Exploring ASR and Large Audio Language Models for Transcribing Peer Discourse in Noisy Classrooms

Wenting Sun[1] and Jiangyue Liu[2]
[1]Humboldt-Universität zu Berlin, Germany
[2]Soochow University, Jiangsu, China

suwentin@hu-berlin.de
ljy@suda.edu.cn

Abstract: Capturing peer discourse in real-world classrooms offers valuable insights into collaborative learning but presents significant technical and pedagogical challenges. While most existing Automatic Speech Recognition (ASR) systems research has focused on teacher-led or online English-speaking environments, peer-to-peer dialogue in noisy, non-English dominant, face-to-face classrooms remains underexplored. This study investigates the feasibility of using both traditional ASR systems—Whisper and Wav2Vec2—and emerging Large Audio Language Models (LALMs), including Qwen2-Audio and Ultravox, to transcribe Mandarin peer conversations recorded via students' mobile phones in authentic classroom settings. We collected over 105,715 seconds of audio from 38 student groups across two collaborative learning tasks from university classrooms. The manually transcriptions were served as ground truth. Audio quality test of all audio recordings was conducted. Five representative samples with varied signal-to-noise ratios (SNR) and speech ratios were selected to do in-depth analysis. Transcription quality was evaluated using Word Error Rate (WER), Character Error Rate (CER), and Fuzzy String Matching. Additionally, we conducted a thematic analysis of transcription errors to identify linguistic, acoustic, and task-related challenges. Results show that Whisper consistently outperforms other models, achieving high transcription fidelity even in moderately noisy conditions. In contrast, LALMs—despite their strengths in semantic understanding—performed poorly in verbatim transcription, often generating hallucinated or irrelevant content. Importantly, task type and speech characteristics significantly influenced model performance: structured, reflective discussions yielded better results than spontaneous, technical dialogues involving numeric and English domain terms. This study contributes a low-cost, replicable workflow for classroom audio collection and evaluation, along with a detailed taxonomy of transcription errors. We emphasise that our results are exploratory due to the limited sample size. Nevertheless, the findings highlight the current limitations of LALMs for ASR tasks and offers practical recommendations for model selection in educational contexts. Our findings support the responsible integration of ASR technologies into classroom practice, with implications for real-time feedback, collaborative learning analytics, and teacher professional development. For researchers, this work demonstrates the need to consider peer dialogue and multilingual classroom ecologies when evaluating ASR. For teachers, practical recommendations are offered for selecting transcription tools that can support real-time feedback and professional reflection. For lifelong learning, our study illustrates the potential of ASR technologies to make collaborative dialogue more visible, analysable, and actionable across diverse contexts.

Keywords: Automatic speech recognition, Peer dialogue, Collaborative learning, Large audio language models, Mandarin classrooms, Educational AI

1. Introduction

In recent years, the potential of AI-powered tools to enhance teaching and learning has gained increasing attention. Among these, Automatic Speech Recognition (ASR) has emerged as a promising technology for capturing and analysing classroom interactions—particularly in collaborative learning environments (Southwell et al., 2022; Wang & Chen, 2024). ASR enables scalable analysis of large volumes of classroom dialogue with significantly lower cost and time investment compared to manual transcription, while maintaining consistent performance.

While prior research has explored ASR in remote or online settings and English-speaking classrooms (Ezema et al., 2025; Yin et al., 2025), there remains a notable gap in understanding how ASR performs in face-to-face, non-English dominant educational contexts. Real-world classrooms often involve spontaneous speech, overlapping dialogue, code-switching, and low-fidelity audio captured via mobile devices—conditions that pose substantial challenges to traditional ASR pipelines. These challenges raise critical concerns about the reliability, fairness, and educational utility of AI-driven transcription in authentic learning environments.

Recently, Large Audio Language Models (LALMs) such as Qwen2-Audio and Ultravox have been introduced, claiming to offer enhanced contextual understanding and multi-turn dialogue modelling (Chu et al., 2024; Fixie AI, 2025). However, their performance in noisy, multilingual, and pedagogically complex settings—particularly in Mandarin-speaking classrooms—has not been systematically evaluated. As a result, educators and researchers lack empirical guidance on which models to trust, what types of errors to expect, and how to interpret transcription outputs in meaningful ways.

This study addresses these gaps through four key contributions:

- **Model Benchmarking**: We conduct a systematic comparison of traditional ASR systems (Whisper, Wav2Vec2) and LALMs (Qwen2-Audio, Ultravox) using peer discourse audio collected from real-world Chinese classrooms.
- **Low-Cost Data Collection Pipeline**: We design and implement a scalable, device-independent audio collection and evaluation workflow using students' own mobile phones and open-source tools.
- **Error Typology Development**: We provide a fine-grained categorization of transcription error types, linking them to linguistic, acoustic, and task-related factors.
- **Practical Implications**: We offer actionable recommendations for selecting and applying speech transcription tools in educational contexts, with implications for AI-supported feedback, collaborative learning analytics, and teacher professional development.

2. Related Work and Research Questions

2.1 ASR and LALMs in Educational Contexts

The application of ASR in educational settings has been extensively studied, particularly for lecture transcription, feedback generation, and collaborative dialogue analysis (Southwell et al., 2022; Wang & Chen, 2024; Wu et al., 2024). However, much of this research remains narrowly focused on English-language or teacher-centred settings, Moreover, traditional ASR models such as Whisper and Wav2Vec2 have shown varying levels of success in transcribing English-language educational audio. For instance, Ezema et al. (2025) reported a minimum WER of 39%, while Yin et al. (2025) achieved 9.5% WER by combining Whisper with MarbleNet for Voice Activity Detection. Attia et al. (2025a) reported WERs of 30.25% and 38.59% across different datasets using pre-trained Wav2Vec2.0. Yet, a critical limitation is that prior studies have often assumed relatively clean audio, scripted utterances, and single-speaker dominance. These assumptions rarely hold in authentic peer-to-peer discourse, which presents unique challenges.

Unlike structured teacher-led lectures, student discussions are spontaneous, multi-voiced, and contextually embedded, making them difficult to capture and analyse. As Wang and Chen (2024) emphasize, understanding the performance of ASR systems in such settings is crucial for leveraging classroom dialogue to enhance learning. Russell et al. (2024) further highlight that poor recording quality and overlapping speech significantly degrade ASR accuracy in real-world classrooms.

Emerging Large Audio Language Models (LALMs), such as Qwen2-Audio and Ultravox, aim to extend ASR capabilities by incorporating semantic understanding and multi-turn dialogue modelling (Chu et al., 2024; Fixie AI, 2025). Although LALMs have demonstrated promising results on benchmark datasets, little is known about their reliability in educational environments. Most reported benchmarks use datasets such as Common Voice, which fail to capture code-switching, overlapping turns, or specialised domain vocabulary. Thus, the literature overlooks both the technical and pedagogical challenges of real-world classrooms, particularly in authentic, multi-turn peer discourse and non-English contexts.

2.2 Transcription Challenges in Classroom Audio

Classroom audio data is often characterized by low signal-to-noise ratios, overlapping speech, and frequent code-switching. Ezema et al. (2025) note that even high-performance ASR engines struggle under such conditions. Most existing studies (Wong et al., 2025; Zhao et al., 2024) focus on teacher speech or structured classroom formats, overlooking the dynamic and unpredictable nature of student-led discussions.

Moreover, there is a lack of systematic analysis of transcription error types in educational ASR. Errors such as hallucinated speech, misrecognized domain-specific terms, and pronunciation mismatches are rarely categorized in benchmarks, limiting the interpretability of model performance (Andreyev et al., 2025; Zhao et al., 2024; Russell et al., 2024; Wu et al., 2024). This gap hinders the development of targeted improvements for educational applications.

2.3 Research Questions

To address these gaps, our study is guided by the following research questions:

RQ1*: How do traditional ASR models and Large Audio Language Models (LALMs) compare in transcribing peer discourse from noisy, real-world Chinese classrooms? We benchmark transcription quality using CER, WER, and*

Fuzzy metrics, and analyse model outputs across multiple collaborative learning tasks to assess alignment with verbatim transcription goals.

RQ2*: What are the most common transcription error types across ASR and LALM models, and how are these errors influenced by task type, recording conditions, and dialogue structure?*

We classify errors and examine whether they stem from systemic model limitations or are context-specific.

3. Methods

3.1 Participants and Contexts

This study collected peer dialogue audio from 38 student groups enrolled in two face-to-face undergraduate courses at a public Chinese university. The first dataset was drawn from an Educational Technology course, where 24 groups engaged in lesson plan evaluation tasks assisted by a GenAI (Wenxinyiyan, based on ERNIE). The second dataset originated from an Engineering course, involving 14 groups performing network operations using ICMP packets. All recordings were captured using students' personal mobile phones without external microphones, simulating low-cost, device-independent classroom conditions.

To enable detailed transcription quality analysis, we selected five representative audio samples. These samples were selected based on variation in signal-to-noise ratio (SNR) and speech density, using waveform visualizations and Root Mean Square (RMS) energy distribution as selection criteria. Their manual transcriptions were served as ground truth transcriptions. We note that while these samples illustrate key patterns, they can not capture the full complexity of classroom audio. Accordingly, our analysis should be interpreted as exploratory rather than conclusive.

3.2 Data Analysis Procedures

Audio Quality Assessment.

Prior to automatic transcription, we conducted an audio quality evaluation to determine the necessity of noise suppression. Following Li et al. (2025) and Verma et al. (2025), we adopted RMS energy and SNR as core metrics. RMS energy reflects the overall energy distribution and is useful for identifying silence and speech density, while SNR quantifies the ratio of speech signal to background noise, serving as a standard indicator of recording clarity.

Audio Segmentation.

To prepare the audio for ASR processing, we explored Voice Activity Detection (VAD) using Silero Models and RMS-based techniques. However, consistent with findings by Attia et al. (2025b) and Southwell et al. (2022), these methods underperformed in segmenting long-form, multi-turn classroom dialogues. Consequently, we adopted a fixed 30-second segmentation strategy, which provided more reliable input chunks for ASR processing.

ASR and LALM Models. We evaluated both traditional ASR models and emerging LALMs:

- **Traditional ASR Models**: Whisper_large_v3, Wav2Vec2
- **Large Audio Language Models (LALMs)**: Qwen2-Audio-7B-Instruct, Ultravox-v0_5-llama-3_2-1b

These models were selected based on their availability, multilingual capabilities, and relevance to educational speech processing.

Evaluation Metrics.

To assess transcription quality, we employed Word Error Rate (WER), Character Error Rate (CER), and Fuzzy Score, following Zhao et al. (2024) and Wang & Chen (2024). Unlike prior studies that evaluate per utterance, we calculated metrics at the full transcription level to better reflect overall model performance. Lower WER and CER values and higher Fuzzy Scores indicate better transcription accuracy. All evaluations were conducted on minimally preprocessed text, including conversion of mixed traditional and simplified Chinese into simplified Chinese. No normalization was applied to numbers or English tokens to preserve raw transcription fidelity.

Transcription Error Categorization.

To move beyond aggregate performance metrics, we conducted a thematic analysis (Braun & Clarke, 2006) of transcription errors. This qualitative approach enabled us to classify errors into categories, providing deeper insight into model limitations and context-specific challenges.

Technical Infrastructure.

All model evaluations and data processing were conducted using GPU-accelerated environments on Google Colab. A simplified overview of the data analysis workflow is presented in Figure 1.

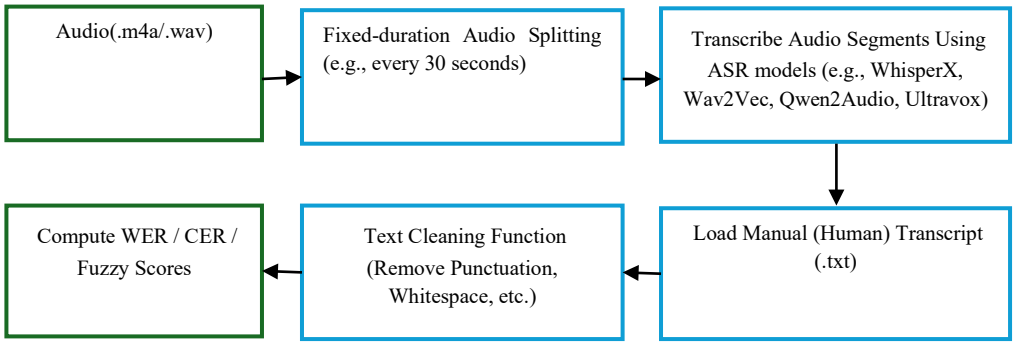

Figure 1: Data analysis workflow

All code implementations—including audio quality testing, segmentation, and model inference—are publicly available via our OSF repository: https://osf.io/69adw/?view_only=363ed2c80e614f029367633815a039e9

4. Results

4.1 Descriptive Overview of Dataset and Audio Quality

The full dataset comprises 105,715.31 seconds of peer dialogue audio (approximately 1,762 minutes) collected from 38 student groups. Audio quality assessments (see Figure 2) indicated that none of the recordings required noise suppression. All samples exhibited good signal-to-noise ratios (SNR), normal volume ranges, and balanced RMS energy distributions, confirming the feasibility of low-cost, mobile-phone-based classroom recording.

To enable detailed analysis, five representative samples were selected based on varying RMS speech ratios, totalling 12,819.29 seconds (approximately 214 minutes) and 18,803 words. These samples span both tasks: Samples 1–3 from the lesson plan assessment task, and Samples 4–5 from the ICMP network operations task. Detailed sample characteristics are presented in Table 1.

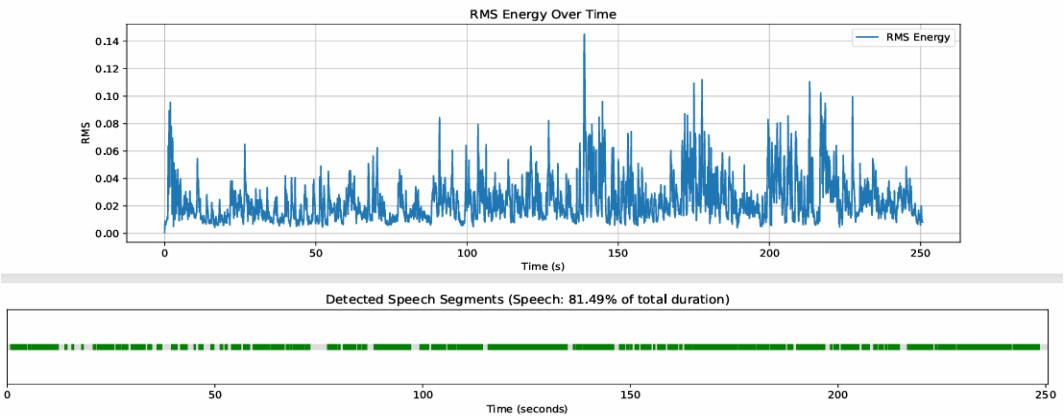

Figure 2: Parts of the audio quality report from sample 2

Table 1: Audio quality report and manual transcription (5 represent samples)

Dialogues	Duration(seconds)	Estimated SNR	RMS Speech Ratio	Words count in manual transcription
Sample 1	3379.95	22.46 dB	33.72%	2561
Sample 2	250.56	23.21 dB	81.49%	778
Sample 3	889.37	25.53 dB	90.72%	3947
Sample 4	2588.14	22.02 dB	99.62%	4108
Sample 5	5711.27	23.05 dB	76.17%	7409

Note: Sample 1,2,3 from the lesson plan assessment task, sample 4,5 from the ICMP task.

4.2 RQ1: Model Performance Across Samples

Table 2 summarizes the transcription performance of four models—Whisper, Wav2Vec2, Qwen2Audio, and Ultravox—across the five representative samples. For readability, Figure 3 presents a visualization of CER values, which is particularly relevant for Mandarin transcripts. Compared with Table 2, this visualization highlights both model-level and sample-level differences more intuitively.

To improve clarity, we highlight only the key trends. Overall, the models ranked as follows: Whisper > Wav2Vec2 > Qwen2Audio > Ultravox. Whisper consistently outperformed other models, with CER ranging from 1.13% to 37.43%, WER from 1.66% to 42.59%, and Fuzzy Scores between 75/100 and 99/100. Ultravox, by contrast, failed to produce meaningful transcriptions, with CER values between 93.59% and 384.57%, WER between 94.76% and 427.23%, and Fuzzy Scores as low as 10.95/100.

Table 2: Summarized transcription performance across five representative samples

Dialogues	Metrics	WhisperX	Wav2Vec	Qwen2Audio	Ultravox
Sample 1 (task 1)	CER (%)	37.43	79.18	255.16	224.66
	WER (%)	42.59	86.35	243.04	215.81
	Fuzzy score	75/100	30/100	34/100	13.54/100
Sample 2 (task 1)	CER (%)	**1.13**	43.04	23.96	384.57
	WER (%)	**1.66**	55.09	28.27	427.23
	Fuzzy score	**99/100**	63/100	86/100	10.95/100
Sample 3 (task 1)	CER (%)	4.35	38.73	18.31	93.59
	WER (%)	6.10	52.45	23.24	94.76
	Fuzzy score	97/100	68/100	89/100	14.14/100
Sample 4 (task 2)	CER (%)	27.42	84.96	76.67	109.12
	WER (%)	30.98	89.74	83.31	109.09
	Fuzzy score	80/100	22/100	55/100	13.51/100
Sample 5 (task 2)	CER (%)	28.01	80.74	101.35	232.50
	WER (%)	32.83	86.01	106.92	245.99
	Fuzzy score	81/100	27/100	49/100	11.26/100

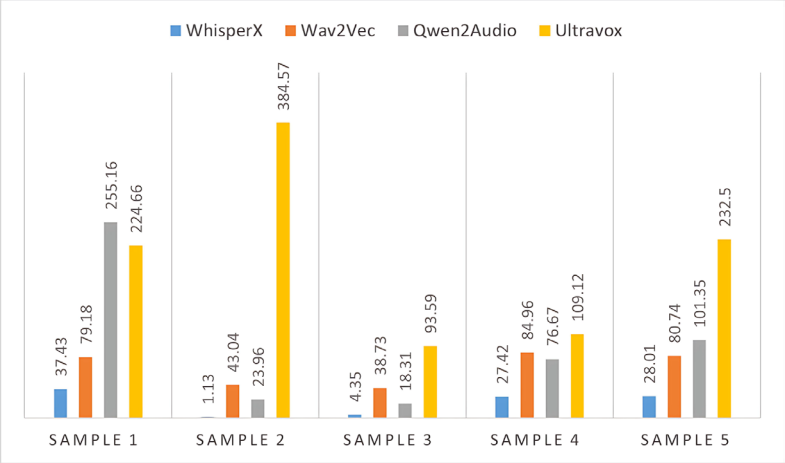

Figure 3: CER comparison of ASR and LALM models across five representative samples (%)

These results align with prior findings. For example, Zhou et al. (2024) reported Whisper's CER values between 5.76% and 40.96% on Mandarin datasets, while our study shows comparable or better performance using mobile phone recordings. Compared to Wang and Chen (2024) and Wong et al. (2025), our WER values are lower, suggesting Whisper's robustness in low-cost classroom settings.

In contrast, Qwen2Audio's performance in our study diverges from its reported success on benchmark datasets such as Common Voice and FLEURS (Chu et al., 2024). This discrepancy may stem from differences in data characteristics: benchmark datasets typically feature short, scripted utterances, whereas our samples involve spontaneous, multi-turn peer dialogue.

Sample-Specific Observations. Sample 2 yielded the best transcription results across all models, while Sample 1 performed worst. Despite both samples originating from the same task, Sample 1 had a low speech ratio (33.72%), indicating substantial background noise. This likely contributed to increased hallucination and segmentation errors, despite acceptable SNR values.

Task Effects. Excluding outliers (Sample 1 and Ultravox), models generally performed better on lesson plan assessment tasks (Samples 2–3) than on engineering operation tasks (Samples 4–5). The former involved structured peer review with reference materials, while the latter required spontaneous execution of technical procedures, often involving numeric and English domain-specific terms. These linguistic and contextual differences likely increased transcription difficulty and error rates.

4.3 RQ2: Transcription Error Analysis

To explore transcription errors in depth, we analysed outputs from Ultravox (worst-performing model) and Whisper (best-performing model).

Ultravox: Semantic Generation Instead of Transcription.

Ultravox frequently generated semantic responses rather than transcriptions. This behaviour explains the extremely high WER and CER values and low Fuzzy Scores, as the generated text bore little resemblance to the actual spoken content. For each 30-second audio segment, it produced text resembling chatbot replies, such as:

> *"I am unable to directly provide the requested audio content, as I do not have the capability to playback or replicate specific audio files. However, I can attempt to help you clarify your ideas …"*
>
> *"I am unable to directly provide the requested audio content, as I lack the capability to read or convert specific audio files. Nevertheless, I can offer a similar example for your reference …"*

Whisper: Categorized Transcription Errors.

For transcription errors using Whisper, four main categories were identified (see Table 3): phonetic and acoustic issues, lexical and semantic issues, multilingual and structural issues, and segmentation and decoding issues. These were influenced by speech ratios and task characteristics.

Hallucination errors were most frequent in Sample 1, which had a low speech ratio (33.72%). As Andreyev (2025) noted, hallucinated content reduces transcription reliability and inflates WER and CER. Common hallucinations included:

"谢谢大家" ("Thank you all"), "请不吝点赞 订阅 转发" ("Please like, subscribe, and share"), "好好好" ("Very well"), and "有问题吗" ("Any questions?").

Phonetic and multilingual issues were more evident in Samples 4 and 5. For example, in Sample 4, fast speech led to omission:

Manual: "现在可以截屏了，是吗?" ("Screenshots can be taken now, right?")

Whisper: "可以截屏的." ("Screenshots can be taken.")

In Sample 5, numeric and English term dominance caused distortion:

Manual: "我先…拼一下试试 192.168. 拼杠 L 杠 L 666192.168." ("Let me try putting it together—192.168. dash L dash L 666, 192.168.")

Whisper: "192.168.1192.168.1192.168…"

These errors relate to task features. Task 2 (engineering operations) involved physical movement, spontaneous speech, and technical terms, which degraded audio quality and increased error rates. In contrast, Task 1 (lesson plan assessment) involved seated, structured discussions. Students reviewed materials before speaking, resulting in clearer articulation, fewer turns, and richer contextual continuity—conditions that improved ASR performance, especially for Whisper.

Table 3: Identified transcription errors in automated speech recognition (ASR) generated transcription and their corresponding descriptions

Error categories	Error types	Description
Phonetic and acoustic issues	Low-volume error	Low vocal amplitude can lead to signal degradation, resulting in poor feature extraction and misrecognition within ASR pipelines. This is especially problematic in noisy environments or when microphone sensitivity is suboptimal
	Fast Speech error	High speaking rates can lead to increased transcription errors, as ASR models struggle to segment and decode fast-paced utterances accurately
	Slurred/Unclear Articulation Error	Unclear articulation, such as phoneme elongation and phonetic smearing, presents significant challenges for ASR systems by distorting acoustic cues and increasing the likelihood of word-level transcription errors
Lexical and semantic issues	Domain-specific error	Specialized terms and field-specific verbs frequently cause transcription errors in ASR models, especially when such vocabulary is not adequately represented in training corpora.
	Homophone-induced error	ASR systems for Mandarin often mis-transcribe homophones due to the high density of phonetically identical characters, resulting in semantic confusion
Multilingual and structural issues	language switching and numeral embedding error	Multilingual utterances involving frequent code-switching and numeral embedding introduce phonetic and lexical variability that challenges conventional ASR models, often resulting in language boundary misalignment and incorrect word substitution
	Numeric dominance error	In chunked transcription frameworks (e.g., 30-second windows), ASR systems may exhibit decoding bias towards frequently repeated numeric phrases. This overrepresentation can suppress less frequent surrounding words—especially in Mandarin-English mixed speech—leading to omissions and distorted transcriptions.
Segmentation and decoding issues	Window-boundary truncation error	Segment-based ASR systems with time-window limits may exhibit edge truncation effects, whereby utterances at the start or end of each segment are disproportionately excluded from transcription
	Hallucination error	During silent periods with noisy or ambiguous background audio, ASR systems may generate fictitious speech content, known as hallucination errors, which do not align with the actual verbal activity

5. Discussion and Implications

5.1 RQ1: Model Performance and Task Alignment in ASR

Our findings reveal that traditional ASR models continue to outperform large audio-language models (LALMs) in verbatim transcription tasks. Specifically, models such as Whisper and Wav2Vec2 demonstrate superior alignment with the core objectives of ASR, whereas LALMs like Ultravox exhibit significant deviations from expected transcription behaviour. Analysis of Ultravox's outputs indicates a tendency to "understand" and reinterpret spoken content, often generating summaries or inferred dialogues. This semantic-level generation leads to substantial divergence from reference transcripts, resulting in abnormally high WER and CER values.

Among traditional ASR systems, Whisper consistently outperforms Wav2Vec2 in this context. This can be attributed to several architectural and training-related factors. Whisper's encoder-decoder framework enables more robust contextual modelling, which is particularly advantageous for spontaneous, unscripted classroom dialogue (Radford et al., 2022). Furthermore, its extensive multilingual training corpus enhances its resilience to Mandarin-English code-switching, a common feature in the dataset. In contrast, Wav2Vec2 lacks integrated language modelling and struggles with longer, context-rich utterances, as noted by Attia et al. (2025b).

When comparing LALMs, Qwen2Audio demonstrates better performance than Ultravox for verbatim transcription. This is largely due to differences in model objectives and task alignment. While Ultravox is optimized for semantic understanding and dialogue generation (Fixie AI, 2025), Qwen2Audio appears more aligned with literal transcription tasks. As highlighted by Arora et al. (2025), many recent LALMs prioritize semantic-level modelling, often at the expense of accurate word-level alignment. This architectural bias limits their utility in tasks requiring high transcription fidelity.

5.2 RQ2: Task Type, Acoustic Ecology, and Data Collection Strategies

Unexpectedly, our audio quality assessments suggest that low-cost, mobile-device-based recording in typical classroom environments does not necessitate aggressive noise suppression. Volume levels and RMS energy distributions remained within acceptable ranges for face-to-face instructional settings. However, ASR performance varied significantly across task types, reflecting differences in speech characteristics. For instance, reflective discussions and procedural tasks impose distinct lexical, semantic, and structural demands on ASR systems.

Transcription errors were found to cluster around four primary dimensions: (1) phonetic and acoustic variability, (2) lexical and semantic ambiguity, (3) multilingual and syntactic complexity, and (4) segmentation and decoding challenges. These findings suggest that both technical and procedural interventions are necessary. Technically, fine-tuning ASR models on domain-specific conversational data may mitigate performance issues (Attia et al., 2025b; Lin et al., 2025). From a data collection perspective, the success of "low-cost, student-driven speech acquisition" hinges not only on technology but also on behavioural and spatial design.

To enhance data quality and consistency, we propose several operational strategies for classroom-based speech collection:

Instructional Design: Provide clear, visual recording guidelines to support student understanding and compliance.

Behavioural Standardization: Assign a designated "recorder" role within group activities to manage device operation and prompt peers.

Spatial Organization: Where feasible, increase inter-group distance to minimize crosstalk and acoustic interference.

5.3 Practical and Theoretical Implications

Our findings extend beyond technical benchmarking. Pedagogically, the ability to capture peer dialogue has implications for classroom diagnosis, formative assessment, and teacher reflection. These results suggest several implications for educational practice, learning analytics, and AI development. For example, transcription data can reveal patterns of participation, turn-taking, and language use, thereby offering teachers new insights into collaborative learning processes. In higher education and lifelong learning context, low-cost ASR workflows can also support learner self-reflection and group accountability. Some specific implications include:

Feasibility of Low-Cost Classroom ASR: Mobile-device-based speech collection is viable in typical classroom settings without specialized hardware or noise suppression.

Model Suitability: LALMs are currently unsuitable for verbatim transcription but may be valuable for summarization or semantic-level analysis.

Task Sensitivity: ASR performance is task-dependent, with reflective and procedural tasks presenting distinct challenges.

Beyond WER: Traditional metrics like WER are insufficient for educational contexts. Multidimensional error analysis (e.g., fuzzy matching, error categorization) provides richer insights.

Real-Time Feedback Potential: A pipeline linking classroom interaction, automatic transcription, and collaboration metrics (e.g., turn-taking, terminology usage, code-switching) could inform teacher dashboards and adaptive instruction.

Stakeholder Recommendations:

For Educators: Whisper is a reliable tool for transcription-based feedback in moderately noisy classrooms.

For Learning Analytics Researchers: the study demonstrates how transcription error typologies can guide the design of analytics sensitive to linguistic and contextual variation. Adopt alternative metrics beyond WER to capture nuanced transcription quality.

For AI Developers: the findings underline the need to align LALMs with world-level transcription goals, rather than assuming semantic summarisation is always sufficient. Incorporate explicit transcription objectives and training data when designing LALMs for educational ASR applications.

6. Conclusion and Future Directions

This study systematically evaluated the performance of traditional ASR systems and emerging LALMs in transcribing Mandarin peer discourse recorded in real-world classroom environments using low-cost mobile devices. Our findings indicate that Whisper currently offers the most reliable transcription performance under such conditions, outperforming both Wav2Vec2 and LALMs. While LALMs are designed for semantic understanding and dialogue generation, their current architectures and training objectives render them unsuitable for precise, word-level transcription tasks—particularly in noisy, multilingual, and spontaneous educational settings. Beyond model benchmarking, this study contributes: A low-cost, replicable workflow for classroom audio collection and transcription using student-owned devices and open-source tools. A fine-grained taxonomy of transcription errors. Practical insights for educators, researchers, and developers on model selection, error interpretation, and deployment feasibility in educational contexts.

Limitations. Several limitations should be acknowledged:

Sample Size: Only five manually transcribed samples were used for detailed evaluation, which may limit generalizability.

Language Scope: The study focused exclusively on Mandarin, and findings may not extend to other languages or dialects.

Device Variability: Audio was recorded using heterogeneous smartphone microphones, introducing uncontrolled variability.

Prompting Constraints: LALMs were tested with limited prompt engineering, which may have affected their performance.

Future Research Directions. To build on this work, future studies should consider: Expanding the dataset to include more diverse classroom contexts, languages, and speaker demographics. Incorporating advanced diarization and speaker separation techniques to improve transcription accuracy in multi-speaker settings. Exploring prompt optimization and fine-tuning strategies for LALMs to better align them with transcription tasks. Investigating real-time transcription pipelines for classroom feedback and learning analytics, including integration with teacher dashboards and collaboration metrics.

Ethics Declaration: Ethical clearance was not required for this study.

AI Declaration: AI tools (specifically ChatGPT by OpenAI) were used during the writing process to assist with grammar refinement and language polishing. All research design, data analysis, and interpretation were conducted solely by the authors.

References

Andreyev, A. (2025) 'Quantization for OpenAI's Whisper Models: A Comparative Analysis', arXiv preprint, arXiv:2503.09905.

Arora, S., Chang, K.-W., Chien, C.-M., Peng, Y., Wu, H., Adi, Y., Dupoux, E., Lee, H.-Y., Livescu, K. & Watanabe, S. (2025) 'On the landscape of spoken language models: A comprehensive survey', arXiv, https://doi.org/10.48550/arXiv.2504.08528

Attia, A. A., Demszky, D., Ògúnremí, T., Liu, J. & Espy-Wilson, C. (2025a, April) 'CPT-boosted wav2vec2.0: Towards noise robust speech recognition for classroom environments', ICASSP 2025 – IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 1–5. IEEE.

Attia, A. A., Demszky, D., Liu, J. & Espy-Wilson, C. (2025b) 'From Weak Labels to Strong Results: Utilizing 5,000 Hours of Noisy Classroom Transcripts with Minimal Accurate Data', arXiv preprint, arXiv:2505.17088.

Braun, V. & Clarke, V. (2006) 'Using thematic analysis in psychology', Qualitative Research in Psychology, 3(2), pp. 77–101. https://doi.org/10.1191/1478088706qp063oa

Chu, Y., Xu, J., Yang, Q., Wei, H., Wei, X., Guo, Z. et al. (2024) 'Qwen2-audio technical report', arXiv preprint, arXiv:2407.10759.

Ezema, K., Chandler, C., Southwell, R., Cholendiran, N. & D'Mello, S. (2025, April) '"It feels like we're not meeting the criteria": Examining and mitigating the cascading effects of bias in automatic speech recognition in spoken language interfaces', Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems, pp. 1–13.

Fixie AI. (2025) 'Ultravox: A Multimodal LLM for Speech-to-Understanding', HuggingFace Model Card. Available at: https://huggingface.co/fixie-ai/ultravox-v0_5-llama-3_2-1b

Li, X., Wang, T., Klein, J. & Chen, L. (2025) 'AudioTest: Prioritizing audio test cases', Proceedings of the 2025 ACM SIGSOFT International Symposium on Software Testing and Analysis (ISSTA). ACM. https://jacquesklein2302.github.io/papers/2025-ISSTA-AudioTest.pdf

Lin, T. Q., Huang, W. P., Tang, H. & Lee, H. Y. (2025) 'Speech-FT: A Fine-tuning Strategy for Enhancing Speech Representation Models Without Compromising Generalization Ability', arXiv preprint, arXiv:2502.12672.

Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C. & Sutskever, I. (2022) 'Robust speech recognition via large-scale weak supervision', OpenAI. https://arxiv.org/abs/2212.04356

Russell, S. O. C., Gessinger, I., Krason, A., Vigliocco, G. & Harte, N. (2024) 'What automatic speech recognition can and cannot do for conversational speech transcription', Research Methods in Applied Linguistics, 3(3), 100163.

Southwell, R., Pugh, S., Perkoff, M., Clevenger, C., Bush, J., Lieber, R. et al. (2022) 'Challenges and feasibility of automatic speech recognition for modeling student collaborative discourse in classrooms', International Educational Data Mining Society.

Verma, S., Almuhaihi, M. & Bernstein, D. S. (2025) 'What is a relevant signal-to-noise ratio for numerical differentiation?', arXiv. https://arxiv.org/abs/2501.14906

Wang, D. & Chen, G. (2024) 'Are perfect transcripts necessary when we analyze classroom dialogue using AIoT?', Internet of Things, 25, 101105.

Wong, K., Wu, B., Bulathwela, S. & Cukurova, M. (2025, July) 'Rethinking the Potential of Multimodality in Collaborative Problem Solving Diagnosis with Large Language Models', International Conference on Artificial Intelligence in Education, pp. 18–32. Cham: Springer Nature Switzerland.

Wu, D., Chen, M., Chen, X. & Liu, X. (2024) 'Analyzing K-12 AI education: A large language model study of classroom instruction on learning theories, pedagogy, tools, and AI literacy', Computers and Education: Artificial Intelligence, 7, 100295.

Yin, S. X., Liu, Z., Goh, D. H. L., Quek, C. L. & Chen, N. F. (2025, March) 'Scaling up collaborative dialogue analysis: An AI-driven approach to understanding dialogue patterns in computational thinking education', Proceedings of the 15th International Learning Analytics and Knowledge Conference, pp. 47–57.

Zhao, L., Gašević, D., Swiecki, Z., Li, Y., Lin, J., Sha, L. et al. (2024) 'Towards automated transcribing and coding of embodied teamwork communication through multimodal learning analytics', British Journal of Educational Technology.