

Navigating AI feedback: Changing perceptions and argumentation skills in science teacher education

Eylem Yıldız-Feyzioğlu^{a *} , Sevil Aymak^a 

^a Aydın Adnan Menderes University, Aydın, Türkiye

Suggested citation: Yıldız-Feyzioğlu, E. & Aymak, S. (2026). Navigating AI feedback: Changing perceptions and argumentation skills in science teacher education. *Journal of Educational Technology and Online Learning*, 9(3), 509-536.

Highlights

- This study explored pre-service science teachers' perceptions of AI-supported argument feedback and how these evolved over time.
- Findings show that while the number of arguments decreased, their quality and use of refutation and multi-component structures improved with AI feedback.
- The study derives design principles for AI-supported feedback, emphasizing holistic argument analysis, speed, interaction, and individualization.

Abstract

This study aimed to examine pre-service teachers' (PST) perceptions of AI-supported argument feedback (AIAF), changes in their perceptions, argument level and component development, process experiences, and to derive principles for AIAF design. The research was conducted in two cycles using the Design-Based Research approach. In the first cycle, prompts based on the Toulmin argumentation model (2003) and Erduran et al. (2004)'s argument level framework was developed for ChatGPT (GPT-4). In the second cycle, these prompts were tested through transcripts of scenario-based discussions with PSTs to provide personalized feedback. Data were collected from semi-structured interviews, discussion transcripts, and AI-generated feedback outputs. The study group consisted of nine science PSTs selected through purposive sampling from a state university in Turkey. The findings showed that the PSTs' perceptions of feedback evolved from a descriptive level to a pedagogical and affective understanding; they began to see AI as a complementary tool. While the quantity of arguments decreased, their quality increased; progress was made in refutation and multi-component argument production. PSTs found AI feedback positive in terms of speed and motivation but limited in terms of personalization. The study recommends the principles of holistic argument analysis, speed, interaction, and individualization for AI-supported feedback design. This study presents original data to the relevant literature, and its findings form the basis for a comprehensive study currently being conducted with a larger sample using a pre-test-post-test design.

Article Info: Research Article

Keywords: *Argumentation Feedback*
Artificial Intelligence Design-Based Research
Pre-Service Teacher

1. Introduction

The ability to evaluate data, synthesize perspectives, and generate innovative solutions for scientific and social problems is fundamental to the production of new knowledge and the reinterpretation of existing knowledge (Nusche et al., 2025; Osborne, 2010). Argumentation cultivates rational thinkers, competent problem solvers, and proactive decision-makers by fostering multiple perspectives through scientific inquiry (Uçar & Demiraslan Çevik, 2020). In this thought process, rather than receiving an argument from a teacher as a cumulative set of factual knowledge (Kuhn, 1993) the learner constructs the connection between the data and the claim by pointing to specific characteristics of the data (Sandoval & Millwood, 2005). Beyond this structural definition, Walton (1989) emphasizes the relational dimension of the argument between the two parties and defines the argument as reasoning that aims to make the other party

* Corresponding author. Department of Mathematics and Science Education, Aydın Adnan Menderes University, 09010, Aydın, Türkiye.

e-mail address: eylemyildiz@adu.edu.tr

Doi: <http://doi.org/10.31681/jetol.1933284>

Received 18 Apr 2026; Accepted 11 Sep 2026

ISSN: 2618-6586. This is an open Access article under the CC BY license.



accept a controversial proposition based on the premises that the other party accepts or can be persuaded to accept. Complementing this dialogical perspective, Kuhn et al. (2013) consider the persuasion process as reasoning in which the other party takes into account the data that will weaken the claim as well as supporting the claim and presents it in a comprehensive and complementary way by making use of more than one complementary source of data. On the other hand, the rapid transmission of both misleading information and argumentation is a gateway to accepting new knowledge after evaluating it applying critical thinking skills (Chotijah et al., 2025). In an era where misleading information spreads rapidly, argumentation is a tool for evaluating information through critical thinking. Therefore, it is essential that preservice teachers (PSTs) are equipped with the necessary pedagogical skills to understand the interaction between science and society and to provide their students with environments for rational decision-making through argumentation (Han-Tosunoglu et al., 2025). This underscores the need for the PSTs to equip themselves with the required pedagogical skills to effectively implement argumentation in their classrooms.

The development of argumentation skills depends on effective teaching methods and feedback practices. Research shows that PSTs can produce more sophisticated arguments as a result of participating in argumentative teaching, support, and feedback sessions (Christenson & Walan, 2023; Çakiroğlu et al., 2025; Kaya, 2013). However, there are significant deficiencies in argumentation skills among students at all educational levels: students often only present claims, fail to provide warrant, or show weak data (Fadillah & Deta, 2020; Sik et al., 2023). PSTs also struggle to teach these skills. For example, they have difficulty evaluating student arguments, especially in constructing warrant, data, and counter-arguments (Aydeniz & Ozdilek, 2016). They tend to make superficial and one-dimensional explanations of socio-scientific issues (Martín-Gámez & Erduran, 2018) and use data mostly to support their own views (Iordanou & Constantinou, 2014). Therefore, learning environments that include individual monitoring and continuous feedback are necessary (Marco-Bujosa et al., 2017).

Data from multiple sources confirms the effect of feedback on the argumentation process is limited. Standard statements (e.g., "good") do not contribute to the learning process (Adalberon, 2021), while individual biases can negatively affect it. PSTs tend to jump to conclusions quickly and may not use critical thinking sufficiently (Ma et al., 2025) and may struggle to generate constructive solutions to problems encountered during argumentation (Chen, 2008). This shows that argumentation requires high-quality and structured feedback.

In recent years, AI-powered tools have become widespread in the process of evaluating student arguments and providing feedback. These tools have been found to reduce student anxiety and improve self-efficacy and the organization of ideas (Kim et al., 2022; Ruwe & Mayweg-Paus, 2023). However, it has also been noted that AI-assisted feedback does not improve PSTS' arguments (Cai et al., 2025; Ruwe & Mayweg-Paus, 2023), and that the reasoning in their arguments may lack originality (Kim et al., 2022). Given PSTS' limited argumentation skills, there is a continuing need for well-structured, personalized, and pedagogically enriched feedback (Nazaretsky et al., 2023).

The role of AI in developing PSTS' argumentation skills is controversial due to the opportunities and limitations it offers. There is a gap in the literature regarding how PSTS interpret and apply AI feedback as well as their developmental needs in this area. Current studies mainly include a comparison of different assessment tools aimed at improving the argumentation skills of pre-service teachers (Rochmah et al., 2025; Guo et al., 2022), examined the effects of technology support to improve argumentation skills (Lin et al., 2026; Ucar-Longford et al., 2024) typically through pre-post comparisons. However, it is not well known how AI feedback affects high-level skills such as rebuttal and counterargument. In addition, how teacher candidates interpret and use this feedback over time has not been sufficiently researched. At this stage, we decided to use design-based research, a long-term research initiative that involves iterative feedback, design, implementation, and redesign, to generate practical solutions to potential problems that we may encounter while developing and implementing our design (Wang, 2020). In this study, in which the context of argumentation comes to the fore, it was aimed to establish a shared balance between the researcher-defined and candidate-defined goals. This balance was pursued while also considering two factors: the PSTS' limited experiences with argumentation and developments that emerged through the feedback

provided in each design cycle (Zinger et al., 2017). Similarly, a number of studies in the literature that use AI chatbots as both a learning and assessment tool to increase teacher candidates' cognitive engagement have revealed the experiences and feedback of teacher candidates through design-based research (Amir, 2026; Jaramillo et al., 2025; Zhang & Liu, 2026). This research aims to contribute to the development of pedagogically valid and needs-responsive AI tools by providing empirical data on the effects of AI tools on PSTS' argumentation skills. This study constitutes the first phase of a broader series of studies, and the findings obtained will guide the next phase, which is being conducted using a pre-test–post-test design.

2. Literature

2.1. *Feedback Dynamics in the Argumentation Process: Functions, Sources, and Pedagogical Implications*

Argumentation is the process of using reasoning to support or question a claim (Van Eemeren et al., 2013). In this process, learners support their claims with data and logic, evaluate counter-arguments, and construct an argument (Martín-Gámez & Erduran, 2018). Toulmin's Argumentation Model is commonly used to explain this structure (Toulmin, 2003). The model consists of claim, data, warrant, backing, qualifier, and rebuttal (Capkinoglu et al., 2020; Wang et al., 2024).

The strength of the Toulmin model lies in its ability to systematically analyze the structure of arguments (Cramer & Kempen, 2022; Erduran et al., 2004). However, the model does not adequately describe how students dynamically evaluate data (Kim & Kim, 2022) or identify implicit assumptions (Macagno & Konstantinidou, 2013). Therefore, it is often used as a scaffold providing a "horizon of possibilities" while acknowledging its limitations (Erduran, 2018). The analytic scheme of Erduran et al. (2004), derived from the model, evaluates argument quality ranging from Level 1 (simple claims) and Level 5 (arguments containing comprehensive refutations). Merely giving students grades is not enough; it is more beneficial for them to receive clear feedback on the strengths and weaknesses of their arguments (Carlile et al., 2018). Consequently, strong and reliable feedback is essential for learners to improve the arguments they produce (Morawski & Budke, 2023; Uçar & Demiraslan Çevik, 2020).

Feedback is a multifaceted concept encompassing information transmission (Wu, 2014), performance measurement, and guiding future development (Hattie & Timperley, 2007; Lipnevich & Panadero, 2021). There is broad consensus that for feedback to be effective, students must actively participate in making sense of it and applying it in their subsequent learning (Carless, 2022; Carless & Boud, 2018; Winstone et al., 2017). Integrating feedback into argumentation instruction is highly effective in developing their argumentation skills in PSTs. For example, the combination of peer feedback and argument mapping has been shown to strengthen PSTs' higher-level skills such as refutation and counter-argument development (Uçar & Demiraslan Çevik, 2020). Similarly, embodied and integrated feedback has promoted data-based counter-argument skills (Hornstein et al., 2025). Furthermore, discussions, peer feedback, and reflective assessment have enabled PSTs to advance from content-based explanations to value-based warrants (Christenson & Walan, 2023). Consequently, well-structured feedback is an essential tool for prospective teachers to construct more critical and nuanced arguments.

2.2. *AI-Supported Feedback: A Viable Pedagogical Solution or a Superficial Approach?*

The expansion of feedback sources now also includes AI as an emerging area of instructional support. AI has developed significantly in machine learning and natural language processing, enabling new learning uses (Say, 2024). Educational applications of these technologies are well-suited for educational implementation by their capability to process written texts and provide instant computerized feedback (Kavitha et al., 2023; Özmütlu, 2021). Along with supporting learners, AI gives teachers powerful means to create content and offer personalized feedback at a previously unmatched level (İşler & Kılıç, 2021). Natural language processing-based systems, in particular, have demonstrated powerful capacity in assessing written arguments, offering automated feedback, and being useful in augmenting formative assessment practices (Engeness & Gamlem, 2025; Hopfenbeck et al., 2023).

AI-supported tools have increasingly been used to provide feedback to students on the process and the product of argumentative writing. AI-supported tools help students at the planning, outlining, and revising

stages of argumentation, resulting in the efficient use of concepts, strategic selection of data, and logical reorganization of arguments (Kim et al., 2022). However, literature reflects a dualistic perspective. For example, AI supported tool was found to reduce students' writing apprehension and enhance their self-efficacy but remained insufficient in helping them elaborate on their arguments more extensively (Ruwe & Mayweg-Paus, 2023). Similarly, while AI-assisted argumentation tools assisted the students in organizing their ideas, their directive and template-driven structures were found to be perhaps confining independent thinking and reasoning (Cai et al., 2025; Kim et al., 2022). Conversely, other research has discovered that students provided with AI-supported feedback on claims, warrants, and data wrote with more components, integrated them in a more complete way, and exhibited significant improvements in the quality of argument (Zhu et al., 2017; Zhu et al., 2020). These improvements were particularly significant for students who started at a lower level and needed several revisions. AI-assisted collaborative argumentation with generative AI also promotes a multiperspectival understanding of argument and allows students to apply argument elements more methodically (Liu & Cui, 2024). Finally, though a tool called ArgueBot provided promising real-time, individualized feedback, researchers emphasized the need for larger studies to evaluate its effectiveness systematically (Wambsganss et al., 2021).

2.3. Comparing the Feedback Capacities of AI Models: Accuracy, Style, and Alignment with Human Feedback

A recent comparative analysis revealed the differences between the level of accuracy and the reliability of feedback given by various tools for AI. For instance, for the analysis of the writings of Chinese language learners, it was found that the tools such as 'ChatGPT' and 'Bard' are superior in the sense of grammatical and formal correctness; whereas 'Claude' showed a very empathetic and expository (Obaidoon & Wei, 2024). A comparative study was conducted by (Uppalapati & Nag, 2024), where the performance of 'ChatGPT,' 'Bard,' 'Bing Chat,' and 'Claude' for questions on medicine and public health was evaluated in the aspects of correctness, comprehensiveness, and clarity. The results showed the consistency of 'ChatGPT' in correctness, 'Bard' for the depth of information, and 'Claude' for the description of the context, but 'Claude' lacked the aspects of correctness. The reason for the consideration of 'ChatGPT' for the analysis might be the consistency in grammatical and formal correctness for the proper analysis of results.

Additionally, studies are available in the literature comparing the quality of the feedback delivered by ChatGPT and human feedback. A study by Steiss et al. (2024) compared the quality of the feedback delivered to students for their written assignments and found human feedback to score higher on the grounds of accuracy and focusing on the major details. However, Yu and Xie (2025) found the quality of the grammatical error feedback delivered by the chatbot to be as high as the quality of the teachers' feedback, yet the feedback delivered additional information about the content and organization of the written piece. Students' feedback implementation rates remained equal for both teachers and the chatbot; the rate for the teachers' feedback stood at 78%, whereas the rate for the chatbot stood at 70%. A study by Escalante et al. (2023) revealed the quality of the chatbot feedback to be equal to the quality of the teachers' feedback in improving the students' writing skills, whereas the study by Usher (2025) revealed the quality of the teachers' feedback to be superior to the chatbot in improving the skills for the complex task.

2.4. Current Study

Studies prove that AI feedback can foster the identification of structural aspects of argument (Cai et al., 2025), but not significantly impact the depth of their arguments (Kim et al., 2022). PSTs need to develop sound argumentation skills, as these will enable them to provide inquiry-based and data-based learning experiences in their classroom. However, this reveals a qualitative imbalance: there are few studies on how PSTs process and use AI feedback to implement these improvements. To address the imbalance, this study is guided by the following research questions:

1. How do PSTs' perceptions of artificial intelligence change throughout their participation in the AI-supported feedback?

2. In what ways does the AI-supported feedback influence PSTs' argumentation levels and the use of argument components in scenario-based discussions?
3. What specific affordances and constraints do PSTs experience when receiving AI-supported feedback during scenario-based argumentation discussions?
4. Which design principles have been identified for the improvement of AI-supported feedback?

3. Methodology

3.1. Research Model/Design

This study applied Design-Based Research (DBR) methodology, with the aim of designing new learning environments and empirically examining their effect Brown (1992). Conducted through researcher-practitioner collaboration, DBR aims to develop practical solutions for real-world educational problems (Anderson & Shattuck, 2012).

DBR involves a four-stage process: (1) analyzing a problem, (2) developing a solution prototype, (3) iterative testing and refinement, and (4) final generalization of principles (Reeves, 2006). Through these iterative cycles, DBR has been shown to systematically enhance both feedback processes and students' argumentative writing skills (Reeves, 2006). Through undergoing such recursive cycles, it has been proven that DBR systematically fosters feedback processes along with students' argumentative writing ability (Gu et al., 2022; Guo et al., 2025; Schwartz et al., 2024). For this reason, an AI-rich pedagogical environment to enhance the argument-writing abilities of PSTs was created in this research. The initial draft of the design was discussed with a small group of PSTs, and based on the feedback received, the design was updated. After resolving these errors, the instructional environment was set. This final version was then implemented with another group of PSTs, and how AI-supported feedback was studied. The AI-supported argumentation feedback (AIAF) for science PSTs is a novel pedagogical tool, which was developed and tested through field applications in this current study.

3.2. Context and Participants

Participants were nine PSTs (7 female, 2 male) who were purposively selected according to criteria of voluntary registration and willingness. Participants were selected from students enrolled in the undergraduate courses "Science Education Programs" (FBO250), and General Physics 1 (FBO101) for which the first author is responsible. At the beginning of the semester, all course participants (approximately 40 individuals) were informed verbally about the purpose, process, rights, and confidentiality conditions of the study, and a written information form was distributed. Following this information session, those who wished to voluntarily participate in the study contacted the first author via email within one week. From among the applicants, nine PSTs who met the predefined cohort criteria (academic class and prior knowledge of argumentation) were selected as the final participants. A total of three different groups of teacher candidates took part in the two cycles of the study. Two of these groups (n=6) consisted of second- and third-year students who had prior knowledge of argumentation gained from the courses they took during their undergraduate education and took part in Cycle 1. The rich and diverse argument components of these groups were used in the iterative development of AIAF prompts. On the other hand, the group participating in Cycle 2 (n=3) consisted of first-year teacher candidates who had no prior knowledge of argumentation. This group was deliberately selected because their lack of prior argumentation knowledge allowed the effect of AI feedback to be observed more clearly, independent of pre-existing knowledge. All of the necessary ethical permissions (Protocol No: 2025/10-XXVII) were obtained and written informed consent was received from all participants prior to the study.

3.2.1. Cohort 1 (Cycle 1: AIAF Prompt Development and Validation)

They were a group of six second- and third-year PSTs who possessed prior knowledge of argumentation theory, having taken the "Argumentation in Science Education." Their background was central to the first cycle, in that their role was to critique, challenge, and authenticate the very first AI-given feedback and prompts and provide expert-level commentary for system development. How the six participants who make

up Cohort 1 were divided into two subgroups and the place of each subgroup in the implementation process are summarized in Table 1.

Table 1.

Cohort 1 (Cycle 1) Participant Groups and Implementation Process

Group	Number of Participants	Scenario	Purpose
Group 1	3 PST	Earthquake scenario	Pilot testing of AIAF's audio/text feedback functions; Detection of code mismatches caused by interruption and social interaction
Group 2	3 PST	Vaccination scenario	Validation of "merging dispersed arguments " and "interactive argument" prompts

3.2.2. Cohort 2 (Cycle 2: Main Implementation with the Validated AIAF)

This cohort consisted of three first-year PSTs (PST-1, PST-2, and PST-3) who were at the initial year of their teacher education. Unlike Cohort 1, they hadn't taken the argumentation-based courses, and therefore their initial experience regarding argumentation was minimal in comparison to others. This novice profile was particularly selected for the main implementation to rigorously test the efficiency of the completed AIAF with a new cohort of PSTs in enabling and promoting argumentation skill from a baseline level. For anonymity and clarity reasons, the elaborate individual analysis in the following sections only presents the analysis of the three Cohort 2 participants.

3.3. Data Collection Instruments

3.3.1. Semi-Structured Interviews

A semi-structured interview, developed by the researchers based on recent literature (e.g., Bauer et al., 2023; Engeness & Gamlem, 2025; Zhou et al., 2024), was used to track changes in participants' knowledge. The guide touched on three topics: (1) views of AI, (2) experiences of AI-facilitated feedback, and (3) evaluation of the strengths and limitations of the AIAF. Pre- and post-interviews (approximately 30 minutes each) were conducted, audio-taped, and transcribed verbatim. Some of the interview questions are summarized in Appendix A.

3.3.2. Scenario-Based Argumentation Discussions

Two scenario-based tasks were developed-one in earthquake engineering and another one in vaccination. Both were specifically constructed to guide PSTs through consideration of the crossing of technical solutions with their socio-ethical implications, for instance, responsibility and equity. The science component involved PSTs in assessing domain-level innovations such as "floating foundation" and "mRNA technology" to construct arguments from scientific data. Upon exploring the social component, the scenarios highlighted society's balances and justice issues. The earthquake scenario revolved around the balance of cost versus building safety, and the vaccine scenario focused on the balance of choice and public health advantages. Surely, these generated arguments regarding mandates for vaccines and access equity. The scenario-based argumentation discussions were recorded using a voice recorder. The recordings were transcribed into written text by researchers and compared with the original recordings for accuracy. This process took approximately one day. The day after the transcription was completed, the participants were invited to individual feedback sessions. The participants' arguments showed distinct differences in length depending on the level. Arguments at the basic level (Level 1) consisted of a single sentence and 4–6 words, while the most developed arguments (Level 5) reached 2–3 sentences and 76 words. Arguments at the intermediate levels (Levels 2–4) ranged from 2–3 sentences to 34–54 words.

3.3.3. *AI-Generated Argument Analysis Reports*

The AIAF generated automated analysis reports for each PST argument, from the refined prompts of Cycle 1. Two important data points were presented in the reports:

- The frequency of Toulmin argument components (warrant, data, claim, etc.).
- The argument's quality according to the framework proposed by Erduran et al. (2004).

These reports were crucial to the computation of inter-rater agreement during the tool's development stage (Cycle 1) and served as the main quantitative data source for monitoring changes in argument structure and quality between the first and second scenarios.

3.4. *Cycle 1: Development and Validation of the AIAF*

3.4.1. *Design*

In the first cycle of this study, the researchers focused on developing appropriate prompts for AI-powered feedback. Prompts to be used with ChatGPT (GPT-4) are designed to evaluate argument components (claim, warrant, data, etc.) and argument levels. The prompts consist of two dimensions. The first dimension aims to determine the components of the argument, and the second dimension aims to evaluate the quality of the argument. In this research, Toulmin's Argumentation Model (1958) was used for the argument components dimension of claims. While some of the studies in the literature based on Toulmin's Argumentation Model only consider whether the components are present (Diola et al., 2025; Lin & Mintzes, 2010), some of which have scored the quality levels of these components; that is, beyond whether the student presented warrants or data, it was examined how detailed, clear and based on scientific principles they were (Hardcastle et al., 2021; Schen, 2013). In other rubrics, alternative explanations and their ability to refute opposing views were measured (Viyanti et al., 2017). In addition, classifications were made from arguments containing only claims to advanced arguments containing multiple strong rebuttals using five-level quality analyses (Fakhriyah et al., 2022; Öztürk & Okumus, 2022). In this study, since it was determined whether these components were present rather than the quality of the argument components of the candidates, it was ensured that the artificial intelligence focused only on the components while writing the prompts. For this purpose, ChatGPT (GPT-4), a pre-trained large language model, is directed to identify the candidates' arguments through structured prompts for the pre-service teachers' arguments.

Erduran et al. (2004), argument level classification was used to evaluate the quality of the argument, which is the second dimension of the claims developed using Toulmin's Argumentation Model (1958). To this end, ChatGPT (GPT-4) is directed to conduct quality assessments through structured prompts for pre-service teachers' arguments. In both prompt sets, ChatGPT was not trained, but only prompted. It was decided to conduct trials to test whether the AI would give voice or written feedback while providing feedback.

During the writing process of the claims, support was received from both field experts (faculty members of the Faculty of Science Education, CEIT and Engineering) and people experienced in artificial intelligence applications; claim writing principles (clear, specific, contextual, step-by-step, specifying length and depth) have been clarified in line with these views.

3.4.2. *Implementation and Initial Evaluation*

In the first applications, group 1 of three pre-service teachers participated in a scenario-based discussion using the scenario on earthquake. To pilot-test the AIAF's feedback functions, data from two of these three participants were used: the AI first attempted to provide feedback to one pre-service teacher using its voice response feature. Then, the written texts of the second pre-service teacher were uploaded to the AI-supported feedback tool as a file, and the tool analyzed the students' argument components within the framework of Toulmin's Argumentation Model (1958) and the argument levels according to the classification of Erduran, Osborne, and Simon (2004).

The voice response feature has been found to be under-functional due to its inability to accurately detect conversations, generate inappropriate responses, and provide timely feedback. As noted above in the second analysis, the functionality of the prompts was tested through written texts, and the deficiencies in the feedback given by AIAF were determined. For this purpose, the similarity between the argument components determined by AIAF and the researcher codes was calculated with percentage agreement. The coding unit was determined as each sentence/statement in the speech transcripts and was independently coded by two researchers according to the components of the Toulmin Argumentation Model (Claim, Data, Warrant, Backing, Qualifier, Rebuttal). Percentage agreement is a metric often used in AI-human comparisons. For example, Prescott et al. (2024) reported a 36–47% agreement between AI-human thematic coding. Similarly, Lippi and Torroni (2015) examined the rate at which different coders similarly labeled the same text in their argument mining studies and reported agreement of 58–67%, and in some cases 84%.

The percentage agreement between the AIAF and the researcher was calculated by dividing the number of units agreed by both coders by the total number of units evaluated at that stage and multiplying by 100. This formula was applied separately at each stage of the pilot testing process. In the sections that follow, each reported agreement rate is presented along with the number of units agreed upon and the total number of units assessed. In this study, it was found that the agreement rate between the AIAF and the researcher was 97% when a student continued the participant's argument without interruption from the beginning to the end of the speech. This rate is based on consensus in 97 out of a total of 100 units evaluated at this stage. In this study, the agreement rate between AIAF and the researcher was found to be 97% only when a student formed an argument by continuing the conversation from the beginning to the end without interruption. However, it was observed that when a student's speech was interrupted by another student in the written text, AIAF did not include the argument components of the first speaking student in the analysis. Therefore, in such cases, the agreement rate between AIAF and the researcher has decreased to 29.76%. This rate is based on consensus in 25 out of a total of 84 units evaluated at this stage. On the other hand, it was determined that when a student contributed to another student's argument or presented a dissenting opinion during the speech, the AIAF did not include these contribution components in the analysis. For example, a participant's statement in support of a data previously presented by a peer was not identified by the AIAF as a separate argumentative contribution, but human coders identified it as a supporting component linked to the ongoing discussion. This has led the tool to consider only individual statements, leaving out argumentative elements arising from social interactions. At this stage, the percentage of agreement between AIAF and the argument components determined by the researcher was found to be 72%. This rate is based on consensus in 144 out of a total of 200 units evaluated at this stage. Coding differences between researchers (e.g., disagreements such as whether a component should be considered separate or combined) were discussed and consensus was reached, and the prompts directed to the AIAF were updated in the process. Due to this ratio, the accuracy of the feedback on the quality of the argument could not be calculated.

3.4.3. Revision

In addition to using the voice response feature of the AIAF, group discussions were tape-recorded, transcribed, and AI was asked to generate feedback based on text transcripts. Nevertheless, listening to the text feedback in audio form remained optional to the participants.

In the initial analyses, the system treated every sentence as an independent unit, regardless of its length or complexity; this led to misclassifications, particularly in the case of long, multi-component arguments that spanned multiple sentences. The tool prompts were altered in two stages. Initially, a new prompt titled "merging dispersed arguments" was developed. This prompt requested the AI to recognize that components of the same argument might be spread across various sentences, to consider inter-sentential relationships, and to assemble all components belonging to one argument. Following the initial update, the updated prompt was validated using a total of 25 argument components derived from the discussion data of a pre-service teacher in group 2. The agreement rate between AIAF classifications and researcher coding was calculated as 96%. This rate is based on consensus in 24 of the 25 units evaluated. It should be noted that this agreement rate reflects the agreement between the AIAF's automated coding and the researcher's

manual coding, and is a different metric from the inter-investigator reliability reported in the thematic analysis section. Percentage agreement was preferred as the metric because it allows for direct comparison with the sample sizes and results reported in previous AI-human comparison studies (Prescott et al., 2024; Lippi & Torroni, 2015) were preferred because they were directly comparable. This improvement enabled the tool to analyze a participant's argument more comprehensively. Subsequently, to include peer interactions in analysis, another prompt called "interactive argument" was added. This prompt was added to recognize a PST's contribution to, or counterargument against, other PST's arguments as new elements of an argument. Following this revision, the percentage of agreement between the argument components identified by the written AIAF and those ascribed by the researcher had climbed to 98.51%. This rate is based on consensus in 132 of the 134 units evaluated. The final agreement rate (98%) exceeds the ranges reported in prior AI-human comparison studies (36–47%, Prescott et al., 2024; 58–84%, Lippi & Torroni, 2015), suggesting that the revised AIAF prompts achieved a notably higher level of coding accuracy. As a result, it became possible to study arguments built in discussion situations as a whole, involving individual and social interaction levels. This revised prompt was then tested using data from the same group of three pre-service teachers, who took part in a discussion on the vaccine scenario at a later point in time, with their conversations being recorded and added to the AIAF. The system was shown to be capable of accurately analyzing both individual arguments and peer conversation. Following the revision phase, full agreement was achieved between the AIAF and the researcher on all 28 arguments, which were evaluated according to the argument level typology of Erduran et al. (2004). The results that were gained during this cycle indicated that analysis of both argument components and argument levels had reached a satisfactory level to be included in the application during the second cycle.

3.5. Cycle 2: Main Implementation with the Validated AIAF

3.5.1. Design

In the second cycle, the goal was to utilize the prompts created and tested in the first cycle to allow the AI to have a conversation with participants and offer individualized feedback. This cycle was carried out with the new group of participants (Cohort 2), who were first-year PSTs.

3.5.2. AI-Supported Argument Feedback Session

After loading the participants' argument transcripts onto the AIAF, each participant was provided with a personal feedback session according to a step-by-step protocol. The session began with an introduction to basic concepts of argument and argumentation, followed by an introduction to Toulmin's Model and Erduran et al.'s (2004) proposed levels of argument quality, which were discussed with detailed examples. PSTs' own arguments were then mapped onto Erduran et al.'s levels and Toulmin components, for instance, highlighting a statement as "Level 2: Claim, Data, Warrant." The argument level assessment was performed automatically by the system. To enable onward development, an example of argument demonstrating development to the next level (e.g., Level 3) was provided. The session concluded with multiple-choice questions, together with follow-up questions to assess understanding more extensively. A sample copy of this feedback mechanism is included in Appendix B.

At the end of the implementation, PSTs were given a comparative development report and motivation tool—a graphical report comparing their performance on the earthquake and vaccine scenarios side-by-side, with the breakdown of argument levels and overall development assessment. Appendix C presents the entire framework for argument development assessment and supportive feedback.

3.5.3. Personalized Action Plans

Following the feedback session, the AI generated personalized action plans along two main axes: Individual Development Suggestions (e.g., writing short arguments from current news, keeping an "argument journal") and Collaborative Development Suggestions (e.g., engaging in discussions by assuming different roles like claim defender or opponent, evaluating peers' arguments). Appendix D presents actual samples of the personalized action plans. The plans were tailored to each participant's level of argumentation and areas needing improvement. Each plan included two components: concrete example statements illustrating the

area needing improvement, and actionable steps to develop that skill. This allowed participants to clearly see both what they needed to improve and how to go about doing so. A before-and-after example of the argument development process of PST-3, one of the participants, is presented in Appendix E.

3.5.4. *Evaluation and Revision*

Individual post-interviews were conducted to elicit participant views on the AIAF process, the ease of understanding of the comparative report, and the viability of the individual action plans. Both qualitative and quantitative data obtained were used in experimenting with the feasibility of the design. The results of the evaluation indicated that the AI dialogues enabled individual interaction and growth. However, to further develop critical thinking capabilities, it was identified that the action plans needed to be augmented with more concrete examples and more applicable to day-to-day life scenarios. Thus, the AI prompts were redesigned for a final time to provide PSTs with strategic questions for group debate and more direct activity prompts for individual study.

3.6. *Data Analysis*

The first, third, and fourth research questions were examined using thematic analysis (Braun & Clarke, 2006; Finlay, 2021). The audio recordings of the interviews were transcribed verbatim and the analysis was carried out manually. The analysis focused only on interview data directly related to the experiences and perceptions of pre-service teachers regarding the AI-powered feedback process. Accordingly, only data segments that address these experiences and perceptions are included in the thematic analysis. Through an inductive approach, the two researchers read the transcripts repeatedly, primarily to gain an in-depth exposure to the data, creating data-driven initial codes without adhering to predetermined categories. The coding unit was defined as the smallest piece of text (unit of meaning) that expresses a single idea, experience, or assessment related to research questions. When a participant expressed more than one independent opinion within the same response, each was treated as a separate coding unit.

The dataset encompasses nine transcripts from two separate sets of interviews with three Cycle 2 participants: six pre- and post-implementation interviews that feed into the theme of "Reconceptualizing AI," and three follow-up interviews that feed into the theme of "PSTs' Experiences and Views on the AIAF." A total of 186 meaning units were identified in the entire dataset, 92 of which were within the scope of the first theme and 94 of them were within the scope of the second theme. The data were analyzed for each pre-service teacher in turn, starting with the first participant. After each participant's data was coded, the resulting codes and candidate themes were continuously compared with those identified in previous transcripts. Following the analysis of the third participant, no new code or theme emerged, and the existing coding framework was found to adequately reflect the full range of participant experiences. For this reason, data collection and analysis for this cycle were terminated, and data saturation was accepted (Fusch & Ness, 2015).

Inter-coder reliability between the two researchers was evaluated separately for each theme by percentage agreement method, taking into account the nature of the coding unit and the limited sample size. Agreement was calculated as 93% for "Reconceptualizing AI" and 91% for "PSTs' Experiences and Views on the AIAF", with an average agreement rate of 92% across the dataset. This value represents only inter-coder (human) agreement and should not be confused with the agreement between the AI Argumentation Feedback (AIAF) classifications and researcher coding reported in section 3.4.3. Coding discrepancies, which were limited in number, were resolved through discussion until consensus was reached through discussion. The agreed coding framework was then used to organize conceptually related codes into broader categories and to develop candidate themes from these categories. These themes have been reviewed iteratively, comparing them with both coded citations and the entire dataset, to ensure internal coherence, clear separation between themes, and alignment with research questions. The thematic structure was refined until the final themes accurately reflected participants' experiences of the AI-powered feedback process.

To answer the second research question, a three-stage analysis integrating quantitative and qualitative methods was employed. In the first stage, the arguments of the pre-service teachers were classified according to the five-step argumentation levels framework developed by Erduran, Simon and Osborne

(2004) by the Artificial Intelligence-Supported Feedback tool (AIAF). In this context, Level 1 represents simple arguments consisting only of claims, and Level 5 represents multi-component high-level arguments that take into account opposing views and include rebuttal and qualitative elements. These classifications created by artificial intelligence were verified by the researchers and included in the analysis process.

In the second stage, the Average Level Score (ALS) was calculated in order to determine the general argumentation tendencies of the teacher candidates. In this process, looking only at the number of arguments was not considered sufficient to reflect the development of the candidates; because this approach ignores the quality level of the components. Indeed, although more argument components were produced in the earthquake scenario and fewer in the vaccine scenario, this quantitative difference alone cannot explain the change in the quality of the arguments. For this reason, ALS was calculated using coefficients corresponding to each argument level (Level 1=1 point, Level 2=2 points, ..., Level 5=5 points); the total score obtained in each scenario was divided by the number of arguments in the relevant scenario and the average argumentation level of the teacher candidates was determined.

To examine qualitative development, argument components (claim, data, warrant, counterargument, etc.) were thematically analyzed; warrants were categorized as functional/economic/health-based, and counterarguments as superficial/multidimensional. In order to increase the validity of the analyses, direct student quotations from the discussion transcripts were also included in the analysis process. These quotations illustrated the most prominent examples of low, medium and high-level arguments from each participant in different scenarios. Reliability was assessed through independent coding by two researchers, and as a result of this process, agreement was achieved at rates of 93% and 91% (average 92%).

As the study was conducted in accordance with the cyclical structure of DBR, the data obtained not only revealed the strengths and weaknesses of the current process but also informed the development of design principles to guide subsequent cycles. These principles, derived from the experiences and views of teacher candidates which were grounded in patterns directly emerging from the data, provided actionable recommendations for improving the system.

4. Findings

4.1. *Reconceptualizing AI: From a Tool of Utility and Risk to a Complementary Aid*

PSTs' perceptions of artificial intelligence (AI) were mixed in the early stages; they saw its potential but expressed serious concerns about reliability and long-term effects. Its value was emphasized by PST-1, who stated that "it makes assignments and can also prove useful for experiments in calculations." PST-2, on the other hand, took a contradictory approach, saying "artificial intelligence makes me happy because it makes my job easier... but when I think about the future, it scares me." On the other hand, PST-3 drew attention to the reliability problem of artificial intelligence and expressed its concerns with the words "it can make reliability and accessibility nervous". Following the intervention, the PSTs' attitudes evolved toward one that was more pedagogically sophisticated and balanced. They have always positioned AI as an adjunct to teachers rather than as a substitute for them. PST-1 was insistent, "it cannot replace the teacher, but it can complement," an opinion precisely stated by PST-2: "AI-supported feedback cannot replace teacher feedback; it can only be a complement." PST-3 also supported this stance, stating that "it has to be seen as a complementary tool; it cannot replace the teacher." While reliability and passivity continued to be issues, they were now formulated in terms of AI as a second and secondary function to human command.

4.2. *Changes in PSTs' Argumentation Levels*

As shown in Table 2, the number of arguments at the lowest level (Level 1) decreased markedly from the earthquake scenario to the vaccination scenario (from 13 to 8), whereas the number of arguments at the highest level (Level 5) more than doubled (from 1 to 3) over the same transition. These findings suggest that pre-service teachers move away from superficial claims and turn to more advanced arguments that take into account opposing views and include a rebuttal component. The limited decline in arguments at the intermediate level (Levels 2 and 3) and the constancy at Level 4 suggest that this development has occurred in a gradual and balanced manner. In addition, although the number of arguments decreased from 28 to 23

in total, the increase in the Average Level Score (ALS) from 2.04 to 2.48 indicates a qualitative progress. This finding suggests that this progress in argumentation quality, together with content differences between scenarios and other instructional elements used in practice, could be one of the possible contributions of AI-supported feedback.

Table 2.

Argument Level Distribution and Average Level Score Across Scenarios

Argument Level	Earthquake Scenario	Vaccination Scenario
Level 1	13	8
Level 2	7	6
Level 3	3	2
Level 4	4	4
Level 5	1	3
Number of Arguments	28	23
Average Level Score	2.04	2.48

Table 3 details the changes in the individual PSTs' ALS values from the earthquake scenario to the vaccine scenario. Although the ALS of every PST rose from the earthquake scenario to the vaccine scenario, the outcome revealed the change among the individual PSTs happened at various levels. For instance, PST-2 indicated the greatest level of enhancement from 2.00 to 2.57, then PST-3 from 1.89 to 2.38, and lastly PST-

1 from 2.25 to 2.50. This development may also reflect the potential contributions of AI-supported feedback, along with the differences between the scenarios and other instructional elements included in the implementation.

Table 3.

Individual participants' development in argumentation levels

PSTs	ALS (Earthquake)	ALS (Vaccination)
PST-1	2.25	2.50
PST-2	2.00	2.57
PST-3	1.89	2.38

4.2.1. Argumentative Development: Analysis of the First PST

In the earthquake and vaccine scenarios, some of PST-1's arguments are low-level constructs consisting of claims, data, and warrants. In the earthquake scenario, the statements "buildings will be needed" and in the vaccine scenario, "something that has increased life expectancy" reflects this situation. Therefore, both arguments have been evaluated at Level 2. At this level, the warrants are formulated as "functional and architectural aesthetics" in the earthquake scenario and "life span extension due to protection from diseases" in the vaccine scenario.

In both scenarios, PST-1 diversified the argument components by comparing advantages and disadvantages and using counterarguments. In the earthquake scenario, "the advantage is that it is suitable for today's architecture and more durable; the disadvantage is the cost. If this system is implemented in most places, the cost will decrease. "This explanation contains the components of claim, data, warrant, counterargument ("high cost"), and qualifier ("if this system is implemented in most places") and corresponds to Level 3. Here, the warrant is explained in economic terms; the counterargument points to the cost disadvantage, while the qualifier indicates a conditional possibility.

PST-1's strongest arguments were again observed in the earthquake scenario. "This project I'm talking about is a whole package... The system in Project 1 may not work or may not be properly monitored. But the pendulum is already a constantly swinging structure..." contains the components of claim, data, warrant, refutation, and support, and is positioned at Level 4. Here, the warrant is based on technical operation, the refutation on the lack of control of the system, and the support on the durability of the pendulum structure.

In the vaccination scenario, the statement “I also agree to be vaccinated... I was saved thanks to my tetanus shot... I’ll wait a year or two for the new vaccine” combines claim, data, warrant, qualifier, and counterargument, reaching Level 5. The warrant is based on personal experience; the refutation is a cautious approach, and the qualifier contains a conditional statement.

Overall, it was observed that PST-1 constructed medium-level arguments in the earthquake scenario and more advanced, multi-component arguments in the vaccine scenario, showing significant progress, especially in the vaccine scenario.

4.2.2. *Argumentative Development: Analysis of the Second PST*

In discussions based on earthquake and vaccine scenarios, many of PST-2’s arguments are low-level constructs based on limited components. In the earthquake scenario, the statement “I also chose the first project because it starts immediately within half a second during the tremor...” is an example of this situation; in the vaccine scenario, the statement “I thought vaccines were important in terms of herd immunity... but they may not have the same effect on everyone” is another example. Since both arguments only contain claim and warrant components, they were evaluated at Level 2. In the earthquake scenario, the warrants are “functional and one-dimensional” (e.g., quick response), while in the vaccine scenario, they are based on “herd immunity.” No counterarguments were observed at this level, and qualifiers were limited. Supporters in the earthquake scenario leaned toward “subjective” (“I think it’s a good advantage”) warrants, while those in the vaccination scenario leaned toward ‘concrete’ (“infant vaccination still protects”) warrants.

In both scenarios, PST-2 increased the variety of arguments by using advantage-disadvantage comparisons and counterarguments. The statement in the earthquake scenario, “It is more affordable than the third project in terms of cost... the only disadvantage is in terms of durability,” demonstrates Level 3 by including claim, data, and counterargument components. The warrants are economic, and the counterarguments are superficially focused on “durability.” In the vaccine scenario, the statement “You watch those who get vaccinated... I would observe whether the vaccine worked” corresponds to Level 3 again with observation-based data. At this level, the earthquake warrants are “economic-practical” and the vaccine warrants are “observational-empirical.”

High-level arguments were observed in both scenarios. In the earthquake scenario, cost- and feasibility-based explanations were supported by refutations that acknowledged opposing elements, exhibiting a structure approaching Levels 4–5. In the vaccine scenario, the statement “Vaccine opponents... but they can be persuaded with numerical data” is at Level 5, containing a claim, data, warrant, and multiple refutations.

Overall, PST-2 showed consistent progress between Levels 3 and 5 in the earthquake and vaccine scenarios, with a decrease in superficial structures; warrants developed in an “economic” manner in the earthquake scenario and in a “social and data-driven” manner in the vaccine scenario.

4.2.3. *Argumentative Development: Analysis of the Third PST*

In earthquake and vaccine scenarios, some of PST-3’s arguments are low-level constructs based on limited components. In the earthquake scenario, the statement “I remember shock absorbers being used in Japan...” consists solely of a claim, while in the vaccine scenario, the statement “I think quality of life will improve and contagion will decrease” contains a claim, reasoning, and data, but no counterarguments. Therefore, the statements were evaluated at Level 1 and Level 2, respectively. The low-level warrants are based on “personal observation” in the earthquake scenario and “social benefit” in the vaccine scenario. While there are no refutations, the qualifiers are limited, and the supporting elements are subjective.

As the process progressed, the argument components of PST-3 diversified. In the earthquake scenario, the statement “The average cost... is 250 kg per square centimeter... I can say that as a disadvantage” is at Level 3, containing a claim, data, warrant, and a weak counterargument. The warrants are based on “technical-economic” factors, while the counterarguments are based on “lack of resilience.” In the vaccine scenario, the statement “I saw that mercury, ether, and aluminum were used in vaccines but are no longer used” is

also at Level 3. Here, the warrant is “information comparison,” while the counterargument is directed at “current safety practices.”

In high-level arguments, PST-3 has taken opposing views into account. In the earthquake scenario, the statement “Our friend PST-1 is right... but not all buildings are the same size, so the first project may be logical” is at Level 4, refuting the opposing view and containing a conditional qualifier. In the vaccine scenario, the statement “80% have been vaccinated, 363 people have fallen ill... we need to talk about the data” is a strong Level 4 structure based on statistical data.

The most advanced arguments were seen in Level 5. In the earthquake scenario, the statement “The first project is weak in terms of resilience... when calculated with the coefficient, the difference increases.” was used, while in the vaccination scenario, the statement “The disease is not the same as it used to be... there is a greater need for vaccination due to population, GMOs, and pollution.” was used. This last argument represented the highest level with multiple reasons, supporting data, and two counterarguments.

In general, PST-3 transitioned from low-level structures to statistical, multifaceted, and counterargument-based arguments; it developed technical-economic reasons in the earthquake scenario and health and social-based reasons in the vaccine scenario.

4.3. PSTs' Experiences and Views on the AIAF

Findings indicate that artificial intelligence-supported feedback (AIAF) may have contributed to the development of metacognitive awareness toward argument analysis skills among preservice teachers. Participants stated that the tool allowed them to “see aspects they hadn't noticed before” and to identify their “gaps in counter-arguments.” One participant noted that this awareness translated into a concrete development plan, saying, “I will construct more cautious arguments in future discussions by using stronger counter-arguments.”

The system's “instant response” and interactive dialogue structure were among the most appreciated features. The text-to-speech feature provided participants with a “more intimate and focused environment.” As one participant highlighted, the system's “continuous questioning on incorrect answers” facilitated learning. The individual action plans presented to the candidates were effective in forming these opinions. The following statement of PST-2 embodies the contribution of individualized feedback to learning: “Analyzing my own sentences and the fact that the analysis was specific to me helped me understand the concepts better.” PSTs predicted that this technology could serve as supportive aid in “regular group work” and “peer discussions” in the classroom. However, criticisms directed at the current system suggesting it “feels like it's repeating the same things to everyone” and that it should offer “more personalized suggestions.”

Despite these positive features, participants agreed that AI cannot replace teachers. Comments such as “Teacher feedback is more effective; AI can only be an assistant” and “receiving feedback from someone you are constantly interacting with is more realistic” highlighted that technology can only have complementary value in a holistic learning environment guided by a teacher.

5. Discussion

Research findings showed that teacher candidates' perceptions of feedback evolved from being a functional tool to gaining a pedagogical and emotional dimension. The perception of artificial intelligence (AI) shifted from the initial benefit-risk dilemma to positioning it as a complementary support. A qualitative improvement in the participants' argument of quality was observed. While speed, interaction, and motivation emerged as AI's strengths, lack of personalization and repetitive expressions were identified as areas for improvement. These results highlight the potential and limitations of integration.

5.1. Positioning AI in Education: From Ambivalent Technology to Complementary Tool

With respect to attitudes towards artificial intelligence, PSTs initially had a combination of hope about its potential usefulness and fear of potential harm. When the intervention ended, they consistently reinterpreted it as a supportive tool that maximizes but does not substitute for the instructor.

PSTs' perception of artificial intelligence (AI) has evolved from an initial benefit-risk dilemma to viewing it as a complementary support tool. Thanks to personalized and interactive feedback, PSTs have moved beyond seeing AI as an uncertain technology and now define it as a functional partner that contributes pedagogically. PSTs view artificial intelligence's facilitation of access to information positively. This finding aligns with the study by Alejandro et al. (2024). That study also found that positive attitudes toward artificial intelligence are reinforced, particularly by perceptions of ease of use and usefulness. However, candidates also express concerns about the diminishing role of teachers and the devaluation of the profession. Haseski (2019) points to a similar pattern, showing that teacher candidates simultaneously express expectations of benefit and concerns about the devaluation of the profession. While previous research generally depicted the current acceptance levels and determining factors of pre-service teachers, this study revealed that pre-service candidates' initial perceptions of uncertainty and anxiety-oriented transformed into a perception of a complementary tool that contributes to the teaching process through experience.

5.2. Advancing Argumentation: A Qualitative Leap in Reasoning

The research reveals that teacher candidates construct simple arguments on the topic of earthquakes and advanced arguments based on health, society, and science on the topic of vaccines. The use of various refutations and qualifiers in the vaccine scenario has increased the candidates' high-level argumentation skills. Previous studies in the literature offer different results. It was determined that the candidates could not produce high-level arguments on vaccine, curfew and distance education, which they directly experienced in many areas of daily life such as education, social life and health (Atabey, 2021). On the contrary, it was revealed that the candidates produced higher-level arguments about topics that they were more familiar with or had prior knowledge about through news, social media, and TV advertisements (Khishfe, 2025). In this study, the previous level of familiarity of the participants with earthquake and vaccine scenarios was not directly measured. Thus, part of the improvement in the quality of arguments observed may have been due to the fact that participants became more familiar with the vaccine scenario through their daily lives (news, social media, health discussions). Additionally, the two scenarios differed in content, which may have caused the candidates to reach different levels of arguments. Due to the socioscientific nature of the vaccine scenario, candidates may have been able to find their justification more easily (Tang et al., 2025). This may explain the fact that despite the decrease in the number of arguments, the arguments have become more qualified.

Throughout the implementation process, it was observed that the candidates developed more qualified and multidimensional structures in their arguments. It is thought that artificial intelligence-supported feedback may contribute to this process. The tool analyzed the components, illustrated strengths and weaknesses with examples, and offered strategies. The feedback process unfolded in three stages. In the first stage, participants were shown, in concrete terms, in which areas they needed to improve so they could identify their missing components. In the second stage, the system presented participants with model arguments by illustrating the target argument component. In the third stage, action plans tailored to each participant's profile clearly showed them what to do in the next argumentation process. This three-stage structure ensured that feedback served not only an evaluative but also a guiding function. It was observed that this three-stage process coincided with the participants moving away from superficiality and establishing argument relations more consciously. It is considered that the personal action plans and exercises presented may have served as a support for critical thinking and may have contributed to the production of high-level arguments. These findings follow a similar direction to recent studies where AI-driven feedback was correlated with pre-service teachers' argument levels. Ba et al. (2025) found that the feedback provided through ChatGPT contributed to candidates' shift from one-dimensional claims to multi-component arguments supported by warrant and rebuttal elements during their debate process. Similarly, the ArgueBot system, developed by

Wambsganss et al. (2021), has achieved a gradual advancement in argument quality through instant feedback.

However, some studies show that the effects are not equally strong in all dimensions. Guo et al. (2024) reported that while they observed improvement in students' use of arguments and warrants in chatbot-supported classroom discussions, there was no significant progress in counterargument and rebuttal skills. Duran (2024) conducted research using tools such as ChatGPT, Gemini, Claude, and Mistral, showed that pre-service teachers structured their claims and data more strongly, but some of them leaned towards superficial explanations rather than in-depth warrant. Scheuer et al. (2010) emphasized that computer-assisted argumentation environments improve argument skills, but this effect is limited without teacher guidance. Additionally, Wambsganss et al. (2022) revealed that the ArgumentFeedback system contributes to students' development of more persuasive and high-quality arguments, but its use may be limited without human guidance due to the algorithm's occasional emphasis on incorrect components. Therefore, the literature generally shows that AI-assisted feedback improves the quality of arguments, but this effect may be limited depending on context, pedagogical prompts, and individual differences.

While a decrease was observed in the number of arguments produced by the candidates in the transition from the first to the second scenario, a significant progress was made in the quality of these arguments. This suggests two possible explanations. Firstly, the AI-powered feedback tool may have helped candidates address their shortcomings and encouraged candidates to deepen their arguments through their individual feedback. This is in line with the idea that it may be more effective to design artificial intelligence as a supportive partner that systematically develops and advances human thoughts rather than a "creativity trigger" that produces different ideas in human-artificial intelligence joint production (Komura & Yamada, 2026). Second, candidates may have omitted weak arguments so as not to diminish the persuasive power of a strong argument (Obermaier & Koch, 2024). Therefore, these findings should be interpreted with caution, as they align with other studies reporting that AI-based feedback is not equally effective across all argumentation components. This preference for teacher support, as reflected in participants' comments (see Section 1.4), may be attributable to the fact that AI interaction follows a standardized structure, applying instructions uniformly rather than adapting to each individual's specific needs (Farrokhnia et al., 2026). The findings suggest that AI-powered feedback may have contributed to the quality of arguments; however, this effect appears to vary based on context, pedagogical guidance, and individual differences.

This study revealed that PSTs found AI feedback valuable in terms of speed, interaction, and motivation, but perceived lengthy explanations and lack of personalization as limitations. These results are in line with similar findings in the literature. As a matter of fact, Jho and Ha (2024) stated that the ChatGPT-based feedback system contributed to their recognition of argument components, but due to their lack of personalization, students saw feedback as a complementary tool rather than a guiding tool. Similarly, Kim et al. (2022), emphasized that while AI-powered prompts enhance students' argument quality, the process cannot be fully personalized due to out-of-context suggestions and technical limitations. Alawad (2025) reported that while significant improvements were observed in students' consecutive written assignments due to AI-powered feedback, their opposing arguments remained weak. Liu and Cui (2024) showed that the use of claims, warrants, and rebuttals increased in ChatGPT-powered group discussions; however, since it is not a substitute for natural human interaction, it has revealed that students see the tool as a supporting element. Similarly, Nguyen and Barbieri (2025) noted that while large language models excel in recognizing claims and support elements in arguments, they fail to rebuttal and provide detailed warrants, making their outputs repetitive and devoid of personalization. When these findings are evaluated together, it is seen that artificial intelligence-supported feedback brings speed, diversity and motivation to the learning process; however, it is clearly seen that it cannot replace teacher feedback both individually and in a group context. Therefore, the candidates' positioning of such tools as functional but complementary coincides with the pedagogical and affective limitations revealed in the literature.

6. Conclusion and implications

Based on these findings, we present four data-based design principles for future AI feedback software like AIAFS. First, AI systems should analyze arguments holistically, not just at the level of individual sentences, but by gathering scattered statements throughout the conversation to provide feedback. This approach prevents fragmented feedback and ensures the student's argument structure is evaluated accurately. This principle, refined during Cycle 1 of this study, is critical to prevent fragmented feedback and ensure a more accurate assessment of a student's overall argumentative structure, thereby addressing the limitation of superficial analysis highlighted by Guerraoui et al. (2023). Second, effective systems should provide concise and personalized suggestions tailored to the individual's learning needs. This directly addresses the limitations identified by our participants and supported by Htaw et al. (2024), moving beyond generalized patterns to provide truly personalized learning pathways. Third, the system's capacity for persistent, dialogic interaction (e.g., follow-up questions after incorrect answers) was a key motivational driver. Future designs should prioritize this supportive and engaging communicative style to sustain engagement and foster a positive learning environment. Fourth, the tool's effectiveness depends on its speed and ability to provide immediate support for the learning process. Beyond these four principles, it is also essential that such tools be positioned as a teaching companion rather than a substitute for the teacher, in order to prevent the devaluation of the teacher's role. It should be noted that the findings of this study were derived from an exploratory DBR cycle conducted with a limited sample. The design principles and preliminary findings obtained are currently being tested in a comprehensive study being conducted with a larger group of participants using a pre-test–post-test design. It is recommended that future research validates these principles in different contexts and with larger samples.

7. Limitations

The small sample of nine PSTs from a single Turkish university is limiting in terms of generalizability but facilitated rich qualitative insight. Another limitation of the study is that the data saturation was obtained with a limited sample (three participants). The fact that no new code or theme appears after the third participant suggests that saturation has been reached. However, this result is not conclusive because the sample is small. Additional findings may occur in a larger sample. Therefore, it is recommended that future studies confirm this finding with larger samples. The tight emphasis was on argument structure, eschewing scientific correctness, and other aspects. Ecological validity was compromised through short-term extracurricular application. Finally, the findings relate to the specific strengths of the GPT-4 model and its prompts, emphasizing the technology-dependent nature of the findings.

References

- Adalberon, E. (2021). Providing assessment feedback to pre-service teachers: a study of examiners' comments. *Assessment & Evaluation in Higher Education*, 46(4), 601-614. <https://doi.org/https://doi.org/10.1080/02602938.2020.1795081>
- Alawad, E. A. (2025). AI-driven analysis: A case study on assessing argumentation proficiency and skill development through sequential submissions. *Linguistic and Philosophical Investigations*, 24(1), 108-117.
- Alejandro, I. M. V., Sanchez, J. M. P., Sumalinog, G. G., Mananay, J. A., Goles, C. E., & Fernandez, C. B. (2024). Pre-service teachers' technology acceptance of artificial intelligence (AI) applications in education. *STEM Education*, 4(4), 445-465. <https://doi.org/https://doi.org/10.3934/steme.2024024>
- Amir, A. (2026). AI-generated dialogic feedback: Designing a pedagogical chatbot grounded in literacy resilience principles. *Education Sciences*, 16(2), 318.
- Anderson, T., & Shattuck, J. (2012). Design-based research: A decade of progress in education research? *Educational researcher*, 41(1), 16-25. <https://doi.org/https://doi.org/10.3102/0013189x11428813>

- Atabey, N. (2021). Science teachers' argument types and supporting reasons on socioscientific issues: Covid-19 pandemic. *International Journal of Psychology and Educational Studies*, 8(2), 214-231.
- Aydeniz, M., & Ozdilek, Z. (2016). Assessing and enhancing pre-service science teachers' self-efficacy to teach science through argumentation: Challenges and possible solutions. *International Journal of Science and Mathematics Education*, 14(7), 1255-1273. <https://doi.org/https://doi.org/10.1007/s10763-015-9649-y>
- Ba, S., Zhan, Y., Huang, L., & Lu, G. (2025). Investigating the impact of ChatGPT-assisted feedback on the dynamics and outcomes of online inquiry-based discussion. *British Journal of Educational Technology*. <https://doi.org/https://doi.org/10.1111/bjet.13605>
- Bauer, E., Greisel, M., Kuznetsov, I., Berndt, M., Kollar, I., Dresel, M., Fischer, M. R., & Fischer, F. (2023). Using natural language processing to support peer-feedback in the age of artificial intelligence: A cross-disciplinary framework and a research agenda. *British Journal of Educational Technology*, 54(5), 1222-1245. <https://doi.org/https://doi.org/10.1111/bjet.13336>
- Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative research in psychology*, 3(2), 77-101. <https://doi.org/https://doi.org/10.1191/1478088706qp063oa>
- Brown, A. L. (1992). Design experiments: Theoretical and methodological challenges in creating complex interventions in classroom settings. *The journal of the learning sciences*, 2(2), 141-178. https://doi.org/https://doi.org/10.1207/s15327809jls0202_2
- Cai, L., Msafiri, M. M., & Kangwa, D. (2025). Exploring the impact of integrating AI tools in higher education using the Zone of Proximal Development. *Education and Information Technologies*, 30(6), 7191-7264. <https://doi.org/https://doi.org/10.1007/s10639-024-13112-0>
- Çakiroğlu, Ü., Güven, O., Yaman, H., & Güven, E. M. (2025). Scientific habits of arguing minds: enhancing online argumentation to foster SHOM. *Educational Technology Research and Development*, 73, 1865–1889. <https://doi.org/10.1007/s11423-025-10474-1>
- Çam, M. B., Çelik, N. C., Güntepe, E. T., & Durukan, Ü. G. (2021). Öğretmen adaylarının yapay zekâ teknolojileri ile ilgili farkındalıklarının belirlenmesi [Determining the awareness of teacher candidates about artificial intelligence technologies]. *Mustafa Kemal Üniversitesi Sosyal Bilimler Enstitüsü Dergisi*, 18(48), 263-285.
- Capkinoglu, E., Yilmaz, S., & Leblebicioglu, G. (2020). Quality of argumentation by seventh-graders in local socioscientific issues. *Journal of Research in Science Teaching*, 57(6), 827-855. <https://doi.org/https://doi.org/10.1002/tea.21609>
- Carless, D. (2022). From teacher transmission of information to student feedback literacy: Activating the learner role in feedback processes. *Active Learning in Higher Education*, 23(2), 143-153. <https://doi.org/https://doi.org/10.1177/1469787420945845>
- Carless, D., & Boud, D. (2018). The development of student feedback literacy: enabling uptake of feedback. *Assessment & Evaluation in Higher Education*, 43(8), 1315-1325. <https://doi.org/https://doi.org/10.1080/02602938.2018.1463354>
- Carlile, W., Gurrapadi, N., Ke, Z., & Ng, V. (2018). Give me more feedback: Annotating argument persuasiveness and related attributes in student essays. *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*,
- Chen, C.-M. (2008). Intelligent web-based learning system with personalized learning path guidance. *Computers & Education*, 51(2), 787-814.
- Chotijah, H. Y., Suparman, S., Kusumaningtyas, D. A., & Raman, A. (2025). Global Research Trends on 21st Century Teacher Competency Using Bibliometric Analysis. *Studies in Learning and Teaching*, 6(2), 305-319.

- Christenson, N., & Walan, S. (2023). Developing pre-service teachers' competence in assessing socioscientific argumentation. *Journal of Science Teacher Education*, 34(1), 1-23. <https://doi.org/https://doi.org/10.1080/1046560x.2021.2018103>
- Cramer, J., & Kempen, L. (2022). Toulmin and beyond: Structuring and analyzing argumentation. In *Proceedings of the Twelfth Congress of the European Society for Research in Mathematics Education (CERME12)*. European Society for Research in Mathematics Education.
- Diola, W. Y., Jalon Jr, J. B., & Prudente, M. S. (2025). Exploring the use of claim-data-reasoning in promoting scientific reasoning skills of elementary school students. *Anatolian Journal of Education*, 10(1), 203-214. <https://doi.org/https://doi.org/10.29333/aje.2025.10115a>
- Duran, V. (2024). Analyzing teacher candidates' arguments on AI integration in education via different chatbots. *Digital Education Review*, 45, 68-83. <https://doi.org/https://doi.org/https://doi.org/10.1344/der.2024.45.68-83>
- Engeness, I., & Gamlem, S. M. (2025). Exploring AI-driven feedback as a cultural tool: A cultural-historical perspective on design of AI environments to support students' writing process. *Integrative Psychological and Behavioral Science*, 59(1), 23. <https://doi.org/https://doi.org/10.1007/s12124-025-09894-8>
- Erduran, S. (2018). Toulmin's argument pattern as a "horizon of possibilities" in the study of argumentation in science education. *Cultural Studies of Science Education*, 13(4), 1091-1099. <https://doi.org/https://doi.org/10.1007/s11422-017-9847-8>
- Erduran, S., Simon, S., & Osborne, J. (2004). TAPping into argumentation: Developments in the application of Toulmin's argument pattern for studying science discourse. *Science education*, 88(6), 915-933. <https://doi.org/https://doi.org/10.1002/sce.20012>
- Escalante, J., Pack, A., & Barrett, A. (2023). AI-generated feedback on writing: Insights into efficacy and ENL student preference. *International Journal of Educational Technology in Higher Education*, 20(1), 57. <https://doi.org/https://doi.org/10.1186/s41239-023-00425-2>
- Fadillah, R., & Deta, U. (2020). The process of developing students' scientific argumentation skill using argument-driven inquiry (ADI) model in senior high school on the topic of elasticity. *Journal of Physics: Conference Series*, 1491(1), 012046. <https://doi.org/https://doi.org/10.1088/1742-6596/1491/1/012046>
- Fakhriyah, F., Rusilowati, A., Nugroho, S. E., Saptono, S., Ridlo, S., Mindyarto, B., & Susilaningsih, E. (2022). The scientific argumentative skill analysis reviewed from the science literacy aspect of pre-service teacher. *International Journal of Evaluation and Research in Education (IJERE)*, 11(4), 2129. <https://doi.org/https://doi.org/10.11591/ijere.v11i4.22847>
- Farrokhnia, M., Latifi, S., Papadopoulos, P. M., Hogenkamp, L., Gijlers, H., Khosravi, H., & Noroozi, O. (2026). Generative AI offers more, but students revise less: comparing the effects of teacher and AI feedback on student essay revisions. *International Journal of Educational Technology in Higher Education*, 23(1), 6.
- Finlay, L. (2021). Thematic analysis: the 'good', the 'bad' and the 'ugly'. *European Journal for Qualitative Research in Psychotherapy*, 11, 103-116. <https://doi.org/https://doi.org/10.24377/EJQRP.article3062>
- Fusch, P. I., & Ness, L. R. (2015). Are we there yet? Data saturation in qualitative research. *The Qualitative Report*, 20(9), 1408-1416. <https://doi.org/10.46743/2160-3715/2015.2281>
- Gu, J., Tang, L., Liu, X., & Xu, J. (2022). Promoting pre-service teacher students' learning engagement: Design-based research in a flipped classroom. *Frontiers in Psychology*, 13, 810275. <https://doi.org/https://doi.org/10.3389/fpsyg.2022.810275>

- Guerraoui, C., Reiser, P., Inoue, N., Mim, F. S., Singh, K., Choi, J., Robbani, I., Naito, S., Wang, W., & Inui, K. (2023). Teach me how to argue: a survey on NLP feedback systems in argumentation In M. Alshomary, C.-C. Chen, S. Muresan, J. Park, & J. Romberg (Eds.), *Proceedings of the 10th Workshop on Argument Mining* (pp. 19-34). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2023.argmining-1.3>
- Guo, K., Li, D., Wang, J., & Chu, S. K. W. (2025). From retrieval to generative models: a design-based research approach to developing a chatbot for argumentative writing. *Innovation in Language Learning and Teaching*, 1-15. <https://doi.org/https://doi.org/10.1080/17501229.2025.2511273>
- Guo, K., Pan, M., Li, Y., & Lai, C. (2024). Effects of an AI-supported approach to peer feedback on university EFL students' feedback quality and writing ability. *The Internet and Higher Education*, 63, 100962. <https://doi.org/https://doi.org/10.1016/j.iheduc.2024.100962>
- Han-Tosunoglu, C., Özer-Aker, F., & Ağlarçı-Özdemir, O. (2025). Transformative Shifts in Informal Reasoning of Pre-Service Teachers in a Real-Time Health Issue. *Journal of Science Teacher Education*, 36(3), 316-350.
- Hardcastle, J. M., Herrmann Abell, C. F., & DeBoer, G. E. (2021). *Validating a claim-data-science idea-reasoning (CESR) framework for use in NGSS assessment tasks* (Grantee Submission, Issue. A. A. f. t. A. o. Science).
- Haseski, H. I. (2019). What do Turkish pre-service teachers think about artificial intelligence? *International Journal of Computer Science Education in Schools*, 3(2), 3-23. <https://doi.org/https://doi.org/10.21585/ijcses.v3i2.55>
- Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of educational research*, 77(1), 81-112. <https://doi.org/https://doi.org/10.3102/003465430298487>
- Hopfenbeck, T. N., Zhang, Z., Sun, S. Z., Robertson, P., & McGrane, J. A. (2023). Challenges and opportunities for classroom-based formative assessment and AI: a perspective article. *Frontiers in Education*, 8, 1270700. <https://doi.org/https://doi.org/10.3389/feduc.2023.1270700>
- Hornstein, J., Greisel, M., & Kollar, I. (2025). Promoting pre-service teachers' data-informed reasoning through peer-feedback: effects of a feedback provision scaffold and a feedback integration scaffold. *Instructional Science*, 1-27. <https://doi.org/https://doi.org/10.1007/s11251-025-09745-4>
- Htaw, M. C., Pipa, D., Sriwattananarothai, N., Pichitpornchai, C., Gubelmann, R., Seufert, S., Niklaus, C., & Handschuh, S. (2024). Argumentative writing software: perceptions of undergraduate students toward artist prototype. In 2024 IEEE 7th Eurasian Conference on Educational Innovation (ECEI) (pp. 92-96). IEEE. <https://doi.org/https://doi.org/10.1109/ECEI60433.2024.10510808>
- Iordanou, K., & Constantinou, C. P. (2014). Developing pre-service teachers' data-based argumentation skills on socio-scientific issues. *Learning and Instruction*, 34, 42-57. <https://doi.org/https://doi.org/10.1016/j.learninstruc.2014.07.004>
- İşler, B., & Kılıç, M. (2021). Eğitimde yapay zekâ kullanımı ve gelişimi [The use and development of artificial intelligence in education]. *Yeni Medya Elektronik Dergisi*, 5(1), 1-11. https://doi.org/10.17932/IAU.EJNM.25480200.2021/ejnm_v5i1001
- Jaramillo, J. J., Chiappe, A., & Delgado, F. S. (2025). From struggle to mastery: AI-powered writing skills in ESL education. *Applied Sciences*, 15(14), 8079.
- Jho, H., & Ha, M. (2024). Towards effective argumentation: design and implementation of a generative AI-based evaluation and feedback system. *Journal of Baltic Science Education*, 23(2), 280-291. <https://doi.org/https://doi.org/10.33225/jbse/24.23.280>
- Kavitha, M., Roobini, S., Prasanth, A., & Sujaritha, M. (2023). Systematic view and impact of artificial intelligence in smart healthcare systems, principles, challenges and applications. In S. H. Tawseef Ayoub Shaikh, Tabasum Rasool, Mohammed Wasid (Ed.), *Machine learning and artificial*

- intelligence in healthcare systems (pp. 25-56). Elsevier. <https://doi.org/https://doi.org/10.1201/9781003265436-2>
- Kaya, E. (2013). Argumentation practices in classroom: Pre-service teachers' conceptual understanding of chemical equilibrium. *International Journal of Science Education*, 35(7), 1139-1158. <https://doi.org/https://doi.org/10.1080/09500693.2013.770935>
- Khishfe, R. (2025). Nature of science conceptions and argumentation skills in different socioscientific contexts. *International Journal of Science Education*, 47(12), 1542-1567.
- Kim, M. K., Kim, N. J., & Heidari, A. (2022). Learner experience in artificial intelligence-scaffolded argumentation. *Assessment & Evaluation in Higher Education*, 47(8), 1301-1316. <https://doi.org/https://doi.org/10.1080/02602938.2022.2042792>
- Kim, N. J., & Kim, M. K. (2022). Teacher's perceptions of using an artificial intelligence-based educational tool for scientific writing. *Frontiers in Education*, 7, 755914. <https://doi.org/https://doi.org/10.3389/educ.2022.755914>
- Komura, K., & Yamada, S. (2026). Deepening ideas vs. exploring new ones: AI strategy effects in human-AI creative collaboration. *PloS one*, 21(1), e0340449.
- Kuhn, D. (1993). Science as argument: Implications for teaching and learning scientific thinking. *Science education*, 77(3), 319-337.
- Kuhn, D., Zillmer, N., Crowell, A., & Zavala, J. (2013). Developing norms of argumentation: Metacognitive, epistemological, and social dimensions of developing argumentative competence. *Cognition and instruction*, 31(4), 456-496.
- Lin, S.-S., & Mintzes, J. J. (2010). Learning argumentation skills through instruction in socioscientific issues: The effect of ability level. *International Journal of Science and Mathematics Education*, 8(6), 993-1017. <https://doi.org/https://doi.org/10.1007/s10763-010-9215-6>
- Lin, Y. R., Jiao, X., & Hung, C. Y. (2026). Enhancing scientific argumentation with visual scaffolding in generative AI-supported environments. *International Journal of Science Education*, 1-26.
- Lipnevich, A., & Panadero, E. (2021). A review of feedback models and theories: Descriptions, definitions, and conclusions. *Frontiers in Education*, 6, 720195. *Frontiers in education*, 6, 720195. <https://doi.org/https://doi.org/10.3389/educ.2021.720195>
- Lippi, M., & Torroni, P. (2015). Argument mining: A machine learning perspective. In E. Black, S. Modgil, & N. Oren (Eds.), *Theory and applications of formal argumentation* (TAFA 2015, Lecture Notes in Computer Science, Vol. 9524, pp. 163–176). Springer. https://doi.org/10.1007/978-3-319-28460-6_10
- Liu, W., & Cui, X. (2024). Generative AI-assisted assisted collaborative argumentation: Implications for the argumentation process and outcome. In *2024 4th International Conference on Educational Technology (ICET)* (pp. 220-225). IEEE. <https://doi.org/https://doi.org/10.1109/ICET62460.2024.10868171>
- Ma, Z.-Q., Kong, L.-Y., Tu, Y.-F., Hwang, G.-J., & Lyu, Z.-Y. (2025). Strengthening collaborative argumentation with interactive guidance: a dialogic peer feedback approach based on the six thinking hats strategy. *Interactive Learning Environments*, 33(1), 296-320. <https://doi.org/https://doi.org/10.1080/10494820.2024.2344051>
- Macagno, F., & Konstantinidou, A. (2013). What students' arguments can tell us: Using argumentation schemes in science education. *Argumentation*, 27(3), 225-243. <https://doi.org/https://doi.org/10.1007/s10503-012-9284-5>

- Marco-Bujosa, L., Gonzalez-Howard, M., McNeill, K., & Loper, S. (2017). Designing and using multimedia modules for teacher educators: Supporting teacher learning of scientific argumentation. *Innovations in Science Teacher Education*, 2(4), 1-16.
- Martín-Gámez, C., & Erduran, S. (2018). Understanding argumentation about socio-scientific issues on energy: a quantitative study with primary pre-service teachers in Spain. *Research in Science & Technological Education*, 36(4), 463-483. <https://doi.org/https://doi.org/10.1080/02635143.2018.1427568>
- Morawski, M., & Budke, A. (2023). Teaching written argumentation to high school students using peer feedback methods—Case studies on the effectiveness of digital learning units in teacher professionalization. *Education Sciences*, 13(3), 268. <https://doi.org/https://doi.org/10.3390/educsci13030268>
- Nazaretsky, T., Mikeska, J. N., & Beigman Klebanov, B. (2023). Empowering teacher learning with AI: Automated evaluation of teacher attention to student ideas during argumentation-focused discussion. In *Proceedings of the 17th International Conference of the Learning Sciences (ICLS 2023)* (pp. 122-132). International Society of the Learning Sciences. <https://doi.org/https://doi.org/10.1145/3576050.3576067>
- Nguyen, N. N., & Barbieri, W. (2025). Mentorship in the age of generative AI: Opportunities and challenges for pre-service teachers. *Education Sciences*, 15(6), 642.
- Nusche, D., Staring, F., Rocha, M., & Montague, C. (2025). Empowered citizens, informed consumers and skilled workers: Designing education and skills policies for a sustainable future. OECD Publishing. <https://doi.org/https://doi.org/10.1787/311cdbeb-en>
- Obaidoon, S., & Wei, H. (2024). ChatGPT, Bard, Bing Chat, and Claude generate feedback for Chinese as foreign language writing: A comparative case study. *Future in Educational Research*, 2(3), 184-204. <https://doi.org/https://doi.org/10.1002/fer3.39>
- Obermaier, M., & Koch, T. (2024). The paradox of argument strength: how weak arguments undermine the persuasive effects of strong arguments. *Scientific Reports*, 14(1), 22244.
- Osborne, J. (2010). Arguing to learn in science: The role of collaborative, critical discourse. *science*, 328(5977), 463-466.
- Özmutlu, A. E. (2021). Doğal dil işleme. In T. Talan & C. Aktürk (Eds.), *Bilgisayar Bilimlerinde Teorik ve Uygulamalı Araştırmalar* [Natural language processing] (pp. 129-154). Efe Akademi Yayınları.
- Öztürk, B., & Okumus, S. (2022). Evaluating prospective elementary school teachers' written and oral arguments on ecology. *International Online Journal of Education and Teaching*, 9(1), 343-360.
- Prescott, M. R., Yeager, S., Ham, L., Rivera Saldana, C. D., Serrano, V., Narez, J., Paltin, D., Delgado, J., Moore, D. J., & Montoya, J. (2024). Comparing the efficacy and efficiency of human and generative AI: qualitative thematic analyses. *JMIR AI*, 3, e54482. <https://doi.org/https://doi.org/10.2196/54482>
- Reeves, D. B. (2006). *The learning leader: How to focus school improvement for better results*. ASCD.
- Rochmah, A., Suyono, S., & Sutoyo, S. (2025). Development of Assessment as Learning Instruments to Improve Chemistry Pre-Service Teachers' Argumentation Skills Using AI-Assisted Collaborative Argumentation Online Whiteboard. *TEM Journal*, 14(4), 3459.
- Ruwe, T., & Mayweg-Paus, E. (2023). “Your argumentation is good”, says the AI vs humans—The role of feedback providers and personalised language for feedback effectiveness. *Computers and education: Artificial intelligence*, 5, 100189. <https://doi.org/https://doi.org/10.1016/j.caeai.2023.100189>
- Sandoval, W. A., & Millwood, K. A. (2005). The quality of students' use of data in written scientific explanations. *Cognition and instruction*, 23(1), 23-55.
- Say, C. (2024). 50 soruda yapay zekâ. *Bilim ve Gelecek Kitaplığı*.

- Schen, M. (2013). A comparison of biology majors' written arguments across the curriculum. *Journal of Biological Education*, 47(4), 224-231. <https://doi.org/https://doi.org/10.1080/00219266.2013.788542>
- Scheuer, O., Loll, F., Pinkwart, N., & McLaren, B. M. (2010). Computer-supported argumentation: A review of the state of the art. *International Journal of Computer-supported collaborative learning*, 5(1), 43-102. <https://doi.org/https://doi.org/10.1007/s11412-009-9080-x>
- Schwartz, G., Herbst, P., & Brown, A. M. (2024). Harnessing asynchronous digital simulations of problem-based lessons to support mathematics teachers' professional development: A design-based approach. *International Journal of Science and Mathematics Education*, 1-34. <https://doi.org/https://doi.org/10.1007/s10763-024-10514-x>
- Sik, Ç. Ç., Benzer, E., & Saparova, S. (2023). Studying writing skills of science teacher candidates on covid-19 pandemic and their opinions regarding the vaccine as a socio-scientific issue. *Educational Policy Analysis and Strategic Research*, 18(4), 215-255. <https://doi.org/https://doi.org/10.29329/epasr.2023.631.10>
- Steiss, J., Tate, T., Graham, S., Cruz, J., Hebert, M., Wang, J., Moon, Y., Tseng, W., Warschauer, M., & Olson, C. B. (2024). Comparing the quality of human and ChatGPT feedback of students' writing. *Learning and Instruction*, 91, 101894. <https://doi.org/https://doi.org/10.1016/j.learninstruc.2024.101894>
- Tang, X., Tan, L., Sadler, T. D., Kong, Y., & Lin, J. (2025). When structure and content of socioscientific argumentation develop in an unbalanced way: A case study. *Science Education*, 109(5), 1464–1483. <https://doi.org/10.1002/sce.21975>
- Toulmin, S. E. (2003). *The uses of argument*. Cambridge University Press.
- Ucar-Longford, B., Hosein, A., & Heron, M. (2024). Improving pre-service teachers' argumentation skills: A holistic online scaffolding design approach. *Journal of digital learning in teacher education*, 40(2), 76-96.
- Uçar, B., & Demiraslan Çevik, Y. (2020). The effect of argument mapping supported with peer feedback on pre-service teachers' argumentation skills. *Journal of Digital Learning in Teacher Education*, 37(1), 6-29. <https://doi.org/https://doi.org/10.1080/21532974.2020.1815107>
- Uppalapati, V. K., & Nag, D. S. (2024). A comparative analysis of AI models in complex medical decision-making scenarios: evaluating ChatGPT, Claude AI, Bard, and Perplexity. *Cureus*, 16, e52485. <https://doi.org/https://doi.org/10.7759/cureus.52485>
- Usher, M. (2025). Generative AI vs. instructor vs. peer assessments: a comparison of grading and feedback in higher education. *Assessment & Evaluation in Higher Education*, 1-16. <https://doi.org/https://doi.org/10.1080/02602938.2025.2487495>
- Van Eemeren, F. H., Grootendorst, R., Johnson, R. H., Plantin, C., & Willard, C. A. (2013). *Fundamentals of argumentation theory: A handbook of historical backgrounds and contemporary developments*. Routledge.
- Viyanti, V., Cari, C., Sunarno, W., & Prasetyo, Z. K. (2017). Level of skill argued students on physics material. *Journal of Physics: Conference Series*, 895(1), 012043. <https://doi.org/https://doi.org/10.1088/1742-6596/895/1/012043>
- Walton, D. N. (1989). Dialogue theory for critical thinking. *Argumentation*, 3(2), 169-184.
- Wambsganss, T., Guggisberg, S., & Söllner, M. (2021). *Arguebot: A conversational agent for adaptive argumentation feedback* Proceedings of the 16th International Conference on Wirtschaftsinformatik (WI 2021), <https://aisel.aisnet.org/wi2021/3>

- Wambsganss, T., Janson, A., & Leimeister, J. M. (2022). Enhancing argumentative writing with automated feedback and social comparison nudging. *Computers & Education*, 191, 104644. <https://doi.org/https://doi.org/10.1016/j.compedu.2022.104644>
- Wang, L., Chen, X., Wang, C., Xu, L., Shadiev, R., & Li, Y. (2024). ChatGPT's capabilities in providing feedback on undergraduate students' argumentation: A case study. *Thinking Skills and Creativity*, 51, 101440. <https://doi.org/https://doi.org/10.1016/j.tsc.2023.101440>
- Wang, Y.-H. (2020). Design-based research on integrating learning technology tools into higher education classes to achieve active learning. *Computers & Education*, 156, 103935.
- Winstone, N. E., Nash, R. A., Parker, M., & Rowntree, J. (2017). Supporting learners' agentic engagement with feedback: A systematic review and a taxonomy of recipience processes. *Educational psychologist*, 52(1), 17-37. <https://doi.org/https://doi.org/10.1080/00461520.2016.1207538>
- Wu, R. (2014). Feedback in distance education: A content analysis of Distance Education: An International Journal, 1980-2013 [Doctoral dissertation, Virginia Polytechnic Institute and State University].
- Yu, H., & Xie, Q. (2025). Generative AI vs. teachers: Feedback quality, feedback uptake, and revision. *Language Teaching Research Quarterly*, 47, 113-137. <https://doi.org/https://doi.org/10.32038/ltrq.2025.47.07>
- Zhang, J. J., Wang, Y. W., Ruan, Q., & Yang, Y. (2024). Digital tourism interpretation content quality: A comparison between AI-generated content and professional-generated content. *Tourism Management Perspectives*, 53, 101279. <https://doi.org/https://doi.org/10.1016/j.tmp.2024.101279>
- Zhang, Y., & Liu, Y. (2026). Designing ChatGPT-mediated feedback activities in EFL writing: a design-based study of the dialogic feedback triangle. *Assessment & Evaluation in Higher Education*, 51(1), 137-160.
- Zhou, X., Zhang, J., & Chan, C. (2024). Unveiling students' experiences and perceptions of artificial intelligence usage in higher education. *Journal of University Teaching and Learning Practice*, 21(6), 126-145. <https://doi.org/https://doi.org/10.53761/xzjprb23>
- Zhu, M., Lee, H.-S., Wang, T., Liu, O. L., Belur, V., & Pallant, A. (2017). Investigating the impact of automated feedback on students' scientific argumentation. *International Journal of Science Education*, 39(12), 1648-1668. <https://doi.org/https://doi.org/10.1080/09500693.2017.1347303>
- Zhu, M., Liu, O. L., & Lee, H.-S. (2020). The effect of automated feedback on revision behavior and learning gains in formative assessment of scientific argument writing. *Computers & Education*, 143, 103668. <https://doi.org/https://doi.org/10.1016/j.compedu.2019.103668>
- Zinger, D., Naranjo, A., Amador, I., Gilbertson, N., & Warschauer, M. (2017). A design-based research approach to improving professional development and teacher knowledge: The case of the Smithsonian Learning Lab. *Contemporary Issues in Technology and Teacher Education*, 17(3), 388-410.

Appendix A: Interview Protocol

General Experiences and Perceptions Regarding Artificial Intelligence

1. What are your feelings about artificial intelligence? (e.g., It makes me anxious, it makes me happy because it makes my work easier, it frightens me, etc.)
2. What are your thoughts on the concept of artificial intelligence?
3. What do you think about the role of artificial intelligence in education? Why?
4. Do you think AI applications will affect the teacher's workload positively or negatively? Why?

5. When you consider Science class and AI applications together, what potential contributions could AI make to this subject?
6. Considering the content of the Science class, in which topics could AI be more beneficial?
7. If you were a teacher, what kind of feedback would you prefer? Would you prefer the feedback you would provide to your students, or the feedback provided by AI? What could be the advantages and disadvantages of these two different types of feedback for a teacher?
8. Now, please answer this from a learner's perspective. Which type of feedback would be more useful for you: AI feedback or teacher feedback?
9. Do you think AI-supported feedback can replace the feedback provided by teachers, or should it be viewed as a complementary tool?

Views on Effectiveness of AI-Supported Feedback

10. How would an educational environment that includes AI feedback affect students' participation rates in activities and the quality of instruction?
11. If AI defended a counter-argument against your own argument, what would your reaction be?
12. Can AI feedback be used to support peer feedback? If so, how could it be used?

Evaluation from the Pre-service Teacher Perspective

13. As a future teacher, do you think you could use AI in your students' learning processes? (e.g., for classroom management, curriculum)
14. What do you think are the advantages and challenges of using AI-supported feedback to identify your students' individual learning needs?
15. How do you think AI-supported feedback will influence student achievement?
16. In your opinion, which is more likely to actively encourage student self-assessment: AI feedback or teacher feedback?
17. Do you think your students would take initiative and set learning goals in AI-supported studies and feedback environments?

Appendix B: Sample AI-Supported Feedback Interaction with PST-1

Hoş geldin Eren! 🌟

Bugün seninle birlikte daha önce grup çalışmasında kurduğun argümanları değerlendirmek değil, argüman kurma ve geliştirme becerini desteklemek, kavramları birlikte gözden geçireceğiz. Hazırsan başlayalım!

🗨️ Önce "argüman" ve "argümantasyon" kavramlarını tanıyalım

- **Argüman**, bir fikri savunmak için oluşturulan ve dayanaklarla desteklenmiş, söylenir ve bunu neden söylediğini açıklarsın.
- **Argümantasyon** ise bu argümanı kurma sürecidir. Bir iddiayı savunurken bileşenlerle nasıl tartıştığını gösterir.

Örnek vereyim:

🗨️ "Deprem bölgelerinde sarkaç gücü projesi kullanılmalı (iddia) çünkü bu (gereke) ve deprem şiddetine karşı en yüksek direnci sağlar (veri)."

Bu ayrımı anladın mı? Hazırsan Toulmin modeline geçelim mi?

👍 🗨️ 🗨️ ...

🗨️ 1. Argüman

Örnek:

"Üçüncü projeyi seçmemin nedeni daha büyük daha geniş binalarda kullanılması. Görsel nüfusu arttırdı için hem daha büyük binalara daha büyük gökdenar gibi binalara ihtiyacımız, görsel estetiği ve mimari de bunu yansıtır."

🗨️ Toulmin Bileşenleri:

- **İddia:** Üçüncü projeyi seçmemin nedeni daha büyük binalarda kullanılması.
- **Veri:** İnsanlık nüfusu arttı, daha büyük binalara ihtiyacımız var.
- **Gereke:** Görsel estetiği ve mimari de bunu yansıtır.

📊 Seviye Değerlendirmesi (Erduran, Simon & Osborne, 2004):

📊 Seviye 2

Çünkü argümanda iddia, veri ve gereke var diyorsun ama görsellik yok.

🗨️ Seviye 5'e Yükseltmek İçin:

Senin başta bu cümleleri eklemek şöyle diyebiliriz:

"Bazı kişiler bu sistemin maliyeti oldukça düşmektedir (görsellik) ama görsellikte karşılaştığımızda da nüfus ve yapı ihtiyacı gibi önünde bulundurulduğunda bu maliyeti karşılanabilir bir düzeydedir (görsellik)."

🗨️ Böylece birden fazla görsellik eklenmiş olur ve Seviye 5'e ulaşırsın.

Hello, I am Eren.

Welcome Eren! 🌟

Today, we will look at the arguments you previously developed in your group work. Our goal is not to evaluate you, but to support your ability to construct and develop arguments. For this, we will first review some fundamental concepts together. If you're ready, let's begin!

🗨️ First, let's get familiar with the concepts of 'argument' and 'argumentation':

- An **Argument** is a statement formed to defend an idea and is supported by grounds. In other words, you state an opinion and explain why you are saying it.
- **Argumentation** is the process of constructing this argument. It shows how you debate an assertion using data, warrants, and other components.

Let me give an example:

🗨️ "The seismic isolation (pendulum power) project should be used in earthquake zones (claim) because it increases durability in large buildings (warrant) and provides the highest resistance against earthquake intensity (data)." Did you understand this distinction? If you're ready, shall we move on to the Toulmin model?

🗨️ 1. Argument Sentence:

The reason I chose the third project is its use in larger, wider buildings. Since the human population is increasing today, there will be a need for larger buildings, like bigger skyscrapers, and today's aesthetics and architecture already reflect this.

🗨️ Toulmin Components:

- **Claim:** The reason I chose the third project is its use in larger buildings.
- **Data:** The human population has increased, and there is a need for larger buildings.
- **Warrant:** Today's aesthetics and architecture also reflect this.

📊 Level Assessment (Erduran, Simon & Osborne, 2004):

📊 Level 2

Because the argument includes a claim, data, and a warrant, but there is no rebuttal.

🗨️ To Raise to Level 5: By adding another sentence of yours, we could say this: 'Some people might think this system is costly (rebuttal) but considering the increasing population and need for construction under today's conditions, this cost is at a level that can be met (rebuttal).'

🗨️ Thus, multiple rebuttals are added, and Level 5 is reached.

Appendix C: Argument Development Assessment and Supportive Feedback

Hoş geldin Eren! 🌟

Bugün seninle birlikte daha önceki grup çalışmasında kurduğun argümanlara bakalım. Amacımız seni değerlendirmek değil, **düşüncelerini nasıl geliştirdiğini birlikte görmek**. İlk olarak **deprem konusundaki** başlangıç seviyene, sonra **aşı konusundaki** her seviyeden bir örnek argümanına, en son olarak da gelişimini gösteren tabloya ve kısa yorumuma bakalım. Hazırsan başlayalım! 🌟

Eren – Argüman Kalitesi Karşılaştırma Tablosu

Seviye	Deprem (İlk)	Aşı (Güncel – Belgede)
Seviye 1	3	3
Seviye 2	2	2
Seviye 3	1	0
Seviye 4	2	2
Seviye 5	0	1

Gelişim Değerlendirmesi – Destekleyici Geri Bildirim

Eren, senin argümanlarında gözle görülür bir ilerleme var. 🌟

Deprem konusundaki çoğunlukla düşüncelerini kısa ve temel düzeyde sunmuşsun. Aşı konusunda ise artık **karşıt görüşleri değerlendiriyor**, örnekler veriyor ve **venilere dayalı güçlü gerekçelerle savlanı** destekliyorsun. Özellikle **Seviye 5** düzeyindeki argümanında, hem bilimsel güveni hem de bireysel özgürlüğü birlikte değerlendirmiş olman çok dikkat çekiydi. 🌟

Bu gelişim senin sadece bilgiye dayalı düşünmediğini, aynı zamanda farklı bakış açılarına da dikkate aldığını gösteriyor. Bilimsel düşünme becerilerini çok güzel geliştiriyorsun. Devam et, çok başarılısın! 🌟

Welcome Eren! 🌟

Today, we'll be looking at the arguments you developed during the previous group work. Our aim is not to evaluate you, but **to observe together how you developed your thoughts**. Firstly, we will look at your **beginner-level argument from your earthquake speech**, then **one example argument from each level in your vaccine speech**, and finally, we will look at the **table showing your development and my brief comment**. If you're ready, I'll begin! 🌟

Eren - Argument Quality Comparison Table

Level	Earthquake (First)	Vaccine (Current)
Level 1	3	3
Level 2	2	2
Level 3	1	0
Level 4	2	2
Level 5	0	1

Developmental Evaluation – Supportive Feedback

Eren, there has been a remarkable improvement in your arguments. 🌟

In your Earthquake discussion, you mostly presented your ideas briefly and at a basic level. However, in your Vaccine discussion, you began to **evaluate opposing views**, provide examples, and support **your claims with strong, data-based reasoning**. Especially in your **Level 5 argument**, your consideration of both scientific trust and individual freedom was truly impressive. 🌟

This progress shows that you are not only thinking based on knowledge but also taking different perspectives into account. You are developing your scientific thinking skills beautifully. Keep it up—you're doing an excellent job! 🌟

Appendix D: Example of Personalized Action Plans

Şimdi son olarak, **yapay zekanın olmadığı ortamlarda** senin argümanlarını nasıl geliştirebileceğini konuşalım mı? Sana özel, küçük ama etkili bir öneri kitapçığı gibi hazırlayalım mı? 🌟

📄 📄 📄 ...

hazır

Prepare it

Harikal! 🌟 İşte senin için özel olarak hazırlanmış, **argüman kurma becerini yapay zeka olmadan da geliştirebileceğin bir öneri rehberi**. Bu küçük el kitabını bir yol haritası gibi düşünebilirsin:

EREN İÇİN ARGÜMAN GELİŞTİRME REHBERİ

🌟 "Daha güçlü argümanlar kurmak senin elinde!"

1. Her zaman bir İDDİA ile başla.

Bir konu hakkında ne düşündüğünü açıkça söyle.

🌟 *İpucu:* "Bence...", "Bu proje daha iyi çünkü..." gibi ifadelerle başlayabilirsin.

📅 Günlük Egzersiz Önerisi:

- Günlük haberlerden bir konu seç.
- Yukarıdaki 6 bileşeni kullanarak kısa bir argüman yaz.
- Her gün bir tane yazarsan 1 haftada 7 güçlü argüman kurmuş olursun! 🌟

Now, finally, should we talk about **how you can improve your arguments in environments without artificial intelligence**? Should I prepare it for you like a small but effective suggestion booklet? 🌟

Great! 🌟 Here is a suggestion guide specially prepared for you, where you can improve your argumentation skills even without artificial intelligence. You can think of this little handbook as a roadmap:

ARGUMENT DEVELOPMENT GUIDE FOR C1

🌟 "Building stronger arguments is in your hands!"

Always start with a CLAIM.

State clearly what you think about a topic.

🌟 *Tip:* You can start with phrases like, "I think...", "This project is better because..."

📅 Daily Exercise Suggestion:

- Select a topic from the daily news.
- Write a short argument using the 6 components listed above.
- If you write one every day, you will have built 7 strong arguments in 1 week! 🌟

Appendix E: Example of One Participant's Argument Development, Before and After

• 1. Original Argument (Raw Version) (Pre-training / Earthquake Scenario):

"I defended the first project. I think the fact that it has an average cost is one of the most important factors, and the fact that it can be used in every building means it doesn't have to be a very large building or a very small building; I think it can be used in an average building."

• 2. Identified Components and Level:

o **Claim:** I defended the first project.

o **Data:** It has an average cost.

o **Justification:** It can be applied to every building type.

o **Support:** This applies to both small and large buildings.

o **Qualitative/Rebuttal:** None

o **Level Assessment (Erduran, Simon & Osborne, 2004):** Level 2

• 3. Feedback Given:

"This argument is at Level 2 because it does not contain a refutation."

• 4. Personalized Action Plan:

Mini Suggestion Package for Improving Your Arguments Even Without AI

- Always provide data for each claim:
- Example: "This system is cheaper because it only requires 3 million TL to build."
- Make your justifications clear:
- Explanations starting with "Because..." strengthen the justification. Add supporting evidence:
- Statements like "Experts also support this view" support your justification. Always include a refutation:

- Addressing opposing views and explaining why you disagree strengthens your argument. Set limits with a qualifying statement:
- Limiting statements like "It is only valid in newly constructed buildings" clarify your claim.

• 5. Subsequent (Improved) Argument (Post-Training and Feedback / Vaccination Scenario):

"80% vaccinated and 363 people became ill, but there were no deaths. When the vaccinated rate dropped to 10%, over 2000 people died. I think it would be more logical to look at this with the data I have later."

• 6. Identified Components and Level:

o **Claim:** Vaccines reduce the mortality rate.

o **Data:** No deaths in an environment where 80% were vaccinated, thousands of deaths in an environment where 10% were vaccinated.

o **Justification:** If the vaccination rate decreases, mortality increases.

o **Support:** Data-driven proposition.

o **Rebuttal 1:** "A decrease in the vaccination rate does not affect the risk" → directly refutes the opposing view.

o **Rebuttal 2:** "This needs to be shown with data" → advocates for thinking based on concrete evidence, not just belief.

o **Level Assessment (Erduran, Simon & Osborne, 2004):** Level 5

o **Feedback Given:** The expanded argument structure is at its highest level with multiple rebuttals.

Based on the feedback received after the earthquake scenario, the participant integrated the data-based rebuttal component into their argument in the vaccine scenario and raised their level.