When AI Meets Early Childhood Education: Large Language Models as Assessment Teammates in Chinese Preschools

Paper ID: 408

Abstract. High-quality teacher-child interaction (TCI) is fundamental to early childhood development, yet traditional expert-based assessment faces a critical scalability challenge. In large systems like China's—serving 36 million children across 250,000+ kindergartens—the cost and time requirements of manual observation make continuous quality monitoring infeasible, relegating assessment to infrequent episodic audits that limit timely intervention and improvement tracking. In this paper, we investigate whether AI can serve as a scalable assessment teammate by extracting structured quality indicators and validating their alignment with human expert judgments. Our contributions include: (1) **TEPE-TCI-370h** (*Tracing Effective Preschool Education*), the first large-scale dataset of naturalistic teacher-child interactions in Chinese preschools (370 hours, 105 classrooms) with standardized ECQRS-EC and SSTEW annotations; (2) We develop *Interaction2Eval*, a specialized LLM-based framework addressing domain-specific challenges—child speech recognition, Mandarin homophone disambiguation, and rubric-based reasoning—achieving up to 88% agreement; (3) Deployment validation across 43 classrooms demonstrating **18**× efficiency gains, highlighting its potential for shifting from annual expert audits to monthly AI-assisted monitoring with targeted human oversight. This work not only demonstrates the technical feasibility of scalable, AI-augmented quality assessment but also lays the foundation for a new paradigm in early childhood education—one where continuous, inclusive, AI-assisted evaluation becomes the engine of systemic improvement and equitable growth.

Keywords: Teacher-child interaction assessment, Educational speech processing, Large language models, Scalable evaluation

1 Introduction

High-quality teacher-child interaction (TCI) is the cornerstone of effective early childhood education (ECE), directly influencing children's cognitive, linguistic, and socio-emotional development [4, 13, 33]. Grounded in Vygotsky's sociocultural theory (1978), learning is understood as a socially mediated process in which language and interaction function as fundamental tools for thinking and development [36]. Especially for preschoolers who are in the early stages of developing abstract thinking and independent learning abilities, interactions with

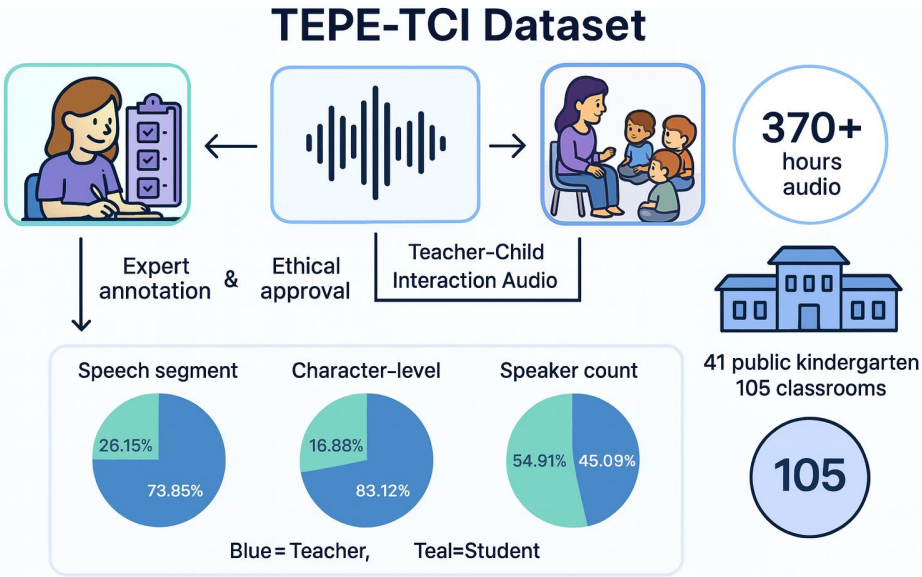

Fig. 1: Overview of the **TEPE-TCI** dataset, comprising over **370** hours of teacher–child interaction audio from **41** public preschools (**105** classrooms). Data were ethically approved, expert-annotated, and analyzed across levels.

adults and peers remain a primary driver of learning and development [23]. Current quality evaluation approach of TCI rely on trained observers using standardized instruments like the Childhood Quality Rating Scale-Emergent Curriculum (ECQRS-EC) [34] and Sustained Shared Thinking and Emotional Wellbeing (SSTEW) [30]. However, this approach requires hours of in-person observation and scoring per classroom and is *fundamentally unscalable* and *low efficiency*—a major challenge for large systems like China's, serving nearly 36 million children across over 250,000 kindergartens [21].

Recent advances in automatic speech recognition (ASR) and large language models (LLMs) suggest a promising alternative: **can we automatically assess teacher-child interaction quality directly from classroom audio?** While this concept seems intuitive, it presents unprecedented technical challenges. Unlike adult speech processing in controlled environments, preschool classrooms feature overlapping child voices, high background noise, non-standard pronunciation, and domain-specific educational terminology [10]. Furthermore, meaningful quality assessment requires understanding subtle pedagogical patterns and contextual nuances that go far beyond simple transcription accuracy.

To bridge this gap, we introduce TEPE-TCI, the first comprehensive TCI dataset in Chinese preschools, enabling **automated teacher-child interaction assessment from classroom audio.** This represents a new challenge at the intersection of speech processing, natural language processing, and educational measurement. Unlike prior work focusing on structured settings like junior or high school dialogues with clear speech and minimal noise [16, 19], our task addresses the complexities of naturalistic preschool interactions, which are

more *spontaneous, noisy,* and *linguistically diverse*. Additionally, our approach emphasizes *language-mediated interaction aspects*, enabling **scalable, rubric-grounded assessment** using real-world classroom recordings. Building on this dataset, we develop Interaction2Eval, an automated assessment framework that combines large-scale naturalistic data collection, specialized speech and language processing, and real-world deployment validation. This work demonstrates both the technical feasibility and practical scalability of AI-augmented quality assessment in early childhood education.

We make three primary contributions:

- We present **TEPE-TCI-370h**, the first comprehensive dataset of naturalistic classroom interactions with expert quality annotations in Chinese preschool contexts, comprising 370 hours of audio from 105 classrooms.
- We develop **Interaction2Eval**, a specialized LLM-based framework addressing domain-specific challenges including child speech recognition, Mandarin homophone disambiguation, and rubric-based reasoning, achieving up to 88% agreement in interaction quality assessment.
- We validate our approach through real-world deployment across 43 classrooms, demonstrating **18**× efficiency gains and the potential for shifting from annual expert audits to continuous AI-assisted monitoring.

2 Related Works

Educational Speech and Dialogue Datasets. While general speech datasets are abundant, educational speech data for early childhood contexts remains scarce, as summarized in Table 1. Most existing children's speech corpora were developed for speech recognition rather than educational assessment, including CSLU Kids [29], CMU Kids [9], PF-STAR [2], and CHILDES [27]. Datasets targeting classroom interactions predominantly focus on K-12 settings. Talk-Moves [32] provides 567 mathematics lesson transcripts with discursive move annotations, NCTE [8] contains elementary classroom recordings with CLASS observation scores, MyST [26] offers 393 hours of conversational speech from grades 3-5 students in tutoring contexts, and SimClass [1] synthesizes classroom audio with ambient noise. For early childhood contexts, resources are notably limited: WSW [31] captures preschool speech via wearables, Playlogue [15] offers 33 hours of adult-child dialogue for naturalistic play, and NCRECE PreK provides pre-kindergarten videos primarily for U.S. research [24, 25]. Notably, these datasets predominantly focus on English-speaking environments and lack standardized quality assessment annotations suitable for teacher-child interaction evaluation. Publicly available children's speech corpora for Chinese are even more limited, particularly for preschool educational contexts. The few existing datasets, including ChildMandarin [42], SingaKids [5], and SLT-CSRC [41], were designed for speech recognition tasks and either lack classroom recordings or contain no quality assessment annotations, which restricts their utility for automated interaction assessment. This gap is particularly significant given the unique challenges in Chinese preschool speech processing, including pervasive

Table 1: Summary of children's speech and educational interaction datasets. K denotes kindergarten, G denotes grade. Diar. indicates speaker diarization and Quality FW indicates quality assessment framework.

Corpus	Language	Age Range	# Speakers	Hours	Style	Diar.	Quality FW	Year
CHIEDE [12]	Spanish	3-6	59	∼8	Conversation	Partial	-	2008
CSLU Kid's Speech Corpus [29]	English	K-G10	1,100	-	Read+Spont.	N	-	2007
CMU Kids Corpus [9]	English	6-11	76	-	Read speech	N	-	1997
PF-STAR Children's Speech [2]	English	4-14	158	14.5	Read Speech	N	-	2005
MyST Corpus [26]	English	G3-G5	1,371	393	Conversation	Y	-	2024
TalkMoves [32]	English	K-12	-	-	Classroom	Y	-	2022
NCTE [8]	English	G4-G5	317	1,660 les.	Classroom	Y	CLASS+MQI	2023
SimClass [1]	English	-	-	391	Simulated	N	-	2025
WSW [31]	English	3-5	17	1,592	Classroom	Y	-	2025
Playlogue [15]	English	Preschool	-	33	Play-based	Y	DPICS	2024
SingaKids [5]	Chinese	7-12	255	75	Reading	N	-	2016
SLT-CSRC C1 [41]	Chinese	7-11	927	28.6	Reading	N	-	2021
SLT-CSRC C2 [41]	Chinese	4-11	54	29.5	Conversation	N	-	2021
ChildMandarin [42]	Chinese	3-5	397	41.3	Conversation	N	-	2024
TEPE-TCI (Ours)	**Chinese**	**3-4**	**2,550**	**370**	**Classroom**	**Y**	**ECQRS-EC+SSTEW**	**2025**

Mandarin homophone ambiguities, domain-specific educational terminology, and culturally-shaped pedagogical practices.

LLMs for Educational Assessment. Large language models have shown substantial progress in educational assessment, with GPT-4 achieving human-comparable grading accuracy [14, 20] and high inter-coder agreement in classroom dialogue analysis [16, 18]. For early childhood education, Wang *et al.* [37] explored LLM-based instructional support evaluation using CLASS protocols [35], while Whitehill *et al.* [38] developed automated the CLASS scoring with utterance-level feedback. However, these approaches focus on a single domain of CLASS framework (e.g., instructional support), rather than capturing the full range of interaction quality. Recent work has explored LLMs for developmental assessment [40], yet comprehensive teacher-child interaction assessment using multi-dimensional professional scales remains unexplored, particularly in non-English contexts where cultural and linguistic factors create distinct challenges.

To our knowledge, no existing system addresses automated ECQRS-EC or SSTEW assessment, nor audio-based teacher-child interaction evaluation in Chinese preschools that spans multiple classroom scenarios. Our work establishes both a new benchmark dataset and initial baselines for this task.

3 Problem Formulation

We formalize **automated teacher-child interaction assessment in preschool environments** as follows: Given a noisy, multi-speaker classroom audio recording A of duration T, the goal is to detect the presence or absence of behavioral indicators defined in standardized educational rubrics such as ECQRS-EC and SSTEW [30, 34]. These rubrics are widely used in ECE to assess the TCI quality based on the occurrence of developmentally supportive teaching behaviors. Each rubric item comprises multiple indicators describing concrete, observable teacher behaviors associated with varying levels of quality - from inadequate to excellent. During 3-hour classroom observation, observers assess whether such behaviors occur and assign scores accordingly. For each indicator i, the system outputs a binary judgment $y_i \in \{0, 1\}$ indicating whether the corresponding behavior

was observed in the interaction. Item-level scores are subsequently derived from indicator patterns following official scoring protocols.

3.1 Technical Challenges

This task introduces several challenges unique to the preschool-based speech processing and educational assessment:

Data Challenges: As reviewed in Section 2, publicly available children's speech corpora for Chinese preschool contexts are extremely limited. Existing datasets either lack naturalistic classroom recordings or contain no standardized quality assessment annotations, making it impossible to train or evaluate automated interaction assessment systems for this domain.

Speech Processing Challenges: Chinese preschool classrooms present compounded acoustic and linguistic difficulties. Acoustically, recordings feature high background noise, overlapping multi-speaker speech, non-standard child pronunciation, and distant-microphone conditions. Linguistically, Mandarin's tonal nature causes extensive homophone ambiguity (*e.g.*, *chénfú*: [sinking/floating] vs. [submission]), while domain-specific educational terminology (*e.g.*, *jìnqū*: [enter learning centers] vs. [go inside]) and teachers' child-directed speech patterns further challenge standard ASR systems.

Assessment Challenges: Moving from transcription to meaningful evaluation requires handling *rubric complexity*—educational assessment scales contain nuanced criteria requiring deep contextual understanding. *Temporal reasoning* is essential as quality judgments must consider interaction patterns across extended time periods. Additionally, *pedagogical knowledge* is also crucial for accurate assessment, requiring understanding of early childhood education principles and developmental appropriateness.

3.2 Task Scope and Limitations

We focus on **language-accessible aspects** of teacher-child interaction—those dimensions that can be reliably evaluated from verbal exchanges alone. While we acknowledge that high-quality interaction encompasses non-verbal elements (gestures, spatial arrangement, materials), our approach addresses the substantial subset of assessment criteria that depend on conversational patterns, questioning strategies, and linguistic scaffolding.

4 Dataset Construction

4.1 Data Collection Protocol

We collected audio using professional recording equipment (iFLYTEK H1 Pro) to capture naturalistic teacher-child interactions across diverse classroom contexts including group activities, free play, outdoor activities, and daily routines.

Scale and Scope: Forty-one preschools participated in this study, spanning three quality tiers as defined by the local education authority: district-level (N=14), municipal-level (N=12), and provincial-level (N=15). These tiers reflect differences in overall institutional quality and operating conditions. From each

preschool, two to three K1 classrooms serving 3 to 4 years old were recruited, with each classroom accommodating approximately 25 students. Eight research assistants conducted data collection over 6 weeks, resulting in 370+ hours of audio from 105 classrooms with average session length of approximately 3.5 hours.

Ethical Consideration: This study was approved by the Ethics Committee of the authors' institution (Approval No. and institution name anonymized for review). Informed consent was obtained from all teachers and parents, who were fully briefed on the study's purpose, procedures, and data use for academic research. Parents retain the right to withdraw their child's data at any time. To protect privacy, speaker diarization retains only role labels (teacher/child) without individual identification. All data are securely stored with access restricted to authorized researchers. The planned public release will include both anonymized transcripts and de-identified audio recordings, with personally identifiable information removed to protect child privacy. The dataset will be licensed for non-commercial academic use only.

4.2 Expert Annotation Process

Professional experts (the same 8 assessors) evaluated the quality of the teacher-child interaction.

Annotation Protocol: Quality assessment followed standardized ECQRS-EC and SSTEW protocols. ECQRS-EC comprises 22 items spanning four domains: Literacy, Mathematics, Science & Environment, and Diversity. SSTEW includes 15 items covering areas such as building independence and emotional wellbeing, language support, and critical thinking. Each item is organized into four performance levels (1=inadequate, 3=minimal, 5=good, 7=excellent), operationalized through behavioral indicators at each level (*e.g.*, Level 3: "Children are allowed to talk among themselves"; Level 7: "Adults scaffold children's conversations"). Our annotation adopted indicator-level binary coding: assessors judged whether each specified behavior was observed (1) or not (0). Item-level scores were subsequently derived from indicator patterns following official scoring protocols: the score was assigned at the highest level for which all indicators were met; if more than 50% of the indicators for the next level were also satisfied, the midpoint between those levels was assigned.

Our study prioritized indicators that are most representative of daily classroom practice and feasible for audio-based observation. As a result, our annotations covered 17 of the 22 ECQRS-EC items and 14 of the 15 SSTEW items, comprising a total of 112 ECQRS-EC indicators and 94 SSTEW indicators.

Quality Assurance: All assessors underwent extensive training on both assessment scales, with inter-rater reliability validation requiring $\kappa > 0.80$ agreement before independent scoring. During the validation phase, two assessors scored each classroom; after achieving reliability thresholds, single assessors completed evaluations. This rigorous process ensured high-quality ground truth annotations essential for automated system development.

4.3 Dataset Characteristics and Processing

Figure 1 presents comprehensive statistics of the **TEPE-TCI-370h** dataset. Speech segment analysis reveals teachers contributing 73.85% of segments while students account for 26.15%. Character-level analysis shows even greater teacher dominance (83.12% vs 16.88%), reflecting typical classroom discourse patterns where teachers provide more extended explanations and instructions. Interestingly, speaker count analysis shows more balanced participation (45.09% teachers vs 54.91% students), indicating active child engagement despite shorter individual contributions. This dataset fills a critical gap as the first large-scale Chinese preschool resource combining naturalistic classroom audio with standardized quality annotations.

5 Interaction2Eval Framework

Building on insights from dataset construction, we develop *Interaction2Eval* (Figure 2), a specialized framework addressing core challenges of automated assessment through three LLM-empowered agents (*Transcription*, *Refinement*, and *Evaluation* Agent)

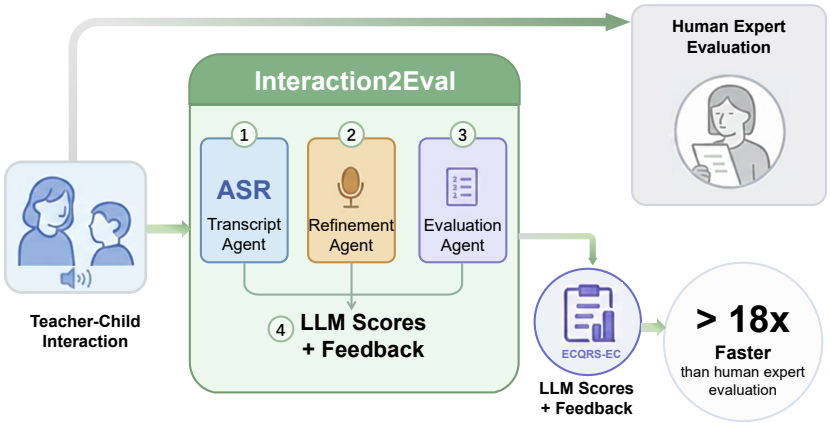

Fig. 2: Overview of the *Interaction2Eval* pipeline.

5.1 Transcription and Refinement Agents

The first stage transforms raw audio into assessment-ready transcripts. Initial ASR processing using Paraformer with diarization and punctuation restoration [3, 11] revealed systematic domain-specific errors: professional transcribers categorized errors as homophones (51.67%), extra words (20.80%), speaker identification (13.72%), punctuation/segmentation (7.75%), and omissions (6.06%). The predominance of homophone errors—significantly higher than general speech tasks—motivated our Refinement Agent design.

Given this challenge, the *Refinement Agent* leverages large language models for context-aware correction. The agent prompt incorporates educational domain knowledge, explicitly stating that the input is preschool classroom speech likely containing homophone errors common in educational settings. To guide

disambiguation, the prompt includes examples of frequent confusion pairs—such as *jìnqū* (enter learning centers) versus *jìnqù* (go inside), and *chénfú* (sinking/floating) versus *chénfú* (submission)—and instructs the model to correct errors while preserving original meaning and speaker attributions. To accommodate context length limits, processing proceeds via sliding window: the model receives transcript segments, produces corrected versions, and corrections are realigned with original timestamps and speaker labels.

5.2 Rubric-Based Evaluation Agent

The *Evaluation Agent* automates the assessment of teacher-child interactions by detecting behavioral indicators defined in the *SSTEW* and *ECQRS-EC* rubrics. We focus specifically on language-accessible indicators—those that can be reliably identified from verbal exchanges, such as whether teachers use open-ended questions or encourage child-initiated dialogue.

To construct the agent, we collaborated with expert assessors to translate rubric specifications into structured prompts. Each prompt provides the model with indicator definitions across performance levels, accompanied by contrastive examples that illustrate both high-quality and low-quality interaction patterns. The model follows an evidence-first reasoning process: it first locates relevant utterances in the transcript, then determines whether the target indicator is present, and finally provides justification grounded in specific textual evidence.

For each indicator, the agent outputs a binary judgment indicating presence or absence, along with the supporting transcript segment. The agent also generates pedagogical suggestions to support teacher professional development. Through iterative prompt refinement with domain experts, we addressed common failure modes including hallucinated evidence and misalignment between detected behaviors and rubric definitions.

6 Experiments and Results

6.1 Transcription Quality Evaluation

To ensure high-quality transcriptions, we compared two leading ASR models, Whisper-large-v3 [28] and FunASR [11], on a 5-hour test set with 16,168 reference characters. Performance was measured using *Character Error Rate (CER)*. As shown in Table 2, raw transcription errors were significant due to homophones, overlapping speech, and domain-specific vocabulary. Whisper-large-v3 exhibited a high raw CER of 35.1%, likely due to suboptimal Mandarin adaptation, while FunASR Paraformer achieved a lower initial CER of 9.9%. Our *Refinement Agent*, powered by the Qwen3-Max [39], reduced errors by addressing homophone ambiguities, lowering Paraformer's CER to **4.3%** (**56.6% relative improvement**) and Whisper's to 23.2%. These results demonstrate the agent's effectiveness, particularly for Mandarin-specific challenges.

Table 2: *CER* results. Δ / $\downarrow$: absolute/relative reduction.

Model	Raw CER (%)	After Refinement (%)	Δ (%) $\downarrow$
Whisper-large [28]	35.1	23.2	-11.9 (↓33.4%)
FunASR Paraformer [11]	9.9	4.3	-5.6 (↓56.6%)

6.2 Error Analysis and Domain Adaptation

Figure 3 visualizes the most frequently misrecognized terms in raw ASR outputs, with term size reflecting error frequency. These errors concentrate heavily in education-specific vocabulary commonly used in teacher-child interactions, including homophonic pairs like (jìn qū: enter learning centers) vs (jìn qù: go inside). This analysis confirms that semantic-level language modeling is critical for accurate transcription in specialized educational settings and validates the importance of our domain-aware refinement approach.

Fig. 3: Frequently misrecognized terms in raw ASR outputs (term size reflects error frequency).

6.3 Assessment Consistency Analysis

To evaluate the robustness of automated scoring across diverse LLM architectures and cultural alignments, we benchmark four state-of-the-art models: two international (GPT-5 [22], Gemini-2.5-pro [7]) and two Chinese-optimized (DeepSeek-v3.1 [17], Qwen3-Max [39]). Results also reveal three key findings:

(1) Chinese-adapted LLMs outperform international counterparts, particularly on both ECQRS-EC and SSTEW dimensions. DeepSeek-v3.1 achieves the highest mean agreement on both scales (87.3% for ECQRS-EC, 87.9% for SSTEW), followed by Qwen3-Max (85.7% and 86.6% respectively). This advantage likely stems from their alignment with Mandarin linguistic patterns, pedagogical terminology, and local classroom discourse norms, highlighting the critical role of *cultural and linguistic grounding* in educational AI systems.

(2) Performance on SSTEW consistently exceeds that on ECQRS-EC across almost all LLMs. This may be attributed to the distinct focuses of the two rubrics: ECQRS-EC focuses on content delivery (*e.g.,* language, early math, science, literacy) while SSTEW indicators focus on how teachers think and talk with children. Therefore, ECQRS-EC requires a more intent-sensitive and curriculum-aligned form of inference, while SSTEW can often be scored

Table 3: Indicator-level agreement between LLM predictions and human expert annotations for ECQRS-EC [34] and SSTEW [30] scales, measured by percentage agreement (%Agr.) and Cohen's Kappa (κ) which quantifies true agreement beyond chance following [6].

		Models							
		GPT-5 [22]		**Gemini-2.5-pro** [7]		**DeepSeek-v3.1** [17]		**Qwen3-Max** [39]	
Scale	**Dimension**	κ	%Agr.	κ	%Agr.	κ	%Agr.	κ	%Agr.
ECQRS-EC	Literacy	0.678	0.844	0.736	0.871	0.705	0.886	0.764	0.882
	Mathematics	0.631	0.836	0.659	0.845	0.721	0.876	0.656	0.853
	Science	0.605	0.823	0.591	0.852	0.704	0.858	0.611	0.836
	Overall Mean	**0.638**	**0.834**	**0.662**	**0.856**	**0.710**	**0.873**	**0.677**	**0.857**
SSTEW	Trust & Self-regulation	0.695	0.859	0.668	0.848	0.782	0.895	0.763	0.903
	Language & Communication	0.718	0.863	0.752	0.878	0.867	0.949	0.825	0.913
	Learning & Critical Think.	0.704	0.820	0.651	0.804	0.648	0.828	0.627	0.815
	Planning & Assessment	0.679	0.837	0.693	0.834	0.665	0.843	0.650	0.831
	Overall Mean	**0.699**	**0.845**	**0.691**	**0.841**	**0.741**	**0.879**	**0.716**	**0.866**

based on observable language structure, affective cues, and scaffolding patterns — areas where language models naturally excel.

(3) Performance is promising but imperfect—leaving room for future work. While top models reach 87.9% agreement (DeepSeek-v3.1 on SSTEW), κ values of 0.71–0.74 indicate moderate-to-substantial but not yet expert-level agreement. For reference, inter-rater reliability among trained human experts in our annotation process was $\kappa \approx 0.82 - 0.85$, suggesting that the best-performing LLMs achieve approximately 85–90% of human expert consistency. This gap highlights the challenge of encoding complex rubric logic (*e.g.,* "sustained shared thinking") into prompts, and underscores TEPE-TCI's role as a *benchmark for improvement*, not a final solution.

These findings validate the feasibility of audio-driven automated assessment, while clearly delineating its current limits—especially for context-heavy constructs like ECQRS-EC.

6.4 Scalability and Efficiency Impact

Our framework effectively reduces assessment time, marking an important step toward scalable deployment. Traditional manual assessment requires approximately 380 minutes per classroom (240 min in-person observation, 20 min indicator coding, 120 min report writing), with expert presence making evaluation inherently sequential. In contrast, Interaction2Eval completes the full pipeline in approximately 21 minutes (5 min audio processing, 12 min transcription and refinement, 4 min evaluation and report generation), achieving an 18× speedup. This efficiency gain stems from two factors: (1) *asynchronous recording* eliminates synchronous observation requirements—teachers record naturally without external observers, enabling parallel assessment at near-zero marginal cost; and (2) *automated processing* reduces manual review through high-accuracy LLM-based refinement (CER 4.3%), requiring expert validation only for flagged ambiguities than full proofreading. At system scale, this translates to substantial resource savings: assessing 100 classrooms monthly would require 633 expert-hours under traditional protocols (100 × 380 min), compared to only 35 hours

with our framework (100 × 21 min), enabling the critical shift from annual audits to continuous monitoring across China's 250,000+ kindergarten system where manual assessment remains logistically infeasible.

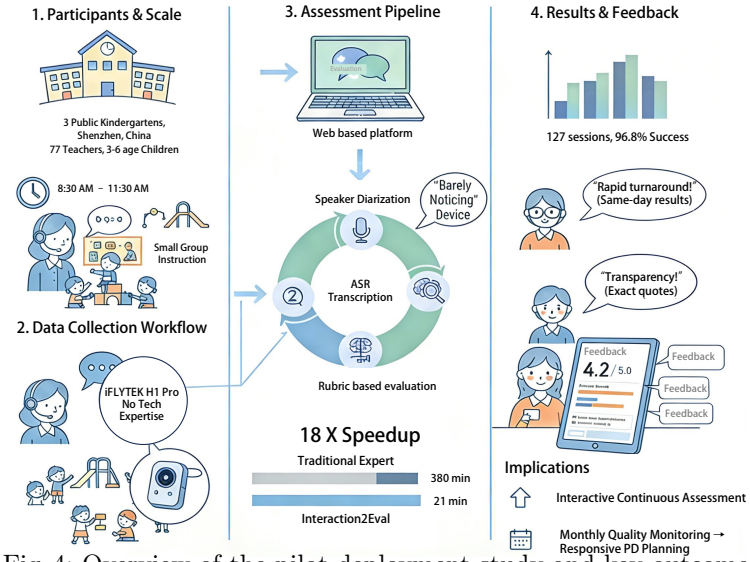

Fig. 4: Overview of the pilot deployment study and key outcomes.

7 From Research to Practice: Deployment Experience

To evaluate real-world viability, we conducted a pilot deployment of Interaction2Eval in operational preschool settings, examining both system performance and human-AI workflow integration (Figure 4).

Participants and Scale. We partnered with 3 public kindergartens in Shenzhen, China, covering 43 classrooms serving 3–6 year-old children. Each classroom had one head teacher, one assistant teacher and 25–35 students. Participating teachers (N=77) received brief training on recording equipment usage but required no technical expertise in AI or assessment protocols.

Data Collection Workflow. Teachers wore unobtrusive recording devices (iFLYTEK H1 Pro) during regular morning sessions (8:30–11:30 AM), capturing naturalistic interactions across activities including circle time, free play, and small group instruction. The wireless design ensured minimal disruption to classroom routines, with teachers reporting "barely noticing" the device after initial adaptation (typically 2–3 days).

Assessment Pipeline. Upon session completion, teachers uploaded audio files to our web-based platform via one-click transfer. The system automatically executed the full pipeline: (1) speaker diarization, (2) ASR transcription, (3) LLM-based refinement, and (4) rubric-based evaluation with indicator-level feedback. Total processing time was approximately 21 minutes for a 3-hour session, compared to 380 minutes for traditional expert observation and scoring.

Results. Over 4 weeks of preliminary deployment, the system processed 127 classroom sessions with a 96.8% success rate (4 sessions required manual intervention due to recording quality issues). Average processing time was 21 minutes per 3-hour session, compared to 380 minutes for traditional manual assessment—an **18× efficiency gain**. This speedup stems from: (1) asynchronous recording eliminating real-time observation requirements, and (2) automated transcription and scoring reducing manual coding time from hours to minutes.

Qualitative Feedback. Post-deployment interviews with 12 teachers and 3 administrators revealed positive reception. Teachers highlighted rapid turnaround and transparency, noting that results were available same-day rather than weeks later and that exact quotes supporting each score made evaluations understandable. A veteran teacher with 22 years of experience described the indicator-level feedback as a "data mirror" that enabled moving from vague intuitions to evidence-based reflection on specific strengths and gaps. Administrators noted the system enabled monthly rather than annual quality monitoring, facilitating more responsive professional development planning.

Limitations Observed. Users correctly identified system boundaries: effective for language-accessible interaction dimensions (dialogue quality, questioning strategies) but unable to assess physical environment or non-verbal engagement—areas requiring multimodal observation. This aligns with our design scope and suggests clear division of labor: AI-assisted continuous monitoring of conversational interactions, complemented by periodic expert evaluation of comprehensive quality including physical and visual aspects.

Implications. These preliminary results validate technical feasibility for scaled deployment. The 18× efficiency gain is not merely quantitative—it enables qualitative transformation from episodic external audits to integrated continuous assessment, supporting the iterative improvement cycles emphasized in contemporary early childhood education quality frameworks.

8 Conclusion and Future Directions

We introduce automated teacher-child interaction assessment from audio as a new research area, contributing **TEPE-TCI-370h**, the first large-scale preschool interaction dataset in China, alongside Interaction2Eval, a specialized LLM-based framework for this challenging problem. Our work demonstrates that high-quality automated assessment is achievable despite significant technical challenges, opening pathways for scalable ECE quality improvement.

Key findings include the critical importance of domain-specific solutions for educational speech challenges (homophones, terminology), the feasibility of expert-level LLM assessment when properly guided by educational rubrics, and the practical scalability of audio-based assessment while maintaining pedagogical validity. Future research directions include multimodal integration (audio-visual), real-time formative feedback, cross-linguistic generalization, and longitudinal studies on educational quality improvement.

References

1. Attia, A.A., Liu, J., Espy-Wilson, C.: Simclass: A classroom speech dataset generated via game engine simulation for automatic speech recognition research. arXiv preprint arXiv:2506.09206 (2025)
2. Batliner, A., Blomberg, M., et al.: The pf_star children's speech corpus (2005)
3. Bredin, H.: pyannote. audio 2.1 speaker diarization pipeline: principle, benchmark, and recipe. In: Interspeech. pp. 1983–1987 (2023)
4. Burchinal, M., Howes, C., Pianta, R.C., et al.: Predicting child outcomes at the end of kindergarten from the quality of pre-kindergarten teaching, instruction, activities, and caregiver sensitivity. Applied Developmental Science (2005)
5. Chen, N.F., et al.: Singakids-mandarin: Speech corpus of singaporean children speaking mandarin chinese. In: Interspeech. pp. 1545–1549 (2016)
6. Cohen, J.: A coefficient of agreement for nominal scales. Educational and psychological measurement (1960)
7. Comanici, G., Bieber, E., Schaekermann, M., et al.: Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. arXiv preprint arXiv:2507.06261 (2025)
8. Demszky, D., Hill, H.: The ncte transcripts: A dataset of elementary math classroom transcripts. In: BEA (2023)
9. Eskenazi, M., Mostow, J., et al.: The CMU kids corpus. Linguistic Data Consortium (1997), `https://catalog.ldc.upenn.edu/LDC97S63`, lDC97S63
10. Foulkes, P., Docherty, G.J., et al.: Sound judgements: perception of indexical features in children's speech. A reader in sociophonetics (2010)
11. Gao, Z., Li, Z., Wang, J., et al.: Funasr: A fundamental end-to-end speech recognition toolkit. In: Interspeech (2023)
12. Garrote, M., Moreno Sandoval, A.: Chiede, a spontaneous child language corpus of spanish. In: LABLITA Workshop (2008)
13. Huang, R., Zheng, H., Siraj, I.: Quality in chinese preschool classrooms: Its structural influencing factors and associations with child development. Early Education and Development (2025)
14. Impey, C., Wenger, M., Garuda, N., et al.: Using large language models for automated grading of student writing about science. International Journal of Artificial Intelligence in Education (2025)
15. Kalanadhabhatta, M., et al.: Playlogue: Dataset and benchmarks for analyzing adult-child conversations during play. IMWUT (2024)
16. Li, X., Han, G., Fang, B., He, J.: Advancing the in-class dialogic quality: Developing an artificial intelligence-supported framework for classroom dialogue analysis. The Asia-Pacific Education Researcher (2025)
17. Liu, A., Feng, B., Xue, B., et al.: Deepseek-v3 technical report. arXiv preprint arXiv:2412.19437 (2024)
18. Long, Y., Luo, H., Zhang, Y.: Evaluating large language models in analysing classroom dialogue. npj Science of Learning (2024)
19. Long, Y., Zhang, Y.: Enhanced classroom dialogue sequences analysis with a hybrid ai agent: Merging expert rule-base with large language models. arXiv preprint arXiv:2411.08418 (2024)
20. Mendonça, P.C., Quintal, F., Mendonça, F.: Evaluating llms for automated scoring in formative assessments. Applied Sciences (2025)

21. Ministry of Education of the People's Republic of China: Statistical bulletin on the development of national education in 2024. http://www.moe.gov.cn/jyb_sjzl/sjzl_fztjgb/202506/t20250611_1193760.html (2025), accessed: 2025-06-11
22. OpenAI: Gpt-5 system card (2025), https://openai.com/index/gpt-5-system-card/, accessed: 2025-08-07
23. Piaget, J., Cook, M., et al.: The origins of intelligence in children, vol. 8. International universities press New York (1952)
24. Pianta, R., Hamre, B., Downer, J., et al.: Early childhood professional development: Coaching and coursework effects on indicators of children's school readiness. Early Education and Development (2017)
25. Pianta, R.: National center for research on early childhood education teacher professional development study (2007-2011) (2016)
26. Pradhan, S., Cole, R., Ward, W.: My science tutor (myst)–a large corpus of children's conversational speech. In: LREC-COLING. pp. 12040–12045 (2024)
27. Pye, C.: The childes project: Tools for analyzing talk (1994)
28. Radford, A., Kim, J.W., Xu, T., et al.: Robust speech recognition via large-scale weak supervision. In: ICML (2023)
29. Shobaki, K., Hosom, J.P., Cole, R.A.: CSLU: Kids' speech version 1.1 (2007). https://doi.org/10.35111/q5tn-8096, https://catalog.ldc.upenn.edu/LDC2007S18, lDC2007S18
30. Siraj, I., et al.: The Sustained Shared thinking and Emotional Well-being (SSTEW) Scale: Supporting process quality in early childhood (2023)
31. Sun, A., Feng, T., et al.: Who said what wsw 2.0? enhanced automated analysis of preschool classroom speech. arXiv preprint arXiv:2505.09972 (2025)
32. Suresh, A., Jacobs, J., Harty, C., et al.: The talkmoves dataset: K-12 mathematics lesson transcripts annotated for teacher and student discursive moves. arXiv preprint arXiv:2204.09652 (2022)
33. Sylva, K., et al.: The effective provision of pre-school education (EPPE) project: Final Report: A longitudinal study funded by the DfES 1997-2004 (2004)
34. Sylva, K., Siraj, I., Taggart, B., Kingston, D.: Early Childhood Quality Rating Scale—Emergent Curriculum (ECQRS-EC). Teachers College Press (2025)
35. Teachstone: The classroom assessment scoring system® (class). https://teachstone.com/class/, accessed: 2025-09-18
36. Vygotsky, L.: Mind in society: The development of higher psychological processes. harvard university press, cambridge, ma (1978)
37. Wang, J., Hankour, K., Zhang, Y., et al.: Classroom observation: Evaluating instructional support automatically in classroom for young children. PRML (2025)
38. Whitehill, J., LoCasale-Crouch, J.: Automated evaluation of classroom instructional support with llms and bows: Connecting global predictions to specific feedback. arXiv preprint arXiv:2310.01132 (2023)
39. Yang, A., Li, A., Yang, B., et al.: Qwen3 technical report. arXiv preprint arXiv:2505.09388 (2025)
40. Yang, Y., Shen, Y., Sun, T., Xie, Y.: Validating the effectiveness of a large language model-based approach for identifying children's development across various free play settings in kindergarten. arXiv preprint arXiv:2505.03369 (2025)
41. Yu, F., Yao, Z., Wang, X., et al.: The slt 2021 children speech recognition challenge: Open datasets, rules and baselines. In: SLT. pp. 1117–1123. IEEE (2021)
42. Zhou, J., et al.: Childmandarin: A comprehensive mandarin speech dataset for young children aged 3-5. In: ACL. pp. 12524–12537 (2025)