Leveraging ASR and LLMs for Automated Scoring and Feedback in Children's Spoken Language Assessments

Natarajan Balaji Shankar[1], *Kaiyuan Zhang*[1], *Andre Mai*[1], *Mohan Shi*[1], *Alaria Long*[2], *Julie Washington*[2], *Robin Morris*[3], *Abeer Alwan*[1]

[1]Dept of Electrical and Computer Engineering, University of California Los Angeles, USA
[2]School of Education, University of California Irvine, USA
[3]Dept of Psychology, Georgia State University, USA

{balaji1312, kaiyuanzhang, andremai, shimohan}@ucla.edu,
{alarial, julie.washington}@uci.edu, robinmorris@gsu.edu, alwan@ee.ucla.edu

Abstract

This paper explores the use of automatic speech recognition (ASR) and large language models (LLMs) for automated scoring and feedback generation in spoken language assessment. We design a three stage pipeline that (1) optimizes ASR hypotheses from student speech, (2) performs task-based scoring using LLMs, and (3) generates natural language feedback justifying each score. We evaluate this pipeline using audio responses from 3rd-8th grade students in the Atlanta, Georgia area, recorded as part of the Test of Narrative Language. Our results show that LLMs can reliably replicate expert annotations while providing interpretable feedback. We further analyze model performance across demographic factors, including dialect and reading proficiency, to assess equity. Our findings demonstrate the promise of ASR and LLMs for robust, explainable, and fair assessment of children's spoken narratives.

Index Terms: Children's Speech, Spoken Language Assessments, Automatic Speech Recognition, Computer-Assisted Language Learning

1. Introduction

Spoken language assessments (SLA) are essential for evaluating oral language skills and diagnosing impairments in educational settings. Recent advances in Automatic Speech Recognition (ASR) and Natural Language Processing (NLP) have enabled scalable, feedback-driven solutions that reduce teacher workload and support PreK-12 instruction (i.e., education from pre-kindergarten through grade 12) [1, 2]. Yet traditional SLA methods remain labor intensive and prone to rater variability. The emergence of large language models (LLMs) marks a paradigm shift in automated assessment. Pretrained on massive text corpora, LLMs like GPT 4o [3] possess strong linguistic priors and can evaluate transcripts for grammar, content, and coherence [4]. Their success in educational tasks such as essay scoring and feedback generation [5, 6] has spurred interest in extending LLMs to spoken assessments, particularly for tasks such as evaluating children's oral narratives. Early work shows that, when paired with robust ASR, LLMs can support holistic evaluation of narrative speech [7].

However, applying LLM-based SLA systems to children's speech introduces unique challenges. Children's speech is marked by high acoustic and linguistic variability due to ongoing developmental changes in articulation and language acquisition [8]. ASR systems tend to perform poorly on child speech, particularly for children who speak non-mainstream dialects such as African American English (AAE), or have language-related disabilities [9, 10]. These disparities are in part due to the underrepresentation of such populations in training data, resulting in systemic ASR biases [11, 12]. The consequence is degraded transcripts that can distort downstream scoring and feedback generation, threatening the fairness and validity of the assessment. Additionally, common ASR practices such as Whisper's tendency to generate hallucinations [13] and normalize speech by removing hesitations [14] can eliminate linguistic features that are informative for SLA tasks.

LLM-based feedback generation introduces further complexity. While promising results have been reported on text-based feedback [15, 16, 17], providing interpretable and pedagogically useful feedback in spoken SLA, particularly for low-resource populations, remains underexplored [18]. In parallel, recent work has evaluated GPT 4o for pronunciation assessment across multiple granularities, demonstrating the potential of large multimodal models for both scoring and feedback on fluency-related dimensions [19].

Prior research in SLA and educational NLP has explored a range of approaches to address these challenges from modeling coherence using ASR transcripts [20] to leveraging pronunciation training systems [21] and adapting essay scoring methods [22]. Multimodal extensions of LLMs have been shown to handle acoustic-text fusion for pronunciation feedback with competitive results [23]. Within child SLA specifically, prior work has explored multitask learning to cope with limited data [24], automatic assessment of prosodic and linguistic markers for reading fluency [25], the use of ASR-derived embeddings [26], and scoring of transcribed spoken responses [27, 28, 29]. Related work on the Test of Narrative Language (TNL) has evaluated BERT-based feature fusion techniques and investigated fairness across dialects, reading levels, and language impairment status [30, 31, 32].

In this work, we present a three-stage pipeline for fair, accurate, and interpretable spoken language assessment scoring using LLMs. We evaluate our system on audio recordings from 3rd-8th grade students in Atlanta, Georgia, performing tasks from the Test of Narrative Language (TNL). In the first stage, we generate and rerank ASR hypotheses to minimize transcription errors. In the second stage, an LLM scores and provides feedback on each test item using rubric-aligned prompts applied to annotated transcripts, augmented with a list of common ASR mistranscriptions. In the final stage, a separate LLM evaluates the quality of the generated feedback. We assess system performance across demographic and ability-based subgroups, comparing model outputs to expert scores and conducting human-in-the-loop evaluations of feedback quality. Our results show that LLMs can produce reliable scores, informative feedback, and equitable performance with no task-specific fine-tuning.

2. Methodology

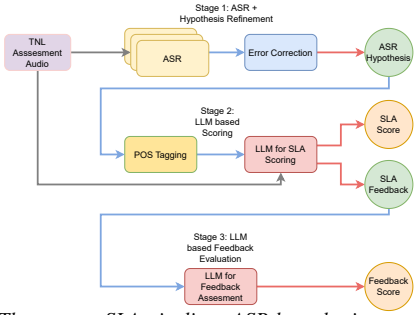

Figure 1: *Three-stage SLA pipeline: ASR hypothesis reranking, LLM-based scoring and feedback generation, and LLM-powered feedback evaluation. (Red → LLMs, Green → Output text, Orange → Output scores)*

2.1. ASR Benchmarking and Hypothesis Correction

Building on prior findings that ASR systems struggle with children's speech particularly for speakers of African American English (AAE) and students with language or reading impairments, we begin by benchmarking a diverse set of commercial ASR offerings such as Azure[1], AWS Transcribe[2], Assembly AI[3], Deepgram[4], ElevenLabs[5], Speechmatics[6], RevAI[7], Gemini[8], and GPT 4o[9], alongside open source models like Whisper [33], Canary Flash [34], and Parakeet v2 [35]. These systems vary in their training data, handling of dialects, and treatment of disfluencies.

To mitigate the transcription errors that persist across systems, we implement a generative hypothesis correction step. Specifically, we aggregate outputs from all ASR systems and apply Generative Speech Error Correction (GenSec) [36, 37] using a Flan-T5 model [38] fine-tuned on ASR correction pairs derived from child speech [39]. Since the model was fine-tuned to operate on five hypotheses, we select the top five ASR outputs (ranked by WER) and feed them into GenSec, which fuses information across hypotheses to produce a corrected transcript. The resulting transcripts serve as inputs to the scoring and feedback stages of our pipeline.

2.2. LLM-Based Narrative Scoring

The second stage of our system involves scoring each narrative response using large language models (LLMs). Our goals are twofold: to evaluate how well LLMs can replicate human scoring and to examine how model type, input modality, and preprocessing strategies affect scoring performance. Unlike prior work on the same dataset, which trained a BERT-based scoring model on ASR transcripts [31], we do not fine-tune or perform task-specific training of any models. Instead, we evaluate LLMs in a zero-shot setting, relying solely on rubric-aligned prompting to assess scoring accuracy. We provide rubric-aligned prompt templates for both scoring and feedback generation, as shown in Table 1

To guide the models, we introduce a lightweight entity annotation step that highlights key linguistic elements in each transcript. Specifically, we use SpaCy's [40] part-of-speech (POS) tagger to embed inline annotations within the ASR output, providing weak supervision to help the LLM attend to relevant grammatical features. In addition, each prompt includes examples of common ASR mistranscriptions, and instructs the model to be tolerant of mispronunciations and minor recognition errors. The model is explicitly prompted to assign a rubric-aligned score for each item and generate accompanying feedback explaining the rationale behind the assigned score.

We evaluate a range of models, including general purpose LLMs (GPT 4o[9], Deepseek v3[10]), and reasoning focused models (Gemini 2.5 Pro[8], GPT o3[9], GPT o4 Mini[9], Claude 3.7[11]). All models are prompted using scoring templates aligned with the assessment rubric, enabling direct comparison across both model size and token cost. Beyond text-only inputs, we also test multimodal scoring capabilities by using models that accept raw audio in addition to text. For multimodal LLMs that support audio input (GPT 4o, Gemini 2.5 Pro), we provide the original speech waveform along with the ASR transcript, allowing the model to leverage acoustic cues in its evaluation. Finally, we conduct an ablation study using human ground truth transcripts. This allows us to isolate the scoring ability of each model in the absence of ASR noise, and to estimate the upper bound of model performance under ideal transcription conditions.

Table 1: *LLM prompt templates for narrative scoring and feedback evaluation*

Score Generation
You are an assistant that scores a child's narrated story based on content. For each Qi, assign 1 point if the child produced the exact target word (including any required inflection or tense), otherwise 0. Be forgiving of clear pronunciation errors, such as {**ASR examples**}. Some target words/phrases are pre-marked with tags such as `<NOUN>...</NOUN>`. After each Qi score, provide feedback Fi explaining what the child said and whether it met the target. Our scoring rubric is as follows: {**scoring rubric**}
Feedback Evaluation
The feedback pertains to a section of the Test of Narrative Language, where the teacher first tells a story, then asks the child to retell it, and evaluates whether the retelling includes specific target words or phrases. As an English teacher, you will evaluate this feedback on three dimensions: {**feedback rubric**}

2.3. Evaluation of Scoring Feedback

In the final stage of our pipeline, we generate natural language feedback for each scored item to explain the assigned score and support instructional use. While LLMs have shown promise in free-text generation, standardized methods for evaluating feedback quality in spoken language assessment remain limited. We adopt a model-based evaluation approach using three LLMs (GPT o3, Gemini 2.5 Pro, and Claude 3.7) as automated raters. Each model receives the student transcript, assessment prompt, and model-generated feedback, and assigns a score from 1 to 10 across three criteria [18]: 1) Explainability: Is the model's rationale for the score clearly presented and understandable to an educator? 2) Usefulness: Does the feedback offer meaningful insights that could help an educator better understand the student's performance or guide instruction? 3) Accuracy: Is the feedback grammatically correct, and does it accurately reflect both the student's response and the scoring criteria?

A representative subset of feedback samples (ensuring coverage across score ranges and demographic groups) is also rated by a trained educational expert using the same rubric. Human and model ratings are directly compared to assess alignment.

[1] https://azure.microsoft.com/
[2] https://aws.amazon.com/transcribe/
[3] https://www.assemblyai.com/
[4] https://deepgram.com/
[5] https://elevenlabs.io/
[6] https://speechmatics.com/
[7] https://rev.ai/
[8] https://gemini.google.com/
[9] https://openai.com/
[10] https://deepseek.com/
[11] https://claude.ai/

While each criterion is rated on a 1-10 scale, we additionally group scores into qualitative bands: 10 indicates perfect alignment, >5 indicates partial satisfaction, and ≤ 5 indicates unsatisfactory feedback. This stratification provides a more interpretable view of how well model-generated feedback meets pedagogical expectations. This dual evaluation enables us to assess the validity of LLM-generated feedback and the potential of LLMs to serve as scalable proxies for expert judgment.

3. Experiments

3.1. Data

This paper uses audio recordings of 184 3rd-8th grade students from Atlanta, Georgia as they perform the "Test of Narrative Language (TNL) - Task 1, Story Retelling" assessment (data collected in [41]) where students are read a story by the test administerer. Of these 184 recordings, 157 were from grades 3 and 4. The students are then asked to retell the story and graded on their ability to use the set of pre-determined test keywords from the original story-telling. These keywords contain story elements (eg. character names, locations, times, important objects, and action verbs) that must be retold in the same verb tense and order to receive credit in the test scoring. For example, if a test item contained the sentence, "**Tim eats** his lunch while **Matt** plays **football**" where the bolded words are the scored keywords, the child will receive points for two of the four keywords if they retell it as "**Tim** played **football** while **Matt ate** lunch," as the word order or tense of the other two keywords are incorrect. Each child's assessment was administered and audio recorded by a trained member of the project staff according to the TNL standardization manual protocols. The recordings were then independently scored by a speech-language pathologist and a second trained speech-language staff member. If disagreements occurred in scoring, the two scorers reviewed the differences to come to a consensus. Each child's score was an integer between 0 and the total number of test keywords. Recordings were taken at the child's school. Audio was recorded in stereo at a sampling rate of 48kHz. All recordings were resampled to mono with a sampling rate of 16kHz for experimentation. Each of the children gave a response with an average length of about 5 min, resulting in approximately 16 total hours of speech. The dataset additionally contains demographic metadata on the students in the following categories: 1) the presence of reading/language impairment, 2) the student's reading ability (good or poor) as measured by standardized reading tests, and 3) the speaker's dialect (either African American English (AAE) or Southern American English) as labeled by the authors according to the procedure in [11].

3.2. Evaluation Metrics

To evaluate the output of each ASR system and our generative correction pipeline, we compute Word Error Rate (WER) between hypothesis and reference transcripts. Both are normalized using Whisper's text normalization pipeline [33] to ensure consistency in casing and punctuation. In addition, we report Speech WER [14], which retains hesitations to more accurately capture transcription performance on spontaneous child speech. For LLM-based scoring, we evaluate the agreement between predicted and expert-annotated scores using classification accuracy. We report Pearson's correlation coefficient (PCC) as an interpretable baseline for linear association, Spearman's rank correlation coefficient (SCC) for ordinal association, and Kendall's Tau (KRC) as a stricter ordinal measure emphasizing pairwise concordance. For generated feedback, we first compute the percent agreement between the feedback content and the judgment of a human/separate LLM evaluator holistically across three dimensions: Explainability, Usefulness, and Accuracy. To assess overall alignment with expert evaluation, we then compute the Quadratic Weighted Kappa (κ) between LLM-predicted and human-assigned scores.

4. Results and Discussion

4.1. ASR Results and Impact of Hypothesis Correction

Table 2: *Zero-shot WER and Speech WER across ASR systems on the TNL recordings. GenSec refers to the output transcript generated by a hypothesis correction model that fuses hypotheses from the top ASR systems. Bold indicates best performance. * denotes open-source models.*

Model	WER (%)	Speech WER (%)
ElevenLabs	11.4	13.5
RevAI	12.1	13.8
Gemini	12.3	14.6
AWS Transcribe	13.1	15.1
Speechmatics	14.3	16.7
Deepgram	13.3	17.1
Canary Flash*	15.6	17.5
Parakeet v2 *	16.8	17.9
Assembly AI	14.1	18.2
Azure	15.3	19.4
GPT 4o	21.0	24.9
Whisper Large v3*	24.5	28.4
GenSec	**11.0**	**12.9**

Table 2 reveals a clear performance gap between commercial and open-source ASR systems in transcribing spontaneous child speech. Commercial systems such as ElevenLabs (11.4% WER) and RevAI (12.1%) consistently outperform open-source alternatives like Whisper Large v3 (24.5%) and Parakeet v2 (16.8%). This disparity suggests that proprietary systems, likely benefiting from large-scale private training corpora, are better equipped to handle the variability in children's speech. Furthermore, when accounting for hesitations via Speech WER, open-source systems degrade more severely, highlighting their limitations on spontaneous child language. Despite these challenges, our generative error-corrected ensemble (GenSec) [39] which fuses hypotheses from the top-performing ASR systems using a fine-tuned Flan-T5 model outperforms all individual models, achieving the lowest WER (11.0%) and Speech WER (12.9%). This demonstrates the value of hypothesis-level fusion in correcting transcription errors and improving transcript quality.

Table 3: *Zero-shot WER and Speech WER breakdown of GenSec transcript by dialect, language impairment status (Lang. Impair.), and reading ability. AAE = African American English, RD = Reading Disability, LI = Language Impairment.*

Category	Group	Count	WER	Speech WER
Dialect	AAE	116	11.7	13.5
	non-AAE	68	9.8	12.0
Lang. Impair.	Control	32	8.2	9.8
	RD Only	60	11.5	13.2
	RD + LI	27	13.6	15.4
Reading Status	Good Readers	152	8.2	9.8
	Poor Readers	32	11.7	13.6
Overall		184	11.0	12.9

Table 3 further examines ASR performance across demographic subgroups following hypothesis correction. While GenSec maintains low overall error rates, significant disparities persist. Transcriptions for speakers of African American English (AAE) show elevated WER (11.7%) compared to non-AAE peers (9.8%), underscoring continued dialect mismatches in training data. Likewise, children with both reading and language impairments (RD + LI) experience the highest WER (13.6%), suggesting that even hypothesis correction does not fully resolve accessibility gaps for learner populations.

Table 4: *Scoring performance of LLMs on the TNL assessment. Acc. = Accuracy, PCC = Pearson's correlation coefficient, SCC = Spearman's rank correlation, KRC = Kendall's Tau.*

Model	Acc.	PCC	SCC	KRC
General Purpose Models				
Deepseek v3	52.17	0.892	0.909	0.778
GPT 4o	33.15	0.618	0.612	0.460
Reasoning Models				
GPT o4 Mini	81.52	0.969	0.964	0.886
Claude 3.7	80.43	0.961	0.955	0.870
Gemini 2.5 Pro	82.07	0.958	0.953	0.858
GPT o3	**86.96**	**0.977**	**0.973**	**0.909**

Table 5: *Scoring performance of multimodal LLMs (GPT 4o, Gemini 2.5 Pro) and best reasoning LLM (GPT o3) under different inputs: ASR transcripts, ASR with audio, and Ground Truth (GT) Transcripts*

Model	Input	Accuracy	PCC	SCC	KRC
GPT 4o	ASR	33.15	0.618	0.612	0.460
	ASR + Audio	33.70	0.687	0.668	0.511
	GT	40.76	0.750	0.721	0.568
Gemini 2.5 Pro	ASR	80.98	0.954	0.946	0.848
	ASR + Audio	82.07	0.967	0.963	0.882
	GT	84.78	0.967	0.961	0.882
GPT o3	ASR	86.96	**0.977**	0.973	**0.909**
	GT	**88.04**	**0.977**	**0.975**	**0.909**

4.2. Evaluating LLMs for Narrative Scoring

Table 4 presents the scoring accuracy of various LLMs on the narrative task. GPT o3 outperforms all other models, achieving 86.96% accuracy and strong correlation with human ratings (PCC = 0.977, SCC = 0.973, KRC = 0.909). Lightweight models such as o4 Mini also perform competitively (81.52% accuracy), suggesting that smaller models, while less aligned with human judgment, may offer reasonable trade-offs between performance and computational cost. The strong performance of GPT o3 suggests that reasoning steps in LLMs contribute to better agreement with rubric-based scoring tasks.

Table 5 evaluates multimodal models that accept additional audio input (GPT 4o, Gemini 2.5 Pro) alongside the best performing model from Table 4 (GPT o3), and shows that transcription quality strongly affects LLM scoring accuracy. We conduct an ablation using ground truth (GT) transcripts to isolate the effect of ASR errors, and find that, as expected, GT consistently yields the highest performance across models, underscoring the sensitivity of LLMs to transcription quality. Notably, GPT o3 shows only a minor drop in accuracy when moving from GT to ASR input (88.04% to 86.96%), suggesting strong robustness to transcription noise, likely because common transcription errors were explicitly included in the prompt. For models that support audio input (GPT 4o, Gemini 2.5 Pro), adding speech provides only modest improvements over transcript-only inputs. GPT o3 remains the most robust across all conditions, while GPT 4o lags significantly, particularly when using ASR transcripts alone.

Table 6 reveals two key trends in model performance across demographic subgroups. First, scoring accuracy with ASR inputs consistently trails GT-based accuracy, closely mirroring subgroup-specific WER, indicating that transcription errors impact downstream scoring. Second, even after controlling for WER, predicted scores for non-mainstream groups show persistent deficits. For example, AAE speakers see no accuracy gain when switching from ASR to GT input, and RD + LI students perform worse than the control group even with perfect transcripts. These discrepancies suggest that current LLM-based scoring may carry latent biases or struggle to generalize to non-standard language patterns, raising fairness concerns that extend beyond ASR quality.

Table 6: *WER and GPT o3 scoring accuracy across student subgroups using ASR and ground-truth (GT) transcripts*

Category	Group	Count	WER	ASR Acc.	GT Acc.
Dialect	AAE	116	11.7	86.21%	86.21%
	non-AAE	68	9.8	88.24%	91.18%
Lang. Impair.	Control	32	8.2	87.50%	92.50%
	RD Only	60	11.5	85.00%	91.67%
	RD + LI	27	13.6	84.78%	88.89%
Reading Status	Good	152	8.2	87.50%	92.50%
	Poor	32	11.7	86.84%	88.16%
Overall		184	11.0	86.96%	88.04%

Table 7: *Percent agreement of scores for generated feedback, and quadratic weighted kappa (QWK) between LLM and human ratings.*

	Human	GPT o3	Claude 3.7	Gemini 2.5 Pro
% agreement	**93**	90	91	90
QWK	–	0.24	0.51	**0.64**

4.3. Evaluation of Scoring Feedback

Table 7 summarizes evaluation results for the feedback generated in Stage 2 of the pipeline (by GPT o3). We assess the quality of this feedback using both human ratings and separate LLMs (GPT o3, Claude 3.7, Gemini 2.5 Pro), following the rubric introduced in Section 3.3. The percent agreement metric reflects whether the evaluator, human or model, judged the feedback as satisfactory across all three criteria: Explainability, Usefulness, and Accuracy. Human ratings indicate that 93% of the feedback was satisfactory, with LLM evaluators closely matching at 90-91%. To capture more fine-grained scoring correlation, we compute Quadratic Weighted Kappa (κ) between the total human rating and model-predicted scores. Gemini 2.5 Pro achieves the highest agreement ($\kappa = 0.64$), while GPT o3 lags with notably lower correlation ($\kappa = 0.24$), despite having generated the original feedback. When analyzing ratings across dimensions, we find that all models consistently undervalue the 'Usefulness' of feedback compared to human evaluators, suggesting limitations in LLMs' ability to assess pedagogical utility. These results point to the potential of LLMs as scalable raters of generated feedback, while also underscoring the need for targeted prompting or calibration to better reflect human priorities in instructional contexts.

5. Conclusion

We present a three-stage pipeline for automated spoken language assessment using LLMs, combining ASR hypothesis refinement, rubric-aligned scoring, and natural language feedback generation. On narrative responses from 3rd–8th graders, commercial ASR systems outperform open-source models, but our generative error correction module (GenSec) achieves the best performance (WER 11.0%). For scoring, GPT o3 yields the highest accuracy (86.96%) and strongest alignment with human ratings (PCC 0.977), with ablations revealing minimal degradation from ASR errors. Feedback is rated favorably by both human and model evaluators (up to 93% agreement; QWK 0.64), though human raters found the feedback more pedagogically useful. These results show that LLM-driven scoring systems can produce reliable scores and meaningful feedback. Unlike traditional SLA systems, ASR-LLM pipelines do not require training separate models for each task, enabling greater flexibility and scalability. While challenges remain, particularly for underrepresented groups (dialects, reading abilities, and language impairments), our findings highlight the promise of ASR-LLM pipelines for scalable, accurate, and equitable evaluation of children's spoken narratives.

6. Acknowledgements

The research is supported in part by the NSF and the IES, U.S. Department of Education (DoE), through Grant R305C240046

to the U. at Buffalo. The opinions expressed are those of the authors and do not represent views of the IES, DoE, or the NSF.

7. References

[1] J. Bryant *et al.*, "How artificial intelligence will impact k–12 teachers," *McKinsey & Company*, 2020.

[2] A. L. Bailey *et al.*, "Addressing bias in spoken language systems used in the development and implementation of automated child language-based assessment," *Journal of Educational Measurement*, 2025.

[3] A. Hurst *et al.*, "Gpt-4o system card," arXiv:2410.21276, 2024.

[4] S. Abdurahman *et al.*, "Perils and opportunities in using large language models in psychological research," *PNAS nexus*, vol. 3, no. 7, 2024.

[5] S. Kim and M. Jo, "Is gpt-4 alone sufficient for automated essay scoring?: A comparative judgment approach based on rater cognition," in *Proceedings of the Eleventh ACM Conference on Learning Scale*, 2024, pp. 315–319.

[6] J. Han *et al.*, "Llm-as-a-tutor in efl writing education: Focusing on evaluation of student-llm interaction," in *Proceedings of the 1st Workshop on Customizable NLP: Progress and Challenges in Customizing NLP for a Domain, Application, Group, or Individual*, 2024, pp. 284–293.

[7] S. Bannò *et al.*, "Can gpt-4 do l2 analytic assessment?" in *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications*, 2024.

[8] S. Lee *et al.*, "Acoustics of children's speech: Developmental changes of temporal and spectral parameters," *JASA*, vol. 105, no. 3, pp. 1455–1468, 1999.

[9] S. Dutta *et al.*, "Challenges remain in building asr for spontaneous preschool children speech in naturalistic educational environments," *Interspeech*, pp. 4322–4326, 2022.

[10] G. Yeung and A. Alwan, "On the difficulties of automatic speech recognition for kindergarten-aged children," *Interspeech*, 2018.

[11] A. Koenecke *et al.*, "Racial disparities in automated speech recognition," *Proceedings of the National Academy of Sciences*, vol. 117, no. 14, pp. 7684–7689, 2020.

[12] A. Johnson *et al.*, "An analysis of large language models for african american english speaking children's oral language assessment," *Journal of Black Excellence in Engineering, Science, & Technology*, vol. 1, 2023.

[13] A. Koenecke *et al.*, "Careless whisper: Speech-to-text hallucination harms," in *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, 2024, pp. 1672–1681.

[14] R. Ma *et al.*, "Adapting an asr foundation model for spoken language assessment," in *9th Workshop on Speech and Language Technology in Education*, 2023, pp. 104–108.

[15] M. Stahl *et al.*, "Exploring llm prompting strategies for joint essay scoring and feedback generation," in *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications*, 2024.

[16] C. Xiao *et al.*, "Human-ai collaborative essay scoring: A dual-process framework with llms," in *Proceedings of the 15th International Learning Analytics and Knowledge Conference*, 2025, pp. 293–305.

[17] S. Gombert *et al.*, "From the automated assessment of student essay content to highly informative feedback: A case study," *International Journal of Artificial Intelligence in Education*, vol. 34, no. 4, pp. 1378–1416, 2024.

[18] N. Phan *et al.*, "Automated content assessment and feedback for finnish l2 learners in a picture description speaking task," in *Interspeech*, 2024, pp. 317–321.

[19] K. Wang *et al.*, "Exploring the potential of large multimodal models as effective alternatives for pronunciation assessment," in *Interspeech*, 2024.

[20] K. Zechner *et al.*, "Automated scoring of speaking tasks in the test of english-for-teaching," *ETS Research Report Series*, vol. 2015, no. 2, pp. 1–17, 2015.

[21] M. Sancinetti *et al.*, "A transfer learning approach for pronunciation scoring," in *2022 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2022, pp. 6812–6816.

[22] M. Uto *et al.*, "Neural automated essay scoring incorporating handcrafted features," in *Proceedings of the 28th International Conference on Computational Linguistics*, Dec. 2020, pp. 6077–6088.

[23] K. Fu *et al.*, "Pronunciation assessment with multi-modal large language models," arXiv:2407.09209, 2024.

[24] I. Baumann *et al.*, "Nonwords pronunciation classification in language development tests for preschool children," *Interspeech*, 2022.

[25] G. Bailly *et al.*, "Automatic assessment of oral readings of young pupils," *Speech Communication*, vol. 138, pp. 67–79, 2022.

[26] Y. Getman *et al.*, "wav2vec2-based Speech Rating System for Children with Speech Sound Disorder," in *Interspeech*, 2022, pp. 3618–3622.

[27] D. Ramesh and S. K. Sanampudi, "An automated essay scoring systems: a systematic literature review," *Artificial Intelligence Review*, vol. 55, no. 3, pp. 2495–2527, 2022.

[28] B. W. Lee *et al.*, "Pushing on text readability assessment: A transformer meets handcrafted linguistic features," in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, Nov. 2021, pp. 10 669–10 686.

[29] R. Gale *et al.*, "Automatic assessment of language ability in children with and without typical development," in *2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society*, 2020, pp. 6111–6114.

[30] H. Veeramani *et al.*, "Towards automatically assessing children's oral picture description tasks," in *9th Workshop on Speech and Language Technology in Education*, 2023, pp. 119–120.

[31] A. Johnson *et al.*, "An equitable framework for automatically assessing children's oral narrative language abilities," *Interspeech*, 2023.

[32] H. Veeramani *et al.*, "Large language model-based pipeline for item difficulty and response time estimation for educational assessments," in *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications*, 2024.

[33] A. Radford *et al.*, "Robust speech recognition via large-scale weak supervision," in *International Conference on Machine Learning, ICML 2023*, vol. 202. PMLR, 2023, pp. 28 492–28 518.

[34] K. C. Puvvada *et al.*, "Less is more: Accurate speech recognition & translation without web-scale data," in *Interspeech*, 2024, pp. 3964–3968.

[35] D. Rekesh *et al.*, "Fast conformer with linearly scalable attention for efficient speech recognition," *2023 IEEE Automatic Speech Recognition and Understanding Workshop*, pp. 1–8, 2023.

[36] S. Bannò *et al.*, "Towards end-to-end spoken grammatical error correction," in *2024 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2024, pp. 10 791–10 795.

[37] K. Manuel *et al.*, "Towards improving asr outputs of spontaneous speech with llms," in *Proceedings of the 20th Conference on Natural Language Processing*, 2024, pp. 339–348.

[38] H. W. Chung *et al.*, "Scaling instruction-finetuned language models," *Journal of Machine Learning Research*, vol. 25, no. 70, pp. 1–53, 2024.

[39] N. B. Shankar *et al.*, "Chser: A dataset and case study on generative speech error correction for child asr," arXiv:2505.18463, 2025.

[40] M. Honnibal *et al.*, "spacy: Industrial-strength natural language processing in python," 2020, available at https://spacy.io/.

[41] E. L. Fisher *et al.*, "Executive functioning and narrative language in children with dyslexia," *American journal of speech-language pathology*, vol. 28, no. 3, pp. 1127–1138, 2019.