Education Tech Research Dev
https://doi.org/10.1007/s11423-025-10545-3

DEVELOPMENT ARTICLE

Discovering technology-aided possibilities for automatic analysis of science teacher questions

Sami Lehesvuori[1,2,3] · **Felipe Urrutia**[4] · **Ville Heilala**[5] · **Roberto Araya**[4] · **Raija Hämäläinen**[1]

Accepted: 19 July 2025
© The Author(s) 2025

Abstract

The research focus on teacher questions is justified by previous research emphasizing the essential role of questions in facilitating meaningful learning in science. Analysis of teacher questions has traditionally been based mainly on manual coding, which is extremely labour intensive. In this study we explore how both machine learning and large language models can be used for this purpose. Whereas machine learning approaches involve supervised training with extensive data, pre-trained large language models operate through prompt engineering. The automatic speech recognition text outputs of 23 physics lessons on the same topic from 23 science teachers were analysed with variety of techniques. The results revealed that the large language model approaches improved with few-shot approaches compared to zero-shot ones. Furthermore, few shot approaches outperformed the supervised machine learning approaches, yet human- and hand-crafted approaches continue to demonstrate their relevance. Implications for science teaching and learning are discussed.

✉ Sami Lehesvuori
sami.lehesvuori@jyu.fi

Felipe Urrutia
furrutia@dim.uchile.cl

Ville Heilala
ville.s.heilala@jyu.fi

Roberto Araya
roberto.araya.schulz@gmail.com

Raija Hämäläinen
raija.h.hamalainen@jyu.fi

[1] The Department of Education, University of Jyväskylä, P.O. Box 35, 40014 Jyväskylä, Finland

[2] Centre for Multilingual Academic Communication, University of Jyväskylä, Jyväskylä, Finland

[3] The Department of Science and Technology Education, University of Johannesburg, Auckland Park, P.O. Box 524, Johannesburg 2006, South Africa

[4] Center for Advanced Research on Education, Institute of Education, University of Chile, Santiago, Chile

[5] Faculty of Humanities and Social Sciences, University of Jyväskylä, P.O. Box 35, 40014 Jyväskylä, Finland

Published online: 29 July 2025

Keywords Teacher questions in science · Large language models · Automatic analysis

Introduction

Student learning in science is enhanced through classroom interaction, with teacher questioning playing a key role in building knowledge, scaffolding understanding, and sustaining engagement (Chin, 2007; Eshach et al., 2013; Scott et al., Scott et al., 2006). However, systematically analyzing teacher questioning and classroom talk at scale poses a significant challenge, largely due to the complexity of classroom interactions and the time-consuming nature of manual coding processes.

In response to these challenges, researchers, practitioners, and professional development leaders are increasingly exploring modern and complementary technology-aided methods for analysing and developing learning interactions (Foster et al., 2024; Hennessy et al., 2020). The labour-intensive nature of traditional methods has necessitated more structured approaches in large-scale studies, often resulting in the loss of relevant information that could be detected through more fine-grained and unstructured methods typically applied to smaller datasets.

Utilizing conversational technologies (e.g., Wu et al., 2025; Yusuf et al., 2025) like pedagogical agents (e.g., Sikström et al., 2022) and digital tutors (e.g., Hobert & Berens, 2024) in authentic classroom settings necessitates that the agents are capable of analyzing spoken human interaction. However, pre-trained large language models (LLMs), developed through advances in natural language processing (NLP) and trained primarily on written language, often face performance challenges when applied to natural spoken data (e.g., Umari et al., 2024). Thus, a multi-method approach was adopted in this study to compare the performance of LLMs with supervised machine learning (ML) approaches and the human-annotated baseline. In this study, LLMs are considered a specific subset of ML, characterized by their pre-training on large-scale text corpora and their capacity to perform prompt-based tasks with little or no additional training. In contrast, traditional ML approaches typically rely on task-specific models trained on labeled examples and guided by features or rules designed for more narrowly defined domains.

The rapid development of AI, particularly through advancements in LLMs, entails enormous possibilities and challenges for education and educational research (Heilala et al., 2025; Kasneci et al., 2023). This study aims to advance the automated analysis of learning interactions by demonstrating how machine learning and AI-supported approaches can be applied to the analysis of teacher questioning. The role instruction in science, and especially teacher talk, play in the learning interactions has been addressed in quantitative studies (Fischer et al., 2014a, 2014b), but the nature of this connection has not yet been sufficiently addressed. This is mainly due to the complexity of the problems related to learning interactions. The technological development can save and direct human resources for solving more sophisticated problems, yet the complexity requires careful human validation of the steps on the way to automatization (Maity et al., 2024).

Traditional supervised machine learning systems improve incrementally through exposure to labelled data, whereas LLMs, pre-trained on vast corpora, demonstrate strong zero- and few-shot generalization capabilities (Kojima et al., 2022). In this study, the ML model is trained on a focused, limited dataset specifically designed for detecting teacher questions. This offers a valuable opportunity to assess its task-specific effectiveness in comparison to the broad, general-purpose knowledge embedded in pre-trained LLMs.

A further challenge arises from the modality of the data. LLMs are predominantly trained and tested on written language (Lee et al., 2024; Umari et al., 2024), which differs substantially from spoken discourse in terms of structure, fluency, and contextual ambiguity. Therefore, their application to transcribed classroom interactions requires careful consideration (Dent & Paul, 2020; Peng et al., 2023). This study tests whether LLMs can effectively interpret authentic, spoken educational data, which is an area that remains underexplored despite its growing relevance in learning analytics.

In general, state-of-the-art technology-supported analysis of interaction has addressed various aspects, such as question authenticity (Jensen et al., 2020; Kelly et al., 2018), instructional activities (Donnelly et al., 2016; Lv et al., 2021, Wang et al., Wang et al., 2023), student roles in online group discussions (Dowell et al., 2019), and emotional support and positive climate (Nojavanasghari et al., 2016). However, there is still limited scholarly work harnessing the technology to study classroom interactions namely in science. Our previous work discussing the use of technology to analyse the content of science teacher talk showed that student learning is linked to how scientific concepts are connected to each other in time (Schlotterbeck et al., 2020). However, while analysis was merely based on the automatic detection of physics content keywords, the role of teacher questions was not addressed. That is, the focus on discourse functions rather than content remained to be explored.

This study investigates the extent to which teacher questions can be automatically detected using machine learning (ML) and pre-trained large language models (LLMs). The focus is placed on teacher questions that support students' meaning making of the taught topic (Mortimer & Scott, 2003). These include both open questions, which invite elaboration and multiple perspectives, and closed questions, which typically aim at brief, factual responses (Chin, 2007). In general, teacher questions in this study are defined as utterances intended to elicit student thinking, check understanding, or guide students in subject-specific learning activities.

By comparing different technology-aided approaches, we aim to evaluate their effectiveness in identifying teacher questioning in classroom discourse. While ML has been applied to various educational phenomena, including teacher questions (Donnelly et al., 2017), limited research has examined how pre-trained LLMs perform in this task compared to models trained on domain-specific data (Fussell et al., 2024). By bridging this gap, the study offers insights into how such technologies can be leveraged to support science teaching and learning through scalable, automated analysis of instructional interactions.

The aim of this study is twofold: first, to explore to what extent AI techniques can be used to detect teacher questions; and second, to examine how these approaches and techniques perform in achieving this goal. While there is extensive research on the manual coding and analysis of science teacher questions (Chin, 2007; Eshach et al., 2013), this study takes an initial step toward more automatic analysis by focusing on the detection of teachers' questions in physics lessons. Before introducing methods, an overview of the relevant theory on teacher questions and technology-aided approaches to detect them are presented.

Theoretical background

Drawing on teacher questions in science

The most fundamental form identified in educational dialogue is the initiation-response-feedback (IRF) pattern of classroom talk (Sinclair & Coulthard, 1975). This triadic pattern,

commonly orchestrated by a teacher and related to a question–answer routine, has dominated interactions for decades (Gall, 1970), particularly in science education (Lemke, 1990; Mercer et al., 2009). Both open and closed questions play a significant role in meaningful learning of science. Open questions promote and stimulate students' participation, leading to more dialogic and extended exchange patterns through which the teacher can obtain information about students' thinking (Chin, 2007; Scott et al., 2006). Meanwhile, closed questions can be effective in specific situations, such as diagnostic questioning that can be used for assessing students' prior knowledge (Ruiz-Primo, 2011). Similarly, authentic questions can be assimilated with open questions to avoid answers being predefined by the teachers (Nystrand et al., 1998). Contrastingly, inauthentic questions are merely based memorization. As brought up, both types of questions hold in the potentiality for enhancing more interactive pedagogy (Wang et al., 2023).

That being said, enhancing the predominant IRF pattern with strategies that foster openness and critical thinking can enhance dialogue among students (Gillies, 2016; Littleton & Mercer, 2013; Sohmer, et al., 2009). Nevertheless, the IRF pattern remains integral to educational discourse, offering a foundational framework adaptable to various instructional objectives. Addressing this, it is crucial for teacher education across disciplines, including science, to grasp the existence of teacher questions. Automatic detection of teacher questions could serve as an effective tool for teachers to reflect on their management of learning interactions (Donnelly et al., 2017).

Indeed, the effectiveness of orchestrating science content and teacher-led learning interactions can be assessed by detecting key elements such as teacher questioning (Chin, 2007). High-quality instructional dialogue relies significantly on science teachers' questioning strategies, particularly for initiating student verbal participation and fostering scientific reasoning (Chin, 2007) and inquiry (Oliveira, 2010, Gillie, 2023; van Zee & Minstrell, 1997).

The use of technology in the analysis of classroom interactions

The automatic detection of teacher questions has primarily relied on ML approaches requiring extensive annotated data for training (Blanchard et al., 2016; Kelly et al., 2018). The previous work includes traditional methods such as open-vocabulary approaches using n-grams and Random Forest classifiers (see Table 1) as well as more advanced techniques like deep transfer learning for natural language processing (Jensen et al., 2020). Acoustic-prosodic features, which are characteristics of speech such as pitch, tone, and rhythm, have also been used with and without linguistic features to analyze spoken language (Donnelly et al., 2017). This study focuses solely on automatic speech recognition (ASR) outputs of teacher talk. Preliminary results from technology-aided teacher question detection are promising (Donnelly et al., 2017) but fully integrating AI remains the next step. Given the limited availability of authentic classroom data for ML training, leveraging pre-trained LLMs is crucial to advancing the analysis of teacher questions and classroom interactions. Enhancing AI accessibility could also improve its practical use for teacher professional development beyond research contexts (Jeon & Lee, 2023).

Despite progress, most existing research on teacher question detection has focused on written language rather than authentic spoken classroom data. For example, Dent and Paul (2020) targeted question detection in Twitter tweets, a context far removed from classroom interactions, while much educational technology research has emphasized detecting and assessing written responses (Chang et al., 2022; Lee et al., 2024).

Table 1 Brief descriptions of the technical vocabulary

Machine Learning (ML) and Natural Language Processing (NLP) Approaches	
BERT (Bidirectional Encoder Representations from Transformers)	An NLP model developed by Google in 2018, BERT uses a transfer learning approach to learn high-dimensional, contextualized representations of words. It's pre-trained using fill-in-the-blank questions and fine-tuned for downstream tasks such as text classification
ProsusAI's Finnish BERT	A language-specific version of BERT optimized for Finnish, achieving state-of-the-art results in various Finnish NLP tasks
Bag of Words (BoW)	A model that represents a text sequence by counting the occurrences of words or sets of words (N-grams). It creates a sparse, interpretable vector where each dimension corresponds to a word's frequency in the text
Hand-Crafted Rules (HCR)	Rules developed by a human researcher for text annotation, used to detect specific patterns like questions in text by identifying explicit questioning words or phrases
ML Classification Models	
Logistic Regression (LR)	A statistical model that classifies inputs by applying a sigmoid function to a linear combination of input features, outputting probabilities for class membership. The model is optimized by minimizing logistic loss
Gradient-Boosted Trees (GBT)	An ensemble learning technique that iteratively builds an ensemble of decision trees, each correcting errors of the previous trees, leading to improved predictive performance
Random Forest Classifier	A Random Forest Classifier is a machine learning model that creates multiple decision trees and makes predictions by combining the results from all the trees
Generative Pre-trained Large Language Models (LLMs)	
GPT-3.5 Turbo	A version of OpenAI's Generative Pre-trained Transformer model, known for its low cost and wide use in various NLP tasks, particularly in educational contexts
GPT-4	An advanced version of the GPT model, expected to solve more complex problems with greater accuracy, albeit at a higher cost
Prompt Engineering and Advanced Techniques	
Prompt Engineering	The practice of designing effective prompts to elicit desired responses from LLMs, crucial for instructing models in specific tasks
Vanilla	The simplest version and prompting technique that can be implemented
Chain of Thought (CoT)	A prompting technique where a model is guided through a step-by-step reasoning process, shown to improve performance in tasks requiring logical and sequential reasoning

Table 1 (continued)

Machine Learning (ML) and Natural Language Processing (NLP) Approaches	
Self-Ask	A model prompting approach where sub-questions are generated and answered first, improving overall performance by breaking down the main question into manageable parts
Experimentation and Metrics	
Pretraining and Fine-tuning	The two-step process in transfer learning where a model is first pre-trained on a large dataset to learn general representations and then fine-tuned on a specific task to optimize its performance
N-gram	A contiguous sequence of n items (words or characters) from a given sample of text, used in models like BoW to capture context within the text
Research and Evaluation	
Zero-shot Learning	The ability of a model to perform a task without any prior specific examples provided during training
Few-shot Learning	The ability of a model to perform a task given a small number of examples during training

Some efforts have extended into automating classroom interaction analysis. Foster et al. (2024), for instance, detected classroom activities from video, and Alic et al. (2022) applied ML to human-constructed transcripts. However, such approaches tend to identify broad instructional formats, such as whole-class versus group work, but struggle with finer discursive features. Xu et al. (2024) found that AI and LLMs could classify major discourse elements, such as speaker diarization, but remained limited in detecting more nuanced elements like teacher feedback. These findings highlight the need for more refined tools. Given the complexity of spoken data, Maity et al. (2024) emphasized that well-crafted prompts paired with human validation are essential for guiding LLMs in accurately identifying key moments.

This emphasis on written language also characterizes recent comparisons between pre-trained LLMs and traditional ML approaches. Studies such as Fussell et al. (2024) and (Yeung, 2025) have primarily focused on written text analysis, while others have explored generative tasks like question creation. Although LLMs are increasingly recognized for their adaptability and effectiveness, their use in analyzing spoken language and classroom discourse remains limited (Tran et al., 2024; Xu et al., 2024). Fussell et al. (2024), for example, compared fine-tuned and pre-trained LLMs with conventional ML models in evaluating students' written lab notes, showing that method effectiveness depends heavily on task context. While traditional ML requires large, annotated datasets and extensive tuning (Yeung, 2025), LLMs benefit from broad pre-training and can be adapted with minimal labeled data. Despite these advantages, ML may outperform LLMs in well-defined domains with clear features and abundant data. Ultimately, selecting the appropriate approach depends on specific research goals, data availability, and task complexity.

While ML and LLM have been applied across educational domains, detecting teacher questions from authentic spoken classroom interactions remains underexplored. This is due to the complexity of spoken language, limited annotated classroom data, and the predominance of LLM training on written text. This study addresses this gap by comparing ML models trained on domain-specific data with pre-trained LLMs to detect

teacher questions from authentic ASR transcriptions. The goal is to advance automated analysis methods that support science teaching, learning, and teacher professional development at scale.

Research questions

The following research question will be addressed:

- To what extent can purposefully trained machine learning approaches and pre-trained large language models detect science teachers' questions in automatic speech recognition outputs of physics lessons?

By answering this question, we aim to understand how these two approaches, one trained specifically on the task (ML) and the other adapted from broad pre-training (LLM), handle the challenge of identifying teacher questions within the context of classroom interaction. Based on the results, the future role of technology in educational research and education will be discussed.

Methods

Terminology

First, to enhance readability, a list of acronyms is provided below:

- AI – Artificial Intelligence
- ASR – Automatic Speech Recognition
- BoW – Bag of Words
- BA – Balanced Accuracy
- BERT – Bidirectional Encoder Representations from Transformers
- CoT – Chain of Thought
- GBT – Gradient-Boosted Trees
- HCR – Hand-Crafted Rules
- IRF – Initiation-Response-Feedback
- LLM – Large Language Model
- LR – Logistic Regression
- ML – Machine Learning
- NLP – Natural Language Processing
- P – Precision
- R – Recall
- SA – Self-Ask
- TP – True Positive
- FP – False Positive
- FN – False Negative
- F1 score – Harmonic Mean of Precision and Recall

Data

The study data consists of 23 physics lessons from 23 different science teachers on the same topic of electricity and electrical energy, typically taught in the ninth grade in Finland. Each lesson was 90 min long. Structuring of the lessons was very similar because teachers tend to follow curriculum-compliant physics text- and workbooks when planning and teaching lessons. A typical lesson included activities such as discussing homework, introducing the day's topic as well as key concepts, experiments, and calculation tasks (Viiri & Helaakoski, 2014).

All the lessons were video recorded with two video cameras, and separate microphones were placed on each side of the classroom. The ASR output of the audio from a portable microphone carried by the teachers is used. The ASR output consists of the recognised speech divided into five-second intervals (Fig. 1.). In other words, all the words recognised within five seconds form a unit of words in a row. Worthwhile to notice that the ASR text output does not include question marks, which is a feature still under development in ASR (Lucas et al., 2023) and could have been used for automatic detection. The overall quality of the purposefully developed ASR has been shown to be sufficient (Lehesvuori et al., 2023). While LLMs have demonstrated an ability to interpret imperfect human language, challenges persist with interactive spoken language (Umari et al., 2024) and in languages that have fewer training resources, such as Finnish (Phan et al., 2024). At the same time, recognizing the limitation of focusing solely on ASR text outputs for analyzing teacher talk, this study emphasizes the teacher's role as an orchestrator in the complex knowledge-building processes and interactions in science. Research indicates that linguistic features are more crucial than missing acoustic-prosodic features when it comes to automatically detecting teacher questions (Donnelly et al., 2017).

The rationale for considering ASR outputs arises from the fact that recording and processing video data in educational research is both time-consuming and fraught with privacy and ethical concerns (Derry et al., 2010). In contrast, handling audio recordings and their associated transcripts is typically more efficient and less demanding in terms of data management (Foster et al., 2024). On the negative side, the information obtained from the videos are still necessary for multimodal analysis of discourse (Kääntä et al., 2021), for example when delving into teacher gestures or proxemic shifts in the classrooms.

Manual annotation of teacher questions

Teacher questions were annotated to ASR transcripts by the first author who followed the coding schemes developed in previous studies (Lehesvuori et al., 2013). Any questions related to the lesson topic and science content were annotated (e.g., "What is the unit of power?", "How many batteries do you need for this lamp?", "What did you get for the result?"). If the question was off topic, it was not annotated. For example, teacher may have been asking when a student needs to leave for health inspection ("When is your doctor's appointment?"). These types of informal questions typically occurred before or after the actual lesson and analysis and did not affect the results since they were excluded from the unit of analysis (i.e., the actual lesson). Manual validation of the technology-driven analysis revealed few off topic instances

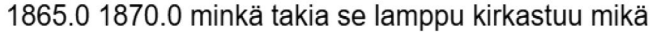


Fig. 1 Example of a single ASR output row including a question. The English translation for the question is "why does the lamp shine brighter"

being detected during the lesson. In overall, without few exceptions, teacher questioning was content-related.

Coding of teacher questions can take place directly from videos (Fischer et al., 2014a, 2014b). The new element here was to place the annotations directly to ASR outputs and interpret the questions only on this information, since this is the same information provided to ML and LLM. The reliability for the manual categorisation (question or not) has been found to be at substantial level between inter-raters when labelled directly with videos (Fischer et al., 2014a, 2014b). Manual inter-rating directly from ASR outputs in this study also resulted in substantial Cohen's Kappa value 0.85. Then, agreement between coding ASR outputs with and without audio resulted in Cohen's Kappa value 0.84. The key difference observed was that some more questions were detected in the audio data. This was mainly due to a few instances where the ASR output either missed or misinterpreted a question word. In total 1480 questions were annotated in ASR outputs with average 64 ($SD = 24$) questions per teacher. Minimum number of annotated questions by a teacher was 27 and maximum 101. Testing the experimental ML and LLM approaches introduced next took place by using the manual annotations (without audio) with final agreements adjusted for the codes initially disagreed upon.

Experimental approaches

Overview of the technical terms and brief descriptions

Before opening up experimental approaches in more detail, a brief description of the technical vocabulary is provided in Table 1 in order to enhance the readability.

Experiments include testing and comparing supervised ML approaches to AI-based LLM approaches with different prompting techniques. We also include metrics about Hand-crafted approaches developed during human annotation of the data. Before introducing the results, a brief overview of the implemented approaches is presented.

Machine learning approaches

BERT is a natural language NLP model developed in 2018 by Google (Devlin et al., 2018). Since then, it has been implemented for various NLP tasks. BERT applies a transfer learning approach in which the model learns high-dimensional, contextualised representations of words using fill-in-the-blank questions. Then, these representations can be adjusted and used for more downstream tasks such as text classification or next-sentence prediction. This tandem of procedures is usually known as pretraining and fine-tuning the model, respectively. Although BERT can be assimilated to LLMs through its use of large-scale pre-training and transformer-based architecture, the major difference lies in its bidirectional encoding and lack of native text generation capabilities. In other words, BERT is primarily designed for understanding and interpreting text, whereas modern LLMs also have the capability to generate coherent and contextually relevant language outputs. The experimentation was conducted by using the large version of ProsusAI's Finnish BERT. The language-specific model can be applied to achieve state-of-the-art results for different Finnish NLP tasks (Virtanen 2019).

The Bag of Words (BoW) model represents a text sequence, such as a sentence or a document, by counting the number of times a particular word or set of words (known as an N-gram) appears in the sequence. Thus, given that each N-gram has its own dimension on

the vector, an advantage of this approach lies in its interpretability and sparsity. This way, it is possible to identify pivotal words based on the coefficients associated in the classifier, given that the coefficients in the classifier are interpretable, as was the case in our experiments. In terms of confirming the validity, the list of words found during experimentation was checked and evaluated by the researcher well familiar with the data. The same researcher conducted the manual annotation. Therefore, the relevance of the words could be confirmed.

Hand-crafted rules (HCR) were developed by the researcher who conducted the manual annotation. A question was annotated to ASR text output when it was explicitly indicated. Explicit questions were often indicated through a questioning word or phrase (e.g., "why", "how", "where", "how many", "what is"). Below are listed some of the rules applied (Note: Since direct translation may compromise linguistic precision and contextual clarity, the full set of rules containing Finnish terms is available upon request):

- If at least one word from set A {how, what, when…(the list continues with typical Finnish questioning words)} appears, then the segment is a question; otherwise, it is not a question
- If at least one phrase from set B {can you recall, do you remember, Does anything come to mind…(the list continues with typical Finnish questioning words and phrases eliciting memorization)} appears, then the segment is a question; otherwise, it is not a question
- If at least one phrase from set C {take a guess, can you guess, can you think of anything…(the list continues with typical Finnish questioning words and phrases that prompt guessing or making assumptions)} appears, then the segment is a question; otherwise, it is not a question

The rules were applied to detect questions, and the one that yielded the best average balanced accuracy across all lessons was selected in the end as the baseline for each respective task. Besides running the BERT and BoW models alone, they were also combined with the HRC to potentially enhance the performance.

Two options were considered for the ML classification model: logistic regression (LR) and gradient-boosted trees (GBT). LR classifies the inputs by applying a one-dimensional affine transformation and then a sigmoid function, returning a number between 0 and 1. These can be interpreted as the probability that the given input is from the positive class. Next, the regression coefficients are optimised by minimising the logistic loss between the predicted probabilities and the ground truth classes. The GBT applies an iteratively generated ensemble of decision trees. In each iteration, the tree is fit over the remaining errors from the past iteration. Therefore, the predictive capability of the model is enhanced with each classifier added, achieving better performance than the Random Forest approach.

Pre-trained language models

LLMs can deal with and generate natural language. In this study, we mainly apply OpenAI API GPT 3.5 turbo, which has (momentarily) been the best-performing model of the GPT 3.5 family. Whereas GPT 3.5 turbo is applied for the whole data set due to its low-cost use. That is, we were able to only test GPT4 (Open AI, 2023) with one lesson only because of the higher costs estimated for the whole data set. It is expected that the newer models can solve more difficult problems with greater accuracy than the previous models. The main

rationale for selection of GPT models was their wide use in the field of education during the time of our investigations. The emergence of GPT-based models marked the beginning of a new era in NLP, where increasingly larger models are trained on massive datasets, enabling them to contextualise with few or even zero examples to generate solutions to a wide range of tasks (Brown et al., 2020). Contemporary LLMs have undergone fine-tuning to follow human instructions in a dialogue format, making them versatile for applications such as summarization, translation, and even text classification (Ouyang et al., 2022).

Today, the term "prompt" refers to the text provided by a user to instruct an LLM on how to perform a specific task in a dialogue-like exchange (see example in Fig. 2.). This paradigm shift is referred to as 'in-context learning,' leading to the development of prompt engineering techniques, which aim to design appropriate prompts for eliciting desired

VANILLA

Input: <context> = mihin sun no
<sentence> = mitäs me teemme ku terveydenhoitaja aikoja kahelle
<context>= et voi mennä joo no

Question: Is the sentence a question? (response yes or no)
Express the answer with the prefix "Then the final answer is:".

((GPT 3.5 Turbo response))

The sentence "mitäs me teemme ku terveydenhoitaja aikoja kahelle" is not a question. It is a statement or a call to action, but there is no direct question in it.

Then the final answer is: No.

SELF-ASK

Input:

<context> = jotai tekee mutta harvemmin sähkömoottoria
<lause> = eihä tuttuja laitteita mitäs sitä
<context> = käytetään usein suihkun jälkeen

Question: Is the sentence a question? (response yes or no)
Warning: --- (questions marks are not in the transcriptions).

((GPT 3.5 Turbo response))

CoT: The sentence provide contains the word "mitäs" -> The sentence is a question.
(one-step reasoning)

CoT: Does the sentence contain a question mark? No. Does the sentence contain a key question word? Yes, the sentence contains the word "mitäs". So, the sentences is a question.

Then the final answer is: Yes.

Fig. 2 Examples of Vanilla zero-shot and Self-ask prompts build on context (<konteksti>) sentence before and after and the sentence with a questioning word (mitäs = what) in the middle (<lause>). Self-Ask builds upon the Chain of Though (CoT) prompting strategy, thus acronym CoT can be found in the response output

responses from LLM (Liu et al., 2021). For example, in this study the simplest prompt tested was (see Fig. 2 for the complete input):

Is the sentence a question?" (response: yes or no).

We refer to this as the Vanilla approach, a term commonly used in information technology to denote the simplest and default configuration. Recent advancements have introduced techniques to enhance response quality, such as Chain of Thought (CoT) (Lee et al., 2024) and Self-Ask (SA) (Press et al., 2023).

The CoT is a step-by-step model that Kojima et al. (2022) applied to solving mathematical problems. In this study, this approach is adopted to detect teacher questions from physics lessons. We use both zero-shot and few-shot strategies to see how LLM performs with and without provided examples. The number of shots was set to five, as research suggests that using more examples leads to only limited improvements in performance (Wei et al., 2022). Increasing examples can sometimes improve performance, but also raises token usage, costs, and environmental impact. Therefore, we used a reasonable initial number consistent with prior work (e.g., Gao et al., 2023; Liévin et al., 2024). Recent research shows that the CoT approach improves the performance of LLM in arithmetic, commonsense, and symbolic reasoning (Wei et al., 2022). Whereas the zero-shot reasoning potentiality of LLMs has been recognised (Kojima et al., 2022), the few-shot approach is also applied to explore whether the classification metrics improve. The prompt for the LLM was:

Is the sentence a question? (response yes or no) Think step by step. The preluding and following lines are considered as context sentences.

Self-Ask builds upon the CoT prompting strategy by taking the reasoning process one step further. Whereas CoT encourages step-by-step thinking, Self-Ask explicitly introduces intermediate sub-questions that guide the model through each reasoning step (Press et al., 2023). The aim is first to generate a sub-question and the answer to this question before the actual main question. This structure can enhance interpretability and accuracy, particularly for tasks requiring layered reasoning or contextual understanding. An interesting feature in the development of the model is the use of eliciting techniques and follow-ups, concepts acknowledged as essential elements of productive instructional dialogue in educational settings. The prompt for the LLM was:

Is the sentence a question? (response yes or no) Think step by step by applying sub questions before the final response" (an example of a generated sub question "Does the sentence include a questioning word?"). The preluding and following lines are considered as context sentences.

As introduced in the above prompts, the context used to determine whether a sentence is a question includes the preceding sentences. Therefore, in addition to the simple prompt, the short context length can make the input more processable for LLM (Sun et al., 2021; Tran et al., 2024).

Metrics for performance evaluation

For the test runs following metrics were investigated: Precision (P), Recall (R), Balanced Accuracy (BA) and Harmonic Mean (F1 score). These metrics have a long tradition in the field of computational data analysis and are widely used for evaluating classification

performance, particularly in tasks involving imbalanced data or complex labelling such as classroom interaction analysis (e.g., Donnelly et al., 2016). Precision is defined as ratio between the number of correctly classified questions (True Positives = TP) and the total number of utterances that were coded (TP + FP). Recall is the ratio between the number of correctly classified questions (TP) and the total number of real positives (TP + False Negatives = FN). Balanced Accuracy (e.g., Brodersen et al., 2010) is defined as the average of the recalls obtained for the positive and negative classification (True Negative). Finally, the F1 score stands for the Harmonic Mean between Precision and Recall. It is particularly informative in contexts where class distributions are imbalanced, as it reflects both precision and recall in a single metric. The formulas can be found in Fig. 3.

Results

The ML and LLM approaches were tested in different settings. First, different LLM (GPT-3.5) prompting approaches were tested with seven lessons (Table 2). Second, selected LLM (GPT-3.5) prompts were compared against ML and Hand-craft approaches with the whole data set (Table 2). Third, and finally, the best performing GPT-3.5 approach was compared against GPT-4 with one test lesson (Table 2).

In the first experiment (see Table 2). Vanilla 0-shot is a baseline for CoT 0-shot and CoT 0-shot is a baseline for CoT 5-shot. We compare CoT and SA to decide which of them we are going to use with the whole data set.

The seven lesson test results show that with GPT-3.5-turbo zero-shots CoT performed better than SA. Then, as expected, the GPT-3.5-turbo five-shot approach was the best-performing classification model. Five-shot prompting approach technique involved five real question examples from the data. The question examples were selected to demonstrate different types from checking and calculations ("How much do you get from that?", "Can you measure power for electric devices?, "How is resistance calculated?") to more challenging and open-ended ones ("How does the voltages change when batteries are connected to series?", "Okay, what does it mean then?"). While we merely aim to cover the task question or not here, the nature

$$^{*}P = \frac{TP}{TP+FP}$$

$$^{**}R = \frac{TP}{TP+FN}$$

$$^{***}BA = \frac{1}{2}\left(\frac{TP}{TP+FN} + \frac{TN}{TN+FP}\right)$$

$$^{****}F1 = 2 \times \frac{P \times R}{P+R}$$

Fig. 3 Metric used for the investigations. (*Precision is the ration between the number of correctly classified positives and the total number of utterances that our model classified as positives; **Recall is the ratio between the number of correctly classified positives and the total number of real positives; ***Balanced Accuracy is the average of the recalls obtained for the positive and negative classes; ****F1 Score is the harmonic mean between precision and recall. *TP* True Positives, *FP* False Positives, *FN* False Negatives, *TN* True Negatives)

Table 2 The precision, recall, balanced accuracy, and F1 score for detecting questions for 7 lessons: Comparison of prompts for large language models

Task		Question vs Non-question			
Features	Classif	P*	R**	BA***	F1****
GPT-3.5-turbo 0-shot	SA*****	0.2387	0.3352	0.6289	0.2789
GPT-3.5-turbo 0-shot	Vanilla	0.2214	0.0577	0.5215	0.0916
GPT-3.5-turbo 0-shot	CoT	0.2818	0.3054	0.6245	0.2931
GPT-3.5-turbo 0-shot after-before context	CoT	0.4239	0.1453	0.5655	0.2164
GPT-3.5-turbo 5-shot	CoT	0.33	0.7970	0.8399	0.4667

CoT Chain of Thought, *SA* Self-ask

*Ratio between the number of correctly classified positives and the total number of utterances that our model classified as positives

**Ratio between the number of correctly classified positives and the total number of real positives

***Balanced accuracy is the average of the recalls obtained for the positive and negative classes

****The harmonic mean between precision and recall

*****Few-shot does not apply for SA, given the nature of the explicit instruction to generate sub-questions. The need for individual shots is made redundant since the examples generated by SA are similar to the few-shots examples

of teacher questioning has been explored further elsewhere (Lehesvuori et al., 2021a). The use of examples seems to improve in-context learning and pushed the model to better performance. All in all, the baseline zero-shot approaches (Vanilla, SA, CoT) seem insufficient in all metrics.

Since the results showed that CoT performs better than SA (both with zero-shot), we selected CoT to perform in five-shot format. We also hypothesize that the performance over the sample generalizes over the whole data. Therefore, we are keeping the CoT as main prompting approach when running the next experiments. The next step was to run selected LLM models with the whole data and compare the results with ML and HCR models (Table 3). Also the combinatory (ML + HCR) models were included in the experiment. In combined-feature approach the classifiers were input with the concatenation of the results from the hand-crafted rules and the automated features.

The results show that while GPT-3.5-turbo performs well, also HCR models are relevant in both sole and combination models (especially BoW + HCR). This raises an interesting challenge for the up-to-date GPT-4 model. However, we were able to run only one lesson for comparison with the best-performing GPT-3.5 model (4) due to the high costs of the GPT-4 (cf. Tran et al., 2024). As discussed later, this limitation highlights the need to make LLMs accessible and affordable to promote fair access (Kasneci et al., 2023).

The one lesson results show GPT-4 outperforming the previous model. Results also indicate that GPT-4 could challenge the HCR with the whole data, yet this experiment cannot be executed here since the high costs Table4.

Table 3 The precision, recall, balanced accuracy, and F1 score for detecting questions for the whole data: Comparison of hand-crafted and automated approaches

Task		Question vs Non-question			
Features	Classif	P*	R**	BA***	F1****
HCR	Human-aided	0.351	0.609	0.77	0.436
BoW	LR	0.185	0.529	0.695	0.27
	GBT	0.172	0.69	0.744	0.272
BERT	LR	0.131	0.567	0.669	0.211
	GBT	0.13	0.517	0.652	0.205
BoW + HCR	LR	0.298	0.7	0.799	0.411
	GBT	0.311	0.742	0.821	0.431
BERT + HCR	LR	0.149	0.615	0.7	0.237
	GBT	0.164	0.71	0.745	0.263
GPT-3.5-turbo 0-shot	Vanilla	0.2137	0.0555	0.5215	0.0881
GPT-3.5-turbo 0-shot	CoT	0.2535	0.2875	0.6179	0.2694
GPT-3.5-turbo 5-shot	CoT	0.2937	0.8036	0.8428	0.4302

BoW Bag of Words, *CoT* Chain of Thought, *GBT* Gradient-boosted trees, *HCR* Hand-crafted rule, *LR* Logistic regression

*Ratio between the number of correctly classified positives and the total number of utterances that our model classified as positives

**Ratio between the number of correctly classified positives and the total number of real positives

***Balanced accuracy is the average of the recalls obtained for the positive and negative classes

****The harmonic mean between precision and recall

Table 4 The precision, recall, balanced accuracy, and F1 score for detecting questions for 1 lesson: Comparison of different large language models

Task		Question vs Non-question			
Features	Classif	P*	R**	BA***	F1****
GPT-4 5-shot	CoT	0.4188	0.9412	0.9145	0.5797
GPT-3.5-turbo 5-shot	CoT	0.36	0.87	0.87	0.51

CoT Chain of thought

*Ratio between the number of correctly classified positives and the total number of utterances that our model classified as positives

**Ratio between the number of correctly classified positives and the total number of real positives

***Balanced accuracy is the average of the recalls obtained for the positive and negative classes

****The harmonic mean between precision and recall

Discussion

This study demonstrated approaches to technology-aided detection of science teacher questions with LLM and ML approaches. The results showed that LLM approaches outperformed standalone supervised ML approaches and achieved parity with combinatorial HCR and ML approaches in detecting science teacher questions from ASR outputs. Expectedly, the few-shot

CoT model performed better than the zero-shot model (Kojima et al., 2021; Tran et al., 2024). While HCR approaches continued to offer value across the dataset, the single-lesson trial with GPT-4 indicates advancements in AI capabilities. Despite the indicated development, different models rely on distinct architectures, which may lead to varying performance across tasks (Shahriar et al., 2024). The research underscores the importance of aligning model selection and configuration with the specific demands of the task, particularly in domains involving complex phenomena under investigations.

When interpreting the results of this study, it is important to recognize that the choice of the most suitable approach and LLM model is highly context-dependent and influenced by multiple factors (Fussell et al., 2024). In addition to evaluating model performance, researchers must consider the availability of data and human resources for machine learning development, as well as the contextual appropriateness of using general-purpose LLMs (Xu et al., 2024). One notable advantage of using accessible, pre-trained LLMs over traditional machine learning approaches is that the latter often require large volumes of authentic, labelled training data, which is frequently limited in educational research settings.

The performance of the models using the selected techniques leaves room for improvement. One potential challenge is that, unlike human-constructed transcriptions, the data consists of automatically generated text outputs from Finnish spoken classroom interaction (Phan et al., 2024), which lack multimodal features (e.g., prosodic cues such as intonation and tone) and punctuation cues like question marks (Donnelly et al., 2017). Accordingly, although LLM may be expected to misinterpret implicit questions as explicit ones without the multimodal cues, the Fig. 2 first example demonstrates an example on how both LLM and human coders agreed on implicit question, which was more of a statement rather than a question (i.e., no question coded).

Despite the limitations, the use of text-based data remains justified due to considerations related to data management and participant safety. Model performance could potentially be enhanced through fine-tuning and the application of more advanced prompt engineering strategies (Fussell et al., 2024). However, emerging evidence suggests that large language models may struggle to interpret complex or layered coding schemes commonly used in classroom discourse analysis (Xu et al., 2024). In particular, more nuanced or hierarchical coding appears to exceed the current capabilities of these models. On the other hand, the use of simple prompts in this study revealed limitations, highlighting the complexity of the data and the need to advance AI capabilities and prompting techniques in tandem.

Reflecting on the technology-aided analysis of classroom interaction, the immediate information about presence or absence of teacher questions in instructions could be essential tool for teachers when reflecting on their practices and potential use of the above approaches. As simplest, automatic detection can easily reveal the extremes (e.g., 27 questions vs. 101 questions in this data), which is useful information for further explorations and reflective discussion for example among peers (cf. Lehesvuori et al., 2021b). Although automated methods cannot yet achieve the accuracy of human coding based on highly developed and validated observational protocols, they can already provide the initial filtering tool for detecting classrooms with potentially richer dialogue. The use and detection of teacher questioning are crucial for fostering more interactive and dialogic teaching in science (Gillies, 2023).

Implications

One implication of the developing LLM technology could be automatic question detection to enhance AI-based tutoring to enrich lecturing and authoritative modes of teaching. Technology can help identifying relevant questions and summarize them for both teachers and students (Kasneci et al., 2023). It aids teachers in reflecting on their questioning strategies and provides students with overviews of key science concepts (cf., Schlotterbeck et al., 2020). Additionally, this process can occur in real-time during lectures, enabling immediate assessment of teacher actions and diagnosis of student contributions. Further development of this technology will significantly advance and amplify educational research with focus on enrichening transmission modes of teaching continuously prevailing in science education (Lyons, 2006; Eshach et al., 2013; Lehesvuori et al., 2018).

Techniques that enable LLMs to consider the semantics of classroom talk can pave the way for more advanced AI applications for adaptive learning and adaptive teaching. In the context of this study, this could mean supporting science teachers in fostering more interactive classroom interactions by providing real-time feedback on their questioning in different in different phases of the lesson as well as in different educational settings (Li et al., 2025). Also, as new conversational technologies like digital tutors (e.g., Hobert & Berens, 2024) are implemented as artificial agents in the classroom, an important prerequisite is their capability to discern spoken utterances and advance techniques toward natural language understanding (NLU) (e.g., Love et al., 2025).

Accordingly, the next step is to automatically distinguish between questions that support authoritative interactions and those that support dialogic interactions. Closed questions, which typically seek brief or factual responses, are often associated with authoritative discourse that emphasizes correctness and a single point of view. In contrast, open-ended questions are more characteristic of dialogic interactions because they invite elaboration, reasoning, and the consideration of multiple perspectives. Automatically identifying these question types can provide deeper insight into the nature of classroom discourse and its alignment with different pedagogical goals. Although, both types play a role in meaningful learning (Mortimer & Scott, 2003; Scott et al., 2006), dialogic interactions are still very infrequent in science, thus there is a call for technology-enhanced implications to facilitate such interactions. This could take place by shifting the focus on detecting more open type of questions that facilitate reasoning, argumentation and productive thinking (Bansal, 2018; Chin, 2007).

In line with the capabilities of LLMs, the ability to formulate valid questions and prompts becomes more crucial than merely responding to them (Johnson et al., 2024). Understanding the role and functions of different types of questions is essential when leveraging the power of LLMs for educational purposes, particularly in crafting prompts (Press et al., 2023). Similarly, LLM-generated feedback should move beyond simple praise to adopt a more process-oriented and feed-forward approach (Estévez-Ayres et al., 2024, Hirunvasiri et al., 2023; Li et al., 2025).

LLMs have received significant attention in educational research (Heilala et al., 2025). However, despite the promising advancements in AI for tutoring purposes, such as Intelligent Tutoring Systems and pedagogical agents (e.g., Estévez-Ayres et al., 2024; Sikström et al., 2022), generative LLMs still have a long way to go in effectively modelling productive educational dialogues in authentic classroom settings through questioning and feedback. Bridging this gap requires extensive multidisciplinary research on the evolving AI technology. While AI techniques currently offer significant benefits

for education and educational research (Jeon & Lee, 2023), these developments are also transforming the roles of teachers and students and raising important social and ethical issues in education (Zhai et al., 2021).

Whereas this study focuses on science teacher questioning, the methods and technologies explored hold potential for application across various teaching and learning contexts. These tools could support a wide range of instructional practices beyond any single subject. Future advancements may extend their usefulness to diverse educational settings, enhancing teaching and learning more broadly.

Conclusions, limitations and future research

In conclusion, the results indicate the complexity of classification tasks from the spoken language compared to written forms of outputs (Lee et al., 2024). Although automated approaches are expectedly more efficient in detecting questions than merely human approaches, the need for relevant examples indicates the need for human input in validation procedures often neglected due to the laborious nature of the processes (Chang et al., 2022). In other words, the interplay between human and technology-led approaches is necessary in studying complex educational phenomena such as discussed in this study.

While recognizing the ongoing evolution of technology, the findings present pertinent information as of the analysis conducted in Autumn 2023. Subsequently, there has been rapid advancement in technology, leading to an expansion in options for large language models (LLMs). We acknowledge the dynamic nature of technological advancements, recognizing that the results obtained in this study may soon become outdated due to this rapid pace of development. Nevertheless, the findings can still serve as a valuable reference for future work. To inform future methodological directions and assess the capabilities of pre-trained LLMs against domain-specific models, this study began with the practical and relevant task of binary coding of teacher questions. This streamlined approach is motivated by concurrent research, including our own, which has shown that large language models (LLMs) often struggle to apply nuanced coding schemes to complex classroom interaction data (Xu et al., 2024), underscoring the value of exploring simpler, prompt-based methods using non–fine-tuned and easily accessible LLMs.

There is a growing concern that the evolution of AI brings risks that conflict with ethical principles, such as equality coming along with open accessibility to explore possibilities of AI. Harnessing the power of AI for good should be open for all regardless of the resources (Kasneci et al., 2023). In the future, there is a need to ensure ethical principles of AI in education, contributing to a responsible and trustworthy digital transformation. Yet, the identified ethical guidelines are mainly abstract (Ryan, 2020), and do not offer practical solutions for teachers. These practical solutions should firmly ground on latest educational research. Our aim is to continue research on science teacher questioning strategies and suggest AI-based implications that facilitate student learning of science. The next step is to use AI to investigate how different types of prompting affect the identification of a larger proportion of questions and the detection of various types of questions asked by science teachers. In general, regarding educational studies, the power of multidisciplinary approaches must be harnessed to develop a merged approach, bringing together the two sides, volume and accuracy.

Acknowledgements 2nd and 4th Authors were supported by ANID/PIA/Basal Funds for Centers of Excellence FB0003/ Support 2024 AFB240004 and ANID Exploración 13240075; 5th Author was was supported by the Academy of Finland under Grant number 353325.

Author contributions All authors contributed to the study conception and design. Material preparation, data collection and analysis were performed by Sami Lehesvuori and Felipe Urrutia. The first draft of the manuscript was written by Sami Lehesvuori and all authors commented on previous versions of the manuscript. All authors read and approved the final manuscript.

Funding Open Access funding provided by University of Jyväskylä (JYU).

Data availability The data of this study are available from the corresponding author upon reasonable request and subject to appropriate permissions. Interested researchers should contact the first author to inquire about access.

Declarations

Funding Kulttuurin ja Yhteiskunnan Tutkimuksen Toimikunta,353325,Raija Hämäläinen,ANID/PIA/Basal Funds for Centers of Excellence FB0003,FB0003,Roberto Araya,FB0003,Felipe Urrutia,ANID Exploración,13240075,Felipe Urrutia,ANID Exploración,13240075,Roberto Araya

Conflict of interest No competing interests.

Ethical approval Following the guidelines of the University of Jyväskylä, an ethics statement was not required for the study.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

References

Alic, S., Demszky, D., Mancenido, Z., Liu, J., Hill, H.C., & Jurafsky, D. (2022). Computationally identifying funneling and focusing questions in classroom discourse. arXiv:2208.04715

Bansal, G. (2018). Teacher discursive moves: Conceptualising a schema of dialogic discourse in science classrooms. *International Journal of Science Education, 40*, 1891–1912

Blanchard, N., Donnelly, P., Olney, A., Samei, B., Ward, B., Sun, X., Kelly, S., Nystrand, M., D'Mello, S. K. (2016). Identifying teacher questions using automatic speech recognition in classrooms. ACLWeb. https://aclanthology.org/W16-3623/

Brodersen, K. H., Ong, C. S., Stephan, K. E., & Buhmann, J. M. (2010). The balanced accuracy and its posterior distribution. Proceedings of the 2010 20th International Conference on Pattern Recognition (ICPR), 3121–3124. https://doi.org/10.1109/ICPR.2010.764

Brown. T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S. (2020). Language models are few-shot learners. arXiv (Cornell University), 2020. https://doi.org/10.48550/arxiv.2005.14165

Chang, L.-H., Kanerva, J., & Ginter, F. (2022). Towards automatic short answer assessment for finnish as a paraphrase retrieval task. ACLWeb, 2022. https://aclanthology.org/2022.bea-1.30/

Chin, C. (2007). Teacher questioning in science classrooms: Approaches that stimulate productive thinking. *Journal of Research in Science Teaching, 44*(6), 815–843. https://doi.org/10.1002/tea.20168

Dent, K.D., & Paul, S.A. (2020). Through the twitter glass: detecting questions in micro-text. arXiv:2006.07732

Derry, S. J., Pea, R. D., Barron, B., Engle, R. A., Erickson, F., Goldman, R., Hall, R., Koschmann, T., Lemke, J. L., Sherin, M. G., & Sherin, B. L. (2010). Conducting video research in the learning sciences: guidance on selection, analysis, technology, and ethics. *Journal of the Learning Sciences, 19*(1), 3–53. https://doi.org/10.1080/10508400903452884

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of deep bidirectional Transformers for language understanding. In *arXiv [cs.CL]*. arXiv. http://arxiv.org/abs/1810.04805

Donnelly, P.J., Blanchard, N., Olney, A.M., Kelly, S., Nystrand, M., & D'Mello, S.K. (2017). Words matter: automatic detection of teacher questions in live classroom discourse using linguistics, acoustics, and context. Proceedings of the Seventh International Learning Analytics & Knowledge Conference

Donnelly, P. J. (2016). Detecting teacher questions in the classroom. *In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, 1*, 564–574. https://doi.org/10.18653/v1/P16-1055

Dowell, N. M., Nixon, T., & Graesser, A. C. (2018). Group communication analysis: A computational linguistics approach for detecting sociocognitive roles in multi-party interactions. *Behavior Research Methods, 51*(3), 1007–1041. https://doi.org/10.3758/s13428-018-1102-z

Eshach, H., Dor-Ziderman, Y., & Yefroimsky, Y. (2014). Question asking in the science classroom: Teacher attitudes and practices. *Journal of Science Education and Technology, 23*, 67–81. https://doi.org/10.1007/s10956-013-9451-y

Estévez-Ayres, I., Callejo, P., Hombrados-Herrera, M. Á., & Delgado Kloos, C. (2024). Evaluation of LLM tools for feedback generation in a course on concurrent programming. *International Journal of Artificial Intelligence in Education*. https://doi.org/10.1007/s40593-024-00406-0

Fischer, H. E., Labudde, P., Neumann, K., & Viiri, J. (2014). *Quality of instruction in physics*. Waxmann Verlag

Fischer, H. E., Neumann, K., Labudde, P., & Viiri, J. (Eds.). (2014b). *Quality of instruction in physics: Comparing Finland*. Waxmann Verlag

Foster, J. K., Korban, M., Youngs, P. A., Watson, G. S., & Acton, S. T. (2024). Automatic classification of activities in classroom videos. *Computers and Education. Artificial Intelligence, 6*, 100207

Fussell, R. K., Flynn, M., Damle, A., Fox, M. F., & Holmes, N. (2024). Comparing large language models for supervised analysis of students' lab notes. *Physical Review Physics Education Research, 21*(1), 010128

Gall, M. D. (1970). The use of questions in teaching. *Review of Educational Research, 40*, 707–721

Gao, Y., Nuchged, B., Li, Y., & Peng, L. (2023). An investigation of applying large language models to spoken language learning. *Applied Sciences (Basel, Switzerland), 14*(1), 224

Gillies, R. M. (2016). Structuring cooperative group work in classrooms. *International Journal of Educational Research, 79*, 97–108. https://doi.org/10.1016/j.ijer.2016.07.004

Gillies, R. M. (2023). Dialogic teaching in a year 5 classroom during cooperative inquiry-based science. *International Journal of Educational Research Open*. https://doi.org/10.1016/j.ijedro.2023.100290

Stefanel Gris, L. R., Marcacini, R., Junior, A. C., Casanova, E., Soares, A., Aluísio, S. M. (2023). Evaluating OpenAI's whisper ASR for punctuation prediction and topic modeling of life histories of the museum of the person. arXiv.org, 2023. https://arxiv.org/abs/2305.14580

Heilala, V., Araya, R., & Hämäläinen, R. (2025). Beyond text-to-text: An overview of multimodal and generative artificial intelligence for education using topic modeling. In SAC '25 : Proceedings of the 40th ACM/SIGAPP Symposium on Applied Computing (pp. 54–63). ACM. https://doi.org/10.1145/3672608.3707764

Hennessy, S., Howe, C., Mercer, N., & Vrikki, M. (2020). Coding classroom dialogue: Methodological considerations for researchers. *Learning, Culture and Social Interaction*. https://doi.org/10.1016/j.lcsi.2020.100404

Hirunyasiri, D., Thomas, D. T., Lin, J., Koedinger, K. R., Aleven, V. (2023). Comparative analysis of GPT-4 and human graders in evaluating praise given to students in synthetic dialogues. arXiv.org. https://doi.org/10.48550/arXiv.2307.02018

Hobert, S., & Berens, F. (2024). Developing a digital tutor as an intermediary between students, teaching assistants, and lecturers. *Educational Technology Research and Development: ETR & D, 72*(2), 797–818. https://doi.org/10.1007/s11423-023-10293-2

Jensen, E., Dale, M., Donnelly, P. J., Stone, C., Kelly, S., Godley, A., & D'Mello, S. K. (2020). Toward Automated Feedback on Teacher Discourse to Enhance Teacher Learning. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*

Jeon, J., & Lee, S. (2023). Large language models in education: A focus on the complementary relationship between human teachers and ChatGPT. *Education and Information Technologies*. https://doi.org/10.1007/s10639-023-11834-1

Johnson, B., Van Campenhout, R., Jerome, B., Castro, M. F., Bistolfi, R., & Dittel, J. S. (2024). Automatic question generation for spanish textbooks: Evaluating Spanish questions generated with the parallel construction method. *International Journal of Artificial Intelligence in Education*. https://doi.org/10.1007/s40593-024-00394-1

Kääntä, L. (2021). Multimodal perspective into teachers' definitional practices: Comparing subject-specific language in physics and history lessons. In S. Kunitz, N. Markee, & O. Sert (Eds.), *Classroom-based conversation analytic research. Educational Linguistics.* Springer

Kasneci, E., Seßler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., Gasser, U., Groh, G. L., Günnemann, 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*. https://doi.org/10.1016/j.lindif.2023.102274

Kelly, S., Olney, A. M., Donnelly, P., Nystrand, M., & D'Mello, S. K. (2018). Automatically measuring question authenticity in real-world classrooms. *Educational Researcher, 47*(7), 451–464. https://doi.org/10.3102/0013189x18785613

Kojima T, Gu SS, Reid M, Matsuo Y, Iwasawa Y. (2022). Large language models are zero-shot reasoners. arXiv:2205.11916 [cs], May 2022. https://arxiv.org/abs/2205.11916

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*. https://doi.org/10.1016/j.caeai.2024.100213

Lehesvuori, S., Schlotterbeck, D., Jiménez, A., Caballero, D., Araya, R., & Hämäläinen, R. (2023). Towards automatic analysis of science classroom talk: Focus on teacher questions. In G. Bansal, & U. Ramnarain (Eds.), Fostering Science Teaching and Learning for the Fourth Industrial Revolution and Beyond (pp. 123–146). IGI Global. Advances in Educational Technologies and Instructional Design Book Series. https://doi.org/10.4018/978-1-6684-6932-3.ch006

Lehesvuori, S., & Ametller, J. (2021). Exploring coherence and authorship in pedagogical link-making in science. *International Journal of Science Education, 43*(17), 2791–2813. https://doi.org/10.1080/09500693.2021.1991599

Lehesvuori, S., Hähkiöniemi, M., Ketonen, L., Lerkkanen, M.-K., Pöysä, S., & Pakarinen, E. (2021). Reflections on dialogicity: Challenges and suggestions by mathematics student teachers. *Learning Culture and Social Interaction*. https://doi.org/10.1016/j.lcsi.2021.100567

Lehesvuori, S., Ramnarain, U., & Viiri, J. (2018). Challenging transmission modes of teaching in science classrooms: Enhancing learner-centredness through dialogicity. *Research in Science Education, 48*(5), 1049–1069. https://doi.org/10.1007/s11165-016-9598-7

Lehesvuori, S., Viiri, J., Rasku-Puttonen, H., Moate, J., & Helaakoski, J. (2013). Visualizing communication structures in science classrooms: Tracing cumulativity in teacher-led whole class discussions. *Journal of Research in Science Teaching, 50*(8), 912–939. https://doi.org/10.1002/tea.21100

Lemke, J. L. (1990). *Talking science: Language, learning, and values*. Ablex Publishing Corporation

Li, K., Lan, J., & Hu, Y. (2025). Comparative analysis of GPT-4.0 and teacher feedback on student-generated questions in the flipped classroom. *Educational Technology Research and Development*. https://doi.org/10.1007/s11423-025-10467-0

Liévin, V., Hother, C. E., Motzfeldt, A. G., & Winther, O. (2024). Can large language models reason about medical questions? Patterns (New York), 5(3), 100943

Littleton, K., & Mercer, N. (2013). *Interthinking: putting talk to work*. Routledge

Liu, P., Yuan, W., Fu, J., Jiang, Z., Hayashi, H., Neubig, G. (2021). Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. arXiv:2107.13586 [cs], 2021. https://arxiv.org/abs/2107.13586

Love, R., Law, E., Cohen, P. R., & Kulić, D. (2025). Teaching a Conversational agent using natural language: Effect on learning and engagement. *International Journal of Artificial Intelligence in Education*. https://doi.org/10.1007/s40593-025-00461-1

Lyons, T. (2006). Different countries, same science classes: Students' experiences of school science in their own words. *International Journal of Science Education, 28*(6), 591–613

Maity, S., Deroy, A., & Sarkar, S. (2024). Exploring the capabilities of prompted large language models in educational and assessment applications. arXiv:2405.11579

Mercer, N., Dawes, L., & Staarman, J. K. (2009). Dialogic teaching in the primary science classroom. *Language and Education, 23*(4), 353–369

Mortimer, E. F., & Scott, P. (2003). *Meaning making in secondary science classrooms*. McGraw-Hill Education

Nojavanasghari, B., Baltrušaitis, T., Hughes, C., & Morency, L. (2016). EmoReact: a multimodal approach and dataset for recognizing emotional responses in children. In *Proceedings of the 18th ACM international conference on multimodal interaction*

Nystrand, M., et al. (1998). *Opening dialogue: Understanding the dynamics of language and learning in the english classroom*. Teachers College Press

Oliveira, A. W. (2010). Improving teacher questioning in science inquiry discussions through professional development. *Journal of Research in Science Teaching, 47*(4), 422–453. https://doi.org/10.1002/tea.20345

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L. E., Simens, M., Askell, A., Welinder, P., Christiano, P. F., Leike, J., & Lowe, R. J. (2022). Training language models to follow instructions with human feedback. arXiv:2203.02155

Peng, L., Nuchged, B., & Gao, Y. (2023). Spoken Language Intelligence of Large Language Models for Language Learning. arXiv:2308.14536

Phan, N., von Zansen, A., Kautonen, M., Voskoboinik, E., Grosz, T., Hilden, R., & Kurimo, M. (2024). Automated content assessment and feedback for Finnish L2 learners in a picture description speaking task. In Interspeech 2024 (pp. 317–321). International Speech Communication Association. Interspeech. https://doi.org/10.21437/interspeech.2024-1166

Press, O., Zhang, M., Min, S., Schmidt, L., Smith, N.A., & Lewis, M. (2022). Measuring and Narrowing the Compositionality Gap in Language Models. arXiv:2210.03350

Ruiz-Primo, M. A. (2011). Informal formative assessment: The role of instructional dialogues in assessing students' learning. *Studies in Educational Evaluation, 37*(1), 15–24. https://doi.org/10.1016/j.stueduc.2011.03.005

Ryan, M. (2020). In AI we trust: Ethics, artificial intelligence, and reliability. *Science and Engineering Ethics*. https://doi.org/10.1007/s11948-020-00228-y

Schlotterbeck, D., Araya, R., Caballero, D., Jimenez, A., Lehesvuori, S., spsampsps Viiri, J. (2020). Assessing teacher's discourse effect on students' learning: A keyword centrality approach. In C. Alario-Hoyos, M. J. Rodríguez-Triana, M. Scheffel, I. Arnedillo-Sánchez, spsampsps S. M. Dennerlein (Eds.), EC-TEL 2020 : Addressing Global Challenges and Quality Education (pp. 102–116). Springer. Lecture Notes in Computer Science. https://doi.org/10.1007/978-3-030-57717-9_8

Scott, P., Mortimer, E. F., & Aguiar, O. G. (2006). The tension between authoritative and dialogic discourse: A fundamental characteristic of meaning making interactions in high school science lessons. *Science Education, 90*, 605–631

Shahriar, S., Lund, B. D., Mannuru, N. R., Arshad, M. A., Hayawi, K., Bevara, R. V., Mannuru, A., & Batool, L. (2024). Putting GPT-4o to the sword: A comprehensive evaluation of language, vision, speech, and multimodal proficiency. *Applied Sciences*. https://doi.org/10.3390/app14177782

Sohmer, R., Michaels, S., O'Connor, M. C., & Resnick, L. B. (2009). Guided construction of knowledge in the classroom. In B. B. Schwarz, T. Dreyfus, & R. Hershkowitz (Eds.), *Transformation of knowledge through classroom interaction* (pp. 105–129). Routledge

Sikström, P., Valentini, C., Sivunen, A., & Kärkkäinen, T. (2022). How pedagogical agents communicate with students: A two-phase systematic review. *Computers & Education, 188*, 104564. https://doi.org/10.1016/j.compedu.2022.104564

Sinclaire, J., & Coulthard, R. M. (1975). *Towards an analysis of discourse: The english used by teachers and pupils*. Oxford University Press

Sun, S., Krishna, K., Mattarella-Micke, A., & Iyyer, M. (2021). Do long-range language models actually use long-range context? arXiv:2109.09115

Tran, N., Pierce, B., Litman, D., Correnti, R., & Matsumura, L.C. (2024). Analyzing large language models for classroom discussion assessment. arXiv:2406.08680

Umair, M., Mertens, J. B., Warnke, L., & de Ruiter, J. P. (2024). Can language models trained on written monologue learn to predict spoken dialogue? *Cognitive Science, 48*(11), e70013. https://doi.org/10.1111/cogs.70013

van Zee, E., & Minstrell, J. (1997). Using questioning to guide student thinking. *Journal of the Learning Sciences, 6*(2), 227–269. https://doi.org/10.1207/s15327809jls0602_3

Viiri, J., & Helaakoski, J. (2014). Content and content structure of physics lessons and students' learning gains: Comparing Finland, Germany, and Switzerland. In H. E. Fischer, P. Labudde, K. Neumann, & J. Viiri (Eds.), *Quality of Instruction in Physics: Comparing Finland, Germany, and Switzerland* (pp. 93–110). Waxmann Verlag

Wang, J., Wang, Y., Wipfli, K., Thacker, B., & Hart, S. (2023). Investigating learning assistants' use of questioning in online courses about introductory physics. *Physical Review Physics Education Research*. https://doi.org/10.1103/PhysRevPhysEducRes.19.010113

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q. V., Zhou, D. (2022). Chain of thought prompting elicits reasoning in large language models. arXiv:2201.11903 [cs], Oct. 2022. https://arxiv.org/abs/2201.11903

Virtanen, A., Kanerva, J., Ilo, R., Luoma, J., Luotolahti, J., Salakoski, T., Ginter, F., & Pyysalo, S. (2019). Multilingual is not enough: BERT for Finnish. arXiv:1912.07076

Xu, S., Huang, X., Lo, C. K., Chen, G., & Jong, M. S. (2024). Evaluating the performance of ChatGPT and GPT-4o in coding classroom discourse data: A study of synchronous online mathematics instruction. *Computers and Education: Artificial Intelligence*. https://doi.org/10.1016/j.caeai.2024.100325

Yeung, S. (2025). A comparative study of rule-based, machine learning and large language model approaches in automated writing evaluation (AWE). In *Proceedings of the 15th international learning analytics and knowledge conference*

Yusuf, H., Money, A., & Daylamani-Zad, D. (2025). Pedagogical AI conversational agents in higher education: a conceptual framework and survey of the state of the art. *Educational Technology Research and Development*. https://doi.org/10.1007/s11423-025-10447-4

Zhai, X., Chu, X., Chai, C. S., Jong, M. S., Istenič, A., Spector, M., Liu, J., Yuan, J., & Li, Y. (2021). A Review of Artificial Intelligence (AI) in Education from 2010 to 2020. *Complex*. https://doi.org/10.1155/2021/8812542

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Sami Lehesvuori is an Adjunct Professor and University Lecturer at the University of Jyväskylä. His recent research explores how technology can be harnessed for the analysis and enhancement of dialogic interactions.

Felipe Urrutia is pursuing a Master's degree in Computer Science at the University of Chile. He holds a degree in Engineering Science, with a mention in Mathematics, from the Faculty of Physical and Mathematical Sciences at the University of Chile. His research expertise lies in applying natural language processing and machine learning to the study of educational phenomena.

Ville Heilala is a Senior Researcher at the University of Jyväskylä focusing on applying computational methods such as artificial intelligence and machine learning to the human sciences.

Roberto Araya is a distinguished professor at the University of Chile, with expertise in Electrical Engineering. Holding a PhD from the University of California, Los Angeles, his primary focus lies in the field of integrated Science, Technology, Engineering, and Math (STEM) education.

Raija Hämäläinen is a Full Professor working in the field of digitalization in and for learning and interaction at the Center for Research for Learning and Teaching at the University of Jyväskylä.