Architecture
Image Studies

Vol.7, Issue 1, pp.119-127, 2026
DOI: https://doi.org/10.62754/ais.v7i1.646

https://journals.ap2.pt/index.php/ais/index

Automated Bloom's Taxonomy Classification of Teacher Questions Using Whisper and GPT-4

Taeryeong Kim[1], Kilhong Joo[2]

Abstract

Effective teacher questioning is central to fostering higher-order thinking, yet manually analyzing classroom discourse for cognitive demand is labor-intensive. This study proposes an automated system that classifies teacher questions into Bloom's revised taxonomy levels by integrating Whisper speech-to-text transcription with GPT-4's zero-shot classification capabilities. The end-to-end pipeline transcribes classroom audio, extracts teacher questions with surrounding context, and assigns one of six cognitive categories: Remember, Understand, Apply, Analyze, Evaluate, and Create. Using 350 questions annotated by three Bloom-trained experts (Fleiss' Kappa = 0.85), the GPT-4–based classifier achieved 64.2% accuracy and a Cohen's Kappa of 0.547 against human labels, outperforming a rule-based keyword baseline. Performance was strongest for higher-order categories such as Analyze and Create (precision ≥ 100%), while most misclassifications occurred between adjacent levels (e.g., Remember–Understand). Bloom-level distributions indicated a predominance of lower-order questions, highlighting the system's potential for providing data-driven feedback to promote balanced cognitive engagement. This work demonstrates the feasibility of scalable, context-sensitive analysis of live classroom questioning, offering practical applications in teacher self-reflection, professional development, and AI-enhanced learning analytics.

Keywords: *Bloom's Taxonomy, GPT, Whisper, STT, Cognitive Level Classification, Teacher Questioning.*

Introduction

Effective teacher questioning plays a crucial role in fostering student engagement, cognitive development, and higher-order thinking in classroom settings. By employing well-designed questions, teachers can guide students beyond simple factual recall toward deeper reasoning, problem-solving, and creative synthesis. To evaluate and improve questioning practices, many educators and researchers adopt Bloom's revised taxonomy, which categorizes cognitive processes into six hierarchical levels: Remember, Understand, Apply, Analyze, Evaluate, and Create [1].

Despite its importance, research consistently demonstrates that most classroom questions are concentrated at lower-order cognitive levels, such as Remember and Understand [2][3]. This imbalance restricts students' opportunities for critical thinking and active learning. To address this challenge, teacher professional development frequently emphasizes self-reflection on questioning strategies. One recommended approach is for teachers to analyze their classroom discourse to evaluate the cognitive levels of their questions [4][1]. However, manually auditing classroom questions is time-consuming, as it requires transcription, question extraction, and expert knowledge of Bloom's taxonomy, thereby creating significant barriers to routine implementation.

Recent advancements in artificial intelligence, particularly in automatic speech recognition (ASR) and large language models (LLMs), provide new opportunities to automate this analysis at scale. Speech recognition systems such as OpenAI's Whisper can transcribe classroom audio into text, while state-of-the-art language models like GPT-4 possess advanced zero-shot reasoning capabilities, allowing them to classify questions based on contextual understanding without requiring task-specific training. Unlike earlier rule-based or supervised machine learning approaches that often rely on

[1] Gyeongin National University of Education, Email: khjoo@ginue.ac.kr, (Corresponding Author)
[2] Gyeongin National University of Education

handcrafted keyword lists or domain-specific training data, GPT-4 can flexibly interpret diverse classroom discourse in real-world teaching scenarios [5][6].

In this study, we develop and evaluate an automated system that integrates Whisper transcription and GPT-4 classification to analyze teacher questions and classify them according to Bloom’s revised taxonomy levels. The system aims to support teacher self-reflection by providing quantitative feedback on questioning practices. We further compare its performance against human expert annotations and a rule-based keyword baseline to assess its classification accuracy and practical applicability in live classroom settings.

This paper makes the following contributions:

□An end-to-end pipeline that automates Bloom-level classification of teacher questions using Whisper ASR and GPT-4.

□Empirical evaluation of GPT-4’s classification performance using expert-labeled classroom data.

□Visualization of Bloom-level distributions to facilitate teacher self-reflection and pedagogical improvement.

The remainder of this paper is organized as follows: Section 2 reviews relevant literature on Bloom’s taxonomy and automated question classification; Section 3 describes the system design and experimental methodology; Section 4 presents and discusses the empirical results; and Section 5 concludes with directions for future work

Related Works

This study is grounded in three interrelated theoretical frameworks that collectively inform the development of an automated teacher-question classification system: (1) Bloom’s Revised Taxonomy of Educational Objectives, (2) Constructive Alignment Theory, and (3) Natural Language Processing and Learning Analytics. These frameworks offer the conceptual foundation for understanding the cognitive levels of teacher questions, their alignment with instructional goals, and the feasibility of automating their analysis through artificial intelligence.

Bloom’s taxonomy, originally introduced by Bloom et al. (1956) and later revised by Anderson and Krathwohl (2001), classifies cognitive processes into six hierarchical categories: Remember, Understand, Apply, Analyze, Evaluate, and Create [4][1]. This framework outlines a progression from lower-order thinking skills, such as memorization, to higher-order thinking, including critical analysis and creative synthesis. In educational practice, Bloom’s taxonomy is widely applied in the design of learning objectives, assessments, and particularly in classroom questioning.

In the context of teacher questioning, each category reflects the level of cognitive demand placed on students. For example, “What is the definition of evaporation?” corresponds to the Remember level, while “Can you design an experiment to test evaporation rates?” corresponds to the Create level. Empirical research has shown that teachers often, and sometimes unintentionally, rely heavily on lower-order questions, which may limit students’ deeper engagement [2][3]. This taxonomy serves as the foundation of our classification system. By categorizing teacher utterances according to Bloom’s levels, we can systematically quantify the cognitive demands of classroom discourse and identify imbalances or gaps in instructional questioning.

Constructive Alignment emphasizes the alignment among three core components of instruction: intended learning outcomes, teaching and learning activities, and assessment tasks. According to this framework, the effectiveness of a learning experience depends on the coherence among these elements. Within this alignment, teacher questions serve not merely as instructional techniques but as real-time reflections of the intended cognitive outcomes. If a lesson aims to foster higher-order thinking but the questions remain focused on factual recall, such misalignment may hinder deep learning. Therefore, analyzing the cognitive levels of teacher questions relative to instructional goals offers a powerful diagnostic tool for evaluating instructional quality.

This study incorporates the principles of Constructive Alignment by examining whether the cognitive levels of teacher questions, as determined by automated classification, align with the intended cognitive objectives of the lesson. Such analysis enables data-informed feedback for teachers and supports professional development efforts aimed at refining instructional design.

Recent advances in Natural Language Processing (NLP) and Learning Analytics (LA) have enabled fine-grained analysis of classroom discourse at scale (Ferguson, 2012). NLP techniques have been applied to classify instructional text — such as assessment items or learning objectives — according to Bloom's taxonomy, utilizing methods ranging from rule-based systems to deep learning [5][6][7][8]. Early rule-based systems (e.g., Haris & Omar, 2015; Yuhana et al., 2019) relied on matching keywords (typically verbs) to corresponding cognitive levels [5][6]. While interpretable and simple, these approaches struggle with language variability and semantic nuance. Machine learning methods — including SVM, Naive Bayes, and k-NN (Yahya et al., 2012) — improved performance, while later deep learning models such as LSTM and BERT-based classifiers (Shaikh et al., 2021; Banujan et al., 2023) achieved over 85% accuracy on structured, written question datasets [7][8]. However, these methods typically focused on written language, which is well-structured and relatively easy to parse. In contrast, spoken classroom dialogue presents unique challenges:

- Speech-to-text (STT) noise and transcription errors
- Non-standard and informal grammar
- Contextual dependencies spanning multiple conversational turns.

To address these challenges, this study employs Whisper, a high-performance STT model, to transcribe real classroom recordings [9]. The extracted questions are then classified using both a rule-based method and GPT-4, a large-scale language model (LLM). GPT-4 represents a paradigm shift in NLP-based classification. Unlike traditional models that require task-specific training data, GPT-4 can perform classification based on its extensive pretrained knowledge through prompt-based instructions. Prior studies (e.g., Zhai & Wang, 2021; Huang et al., 2021) have demonstrated the viability of LLMs for educational discourse analysis, but few have explored their application to live classroom dialogue.

Recent studies have explored the educational impact of integrating STT technologies directly into classroom activities. For instance, Kim and colleagues (2024) examined a speech-to-text-based class format and found that real-time transcription not only improved learner comprehension but also facilitated active engagement by reducing cognitive load related to note-taking. This aligns with our approach, where Whisper-based transcription forms the foundation for downstream Bloom-level question classification.

In this study, we leverage GPT-4's capability to interpret the intent of questions within context, hypothesizing that it can classify teacher questions into Bloom's six levels with high validity. Its performance is compared against both human expert ratings and a rule-based baseline, with agreement evaluated using metrics such as accuracy and Cohen's Kappa. This comparative approach not only validates GPT-4's effectiveness but also demonstrates the feasibility of scalable, cost-efficient feedback systems to support reflective teaching practice.

Automated Bloom's Taxonomy Classification of Teacher Questions

This study develops a fully automated end-to-end pipeline that classifies teacher questions extracted from classroom audio recordings based on Bloom's revised taxonomy. The pipeline integrates multiple components: speech transcription, question extraction, both rule-based and GPT-based classification, and performance evaluation using human-annotated ground truth data. This section describes the overall system architecture, details the data collection and preprocessing procedures, explains the classification methodologies, and outlines the evaluation strategy..

System Architecture Overview

The proposed system consist of the following key modules:

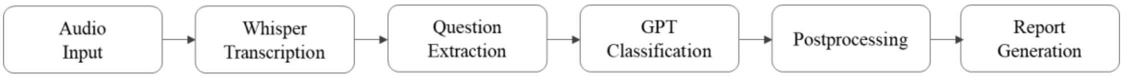


Fig. 1. System Pipeline Diagram

The system executes an end-to-end pipeline that automatically classifies teacher questions from classroom audio recordings based on Bloom's revised taxonomy. The pipeline includes five primary modules:

The first module is 'Audio Transcription'. Classroom audio recordings are transcribed using OpenAI's Whisper speech recognition model, which is capable of producing highly accurate transcripts even under the variable acoustic conditions typical of real-world classrooms. Second module is 'Question Extraction'. Teacher questions are automatically extracted from the transcripts using rule-based syntactic and lexical pattern matching, including the detection of question markers, interrogative forms, and discourse cues that indicate questioning intent. The third module is ‘Question Classification’. The extracted questions are classified into Bloom’s six cognitive levels using two parallel methods:

- A rule-based baseline classifier assigns Bloom levels by matching predefined keyword lists to specific cognitive categories. While offering interpretability and simplicity, this method lacks the ability to process contextual nuance.
- A GPT-4-based classifier, which represents the primary innovation of this study. The GPT-4 model processes each question together with its surrounding discourse context and applies zero-shot prompt-based classification, leveraging its extensive pretrained knowledge and contextual reasoning capabilities to assign appropriate Bloom levels.

The fourth module is ‘Post-Processing and Label Normalization’. The classification outputs are normalized to standardized Bloom level labels to ensure consistency across classification methods. This normalization process resolves synonymous expressions and minor formatting variations. Last module is ‘Evaluation and Visualization’. The classification results are compared against expert human annotations. Quantitative evaluation metrics, including accuracy and Cohen’s Kappa, are calculated, and cognitive level distributions are visualized to provide actionable feedback that supports teacher reflection and professional development.

Human Annotation for Ground Truth

The dataset comprises 12 real-world secondary school classroom recordings collected in Korea, covering a diverse range of subjects and instructional styles. Each class session lasted approximately 20 to 30 minutes, resulting in a total of approximately 5 to 6 hours of classroom discourse. From these recordings, a total of approximately 350 teacher questions were extracted and subjected to cognitive level classification. The extracted questions were independently annotated by three educational experts trained in Bloom’s revised taxonomy, with consensus labels determined through majority agreement. The inter-annotator reliability, measured by Fleiss’ Kappa, reached 0.85, indicating substantial agreement and supporting the validity of the human-annotated ground truth.

Audio Recording and Transcription

Audio data were recorded from various Elementary and Secondary-level classrooms in Korea, covering a range of subject areas. Each recording session lasted approximately 25 minutes and employed a unidirectional lapel microphone to isolate teacher speech from background noise. Transcription was performed using OpenAI’s Whisper (base model), which offers an optimal balance between speed and transcription accuracy for Korean language inputs. Whisper produced time-aligned, speaker-specific transcripts with minimal transcription errors, facilitating reliable downstream processing for subsequent analysis. The question analysis pipeline in this study processes input speech through a speech recognition module before connecting it to the information processing stage. This flow is similar to the information selection agent–based speech web architecture proposed by Kwon and Kinoshita (2006), with the key difference being that the proposed system employs a state-of-the-art deep learning language model to automatically classify the cognitive level of teacher questions [10].

Question Extraction

Teacher questions were extracted from transcripts using a hybrid approach:

- Syntactic cues: Sentences ending in question marks
- Lexical patterns: Interrogatives such as “why, what, how, when, where” in Korean
- Positional filtering: Questions framed by discourse context

To preserve semantic context for classification, each extracted question was paired with its immediately preceding and following sentences. This process generated a corpus of question-context pairs, which were used for both the rule-based and GPT-based classification modules.

Rule-Based Classification of Cognitive Levels

Each question in the corpus was classified using a custom rule-based system. This system consisted of:

- Action verb mapping: Direct association of question verbs with Bloom’s cognitive categories (e.g., “define” mapped to Remember; “compare” mapped to Analyze)

- Question structure templates: Syntactic pattern matching based on common sentence structures (e.g., questions such as “What are the differences between…” assigned to Analyze)

- Semantic cue identification: Phrases indicating justification, reflection, or problem-solving

This rule-based approach, grounded in the language of instructional objectives, allowed for transparent and interpretable classification. The results were exported as structured data for subsequent comparison. Although a rule-based classifier was initially developed as a baseline for comparison, its accuracy was considerably lower than the GPT-4 model, particularly due to its inability to handle the linguistic diversity and contextual nuances inherent in natural classroom discourse. As such, we chose not to include detailed quantitative results for the rule-based system and focused the analysis on the GPT-based approach.

GPT-Based Classification of Cognitive Levels

The GPT-4 model was employed for classification using the OpenAI API. Each classification prompt included the following components:

- A brief description of Bloom’s taxonomy, provided in Korean to align with the language of the classroom discourse

- The extracted question along with its surrounding discourse context

- Explicit output formatting instructions (e.g., “Evaluate - requires judgment”)

GPT-4 typically responded with the assigned Bloom level followed by a brief justification for its decision. The model outputs were subsequently parsed and normalized to standardized English labels (e.g., “이해” converted to “Understand”) to ensure consistency across classifications.

Label Normalization and Cleaning

To ensure consistency across both classification methods, a label normalization process was applied to the outputs of both the GPT-based and rule-based classifiers. All labels were mapped to the six standard Bloom categories, resolving synonym variations (e.g., “Knowledge” remapped to “Remember”) and removing any ambiguous or mixed-level responses. This normalization procedure ensured valid cross-method comparisons and minimized classifier-specific labeling inconsistencies.

Comparative Evaluation with Human Benchmark

The classification outputs from both the rule-based and GPT-based systems were evaluated against the human-annotated ground truth dataset. Three primary evaluation metrics were used:

- Accuracy: The proportion of classifications that matched the human-assigned Bloom levels

- Cohen’s Kappa: A statistical measure of inter-rater agreement that accounts for agreement occurring by chance

- Confusion Matrix: A detailed visualization of misclassification patterns across Bloom’s cognitive levels

Together, these evaluation metrics provided a comprehensive quantitative assessment of each classifier’s reliability and diagnostic utility within authentic teaching contexts.

Visualization and Reporting

Finally, the classification results were visualized through various graphical representations, including cognitive level distribution charts, method-wise performance comparisons, and heatmaps of confusion matrices. These visualizations were generated using custom-developed Python scripts and libraries. In addition, representative sample questions for each Bloom level were compiled into summary reports to support teacher reflection and professional development. A summary dashboard was also created to visualize cognitive demand trends across different lessons, enabling instructors to monitor and adjust their questioning strategies over time.

Result and Discussion

The GPT-based classifier achieved an overall accuracy of 64.2% and a Cohen’s Kappa score of 0.547 when compared to expert human annotations across all extracted classroom questions. These results reflect moderate agreement, demonstrating that GPT-4 can approximate expert-level Bloom taxonomy classification despite the absence of domain-specific training. Compared to the rule-based keyword matching baseline evaluated in preliminary stages, GPT-4 substantially improved both classification accuracy and consistency of agreement with human raters.

Table 1. Overall Classification Performance of the GPT-Based Bloom Taxonomy Classifier (Aggregate Across All Lessons)

Metric	Value
Accuracy	64.2%
Cohen’s Kappa	0.547

Table 2 presents the model’s precision, recall, and F1-score for each Bloom level category. The performance varied across cognitive levels, reflecting both the inherent challenges of distinguishing between adjacent levels and GPT-4’s strengths in contextual reasoning and interpretation.

Table 2. Per-Level Classification Performance (Precision / Recall / F1-Score)

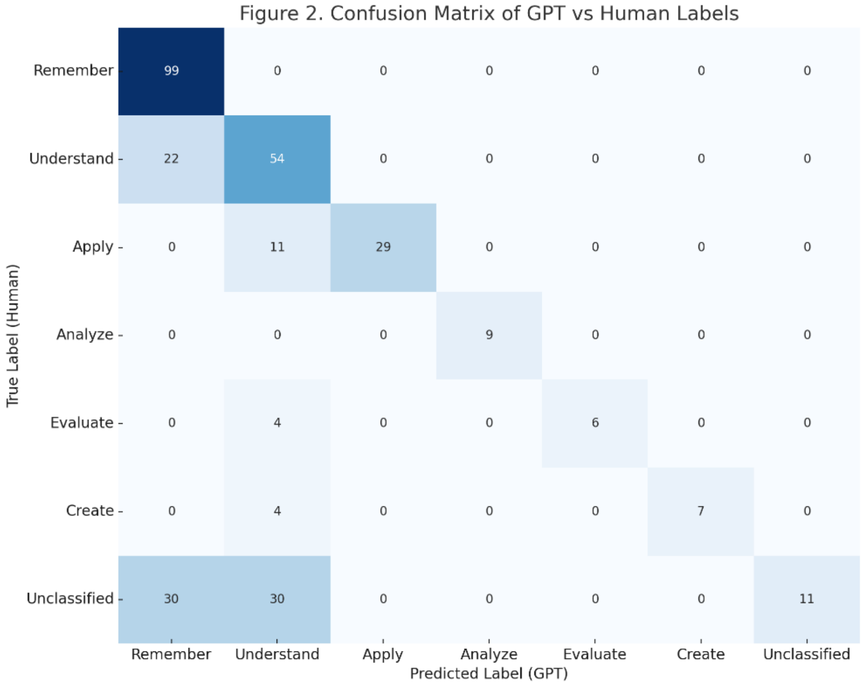
Bloom Level	Precision	Recall	F1-Score
Remember	64.7%	85.3%	74.0%
Understand	47.6%	71.1%	56.9%
Apply	76.9%	72.5%	74.6%
Analyze	100.0%	75.0%	85.7%
Evaluate	72.7%	66.7%	69.6%
Create	100.0%	14.3%	24.2%

The model exhibited particularly high precision for higher-order categories such as Analyze and Create, successfully identifying these levels when confident. However, recall for these complex levels remained moderate, suggesting that the model employed a conservative classification strategy, favoring precision over sensitivity when addressing more cognitively demanding questions. For the Apply and Evaluate categories, the model demonstrated relatively balanced precision and recall, indicating stable classification performance for mid-level cognitive processes.

The Understand category posed the greatest classification challenge. Although recall was relatively high, precision was considerably lower, suggesting frequent over-classification of questions into the Understand level, even when they may have belonged to adjacent categories such as Remember or Apply. This difficulty highlights the inherent ambiguity of distinguishing comprehension-level questions from factual recall and application-oriented prompts, particularly within the fluid nature of spontaneous classroom dialogue.

To further investigate classification errors, Figure 2 presents the confusion matrix comparing GPT-4 predictions to expert annotations. Most misclassifications occurred between adjacent Bloom levels, particularly between Remember and Understand, and between Understand and Apply. These patterns are consistent with previous findings, underscoring that the boundaries between these levels are inherently subtle and highly context-dependent.

The model demonstrated strong agreement for the Remember category, accurately capturing a large proportion of factual recall questions. In contrast, for Unclassified items—cases where human experts refrained from assigning a Bloom level—the model tended to force assignments into one of the Bloom categories, resulting in poor recall performance for the Unclassified class.



	Remember	Understand	Apply	Analyze	Evaluate	Create	Unclassified
Remember	99	0	0	0	0	0	0
Understand	22	54	0	0	0	0	0
Apply	0	11	29	0	0	0	0
Analyze	0	0	0	9	0	0	0
Evaluate	0	4	0	0	6	0	0
Create	0	4	0	0	0	7	0
Unclassified	30	30	0	0	0	0	11

Figure 2. Confusion Matrix of GPT vs Human Labels

The overall Bloom level distribution classified by GPT-4 across all lessons is displayed in Figure 3. Consistent with prior research, the majority of classroom questions were concentrated in lower-order cognitive categories, with Remember and Understand comprising most of the teacher discourse.

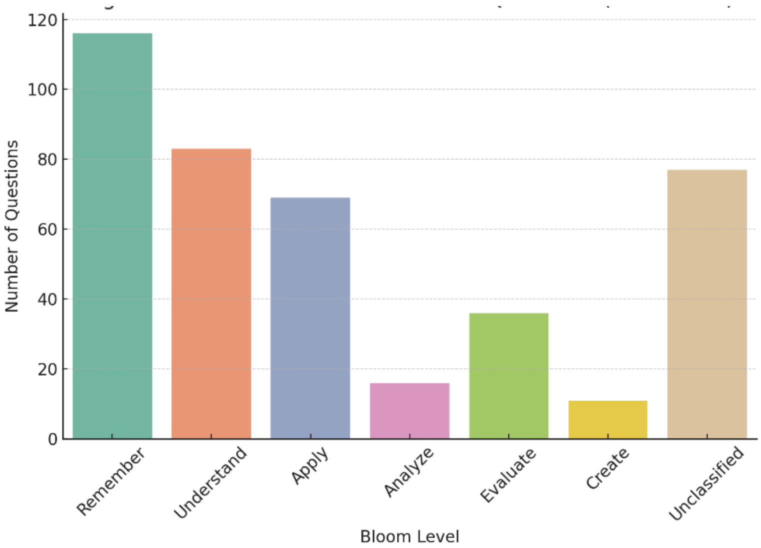


Figure 3. Bloom Level Distribution of All Questions (GPT-based)

This automated distribution offers valuable feedback for teachers by enabling systematic reflection on their questioning strategies. Recognizing a disproportionate reliance on lower-order questions may encourage teachers to intentionally design lessons that incorporate higher-order questions, thereby fostering student reasoning, evaluation, and creativity. Consequently, the system provides a scalable, data-driven mechanism to support continuous teacher professional development and instructional enhancement. These findings align with broader research demonstrating the positive educational impact of generative AI tools. For instance, Kim and Ryu (2024) reported that integrating generative AI into nursing education enhanced learners' academic performance, motivation, self-efficacy, and learning attitudes, while also being favorably perceived as a class management and assessment tool. This suggests that AI-driven systems, such as our GPT-4–based question classification framework, can potentially contribute not only to instructional diagnostics but also to fostering positive learner outcomes [11].

Conclusion

This study developed and validated an automated system for classifying teachers' classroom questions according to cognitive levels, utilizing speech recognition, large language models, and educational theory. By combining Whisper-based transcription with GPT-4's advanced language understanding capabilities, the system analyzes recorded classroom discourse and assigns Bloom’s revised taxonomy levels to teacher-generated questions. In empirical evaluations against expert human annotations, the GPT-based classifier demonstrated moderate agreement with human coders (accuracy: 64.2%; Cohen’s Kappa: 0.547), substantially outperforming conventional rule-based keyword matching approaches. The system successfully identified higher-order categories such as Analyze, Evaluate, and Create with strong precision, while also revealing persistent challenges in differentiating between Understand and Remember levels—an area often ambiguous even for human raters.

The findings reaffirm previous research indicating that most classroom questioning remains concentrated in lower-order categories such as factual recall and basic comprehension. Unlike purely descriptive studies, however, the proposed system offers teachers actionable, individualized feedback on their questioning patterns through automated analysis. By providing clear summaries and visual representations of cognitive level distributions, the system lowers the barrier for teachers to engage in continuous reflective practice, enabling them to identify imbalances in their instructional discourse and intentionally incorporate more cognitively demanding questions. Such data-informed reflection holds promise for fostering higher-order thinking and deeper student engagement.

From a methodological perspective, this research contributes to the growing body of evidence that large language models such as GPT-4 can serve as powerful tools for educational observation, capable of interpreting nuanced instructional language in real-world classroom settings. Unlike rule-based or supervised learning models, GPT-4 operates without task-specific training data, applying general world knowledge and contextual reasoning to classification tasks—even in noisy, spontaneous classroom dialogue. These capabilities open multiple avenues for future expansion in AI-driven teaching analytics, including evaluations of teacher feedback, student questioning patterns, dialogue complexity, and broader instructional interactions.

Several limitations remain. This study was conducted on a relatively small dataset consisting of three classroom recordings. Future research should apply the system to larger and more diverse educational contexts to further validate its generalizability. In addition, the classifier’s occasional overconfidence in assigning Bloom levels to ambiguous or uncategorizable utterances highlights the need for improved uncertainty calibration and abstention mechanisms. In particular, we observed that GPT-4 tended to assign Bloom levels even for utterances that human experts refrained from categorizing (i.e., Unclassified cases). This behavior reflects a common property of large language models—namely, their tendency toward overconfidence in forced classification tasks, even when uncertainty is high. Addressing this tendency through confidence calibration or abstention mechanisms may further enhance the system's robustness in real-world applications. Addressing these challenges will be essential to ensure robust performance across varying classroom environments, languages, and subject areas.

Ultimately, this study demonstrates the feasibility and potential of integrating advanced AI systems into teacher professional development. By enabling teachers to audit their own questioning practices with minimal additional effort, such systems function as virtual instructional coaches—continuously observing, analyzing, and providing feedback after each lesson. As speech recognition and language models continue to evolve, scalable and accessible platforms may emerge that empower teachers, regardless of location or resources, to continually refine their instructional practices and better support 21st-century learning objectives. At its core, this research embodies a meaningful synergy between artificial intelligence and pedagogy—leveraging technology not to replace teachers, but to enhance their capacity for reflective, student-centered teaching.

References

[1] Anderson, L. W., & Krathwohl, D. R. (Eds.). (2001). A taxonomy for learning, teaching, and assessing: A revision of Bloom’s taxonomy of educational objectives. New York, NY: Longman.

[2] Chin, C. (2006). Classroom interaction in science: Teacher questioning and feedback to students' responses. International Journal of Science Education, 28(11), 1315–1346. https://doi.org/10.1080/09500690600621100

[3] Yahya, A. A., Husain, W. B., & Nazri, M. M. (2012). Automated categorization of questions according to Bloom's taxonomy. Procedia - Social and Behavioral Sciences, 59, 297–303. https://doi.org/10.1016/j.sbspro.2012.09.278

[4] Bloom, B. S., Engelhart, M. D., Furst, E. J., Hill, W. H., & Krathwohl, D. R. (1956). Taxonomy of educational objectives: The classification of educational goals. Handbook I: Cognitive domain. New York, NY: David McKay Company.

[5] Haris, S. S., & Omar, N. (2015). Bloom's taxonomy question categorization using rules and N-gram approach. Journal of Theoretical and Applied Information Technology, 76(3), 401–407.

[6] Yuhana, N., Husain, W., Yahya, A., & Nazri, M. (2019). A hybrid approach for Bloom taxonomy question classification using semantic similarity and neural network. Journal of Physics: Conference Series, 1179(1), 012056. https://doi.org/10.1088/1742-6596/1179/1/012056

[7] Shaikh, A., Daudpotta, S. M., & Imran, A. S. (2021). Bloom's learning outcomes' automatic classification using LSTM and pretrained word embeddings. IEEE Access, 9, 117887–117909. https://doi.org/10.1109/ACCESS.2021.3106443

[8] Banujan, K., Kumara, S., Prasanth, S., & Ravikumar, N. (2023). Automatic Bloom's taxonomy–based question classification using BERT embeddings. IEEE Access, 11, 1–14. https://doi.org/10.1109/ACCESS.2023.3244444

[9] OpenAI. (2023). GPT-4 technical report. arXiv:2303.08774. https://doi.org/10.48550/arXiv.2303.08774

[10] Kwon, H. J., & Kinoshita, J. (2013). Novel speech web architecture based on information selection agent. International Journal of Advanced Culture Technology, 1(1), 11–14.

[11] Kim, H., & Ryu, Y. (2024). A study on nursing education using generative AI: A scoping review. Journal of Convergence Knowledge, 12(4), 145–158.