Rachmatullah *et al.*
Disciplinary and Interdisciplinary Science Education Research (2026) 8:3
https://doi.org/10.1186/s43031-025-00151-x

Disciplinary and Interdisciplinary Science Education Research

RESEARCH Open Access

Exploring large language model's capabilities in identifying science teacher PCK using lesson plans and open-ended questions

Arif Rachmatullah[1*], Shaishav Tayde[2], Nonye Alozie[1], Soonhye Park[3] and Selin Capan[1]

Abstract
Pedagogical content knowledge (PCK) has been a cornerstone of science teacher education research, yet its practical application remains limited because of the non-standardized, time-intensive, and labor-intensive nature of PCK data collection and analysis. This study explores the potential of large language models (LLMs) to identify science teachers' PCK levels on the topic of photosynthesis using open-ended responses and lesson plans. Iterative cycles of training and testing LLMs to assess various PCK components were conducted, introducing an innovative approach that utilized synthetic responses to train the models, which were subsequently validated with actual teacher responses. Findings indicate that synthetic data effectively trained LLMs to identify teacher PCK levels, though performance varied across PCK components. For instance, some models demonstrated strong performance in assessing Knowledge of Instructional Strategies and Representations, as well as Knowledge of Assessment of Science Learning, but struggled with Knowledge of Student Understanding. The study also examined the relationships between teacher characteristics (e.g. self-efficacy, years of experience, and National Board Certification) and PCK levels identified by both humans and LLMs. Results showed some alignment in correlations for particular PCK components, though consistency varied across models. Furthermore, the human-machine reliability for identifying PCK levels from lesson plans approached human-human reliability, with some values exceeding 0.80. These findings highlight the significant potential of LLMs have to advance and scale science teacher PCK research by incorporating multiple data sources. Challenges and opportunities associated with identifying PCK levels using LLMs are discussed, providing implications for future research and science teacher education.

Keywords PCK, Large language model, Lesson plans, Open-ended responses, Photosynthesis

*Correspondence:
Arif Rachmatullah
arif.rachmatullah@sri.com; arifrrach@gmail.com
[1]SRI International, Menlo Park, USA
[2]University of California Irvine, Irvine, USA
[3]North Carolina State University, Raleigh, USA

Introduction

A substantial body of research has demonstrated that teacher quality is one of the most influential factors affecting student learning outcomes (Burroughs et al., 2019; Hattie, 2012). In science education, effective learning goes beyond the acquisition of content knowledge—it encompasses the development of scientific skills, interests, attitudes, and identities (OECD, 2023). This broader conception of science learning has reshaped the criteria for what constitutes a high-quality science teacher. It is no longer sufficient for teachers to possess deep content knowledge or pedagogical skills alone; rather, they must be able to organize, represent, and adapt science content in ways that are responsive to the diverse needs, interests, backgrounds, and skill levels of their students (Park & Oliver, 2008b).

This specialized knowledge base, known as pedagogical content knowledge (PCK; Shulman, 1986, 1987), represents a unique synthesis of content and pedagogy that is central to effective teaching. Shulman (1987) introduced the concept of PCK as a "special amalgam of content and pedagogy that is uniquely the province of teachers" (p. 8). Situated at the intersection of what to teach and how to teach it, PCK enables teachers to anticipate student difficulties, address misconceptions, and adapt instruction to diverse learners (Cochran, 1992; Loughran et al., 2004). Research shows that PCK supports reform-oriented instruction, enhances teaching quality, and improves science learning outcomes (Kulgemeyer & Riese, 2018; Park et al., 2011; Sadler et al., 2013). Reflecting this evidence, many educational reform documents describe PCK as a knowledge base necessary for effective teaching of a subject (e.g., National Research Council[NRC], 2012; National Council for Accreditation of Teacher Education, 2008). Accordingly, PCK has been proposed as a key indicator for evaluating teaching expertise and effectiveness (Park et al., 2018).

Due to its complexity, researchers have often used multiple data sources to capture a teacher's PCK, including artifacts from teaching tasks, lesson observations, interviews, and written tests, surveys, or questionnaires (Chan & Hume, 2019; Großmann & Krüger, 2024; Park & Chen, 2012). However, researchers have relied on just one or two data sources, limiting the comprehensiveness of PCK that can be identified (Chan & Hume, 2019). While using multiple data sources provides a fuller view, they are labor-intensive and difficult to scale (Park & Suh, 2015), which poses constraints on large-scale research and practices that link PCK, teaching practice, and student learning.

In light of these challenges, emerging artificial intelligence (AI) technologies—particularly large language models (LLMs)—present exciting new possibilities for scalable PCK assessment. LLMs, a form of natural language processing (NLP), can generate, classify, and interpret human-like language using vast text corpora (Kasneci et al., 2023).; While LLMs have been increasingly used to assess student learning, their application to evaluating teacher knowledge—especially complex, context-dependent constructs like PCK—remains limited. Early research suggests promise but also highlights critical concerns around scoring reliability, prompt sensitivity, and interpretive alignment between AI and human raters (Zhai, 2020 Yang et al., 2025; Fang et al., 2023).

In this regard, the current study investigates the capacity of LLMs to identify levels of science teachers' PCK through both synthetic and authentic responses. Specifically, we examine the extent to which LLM-generated responses resemble those of actual teachers, the reliability of LLM scoring across PCK components, and the degree of alignment between human and LLM-generated scores. In addition, we evaluate whether PCK levels identified by LLMs exhibit expected correlations with teacher characteristics, including self-efficacy, teaching experience, and advanced licensure, as hypothesized based on existing PCK literature. The research questions guiding this study are:

RQ1 To what extent do LLMs appropriately identify the levels of science teacher PCK in lesson plans and open-ended responses, compared with human labeling?

RQ2 How do the relationships between science teachers' PCK levels identified by LLMs and other teacher characteristics (e.g., self-efficacy, years of experience, and National Board Certification) compare with the relationships between these characteristics and PCK levels identified by human coders?

This study is designed as a *proof of concept* that explores the feasibility of applying LLMs to identify teachers' PCK. Proof-of-concept studies in educational AI often employ modest sample sizes to establish methodological viability before scaling up (e.g., Zhai et al., 2020; Yang et al., 2025). Our dataset–comprising 53 lesson plans and 115 teacher responses–aligns with comparable early-phase AI-in-education and PCK validation studies (e.g., Park et al., 2018; Großmann & Krüger, 2024). These sample sizes are sufficient to demonstrate the feasibility, validity trends, and methodological implications of using LLMs for PCK identification, while also providing a practical foundation for larger, confirmatory studies.

Theoretical background

LLMs in educational assessment: opportunities and challenges

In recent years, LLMs such as GPT-4, LLaMA, and Mistral have gained traction as powerful tools for automating educational assessments (Kasneci et al., 2023). Much

of the research has focused on evaluating student learning, particularly through open-ended tasks like written explanations and reflections (Bonner et al., 2023; Baral et al., 2024; Matelsky et al., 2023). In science education, machine learning models have also been used to assess students' learning as well as scientific modeling abilities among others (Zhai et al., 2020, 2022). These tools show great potential in generating rubric-aligned evaluations, delivering timely feedback, and supporting personalized instruction at scale (Bewersdorff et al., 2025; Capdehourat et al., 2024; Liu & Li, 2024; Mizori et al., 2025).

The efficiency gains are particularly valuable in large-scale contexts, where LLMs can reduce the time spent on grading, summarizing, and generating feedback (Bonner et al., 2023). However, despite their promise, LLM-based assessments and evaluations raise concerns around reliability and fairness. Key challenges include surface-level linguistic interpretations, algorithmic bias, and the generation of hallucinated or misleading content (Zhai & Nehm, 2023; Yan et al., 2023). These limitations pose risks to validity and teacher trust, especially when LLMs are used in high-stakes or professional learning settings (Holstein et al., 2019; Yang et al., 2025). To responsibly integrate LLMs into educational practice, transparency, fairness, and construct validity must be prioritized (Zhai & Krajcik, 2022; Bewersdorff et al., 2025).

While much research has examined the use of LLMs for student assessment, their application to teacher knowledge—and PCK in particular—remains limited. Assessing PCK poses unique challenges: it is context-dependent, dynamic, and expressed through nuanced pedagogical reasoning. Recent studies have begun exploring how LLMs might be adapted for this purpose. Yang et al. (2025), for instance, investigated how GPT-4 performed on video-based constructed-response tasks focused on analyzing student thinking and evaluating teacher responsiveness-two critical sub-constructs of PCK. They found that LLMs matched human raters in reliability and scoring efficiency but also introduced construct-irrelevant variance (CIV), such as prompt sensitivity and rater leniency, which aligns with an earlier study using machine learning model – a traditional supervised model – by Zhai et al. (2020). These findings highlight the need for calibrated scoring logic and interpretive alignment between human and AI scorers (Yang et al., 2025).

Moreover, LLMs behave differently from traditional supervised models. Whereas supervised models tend to be conservative and rigid, LLMs often display greater leniency and are more sensitive to prompt phrasing (Yang et al., 2025; Wu et al., 2025). If left unaddressed, such inconsistencies could distort score interpretations—particularly in domains like PCK that rely on complex, interpretive judgments. Techniques such as rubric conditioning, chain-of-thought prompting, and few-shot learning have shown promise in enhancing the transparency and fidelity of LLM-based assessments (Lee et al., 2024; Fang et al., 2023).

Despite these limitations, emerging literature points to the transformative potential of LLMs in teacher assessment, including generating synthetic training data, modeling pedagogical reasoning, and providing formative feedback (Zhai et al., 2020; Bonner et al., 2023; Fang et al., 2023). Taken together, these developments suggest that with appropriate theoretical grounding and transparency safeguards, LLMs can support scalable assessment of complex constructs like PCK.

Pentagon model of PCK for teaching science

Since Shulman (1986) first introduced the concept of PCK, numerous studies have expanded and elaborated the construct, proposing various PCK models which constitute diverse components (e.g., Carlson et al., 2019; Gess-Newsome, 2015; Magnusson et al., 1999; Park & Oliver, 2008b). However, despite variations in extant PCK models in terms of PCK components and labels or descriptions of these components, the PCK models generally elaborated upon and expanded on Shulman's (1986) two key components by modifying them or adding new components: (1) knowledge of instructional strategies and representations of subject matter that make it comprehensible to students, and (2) knowledge of students' understanding including conceptions, preconceptions, and learning difficulties of that subject matter (van Driel et al., 1998; Chan & Hume, 2019). For instance, Grossman (1990) extended Shulman's model by incorporating two additional elements—teachers' curricular knowledge and their conceptions of the purposes of teaching the subject—alongside expanded definitions of the original components. Magnusson et al. (1999) later refined this model, drawing on Tamir's (1988) framework, by introducing a fifth component: knowledge of assessment. While Grossman's model is intended for general application across subject areas, Magnusson et al.'s (1999) model is discipline-specific, tailored for science education. In their model, PCK consists of five distinct components: (1) orientations toward teaching science (OTS), (2) knowledge of students' understanding of science (KSU), (3) knowledge of instructional strategies and representations (KISR), (4) knowledge of science curricula (KSC), and (5) knowledge of assessment (KAs) related to scientific literacy.

Because of its explicit attention to science teaching, the Magnusson et al.'s (1999) model has become widely adopted in science education research. However, Park and Oliver (2008b) noted a limitation in this model: it treats the five components as separate and fails to highlight the dynamic interplay among them. Research has since shown that the coherence and integration of these

components significantly influence the overall quality of a teacher's PCK (Aydin & Boz, 2013; Barendsen & Henze, 2019; Park & Chen, 2012). To address this, Park and Oliver (2008b) introduced a revised representation—a pentagon model—placing PCK at the center and emphasizing the interrelated nature of the five components. They conceptualized PCK not as a static sum of parts, but as an integrated, synergistic system in which components interact in complex ways. This model was first informed by an extensive literature review and then refined through empirical validation (Park & Oliver, 2008a, 2008b). Their subsequent studies supported that reflection-in-action and reflection-on-action were key mechanisms through which integration occurred, strengthening the internal coherence of PCK as a construct (Park & Oliver, 2008a; Park & Chen, 2012). Given this study's focus on the assessment of PCK in science teachers, we adopt the pentagon model for its strong empirical grounding and relevance to science education. Figure 1 summarizes this conceptual framework, showing how PCK components are interconnected.

To measure PCK in such a dynamic and integrated manner, researchers must first identify and code specific PCK components within each data source, and then analyze how these components connect across sources. This process is typically time-consuming and resource-intensive. In this regard, Park and Suh (2019) suggested that a computer-based analysis using the PCK map approach could help address this issue. Building on their suggestion, this study attempted to reduce the manual effort and time required, by training LLMs that can automatically identify PCK components and assess their levels. Given this focus, this study has great potential to pave the way for AI-based technologies, such as LLMs, to help researchers in scaling up PCK studies. By enabling quick, valid, reliable, and comprehensive identification of PCK components, AI can help explore unresolved research questions related to PCK that have been hindered by the limitations of existing instruments.

Hypothesized relationships between PCK and Teacher characteristics

By adopting the pentagon model, we aim to identify the levels of four PCK components except orientations toward teaching science (OTS) using the LLMs, while assessing OTS separately through the Epistemic Orientations toward Teaching Science for Knowledge Generation (EOTS-KG) survey (Suh et al., 2022). This approach was chosen because OTS reflects broader beliefs about the purposes of science teaching as well as beliefs about teaching and learning in general (Friedrichsen et al., 2011; Magnusson et al., 1999), which are difficult to infer from the types of data used to assess PCK components through LLMs in this study (i.e., lesson plans and open-ended responses). The EOTS-KG survey was deemed appropriate given its strong alignment with the conceptualization of OTS (Suh & Park, 2017; Suh et al., 2022). Specifically, EOTS-KG consists of four interconnected dimensions that shape instructional practices for fostering knowledge generation, falling along a spectrum where beliefs are teacher-centered at one end and learner-centered at the opposite end: (1) epistemic alignment, (2) authority relations in learning, (3) views on the nature of knowledge, and (4) beliefs about student ability (Suh et al., 2022). Epistemic alignment reflects the coherence among a teacher's beliefs about knowledge, learning, and instruction. Authority relations in learning relate to whether knowledge is viewed as being transmitted from an authoritative source or being constructed by the individual themselves. Beliefs about the nature of knowledge refer to whether knowledge is seen as fixed or evolving.

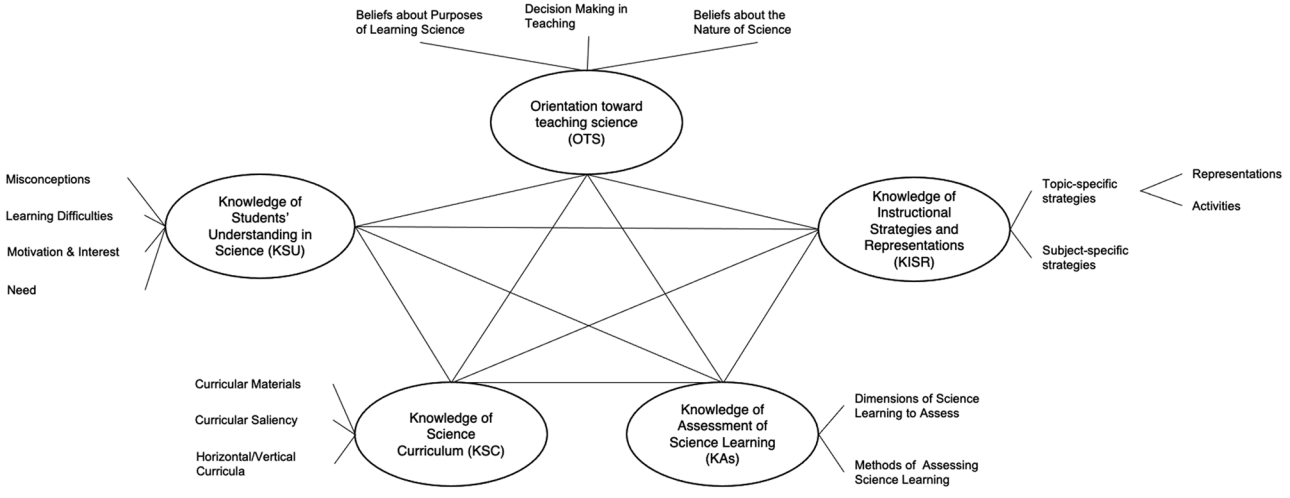

Fig. 1 Conceptual framework of the study based on the Pentagon Model of PCK adapted from Park and Oliver (2008b)

Finally, beliefs about student ability capture views on students' ability to learn.

Although LLMs were not used to assess OTS directly, we tested the hypothesized relationship between OTS, as measured by EOTS-KG, and the other PCK components identified by LLMs. This served two purposes: to validate the EOTS-KG's ability to assess OTS and to examine whether the PCK levels identified by the LLMs align with theoretical expectations drawn from existing PCK literature. Previous studies employing the pentagon model or similar frameworks indicate that OTS is positively associated with other PCK components and overall PCK development (Boesdorfer & Lorsbach, 2014; Demirdöğen, 2016; Faikhamta, 2013; Park & Oliver, 2008b). Therefore, we expect the PCK levels identified by LLMs to be correlated with OTS, at least to the same degree as the human-observed PCK levels. However, Demirdöğen (2016) found that only specific elements of OTS, particularly beliefs about science teaching and learning, as conceptualized as epistemic alignment in Suh et al.'s (2022) framework, were related to KISR. This finding suggests that OTS, in its entirety, may not strongly correlate with KISR in the current study.

One of the teacher characteristics that research has investigated in relation to PCK is self-efficacy for teaching science topics, which Park and Oliver (2008b) included in their hexagon model of science teacher PCK. While Uzuntiryaki-Kondakci et al. (2017) reported qualitative association between self-efficacy for teaching science and PCK, Thomson et al. (2017) found no statistically significant correlation between them. Therefore, we hypothesize that both human-observed and LLM-identified levels of PCK components would not show significant correlations with self-efficacy in teaching science.

In addition, several studies have examined the relationship between PCK and teacher characteristics such as years of science teaching experience and advanced licensure (e.g., National Board Certification). Research has suggested that more experienced teachers tend to have higher levels of PCK (Chan & Yung, 2018; Smit et al., 2017; Van Driel et al., 1998). However, Park et al. (2020) noted that teaching experience alone might not always predict PCK levels. Similarly, advanced licensure, such as National Board Certification or other science subject-specific certifications, has been shown to positively related to PCK development (Park & Oliver, 2008a; Park et al., 2020). Based on these findings, we hypothesize that both human-observed and LLM-identified PCK levels would reflect similar relationships with teaching experience and advanced licensure.

Methods

To address the research questions regarding the feasibility of using LLMs to identify teacher PCK, we used two different data sources: open-ended responses in the PCK assessment and lesson plans. Our focus was on four key components of PCK: KSU, KISR, KSC, and KAs. While synthetic responses were used to train the models for open-ended response before testing and validating them against actual teacher responses, actual lesson plans collected from open-source online platforms and teacher professional online communities were used to train the models for lesson plans. All procedures were reviewed and approved by the SRI International Institutional Review Board (IRB), which determined the study as exempt under educational research criteria.

Developing, testing, and validating the lesson plan scoring model

We gathered a total of 53 lesson plans on the topic of photosynthesis, from various open-source online platforms, including teacher forums, educational resource websites (e.g., http://www.teacherspayteachers.com), and personal teacher webpages. To be included in our dataset, lesson plans had to meet the following criteria: (1) focused on photosynthesis, (2) written in English, (3) downloadable in full text, and (4) explicitly intended for K–12 classroom use. The lesson plans ultimately selected covered a range of grade levels from 3rd through 12th grade. These lesson plans were selected from open-access educational and teacher personal teaching blogs that explicitly allowed public download and reuse for research and educational purposes. All materials were screened to ensure compliance with copyright and licensing norms. We did not restrict our selection to specific grade levels. This decision was grounded in our conceptualization of PCK, which emphasizes the interconnectedness of its components rather than the specific depth or breadth of the scientific content taught. While prior research has underscored the role of content knowledge in shaping PCK (e.g., Kirschner et al., 2016; Sorge et al., 2019;), we argue that content coverage required for particular grade-level expectations is sufficient to evaluate PCK quality using the pentagon model, where the focus lies in how PCK components relate to and reinforce one another across contexts.

To evaluate the quality of these lesson plans, we used the validated rubrics developed by Großmann and Krüger (2024) based on the pentagon model of PCK. The authors developed the rubrics based on the assumption that teachers' levels of PCK are correlated with the quality of lesson plans. The rubrics break down each component of the PCK pentagon model into 24 finer grained subcomponents and are evaluated on a scale of 1 (not present), 2 (sufficient), and 3 (excellent), making them

adaptable for measuring the quality of lesson plans. In addition, the rubrics were validated using lesson plans created by in- and pre-service teachers, with interrater reliability values ranging from 0.57–1.00. However, we found that five components specific to pre-service teachers' lesson plans are not suitable for our collected lesson plans written by in-service science teachers. For example, *progression throughout a unit* subcomponent refers to how teachers identify the way multiple lesson plans within a unit build upon each other to achieve the unit's goals. We found that in-service teachers typically do not include such a section, as they tend to create lesson plans one by one. In addition, given that out current work is a proof of concept we focus more on certain key subcomponents of each PCK component.

Consequently, only nine subcomponents, as shown in Table 1, were used to provide a proof of concept of models to identify PCK using lesson plans. Two researchers independently coded 53 collected lesson plans using the modified rubrics. Another researcher was invited to calculate the reliability between the two coders using Gwet AC1 (Gwet, 2008). Then the three researchers met to discuss and resolve discrepancies until reaching a consensus.

To develop each automated scoring model to assess the quality of the same lesson plans for each selected subcomponent, given the limited number of 53 lesson plans, we started with using zero-shot learning. In this approach, no examples were provided to the LLMs, and the Explanation-Prediction (E-P) approach (Huang et al., 2023) was applied to assess initial model performance. Then, we transitioned to few-shot learning, using examples generated by the initial zero-shot learning model. These examples served as a foundation for helping the model grasp the complexities involved in scoring such content. The prompt engineering approach we used in this stage is a combination of two aspects: chain of thought and heuristics in a hierarchical decision-making structure. This combination made the automated scoring model establish a clear understanding of the ranking criteria coupled with the examples provided to make its decision-making more accurate (Amatriain, 2024). Finally, the trained automated scoring models were used to score the remaining lesson plans on the nine selected subcomponents. Additionally, to assess the performance of the model relative to human labeling, we used Gwet's AC1 with quadratic weighting (Gwet, 2008) for both human-human and machine-human reliability.

Developing, testing, and validating models for identifying PCK using open-ended questions

Phase 1. Training LLMs using synthetic data

We developed a PCK measure by adapting the PCK assessment developed and validated by Park et al. (2018) that aimed to measure teacher PCK for teaching photosynthesis. While Park et al.'s (2018) assessment focused on only two components of PCK – KSU and KISR – we added the Knowledge of Assessment (KAs) component to our measure to expand its scope. Our initial PCK measure utilizes six text-based classroom scenarios used in

Table 1 Selected PCK subcomponents from Großmann and Krüger's (2024) rubrics

PCK Component	PCK Subcomponent	Subcomponent Code	Description
KSC	Intended learning outcome in the lesson	O3	Evaluates how well the learning outcome is stated addressing what students will achieve or improve as a result of the lesson.
	Choice of topic	C2	Evaluates the relevance and justification for selecting the lesson topic to foster the intended competencies or learning outcomes by considering its alignment with curriculum demands, societal relevance, and educational principles.
KSU	Learning difficulties	S4	Assesses whether potential challenges (e.g., language barriers or topic-specific challenges) students may face during the lesson are identified.
	Methodical skills	S5	Assesses whether students' skills (e.g., science and engineering practices) needed to achieve the learning outcome are identified.
KISR	Structure of the development of competencies	I1	Evaluates whether the lesson is designed to build competencies progressively through its phases, with clear connections between tasks and learning objectives throughout all parts of the lesson.
	Lesson structure	I2	Evaluates the organization of the lesson, specifically the logical sequencing of phases and instructional steps, and whether they contribute effectively to achieving the learning outcomes.
	Suitability of the methods	I3	Evaluates whether the teaching methods and activities selected for the lesson are appropriate for achieving the intended content or process-related learning outcomes.
KAs	Transparency performance expectations	A1	Assesses whether and how clearly the performance expectations for the intended learning outcomes (content, form, and scope) are communicated to students, including the extent to which relevant and specific criteria are shared to guide students in understanding what is expected.
	Products of students' learning	A2	Evaluates how the lesson concludes with the presentation and discussion of students' learning products.

Park et al. (2018), in which students demonstrate alternative conceptions and learning difficulties related to photosynthesis. With each scenario, respondents are prompted to answer six open-ended questions related to the three target PCK components, two questions for each component. This initial measure was distributed via email to approximately 5,000 teachers across various U.S. states. Despite offering monetary compensation and conducting active recruitment efforts, only 18 complete responses returned, primarily due to the length and cognitive demands of the PCK questions. To address this challenge and increase teacher participation, we decided to streamline the measure by focusing on a single scenario with six questions, as shown in Fig. 2.

Considering the aforementioned low return rate, we followed the UK Department for Education (2024) recommendation to use synthetic data to train and test LLMs, especially when education data are limited. Synthetic data refer to responses generated by LLMs when prompted to answer the questions in the PCK measure. Aa a result, we generated a wide range of synthetic responses with LLMs such as Llama, Mistral, and Zephyr to provide diverse examples for the model to learn from. Figure 3 illustrates the process used to train LLMs to generate high-quality synthetic responses.

Specifically, to obtain synthetic data, we used the scenario and questions from the final PCK measure as inputs and leveraged pretrained LLMs for response synthesis. The prompts for the LLMs were developed iteratively: We manually refined the prompt template, analyzed the LLM outputs, and adjusted the prompts to better resemble real teacher responses. Each prompt sequence included a system prompt that defined the teacher's grade level, years of experience, and teaching context, followed by user prompts presenting the classroom scenario and open-ended PCK questions. Four researchers evaluated whether the generated responses were AI-like, and we further engineered the prompt template to improve the quality of the synthetic data. This iterative process involved revising phrasing, specificity, and examples to reduce generic or overly polished outputs and to encourage realistic, varied reasoning patterns. To promote transparency and replicability, we provide representative system and user prompts, as well as sample model outputs and associated PCK rubric scores, in Appendix A. The prompts were systematically varied to reflect different levels of teaching experience and to elicit diverse reasoning patterns aligned with the PCK components. We acknowledge that despite iterative refinement, prompts may still embed assumptions or biases that influence model responses, a limitation discussed further in the Discussion section.

A total of 188 sets of synthetic data, comprising 1,128 responses to six questions, were generated and manually coded using the scoring rubric exemplified in Fig. 2. We then trained and tested transformer-based Bidirectional Encoder Representations from Transformers (BERT) models using the 80:20 split to identify the sophistication level of each PCK component in each response. We used 80% labeled data to train the model and 20% to validate the models. BERT models were selected for their ability to adapt semantics to specific application contexts (Rogers et al., 2021). We trained four BERT models—RoBERTa, BERTxnli, EduBERT, and FineWeb—to identify the PCK levels of the generated synthetic response data on a scale from 1 to 3. Descriptions of these models

Scenario: Ms. Hammer was wrapping up the lesson on photosynthesis by writing the following summary equation on the board:

$$6\,CO_2 + 6\,H_2O \xrightarrow{\text{Sunlight, Chlorophyll}} 6\,O_2 + C_6H_{12}O_6\ \text{(glucose)}$$

Ms. Hammer initiated a discussion focusing on the equation to gauge students' understanding of photosynthesis. Below is an excerpt of their discussion.

April: Sunlight is necessary to warm the plants to do photosynthesis. During photosynthesis, plants absorb heat energy from the sun.

Bill: That's correct. This process occurs in the green pigments within chloroplasts. Then, when there is no light, heat energy is converted into chemical energy, producing glucose.

Cindy: Glucose, along with other nutrients, is also absorbed through the roots from the soil. They serve as food, helping the plants to grow.

Assessment Question Category	PCK Construct
KSU1. Identifying students' conceptual challenges KSU2. Reasoning about why students have such challenges	Knowledge of Student Understanding (KSU)
KISR1. Proposing instructional approaches KISR2. Reasoning about the proposed instructional approaches	Knowledge of Instructional Strategies and Representations (KISR)
KA1. Proposing assessment methods and timing, and reasoning about the proposed assessments KA2. Proposing criteria for assessing students' understanding	Knowledge of Science Assessments (KAs)

Assessment Question Category	PCK Construct
Identifying students' conceptual challenges	KSU1

Level	Rubric Description	Assigned Score
	Missing or responses like "See above" or "Not applicable".	M
	AI-like responses/responses that don't like real teachers' responses	99
Limited	Identifying **ONE** unique target misconception/conceptual challenge OR ONLY general/irrelevant misconception.	1
Moderate	Identifying **TWO** unique target misconceptions/conceptual challenges.	2
Sufficient	Identifying **at least THREE** unique target misconceptions/conceptual challenges.	3

Sample Response: Based on the explanations provided by the students, it seems that April may have a misconception that sunlight is necessary to warm the plants in order to do photosynthesis, while Bill and Cindy may have difficulty understanding the role of glucose as a source of energy for plant growth.

Fig. 2 Scenario and questions in the PCK measure along with an example rubric

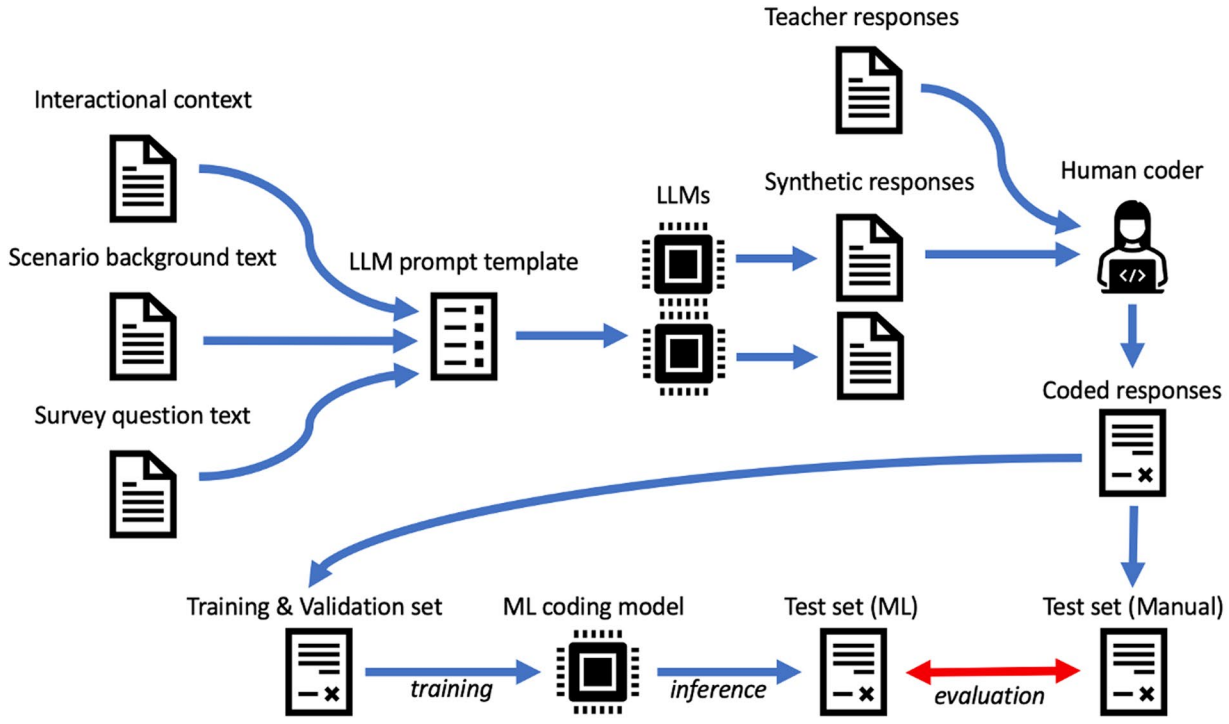

Fig. 3 Process for generating synthetic data and training the automated scoring model

Table 2 Trained and tested LLMs

LLM	Description
Robustly Optimized BERT Pretraining Approach (RoBERTa)	An enhanced variant of BERT that improves the pretraining process by using a larger dataset and skipping a task called Next Sentence Prediction, leading to better performance on NLP benchmarks (Liu, 2019).
BERTxnli	A version of BERT fine-tuned for natural language inference (NLI), designed to handle multilingual and cross-lingual understanding by training on language pairs, making it suitable for translation and inference tasks (Conneau et al., 2018).
EduBERT	A specialized BERT model tailored for educational contexts, focusing on understanding education data and facilitating tasks such as automated feedback generation and interpreting student responses in assessments (Clavié & Gal, 2019).
FineWeb	An LLM refined specifically for extracting information from web data, with an emphasis on understanding webpages and web-specific content, making it valuable for information retrieval, summarization, and content analysis (Penedo et al., 2024).

are provided in Table 2. We selected these transformer-based BERT models due to their strong contextual understanding, established reliability in automated scoring tasks, and adaptability to small, domain-specific datasets. Their diverse pretraining corpora, ranging from educational text (EduBERT) to multilingual inference data (BERTxnli), allowed comparative evaluation of model architectures for PCK analysis. To assess the performance of these models, we used Gwet's AC1 with quadratic weighting for both human-human and machine-human reliability on the synthetic data. Discrepancies among human coders were discussed and resolved before using the labels to train and test the models.

Phase 2. Testing LLMs pre-trained with synthetic data using actual teacher responses

After training and validating the models on synthetic data, we tested their performance on actual teacher response data. The PCK measure consisting of one scenario and six questions, used to train the models, were distributed via email and 115 U.S. science teachers completed the PCK measure as well as other surveys (i.e., self-efficacy in teaching photosynthesis, epistemic orientation toward teaching science, and demographic questions) used in Phase 3. The collected responses were analyzed to determine whether the four LLMs trained on synthetic data could reliably identify teachers' PCK based on their actual responses to open-ended questions. To evaluate human-machine reliability on the actual teacher responses, we calculated Gwet's AC1 with quadratic weighting. For PCK components with low human-machine reliability (<0.70), we conducted a thematic analysis (Saldaña, 2021) to identify patterns and reasons behind the models' misalignment with human observations. Two researchers who had previously coded and scored the responses independently examined separate sets of responses where discrepancies between human

and machine scores occurred. Each researcher identified potential reasons for the misalignment, focusing on differences in interpretation and scoring. The researchers then met to discuss their findings and used a selective coding process to generate overarching themes. Although we did not conduct double-coding, we used collaborative discussion to support the trustworthiness of the findings.

Phase 3. Validating the LLMs generated PCK levels of actual teacher responses through correlation

To validate the machine-generated PCK levels and answer the second research question, we conducted Pearson and Spearman correlation tests. These tests explored the relationships between the PCK levels identified by the LLMs and the aforementioned teacher characteristics such as self-efficacy in teaching photosynthesis, years of science teaching experience, and National Board Certification. We then compared these relationships to those established using PCK levels identified by human evaluators. We modified the T-STEM science scale validated by Unfried et al. (2022) by changing "science" in the item wording to "photosynthesis and cellular respiration" to measure teachers' self-efficacy in teaching photosynthesis. This scale consisted of eight 5-point Likert-type items with Cronbach's alpha of 0.97. To measure teachers' OTS, we used forty-four 5-point Likert items developed and validated by Suh et al. (2022) with Cronbach's alpha of 0.81. Regarding the demographic information of the 115 teachers, the average years of science teaching experience was 15.46 ($SD = 9.21$), with a range from 0 to 42 years. Among the participants, 61% identified as female, 36% as male, and 3% preferred not to disclose their gender. Additionally, 15% of the teachers held National Board Certification. The majority of teachers identified as White (70%), followed by Asian (11%), Hispanic (9%), and Black/African American (4%). Regarding grade level taught, 63% of the participants were high school teachers, 18% were middle school teachers, 4% were elementary school teachers, and 4% were either K–12 teachers or did not provide this information.

Table 3 Comparison of human-human and human-machine reliability of lesson plans subcomponent scoring (PA = percent agreement)

PCK Component	Subcomponent Code	Human-Human		Human-Machine	
		PA (%)	AC1	PA (%)	AC1
KSC	O3	93.63	0.870	97.5	0.947
	C2	94.83	0.894	90.9	0.750
KSU	S4	91.38	0.749	97.0	0.927
	S5	90.52	0.839	96.1	0.929
KISR	I1	87.93	0.752	88.4	0.743
	I2	96.55	0.954	97.4	0.968
	I3	91.38	0.847	91.8	0.856
KAs	A1	98.30	0.966	98.0	0.960
	A2	93.63	0.870	96.1	0.924

Findings

Evaluating LLMs' accuracy in identifying science teacher PCK levels compared with human labeling (RQ1)

To address the first research question, we report findings from human-human and human-machine reliability analyses for both lesson plans and open-ended responses. Additionally, we present thematic analyses of the patterns and contexts where the models, pretrained on synthetic data, failed to accurately identify levels of PCK when tested with actual teacher responses.

Lesson plans

As illustrated in Table 3, most human-machine reliability values across all components surpassed the human-human reliability values. This was particularly evident for the KSU subcomponents, where human-human reliability values were below 0.90, but human-machine reliability values exceeded 0.90. These results indicate that the models effectively assessed the quality of potential students' learning difficulties (S4) and necessary methodological skills students need to learn (S5) as described in the lesson plans.

Similarly, for all subcomponents of KAs, human-machine reliability values were also above 0.90, suggesting that the models performed well in evaluating both the transparency of performance expectations and the products of students' learning as presented by teachers in the lesson plans. However, for KSC, while the model outperformed human-human reliability on O2 (assessing the quality of learning objectives in the lesson plans), the model for C2 performed poorly (C2: evaluating the relevance and justification for selecting the lesson topic to foster intended competencies and learning outcomes, including its alignment with curriculum demands, societal relevance, and educational principles).

Additionally, for KISR, the models performed as well as human coders on I2 and I3. These components assess the logical sequencing of phases and instructional steps, their effectiveness in achieving learning outcomes, and the appropriateness of teaching methods and activities for the intended content or process-related learning outcomes. However, the models struggled with I1, which evaluates whether the lesson is designed to build competencies progressively through its phases, with clear connections between tasks and learning objectives throughout the lesson.

These findings suggest that while the models excel at analyzing explicit criteria and assessing certain aspects of lesson plans, they face challenges in identifying reasoning patterns and connections between tasks and learning objectives throughout all phases of the lesson in the

Table 4 Comparison of human-human and human-machine reliability values for synthetic responses

Question	Human-Human		FineWeb		EduBERT		BERTxnli		RoBERTa	
	PA(%)	AC1	PA(%)	AC1	PA(%)	AC1	PA(%)	AC1	PA(%)	AC1
KSU1	92.42	0.807	86.11	0.692	89.81	0.748	73.15	0.284	88.89	0.724
KSU2	91.08	0.808	88.46	0.755	84.62	0.669	87.50	0.745	85.58	0.706
KISR1	92.19	0.766	93.52	0.820	89.81	0.738	89.81	0.734	86.11	0.594
KISR2	94.46	0.919	94.00	0.887	95.00	0.927	91.00	0.814	92.00	0.818
KAs1	94.87	0.875	94.23	0.844	97.12	0.931	97.12	0.931	89.42	0.736
KAs2	90.16	0.705	87.96	0.644	90.74	0.737	86.11	0.589	91.67	0.755

Table 5 Comparison of human-machine reliability values for actual Teacher responses

Question	FineWeb		EduBERT		BERTxnli		RoBERTa	
	PA(%)	AC1	PA(%)	AC1	PA(%)	AC1	PA(%)	AC1
KSU1	83.11	0.511	75.00	0.302	74.34	0.268	82.45	0.499
KSU2	79.60	0.566	78.95	0.564	81.14	0.565	83.11	0.620
KISR1	88.27	0.712	87.17	0.679	85.18	0.620	89.82	0.721
KISR2	96.90	0.967	97.57	0.974	96.46	0.958	95.80	0.954
KAs1	93.58	0.917	94.69	0.932	95.35	0.940	76.11	0.607
KAs2	94.03	0.914	92.70	0.901	92.48	0.891	93.81	0.899

lesson plans. This may indicate limitations in the model's ability to process nuanced pedagogical reasoning and coherence in instructional design.

Open-ended responses: training and testing using synthetic responses

Table 4 summarizes the results of testing the pretrained models with synthetic data and comparing their performance with human-human reliability. As indicated in the table, most reliability values among human coders ranged from 0.705 to 0.919, indicating acceptable to satisfactory levels (DeVellis & Thorpe, 2021). For the first question (KSU1) in the PCK measure, which asks to identify students' conceptual challenges with photosynthesis, EduBERT achieved the highest human-machine reliability, closely matching human-human reliability. This was followed by RoBERTa and FineWeb, which demonstrated reliability values similar to EduBERT. By contrast, BERTxnli performed poorly on this question.

For the second question (KSU2) in the PCK measure, which focuses on identifying the reasons behind the conceptual challenges identified in KSU1, EduBERT performed worse than the other models. In this case, the other models achieved reliability values above 0.70, with FineWeb emerging as the best-performing model among the four. Interestingly, FineWeb also performed the best—and even outperformed human coders—on the third question (KISR1), which asks respondents to describe pedagogical approaches to help students overcome the challenges identified in KSU1. FineWeb achieved a reliability value above 0.80, while EduBERT and BERTxnli produced values closer to human-human reliability. However, for the fourth question (KISR2), which prompts respondents to justify the rationale behind the pedagogical approaches mentioned in KISR1, EduBERT performed the best among all models, achieving a human-machine reliability value of 0.927, surpassing the human-human reliability value of 0.919. The other models also performed well on this question, with reliability values exceeding 0.80. Notably, when coding the synthetic responses for KISR2, we observed a significant imbalance in the variability of levels, with responses heavily concentrated at Levels 1 and 2, and only 4% categorized as Level 3.

For the final two questions measuring knowledge of science learning assessments (KAs), EduBERT and BERTxnli performed the best on KAs1, which focuses on identifying types of assessments suitable for measuring students' understanding of photosynthesis. Both models surpassed the human-human reliability value. On KAs2, which examines criteria for determining whether students still face conceptual challenges, EduBERT and RoBERTa performed the best, achieving reliability values above 0.70 and surpassing human-human reliability.

These findings underscore the importance of addressing and resolving discrepancies among human coders before using their labeled synthetic responses to train the LLMs. In several instances, the models outperformed human coders when labeling synthetic responses, highlighting the potential of LLMs to provide reliable evaluations for specific PCK components.

Open-ended responses: LLM performance using actual teacher responses

When the pretrained and tested LLMs with synthetic data were used to identify PCK levels in actual teacher responses, as shown in Table 5, all four models performed well overall, with human-machine reliability

values mostly above or around 0.90 for KISR2, KAs1, and KAs2. However, RoBERTa underperformed on KAs1 compared with the other models.

Interestingly, the models performed significantly worse on KSU1 and KSU2 when tested on actual teacher responses compared with their performance on synthetic data, where reliability values were above 0.65 and approached 0.70. With actual responses, reliability values for KSU1 and KSU2 mostly ranged between 0.50 and 0.60. Notably, EduBERT and BERTxnli performed the worst on KSU1, with reliability values of only 0.302 and 0.268, respectively. Similarly, the models' performance on KISR1 was lower than when tested with synthetic data, with the exception of RoBERTa, which performed better when using synthetic data. Based on these results, FineWeb emerged as the most reliable and stable model across the different components.

Our qualitative analysis of the responses and patterns of levels assigned by the LLMs for KSU1, KSU2, and KISR1 identified specific themes for each of the models, especially for where the models failed. Table 6 provides specific examples of challenges encountered by each model for these three questions.

For KSU1, three key themes emerged, highlighting challenges most models faced. The first challenge was the models' tendency to assign higher levels of PCK for

Table 6 Challenges faced by each llm for questions KSU1, KSU2, and KISR1

FineWeb	EDUBERT	BERTxnli	RoBERTa
KSU1			
Evaluated responses with incomplete or ungrammatical sentences poorly, even if the conceptual challenges were correct.	Over-scored general responses using correct terminology, even when lacking detail.	Evaluated generic mentions of concepts (e.g., "the role of X") as correct conceptual challenges, against expectations.	Often rated vague phrases like "the role of X" or incomplete sentences higher than expected.
Relied heavily on common terminology from the training set, missing unconventional phrasing.	Showed inconsistency when scoring sentences without explicit mentions of misconceptions or correct conceptions.	Struggled with nuanced differences between concepts (e.g., light vs. heat energy) and occasionally undercounted or overcounted misconceptions.	Miscounted multiple misconceptions described in one sentence or similar language used for different scenarios.
Overcounted conceptual challenges when phrasing was reiterated and misclassified unrelated incorrect conceptual challenges as relevant.	Counted reiterations of the same misconception as separate instances.	Failed to evaluate correct conceptual challenges not explicitly framed in terms of conceptual challenges.	Struggled with technical terminology or linking less obvious related terms to relevant misconceptions, leading to score inconsistencies.
KSU2			
Had difficulty identifying abstract or complex ideas unless explicitly stated.	Struggled with indirect references to abstract or complex ideas.	Over-scored general phrases like "this can be hard for students to understand" without deeper explanation.	Tended to score phrases and incomplete sentences as 1, which worked better for B than A but still caused inconsistencies.
Required detailed articulation for responses about students' personal experiences or background information to be scored correctly.	Had issues recognizing less formal or indirect mentions of student-centered reasons unless explicitly connected to the question.	Struggled to consistently flag valid reasons unless they were part of a complete sentence using subject-specific terminology.	Most sensitive to grammatical structure among the models, leading to difficulty with indirect allusions or unstructured responses.
Missed responses that alluded to complexity or abstraction without directly using key terms.	Was overly reliant on grammatical structure and photosynthesis terminology for accurate scoring.	Rarely recognized student-centered reasons unless explicitly linked to a misconception or concept.	Scored general instructional challenges, unrelated to target misconceptions, as valid reasons, occasionally inflating scores.
KISR1			
Struggled to identify strategies described in sentences where misconceptions and activities were in separate clauses or indirectly connected.	Focused on identifying clear strategies through action verbs (e.g., "we'll do a lab activity"), but often missed those described without explicit labels.	Over-scored responses that listed multiple explanations or questions, counting each as a separate strategy, even if not detailed.	Failed to recognize experiments or activities without explicit labels or details.
Tended to count multiple misconception-activity links as separate strategies, inflating scores.	Scored responses with detailed terminology higher, even if the strategy was not explicitly linked to misconceptions.	Frequently missed experiment descriptions that lacked explicit labeling or detailed explanation.	Over-scored responses that mentioned multiple misconceptions or vague references without specific strategies.
Missed strategies described with less common terminology or phrased indirectly.	Missed indirect references to activities or experiments more often than other models.	Tended to give higher scores to responses using common terminology, even if the response lacked depth.	Sentence structure played a significant role in scoring, with grammatically correct responses receiving disproportionately higher scores.

vague mentions or terminology without detailed explanations. For example, BERTxnli identified general conceptual phrases such as "the role of sunlight" as a conceptual challenge. However, according to the rubric and the labeled training data, such general mentions of concepts should not be considered conceptual challenges. Additionally, the models struggled to differentiate between nuanced concepts such as light and heat energy, leading to occasional undercounts or overcounts of conceptual challenges. The second challenge involved reiteration and redundancy. Conceptual challenges phrased in different ways were sometimes counted multiple times, such as "Some misconceptions I notices is that students think that heat is part of the process and that the plant need to heat up," counted as having two conceptual challenges instead of one. In contrast, closely related challenges with similar phrasing but occurring in different contexts were often counted as a single mention of conceptual challenge. This inconsistency frequently resulted in a lower level of KSU1 being assigned. The third challenge involved terminology recognition. The models inconsistently linked related terms (e.g., associating 'nutrients' and 'food molecules' with glucose or 'ATP' with energy misconceptions), leading to inaccuracies in PCK identification. This issue may have stemmed from the synthetic training data, as responses using less common phrasing or terminology—such as 'nutrients' and 'food molecules'—were often overlooked.

For KSU2, three major themes were identified in relation to the models' struggles in evaluating teacher responses accurately. The first issue was associated with key terms and phrasing. The models had difficulty identifying key terms such as "abstract," "multi-step process," and "hard to visualize" unless these terms were explicitly stated in clear, grammatical sentences. Indirect references or allusions to these ideas (e.g., "not intuitive to think about the world from a chemical perspective") were often missed, despite the rubric and training with synthetic data emphasizing that such reasons should be counted as valid. The second challenge involved recognizing student-centered reasons. The models inconsistently flagged reasons related to students' personal experiences or counted oversimplified background knowledge as valid, which led to instability in their evaluations. Finally, the third challenge pertained to sentence structure sensitivity. Responses written as incomplete or ungrammatical phrases were frequently rated as Level 1, regardless of the validity and relevance of their content.

For KISR1, similar themes to those found in KSU1 and KSU2 became apparent, particularly regarding sentence structure sensitivity and terminology. The models tended to assign lower levels to incomplete or ungrammatical sentences, even when valid strategies were described. Conversely, grammatically correct but vague or overly general responses were sometimes assigned higher levels. The models also heavily relied on specific terminology from the training dataset/synthetic data (e.g., "sunlight," "glucose") and often overlooked responses that used less common or indirect terms (e.g., "the product of photosynthesis"). A unique theme identified in KISR1 was related to the recognition of certain pedagogical approaches, particularly those involving experiments, demonstrations, or activities. All models struggled to recognize these strategies unless explicitly labeled as "experiment," "demo," or "activity." This challenge was especially evident in responses that lacked complete sentences or additional details.

For KISR2, another major theme surfaced. The models struggled to recognize a single strategy that addressed multiple conceptual challenges simultaneously. Instead, the models often evaluated these responses as representing multiple valid strategies, assigning them a higher level, despite the rubric indicating that such responses should be counted as one strategy dealing with multiple conceptual challenges.

Comparing relationships between teachers' PCK levels identified by LLMs and humans with key teacher characteristics (RQ2)

To test correlations between the PCK levels identified by LLMs and humans, respectively, and selected teacher characteristics, scores from two sub-questions measuring each PCK component were combined to generate a composite mean score for each component. As shown in Table 7, the human-generated KSU score was positively and significantly correlated ($p < 0.05$) with EOTS-KG as a measure of OTS and years of teaching experience, but it was not significantly correlated with self-efficacy for teaching photosynthesis or having National Board Certification. Similarly, KSU scores generated by FineWeb showed positive and significant correlations with EOTS-KG and years of science teaching experience, with correlation coefficients closely matching those of the human-generated scores. However, KSU scores from other models did not show significant correlations with these constructs.

Regarding KISR, we found that the human-generated KISR score was positively and significantly correlated with years of science teaching experience and having National Board Certification, but it was not correlated with self-efficacy for teaching photosynthesis or EOTS-KG. FineWeb-generated KISR scores also showed positive and significant correlations with years of science teaching experience and National Board Certification, with correlation coefficients and p-values closely aligning with those of human-generated scores. While scores from EduBERT and RoBERTa were also significantly correlated with these variables, their correlation coefficients

Table 7 Results of correlation tests

PCK Component	LLM	Self-Efficacy for Teaching Photosynthesis		Epistemic Orientation Toward Teaching Science-knowledge generation (EOTS-KG)		Years of Science Teaching Experience		Having National Board Certification*	
		r(113)	*p*	*r*(113)	*p*	*r*(113)	*p*	*r*(113)	*p*
KSU	**Human**	**−0.051**	**0.592**	**0.199**	**0.034**	**0.246**	**0.008**	**0.044**	**0.644**
	EduBERT	0.122	0.193	0.033	0.729	−0.029	0.755	0.137	0.144
	RoBERTa	−0.192	0.040	0.174	0.064	0.088	0.350	−0.042	0.659
	FineWeb	0.059	0.529	0.262	0.005	0.197	0.035	−0.004	0.970
	BERTxnli	0.083	0.381	0.059	0.534	0.064	0.499	0.031	0.744
KISR	**Human**	**0.048**	**0.610**	**0.166**	**0.080**	**0.229**	**0.015**	**0.259**	**0.006**
	EduBERT	0.214	0.021	0.113	0.233	0.180	0.055	0.290	0.002
	RoBERTa	0.143	0.126	0.083	0.382	0.174	0.062	0.190	0.042
	FineWeb	0.125	0.182	0.109	0.248	0.230	0.014	0.264	0.004
	BERTxnli	0.074	0.432	−0.002	0.981	0.065	0.492	0.165	0.079
KAs	**Human**	**−0.093**	**0.327**	**0.208**	**0.028**	**0.066**	**0.491**	**0.063**	**0.509**
	EduBERT	0.094	0.317	−0.013	0.891	0.048	0.610	0.052	0.580
	RoBERTa	0.003	0.973	0.105	0.265	0.119	0.204	0.159	0.089
	FineWeb	0.010	0.917	0.135	0.153	0.084	0.373	0.124	0.187
	BERTxnli	0.035	0.714	0.183	0.051	0.064	0.500	0.109	0.246

Note: *Spearman correlation test

and *p*-values were less aligned with the human-generated scores compared to FineWeb.

For KAs scores, the human-generated score was positively and significantly correlated only with EOTS-KG and not significantly correlated with self-efficacy for teaching photosynthesis, years of science teaching experience, or National Board Certification. None of the models replicated this result, as the correlations between LLM-generated KAs scores and EOTS-KG were all nonsignificant. Among the nonsignificant correlations, EduBERT, FineWeb, and BERTxnli produced correlation coefficients and *p*-values that were closer to the human-generated scores for years of science teaching experience. Additionally, EduBERT showed correlation coefficients and *p*-values closer to human-generated scores for the correlation with National Board Certification. Taken together, these results suggest that no single LLM should be used to identify and evaluate science teachers' PCK based on open-ended responses. Instead, these findings indicate that a system combining multiple models as an ensemble may be more effective for evaluating science teachers' PCK.

Discussion

This study represents an initial effort to explore the capabilities of LLMs in identifying levels of science teachers' PCK. Specifically, we aimed to develop, train, test, and validate LLMs to assess the levels of PCK components using lesson plans and open-ended responses to PCK-related questions. While we made progress in leveraging LLMs to identify PCK levels, we also identified areas for improvement.

This proof-of-concept study contributes to two interrelated fields. First, it advances PCK research by demonstrating a scalable and multi-source method for assessing teachers' knowledge using AI-based text analysis. Second, it extends the AI-in-education literature by applying LLMs to a complex, context-dependent construct—teacher PCK—showing both their promise and their current limitations. Together, these contributions highlight the potential of LLMs not only as analytic tools but also as catalysts for rethinking how teacher knowledge can be assessed and supported at scale. The following sections discuss these contributions in greater detail, moving from specific findings about reliability and data sources to broader implications for modeling, integration, and ethics.

Reliability and construct alignment in lesson plan scoring

In developing and testing LLMs to identify PCK levels using lesson plans, we achieved relatively high reliability between human and machine assessments. However, challenges arose in achieving high reliability for identifying reasoning patterns and establishing clear connections between tasks and learning objectives throughout all phases of the lesson as described in the lesson plans. These challenges were anticipated, as we did not employ more advanced models, such as BERT, capable of processing nuanced pedagogical reasoning and complex textual patterns. Even with advanced models like BERT, studies such as Mondorf and Plank (2024) have shown that LLMs often rely on surface-level patterns rather than genuine reasoning. This reinforces prior concerns in the literature about AI's difficulty in modeling context-sensitive reasoning (Zhai et al., 2020; Zhai & Nehm, 2023).

Additionally, the limited number of lesson plans available for training and testing ($n = 53$) further constrained model development. This highlights a broader issue in teacher education research, where collecting diverse teacher artifacts remains challenging. Despite these limitations, our models performed well in assessing explicit, structured criteria in lesson plans. These findings establish a foundation for exploring how LLMs can capture more interpretive, coherence-related features of instructional design, providing a transition to the next focus—how different data sources and training strategies influence model performance.

Integrating synthetic and authentic data for PCK assessment

Assessing PCK is cognitively demanding, which became evident when we initially used the Park et al. (2018) assessment comprising six scenarios. Many teachers dropped out or failed to complete the assessment because of its length. Consequently, we opted to use a single scenario with more targeted questions assessing various PCK components. Synthetic data, as recommended by the UK Department for Education (2024), proved helpful for training AI models to identify PCK levels in educational settings.

However, challenges remained, including imbalanced representation of PCK components in the generated data, even after prompt engineering. This indicates the need for more advanced prompt-engineering techniques to train models effectively to generate more diverse synthetic data. While models trained on synthetic data performed well, they struggled when tested on actual teacher responses. For instance, EduBERT, which performed best on synthetic data, showed decreased accuracy with actual responses. This aligns with concerns raised in prior work about generalizability and depth of LLM-generated content (Bonner et al., 2023; Fang et al., 2023), which emphasizes the importance of pairing synthetic data with authentic, context-rich responses during model development—particularly when assessing constructs as nuanced as PCK.

Taken together, the lesson-plan and open-response analyses reveal a consistent theme: LLMs are most effective when dealing with explicit, structured indicators of pedagogical quality but less reliable when interpreting implicit reasoning. The next section explores this challenge in greater depth by examining how the interconnected nature of PCK components complicates automated modeling.

Modeling pedagogical reasoning and interconnected PCK components

An important insight from this study is that LLMs performed unevenly across PCK components, with notably lower reliability for Knowledge of Student Understanding (KSU) compared to KISR and KAs. This variability may not be merely a technical limitation but may reflect deeper differences in the cognitive demands and linguistic visibility of different PCK dimensions. KSU involves two core sub-tasks: (1) identifying students' conceptual challenges and (2) explaining the possible sources or reasons behind those misconceptions. Our findings showed that LLMs struggled most with the first task—accurately identifying misconceptions—when working with actual teacher responses, especially when responses included vague phrasing or lacked standard terminology. This aligns with prior research on conceptual nuance showing that LLMs are limited in their ability to reason beyond surface-level textual cues (Zhai & Nehm, 2023; Mondorf & Plank, 2024).

The second KSU task, which requires inferring underlying causes of misconceptions (e.g., abstraction or prior experience), was somewhat more reliably scored, though still below the accuracy observed for other components like KAs. These findings suggest that certain types of teacher reasoning—especially those that are implicit, nuanced, or inferential—remain difficult for LLMs to detect reliably.

This difficulty also raises broader questions about how well discrete, rubric-based scoring approaches capture the inherently dynamic and integrated nature of PCK. The pentagon model (Park & Oliver, 2008b) conceptualizes PCK components as mutually reinforcing and interconnected. However, operational scoring often treats each component as an independent category, which can obscure interactions such as when teachers justify instructional strategies based on their understanding of student thinking. This critique parallels issues raised by Yang et al. (2025) in video-based PCK assessments, where LLMs misinterpreted instructional decisions when separated from the underlying student reasoning.

Recognizing these challenges points directly to the need for more integrated modeling frameworks, discussed next, that combine complementary models and data sources to capture the multidimensional nature of PCK more accurately.

Toward ensemble and integrated modeling approaches

While FineWeb stood out as the most consistent model in replicating human-generated scoring patterns, no single LLM performed equally well across all PCK components. This finding highlights the need for ensemble approaches that combine the strengths of multiple models. For instance, a weighted-voting ensemble (Mienye & Sun, 2022) could assign greater weight to models that perform well on specific components like KSU, while a stacking approach could use a meta-model to optimize predictions across models (Polikar, 2012). Research in

ensemble modeling has shown that such systems can lower error rates and improve stability (Zhou, 2021).

Building on the observed variation in model performance, ensemble or hybrid systems may offer a practical pathway toward representing PCK as an integrated construct, aligning better with the theoretical coherence emphasized in the pentagon model. Future studies should test ensemble configurations for PCK scoring and explore integrating lesson-plan and open-response models to create unified teacher-knowledge profiles. These integrated approaches naturally raise questions of scale, ethics, and interpretability—issues addressed in the following section.

Limitations and future directions

Although this proof-of-concept study demonstrates encouraging outcomes, several limitations should be acknowledged. First, even after iterative prompt engineering, the responses generated by LLMs may embed subtle biases or lack the cognitive depth found in real teacher writing. LLMs also tend to overfit to surface-level linguistic cues derived from scoring rubrics (Mondorf & Plank, 2024), which can hinder their ability to detect deeper pedagogical reasoning. Second, our dataset—53 lesson plans and 115 open-ended responses—limits generalizability. While sufficient for an initial proof of concept, larger and more diverse datasets are needed to support model development across instructional contexts, teaching styles, and grade levels.

Third, ethical and validity-related concerns remain. The risk of algorithmic bias, especially if training data reflect narrow instructional norms, raises questions about equity and fairness (Holstein et al., 2019; Popenici & Kerr, 2017). Scoring transparency also remains a challenge, as LLM-generated assessments may lack interpretability for teachers or evaluators. Most importantly, teacher agency could be undermined if such tools are used in high-stakes or evaluative contexts without teacher input, critical oversight, or alignment with local educational values. Moving forward, future research should expand datasets, incorporate multimodal evidence (e.g., classroom video and audio), and experiment with ensemble and hybrid human–AI scoring systems. These efforts will strengthen construct validity, fairness, and interpretability, paving the way for responsible and equitable use of AI in teacher assessment.

Conclusion

In summary, this study provides initial empirical evidence that LLMs can identify elements of science teachers' PCK with promising levels of reliability–particularly for explicit, structured aspects of instructional design and assessment–while also revealing persistent challenges in modeling nuanced pedagogical reasoning. By integrating synthetic and authentic data, comparing multiple model architectures, and aligning findings with the pentagon model of PCK, this work demonstrates both the potential and the boundaries of current LLMs in capturing teacher PCK.

The models developed in this study could evolve into technological tools enabling science teachers to reflect on their PCK and enhance their instructional practices. Such tools would also benefit science teacher educators and professional development providers by identifying areas for PCK growth, allowing for more targeted programs that address teachers' diverse needs and dispositions (Bae et al., 2020; Rachmatullah et al., 2023). Furthermore, these tools can support researchers in analyzing large-scale PCK data, advancing our understanding of PCK and its applications in the classroom. Ultimately, this proof-of-concept investigation underscores that the intersection of PCK research and AI has the potential to transform how teacher PCK is studied, assessed, and supported. Continued collaboration between science education researchers and AI developers will be essential for realizing this vision responsibly and effectively.

Supplementary information

The online version contains supplementary material available at https://doi.org/10.1186/s43031-025-00151-x.

Supplementary Material 1

Acknowledgements

This study was funded by SRI International through its internal research and development (IRAD) program. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of SRI International. We also thank John Niekrasz, Marta Mielicki, Hui Yang, Sophia Ouyang, Elica Sharifnia, and all teachers who supported the project.

Author contributions

AR directed the project, analyzed the data, and drafted the initial version of the manuscript. ST trained and ran the LLMs and wrote the method sections on this. NA co-directed the project, conceptualized the study, and reviewed the manuscript. SP developed the assessment used in the study, wrote the theoretical framework section, and reviewed the manuscript. SC analyzed the qualitative data and reviewed the manuscript.

Funding

This study was funded by SRI International through its internal research and development (IRAD) program.

Data availability

The datasets used and/or analyzed during the current study are available from the corresponding author on reasonable request.

Declarations

Ethics approval and consent to participate

Ethics approval for this study was reviewed and waived by the Institutional Review Board (IRB) at SRI International, as the study met the criteria for exemption. Informed consent was obtained from all participants in accordance with institutional guidelines.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Received: 11 July 2025 / Accepted: 7 November 2025
Published online: 14 January 2026

References

Amatriain, X. (2024). Prompt design and engineering: Introduction and advanced methods. arXiv. https://doi.org/10.48550/arXiv.2401.14423

Aydin, S., & Boz, Y. (2013). The nature of integration among PCK components: A case study of two experienced chemistry teachers. *Chemistry Education Research and Practice*, *14*(4), 615–624.

Bae, C. L., Hayes, K. N., & DeBusk-Lane, M. (2020). Profiles of middle school science teachers: Accounting for cognitive and motivational characteristics. *Journal of Research in Science Teaching*, *57*(6), 911–942.

Baral, S., Worden, E., Lim, W. C., Luo, Z., Santorelli, C., Gurung, A.,& Heffernan, N. (2024). Automated feedback in math education: A comparative analysis of LLMs for open-ended responses. arXiv preprint arXiv:2411.08910.

Barendsen, E., & Henze, I. (2019). Relating teacher PCK and teacher practice using classroom observation. *Research in Science Education*, *49*(5), 1141–1175. https://doi.org/10.1007/s11165-017-9637-z

Bewersdorff, A., Hartmann, C., Hornberger, M., Seßler, K., Bannert, M., Kasneci, E.,& Nerdel,C (2025). Taking the next step with generative artificial intelligence: The transformative role of multimodal large language models in science education. *Learning and Individual Differences*, *118*, 102601. https://doi.org/10.1016/j.lindif.2024.102601

Boesdorfer, S., & Lorsbach, A. (2014). PCK in action: Examining one chemistry teacher's practice through the lens of her orientation toward science teaching. *International Journal of Science Education*, *36*(13), 2111–2132. https://doi.org/10.1080/09500693.2014.909959

Bonner, E., Lege, R., & Frazier, E. (2023). Large language model-based artificial intelligence in the language classroom: Practical ideas for teaching. *Teaching English with Technology*, *23*(1), 23–41.

Burroughs, N., Gardner, J., Lee, Y., Guo, S., Touitou, I., Jansen,& K., Schmidt, W. (2019). A review of the literature on teacher effectiveness and student outcomes. In *Teaching for excellence and equity: Analyzing teacher characteristics, behaviors and student outcomes with TIMSS* (pp. 7–17). Springer.

Capdehourat, G., Lorenzo, B., Porteiro, P.,& Trigo, J. (2024, November). Leveraging LLMs for K-12 education assessment: An open-ended question answering use case. In *2024 IEEE URUCON* (pp. 1–4). IEEE. https://doi.org/10.1109/URUCON63440.2024.10850479

Carlson, J., Daehler, K. R., Alonzo, A., Barendsen, E., Berry, A., Borowski, A., Carpendale, J., Chan, K. K. H., Cooper, R. Friedrichsen, P., Gess-Newsome, J., Henze-Rietveld, I., Hume,A., Kirschner, S., Liepertz, S., Loughran, J., Mavhunga, E., Neumann, K., Nilsson, P.,… Wilson, C. (2019). The refined consensus Model of pedagogical content knowledge in science education. In A. Hume, R. Cooper, A. Borowski (Eds.), *Repositioning pedagogical content knowledge in teachers' knowledge for teaching science* (pp. 77–92). Springer.

Chan, K. K. H., & Hume, A. (2019). Towards a consensus model: Literature review of how science teachers' pedagogical content knowledge is investigated in empirical studies. In A. Hume, R. Cooper, & A. Borowski (Eds.), *Repositioning pedagogical content knowledge in teachers' knowledge for teaching science* (pp. 3–76). Springer.

Chan, K. K. H., & Yung, B. H. W. (2018). Developing pedagogical content knowledge for teaching a new topic: More than teaching experience and subject matter knowledge. *Research in Science Education*, *48*, 233–265. https://doi.org/10.1007/s11165-016-9567-1

Clavié, B., & Gal, K. (2019). *EduBERT: Pretrained deep language models for learning analytics*. arXiv. https://doi.org/10.48550/arXiv.1912.00690

Cochran, K. F. (1992). Pedagogical content knowledge: Teachers' transformations of subject matter (Monograph No. 5). *National Association for Research in Science Teaching. Research Matters … to the Science Teacher*.

Conneau, A., Lample, G., Rinott, R., Williams, A., Bowman, S. R., Schwenk, H., & Stoyanov, V. (2018). *XNLI: Evaluating cross-lingual sentence representations*. arXiv. https://doi.org/10.48550/arXiv.1809.05053

Demirdöğen, B. (2016). Interaction between science teaching orientation and pedagogical content knowledge components. *Journal of Science Teacher Education*, *27*(5), 495–532. https://doi.org/10.1007/s10972-016-9472-5

DeVellis, R. F., & Thorpe, C. T. (2021). *Scale development: Theory and applications*. Sage publications.

Faikhamta, C. (2013). The development of in-service science teachers' understandings of and orientations to teaching the nature of science within a PCK-based nos course. *Research in Science Education*, *43*, 847–869. https://doi.org/10.1007/s11165-012-9283-4

Fang, L., Lee, G. G., & Zhai, X. (2023). Using GPT-4 to augment unbalanced data for automatic scoring. arXiv preprint arXiv:2310.18365.

Friedrichsen, P., van Driel, J. H., & Abell, S. K. (2011). Taking a closer look at science teaching orientations. *Science Education*, *95*(2), 358–376.

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.), *Re-examining pedagogical content knowledge in science education* (pp. 28–42). Routledge.

Grossman, P. L. (1990). *The making of a teacher: Teacher knowledge and teacher education*. Teachers College Press.

Großmann, L., & Krüger, D. (2024). Assessing the quality of science teachers' lesson plans: Evaluation and application of a novel instrument. *Science Education*, *108*(1), 153–189. https://doi.org/10.1002/sce.21832

Gwet, K. L. (2008). Computing inter-rater reliability and its variance in the presence of high agreement. *The British Journal of Mathematical and Statistical Psychology*, *61*(1), 29–48. https://doi.org/10.1348/000711006X126600

Hattie, J. (2012). *Visible learning for teachers: Maximizing impact on learning*. Routledge.

Holstein, K., Wortman Vaughan, J., Daumé III, H., Dudik, M.,& Wallach, H. (2019, May). Improving fairness in machine learning systems: What do industry practitioners need? In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (pp. 1–16).

Huang, S., Mamidanna, S., Jangam, S., Zhou, Y., & Gilpin, L. H. (2023). Can large language models explain themselves? A study of LLM-generated self-explanations. arXiv. https://doi.org/10.48550/arXiv.2310.11207

Kasneci, E., Seßler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., Gassner, U., Groh, G., Günnemann, S., Hüllermeier, E., Krusche, S., Kutyiok, G., Michaeli, T., Nerdel, C., Pfeffer, J., Poquet, O., Sailer, M., Schmidt, A., Seidel, Seidel, T.,…Kasneci, G. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and Individual Differences*, *103*, Article 102274. https://doi.org/10.1016/j.lindif.2023.102274

Kirschner, S., Borowski, A., Fischer, H. E., Gess-Newsome, J., & von Aufschnaiter, C. (2016). Developing and evaluating a paper-and-pencil test to assess components of physics teachers' pedagogical content knowledge. *International Journal of Science Education*, *38*(8), 1343–1372.

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

Lee, G. G., Latif, E., Wu, X., Liu, N., & Zhai, X. (2024). Applying large language models and chain-of-thought for automatic scoring. *Computers and Education: Artificial Intelligence*, *6*, 100213. https://doi.org/10.1016/j.caeai.2024.100213

Liu, Y. (2019). *RoBerta: A robustly optimized BERT pretraining approach*. arXivarXiv. https://doi.org/10.48550/.1907.11692

Liu, Z., & Li, M. (2024). Large language model for assignment feedback on open-ended subjective questions. *International Educational Research*, *7*(3), 1–10. https://doi.org/10.30560/ier.v7n3p1

Loughran, J., Mulhall, P.,& Berry, A. (2004). In search of pedagogical content knowledge in science: Developing ways of articulating and documenting professional practice. *Journal of Research in Science Teaching*, *41*(4), 370–391. https://doi.org/10.1002/tea.20007

Magnusson, S., Krajcik, L., & Borko, H. (1999). Nature, sources and development of pedagogical content knowledge. In J. Gess-Newsome & N. G. Lederman (Eds.), *Examining pedagogical content knowledge* (pp. 95–132). Kluwer Academic Publishers.

Matelsky, J. K., Parodi, F., Liu, T., Lange, R. D.,& Kording, K. P. (2023). A large language model-assisted education tool to provide feedback on open-ended responses. arXiv preprint arXiv:2308.02439.

Mienye, I. D., & Sun, Y. (2022). A survey of ensemble learning: Concepts, algorithms, applications, and prospects. *IEEE Access*, *10*, 99129–99149. https://doi.org/10.1109/ACCESS.2022.3207287

Mizori, R., Sadiq, M., Ahmad, M. T., Siu, A., Ahmad, R. R., Yang, Z.,… Galloway, J(2025). Stem exam performance: Open-versus closed-book methods in the large

language model era. *The Clinical Teacher*, *22*(1), e13839. https://doi.org/10.1111/tct.13839
Mondorf, P., & Plank, B. (2024). Beyond accuracy: Evaluating the reasoning behavior of large language models–a survey. arXiv preprint arXiv:2404.01869.
National Council for Accreditation of Teacher Education. (2008). *Professional standards for the accreditation of teacher preparation institutions*. ERIC Clearinghouse.
National Research Council. (2012). *A framework for K-12 science education: Practices, crosscutting concepts, and core ideas*. National Academies Press.
OECD. (2023). *PISA 2025 science framework (draft)*. https://pisa-framework.oecd.org/science-2025
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. https://doi.org/10.1002/tea.21022
Park, S., Choi, A., & Reynolds, W. M. (2020). Cross-national investigation of teachers' pedagogical content knowledge (PCK) in the US and South Korea: What proxy measures of teacher quality are related to PCK? *International Journal of Science Education*, *42*(15), 2630–2651. https://doi.org/10.1080/09500693.2020.1823046
Park, S., Jang, J. Y., Chen, Y. C., & Jung, J. (2011). Is pedagogical content knowledge (PCK) necessary for reformed science teaching?: Evidence from an empirical study. *Research in Science Education*, *41*, 245–260. https://doi.org/10.1007/s11165-009-9163-8
Park, S., & Oliver, J. S. (2008a). National Board Certification (NBC) as a catalyst for teachers' learning about teaching: The effects of the NBC process on candidate teachers' PCK development. *Journal of Research in Science Teaching*, *45*(7), 812–834. https://doi.org/10.1002/tea.20234
Park, S., & Oliver, J. S. (2008b). Revisiting the conceptualisation of pedagogical content knowledge (PCK): PCK as a conceptual tool to understand teachers as professionals. *Research in Science Education*, *38*, 261–284. https://doi.org/10.1007/s11165-007-9049-6
Park, S., & Suh, J. (2015). Trajectory from portraying toward assessing PCK: Drives, dilemmas, and directions for future research. In A. Berry, P. Friedrichsen, & J. Loughran (Eds.), *Re-examining pedagogical content knowledge in science education* (pp. 104–119). Routledge.
Park, S., & Suh, J. (2019). The PCK map approach to capturing the complexity of enacted PCK (ePCK) and pedagogical reasoning. In A. Hume, R. Cooper, & A. Boroswski (Eds.), *Repositioning pedagogical content knowledge in teachers' professional knowledge* (pp. 185–197). Springer.
Park, S., Suh, J., & Seo, K. (2018). Development and validation of measures of secondary science teachers' PCK for teaching photosynthesis. *Research in Science Education*, *48*, 549–573. https://doi.org/10.1007/s11165-016-9578-y
Penedo, G., Kydlíček, H., Lozhkov, A., Mitchell, M., Raffel, C., Von Werra, L., & Wolf, T. (2024). The FineWeb datasets: Decanting the web for the finest text data at scale. arXiv. https://doi.org/10.48550/arXiv.2406.17557
Polikar, R. (2012). *Ensemble learning. Ensemble machine learning: Methods and applications*. *Cham*: *Springer*.
Popenici, S. A., & Kerr, S. (2017). Exploring the impact of artificial intelligence on teaching and learning in higher education. *Research and Practice in Technology Enhanced Learning*, *12*(1), 1–13. https://doi.org/10.1186/s41039-017-0062-8
Rachmatullah, A., Hinckle, M., & Wiebe, E. N. (2023). The role of teachers' self-efficacy beliefs and habits in differentiating types of K–12 science teachers. *Research in Science Education*, *53*(2), 337–355.
Rogers, A., Kovaleva, O., & Rumshisky, A. (2021). A primer in BERTology: What we know about how BERT works. *Transactions of the Association for Computational Linguistics*, *8*, 842–866.
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
Saldaña, J. (2021). *The coding manual for qualitative researchers (3rd ed.)*. Sage Publications.
Shulman, L. (1987). Knowledge and teaching: Foundations of the new reform. *Harvard Educational Review*, *57*(1), 1–23.
Shulman, L. S. (1986). Those who understand: Knowledge growth in teaching. *Educational Researcher*, *15*(2), 4–14.
Smit, R., Weitzel, H., Blank, R., Rietz, F., Tardent, J., & Robin, N. (2017). Interplay of secondary pre-service teacher content knowledge (ck), pedagogical content knowledge (PCK) and attitudes regarding scientific inquiry teaching within teacher training. *Research in Science & Technological Education*, *35*(4), 477–499. https://doi.org/10.1080/02635143.2017.1353962
Sorge, S., Kröger, J., Petersen, S., & Neumann, K. (2019). Structure and development of pre-service physics teachers' professional knowledge. *International Journal of Science Education*, *41*(7), 862–889. https://doi.org/10.1080/09500693.2017.1346326
Suh, J., & Park, S. (2017). Exploring the relationship between pedagogical content knowledge (PCK) and sustainability of an innovative science teaching approach. *Teaching & Teacher Education*, *64*, 246–259.
Suh, J. K., Hwang, J., Park, S., & Hand, B. (2022). Epistemic orientation toward teaching science for knowledge generation: Conceptualization and validation of the construct. *Journal of Research in Science Teaching*, *59*(9), 1651–1691. https://doi.org/10.1002/tea.21769
Tamir, P. (1988). Subject matter and related pedagogical knowledge in teacher education. *Teaching and Teacher Education*, *4*(2), 99–110. https://doi.org/10.1016/0742-051X(88)90011-X
Thomson, M. M., DiFrancesca, D., Carrier, S., & Lee, C. (2017). Teaching efficacy: Exploring relationships between mathematics and science self-efficacy beliefs, PCK and domain knowledge among preservice teachers from the United States. *Teacher Development*, *21*(1), 1–20. https://doi.org/10.1080/13664530.2016.1204355
UK Department for Education. (2024). Use cases for Generative AI in education. Building a proof of concept for Generative AI feedback and resource generation in education contexts: Technical report. https://assets.publishing.service.gov.uk/media/66cdb0cc239c5e6b4dc0533a/Use_cases_for_generative_AI_in_education_technical_report.pdf
Unfried, A., Rachmatullah, A., Alexander, A., & Wiebe, E. (2022). An alternative to STEBI-A: Validation of the T-STEM science scale. *International Journal of STEM Education*, *9*, Article 24. https://doi.org/10.1186/s40594-022-00339-x
Uzuntiryaki-Kondakci, E., Demirdöğen, B., Akın, F. N., Tarkin, A., & Aydın-Günbatar, S. (2017). Exploring the complexity of teaching: The interaction between teacher self-regulation and pedagogical content knowledge. *Chemistry Education Research and Practice*, *18*(1), 250–270.
van Driel, J. H., Verloop, N.,& De Vos, W. (1998). Developing science teachers' pedagogical content knowledge. *Journal of Research in Science Teaching*, *35*(6), 673–695. https://doi.org/10.1002/(SICI)1098-2736(199808)35:6%3C673:AID-TEA5%3E3.0.CO;2-J
Wu, X., Saraf, P. P., Lee, G., Latif, E., Liu, N., & Zhai, X. (2025). Unveiling scoring processes: Dissecting the differences between LLMs and human graders in automatic scoring. *Technology, Knowledge, and Learning*, 1–16. https://doi.org/10.1007/s10758-025-09836-8
Yan, D., Fauss, M., Hao, J.,& Cui, W. (2023). Detection of AI-generated essays in writing assessments. *Psychological Test and Assessment Modeling*, *65*(1), 125–144
Yang, Y., Wang, S., & Zhai, X. (2025). Using large language models to assess teachers' pedagogical content knowledge. In *Proceedings of the International Conference on AI in Education (Workshop)* (pp. 1–14), Palermo, Italy.
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*, Article 100916. https://doi.org/10.1016/j.stueduc.2020.100916
Zhai, X., He, P., & Krajcik, J. (2022). Applying machine learning to automatically assess scientific models. *Journal of Research in Science Teaching*, *59*(10), 1765–1794. https://doi.org/10.1002/tea.21773
Zhai, X., & Krajcik, J. (2022). Pseudo ai bias. arXiv preprint arXiv:2210.08141.
Zhai, X., & Nehm, R. H. (2023). Ai and formative assessment: The train has left the station. *Journal of Research in Science Teaching*, *60*(6), 1390–1398.
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
Zhou, Z. H. (2021). *Machine learning*. Springer nature.

Publisher's Note