

Article

An Exploratory Comparison of Early AI Feedback Attempts on Preservice Physics Teachers' Reflective Writing

Lukas Mientus ^{1,*}, Peter Wulff ², Anna Nowak ³ and Andreas Borowski ⁴

¹ Physics Education Research, Faculty of Natural Science, Institute of Physics, Otto von Guericke University, Magdeburg 39106, Germany

² Physics and Physics Education Research, Faculty of Social Science and Natural Science, Institute for Chemistry, Physics, and Technology, Ludwigsburg University of Education, Ludwigsburg 71602, Germany

³ Center for Teacher Training and Education Research, University of Potsdam, Potsdam 14476, Germany

⁴ Physics Education Research, Faculty of Science, Institute of Physics and Astronomy, University of Potsdam, Potsdam 14476, Germany

* Correspondence: lukas.mientus@ovgu.de

Abstract

Reflective writing is a core component of teacher education, especially during practical internships. However, providing high-quality feedback on reflections is resource-intensive. This study examines descriptively observable associations between an early dual-feedback approach combining basic (automated) and elaborate (human-generated) feedback and structural features of preservice physics teachers' reflective writing, prior to the widespread adoption of generative AI in education. Using an exploratory, non-equivalent, non-concurrent cohort design, we analyzed participant-level aggregates of written reflections from a non-intervention cohort ($N = 22$) and an intervention cohort ($N = 32$), applying a validated reflection-supporting model to assess structural composition and discursive elements of reflective writing. In the intervention, basic feedback was generated by a previously validated BERT-based machine learning model focusing on structural reflection elements, while elaborate feedback addressed content-related and pedagogical depth. In this study, the automated model was employed as an analytic measurement instrument drawing on validation work demonstrating its transferability across comparable reflection contexts. Quantitative analyses did not reveal systematic longitudinal growth in indicators of reflective writing quality in either cohort. Across comparable measurement points, descriptively different structural reflection profiles were observed between cohorts, without permitting causal or developmental interpretations. Feedback acceptance was high overall, although structural AI feedback was perceived as less personalized and less useful. These findings highlight the descriptive value of early, non-generative AI-based approaches for scalable structural diagnostics of reflective writing, while underscoring the continued importance of human-generated, content-focused feedback. The study establishes an empirical baseline for evaluating contemporary generative AI-based feedback systems in teacher education.

Keywords: reflective writing; teacher education; feedback intervention; machine learning; preservice physics teachers

Academic Editors: Daniel Muijs, Roberto Capone, Ana Isabel Sacristán, Maria Giuseppina Adesso and Mario Lepore

Received: 25 July 2025

Revised: 14 January 2026

Accepted: 30 January 2026

Published: 12 February 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

1. Introduction

Developing reflective competence is widely regarded as a key goal in teacher education, especially during practice-oriented internships (Brouwer & Korthagen, 2005; Gröschner et al., 2013; Schubarth et al., 2009). Reflection enables preservice teachers to connect theoretical knowledge with their classroom experiences and engage in continuous professional development (Schön, 1983; von Aufschnaiter, 2023). However, fostering high-quality reflection is challenging, particularly in large cohorts where personalized feedback on reflective writing is often constrained by limited instructional resources (Narciss, 2006; Poldner et al., 2014).

Digital technologies offer new opportunities to scale support for reflective writing. Advances in natural language processing and machine learning (ML) have made it possible to generate real-time, structured feedback on student texts (Buckingham Shum et al., 2017; Ullmann, 2019). While these technologies cannot replace human judgment in complex pedagogical matters, they may assist by automating feedback on formal structure or prompting deeper reasoning. A combined approach—where ML tools support instructors by analyzing structural aspects of reflection and freeing them to focus on content-related feedback—could offer a pragmatic solution in teacher education.

This study investigates such a dual-feedback approach using a non-equivalent, non-concurrent observational comparison of two cohorts of preservice physics teachers. The design is exploratory and does not allow for causal inferences about the impact of the intervention. Nevertheless, it can highlight potentials of early AI methods – even before the release of the generative AI Chat GPT-3.5 in November 2022. This paper focuses on whether this structured feedback model (based on BERT, an early and less sophisticated AI model) is associated with differences in the quality of reflective writing over the course of an internship. Measuring text structure and the presence of higher-order reasoning this study aims to establish an empirical baseline for evaluating AI-based feedback approaches to reflective writing in science teacher education.

2. Theory

2.1. Pedagogical Reasoning and Reflection-on-Action

Beyond inherent professional knowledge a teacher brings into a teaching situation, it is crucial to recognize the significance of professional perception regarding teaching elements and their theoretically grounded analysis. A teachers' professional perception of a teaching situation is essential for acquiring or developing professional competencies (Gruber, 2021; van Es & Sherin, 2021). Reflection links the two elements of knowledge and perception and is said to therefore stimulate a sustainable intrinsic professionalization process (Roters, 2012). In terms of developing the professional teaching career, teachers' reflective competencies can thus be identified as a prerequisite for continued professional development and learning (Combe & Kolbe, 2008). Reflection is emphasized as a central building block for the professional development of teachers (Christof et al., 2018; Korthagen & Kessels, 1999; Schön, 1983; Sorge et al., 2019) and appears to be indispensable for the development of teachers' teaching personality (Rothland, 2021). National and international teacher training standards feature reflection on a regular basis (e.g., KMK, 2004, 2019; NBPTS, 2016) as well as teacher education programs (Darling-Hammond, 2012). The promotion of the reflective ability thus establishes itself as a significant goal of teacher education (Browning & Korthagen, 2021), although empirical findings remain inconclusive regarding the extent to which reflection directly improves teaching performance (Collin et al., 2013; Häcker, 2019; Labott & Reintjes, 2022; Wyss, 2013).

On the one hand, reflection can be seen as a connection between teaching in class and new planning (Korthagen, 2001; Nordine et al., 2021), while on the other reflection is not intended to be separate from the classroom action (Alonzo et al., 2019). As a process, however, there is a risk of it becoming a mere analysis (von Aufschnaiter, 2023). Although a detailed lesson analysis can demonstrate a deep understanding of the reflected situation, only a personal reference in a reflection (e.g., by deriving consequences for improving the teaching or the lesson) is able to stimulate professional development (von Aufschnaiter et al., 2019). The content, goals, and methods for reflection are diverse (Christof et al., 2018) and can also depend on reflection-related prior knowledge (e.g., a reflection-supporting prompt or a stimulated recall). Suitable questions that structure the reflection process or appropriate prompts and methods (Wyss, 2018), may prove helpful for reflecting.

During their professional development, teachers are envisioned as reflective practitioners, i.e., professionals who continuously and constantly monitor their performance based on experience and feedback and adapt it to the situation (Abels, 2011; Altrichter, 2000; Schön, 1983). In the course of teaching internships, this can arise as a major aim of lecturers' support. However, since a reflection of an experienced teaching situation is often preceded by a form of analysis (Meschede, 2014), supporting reflective reasoning using feedback ought to be helpful, as a teacher can be competent in analyzing teaching without having to be able to reflect competently ("analytical practitioners" (von Aufschnaiter, 2023, p. 40)). Hence, it is necessary to operationalize reflection-related pedagogical reasoning. According to Nowak et al. (2019) and further Nowak (2023) following (among others) Korthagens' (2001) ALACT-Model^{Error! Reference source not found.}, a beneficial reflection can be constructed in the following ways: using (1) Circumstances of the teaching situation and Descriptions of teachers' and students' acting and interactions in the specific situation to reflect on, (2) Evaluations of how a person (teacher her/himself or observer) perceives the teaching situation (and why), (3) Alternative actions for the activities observed, where the teacher should consider what they could have done differently to make the perceived experiences better, or (4) Consequences a teacher can draw for her/his professional development. In line with Alonzo et al. (2019), Korthagens' Act-of-Trial (i.e., the performance in class) is conceptually separate from the reflective reasoning process itself in the reflection-supporting model of Nowak et al. (2019), so feedback allows to address either teachers' theory-based planning, classroom teaching, or the reflection-on-action itself. Following this operationalization, it becomes clear that pedagogical reasoning can be divided into descriptive elements (circumstances and descriptions) and discursive elements (evaluations, alternatives, consequences). In reference to Wyss (2018) written reflections with less descriptive elements indicate a reflection of higher quality or can represent teachers' higher-order reasoning.

2.2. Required Content for Feedback on Written Reflections

Feedback on written reflections should address structural and content-related aspects (i.e., basic and elaborate feedback, see Narciss, 2006, 2018) to activate reflective competence during professional development (Poldner et al., 2014). Emphasizing structural aspects may be particularly important for novices in teacher training programs, as a structured approach is a necessary prerequisite for the development of expertise (Korthagen & Kessels, 1999) and serves as a basis for higher-quality reflection processes (Hatton & Smith, 1995). Integrating computer-based analyses ensures that preservice teachers receive personalized feedback promptly (Buckingham Shum et al., 2017). In addition, content-related feedback is essential, which explicitly assesses the identified occasions for reflection and thus includes theoretical knowledge

(Leonhard & Rihm, 2011), providing a basis for an in-depth examination of one's knowledge. Following Hattie and Timperley (2007), effective feedback should encompass inquiries about the purpose of the feedback (Feed Up), the implementation of the task or a status quo (Feed Back), and a concrete next step or an impulse for professional development (Feed Forward).

However, it is still unclear how basic (on argumentation structure) and elaborate (more in depth) feedback in this manner can effectively and efficiently support the integration of experience and knowledge in teaching practice. In particular, feedback on pre-service teachers' reflections in teaching is often time-consuming, considering lecturers supervise many teachers. In contrast, computers can provide instantaneous feedback, which might raise teachers' acceptance of tasks such as reflective writing. For these reasons, we introduced an innovative approach by implementing structural feedback using computer-based systems (Lai & Calandra, 2007). These early attempts were able to generate feedback using structural analyses of written reflections—even without generative AI. It allows lecturers to focus their full attention on the content-related aspects of the reflections to give elaborate feedback. But is this effort justified?

2.3. Large Language Models and Generative AI

The rise of large language models (LLMs) in natural language processing has been driven by a rapid evolution from encoder-based contextual models to large-scale generative systems. One of the foundational breakthroughs was BERT (Bidirectional Encoder Representations from Transformers), introduced by Devlin et al. (2018), which used bidirectional Transformer encoders and masked language modeling to generate deep contextual embeddings, but was not designed for text generation. Its architecture and pre-training approach strongly influenced subsequent work in Natural Language Processing (NLP). Following BERT, autoregressive Transformer models such as GPT-2 (OpenAI, 2019) demonstrated that scaling up transformer-based language models could yield impressive text generation capabilities. The next major step came with GPT-3 (Brown et al., 2020), which scaled to 175 billion parameters and exhibited few-shot and zero-shot learning capacities, far surpassing earlier models in flexibility and fluency. GPT-3 marked a turning point toward generative AI as a broadly useful tool. Finally, by late 2022, the public release of ChatGPT (OpenAI, 2022) made conversational generative AI widely accessible, catalyzing interest in AI-mediated writing, dialogue, and feedback systems. This historical trajectory—from encoder-only contextual models to large-scale generative and conversational LLMs—frames this study as an early exploration of automated feedback, predating the generative AI wave. This study predates generative AI for feedback reflective writing but provides a baseline for later work with LLMs.

From a methodological perspective, it is important to distinguish between generative and analytic uses of large language models. While recent generative systems are designed to produce adaptive feedback texts, earlier encoder-based models such as BERT were developed primarily for representation learning and classification tasks (Wulff et al., 2022). Prior validation studies demonstrated that, among several deep learning architectures, the BERT-based approach achieved the most robust performance for classifying reflective writing according to theory-driven category systems (Wulff et al., 2022). This advantage is commonly attributed to the transformer architecture underlying BERT, which enables models to attend to long-range dependencies between tokens within a text and to integrate contextual information across entire sentences or paragraphs (Vaswani et al., 2017; Alammari, 2018). In addition, positional encodings allow the model to retain information about the sequential structure of the input, which is particularly relevant for modeling the progression of reflective

argumentation (Devlin et al., 2018). On this basis, pretrained models such as BERT can be empirically validated as analytical instruments for reflective writing, even across institutional contexts, provided that a stable theoretical framework underlies the classification (Sorge et al., 2025). In this sense, the present study does not conceptualize BERT as an intelligent tutor or feedback generator but as a structural measurement device that enables scalable, theory-based diagnostics of reflective writing. This analytic role is fundamentally different from generative feedback approaches and should therefore be evaluated according to criteria of classification validity and contextual transferability rather than adaptability to individual situations. The study thus aligns with recent calls to treat large language models in educational research as epistemic tools whose function and limits must be clearly specified (Sorge et al., 2025).

2.4. Research Questions

This study explores descriptively observable associations between early AI-supported feedback processes and structural features of preservice physics teachers' reflective writing during internships. Specifically, the intervention was designed to provide rapid, individualized feedback on written reflections, combining both algorithm-generated and instructor-generated responses. This approach responds to preservice teachers' stated need for timely, actionable feedback during field experiences. Rather than offering holistic assessments, the intervention disaggregates feedback into structural (basic) and content-related (elaborate) components to promote more targeted reflection.

While this approach is promising, empirical research on hybrid feedback systems in teacher education—especially those combining AI-based and human-generated inputs—remains limited. This study contributes to closing that gap by comparing development in teachers' reflection across cohorts and exploring user perceptions of feedback forms.

RQ1: To what extent do descriptively observable differences in the structural composition of reflective writing coincide with participation in different internship feedback formats?

This question focuses on measurable changes in the reflective texts of preservice teachers who received structured feedback during the intervention versus those in a typical (business-as-usual) supervision model. The analysis examines both group-level differences and within-group development over time. Given the inconsistent findings in previous literature—some showing growth in reflective depth (e.g., Fund et al., 2002; Klempin, 2021), others reporting no significant change (e.g., Meißner et al., 2020; Wyss, 2013)—we descriptively document patterns of similarity and difference between cohorts without testing inferential hypotheses.

RQ2: How do preservice teachers perceive the usefulness and personalization of AI-based basic versus human-generated elaborate feedback?

In addition to textual outcomes, this study explores preservice teachers' perceptions of the feedback they received. RQ1 focuses exclusively on objectively measurable textual outcomes in reflective writing. RQ2, in contrast, addresses participants' subjective perceptions of the feedback they received. These two perspectives are analytically independent and are interpreted separately. Drawing on their evaluations, we examine whether the immediacy of basic AI-generated feedback compensates for its lack of nuance, and how it compares to the perceived value of more elaborate, human-generated input. We expect that teachers will favor feedback that is both rapid and personalized, but may view overly simplified feedback as less instructive. While direct equivalence between internship contexts is not assumed, the design includes descriptive comparisons of cohort characteristics and supervision settings. These contextual differences are addressed in the Section 3 to support interpretation of the intervention's effects.

3. Methods

3.1. Sample and Cohort

This study used a non-equivalent, non-concurrent, exploratory observational cohort design comparing written reflections from two cohorts of preservice physics teachers at the University of Potsdam, Germany. The analysis is exploratory and descriptive, and between-cohort differences cannot be interpreted as causal effects of the feedback concept. Because the cohorts were embedded in different semesters and structural conditions (e.g., COVID-19 disruptions, varying seminar contact time), the comparison provides contextualized descriptive insights rather than a controlled causal evaluation of the feedback approach. Written reflections from two cohorts of preservice physics teachers at the University of Potsdam were analyzed:

- Non-intervention cohort: $N = 22$ (three semesters from 2017–2018)
- Intervention cohort: $N = 32$ (five semesters from 2020–2023)

The students of both cohorts usually participate in the teaching internship in their last year of university-based studies and studied physics and education since at least four years. The internship lasted 14 weeks in both groups and was accompanied by seminar sessions before, during, and after the school placement. While structural similarities existed between the groups, context factors such as COVID-19 and less seminar contact time in the intervention cohort limit direct comparability. Furthermore, students in the non-intervention group wrote six written reflections during the internship, while students in the intervention cohort wrote only three reflections each, which limits longitudinal comparability because stability effects may partly result from fewer measurement occasions rather than genuine differences in reflective development.

No systematic demographic or background data (e.g., age, gender, prior teaching experience, digital skills, initial writing competence) were collected for the cohorts. Therefore, baseline comparability between cohorts cannot be established. The cohort comparison must consequently be interpreted as descriptive and exploratory rather than inferential. No additional background, demographic, or competence-related data exist beyond the written reflections and acceptance survey responses reported in this study. Consequently, it is impossible to reconstruct baseline equivalence retrospectively, and no further descriptive characterization of the cohorts can be provided.

3.2. Intervention

As an intervention the feedback concept “ReFeed” was implemented in a teaching internship at the University of Potsdam, whereby feedback was given on written reflections of the preservice physics teachers (Mientus et al., 2021). The intervention was conceptually grounded in the feedback model of Hattie and Timperley (2007) to structure feedback on the elements and content of preservice teachers’ written reflections. In the preliminary seminar, the teachers received instructions on the reflection-supporting model of Nowak et al. (2019) to clarify how to reflect in a structured way (Feed Up). During the teaching practice, teachers submit written reflections on situations from their lessons three times. After each submission, teachers receive two documents of written feedback (see Figure 1). All documents are then stored in a secure drop box and shared between teachers and lecturers using pseudonyms. Examples of these documents can be viewed at the following link: lukasmientus.github.io/refeed/example-feedback.pdf.

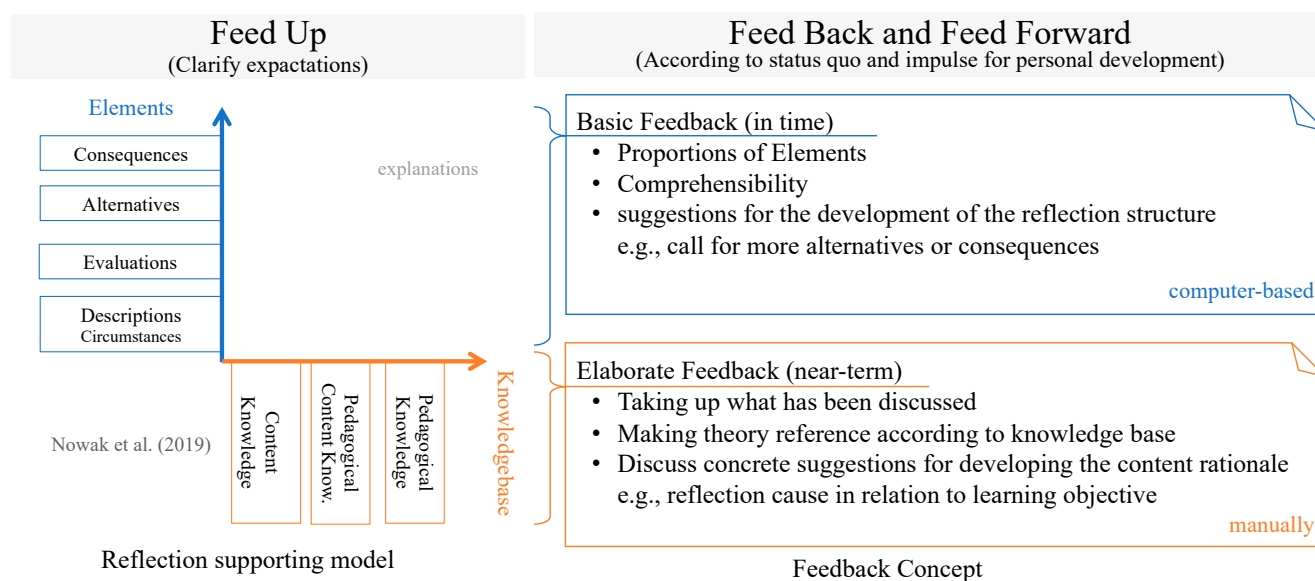


Figure 1. Feedback Concept including Feed Up, Feed Back and Feed Forward.

Teachers first receive a computer-generated feedback document (basic feedback) on implementing the structural elements of the reflection-supporting model within one day. For this, Wulff et al. (2020) trained an ML algorithm (based on BERT) to code the reflection elements of Nowak et al. (2019) (i.e., Circumstances, Descriptions, Evaluations, Alternatives, Consequences) in written reflections sentence-wise. Using this automated classification the basic feedback includes percentages, indicating to what extent each discourse element of the reflection-supporting model the teachers formulated in their written reflection (Feed Back). Furthermore, the computer-based algorithm provides specific suggestions, i.e., to focus more intensely on alternatives or consequences for further reflection (Feed Forward). This scalable and cost-effective computer-based feedback may support teachers to learn how to implement the reflection-supporting model of Nowak et al. (2019). Thus, it offers a potentially valuable resource for engaging in higher-order reflection processes.

Beyond basic feedback, lecturers provide significant support to teachers by giving a second feedback document (elaborate feedback) that addressed the substance of their reflections in a short time (after 14 days at the latest). Considering the structural analysis of the computer-based learning algorithm, lecturers identify critical discourses of teachers' pedagogical reasoning in the written reflection. In the elaborate feedback, occasions for reflection are singled out based on a normative concept (e.g., basics of teaching quality, scientific experimentation cycle, learning goal orientation, or reference to teacher ideas) (Feed Back). This framework encourages teachers to deepen their reflection process in terms of content, for example, by questioning the extent to which a selected focus of an experimental situation is conducive to the learning objective (Feed Forward). For this elaborate feedback a simple guideline of keywords and the individual basic feedback was used by the lecturers who have teaching expertise of more than three years.

The pedagogical approach aims to employ learning-effective, cost-efficient, and reliable feedback under authentic internship conditions. With our feedback concept we aimed to provide structured support for reflective writing during the internship. This explorative study compares the development with our concept and a second cohort that was not accompanied by the intervention.

3.3. Instruments

3.3.1. Validity and Transferability of the BERT-Based Classification Model

Wulff et al. (2020) investigated the use of machine learning (ML) in conjunction with natural language processing (NLP) to enable analytic, formative assessment of written reflections by physics (domain experts) and non-physics (domain novices) preservice teachers on a standardized teaching vignette (free fall physics lesson). They reused a pretrained BERT-based language model originally trained on physics preservice teacher data to classify sentence segments into reflection-supporting model categories (Circumstances, Description, Evaluation, Alternatives, Consequences) and compared performance before and after fine-tuning with non-physics context data. Their pretrained model (ML-base) applied to the non-physics test data achieved a macro F1 score of 0.52 and weighted F1 of 0.67 with Cohen's $\kappa \approx 0.53$ –0.59; after further fine-tuning (ML-finetuned), performance improved to a macro F1 of 0.58 and weighted F1 of 0.74 with Cohen's $\kappa \approx 0.63$ –0.64, indicating substantial human–machine agreement for filtering higher-level reasoning elements. Using the fine-tuned model to extract higher-level reasoning sentences, Wulff et al. (2020) applied BERT embeddings to identify qualitatively interpretable topics that differentiated physics-specific versus general content and found that longer texts included more physics-specific clusters, with topic distributions and coherence metrics reflecting expert-like writing. Further, human–machine agreement on text quality indicators (e.g., topic presence, text length) ranged from fair to poor depending on raters' familiarity, underscoring challenges in manual evaluation relative to ML-based measures.

Empirical work has demonstrated that transformer-based language models such as BERT can be employed as robust analytic instruments for reflective writing when grounded in a stable theoretical framework and applied to a consistent text genre. Sorge et al. (2025) evaluated the transferability and diagnostic validity of the BERT-based classifier of Wulff et al. (2021). Trained in the original context (Context A) and applied to an independent context (Context B) without additional fine-tuning, the model achieved substantial agreement with human coding (weighted F1 = 0.72). When training data from both contexts were combined, performance increased to a weighted F1 of 0.77, indicating that contextual fine-tuning yields incremental improvements but is not required for valid application. The reliability of the human gold standard was high in both contexts (Cohen's $\kappa > 0.73$ in Context A; $\kappa > 0.95$ in Context B), providing a strong benchmark for model evaluation. Beyond sentence-level classification accuracy, the model was successfully used to aggregate discourse proportions and trace longitudinal shifts in reflection structure, revealing measurable changes from descriptive to evaluative components across multiple writing tasks.

All analytic instruments used in this study have been validated and documented extensively in prior publications, which include implementation details, performance benchmarks, and methodological transparency; therefore, no additional technical documentation is reproduced here.

Together with prior validation results (Wulff et al., 2022), these findings support the use of a once-fine-tuned BERT model as a transferable measurement instrument for analyzing structural features of reflective writing. Accordingly, in the present study the classifier was employed as a validated analytic tool rather than a model under development, and no additional performance metrics (e.g., F1-scores) were computed for the current dataset (Wulff et al., 2022; Sorge et al., 2025).

3.3.2. Analyzing Written Reflections (RQ1)

Building on these architectural affordances, the classification algorithm of Wulff et al. (2022) assigns each sentence of a written reflection to one of the discourse elements of

the reflection-supporting framework (Nowak et al., 2019) on a sentence-by-sentence basis. In doing so, the model operationalizes reflective reasoning as a distribution of theoretically defined discourse elements rather than as a semantic interpretation of individual teaching situations. This allows complex reflection processes to be represented in a structured and comparable form across texts and cohorts.

Assuming a theoretically motivated target distribution of discourse elements in a written reflection of high quality (Mientus et al., 2023a), the text starting with 35% Circumstances and Descriptions, following by 35% Evaluations and ends with 15% Alternatives and 15% Consequences. Standardizing this distribution on the text length of a written reflection, Mientus et al. (2023a) validated an automated quality indicator (Level of Structure). The Level of Structure (LOS) is computed analogously to a Cohen's κ Coefficient of agreement between each written reflection and the assumed standard distribution. Logically, the LOS can take values from -1 (maximally reversed order), through 0 (maximal disorder), up to 1 (perfect order and distribution). Typically, LOS values range from -0.1 to 0.7 (Mientus et al., 2023b). This indicator allows quality estimation independent of text length, which is a common proxy for quality in written reflections (Chodorow & Burstein, 2004; Leonhard & Rihm, 2011).

In summary, for each written reflection the following data were available for analysis: (1) text length, (2) absolute proportions of discourse elements according to Nowak et al. (2019), (3) relative proportions of descriptive (Circumstances and Descriptions) and discursive (Evaluations, Alternatives, and Consequences) elements, and (4) the Level of Structure as a structural quality indicator (Mientus et al., 2023a). Effect sizes were reported as rank-biserial correlations (r_{rb}) for Wilcoxon rank-sum tests and as marginal and conditional R^2 for mixed-effects models.

3.3.3. Feedback Acceptance Survey (RQ2)

Alongside implementing the feedback in the intervention cohort, a validated acceptance survey was administered after each of the two feedback phases. The instrument comprised Likert-scaled items across five dimensions (see Wulff et al. (2021); items available under lukasmientus.github.io/refeed/acceptance-survey.pdf): (1) Effectiveness (perceived impact on improvement), (2) Usefulness (stimulation of reflective thinking), (3) Personalization (individual relevance to one's reflection), (4) Subjective Accuracy (perceived correctness of the feedback), and (5) Values (importance attributed to reflection and reflective writing after receiving feedback). The acceptance instrument comprised four primary Likert-type scales ($1 = \text{"strongly disagree"}$ to $4 = \text{"strongly agree"}$), each measured with 2–4 items such as the example item for Effectiveness ("The feedback helped me better understand how to structure my reflection"). The internal consistency of these scales can be reported using Cronbach's α with values indicating acceptable reliability for Effectiveness ($\alpha = 0.73$), Usefulness ($\alpha = 0.85$), and Subjective Accuracy ($\alpha = 0.82$), while the Personalization scale showed lower reliability ($\alpha = 0.55$), consistent with its narrower item set.

4. Results

4.1. Development of Written Reflections in Comparison (RQ1)

4.1.1. Distributional Properties and Initial Group Differences

Analogous to the comparison of both cohorts in terms of internship structure, this section compares both cohorts with respect to (1) text length, (2) discourse elements, and (3) Level of Structure (LOS) as a quality indicator. For the between-cohort comparison, analyses were conducted at the participant level to account for the nested data structure (multiple reflections per preservice teacher). Specifically, one aggregate value per participant was computed by taking the median across all available reflections. Accordingly, the

comparison is based on $N = 22$ preservice teachers in the business-as-usual cohort and $N = 32$ preservice teachers in the intervention cohort.

Shapiro–Wilk tests indicated significant deviations from normality for sentence counts in both cohorts (participant-level medians; business-as-usual: $W < 0.95$, $p < 0.05$; intervention: $W < 0.95$, $p < 0.05$). Accordingly, all between-cohort comparisons were treated as non-parametric, and results are reported using medians and interquartile ranges (IQRs) rather than means and standard deviations. Table 1 presents descriptive statistics based on participant-level medians (one value per preservice teacher). On the participant level, sentence counts showed no meaningful differences between cohorts (median_{intervention} = 44.25; median_{non-intervention} = 34.25; $p = 0.411$), indicating comparable overall text length.

Table 1. Comparison of reflection characteristics for both cohorts based on participant-level medians (one value per preservice teacher). Note. Values represent medians aggregated per participant (median across all reflections per preservice teacher) *.

Cohort	Non-Intervention (N = 22)		Intervention (N = 32)		MWU-Test (Participant-Level Aggregates)		
	Median	iqr	Median	iqr	U	p	R(rb)
Sentences	36	27	44	29	423.5	0.411	0.13
Circumstances (in %)	33	14	22	12	105.0	6.641	−0.72
Descriptions (in %)	34	14	40	12	565.0	0.001	0.51
All descriptive elements in sum	67	15	63	12	248.0	0.035	−0.34
Evaluations (in %)	21	10	24	12	485.0	0.064	0.30
Alternatives (in %)	6	8	7	6	439.5	0.275	0.18
Consequences (in %)	4	3	6	45	533.5	0.008	0.43
All discursive elements in sum	33	15	38	12	500.0	0.035	0.34

* Note. All variables were aggregated at the participant level by computing the median across all available reflections per preservice teacher. Between-cohort comparisons were conducted on these participant-level aggregates to account for the nested structure of the data (multiple reflections per participant). Effect sizes are reported as rank-biserial correlations (r_{rb}).

Regarding discourse composition, clear distributional differences emerged between cohorts. The proportion of Circumstances was substantially higher in the business-as-usual cohort, whereas the intervention cohort showed higher proportions of Descriptions, Evaluations, and Consequences. Differences in Alternatives were not statistically meaningful. When discourse elements were aggregated, the intervention cohort exhibited a lower proportion of descriptive elements (Circumstances + Descriptions) and a correspondingly higher proportion of discursive elements (Evaluations + Alternatives + Consequences). These patterns reflect a structural shift in reflective focus rather than differences in text length.

Between-cohort comparisons were conducted using Wilcoxon rank-sum tests on participant-level aggregates; effect sizes are reported as rank-biserial correlations (r_{rb}). Medium to large effects were observed for several indicators. In particular, strong differences emerged for Circumstances ($r_{rb} = -0.72$) and Descriptions ($r_{rb} = 0.51$), as well as for the Level of Structure (LOS; $r_{rb} = 0.45$). Differences for Consequences (absolute and relative) were of medium magnitude, whereas effects for Evaluations were small to medium. Differences in Alternatives and sentence count were negligible. At the participant level, reflections associated with the intervention cohort showed a different structural profile, characterized by a lower emphasis on contextual framing and a higher emphasis on descriptive, evaluative, and consequential elements, accompanied by higher LOS values.

4.1.2. Longitudinal Development Across the Internship

Text length did not change significantly across measurement points in either cohort. In the business-as-usual cohort, median sentence counts showed a gradual decrease across reflections, whereas in the intervention cohort median text length remained relatively stable across the three submissions. However, given the unequal number of reflection occasions (six vs. three), longitudinal comparisons between cohorts are structurally constrained. Figure 2 illustrates the distribution of descriptive and discursive elements across reflection occasions. Across the first three reflection occasions, participant-level median proportions indicate that the business-as-usual cohort consistently exhibited a higher proportion of descriptive elements, whereas the intervention cohort showed a comparatively higher share of discursive elements from the outset. These patterns were also reflected in the Level of Structure (LOS) indicator. Median LOS values were consistently higher in the intervention cohort at all comparable measurement points. In both cohorts, LOS values declined over time. This decline was descriptively observable but did not eliminate the between-cohort difference present at the first reflection. Taken together, longitudinal patterns indicate descriptively observable between-cohort differences in reflection structure across comparable measurement points, without allowing conclusions about stability, growth, or intervention effects. No statistically significant within-cohort changes were observed for text length, discourse composition, or LOS. All longitudinal trends should therefore be interpreted descriptively.

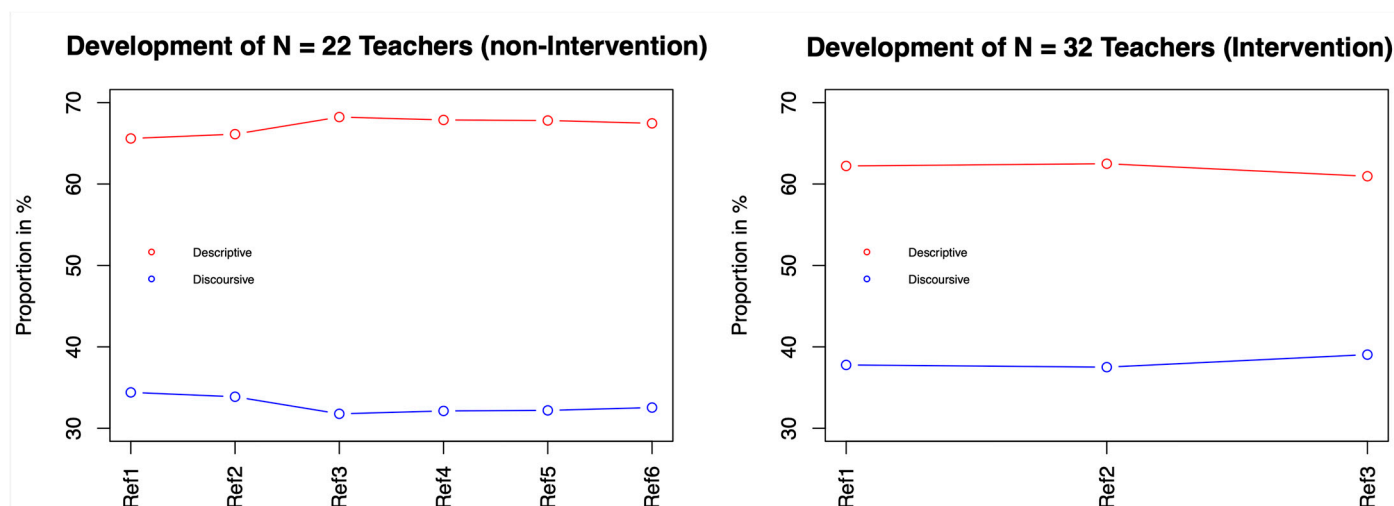


Figure 2. relative distributions of descriptive and discursive elements across reflection occasions for both cohorts.

The development of the Level of Structure over time did not change significantly in either cohort. To account for repeated measures, mixed-effects models were estimated separately for each cohort. Model fit was quantified using marginal and conditional R^2 values. In the intervention cohort, fixed effects explained approximately 27% of the variance in LOS (marginal $R^2 = 0.27$), while the full model including random effects explained 64% (conditional $R^2 = 0.64$). In the business-as-usual cohort, corresponding values were marginal $R^2 = 0.25$ and conditional $R^2 = 0.48$. Across both cohorts, a substantial proportion of variance was attributable to between-individual differences, indicating heterogeneous reflection trajectories. Figure 3 illustrates descriptive changes in LOS across reflection occasions. In both cohorts, LOS values showed a downward trend over time; however, the intervention cohort maintained higher LOS values across all comparable measurement points.

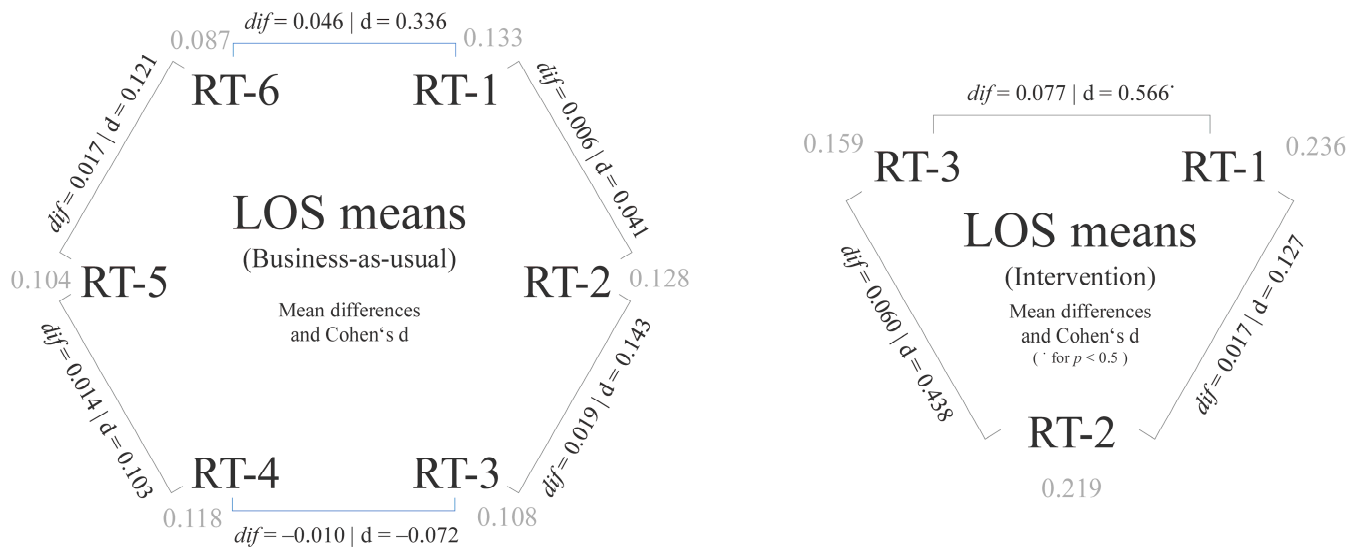


Figure 3. Development of the Level of Structure (LOS) across reflection occasions in both cohorts. (Reflection texts—RT-1 ... RT-6).

4.1.3. Cluster-Based Analysis of Structural Quality

To account for heterogeneity in baseline reflection structure, reflections were grouped using baseline LOS values. This analysis is included as a methodological illustration rather than as a substantive result. Clustering solutions differed by cohort and yielded uneven cluster sizes, particularly in the intervention cohort, where one cluster comprised only three participants. Mixed-effects models fitted separately for each cohort indicated descriptive differences in LOS levels between clusters. However, these patterns are highly sensitive to small and uneven cluster sizes. In both cohorts, the highest-LOS clusters exhibited declines over time, whereas lower-LOS clusters showed smaller changes or slight increases. Taken together, the cluster-based illustrations highlight heterogeneous and highly case-sensitive trajectories in LOS, may be driven by declines in initially high-LOS clusters and comparatively stable or slightly increasing trends in lower-LOS clusters. Given the exploratory design, non-equivalent cohorts, and small cluster sizes, these findings are descriptive and should be interpreted with substantial caution. The illustration in Figure 4 shows the LOS per person (thin lines) and the clusters, represented by the mean (thick line) and standard deviation range (colored area).

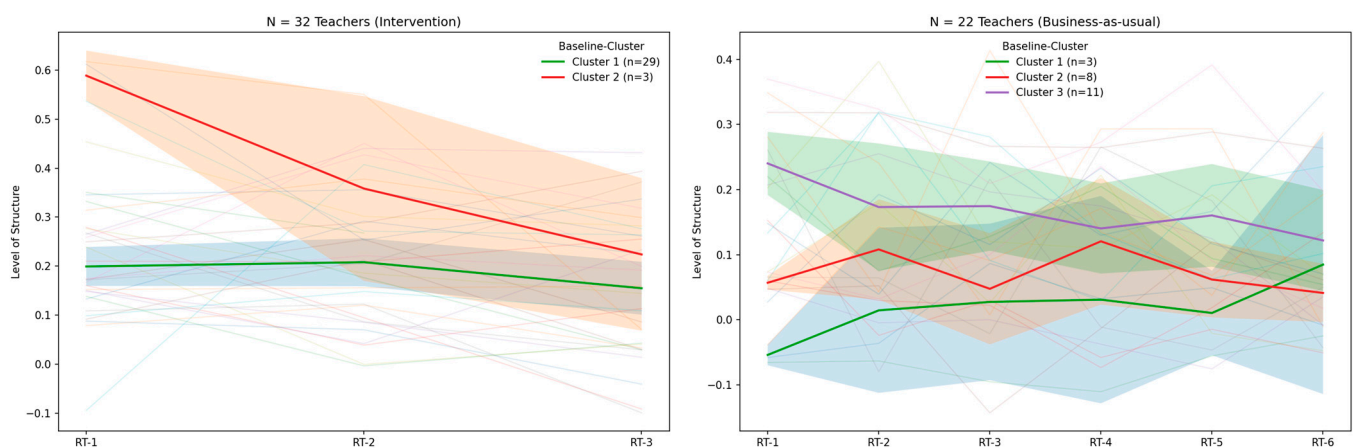


Figure 4. Cluster-based development of the Level of Structure (LOS) in both cohorts.

4.2. Feedback Acceptance (RQ2)

An acceptance survey assessed how preservice physics teachers evaluate the basic and elaborate feedback in the intervention cohort. Preservice physics teachers' perceptions of the two feedbacks are illustrated in Figure 5. In boxplots, the statistical values from the acceptance survey become obvious. The diagram is based on data from five semesters from October 2020 to March 2023. The value attributed to reflective writing is emphasized irrespective of which feedback was provided previously (Wilcoxon rank-sum test: $W = 445$, $p = 0.130$). Additionally, preservice physics teachers rate the subjective accuracy of both types of feedback comparably high ($W = 334$, $p = 0.004$). However, structural feedback receives notably more negative ratings in terms of effectiveness ($W = 120$, $p < 0.001$), usefulness ($W = 117$, $p < 0.001$), and personalization ($W = 710$, $p < 0.001$). Elaborate (human-generated) feedback received significantly higher ratings for effectiveness, usefulness, and personalization compared to basic (AI-generated) feedback, according to Wilcoxon rank-sum tests.

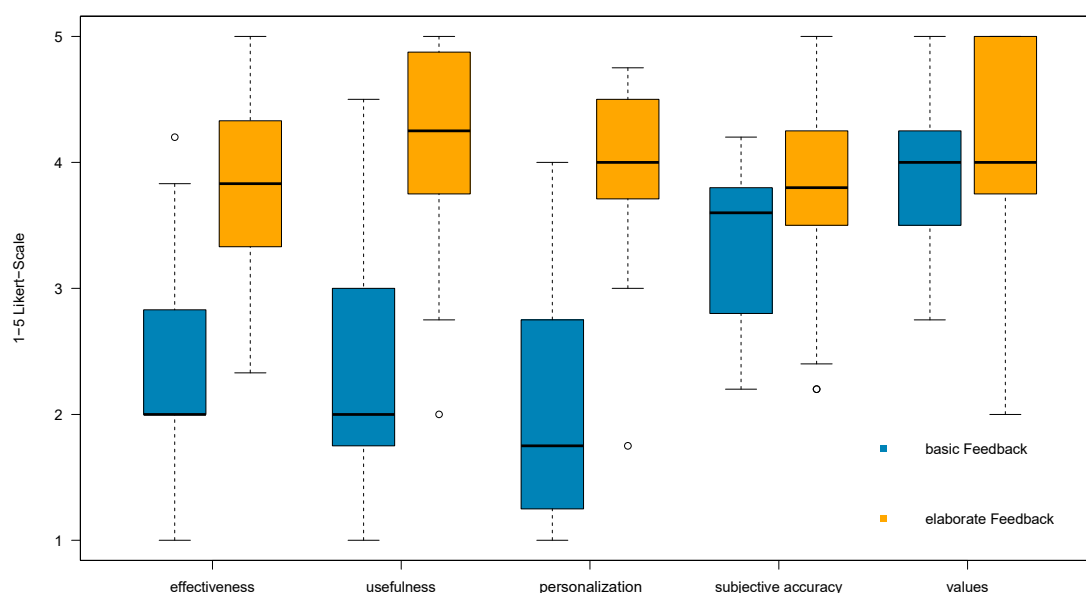


Figure 5. Acceptance of basic and elaborate feedback in the intervention cohort.²

5. Discussion, Limitations, and Conclusions

5.1. Discussion of Results

This study investigated the integration of early AI-supported feedback for preservice physics teachers on reflective writing during teaching internships. Participant-level analyses revealed systematic differences in the structural composition of reflective writing between cohorts. Median reflection profiles of preservice teachers in the intervention cohort were characterized by lower proportions of contextual Circumstances and higher proportions of Descriptions, Evaluations, and Consequences, accompanied by higher values on the Level of Structure (LOS). These findings are consistent with theoretical assumptions that structured feedback is associated with shifts in reflective focus toward discursive reasoning elements (e.g., Thammasitboon et al., 2018). Importantly, these participant-level differences emerged independently of median text length per preservice teacher, suggesting qualitative rather than quantitative variation in reflective writing. Across both cohorts, no systematic longitudinal increase in reflective quality was observed. Instead, reflection indicators showed stable or declining trajectories over time, irrespective of feedback condition. Nevertheless, across comparable measurement points, participant-level median LOS values were descriptively higher in the intervention cohort. These differences were

present from the first reflection onward and remained descriptively stable throughout the internship. Given the non-equivalent, non-concurrent cohort design, these findings must be interpreted as descriptive associations rather than effects of the intervention. In particular, it cannot be determined whether the observed differences reflect the feedback approach itself, cohort-specific characteristics, or contextual conditions of the respective internship phases.

At the same time, the pattern of results were associated a structural shift in reflective focus rather than linear developmental growth. The intervention cohort's reflections showed less emphasis on situational framing and more emphasis on evaluative and consequential reasoning, which aligns with conceptualizations of higher-order reflective writing (Wyss, 2018; von Aufschnaiter et al., 2019). Notably, no corresponding increase was observed for Alternatives, indicating that reflective depth was expressed primarily through evaluation and consequence-building rather than through the generation of multiple action options. Cluster-based analyses further indicated heterogeneous reflection trajectories within both cohorts. In particular, participants with initially high LOS values tended to show declining trajectories, whereas lower-LOS clusters exhibited comparatively stable or slightly increasing patterns. Given small and uneven cluster sizes, these findings are illustrative and highlight individual variability rather than systematic intervention effects.

This interpretation is consistent with recent findings by Sorge et al. (2025), who showed that ML-based structural feedback primarily supports diagnostic transparency and reflective orientation rather than linear growth in reflection quality. From this perspective, the present findings suggest that early AI-supported feedback may be associated with structurally different reflection profiles without necessarily fostering progressive development during practice-intensive phases.

The developmental patterns observed in the present study differ from findings reported by Sorge et al. (2025), who investigated AI-supported feedback in a seminar-based university course. In their study, repeated reflective writing combined with automated structural feedback was associated with a gradual shift from descriptive toward more evaluative reflection components. In contrast, the present results show no clear longitudinal increase in reflective quality as measured by the Level of Structure (LOS). Instead, between-cohort differences remained descriptively stable over time.

A central explanation for this divergence lies in the different instructional contexts. Sorge et al. (2025) implemented their intervention in a seminar setting explicitly designed to support reflective writing, using video-based cases and repeated low-stakes writing tasks. By contrast, both cohorts in the present study produced their reflections during extended school-based practice phases. Reflective writing in internships is embedded in demanding professional situations characterized by time pressure, emotional involvement, and competing instructional responsibilities, which have been shown to constrain reflective elaboration (Voss & Kunter, 2020; Wyss, 2013). Under such conditions, reflection writing may be shaped more strongly by situational demands than by feedback intensity.

Accordingly, the present findings indicate that differences in reflection structure observed between cohorts cannot be straightforwardly interpreted as developmental effects. Rather, they point to context-dependent patterns of reflective engagement that may be shaped by instructional design, feedback framing, and internship conditions.

Preservice teachers consistently valued the reflective writing task. Yet, their evaluations revealed clear differences between types of feedback. Elaborate, human-generated feedback was rated as more effective, useful, and personalized. In contrast, computer-generated structural feedback, although immediate, was often perceived as impersonal and lacking depth. This discrepancy highlights a central tension between scalability and perceived pedagogical value in early AI-supported feedback systems. While generative

AI offers greater potential to address these issues, the present study provides an empirical baseline against which later developments can be evaluated.

5.2. Limitations

Several limitations must be considered. First, the exclusive reliance on written reflections captures only part of reflective competence. Internship phases often involve high professional demands, which may interfere with sustained reflective engagement (Bönke et al., 2024). Thus, declines in reflection quality may reflect contextual overload rather than shortcomings of the feedback approach.

Second, the non-equivalent and non-concurrent cohort design constitutes a major limitation. The two cohorts were embedded in different institutional, temporal, and societal contexts (2017–2018 vs. 2020–2023), including substantial disruptions related to the COVID-19 pandemic. These contextual differences may have affected supervision intensity, teaching opportunities, workload, and reflective engagement. No systematic demographic or background data (e.g., age, gender, prior teaching experience, digital skills, or initial writing competence) were collected for either cohort. As a result, baseline equivalence between cohorts cannot be established, and all observed differences must be interpreted as descriptive associations rather than effects of the feedback intervention. Moreover, longitudinal comparability is structurally constrained by the different numbers of reflection tasks across cohorts (three vs. six). Apparent stability or decline effects may therefore partly reflect measurement design rather than genuine developmental trajectories.

A further limitation concerns the involvement of study authors as lecturers in both cohorts. Although comparable instructional intentions guided supervision, lecturer involvement may have influenced students' engagement with reflective writing or their perceptions of feedback.

Finally, reflection quality was operationalized primarily through structural and distributional features of written reflections. While prior validation studies support the analytic use of the BERT-based model for this purpose, such indicators necessarily underrepresent content relevance, personal meaning, and pedagogical nuance (Jay & Johnson, 2002). Accordingly, the measures used capture structural aspects of reflective reasoning rather than holistic reflective competence.

5.3. Conclusions and Perspectives

This exploratory study does not provide evidence for the effectiveness of AI-supported feedback but illustrates how early AI-based tools can be used to descriptively analyze structural features of reflective writing under authentic internship conditions. Although the intervention cohort produced reflections with a higher proportion of discursive elements, no systematic within-cohort development was observed over time. This suggests that structured prompting may be associated with differences in reflective focus and organization without necessarily fostering cumulative growth during internship phases. Rather than indicating developmental effects, the findings point to context-dependent reflection patterns shaped by instructional framing, feedback design, and internship demands.

Cluster-based analyses further suggested heterogeneous individual trajectories, with declines among initially high-LOS writers and comparatively stable patterns among lower-LOS writers. These patterns underline the importance of accounting for individual variability when interpreting reflective development. Taken together, the present findings do not support claims about the effectiveness of AI-supported feedback in accelerating reflective development. Instead, they highlight the descriptive value of AI-based approaches for identifying structural patterns in reflective writing under authentic internship conditions.

Preservice teachers preferred feedback that was timely, elaborate, and personalized. While automated feedback ensured immediacy, its structural focus limited perceived usefulness. This finding underscores a persistent tension between scalability and pedagogical richness in automated feedback systems. Future feedback systems, particularly those based on generative AI, may better address this tension between scalability and personalization.

In summary, this study positions early, non-generative AI-based feedback as a scalable diagnostic tool for analyzing reflective writing rather than as an intervention that demonstrably improves reflective competence. Its contribution lies in enabling theory-based structural diagnostics and in establishing an empirical baseline against which contemporary generative AI feedback systems in teacher education can be evaluated.

Author Contributions: Conceptualization, L.M., P.W., A.N., & A.B.; methodology, L.M.; software, L.M.; validation, L.M., P.W., & A.N.; formal analysis, L.M.; investigation, L.M., P.W., A.N., & A.B.; resources, A.B.; data curation, L.M., P.W., & A.N.; writing—original draft preparation, L.M.; writing—review and editing, L.M., P.W., A.N., & A.B.; visualization, L.M.; supervision, A.B.; project administration, A.B.; funding acquisition, A.B. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by German Federal Ministry of Education and Research (BMBF) grant numbers 01JA1816 and 01JA23M05A. The APC was funded by the University of Potsdam.

Institutional Review Board Statement: All procedures performed in studies involving human participants or on human tissue were in accordance with the ethical standards of the institutional and/or national research committee and with the 1975 Helsinki declaration and its later amendments or comparable ethical standards. Informed consent was obtained from all individual participants included in the study. Based on a recommendation by the Institutional Review Board, ethical review and approval were waived for this study due to the strict non-collection of detailed personal data.

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: For data availability please contact the first author.

Conflicts of Interest: The authors declare no conflict of interest.

Notes

1. ALACT: (1) Action, (2) Looking back on the action, (3) getting Awareness of essential aspects, (4) Creating alternatives methods of Action and (5) Starting a new Action with a Trial of alternatives.
2. Boxes represent the interquartile range (IQR), the central line indicates the median, and whiskers correspond to $\pm 1.5 \times \text{IQR}$; outliers beyond this range are displayed as individual points. Sample sizes per time point are provided in Table 1/described in Section 3.1.

References

- Abels, S. (2011). *LehrerInnen als „Reflective Practitioner“*. Reflexionskompetenz für einen demokratieförderlichen Naturwissenschaftsunterricht. vs. Verlag.
- Alammar, J. (2018). *The illustrated transformer*. Available online: <https://jalammar.github.io/illustrated-transformer/> (accessed on 9 January 2026).
- Alonzo, A., Berry, A., Nilsson, P. (2019). Unpacking the Complexity of Science teachers' PCK in Action. Enacted and Personal PCK. In A. Hume, R. Cooper & A. Borowski (Eds.), *Repositioning pedagogical content knowledge in teachers' professional knowledge* (pp. 271–286). Springer. https://doi.org/10.1007/978-981-13-5898-2_12.
- Altrichter, H. (2000). Handlung und reflexion bei Donald Schön. In G.H. Neuweg (Eds.), *Wissen — Können — Reflexion* (pp. 201–221). Studien-Verlag.

- Bönke, N., Klusmann, U., Kunter, M., Richter, D., & Voss, T. (2024). Long-term changes in teacher beliefs and motivation: Progress, stagnation or regress? *Teaching and Teacher Education*, 141, 104489. <https://doi.org/10.1016/j.tate.2024.104489>.
- Brouwer, N., & Korthagen, F. A. J. (2005). Can teacher education make a difference? *American Educational Research Journal*, 42, 153–224.
- Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., & Amodei, D. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33, 1877–1901. <https://doi.org/10.48550/arXiv.2005.14165>.
- Browning, T. D., & Korthagen, F. A. J. (2021). The winding road of student teaching: Addressing Uncertainty with core reflection. *European Journal of Teacher Education*, 46(4), 621–638. <https://doi.org/10.1080/02619768.2021.1973421>.
- Buckingham Shum, S., Sándor, Á., Goldsmith, R., Bass, R., & McWilliams, M. (2017). Towards reflective writing analytics: Rationale, methodology, and preliminary results. *Journal of Learning Analytics*, 4(1), 58–84. <https://doi.org/10.18608/jla.2017.41.5>.
- Chodorow, M., & Burstein, J. (2004). Beyond essay length. Evaluating e-rater's performance on Toefl essays. *ETS. Research Report Service*, 2004, i-38. <https://doi.org/10.1002/j.2333-8504.2004.tb01931.x>.
- Christof, E., Köhler, J., Rosenberger, K., & Wyss, C. (2018). *Mündliche, schriftliche und theatrale Wege der Praxisreflexion: Beiträge zur professionalisierung pädagogischen handelns* (1st ed.). Hep der Bildungsverlag.
- Collin, S., Karsenti, T., & Komis, V. (2013). Reflective practice in initial teacher training: Critiques and perspectives. *Reflective Practice*, 14(1), 104–117. <https://doi.org/10.1080/14623943.2012.732935>.
- Combe, A., & Kolbe, F.-U. (2008). Lehrerprofessionalität: Wissen, Können, Handeln. In W. Helsper & J. Böhme (Eds.), *Handbuch der schulforschung* (pp. 857–875). Springer.
- Darling-Hammond, L. (2012). *Powerful teacher education: Lessons from exemplary programs*. John Wiley & Sons.
- Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv*, arXiv:1810.04805. <https://doi.org/10.48550/arXiv.1810.04805>.
- Fund, Z., Court, D., & Kramarski, B. (2002). Construction and Application of an Evaluative Tool to Assess Reflection in Teacher-Training Courses. *Assessment & Evaluation in Higher Education*, 27(6), 485–499. <https://doi.org/10.1080/0260293022000020264>.
- Gröschner, A., Schmitt, C., & Seidel, T. (2013). Veränderung subjektiver Kompetenzeinschätzungen von Lehramtsstudierenden im Praxissemester. *Zeitschrift für Pädagogische Psychologie*, 27(1/2), 77–86. <https://doi.org/10.1024/1010-0652/a000090>.
- Gruber, H. (2021). Reflexion. Der Königsweg zur Expertise-Entwicklung. *Journal für LehrerInnenbildung*, 21(1), 108–117. <https://doi.org/10.35468/jlb-01-2021-10>.
- Hattie, J., & Timperley, H. (2007). The Power of Feedback. *Review of Educational Research*, 77(1), 81–112. <https://doi.org/10.3102/003465430298487>.
- Hatton, N., & Smith, D. (1995). Reflection in Teacher Education: Towards Definition and Implementation. *Teaching and Teacher Education*, 11(1), 33–49. [https://doi.org/10.1016/0742-051X\(94\)00012-U](https://doi.org/10.1016/0742-051X(94)00012-U).
- Häcker, T. H. (2019). Reflexive Professionalisierung. Anmerkungen zu dem ambitionierten Anspruch, die Reflexionskompetenz angehender Lehrkräfte umfassend zu fördern. In M. Degeling, N. Franken, S. Freund, S. Greiten, D. Neuhaus, & J. Schellenbach-Zell (Eds.), *Herausforderung Kohärenz: Praxisphasen in der universitären Lehrerbildung: Bildungswissenschaftliche und fachdidaktische Perspektiven* (pp. 81–96). Verlag Julius Klinkhardt.
- Jay, J. K., & Johnson, K. L. (2002). Capturing complexity: A typology of reflective practice for teacher education. *Teaching and Teacher Education*, 18(1), 73–85. [https://doi.org/10.1016/S0742-051X\(01\)00051-8](https://doi.org/10.1016/S0742-051X(01)00051-8).
- Klempin, C. (2021). Zu Entwicklung und Messung von Reflexionstiefe und -breite von Lehramtsstudierenden. Eine mixed methods interventionsstudie. *Journal für LehrerInnenbildung*, 21(1), 76–85. <https://doi.org/10.35468/jlb-01-2021-07>.
- KMK (Sekretariat der Ständigen Konferenz der Kultusminister der Länder in der Bundesrepublik Deutschland). (2004). *Standards für die Lehrerbildung: Bildungswissenschaften*. Sekretariat der Ständigen Konferenz der Kultusminister der Länder in der Bundesrepublik Deutschland.
- KMK (Sekretariat der Ständigen Konferenz der Kultusminister der Länder in der Bundesrepublik Deutschland). (2019). *Standards für die Lehrerbildung: Bildungswissenschaften (Beschluss der Kultusministerkonferenz vom 16.12.2004 i. d. F. vom 16.05.2019)*. Available online: www.kmk.org (accessed on 9 January 2026).
- Korthagen, F. A. J. (2001). *Linking practice and theory. The pedagogy of realistic teacher education*. Lawrence Erlbaum Associates.
- Korthagen, F. A. J., & Kessels, J. (1999). Linking theory and practice: Changing the pedagogy of teacher education. *Educational Research*, 28(4), 4–17.

- Labott, D., & Reintjes, C. (2022). Unvereinbarkeit von Bewertung und Reflexionsaufgaben in der Lehrer*innenbildung. In C. Reintjes, & I. Kunze (Eds.), *Reflexion und reflexivität in unterricht, schule und lehrer: Innenbildung* (pp. 170–184). Klinkhardt. <https://doi.org/10.35468/5969-10>.
- Lai, G., & Calandra, B. (2007). Using online scaffolds to enhance preservice teachers' reflective journal writing: A qualitative analysis. *International Journal of Technology in Teaching and Learning*, 3(3), 66–81.
- Leonhard, T., & Rihm, T. (2011). Erhöhung der Reflexionskompetenz durch Begleitveranstaltungen zum Schulpraktikum? Konzeption und Ergebnisse eines Pilotprojekts mit Lehramtsstudierenden. *Lehrerbildung auf dem Prüfstand*, 4(2), 240–270.
- Mientus, L., Wulff, P., Nowak, A., & Borowski, A. (2021). ReFeed: computerunterstütztes Feedback zu Reflexionstexten - Ein Lehrkonzept zur Förderung der Reflexionskompetenz angehender Physiklehrkräfte an der Universität Potsdam. In: M. Kubsch, S. Sorge, J. Arnold, & N. Graulich (Eds.): *Lehrkräftebildung neu gedacht—Ein praxishandbuch für die lehre in den naturwissenschaften und deren didaktiken* (pp. 160–165). Waxmann. <https://doi.org/10.31244/9783830993490>.
- Mientus, L., Wulff, P., Nowak, A., & Borowski, A. (2023a). Fast-and-frugal means to assess reflection-related reasoning processes in teacher training—Development and evaluation of a scalable machine learning-based metric. *Zeitschrift für Erziehungswissenschaften*. <https://doi.org/10.1007/s11618-023-01166-8>.
- Mientus, L., Wulff, P., Nowak, A. & Borowski, A. (2023b). Algorithmen als Dozierende? Akzeptanz von KI-basierten Lernangeboten in der Physik-Lehrkräftebildung. In: J. Hermanns (Eds.), *PSI-Potsdam—Ergebnisbericht zu den aktivitäten im rahmen der qualitätsinitiative lehrerbildung (2019–2023)* (pp. 117–129). Universitätsverlag Potsdam. <https://doi.org/10.25932/publishup-60187>.
- Meißner, C., Klempin, C., Dohrmann, R., & Nordmeier, V. (2020). Veränderung der Reflexionskompetenz im Lehr-Lern-Labor. In S. Habig (Eds.), *Naturwissenschaftliche kompetenz in der gesellschaft von morgen. Gesellschaft für didaktik der chemie und physik, jahrestagung in Wien 2019* (pp. 685–688). Universität Duisburg-Essen.
- Meschede, N. (2014). Professionelle Wahrnehmung der inhaltlichen Strukturierung im naturwissenschaftlichen Grundschulunterricht. Theoretische Beschreibung und empirische Forschung. In M. Hopf, H. Niedderer, M. Ropohl, & E. Sumfleth (Eds.), *Studien zum physik- und chemielernen* (Vol. 163). Logos-Verlag.
- Narciss, S. (2006). *Informatives tutorielles Feedback: Entwicklungs- und Evaluationsprinzipien auf der Basis instruktionspsychologischer Erkenntnisse* (Vol. 56). Waxmann.
- Narciss, S. (2018). Feedbackstrategien für interaktive Lernaufgaben. In Niegemann, H., Weinberger, A. (Eds.), *Lernen mit bildungstechnologien*. Springer Reference Psychologie. https://doi.org/10.1007/978-3-662-54373-3_35-1.
- NBPTS (National Board for Professional Teaching Standards). (2016). *What teachers should know and be able to do*. National Board for Professional Teaching Standards.
- Nordine, J., Sorge, S., Delen, I., Evans, R., Juuti, K., Lavonen, J., Nilsson, P., Ropohl, M., & Stadler, M. (2021). Promoting coherent science instruction through coherent science teacher education: A model framework for program design. *Journal of Science Teacher Education*, 32(8), 911–933. <https://doi.org/10.1080/1046560X.2021.1902631>.
- Nowak, A., Kempin, M., Kulgemeyer, C., & Borowski, A. (2019). Reflexion von Physikunterricht. In C. Maurer (Eds.), *Naturwissenschaftliche bildung als grundlage für berufliche und gesellschaftliche teilhabe. Gesellschaft für didaktik der chemie und physik, jahrestagung in Kiel 2018* (Vol. 11, pp. 838–841). Universität Regensburg.
- Nowak, A. (2023). Untersuchung der Qualität von Selbstreflexionstexten zum Physikunterricht. Entwicklung des Reflexionsmodells REIZ. In M. Hopf & M. Ropohl (Eds.), *Studien zum physik- und chemielernen*. Logos. <https://doi.org/10.30819/5739>.
- OpenAI. (2019). *Better language models and their implications (GPT-2)*. OpenAI Blog.
- OpenAI. (2022). *ChatGPT: Optimizing language models for dialogue*. OpenAI Blog.
- Poldner, E., van der Schaaf, M., Simons, P. R.-J., van Tartwijk, J., & Wijngaards, G. (2014). Assessing student teachers' reflective writing through quantitative content analysis. *European Journal of Teacher Education*, 37(3), 348–373.
- Roters, B. (2012). *Professionalisierung durch reflexion in der lehrerbildung*. Waxmann.
- Rothland, M. (2021). Die „Lehrerpersönlichkeit“: Das Geheimnis des Lehrberufs?. *Die Deutsche Schule*, 113(2), 188–198.
- Schön, D. A. (1983). *The reflective practitioner: How professionals think in action*. Basic Books.
- Schubarth, W., Speck, K., Seidel, A., & Wendland, M. (2009). Unterrichtskompetenzen bei Referendaren und Studierenden. Empirische Befunde der Potsdamer Studien zur ersten und zweiten Phase der Lehrerausbildung. *Lehrerbildung auf dem Prüfstand* 2(2), 304–323. <https://doi.org/10.25656/01:14705>.
- Sorge, S., Stender, A., & Neumann, K. (2019). The development of science teachers' professional competence. In A. Hume, R. Cooper, & A. Borowski (Eds.), *Repositioning pedagogical content knowledge in teachers' knowledge for teaching science* (pp. 149–164). Springer. https://doi.org/10.1007/978-981-13-5898-2_6.

- Sorge, S., Wulff, P. & Kubsch, M. (2025) Using a large language model to provide individualized feedback for pre-service physics teachers' written reflections. *Disciplinary and Interdisciplinary Science Education Research* 7, 25. <https://doi.org/10.1186/s43031-025-00145-9>.
- Thammasitboon, S., Rencic, J., Trowbridge, R., Olson, A., Sur, M., & Dhaliwal, G. (2018). The Assessment of Reasoning Tool (ART): Structuring the conversation between teachers and learners. *Diagnosis*, 5(4), 197–203. <https://doi.org/10.1515/dx-2018-0052>.
- Ullmann, T. D. (2019). Automated analysis of reflection in writing: Validating machine learning approaches. *International Journal Artificial Intelligence in Education*, 29(2), 217–257.
- van Es, E. A., & Sherin, M. G. (2021). Expanding on prior conceptualizations of teacher noticing. *ZDM Mathematics Education*, 53(1), 17–27. <https://doi.org/10.1007/s11858-020-01211-4>.
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. Advances in neural information processing systems. *arXiv*, arXiv:1706.03762. <https://doi.org/10.48550/arXiv.1706.03762>.
- von Aufschnaiter, C. (2023): Reflexive Professionalisierung: Zentral—Vielschichtig—Herausfordernd. In L. Mientus, A. Nowak, & C. Klempin (Eds.), *Reflexion in der Lehrkräftebildung—Empirisch, phasenübergreifend, interdisziplinär*. Universitätsverlag Potsdam. <https://doi.org/10.25932/publishup-59171>.
- von Aufschnaiter, C., Fraij, A., & Kost, D. (2019). Reflexion und Reflexivität in der Lehrerbildung. *Herausforderung Lehrer*innenbildung—Zeitschrift zur Konzeption, Gestaltung und Diskussion (HLZ)*, 2(1), 144–159. <https://doi.org/10.4119/HLZ-2439>.
- Voss, T., & Kunter, M. (2020). “Reality shock” of beginning teachers? Changes in teacher candidates' emotional exhaustion and constructivist-oriented beliefs. *Journal of Teacher Education*, 71(3), 292–306. <https://doi.org/10.1177/0022487119839700>.
- Wulff, P., Buschhüter, D., Westphal, A., Nowak, A., Becker, L., Robalino, H., Stede, M., & Borowski, A. (2020). Computer-Based Classification of Preservice Physics Teachers' Written Reflections. *Journal of Science Education and Technology*, 30(1), 1–15. <https://doi.org/10.1007/s10956-020-09865-1>.
- Wulff, P., Mientus, L., Nowak, A. and Borowski, A. (2021) ‘Stärkung praxisorientierter hochschullehre durch computerbasierte rückmeldung zu reflexionstexten in der physikdidaktik’, *die hochschullehre* (Vol. 7, pp. 93–99). Wbv Publikation. <https://doi.org/10.3278/HSL2111W>.
- Wulff, P., Buschhüter, D., Westphal, A., Mientus, L., Nowak, A., & Borowski, A. (2022). Bridging the Gap Between Qualitative and Quantitative Assessment in Science Education Research with Machine Learning—A Case for Pretrained Language Models-Based Clustering. *Journal of Science Education and Technology*, 31, 490–513. <https://doi.org/10.1007/s10956-022-09969-w>.
- Wyss, C. (2013). *Unterricht und reflexion. Eine mehrperspektivische untersuchung der unterrichts- und reflexionskompetenz von lehrkräften*. Waxmann.
- Wyss, C. (2018). Mündliche, kollegiale Reflexion von videografiertem Unterricht. In E. Christof, J. Köhler, K. Rosenberger, & C. Wyss (Eds.), *Mündliche, schriftliche und theatrale wege der praxisreflexion: Beiträge zur professionalisierung pädagogischen handelns* (1st ed., pp. 15–49). hep Verlag.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.