frontiers | Frontiers in Education

TYPE Original Research
PUBLISHED 09 January 2023
DOI 10.3389/feduc.2022.1061461

Check for updates

Enhancing writing analytics in science education research with machine learning and natural language processing—Formative assessment of science and non-science preservice teachers' written reflections

Peter Wulff[1]*, Andrea Westphal[2], Lukas Mientus[3], Anna Nowak[3] and Andreas Borowski[3]

[1]Physics and Physics Education Research, Heidelberg University of Education, Heidelberg, Germany, [2]Department of Educational Research, University of Greifswald, Greifswald, Mecklenburg-Vorpommern, Germany, [3]Physics Education Research Group, University of Potsdam, Potsdam, Brandenburg, Germany

OPEN ACCESS

EDITED BY
Xiaoming Zhai,
University of Georgia,
United States

REVIEWED BY
Zhi Liu,
Central China Normal University, China
Lap-Kei Lee,
Hong Kong Metropolitan University, China

*CORRESPONDENCE
Peter Wulff
✉ peter.wulff@ph-heidelberg.de

SPECIALTY SECTION
This article was submitted to
STEM Education,
a section of the journal
Frontiers in Education

RECEIVED 04 October 2022
ACCEPTED 15 December 2022
PUBLISHED 09 January 2023

CITATION
Wulff P, Westphal A, Mientus L, Nowak A and Borowski A (2023) Enhancing writing analytics in science education research with machine learning and natural language processing—Formative assessment of science and non-science preservice teachers' written reflections. *Front. Educ.* 7:1061461.
doi: 10.3389/feduc.2022.1061461

COPYRIGHT
© 2023 Wulff, Westphal, Mientus, Nowak and Borowski. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

Introduction: Science educators use writing assignments to assess competencies and facilitate learning processes such as conceptual understanding or reflective thinking. Writing assignments are typically scored with holistic, summative coding rubrics. This, however, is not very responsive to the more fine-grained features of text composition and represented knowledge in texts, which might be more relevant for adaptive guidance and writing-to-learn interventions. In this study we examine potentials of machine learning (ML) in combination with natural language processing (NLP) to provide means for analytic, formative assessment of written reflections in science teacher education.

Methods: ML and NLP are used to filter higher-level reasoning sentences in physics and non-physics teachers' written reflections on a standardized teaching vignette. We particularly probe to what extent a previously trained ML model can facilitate the filtering, and to what extent further fine-tuning of the previously trained ML model can enhance performance. The filtered sentences are then clustered with ML and NLP to identify themes and represented knowledge in the teachers' written reflections.

Results: Results indicate that ML and NLP can be used to filter higher-level reasoning elements in physics and non-physics preservice teachers' written reflections. Furthermore, the applied clustering approach yields specific topics in the written reflections that indicate quality differences in physics and non-physics preservice teachers' texts.

Discussion: Overall, we argue that ML and NLP can enhance writing analytics in science education. For example, previously trained ML models can be utilized in further research to filter higher-level reasoning sentences, and thus provide science education researchers efficient mean to answer derived research questions.

KEYWORDS

machine learning, natural language processing, teachers' reflections, assessment, deep learning

Introduction

Science education researchers routinely use writing assignments either for assessment or facilitating learning (Prain and Hand, 1996; Chen et al., 2013; Docktor et al., 2016). Empirical evidence supports that writing assignments, in general, can be utilized to enhance learning processes and evaluate cognitive processes (Bangert-Drowns et al., 2004; Graham and Perin, 2007; Kellogg, 2008). Writing assignments are most effective when they are coupled with meta-cognitive guidance on how to write (Bangert-Drowns et al., 2004). As such, they have been implemented in science teacher education to enhance, among others, teachers' reflection processes (Hume, 2009), facilitate conceptual understanding about force and motion (Chen et al., 2013), or developing students' critical thinking skills (Stephenson and Sadler-McKnight, 2016). Written language artifacts such as essays can thus provide evidence on learners' knowledge, understanding, and learning processes. Yet, existing studies barely engage in analytic, formative assessment of the writing, but rather utilize holistic, summative assessment in the form of scores and group comparisons with reference to these scores. A finer-grained, analytic and formative assessment of written language products can potentially provide more detailed evidence on represented knowledge in learners' texts. Advances in computer methods, namely in the field of artificial intelligence research such as machine learning (ML) and natural language processing (NLP), have been considered promising means to extend analytic, formative assessment of language artifacts (Burstein, 2009; Buckingham Shum et al., 2017). While many studies in science education examined potentials of ML and NLP to score responses based on human annotated datasets (Ha et al., 2011; Zhai et al., 2020), it is less clear in what ways ML and NLP can be used in more explorative ways in science teacher education to enable formative assessment.

This study therefore seeks to utilize ML and NLP as analytic, formative assessment tools for science and non-science preservice teachers' written reflections. ML and NLP methods are used to filter higher-level reasoning segments in the physics and non-physics preservice teachers' written reflections and identify differences in represented knowledge in the texts between both teacher groups. To automatically filter higher-level reasoning segments, a formerly trained ML model was used to classify segments in teachers' responses and extract segments on higher-level reflection-related reasoning. Then, a ML-based clustering approach is used to cluster these segments with the goal to extract expert-novice differences in the texts.

Assessing written language artifacts in science education research

Language is intricately linked to scientific literacy, science learning, and science teaching, because language provides a generic medium for interpreting experiences and communicating ideas (Norris and Phillips, 2003; Carlsen, 2007; Halliday and Matthiessen, 2007). In particular, personal writing was found to be effective for learning (Smyth, 1998; Bangert-Drowns et al., 2004). Consequently, science education researchers used writing assignments to facilitate development of conceptual understanding, critical thinking, and reflective thinking, among others, and for assessment. Facilitating conceptual understanding and writing quality was accomplished with the Science Writing Heuristic (SWH). Cronje et al. (2013) used SWH to support their biology undergraduates' writing. They provided the students with specific prompting questions that they could use to write up their lab report. Students who received the SWH instruction scored significantly higher on their writing assignments.

Personal writing is also used in science teacher education to facilitate professional development. A commonly used method is reflective journal writing, where science preservice teachers are meant to connect their professional knowledge to interpret teaching experiences (Bain et al., 1999; Hume, 2009). Reflecting on teaching experiences was singled out as a feature for effective teacher education programs (Darling-Hammond, 2012). Reflection can be considered as a deliberate and structured thinking process that requires evaluation of one's own professional development with the goal to personally grow or improve one's teaching (Hatton and Smith, 1995; Von Aufschnaiter et al., 2019; Jung et al., 2022). Reflection processes in teacher education often relate to teaching experiences in classrooms (Jung et al., 2022). Hence, teachers' reflective competencies comprise the noticing and description of relevant classroom events, interpreting them, and learning from them (Korthagen and Kessels, 1999; Korthagen, 2005).

Noticing and interpreting learning-relevant classroom events is then linked with science teachers' professional knowledge and beliefs (Carlson et al., 2019; Wulff P. et al., 2022). In mathematics and science teacher education, professional knowledge, subject matter knowledge, and knowledge of student understanding/misconceptions have been singled out as particularly important for effective teaching (Park and Oliver, 2008; Sadler et al., 2013). Moreover, science experts' content knowledge tends to be well interconnected and coherent (Koponen and Pehkonen, 2010; Nousiainen and Koponen, 2012). This knowledge base, among others, allows expert science teachers to notice relevant classroom events and interpret them (Todorova et al., 2017; Chan et al., 2021). Novice science teachers, on the other hand, oftentimes lack the adequate professional knowledge to notice the substance of students' responses (Hume, 2009; Levin et al., 2009; Talanquer et al., 2015; Sorge et al., 2018).

To improve noticing and reflective competencies, novice teachers engage in teaching practices and reflect their teaching (Korthagen and Kessels, 1999; Wenner and Kittleson, 2018). They ideally receive guidance from instructors on how to move towards deeper levels of reflection (Lin et al., 1999). Guidance is typically provided through reflection-supporting models (Poldner et al., 2014). These models single out elements of reflections. Common elements include observation, interpretation, inference on causes, alternative modes of action, and consequences (Korthagen and Kessels, 1999; Poldner et al., 2014; Aeppli and Lötscher, 2016; Ullmann, 2019). Nowak et al. (2019) devised a reflection-supporting model which differentiates reflection elements that are important categories and should be addressed in a written reflection. In this model, preservice teachers are instructed to begin with outlining circumstances of the relevant teaching situation. Next, they describe the teaching situation and evaluate relevant aspects of it with help of their professional knowledge. Finally, the science teachers outline alternatives for their decisions and devise consequences for their professional growth. While circumstances and observations form the basis for reflections, evaluation, alternatives, and consequences can be considered higher-level reasoning elements (Seidel and Stürmer, 2014; Kleinknecht and Gröschner, 2016).

To interface the reflection-supporting models and reflective thinking processes, teachers are typically instructed to verbalize their observations and decision-making processes according to reflection-supporting models (Mena-Marcos et al., 2013; Poldner et al., 2014; Wenner and Kittleson, 2018). Verbalizations can be collected in many different ways such as logbooks, portfolios, reports, or diaries (Hatton and Smith, 1995; Loughran and Corrigan, 1995). Overall, verbalizations in written form were found to provide rich evidence for reflection processes (Hatton and Smith, 1995). Written reflections are mostly scored in holistic, summative form (Poldner et al., 2014). Holistic assessment are characterized by aggregate evaluations of language and ideas that oftentimes contain several conceptual components (Jescovitch et al., 2021).

Challenges in assessing reflective competencies in written reflections

Researchers repeatedly documented difficulties with assessing written reflections. Kost (2019) was forced to apply consensus coding for analyzing physics teachers' reflections, because the human interrater agreements were rather low in the context of classifying physics teachers' reflections. Also in a setting in science education, Abels (2011) ended up with a coding process that comprised six stages with many raters involved to reach agreement. Agreements between the raters in the first five circles were rather low, caused by the inferential nature of the task. Kember et al. (1999) first reached reasonable agreement for eight raters, and later acceptable agreement between four raters for level of reflection. They noticed that disagreements resulted from different interpretations from the written reflections and they suggest to only employ project-intern raters. Sparks-Langer et al. (1990) note about their coding that "[u]sing a one-level difference in codes as acceptable, the two raters' interview scores matched in 81 percent of the cases. Reliability was less satisfactory for the journal data, possibly because the questions in the original journal format did not elicit the kind of thinking we were coding" (p. 27). In the discussion, the authors suggest a multi-dimensional coding manual to cope with coding issues. In sum, accurate and reliable manual coding of reflections was difficult, because language in general is ambiguous and human raters' project-intern expertise might be necessary. Given these challenges, Leonhard and Rihm (2011) content that their content analyses (i.e., reaching human interrater agreement and developing a coding manual) were not scalable across contexts. Ullmann (2019) argued that human resources available in teacher training programs are a major bottleneck to provide preservice teachers opportunities for feedback on their reflection.

Advances in computer software and hardware increasingly enable the effective and efficient processing of language data (Goldberg, 2017). Hence, computer methods became popular to analyze written reflections, partly to address the abovementioned challenges and partly to explore novel potentials for inquiry (Buckingham Shum et al., 2017). Ullmann (2019) summarizes approaches to computer-based assessment of written reflections into the categories dictionary-based, rule-based, and ML. Ullmann (2017) showed that predefined dictionaries could be well used to detect some elements in a reflection-detection model, such as experience, with fair accuracy. Rule-based approaches typically use researcher-defined dictionaries in conjunction with rules to identify elements in reflection-supporting models (Gibson et al., 2016; Buckingham Shum et al., 2017). The accuracies in detecting categories in rule-based approaches were similar to dictionary-based approaches, and, overall, rather low (Ullmann, 2019). Both approaches also require hand-crafting of relevant features (e.g., terms in the dictionary) prior to data analysis. As a more inductive inquiry method, ML has the potential to automate these tasks and reach more accurate results.

ML- and NLP-based language assessment in science education

Assessment methods for written language artifacts such as written reflections in science education require principled methods that allow for rich hypotheses spaces, given the complexity of language (Lieberman et al., 2007; Mainzer, 2009). Commonly used quantitative statistical methods such as parametric, linear models in the stochastic data modeling paradigm (see: Breiman, 2001) are mostly incapable to model the complex relationships that are characteristic for language artifacts. Qualitative methods such as content analysis, on the other hand, require substantial human resources and, thus, do not scale well. ML and NLP have been proposed to be suitable methods to algorithmically model complex processes and phenomena (Breiman, 2001). ML refers to inductive problem solving by computers, i.e., algorithms learn from data (Marsland, 2015; Rauf, 2021).

ML and NLP have been adopted in a variety of contexts in science education research with written language artifacts. Science education researchers used ML models to assess complex constructs in constructed responses (Ha et al., 2011; Wulff et al., 2020; Zhai et al., 2022), explore patterns in large language-based datasets (Odden et al., 2021; Wulff P. et al., 2022), or develop means for automated guidance and feedback (Donnelly et al., 2015; Zhai et al., 2020). In these studies, it has been shown that ML and NLP could reliably and validly classify argumentations, explanations, or reflections of middle-school, high-school, and university students and teachers. Despite these successful applications, there remains the challenge that ML models especially in the deep-learning contexts require excessively large training datasets that are seldomly available in science education research.

The learning paradigm of transfer learning, i.e., the application of pretrained ML models in novel contexts with the goal to fine-tune the models with new data, is a promising path to mitigate data requirements, enhance generalizability of ML models, and spare resources. Successful transfer learning approaches include few-shot learning, where pretrained language models in one domain can be used as the backbone for novel tasks such as classification in another domain. This was also implemented in science education research. For example, Carpenter et al. (2020) and Wulff M. et al. (2022) could demonstrate that utilizing pretrained ML models (here: language models) that were developed by other researchers in more generic research contexts could be used to enhance performance of ML models in science education-specific tasks such as scoring reflective depth and breadth of middle-school students', preservice teachers', or in-service teachers' written responses. This opens up perspectives for science education research to develop models in one research context—say physics education—and share these models in other fields—say educational psychology (or vice versa)—in order to jointly develop models and address more encompassing research questions. However, to what extent transfer learning with ML models works well across educational research contexts has not yet been tested.

Also for assessing written reflections ML yielded most promising results (Buckingham Shum et al., 2017; Ullmann, 2019; Nehyba and Štefánik, 2022; Wulff M. et al., 2022). Wulff et al. (2020) showed that supervised ML approaches with shallow ML models facilitated automated classification of reflection elements in preservice physics teachers' written reflections with acceptable accuracy. However, the generalizability of the models was rather poor. Advances in deep-learning helped to improve generalizability. For example, the development of pretrained ML-based language models became possible. Following the distributional hypothesis in linguistics that meaning in language largely results from the context (Harris, 1954), language models are oftentimes trained with the objective to predict the context words, given a sequence of words. NLP researchers trained large language models based on self-supervised learning regimes (e.g., masked language modeling, Nehyba and Štefánik, 2022) with large written language repositories such as the Internet to detect regularities in language and use them in down-stream tasks such as sentiment classification of sentences (Jurafsky and Martin, 2014; Ruder, 2019). Once trained, language models capture regularities in language (grammar, semantics) that can be finetuned in a paradigm called transfer learning for downstream tasks such as analogical reasoning (Mikolov et al., 2013; Ruder, 2019). Finetuning, then, can refer to using data from a novel task and the pretrained language model as a backbone. The weights in the language model will then be adjusted to also excel at the other task. In particular, some language models are capable of predicting next words based on a sequence of words, a capability that resembles human linguistic competence (Devlin et al., 2018) and cognitive processes such as predictive inference (Adams et al., 2013).

Using pretrained language models as the backbone for simpler ML models (e.g., classification models) was found to improve task performance for the simpler ML models. For example, the bidirectional encoder representations for transformers (BERT) architecture is trained to predict next words in forwards and backwards direction (bidirectional). Utilizing BERT as the backbone for further NLP tasks such as classification typically improved performance (Devlin et al., 2018). Wulff M. et al. (2022) could show that utilizing BERT for reflective writing analytics in science teacher education could boost classification accuracy and generalizability. In line with these findings, Nehyba and Štefánik (2022) found that a language model outperformed shallow ML models in classifying reflection in general educational contexts and Liu et al. (2022) found that a convolutional BERT-CNN substantially improved cognitive engagement recognition. Also, Carpenter et al. (2020) showed that pretrained language models yielded the best classification performance for reflective depth of middle-school students' responses in a game-based microbiology learning environment. Pretrained language models could not only help to improve classification accuracy, but also to identify and cluster science teachers' responses in unsupervised ML approaches. Wulff P. et al. (2022) showed that unsupervised ML models in conjunction with pretrained language models such as BERT could also be used to explore themes that the physics teachers addressed in their written responses which related to classroom events and teachers' noticing. More general and more physics-specific themes could be differentiated. ML, and pretrained language models in particular, have proven to be effective and efficient methods to advance reflective writing analytics through supervised and unsupervised approaches. While ML and NLP cannot not necessarily resolve the challenges around high-inferential coding categories and ambiguity in teachers' language in reflective writing analytics in science education, they might well facilitate the implementation of reliable and valid coding for well-defined (sub-)tasks and thus enable researchers to answer derived research questions. It is also unclear to what extent these ML methods could be used to facilitate analytic, formative assessment of written reflections, e.g., to extract quality indicators to differentiate expert and novice written reflections.

This study

The goal of this study is to explore capabilities of ML and NLP to formatively assess physics and non-physics preservice teachers' written reflections. We first examined classification accuracy of a pretrained language model (BERT) to filter segments of higher-level reasoning in the written reflections according to the reflection-supporting model:

> *RQ1: To what extent can a pretrained language model, that was trained in a physics education research context with physics preservice teachers (domain expert), accurately classify non-physics preservice teachers'* (i.e., *domain novice) written reflections? To what extent can the classification accuracy be enhanced by finetuning the ML model in the novel context?*

The physics and non-physics preservice teachers potentially notice, describe, and evaluate the standardized teaching situation differently, given their differences in professional content knowledge and pedagogical content knowledge. We then examined to what extent the ML models can be used to formatively assess differences in the written reflections between physics and non-physics preservice teachers:

> *RQ2: To what extent can the pretrained language model be used for formative assessment purposes of the physics and non-physics preservice teachers' written reflections on a teaching situation depicted in a standardized video vignette?*

We further subdivide RQ2 into the following explorative sub-questions:

> *RQ2a: In what ways can a clustering approach with higher-level reasoning elements extract quality indicators for evaluating the written reflections?*
> *RQ2b: To what extent do human raters assess segments similar compared to the machine?*

Materials and methods

Samples

An important component for this study is the differentiation of physics (i.e., domain experts) and non-physics (domain novices) preservice teachers' written reflections. Hence, throughout this study two research contexts are differentiated: physics context and non-physics context. All preservice teachers in both groups were instructed to write a reflection on the basis of the reflection-supporting model either on a video vignette or on their own teaching experiences. Data was collected in multiple university courses. Courses included bachelor and master's courses such as teaching internships or regular content-based seminars (referenced as BA/MA seminar/internship in the remainder). Table 1 shows an overview and the descriptive statistics for the written reflections, disaggregated by subsample.

The group physics context comprises several phases of data collection exclusively with physics preservice teachers and few in-service physics teachers. The physics preservice teachers either reflected on their own teaching experiences in school internships or they reflected on a standardized teaching situation in a video vignette. We collected data over several years in various university courses in different universities in Germany (called University A/B/C/D in the remainder). The group physics context comprised $N = 81$ preservice teachers in 10 subsamples (see Table 1) who reflected either on a video vignette or their own physics lessons during university-based teaching internships. Ages ranged from 21.0 years on average to 29.6 years on average. Overall average age for preservice physics teachers was 24.6 years.

In the non-physics context preservice teachers in a general educational seminar were instructed to write a reflection on the video vignette only. The non-physics context (seminar "University B/SS2021/BA internship (education)" in Table 1) comprised $N = 68$ preservice teachers who reflected on a teaching situation depicted in a standardized video vignette. Preservice teachers in the non-physics context were on average 23.0 years old.

Differences between both contexts are already apparent when considering segments (i.e., sentences) per document and mean words per segment. The non-physics preservice teachers scored in the lower half of the distributions for segments per document, 7.69 and 17.7, respectively whereas the median (SD) values were 9.2 (3.3) and 18.6 (6.3). The type-token-ratio was also lowest for the non-science context sample, 0.22, against a median (SD) values of 0.40 (0.10). This means that these students used a more unspecific language (i.e., less unique words). Linguists posit that the type-token-ratio can be indicative of the acquired vocabulary by a person (Youmans, 1990). Hence, this can be seen as evidence that the non-science students had less domain-specific vocabulary. This can be expected, because they were no domain experts, and domain experts tend to know more specific vocabulary that they can use.

Video vignette

The main focus material of this study is a video vignette depicting a standardized teaching situation, which was developed to provide preservice physics teachers with suitable material to reflect upon. A video vignette presents viewers a short, problem-oriented teaching situation (Billion-Kramer et al., 2020). Preservice teachers are put in a position to judge and advocate teaching by others (Oser et al., 2010). The video vignette used in the present study depicts an introductory

TABLE 1 Descriptive overview of the various samples that were considered in this study.

Seminar (location/ semester/ seminar)	*N*	Segments	Segments/ document	Mean words per seg	(SD)	NaNs	Type-token-ratio	Mean age
University C/ WS202021/ unknown	5	63	12.6	19.7	8.1	0	0.44	-
University B/ SS2021/BA internship (education)	68	523	7.69	17.7	8.2	16	0.22	23.0
University A/ SS2021/MA Seminar (physics)	4	49	12.25	16.4	6.4	0	0.48	-
University B/ SS2020/unknown	8	52	6.5	32.9	21.9	6	0.4	29.6
University B/ SS2021/MA internship (physics)	1	13	13.0	18.6	5.5	0	0.69	-
University B/ SS2021/MA internship (physics)	7	36	5.14	15.5	7.6	0	0.57	-
University B/ WS201920/ unknown	13	148	11.38	23.8	13.6	7	0.33	22.0
University B/ WS202021/MA internship (physics)	5	46	9.2	15.8	7.9	0	0.51	-
University B/ WS202021/MA internship (physics)	5	81	16.2	15.1	8.1	1	0.43	22.0
University B/ WS202021/BA internship (physics)	3	24	8.0	19.8	10.5	1	0.55	21.0
University D/ SS2020/unknown	30	240	8.0	31.8	20.3	10	0.25	25.2
Median (SD)	5.0 (19.7)	52.0 (150.0)	9.2 (3.3)	18.6 (6.3)	8.1 (5.6)	1.0 (5.4)	0.4 (0.1)	22.5 (3.2)

WS, Winter semester (October to March); SS, summer semester (April to September).

physics lesson covering free fall and factors influencing the free falling movement (see: Wulff P. et al., 2022). The vignette depicts a teacher (an intern in a preparatory teacher course) who starts the lesson with several small experiments. He intends to demonstrate to the students that free fall behavior is independent of mass. The students notice the relevance of air resistance for free falling movement. The teachers then demonstrates a vacuum tube experiment, where a feather and a screw (in an evacuated vacuum tube) move at the same pace. Subsequently, the teacher instructs the students to write down a definition of free fall and to devise own experiments on how to experimentally determine whether free falling movement is at constant velocity or accelerated.

The pedagogical value of the video vignette lies in the fact that it presents an authentic, complex teaching situation that implements some known challenges in physics and science teaching. For example, the teacher does not adequately control experimental variables when letting the objects fall in the shown experiments. A student notices this, however, the teacher fails to notice the substance of the students' remark and glosses over it. In general, the lesson is rather ill-structured (redundant experiments) and the teacher performs all experiments himself leaving few opportunities for cognitive engagement of the students. When asking the students to pose hypotheses, however, the teacher evaluates the hypotheses unsystematically leaving multiple hypotheses conflated with each other. At multiple times there are thoughtful comments by the students where the teacher fails to engage with the very substance of these remarks. For example, a student asks an intricate question on consequences if an object is accelerated *ad infinitum* (involving the speed of light). Another student asks why it is called free fall if a parachute jumper jumps off an airplane. In both cases, the teacher does not pick up on the opportunity to relate these questions to the lesson's contents.

We emphasize that these challenges are rather germane to novice teachers' physics teaching, rather than attribute to failures by the individual teacher. For example, cognitive overload in early teaching raises the burden for teachers to adequately respond to the substance in students' responses *in situ* (Levin et al., 2009). Moreover, control of variables is an intricate concept that is even more difficult to implement in practice—especially with short experiments that are meant to demonstrate phenomena rather than experiments where the entire experimental cycle is implemented. Hence, the video vignette presents novice teachers with opportunities to notice relevant challenges in physics teaching and propose alternatives.

Analyses

To assess in what ways ML and NLP can be utilized for analytic, formative assessment of written reflections on the video vignette we employ ML models that were trained in prior research studies and finetune them according to our specific goals (see Figure 1).

Filtering higher-level reasoning segments in physics and non-physics preservice teachers' written reflections (RQ1)

To filter higher-level reasoning, we reuse a ML model that was trained in a prior study based on preservice physics teachers' written reflections (Wulff M. et al., 2022) and evaluate to what extent this ML model (**ML-base**) could be used to accurately classify segments in non-physics preservice teachers' written reflections. Classification categories comprised the elements in the reflection-supporting model: Circumstances, Description, Evaluation, Alternatives, and Consequences. Human interrater agreement for these categories is typically substantial (Wulff M. et al., 2022). Sentences in the written reflections (e.g., "The teacher picked up the vacuum tube and demonstrated the free falling movement") were annotated by three independent human raters, which were student assistants who were trained to classify segments according to the reflection-supporting model. Each of the raters coded a subset of the new data according to the elements in the reflection-supporting model. Interrater agreements, as measured by Cohen's kappa, were substantial (ranging from 0.70 to 0.77). The entire dataset is then annotated and split (randomly) into train (75%) and test (25%) dataset.

We then further finetune this ML model with data from the non-physics context (**ML-finetuned**) and examine if classification performance can be improved. Evaluating the performance of the ML models will be achieved through cross-validation where generalizability of the ML model is tested by applying it to unseen test data. In cross-validation, the ML model is trained on a training dataset and tested on a held-out test dataset that the model did not see in the training phase. Hence, the predictive

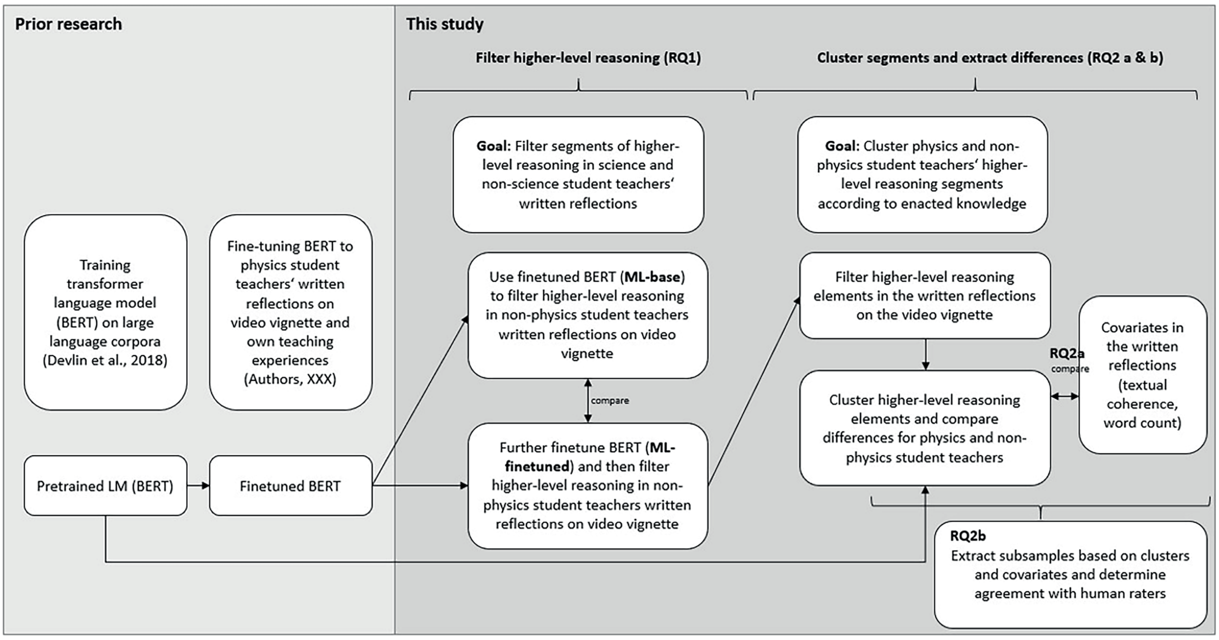
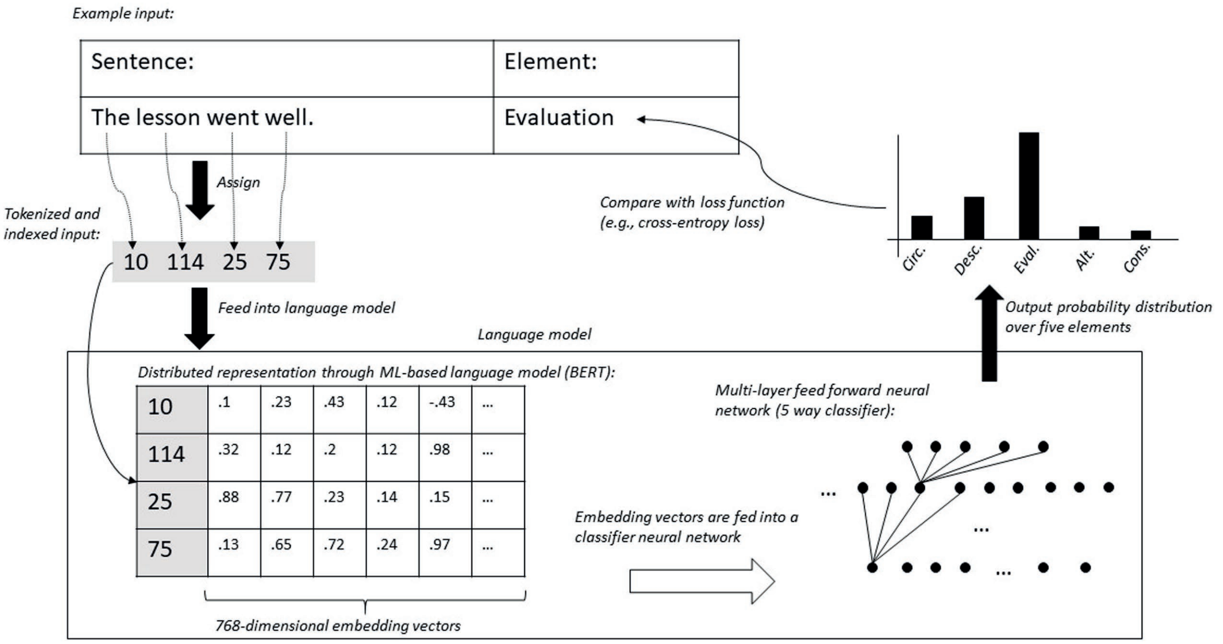

Prior research
This study
Filter higher-level reasoning (RQ1)
Cluster segments and extract differences (RQ2 a & b)
Goal: Filter segments of higher-level reasoning in science and non-science student teachers' written reflections
Goal: Cluster physics and non-physics student teachers' higher-level reasoning segments according to enacted knowledge
Training transformer language model (BERT) on large language corpora (Devlin et al., 2018)
Fine-tuning BERT to physics student teachers' written reflections on video vignette and own teaching experiences (Authors, XXX)
Use finetuned BERT (ML-base) to filter higher-level reasoning in non-physics student teachers written reflections on video vignette
compare
Further finetune BERT (ML-finetuned) and then filter higher-level reasoning in non-physics student teachers written reflections on video vignette
Filter higher-level reasoning elements in the written reflections on the video vignette
Cluster higher-level reasoning elements and compare differences for physics and non-physics student teachers
RQ2a
compare
Covariates in the written reflections (textual coherence, word count)
Pretrained LM (BERT)
Finetuned BERT
RQ2b
Extract subsamples based on clusters and covariates and determine agreement with human raters

FIGURE 1
Overview of research process to answer RQ1 and RQ2.


Example input:
Assign
Tokenized and indexed input:
10 114 25 75
Feed into language model
Language model
Distributed representation through ML-based language model (BERT):
768-dimensional embedding vectors
Embedding vectors are fed into a classifier neural network
Multi-layer feed forward neural network (5 way classifier):
Output probability distribution over five elements
Compare with loss function (e.g., cross-entropy loss)
Circ.
Desc.
Eval.
Alt.
Cons.

Sentence:	Element:
The lesson went well.	Evaluation

10	.1	.23	.43	.12	-.43	...
114	.32	.12	.2	.12	.98	...
25	.88	.77	.23	.14	.15	...
75	.13	.65	.72	.24	.97	...

FIGURE 2
Detailed representation of the classification pipeline for the ML-based probalistic language model.

capabilities of the ML model can be assessed through the accuracy with which unseen examples can be correctly classified. Typically, classification performance on test data drops because the training data in real-world applications is complex and cannot exhaustively represent the domain of interest. Less represented training examples result in performance decreases, such that training data typically more accurately captures well-represented relationships (Christian, 2021).

In our case, the input segments were sentences extracted through the spaCy library in Python (Honnibal and Montani, 2017) from the written reflections that are manually coded according to the elements of the reflection-supporting model (see Figure 2). The words were tokenized (word-piece tokenization) based on a predefined vocabulary into word pieces in order to avoid very rare words to occur in the language model (see Figure 2). This has been found to improve model performance for text translation tasks (Wu et al., 2016). Each index in the input then refers to a so-called embedding vector of predefined dimension (e.g., 300). The embedding vectors are forwarded into the ML-based language model (see Figure 1). So-called transformer-based language models became the *de-facto* standard for language modelling in various tasks (Devlin et al., 2018). Language data can be characterized by long-range dependencies, as compared to image data, where rather only short-range dependencies occur (Vaswani et al., 2017). Transformer-based language models make use of an attention mechanism that help input tokens to decide which other tokens to attend to (a feature of language is that pronouns refer to other words, for example). It has then been shown that additional neural network layers (using the transformer model as backbone) can use the transformer representations of the language input to perform tasks such as classification (Devlin et al., 2018; Ostendorff et al., 2019).

In RQ1 we will reuse BERT that had been fine-tuned in a former project with the physics context data and apply this ML model (called: **ML-base**) to the non-physics preservice teachers' data who received slightly different instruction (focus on cognitive activation) for the same video vignette. We used BERT that was open-sourced by Google research (Devlin et al., 2018) and trained for German language by deepsetAI. All language inputs to this model were tokenized on the basis of 30,000 unique tokens, the standard word piece tokenizer. The embedding dimensionality for the tokens was 768. The BERT model was used with default configuration (base version: 12 attention layers, 12 hidden layers, 200 tokens maximum sequence length[1]). We found that this model could be used to classify elements in the reflection-supporting model in written reflections with substantial human-computer-agreement (F1 score on held-out test data: 0.81; Cohen's kappa: 0.74, see: Wulff M. et al., 2022). Note that F1 score is an aggregate measure for precision and recall (find a conclusive discussion here: Jurafsky and Martin, 2014). F1 ranges from 0 to 1, with 1 being perfect performance of the classifier. In this study, we will apply **ML-base** without further finetuning to the dataset of non-physics preservice teachers. If the pretrained model could accurately classify these written reflections, this would indicate that we could apply a once developed ML model in a novel research context. Afterwards, we will further finetune **ML-base** on

[1] Please find details here: https://huggingface.co/bert-base-cased/blob/main/config.json (Accessed November 28, 2022).

the new data with the objective to accurately classify elements in the reflection-supporting model. We then compare the performance of the pretrained (**ML-base**) and further finetuned model (called: **ML-finetuned**). If the further finetuning improves the classification performance (which could be expected, see Ruder, 2019), we would use **ML-finetuned** to filter higher-level reasoning elements in the physics and non-physics preservice teachers written reflections.

Clustering higher-level reasoning elements and extracting enacted knowledge (RQ2a)

Next, we set out to cluster the higher-level reasoning elements with unsupervised clustering algorithms and pretrained embeddings with the goal to extract features such as represented knowledge that can differentiate physics and non-physics preservice teachers' writing (see Figure 1). We hypothesized that physics preservice teachers used more physics-related clusters compared to the non-physics preservice teachers (Norris and Phillips, 2003; Yore et al., 2004). Clustering of the segments was done with as follows: (a) the filtered segments (RQ1) were contextually embedded through BERT (see Figure 1), (b) the contextualized embeddings were reduced in dimensionality to computationally ease the clustering process, and (c) the reduced contextual embeddings were clustered with an unsupervised ML algorithm.

To calculate contextualized embeddings (a), the Python library *sentence transformer* was used (Reimers and Gurevych, 2019). These contextualized embeddings have 768 dimensions. From a computational perspective, it is reasonable to reduce dimensionality. In line with the exemplary use case by Grootendorst (2020), we (b) utilized uniform manifold approximation and projection (UMAP) to reduce the dimensionality of the embeddings (McInnes et al., 2018). UMAP was found to efficiently reduce high-dimensional data by keeping local structure, which is desirable in our context (Grootendorst, 2020). UMAP involves several crucial hyperparameters that control the resulting embeddings vectors. First, number of neighbours controls the scope (local versus global) of the structure which the algorithm is looking at.[2] This hyperparameter was set to 5, because this is a trade-off between looking at local and global structure and was found to be appropriate for the context of written reflections in physics (Wulff P. et al., 2022). Minimal distance controls the tightness of points. The default value of 0.1 is kept in our case. Finally, number of components controls the dimension of the target embeddings. We reduced the 768 dimensions to 10 for further processing (and ultimately to 2 for visual inspection). Finally, we (c) used a density-based clustering technique, hierarchical density-based spatial clustering of applications with noise (HDBSCAN; Campello et al., 2013; McInnes et al., 2017), to group the evaluation segments. HDBSCAN determines dense volumes in the embedding space and extract clusters based on the stability over density variation and noise datapoints (Kriegel et al., 2011; Campello et al., 2013; Wulff P. et al., 2022). An important hyperparameter[3] is minimal cluster size that determines the smallest possible value for instances in one cluster. Here, a value of 10 is chosen, given that we had approx. 1,200 segments, which would allow up to 120 clusters to be formed.

To examine if the extracted clusters can be used for formative assessment, expertise-related covariates were considered (see Figure 1). Chodorow and Burstein (2004) showed that text length (as measured through word count) in regression models could account for up to 60% of the variance in essay scores (see also: Fleckenstein et al., 2020). Rafoth and Rubin (1984, p. 447) concluded that "composition length is well established as the single most powerful of composition quality ratings." Leonhard and Rihm (2011) report a significant relationship of text length (number of symbols) with reflective depth ratings. They report high correlations (from $r = 0.26$ to $r = 0.76$, $p < 0.05$) between number of symbols and reflective breath and reflective depth (similar findings in: Carpenter et al., 2020; Krüger and Krell, 2020). Hence, domain experts likely compose longer texts as compared to domain novices. Word count will be calculated as the number of words in a written reflection.

Moreover, experts are likely more capable of engaging in elaborate processes of knowledge retrieval from long-term memory and transformation of knowledge towards rhetorical goals (integrating author and text representation), rather than more direct transmission (knowledge telling) in the form of 'think-say' or 'what-next', which is a rather linear process (Kellogg, 2008; Galbraith, 2009; Baaijen and Galbraith, 2018). The writing process is intricately involved with episodic and semantic memory, and, hence, the types of knowledge which are differently organized for domain experts compared to novices (Jong and Ferguson-Hessler, 1986). Consequently, we expected domain experts' texts to have a higher degree of coherence between the sentences (McNamara et al., 1996; Crossley et al., 2016). A coherence indicator was calculated on the basis of the contextualized segment embeddings *via* sentence transformers and **ML-base**. Similar segments will be close in distance in embedding space, hence the cosine similarity between two sentences that are semantically related will be high. We calculated within each written reflection all mutual cosine similarities between all sentences. All sentence similarities above the 0.80 quantile were considered similar to each other. Related sentences will be represented by means of a link between them in our graphical representations below.

Determine human-machine agreement for quality indicators (RQ2b)

To externally validate the quality indicators (clusters, word count, and textual coherence), a randomly selected subsample

[2] See documentation and tutorial here: https://umap-learn.readthedocs.io/en/latest/parameters.html (Accessed November 27, 2022).

[3] For examples and tutorial see: https://nbviewer.org/github/scikit-learn-contrib/hdbscan/blob/master/notebooks/How%20HDBSCAN%20Works.ipynb (Accessed November 27, 2022).

comprising high quality texts and another randomly selected subsample comprising low quality texts was compared with manual ratings. To do so, we randomly selected five sentences in each of the 20 lowest and 20 highest rated written reflections based on their text length and their number of sentences that were grouped into domain-specific clusters. Agreement between human labels ("high" versus "low") with automatically determined labels was calculated. Three independent raters (including the first author) received a spreadsheet with the 5 sampled sentences for the 40 different written reflections. All three raters had a physics background. Rater A was the first author and knew the proportions (each 50%) of lower and higher scored segments. Raters B and C were researchers (graduate students) involved in the project and knew the observed teaching situation, but not the proportions. Rater D was an independent rater (graduate student) who was not involved in the project and did not know the proportions. Interrater agreements were calculated through Cohen's kappa, which is a commonly used metric for chance-corrected agreement among any two raters. Cohen's kappa values over 0.75 indicate excellent agreement, values of 0.40 to 0.75 indicates fair to good agreement, and below 0.40 indicate poor agreement (Fleiss et al., 1981).

Results

> *RQ1: To what extent Can a pretrained language model, that Was trained In a physics education research context with physics preservice teachers (domain expert), accurately classify non-physics preservice teacher's (i.e., domain novice) written reflections? To what extent Can The classification accuracy Be enhanced By finetuning The ML model In The novel context?*

To examine to what extent pretrained ML models could be utilized to filter higher-level reasoning segments (i.e., evaluations) from non-physics preservice teachers' written reflections, the different BERT models (**ML-base** and **ML-finetuned**, see Figure 1) were used to classify segments in the written reflections according to the elements in the reflection-supporting model. To evaluate performance of **ML-base**, **ML-base** was fit on the test data without any further adjustment (i.e., finetuning) of the model weights. Table 2 shows the performance metrics for applying the validated model to the unseen novel data, namely a subset of the non-physics preservice teachers written reflections. Overall, as judged by the macro and weighted F1 score, the accuracy of the model predictions with the human ratings is acceptable, 0.52 and 0.67, respectively. Note that in the novel research context, no circumstances were coded by the human raters because outlining circumstanc es was not part of the task instruction in the video vignette. This hampers accuracy for the macro F1 score. The weighted F1 score accounts for this issue. As we commonly see with ML models, and inductive learning more generally, classification performance is correlated with the support (i.e., samples for testing from each category). As such, descriptions can be classified best, whereas alternatives worst. The Cohen's kappa values were 0.53 and 0.59 for the accuracy with and without circumstances, respectively. **ML-base** performed notably better in the original research context. In the original context, **ML-base** performed at 0.81 for weighted F1 (Wulff M. et al., 2022), compared to 0.67 in this context. We now further trained the **ML-base** with the non-physics preservice teachers' written reflections, i.e., further finetune the model, to eventually improve performance.

TABLE 2 Performance of ML-base on test dataset from non-physics preservice teachers.

	Precision	Recall	f1-score	support
Alternatives	0.59	0.59	0.59	34
Description	0.80	0.71	0.75	99
Evaluation	0.57	0.82	0.67	71
Consequences	0.97	0.43	0.60	76
Circumstances	0.00	0.00	0.00	0
Micro avg	0.65	0.65	0.65	280
Macro avg	0.59	0.51	0.52	280
Weighted avg	0.76	0.65	0.67	280
Samples avg	0.65	0.65	0.65	280

TABLE 3 Performance of ML-finetuned on test dataset from non-physics preservice teachers.

	Precision	Recall	f1-score	Support
Alternatives	0.72	0.62	0.67	34
Description	0.77	0.75	0.76	99
Evaluation	0.63	0.75	0.68	71
Consequences	0.85	0.74	0.79	76
Circumstances	0.00	0.00	0.00	0
Micro avg	0.73	0.73	0.73	280
Macro avg	0.59	0.57	0.58	280
Weighted avg	0.75	0.73	0.74	280
Samples avg	0.73	0.73	0.73	280

In order to fine-tune **ML-base**, we used the training data. We trained the model for 10 epochs with a batch size of 4. As previously done, we used the Adam optimizer with a learning rate of 5e-7. Table 3 show the classification performance for **ML-finetuned**. A noticeable improvement in classification performance was found. The macro and weighted F1 scores improved to 0.58 and 0.74. We still find the correlation with support and classification accuracy. Some categories, such as description, consequences, and evaluation could be labeled with good accuracy. The Cohen's kappa values for the overall classification improved to 0.63 and 0.64 with and without circumstances, respectively. These values indicate substantial human-computer agreement (Landis and Koch, 1977).

> *RQ2: To what extent Can The pretrained language model Be used for formative assessment purposes of The physics and non-physics teachers' written reflections On a teaching situation depicted In a standardized video vignette?*
> *RQ2a: In what ways can a clustering approach with higher-level reasoning elements extract quality indicators for evaluating the written reflections?*

The ML model **ML-finetuned** could now be used to filter higher-level reasoning elements (here: evaluations) in physics and non-physics preservice teachers written reflections on the video vignette. Based on the extracted evaluations, we then applied a clustering approach to extract facets of enacted knowledge in the evaluations. By investigating the most representative words from each topic, the following discernable topics were identified (see Table 4). To facilitate interpretation, Figure 3 displays a two-dimensional representation of the extracted topics. Each point in this two-dimensional space represents a single sentence from a preservice teacher. The color-coding differentiates the topics. The lines represent two sentences that are semantically similar, i.e., have a high cosine similarity in a students' written reflection. The grey dots refer to noisy sentences, e.g., sentences that are too general or include multiple topics. The left side of Figure 3 refers to written reflections with a length below median, whereas the right side refers to written reflections with a length above median.

Qualitative and quantitative differences in topic distribution and textual coherence can be seen. In particular, the longer texts include more physics-specific topics and have a higher score (rho) for textual coherence. Rho refers to the density of connections that a student made in a written reflection (i.e., "number of connections by student X"/"number of sentences by student X"). Noticeably, the shorter texts more often refer to topics 11 and 4. Both topics refer to rather generic observations such as observably active participation especially by male students. The longer texts more often include topics 15, 16, and 17. These topics relate to physics-specific events such as the students' question on whether parachute jump is free fall, the vacuum tube experiment, and a question related to the speed of light. Noticing and reasoning about these topics arguably requires more physics knowledge and would be more characteristic for expert-like written reflections. Quantitative differences in topic proportions between the groups were calculated with Mann–Whitney-U rank sum tests. Mann–Whitney-U rank sum test is oftentimes used in language analytics, because words and sentences are not normally distributed (Kelih and Grzybek, 2005). Given the Bonferroni correction for multiple tests, p values smaller than 0,003 (i.e., 0,05/19) can be considered significant. We found that the longer reflections included significantly more physics-specific topics (see Table 5).

TABLE 4 Five most representative words for each extracted topic.

Topic	Most representative words	English translation
−1	sus, lehrer, lehrkraft, schüler, experiment	students, teacher, student, experiment
0	freundlich, art, herrscht, sicher, lehrkraft	friendly, type, ruling, safe, teacher
1	lehrers, positiv, handlungen, empfinde, aufgefallen	teacher, positive, actions, feel, noticed
2	denke, kompetenzen, guten, notwendigen, mitbringe	think, skills, good, necessary, bring
3	negativ, faktoren, beteiligung, kleinere, selben	negative, factors, participation, smaller, same
4	aufgefallen, negativ, klasse, sus, unruhe	noticed, negative, class, students, tumult
5	sus, untereinander, melden, finde, klar	students, among each other, raise hand, find, clear
6	finde, positiv, lehrkraft, ausreden, lässt	to find, positive, teacher, let talk
7	bewerten, antworten, lässt, lehrkraft, fragen	assess, answer, let, teacher, ask
8	sus, scheinen, meist, sicherlich, weiterhin	students, seem, mostly, certainly, continue
9	bezüglich, förderlich, gesamten, späteren, gespräch	regarding, beneficial, entire, later, conversation
10	gelungen, experiment, experimente, hypothesen, durchgeführt	successful, experiment, experiments, hypotheses, executed
11	lehrer, unterricht, schüler, gelingt, kognitiv	teacher, teaching, students, succeeds, cognitive
12	grundlage, zudem, art, ausdrücken, vertrauen	basis, moreover, kind, to express, to trust
13	unterricht, aktiv, mitgearbeitet, führung, beteiligt	teaching, active, participating, leadership, involved
14	klasse, vorne, interessiert, zeigt, interesse	class, front, interested, shows, interest
15	frage, fall, fallschirm, boden, lehrer	question, fall, parachute, ground, teacher
16	luftwiderstand, fall, lichtgeschwindigkeit, einstieg, thema	air resistance, fall, speed of light, introduction, topic
17	vakuumröhre, vakuum, röhre, versuch, luft	vacuum tube, vacuum, tube, experiment, air
18	eher, sus, erkennen, experimente, ergebnisse	rather, students, recognize, experiments, results

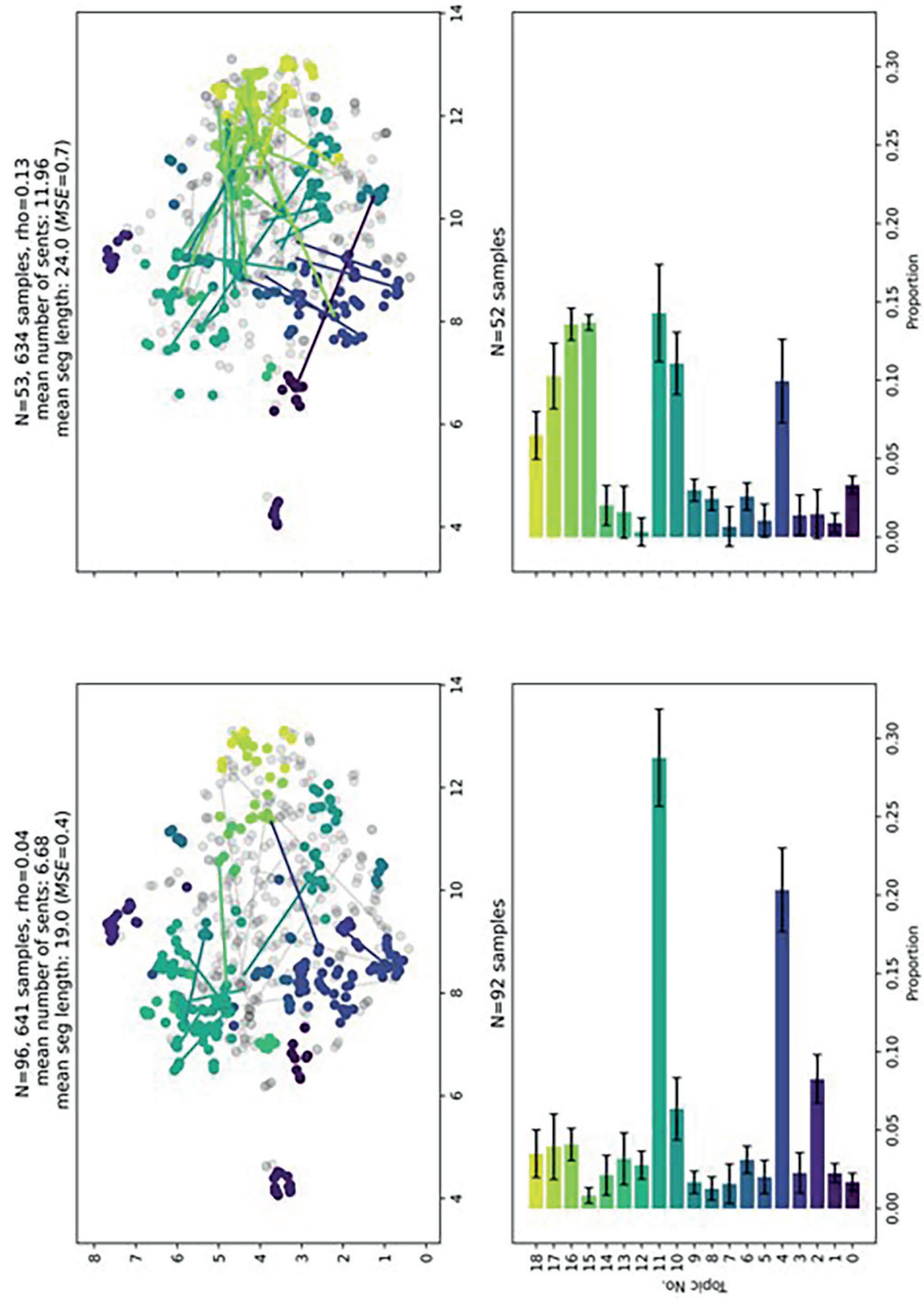


FIGURE 3
Representation space of the clusters and highlighted relationships between sentences (top row). Barplot for average proportions of topics in the evaluations for pre-service teachers with low (left) and high (right) text lengths. Color coding refers to the different topics.

As expected, we find that from the non-physics preservice teachers, 85% were in the lower scoring group (split by document length only). We also extended the splitting criterion and used document length in conjunctions with sum of addressed topics 15, 16, and 17 in a document, because these three topics were specifically related to physics-specific contents and would most likely be indicative of domain expertise. This aggregate score[4] was then split into a 5-point scale. The non-physics preservice teachers scored on average

[4] Aggregate score=(Document length (on log scale))+(Sum of topics 15, 16, and 17). These values were then z-standardized and grouped into five equally spaced quantiles.

TABLE 5 Mann-Whitney-U rank-sum tests for the different topics and groups (low and high text lengths).

Topic	*U*	*p*
0	1,943	0.001
1	2,376	0.457
2	2,023	0.015
3	2,347	0.355
4	2,365	0.454
5	2,340	0.305
6	2,302	0.248
7	2,373	0.427
8	2015	0.002
9	2,328	0.265
10	1,843	0.002
11	2,262	0.288
12	2,296	0.172
13	2,343	0.357
14	2,266	0.107
15	1,834	<0.001
16	1,779	<0.001
17	1,733	<0.001
18	1,802	<0.001

2.13 (*SD* = 1.15), whereas the physics preservice teachers scored on average 3.73 (*SD* = 1.20).

RQ2b: To what extent do human raters assess segments similar compared to the machine?

It would now be possible to use document length and the physics-specific topics as quality indicators for automated, formative assessment purposes. To evaluate to what extent human raters would similarly differentiate written reflections based on document length and physics-specific contents (i.e., topics 15, 16, and 17), three independent raters who were not familiar with the analyses (except for rater A) scored randomly sampled written reflections into either of two categories, low and high quality. To avoid that human raters use document length as a proxy criterion for their quality rating, we rather randomly sampled five sentences from each reflection. Overall, 40 written reflections were scored (20 lower and 20 higher quality). Table 6 shows to what extent the human ratings agreed with the results of the ML model ratings. Rater A had the highest agreement with the ML-based ratings (0.65). This can be expected given that rater A knew the relevant criteria (document lengths, and physics topics) that were used to score the texts. Agreements for raters B (0.49), C (0.26), and D (0.29) dropped noticeably. Rater B's agreement with the ML-based score was, however, higher compared to raters C and D. This might be attributed to the higher familiarity of rater B with the context of written reflections and the standardized teaching situation. Note also that in any case the Cohen's kappa values increased if only ratings were considered that were judged as certain by the raters (see second value in Table 6). This might be result from the fact that it is sometimes difficult, even impossible, to judge quality based on only five sampled sentences.

TABLE 6 Computer-human agreement on coding the computer-scored written reflections.

	Computer	Rater A	Rater B	Rater C	Rater D
Computer	-	0.65; 0.66; 24 samples	0.49; 0.64; 16 samples	0.26; 0.03; 13 samples	0.29; 0.47; 15 samples
Rater A		-	0.35; 0.55; 25 samples	0.25; 0.46; 19 samples	0.27; 0.51; 28 samples
Rater B			-	0.24; 0.38; 11 samples	0.37; 0.51; 21 samples
Rater C				-	0.25; 1.00; 12 samples
Rater D					-

First value is the raw Cohen's kappa. Second value is Cohen's kappa for ratings that were judged as certain. Third value is the number of samples that were judged as certain.

Summary

Writing assignments such as reflective writing in science teacher education are widely used methods to enhance science learning and assessment of competencies. While typically rather holistic, summative assessment is used to score writing assignments, ML and NLP methods have been argued to facilitate analytical, formative assessment. Analytical, formative assessment would be desirable given that it can be used to provide feedback on how to improve task performance, rather than text quality. In this study we explored potentials and challenges of utilizing ML and NLP to advance formative assessment in science teacher education for reflective writing.

In RQ1 we used ML models to filter higher-level reasoning segments in physics and non-physics preservice teachers' written reflections on a video vignette. We found that a previously trained ML model (**ML-base**) that was reused in the present study yielded acceptable performance to filter higher-level reasoning segments. This performance could be noticeably improved by further finetuning the ML model with training data from the non-physics preservice teachers to reach substantial human-machine agreement (**ML-finetuned**). Hence, **ML-finetuned** can be readily used to filter segments from the physics and non-physics preservice teachers. Finetuning ML models has been widely employed in the context of deep learning research (Brazdil et al., 2022) and even teacher education (Nehyba and Štefánik, 2022; Wulff M. et al., 2022). ML researchers showed that ML-based language models have the capacity to transfer to novel situations, also in reflective writing analytics (Nehyba and Štefánik, 2022). Our findings in RQ1 are in line with these results. We could show that reusing an ML model with a different student population (non-physics preservice teachers) was possible and that further finetuning the ML model with data from the novel context could improve classification performance. The finetuned ML model could be used for formative assessment, e.g., automated, instantaneous identification of elements in the reflection-supporting model in preservice teachers written reflections. It can also be used as a tool for science education researchers to answer derived research questions, by implementing reliable coding for a subtask.

In RQ2 we employed the finetuned ML model to filter higher-level reasoning elements and cluster them to identify quality indicators in the preservice teachers' written reflections. In RQ2a it was examined to what extent clustering of the higher-level reasoning elements yielded interpretable topics that correlate with other quality indicators such as text length. We used BERT in conjunction with UMAP and HDBSCAN to cluster the segments. Furthermore, the written reflections were median split with regards to text length (word count). The extracted topics could be distinguished to relate to more general and more physics-specific contents in the video vignette. Furthermore, the longer written reflections included more physics-specific topics compared to shorter written reflections. Moreover, the groups of physics and non-physics preservice teachers were distributed unequally across the longer and shorter written reflections. The physics preservice teachers wrote more expert-like, e.g., they included more physics-specific topics, wrote on average longer and more coherent reflections. Findings in writing analytics and noticing research buttress these findings. Expert teachers notice more learning-relevant events when observing, given that they have a more elaborate professional knowledge base for interpretation (Chan et al., 2021). Experts' writing is also more coherent, given, among others, their elaborate knowledge base (Kellogg, 2008). Our findings mirror these findings for the particular context of reflecting on a physics teaching situation in a video vignette. This clustering approach alongside the coherence metric can be well used as formative assessment tools. Formative assessment could be related to the specific topics that the preservice teachers include in their evaluations of a lesson and which they missed out on other topics.

We finally examined in RQ2b to what extent human raters were also able to distinguish written reflections in the same way the machine did (based on text length and physics topics). All human-machine agreements values (Cohen's kappa) were positive. Hence, document length and addressed physics topics relate to some extent to human judged text quality. However, degree of human-machine agreement ranged from fair to poor agreement, depending on the familiarity of the human rater with the research context. Familiarity with the written reflections and the standardized video vignette seemingly helped to raise human-machine agreement. Moreover, even the human raters did not agree with each other on the text quality. Hence, the text excerpts are probably too short to provide all the necessary information to determine text quality. An extended validation procedure would be needed to determine to what extent simple criteria such as document length and addressed physics topics alone could be used to automatically score preservice teachers written reflections.

Limitations

Our study has several limitations that relate to (1) the experimental control and variations of this study, (2) explainability of ML model decisions and implicit bias, and (3) implications resulting from the observed group differences. (1) Experimental setup is crucial in studies on ML. Even though there is theoretical progress to understand ML algorithms, most algorithms are too complex to be formally analyzed (Langley, 1988; Engel and van Broeck, 2001). Hence, the empirical component in ML studies is important, and independent variables should be systematically controlled and evaluated with regards to a dependent variable (typically performance). Different ML algorithms are oftentimes independently varied in order to find most promising performance for a specific ML algorithm. In our study, however, we rather constrained analyses to one ML algorithm, i.e., the transformer-based language model BERT with a classification ML algorithm on top. Recent progress in transformer architectures (e.g., Robustly Optimized BERT Pre-training Approach: RoBERTa) makes it likely that there may be even more performant alternative ML algorithms that could be used in future research. Given that our main goal in RQ1 was the filtering with a previously finetuned BERT model and answering derived research questions, we did not consider alternative implementations in this context. To determine the contextualized embeddings in RQ2a, we suggest that future research either try to calculate the embeddings based on the finetuned BERT model, or try different contextualized embeddings such as GloVe or ELMo (Carpenter et al., 2020). Moreover, it should also be evaluated to what extent accuracy for the further finetuned ML model in the original research context is changed. ML researchers documented a phenomenon called catastrophic forgetting, where ML models eventually decrease performance in old tasks, once a new task is acquired (McCloskey and Cohen, 1989). However, language models such as BERT seem to be more robust in these regards (Devlin et al., 2018).

Finding best performing ML algorithms for specific problems also involves systematically evaluating performance for different hyperparameter configurations. Our chosen experimental setup included multiple hyperparameters that relate to the contextualized embeddings through sentence transformers and BERT, the dimensionality reduction through UMAP, and the clustering through HDBSCAN. Rather than systematically varying all hyperparameters we heuristically chose values based on prior studies with similar research goals (Grootendorst, 2020; Wulff P. et al., 2022) and the tutorials referenced in the footnotes above. We therefore cannot exclude the possibility that there exist hyperparameter configurations which yield more interpretable topics. However, choosing the hyperparameters heuristically is advantageous from a sustainability perspective, given the compute resources that can be necessary for training and finetuning especially deep learning language models (Strubell et al., 2019).

(2) Even though the ML models reached substantial agreement with human raters and well interpretative and diagnostically valuable topics, we did not investigate on what grounds the classifications and clusters were reached. Hence, the ML models remain black-boxes in our context. Pretrained language models are commonly trained on language sources such as Wikipedia and the Internet where all sorts of gender and racial biases in language are present (Bhardwaj et al., 2020). Similar implicit biases were also found to be present in the pretrained language models (Caliskan et al., 2017). Our models would have to be examined with regards to gender biases or similar biases. However, necessary covariates need to be collected which was not part of this study.

(3) Even though we found significant group differences between physics and non-physics preservice teachers' written reflections, we stress that these findings do not reflect the competencies of the students in the respective groups. We merely used the different populations to showcase potentials of the employed ML and NLP methods to enable formative assessment. Both groups of students differed in relevant covariates (age, instruction, subjects) that have not been controlled for.

Enhancing writing analytics in science education research

An important part of research in science education engages with the development of reliable and valid assessment instruments that are ideally shared across research contexts as measuring instruments. This typically involves the development of coding rubrics. Once the rubric is meant to be used in a different context, human raters have to be trained, and coding performance becomes a function of expertise levels and other circumstances. In this study we explored a way in which ML algorithms acquire the capacity to perform the coding, and thus can function as an interface to connect different research contexts. Once trained ML models can be shared across contexts, reimplemented, and further finetuned by which they improve their performance. These capabilities might enhance science education research processes, where ML models are trained in one context, and further finetuned and improved in different contexts. This paradigm is called transfer learning or meta learning in ML research. Different contexts can relate to sample characteristics (e.g., expertise level, language), or task characteristics (e.g., scientific practice). A rather general template how this research with ML models can be done in science education is presented in the form of a flow chart in Figure 4. (Science) education researchers used pretrained language models to enhance classification performance (Carpenter et al., 2020; Liu et al., 2022; Wulff M. et al., 2022) or to cluster responses (Wulff P. et al., 2022). This study followed up on this research and extended previously used pretrained ML models to answer derived research questions and cluster them. Follow up research should evaluate to what extent transfer across tasks is also possible with these pretrained language models. The versatility of language models to form the backbone for different language-related tasks and the importance of writing assignments in science education motivate this path to be further explored.

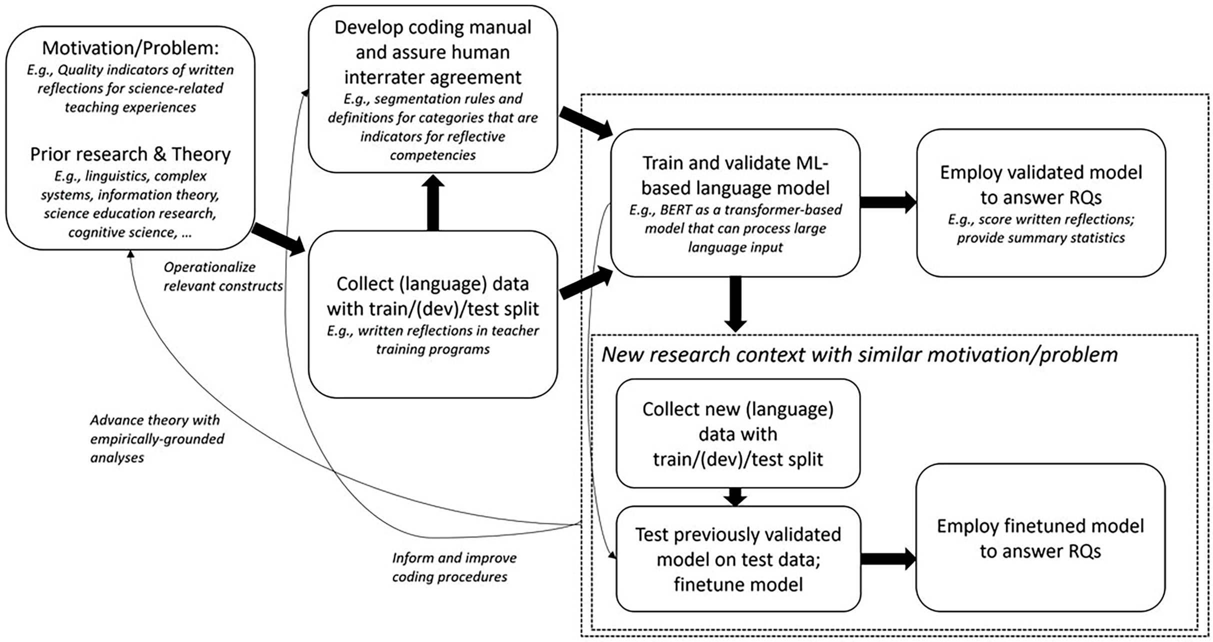

Motivation/Problem:
E.g., Quality indicators of written reflections for science-related teaching experiences
Prior research & Theory
E.g., linguistics, complex systems, information theory, science education research, cognitive science, …
Operationalize relevant constructs
Develop coding manual and assure human interrater agreement
E.g., segmentation rules and definitions for categories that are indicators for reflective competencies
Collect (language) data with train/(dev)/test split
E.g., written reflections in teacher training programs
Train and validate ML-based language model
E.g., BERT as a transformer-based model that can process large language input
Employ validated model to answer RQs
E.g., score written reflections; provide summary statistics
New research context with similar motivation/problem
Collect new (language) data with train/(dev)/test split
Test previously validated model on test data; finetune model
Employ finetuned model to answer RQs
Advance theory with empirically-grounded analyses
Inform and improve coding procedures

FIGURE 4
Research pipeline to employ and transfer ML-based language model.

Data availability statement

The datasets presented in this article are not readily available because the raw data cannot be shared. However, model sharing can be requested by contacting the corresponding author. Requests to access the datasets should be directed to peter.wulff@ph-heidelberg.de.

Ethics statement

Ethical review and approval was not required for the study on human participants in accordance with the local legislation and institutional requirements. The patients/participants provided their written informed consent to participate in this study.

Author contributions

PW: manuscript writing and data analysis. AW, LM, AN, and AB: manuscript revision and data collection. All authors contributed to the article and approved the submitted version.

Funding

The article processing charge was funded by the Baden-Württemberg Ministry of Science, Research and the Arts and the Heidelberg University of Education in the funding programme Open Access Publishing. The research was partly supported by the German Federal Ministry of Science and Education (BMBF), grant number: 01JA1816. This project is part of the "Qualitätsoffensive Lehrerbildung," a joint initiative of the Federal Government and the Länder which aims to improve the quality of teacher training. The program is funded by the Federal Ministry of Education and Research. The authors are responsible for the content of this publication.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Abels, S. (2011). *LehrerInnen als 'Reflective Practitioner': Reflexionskompetenz für einen demokratieförderlichen Naturwissenschaftsunterricht [Teachers as reflective practitioners]* (*1st*). Wiesbaden: VS Verl. für Sozialwiss.

Adams, R. A., Shipp, S., and Friston, K. J. (2013). Predictions not commands: active inference in the motor system. *Brain Struct. Funct.* 218, 611–643. doi: 10.1007/s00429-012-0475-5

Aeppli, J., and Lötscher, H. (2016). EDAMA - Ein Rahmenmodell für Reflexion. *Beiträge Zur Lehrerinnen- Und Lehrerbildung* 34, 78–97. doi: 10.25656/01:13921

Baaijen, V. M., and Galbraith, D. (2018). Discovery through writing: relationships with writing processes and text quality. *Cogn. Instr.* 36, 199–223. doi: 10.1080/07370008.2018.1456431

Bain, J. D., Ballantyne, R., Packer, J., and Mills, C. (1999). Using journal writing to enhance student teachers' reflectivity during field experience placements. *Teach. Teach.* 5, 51–73. doi: 10.1080/1354060990050104

Bangert-Drowns, R. L., Hurley, M. M., and Wilkinson, B. (2004). The effects of school-based writing-to-learn interventions on academic achievement: a meta-analysis. *Rev. Educ. Res.* 74, 29–58. doi: 10.3102/00346543074001029

Bhardwaj, R., Majumder, N., and Poria, S. (2020). Investigating gender bias in BERT. ArXiv [Preprint].

Billion-Kramer, T., Lohse-Bossenz, H., Dörfler, T., and Rehm, M. (2020). Professionswissen angehender Lehrkräfte zum Konstrukt Nature of Science (NOS): Entwicklung und Validierung eines Vignettentests (EKoL-NOS). *Zeitschrift Für Didaktik Der Naturwissenschaften* 26, 53–72. doi: 10.1007/s40573-020-00112-z

Brazdil, P. B., van Rijn, J. N., Soares, C., and Vanschoren, J. (2022). *Metalearning: Applications to Automated Machine Learning and Data Mining* (*2nd*). Cham: Springer.

Breiman, L. (2001). Statistical modeling: the two cultures. *Stat. Sci.* 16, 199–231. doi: 10.1214/ss/1009213726

Buckingham Shum, S., Sándor, Á., Goldsmith, R., Bass, R., and McWilliams, M. (2017). Towards reflective writing analytics: rationale, methodology and preliminary results. *J. Learn. Anal.* 4, 58–84. doi: 10.18608/jla.2017.41.5

Burstein, J. (2009). "Opportunities for natural language processing research in education" in *Springer Lecture Notes in Computer Science*. ed. A. Gebulkh (New York NY: Springer), 6–27.

Caliskan, A., Bryson, J. J., and Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. *Science* 356, 183–186. doi: 10.1126/science.aal4230

Campello, R. J., Moulavi, D., and Sander, J. (2013). "Density-based clustering based on hierarchical density estimates" in *Advances in Knowledge Discovery and Data Mining*. eds. J. Pei, V. S. Tseng, L. Cao, H. Motoda and G. Xu (Berlin, Heidelberg: Springer Berlin Heidelberg), 160–172.

Carlsen, W. S. (2007). "Language and science learning" in *Handbook of Research on Science Education*. eds. S. K. Abell and N. Lederman (Mawhah, NJ: Lawrence Erlbaum Associates Publishers)

Carlson, J., Daehler, K., Alonzo, A., Barendsen, E., Berry, A., Borowski, A., et al. (2019). "The refined consensus model of pedagogical content knowledge" in *Repositioning Pedagogical Content Knowledge in Teachers' Professional Knowledge*. eds. A. Hume, R. Cooper and A. Borowski (Singapore: Springer)

Carpenter, D., Geden, M., Rowe, J., Azevedo, R., and Lester, J. (2020). "Automated analysis of middle school students' written reflections during game-based learning" in *Artificial Intelligence in Education*. eds. I. I. Bittencourt, M. Cukurova, K. Muldner, R. Luckin and E. Millán (Cham: Springer International Publishing), 67–78.

Chan, K. K. H., Xu, L., Cooper, R., Berry, A., and van Driel, J. H. (2021). Teacher noticing in science education: do you see what I see? *Stud. Sci. Educ.* 57, 1–44. doi: 10.1080/03057267.2020.1755803

Chen, Y.-C., Hand, B. B., and McDowell, L. (2013). The effects of writing-to-learn activities on elementary students' conceptual understanding: learning about force and motion through writing to older peers. *Sci. Educ.* 97, 745–771. doi: 10.1002/sce.21067

Chodorow, M., and Burstein, J. (2004). Beyond essay length: Evaluating e-rater's performance on Toefl essays. *ETS Res. Rep. Ser.* 2004:i-38. doi: 10.1002/j.2333-8504.2004.tb01931.x

Christian, B.. (2021). *The Alignment Problem: How Can Machines Learn Human Values?* London: Atlantic Books.

Cronje, R., Murray, K., Rohlinger, S., and Wellnitz, T. (2013). Using the science writing heuristic to improve undergraduate writing in biology. *Int. J. Sci. Educ.* 35, 2718–2731. doi: 10.1080/09500693.2011.628344

Crossley, S. A., Muldner, K., and McNamara, D. S. (2016). Idea generation in student writing. *Writ. Commun.* 33, 328–354. doi: 10.1177/0741088316650178

Darling-Hammond, L. (2012). *Powerful Teacher Education: Lessons from Exemplary Programs* (*1st.*). San Francisco, CA: Jossey-Bass.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). BERT: pre-training of deep bidirectional transformers for language understanding. ArXiv, 1810.04805 [Preprint].

Docktor, J. L., Dornfeld, J., Frodermann, E., Heller, K., Hsu, L., Jackson, K. A., et al. (2016). Assessing student written problem solutions: a problem-solving rubric with application to introductory physics. Physical review. *Phys. Educ. Res.* 12:10130. doi: 10.1103/PhysRevPhysEducRes.12.010130

Donnelly, D. F., Vitale, J. M., and Linn, M. C. (2015). Automated guidance for thermodynamics essays: critiquing versus revisiting. *J. Sci. Educ. Technol.* 24, 861–874. doi: 10.1007/s10956-015-9569-1

Engel, A., and van den Broeck, C.. (2001). *Statistical Mechanics of Learning*. Cambridge: Cambridge University Press.

Fleckenstein, J., Meyer, J., Jansen, T., Keller, S., and Köller, O. (2020). Is a long essay always a good essay? The effect of text length on writing assessment. *Front. Psychol.* 11:562462. doi: 10.3389/fpsyg.2020.562462

Fleiss, J. L., Levin, B., and Paik, M. C. (1981). The measurement of interrater agreement. *Stat. Methods Rates Proportions* 2, 212–236.

Galbraith, D. (2009). Writing as discovery. *Br. J. Educ. Psychol.* 2, 5–26. doi: 10.1348/978185409X421129

Gibson, A., Kitto, K., and Bruza, P. (2016). Towards the discovery of learner metacognition from reflective writing. *J. Learn. Anal.* 3, 22–36. doi: 10.18608/jla.2016.32.3

Goldberg, Y. (2017). Neural network methods for natural language processing. in *Synthesis Lectures on Human Language Technologies*. San Rafael, CA: Morgan and Claypool.

Graham, S., and Perin, D. (2007). A meta-analysis of writing instruction for adolescent students. *J. Educ. Psychol.* 99, 445–476. doi: 10.1037/0022-0663.99.3.445

Grootendorst, M. (2020). Topic modeling with BERT. Available at: https://towardsdatascience.com/topic-modeling-with-bert-779f7db187e6 (Accessed December 21, 2022).

Ha, M., Nehm, R. H., Urban-Lurain, M., and Merrill, J. E. (2011). Applying computerized-scoring models of written biological explanations across courses and colleges: prospects and limitations. *CBE Life Sci. Educ.* 10, 379–393. doi: 10.1187/cbe.11-08-0081

Halliday, M. A. K., and Matthiessen, C. M. I. M. (2007). *An Introduction to Functional Grammar* (*3rd*). London: Hodder Education.

Harris, Z. S. (1954). Distributional structure. *Word* 10, 146–162. doi: 10.1080/00437956.1954.11659520

Hatton, N., and Smith, D. (1995). Reflection in teacher education: towards definition and implementation. *Teach. Teach. Educ.* 11, 33–49.

Honnibal, M., and Montani, I. (2017). *spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing*. Available at: https://spacy.io/

Hume, A. (2009). Promoting higher levels of reflective writing in student journals. *High. Educ. Res. Dev.* 28, 247–260. doi: 10.1080/07294360902839859

Jescovitch, L. N., Scott, E. E., Cerchiara, J. A., Merrill, J., Urban-Lurain, M., Doherty, J. H., et al. (2021). Comparison of machine learning performance using analytic and holistic coding approaches across constructed response assessments aligned to a science learning progression. *J. Sci. Educ. Technol.* 30, 150–167. doi: 10.1007/s10956-020-09858-0

Jong, T. D., and Ferguson-Hessler, M. G. (1986). Cognitive structures of good and poor novice problem solvers in physics. *J. Educ. Psychol.* 78, 279–288. doi: 10.1037/0022-0663.78.4.279

Jung, J., Lu, Y.-H., and Ding, A.-C. E. (2022). How do prompts shape preservice teachers' reflections? A case study in an online technology integration class. *J. Teach. Educ.* 73, 301–313. doi: 10.1177/00224871211056936

Jurafsky, D., and Martin, J. H. (2014). "Speech and language processing" in *Always Learning. 2nd* ed (Harlow: Pearson Education)

Kelih, E., and Grzybek, P. (2005). Satzlänge: Definitionen, Häufigkeiten, Modelle (Am Beispiel slowenischer Prosatexte) [Sentence length: definitions, frequencies, models]. *LDV-Forum* 20, 31–51. doi: 10.21248/jlcl.20.2005.74

Kellogg, R. T. (2008). Training writing skills: a cognitive developmental perspective. *J. Writing Res.* 1, 1–26. doi: 10.17239/jowr-2008.01.01.1

Kember, D., Jones, A., Loke, A., McKay, J., Sinclair, K., Tse, H., et al. (1999). Determining the level of reflective thinking from students' written journals using a coding scheme based on the work of Mezirow. *Int. J. Lifelong Educ.* 18, 18–30. doi: 10.1080/026013799293928

Kleinknecht, M., and Gröschner, A. (2016). Fostering preservice teachers' noticing with structured video feedback: results of an online- and video-based intervention study. *Teach. Teach. Educ.* 59, 45–56. doi: 10.1016/j.tate.2016.05.020

Koponen, I. T., and Pehkonen, M. (2010). Coherent knowledge structures of physics represented as concept networks in teacher education. *Sci. Educ.* 19, 259–282. doi: 10.1007/s11191-009-9200-z

Korthagen, F. A. (2005). Levels in reflection: core reflection as a means to enhance professional growth. *Teach. Teach.* 11, 47–71. doi: 10.1080/1354060042000337093

Korthagen, F. A., and Kessels, J. (1999). Linking theory and practice: changing the pedagogy of teacher education. *Educ. Res.* 28, 4–17. doi: 10.3102/0013189X028004004

Kost, D. (2019). Reflexionsprozesse von Studierenden des Physiklehramts. Dissertation. Justus-Liebig-University in Gießen.

Kriegel, H.-P., Kröger, P., Sander, J., and Zimek, A. (2011). Density-based clustering. *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.* 1, 231–240. doi: 10.1002/widm.30

Krüger, D., and Krell, M. (2020). Maschinelles Lernen mit Aussagen zur Modellkompetenz. *Zeitschrift Für Didaktik Der Naturwissenschaften* 26, 157–172. doi: 10.1007/s40573-020-00118-7

Landis, J. R., and Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics* 33, 159–174. doi: 10.2307/2529310

Langley, P. (1988). Machine learning as an experimental science. *Mach. Learn.* 3, 5–8. doi: 10.1007/BF00115008

Leonhard, T., and Rihm, T. (2011). Erhöhung der Reflexionskompetenz durch Begleitveranstaltungen zum Schulpraktikum? Konzeption und Ergebnisse eines Pilotprojekts mit Lehramtsstudierenden. *Lehrerbildung Auf Dem Prüfstand* 4, 240–270. doi: 10.25656/01:14722

Levin, D. M., Hammer, D., and Coffey, J. E. (2009). Novice Teachers' attention to student thinking. *J. Teach. Educ.* 60, 142–154. doi: 10.1177/0022487108330245

Lieberman, E., Michel, J.-B., Jackson, J., Tang, T., and Nowak, M. A. (2007). Quantifying the evolutionary dynamics of language. *Nature* 449, 713–716. doi: 10.1038/nature06137

Lin, X., Hmelo, C. E., Kinzer, C., and Secules, T. (1999). Designing technology to support reflection. *Educ. Technol. Res. Dev.* 47, 43–62. doi: 10.1007/BF02299633

Liu, S., Liu, S., Liu, Z., Peng, X., and Yang, Z. (2022). Automated detection of emotional and cognitive engagement in MOOC discussions to predict learning achievement. *Comput. Educ.* 181:104461. doi: 10.1016/j.compedu.2022.104461

Loughran, J., and Corrigan, D. (1995). Teaching portfolios: a strategy for developing learning and teaching in preservice education. *Teach. Teach. Educ.* 11, 565–577. doi: 10.1016/0742-051X(95)00012-9

Mainzer, K. (2009). Challenges of complexity in the 21st century. *Evol. Inst. Econ. Rev.* 6, 1–22. doi: 10.14441/eier.6.1

Marsland, S. (2015). "Machine learning: an algorithmic perspective" in *Chapman & Hall / CRC Machine Learning & Pattern Recognition Series. 2nd* ed (Boca Raton, FL: CRC Press)

McCloskey, M., and Cohen, N. J. (1989). "Catastrophic interference in connectionist networks: the sequential learning problem" in *Psychology of learning and motivation*. ed. G. H. Bower, vol. *24* (Cambridge, MA: Academic Press), 109–165.

McInnes, L., Healy, J., and Astels, S. (2017). Hdbscan: hierarchical density based clustering. *J. Open Source Softw.* 2:205. doi: 10.21105/joss.00205

McInnes, L., Healy, J., Saul, N., and Grossberger, L. (2018). UMAP: uniform manifold approximation and projection. *J. Open Source Softw.* 3:861. doi: 10.21105/joss.00861

McNamara, D., Kintsch, E., Butler Songer, N., and Kintsch, W. (1996). Are good texts always better? Interactions of text coherence, background knowledge, and levels of understanding in learning from text. *Cogn. Instr.* 14, 1–43. doi: 10.1207/s1532690xci1401_1

Mena-Marcos, J., García-Rodríguez, M.-L., and Tillema, H. (2013). Student teacher reflective writing: what does it reveal? *Eur. J. Teach. Educ.* 36, 147–163. doi: 10.1080/02619768.2012.713933

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013). "Distributed representations of words and phrases and their compositionality," *Advances in Neural Information Processing Systems (Bd. 26)*. eds. C. J. Burges, L. Bottou, M. Welling, Z. Ghahramani and K. Q. Weinberger (Hrsg.) (Curran Associates, Inc.).

Nehyba, J., and Štefánik, M. (2022). Applications of deep language models for reflective writings. *Educ. Inform. Technol.* doi: 10.1007/s10639-022-11254-7

Norris, S. P., and Phillips, L. M. (2003). How literacy in its fundamental sense is central to scientific literacy. *Sci. Educ.* 87, 224–240. doi: 10.1002/sce.10066

Nousiainen, M., and Koponen, I. T. (2012). Concept maps representing knowledge of physics: connecting structure and content in the context of electricity and magnetism. *Nordic Stud. Sci. Educ.* 6, 155–172. doi: 10.5617/nordina.253

Nowak, A., Kempin, M., Kulgemeyer, C., and Borowski, A. (2019). "Reflexion von Physikunterricht [Reflection of physics lessons]," in *Naturwissenschaftliche Bildung als Grundlage für berufliche und gesellschaftliche Teilhabe.* (S. 838). Jahrestagung in Kiel 2018. Ed. C. Maurer (Regensburg: Gesellschaft für Didaktik der Chemie und Physik).

Odden, T. O. B., Marin, A., and Rudolph, J. L. (2021). How has science education changed over the last 100 years? An analysis using natural language processing. *Sci. Educ.* 105, 653–680. doi: 10.1002/sce.21623

Oser, F. K., Forster-Heinzer, S., and Salzmann, P. (2010). Die Messung der Qualität von professionellen Kompetenzprofilen von Lehrpersonen mit Hilfe der Einschätzung von Filmvignetten: Chancen und Grenzen des advokatorischen Ansatzes. *Unterrichtswissenschaft* 38, 5–28.

Ostendorff, M., Bourgonje, P., Berger, M., Moreno-Schneider, J., Rehm, G., and Gipp, B. (2019). Enriching BERT with knowledge graph Embeddings for document classification. ArXiv (1909.08402v1) [Preprint].

Park, S., and Oliver, J. S. (2008). Revisiting the conceptualisation of pedagogical content knowledge (PCK): PCK as a conceptual tool to understand teachers as professionals. *Res. Sci. Educ.* 38, 261–284. doi: 10.1007/s11165-007-9049-6

Poldner, E., van der Schaaf, M., Simons, P. R.-J., van Tartwijk, J., and Wijngaards, G. (2014). Assessing student teachers' reflective writing through quantitative content analysis. *Eur. J. Teach. Educ.* 37, 348–373. doi: 10.1080/02619768.2014.892479

Prain, V., and Hand, B. B. (1996). Writing for learning in secondary science: rethinking practices. *Teach. Teach. Educ.* 12, 609–626. doi: 10.1016/S0742-051X(96)00003-0

Rafoth, B. A., and Rubin, D. L. (1984). The impact of content and mechanics on judgments of writing quality. *Writ. Commun.* 1, 446–458. doi: 10.1177/0741088384001004004

Rauf, I. A. (2021). *Physics of data science and machine learning.* Boca Raton: CRC Press.

Reimers, N., and Gurevych, I. (2019). Sentence-BERT: Sentence Embeddings using Siamese BERT-networks: Association for Computational Linguistics. in *Proceedings of the 2019 conference on empirical methods in natural language processing.* Stroudsburg, PA: Association for Computational Linguistics.

Ruder, S. (2019). Neural transfer learning for natural language processing. Dissertation. Ireland: National University of Ireland.

Sadler, P. M., Sonnert, G., Coyle, H. P., Cook-Smith, N., and Miller, J. L. (2013). The influence of teachers' knowledge on student learning in middle school physical science classrooms. *Am. Educ. Res. J.* 50, 1020–1049. doi: 10.3102/0002831213477680

Seidel, T., and Stürmer, K. (2014). Modeling and measuring the structure of professional vision in preservice teachers. *Am. Educ. Res. J.* 51, 739–771. doi: 10.3102/0002831214531321

Smyth, J. M. (1998). Written emotional expression: effect sizes, outcome types, and moderating variables. *J. Consult. Clin. Psychol.* 66, 174–184. doi: 10.1037//0022-006x.66.1.174

Sorge, S., Neumann, I., Neumann, K., Parchmann, I., and Schwanewedel, J. (2018). Was ist denn da passiert? *MNU J.* 6, 420–426.

Sparks-Langer, G. M., Simmons, J. M., Pasch, M., Colton, A., and Starko, A. (1990). Reflective pedagogical thinking: how can we promote it and measure it? *J. Teach. Educ.* 41, 23–32. doi: 10.1177/002248719004100504

Stephenson, N. S., and Sadler-McKnight, N. P. (2016). Developing critical thinking skills using the science writing heuristic in the chemistry laboratory. *Chem. Educ. Res. Pract.* 17, 72–79. doi: 10.1039/C5RP00102A

Strubell, E., Ganesh, A., and McCallum, A. (2019). Energy and policy considerations for deep learning in NLP. ArXiv [Preprint].

Talanquer, V., Bolger, M., and Tomanek, D. (2015). Exploring prospective teachers' assessment practices: noticing and interpreting student understanding in the assessment of written work. *J. Res. Sci. Teach.* 52, 585–609. doi: 10.1002/tea.21209

Todorova, M., Sunder, C., Steffensky, M., and Möller, K. (2017). Pre-service teachers' professional vision of instructional support in primary science classes: how content-specific is this skill and which learning opportunities in initial teacher education are relevant for its acquisition? *Teach. Teach. Educ.* 68, 275–288. doi: 10.1016/j.tate.2017.08.016

Ullmann, T. D. (2017). "Reflective writing analytics. Empirically Determined Keywords of Written Reflection." *ACM International Conference Proceeding Series.* LAK '17 Proceedings of the Seventh International Learning Analytics & Knowledge Conference. 163–167.

Ullmann, T. D. (2019). Automated analysis of reflection in writing: validating machine learning approaches. *Int. J. Artif. Intell. Educ.* 29, 217–257. doi: 10.1007/s40593-019-00174-2

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., et al. (2017). "Attention is All you Need," *Advances in Neural Information Processing Systems* (Bd. 30). eds. I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan et al. (Hrsg.) (Curran Associates, Inc.).

Von Aufschnaiter, C., Fraij, A., and Kost, D. (2019). Reflexion und Reflexivität in der Lehrerbildung. *Challenge Teach. Train. J. Concept. Design. Discussion* 2, 144–159. doi: 10.4119/UNIBI/HLZ-144

Wenner, J. A., and Kittleson, J. (2018). Focused video reflections in concert with practice-based structures to support elementary teacher candidates in learning to teach science. *J. Sci. Teach. Educ.* 29, 741–759. doi: 10.1080/1046560X.2018.1512362

Wu, Y., Schuster, M., Chen, Z., Le, V. Q., Norouzi, M., Macherey, W., et al. (2016). Google's neural machine translation system: bridging the gap between human and machine translation. ArXiv [Preprint].

Wulff, P., Buschhüter, D., Nowak, A., Westphal, A., Becker, L., Robalino, H., et al. (2020). Computer-based classification of preservice physics teachers' written reflections. *J. Sci. Educ. Technol.* doi: 10.1007/s10956-020-09865-1

Wulff, P., Buschhüter, D., Westphal, A., Mientus, L., Nowak, A., Borowski, A., et al. (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.* doi: 10.1007/s10956-022-09969-w

Wulff, M., Mientus, C., Nowak, M., and Borowski, K. (2022). Utilizing a Pretrained Language Model (BERT) to Classify Preservice Physics Teachers' Written Reflections. *Int. J. Artif. Intell. Educ.* doi: 10.1007/s40593-022-00290-6

Yore, L. D., Hand, B. B., Goldman, S. R., Hildebrand, G. M., Osborne, J. F., Treagust, D. F., et al. (2004). New directions in language and science education research. *Read. Res. Q.* 39, 347–352. doi: 10.1598/RRQ.39.3.8

Youmans, G. (1990). Measuring lexical style and competence: the type-token vocabulary curve. *Style* 24, 584–599.

Zhai, X., He, P., and Krajcik, J. S. (2022). Applying machine learning to automatically assess scientific models. *J. Res. Sci. Teach.* 59, 1765–1794. doi: 10.1002/tea.21773

Zhai, X., Yin, Y., Pellegrino, J. W., Haudek, K. C., and Shi, L. (2020). Applying machine learning in science assessment: a systematic review. *Stud. Sci. Educ.* 56, 111–151. doi: 10.1080/03057267.2020.1735757