Developing an LLM-specific Coding Framework for Prediction and Feedback Modeling of Science Teachers' Instructional Expertise

Jin Eun Yoo
Korea National University of Education
Department of Education
250 Taeseongtabyeon-ro, Cheongju, Republic of Korea
jeyoo@knue.ac.kr

Nam-Hwa Kang
Korea National University of Education
Department of Physics Education
250 Taeseongtabyeon-ro, Cheongju, Republic of Korea
nama.kang@knue.ac.kr

Suna Ryu
Korea National University of Education
Department of Chemistry Education
250 Taeseongtabyeon-ro, Cheongju, Republic of Korea
sunaryu@knue.ac.kr

Jun-ki Lee
Jeonbuk National University
Institute of Science Education
567 Baekje-daero, Jeonju, Republic of Korea.
junki@jbnu.ac.kr

Youngsun Kwak
Korea National University of Education
Department of Earth Science Education
250 Taeseongtabyeon-ro, Cheongju, Republic of Korea
kwak@knue.ac.kr

Hyeong Gwan Kim
Korea National University of Education
Department of Education
250 Taeseongtabyeon-ro, Cheongju, Republic of Korea
milphy@nate.com

Abstract. *This study introduces a novel coding framework for classroom discourse tailored to large language models (LLMs). To address limitations of prior studies relying on global-level observation tools, this framework adopts utterance-level analysis and fine-grained coding. Predicate-level chunking, coding, and rating were collaboratively conducted using a dual-expertise approach, involving teams of science education faculty and experienced teachers, across 125 science sessions from 27 teachers. Over the course of the process, a glossary of terms was also developed via a primarily bottom-up approach. Quadratic Weighted Kappa results demonstrated reasonable rater consistency, supporting the validity of the framework.*

Keywords. Large Language Models (LLMs), classroom discourse, science education, instructional analysis, utterance-level coding, coding framework, rater consistency, teacher expertise, science teacher

1 Introduction

Teachers play a pivotal role in education. It is widely acknowledged that 'the quality of education cannot exceed the quality of its teachers,' underscoring how teachers' instructional and assessment practices influence student achievement. In particular, teacher language has long functioned as a primary tool for instructional delivery in classroom discourse. For example, the specific lexical choices, the timing of verbal interventions, and the form of feedback provided can have a decisive impact on students' learning. Classroom discourse studies are grounded in this perspective. Experts in traditional studies typically evaluate one or more complete lessons—either through direct observation or via recorded sessions—using established observation frameworks. While such approaches allow for a holistic evaluation of instruction, they heavily rely on subjective judgment, lack scalability, and tend to concentrate on exemplary teachers—thus limiting their applicability to the

practical needs of teachers with diverse backgrounds and varying levels of experience.

Recent advances in large language models (LLMs) have opened new possibilities for analyzing and modeling classroom discourse in ways that were not feasible with traditional observation-based approaches. While LLM-based studies are increasingly applied to classroom discourse analysis, they continue to rely on structured coding schemes such as CLASS (Classroom Assessment Scoring System) or MQI (Mathematical Quality of Instruction)—widely used in traditional observation research. These frameworks were designed for global-level lesson analysis, not for the local, utterance-level applications. This misalignment presents a critical challenge to the proper granularity required for effective LLM implementation.

To address this issue, the present study aims to develop a coding scheme tailored to LLM analytical units in modeling teacher expertise prediction and feedback. Specifically, this study integrates authentic classroom discourse from in-service secondary science teachers with state-of-the-art deep learning techniques and instructional analysis methods to construct both theoretical and empirical models for predicting teacher expertise and generating formative feedback. This manuscript reports year-one findings of a three-year project, focusing on the development of the coding scheme for LLM modeling.

Conventionally, raters are responsible for assigning scores or ratings—such as on a 1-to-5 scale—whereas coders categorize or classify data according to predefined themes or categories. In the present study, the act of chunking teacher utterances based on semantic coherence is defined as coding, while assigning a numerical score to each chunked segment is regarded as rating. As the same individuals performed both tasks in sequence, the terms ‘rater’ and ‘coder’ are used interchangeably throughout this study.

2 Literature Review

A number of LLM-based studies on classroom discourse have utilized transcript data provided by the National Center for Teacher Effectiveness (NCTE). The NCTE dataset consists of classroom observation records collected and annotated between 2010 and 2013 from four school districts in the New England region. It captures mathematics instruction delivered by 317 elementary teachers to 4th and 5th grade students, most of whom came from low-income and historically marginalized backgrounds (Kane et al., 2015).

A distinguishing feature of the dataset is its inclusion of 1,660 annotated transcripts—each corresponding to a 45- to 60-minute instructional session—evaluated by expert raters using two validated classroom observation protocols: the Classroom Assessment Scoring System (CLASS; Pianta et al., 2008) and the Mathematical Quality of Instruction (MQI; Hill et al., 2008). Several prior studies, including Alic et al. (2022), Wang and Demszky (2023), Xu et al. (2024), and Hardy (2025), have employed the built-in annotated data from NCTE. Accordingly, a brief overview of the two rating instruments—CLASS and MQI—is presented below.

2.1 CLASS and MQI

The CLASS is an observational tool designed for evaluating instructional quality in K–12 classrooms. It involves segmenting a lesson into observation cycles of 15 to 20 minutes, typically conducted over two to six cycles per session. Each cycle is scored on a 7-point scale (ranging from 1 to 7) across 10 to 11 items, organized under three core domains: emotional support, classroom organization, and instructional support. CLASS emphasizes the quality of teacher–student interactions and intentionally excludes curriculum-specific content and the physical classroom environment from its evaluation criteria.

Similarly, the MQI is a structured classroom observation protocol specifically designed to assess the mathematical rigor and interactive quality of instruction in K–12 mathematics settings. A complete mathematics lesson is divided into segments of approximately 5 to 7 minutes, which are individually rated and then aggregated to yield a composite session score. The MQI rubric comprises five dimensions: richness of the mathematics, errors and imprecision, working with students and mathematics, student participation in meaning-making and reasoning, and connections between classroom work and broader mathematical ideas. Each dimension is rated using a three-level ordinal scale (low, mid, high).

2.2 Studies Using Built-in NCTE Ratings without Additional Human Raters

Among previous studies, Xu et al. (2024) and Hardy (2025) leveraged the built-in annotations provided by the NCTE dataset, without recruiting additional human raters. Xu et al. (2024) applied Transformer-based models—including BERT, DistilBERT, XL-Net, RoBERTa, and Llama2-7B with QLoRA—to 9,886 segments of MQI transcripts, each approximately 7.5 minutes in length. Based on performance metrics such as F1 score, majority F1, and Spearman’s rho, the results indicated that LLM-based models performed comparably to human raters on tasks requiring lower levels of pedagogical reasoning. However, their performance declined significantly in evaluating more complex teaching practices. Notably, the study also found that using only teacher utterances as input yielded results similar to those obtained when both teacher and student utterances were included.

Hardy (2025), on the other hand, conducted a secondary analysis focusing on four of the thirteen MQI items. The study compared the rating performance of GPT-based models and encoder-based models using evaluation criteria such as concordance,

confidence, validity, bias, fairness, and helpfulness. Analytical methods included assessments of rater consistency, generalizability and decision studies, and hierarchical rater modeling. Overall, encoder-based models outperformed GPT-based models and achieved performance levels comparable to those of human raters.

2.3 Studies Involving Additional Human Raters and Calibration Procedures

In some studies utilizing the NCTE dataset, researchers additionally recruited human raters and implemented rater training procedures, depending on the research objectives. For instance, Alic et al. (2022) focused on identifying teachers' use of *focusing* and *funneling* questions and demonstrated that a supervised RoBERTa model achieved a strong linear correlation of 0.76 with expert-coded labels. To construct the dataset, they applied three sampling criteria—mathematical relevance, follow-up on a prior student utterance, and inclusion of a question—to select 2,348 examples of teacher–student exchanges. These were then annotated using three MQI items (student-provided explanations; overall student participation and meaning-making and reasoning; and mathematical quality of instruction), with a three-category coding scheme (0 = not meeting criteria; 1 = funneling; 2 = focusing). Thirteen raters underwent structured rater training and calibration sessions, resulting in a Fleiss' Kappa of 0.415 on the randomly assigned utterance pairs.

In comparison, Wang and Demszky (2023) used the built-in annotations provided in the NCTE dataset but recruited two additional human raters to compare against ChatGPT-generated responses. Specifically, they randomly selected ten transcripts each for the CLASS and MQI instruments from the total of 1,660 transcripts. The utterances within each segment were grouped into equal-sized bins. Two mathematics teachers were then tasked with evaluating ChatGPT's zero-shot performance on a range of teacher coding tasks. Depending on the task, three or four evaluation criteria were used on a 3-point scale (yes, somewhat, no). However, the study did not provide detailed information regarding the training of these raters or how discrepancies in the actual ratings were addressed. According to Spearman's rho analyses, ChatGPT performed comparably to human raters on relatively simple instructional tasks but failed to capture more complex teaching practices effectively.

2.4 Rationale of the Study

Traditional classroom analysis studies have typically involved the collection of instructional sessions conducted by confident teachers in controlled, evaluative settings. These studies generally rely on structured observation frameworks and domain experts who analyze one or more sessions using extended time intervals—such as full-class sessions or 15-minute blocks—as the primary unit of analysis. In contrast, LLM-based studies, including the present one, analyze classroom discourse in the form of text data, often focusing on teacher utterances.

Despite their strengths in processing large-scale language data, prior research indicates that LLMs perform well on relatively simple, low-inference tasks but fall short of human-level performance on more cognitively demanding or pedagogically nuanced tasks (Wang & Demszky, 2023; Xu et al., 2024). These limitations may stem from two critical challenges: the absence of an LLM-specific evaluation framework and the misalignment between traditional units of analysis and those suitable for LLM-based approaches.

To enhance the effectiveness of LLMs in classroom discourse analysis, it is therefore necessary to shift the unit of analysis toward smaller, utterance-level segments and to develop evaluation frameworks tailored to the characteristics of LLMs. However, to date, no study has proposed a coding framework explicitly designed for LLM-based approaches. Most existing studies remain secondary analyses of data originally collected under traditional observation paradigms, typically applying global-level frameworks such as CLASS and MQI. Even in cases involving newly collected data, no novel coding schemes have been introduced that align with the analytical granularity required for effective LLM implementation.

Due to the fundamental difference in analytical approach, it was not feasible to directly apply existing science education observation frameworks. Therefore, it became essential to develop and implement a coding scheme tailored to LLM-based modeling. To conclude, a coding framework for full-class instructional sessions should be developed, incorporating coding schemes for smaller, semantically meaningful segments to enable more effective implementation of LLM-based analysis. Table 1 presents the comparison between the present study and traditional classroom discourse research.

Table 1. Comparison between the Present Study and Traditional Research

	Present Study	Traditional Research
Data	classroom transcripts composed of teacher utterances (text data)	audio-visual data, typically from direct observation
Data collection	authentic teaching practices	often collected from controlled settings
Unit of analysis	utterance-level chunked segments	full-class session or minute blocks; research question-specific

		chunking (e.g., teacher-student utterance pairs)
Coding framework	no clearly established or widely agreed-upon coding framework	structured frameworks (e.g., CLASS, MQI)
Purpose	prediction and feedback modeling of teachers' instructional expertise	evaluation and feedback on the given lesson

3 Methods

3.1 Materials

A total of 125 classroom sessions were collected from 27 middle school science teachers during the second semester of 2024 and the first semester of 2025. In South Korea, middle school science teachers typically major in one of the four subjects—physics, chemistry, life science, or earth science—but are required to teach all areas. Each participating teacher recorded at least three instructional sessions, each approximately 45 minutes long, in both their major and non-major subject areas.

Table 2 presents the distribution of instructional sessions by teachers' majors and the subjects they taught. The table details the number of participating teachers for each major and breaks down the total number of sessions they conducted across four subject areas: Physics, Chemistry, Life Science, and Earth Science. Percentages indicate the proportion of sessions within each major that were devoted to each subject. Notably, teachers frequently taught subjects outside their major fields, reflecting the interdisciplinary nature of middle school science instruction in the given context.

Table 2. Number of Teachers and Instructional Sessions by Academic Major and Subject Area

Major	# of Teachers (%)	Subjects Taught	Sessions (%)
Physics	10 (37.01%)	Physics	12 (29.27)
		Chemistry	9 (21.95)
		Life Science	14 (34.15)
		Earth Science	6 (14.63)
		Subtotal	41(100%)
Chemistry	4 (14.81%)	Physics	4 (20.00)
		Chemistry	8 (40.00)
		Life Science	6 (30.00)
		Earth Science	2 (10.00)
		Subtotal	20(100%)
Life Science	8 (29.63%)	Physics	6 (16.22)
		Chemistry	14 (37.84)
		Life Science	14 (37.84)
		Earth Science	3 (8.11)
		Subtotal	37(100%)
Earth Science	5 (18.52%)	Physics	9 (33.33)
		Chemistry	3 (11.11)
		Life Science	3 (11.11)
		Earth Science	12 (44.44)
		Subtotal	27(100%)
Total	27 (100%)		125

The collected audio data were transcribed into text data using a speech-to-text (STT) software, and each participating teacher manually corrected their transcript. For each subject area, experienced teachers with acknowledged instructional expertise were nominated to serve as coders. Eight teacher-raters participated in the study, with two each holding undergraduate degrees in physics, chemistry, life science, and earth science education. Their teaching experience ranged from over 6 to 22 years, and all were either enrolled in or had completed a master's or doctoral degree in science education. Several had additional experience in national assessment development, teacher training, or instructional material review, demonstrating both subject-matter expertise and familiarity with evaluation practices. Of note, the teacher-coders contributed solely to the coding process and were not among those who provided instructional recordings.

Based on the transcribed data, paired teams of university faculty and coders—each representing the four subjects of science—collaboratively constructed instructional data. The instructional dataset consisted of four components: *input*, *theme*, *output*, and *rating*. For each meaningful teacher utterance selected from the corrected transcript, the utterance was entered into the *input* column, with its corresponding thematic category and qualitative evaluation recorded in the *theme* and *output* columns, respectively. A numerical rating on a scale of 1 to 5 was also assigned in the *rating* column (1: educationally inappropriate utterance; 2: utterance requiring improvement; 3: average-level utterance; 4: educationally effective utterance; 5: educationally exemplary utterance).

This study retained only teacher utterances in the transcripts, excluding all student speech. The inclusion of student utterances in classroom recordings entails logistical complexities, such as the deployment of multiple microphones, and introduces additional ethical considerations, including more stringent IRB procedures due to the involvement of minors. Notably,

Xu et al. (2024) demonstrated that, even for student-related variables, using teacher utterances alone as input for large language models produced results that were comparable to—or for some MQI variables exceeded—those obtained when student speech was also included.

3.2 Establishing a Robust Coding Framework

3.2.1 Chunking Units

In studies involving relatively short texts—such as essay-type prompt-response pairs or teacher–student question–answer exchanges—the chunking unit is generally well-defined. However, when analyzing teacher discourse over the course of a 45-minute instructional session, determining appropriate analytical units for LLM-based analysis presents a more nuanced and challenging task.

Considering data-processing efficiency and computational cost depending on chunking units in LLM-based modeling, two distinct chunking approaches were initially applied: small-unit chunking, which involved detailed coding at the sentence level, and large-unit chunking, which involved chunking at the paragraph level. Comparison of the two chunking methods was also one of the research questions. However, coders reported substantial difficulty in consistently distinguishing between the two chunking units during the actual coding process. Consequently, considerable discrepancies were observed across subject areas, prompting the revision of the initial coding scheme.

A review of the recorded instructional sessions further revealed notable variation in teachers' speech patterns. While some teachers spoke in long, complex utterances, others used brief sentences, phrases, or even single words. The STT software segmented the audio based on pause detection, which led to substantial variability across sessions and speakers—some segments appeared as full paragraphs, whereas others consisted of only one or two words.

To ensure consistency and support fine-grained feedback generation appropriate for LLM-based modeling, chunking was conducted at the predicate level wherever possible. In instances where teachers produced extended utterances without natural pauses, manual chunking was applied based on semantic coherence. Moreover, utterances that had already been segmented into separate lines by the software were not merged during the coding process.

3.2.2 The Need for Systematic Coding and Rating

A commonly noted limitation of LLM-based models is their tendency to generate overly generic feedback—functionally akin to returning the median in statistical terms. Previous studies have reported that response generated by these models are often characterized as 'not novel or insightful' (e.g., Wang & Demszky, 2023). In recognition of this limitation, the present study allowed coders the flexibility to develop and assign themes during the coding process. Nonetheless, pronounced differences emerged in the descriptive statistics of the number of themes and the distribution of ratings across subject areas.

The discrepancies observed in chunking units and theme development during the initial coding process highlighted the necessity of systematic coding and rating and the development of a standardized glossary of terms. To ensure the validity of subsequent analyses, two raters were assigned per subject area, and a targeted, systematic coding and rating approach was implemented. The specific procedures are detailed in the following section.

3.3 Coding and Rating Procedures

3.3.1 Glossary of Terms Development: A Primarily Bottom-Up Approach

The development of the glossary followed a primarily bottom-up approach, albeit subtly informed by theoretical background and prior disciplinary training. The initial coding yielded 172 terms in physics, 71 in chemistry, 63 in life science, and 373 in earth science. When organized in alphabetical order, the total number of terms reached 678. However, many redundancies were identified due to variations in particle usage (e.g., checking students' level of understanding vs. checking student understanding) as well as semantically identical (e.g., self-directed learning vs. autonomous learning) or closely related expressions (e.g., task presentation vs. task guidance). Through collaborative and iterative refinement rounds involving four science education professors—each representing one of the four subject areas—and one professor specializing in educational evaluation, the initial glossary was consolidated into 223 terms. Each science education professor then developed exemplary annotation responses based on the initial glossary, which were used in multiple rounds of coding and rating. Feedback from these sessions led to further refinements of both the glossary and the coding scheme.

Finally, the revised glossary comprises 144 finalized themes, each accompanied by a simple definition, a list of thesaurus terms, and a higher-order category. These categories specify the instructional phase and functional role associated with each theme and provide corresponding rating guidelines. For example, themes under the category 'Basic Instructional Management' (e.g., lesson introduction, closure, review of prior learning) are capped at a maximum rating of 3 points, whereas more pedagogically important themes such as 'inducing cognitive conflict for conceptual change' may warrant ratings of 4 or 5 points.

3.3.2 Coding and Rating: A Dual-Expertise Approach

The academic expertise of science education faculty is a sine qua non in the evaluation of science teachers' instructional and assessment practices. This need is particularly salient in the present study, which developed and implemented a novel *LLM-specific instructional expertise coding framework*. Equally critical, however, is the experiential knowledge of experienced science teachers who continuously engage in, and reflect on, instructional and assessment practices within authentic classroom settings. The study thus recognizes the complementary and essential roles of both theoretical expertise in science education and the practical expertise of in-service science teachers.

To operationalize this dual-expertise approach, coding and rating were conducted by a team composed of one science education faculty member and two experienced science teachers per subject, with each member's disciplinary and practical knowledge fully respected and leveraged. The feedback obtained throughout this process informed the refinement of the thematic framework and coding manual, including the glossary of terms, thereby enhancing the consistency of the ratings and contributing the validity of the approach.

The process proceeded in several iterative stages. In Step 1, a faculty member with expertise in educational evaluation provided a comprehensive set of guidelines outlining the procedures for coding and rating, including protocols for constructing the instructional data set.

In Step 2, each of the four science education faculty members independently selected one instructional session within their respective major subject areas (45 minutes in length) and produced subject-specific annotations. The two teacher-coders then independently coded the same session. Their results were iteratively reviewed through collaborative discussion to reach consensus within the team. Discrepancies were of particular concern, and any related misconceptions or inappropriate coding decisions were carefully examined.

Step 3 involved full-faculty meetings. Issues arising during the Step 2—including challenges, cautions, and suggestions—were addressed, and annotations on common factors (e.g., Basic Instructional Management) were collaboratively generated to ensure consistency across subjects.

In Step 4, the common-factor annotations derived in Step 3 were fed back to the respective teams. Incorporating the feedback, each team generated subject-specific annotations within their domain, producing a finalized annotation set for one full instructional session, known as the consensus version. The procedures and efforts used to generate this finalized set were subsequently applied to the annotation of the remaining sessions. Fig. 1 illustrates the procedures.

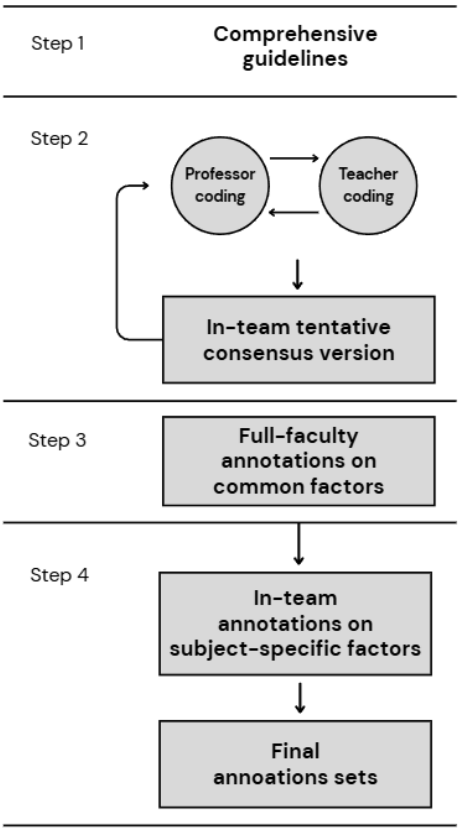


Figure 1. Coding and Rating Procedures

3.4 Analysis Methods

This study employed Quadratic Weighted Kappa (QWK) to assess rater consistency. QWK is an extension of Cohen's Kappa that adjusts for agreement occurring by chance. Unlike Cohen's Kappa, which assigns equal weight to all disagreements, QWK applies quadratic weights to discrepancies, such that larger differences between ratings incur greater penalties.

Equation 1 expresses the unweighted form of Cohen's Kappa, which is calculated as one minus the ratio between q_o and q_c (Cohen, 1968). q_o is the proportion of observed disagreement ($p_o = 1 - q_o$), and q_c is the proportion of disagreement by chance ($p_c = 1 - q_c$).

$$\kappa = 1 - \frac{q_o}{q_c}. \quad (1)$$

Suppose item i is rated by rater 1(j) and rater 2(k) on a K-ordinal scale. The weighted Kappa incorporates weights to the unweighted Kappa as in equation 2 (Vanbelle, 2016). The sole distinction between the equations is whether the disagreement rates are weighted. Equation 2 provides a general formulation of the weighted Kappa.

$$\kappa^w = 1 - \frac{q_o^w}{q_c^w}, \quad (2)$$

where $q_o^w = \sum_{j=1}^{K}\sum_{k=1}^{K} v_{jk} \cdot p_{jk}$ and $q_c^w = \sum_{j=1}^{K}\sum_{k=1}^{K} v_{jk} \cdot p_{j\cdot} \cdot p_{\cdot k}$.

When the quadratic weights defined in equation 3 are applied to equation 2, the resulting coefficient is referred to as the Quadratic Weighted Kappa (QWK).

$$v_{jk} = \left(\frac{j-k}{K-1}\right)^2, \quad (3)$$

Where $0 \le v_{jk} \le 1,\ v_{jj} = v_{kk} = 0\ (j,k = 1,\dots,K)$.

Although Spearman's rank-order correlation (rho) is also frequently used to examine rater consistency in LLM research, it measures the monotonic relationship between raters by converting ratings into ranks. In contrast, QWK directly evaluates the degree of actual agreement by preserving the original rating values. Therefore, in contexts such as the present study—where tied scores are prevalent—Spearman's rho is not among the most suitable statistics, as it neither adjusts for chance agreement nor performs robustly in the presence of numerous ties.

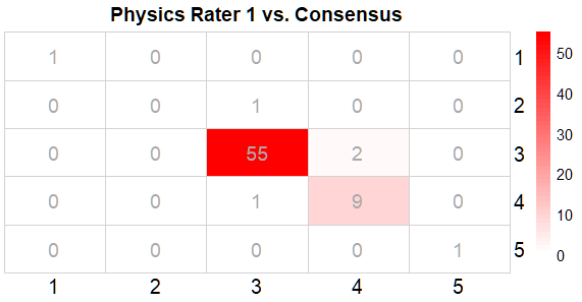
QWK ranges from –1 to 1, with values closer to 1 indicating stronger agreement between raters, and a value of 0 suggesting agreement no better than chance. The QWK results were calculated for every iteration, and shared with the coding team, which contributed to improving rater consistency. Specifically, the consensus rating from the final round was regarded as the ground truth. Accordingly, the QWK between the consensus and each individual rater in the final round was calculated and presented, accompanied by heatmaps visualizing the alignment between each rater and the consensus. The R packages *Metrics* and *pheatmap* were used to compute QWK and to generate heatmaps, respectively.

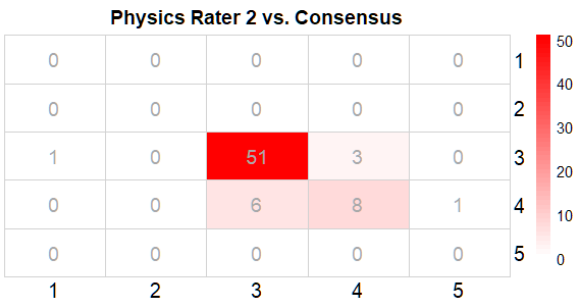
4 Results

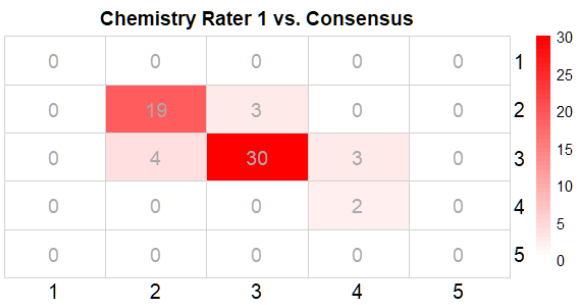
After a full iteration of Steps 1 through 4 in Fig. 1, the within-subject rater consistency results were summarized in Table 3. Specifically, the QWK between each rater and the consensus rating was calculated. Based on the interpretation criteria proposed by Landis and Koch (1977), all four subjects exhibited at least moderate agreement, with some reaching levels categorized as substantial or even almost perfect. These outcomes are further illustrated through heatmaps (Fig. 2).

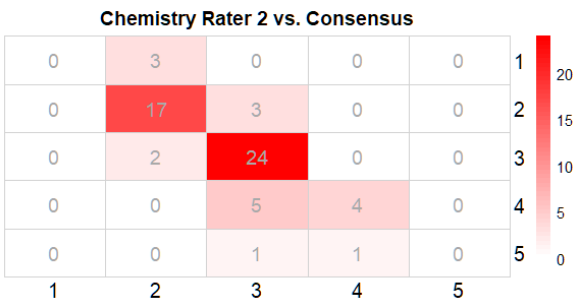
Table 3. QWK between Raters and Consensus by Subject

Subject	Rater 1 vs. Consensus	Rater 2 vs. Consensus
Physics	0.89	0.52
Chemistry	0.75	0.74
Life Science	0.41	0.69
Earth Science	0.96	0.98









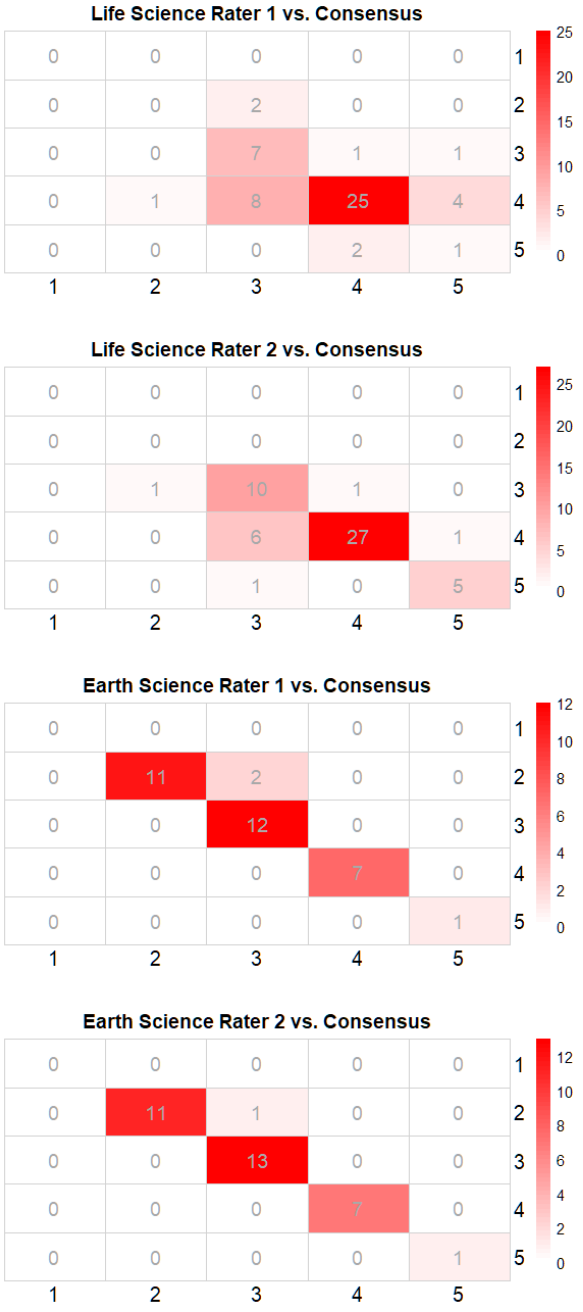


Figure 2. Rater–Consensus Agreement Heatmaps by Subject

5 Discussion

In response to the limitations identified in previous LLM research, this study designed and implemented predicate-level chunking, coding, and rating wherever possible. However, this fine-grained approach introduced a critical challenge: such segmenting could at times obscure the overarching instructional intent. For instance, when explanatory discourse is structured through rhetorical questioning followed by self-answering, the isolated coding of teacher utterances—coupled with the absence of student responses—can lead to mislabeling. To address this issue, coders engaged in repeated, holistic reviews of entire lessons to contextually infer the teacher's pedagogical intent prior to conducting coding and rating. Continuous refinement of the coding scheme and iterative updates to the coding guidelines were integral to the methodological rigor of this study.

Moreover, to elevate moderate levels of rater consistency observed in some subject areas to substantial or near-perfect agreement, additional coder training appears warranted. In this study, inter-rater consistency was calculated at the session level using rater pairs within each subject. Future research will involve re-running the cross-subject calibration process for shared coding factors (corresponding to Step 3 in Fig. 1) and subsequently evaluating consistency across all eight raters. This will support a more nuanced and systematic refinement of the coding and rating protocol.

To date, LLM research on modeling teacher instructional expertise has predominantly relied on traditional, global evaluation frameworks such as CLASS and MQI, resulting in alignment limitations for fine-grained analysis. To address this gap, we introduce a predicate-level, utterance-centric coding framework developed through a dual-expertise approach that combines science education faculty with experienced teachers. Specifically, teacher utterances are chunked into smaller units of analysis, and the chunked segments are coded and rated according to instructional expertise glossary, which were developed as a part of the process via a primarily bottom-up approach. The Quadratic Weighted Kappa results demonstrate reasonable rater consistency.

This study lays the groundwork for domain-adaptive LLM research. Future work will integrate the coding outputs into LLM pipelines for quality feedback modelling, leveraging and benchmarking latest techniques such as data augmentation and prompt engineering to identify optimal strategies for teacher expertise modeling. While the present framework was developed in the context of science education, exploring its adaptation to other subjects (e.g., social sciences, language, and mathematics courses) would be highly worthwhile. Depending on the specific subject domain, some areas may lend themselves more readily to LLM modeling, whereas others may require the integration of additional factors to improve prediction. Extending the framework to such areas could provide critical insights into its flexibility and broader applicability. Ultimately, this line of research aspires to advance both the methodological rigor and practical impact of LLM-based educational analytics, fostering more effective, data-driven support for teaching and learning.

Acknowledgments

This work was supported by the Ministry of Education of the Republic of Korea and the National Research Foundation of Korea (NRF-2024S1A5C3A0104264212).

References

Alic, S., Demszky, D., Mancenido, Z., Liu, J., Hill, H., & Jurafsky, D. (2022). Computationally identifying funneling and focusing questions in classroom discourse. *arXiv.* https://doi.org/10.48550/arXiv.2208.04715

Cohen, J. (1968). Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. *Psychological Bulletin, 70*(4), 213–220.

Hardy, M. (2025). "All that glitters": Techniques for evaluations with unreliable model and human annotations. *Findings of the Association for Computational Linguistics: NAACL 2025*, 2250–2278.

Hill, H., Blunk, M., Charalambous, C., Lewis, J., Phelps, G., Sleep, L., & Ball, D. (2008). Mathematical knowledge for teaching and the mathematical quality of instruction: An exploratory study. *Cognition and Instruction, 26*(4), 430-511.

Kane, T., Hill, H., & Staiger, D. (2015). National center for teacher effectiveness main study. *Inter-university Consortium for Political and Social Research*. icpsr36095-v2.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33*(1), 159-174.

Vanbelle S. (2016). A new interpretation of the weighted kappa coefficients. *Psychometrika, 81*(2), 399-410.

Wang, R., & Demszky, D. (2023). Is ChatGPT a good teacher coach? Measuring zero-shot performance for scoring and providing actionable insights on classroom instruction. *arXiv.* https://doi.org/10.48550/arXiv.2306.03090

Xu, P., Liu, J., Jones, N., Cohen, J., & Ai, W. (2024). The promises and pitfalls of using language models to measure instruction quality in education. *arXiv.* https://doi.org/10.48550/arXiv.2404.02444