LLM-Human Alignment in Evaluating Teacher Questioning Practices: Beyond Ratings to Explanation

Ruikun Hou[1,2], Tim Fütterer[2], Babette Bühler[1], Patrick Schreyer[3], Peter Gerjets[4], Ulrich Trautwein[2], Enkelejda Kasneci[1],

[1]Technical University of Munich, [2]University of Tübingen,
[3]University of Kassel, [4]Leibniz-Institut für Wissensmedien

{ruikun.hou, babette.buehler, enkelejda.kasneci}@tum.de, {tim.fuetterer, ulrich.trautwein}@uni-tuebingen.de
schreyer@uni-kassel.de, p.gerjets@iwm-tuebingen.de

Abstract

The systematic assessment of teaching quality through classroom observation is a critical yet challenging task in educational research. Traditionally, trained raters evaluate instructional practices by analyzing classroom interactions (e.g., watching videos and annotating transcripts) based on structured protocols. Whereas the potential of using Large Language Models (LLMs) to automate teaching quality assessment is increasingly being explored, few studies have examined the underlying reasoning behind those generated holistic scores, which could provide specific feedback for raters and teachers. In this study, we investigate the alignment between LLM- and human-generated assessments of teacher questioning practices, focusing on both quality rating agreement and evidence selection overlap. Specifically, advanced GPT models (GPT-4o and o1) were prompted using Chain-of-Thought (CoT) reasoning to analyze transcripts sequentially by extracting textual evidence, classifying question types, and assigning ratings. Analyzing 28 lesson transcript segments from the Global Teaching Insights study, each carefully annotated with highlights related to questioning practices, we found that CoT prompting generally improved both rating and evidence alignment compared to basic instructions. Under CoT reasoning, GPT-4o achieved the highest Quadratic Weighted Kappa score of 0.33 for quality rating agreement, whereas o1-extracted evidence yielded the highest character-level Intersection over Union of 0.14 with human transcript annotations. Qualitative analyses revealed that LLM and human annotations aligned in identifying explicit questioning forms, but they differed in annotation scope and granularity. Our study highlights LLMs' potential to enhance the explainability of rating decisions, assist manual assessment by highlighting relevant discourse evidence, and suggest possible approaches to offer teachers specific feedback that goes beyond numerical scores.

1 Introduction

Observing teaching practices in classrooms provides a crucial approach to assessing teaching quality and promoting teachers' professional development (Pianta and Hamre, 2009; Seidel and Shavelson, 2007; Praetorius et al., 2025). To systematically evaluate the quality of teacher-student interactions, multiple classroom observation protocols have been developed over the past decades, such as the Classroom Assessment Scoring System (CLASS) (Pianta et al., 2008) and the Protocol for Language Arts Teaching Observations (PLATO) (Grossman et al., 2013). These structured protocols typically assess multiple facets of teaching dynamics, e.g., instructional practices, emotional support, and classroom management. To code such protocols, trained raters capture important events of classroom interactions from videotaped lessons and assign scores to pre-defined teaching quality dimensions based on the observed evidence within a lesson segment. Due to the complex nature of classroom interactions, this manual observation process often requires substantial human effort and time. In addition, raters typically undergo intensive training and pass quality control checks to ensure rating reliability. Given these resource-intensive demands, automated assessment approaches using artificial intelligence (AI) techniques could enhance both the scalability of classroom observation studies and the frequency of feedback provided to teachers.

Rapid advancements in Large Language Models (LLMs) have demonstrated remarkable capabilities in understanding and analyzing natural language, leading to their expanding applications across diverse fields (Yang et al., 2024). In educational contexts, LLMs have been explored for various tasks to foster teaching and learning (Kasneci et al., 2023), including essay writing assessment (Seßler et al., 2023, 2024), teachable agents for programming (Ma et al., 2024), and instructional plan genera-

Proceedings of the Artificial Intelligence in Measurement and Education Conference (AIME-Con) – Volume 1: Full Papers, pages 239–249

tion (Hu et al., 2024). Meanwhile, the emerging paradigm of "LLMs-as-Judges" has gained increasing attention for leveraging LLMs as evaluators in complex tasks (Gu et al., 2024), presenting new possibilities for supporting classroom observation procedures. Recent studies have explored the use of LLMs for assessing teaching practices through the analysis of lesson transcripts (Wang and Demszky, 2023; Tran et al., 2024). Whereas in these studies, LLMs are prompted to provide supporting evidence before generating ratings, their evaluations focus primarily on the consistency between model-generated and human-assigned scores, leaving the alignment of evidence-based rationales unexplored. As classroom discourse often involves complex interactions, understanding how LLMs analyze teaching dynamics and justify their rating decisions is critical for validating their assessment capabilities and ensuring the interpretability and trustworthiness of automated evaluations.

In this study, we investigate rating agreement and the correspondence between LLM-identified and human-documented evidence in assessing teacher questioning practices. The analysis is based on lesson transcript data from the German subset of Global Teaching Insights (GTI) (OECD, 2020), a large-scale international classroom observation study. These transcripts document authentic mathematics instruction dialogues and contain detailed annotations from trained raters, including highlighted text spans that reflect specific teaching practices. We focus on the Questioning component within the GTI observation protocol (Bell et al., 2018a), as questioning practices can be effectively analyzed using text transcripts alone, whereas other dimensions (e.g., social-emotional support) may rely more on non-verbal cues such as visual behaviors. Furthermore, teacher questioning serves as a prominent feature of classroom discourse, and effective questioning practices play a pivotal role in student cognitive engagement and learning outcomes (Redfield and Rousseau, 1981; Chin, 2007). The GTI Questioning component reflects the cognitive demand of teacher questions, emphasizing how they engage students at different levels of cognitive complexity and depth of processing information. To automate assessment, we leverage the zero-shot capabilities of advanced GPT models (GPT-4o and o1), using Chain-of-Thought (CoT) reasoning (Wei et al., 2022) to guide them in identifying key evidence, categorizing question types, and assigning ratings. Afterward, we analyze human and LLM alignment in holistic ratings, evidence overlap at both character and span levels, and questioning classification. This comprehensive analysis reveals how automated assessment approaches correspond to authentic human rating practices, providing insights into LLMs' potential to support manual observation processes and offer teachers concrete feedback.

2 Related Work

2.1 Automated Questioning Identification

Early research in automated identification of teacher questions focused on analyzing classroom discourse transcripts and audio recordings. (Donnelly et al., 2017) trained machine learning (ML) models that combined linguistic, acoustic, and context features to identify whether a teacher utterance constitutes a question. (Kelly et al., 2018) developed automated methods to estimate the proportion of authentic questioning in a class period. Recent studies have advanced beyond binary question detection to analyze specific questioning strategies. (Alic et al., 2022) employed both supervised and unsupervised ML approaches to distinguish between "funneling" questions that guide students toward normative answers and "focusing" questions that encourage deeper thinking. Similarly, (Datta et al., 2023) leveraged pre-trained language models to classify teacher questions into four categories: probing, procedural, expository, and others. Moreover, (Kupor et al., 2023) fine-tuned GPT models to identify various instructional talk moves, including questioning strategies such as eliciting and probing student ideas. Whereas these studies demonstrated advances in automated questioning identification, they primarily treated questions as isolated instances rather than analyzing patterns of questioning practices within the broader context of classroom discourse.

2.2 LLMs in Teaching Practice Assessment

With recent progress in LLMs, researchers have explored their potential for automated assessment of teaching practices by analyzing classroom transcripts. One of the earliest studies in this area was conducted by (Wang and Demszky, 2023), who leveraged GPT-3.5's zero-shot capability with different prompting methods to evaluate several aspects of teaching practices based on the CLASS and Mathematical Quality Instruction (MQI) (Hill et al., 2008) frameworks. In addition to score

prediction, they prompted GPT-3.5 to identify exemplary and problematic examples for each assessed dimension and to generate suggestions for how teachers could enhance student reasoning within the given classroom discourse. Analyzing a dataset of 100 transcript segments, they found that model-predicted scores showed generally poor alignment with human-assigned ratings and that employing CoT reasoning did not improve performance. Following this direction, (Hou et al., 2024) demonstrated that the more advanced GPT-4 model achieved superior zero-shot performance compared to GPT-3.5 in assessing classroom social-emotional support levels. Moreover, (Tran et al., 2024) examined how different task formulations affect LLMs' performance in evaluating classroom discussion quality, finding that explicitly guiding LLMs to extract relevant dialogue turns improved rating accuracy. Additionally, (Whitehill and LoCasale-Crouch, 2024) proposed a novel approach to assessing CLASS scores by prompting Llama 2 (Touvron et al., 2023) to identify behavioral indicators in individual teacher utterances, then aggregating these to predict holistic scores via linear regression. Their results showed that this automated approach could approach human inter-rater reliability while providing explanations at the utterance level.

Whereas existing studies explored prompting LLMs to extract relevant utterances beyond rating scores, they either focused solely on rating performance (Hou et al., 2024; Tran et al., 2024) or validated the extracted evidence through external reviewers (such as recruited teachers in (Wang and Demszky, 2023) and authors themselves in (Whitehill and LoCasale-Crouch, 2024)) rather than examining their alignment with trained raters directly involved in the protocol coding process. In this context, our study contributes to this line of research by (1) conducting a comprehensive analysis of both rating alignment and evidence selection overlap between LLM outputs and human annotations, (2) examining whether CoT reasoning enhances LLMs' capability to replicate human assessment procedures, and (3) providing insights into the similarities and differences between LLM and human approaches in extracting discourse evidence to reason their rating decisions.

3 Methodology

As illustrated in Figure 1, our study investigates the extent to which LLM reasoning aligns with manual annotations in assessing teacher questioning practices, comparing both rating assignments and the evidence selected from lesson transcripts to justify these decisions. This fosters the comprehension of LLM explainability, which is critical for the trustworthy use of AI in educational contexts.

3.1 Dataset

Our analysis utilized data from Germany, one of the participating countries in the Global Teaching Insights (GTI) study (formerly known as Teaching and Learning International Survey–Video (TALIS–Video)) (OECD, 2020). The GTI study systematically collected extensive classroom data on the teaching of quadratic equations and conducted a comprehensive global analysis of effective instructional practices. The German data includes 100 lesson recordings from 50 classrooms with 1,140 students across 38 schools. Anonymized lesson transcripts were created manually from the videos, with timestamps and speaker identifiers (e.g., "L" for teachers, "S01", "S02" for students).

Grounded in the GTI video observation protocol (Bell et al., 2018a), each lesson was divided into 16-minute segments, with each segment rated by intensively trained raters on a 1-4 scale across six domains. Each domain contains three components of instructional practice. Our study focused on the Questioning component, an important element of the Discourse domain. Effective questioning practices that facilitate student learning require students to engage across multiple levels of cognitive reasoning, with particular emphasis on higher-order thinking skills (Henningsen and Stein, 1997). To this end, the GTI Questioning component evaluates the cognitive demands of teacher questions, categorizing them into three types: (1) questions that prompt students to recall information, report answers, provide yes/no responses, or define terms; (2) questions that require students to summarize, explain, classify, or apply rules, processes, or formulas; and (3) questions that challenge students to analyze, synthesize, justify, or conjecture. Table 1 presents examples of each question type. The four-point rating scale reflects the relative emphasis and distribution of these question types throughout a lesson segment, with higher ratings indicating a greater proportion of more cognitively demanding questions (see Figure 3).

During video observation, raters were instructed to use the lesson transcripts as auxiliary tools, in which they highlighted relevant text spans and

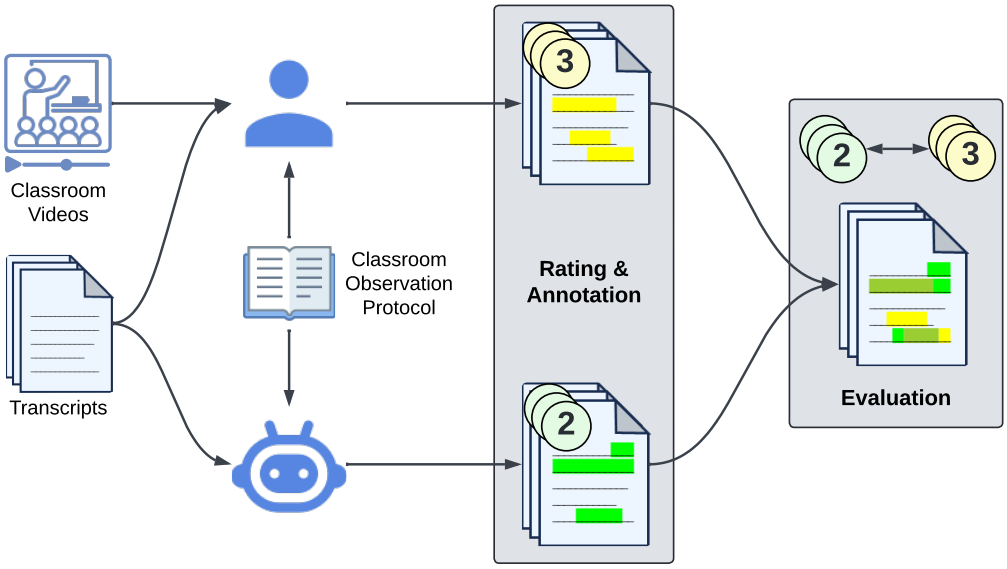

Figure 1. An overview of our study. LLMs are prompted to evaluate instructional questioning practices on a four-point scale and identify relevant evidence excerpts from transcripts. We then analyze the alignment between LLMs' assigned ratings and text selections with those provided by a trained rater.

Questions that request students to...	**Examples**
recall, report an answer, provide yes/no answers, and/or define terms	(1) What did you get Patrick? (2) What is the equivalence principle? (3) What is A? B? C? (4) Did you understand that explanation? (5) Do you remember what we did yesterday?
summarize, explain, classify, or apply rules, processes, or formulas	(1) Can you tell me how did you get this answer? (2) Let's see if substituting 4 and 8 each into x in equation 2 would work. Why do we substitute 4 and 8? (3) How many conditions do the roots of quadratic equations with one unknown have? What are they?
analyze, synthesize, justify, or conjecture	(1) The perimeter of a rectangle is 20. What is the area of the rectangle? (2) What is the pattern you notice across the three problems we just solved? Look carefully. (3) Can you explain why you disagree? Why do you think completing the square is a more efficient approach than just using the quadratic equation for number 4 on the board?

Table 1. Three questioning types and their examples (Bell et al., 2018b).

```
0:07:40  T:  yes exactly no that is thus the case where we have no
             solution when do we get exactly one solution from our
             quadratic equation DC_QUEST(recall) S05
```

Figure 2. Visualization of an authentic manual annotation example (translated word-for-word from German to English for readability). DC_QUEST means the Questioning component in the Discourse domain. T: Teacher; S05: Student.

specified the corresponding components. As these transcript annotations served primarily as working notes to support the rating process rather than as systematically standardized documentation, individual raters could vary in their annotation styles. Some raters provided detailed evidence highlights for certain components while annotating others sparsely, reflecting individual documentation preferences. This authentic variation, although valuable for understanding real-world assessment practices, presented challenges for conducting a unified analysis across all raters. Therefore, to enable a focused analysis of questioning practices, we selected data from one rater who provided detailed textual evidence for the Questioning component, including fine-grained annotations that categorized questioning strategies. This resulted in a dataset of 8 lessons containing 28 segments and 149 highlighted text spans. Figure 2 illustrates an annotation example. Additionally, as each segment was assessed by two randomly assigned raters, we calculated the selected rater's agreement with others who evaluated the same set of segments for the Questioning component, resulting in a Quadratic Weighted Kappa (QWK) score of 0.48. To contextualize this agreement level, we extended the leave-one-rater-out analysis to all 14 raters over the whole GTI Germany dataset. The average QWK score across all raters was 0.23, indicating that the chosen rater exhibited relatively high consistency with others in evaluating questioning practices. Moreover, we processed the annotated transcripts using INCEpTION (Klie et al., 2018), an open-source text annotation platform, to convert the original annotations into a structured dataset by pairing highlighted text spans with corresponding question categories for subsequent comparison with LLM annotations.

3.2 ChatGPT Zero-Shot Annotation

We employed two state-of-the-art GPT models[1], GPT-4o (*gpt-4o-2024-11-20*) and o1 (*o1-2024-12-17*), to automatically assess questioning practices while extracting supporting evidence from classroom dialogues. For each model, we investigated two zero-shot prompting strategies: a basic prompt and Chain-of-Thought (CoT) (Wei et al., 2022) reasoning. The basic prompt (see Figure 3) positioned LLMs as expert raters in evaluating teaching practices and outlined multiple rating guidelines, including examples of the three questioning types (see Table 1) and detailed scoring criteria for each rating level. Additionally, the prompt incorporated a 16-minute segment transcript for evaluation and instructed LLMs to provide a rating score along with a list of supporting evidence in the form of verbatim excerpts. Model responses were required in JSON format to ensure consistency and facilitate systematic analysis. The CoT prompt maintained all elements of the basic prompt but introduced explicit procedural analysis steps shown in Figure 4. Besides rating assignment and evidence extraction, the model was required to classify each identified evidence excerpt into one of three predefined question categories in its JSON response. This additional categorization enabled a more granular analysis of instructional questioning practices.

We accessed both GPT models via the OpenAI API and used default hyperparameters for inference. Due to the variability in LLM outputs, we conducted three independent runs for each experimental setting and averaged their evaluation results.

3.3 Evaluation Metrics

To understand how closely automated assessments correspond to authentic human annotations in evaluating teacher questioning practices, we examined three aspects: (1) rating alignment, (2) textual evidence overlap, and (3) questioning categorization consistency. The used metrics are described below.

(1) First, we utilized Quadratic Weighted Kappa (QWK) to measure agreement between model-generated and human-assigned scores. QWK accounts for the ordinal nature of rating levels by penalizing larger disagreements more heavily than smaller ones. (2) Second, we explored the degree of overlap between model-extracted and human-highlighted evidence spans at both character and span levels. At the character level, for a given transcript segment, we mapped each evidence excerpt (from either the model or the human rater) to its respective position in the transcript by identifying its start and end character indices. Afterward, we calculated the Intersection over Union (IoU) between model-extracted and human-annotated evidence sets, defined as the ratio of overlapping characters relative to the total number of characters covered by either set. However, interactive classroom discourse differs from structured written text, often involving fragmented, overlapping, and context-dependent utterances. As a result, annotators could include varying amounts of surrounding context when marking the same evidence. This in-

[1] https://platform.openai.com/docs/models

Task
You are an expert in evaluating the quality of classroom interactions based on lesson transcripts. You will be provided with a German transcript of a mathematics lesson segment focusing on quadratic equations. The transcript includes timestamps and speaker annotations, with 'L' indicating the teacher and 'S' followed by an ID number identifying anonymous students. Your task is to assess the teaching quality dimension of 'Questioning', which evaluates the nature of the questions asked by teachers that request students engage in a range of types of cognitive reasoning.

Important Note
* Rhetorical questions (i.e., questions the teacher poses and either does not answer or answers him or herself) should not be counted during rating.
* The rater should focus on what kinds of questions characterize the segment.
* Here are three types of questions with examples:
    {*Examples of three question types*}

Rating Scale (1-4, low to high)
* Score 1: Questions generally request students recall, report an answer, provide yes/no answers, and/or define terms.
* Score 2: Questions generally request students recall, report an answer, provide yes/no answers, and/or define terms, although there are some questions that request students summarize, explain, classify, or apply rules, processes, or formulas.
* Score 3: Despite a few questions that request students recall, report, and/or define, most questions request that students summarize, explain, classify, or apply rules, processes, or formulas. There may be a small number of questions that request students analyze, synthesize, justify, or conjecture.
* Score 4: Questions request a mixture of recall, reporting, defining, summarizing, explaining, classifying, applying rules, processes, or formulas, analyzing, synthesizing, justifying, and/or conjecturing, but the emphasis is on questions that request students analyze, synthesize, justify, or conjecture.

Instructions
Provide a JSON response with a rating and a list of key evidence supporting your rating:
{
  "rating": <integer score>,
  "evidence": ["<exact quote 1>", "<exact quote 2>", ...]
}
Note: Use exact character-for-character text spans from the provided transcript as complete evidence. Do NOT modify, paraphrase, abbreviate, or omit any content (including punctuation and formatting).

Transcript
Below is the transcript to be rated, enclosed in triple backticks:
```{*Transcript*}```

Figure 3. Basic prompt including comprehensive coding rubrics and response instructions. {*Examples of three question types*} can be found in Table 1.

...
Instructions
Analyze the transcript following these steps:
1. Read the transcript carefully,
2. Identify teacher questions maching the three aforementioned types,
3. Rate the transcript based on the coding rubrics.
Provide a JSON response with a rating and a list of relevant teacher questions along with their types:
{
  "rating": <integer score>,
  "evidence":
    [{"question": <exact quote 1>, "type": <1, 2, or 3>}, ...]
}
Note: Use exact character-for-character text spans from the provided transcript as complete evidence. Do NOT modify, paraphrase, abbreviate, or omit any content (including punctuation and formatting).
...

Figure 4. Chain-of-Thought prompt (sharing identical rating guidelines with the basic prompt).
```

herent variation in annotation granularity made exact character-level alignment overly restrictive. To address this, we adopted a more flexible matching criterion at the span level. A model-extracted span was considered a match with a manual annotation if they overlapped within the same general region of the transcript. Based on these counts, we calculated Precision (i.e., proportion of matches among model-extracted spans), Recall (i.e., proportion of matches among human-annotated spans), and F1-Score (i.e., harmonic mean of Precision and Recall). The resulting metrics were averaged across all transcript segments. (3) Finally, for matched evidence spans, we evaluated the classification of question types utilizing weighted Precision (i.e., how many of model-predicted question types were correct), Recall (i.e., how many of human-labeled question types were identified), and F1-Score.

4 Results

Table 2 presents the results on the consistency between LLM- and human-assigned ratings. GPT-4o with the CoT prompt achieved the highest agreement with human ratings (QWK=0.33), yielding notable improvement over its basic prompt counterpart. In contrast, the o1 model showed similar QWK scores for both basic and CoT prompts. Subsequently, we analyzed the overlap between model-extracted and human-annotated evidence at both character and span levels, along with their question type categorization. The results are summarized in Table 3. Regarding the selection of textual evidence, CoT prompting generally yielded higher character-level overlap than the basic prompt across both models, with the o1 CoT variant achieving the highest IoU of 0.14. At the span level, o1 resulted in stronger alignment across all metrics compared to GPT-4o under both prompting conditions. Similar to the rating results, the o1 model maintained a consistent F1 score of 0.38 between basic and CoT prompts. For question type categorization, under CoT prompting, the o1 model achieved a higher F1-Score than GPT-4o in distinguishing questioning evidence across three levels of cognitive demand.

In addition to the quantitative analysis, Figure 5 illustrates an exemplary comparison between human and LLM (o1-CoT) evidence selections from a transcript excerpt. The excerpt presents classroom discourse at the start of a math lesson, where the teacher reviews quadratic equations through a series of questions. The visualization captures both areas of convergence and divergence between human and model annotations. The overlapping regions (lime green) indicate alignment in identifying explicit instructional questioning while also revealing variations in the amount of surrounding context included by annotators when highlighting the same evidence. Whereas model-specific annotations (pure green) frequently include direct references to individual students (e.g., "S18") and focus on explicit computational prompts such as "yes and q," human-only annotations (yellow), such as "now we substitute this into...," tend to reflect a more contextualized and indirect questioning approach directed at students. Moreover, this comparison reveals that both model and human annotations occasionally overlook certain questions, leaving them unmarked and thus excluded from the assessment (e.g., "-6/2 is how much").

5 Discussion

In this study, we explored the zero-shot capabilities of advanced LLMs in evaluating instructional questioning practices from classroom transcripts. Our comparative analysis between LLM judgments and authentic human annotations revealed both areas of alignment and notable discrepancies in assessment approaches. The overall inter-rater agreement (QWK=0.23, see Sect. 3.1) in the GTI Germany dataset for the Questioning component underscores the inherent subjectivity of classroom observation over 16-minute instruction segments, suggesting that even trained human raters may differ in their interpretations of questioning practices. Within this context, the alignment between LLMs and the selected rater (QWK up to 0.33), whereas falling below the selected rater's agreement with other raters (QWK=0.48) on the same segment set, indicates the potential of LLMs to serve as assistive tools in classroom observation. Our results showed a moderate overlap in evidence selections (IoU up to 0.14, F1 up to 0.38) between LLM outputs and human annotations. Through analysis of multiple transcript examples, we observed that, whereas LLMs and the human rater aligned in identifying explicit instructional questions, they differed in annotation scope and granularity. LLMs tended to provide more comprehensive coverage of questioning instances throughout the discourse and include more surrounding dialogue around each question. It is important to consider the nature of human-annotated transcripts, which were created as prac-

Model	Prompt Type	QWK
Human Raters	-	0.48
GPT-4o	Basic	0.17
	CoT	0.33
o1	Basic	0.22
	CoT	0.23

Table 2. Results of agreement between LLM-generated scores and human-assigned ratings. For human raters, QWK is computed by comparing the chosen rater's scores with those of other raters who evaluated the same set of segments.

Model	Prompt Type	Evidence Overlap				Question Categorization		
		IoU	Recall	Precision	F1	Recall	Precision	F1
GPT-4o	Basic	0.06	0.14	0.23	0.18	-	-	-
	CoT	0.10	0.21	0.29	0.24	0.44	**0.67**	0.50
o1	Basic	0.09	0.33	**0.45**	**0.38**	-	-	-
	CoT	**0.14**	**0.52**	0.29	**0.38**	**0.61**	0.55	**0.58**

Table 3. Results of LLM and human alignment in evidence selections and question categorization. Bold values indicate the highest alignment for each metric.

0:00:00 T: I wish you all a very good morning (C: good morning) so we continue with quadratic equations we started in the last lesson solving these equations using the solution formula (.) do you still remember what the solution formula is if not take a look in your folder in the notebook S18 (S18: x1 x2 = uhm =-p/2²+/- uhm √(p/2)²-q) exactly you have already solved some exercises that were in normal form let's do one more example x²+6x+8=0 (.) what is p in this equation (.) S06 (S06: 6) yes and q S06 (S06: 8) yes (.) now we substitute this into the solution formula who will dictate the equation for me S15 (S15: uhm x1,2 uhm =- uhm 6/2) mhm (S15: +/- uhm √ uhm (uhm 6/2)²-8) exactly -6/2 is how much S17 (S17: - 3) -3 (.) how large is this term use a calculator if necessary (.) 6/2²-8 S19 (S19: 1) yes so -3+/-1 (.) what is then x1 S17 (S17: -2) say the full calculation (S17: uhm -3+1) exactly (S17: equals minu-) and x2 (S17: -3-1 (.) =-4) correct so our solution set is -2 and -4

Figure 5. Visualization of evidence overlap between human and model (o1-CoT) annotations in a transcript excerpt (translated from German to English). Yellow: human-only annotations; Pure green: model-only annotations; Lime green: overlapping selections. T: Teacher; C: Class; S06, S18, ...: Student.

tical notes rather than exhaustive documentation. Given that human raters assessed multiple components simultaneously, they may prioritize salient evidence that most effectively supports their evaluations, potentially omitting trivial or redundant instances. For example, in Figure 5, whereas the model identified some short questions, human annotations tended to focus on more substantive spans that warranted explicit documentation. This selective nature reflects authentic annotation practices in real-world classroom observation settings.

Moreover, our findings revealed distinct patterns in how GPT-4o and o1 performed under basic and CoT prompting. For GPT-4o, CoT reasoning resulted in higher agreement with human annotations across all metrics compared to the basic instruction, indicating that explicit guidance through structured reasoning steps helps LLMs better approximate human assessment practices. In contrast, o1 yielded comparable alignment levels between basic and CoT prompts, possibly attributed to o1's intrinsic reasoning mechanism that enables it to infer and apply implicit procedural analysis from basic instructions paired with rating guidelines. Further, whereas GPT-4o with CoT achieved a higher rating agreement than o1, o1 led to greater consistency with human annotations in evidence selections and question type classification. This discrepancy suggests potential differences in how human raters and LLMs translate identified evidence into holistic scores, indicating the complexity of teaching quality assessment. Additionally, o1 generates numerous intermediate tokens during its internal reasoning process before producing a final response, resulting in longer inference time and higher API costs. Thus, GPT-4o with CoT prompting might present a more balanced solution, offering a trade-off between accuracy and efficiency.

Our findings suggest promising implications for integrating LLMs into classroom observation. Unlike traditional automated methods (Ramakrishnan et al., 2021; James et al., 2018) that often lack explainability, LLMs can explicitly identify relevant discourse evidence to support their ratings. Thus, they could serve as complementary raters by suggesting highlights in transcripts for human raters to validate or refine. This human-in-the-loop approach would reduce the manual workload of sifting through lengthy transcripts while maintaining expert-level judgment. Beyond assisting manual observation, LLM-generated evidence-based assessments could form the basis for systems that provide teachers with timely feedback, offering both holistic ratings and representative examples of teaching practices (e.g., high-quality questions that promote students' high-order thinking). Moreover, these automated analyses could be valuable as training materials for novice raters by providing annotated instances of different questioning types and their impact on instruction. However, given the high-stakes nature of teaching quality assessment and potential algorithmic biases, it is crucial to recognize that LLMs should complement, rather than replace, manual coding or professional development resources. This allows raters and teachers to choose how to utilize these automated annotations to refine professional skills at their own pace.

One limitation of our study is its focus solely on the GTI Questioning component. While this verbally-oriented practice suits transcript analysis, the generalizability of our findings to other instructional components (e.g., teacher feedback) remains to be explored. Moreover, although the selected rater exhibited above-average agreement with peers, expanding the dataset to include more raters would allow for a more comprehensive understanding of human annotation variability. Further, while CoT reasoning showed promising results, future research could benefit from employing more sophisticated promoting engineering strategies, like in-context learning. Additionally, with the rise of open-source models such as Llama (Touvron et al., 2023), future work could explore their applications to classroom observation tasks, offering potential alternatives to closed-source models and enabling large-scale studies at reduced operational costs.

6 Conclusion

This study examines the alignment between LLM- and human-generated assessments of teacher questioning practices, involving both rating assignments and evidence extraction. CoT prompting proves effective in guiding LLMs to approximate human assessment procedures. Although LLM and human annotations exhibit different patterns in granularity and context inclusion, these variations highlight complementary approaches to identifying instructional events. Our findings suggest LLMs' potential to enhance the explainability and trustworthiness of rating decisions and to facilitate manual observation and foster teacher professional growth in future applications.

References

Sterling Alic, Dorottya Demszky, Zid Mancenido, Jing Liu, Heather Hill, and Dan Jurafsky. 2022. Computationally identifying funneling and focusing questions in classroom discourse. *arXiv preprint arXiv:2208.04715*.

Courtney Bell, Yi Qi, Margaret Witherspoon, Mariana Barragan, and Heather Howell. 2018a. Annex a: Talis video observation codes: Holistic domain ratings and components. In *Global Teaching Insights: Technical Report*. OECD.

Courtney Bell, Yi Qi, Margaret Witherspoon, Mariana Barragan, and Heather Howell. 2018b. Annex a: Talis video training notes: Holistic domain ratings and components. In *Global Teaching Insights: Technical Report*. OECD.

Christine Chin. 2007. Teacher questioning in science classrooms: Approaches that stimulate productive thinking. *Journal of Research in Science Teaching: The Official Journal of the National Association for Research in Science Teaching*, 44(6):815–843.

Debajyoti Datta, James P Bywater, Maria Phillips, Sarah Lilly, Jennifer L Chiu, Ginger S Watson, and Donald E Brown. 2023. Classifying mathematics teacher questions to support mathematical discourse. In *International Conference on Artificial Intelligence in Education*, pages 372–377. Springer.

Patrick J Donnelly, Nathaniel Blanchard, Andrew M Olney, Sean Kelly, Martin Nystrand, and Sidney K D'Mello. 2017. Words matter: automatic detection of teacher questions in live classroom discourse using linguistics, acoustics, and context. In *Proceedings of the Seventh International Learning Analytics & Knowledge Conference*, pages 218–227.

Pam Grossman, Susanna Loeb, Julie Cohen, and James Wyckoff. 2013. Measure for measure: The relationship between measures of instructional practice in middle school english language arts and teachers' value-added scores. *American Journal of Education*, 119(3):445–470.

Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, and 1 others. 2024. A survey on llm-as-a-judge. *arXiv preprint arXiv:2411.15594*.

Marjorie Henningsen and Mary Kay Stein. 1997. Mathematical tasks and student cognition: Classroom-based factors that support and inhibit high-level mathematical thinking and reasoning. *Journal for research in mathematics education*, 28(5):524–549.

Heather C Hill, Merrie L Blunk, Charalambos Y Charalambous, Jennifer M Lewis, Geoffrey C Phelps, Laurie Sleep, and Deborah Loewenberg Ball. 2008. Mathematical knowledge for teaching and the mathematical quality of instruction: An exploratory study. *Cognition and instruction*, 26(4):430–511.

Ruikun Hou, Tim Fütterer, Babette Bühler, Efe Bozkir, Peter Gerjets, Ulrich Trautwein, and Enkelejda Kasneci. 2024. Automated assessment of encouragement and warmth in classrooms leveraging multimodal emotional features and chatgpt. In *International Conference on Artificial Intelligence in Education*, pages 60–74. Springer.

Bihao Hu, Longwei Zheng, Jiayi Zhu, Lishan Ding, Yilei Wang, and Xiaoqing Gu. 2024. Teaching plan generation and evaluation with gpt-4: Unleashing the potential of llm in instructional design. *IEEE Transactions on Learning Technologies*.

Anusha James, Mohan Kashyap, Yi Han Victoria Chua, Tomasz Maszczyk, Ana Moreno Núñez, Rebecca Bull, and Justin Dauwels. 2018. Inferring the climate in classrooms from audio and video recordings: a machine learning approach. In *IEEE International Conference on Teaching, Assessment, and Learning for Engineering*, pages 983–988.

Enkelejda Kasneci, Kathrin Seßler, Stefan Küchemann, Maria Bannert, Daryna Dementieva, Frank Fischer, Urs Gasser, Georg Groh, Stephan Günnemann, Eyke Hüllermeier, and 1 others. 2023. Chatgpt for good? on opportunities and challenges of large language models for education. *Learning and individual differences*, 103:102274.

Sean Kelly, Andrew M Olney, Patrick Donnelly, Martin Nystrand, and Sidney K D'Mello. 2018. Automatically measuring question authenticity in real-world classrooms. *Educational Researcher*, 47(7):451–464.

Jan-Christoph Klie, Michael Bugert, Beto Boullosa, Richard Eckart De Castilho, and Iryna Gurevych. 2018. The inception platform: Machine-assisted and knowledge-oriented interactive annotation. In *Proceedings of the 27th international conference on computational linguistics: System demonstrations*, pages 5–9.

Ashlee Kupor, Candice Morgan, and Dorottya Demszky. 2023. Measuring five accountable talk moves to improve instruction at scale. *arXiv preprint arXiv:2311.10749*.

Qianou Ma, Hua Shen, Kenneth Koedinger, and Sherry Tongshuang Wu. 2024. How to teach programming in the ai era? using llms as a teachable agent for debugging. In *International Conference on Artificial Intelligence in Education*, pages 265–279. Springer.

OECD. 2020. *Global Teaching InSights: A Video Study of Teaching*. OECD, Paris.

Robert C Pianta and Bridget K Hamre. 2009. Conceptualization, measurement, and improvement of classroom processes: Standardized observation can leverage capacity. *Educational researcher*, 38(2):109–119.

Robert C Pianta, Karen M La Paro, and Bridget K Hamre. 2008. *Classroom Assessment Scoring System™: Manual K-3.* Paul H Brookes Publishing.

Anna-Katharina Praetorius, Charalambos Y Charalambous, Svenja Vieluf, Mirjam Steffensky, Richard Göllner, and Benjamin Fauth. 2025. Rethinking teaching-quality research: a reflection on the role of core working assumptions and possible pathways for future research. *School Effectiveness and School Improvement*, 36(2):314–334.

Anand Ramakrishnan, Brian Zylich, Erin Ottmar, Jennifer LoCasale-Crouch, and Jacob Whitehill. 2021. Toward automated classroom observation: Multimodal machine learning to estimate class positive climate and negative climate. *IEEE Transactions on Affective Computing*.

Doris L Redfield and Elaine Waldman Rousseau. 1981. A meta-analysis of experimental research on teacher questioning behavior. *Review of educational research*, 51(2):237–245.

Tina Seidel and Richard J Shavelson. 2007. Teaching effectiveness research in the past decade: The role of theory and research design in disentangling meta-analysis results. *Review of educational research*, 77(4):454–499.

Kathrin Seßler, Maurice Fürstenberg, Babette Bühler, and Enkelejda Kasneci. 2024. Can ai grade your essays? a comparative analysis of large language models and teacher ratings in multidimensional essay scoring. *arXiv preprint arXiv:2411.16337*.

Kathrin Seßler, Tao Xiang, Lukas Bogenrieder, and Enkelejda Kasneci. 2023. Peer: Empowering writing with large language models. In *European Conference on Technology Enhanced Learning*, pages 755–761.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, and 1 others. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Nhat Tran, Benjamin Pierce, Diane Litman, Richard Correnti, and Lindsay Clare Matsumura. 2024. Analyzing large language models for classroom discussion assessment. *arXiv preprint arXiv:2406.08680*.

Rose E Wang and Dorottya Demszky. 2023. Is chatgpt a good teacher coach? measuring zero-shot performance for scoring and providing actionable insights on classroom instruction. *arXiv preprint arXiv:2306.03090*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, and 1 others. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Jacob Whitehill and Jennifer LoCasale-Crouch. 2024. Automated evaluation of classroom instructional support with llms and bows: Connecting global predictions to specific feedback. *Journal of Educational Data Mining*.

Jingfeng Yang, Hongye Jin, Ruixiang Tang, Xiaotian Han, Qizhang Feng, Haoming Jiang, Shaochen Zhong, Bing Yin, and Xia Hu. 2024. Harnessing the power of llms in practice: A survey on chatgpt and beyond. *ACM Transactions on Knowledge Discovery from Data*, 18(6):1–32.