




www.ijtes.net

AI in STEM Teacher Education: Inquiry into Capabilities of an Emerging Technology

Peter Wulff 
Heidelberg University of Education, Germany

Lukas Mientus 
University of Potsdam, Germany

Anna Nowak 
University of Potsdam, Germany

Andreas Borowski 
University of Potsdam, Germany

To cite this article:

Wulff, P., Mientus, L., Nowak, A., & Borowski, A. (2025). AI in STEM teacher education: Inquiry into capabilities of an emerging technology. *International Journal of Technology in Education and Science (IJTES)*, 9(4), 597-618. <https://doi.org/10.46328/ijtes.5105>

The International Journal of Technology in Education and Science (IJTES) is a peer-reviewed scholarly online journal. This article may be used for research, teaching, and private study purposes. Authors alone are responsible for the contents of their articles. The journal owns the copyright of the articles. The publisher shall not be liable for any loss, actions, claims, proceedings, demand, or costs or damages whatsoever or howsoever caused arising directly or indirectly in connection with or arising out of the use of the research material. All authors are requested to disclose any actual or potential conflict of interest including any financial, personal or other relationships with other people or organizations regarding the submitted work.



This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.

AI in STEM Teacher Education: Inquiry into Capabilities of an Emerging Technology

Peter Wulff, Lukas Mientus, Anna Nowak, Andreas Borowski

Article Info

Article History

Received:

9 November 2024

Accepted:

23 April 2025

Keywords

Artificial intelligence

Teacher education

Professional knowledge

Abstract

Important prerequisites for effective teaching in science, technology, engineering, and mathematics (STEM) are outlined in the refined consensus model of pedagogical content knowledge: Teachers need to become able to apply their pedagogical content knowledge in practice, called enacted pedagogical content knowledge. To support pre-service STEM teachers to develop enacted pedagogical content knowledge, scholars in STEM fields need to implement opportunities of lesson planning, teaching, and reflecting upon teaching, and provide assessments and guidance for professional development. This theoretical article reviews advances in artificial intelligence-based methods in reference to applications of these methods in pre-service STEM teacher education with a focus on the refined consensus model for pedagogical content knowledge. A number of selected studies provides valuable insights into opportunities and challenges of applying AI in STEM teacher education. We found that AI technologies, based on their differing degrees of sophistication, provide different affordances with regards to research and professional development. AI technologies already prevail STEM teacher education and STEM learning more generally. To enhance STEM teacher education with AI technologies, affordances of the technologies with reference to potentials for automation and feedback have to be recognized. The refined consensus model provides a valuable lens to conceptualize these affordances.

Introduction

Teachers in science, technology, engineering, and mathematics (STEM) fields need well-structured and applicable professional knowledge to design lessons and enact effective teaching (Abell, 2007; Baumert & Kunter, 2011; Magnusson et al., 1999; Sadler et al., 2013). The structure and developmental mechanisms for this professional knowledge, the pedagogical content knowledge in particular, are synthesized in the refined consensus model (RCM) of pedagogical content knowledge (PCK). Well-structured PCK and enacting lesson planning, teaching, and reflecting upon are prerequisites to develop enacted PCK that enables the pre-service teachers in STEM fields (PSTs) to professionally act later on in their profession (Gess-Newsome, 2015; Gess-Newsome et al., 2017). STEM teacher education is therefore organized in such a way that PSTs develop their professional knowledge in a university phase and are then required to engage in increasingly complex and authentic teaching enactments in school contexts (Clarke & Hollingsworth, 2002; Grossman et al., 2009; Zeichner, 2010). To allow PSTs to

continuously improve their teaching, it is important that they develop a contextualized and applicable PCK ('enacted') and apply it in practice (Alonzo et al., 2019). Furthermore, PSTs need to develop the ability to reflect upon their teaching experiences and ground their professional growth in evidence of student learning outcomes (Clarke & Hollingsworth, 2002; Korthagen, 1999; Korthagen & Kessels, 1999). However, either developing contextualized and applicable PCK, as well as the capability to reflect upon one's teaching experiences is difficult for PSTs and guidance, e.g., through feedback, is an important factor to promote professional growth (Carlson et al., 2019; Mena-Marcos et al., 2013). Therefore, it is important to provide (technology-)rich learning opportunities for PSTs to apply PCK and reflect upon it alongside with individualized feedback (Hattie & Timperley, 2007; Kluger & DeNisi, 1996; Lai & Calandra, 2010; Schön, 1983).

Facilitating PSTs to develop the required professional skills requires substantial and substantive resources from instructors and researchers (Zhai, Haudek, Shi, et al., 2020). For example, developing professional skills and knowledge is expected to be most effective when timely feedback is provided for teaching enactments, however, this is demanding for instructors, and not often done (effectively) in practice (Clarke & Hollingsworth, 2002; Hattie & Timperley, 2007; Lai & Calandra, 2010). Moreover, assessing PCK and practices such as lesson planning and reflection reports, as well as providing adaptive and individualized feedback, is conceptually and practically challenging, given that teachers produce large amounts of planning and reflection documents throughout their training, and these documents entail threads and discourses that both longitudinally and concurrently relate to each other. As such, an amount of complexity is created that easily escapes the capacity and capabilities of instructors who operate under various constraints such as limited time.

Technologies, in particular artificial Intelligence (AI)-enhanced technologies, can play an important role in providing teachers, instructors and researchers in STEM teacher education contexts evidence-centered means for assessment and individualized feedback. AI-based methods have proven to be valuable resources for educational research (Nelson et al., 2021; Salas-Pilco et al., 2022). In addition, AI-based methods have also been shown to be valuable tools to assess professional competencies (M. Liu et al., 2019; Ullmann, 2019; Zhai, Yin, et al., 2020). Recently advanced generative AI (GenAI) tools even extended the capabilities for assessment and feedback (Kasneci et al., 2023). This theoretical study seeks to answer the following research question in the context of the RCM as our guiding framework: In what domains and in what ways have (Gen)AI methods be utilized in teacher education in STEM fields, and what potentials and challenges can be derived from these applications.

Professional Competencies in Science Teacher Education

The teaching profession is characterized by the necessity for teachers to act and decide under uncertainty (Aeppli & Löttscher, 2016). To ground their decisions, teachers rely on their attitudes, beliefs, and professional knowledge, which is commonly differentiated into content knowledge, pedagogical knowledge, as well as PCK as the "amalgam" between domain-specific knowledge and pedagogical insights (Shulman, 1986), or an "amalgam of instructional strategies, content representations and content knowledge" (Kind & Chan, 2019, p. 968). Arguably, the structure and development of PCK can be seen as a particular provenance of discipline-based educational research in STEM fields as outlined in the RCM (Carlson et al., 2019). PCK, it has been shown, develops both in

quantity and structure during PST education (Baumert et al., 2010; Krauss et al., 2008; Sorge, Kröger, et al., 2019; Sorge, Stender, & Neumann, 2019). For primary PSTs it has been shown that PCK is linked to professional perception of situations (Meschede et al., 2017). (Enacted) PCK eventually develops through continuous lesson planning, teaching, and reflection (Carlson et al., 2019; Mientus et al., 2022). Planning, teaching, and reflection hinge on personal PCK (reflection is unproductive if it solely relies on personal experience, as argued in Williams and Grudnoff (2011)), where reflection can be defined as a structured thinking process in which individual dispositions (e.g., the PCK) are related to situation-specific affordances and circumstances with the goal to advance one's competencies and teaching (von Aufschnaiter et al., 2019).

The RCM synthesizes these findings into an overarching framework that integrates prior models for PCK in science education (Carlson et al., 2019). While the RCM was developed in the context of science education, similar conceptualizations exist for professional knowledge (integration of different knowledge facets, collective versus personal knowledge, ...), and the importance of attitudes and beliefs also in mathematics, engineering, and technology education (Krauss et al., 2008; Love & Hughes, 2022; Mishra & Koehler, 2006; Neuweg, 2014; Tatto et al., 2008). The RCM posits three realms of PCK, collective, personal, and enacted PCK (see simplified version of the RCM depicted in Figure 1). Collective PCK can be understood as the agreed upon knowledge in the community of practicing scholars in STEM fields (e.g., students misconceptions or learning difficulties). Personal PCK is the individual knowledge that teachers in STEM fields might have and that they utilize for lesson planning and reflection. Moreover, their attitudes and beliefs significantly impact planning and reflections, as well as enacting (enacted PCK) lessons. We will use the RCM in the following as the guiding framework for effective professional development programs in STEM education. Several domains of the RCM will be differentiated: assessing the realms of knowledge (collective, personal, and enacted PCK), and enhancing professional development of teachers in STEM fields.

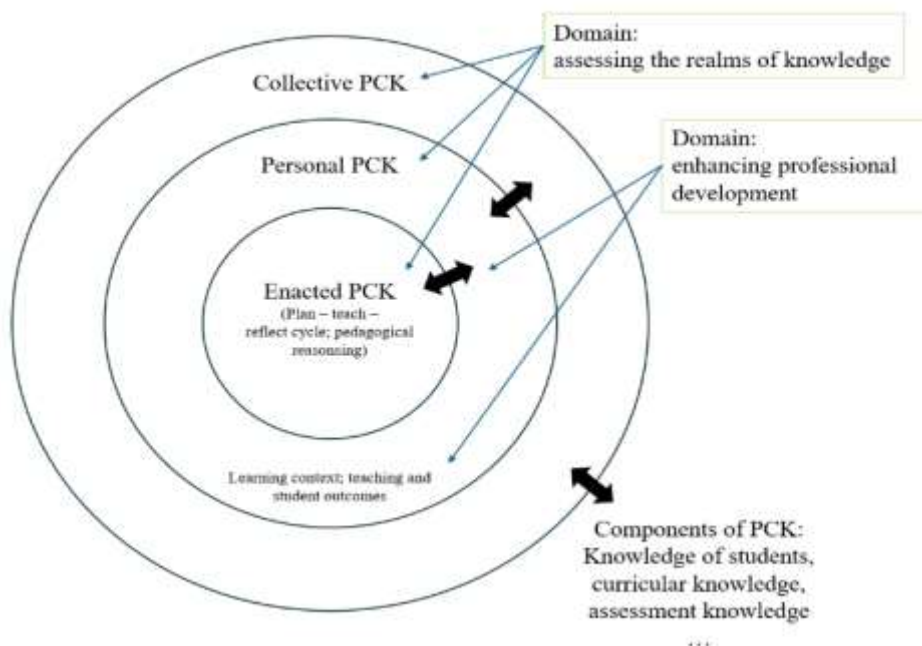


Figure 1. Simplified Version of the RCM as Derived in Carlson et al. (2019). The Identified Domains Are Depicted as Boxes

Prevailing Challenges in STEM Teacher Education

PCK (as operationalized in the RCM) structures professional development programs, and it has been empirically shown that professional development interventions and programs can foster development of PCK, attitudes, and beliefs of PSTs (Gess-Newsome et al., 2017; Sorge, Kröger, et al., 2019). However, the exact mechanisms are not well understood, given that either collective PCK and enacted PCK are multifaceted, complex constructs, which depend on, among others, context, topic, and grade-level. For example, PCK entails multiple, highly-interrelated dimensions such as knowledge of student understanding and knowledge of instructional strategies and representations (Carlson et al., 2019; Park & Oliver, 2008). Yet, neither student understanding, nor the concept of representation are unambiguous concepts. For example, students' understandings in domains such as evolution, mechanics, or thermodynamics are quite different from one other. Models of conceptual development and conceptual change explain facets of student understanding (diSessa, 2018), but there is arguably no monolithic knowledge of student understanding. Consequently, it is challenging to assess such dimensions of PCK, given that knowledge of student understanding can manifest in questioning strategies in science lessons, or in designing tasks in specific ways. Valid assessment of PCK might therefore entail multiple sources of evidence of classroom interactions, as well as constructed responses to questionnaires or vignette assessments (Zhai, Yin, et al., 2020). Educational researchers argued for the inception of performance assessments (Kulgemeyer & Riese, 2018). In such assessments, multiple sources of evidence have to be integrated in much more involved ways compared to simple closed-form questionnaires. For example, language utterances (which are complex and multi-faceted) can be utilized as sources of evidence for PCK (as in the PCK maps approach, see Park and Chen (2012)).

Besides improved forms of assessment, pre-service teachers in STEM fields need to receive feedback for their task-performance and teaching enactments that should be continuous, adaptive, and timely (Clarke & Hollingsworth, 2002; Hattie & Timperley, 2007). Moreover, feedback should be grounded in established theory. For example, reflections are oftentimes scaffolded with reflection-supporting models (Kost, 2019; Lai & Calandra, 2010; A. Nowak et al., 2019). Reflection is typically initiated by an unexpected event or a problem (Dewey, 1933; Rodgers, 2002). This event or problem is to be described in a way that is as objective as possible. Also, circumstantial factors (e.g., the learning objectives, grade level, etc.) have to be outlined. Afterwards, reasons for this event occurring have to be singled out (e.g., particular question of a student that might hint at difficulties of understanding the subject matter). Finally, alternative modes of action, as well as consequences for personal professional development, must be outlined (Korthagen, 1999; Korthagen & Kessels, 1999). There are several well-known issues that PSTs face when reflecting upon their teaching enactments. For example, particular problematic situations might not be recognized, or students reflect their teaching enactments in a rather evaluative mode, as compared to a more objective way (Korthagen & Kessels, 1999). Assessing verbal reports of reflections is daunting and requires empirically grounded means of analysis. Providing feedback and guidance for such learning products is even more challenging, given that this takes up substantial resources.

With the recent progress in computer-based assessment and AI, novel opportunities appear for research and support of professional competencies such as PSTs' PCK and reflections. For example, AI-based technologies were found to be capable of analysis of complex data sets, automated assessment, and individualized feedback,

which offer novel potentials for PST education (Xu & Ouyang, 2022). At the same time, these technologies are riddled with challenges such as lack of transparency, unreliable information extraction, and bias. These challenges are severe for high-stakes testing environments, as encountered in educational institutions (Krist et al., 2025).

Artificial Intelligence, Machine Learning, and Natural Language Processing

The study of AI originated in the mid-20th century and its workhorse method was termed machine learning (ML) (T. Mitchell, 1997; Samuel, 1959). ML, as a statistical learning approach, enables computers to learn patterns and relationships from data and enables them to generate reasonable output. Early applications of ML included learning weights in an artificial neural networks, the (multi-layer) perceptron, to classify data (ROSENBLATT, 1958). Later on, ML algorithms were used for computerized language processing, namely natural language processing (NLP). Early approaches included retrieval of information from text with ML algorithms by transforming text into vectors and calculating similarity (Deerwester et al., 1990). Two widely used ML learning approaches are supervised and unsupervised learning (Zhai, Haudek, Shi, et al., 2020; Zhai, Yin, et al., 2020). In the former, the labels are provided along with some examples. The latter is characterized only by examples without labels.

An intricate problem remained with capturing natural language with computers. Natural language is quite complex to analyze, even rather simple constructed response items comprise large outcome spaces (Meurers, 2012). For example, a typical learner can choose from tremendous resources to generate a response: S/he can choose from a vocabulary (mental lexicon) that might comprise some thousand words to produce language as well as mental rules (M. A. Nowak et al., 2001), that constrain allowed sentences to be produced. Such situations are characterized by combinatorial explosion, or "curse of dimensionality" (Bishop, 2006). A consequence of this large outcome space is the sparsity with which any combination of features in the data set is present in the outcome space. In ML research this relates to the problem of generalization. While it has been verified that i.i.d. (independently, identically distributed) generalization is feasible for deep learning models, it is often unclear what true generalization means in different contexts (Note that in high-dimensional (>100 dimensions) data sets, any problem amounts to be extrapolation rather than interpolation, see Balestrierio et al. (2023)). However, the surprise for researchers in ML and AI was that either the "curse of dimensionality" and related problems of high-dimensional optimization are no real challenges but rather affordances for deep learning (Sejnowski, 2024).

Regularities in language help capture the complexity (M. A. Nowak et al., 2001). For example, science language is organized around core concepts such as energy, or system, and exhibits regularities that suggest natural language to be a complex system (Wulff, 2023). Complex systems are comprised of parts, e.g., individual words, that form a whole, e.g., a text. It was then shown that words in language exhibit certain distributional properties: the frequency of words scales with the rank according to a power law distributions, and shorter words occur more often compared to longer words, which can be explained by information-theoretic principles such as least-effort (Zanette, 2014). These are patterns that are extraordinarily difficult for human researchers to see. Yet, AI-based tools were also found to excel at language analysis. One particular NLP application that was found to excel at language-related tasks were language models, which seek to predict a token based on a sequence of text (Goldberg,

2017). Language models became largely possible with advances in deep learning, i.e., the training of artificial neural networks with many layers and millions of parameters. In particular, mapping discrete symbols (such as tokens or words) to continuous, lower-dimensional vectors (called embeddings) became a facilitator of performance improvements in many NLP tasks (Goldberg, 2017). Later, large language models (LLMs) improved upon language models, mainly by novel architectures. In particular, an attention mechanism in the transformer language model architecture was a major improvement over existing architectures (Vaswani et al., 2017). After training these models on vast corpora of data, transformer-based LLMs are capable to attend to important words in a sequence and, in an auto-regressive manner, generate probable next tokens. Several parameters can be adjusted by the user to accommodate for the degree of determinism with which the tokens are chosen from the probability distribution.

In sum, (generative) AI tools such as ML algorithms or LLMs can leverage novel potentials for STEM education research and instruction such as assessment and feedback for different domains in the RCM such as collected or enacted PCK. Yet, as exciting as the potentials may be, ML algorithms are inherently statistical by nature, and thus model validation is an intricate and complex problem. We will now review applications of ML and LLMs in pre-service STEM teacher education and reflect upon potentials and challenges given the details outlined in the reviewed studies.

Applications of ML in (STEM) Teacher Education

Supervised and unsupervised ML have been employed in STEM education research. An early application of ML was presented by Wang et al. (2008), who combined ML and NLP to evaluate problem solving performance earth science education. They manually classified open-ended responses and found that ML algorithms could predict category membership approximately equally well as human raters could (Bleckmann & Friege, 2023; Donnelly et al., 2015; Krüger & Krell, 2020; H.-S. Lee et al., 2019; Linn et al., 2014; O. L. Liu et al., 2014; Mao et al., 2018; Nehm et al., 2012; Sherin, 2013; Wulff et al., 2021; Zhai et al., 2022; Zhu et al., 2017). ML and NLP have then been used in in-service and pre-service (STEM) teacher education (Salas-Pilco et al., 2022). Mostly supervised ML learning techniques were used to classify teachers' responses in specific task environments (Xu & Ouyang, 2022; Zhai, Haudek, Shi, et al., 2020). For example, Wahlen et al. (2020) use a supervised ML approach to automatically classify the PCK of economy teachers. They found that a good human-machine agreement was achievable. Zhai, Haudek, Stuhlsatz, and Wilson (2020) utilized supervised ML to score in-service science teachers' PCK-based constructed responses to video-clips of elementary science lessons. They found that the ML algorithms could approximate the human scoring. Moreover, the ML algorithms were more consistent in scoring different scenarios compared to the human raters. These findings resonate with other studies that applied ML methods, showing that simple classification and scoring problems can be imitated by machines (Wang et al., 2008). What becomes apparent in this study is that a substantial amount of manual coding is required, as well as technical expertise to train the ML algorithms. These then are only valid for the specific scenarios, and it is not clear if they are also valid for PST responses. In reference to the RCM, traditional ML offer the possibility to reliably assess personal PCK even with constructed-responses. Constructed response items have been argued to be more appropriate for validly assessing complex constructs (Nehm et al., 2012; von Aufschnaiter et al., 2019).

As such, ML provides a path toward implementing performance assessment at scale (note that automatic transcription of recorded voice is a task AI technologies excel at, which could be utilized in recording of teaching enactments).

Also text data and different professional skills can be assessed with traditional ML. Instructors and researchers recognized the tremendous efforts that are involved with assessing reflections in teacher education (Leonhard & Rihm, 2011; Solopova et al., 2023; Ullmann, 2019). This often leads to the neglect of high-quality feedback for reflections, which arguably hinders PSTs in actively learning from their teaching experiences. Education researchers devised process models as proposals for what a complete reflection entails. Typically, levels (depths) and elements (breadth) for reflections are differentiated (Poldner et al., 2014). Assessing the depth of reflections is rather difficult, and holistic assessment of reflections is required (Poldner et al., 2014). This process (as any coding process) is conceptually challenging (Biernacki, 2014), and researchers resorted to more analytical, low-inferential coding, which can be automated readily. (Ullmann, 2019) could show that simple ML models could be utilized to classify sentences in reflective writing in educational contexts. Wulff et al. (2021) utilized ML models in the context of PSTs' written reflections, and showed that sentences which are classified according to a reflection-supporting model by A. Nowak et al. (2019) could be input into an ML model which then relates them to elements in the reflection-supporting model. Reflection is a lever for professional development as outlined in the RCM. Consequently, ML and AI technologies can play an important role to strengthen assessment and potentially feedback in this domain in the RCM.

ML and NLP also excel at extracting patterns in high-dimensional, complex data sets (Hastie et al., 2008). This is typically done with unsupervised ML methods (Odden et al., 2019). To exploratively analyze complex, constructed-response data, unsupervised ML approaches have been utilized. Copur-Gencturk et al. (2023) used an unsupervised ML approach to identify topics in constructed responses about proportions by mathematics teachers. They conclude that the unsupervised ML approach could capture relevant nuances in the mathematics teachers' (as cross-validated with a qualitative content analysis), and that these nuances are important for performance. Similarly, Cutumisu and Guo (2019) applied an unsupervised ML approach to identify themes in pre-service science teachers' reflections and found that many students reflected upon their positive experience (Topic 3). This topic modeling approach also allowed the authors to assign students reflections to topics, and evaluate the share of different topics across the corpus of reflections. Such a "positivity bias" also appeared in Wulff et al. (2021).

Overall, for assessment of constructed response items, traditional ML and NLP can provide valuable resources (Nehm et al., 2012). STEM educators increasingly embrace knowledge-in-use assessments and performance tests, given the importance to assess competencies in authentic (performance) situations (Harris et al., 2019; Kulgemeyer & Riese, 2018), and they utilize ML and AI technologies to augment these assessments (Zhai, Haudek, Stuhlsatz, & Wilson, 2020; Zhai et al., 2022). In-depth insights into reasoning processes and knowledge application are expected from constructed response assessments (O. L. Liu et al., 2014), which allows science teacher education researchers engaged with the RCM to accurately and saleably assess performance-related skills such as enacted PCK or reflection. While traditional ML models might not generalize well across different

scenarios (that require wide transfer), human raters also face flaws. Human raters require resources each year to complete the manual coding, which is a benefit of the machines (Nehm et al., 2012). Moreover, the expertise of different human raters can vary, and effects of fatigue restrict the amount of coding that can be performed in a certain period of time (Zhai, Yin, et al., 2020).

Deep Learning Applications in (Science) Teacher Education

Advances in ML mainly related to processing larger data sets, with more sophisticated algorithms, in a faster way. This allowed researchers to train deep learning models (not to be confused with the concept “deep learning” in teacher education) which typically outperform more traditional ML models (Goodfellow et al., 2016). However, deep learning models also introduce more complexity and hence model decisions are typically less interpretable (Goodfellow et al., 2016). The improvements in performance were also found when applying deep learning models in PST education. In assessment of PSTs' written reflection it was shown that the generalization performance of the traditional ML models (here: multinomial logistic regression) was rather poor. Use of deep learning-based models, here LLMs, could substantially improve classification performance and generalizability, either across teachers, institutions, and even subjects (Carpenter et al., 2021; Carpenter et al., 2020; Wulff, Mientus, et al., 2022). These models were then used in practice to guide PSTs in their reflective writing by providing feedback on the structure of the reflection and identifying opportunities for improvement (Mientus et al., 2023). Moreover, once developed, ML models that are based on LLMs could be further improved in novel contexts (Wulff et al., 2023). This opens up the possibility to co-constructively and cumulatively develop and refine coding rubrics that are applicable across research contexts.

Deep learning ML models, and LLMs in particular, can also be used for unsupervised ML analyses. Given the link of reflection to prior experiences, analysis of reflections by PSTs can yield insights into the noticing process of PSTs. For example, Wulff, Buschhüter, et al. (2022) utilized LLMs to transform written reflections of PSTs into vector representations to be further processed. Transformer-based LLMs (Devlin et al., 2018) were utilized to retrieve such representations that are sensitive to word senses (Reimers & Gurevych, 2019; Wiedemann et al., 2019). Furthermore, clustering approaches can then be utilized to meaningfully differentiate different topics in language data such as reflection reports. (Solopova et al., 2023) designed a feedback AI, based on didactic theory for pre-service teachers. They combined supervised ML algorithms and unsupervised ML algorithms to design a multi-faceted automated feedback based on AI.

With regards to the identified domains of the RCM, deep learning improves upon the capabilities of ML for assessment. The capabilities of LLMs and unsupervised ML also allow for systematic analysis of collective PCK by concurrently analyzing large corpora, e.g., textbook of PCK. Deep learning methods then enable the integrated analysis of collective and personal PCK. However, deep learning approaches oftentimes lack easy methods for understanding model decisions, which poses a challenge for STEM education research given that validity concerns oftentimes need full disclosure of why certain decisions were made. Particularly in high-stakes educational environments applications of these technologies needs to be regulated (Wulff et al., 2025).

Generative AI in (Science) Teacher Education

Generative AI (either vision models or language models) offer novel potentials for analyzing data sets, as well as generative tasks such as providing adaptive guidance. In general, generative language models such as GPT-3.5 and GPT-4 (generative pre-trained transformers) were shown to capture a large amount of domain knowledge, such as conceptual knowledge in physics (Kieser et al., 2023; Kortemeyer, 2023; West, 2023), or capabilities in mathematics and engineering-related quantitative reasoning (Lewkowycz et al., 2022). Cooper (2023) applied a self-study methodology and found that ChatGPT (an often used generative LLM based on GPT models) could meaningfully respond to education-related questions in STEM insofar as it included research themes in the educational research literature. Zhai (2023) found that ChatGPT could be utilized to help with writing an academic research paper on AI in education. However, pre-service and in-service teachers only have limited experience with and knowledge about these (Gen)AI technologies (partly due to lacking infrastructure) (NYAABA & Zhai, 2024; Sperling et al., 2024).

Related to teachers' professional skills, Li et al. (2023) investigated whether GenAI LLMs (ChatGPT) could write reflectively and generate high-quality reflections. The authors found that ChatGPT generated reflections of higher quality compared to the students' reflections, as assessed through an eight-category rubric. Moreover, they found that human expert raters could not reliably differentiate reflective writing generated by the students versus ChatGPT. This resonates with findings that ChatGPT-generated physics essays reach student quality and cannot be detected as artificially generated (Yeadon et al., 2023). These findings raise the concern that students might outsource important reflective tasks to GenAI and fail to reflect their own learning processes. Furthermore, G.-G. Lee and Zhai (2024) enabled PST to utilize ChatGPT to design science lesson plans. The designed lesson plans by ChatGPT were found to be variable with regards to teaching and learning strategies, and domains. Thus, ChatGPT could function as an individualized assistant for PSTs to plan lessons or as a tool to generate learning materials. Similarly, Küchemann et al. (2023) allow PST to use ChatGPT to design instructional tasks, and the research by Kieser et al. (2023) indicates that GenAI tools could be utilized to test one's designed tests and instructional materials. In the study by Küchemann et al. (2023) it was found that tasks by PSTs who used ChatGPT were equally accurate compared to tasks designed by PST who used a textbook. However, the textbook group achieved higher clarity and more meaningful contexts. Given that ChatGPT is not a dedicated resources for physics-specific purposes this is quite interesting for instructional settings and highlights the usefulness of ChatGPT for PST.

In fact, specifically instructing (i.e., prompting) GenAI LLMs such as GPT can improve outcomes and ground them towards the expectations of researchers. Even simple prompts such as "let's think step by step" (or, for that matter, imagine being a Star Trek character when solving math problems) were found to improve the accuracy of the output of the generative LLM (Polverini & Gregorcic, 2024a; Wei, 2022). The basic idea behind prompting strategies is to breakdown tasks into subtasks, and make explicit the chain of "thoughts" involved in reaching a certain solution (Khot et al., 2023). More particularly, Wan and Chen (2024) showed that extensive prompting for feedback on a physics problem solution could enable the generative LLM to provide useful and accurate feedback for learners. Important facets of the designed prompting template were: 1) provide an expert solution, 2) provide

examples, and 3) describe how the feedback should be designed. This prompting strategy then enabled the LLM to generate feedback which was equally accurate and more personally useful, as evaluated by the feedback-receiving students.

GenAI tools such as generative LLMs enables a multitude of applications related to the domains in the RCM. For example, designing prompting templates can be valuable in reflective writing analytics, since it can be used to better enable a generative LLM to provide targeted feedback for PSTs' written reflections and thus help improve the implementation of the teaching cycle in the RCM. This then could be leveraged as a resource to support PST to reflect in teaching internships. Moreover, generative LLMs can be used by PST as a resource to probe their personal PCK, e.g., their knowledge of student understanding, methods, or instructional strategies. To harness the full potentials of these genAI technologies, STEM education researchers should engage in designing effective templates that facilitate feedback grounded in theoretical constructs such as PCK.

Conclusions and Future Directions

In Figure 1 we outlined the domains assessing the realms of PCK and enhancing professional development for teachers in STEM. We argue that in both realms AI methods can provide unique potentials that already have been partly realized in the reviewed studies, but need to be systematically taken up and expanded upon. With regards to assessing the realms of PCK, (gen)AI, and ML in particular allow researchers to extract patterns in complex data (unsupervised ML), and automatically score data according to established coding rules (supervised ML). These two ML approaches have been utilized in STEM teacher education to explore themes in PST's PCK, automatically score constructed PCK responses, extract topics and themes in constructed responses, and assess reflective thinking processes. Deep learning ML models then allowed educational researchers to increase accuracy in coding, as well as make more informed thematic maps of PST's PCK. Moreover, large representative document corpora can be explored with AI methods and therefore facilitate systematic curriculum development and extraction of students' ideas. With tools such as automated transcription (AI based) and NLP, researchers have now tools to concurrently analyze large language corpora with reference to PCK and thus identify relationships between personal and enacted PCK.

Related to enhancing professional development of STEM teachers, AI methods can foster knowledge exchange and analysis of teaching and student outcomes as outlined in Figure 1. For example, genAI can be utilized to design instructional tasks, and even generate lesson plans. The scope of possible tasks is even broader, and specific prompting of generative genAI models was shown to be able to improve the generated outputs towards pre-specified instruction. These capabilities can enhance adaptive, individualized feedback, which is crucial for teachers' professional development (Clarke & Hollingsworth, 2002). GenAI models can also be used to integrate student outcomes into prompting and thus help ground teachers' reflections to actual learning outcomes. This provides entirely novel ways of generating integrated feedback, however, research is only at its infancy. Development of professional skills such as lesson planning and reflection can be enhanced with genAI. For example, GenAI models can be utilized to generate lesson plans (G.-G. Lee & Zhai, 2024). However, they can also be used in the reverse direction: to identify themes and knowledge elements in lesson plans by PSTs. This

hinges on specific prompting strategies, which researchers have to develop. Existing theory and models for either lesson planning, and reflection are crucial for prompt design, given that they can inform the design of meaningful prompts (Solopova et al., 2023). This then allows for adaptive, computerized feedback that has long been shown to be very effective regarding learning outcomes (Lai & Calandra, 2010; VanLehn, 2011). The recent advances in GenAI substantially boost performance and versatility of automated dialogue systems that could provide teachers their personal agents to improve professional skills.

Considering Limitations

AI approaches are no silver bullets and particular care with applying them has to be taken, especially in (STEM) teacher education where decisions eventually are very consequential for the individual teachers and their prospective students (as they become multipliers). Above we highlighted that ML algorithms are statistical learning procedures. This relates to the most substantial limitation of these approaches: they are ultimately constrained by the given training data set. Even though LLMs are capable to generate novel output (e.g., sentences that are not in the training data set), ML algorithms lack an appropriate world model, which makes them fail glaringly in tasks that can be trivial for humans (Bahdanau et al., 2019; M. Mitchell, 2023; M. Mitchell & Krakauer, 2023; M. Mitchell et al., 2023). While Lake and Baroni (2023) showed that certain genAI models are capable to systematically generalize (i.e., infer underlying rules of a system such as a grammar) to some extent, M. Mitchell et al. (2023) showed that similar genAI models fail at abstraction tasks, and thus lack fundamental capabilities that humans have. Researchers showed that the underlying machinery of GenAI LLMs is all but comprehensible.

The lack of an appropriate world model and spoiled training data in genAI lead to limitations such as false factual knowledge (Gregorcic & Pendrill, 2023), limitations in multi-modal capabilities (Polverini & Gregorcic, 2024b), as well as limitations related to problem solving capabilities (Kieser & Wulff, 2024). In fact, oftentimes these models simply regurgitate what was seen in training (Bender et al., 2021), which could resemble a misconception about conceptual science knowledge. This raises the important question of to what extent these models with specific training regimes can be creative or innovative, which many researchers doubt (Browning & LeCun, 2022). "It's extremely typical of machine learning that it manages to do a good job of getting things 'roughly right'. But nailing the details is not what machine learning tends to be good at" (Wolfram, 2024). Wolfram (2024) further argues that ML algorithms such as artificial neural networks will find pockets/attractors in their sensory spaces (e.g., chemical space), however, humans will only be able to detect the sensible, or interpretable, ones, i.e., those for which an experiential basis exists. This creates an interface problem where humans might not be able to make sense of innovative patterns that eventually might be stored in the ML models. As long as generative LLMs are grounded in experiential data, they might accomplish certain tasks satisfactorily. However, once reliable knowledge and reasoning capabilities become necessary, these models were shown to be brittle (M. Mitchell, 2023). Even worse, they reproduce biases and stereotypes related to ethnicity, race, or gender that are present in the training data (Christian, 2021), fundamentally related to aleatoric and epistemic uncertainty in ML (Wulff et al., 2025). Although many efforts have been made to reverse engineer these failures, progress has been modest to our estimation. Alongside issues of privacy, proprietary rights, and closed-source software (at least for GPT

models, somewhat ironically related by a company called OpenAI), application in educational practice should be severely limited and closely monitored.

With regards to STEM teacher education it is particularly problematic that no high-quality training data is provided for genAI models which encapsulates learning difficulties, knowledge of students understanding, and other facets of established educational theory in STEM fields. At least, we cannot assure that other training data that encapsulates common misconceptions about STEM-related phenomena overrides the knowledge gained in educational research. As such, critical investigation of the employed models and careful prompt engineering become all the more relevant.

Concluding Remarks

In sum, AI and ML methods are beginning to be used in PST education and can provide valuable resources to automate assessment and provide individualized feedback, and thus potentially enhance professional development of PSTs. With careful consideration for the specific affordances and limitations, AI and ML methods can be expected to provide novel capabilities for enhancing professional development of STEM teachers with reference to the specific domains in the RCM. GenAI tools offer particular opportunities for assessment and adaptive feedback. However, precision of genAI for specific tasks is limited, especially when it comes to tasks that require high levels of expertise (Solopova et al., 2023).

Many tasks in professional development programs in teacher education in STEM arguably require medium levels of expertise, such as providing feedback on the breadth of teachers' reflections, or the structure of their lesson plans. Even such structural (potentially superficial feedback) can have positive effects even on teachers understanding of certain tasks and requirements, which could streamline professional practices where PSTs rightfully complain that they often lack a precise knowledge of the requirements. For instructors this could exempt them from repetitive tasks such as specifying the task requirements each and every time.

Appropriately motivating and justifying the use of genAI in educational contexts will become a crucial means to raise awareness and engagement with this important new technology. Nazaretsky et al. (2022) designed a professional development program for fostering teachers AI-related understanding. The authors identify confirmation bias and trust as a crucial components to address to motivate teacher to adopt novel technologies. By no means should we encourage teachers to trust any AI technology, however, we eventually should foster a mindset to explore potentials and critically reflect areas for application and limitations.

Researchers argue that LLMs can particularly complement and augment humans in (explorative) ideation processes, whereas verification of solutions is not a strong suit of these particular AI agents. Such considerations can guide efforts to utilize AI to enhance PST professional development. Empowering PSTs in STEM fields to utilize genAI, which are enriched by researchers and instructors templates, could help reduce the uncertainty that comes along with the teaching profession and help teachers to focus on their actual profession: effectively teaching STEM subjects.

Abbreviations

AI	Artificial intelligence
GenAI	Generative artificial intelligence
LLM	Large language model
ML	Machine learning
NLP	Natural language processing
PCK	Pedagogical content knowledge
PST	Pre-service teachers in STEM fields
RCM	Refined consensus model

References

- Abell, S. K. (2007). Research on Science Teacher Knowledge. In S. K. Abell & N. Lederman (Eds.), *Handbook of research on science education*. Lawrence Erlbaum Associates Publishers.
- Aeppli, J., & Lötscher, H. (2016). EDAMA - Ein Rahmenmodell für Reflexion. *Beiträge Zur Lehrerinnen- Und Lehrerbildung*, 34(1), 78–97. <https://doi.org/10.25656/01:13921>
- Alonzo, A. C., Berry, A., & Nilsson, P. (2019). Unpacking the Complexity of Science teachers' PCK in Action: Enacted and Personal PCK. In A. Hume, R. Cooper, & A. Borowski (Eds.), *Repositioning Pedagogical Content Knowledge in Teachers' Professional Knowledge* (pp. 271–286). Springer.
- Bahdanau, D., Murty, S., Noukhovitch, M., Nguyen, T. H., Vries, H. d., & Courville, A. (2019). Systematic Generalization: What Is Required and Can It Be Learned? *ArXiv*.
- Balestrierio, R., Pesenti, J., & LeCun, Y. (2023). Learning in High Dimension Always Amounts to Extrapolation. *ArXiv*.
- Baumert, J., & Kunter, M. (2011). Das Kompetenzmodell von COACTIV. In M. Kunter, J. Baumert, W. Blum, U. Klusmann, S. Krauss, & M. Neubrand (Eds.), *Professionelle Kompetenz von Lehrkräften: Ergebnisse des Forschungsprogramms COACTIV* (pp. 29–53). Waxmann.
- Baumert, J., Kunter, M., Blum, W., Brunner, M., Voss, T., Jordan, A., Klusmann, U., Krauss, S., Neubrand, M., & Tsai, Y.-M. (2010). Teachers' Mathematical Knowledge, Cognitive Activation in the Classroom, and Student Progress. *American Educational Research Journal*, 47(1), 133–180. <https://doi.org/10.3102/0002831209345157>
- Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the Dangers of Stochastic Parrots. *FAccT*, 610–623. <https://doi.org/10.1145/3442188.3445922>
- Biernacki, R. (2014). Humanist Interpretation Versus Coding Text Samples. *Qualitative Sociology*, 37(2), 173–188. <https://doi.org/10.1007/s11133-014-9277-9>
- Bishop, C. M. (2006). *Pattern recognition and machine learning*. Information Science and Statistics. Springer Science+Business Media LLC. <https://www.microsoft.com/en-us/research/uploads/prod/2006/01/Bishop-Pattern-Recognition-and-Machine-Learning-2006.pdf>

- Bleckmann, T., & Friege, G. (2023). Concept maps for formative assessment: Creation and implementation of an automatic and intelligent evaluation method. *Knowledge Management & E-Learning: An International Journal*, 433–447. <https://doi.org/10.34105/j.kmel.2023.15.025>
- Browning, J., & LeCun, Y. (2022). AI And The Limits Of Language. <https://www.noemamag.com/ai-and-the-limits-of-language/>
- Carlson, J., Daehler, K., Alonzo, A. C., Barensen, E., Berry, A., Borowski, A., Carpendale, J., Chan, K., Cooper, R. [Rebecca], Friedrichsen, P., Gess-Newsome, J., Henze-Rietveld, I., Hume, A., Kirschner, S., Liepertz, S., Loughran, J., Mavhunga, E., Neumann, K., Nilsson, P., . . . Wilson, C. D. (2019). The Refined Consensus Model of Pedagogical Content Knowledge. In A. Hume, R. Cooper, & A. Borowski (Eds.), *Repositioning Pedagogical Content Knowledge in Teachers' Professional Knowledge*. Springer.
- Carpenter, D., Cloude, E., Rowe, J., Azevedo, R., & Lester, J. (2021). Investigating Student Reflection during Game-Based Learning in Middle Grades Science: International Learning Analytics and Knowledge Conference (LAK21), 280–291. <https://doi.org/10.1145/3448139.3448166>
- Carpenter, D., Geden, M., Rowe, J., Azevedo, R., & Lester, J. (2020). Automated Analysis of Middle School Students' Written Reflections During Game-Based Learning. In I. I. Bittencourt, M. Cukurova, K. Muldner, R. Luckin, & E. Millán (Eds.), *Artificial Intelligence in Education* (pp. 67–78). Springer International Publishing.
- Christian, B. (2021). *The Alignment Problem: How Can Machines Learn Human Values?* Atlantic Books. <https://ebookcentral.proquest.com/lib/kxp/detail.action?docID=6294690>
- Clarke, D., & Hollingsworth, H. (2002). Elaborating a model of teacher professional growth. *Teaching and Teacher Education*, 18(8), 947–967. [https://doi.org/10.1016/S0742-051X\(02\)00053-7](https://doi.org/10.1016/S0742-051X(02)00053-7)
- Cooper, G. (2023). Examining Science Education in ChatGPT: An Exploratory Study of Generative Artificial Intelligence. *Journal of Science Education and Technology*, 32(3), 444–452. <https://doi.org/10.1007/s10956-023-10039-y>
- Copur-Gencturk, Y., Choi, H.-J., & Cohen, A. (2023). Investigating teachers' understanding through topic modeling: a promising approach to studying teachers' knowledge. *Journal of Mathematics Teacher Education*, 26(3), 281–302. <https://doi.org/10.1007/s10857-021-09529-w>
- Cutumisu, M., & Guo, Q. (2019). Using Topic Modeling to Extract Pre-Service Teachers' Understandings of Computational Thinking From Their Coding Reflections. *IEEE Transactions on Education*, 62(4), 325–332. <https://doi.org/10.1109/TE.2019.2925253>
- Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. (1990). Indexing by Latent Semantic Analysis.
- Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *ArXiv*, 1810.04805.
- Dewey, J. (1933). *How we think: A restatement of the relation of reflective thinking to the educative process* ((New ed.)). Heath.
- diSessa, A. A. (2018). Computational Literacy and “The Big Picture” Concerning Computers in Mathematics Education. *Mathematical Thinking and Learning*, 20(1), 3–31. <https://doi.org/10.1080/10986065.2018.1403544>

- Donnelly, D. F., Vitale, J. M., & Linn, M. C. (2015). Automated Guidance for Thermodynamics Essays: Critiquing Versus Revisiting. *Journal of Science Education and Technology*, 24(6), 861–874. <https://doi.org/10.1007/s10956-015-9569-1>
- Gess-Newsome, J. (2015). A model of teacher professional knowledge and skill including PCK: Results of the thinking from the PCK summit. In A. Berry, P. Friedrichsen, & J. Loughran (Eds.), *Teaching and learning in science series. Re-examining pedagogical content knowledge in science education*. Routledge.
- Gess-Newsome, J., Taylor, J. A., Carlson, J., Gardner, A. L., Wilson, C. D., & Stuhlsatz, M. A. M. (2017). Teacher pedagogical content knowledge, practice, and student achievement. *International Journal of Science Education*, 1–20. <https://doi.org/10.1080/09500693.2016.1265158>
- Goldberg, Y. (2017). *Neural Network Methods for Natural Language Processing. Synthesis Lectures on Human Language Technologies*. Morgan and Claypool.
- Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT Press. <http://www.deeplearningbook.org/>
- Gregorcic, B., & Pendrill, A.-M. (2023). ChatGPT and the frustrated Socrates. *Physics Education*, 58(3), 35021. <https://doi.org/10.1088/1361-6552/acc299>
- Grossman, P. L., Compton, C., Igra, D., Ronfeldt, M., Shahan, E., & Williamson, P. W. (2009). Teaching practice: A cross-professional perspective. *Teachers College Record*, 111(9), 2055–2100.
- Harris, C., Krajcik, J. S., Pellegrino, J. W., & DeBarger, A. H. (2019). Designing Knowledge-In-Use Assessments to Promote Deeper Learning. *Educational Measurement: Issues and Practice*, 38(2), 53–67.
- Hastie, T., Tibshirani, R., & Friedman, J. (2008). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer.
- Hattie, J., & Timperley, H. (2007). The Power of Feedback. *Review of Educational Research*, 77(1), 81–112. <https://doi.org/10.3102/003465430298487>
- Kasneci, E., Sessler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., Gasser, U., Groh, G., Günemann, S., Hüllermeier, E., Krusche, S., Kutyniok, G., Michaeli, T., Nerdel, C., Pfeffer, J., Poquet, O., Sailer, M., Schmidt, A., Seidel, T., . . . Kasneci, G. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and Individual Differences*, 103, 102274. <https://doi.org/10.1016/j.lindif.2023.102274>
- Khot, T., Trivedi, H., Finlayson, M., Fu, Y., Richardson, K., Clark, P., & Sabharwal, A. (2023). Decomposed Prompting: A Modular Approach for Solving Complex Tasks. *ArXiv*.
- Kieser, F., & Wulff, P. (2024). Using large language models to probe cognitive constructs, augment data, and design instructional materials. In M. S. Khine (Ed.), *Machine Learning in Educational Sciences: Approaches, Applications and Advances*. Springer Nature. https://doi.org/10.1007/978-981-99-9379-6_14
- Kieser, F., Wulff, P., Kuhn, J., & Küchemann, S. (2023). Educational data augmentation in physics education research using ChatGPT. *Physical Review Physics Education Research*, 19(2), 1–13. <https://doi.org/10.1103/PhysRevPhysEducRes.19.020150>

- Kind, V., & Chan, K. K. H. (2019). Resolving the amalgam: connecting pedagogical content knowledge, content knowledge and pedagogical knowledge. *International Journal of Science Education*, 41(7), 964–978. <https://doi.org/10.1080/09500693.2019.1584931>
- Kluger, A. N., & DeNisi, A. (1996). The effects of feedback interventions on performance: A historical review, a meta-analysis, and a preliminary feedback intervention theory. *Psychological Bulletin*, 119(2), 254–284. <https://doi.org/10.1037/0033-2909.119.2.254>
- Kortemeyer, G. (2023). Could an Artificial-Intelligence agent pass an introductory physics course? *Physical Review Physics Education Research*, 19(1), Article 010132, 15. <https://doi.org/10.1103/PhysRevPhysEducRes.19.010132>
- Korthagen, F. A. (1999). Linking Reflection and Technical Competence: the logbook as an instrument in teacher education. *European Journal of Teacher Education*, 22(2-3), 191–207. <https://doi.org/10.1080/0261976899020191>
- Korthagen, F. A., & Kessels, J. (1999). Linking Theory and Practice: Changing the Pedagogy of Teacher Education. *Educational Researcher*, 28(4), 4–17.
- Kost, D. (2019). *Reflexionsprozesse von Studierenden des Physiklehramts: Dissertation at Justus-Liebig-University in Gießen*.
- Krauss, S., Baumert, J., & Blum, W. (2008). Secondary mathematics teachers' pedagogical content knowledge and content knowledge: validation of the COACTIV constructs. *ZDM*, 40(5), 873–892. <https://doi.org/10.1007/s11858-008-0141-9>
- Krist, C., Kubsch, M., & Wulff, P. (2025). Human-Machine Interactions in Machine Learning Modeling: The Role of Theory. In P. Wulff, M. Kubsch, & C. Krist (Eds.), *Applying Machine Learning in Science Education Research: When, How, and Why?* (pp. 143–154). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-74227-9_8
- Krüger, D., & Krell, M. (2020). Maschinelles Lernen mit Aussagen zur Modellkompetenz. *Zeitschrift Für Didaktik Der Naturwissenschaften*, 26(1), 157–172. <https://doi.org/10.1007/s40573-020-00118-7>
- Küchemann, S., Steinert, S., Revenga, N., Schweinberger, M., Dinc, Y., Avila, K. E., & Kuhn, J. (2023). Physics task development of prospective physics teachers using ChatGPT. *ArXiv*.
- Kulgemeyer, C., & Riese, J. (2018). From professional knowledge to professional performance: The impact of CK and PCK on teaching quality in explaining situations. *Journal of Research in Science Teaching*, 55(10), 1393–1418. <https://doi.org/10.1002/tea.21457>
- Lai, G., & Calandra, B. (2010). Examining the effects of computer-based scaffolds on novice teachers' reflective journal writing. *Educational Technology Research and Development*, 58(4), 421–437. <https://doi.org/10.1007/s11423-009-9112-2>
- Lake, B. M., & Baroni, M. (2023). Human-like systematic generalization through a meta-learning neural network. *Nature*, 623.
- Lee, G.-G., & Zhai, X. (2024). Using ChatGPT for Science Learning : A Study on Pre-service Teachers' Lesson Planning. Advance online publication. <https://doi.org/10.13140/RG.2.2.13711.76965>
- Lee, H.-S., Pallant, A., Pryputniewicz, S., Lord, T., Mulholland, M., & Liu, O. L. (2019). Automated text scoring and real-time adjustable feedback: Supporting revision of scientific arguments involving uncertainty. *Science Education*, 103(3), 590–622. <https://doi.org/10.1002/sce.21504>

- Leonhard, T., & Rihm, T. (2011). Erhöhung der Reflexionskompetenz durch Begleitveranstaltungen zum Schulpraktikum? Konzeption und Ergebnisse eines Pilotprojekts mit Lehramtsstudierenden. *Lehrerbildung Auf Dem Prüfstand*, 4(2), 240–270. <https://doi.org/10.25656/01:14722>
- Lewkowycz, A., Andreassen, A., Dohan, D., Dyer, E., Michalewski, H., Ramasesh, V., Slone, A., Anil, C., Schlag, I., Gutman-Solo, T., Wu, Y., Neyshabur, B., Gur-Ari, G., & Misra, V. (2022). Solving Quantitative Reasoning Problems with Language Models. *ArXiv*.
- Li, Y., Sha, L., Yan, L., Lin, J., Raković, M., Galbraith, K., Lyons, K., Gašević, D., & Chen, G. (2023). Can large language models write reflectively. *Computers and Education: Artificial Intelligence*, 4, 100140. <https://doi.org/10.1016/j.caeai.2023.100140>
- Linn, M. C., Gerard, L., Ryoo, K., McElhaney, K., Liu, O. L., & Rafferty, A. N. (2014). Education technology. Computer-guided inquiry to improve science learning. *Science (New York, N.Y.)*, 344(6180), 155–156. <https://doi.org/10.1126/science.1245980>
- Liu, M., Buckingham Shum, S., Mantzourani, E., & Lucas, C. (2019). *Evaluating Machine Learning Approaches to Classify Pharmacy Students' Reflective Statements: International Conference on Artificial Intelligence in Education*.
- Liu, O. L., Brew, C., Blackmore, J., Gerard, L., Madhok, J., & Linn, M. C. (2014). Automated Scoring of Constructed-Response Science Items: Prospects and Obstacles. *Educational Measurement: Issues and Practice*, 33(2), 19–28. <https://doi.org/10.1111/emip.12028>
- Love, T. S., & Hughes, A. J. (2022). Engineering pedagogical content knowledge: examining correlations with formal and informal preparation experiences. *International Journal of STEM Education*, 9(1). <https://doi.org/10.1186/s40594-022-00345-z>
- Magnusson, S., Krajcik, J. S., & Borko, H. (1999). Nature, sources, and development of pedagogical content knowledge for science teaching. In J. Gess-Newsome & N. G. Lederman (Eds.), *Examining pedagogical content knowledge: The construct and its implication for science education* (pp. 95–132). Kluwer Academic.
- Mao, L., Liu, O. L., Roohr, K., Belur, V., Mulholland, M., Lee, H.-S., & Pallant, A. (2018). Validation of Automated Scoring for a Formative Assessment that Employs Scientific Argumentation. *Educational Assessment*, 23(2), 121–138. <https://doi.org/10.1080/10627197.2018.1427570>
- Mena-Marcos, J., García-Rodríguez, M.-L., & Tillema, H. (2013). Student teacher reflective writing: what does it reveal? *European Journal of Teacher Education*, 36(2), 147–163. <https://doi.org/10.1080/02619768.2012.713933>
- Meschede, N., Fiebranz, A., Möller, K., & Steffensky, M. (2017). Teachers' professional vision, pedagogical content knowledge and beliefs: On its relation and differences between pre-service and in-service teachers. *Teaching and Teacher Education*, 66, 158–170. <https://doi.org/10.1016/j.tate.2017.04.010>
- Meurers, D. (2012). Natural Language Processing and Language Learning. In C. A. Chapelle (Ed.), *The Encyclopedia of Applied Linguistics: 10 Volume Set* (1., Auflage). Wiley, J.
- Mientus, L., Hume, A. C., Wulff, P., Meiners, A., & Borowski, A. (2022). Modelling STEM Teachers' Pedagogical Content Knowledge in the Framework of the Refined Consensus Model: A Systematic Literature Review. *Education Sciences*, 12, 1–25. <https://doi.org/10.25932/PUBLISHUP-56912>

- Mientus, L., Wulff, P., Nowak, A., & Borowski, A. (2023). Fast-and-frugal means to assess reflection-related reasoning processes in teacher training—Development and evaluation of a scalable machine learning-based metric. *Zeitschrift Für Erziehungswissenschaft*. Advance online publication. <https://doi.org/10.1007/s11618-023-01166-8>
- Mishra, P., & Koehler, M. J. (2006). Technological Pedagogical Content Knowledge: A Framework for Teacher Knowledge. *Teacher College Records*, 108(6), 1017–1054.
- Mitchell, M. (2023). AI's challenge of understanding the world. *Science*, 382(6671), eadm8175. <https://doi.org/10.1126/science.adm8175>
- Mitchell, M., & Krakauer, D. C. (2023). The debate over understanding in AI's large language models. *Proceedings of the National Academy of Sciences of the United States of America*, 120(13), e2215907120. <https://doi.org/10.1073/pnas.2215907120>
- Mitchell, M., Palmarini, A. B., & Moskvichev, A. (2023). Comparing Humans, GPT-4, and GPT-4V On Abstraction and Reasoning Tasks. *ArXiv*.
- Mitchell, T. (1997). *Machine learning*. McGraw-Hill Education.
- Nazaretsky, T., Ariely, M., Cukurova, M., & Alexandron, G. (2022). Teachers' trust in AI -powered educational technology and a professional development program to improve it. *British Journal of Educational Technology*, 53(4), 914–931. <https://doi.org/10.1111/bjet.13232>
- Nehm, R. H., Ha, M., & Mayfield, E. (2012). Transforming Biology Assessment with Machine Learning: Automated Scoring of Written Evolutionary Explanations. *Journal of Science Education and Technology*, 21(1), 183–196. <https://doi.org/10.1007/s10956-011-9300-9>
- Nelson, L. K., Burk, D., Knudsen, M., & McCall, L. (2021). The future of coding: A comparison of hand-coding and three types of computer-assisted text analysis methods. *Sociological Methods & Research*, 50(1), 202–237.
- Neuweg, G. H. (2014). Das Wissen der Wissensvermittler. In E. Terhart, H. Bennewitz, & M. Rothland (Eds.), *Handbuch der Forschung zum Lehrerberuf* (2. überarbeitete und erweiterte Auflage, pp. 451–477). Waxmann.
- Nowak, A., Kempin, M., Kulgemeyer, C., & Borowski, A. (2019). Reflexion von Physikunterricht [Reflection of physics lessons]. In C. Maurer (Ed.), *Naturwissenschaftliche Bildung als Grundlage für berufliche und gesellschaftliche Teilhabe. Jahrestagung in Kiel 2018* (p. 838). Gesellschaft für Didaktik der Chemie und Physik.
- Nowak, M. A., Komarova, N. L., & Niyogi, P. (2001). Evolution of Universal Grammar. *Science*, 291, 114–118.
- NYAABA, M., & Zhai, X. (2024). Generative AI Professional Development Needs for Teacher Educators. *Journal of AI*, 8(1), 1–13. <https://doi.org/10.61969/jai.1385915>
- Odden, T. O. B., Lockwood, E., & Caballero, M. D. (2019). Physics computational literacy: An exploratory case study using computational essays. *Physical Review Physics Education Research*, 15(2). <https://doi.org/10.1103/PhysRevPhysEducRes.15.020152>
- Park, S., & Chen, Y.-C. (2012). Mapping out the integration of the components of pedagogical content knowledge (PCK): Examples from high school biology classrooms. *Journal of Research in Science Teaching*, 49(7), 922–941.

- Park, S., & Oliver, J. S. (2008). Revisiting the Conceptualisation of Pedagogical Content Knowledge (PCK): PCK as a Conceptual Tool to Understand Teachers as Professionals. *Research in Science Education*, 38(3), 261–284. <https://doi.org/10.1007/s11165-007-9049-6>
- Poldner, E., van der Schaaf, M., Simons, P. R.-J., van Tartwijk, J., & Wijngaards, G. (2014). Assessing student teachers' reflective writing through quantitative content analysis. *European Journal of Teacher Education*, 37(3), 348–373. <https://doi.org/10.1080/02619768.2014.892479>
- Polverini, G., & Gregorcic, B. (2024a). How understanding large language models can inform the use of ChatGPT in physics education. *European Journal of Physics*, 45(2), 25701. <https://doi.org/10.1088/1361-6404/ad1420>
- Polverini, G., & Gregorcic, B. (2024b). Performance of ChatGPT on the Test of Understanding Graphs in Kinematics. *ArXiv*.
- Reimers, N., & 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*. Association for Computational Linguistics. <https://arxiv.org/abs/1908.10084>
- Rodgers, C. (2002). Defining Reflection: Another Look at John Dewey and Reflective Thinking. *Teachers College Record*, 104(4), 842–866. <https://doi.org/10.1111/1467-9620.00181>
- ROSENBLATT, F. (1958). The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review*, 65(6), 386–408. <https://doi.org/10.1037/h0042519>
- Sadler, P. M., Sonnert, G., Coyle, H. P., Cook-Smith, N., & Miller, J. L. (2013). The Influence of Teachers' Knowledge on Student Learning in Middle School Physical Science Classrooms. *American Educational Research Journal*, 50(5), 1020–1049. <https://doi.org/10.3102/0002831213477680>
- Salas-Pilco, S., Xiao, K., & Hu, X. (2022). Artificial Intelligence and Learning Analytics in Teacher Education: A Systematic Review. *Education Sciences*, 12(8), 569. <https://doi.org/10.3390/educsci12080569>
- Samuel, A. L. (1959). Some studies in Machine Learning Using the Game of Checkers. *IBM Journal*.
- Schön, D. A. (1983). *The reflective practitioner: How professionals think in action*. Basic Books. <http://www.loc.gov/catdir/enhancements/fy0832/82070855-d.html>
- Sejnowski, T. J. (2024). *Chatgpt and the future of AI: The deep language revolution*. The MIT Press.
- Sherin, B. (2013). A Computational Study of Commonsense Science: An Exploration in the Automated Analysis of Clinical Interview Data. *Journal of the Learning Sciences*, 22(4), 600–638. <https://doi.org/10.1080/10508406.2013.836654>
- Shulman, L. S. (1986). Those Who Understand: Knowledge Growth in Teaching. *Educational Researcher*, 15(2), 4–14.
- Solopova, V., Rostom, E., Cremer, F., Gruszczynski, A., Witte, S., Zhang, C., López, F. R., Plöbl, L., Hofmann, F., Romeike, R., Gläser-Zikuda, M., Benz Müller, C., & Landgraf, T. (2023). Papagai: Automated Feedback for Reflective Essays. In D. Seipel & A. Steen (Eds.), *Lecture notes in computer science Lecture notes in artificial intelligence: Vol. 14236. Ki 2023: Advances in artificial intelligence: 46th German conference on AI, Berlin, Germany, September 26-29, 2023 : Proceedings* (Vol. 14236, pp. 198–206). Springer. https://doi.org/10.1007/978-3-031-42608-7_16


- Sorge, S., Kröger, J., Petersen, S., & Neumann, K. (2019). Structure and development of pre-service physics teachers' professional knowledge. *International Journal of Science Education*, 28(10), 1–28. <https://doi.org/10.1080/09500693.2017.1346326>
- Sorge, S., Stender, A., & Neumann, K. (2019). The Development of Science Teachers' Professional Competence. In A. Hume, R. Cooper, & A. Borowski (Eds.), *Repositioning Pedagogical Content Knowledge in Teachers' Professional Knowledge*. Springer.
- Sperling, K., Stenberg, C.-J., McGrath, C., Åkerfeldt, A., Heintz, F., & Stenliden, L. (2024). In search of artificial intelligence (AI) literacy in teacher education: A scoping review. *Computers and Education Open*, 6, 100169. <https://doi.org/10.1016/j.caeo.2024.100169>
- Tatto, M. T., Schwille, J., Senk, S., Ingvarson, L., Peck, R., & Rowley, G. (2008). *Teacher Education and Development Study in Mathematics (TEDS-M): Policy, practice, and readiness to teach primary and secondary mathematics. Conceptual framework*. Teacher Education and Development International Study Center, College of Education, Michigan State University.
- Ullmann, T. D. (2019). Automated Analysis of Reflection in Writing: Validating Machine Learning Approaches. *International Journal of Artificial Intelligence in Education*, 29(2), 217–257. <https://doi.org/10.1007/s40593-019-00174-2>
- VanLehn, K. (2011). The Relative Effectiveness of Human Tutoring, Intelligent Tutoring Systems, and Other Tutoring Systems. *Educational Psychologist*, 46(4), 197–221. <https://doi.org/10.1080/00461520.2011.611369>
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, \. u., & Polosukhin, I. (2017). Attention is All you Need. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, & R. Garnett (Eds.), *Advances in Neural Information Processing Systems*. Curran Associates, Inc. <https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf>
- von Aufschnaiter, C., Fraij, A., & Kost, D. (2019). Reflexion und Reflexivität in der Lehrerbildung. Advance online publication. <https://doi.org/10.4119/UNIBI/HLZ-144> (144-159 Seiten / Herausforderung Lehrer_innenbildung - Zeitschrift zur Konzeption, Gestaltung und Diskussion, Bd. 2 Nr. 1 (2019): Herausforderung Lehrer_innenbildung - Ausgabe 2).
- Wahlen, A., Kuhn, C., Zlatkin-Troitschanskaia, O., Gold, C., Zesch, T., & Horbach, A. (2020). Automated Scoring of Teachers' Pedagogical Content Knowledge – A Comparison Between Human and Machine Scoring. *Frontiers in Education*, 5, Article 149. <https://doi.org/10.3389/feduc.2020.00149>
- Wan, T., & Chen, Z. (2024). Exploring Generative AI assisted feedback writing for students' written responses to a physics conceptual question with prompt engineering and few-shot learning. *Physical Review Physics Education Research*.
- Wang, H.-C., Chang, C.-Y., & Li, T.-Y. (2008). Assessing creative problem-solving with automated text grading. *Computers & Education*, 51(4), 1450–1466. <https://doi.org/10.1016/j.compedu.2008.01.006>
- Wei, J. (2022). Emergent Abilities of Large Language Models.
- West, C. G. (2023). AI and the FCI: Can ChatGPT Project an Understanding of Introductory Physics? *ArXiv*.
- Wiedemann, G., Remus, S., Chawla, A., & Biemann, C. (2019). Does BERT Make Any Sense? Interpretable Word Sense Disambiguation with Contextualized Embeddings. *ArXiv*.

- Williams, R., & Grudnoff, L. (2011). Making sense of reflection: a comparison of beginning and experienced teachers' perceptions of reflection for practice. *Reflective Practice*, 12(3), 281–291. <https://doi.org/10.1080/14623943.2011.571861>
- Wolfram, S. (2024). *Can AI Solve Science?* <https://writings.stephenwolfram.com/2024/03/can-ai-solve-science/>
- Wulff, P. (2023). Network analysis of terms in the natural sciences insights from Wikipedia through natural language processing and network analysis. *Education and Information Technologies*, 28, 14325–14346. <https://doi.org/10.1007/s10639-022-11531-5>
- Wulff, P., Buschhüter, D., Nowak, A., Westphal, A., Becker, L., Robalino, H., Stede, M., & Borowski, A. (2021). Computer-Based Classification of Preservice Physics Teachers' Written Reflections. *Journal of Science Education and Technology*, 30, 1–15. <https://doi.org/10.1007/s10956-020-09865-1>
- Wulff, P., Buschhüter, D., Westphal, A., Mientus, L., Nowak, A., & Borowski, A. (2022). Bridging the Gap Between Qualitative and Quantitative Assessment in Science Education Research with Machine Learning — A Case for Pretrained Language Models-Based Clustering. *Journal of Science Education and Technology*, 31, 490–513. <https://doi.org/10.1007/s10956-022-09969-w>
- Wulff, P., Kubsch, M., & Krist, C. (Eds.). (2025). *Springer Texts in Education. Applying Machine Learning in Science Education Research: When, How, and Why?* (1st ed. 2025). Springer Nature Switzerland, Imprint: Springer. <https://doi.org/10.1007/978-3-031-74227-9>
- Wulff, P., Mientus, L., Nowak, A., & Borowski, A. (2022). Utilizing a Pretrained Language Model (BERT) to Classify Preservice Physics Teachers' Written Reflections. *International Journal of Artificial Intelligence in Education*. Advance online publication. <https://doi.org/10.1007/s40593-022-00290-6>
- Wulff, P., Westphal, A., Mientus, L., Nowak, A., & 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. *Frontiers in Education*, 7, 1–18. <https://doi.org/10.3389/educ.2022.1061461>
- Xu, W., & Ouyang, F. (2022). The application of AI technologies in STEM education: a systematic review from 2011 to 2021. *International Journal of STEM Education*, 9(1). <https://doi.org/10.1186/s40594-022-00377-5>
- Yeadon, W., Inyang, O.-O., Mizouri, A., Peach, A., & Testrow, C. P. (2023). The death of the short-form physics essay in the coming AI revolution. *Physics Education*, 58(3), 35027. <https://doi.org/10.1088/1361-6552/acc5cf>
- Zanette, D. (2014). Statistical pattern in written language. *ArXiv*, 1412.3336.
- Zeichner, K. M. (2010). Rethinking the connections between campus courses and field experiences in college- and university-based teacher education. *Journal of Teacher Education*, 61(1-2), 89–99.
- Zhai, X. (2023). ChatGPT User Experience: Implications for Education. *SSRN*.
- Zhai, X., Haudek, K., Shi, L., Nehm, R., & Urban-Lurain, M. (2020). From substitution to redefinition: A framework of machine learning-based science assessment. *Journal of Research in Science Teaching*, 57(9), 1430–1459. <https://doi.org/10.1002/tea.21658>
- Zhai, X., Haudek, K. C., Stuhlsatz, M. A., & Wilson, C. (2020). Evaluation of construct-irrelevant variance yielded by machine and human scoring of a science teacher PCK constructed response assessment. *Studies in Educational Evaluation*, 67, 100916. <https://doi.org/10.1016/j.stueduc.2020.100916>


- Zhai, X., He, P., & Krajcik, J. S. (2022). Applying machine learning to automatically assess scientific models. *Journal of Research in Science Teaching*.
- Zhai, X., Yin, Y., Pellegrino, J. W., Haudek, K. C., & Shi, L. (2020). Applying machine learning in science assessment: a systematic review. *Studies in Science Education*, 56(1), 111–151. <https://doi.org/10.1080/03057267.2020.1735757>
- Zhu, M., Lee, H.-S., Wang, T., Liu, O. L., Belur, V., & Pallant, A. (2017). Investigating the impact of automated feedback on students' scientific argumentation. *International Journal of Science Education*, 39(12), 1648–1668. <https://doi.org/10.1080/09500693.2017.1347303>

Author Information


Peter Wulff

 <https://orcid.org/0000-0002-5471-7977>
Heidelberg University of Education
Germany
Contact e-mail: peter.wulff@ph-heidelberg.de


Lukas Mientus

 <https://orcid.org/0000-0001-5344-4770>
University of Potsdam
Germany

Anna Nowak

 <https://orcid.org/0000-0002-6890-3463>
University of Potsdam
Germany

Andreas Borowski

 <https://orcid.org/0000-0002-9502-0420>
University of Potsdam
Germany
