyed.neshaei@epfl.ch, EPFL, Lausanne, Switzerland; Antonia Tolzin, antonia.tolzin@uni-kassel.de, Research Center for IS Design (ITeG), University of Kassel, Kassel, Germany; Yvonne Berkle, yvonne.berkle@rptu.de, RPTU University Kaiserslautern-Landau, Landau, Germany; Miriam Leuchter, miriam.leuchter@rptu.de, RPTU University Kaiserslautern-Landau, Landau, Germany; Jan Marco Leimeister, janmarco.leimeister@unisg.ch, University of St.Gallen, St.Gallen, Switzerland and Research Center for IS Design (ITeG), University of Kassel, Kassel, Germany; Andreas Janson, andreas.janson@unisg.ch, University of St.Gallen, St.Gallen, Switzerland; Thiemo Wambsganss, thiemo.wambsganss@bfh.ch, Institute for Digital Technology Management, Bern University of Applied Sciences, Bern, Switzerland.

2573-0142/2025/4-ARTCSCW125 $15.00
https://doi.org/10.1145/3711023

Helps Students Learn from their Mistakes and Write Better Arguments. *Proc. ACM Hum.-Comput. Interact.* 9, 2, Article CSCW125 (April 2025), 32 pages. https://doi.org/10.1145/3711023

1 INTRODUCTION

With the progress toward more distributed ways of working, the importance of organizations being able to support, adapt, and innovate collaborative working environments is steadily growing [97]. Among other factors, such environments usually reward continuous learning, actively support the learning efforts of the individuals working in organizations, and enable them to meet the missions of the organization together [37]. As a result and due to the rise in the usage of artificial intelligence (AI)-supported working environments among individuals worldwide, the field of computer-supported cooperative work (CSCW) has centered around novel and effective ways of supporting and instructing users with regards to the ever-changing landscape of working environments (e.g., research ranging from more established works [19, 116] up to recent research considering the role of AI in CSCW [61, 64, 153]). Specifically, in collaborative work environments, interdisciplinary and creative tasks are becoming more relevant and demand higher-order thinking skills such as critical reasoning and problem-solving skills [131]. This shift of skills is a result of an ever-increasing amount of readily available data, entailing a shift in educational institutions, which need to equip learners with the necessary skill set for a future career and thus have to adapt their teaching methods and learning material. One subclass of higher-order thinking skills is the skill of arguing in a structured way; *argumentation* is considered a 'powerful vehicle' to persuade audiences [108]. A well-written argumentation can be used to reach a shared understanding and good argumentative skills are also conducive to critical thinking and creating an argumentation that helps to acquire knowledge about a topic [108]. Although studies have shown that basic skill levels of argumentation are acquired at a very young age early [26, 83], the development of these skills is closely tied to the educational level [62]. Academic achievements are often closely linked to students' interactions with the teacher and the feedback the students receive [2]. However, the settings in traditional classrooms are challenging, as teachers do not have the resources to focus on individual students. While argumentation feedback has been shown to be successful in helping students in a variety of levels to write higher-quality argumentative texts [65, 68, 93, 129], providing personalized argumentative writing feedback by instructors at scale to the students is infeasible, especially in large-scale classrooms and educational curricula [59].

Computer-based intelligent and interactive writing assistants can mitigate this issue by helping in the process of providing feedback at scale to multiple users at the same time [71]. Thus, to support teachers and reduce their workload, recent research in the fields of education and CSCW utilizes natural language processing (NLP) by providing intelligent and adaptive feedback [43, 136, 151]. With modern ML and NLP methods [149] (including text classification [23] or text generation [14]), there is the possibility to implement adaptive learning tools that support learners in their individual learning paths while decreasing the load on teachers. Current digital learning approaches mainly analyze argumentative texts based on the overall structure and number of arguments and support students with adaptive feedback [39, 43]. Research has shown that addressing learner errors can be an effective learning method and be beneficial to learners, as engaging with errors encourages active learning [85, 106]. Learner errors need to be followed by corrective feedback which provides a corresponding explanation of the error and solution to be effective [95], as otherwise, presenting learners with their errors can lead to frustration, decreased motivation, and lower learning outcomes [112]. However, past research lacks studies that investigate the potential of learning from argumentation errors [93, 107].

Our work thus addresses the lack of the use of learner errors in argumentative learning tools. In this research, we develop our design requirements and provide design principles for a collaborative

and cooperative digital learning tool called **ALure** (**A**rgumentation **L**earning through Fail**ure**) that supports learners in their individual learning journey to improve their argumentation skills. We aim to create a learning environment that uses and adapts to learner errors. Thus, we formulate our research question (RQ) as follows:

How can we design a digital learning tool based on the paradigm of *learning from errors* to effectively train students' argumentation skills?

To achieve this goal and answer our RQ, we develop our tool as a user-centric web application. We design two versions of our application with the *learning from errors* paradigm in mind, identifying the student errors in different ways. The first version (*ALure A*) runs a fine-tuned BERT model using text classification based on a novel corpora we compile. This version finds the argumentative errors in the student-written text using the fine-tuned classifier and highlights the errors in the text area. The second version (*ALure B*) works with the GPT-3.5 model using the text generation method. This version includes prompts sent to GPT-3.5 that incorporate information regarding the *learning from errors* paradigm, reflecting the errors made by the students in the generated responses[1]. To the best of our knowledge, ALure is the first web-based intelligent *argumentative* writing assistant that uses two interfaces with support for both modalities of text classification (using BERT) and text generation (using GPT), allowing us to compare the effects of the two modalities among students with the same general user interaction and interface.

To evaluate how our tool helps students in learning argumentation skills, we test it in a large-scale lecture in a Western European university with 305 students. We measure the argumentation skills of the students before and after using ALure using the standardized test provided by Berkle et al. [8]. We compare the two versions of ALure with a control group that receives standard instructional videos also used in the previous course offerings to learn argumentation from, instead of using any intelligent writing assistant with feedback.

The results of our evaluation show a learning gain for all three groups, i.e., increased scores for both argument structure and fallacies[2] in our post-test compared to our pre-test, showing the effectiveness of the *learning from errors* paradigm in argumentative writing. Specifically, in *argument structure*, *ALure B* had the most learning gain, followed closely by video instructions, and with *ALure A* at the third position. In *fallacies*, both *ALure B* and *ALure A* led to the most learning gain in order, leaving video instructions at a distant third place with even a slight decrease in test scores. We also find a significant difference in results for ALure B from the pre-test to the post-test.

We contribute to the writing assistants literature and complement the previous CSCW research works on argumentation and persuasive writing support [136, 150, 152] by A) design insights and principles for implementing AI-based models for argumentative writing education using the *learning from errors* paradigm, and B) comparing a generative model (usable out-of-the-box but uncontrollable and prone to hallucination) versus integrating the predictions of a small model (controllable with a set of pre-defined outputs but with relatively lower performance and need for data annotation) into our user interface. Our work sheds light on the similarities and differences of using text generation and text classification NLP modalities when embedded in argumentation learning writing assistants, as well as on the effects of cooperative and collaborative intelligent writing assistants on argumentation learning designed and implemented using the *learning from errors* paradigm. As a result, our work has the potential to shape the general direction for future

[1]For the GPT group, we did not develop or fine-tune our custom generative large language model (and rather used GPT with prompt engineering).

[2]We define *fallacies* as issues often found in student argumentative texts, namely overgeneralization, circular fallacies, logical fallacies, and passages with low relevance. For more details, see Section 2.1.

CSCW research in the domain of helping future employees in their argumentative skills, readying them for the ever-changing collaborative work environments.

2 RELATED WORKS AND THEORETICAL BACKGROUND

Our work was inspired by the previous works on argumentation learning and related background theories underpinning our study, as well as intelligent writing assistants, especially those that help learners to improve their argumentation skills.

2.1 Argumentation Learning and Fallacies

The organizations worldwide are moving towards distributed working methods, necessitating collaborative working environments in which learners are actively supported to reach organizational goals together [37, 97]. The creative tasks in such environments necessitate educational programs to provide learners with a set of relevant skills for their future careers, with a focus on critical reasoning and higher-order thinking.

As a concrete necessary skill to possess, the task of *argumentation* is considered useful not only for persuasive purposes, but also as a mechanism to 'achieve goals, clarify a doubt, decide, solve a conflict, amplify knowledge, etc.' [101]. Training argumentative skills implicitly involves the training of other skills such as acquiring basic knowledge about a topic, critical reasoning, and problem-solving [108]. The latter two are considered higher-order thinking skills, which are considered to require higher amounts of cognitive processing and are thus more complex to teach and learn than basic knowledge skills [6]. Argumentative skills are thus important in both the work environment as well as for training a wide range of other skills. Therefore, improving the argumentation skills of students is an essential but challenging factor in education. To train argumentation skills, the subsequent sketched pedagogical body of literature provides promising entry points.

Argumentation theory (i.e., the study of argumentation) has been applied in a variety of fields, such as pedagogy or machine learning [18, 109]. Consequently, there are many existing argumentation models, but in the scope of this paper, only monological models, where a single person constructs arguments and draws conclusions [9], are of interest, as digital argumentative learning tools typically focus on monological arguments (for exceptions, see Latifi et al. [67]). The overall structure of arguments consists of *claims* and *premises*, where claims represent debatable opinions and positions that are supported or attacked by premises [96, 128, 133].

A well-written argumentation follows the formal structure of (simplified) theoretical models and is also characterized in the sense that it does not contain argumentative *fallacies*, as they weaken an argument or can even lead to incorrect statements [130]. It is important to note that there is an ongoing debate about argumentative fallacies in the literature [100, 130] and there is currently no consensus on how to define fallacies and on the exact terminology. For the sake of simplicity, we define fallacies as the four fallacy types often found in student texts and covered by Berkle et al. [8], including 1) overgeneralization [7, 53], 2) circular fallacies [41, 102], 3) logical fallacies [15, 80], and 4) low relevance [22, 49, 134]. The first three types of fallacies are structural and violate the internal consistency of fallacies [89], while low relevance is a fallacy with regard to the content of an argument. *Overgeneralizations* use one specific example to draw general conclusions and are thus invalid, e.g. 'My last paper got rejected, so this one will certainly be rejected too.' In *circularities*, the premise repeats the same statement the claim has made, without providing additional information, e.g. 'Smoking is unhealthy (claim), as it is harmful (premise).' *Logical* fallacies violate rules of formal logic, such as the following: 'All cats have four legs. Rex has four legs, so Rex must be a cat.'

2.2 The Paradigm of *Learning From Errors*

With the rising importance of learning tools to be used in educational environments, researchers have covered a set of variations in learning tools to support students more effectively. Specifically, Metcalfe [85] has explained the paradigm of *learning from errors* in learning sciences, stating that "errorful learning followed by corrective feedback" (see page 1) can be useful and beneficial for learners. Historically, making errors has been considered partly detrimental to the ongoing learning process [3, 5, 115]. However, more recent works in HCI and CSCW show promising results in allowing students to interactively practice and make errors, then providing corrective feedback, to help them learn better compared to the traditional reading or video watching-based methods of learning [69, 73, 85, 136, 147].

In particular, allowing learners to correct errors they make as a part of the *learning from errors paradigm* leads to critical and constructive reasoning about the task they are undertaking. This can improve their learning gains, as it encourages learning beyond pure knowledge acquisition and in a hands-on practice session [85]. Highlighting mistakes has the potential to reduce learners' uncertainty and provide them with strategies on how to improve their skills and correct their mistakes in the future [85]. More specifically, the *learning from errors* paradigm is connected with the *cognitive load theory* [58, 124], which states that learners only have limited cognitive resources that need to be addressed adequately. Thus, presenting large sets of reading material can be detrimental to the learning progress of the students as management of cognitive load might fail [50]. On the other hand, the *learning from errors* paradigm allows the students to practically learn a task using a hands-on interactive session with less cognitive load than traditional learning approaches, while receiving corrective feedback [85]; in other words, highlighting errors has the possibility to guide the attention of the students and reduce their cognitive load.

In the *learning from errors* paradigm, the validity, relevancy, and usefulness of the *corrective feedback* have an important role, in addition to merely identifying the learner errors. Previous works have explored a variety of methods to help learners by providing feedback; for example, Ellis [27] highlights the need for both positive and negative feedback to foster motivation for continuing learning. As another example, the work of Bitchener et al. [10] compares a range of methods for providing feedback in a language learning educational setting, and the work of Knight et al. [59] explores providing automated formative feedback on the rhetorical moves in the student writings. Researchers have also explored the best approaches to incorporate feedback in the design of intelligent computer-based tools; for example, Weber et al. [147] classify feedback based on the specific related legal construct in the students' case solutions, or Shi et al. [111] show feedback to the users in the form of hint recommendations for brainstorming, as well as writing refinements.

2.3 Technology-Enhanced Systems for Argumentation Skill Development

Pedagogical strategies for argumentation instruction face significant hurdles. Jonassen and Kim [54] indicate three primary obstacles: (i) deficient pedagogical proficiency among teachers to cultivate argumentation, (ii) limited instructional time due to curricular demands, and (iii) learners' inadequate prior knowledge. As a result, several scholars advocate for prioritizing argumentation skills within formal educational frameworks [25, 63]. Usually, students acquire argumentation abilities through peer or educator interactions, but this approach often lacks individualized guidance. Sustained feedback and personalized instruction are essential for argumentation training to be effective [44, 132].

Due to the increasing use of computer-based tools among working individuals globally, the literature on CSCW has covered exploring effective and useful methods to support users in the evolving landscape of working environments [19, 116]. As a result, interdisciplinary researchers

from educational technology and specifically CSCW have explored how digital assistants can mitigate these gaps and help with students' argumentation learning [150]. Such technology-driven systems offer multiple advantages including scalability, consistency, and greater accessibility compared to traditional educational methods [107]. Researchers have investigated such assistants across a wide range of domains, including law [99, 146], business reviews [140], and conversational argumentation [138], among others. Specifically, previous works in CSCW have considered various aspects of intelligent systems helping students with their argumentation. For example, the work of Wambsganss et al. [136] showed an application of computer-supported tools for helping students in an interactive setting in argumentative writing; they presented how students can be nudged by argumentation evaluation based on their written text. As another example, a previous research [150] covered the design of an interactive visual system which provided guidance on persuasive strategies based on examples. Additionally, Yim et al. [152] explored the quality of student-written persuasive texts in their study, comparing various synchronous collaboration styles in their research.

In particular, CSCW research has regularly discussed the topic of computer-supported collaborative learning (CSCL) [16, 47], which is a field concerning how computer-based technologies can support students in their learning process [24]. Previous works in CSCW and CSCL [107, 140] specify different types of intelligent systems for argumentation learning: *discussion scripting approaches* which supply learners with predefined argumentative components based on the script theory of guidance [31, 107], thereby facilitating structured dialogues; *representational guidance approaches* in which learners are aided by visual depictions of their argumentative frameworks, aimed at enhancing individual reasoning, collaboration, and learning outcomes [91, 99]; and *adaptive support approaches* including systems delivering customized pedagogical feedback, as well as offering suggestions and guidelines for refining argumentation [1, 118, 140]. However, most of these learning tools are situated in domain-specific exercises. Especially for the adaptive systems, mostly supervised modeling approaches have been used to provide students with individual feedback, limited to one domain. In fact, research on domain-independent adaptive support for argumentation learning is rather scarce. Also, to the best of our knowledge, we did not find many approaches that compared large language models (LLMs) using prompting approaches versus text classification models to provide students with adaptive feedback on their argumentative writing.

2.4 Intelligent Writing Assistants

In this research, we follow the definition from previous works [36, 71, 111] and define an *intelligent writing assistant* as any software which supports learners, or more generally any user, in composing, rewriting, or generating new ideas related to their writing task. An example of a writing assistant is Jasper AI[3] which can help with generating written product descriptions or provide ideas for blog posts. Another example is ProWritingAid[4] which provides style improvements, grammar checking, and rephrasing. Other similar systems, such as Grammarly[5] and Linguix[6] are also considered examples of writing assistants. Specifically, writing assistants have been among the common topics in CSCW literature, with researchers exploring how writers can benefit and get support from computer tools helping them in various writing phases [105, 110, 122].

The recent advances with Transformer-based models [149] have enabled a whole new diverse set of writing assistants to be able to support users and learners in writing their texts. Previous researchers, in particular, have explored how to leverage writing assistants to help users in their writing, in the domains of stories [94], administrative texts [33], creative writing [123], lyrics [145],

[3]https://www.jasper.ai
[4]https://prowritingaid.com
[5]https://grammarly.com
[6]https://linguix.com

legal case solutions [146, 147], emotional and emphatic writing [142], peer reviews [40, 90, 120], procedural writing [84], and mental health [98], to name a few.

Writing assistants have also specifically been used in the domain of argumentation learning to help learners. For example, Wambsganss et al. [140] found in their research that students using their argumentation feedback system, AL, wrote more convincing text with better argumentation quality. Previous works also show the effectiveness of their argumentation learning writing assistant in the domain of English language learning [135]. Wambsganss and Niklaus [139] introduced an argumentation annotation approach for student business model pitches and evaluated a writing assistant based on classification models trained on their data in a real-world scenario. Afrin et al. [1] have designed and evaluated an argumentation writing assistant that helps learners to recognize their revisions of argumentative essays. Researchers have also explored embedding argumentation learning writing assistants in conversational agents [138].

In particular, many prior approaches have utilized generative models (e.g., GPT) to provide user support [71]. Such models come with several benefits, such as high zero-shot performance rooting in their pretraining process [11, 60], natural and human-like responses [154], and easy deployment to various downstream tasks without the need for separate human annotation processes [126]. However, they are known to come with issues including hallucinating and low controllability over the range of possible outputs [51], limiting the relevance and usefulness of their responses in educational scenarios. As a result, many works continue to also embed task-specific models, e.g., BERT for text classification, in their writing assistants [71]. However, previous works comparing and contrasting both approaches (generative and task-specific models), specifically in the task of argumentative writing, are rare in the literature.

Moreover, to the best of our knowledge, there is a lack of previous work on designing and implementing argumentation learning writing assistants, based on the *learning from errors* paradigm, as well as investigating how more recent NLP models (e.g., GPT) are able to be effective in teaching argumentation skills to learners, compared to the classifier-based approaches with task-specific models done in previous works [135, 139, 140]. While previous works have explored using GPT in argumentation learning environments [121], the embedding of GPT in a writing assistant based on our design requirements and the *learning from errors* paradigm, as well as evaluating the learning gains of such system in comparison with the more traditional approach of classifier models, is rare in the literature. Thus, we aim to investigate the possibilities and effects of a domain-independent argumentation writing assistant helping students develop their argumentation skills.

3 DESIGN OF ALURE

In this research work, we build our intelligent writing assistant called ALure, which allows us to provide learners with adaptive guidance in their individual argumentation learning progress and to improve their argumentative skills, in light of the *learning from errors* paradigm. We specifically consider our target group as students who are currently in training to become employees in work environments later on as instructors. As a result, we consider *designing for learners* in the education sector as our principle throughout the design and implementation process of ALure.

3.1 Design Requirements for ALure

To deduce design requirements for ALure, we use a two-fold approach: On the one hand, we derive requirements from a comprehensive literature review, following the theory-driven approach as defined by Briggs [13]. Following this approach, we believe that using corrective feedback and the *learning from errors* paradigm in our writing assistant can enable students to improve their argumentation skills. On the other hand, we conducted a learner-centered requirement analysis using requirement analyses performed by a total of 43 business students (26 female and 17 male;

average age = 21.02, SD = 1.30) attending an introductory course in business informatics at a Western European university[7]. We asked the students to derive requirements for a digital argumentative learning tool in terms of user, use, and utility principles [12]. We first described the study goal to the students and asked them to provide opinions on their self-proclaimed requirements for an argumentative writing assistant. They were asked to reflect and think about how such a tool would fit in their educational curricula, how they can best make use of such a tool, what benefits the tool can bring to them in their argumentation learning process, and what features the tool should have to better satisfy their argumentation learning goal. We did not take the requirements from the student responses as-is; in fact, we utilized this student input process as a form of "understanding users" in terms of qualitative research [20]. In doing so, we followed previous works on designing argumentation writing assistants [138, 140] that also derived part of the design requirements of their systems from the students. Two learning sciences researchers among the authors of this paper (one an expert in teaching argumentative writing, and the other both in argumentation and in machine learning research) read the student responses carefully together in a workshop and clustered them into user requirements using a semi-deductive[8] thematic analysis, fixing any disagreements as they happened. They distilled seven requirements in total, provided in Table 1.

As seen in Table 1, R1 addresses the use of errors for different learning tactics (similar to what was described in Section 2.2). First, the setting needs to provide the learner with the opportunity to make mistakes. Building upon this, the learner should be *allowed* to make mistakes, which is translated into R2. This is informed by Metcalfe [85], who mentions "if the goal is optimal performance in high-stakes situations, it may be worthwhile to allow and even encourage students to commit and correct errors in low-stakes learning situations, rather than assiduously avoiding errors at all costs." As a result, mistakes can also be used to dynamically provide the learner with adaptive strategies, as a part of the general *learning from errors* paradigm. The next requirements, R3 and R4, address the idea that providing the learner with adequate theoretical material about argumentation and argumentative errors improves the student's ability to adhere to desired learning paths. Also, R5 deals with the use of adaptive guidance to reduce the extraneous load on the learner. Requirements analysis with student analyses revealed that learners are also aware of the risk of frustration when working with their errors. These aspects thus focus on bringing attention to the positive aspects of their argumentation. As this is in line with previous research on how to decrease frustration, we draw one requirement from the qualitative data collection from students (R6).

Further analysis showed that learners wish to observe their learning progress. According to research from cognitive load theory, an important factor in reducing extraneous overload is to continuously present the learner with their learning goal and progress, as uncertainty about the performance often leads to increased frustration in the learner and thus can negatively impact the learning performance [112, 124]. Providing the learner with information about their progress, their goal, and information on how to achieve this goal helps the learner to stay motivated and engaged [32, 112]. Therefore, as this student requirement is in line with previous research, we formulate a corresponding requirement (R7).

We specifically note that our extracted requirements are considered as standards in writing support assistants, and have been mostly used previously to support learners (e.g., [147] for *learning from errors* requirements, or [137, 143] for *cognitive load theory* requirements, among others). However, we consider our work among the first to investigate how our requirements apply

[7]All students provided consent for participating in our research.

[8]The researchers had the background work in mind (specifically, the research cited in Table 1), but were open to see student responses not matching any specific previously-investigated work.

Category	Requirements	Source
Learning from Errors	R1) The system should identify types of argumentative errors.	Learner-inspired Design ∗
	R2) The tool should provide strategies on how to correct argumentative errors.	Learner-inspired Design Metcalfe [85]
Argumentation Theory	R3) The tool should give reference points for good argumentations.	Shute [112]
	R4) The tool should give background knowledge about the anatomy of an argument.	Moreno [87]
Cognitive Load Theory	R5) The tool should provide guidance on how to resolve argumentative errors.	Sweller and Cooper [125]
	R6) The tool should give positive, encouraging guidance.	Learner-inspired Design
Learning Progress	R7) The system should track learner progress and visualize various elements of the learning progress.	Learner-inspired Design Shute [112]

Table 1. Our collected design requirements for ALure. ∗ : By which, we mean the insights extracted from students in the qualitative data collection process.

differently to our two different design and technical modalities of classification and generation models, in the context of argumentative writing.

3.2 Interface Design of ALure

Based on our collected design requirements (see Table 1) and taking inspiration from the design science research literature (e.g., [38]), we deduced design principles for ALure. Two researchers with multiple years of expertise in designing user-centric systems for argumentation skills, who were among the authors of this paper, critically investigated the design requirements and introduced design principles by grouping them into five main themes, spanning across the four categories of *learning from errors*, *argumentation theory*, *cognitive load theory*, and *learning progress*. Table 2 presents the formulated design principles with the corresponding requirements. Specifically, the design principles in Table 2 were extracted as below:

- **DP1 (Active Learning):** Metcalfe [85] mentions the positive link between engagement with errors and exploratory active learning. Other works also have investigated the role of *learning from errors* in active learning of the students [73]. As a result, we formulated DP1 based on the requirements R1 and R2 in the *learning from errors* category. This design principle refers to supporting students with instruction to correct their errors, thus increasing their argumentation quality.
- **DP2 (Argumentation Anatomy):** This design principle, being based on requirements R3 and R4, refers to supporting learners by integrating explanations of argument anatomy and fallacies (to see the full explanations we used, refer to the prompts in the appendix). It has been classified into the *argumentation theory* category, in light of previous works on argumentation theory advising users on how to follow the correct argumentation structures [34, 35].
- **DP3 (Domain Knowledge):** This design principle refers to the additional learning material provided to the students (see the prompts in the appendix for the explanations we used in our tool) and is based on requirements R3, R4, and R5. It has also been classified into the

argumentation theory category, with regards to previous argumentation theory research highlighting the positive role of additional learning material[9] [29, 42].

- **DP4 (Ensure High Motivation):** This design principle, based on requirement R6, mentions highlighting not only the negative (e.g., errors) but also the positive (e.g., the identified well-connected premises and claims) elements of the argumentative text written by the learner. It has been classified into the *cognitive load theory* category, as previous works have explored correlations between lower cognitive load and higher engagement and motivation [28, 30, 82].
- **DP5 (Learning Progress):** This design principle, based on requirement R7, refers to providing visual elements such that students can be supported in their learning progress. It has been classified into a single-member category with the same name. Specifically, prior research works have investigated the role of visual and/or multimedia elements in helping students monitor learning progress [72, 112], serving as a motivation for formulating this design principle.

Category	Design Principle	Description	Addressed Requirements
Learning from Errors	**DP1:** Active Learning	To increase learners' argumentation quality, support them with instruction on how to correct argumentative errors.	R1, R2
Argumentation Theory	**DP2:** Argumentation Anatomy	To support learners in their understanding of a high-quality argumentation, integrate explanations of an argument's anatomy and any types of argumentative fallacies.	R3, R4
	DP3: Domain Knowledge	To support learners in the understanding of the relevant domain knowledge, provide further learning material.	R3, R4, R5
Cognitive Load Theory	**DP4:** Ensure High Motivation	To avoid frustration, highlight the positive elements of the argumentative learner text.	R6
Learning Progress	**DP5:** Learning Progress	To support learners with guidance on their learning progress, provide visual elements for the learning progress.	R7

Table 2. Design principles for our intelligent argumentative learning tool ALure.

We implemented ALure based on the design principles extracted and presented above. Specifically, we implemented two versions of ALure (see Figure 1 for two screenshots showing our two versions):

- **ALure A:** This version of ALure uses a BERT text classifier model in the back-end, finding claims, premises, and argumentative fallacies regarding missing premises (the technical

[9]While this design principle is classified in the *argumentation theory* category, it also addresses requirement R5 which was classified in the *cognitive load theory* category in Table 1. This is because while the guidance to resolve errors can help students reduce their cognitive load, they can also be considered as part of the general framework of the guidance provided to the user based on argumentation theory.

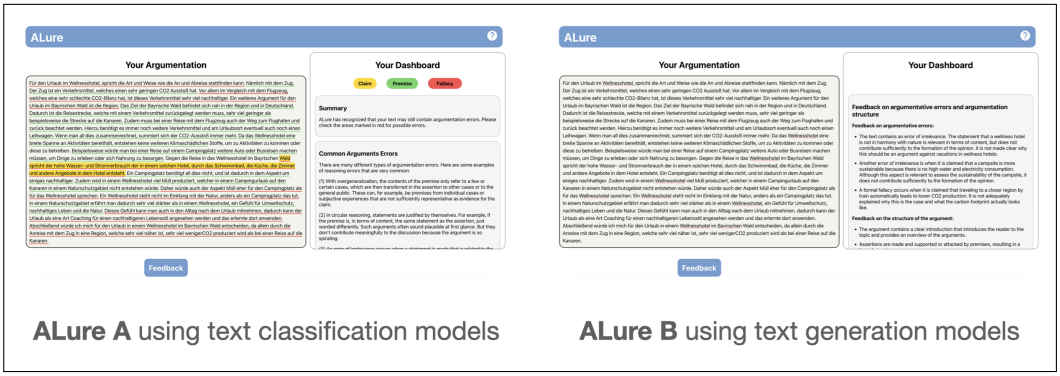

Fig. 1. The two different versions of ALure. ALure A works with our fine-tuned BERT classification models in the back-end, color-coding the claims, premises, and errors in the input text. ALure B works with the GPT-3.5 API provided by OpenAI along with our prompts, and supports learners by means of providing holistic feedback on the input text in a textual form. The screenshots are translated from German to English for presentation in this paper.

details are described in the subsequent sections). Based on the outputs of the classifier model, we mark sentences as being a premise (green), a claim having a supporting premise (yellow), or a claim *without* any supporting premise (red), based on **DP4** and **DP5**. To satisfy the requirements of the *learning from errors* paradigm, we also show a summary based on the errors (red) identified in the text, as well as providing instructions (i.e., *corrective feedback*) to correct argumentative errors (**DP1**), textual explanations (**DP2**), and further learning material for the user (**DP3**). A screenshot of ALure A can be seen on the left side of Figure 1.

- **ALure B:** This version of ALure uses a GPT-3.5 text generation model in the back-end, asking GPT to provide feedback concerning our design principles (**DP1**, **DP2**, and **DP4**) as well as the text provided by the learner in the tool (the technical details are described in the subsequent sections). Due to GPT being a generative model, we do not color-code individual sentences (**DP5**) in ALure B, but rather report the general feedback provided by the GPT-3.5 Application Programming Interface (API) from OpenAI. However, the principles of the *learning from errors* paradigm are formulated in the prompt itself, causing GPT to respond with identified errors in the student text, accompanied by corrective feedback. We include further learning material as well (**DP3**), similar to ALure B. A screenshot of ALure B can be seen on the right side of Figure 1.

Building on the insights from previous works mentioned in Section 2.4 and to find which types of models can be useful in future argumentative writing assistants using the *learning from errors* paradigm, we designed ALure in two versions to compare the performance and effectiveness of BERT vs GPT (i.e., a fine-tuned task-specific smaller classifier versus a general domain-independent text generation LLM). We specifically used BERT as the basis for our text classification model in ALure A, as 1) the model was lightweight enough to run on our available resources, 2) it was fast enough to fine-tune on our data and iterate on the model, and 3) it has been used and shown to be practical in many prior works on writing assistants (e.g., [111, 138, 147]).

3.3 Implementation of ALure

We implemented ALure as a web application, supported to be used on desktop browsers. We implemented our front-end with React and our back-end with the Flask framework using Python language.

3.3.1 ALure A. To implement ALure A using the classification approach, we needed data to train the models. Thus, we aimed to collect field data to train our models to classify premises, claims, and their relation. The data for claims and premises allows us to identify the argumentative components in students' texts, while the relation data can allow us to identify missing relations and provide them to the users in the lights of the *learning from errors* paradigm. The data was collected in two subsequent educational semesters from 346 students (253 females, 86 males, 1 non-binary, and 6 non-specified; average age = 23.31, SD = 2.77) in the fields of *elementary school education* (141 students; 118 females, 19 males, and 4 non-specified; average age = 22.61, SD = 2.53), *special education* (137 students; 102 females, 32 males, 1 non-binary, and 2 non-specified; average age = 23.70, SD = 2.67), and *economics* (68 students; 33 females and 35 males; average age = 23.99, SD = 3.14) at two Western European universities[10]. The users had to discuss two positions of a debate (given by means of scenarios) in a coherent text by means of pro and contra arguments in order to finally decide on one of the two positions. In total, there were five such scenarios on different topics: two educational topics, two on sustainability, and one on a topic related to economics.

Three individuals annotated the data, identified the premises and the claims, and also specified if any given premise supports a claim in a premise-claim pair, using the Tagtog tool. The annotations were based on a rigorous annotation guideline which we carefully defined. The guideline includes detailed definitions and examples of claims, premises, and their relations, defining the underlying schema, based on the work of Toulmin et al. [128] and following the guidelines provided by Wambsganss et al. [141] as well as Stab and Gurevych [117] as the methodology for our annotation. In total, after filtering all entries with less than five characters, our dataset included 5215 sentences: 2351 sentences were labeled as premise, 1871 as claim, and 993 as none. Each entry (i.e., sentence) on average included 18.34 words (SD = 10.26), and 126.69 characters (SD = 71.17). Thus, the corpus had a total of 95645 words and 660705 characters. For the premise-claim relations, the corpus included 1480 non-related and 2062 related pairs.

To find the reliability of the annotations by the three annotators, we calculate Krippendorff's α and Fleiss' κ on the annotations of the claims and premises, as well as the claim-premise relations. Regarding the annotation of the claims and premises, we found a Krippendorff's α of 0.77 and a Fleiss' κ of 0.77 (suggesting a *Substantial* agreement). Regarding the annotation of the claim-premise relations, we found a Krippendorff's α of 0.40 and a Fleiss' κ of 0.39 (suggesting a *Fair* agreement) [66]. While the agreements for the claim-premise *relations* are not high, they are generally in line with values reported for *relation annotation* by other works in ML-based writing support assistants (e.g., Fleiss' κ of 0.3979 for claim-premise relation annotation reported by Weber et al. [146]).

We then took the annotated data, pre-processed it using Spacy[11], and fine-tuned two BERT models [23] based on the data using one epoch of training for each model. We used a warmup ratio of 0.06, a maximum sequence length of 128, a batch size of 8, and a learning rate of 4e-5. The descriptions of the two models are as follows:

- **Premise or Claim Classifier:** This is a three-class model which classifies any input sentence as being a premise, claim, or none of them. We use the output of this model for both identifying

[10]We collected the data from a diverse set of sources to make it usable and beneficial for evaluating the potential of providing feedback across a diverse set of distinct domains. In the current study, we only focus on one specific downstream task and leave the rest for future work.
[11]https://spacy.io

and color-coding claims and premises in ALure A, as well as to provide them as inputs to the next model.

- **Premise-claim Relation Classifier:** This is a two-class (binary) model that gets a claim and a premise, classified using the previous model, as inputs and classifies if the premise supports the claim or not. We use the output of this model in ALure A to find unsupported claims and mark them in the interface with the red color. The red underlines in the text provide the identified errors to the learners in the framework of the *learning from errors* paradigm.

The precision, recall, and F1 score / accuracy for each class in the two models are provided in Table 3.

Model Name	Model Type	Class	Precision	Recall	F1 Score / Accuracy (for binary)
Premise or Claim Classifier	3-class BERT	Premise	0.65	0.76	0.70
		Claim	0.64	0.56	0.60
		None	0.85	0.65	0.73
Premise-claim Relation Classifier	Binary BERT	-	0.93	0.93	0.93

Table 3. The precision, recall, and F1 score (accuracy for the binary model) for each of the models we fine-tuned and embedded in ALure A.

3.3.2 ALure B. On the other hand, to implement ALure B using the text generation approach, we used the API provided by OpenAI to access GPT-3.5 (which is also the model behind the initial version of ChatGPT [114]). Prompt engineering has been shown to be effective for utilizing the knowledge encoded in the parameters of domain-independent LLMs for a variety of downstream tasks [17, 76, 104]. As a result, we chose to design a prompt following the design principles as mentioned in Table 2; specifically[12]: prompting GPT to give useful feedback and instructions (DP1)[13], integrating explanations of the anatomy of argumentation and fallacies (DP2)[14], and also mentioning the positive aspects (rather than only feedback *against* the writing) (DP4)[15]. Specifically, the paradigm of *learning from errors* was explicitly included in the prompt design to ensure GPT outputs the learner errors as well as the relevant corrective feedback[16]. Additionally, following prior works using prompt engineering [17, 86], GPT was instructed at the beginning of the prompt to act as an *expert* in our task (i.e., writing argumentative texts) to achieve reasonable performance[17]. The prompt also included instructions on outputting text following certain structures (e.g., HTML tags) for better appearance when being shown in the front-end of ALure[18].

[12]DP3 was not included in the prompt; rather, we included extra material in a separate area in our tool. Also, we did not utilize DP5 in ALure B, the effects of which are described in the Discussion section.

[13]For example, "*Give feedback on argumentative errors in the text*", "*In your feedback, focus on the ability to persuade, the structure of the argument in terms of claims and premises, possible argumentative errors, and not on the content*", etc.

[14]For example, "*Typically, a new argument begins with an assertion. This claim must be supported or attacked below, using premises for this purpose*", "*A strong argumentative text ends the discussion with a summary. All arguments made are presented again to give the reader a final overview.*", etc.

[15]For example, "*Also give three reasons for and against the structure of the argument.*", etc.

[16]For example, "*You are also trained to recognize argumentative errors*", "*Give feedback on argumentative errors in the text*", etc.

[17]"*Imagine you are an expert at writing argumentative texts.*"

[18]For example, "*Format the feedback as an unsorted HTML list*", etc.

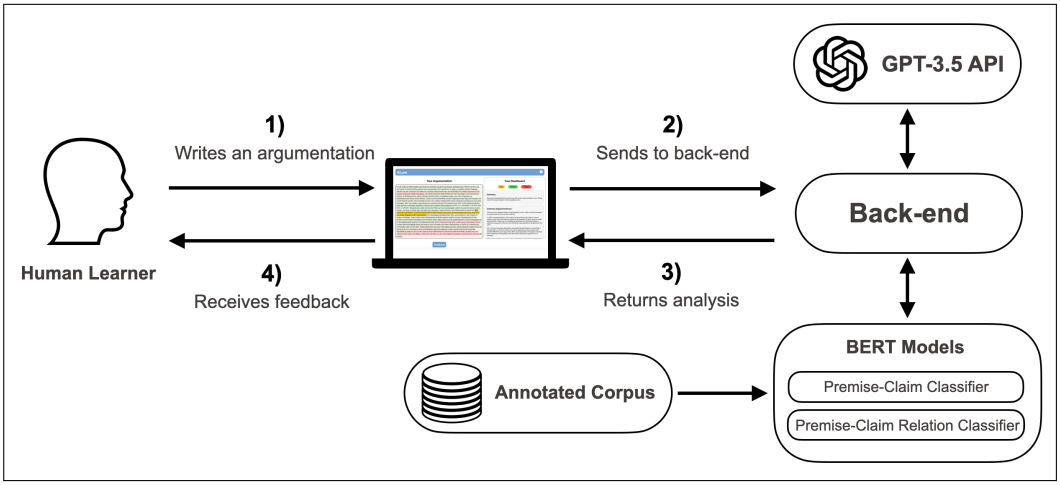

Fig. 2. The architectural flow of ALure. The learner begins inputting their argumentative text into ALure. The text is sent to the back-end, then forwarded either to our BERT models fine-tuned on our collected corpus or to GPT-3.5 API by OpenAI, depending on the version of ALure (A and B, respectively) the learner is using. The results from the back-end are presented to the learner in the user interface of ALure, including the errors made by the learners and the corresponding corrective feedback as a part of the *learning from errors* paradigm.

To design the prompt, two authors of this paper (one being an expert in argumentative writing instruction, and the other both in argumentation and in machine learning research) collaboratively designed the prompt in the German language during a workshop together. They used the OpenAI Playground to try a set of different prompts and evaluated the outputs from GPT critically. Also, the annotation guideline for the implementation of ALure A was utilized as an inspiration for setting up the prompt. They then agreed on a final prompt to be included in ALure B. The final prompt was applied to 20 input texts randomly selected from our dataset, and the two authors evaluated the output given by GPT-3.5 and considered it valid and correct. Thus, we confirmed that our prompt led to relevant suggestions for improvements and feedback on the argumentative texts submitted by the user. The full prompt text can be seen in the appendix.

The architectural flow of how ALure works can be seen in Figure 2.

4 EXPERIMENTAL EVALUATION

In this section, we plan to find the answer to our research question (as presented in Section 1) by conducting a study involving ALure. This enables us to see how the two different versions of ALure provide effects differently in terms of the argumentation learning gains of students. We first present the demographics of the participants and then the details of our study.

4.1 Participants

We conducted our experiment in a large-scale lecture on 305 students in a Western European university, split into three groups for ALure A, ALure B, and the control group. All students were pre-service teachers. All of the students participated voluntarily in our research. Originally, 348 students participated in our study, but a total of 43 participants were removed from further analyses (after the experiment ended and before starting to analyze data), out of which 20 did not participate in the post-test, and 23 did not pass the treatment check. The purpose of the treatment check was to verify whether the intervention was carried out by the participants. For this purpose, 7 people

who did not submit all 4 of their argumentative texts during the intervention phase, as well as 16 people who answered "0" at least once when asked how often they got feedback in ALure, were excluded. Thus, 305 participants were included in our sample. The students had a mean age of 23.32 (SD = 3.19). Out of all the 305 students, 50 identified as male, and 255 identified as female. The courses in which the experiment was conducted were mandatory to attend, but all students participated in this research voluntarily and could withdraw from participation until the end of the study. The data were anonymized before analysis, and the conduct of the study was approved by the ethics committee. The participants were pedagogy students enrolled in a mandatory course at a university and were assigned randomly into one of the three groups:

- **First treatment group (*ALure A*)**: Comprised of 81 students (17 males, 64 females); mean age 23.76 (SD = 3.41). This group employed a BERT-based classifier model to aid in the argumentative writing process.
- **Second treatment group (*ALure B*)**: Comprised of 86 students (12 males, 74 females); mean age 24.23 (SD = 3.57). This group used the GPT-3.5 generative model to aid in the argumentative writing process.
- **Control group (*Video*)**: Comprised of 138 students (21 males, 117 females); mean age 22.57 (SD = 2.58). Participants in this group viewed an instructional video prepared by the course instructors, covering key aspects of argumentative structure and fallacies. We specifically used a *video watching* approach for the control group, instead of, e.g., another writing assistant, as it has been the standard way of how students learned argumentation in the curriculum of this course in our university over multiple semesters. The video and its contents were provided by the course instructors collectively and refined over multiple offerings of the course over subsequent semesters. After watching the video, the students were asked to mark their arguments themselves when writing a text without any help of an argumentative writing support tool (i.e., no direct *learning from errors* paradigm support using any tool happened in the control group).

4.2 Study Procedure

The study was designed to unfold over a six-week period, incorporating a pre-test, an intervention phase, and a post-test for evaluating the efficacy of the ALure tool in enhancing argumentative skills. Each participant was required to write four argumentative essays over the course of four weeks (Weeks 2-5).

(1) **Week 1 - Pre-test:** All participants undertook a pre-test based on the argumentation exam developed by Berkle et al. [8]. This exam gauged the initial argumentative writing skills of the participants, focusing on structure and fallacies.
(2) **Weeks 2-5 - Intervention:** During these weeks, participants were mandated to interact at least four times with their designated tool (either ALure A, ALure B, or the Video lecture on "Arguing convincingly according to Toulmin" which was also used in the previous offerings of the course). Participants were tasked with writing one essay each week (two essays on the topics of pedagogy and two essays on sustainability after the four weeks).
   - Students using ALure A and ALure B received adaptive feedback based on their argumentative structure and fallacies and were able to revise their text accordingly. For the control group, after writing their initial text for each scenario, they were tasked to identify all claims and premises in their written text. Subsequently, they revised their original text, too. However, the control group did not receive any adaptive feedback.
(3) **Week 6 - Post-test:** In the final week, a post-test identical to the pre-test was administered to measure improvements in argumentative skills.

4.3 Measurement

(1) **Pre-Test and Post-Test:** The test instrument adapted from Berkle et al. [8] was used as a pre and post-test for argumentation skill development. This test was divided into two sections:
   - *Fallacy Recognition*: Featured 17 items designed to assess the ability to identify fallacious arguments.
   - *Argument Structure Recognition*: Included 13 items aimed at understanding the participants' grasp of proper argumentative structures.

   To mitigate bias, the argument structure section was presented after the fallacy recognition part. Items were organized into blocks representing three different subject domains (pedagogy, economics, and sustainability), which were randomized to avoid sequence effects.

(2) **Sociodemographic Questionnaire:** Following the post-test, a questionnaire was administered to collect sociodemographic data. However, the main focus of this study remains on the pre-test and post-test scores.

4.4 Analysis

For statistical data analysis, we used the program R version 4.1.0 and conducted a mixed between-within Analysis of Variance (ANOVA) as this type of ANOVA is used to analyze the effects of independent variable(s) (groups: *ALure A*, *ALure B*, and *Video*) on a dependent variable, while taking into consideration the potential influence of repeated measurements (pre-test and post-test) on the same subjects. With this method, we can investigate if the three groups differ in their Argument Structure or Fallacy Scores (*main effect*) and if the three groups differ based on their mean Argument Structure or Fallacy Scores change over the two measurement points (*interaction effect*).

5 RESULTS

Internal consistency for the questions regarding the *argument structures* in the test [8] was *high*, with Cronbach's $\alpha = 0.826$ in the pre-test and Cronbach's $\alpha = 0.853$ in the post-test. Internal consistency for the *fallacy* score was considered *acceptable*, with Cronbach's $\alpha = 0.702$ in the pre-test and Cronbach's $\alpha = 0.728$ in the post-test. We assume normality, enabling us to conduct ANOVA analysis, due to the large size of the groups.

5.1 Fallacies

Concerning the fallacy scores from the tests, ANOVA revealed a statistically significant interaction between the effects of the group (ALure A, ALure B, Video) and the two measurement points (pre-test and post-test) ($F(2, 302) = 5.09$, $p < 0.01$). The three groups differ in how their mean fallacy scores change between the two measurement points. Simple main effects analysis showed that the groups did not have a statistically significant effect on the fallacy scores ($F(2, 302) = 0.49$, $p = 0.61$). However, simple main effects analysis showed that measurement points did have a statistically significant effect on fallacy scores ($F(2, 302) = 10.81$, $p < 0.01$).

To have a closer look at the results, post-hoc analyses were conducted. The one-way ANOVAs for the pre-test ($F(2, 302) = 0.55$, $p = 0.58$) and for the post-test ($F(2, 302) = 0.23$, $p = 0.10$) were not significant. The three groups did not differ in their fallacy scores, neither in the pre-test nor in the post-test. Post-hoc Tukey tests showed that the two measurement points differed significantly at $p < 0.01$, but the three pairwise comparisons of the groups did not differ significantly (ALure A - ALure B, ALure A – Video, ALure B – Video: $p > 0.05$). Furthermore, pairwise comparisons showed a significant difference in the ALure B group from the pre-test to the post-test ($p < 0.01$). Also, the mean score for the groups ALure A and ALure B increased from the pre-test to the post-test, with the most increase belonging to ALure B (+34.62%), then ALure A (+19.06%). On the other hand, the

mean score in the video group slightly decreased from the pre-test to the post-test (-3.11%). The comparison of the learning gains among groups for the fallacy scores can be seen in Table 4.

5.2 Argument Structure

ANOVA revealed that there was no statistically significant interaction between the effects of the group (ALure A, ALure B, and Video) and the two measurement points (pre-test and post-test) ($F(2, 302) = 0.58$, $p = 0.56$). The three groups did not differ in how their mean argument structure scores changed between the two measurement points. Simple main effects analysis showed that the groups did not have a statistically significant effect on the argument structure scores ($F(2, 302) = 1.06$, $p = 0.35$). However, simple main effects analysis showed that measurement points again had a statistically significant effect on argument structure scores ($F(2, 302) = 39.28$, $p < 0.001$).

To have a closer look at the results, post-hoc analyses were conducted. The one-way ANOVAs for the pre-test ($F(2, 302) = 1.64$, $p = 0.20$) and for the post-test ($F(2, 302) = 0.52$, $p = 0.60$) were not significant. The three groups did not differ in their argument structure scores, neither in the pre-test nor in the post-test. Post-hoc Tukey tests showed that the two measurement points differed significantly at $p < 0.01$, but the three pairwise comparisons of the groups did not differ significantly (ALure A - ALure B, ALure A – Video, ALure B – Video: $p > 0.05$). Furthermore, pairwise comparisons showed significant increases in the Video ($p < 0.001$) and the ALure B ($p = 0.002$) groups from the pre-test to the post-test. Also, the mean score for all three groups (ALure A, ALure B, and Video) increased from the pre-test to the post-test, with the most increase belonging to ALure B (+38.26%), then Video closely following it (+33.39%), and ALure A coming at the third position (+19.75%). The comparison of the learning gains among groups for the argument structure scores can be seen in Table 4.

We also evaluated the *total* learning gain, concerning the sum of the scores in the two metrics investigated by the test we used (fallacy and argument structure) and report the results in Table 4. Based on the sum of the scores, the most learning gain belongs to ALure B (+36.79%), then Video (+20.20%), and finally, ALure A closely following (+19.52%).

The plots comparing different study groups across pre-test and post-test can be seen in Figure 3.

Test Item	Group	Pre-test		Post-test		Learning Gain
		Mean	SD	Mean	SD	
Fallacies	ALure A	3.20	2.90	3.81	2.98	+19.06%
	ALure B	2.86	2.70	3.85	3.03	+34.62%
	Video	3.22	2.48	3.12	2.70	-3.11%
Argument Structure	ALure A	6.43	5.48	7.70	6.11	+19.75%
	ALure B	4.94	5.00	6.83	6.24	+38.26%
	Video	5.69	5.42	7.59	6.29	+33.39%
Sum	ALure A	9.63	6.64	11.51	7.56	+19.52%
	ALure B	7.80	6.63	10.67	8.17	+36.79%
	Video	8.91	6.60	10.71	7.48	+20.20%

Table 4. The results from our experiment, comparing the argumentation learning gains of the participants in ALure A, ALure B, and the Video control group from the pre-test to the post-test.

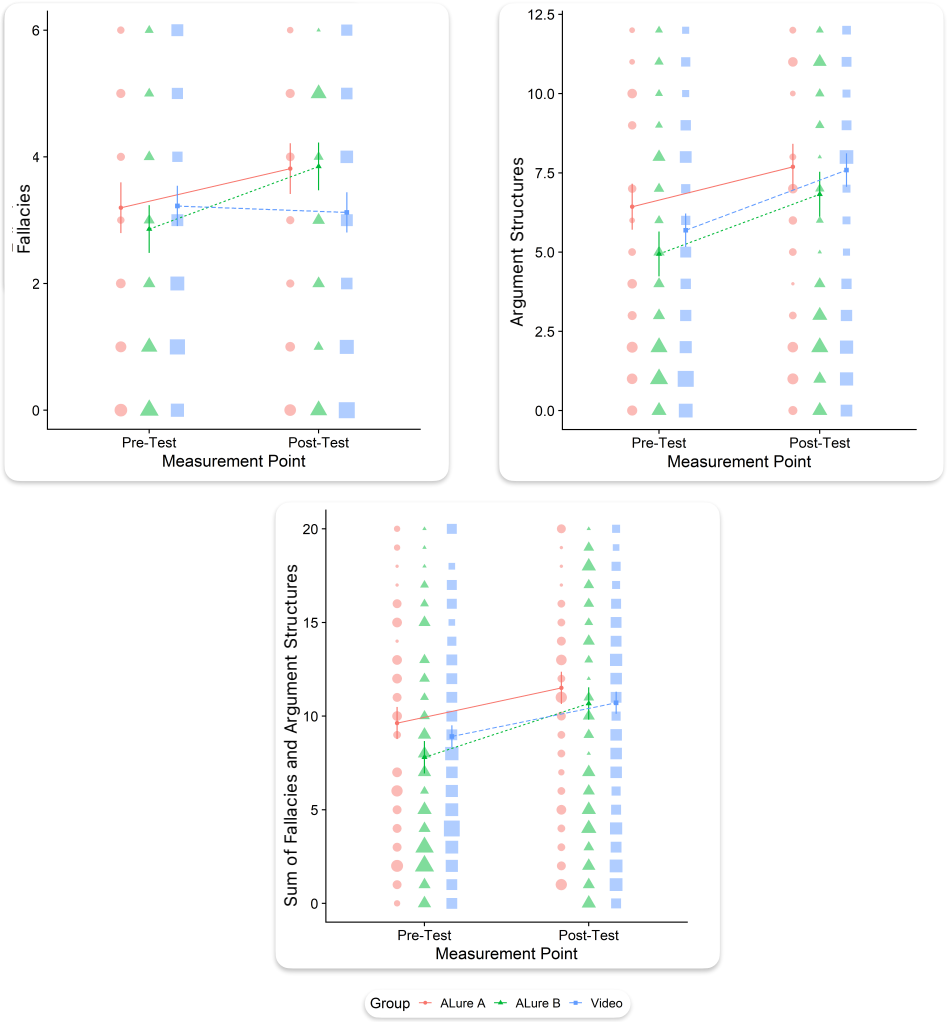

Fig. 3. The plots comparing our different study groups across pre-test and post-test. Our results show overall learning gains when either of the two approaches of text classification or text generation are used.

6 DISCUSSION

In our research, we investigated how we can design a cooperative and collaborative digital learning tool, incorporating the *learning from errors* paradigm, to instruct students on their argumentation skills. The aim of our study was to explore the effects of a domain-independent writing assistant for argumentation learning on helping students gain argumentation skills. Thus, we extended upon prior works on providing formative feedback to students (e.g., [59]) by using the *learning from errors* paradigm to guide us as a theory principle for providing argumentation feedback to learners. With regards to the design of the backbone of our system to help students in argumentation,

we observed that LLMs for text generation can be viable alternatives, as they come with high zero-shot performances, natural responses, and less need for human annotation. However, their low controllability and hallucination could possibly harm student learning in our educational scenario of argumentative writing. As such, we aimed to explore both approaches of 1) preparing and annotating a corpus that builds from different domains, such as business, pedagogy, and sustainability, according to our rigorous annotation guideline, and 2) prompting LLMs with a few-shot approach for providing holistic feedback on the argumentative texts provided by the learners.

In order to do so, we collected our design requirements for designing ALure, our user-centric web application as our argumentation learning writing assistant. We compiled design principles out of the requirements and used them in lights of the *learning from errors* paradigm to implement ALure in two versions: one with a classifier BERT model and the other using the GPT-3.5 model. The evaluation of the two versions of ALure as well as a control group using instructor-prepared standard videos also used in previous course offerings over multiple semesters as their method of learning argumentative writing in a large-scale lecture study shows learning gains in all of the groups. Thus, we provide promising results on using intelligent argumentation writing assistants as alternatives to traditional methods of argumentation learning instruction (e.g. instructor-designed video lectures) in real-world environments when using *errors* with corrective feedback as a means for student support with minimal cognitive load.

We found generally close learning gains obtained by both approaches of text classification (indicating premises, claims, and errors in the text, along with static feedback on the text) and text generation (providing general dynamically generated feedback on the argument structure and fallacies), showing the benefits of incorporating our design principles and the *learning from errors* paradigm (see Table 2) in our design. We specifically observed how incorporating active learning (**DP1**), constructive feedback and explanations (**DP2** and **DP4**), and further learning material (**DP3**) led to learning gains in both of our study groups. Thus, we consider our work to be beneficial to future research on designing collaborative argumentation writing assistants, as an evaluation of the impact of the different model modalities on learners, in light of the design principles we provide and the paradigm of *learning from errors*.

Although the difference between the two groups was not significant, ALure B (the group using generative models) had a higher learning gain for both fallacies and argument structure. Moreover, ALure B led to a significant difference between scores in both fallacies and argument structure scores from pre- to post-test, as opposed to ALure A. This is while ALure B did not employ visual indicators in the text for feedback (see Section 3.2 and **DP5** in Table 2), and rather presented feedback only in the format of text returned by the GPT-3.5 API. This suggests the helpfulness and relevancy of adaptively generated feedback on learning, as opposed to conducting inference using smaller pre-trained categorical models, even when visual cues are rare or not present. Similar to the findings of previous intelligent assistants using LLMs [48, 52, 56, 71], we believe the better results in ALure B come from the higher versatility and expression power of the responses from text generation model, prompting the students to reflect on their argumentative writing. This is in line with prior works showing a higher versatility and more naturalness in the responses from text generation models, such as GPT [4, 56, 70, 103, 148]. In addition to the higher naturalness and versatility of the responses, GPT-based models commonly outperform the prior models in reasoning tasks [75], suggesting that the outputs returned by GPT are more factual and more relevant to the text. Prior works have shown that GPT is also useful in generating relevant and to-the-point error descriptions in assistants providing intelligent feedback to students [113] as well as in being able to retrieve information from input texts with the potential to be used in educational scenarios [21]. This further supports the additional benefit of using GPT we observed

in argumentative writing with ALure. On the other hand, for providing feedback, ALure A sticks to pre-defined texts and highlights in the student writing (due to the lack of a text *generation* model outputting content that the students can easily connect themselves with). The pre-defined texts can get repetitive or unnatural for the student over multiple interactions (i.e., over subsequent weeks). This is while the corrective feedback in the *learning from errors* paradigm should provide new information to the students at each time to increase their learning gains; repetitive feedback and responses with limited amounts of personalization have been known to limit the positive effects on the students' learning [45, 81]. Thus, the pre-defined limited-by-nature set of possible feedback modalities, provided by ALure A, would not be as effective as the versatile endless set of possible feedback responses provided by ALure B, in terms of learning gains and post-test performance. As a summary, we suggest future researchers consider including generative AI-based models in their writing assistants in order to facilitate the experiences of the students using the tool, as well as providing more relevant suggestions and error descriptions to the users.

With that said, implementing AI-based models in educational settings and writing assistants comes with its own set of challenges. As noted by Lee et al. [71] in their design space, there have been recent efforts in the domain of writing assistants to consider the possible risks of AI models and LLMs (e.g., GPT) regarding the factuality of the responses, biases found in the model outputs, and the adherence of the assistants to ethical standards and social norms [46, 144]. As a result, while both AI-based approaches turned out to be helpful for argumentative writing in ALure based on our results, and no particular harmful outcome was raised by students during the course of using ALure, the effects of our models on the more general argumentative writing ecosystem and conventions [71] when implemented in real-world scenarios necessitates further larger-scale exploration.

In summary, our results also shed light on the supporting role of incorporating LLMs into educational scenarios (as also discussed in previous works, e.g., [55, 77, 92, 119]), for the domain of argumentative writing.

6.1 Theoretical and Practical Contributions

The literature in CSCW has long advocated for the applications of computer-based intelligent tools to mediate and help with *social structures* [79]. As a result, due to the importance of successful argumentation in collaborative work environments, computer systems helping with this process can be proven useful among students. Thus, multiple previous CSCW research works have explored argumentative and persuasive writing support using computer-based systems [136, 150, 152]. Moreover, the literature on CSCL provides the paradigm of *learning from errors* which is explored rarely in previous CSCW research on argumentation support . Additionally, the effects of generating holistic feedback on the argumentative text by means of more recent NLP models, such as GPT, compared to the classification models, have not been thoroughly investigated and researched in previous works. More specifically, although GPT has been embedded in argumentation learning scenarios in previous works [121], comparing how GPT-enabled systems help students learn argumentation skills with the more traditional classifier-based approaches, by being embedded in a user-centric writing assistant with the *learning from errors* paradigm, is lacking in the literature.

Thus, this research makes several contributions to the literature on argumentation learning and collaborative intelligent writing assistants in the domain of HCI and CSCW. First, we collect design requirements and extract design principles for our intelligent tool ALure to help learners in their argumentation skills using the *learning from errors* paradigm from the CSCL literature. The extracted design principles can be used as a basis for future research on designing and developing intelligent writing assistants for argumentation skills learning. Additionally, based on the design principles, we implement two different versions of our tool, one based on a classifier model (BERT)

and the other based on a generative model (GPT), and test them in a large-scale lecture experiment to find out the learning gains of the students. We compare the results of an argumentation test before and after learners use our tool, besides a control group watching instructional videos on argumentation learning. After analyzing the results obtained from our experiments, we investigate the amount of learning gains in all groups concerning argumentation structure and fallacies. We report the differences among groups and discuss the similarities and differences of using different NLP approaches and modalities to implement our collaborative writing assistant, along with how they provide benefits to learners in terms of argumentation learning gains.

As a result, in this work, we provided design principles for implementing AI-based models in argumentation support using learner errors and fallacies, and compared a GPT text generation model versus the integration of the predictions from a smaller task-specific BERT classifier into the interface of ALure. Our comparison showed the specific nuances of replacing more traditional classifier-based writing assistants with relatively domain-independent text generation models in their backbone in the domain of argumentation support. Thus, findings from our research have the potential to form the future direction of argumentative writing assistants, specifically in domains with less human-labeled data and limited usability of classification-based models.

6.2 Limitations and Future Work

However, our work also comes with several limitations as well as calls for future work. To begin with, we conducted our experiment in a single German-language Western European university with a limited number of students in a certain domain, thus we ask future researchers to also conduct our experiment in more diverse environments, in terms of domain, language, region, and the count of participants, to make our claims more generalizable. Also, we focused on only a single argumentation learning test as our pre- and post-test, as well as *video* tutoring for our control group (as opposed to alternatives such as hands-on projects), while future works can embed a more diverse range of tests measuring argumentation learning from different aspects and viewpoints or over a longer period of time, as well as also possibly considering other differences among groups (e.g. time taken to produce text).

For our classification-based version of ALure (ALure A), we fine-tuned BERT models. However, we call for future work to also use newer Transformer-based models and improve upon the precision and recall of the models compared with what we obtained, along with embedding them in a writing assistant based on our design principles and observing the differences in learning gains. Additionally, we hypothesize that with a better inter-rater agreement in the data annotation process in future studies (specifically for the premise-claim relations), the modeling performance will also increase, and thus the feedback provided to the students will be of a higher quality. On a related note, we propose *fine-tuning* text generation models (e.g., GPT) to possibly achieve higher-quality and more relevant argumentation feedback, to future researchers. Also, currently, while the LLM we used for ALure B is naturally domain-independent, a prompt engineering process based on domain knowledge was still necessary for ALure B to be able to provide relevant responses. We thus call for future work to explore the dimension of alleviating the need for domain experts to carefully design prompts, when integrating LLMs in argumentative writing assistants. We also call for future work to investigate the potential downsides of incorporating generative models, such as biases [57, 78, 144] or hallucinations [74, 88], in argumentative writing assistants.

Moreover, we have not investigated the *interaction* data of students interacting with ALure to explore how the *learning from errors* paradigm has helped students in their argumentative writing instruction. Unfortunately, capturing interaction data was not initially included in the agreement with the ethics board of our university for this study, and was not conveyed to the students when they were using ALure. As a result, we were not allowed to capture interaction data and more

technical variables. With that said, we believe that analyzing more variables of data (e.g., time taken to write, clickstreams, typing logs, the semantic and quantitative changes of the texts after each round of asking ALure for feedback, etc.) could be valuable for future work to better explore the nuanced effects of the *learning from errors* paradigm and our design principles on how students interact with our argumentative writing tool. Further, such an analysis would also provide novel insights into the meta-cognitive demands of rich feedback provided by LLMs during learning processes and how to improve the sense-making of learners from generative natural language text feedback [127].

7 CONCLUSION

Organizations increasingly need to support collaborative working environments, necessitating educational institutions to equip students with the necessary skill sets for their future professions. Specifically, *argumentative writing* is a skill positively associated with critical thinking. While previous CSCW works have explored providing argumentation support to learners in the form of intelligent writing assistants, recent CSCL research works also suggest the positive effects of the *learning from errors* paradigm on students' learning gains. In this work, continuing the research direction laid out by previous CSCW works in the domain of argumentative and persuasive writing, we designed and implemented ALure, our intelligent writing assistant, based on the *learning from errors* paradigm, to help students in their argumentation learning process. To do so, we collected design requirements from students and compiled our design principles for our argumentation learning writing assistant based on the requirements. We implemented two different versions of ALure, one using text classification (BERT) models for direct classification of claims, premises, and their relations, and the other using text generation (GPT) models for providing holistic feedback on the argumentative text. We compared the two versions, along with a control group watching an instructor-prepared argumentation learning instructional video, in a large-scale lecture with 305 students. We investigated the learning gains in all three groups in terms of argument structure and fallacies, and discussed the differences among groups. Our work sheds light on the potential of recent advances in CSCW, HCI, and NLP research (such as generative models) to help students learn argumentation skills and prepare them for the ever-changing collaborative working environments.

ACKNOWLEDGEMENTS

The results presented were partially developed in the research projects: Komp-HI funded by the German Federal Ministry of Education and Research (BMBF, grant 16DHBKI073), and Managing the Algorithm: Prompt Engineering for AI-based Systems as an Emerging Business Skill by the Swiss National Science Foundation (SNSF, grant number: 221281). We thank the BMBF and SNSF for supporting our research. Further, we thank Denise Löfflad for her input on earlier versions of this paper.

REFERENCES

[1] 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*. 1–13.

[2] Joseph P Allen, Robert C Pianta, Anne Gregory, Amori Yee Mikami, and Janetta Lun. 2011. An interaction-based approach to enhancing secondary school instruction and student achievement. *Science (New York, N.Y.)* 333, 6045 (2011), 1034–1037. Publisher: American Association for the Advancement of Science.

[3] David P Ausubel. 1968. A Cognitive view. *Educational psychology* (1968). Publisher: Holt, Reinehert & Winston.

[4] Sanghwan Bae, Donghyun Kwak, Sungdong Kim, Donghoon Ham, Soyoung Kang, Sang-Woo Lee, and Woomyoung Park. 2022. Building a role specified open-domain dialogue system leveraging large-scale language models. *arXiv preprint arXiv:2205.00176* (2022).

[5] Albert Bandura and others. 1986. Social foundations of thought and action. *Englewood Cliffs, NJ* 1986, 23-28 (1986), 2.

[6] Miri Barak and Yehudit Judy Dori. 2009. Enhancing higher order thinking skills among inservice science teachers via embedded assessment. *Journal of Science Teacher Education* 20, 5 (2009), 459–474. Publisher: Taylor & Francis.

[7] Klaus Bayer. 2007. *Argument und argumentation: logische grundlagen der argumentationsanalyse.* Vandenhoeck & Ruprecht.

[8] Yvonne Berkle, Lukas Schmitt, Antonia Tolzin, Andreas Janson, Thiemo Wambsganss, Jan Marco Leimeister, and Miriam Leuchter. 2023. Measuring university students' ability to recognize argument structures and fallacies. *Frontiers in Psychology* 14 (2023). Publisher: Frontiers Media SA.

[9] Philippe Besnard and Anthony Hunter. 2008. *Elements of argumentation.* Vol. 47. MIT press Cambridge.

[10] John Bitchener, Stuart Young, and Denise Cameron. 2005. The effect of different types of corrective feedback on ESL student writing. *Journal of second language writing* 14, 3 (2005), 191–205. Publisher: Elsevier.

[11] Beatriz Borges, Negar Foroutan, Deniz Bayazit, Anna Sotnikova, Syrielle Montariol, Tanya Nazaretzky, Mohammadreza Banaei, Alireza Sakhaeirad, Philippe Servant, Seyed Parsa Neshaei, and others. 2024. Could ChatGPT get an engineering degree? Evaluating higher education vulnerability to AI assistants. *Proceedings of the National Academy of Sciences* 121, 49 (2024), e2414955121. Publisher: National Academy of Sciences.

[12] Walter Brenner, Dimitris Karagiannis, Lutz Kolbe, Jens Krüger, Larry Leifer, Hermann-Josef Lamberti, Jan Marco Leimeister, Hubert Österle, Charles Petrie, Hasso Plattner, and others. 2014. User, use & utility research: the digital user as new design perspective in business and information systems engineering. *Wirtschaftsinformatik* 56 (2014), 65–72. Publisher: Springer.

[13] Robert O Briggs. 2006. On theory-driven design and deployment of collaboration systems. *International Journal of Human-Computer Studies* 64, 7 (2006), 573–582. Publisher: Elsevier.

[14] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, and others. 2020. Language models are few-shot learners. *Advances in neural information processing systems* 33 (2020), 1877–1901.

[15] Georg Brun and Gertrude Hirsch Hadorn. 2021. *Textanalyse in den Wissenschaften: Inhalte und Argumente analysieren und verstehen.* Vol. 4. vdf Hochschulverlag AG.

[16] Jake Rowan Byrne and Brendan Tangney. 2010. CAWriter: A CSCW/CSCL tool to support research students' academic writing. In *Proceedings of HCI 2010.* BCS Learning & Development.

[17] Banghao Chen, Zhaofeng Zhang, Nicolas Langrené, and Shengxin Zhu. 2023. Unleashing the potential of prompt engineering in Large Language Models: a comprehensive review. *arXiv preprint arXiv:2310.14735* (2023).

[18] Oana Cocarascu and Francesca Toni. 2016. Argumentation for machine learning: A survey.. In *COMMA*. 219–230.

[19] Betty Collis. 1993. Cooperative learning and CSCW: Research perspectives for internetworked educational environments. In *Invited papers of the international working conference of IFIP WG3. 3" lessons from learning"*.

[20] Alan Cooper, Robert Reimann, and David Cronin. 2007. *About face 3: the essentials of interaction design.* John Wiley & Sons.

[21] Sagnik Dakshit. 2024. Faculty perspectives on the potential of RAG in computer science higher education. *arXiv preprint arXiv:2408.01462* (2024).

[22] T Damer. 2008. *Attacking faulty reasoning: A practical guide to fallacy-free arguments.* Nelson Education.

[23] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).

[24] Pierre Dillenbourg and Frank Fischer. 2007. Computer-supported collaborative learning: The basics. *Zeitschrift für Berufs-und Wirtschaftspädagogik* 21 (2007), 111–130.

[25] Rosalind Driver, Paul Newton, and Jonathan Osborne. 2000. Establishing the norms of scientific argumentation in classrooms. *Science education* 84, 3 (2000), 287–312. Publisher: Wiley Online Library.

[26] Ann R Eisenberg and Catherine Garvey. 1981. Children's use of verbal strategies in resolving conflicts. *Discourse processes* 4, 2 (1981), 149–170. Publisher: Taylor & Francis.

[27] Rod Ellis. 2009. Corrective feedback and teacher development. *L2 Journal: An electronic refereed journal for foreign and second language educators* 1, 1 (2009).

[28] Paul Evans, Maarten Vansteenkiste, Philip Parker, Andrew Kingsford-Smith, and Sijing Zhou. 2024. Cognitive load theory and its relationships with motivation: a self-determination theory perspective. *Educational Psychology Review* 36, 1 (2024), 7. Publisher: Springer.

[29] Linda Hania Fasha and Wahyu Sopandi. 2024. Analysis of teaching material: To what extent are students' argumentation learning. In *International conference on teaching, learning and technology (ICTLT 2023).* Atlantis Press, 285–292.

[30] David F Feldon, Gregory Callan, Stephanie Juth, and Soojeong Jeong. 2019. Cognitive load as motivational cost. *Educational Psychology Review* 31 (2019), 319–337. Publisher: Springer.

[31] Frank Fischer, Ingo Kollar, Karsten Stegmann, and Christof Wecker. 2013. Toward a script theory of guidance in computer-supported collaborative learning. *Educational psychologist* 48, 1 (2013), 56–66. Publisher: Taylor & Francis.

[32] Sandra L Fisher and J Kevin Ford. 1998. Differential effects of learner effort and goal orientation on two learning outcomes. *Personnel Psychology* 51, 2 (1998), 397–420. Publisher: Wiley Online Library.

[33] Thomas François, Adeline Müller, Eva Rolin, and Magali Norré. 2020. AMesure: a Web platform to assist the clear writing of administrative texts. In *Proceedings of the 1st conference of the asia-pacific chapter of the association for computational linguistics and the 10th international joint conference on natural language processing: System demonstrations*. 1–7.

[34] James B Freeman. 2011. *Argument structure:: Representation and theory*. Vol. 18. Springer Science & Business Media.

[35] James B Freeman. 2011. *Dialectics and the macrostructure of arguments: A theory of argument structure*. Vol. 10. Walter de Gruyter.

[36] Katy Gero, Alex Calderwood, Charlotte Li, and Lydia Chilton. 2022. A design space for writing support tools using a cognitive process model of writing. In *Proceedings of the first workshop on intelligent and interactive writing assistants (In2Writing 2022)*. 11–24.

[37] Jim E Greer, Gordon Mccalla, Jason A Collins, Vive S Kumar, Paul Meagher, and Julita Vassileva. 1998. Supporting peer help and collaboration in distributed workplace environments. *International Journal of Artificial Intelligence in Education* 9 (1998), 159–177.

[38] Shirley Gregor, Leona Chandra Kruse, and Stefan Seidel. 2020. Research perspectives: the anatomy of a design principle. *Journal of the Association for Information Systems* 21, 6 (2020), 2.

[39] Matej Guid, Martin Možina, Matevž Pavlič, and Klemen Turšič. 2019. Learning by arguing in argument-based machine learning framework. In *Intelligent tutoring systems: 15th international conference, ITS 2019, kingston, jamaica, june 3–7, 2019, proceedings 15*. Springer, 112–122.

[40] Andreas Göldi, Thiemo Wambsganss, Seyed Parsa Neshaei, and Roman Rietsche. 2024. Intelligent support engages writers through relevant cognitive processes. In *Proceedings of the CHI conference on human factors in computing systems*. 1–12.

[41] Ulrike Hahn, Mike Oaksford, and Adam Corner. 2005. Circular arguments, begging the question and the formalization of argument strength. In *Proceedings of AMKLC'05, international symposium on adaptive models of knowledge, language and cognition*. 34–40.

[42] A Hakim, I Sahmadesti, and S Hadisaputra. 2020. Promoting students' argumentation skill through development science teaching materials based on guided inquiry models. In *Journal of physics: Conference series*, Vol. 1521. IOP Publishing, 042117. Number: 4.

[43] Jeongyun Han, Kwan Hoon Kim, Wonjong Rhee, and Young Hoan Cho. 2021. Learning analytics dashboards for adaptive support in face-to-face collaborative argumentation. *Computers & Education* 163 (2021), 104041. Publisher: Elsevier.

[44] John Hattie and Helen Timperley. 2007. The power of feedback. *Review of educational research* 77, 1 (2007), 81–112. Publisher: Sage Publications Sage CA: Thousand Oaks, CA.

[45] Matthew Jensen Hays, Nate Kornell, and Robert A Bjork. 2010. The costs and benefits of providing feedback during learning. *Psychonomic bulletin & review* 17 (2010), 797–801. Publisher: Springer.

[46] Md Naimul Hoque, Bhavya Ghai, and Niklas Elmqvist. 2022. DramatVis Personae: Visual text analytics for identifying social biases in creative writing. In *Proceedings of the 2022 ACM designing interactive systems conference*. 1260–1276.

[47] JWDL Hsiao. 1996. CSCL theories. *Austin: The University of Texas* (1996).

[48] Minlie Huang, Xiaoyan Zhu, and Jianfeng Gao. 2020. Challenges in building intelligent open-domain dialog systems. *ACM Transactions on Information Systems (TOIS)* 38, 3 (2020), 1–32. Publisher: ACM New York, NY, USA.

[49] Patrick Hurley. 2011. *A concise introduction to logic*. Nelson Education.

[50] Andreas Janson, Matthias Söllner, and Jan Marco Leimeister. 2020. Ladders for learning: is scaffolding the key to teaching problem-solving in technology-mediated learning contexts? *Academy of Management Learning & Education* 19, 4 (2020), 439–468. Publisher: Academy of Management Briarcliff Manor, NY.

[51] Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *Comput. Surveys* 55, 12 (2023), 1–38. Publisher: ACM New York, NY.

[52] Eunkyung Jo, Daniel A Epstein, Hyunhoon Jung, and Young-Ho Kim. 2023. Understanding the benefits and challenges of deploying conversational AI leveraging large language models for public health intervention. In *Proceedings of the 2023 CHI conference on human factors in computing systems*. 1–16.

[53] Ralph Henry Johnson and J Anthony Blair. 2006. *Logical self-defense*. Idea.

[54] David H Jonassen and Bosung Kim. 2010. Arguing to learn and learning to argue: Design justifications and guidelines. *Educational Technology Research and Development* 58 (2010), 439–457. Publisher: Springer.

[55] Majeed Kazemitabaar, Xinying Hou, Austin Henley, Barbara J Ericson, David Weintrop, and Tovi Grossman. 2023. How novices use LLM-Based code generators to solve CS1 coding tasks in a self-paced learning environment. *arXiv preprint arXiv:2309.14049* (2023).

[56] Taewan Kim, Seolyeong Bae, Hyun Ah Kim, Su-woo Lee, Hwajung Hong, Chanmo Yang, and Young-Ho Kim. 2024. MindfulDiary: Harnessing large language model to support psychiatric patients' journaling. In *Proceedings of the CHI conference on human factors in computing systems*. 1–20.

[57] Hannah Rose Kirk, Yennie Jun, Filippo Volpin, Haider Iqbal, Elias Benussi, Frederic Dreyer, Aleksandar Shtedritski, and Yuki Asano. 2021. Bias out-of-the-box: An empirical analysis of intersectional occupational biases in popular generative language models. *Advances in neural information processing systems* 34 (2021), 2611–2624.

[58] Paul A Kirschner, John Sweller, Femke Kirschner, and Jimmy Zambrano R. 2018. From cognitive load theory to collaborative cognitive load theory. *International journal of computer-supported collaborative learning* 13 (2018), 213–233. Publisher: Springer.

[59] Simon Knight, Antonette Shibani, Sophie Abel, Andrew Gibson, and Philippa Ryan. 2020. AcaWriter: A learning analytics tool for formative feedback on academic writing. *Journal of Writing Research* (2020). Publisher: ARLE (International Associaton for Research in L1 Education).

[60] Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *Advances in neural information processing systems* 35 (2022), 22199–22213.

[61] Yubo Kou and Xinning Gui. 2020. Mediating community-AI interaction through situated explanation: the case of AI-Led moderation. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW2 (2020), 1–27. Publisher: ACM New York, NY, USA.

[62] Deanna Kuhn. 2001. How do people know? *Psychological science* 12, 1 (2001), 1–8. Publisher: SAGE Publications Sage CA: Los Angeles, CA.

[63] Deanna Kuhn. 2005. *Education for thinking*. Harvard University Press.

[64] Chinmay Kulkarni, Tongshuang Wu, Kenneth Holstein, Q Vera Liao, Min Kyung Lee, Mina Lee, and Hariharan Subramonyam. 2023. LLMs and the infrastructure of CSCW. In *Companion publication of the 2023 conference on computer supported cooperative work and social computing*. 408–410.

[65] Yau Wai Lam, Khe Foon Hew, and Kin Fung Chiu. 2018. Improving argumentative writing: Effects of a blended learning approach and gamification. (2018). Publisher: University of Hawaii National Foreign Language Resource Center.

[66] J Richard Landis and Gary G Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics. Journal of the International Biometric Society* (1977), 159–174. Publisher: JSTOR.

[67] Saeed Latifi, Omid Noroozi, Javad Hatami, and Harm JA Biemans. 2021. How does online peer feedback improve argumentative essay writing and learning? *Innovations in Education and Teaching International* 58, 2 (2021), 195–206. Publisher: Taylor & Francis.

[68] Saeed Latifi, Omid Noroozi, and Ebrahim Talaee. 2023. Worked example or scripting? Fostering students' online argumentative peer feedback, essay writing and learning. *Interactive Learning Environments* 31, 2 (2023), 655–669. Publisher: Taylor & Francis.

[69] Martin Lauzier and Annabelle Bilodeau Clarke. 2024. Linking learning goal orientation to learning from error: the mediating role of motivation to learn and metacognition. *European Journal of Training and Development* 48, 5/6 (2024), 485–500. Publisher: Emerald Publishing Limited.

[70] Minha Lee, Sander Ackermans, Nena Van As, Hanwen Chang, Enzo Lucas, and Wijnand IJsselsteijn. 2019. Caring for Vincent: a chatbot for self-compassion. In *Proceedings of the 2019 CHI conference on human factors in computing systems*. 1–13.

[71] Mina Lee, Katy Ilonka Gero, John Joon Young Chung, Simon Buckingham Shum, Vipul Raheja, Hua Shen, Subhashini Venugopalan, Thiemo Wambsganss, David Zhou, Emad A Alghamdi, and others. 2024. A design space for intelligent and interactive writing assistants. In *Proceedings of the CHI conference on human factors in computing systems*. 1–35.

[72] Philipp Leitner, Martin Ebner, Hanna Geisswinkler, and Sandra Schön. 2021. Visualization of learning for students: A dashboard for study progress–development, design details, implementation, and user feedback. In *Visualizations and dashboards for learning analytics*. Springer, 423–437.

[73] Evelyn Palominos Letelier. 2022. *Transforming errors into learning opportunities in simulation-based learning*. University of Technology Sydney (Australia).

[74] Junyi Li, Xiaoxue Cheng, Wayne Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. 2023. HaluEval: A large-scale hallucination evaluation benchmark for large language models. *arXiv e-prints* (2023), arXiv–2305.

[75] Hanmeng Liu, Ruoxi Ning, Zhiyang Teng, Jian Liu, Qiji Zhou, and Yue Zhang. 2023. Evaluating the logical reasoning ability of ChatGPT and GPT-4. *arXiv preprint arXiv:2304.03439* (2023).

[76] Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2023. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *Comput. Surveys* 55, 9 (2023),

1–35. Publisher: ACM New York, NY.
[77] Chung Kwan Lo. 2023. What is the impact of ChatGPT on education? A rapid review of the literature. *Education Sciences* 13, 4 (2023), 410. Publisher: MDPI.
[78] Li Lucy and David Bamman. 2021. Gender and representation bias in GPT-3 generated stories. In *Proceedings of the third workshop on narrative understanding*. 48–55.
[79] Kalle J Lyytinen and Ojelanki K Ngwenyama. 1992. What does computer support for cooperative work mean? A structurational analysis of computer supported cooperative work. *Accounting, Management and Information Technologies* 2, 1 (1992), 19–37. Publisher: Elsevier.
[80] Winfried Löffler. 2008. *Einführung in die logik*. Vol. 18. W. Kohlhammer Verlag.
[81] Uwe Maier and Christian Klotz. 2022. Personalized feedback in digital learning environments: Classification framework and literature review. *Computers and Education: Artificial Intelligence* 3 (2022), 100080. Publisher: Elsevier.
[82] Andrew J Martin and Paul Evans. 2018. Load reduction instruction: Exploring a framework that assesses explicit instruction through to independent learning. *Teaching and Teacher Education* 73 (2018), 203–214. Publisher: Elsevier.
[83] Douglas W Maynard. 1985. How children start arguments. *Language in society* 14, 1 (1985), 1–29. Publisher: Cambridge University Press.
[84] Paola Mejia-Domenzain, Jibril Frej, Seyed Parsa Neshaei, Luca Mouchel, Tanya Nazaretsky, Thiemo Wambsganss, Antoine Bosselut, and Tanja Käser. 2024. Enhancing procedural writing through personalized example retrieval: a case study on cooking recipes. *International Journal of Artificial Intelligence in Education* (2024), 1–37. Publisher: Springer.
[85] Janet Metcalfe. 2017. Learning from errors. *Annual review of psychology* 68 (2017), 465–489. Publisher: Annual Reviews.
[86] Jiřı Milička, Anna Marklová, Klára VanSlambrouck, Eva Pospıšilová, Jana Šimsová, Samuel Harvan, and Ondřej Drobil. 2024. Large language models are able to downplay their cognitive abilities to fit the persona they simulate. *PLOS One* 19, 3 (2024), e0298522. Publisher: Public Library of Science San Francisco, CA USA.
[87] Roxana Moreno. 2004. Decreasing cognitive load for novice students: Effects of explanatory versus corrective feedback in discovery-based multimedia. *Instructional science* 32, 1-2 (2004), 99–113. Publisher: Springer.
[88] Dor Muhlgay, Ori Ram, Inbal Magar, Yoav Levine, Nir Ratner, Yonatan Belinkov, Omri Abend, Kevin Leyton-Brown, Amnon Shashua, and Yoav Shoham. 2023. Generating benchmarks for factuality evaluation of language models. *arXiv preprint arXiv:2307.06908* (2023).
[89] Hannes Münchow, Tobias Richter, Sarah von der Mühlen, and Sebastian Schmid. 2019. The ability to evaluate arguments in scientific texts: Measurement, cognitive processes, nomological network, and relevance for academic success at the university. *British Journal of Educational Psychology* 89, 3 (2019), 501–523. Publisher: Wiley Online Library.
[90] Seyed Parsa Neshaei, Roman Rietsche, Xiaotian Su, and Thiemo Wambsganss. 2024. Enhancing peer review with AI-powered suggestion generation assistance: Investigating the design dynamics. In *Proceedings of the 29th international conference on intelligent user interfaces*. 88–102.
[91] E Michael Nussbaum, Denise L Winsor, Yvette M Aqui, 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 (2007), 479–500. Publisher: Springer.
[92] Josephine Oranga. 2023. Benefits of artificial intelligence (ChatGPT) in education and learning: Is Chat GPT helpful? *International Review of Practical Innovation, Technology and Green Energy (IRPITAGE)* 3, 3 (2023), 46–50.
[93] Jonathan F Osborne, J Bryan Henderson, Anna MacPherson, Evan Szu, Andrew Wild, and Shi-Ying Yao. 2016. The development and validation of a learning progression for argumentation in science. *Journal of research in science teaching* 53, 6 (2016), 821–846. Publisher: Wiley Online Library.
[94] Hiroyuki Osone, Jun-Li Lu, and Yoichi Ochiai. 2021. BunCho: ai supported story co-creation via unsupervised multitask learning to increase writers' creativity in Japanese. In *Extended abstracts of the 2021 CHI conference on human factors in computing systems*. 1–10.
[95] Amy A Overman, Joseph DW Stephens, and Mary F Bernhardt. 2021. Enhanced memory for context associated with corrective feedback: evidence for episodic processes in errorful learning. *Memory (Hove, England)* 29, 8 (2021), 1017–1042. Publisher: Taylor & Francis.
[96] Joonsuk Park and Claire Cardie. 2014. Identifying appropriate support for propositions in online user comments. In *Proceedings of the first workshop on argumentation mining*. 29–38.
[97] Harshada Patel, Michael Pettitt, and John R Wilson. 2012. Factors of collaborative working: A framework for a collaboration model. *Applied ergonomics* 43, 1 (2012), 1–26. Publisher: Elsevier.
[98] Zhenhui Peng, Qingyu Guo, Ka Wing Tsang, and Xiaojuan Ma. 2020. Exploring the effects of technological writing assistance for support providers in online mental health community. In *Proceedings of the 2020 CHI conference on human factors in computing systems*. 1–15.

[99] Niels Pinkwart, Kevin Ashley, Collin Lynch, and Vincent Aleven. 2009. Evaluating an intelligent tutoring system for making legal arguments with hypotheticals. *International Journal of Artificial Intelligence in Education* 19, 4 (2009), 401–424. Publisher: IOS Press.

[100] Iyad Rahwan and Guillermo R Simari. 2009. *Argumentation in artificial intelligence*. Vol. 47. Springer.

[101] Eddo Rigotti and Sara Greco Morasso. 2009. Argumentation as an object of interest and as a social and cultural resource. *Argumentation and education: Theoretical foundations and practices* (2009), 9–66. Publisher: Springer.

[102] Lance J Rips. 2002. Circular reasoning. *Cognitive science* 26, 6 (2002), 767–795. Publisher: Wiley Online Library.

[103] Stephen Roller, Emily Dinan, Naman Goyal, Da Ju, Mary Williamson, Yinhan Liu, Jing Xu, Myle Ott, Kurt Shuster, Eric M Smith, and others. 2020. Recipes for building an open-domain chatbot. *arXiv preprint arXiv:2004.13637* (2020).

[104] Pranab Sahoo, Ayush Kumar Singh, Sriparna Saha, Vinija Jain, Samrat Mondal, and Aman Chadha. 2024. A systematic survey of prompt engineering in large language models: Techniques and applications. *arXiv preprint arXiv:2402.07927* (2024).

[105] Bahareh Sarrafzadeh, Sujay Kumar Jauhar, Michael Gamon, Edward Lank, and Ryen W White. 2021. Characterizing stage-aware writing assistance for collaborative document authoring. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW3 (2021), 1–29. Publisher: ACM New York, NY, USA.

[106] Inge Scheper, Ellen RA de Bruijn, Dirk Bertens, Roy PC Kessels, and Inti A Brazil. 2019. The impact of error frequency on errorless and errorful learning of object locations using a novel paradigm. *Memory (Hove, England)* 27, 10 (2019), 1371–1380. Publisher: Taylor & Francis.

[107] Oliver Scheuer, Frank Loll, Niels Pinkwart, and Bruce M McLaren. 2010. Computer-supported argumentation: A review of the state of the art. *International Journal of Computer-supported collaborative learning* 5 (2010), 43–102. Publisher: Springer.

[108] Baruch B Schwarz. 2009. Argumentation and learning. *Argumentation and education: Theoretical foundations and practices* (2009), 91–126. Publisher: Springer.

[109] Baruch B Schwarz, Lauren B Resnick, and Michael J Baker. 2017. *Dialogue, argumentation and education: History, theory and practice*. Cambridge University Press.

[110] Hua Shen, Chieh-Yang Huang, Tongshuang Wu, and Ting-Hao Kenneth Huang. 2023. ConvXAI: Delivering heterogeneous AI explanations via conversations to support human-AI scientific writing. In *Companion publication of the 2023 conference on computer supported cooperative work and social computing*. 384–387.

[111] Shuming Shi, Enbo Zhao, Duyu Tang, Yan Wang, Piji Li, Wei Bi, Haiyun Jiang, Guoping Huang, Leyang Cui, Xinting Huang, and others. 2022. Effidit: Your ai writing assistant. *arXiv preprint arXiv:2208.01815* (2022).

[112] Valerie J Shute. 2008. Focus on formative feedback. *Review of educational research* 78, 1 (2008), 153–189. Publisher: Sage Publications.

[113] Anjali Singh, Christopher Brooks, and Xu Wang. 2024. The impact of student-AI collaborative feedback generation on learning outcomes. In *AI for education: Bridging innovation and responsibility at the 38th AAAI annual conference on AI*.

[114] Adish Singla. 2023. Evaluating ChatGPT and GPT-4 for visual programming. *arXiv preprint arXiv:2308.02522* (2023).

[115] Burrhus Frederic Skinner. 1965. *Science and human behavior*. Simon and Schuster. Number: 92904.

[116] Dave Snowdon, Elizabeth F Churchill, and Alan J Munro. 2001. Collaborative virtual environments: Digital spaces and places for CSCW: An introduction. In *Collaborative virtual environments: Digital places and spaces for interaction*. Springer, 3–17.

[117] Christian Stab and Iryna Gurevych. 2017. Parsing argumentation structures in persuasive essays. *Computational Linguistics* 43, 3 (2017), 619–659. Publisher: MIT Press One Rogers Street, Cambridge, MA 02142-1209, USA journals-info.

[118] Christian Stab and Iryna Gurevych. 2017. Recognizing insufficiently supported arguments in argumentative essays. In *Proceedings of the 15th conference of the European chapter of the association for computational linguistics: Volume 1, long papers*. 980–990.

[119] Jennifer L Steele. 2023. To GPT or not GPT? Empowering our students to learn with AI. *Computers and Education: Artificial Intelligence* 5 (2023), 100160. Publisher: Elsevier.

[120] Xiaotian Su, Thiemo Wambsganss, Roman Rietsche, Seyed Parsa Neshaei, and Tanja Käser. 2023. Reviewriter: AI-generated instructions for peer review writing. In *Proceedings of the 18th workshop on innovative use of NLP for building educational applications (BEA 2023)*. 57–71.

[121] Yanfang Su, Yun Lin, and Chun Lai. 2023. Collaborating with ChatGPT in argumentative writing classrooms. *Assessing Writing* 57 (2023), 100752. Publisher: Elsevier.

[122] Lu Sun, Stone Tao, Junjie Hu, and Steven P Dow. 2024. MetaWriter: Exploring the potential and perils of AI writing support in scientific peer review. *Proceedings of the ACM on Human-Computer Interaction* 8, CSCW1 (2024), 1–32. Publisher: ACM New York, NY, USA.

[123] Ben Swanson, Kory Mathewson, Ben Pietrzak, Sherol Chen, and Monica Dinalescu. 2021. Story centaur: Large language model few shot learning as a creative writing tool. In *Proceedings of the 16th conference of the European chapter of the association for computational linguistics: System demonstrations*. 244–256.

[124] John Sweller. 1988. Cognitive load during problem solving: Effects on learning. *Cognitive science* 12, 2 (1988), 257–285. Publisher: Elsevier.

[125] John Sweller and Graham A Cooper. 1985. The use of worked examples as a substitute for problem solving in learning algebra. *Cognition and instruction* 2, 1 (1985), 59–89. Publisher: Taylor & Francis.

[126] Zhen Tan, Alimohammad Beigi, Song Wang, Ruocheng Guo, Amrita Bhattacharjee, Bohan Jiang, Mansooreh Karami, Jundong Li, Lu Cheng, and Huan Liu. 2024. Large language models for data annotation: A survey. *arXiv preprint arXiv:2402.13446* (2024).

[127] Lev Tankelevitch, Viktor Kewenig, Auste Simkute, Ava Elizabeth Scott, Advait Sarkar, Abigail Sellen, and Sean Rintel. 2024. The metacognitive demands and opportunities of generative AI. In *Proceedings of the CHI conference on human factors in computing systems*. 1–24.

[128] Stephen Toulmin, Richard D Rieke, and Allan Janik. 1979. An introduction to reasoning. (1979).

[129] Anahuac Valero Haro, Omid Noroozi, Harm JA Biemans, and Martin Mulder. 2019. The effects of an online learning environment with worked examples and peer feedback on students' argumentative essay writing and domain-specific knowledge acquisition in the field of biotechnology. *Journal of Biological Education* 53, 4 (2019), 390–398. Publisher: Taylor & Francis.

[130] Frans H Van Eemeren, Rob Grootendorst, and Tjark Kruiger. 2019. *Handbook of argumentation theory: A critical survey of classical backgrounds and modern studies*. Vol. 7. Walter de Gruyter GmbH & Co KG.

[131] Jan Vom Brocke, Wolfgang Maaß, Peter Buxmann, Alexander Maedche, Jan Marco Leimeister, and Günter Pecht. 2018. Future work and enterprise systems. *Business & Information Systems Engineering* 60 (2018), 357–366. Publisher: Springer.

[132] Lev Semenovich Vygotsky and Michael Cole. 1978. *Mind in society: Development of higher psychological processes*. Harvard university press.

[133] Henning Wachsmuth, Manfred Stede, Roxanne El Baff, Khalid Al Khatib, Maria Skeppstedt, and Benno Stein. 2018. Argumentation synthesis following rhetorical strategies. In *Proceedings of the 27th international conference on computational linguistics*. 3753–3765.

[134] Douglas Walton. 2004. Classification of fallacies of relevance. *Informal logic* 24, 1 (2004).

[135] Thiemo Wambsganss, Andrew Caines, and Paula Buttery. 2022. 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)*. 134–140.

[136] Thiemo Wambsganss, Andreas Janson, Tanja Käser, and Jan Marco Leimeister. 2022. Improving students argumentation learning with adaptive self-evaluation nudging. *Proceedings of the ACM on Human-Computer Interaction* 6, CSCW2 (2022), 1–31. Publisher: ACM New York, NY, USA.

[137] Thiemo Wambsganss, Andreas Janson, Matthias Söllner, Ken Koedinger, and Jan Marco Leimeister. 2024. Improving students' argumentation skills using dynamic machine-learning–based modeling. *Information Systems Research* (2024). Publisher: INFORMS.

[138] 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*. 1–13.

[139] Thiemo Wambsganss and Christina Niklaus. 2022. Modeling persuasive discourse to adaptively support students' argumentative writing. In *Proceedings of the 60th annual meeting of the association for computational linguistics (volume 1: Long papers)*. 8748–8760.

[140] 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 2020 CHI conference on human factors in computing systems*. 1–14.

[141] Thiemo Wambsganss, Christina Niklaus, Matthias Söllner, Siegfried Handschuh, and Jan Marco Leimeister. 2020. A corpus for argumentative writing support in German. *arXiv preprint arXiv:2010.13674* (2020).

[142] Thiemo Wambsganss, Christina Niklaus, Matthias Söllner, Siegfried Handschuh, and Jan Marco Leimeister. 2021. Supporting cognitive and emotional empathic writing of students. *arXiv preprint arXiv:2105.14815* (2021).

[143] Thiemo Wambsganss and Roman Rietsche. 2019. Towards designing an adaptive argumentation learning tool.. In *ICIS*.

[144] Thiemo Wambsganss, Xiaotian Su, Vinitra Swamy, Seyed Parsa Neshaei, Roman Rietsche, and Tanja Käser. 2023. Unraveling downstream gender bias from large language models: a study on AI educational writing assistance. In *Findings of the association for computational linguistics: EMNLP 2023*, Houda Bouamor, Juan Pino, and Kalika Bali (Eds.). Association for Computational Linguistics, Singapore, 10275–10288. https://doi.org/10.18653/v1/2023.findings-

emnlp.689
[145] Kento Watanabe, Yuichiroh Matsubayashi, Kentaro Inui, Tomoyasu Nakano, Satoru Fukayama, and Masataka Goto. 2017. Lyrisys: An interactive support system for writing lyrics based on topic transition. In *Proceedings of the 22nd international conference on intelligent user interfaces*. 559–563.
[146] Florian Weber, Thiemo Wambsganss, Seyed Parsa Neshaei, and Matthias Söllner. 2023. Structured persuasive writing support in legal education: A model and tool for german legal case solutions. In *Findings of the association for computational linguistics: ACL 2023*. 2296–2313.
[147] Florian Weber, Thiemo Wambsganss, Seyed Parsa Neshaei, and Matthias Söllner. 2024. LegalWriter: An intelligent writing support system for structured and persuasive legal case writing for novice law students. In *Proceedings of the CHI conference on human factors in computing systems*. 1–23.
[148] Jing Wei, Sungdong Kim, Hyunhoon Jung, and Young-Ho Kim. 2024. Leveraging large language models to power chatbots for collecting user self-reported data. *Proceedings of the ACM on Human-Computer Interaction* 8, CSCW1 (2024), 1–35. Publisher: ACM New York, NY, USA.
[149] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, and others. 2019. HuggingFace's Transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771* (2019).
[150] 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 ACM on Human-Computer Interaction* 6, CSCW2 (2022), 1–30. Publisher: ACM New York, NY, USA.
[151] Haoran Xie, Hui-Chun Chu, Gwo-Jen Hwang, and Chun-Chieh Wang. 2019. Trends and development in technology-enhanced adaptive/personalized learning: A systematic review of journal publications from 2007 to 2017. *Computers & Education* 140 (2019), 103599. Publisher: Elsevier.
[152] Soobin Yim, Dakuo Wang, Judith Olson, Viet Vu, and Mark Warschauer. 2017. Synchronous collaborative writing in the classroom: undergraduates' collaboration practices and their impact on writing style, quality, and quantity. In *Proceedings of the 2017 ACM conference on computer supported cooperative work and social computing*. 468–479.
[153] Rui Zhang, Nathan J McNeese, Guo Freeman, and Geoff Musick. 2021. "An ideal human" expectations of AI teammates in human-AI teaming. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW3 (2021), 1–25. Publisher: ACM New York, NY, USA.
[154] Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, and others. 2023. A survey of large language models. *arXiv preprint arXiv:2303.18223* (2023).

APPENDIX

Texts shown in ALure A[19]

Tooltip describing claim: Claims are marked yellow in the text. Typically, a new argument begins with a claim. This claim must then be supported or attacked, using premises for this.

Tooltip describing premise: Premises are marked green in the text. Strong arguments are characterized by claims being supported or attacked by at least two premises.

Tooltip describing error: Possible errors in the argumentation are marked in red in the text.

The text shown if the learner writing does not have errors recognized by the model: ALure has recognized that your text probably does not contain errors. Check whether you can find any errors yourself.

The text shown if the learner writing has errors recognized by the model: ALure has recognized that your text may still contain errors. Please check the areas marked in red for possible errors.

Text indicating the types of argumentation errors shown to the user: Strong arguments are plausible and do not contain any arguments. There are many different types of arguments. Here are four examples that occur very frequently: (1) In overgeneralization,

[19]We machine-translated all texts into English for putting them in our paper. The texts are included in the same formatting as it was used in ALure, including, for example, the HTML tags we used in our prompt.

the content of the premise only refers to a few or specific cases (e.g. subjective experiences) that are not sufficiently representative as evidence for the claim.
Example: Dogs that bark do not bite. My dog barks often and has never bitten anyone. (2) In circular reasoning, statements are justified by themselves because the premise is the same statement as the claim, just formulated differently. Such arguments often sound plausible at first glance. However, they do not contribute meaningfully to the discussion because the argument is in a state of crisis.
Example: If you have a dog, you often go for a walk because you have to walk a dog every day. (3) Irrelevance occurs when a statement misses the core topic of the argument. At first glance, such arguments often seem convincing because they can certainly strengthen the intended position. In terms of content, however, they have no value for the argument, but rather distract the reader from the actual topic.
Example: A dog is a lot of work, for which you also have to have the time. Having to go for a walk is not much fun, especially in bad weather. (4) In the formal fallacy, the claim is derived from the premises as a logical consequence, although the premises do not logically allow this conclusion. The error here is not in the content, but in the fact that the logical form of the argument is not correct.
Example: The mail must have been there already. Bello barked earlier and he always barks when he sees the postman.
Even if Bello always barks when the mail arrives, he does not necessarily only bark when the mail arrives. There could just as well have been another reason for the barking.

Explanation of arguments: An argument is about discussing the advantages and disadvantages of a controversial topic. In this way, you can form your own opinion on the topic and convince others of it. An argument therefore needs arguments that look at the subject of discussion from different perspectives. An argument always consists of several statements: an assertion and at least one premise, or better still, several premises. The assertion is a statement about which people can have different opinions. Premises are statements that plausibly explain this assertion. Statements that are undisputed, such as verifiable facts, are particularly suitable here. A convincing argument should contain around 2-3 supporting premises that explain in a comprehensible way why the assertion is correct. In addition, the persuasiveness of the argument can be increased by adding a refuting premise. A refuting premise is a statement that speaks against the assertion. It shows that you are aware that there are exceptions and that the statement does not have to be true in all circumstances. Example:
 Statement: The neighbor probably has a new dog.
 Premise (supporting): You can hear it barking next door.
 Premise (supporting): He recently said he would like a new dog.
 Premise (refuting): It could just be a visitor with a dog.

GPT Prompt used in ALure B

"Imagine you are an expert at writing argumentative texts. They are particularly trained to recognize the structures in arguments. The structure of good arguments looks like this. A good argumentative text begins with an introduction. You are welcome to give an overview of the arguments that will be discussed. The introduction serves to introduce the reader to the topic. A new topic paragraph can be introduced with a non-argumentative sentence that prepares the reader for the new topic. Typically, a new argument begins with an assertion. This claim must be supported or attacked below, using premises for this purpose. Strong arguments are characterized by claims being supported or attacked by at least two premises. For better understanding, non-argumentative text can also be used in a discussion. However, care must be taken to ensure that the text largely consists of argumentative text, otherwise the significance will be weakened. A strong argumentative text ends

the discussion with a summary. All arguments made are presented again to give the reader a final overview.

You are also trained to recognize argumentative errors, because the persuasiveness of an argumentative text depend on both its structure and whether certain argumentation errors are present or not. An argument error occurs when:

1. a premise is used as an explanation for a claim, but does not (sufficiently) explain the claim,

2. a claim is drawn as a conclusion from a premise from which this conclusion cannot be (sufficiently) derived, or

3. a claim is made that does not (sufficiently) contribute to the formation of the statement.

An argumentation error lies in the relationship between two argumentation components, e.g. between a claim and a premise. There are four errors here:

(1) An error of irrelevance occurs when a statement is made that is related to the topic of the argument but misses its core content. At first glance, such arguments often seem convincing, as they can certainly strengthen the intended position. In terms of content, however, they have no value for the argument, but rather distract the reader from the actual topic. Here is an example, the content is irrelevant:

A disadvantage of the dishwasher is that it is very expensive.

(2) In the formal fallacy, an argument component is related to a conclusion that cannot be formally derived from it. For this purpose, the individual statements of the relevant argument components can be translated into a formal language. An example would be as follows: This example contains the following statements, but the content is irrelevant: Statement: There are dishwashers with Eco programs

Conclusion: The dishwasher is more environmentally friendly than handwashing.

(3) With overgeneralization, a fact that only relates to a few or certain cases is transferred to other cases or to the general public. An example would be as follows, but the content is irrelevant: There are different models today that also have certain eco programs. If we buy a new machine, we will definitely save more than with a used machine.

(4) In circular reasoning, a statement is justified by itself. These are often redundant statements, in different words. An example would be as follows, but the content is irrelevant: In the flipped classroom method, there are often difficulties in implementing the teaching of material with regard to the heterogeneity of a class. There are difficulties regarding internal differentiation, especially when conveying material. On the other hand, using the method does not result in a homogeneous transfer of material.

Give feedback on argumentative errors in the text. Limit yourself to three feedback points. Explicitly address whether and which argumentative errors the text has. Also give three reasons for and against the structure of the argument. In your feedback, focus on the ability to persuade, the structure of the argument in terms of claims and premises, possible argumentative errors, and not on the content. It's about the following text:"

[text here]

"Always write feedback in German. Format the feedback as an unsorted HTML list with a minimum of 150 and a maximum of 180 words. All reasons are brought to an end. Please do not write additional feedback paragraphs after the feedback about the errors and the six reasons. The reasons for and against the argument are only listed after feedback on argumentative errors and must not each relate to the same aspects. A reason for the argument structure should not be a reason against the structure. An example of how the feedback should be structured, where only the structure should be used: Feedback on argumentative errors: <ul> The components of this argument all represent the same statement in terms of content, they are just each worded a little differently. No component provides information as to why this statement can

be accepted. Rather, it is said: “It is so because it is so.” From an argumentative point of view, such circular reasoning is of no value, although they are usually correct from a formal and logical point of view.   The premise here is simply that there are newer models with Eco programs. A conclusion that you might be able to save more with a newer machine than with an older machine would initially be understandable. In the claim, however, it is now presented as if all new machines are inevitably equipped with an Eco program. It is also assumed that the used machine does not have this, although it is not at all clear whether it belongs to the generation of machines with Eco programs. The statement that there are machines with an Eco program is transferred to all new devices and is categorically excluded for used machines. This is an overgeneralization.   The claim is relevant here for the formation of the statement. Two premises follow, both of which reinforce the impact on the environment caused by the creation of waste by the dishwasher. If you take a closer look at the statements, you will notice that the claim is about the environmental impact of disposing of the dishwasher. The second premise, that dishwasher tabs create a lot of waste, has nothing to do with the later disposal of the machine. It is therefore completely irrelevant to the aspect of disposal addressed in the claim. There is therefore an error of irrelevance here.   Reasons for the structure of the argument: <ul> Reason 1   Reason 2 </ li> Reason 3 Reasons against the structure argument: <ul> Reason 1 Reason 2 < li>Reason 3 ”

Received January 2024; revised July 2024; accepted October 2024