International Journal of Artificial Intelligence in Education (2025) 35:282–303
https://doi.org/10.1007/s40593-024-00435-9

ARTICLE

Supporting Literacy Assessment in West Africa: Using State-of-the-Art Speech Models to Assess Oral Reading Fluency

Owen Henkel[1] · **Hannah Horne-Robinson**[2] · **Libby Hills**[3] · **Bill Roberts**[4] · **Josh McGrane**[5]

Accepted: 13 October 2024 / Published online: 28 January 2025

Abstract

This paper reports on a set of three recent experiments utilizing large-scale speech models to assess the oral reading fluency (ORF) of students in Ghana. While ORF is a well-established measure of foundational literacy, assessing it typically requires one-on-one sessions between a student and a trained rater, a process that is time-consuming and costly. Automating the assessment of ORF could support better literacy instruction, particularly in education contexts where formative assessment is uncommon due to large class sizes and limited resources. This research is among the first to examine the use of the most recent versions of large-scale speech models for ORF assessment in the Global South.We find that the best performing model, Whisper V2, with no additional fine-tuning, produces transcriptions of Ghanaian students reading aloud with a Word Error Rate of 10.3. When these transcriptions are used to produce fully automated ORF scores, they closely align with scores generated by expert human raters, with a correlation coefficient of 0.98. These results were achieved on a representative dataset (i.e., students with regional accents, recordings taken in actual classrooms), using a free and publicly available speech with no additional fine-tuning. This model's strong performance on real-world classroom data, combined with its accessibility and simplified implementation, suggests potential for scaling ORF assessment in lower-resource, linguistically diverse educational contexts.

Keywords Automatic Speech Recognition · Educational assessment · Foundational Literacy

Extended author information available on the last page of the article

Introduction

UNESCO estimates the number of illiterate individuals to be over 800 million, nearly 10% of the global population, the majority of whom live in the Global South (Pritchett, 2013). Education systems require multiple components to effectively teach reading, with regular and high-quality assessment of basic reading skills being a critical element. Frequent formative assessment has been shown to be an important component of effectively teaching students to read fluently, as it can inform classroom-level teaching practice, providing real-time feedback to students and teachers (Black & Wiliam, 2009; Nag, 2017).

One of the most widely used measures of foundational literacy skills is oral reading fluency (ORF),which is generally considered a reliable predictor of reading comprehension, particularly in early grades (Fuchs et al., 2001; Spear-Swerling, 2006). However, assessing ORF typically requires a one-to-one interaction between the assessor and the student for several minutes, meaning there is considerable time, cost, and complexity associated with administering and grading these tasks (Magliano & Graesser, 2012). For governments and educational institutions in many countires, conducting these types of literacy assessments regularly can be particularly challenging due to large rural populations, unreliable infrastructure, and weak organizational capacity (Dubeck & Gove, 2015). Given the fact that even the largest scale reading assessments are currently administered infrequently and to a small fraction of students in each country, automating ORF assessment would be a valuable development.

Automatic Speech Recognition (ASR) models may be able to automate ORF assessment, however, ASR models have previously failed to perform satisfactorily on child voice datasets (Jain, Barcovschi, Yiwere, Bigioi, et al., 2023a, b). This has primarily been due to several factors that differentiate children's speech from adult speech, such as a higher pitch, developing vocabulary, and irregular pronunciation (Gerosa et al., 2009). In the context of LMICs, there are added complications, most notably limited data representing regional accents, and restricted access to recording equipment (Feng et al., 2024; Nouza et al., 2023). However, the rapidly evolving capabilities of the newest generation of large language, speech, and multi-modal models (e.g., GPT-4, BardGemini, Claude, Whisper V2, and wav2vec 2.0) may have changed this. The growth in the size of training datasets and the number of model parameters appears to have dramatically improved their overall performance benchmark and ability to generalize to new datasets, and in some cases has led to the emergence of new and unexpected abilities (Stiennon et al., 2022; Wei et al., 2022). In the specific case of large-scale speech models there have been considerable advances in ASR technology, particularly with open-source models such as wav2vec 2.0 and Whisper V2, which have demonstrated human-level performance across a range of datasets. Based on these results, it seemed plausible that large-scale speech models could potentially assess ORF, however, there has been limited research on the potential use of this new generation of language and speech models for literacy assessment, and the research that exists has been conducted almost exclusively in high-income countries. It is critical to ensure that less well-resourced education

systems can also harness the potential advantages of these technologies. This study addresses this gap in the literature in three ways.

First, we introduce a publicly available dataset, the University of Oxford & Rising Academies Ghanaian Student Oral Reading Fluency Dataset (hereafter referred to as the Ghana ORF Dataset). This dataset consists of the results from ORF tasks, administered to 130 students between the ages of 9 and 18 in Ghana. The Ghana ORF dataset contains the original passages, the audio recording of students reading these passages, and high-quality human transcriptions. Secondly, this paper provides the first empirical research on the ability of the most recent generation of open-source ASR models to transcribe the speech and oral reading of Ghanaian students. Third, this study compares automatically generated WCPM scores to those produced by expert humans. By triangulating the findings from these three experiments—testing model performance on a representative Ghanaian dataset, benchmarking against a U.S. dataset to examine accent and environment effects and validating automated scoring against human raters—we aim to better understand the feasibility and limitations of using large-scale speech models for automatic ORF assessment. The top-line finding—that the best-performing model produces ORF scores highly similar to those of human raters—has important implications for educational assessment, and particularly for the formative assessment of reading ability in the Global South.

Prior Work

ASR Model Architectures and Benchmarks

Automatic Speech Recognition systems have been an active area of research for a long time, but in the past 5 years innovative approaches to model architecture have led to dramatic performance increases. While ASR has been commonplace in various use cases from telephone call-routing, to captioning live television, to transcribing communications, these were typically tightly controlled contexts, and ASR models often struggled when asked to deal with new datasets (Adams, 2011). This was primarily because the most common approach for training ASR models was using a supervised training paradigm, where models were trained using collections of high-quality corpora of paired audio records and human generated transcripts (Jain, Barcovschi, Yiwere, Corcoran, et al., 2023a; Nouza et al., 2023). However, this approach limited the size and diversity of the datasets that were available to train models and resulted in a situation where ASR models matched or exceeded human-level performance on canonical ASR evaluation datasets but have struggled with overfitting and domain shift when tested on other datasets or in "real-world" settings (Geirhos et al., 2020; Panayotov et al., 2015; Radford et al., 2022).

In response to these challenges, other researchers proposed unsupervised pre-training methods, most notably wav2vec 2.0 (Baevski et al., 2020). Unsupervised learning approaches have the advantage of being trained on audio without requiring human labels, thus unlocking large datasets of unmarked speech for use and has made it feasible to train models on up to 1,000,000 h of audio (Zhang et al.,

2022). However, this approach comes with a notable drawback; because the model is trained only on input speech, a fine-tuning phase is needed for them to perform tasks such as speech transcription, often be a complex process that requires a skilled practitioner to implement (Radford et al., 2022).

Recognizing the strengths and limitations of both approaches, more recent research has started to explore an approach dubbed "lightly supervised" which uses both unsupervised and supervised data, relax the need for all data to be human-validated transcripts and allowing the use of audio transcriptions found on the internet, such as subtitles for movies or shows and audio books (Radford et al., 2022). The largest and most successful implementation of this approach to date is Whisper V2, which was trained on over 680,000 h of audio from diverse recording setups, environments, speakers, and languages. When tested across a variety of ASR datasets, Whisper V2 purported to outperform wav2vec 2.0, achieving an average Word Error Rate (WER) of 12.8 compared to 29.3 (Radford et al., 2022).

Bias and Differential Performance in ASR

The issue of algorithmic bias has been a concern for many years, with research dating back to the 1960s highlighting the potential for bias in educational testing (Baker & Hawn, 2022). In recent literature, researchers have used various terms to describe bias, including unfair, discriminatory, and demographic disparity (Mehrabi et al., 2021; Mitchell et al., 2021). Some sources of bias are related to insufficient accurate datasets that are used to train models.

Algorithmic bias has been observed in various educational contexts, including dropout prediction, course failure prediction, automated essay scoring, language proficiency assessment, and student emotion detection" (Jiang & Pardos, 2021; Kizilcec & Lee, 2021; Loukina et al., 2014). A previous example of this could be the inclusion of student demographics as predictors of grades, which can maintain bias by predicting lower grades for certain groups (Jiang & Pardos, 2021).

In the specific case of ASR, differences in accents can also impact the accuracy of transcription, as training datasets may not account for regional differences in pronunciation. For instance, in 2020, researchers found that large-scale commercially available ASR models had a significantly higher WER for black Americans than for white Americans—a fact they attributed to the datasets used to train the models (Koenecke et al., 2020). While there is some evidence that the most recent generation of ASR models have been trained on larger and more diverse datasets, thereby performing better on various cases, they still may not fully account for regional differences in accent and pronunciation. However, given the large number of potentially impacted groups and the sizeable, and sometimes non-public, datasets used to train the models, it can be challenging to determine whether an algorithm functions equally well for different groups or if it perpetuates bias (Tempelaar et al., 2020). For these reasons is it critical to have diverse data sources and model evaluation datasets, particularly for linguistically or culturally underrepresented groups, to address algorithmic bias effectively (Suresh & Guttag, 2021).

ASR in Educational Settings

While some limited applications of ASR in educational contexts were used in the past, they were mainly restricted to adult language learning products. This was likely because there are several differences between adults' and children's speech – both acoustically and linguistically – that create extra challenges to build accurate ASR for children. The acoustic frequency of children's speech is higher than adults because of their shorter vocal tracts, and children are also more likely to use imaginative words, ungrammatical phrases, incorrect pronunciations, all of which present challenges for language modelling (Gray et al., 2014). Due to the limited scale of high-quality child speech datasets available, it has historically been challenging to adapt ASR models to these characteristics. Additionally, most datasets are collected in classrooms, and background noise is ubiquitous and difficult to eliminate from recordings without moving students to a separate room, which would be unfeasible. For these reasons until the last few years, ASR was often considered too complex, time-consuming and expensive to implement in a classroom with young students (Cincarek et al., 2009; Gerosa et al., 2009; Gray et al., 2014). However, with recent progress in ASR for adult speech (with top performing models now rivaling human level performance on many tasks, including transcription), there has been an increase in its application to non-adult educational settings (Baevski et al., 2020; Radford et al., 2022).

A variety of approaches have attempted to enhance the performance of automatic child speech recognition systems, and some context-specific finetuned speech models have shown improved performance (Bolaños et al., 2013a, b; Fan et al., 2022). However, their implementation in real-world educational environments is hindered by two main challenges. Firstly, fine-tuning can be a complicated process that demands advanced technical skills, and the acquisition and labeling of the fine-tuning dataset remains an expensive and labor-intensive undertaking (Jain, Barcovschi, Yiwere, Corcoran, et al., 2023a; Radford et al., 2022). Secondly, models that demonstrate human level performance when trained on a specific dataset can struggle to perform well when tested on a new dataset (Geirhos et al., 2020).

Research in fine-tuning self-supervised speech model learning have shown improvements in child speech recognition. In one set of experiments, Jain et al., (2023a, b) explored using wav2vec 2.0 with different pretraining and fine-tuning approaches. Their models outperformed the unmodified wav2vec 2.0 on child speech, when tested on well-established student speech corpora, such as MyST Corpus, PFSTAR dataset, and CMU Kids dataset. They reported that, with as little as 10 h of child speech data for fine-tuning, they were able to achieve a WER below 15, on par with adult speech. In a subsequent study, they compared WhisperV2, a large ASR model released by OpenAI, with fine-tuned self-supervised models like wav2vec 2.0. Whisper, trained with significantly more data and multilingual resources, was expected to improve child speech recognition through fine-tuning. The results showed significant improvements in ASR performance compared to the non-fine-tuned Whisper model, reaching parity with human raters. We consider these results (summarized in Fig. 1 below) to be the best results to-date in child speech transcription. However, it remains unclear how ASR models like Whisper would perform in

Story Text Being Read	Human Transcript of Student Reading (Errors marked in yellow)
A farmer went out one day to search for a lost calf. He went to the valley and searched by the riverbed, among the reeds, behind the rocks and in the rushing water. He climbed the slopes of the high mountain with its rocky cliffs. He looked behind a large rock in case the calf had huddled there to escape the storm. And that was where he stopped. There, on a ledge of rock, was a most unusual sight. An eagle chick had hatched from its egg a day or two earlier, and had been blown from its nest by a terrible storm. He reached out and cradled the chick in both hands. He would take it home and care for it.	a farmer went out one day to search for a lost scarf he went to the valley and searched by the riverbed among the reeds behind the rocks and in the rushing water he climbed the slopes of the high mountain with its rocky cliffs he locked he looked behind a large rock in a case the calf had head headed there to escape the storm and that was where the and that was where he stopped there, on a ladge ledge of rock, was a most usual sight an eagle chick had hatched from its egg a day or two earlier and had been blowned from its nest by a tr-tribal storm he reached out and cr-crawded the chick in both hands he would take it home and care for it
122 words in original text	15 Student reading errors, 67 seconds long
Audio file of student reading	

Fig. 1 Example of Passage and Human Scoring

Table 1 Performance of ASR Model on Child-Specific Datasets Measured with WER

	MyST Corpus	PFSTAR dataset	CMU Kids dataset
wav2vec 2.0 large (10 h finetuning)	27.17	3.50	21.35
wav2vec 2.0 large (65 h finetuning)	7.42	2.99	14.18
Whisper V2 Large (unmodified)	25.00	73.68	12.69
Whisper V2 Large (10 h finetuning)	15.79	2.88	15.22
Whisper V2 Large (65 h finetuning)	13.34	4.17	17.11

Note: Adapted from Jain et. al. 2023a

scenarios which combined children's speech, regional accents, and environmental noise (e.g., background conversations Table 1).

Assessing Reading Fluency

Reading fluency refers to the ability to read a text accurately, with proper expression, and at a sufficient rate so that working memory is not overloaded (National Reading Panel, 1998). Non-fluent readers may read words correctly but slowly and

laboriously, with a stilted and monotonous tone, and struggle with word grouping (Fuchs et al., 2001; Rasinski et al., 2009). As readers become more fluent, the require less working memory to recognize words, allowing them to allocate more attention to maintaining the text in working memory, accessing background knowledge, and making inferences (LaBerge & Samuels, 1974).

Empirical research has demonstrated that without sufficient automaticity, it is impossible to retain a sentence in working memory. For example, if a reader cannot complete a sentence within approximately twelve seconds, they will have forgotten the beginning of the sentence by the end. Much of the research on reading fluency focuses on ORF due to its practicality (Hasbrouck & Tindal, 2006) as there are challenges in measuring silent reading fluency, particularly in classroom settings.

While ORF can be assessed in various ways, one of the most widely used measures is Words Correct Per Minute (WCPM) (Roehrig et al., 2008). This typically involves asking a student to read a passage aloud for a set period, while the rater marks reading errors. When the student finishes, the errors are tallied, subtracted from the total number of words in the original text, and then divided by the time the student spent reading. The resulting score, WCPM, is a combined measure of accuracy and rate, with higher scores considered to be indicative of stronger ORF. These assessments are typically conducted one-on-one with a student and a trained rater, making it time-consuming and costly to administer. Therefore, there are potential benefits of using ASR to partially or fully automate ORF assessments (Duong et al., 2011).

Existing Oral Reading Fluency Datasets

While there are numerous datasets that have been used to evaluate ASR of adult speech, there are far fewer that explicitly focus on children's speech. Some of the best-known include MyST Corpus, PFSTAR dataset, and CMU Kids dataset. However, these datasets are typically focused on children's speech rather than oral reading. This distinction is important because when assessing ORF, it is crucial for the transcripts to accurately preserve reading errors. In normal ASR, the goal of the task is typically to produce a readable transcript, and the models themselves are often trained on combined audio and transcripts. These transcripts, however, often omit repetitions and mispronunciations. While this is a reasonable decision for most ASR applications, in the context of ORF, it is precisely these errors that provide critical information about a student's reading fluency.

A further complication is that differences in accents can have large impacts on transcription quality. For example, a typical American pronounces the word "water" (i.e., H2O) with a soft "t" in the middle and "schwa" at the end, whereas in West Africa, the more common pronunciation would have both harder "t" sounds in the middle and a more distinct "er" ending. While both pronunciations are correct depending on the context, a model that has been trained primarily on audio from the United States will be less likely to correctly transcribe the word when it is spoken with a West African accent.

To our knowledge, there are no publicly available datasets of children's voices speaking English that (a) focus on students' oral reading rather than speaking and (b) are from countries other than the United States, Canada, or the United Kingdom.

Current Study

The goal of this study is threefold: (1) to create and share a novel dataset of students in Ghana completing ORF tasks, (2) to evaluate the performance of open-source ASR models in transcribing audio clips of Ghanaian students reading, and (3) to determine if these transcripts could be used to automatically produce accurate estimates of students' ORF. The novel dataset introduced in this paper is important because it addresses the scarcity of datasets featuring student oral reading, particularly from Anglophone LMICs. The three experiments aim to better understand the feasibility of using ASR as part of a pipeline to automatically estimate ORF. Additionally, because this dataset has demographic information linked to each response, we can investigate whether the models grade students' answers differently based on their background attributes (age, gender, home language ability) compared to human raters.

Rising Academies and Oxford University Oral Reading Fluency Dataset

This dataset is comprised of a series of literacy assessments conducted with students from Rising Academies, a school network in Ghana. These schools are located the peri-urban periphery of Accra, and the student come from families working primarily in the informal sector, with annual household incomes ranging from $1,000 -$5,000 PPP. Students aged between 13 and 18 were asked to complete various tasks aimed at assessing the reading abilities. They were first given a quick screening which determined that 32 of them were nonreaders. The remaining 130 students were asked to read a series of short passages aloud. All the reading passages were taken from the 2016 PrePIRLS (now known as PIRLS Literacy), an international comparative study that assesses the reading skills of fourth-grade students in 60 countries, and which was specifically designed for application in LMICs and nations with lower literacy rates (Methods & Procedures in PIRLS, 2016, 2018). The ORF tasks were audio-recorded and later transcribed by human annotators. Table 2 shows an overview of the dataset, including student characteristics.

Table 2 Key Characteristics of Ghana ORF Dataset

Total Students	162
Pre-readers vs Readers	32 | 130
Age: Average, min, max	13, 9, 18
Speak English at Home: count (percent)	35 (15%)
Female vs Male: count (percent)	88 (68%) / 42 (32%)
Total Recordings	239

To generate high-quality transcriptions one of the authors listened to the recording while also reading along with the source story, to produce a faithful transcript of the student reading – paying careful attention to included added or repeated words. During this process the were able to pause and relisten to audio files. We estimate that this careful transcription took approximately 5 min per story. As a quality control measure a second author conducted the same process independently on a subset of the audio files, but we found no systematic differences in transcripts. Figure 1 shows an example transcript.

Experimental Design

The study conducted three interrelated experiments to investigate the feasibility and efficacy of using large-scale speech models to automatically assess ORF in Ghana. The experimental design was structured to systematically address the key challenges associated with automating ORF assessment in low-resource, linguistically diverse educational contexts.

In the **first experiment**, we evaluated the performance of two state-of-the-art speech models, Whisper V2 and wav2vec 2.0, on a novel dataset of Ghanaian students' oral reading recordings. This dataset, the University of Oxford and Rising Academies Reading Comprehension Dataset (Ghana ORF Dataset), provided a representative sample of student recordings with regional accents and background classroom noise. By testing the models' ability to accurately transcribe these recordings without any fine-tuning, we aimed to evaluate the feasibility of using off-the-shelf speech models for ORF assessment in Ghana. The **second experiment** sought to contextualize the models' performance on the Ghana ORF Dataset by comparing it to their performance on a dataset of American students' oral reading recordings, the CU Kids Read and Summarized Story Corpus (CU Dataset). This comparison allowed us to better understand the impact of accent and recording environment on transcription accuracy, factors that could potentially limit the effectiveness of using speech models for ORF assessment in diverse educational settings. In the **third experiment**, we used the transcriptions generated by Whisper V2 in the first experiment to automatically calculate students' ORF scores using the WCPM metric. We then compared these fully automated scores to scores generated by expert human raters. This experiment aimed to validate the feasibility of using speech model transcripts to generate reliable ORF scores, a critical step in automating the assessment process.

Measuring Model Performance

In the context of this study, it is crucial to distinguish between the two key measures employed: Word Error Rate (WER) and WCPM. WER is a standard metric used to evaluate the accuracy of ASR systems. It quantifies the discrepancy between a reference transcript (usually a human-generated "ground truth") and the transcript generated by the ASR model. Formally WER is defined as:

$$WER = I + D + S/N$$

Where I, D, and S represent the number of word insertions, deletions, and substitutions, identified in the candidate text, and N is the total number of words in the source text (Ali & Renals, 2018). This number is often then multiplied by 100 for ease of interpretation. Hence, WER scores have a definitional minimum score of 0 which indicates that the two documents are identical. While WER can technically exceed 100, the typical range is between 0 and 100, with lower scores being considered better.

In the first and second experiments of this study, WER was used to evaluate the performance of Whisper V2 and wav2vec 2.0 on the Ghana ORF Dataset and the CU Dataset, respectively. These experiments aimed to determine the feasibility of using off-the-shelf speech models for accurately transcribing oral reading recordings in diverse educational settings.

While WER is an essential measure of transcription accuracy, it does not fully capture the nuances of ORF. WCPM is a widely used measure of ORF that combines aspects of reading accuracy and reading rate. It is calculated by subtracting the number of reading errors from the total number of words in each passage and then dividing this value by the time taken to read the passage in minutes.

$$WCPM = (totalwordsinstory - totalerrors) \times (60/durationofaudioclipinseconds)$$

The inclusion of WCPM in this study is crucial because it provides a more comprehensive evaluation of the potential for automating ORF assessment. While WER gives us an indication of the accuracy of the speech models in transcribing student recordings, WCPM allows us to evaluate whether these transcriptions can be used to generate reliable measures of ORF, which is essential for practical applications of ASR in educational settings.

Results

Evaluating Speech Models' Performance on the Ghana ORF Dataset (Experiment 1)

To better evaluate the performance of two state-of-the-art open-source ASR models, namely Whisper V2 large and wav2vec 2.0, we conducted a comparative analysis using a subset of the Ghana ORF Dataset. Our selection consisted of 60 audio files chosen based on their audio quality and the completeness of the stories narrated by the students. This curated subset represented a best-case scenario for evaluating the underlying model performance, as the audio clips generally had better audio quality, minimal background noise, and more fluent student readers. It is important to note that we did not fine-tune or modify these models, using them "out-of-the-box" for this evaluation.

To ensure a fair comparison, we standardized the text of all three transcriptions by removing punctuation and converting the text to lowercase. We then compared

Table 3 WER by model on Ghana ORF Dataset

	Whisper V2	wav2vec 2.0
Average WER	10.3	27.3
Range—Max | Min	25 | 1	69 | 5
Standard Deviation	5.5	14.4

the model-generated transcriptions to the high-quality human transcriptions, which served as the ground truth. The WER was calculated using JiWER, a simple and fast Python package designed to evaluate ASR systems. JiWER calculates the WER of the model transcriptions relative to the ground truth reference text, which, in this case, was the expert human transcription. The results of this analysis are summarized in Table 3 below.

Whisper V2 achieved a WER of 10.3, which is typically considered strong performance when transcribing real-world speech. This result is even more notable given the historical challenges of transcribing children's speech accurately. Furthermore, the WER of 10.3 surpassed the average performance (12.8) of Whisper V2 across a variety of ASR benchmarks, highlighting its robustness and adaptability. In contrast, wav2vec 2.0 demonstrated significantly higher error rates, with an average WER of 27.3, suggesting that Whisper V2 or similar models would be preferable for transcription tasks in this context.

None-the-less, the WER of 10.3 indicates that open-source ASR models have dramatically improved compared to previous performance benchmarks for assessing children's speech with regional accents. This improvement suggests that these models may have reached a level of accuracy sufficient for use in educational applications. To better understand this potential, we first investigate whether the model's performance varies depending on students' accents. Subsequently, we explore whether the WER is low enough to generate reliable estimates of students' ORF. By addressing these questions, we aim to provide insights into the feasibility and reliability of employing state-of-the-art ASR models in educational settings, particularly for assessing and supporting students' reading skills.

Manually inspecting the differences between how the model transcribed the audio files was also informative. Qualitatively, wav2vec 2.0 appears to preserve reading errors more faithfully (e.g., "he locked, he looked", "had head headed"), whereas Whisper V2 more often corrects repetitions and reading errors. Furthermore, wav2vec 2.0 appears to transcribe some words phonetically, relative to an American accent. For instance, the way the Ghanaian student says the word "rushing" is phonetically quite similar how an average American might say the word "Russian". Other examples include "rog", for "rock" and the space between "river" and "bed".

These differences are likely attributable to distinct training approaches. wav2vec 2.0's purely unsupervised training approach, makes it more sensitive to the acoustic properties of the reading. In contrast, Whisper V2's lightly supervised approach means that the model is trained on transcript-audio pairs, which often omit repeated words, and capture what a speaker was trying to say, rather than record every verbal stumble (Radford et al., 2022). Whisper V2 has likely learned this from its training

Whisper V2 large	**wav2vec 2.0** (transcription differences highlighted in yellow)
a farmer went out one day to search for a lost scarf he went to the valley and searched by the riverbed among the reeds behind the rocks and in the rushing water he climbed the slopes of the high mountain with its rocky cliffs he looked behind a large rockin a case the calf had headed there to escape the storm and that	a farmer went out one day to search for a lost scarf he went to the valley andserchedby the river bedamong the reeds behind the rocks and in therussian water he climbed the slopes of the high mountain with its rocky cliffs he locked he looked behind a largerog in a case the calf hadheadheaded there to escape the storm and that
Audio file of student reading	

Fig. 2 Alternate Transcripts of Same Audio Clip

data, and the transcripts may edit out repetitions or verbal stumbles. Artifacts of this can be seen by comparing transcripts, with differences highlighted in yellow (Fig. 2).

Evaluating Speech Models' Performance on CU Dataset (Experiment 2)

The recordings in the Ghana ORF Dataset differ from many other children-specific ASR datasets in that the recordings were made in classrooms, most students spoke English as a second (or third) language, and the local accent is distinct from many North American accents. Prior research has shown that older ASR models transcribing adult speech demonstrated weaker performance when transcribing audio of people whose accents are regional, non-standard, or simply not well represented in the training corpora (Koenecke et al., 2020). As the student reading in the Ghana ORF had regional accents we wanted to investigate whether these models had an easier time handling North American students' ORF recordings.To do so, we used the CU Kids Read and Summarized Story Corpus, which contains speech recordings and transcripts from American children completing ORF and reading comprehension assessments. The participants were English-speaking elementary students in grades 3–5, approximately 8–12 years old. For the purposes of this study, we focused on the audio of students reading stories aloud Fig. 3.

To better understand how transcription accuracy was affected by factors such as age, ability, and accent, we evaluated Whisper V2 on transcribed recordings from the CU Dataset. We selected 45 audio recordings, representing a mix of age and ability levels, which already had high-quality human-generated transcripts. The WER of the model-generated transcript was calculated relative to the human transcript, which we considered the ground truth. These results were then compared to the Ghana ORF files from Experiment 1. This comparison allowed us to explore whether Whisper V2 might potentially perform better with American accents than Ghanaian ones.

As shown above in Table 4, the model exhibited very similar performance between the two datasets, with a slightly lower WER on the CU Dataset compared to the Ghana ORF Dataset. Unlike previous studies that reported large gaps

CU Kids Read and Summarized Story Corpus	
Text Being Read	Human Transcript of Audio (Reading errors marked inyellow)
Tim liked to go outside on sunny days. He liked to feel the cool breeze. He liked to hear the wind. He liked to see the white, puffy clouds. Today clouds were racing across the sky. Clouds never stay in one place. Tim wondered where the clouds go. "White puffy clouds come on sunny days," Tim said. "What else do I know about clouds?" Tim thought about last Monday. Monday was a foggy day. "What is fog?" Tim asked his Mom. "Fog is a layer of clouds near the ground," Mom said. "Where does the fog go?"Tim asked.	Tim liked to go outside on sunny days. He liked to feel the the cool breeze. He liked to hear the wi wind. He liked to see the white, fluffy clouds. Today clouds were racing across the sky. Clouds never stay in one place. Tim wondered where the clouds go. "White fluff y clouds come on sunny days," Tim said. "What else do I know about clouds?" Tim thought about last Monday. Monday was a foggy day. "What is fog?" Tim asked his Mom. "Fog is a layer of clouds near the ground," Mom said. "Where does the fog go?" Tim asked.
99 Total in original text	4 Student reading errors
Audio file of student reading	

Fig. 3 Example of ORF Passage and Scoring

Table 4 Whisper V2 WER by Dataset

	Ghana ORF Dataset	CU Dataset
Average WER	10.3	10.2
Range—Max | Min	25 | 1	31.0 | 0.0
Standard Deviation	5.5	6.1

in WER—for example, Koenecke et al. (2020) reported a gap in WER of over 20 between Black and White speakers—in this case, we found only a small difference. This notable improvement in accuracy may be attributed to the larger, more varied, multilingual datasets on which the newest generation of speech models were trained. However, given the small size of the dataset and the importance of addressing bias and equity in speech models, we consider these results provisional and believe that this area can benefit from continued research.

Automatically Calculating WCPM From a Transcript (Experiment 3)

Experiments 1 and 2 demonstrate the potential for using ASR in assessing ORF. However, strong WER performance does not always translate directly to improved performance on downstream tasks. To better understand whether open-source ASR models can effectively support ORF assessment, we conducted a final experiment with two primary objectives. First, we tested the transcription performance of

Table 5 Human vs Model Generated Estimates of Student Reading Errors

	Expert Human	**Fully Automated**
WER—**Average** | Max | Min | StDev	**0.14** | 0.0 | 0.61 | 0.11	**0.12** | 0.0 | 0.77 | 0.11
WER—Average Difference Per Item	0.05	
Correlation Coefficient	0.79	

Whisper V2 on a slightly larger subset of the Ghana ORF Dataset. This was the original set of 60 audio file used in experiments 1 and 2 and an additional 95 recording, for a total of 165 ORF assessment. This additional 95 included recordings with more background noise and weaker readers, we allowed us to evaluate the model's robustness in more challenging, "real-world" scenarios.

To fully automate the scoring of WCPM we first used WhisperV2 to transcribe student audio files, and then JiWER Python package to calculate the WER between the original story text and the transcript the original story text. We compare WER calculation for both the model transcript and the human transcript, relative to the original text. In this case, the Word Error Rate of the expert human transcripts shouldn't be seen as transcription errors, but instead as accurately capturing student reading mistakes. Hence, for this question we are looking to see how similar the respective WER are between the human transcripts and the model transcripts. Table 5 presents the aggregate results, including the average, minimum value, maximum value, and standard deviation of the WER for Whisper V2 compared to human values.

The overall error rates are very similar, indicating they produce similar transcriptions, and the difference on a per-file basis is small. Additionally, the WER results between human and model transcriptions are strongly correlated. We take these results as an indication that Whisper V2 is producing broadly similar but not identical transcriptions to that of a human rater. However, it is possible that even this small discrepancy could impact the estimate of ORF. Hence, the next step was to compare the human and model-generated WCPM scores. To calculate and automatically generated WCPM scores we divided the number of correct words in the model-generated transcript by the length of the audio files (the time spent reading in minutes) to derive a WCPM score.

$$WCPM = (totalwordsinstory * (1 - WER)) \times (60/durationofaudioclipinseconds)$$

To calculate human generated WCPM scores, we used took the high quality human transcriptions of these audio files and counted the student's reading errors according to the standard WER protocol (insertions, deletions, and substitutions). Then the number of errors was subtracted from the total story length and was then divided by the duration of the audio clip in seconds and multiplied by 60.

$$Human\ WCPM = (total\ words\ in\ story - (errors)) \times (60/duration\ of\ audio\ clip\ in\ seconds)$$

As a quality control measure a second author, conducted the same process independently on a subset of the transcriptions, but we found no systematic

Table 6 Human vs Model Generated Estimates of WCPM

	Expert Human	**Fully Automated**
WER—**Average** | Max | Min | StDev	**91.7** | 0.0 | 101 | 14.8	**93.5** | 9.4 | 170 | 31.6
WCPM—Average Difference	4.82	
Correlation Coefficient	0.98	

differences in the estimates. Because the other values in WCPM (length of story and time spent reading) are objective, any discrepancies between expert human raters (our gold standard measure) and the model estimates will be attributable to either the transcription itself or the error calculation rate. This allowed us to compare the automatically generated scores to those made by expert human raters.

Table 6 reports the agreement between the automated and human-generated scores, we can better understand the feasibility of employing open-source ASR models in ORF assessment. The results of the experiment demonstrate that the performance of the fully automated approach using Whisper V2 and the expert human raters was very similar in assessing ORF. On average, the human raters were slightly stricter, identifying more errors, resulting in a higher WER and, consequently, a lower WCPM rate. However, the magnitude of the discrepancy—5 WCPM – is considered small in the context of ORF tests. Further, automated scores and the expert human scores were highly correlated (0.98) high, suggesting that the same students were receiving similar scores relative to one another, which is arguably the most the most important measure in terms of practical implications.

Interestingly, the WCPM estimates were more highly correlated than were the raw transcripts themselves. One likely explanation for this is because the WCPM measure considers not only reading accuracy but also reading rate, transcription errors introduced by the ASR will be a relatively smaller portion of the overall score.

Taken as a whole, we consider the results are very encouraging and suggest that it may be feasible to use open-source ASR models to assess ORF. While the automated measure of ORF does not match the expert generated score exactly, WCPM is a directional and approximate measure of reading ability, and one that typically fluctuates even when conducted with the same student, meaning that small measurement fluctuations are unlikely to hurt the practical use of the results. Additionally, ORF is almost always used as a formative or diagnostic measure of reading ability rather than a summative one, so there is less chance that small measurement errors will negatively impact students. In sum, while there are numerous open questions, these find these results are encouraging, particularly because of the dataset on which it was tested. However, given the strong theoretical and empirical basis for considering both rate and accuracy in ORF assessment, further research should investigate the specific nature of discrepancies between human and model transcripts. Overall, these results are especially encouraging considering the diverse student population represented in the dataset.

Analyzing Potential Bias

Bias and fairness in AI models, particularly in educational contexts, have been extensively studied in the literature (see Baker & Hawn 2022 for a comprehensive review). However, the definitions of fairness and bias in AI research remain complex and often contested. In this study, we operationalize bias as a situation where the model exhibits a higher error rate when transcribing speech of students from specific socio-economic or demographic groups compared to human raters. For instance, if the model consistently produced transcripts with higher WER for female students than for male students, relative to human rater scores, this might indicate a form of bias.

The underlying reasons for such bias in ASR models are multifaceted and subject to debate. A prevalent explanation suggests that due to the composition of training corpora, these models may perform better on audio from certain groups, such as native English speakers, compared to non-native speakers. It is crucial to emphasize that our evaluation focuses not on the total error estimate, which may inherently differ between native and non-native speakers, but rather on the discrepancy in reading error estimates between the model and human raters. While it is possible that such discrepancies could reflect implicit biases held by human raters, we consider this less likely given the relatively objective nature of the task (identifying oral reading errors from audio recordings) and our rigorous coding procedures, which involved listening to each clip multiple time. Nevertheless, a thorough examination of potential rater bias is a substantial question in its own right and falls outside the scope of this study.

In our analysis, we examine the differences between model estimates and human estimates of both WER and Words Per Minute (CWPM) across two salient demographic groups: gender and native language status. In both scenarios, the difference between group averages of model estimates and human estimates is substantially less than one standard deviation. This indicates that we cannot reject the null hypothesis, which posits no systematic difference in the model's performance across these groups. Consequently, we found no statistically significant evidence that the model performs worse (i.e., produces less accurate transcriptions) based on a student's gender or home language. While this is encouraging, given the relatively small size of the as well the ethical implications of this question, we treat our results with caution, and believe that these types of checks should be conducted regularly Table 7.

Table 7 Average Difference Between Model Score and Human Score by Demographic Group

	WER	WCPM
Male | Female | StDev	0.05 | 0.03 | 0.05	5.9 | 4.2 | 4.0
English |Other | Diff | StDev	0.05 | 0.06 | 0.05	5.6 | 4.7 | 4.0

Discussion

This paper presents a new dataset of the ORF of students in Ghana and explores the feasibility of using open-source speech recognition models to automatically assess ORF, through a set of three experiments utilizing large-scale speech models. We found that Whisper V2, without additional fine-tuning, achieved impressive accuracy in transcribing Ghanaian students' readings, with a low WER of 10.3. Notably, the automated ORF scores derived from these transcriptions showed a strong correlation (0.98) with expert human raters' scores, demonstrating the model's effectiveness on real-world classroom data with regional accents.

Implications

ORF is a crucial component of reading assessment, particularly for beginning and intermediate readers. However, traditional ORF assessments can be time-consuming and resource-intensive, especially in educational settings with large student populations, limited staff, and competing priorities. Automating ORF assessment using technology, such as ASR, could enable more frequent and widespread monitoring of student progress, allowing educators to identify and address reading difficulties more quickly and effectively. This study's findings have several important implications.

First, top-performing open-source models can now transcribe children's oral reading with a relatively high level of accuracy, which had been a major challenge until recently. The Whisper V2 model achieved a WER of 10.3 on the Ghana ORF Dataset without any fine-tuning, a level of accuracy comparable to state-of-the-art adult speech recognition from just a few years ago. Second, the models handle Ghanaian accents equally well as American accents, without any additional fine-tuning. This is a notable finding, as previous research had shown ASR models to be more accurate with overrepresented accents. The study found very similar performance between the Ghana ORF Dataset and the American CU Dataset, with average WERs of 10.3 and 10.2 respectively. This contrasts with earlier studies that reported significant disparities in ASR accuracy based on accent. While we consider these finding provisional, they do suggest that newer ASR models may be more able to handle diverse accents and points to their potential use in multilingual and multicultural educational contexts. Third, when used to calculate WCPM, a widely used measure of ORF that combines both reading accuracy and reading rate, the models produced results extremely similar those of expert human raters. The study found a high correlation (0.98) between automated WCPM scores and those generated by expert human raters.

Given its high performance, accessibility, and lack of need for complex training, this approach using large-scale speech models for ORF assessment shows promise for implementation and scaling in diverse, lower-resource educational contexts. The combination of accurate transcription, ability to handle diverse accents, and production of reliable ORF scores suggests that this technology could significantly enhance literacy assessment practices, particularly in settings where regular, individualized assessment has been challenging due to resource constraints.

Limitations

While this study provides valuable insights into the potential of ASR models for the automated assessment of ORF, it is not without its limitations. Firstly, the study was conducted using only two specific ASR models, Whisper V2 and wav-2vec 2.0. While these models are considered state-of-the-art, the findings may not be generalizable to other ASR models. Secondly, the study used a relatively small sample size from two specific datasets, one from Ghana and one from the United States. Finally, these experiments only examine the potential for automating ORF assessment in English.

Further Research

One key direction is to develop and curate large, diverse datasets that represent the wide range of accents, languages, and educational contexts found in the Global South. By training and testing speech models on these datasets, we can enhance their robustness and generalizability, ensuring that they can accurately assess ORF across various populations. One critical issue is the need to address potential model biases, particularly those related to underrepresented accents and languages. Future research should focus on developing bias detection and mitigation strategies, such as data augmentation techniques or adversarial training approaches, to ensure that automated assessments are fair and equitable across diverse student populations. We hope that by making this dataset available for research purposes, other researchers can contribute to this important topic. Additionally, it is crucial to understand the practical implications and challenges of implementing these technologies in resource-limited contexts. This includes examining the technical infrastructure and support required to deploy and maintain automated ORF assessment systems, as well as the cost-effectiveness and sustainability of these solutions.

Conclusion

The primary objective of this study was to explore the feasibility and efficacy of employing ASR models for the automated assessment of ORF. The study was particularly focused on addressing two central challenges associated with the automation of ORF assessment in LMICs: the quality of transcription and the conversion of transcription into WCPM. The experimental design of the study utilized two state-of-the-art ASR models, namely Whisper V2 and wav2vec 2.0, to transcribe audio files of students' oral reading from two distinct datasets: a novel dataset from Ghana and the CU Datasets and Summarized Story Corpus from the United States. The transcriptions generated by these models were subsequently used to calculate a WCPM score for each student, which was then compared to the score derived from human generated transcripts.

The key findings of the research are multifaceted and provide valuable insights into the potential of ASR models in the context of ORF assessment. The results suggest that ASR models have the potential to accurately transcribe student oral reading and calculate WCPM scores, thereby enhancing the efficiency of ORF assessment and reducing the reliance on human raters. This could be particularly beneficial in LMICs, where the systematic and regular assessment of literacy skills can be challenging due to factors such as large rural populations, unreliable infrastructure, and limited organizational capacity.

In the broader context of formative assessment, the findings of this research suggest that ASR models could be employed to assist in the assessment of student work across a variety of contexts. The demonstrated ability of these models to accurately transcribe student speech and calculate performance metrics could be applied to other areas of assessment, such as oral presentations or language proficiency tests. However, the research also underscores the need for further investigation to address the challenges associated with model bias against underrepresented accents and the technical complexity of fine-tuning these models. The findings of this study, therefore, not only contribute to the existing body of knowledge on the application of ASR models in educational settings but also highlight potential avenues for future research in this area.

Declarations

Competing Interests The first author has an ongoing research partnership with Rising Academies and works as a consultant on a project related to developing a conversational agent to support students' math skills. This study is an extension of his doctoral dissertation and predates the consulting relationship. The second author participated in the marking and analysis of transcripts while still a student at University of Oxford, however in over the course of the past year, she has joined Rising Academies as Research and Assessment Manager.

References

Adams, M. (2011). *Technology for developing children's language and literacy—Bringing speech recognition to the classroom*. The Joan Ganz Cooney Center at Sesame Workshop.

Ali, A., & Renals, S. (2018). Word Error Rate Estimation for Speech Recognition: E-WER. *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics* (Volume 2: Short Papers), 20–24. https://doi.org/10.18653/v1/P18-2004

Baevski, A., Zhou, H., Mohamed, A., & Auli, M. (2020). wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations (arXiv:2006.11477). arXiv. https://doi.org/10.48550/arXiv.2006.11477

Baker, R. S., & Hawn, A. (2022). Algorithmic bias in education. *International Journal of Artificial Intelligence in Education, 32*(4), 4. https://doi.org/10.1007/s40593-021-00285-9

Black, P., & Wiliam, D. (2009). Developing the theory of formative assessment. *Educational Assessment, Evaluation and Accountability, 21*(1), 5–31. https://doi.org/10.1007/s11092-008-9068-5

Bolaños, D., Cole, R. A., Ward, W. H., Tindal, G. A., Hasbrouck, J., & Schwanenflugel, P. J. (2013a). Human and automated assessment of oral reading fluency. *Journal of Educational Psychology, 105*(4), 1142–1151. https://doi.org/10.1037/a0031479

Bolaños, D., Cole, R. A., Ward, W. H., Tindal, G. A., Schwanenflugel, P. J., & Kuhn, M. R. (2013b). Automatic assessment of expressive oral reading. *Speech Communication, 55*(2), 221–236. https://doi.org/10.1016/j.specom.2012.08.002

Chan, W., Park, D., Lee, C., Zhang, Y., Le, Q., & Norouzi, M. (2021). SpeechStew: Simply Mix All Available Speech Recognition Data to Train One Large Neural Network (arXiv:2104.02133). arXiv. http://arxiv.org/abs/2104.02133

Cincarek, T., Gruhn, R., Hacker, C., Nöth, E., & Nakamura, S. (2009). Automatic pronunciation scoring of words and sentences independent from the non-native's first language. *Computer Speech & Language, 23*(1), Article 1. https://doi.org/10.1016/j.csl.2008.03.001

Dubeck, M. M., & Gove, A. (2015). The early grade reading assessment (EGRA): Its theoretical foundation, purpose, and limitations. *International Journal of Educational Development, 40*, 315–322. https://doi.org/10.1016/j.ijedudev.2014.11.004

Duong, M., Mostow, J., & Sitaram, S. (2011). Two methods for assessing oral reading prosody. *ACM Transactions on Speech and Language Processing, 7*(4), 1–22. https://doi.org/10.1145/1998384.1998388

Fan, R., Zhu, Y., Wang, J., & Alwan, A. (2022). Towards Better Domain Adaptation for Self-Supervised Models: A Case Study of Child ASR. *IEEE Journal of Selected Topics in Signal Processing, 16*(6), 1242–1252. https://doi.org/10.1109/JSTSP.2022.3200910

Feng, S., Halpern, B. M., Kudina, O., & Scharenborg, O. (2024). Towards inclusive automatic speech recognition. *Computer Speech & Language, 84*, 101567. https://doi.org/10.1016/j.csl.2023.101567

Fuchs, L. S., Fuchs, D., Hosp, M. K., & Jenkins, J. R. (2001). Oral Reading Fluency as an Indicator of Reading Competence: A Theoretical, Empirical, and Historical Analysis. *Scientific Studies of Reading, 5*(3), 239–256. https://doi.org/10.1207/S1532799XSSR0503_3

Geirhos, R., Jacobsen, J.-H., Michaelis, C., Zemel, R., Brendel, W., Bethge, M., & Wichmann, F. A. (2020). Shortcut learning in deep neural networks. *Nature Machine Intelligence, 2*(11), 665–673. https://doi.org/10.1038/s42256-020-00257-z

Gerosa, M., Giuliani, D., Narayanan, S., & Potamianos, A. (2009). A review of ASR technologies for children's speech. Proceedings of the 2nd Workshop on Child, Computer and Interaction - WOCCI '09, 1–8. https://doi.org/10.1145/1640377.1640384

Gray, G., McGuinness, C., & Owende, P. (2014). *An application of classification models to predict learner progression in tertiary education* (pp. 549–554). IEEE. https://doi.org/10.1109/IAdCC.2014.6779384

Hasbrouck, J., & Tindal, G. A. (2006). Oral Reading Fluency Norms: A Valuable Assessment Tool for Reading Teachers. *The Reading Teacher, 59*(7), 636–644. https://doi.org/10.1598/RT.59.7.3

Jain, R., Barcovschi, A., Yiwere, M., Corcoran, P., & Cucu, H. (2023a). Adaptation of Whisper models to child speech recognition. https://doi.org/10.48550/ARXIV.2307.13008

Jain, R., Barcovschi, A., Yiwere, M. Y., Bigioi, D., Corcoran, P., & Cucu, H. (2023b). A WAV2VEC2-Based Experimental Study on Self-Supervised Learning Methods to Improve Child Speech Recognition. *IEEE Access, 11*, 46938–46948. https://doi.org/10.1109/ACCESS.2023.3275106

Jiang, W., & Pardos, Z. A. (2021). Towards equity and algorithmic fairness in student grade prediction. In *Proceedings of the 2021 AAAI/ACM conference on AI, ethics, and society* (pp. 608–617). https://doi.org/10.1145/3461702.3462623

JiWER. (n.d.). [Computer software]. https://jitsi.github.io/jiwer/

Kizilcec, R. F., & Lee, H. (2021). Algorithmic fairness in education. *arXiv*, arXiv:2007.05443. https://doi.org/10.48550/arXiv.2007.05443

Koenecke, A., Nam, A., Lake, E., Nudell, J., Quartey, M., Mengesha, Z., Toups, C., Rickford, J. R., Jurafsky, D., & Goel, S. (2020). Racial disparities in automated speech recognition. *Proceedings of the National Academy of Sciences, 117*(14), 7684–7689. https://doi.org/10.1073/pnas.1915768117

LaBerge, D., & Samuels, J. (1974). Toward a Theory of Automatic Information Processing in Reading. *Cognitive Psychology, 6*.

Loukina, A., Zechner, K., & Chen, L. (2014). Automatic evaluation of spoken summaries: The case of language assessment. In *Proceedings of the ninth workshop on innovative use of NLP for building educational applications* (pp. 68–78). https://doi.org/10.3115/v1/W14-1809

Magliano, J. P., & Graesser, A. C. (2012). Computer-based assessment of student-constructed responses. *Behavior Research Methods, 44*(3), 608–621. https://doi.org/10.3758/s13428-012-0211-3

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A survey on bias and fairness in machine learning. *ACM Computing Surveys, 54*, 1–35. https://doi.org/10.1145/3457607

Methods and procedures in PIRLS 2016. (2018). TIMSS & PIRLS.

Mitchell, S., Potash, E., Barocas, S., D'Amour, A., & Lum, K. (2021). Algorithmic fairness: choices, assumptions, and definitions. *Annual Review of Statistics and its Application, 8*(1), 141–163. https://doi.org/10.1146/annurev-statistics-042720-125902

Nag, S. (2017). Assessment of Literacy and Foundational Learning in Developing Countries. HEART (Health & Education Advice & Resource Team).

National Reading Panel. (1998). Teaching children to read: Report of the Subgroups. National Reading Panel.

Nouza, J., Mateju, L., Cerva, P., & Zdansky, J. (2023). Developing State-of-the-Art End-to-End ASR for Norwegian. In K. Ekštein, F. Pártl, & M. Konopík (Eds.), Text, Speech, and Dialogue (Vol. 14102, pp. 200–213). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-40498-6_18

Panayotov, V., Chen, G., Povey, D., & Khudanpur, S. (2015). Librispeech: An ASR corpus based on public domain audio books. 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 5206–5210. https://doi.org/10.1109/ICASSP.2015.7178964

Perfetti, C., & Stafura, J. (2014). Word Knowledge in a Theory of Reading Comprehension. *Scientific Studies of Reading, 18*(1), 22–37. https://doi.org/10.1080/10888438.2013.827687

Pritchett, L. (2013). The Rebirth of Education: Schooling Ain't Learning. Center for Global Development.

Rasinski, T., Rikli, A., & Johnston, S. (2009). Reading Fluency: More Than Automaticity? More Than a Concern for the Primary Grades? *Literacy Research and Instruction, 48*(4), 350–361. https://doi.org/10.1080/19388070802468715

Roehrig, A. D., Petscher, Y., Nettles, S. M., Hudson, R. F., & Torgesen, J. K. (2008). Accuracy of the DIBELS Oral Reading Fluency Measure for Predicting Third Grade Reading Comprehension Outcomes. *Journal of School Psychology, 46*(3), 343–366. https://doi.org/10.1016/j.jsp.2007.06.006

Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., & Sutskever, I. (2022). Robust Speech Recognition via Large-Scale Weak Supervision (arXiv:2212.04356). arXiv. https://doi.org/10.48550/arXiv.2212.04356

Spaull, N., Pretorius, E., & Mohohlwane, N. (2020). Investigating the comprehension iceberg: Developing empirical benchmarks for early-grade reading in agglutinating African languages. South African Journal of Childhood Education, 10(1). https://doi.org/10.4102/sajce.v10i1.773

Spear-Swerling, L. (2006). Children's Reading Comprehension and Oral Reading Fluency in Easy Text. *Reading and Writing, 19*(2), 199–220. https://doi.org/10.1007/s11145-005-4114-x

Suresh, H., & Guttag, J. V. (2021). A framework for understanding sources of harm throughout the machine learning life cycle. *Equity and Access in algorithms, mechanisms, and optimization* (pp. 1–9). https://doi.org/10.1145/3465416.3483305

Stiennon, N., Ouyang, L., Wu, J., Ziegler, D. M., Lowe, R., Voss, C., Radford, A., Amodei, D., & Christiano, P. (2022). Learning to summarize from human feedback (arXiv:2009.01325). arXiv. https://doi.org/10.48550/arXiv.2009.01325

Tempelaar, D., Rienties, B., & Nguyen, Q. (2020). Subjective data, objective data and the role of bias in predictive modelling: Lessons from a dispositional learning analytics application. *PLoS ONE, 15*(6). https://doi.org/10.1371/journal.pone.0233977

Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., Yogatama, D., Bosma, M., Zhou, D., Metzler, D., Chi, E. H., Hashimoto, T., Vinyals, O., Liang, P., Dean, J., & Fedus, W. (2022). Emergent Abilities of Large Language Models (arXiv:2206.07682). arXiv. https://doi.org/10.48550/arXiv.2206.07682

Zhang, Y., Park, D. S., Han, W., Qin, J., Gulati, A., Shor, J., Jansen, A., Xu, Y., Huang, Y., Wang, S., Zhou, Z., Li, B., Ma, M., Chan, W., Yu, J., Wang, Y., Cao, L., Sim, K. C., Ramabhadran, B., … Wu, Y. (2022). BigSSL: Exploring the Frontier of Large-Scale Semi-Supervised Learning for Automatic Speech Recognition. IEEE Journal of Selected Topics in Signal Processing, 16(6), 1519–1532. https://doi.org/10.1109/JSTSP.2022.3182537

Authors and Affiliations

Owen Henkel[1] · Hannah Horne-Robinson[2] · Libby Hills[3] · Bill Roberts[4] · Josh McGrane[5]

✉ Owen Henkel
owen.henkel@education.ox.ac.uk

Hannah Horne-Robinson
hannah.horne-robinson@risingacademies.com

Libby Hills
libby.hills@jacobsfoundation.org

Bill Roberts
bill@legiblelabs.com

Josh McGrane
joshua.mcgrane@unimelb.edu.au

1 University of Oxford, Oxford, UK

2 Rising Academies, Accra, Ghana

3 Jacobs Foundation, Zurich, Switzerland

4 Legible Labs, New York, USA

5 University Melbourne, Parkville, Australia