Using Large Language Models and Speech-to-Text Models to Facilitate the Assessment of Basic Literacy in Ghana

Owen Henkel

St Antony's College, University of Oxford

Submitted to the University of Oxford in partial fulfilment of the requirements for the Degree of Doctor of Philosophy in Education

2025

Supervisors:

Associate Professor Joshua McGrane

Associate Professor Scott Hale

Table of Contents

ABSTRACT **6**

CHAPTER 1 – INTRODUCTION **9**

1.1 Outline of Thesis 11

CHAPTER 2 - LITERATURE REVIEW **13**

2.1 Components of Reading Ability 13
2.2 Assessing Reading Ability 19
2.3 Recent Advances in NLP and ASR 25
2.4 Using NLP and ASR to Assess Reading Ability 31
2.5 Conclusion 37

CHAPTER 3 – METHODOLOGY **39**

3.1 Measuring Accuracy and Agreement 39
3.3 Collection of Novel Datasets 42
3.3 ROARS Dataset 44
3.4 Ghanaian Student Oral Reading Fluency Dataset 46
3.5 CU Dataset 47
3.6 Disclosure Statement 48

CHAPTER 4 – CAN SPEECH MODELS ASSESS ORAL READING FLUENCY? **49**

4.1 Introduction 49
4.2 Prior Work 50
4.3 Current Study 53
4.4 Experiment 1 - Evaluating Speech Models' Performance on the Ghana ORF Dataset 57
4.5 Experiment 2 - Evaluating Speech Models' Performance on CU Dataset 60
4.6 Experiment 3 - Automatically Calculating WCPM From a Transcript 61
4.7 Analysing Potential Bias 64
4.8 Discussion 65
4.9 Conclusion 67

CHAPTER 5 - CAN LLMS GRADE SHORT ANSWER READING COMPREHENSION QUESTIONS? **69**

5.1 Introduction 69
5.2 Prior Work 71
5.3 Current Study 76
5.4 Results 80
5.5 Discussion 86
5.6 Conclusion 88

CHAPTER 6 - CAN LLMS GRADE STORY RETELL **90**

6.1 Introduction 90
6.2 Prior Work 91
6.3 Dataset and Human Grading 94

6.4 Experiment 1 - Inter-Rater Agreement on Story Retell 96
6.5 Experiment 2 – Convergent Validity of Story Retell 98
6.6 Experiment 3 - Evaluating LLMs' Ability to Score Student Story Retells 102
6.7 Discussion 104
6.8 Conclusion 106

CHAPTER 7 - DISCUSSION 108

7.1 Summary of Findings 108
7.2 Implications 109
7.3 Limitations 111
7.4 Directions for Further Research 115
7.5 Conclusions 117

CHAPTER 8 - REFERENCES 119

APPENDICES 144

Appendix A: Oral Reading Fluency Passages 144
Appendix B: Reading Comprehension Passages 146
Appendix C: CUREC Application 152
Appendix D: Abstract of Published Article 1 168
Appendix E: Abstract of Published Article 2 169

List of Tables

Table 3.1 Key Characteristics of ROARS dataset 44

Table 4.1 Performance of ASR Model on Child-Specific Datasets Measured with WER 52

Table 4.2 WER by model on Ghana ORF Dataset 58

Table 4.3 Whisper v2 WER by Dataset 61

Table 4.4 Human vs Model Generated Estimates of Student Reading Errors 62

Table 4.5 Human vs Model Generated Estimates of WCPM 63

Table 4.6 Average Difference Between Model Score and Human Score by Demographic Group 64

Table 5.1 Summary of Publicly Available Short-Answer Datasets 75

Table 5.2 Performance of Four Varieties of GPT Models on two-class Task 81

Table 5.3 Performance of Four Varieties of GPT Models on three-class Task 82

Table 5.4 Performance Across Different Approaches to Scoring Responses 83

Table 5.5 Misclassification Rates by Demographic Group 85

Table 6.1 Comparison of Average Score by Rubric and Rater 97

Table 6.2 Comparison of Interrater Agreement by Rubric and Measure 98

Table 6.3 Correlation Between Different Measures of Story Retell 100

Table 6.4 Correlation Between Measures of Retell, Oral Reading Fluency and Comprehension 101

Table 6.5 Performance of baseline prompt Using GPT4o-2024-05-13 103

Table 6.6 Model Human Agreement compared to Human-to-Human Agreement 104

List of Figures

Figure 2.1 Similarities Across Influential Stage-Based Models of Reading Development 15

Figure 2.2 Processes Comprising Reading Comprehension 18

Figure 2.3 Definitions for Different Assessment Tasks Related to Whole Story Comprehension 22

Figure 3.1 Assessment Types and Scoring Procedures 43

Figure 3.2 Example of Short Answer Scoring 45

Figure 3.3 Example of Retell Scoring 46

Figure 3.4 Example of Passage and Human Scoring 47

Figure 3.5 Example of Human Transcriptions 48

Figure 4.1 Example of Passage and Human Scoring 58

Figure 4.2 Alternate Transcripts of Same Audio Clip 59

Figure 5.1 Examples of Student Answers and Ratings 77

Figure 5.2 System Prompts Used in Experiment 79

Figure 6.1 Original Rubrics 94

Figure 6.2 Adapted Rubrics 95

Figure 6.3 Examples of different rater's scoring by rubric 96

Abstract

This dissertation examines how recent advances in artificial intelligence, particularly in Natural Language Processing (NLP) and Automatic Speech Recognition (ASR), could enhance the assessment of reading ability in English in West Africa. Although regular diagnostic and formative assessment has proven essential for effective reading instruction, the substantial time and expertise needed to conduct and evaluate reading assessments often restricts their routine implementation, especially in resource-constrained education systems. Recent innovations in Natural Language Processing (NLP) and Automatic Speech Recognition (ASR) offer promising new approaches for making reading assessment more accessible and scalable in low-resource environments.

To investigate this potential, this research develops and analyzes two novel datasets gathered in partnership with Rising Academies, a network of schools in Ghana. The primary dataset encompasses results from various reading assessment tasks administered to 162 students between the ages of 9 and 18, including audio recordings of oral reading, written and spoken responses to comprehension questions, and story retells. The secondary dataset contains high-quality human transcriptions of these same students' performance on the oral reading fluency tasks, enabling comprehensive analysis of reading errors and fluency patterns. All assessment materials were selected from released items from the 2016 PrePIRLS assessment, which was specifically designed for use in low and middle-income countries. Using these datasets, I then conducted three interconnected empirical studies.

The first study evaluates the viability of using state-of-the-art ASR models to assess oral reading fluency. Using the Ghanaian student dataset, the research determined that the Whisper v2 model could transcribe student reading with high accuracy (Word Error Rate of 10.3%) without any fine-tuning or adaptation. Significantly, the model's performance remained consistent across different student populations, demonstrating similar accuracy when transcribing recordings from both Ghanaian and American students. When these transcriptions were used to calculate Words Correct Per Minute scores, they demonstrated strong agreement with expert human raters (correlation of 0.98).

The second study explores the capacity of Large Language Models to evaluate short answer reading comprehension questions. Drawing from a dataset of over 1,000 student responses, the research established that GPT-4 could achieve accuracy levels matching those of expert human raters. With minimal prompt engineering, the model attained a Quadratic Weighted kappa of 0.91 in the three-class condition and a Linear Weighted kappa of 0.87 in the two-class condition, surpassing existing benchmarks in automated short answer grading.

The third study assesses the effectiveness of LLMs in grading story retell tasks while simultaneously examining the reliability and validity of various retell measures. The research documented strong inter-rater reliability among human raters using both three-class and five-class rubrics (Kendall's W of 0.81 and 0.85 respectively). It also revealed moderate correlations between different measures of retell and other indicators of reading ability, suggesting that retell tasks capture unique aspects of reading comprehension. In evaluating LLM performance, the study found that GPT-4 could effectively replicate human scoring, achieving agreement levels (QWK of 0.82) that nearly matched human-to-human agreement (0.85).

Several themes emerged across these studies. First, the latest generation of AI models exhibited remarkably robust performance across different assessment types without requiring task-specific fine-tuning or extensive prompt engineering. This marks a substantial improvement over previous approaches that demanded considerable technical expertise and task-specific training data. Second, throughout all three studies, the automated approaches reached levels of agreement with human raters that paralleled inter-rater agreement among experts, indicating these technologies could serve as reliable tools for diagnostic and formative assessment purposes. Third, the studies consistently showed minimal evidence of systematic bias in model performance based on student demographics, though this finding warrants careful interpretation given the relatively limited sample sizes.

The research underscores the crucial role of assessment design in successful automation. Specifically, more detailed rubrics that identify meaningful distinctions in student performance appear particularly well-suited to automated scoring. Furthermore, the studies illuminate how different assessment formats may capture distinct aspects of reading ability, reinforcing the

importance of employing multiple assessment types to construct a comprehensive picture of student reading proficiency.

These findings contribute to an understanding of how AI can support literacy assessment across diverse educational settings. While previous studies have shown AI's potential for educational assessment in well-resourced environments, this dissertation provides pioneering empirical evidence for the feasibility of applying these technologies to support reading assessment. The consistently strong performance across various assessment types and student populations suggests that recent AI advances could facilitate more regular and comprehensive evaluation of reading ability, particularly in contexts where traditional assessment methods face resource constraints. By demonstrating the viability of automated scoring across multiple assessment types while identifying areas requiring further study, this research enhances our understanding of how artificial intelligence can strengthen literacy assessment in diverse educational contexts.

Chapter 1 – Introduction

Effective reading instruction depends on several key components, with regular, high-quality assessment of basic reading skills being crucial because it allows educators to identify students' abilities, monitor their progress, and tailor their instructional approaches accordingly. Classroom-based reading assessment falls into three categories: diagnostic assessment identifies students' existing knowledge before instruction begins (Bejar, 1984), formative assessment provides ongoing feedback during learning (Black & Wiliam, 2009), and summative assessment evaluates learning outcomes against standards at the conclusion of an instructional period (Bennett, 2011). Ideally, these three forms of assessment work together in a complementary fashion to provide comprehensive insights into student learning. In practice, however, diagnostic and formative assessments—though critical for effective teaching—are frequently neglected or conducted infrequently due to time constraints (Bennett, 2011; Landauer et al. 2009). When these assessments are conducted, the time-intensive nature of traditional literacy assessment has resulted in an overreliance on multiple-choice questions, which often create a disconnect between assessment tasks and real-world reading expectations. This disconnect has led to compelling arguments for using open-ended formats which can evaluate dimensions of reading comprehension that closed-response questions cannot capture (Pearson & Hamm, 2006). Advocates for open-ended formats—including short answer, story retell, and summarization—highlight their greater face validity, reduced influence of test-taking strategies, lower risk of floor effects, and better alignment with specific comprehension processes.

These benefits are particularly valuable in resource-constrained education systems, such as those in Anglophone Sub-Saharan Africa (e.g., Ghana) where this dissertation's empirical work is centered. While diagnostic and formative assessment are critically needed in these contexts, they are rarely implemented effectively due to resource constraints. Widely recognized assessments such as Early Grade Reading Assessment (EGRA), Annual Status of Education Reports (ASER), and Uwezo have been developed to address design issues of typical multiple-choice tests. However, these assessments are still mainly conducted to evaluate the progress of an entire country's education system and reach only a small fraction of students in each country. While useful for policy development and monitoring education systems, these assessments

cannot inform classroom practice due to the substantial time, cost, and complexity of administration and grading (Wagner et al., 2012; Gove & Cvelich, 2011).

Recent advances in Natural Language Processing (NLP) and Automatic Speech Recognition (ASR) technologies offer promising opportunities for automating diagnostic and formative reading assessments, particularly for open-ended assessment formats. These technologies could potentially automate the scoring of open-ended reading tasks—including oral reading fluency, short answer responses, and story retell—making frequent, high-quality assessment feasible even in resource-constrained environments. While these technologies might have implications for all assessment types, this dissertation explicitly avoids applications for summative assessment due to the higher stakes and potential for model errors to result in longer-term negative impacts on students. Although both diagnostic and formative assessment types serve complementary functions within classroom instruction, diagnostic assessment is likely more feasible to implement due to its less frequent administration requirements, and diminished teacher training demands. Formative assessment, by contrast, necessitates regular technology availability, rapid feedback mechanisms, and extensive professional development—constraints that present significant implementation barriers in resource-limited contexts. Nevertheless, throughout this work, I conceptualize diagnostic and formative assessment as functionally equivalent for research purposes, both representing low-stakes, school-level assessment modalities designed primarily to inform instructional decision-making (Wagner et al., 2012).

This dissertation addresses a significant research gap by exploring whether NLP and ASR can meaningfully support the diagnostic and formative assessment of reading ability in English in West Africa. I focus specifically on three widely used open-response format types: oral reading fluency, short answer questions, and story retell. My research approach combines the collection of original datasets of Ghanaian students' responses to all three assessment types with empirical experiments comparing how effectively different NLP and ASR approaches can evaluate these responses relative to human raters. While the empirical research is centered in Ghana, the findings may have implications for other Anglophone West-African countries given regional linguistic, cultural, and educational similarities, though such generalizations must be approached with appropriate caution. Through this work, I aim to determine whether current model performance is sufficient for practical implementation in ongoing, formative assessments,

and how this threshold varies for different assessment types and educational contexts. This research contributes not only novel datasets and empirical findings but also practical insights into how AI-powered assessment tools might transform reading instruction in resource-constrained environments.

1.1 Outline of Thesis

This dissertation aims to explore whether NLP and ASR can meaningfully support the diagnostic and formative assessment of reading ability in English in West Africa. More specifically, I focus on three of the most widely used open-response format types: oral reading fluency, short answer questions, and story retell. In exploring this topic, I touch on several interlaced research questions relating to evaluating and improving human grading of reading assessment tasks (e.g., ORF, comprehension, retell), exploring how well current NLP and ASR models can mimic human grading of these tasks, whether the level of model performance is sufficient for practical implementation in ongoing, diagnostic and formative assessments, and how this threshold varies for different assessment types and educational contexts.

Chapter 2 provides a detailed literature review, including exploring the background concepts necessary to understand the subsequent chapters. It covers topics such as the theoretical and practical aspects of assessing reading ability, the unique challenges faced by Low and Middle-Income Countries (LMICs) in assessing reading ability, the difficulties of conducting ongoing diagnostic and formative assessments on a large scale, and recent advances in NLP and ASR, with a specific focus on LLMs and other foundation models. This chapter also examines the opportunities and risks involved in assessing reading ability.

Chapter 3 provides a detailed overview of the methodologies used in the empirical papers, including explaining the techniques that could not be extensively discussed in the papers themselves. This chapter also details the design and collection of two novel datasets, which were developed in collaboration with Rising Academies, a school network in Ghana. These datasets comprise the results of 162 students aged 13-18 completing various reading assessment tasks.

The two datasets, now publicly available, aim to facilitate research on AI solutions for diverse educational contexts.

Chapter 4 focuses on the use of story retell as a method for assessing reading comprehension and explores the potential of LLMs in grading these assessments. It examines the impact of different rubrics on inter-rater reliability among human experts, the convergent validity of various retell measures, and the capacity of LLMs to grade story retells similarly to human raters. An earlier version of this chapter has been accepted for presentation at the 2025 conference Winter Text and Discourse.

Chapter 5 presents findings from a study using LLMs to grade short answer question responses from the novel Ghanaian dataset. This work is novel due to its empirical evaluation of generative LLMs' performance on grading short answer reading comprehension questions using real student data. An expanded version of this paper has been published in a special issue of the Journal of Artificial Intelligence in Education (Henkel et al., 2024).

Chapter 6 presents a study addressing two central challenges in automating ORF measurement: transcription quality and the conversion of transcription into words correct per minute (WCPM). The chapter compares the performance of different ASR models on corpora of student oral reading, including an existing corpus from American students and the novel dataset of Ghanaian students. An expanded version of this paper has been published in a special issue of the Journal of Artificial Intelligence in Education (Henkel et al., 2025).

Finally, Chapter 7 discusses and summarizes the findings, including expanding and synthesizing various topics raised in and across the empirical chapters. Moreover, it addresses the various implications and limitations of the research studies, and explores broader issues, including the importance of future research on interpretability and real-world application of these AI-powered assessment tools.

Chapter 2 - Literature Review

Reading development and assessment are complex domains that draw on multiple research traditions, from cognitive psychology to educational measurement. This chapter examines four key areas of literature that inform this research: (1) the theoretical foundations of reading development, (2) approaches to assessing reading ability, (3) recent advances in Natural Language Processing (NLP) and Automatic Speech Recognition (ASR), and (4) the emerging applications of these technologies in educational assessment. Given that this dissertation focuses specifically on reading in English, particular attention is paid to research on alphabetic writing systems while acknowledging that findings may not generalize fully to other languages and scripts (Nag, 2017).

2.1 Components of Reading Ability

Reading, in the context of this dissertation, refers specifically to the comprehension of written English text through both decoding of individual words and comprehension of the connected text. This section examines how children typically progress from early decoding skills to fluent reading and comprehension. Understanding these components of reading ability and their development is crucial for evaluating the potential of automated assessment approaches to support literacy instruction.

2.1.1 Learning to Read

Learning to read is often described as a series of interrelated but broadly sequential developmental stages through which children typically move (Castles et al., 2018; Ehri, 1995). For those learning to read in alphabetic systems, the first step is mastering the fundamental connection between written symbols and speech sounds—known as the alphabetic principle (Castles & Coltheart, 2004). This is a complex task because the graphic characteristics of letters are fundamentally arbitrary, meaning there is no innate correspondence between them and the sounds they represent (Nation et al., 2001; Wimmer & Goswami, 1994). Consequently, nearly all children require assistance by way of explicit teaching in learning to read.

Once a learner can decode words (i.e., the symbol-to-sound relationship) they will be able to match printed words into spoken language, meaning they can use their pre-existing knowledge

of the language to access meaning (Cunningham & Stanovich, 1997; Ouellette, 2006). When first learning to read, word recognition skills are the main factors distinguishing stronger and weaker readers, but as decoding becomes increasingly accurate and automatic, the correlation between reading ability and decoding skills tends to decrease, while the correlation between reading comprehension and language comprehension (e.g., listening to stories read, or spoken aloud) increases (Carver, 1998; Gough & Tunmer, 1986; Schwanenflugel et al., 2004).

The transition from reading words primarily via alphabetic decoding to recognizing words and phrases effortlessly (i.e., reading fluently) is critical in order for readers to successfully comprehend more complex texts; otherwise the pace of reading will be too slow and laborious to dedicate cognitive resources to memory for text and accessing relevant background knowledge (August et al., 2006; Catts et al., 2006; Reed & Vaughn, 2012). Relatedly, as readers master the basic skills of word recognition, reading comprehension comes to depend relatively more on language-comprehension abilities, and this is consistent with the shift that typically occurs in the middle elementary school grades from "learning to read" to "reading to learn" (Chall, 1983; Lervåg et al., 2018).

Because reading encompasses multiple interconnected elements, researchers have developed diverse frameworks to illuminate its various dimensions. Stage-based models of reading development are empirically supported by the fact that proficiency in early reading stages is predictive of later stages (Ehri, 1995). For example, letter knowledge has been consistently shown to be a strong predictor of early word reading, while word reading has been shown to predict reading fluency, and reading fluency to predict reading comprehension (Cunningham & Stanovich, 1997; Nation et al., 2010). A few of the most influential models are summarized below in Figure 2.1.

Figure 2.1
Similarities Across Influential Stage-Based Models of Reading Development

<table>
<tr><th></th><th></th><th>Chall (1983)</th><th>Firth (1985)</th><th>Ehri (1995)</th></tr>
<tr><td>Pre-reading</td><td>Read words by remembering visual or contextual cues.</td><td rowspan="2">Stage 0: Memory contextual guessing</td><td rowspan="2">Logographic</td><td>Pre-alphabetic</td></tr>
<tr><td>Early Reading</td><td>Acquire letter knowledge and use it to remember how to read words by forming partial connections in memory. Use the sound values of some letters to form connections between spellings and sounds.</td><td>Partial alphabetic</td></tr>
<tr><td>Decoding</td><td>Form connections between all of the graphemes in spellings and the phonemes in pronunciations to remember how to read words.</td><td>Stage 1: Decoding, attending to letters/sounds</td><td>Alphabetic</td><td>Full alphabetic</td></tr>
<tr><td>Fluency</td><td>Recognize the pronunciations and meanings of written words immediately upon seeing them without expending any attention or effort decoding the words.</td><td>Stage 2: Fluency, consolidation</td><td>Orthographic</td><td>Consolidated alphabetic, automaticity</td></tr>
<tr><td>Proficiency/ Comprehension</td><td colspan="4">Show an overall understanding of the text, including inferential as well as literal information. Be able to extend ideas in the text through simple inferences, make connections among and relate ideas in the text to personal experiences, and draw conclusions.</td></tr>
</table>

Note: Adapted from Ehri, 2005

However, it would be a mistake to think of reading development as strictly hierarchical; it is best understood as an interactive process that combines bottom-up skills and top-down processes at any developmental or ability level (Perfetti & Adlof, 2012). Precisely because reading is a complex skill that must be explicitly taught, it is quite common for learners to struggle with different aspects of reading due to cognitive abilities, language, or learning environment. While pre-reading, early reading, and decoding - as laid out in Figure 2.1 above - are clearly prerequisites for learning to read, this dissertation focuses on the subsequent stages of fluency and comprehension.

2.1.2 Reading Fluency

Reading fluency is the ability to read a text accurately, with proper expression, and at a sufficient rate so that working memory is not overburdened (National Reading Panel, 1998). Non-fluent readers can read words correctly but do so slowly and laboriously, with an awkward or

monotonous tone, while struggling to group words appropriately (Fuchs et al., 2001; Rasinki, 2004; Spear-Swerling, 2006). Fluent reading is an important component of reading ability because it demonstrates both word knowledge and automaticity, each of which are key components of successful reading (Perfetti & Hogaboam, 1975; Rasinski et al., 2009). As word reading becomes increasingly efficient and fluent, it uses fewer cognitive resources, which can then instead be allocated to maintaining text in working memory, accessing background knowledge, and making inferences (LaBerge & Samuels, 1974; Wolf et al., 1986). Various empirical research has shown that without sufficient automaticity, it is impossible to retain a sentence in working memory (August et al., 2006; Catts et al., 2006; Schwanenflugel et al., 2004).

Much of the research on the relationship between reading fluency and reading comprehension focuses on Oral Reading Fluency (ORF) because of the practical difficulties involved in the measurement of silent reading fluency, particularly in classrooms or other field situations (Hasbrouck & Tindal, 2006). Empirical studies have shown strong correlations between ORF and reading comprehension and ORF is generally considered a reliable predictor of reading comprehension, particularly in early grades (Fuchs et al., 2001; Roehrig et al., 2008; Spear-Swerling, 2006). For instance, Fuchs et al. (2001) found that ORF is strongly associated with conventional measures of reading comprehension, with correlations ranging from .80 to 90. Moreover, research has shown that ORF in third graders predicted fourth-grade reading comprehension and accounted for a larger percentage of variance than any factor except third-grade reading comprehension itself (Spear-Swerling, 2006). In addition, this relationship has been shown both in high stakes state-mandated exams, as well as a variety of different assessment formats, such as Multiple-choice, open-response, cloze, and story summarization tests (Deno et al., 2001; Wolf & Katzir-Cohen, 2001). It should also be noted that ORF becomes less predictive of reading comprehension as students develop greater reading proficiency and word-identification accuracy (Jenkins et al., 2003; Kim et al., 2010). This non-linear relationship is attributed to the fact that at the beginning stages of reading development, both ORF and reading comprehension are sensitive to individual differences in word identification (Hoover & Gough, 1990; Valencia et al., 2010). However, once basic reading skills have become automatic and reading is relatively fast and accurate, reading comprehension begins to depend relatively

more heavily on language comprehension, and as a result, ORF becomes less predictive (Fuchs et al., 2001; Spear-Swerling, 2006).

2.1.3 Reading Comprehension

Reading comprehension, as a psychological construct, refers to a reader's ability to extract and create meaning from written text. There are numerous theoretical models describing reading comprehension. One of the best known and most widely cited is the Simple View of Reading, which posits that successful reading is a product of two elements: word recognition (reading words accurately and fluently, accessing lexical representations), and language comprehension (vocabulary, grammar and syntax). Various empirical studies have shown that decoding and linguistic comprehension each predicts reading comprehension (Gough & Tunmer, 1986). Another influential model is Kintsch's (1988) Construction-Integration model, which describes three levels of comprehension: literal understanding of the words and phrases, the "gist" understanding of what is being read, and a more complex situational model. These models, along with numerous others beyond the scope of this review, generally agree that reading comprehension is a complex construct involving the coordination of a number of integrated processes, whereby readers construct mental models of text meaning at multiple levels (Castles et al., 2018; Hulme & Snowling, 2011; Perfetti & Adlof, 2012; Vellutino et al., 2007).

While most theoretical models of reading comprehension refer to it as a unified construct, its assessment in practice involves measuring various subskills—from understanding literal meaning, to reasoning about textual information, to evaluating claims. Different assessment formats (e.g., multiple-choice, story retell, short answer, and summarization) attempt to measure this construct, each capturing different aspects of comprehension ability. However, psychometric research has found that different assessments purporting to measure reading comprehension often show lower than expected correlations with each other, suggesting they may be tapping into distinct subcomponents of comprehension or introducing construct-irrelevant variance through format effect (Cain & Oakhill, 2006; Gregory, 2007). Hence, when assessing reading comprehension, it is important to be specific what one is referring, and to verify that different assessment methods produce correlated results (Collins et al., 2020).

In this research, I adopt the definition of reading comprehension proposed in the PISA for Development Reading Framework (2018), summarized in Figure 2.2. I choose this framework because it articulates increasingly sophisticated levels of reading comprehension in terms of the associated student capabilities and is focused on assessment of reading comprehension in English - and other romance languages - in LMICs.

Figure 2.2
Processes Comprising Reading Comprehension

Process	**Readers are able to…**
Access and Retrieve Information	Identify in print individual words that would occur in the everyday listening lexicon of average adult speakers of the language Combine words in order to parse sentences and represent their literal meaning. Locate explicitly stated individual pieces of information, in large passages.
Literal Comprehension	Comprehend explicitly stated information that may be found in sentences or passages. Combine the meaning of small sets of sentences in order to form an internal representation of simple descriptions or narrations. Form a representation of the information contained across multiple sentences, connecting the idea units and structuring them in memory.
Form a Broad Understanding	Consider the text as a whole or in a broad perspective. Demonstrate initial understanding by identifying the main topic or message or by identifying the general purpose or use of the text. Connect various pieces of information to make meaning, comparisons of degree, or understanding cause and effect relationships.
Develop an Interpretation	Go beyond the literal meaning of a passage or a text and identify the underlying assumptions or implications. Identify and list supporting evidence, and compare and contrast information, draw together two or more pieces of information from the text
Reflect on and Evaluate the Content of a Text	Connect information in a text to knowledge from outside sources. Assess the claims made in the text against their own knowledge of the world. Articulate and defend their own points of view
Reflect on and Evaluate the Form of a Text	Consider a text objectively and evaluate its quality and appropriateness. Evaluate how successful an author is in portraying some characteristic or persuading a reader.

Note: Adapted from the PISA for Development Reading Framework (2018)

The first three sub-processes—accessing and retrieving information, literal comprehension, and forming a broad understanding—and the associated descriptions of student abilities closely correspond both with definitions of functional literacy and the types of cognitive

abilities required by reading comprehension questions found in hybrid assessments and Large-scale Assessments (LSAs) targeted for emerging readers (Keenan et al., 2008; Kirsch & Guthrie, 1977). For this reason, when discussing "reading comprehension" in subsequent sections, I will primarily be using the term in the narrower sense of these three sub-processes and their associated skills.

2.2 Assessing Reading Ability

Assessment is the process of obtaining information on students' knowledge and abilities in order to give feedback to the student or teacher, judge instructional effectiveness, and inform policy (Braun et al., 2006). It is important to assess students' underlying skills in order to monitor education provision and inform decisions regarding curriculum, school management, and resource allocation and to inform teaching practice at the classroom level (Nag, 2017; Wagner et al., 2012). In the context of reading specifically, assessment is particularly important because reading skills are foundational for academic success across subject areas, and early identification of reading difficulties can prevent the cascading negative effects of reading failure (Cunningham & Stanovich, 1997).

2.2.1 Assessing Reading Fluency

While ORF can be evaluated in various ways, one of the most widely used measures is words correct per minute (WCPM). This is calculated by having a student read a passage for one minute while an assessor marks incorrect words, after which the correctly read words are tallied (Piper et al., 2016). Proponents of WCPM emphasise that it can be used with different texts, provided the texts are of similar difficulty, and that it is relatively straightforward to administer and produces easy to understand results (Good et al., 2001; Valencia et al., 2010). However, one of the main criticisms of using WCPM is that it only operationalizes two of the three elements in the commonly accepted definition of fluency—speed and accuracy—while ignoring prosody. Some argue that when a student's ability is measured by reading speed, teachers and students may fixate on reading fast, rather than reading the text at a normal rate with correct intonation, which in turn may actually interfere with comprehension, undermining the goal of students reading for understanding (Bolaños et al., 2013). Moreover, some empirical research has found that stronger prosody early in school is associated with higher levels of later reading fluency over

and above word-rate based measures (Benjamin & Schwanenflugel, 2010; Kuhn et al., 2010; Miller & Schwanenflugel, 2008).

Assessments that attempt to include prosody typically entail a trained rater listening to multiple children performing the same task and assigning them scores based on a pre-designed rubric. The National Assessment of Educational Progress *(NAEP)* Oral Reading Fluency Scale is one of the most commonly used measures, employing a 4-point rubric where raters evaluate children from reading primarily word-by-word at the low end to reading in meaningful phrase groups at the upper end (Daane et al., 2002). Annual Status of Education Reports (ASER) uses a different approach for evaluating ORF that asks a pair of highly trained assessors to rate whether the child reads fluently or haltingly, more specifically, if the student sounds like they are "reading the words as a sentence" (R. Banerji, personal communication, 2020). ASER conducted internal research explicitly comparing WCPM measures to expert raters and find high (over 0.9) correlation between them (R. Banerji, personal communication, 2020). Nonetheless, there is debate about the reliability of this approach because of observed low inter-rater reliability scores, typically 0.6 to 0.7. However, the alternate approach, directly specifying linguistic characteristics of prosody (e.g., intonation, pacing, appropriate pauses in between words, etc.) has the same inter-rater reliability challenges and is also time-consuming and technically challenging to administer (Cowie et al., 2002). As a result, improving the assessment of prosody and developing better understanding of its role in reading comprehension has been identified as an important area for further research (Kuhn et al., 2010; Roehrig et al., 2008).

2.2.2 Assessing Reading Comprehension

A variety of question formats are used to assess reading comprehension, each with its own benefits and limitations (Bellinger & DiPerna, 2011). Closed response questions, including multiple-choice, true/false, and fill-in-the-blank, have been the dominant method for assessing reading comprehension for several decades, likely due to their cost-effectiveness and simplicity to administer and score (Pearson & Hamm, 2006). However, they also have several limitations. First, they are time-consuming to develop, and questions cannot be reused if test security is a concern. Second, students' responses may be more related to their background knowledge or test-taking strategies than to actual comprehension. There are examples of individuals

performing significantly better than chance when answering the questions without even having read the corresponding passages (Magliano & Graesser, 2012; Magliano & Millis, 2003). Third, some claim that these closed response questions may lack face validity because of the disconnect between what students are asked to do on an exam and what they are expected to do in day-to-day circumstances. Specifically that they show whether students recognize meanings that test developers construct from text, but not whether students can independently extract meaning from a text (Jenkins et al., 2004; Landauer et al., 1998).

These limitations may become more serious in contexts where students come from multilingual environments and are not already fluent in the language in which they are learning to read (Klaas & Trudell, 2011; Spaull et al., 2020). For instance, a student who has adequate decoding skills but lacks sufficient understanding of the language they are reading, or one with adequate decoding and comprehension skills but a lack of familiarity with multiple-choice question formats, might score similarly to a student with weak decoding skills. Indeed, one of the main arguments for open-ended formats is that they can evaluate more varied aspects of reading comprehension than closed response questions can (Pearson & Hamm, 2006).

Open-ended comprehension questions require a student to read or listen to a passage and then answer questions, recall key elements, or construct a summary in their own words (Shapiro et al., 2014). Unlike closed response questions, there is no single correct answer, and grading is usually conducted by raters using a rubric (Reed & Vaughn, 2012). Some researchers argue that open-ended questions are appealing because they rely on students actively reconstructing text, and the close correspondence with real-world applications of reading comprehension, i.e., they have greater face validity (Bishop & Adams, 1990; Cain & Oakhill, 2006). They also argue that they are better suited to evaluate more complex facets of reading comprehension, such as forming a broad understanding of a text, developing an interpretation of it, and/or evaluating its content (OECD, 2018; Pearson & Hamm, 2006). The three most widely used types of open-ended questions are short answer, story retell, and summarization. Short answer or constructed response items require the student to read a passage and then answer a specific question about the passage using their own words in a few sentences (Burrows et al., 2015). Correct answers are

seen as prima facie evidence of comprehension, as there is no way of using process of elimination to identify the most plausible of a set of options (van den Bergh, 1990).

While the terms story retell, story recall, summarization, and paraphrasing are sometimes used interchangeably in the literature, it is important to note their conceptual distinctions as related but distinct skills. Kintsch and van Dijk (1978) emphasize that retelling, recalling, summarizing, and paraphrasing are considered distinct skills that require different degrees of telling or transforming knowledge.

Figure 2.3
Definitions for Different Assessment Tasks Related to Whole Story Comprehension

Task Type	Definition
Retell	The content in the passage in sequence
Recall	The key content from the passage, close to original wording, not necessarily in sequence
Summarize	The main ideas and significant details of the passage in a shortened organized form
Paraphrase	Restating a particular part of the content in different words, likely involves the main idea

Story retell involves a comprehensive recounting of the text, including details and maintaining the original sequence of events, while story recall focuses on retrieving specific information or events without necessarily preserving the original structure. Story summarization requires students to distil the main ideas and key points of the text, often involving a higher level of abstraction and transformation of the original content. Story summarization requires students to read a passage and describe the key ideas and concepts in a coherent way, while at the same time eliminating superfluous details (Kintsch & van Dijk, 1978; Spirgel & Delaney, 2016). Various studies have shown a strong association between summarization tasks and other measures of comprehension (Head et al., 1989; Perin & Lauterbach, 2018; Winograd, 1984).

2.2.3 Language Specific Considerations in Reading Assessment

When encountering a child with poor reading skills, there are multiple potential causes and it may be difficult to determine if the root cause is inadequate prior instruction, a lack of proficiency in the language, issues related to family life, or in rare cases an underlying cognitive impairment (Alderson et al., 2016; Geva & Siegel, 2000; Kormos et al., 2019). Children

experiencing inadequate reading instruction are less likely to develop strong decoding skills, which restricts their opportunities for the necessary practice to avoid being caught in a self-perpetuating cycle of learning loss (Cunningham & Stanovich, 1997). L2 learners, those whose first language is different than the language of instruction, face additional challenges when learning to read.

These children typically have smaller vocabularies in the language of instruction and less relevant background knowledge in relation to the texts they are reading (Geva & Massey-Garrison, 2013). Consequently, even with strong decoding and word recognition, these students will not be able to access the thousands of words already present in the oral vocabularies and the rich understanding of the grammar and syntax of a language they already speak. In the case of multilingual educational contexts such as Ghana, the larger majority of students are not fully proficient in the language in which they are being taught to read. As a result, it can be difficult to determine if reading comprehension difficulties are due to inadequate prior instruction in decoding skills or a lack of proficiency in L2 (i.e., language of instruction).

These challenges can also be further complicated by the specific characteristics of L1 and the language of instruction. English is generally considered to have a deep orthography relative to other alphabetic languages, which means it takes longer to learn to decode (Hanley et al., 2004; Wimmer & Goswami, 1994). Several studies of L1 learning to read in languages with shallow orthographies reported high accuracy (i.e., greater than 95% scores) for word recognition by the end of Grade 1, while children learning to read English with its relatively deep orthography scored no more than 70% correct in word reading (Frith, 2017). However, most of the research on ORF has been confined to English speakers reading in their native tongue. While most research shows that poor fluency and word recognition rates in beginning English readers are closely associated with both current and future reading comprehension, the strength of this relationship varies significantly by language and there is no set fluency rate at which comprehension is demonstrated across languages (Bartlett et al., 2015; Dowd & Pisani, 2013; Good et al., 2001). For this reason, some researchers argue that English is something of an outlier in terms of spelling–sound correspondence, which might mean that reading science research conducted on English speakers has less relevance for other languages (Share, 2008).

2.2.4 Large Scale Assessment of Reading Ability

Many countries periodically conduct large-scale-assessment (LSAs) to evaluate educational progress at the national level. National large-scale assessments, such as the National Assessment of Educational Progress *(*NAEP) in the United States, generally focus on capturing information about students in a single educational system, whereas international large-scale assessments, such as the Progress in Reading Literacy Study (PIRLS) and the Programme for International Student Assessment (PISA), are typically undertaken primarily to allow for cross-national comparisons of educational achievement and intra-country analyses of student proficiency in foundational skills and core subjects (Lockheed, 2008). While LSAs may be helpful in monitoring and comparing relative academic performance and student achievement at the national level, they are poorly suited for assessing literacy and tracking student progress. First, because they are administered infrequently and to small sample groups, it is not possible for the results to be used at the classroom level where they could help teachers identify struggling students, diagnose their difficulties, and select the appropriate strategies to support them.

Large-scale formative assessments, also referred to as hybrid assessments, aim to combine the methodologies of large-scale standardized assessments with in-class formative assessments (Braun et al., 2006; Nag, 2017). Proponents argue that they can be an important complement to LSAs because of their ability to be deployed and adjusted more frequently, to be adapted for specific cultural and linguistic contexts, and to be easily understood and administered by local educators. Perhaps most importantly, the assessment results can help teachers adapt their teaching methods (Dowd & Pisani, 2013). Large-scale formative assessments typically focus on fundamental skills and are administered individually, consist of a group of assessment tasks, with specific tasks selected based on the situation, and primarily evaluate foundational reading skills in a simple, quick, cost-effective, and easy-to-administer manner. This allows for prompt release and reporting of results in a format that is easily understood by all stakeholders involved (Dowd & Pisani, 2013).

The popularity of hybrid assessments is demonstrated by Nag's (2017) review of literacy assessments in LMICs, which identified over 300 unique reading assessments in more than 50 languages, many of which shared key characteristics of hybrid assessments. Two well-known and extensively researched hybrid assessments are ASER and EGRA, both of which include open-ended reading tasks. ASER has been conducted annually for over a decade, assessing over 600,000 children in approximately 16,000 villages across India (Banerji, 2016). EGRA, developed in 2006 to address the lack of literacy assessment tools in LMICs, consists of a toolkit of 14 subtests that evaluate foundational literacy skills in the early years of primary school across various alphabetic languages, and is one of the most widely administered assessments of basic skills worldwide (Dubeck & Gove, 2015). However, as discussed above, assessment systems based on open-ended responses are challenging to implement at scale because they require extensive training and calibration of raters and are therefore resource intensive, which in turn limits their use.

While hybrid assessments like ASER and EGRA have made important contributions to tracking students' reading levels, their reliance on human administration makes them operationally complex and expensive to implement. The limitations of these assessment systems reflect a broader challenge in educational measurement: balancing the need for regular, individualized assessment against practical constraints of time and resources. This tension is particularly acute in low-resource settings, where the costs and complexities of administering individual assessments often restrict their use to periodic sampling rather than ongoing diagnostic and formative assessment. Recent technological advances, particularly in artificial intelligence and speech recognition, may offer new solutions to this longstanding challenge. Recent advances in Natural Language Processing (NLP) and Automatic Speech Recognition (ASR), particularly the emergence of large language models and improved speech recognition systems, suggest new possibilities for making reading assessment more accessible and scalable in low-resource contexts.

2.3 Recent Advances in NLP and ASR

NLP, which is typically focused on the written word, and the closely related field of ASR, which focuses on the spoken word, entail the analysis of human language by computers (McNamara et

al., 2017). While the application of NLP and ASR to education has been a topic of research for at least two decades, rapidly improving infrastructure and computing power, coupled with important and theoretical advances have led to increased interest among researchers and practitioners (McNamara et al., 2017).

2.3.1 Historical Progress of NLP

The earliest research in NLP dates back to the 1960s, though initial approaches could only handle text produced under tightly specified conditions, and computing was still decades away from being able to evaluate open response items effectively (Foltz et al., 2020; Page, 1966). In the 1990s, research began incorporating machine learning (ML) algorithms. Unlike pre-specified rules, these algorithms could independently identify patterns in example data and generate predictions for new data, and included approaches like bag-of-words, N-gram, and Term Frequency-Inverse Document Frequency, which compare the frequency and co-occurrence of words in responses (Holmlund et al., 2020). In the 2000s, these approaches were augmented by newer approaches such as Latent Semantic Analysis, which aims to capture the semantic relationships between words and represent them numerically. It does this by converting words into numerical values, thereby reducing their dimensionality, and word relationships are then represented in a matrix (Landauer et al., 1998; Torfi et al., 2020). When applied to texts, the mathematical distance between them can be calculated, with smaller distances indicating similar responses. For instance, in automatic essay grading, better scores are awarded to essays more similar to highly rated human-evaluated example texts (Dumais, 2005; Foltz et al., 2000). Despite theoretical progress, practical applications of these approaches to essay scoring that performed at human rater-like levels were still limited into the 2010s.

A significant development came with the introduction of word embedding and Word2Vec (Mikolov et al., 2013). Word embeddings represent words numerically in relation to one another, measuring similarity between words rather than between whole texts. Word2Vec was trained using two tasks: Continuous bag of words and Skip Gram. It was one of the first NLP approaches to use neural networks to generate word embeddings through unsupervised learning, allowing the model to be trained on unlabelled text from large online corpora like Wikipedia rather than relying on human-labelled datasets. The success of Word2Vec led to the

development of other word-embedding models such as GloVe in 2014, and FastText in 2016 (Joulin et al., 2017; Pennington et al., 2014).

Concurrent research explored applying Recurrent Neural Networks (RNNs) to language modelling. RNNs specialize in sequential data, making them suitable for language processing compared to approaches that ignore word order or only consider limited sequences (Cheng et al., 2015; Taghipour & Ng, 2016). Long Short-Term Memory, a popular RNN variant, provided an efficient way to process input sequences while maintaining information about previous content (Mueller & Thyagarajan, 2016). One of the highest-performing LSTM models was ELMo (Peters et al., 2018). Despite their progress, RNN-based approaches faced a fundamental computational limitation: their sequential nature meant that each word had to be processed one after another, with the model tracking relationships between words that could be far apart in the sentence. This inherently sequential processing made it difficult to parallelize the computations across multiple processors, creating a bottleneck that limited both training speed and the model's ability to handle longer texts (Vaswani et al., 2017). An important advance was the transformer architecture, which was easier to parallelize, resulting in faster training. It also improved the model's ability to retain information about long-range dependencies (Vaswani et al., 2017). This new architecture allowed models to be trained on enormous text corpora consisting of billions of words, which significantly enhanced performance. This in turn enabled a new paradigm of model-training, namely, transfer learning.

Building on these advances, researchers in the late 2010s proposed BERT (Bidirectional Encoder Representations from Transformers), which quickly became state-of-the-art on numerous benchmark tasks (Devlin et al., 2019). Using a transformer architecture, the model could be pre-trained on huge text corpora—by some estimates over 3 billion words—producing strong results on generic NLP tasks. The model could then be fine-tuned with task-specific data to optimize results for specific problems (Sung et al., 2019). While BERT's introduction sparked the development of numerous early transformer variants like RoBERTa and GPT-2, the field has evolved rapidly since then, with more recent models demonstrating dramatically improved capabilities, as will be discussed in section 2.3.2 on the emergence of modern large language models.

2.3.2 Importance of Transfer learning to NLP Capabilities

Transfer learning is a machine learning technique that usually consists of two steps: pre-training and fine-tuning. During the pre-training phase, a neural network model conducts unsupervised learning on a large-scale general dataset to establish model weights. Next, during the fine-tuning phase, the model is trained using supervised learning on a smaller, carefully constructed dataset (Kojima et al., 2022). In the specific case of NLP models using a transformer architecture, during the pre-training phase, the model is exposed to a vast amount of unlabelled text data, such as books, articles, or web pages. The model learns to predict missing words or sentences in this unlabelled data, effectively capturing the statistical patterns, grammar, and contextual relationships in the language (Devlin et al., 2019). This task is often called masked language modelling or next-token prediction. The objective of pre-training is to build a strong language model that encodes comprehensive knowledge about language syntax and semantics.

The second stage in the transfer learning paradigm, fine-tuning, typically involves two steps. First, the pre-trained model is taken, and specific layers or parameters are unfrozen, allowing them to be modified during the fine-tuning process. Second, the model is then fine-tuned on a task-specific dataset (Kojima et al., 2022; Raffel et al., 2020). This process significantly reduces the necessary amount of training and allows the model to adapt to the specifics of the target task and domain. The transfer learning approach offers an alternative to the traditional statistical NLP approach, which requires significant time and expense for training task-specific models on labelled datasets. Through exposure to an extensive text corpus, many scholars contend that the model develops sophisticated internal representations capable of capturing meaningful patterns in language and context (Raffel et al., 2020; Wei et al., 2022). However, while models are usually task-agnostic in architecture, adapting them for new tasks still requires task-specific datasets that are usually composed of thousands of examples (Brown et al., 2020). Building upon these acquired language representations, the subsequent fine-tuning stage modifies the model to execute specific tasks within a target domain using smaller datasets with task-specific labels.

While transfer learning based approaches have demonstrated high accuracy when evaluating specific tasks and represent substantial improvement relative to preceding approaches,

their development requires significant technical expertise, which is often unavailable in most schools. Even in situations where a school may have access to such expertise and can develop such a model, a labelled dataset of a few hundred examples of correct and incorrect student answers is required, and the model must be fine-tuned each time the question students answer changes, an example of the challenges of generalizability (Camus & Filighera, 2020; Perez et al., 2021). In the context of transfer learning, generalizability refers to a model's ability to perform well on unseen or previously unseen data, beyond the data it was trained on. Models that generalize well can adapt to different situations, data distributions, and domains, providing more reliable and practical solutions (Blanchard et al., 2011). However, there are several situations where models may not generalize well. Small variations in data distribution between the source domain (where the model undergoes training or fine-tuning) and the target domain (where the model is deployed) can result in degraded model performance. Domain shift presents a particularly difficult challenge to overcome because even tasks that appear highly similar in their fundamental structure and objectives may contain subtle distinctions that are not readily apparent yet considerably impact the model's effectiveness (Elsahar & Gallé, 2019; Zhao et al., 2020).

2.3.3 Recent Breakthroughs with LLMs

The more recent generation of Large Language Models (LLMs) such as GPT-4, Claude, Llama, Mistral, Gemini, etc., were trained in a similar manner to previous generations of LLMs, but utilized significantly larger datasets and dramatically increased the number of parameters in the model, in some cases by more than an order of magnitude (Chowdhery et al., 2023; Stiennon et al., 2022). Additionally, following the pre-training stage, these models underwent "instruction fine-tuning" to enhance their ability to follow human instructions and align their output with human preferences (Stiennon et al., 2022). Although the implementation details differ somewhat across model architectures, the overall outcome was a significant improvement in accessibility, enabling users without technical expertise to interact through natural language commands (known as prompting), along with a notable capacity to adapt to unfamiliar tasks, frequently with limited example exposure (Ouyang et al., 2022).

Importantly, these general LLMs can perform various linguistic tasks that traditionally required task-specific, fine-tuned LLMs (Kojima et al., 2022; Wei et al., 2022). They can do so

by relaying instructions and a limited number of examples. Thus, they promise a solution to both key limitations of the transfer learning-based approach: the complexity of the fine-tuning process and limited generalizability. Moreover, there is also a rapidly growing body of evidence that these models can complete evaluation tasks on novel datasets with only minimal prompt engineering (Gilardi et al., 2023; Kuzman et al., 2023). This capability has significant implications for various applications, including automated grading. Instead of developing a custom-trained model for each specific grading task, users can now provide the LLM with a description of the grading criteria and a few examples, enabling it to assess student responses effectively (Kortemeyer, 2023; Mizumoto & Eguchi, 2023; Schneider et al., 2023). Cohn et al. (2024) explored using an LLM to grade short answer responses in high school science. Their approach combined several techniques to improve accuracy: they showed the model a few examples of correctly graded answers (few-shot learning), had humans review and correct the model's initial grades to help it improve (active learning), and asked the model to explain its reasoning step-by-step when making grading decisions (chain-of-thought reasoning). Using these techniques together, they found that GPT-4 could successfully score student answers at a level comparable to human graders.

2.3.4 Recent Progress in ASR

While ASR has been commonplace in various use cases including telephone call-routing, captioning live television, and transcribing communications, these are all tightly controlled contexts, and ASR models historically struggled when asked to deal with new datasets and tasks (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 et al., 2023b; Nouza et al., 2023). However, this approach limited the size and diversity of the datasets that were available to train models. This created a problematic situation: while the models performed extremely well on the specific types of audio recordings they were trained on, they often performed poorly when encountering new types of recordings or real-world conditions that differed from their training data. For example, a model trained primarily on clear recordings of American English speakers might struggle with background noise or different accents, even

though it showed excellent performance on its original test dataset (Geirhos et al., 2020; Radford et al., 2022).

In response to these challenges, other researchers proposed unsupervised pre-training methods, most notably wav2vec 2.0. The wav2vec 2.0 approach uses unsupervised learning that can be trained on audio without requiring human labels, thereby unlocking large datasets of unmarked speech and making it feasible to train models on up to a million hours of audio (Zhang et al., 2022). However, this approach has a notable drawback; the model is only trained on input speech, so it requires a fine-tuning phase to perform tasks such as speech transcription, which is often a complex process requiring a skilled practitioner (Radford et al., 2022). More recent research has explored a "lightly supervised" approach, which uses both unsupervised and supervised data and relaxes the need for human-validated transcripts. This allows for 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 hours of audio from diverse recording setups, environments, speakers, and languages. When tested across a variety of ASR datasets, Whisper v2 demonstrated significantly better accuracy than wav2vec 2.0, making about half as many transcription errors. Specifically, Whisper v2 achieved an average Word Error Rate of 12.8, meaning it incorrectly transcribed roughly 13 out of every 100 words, compared to wav2vec 2.0's error rate of 29.3, or about 29 mistakes per 100 words (Radford et al., 2022). This improvement in accuracy represents a substantial step forward in making automated transcription more reliable for real-world applications, including educational assessment.

2.4 Using NLP and ASR to Assess Reading Ability

Early research applying NLP and ASR techniques in educational settings was predominantly related to surface-level text processing, including spelling and grammar checkers, translation tools, and plagiarism detectors (Magliano & Graesser, 2012; Shermis et al., 2010). As research progressed and it became possible for computers to analyse text at the level of semantics and discourse structure, NLP was used as part of intelligent tutoring systems, particularly those that interact with students via dialogue (McNamara et al., 2017). Subsequently, research was carried out in adjacent areas such as virtual agents, student support chatbots, and discourse analysis in

online collaborative learning environments, such as massive open online courses (MOOCs) (Graesser et al., 2000; Ma et al., 2014; McArthur et al., 2005). Another widely researched application of NLP in education is automated scoring, including the scoring of open-ended responses to assessment questions (Burrows et al., 2015; Foltz et al., 2020). Automated scoring can be understood as automatically converting students' performance on educational tasks into relevant evidence to make inferences about the students' abilities (Foltz et al., 2020). For closed-response questions this is relatively easy to accomplish as there are already objective answers that can be scored by computers or other specialized machines (DiCerbo et al., 2020). In contrast, open-ended or performance-based assessment requires learners to produce written or spoken responses, which can potentially provide richer and more nuanced evidence regarding learners' skills, but these responses are far more difficult to interpret and have historically required human grading (DiCerbo et al., 2020).

Interest has been increasing over the past few years in one potential solution—using Natural Language Processing (NLP) to help with the assessment of reading fluency and comprehension. A particularly promising approach is to take large language models (LLMs), which have been pre-trained on large text corpora and have encoded more general statistical properties of language and fine-tune them on task and domain-specific data so that they can accurately evaluate student responses. This approach has been discussed in a wide range of domains, including law, medicine, and biology. It has also shown some promising initial results in the field of educational assessment (Fernandez et al., 2023; Ullmann, 2019).

2.4.1 Automatically Assessing Reading Comprehension

Open-ended questions are well-suited to assessing beginner-readers' comprehension, but the time, cost, and complexity associated with administering and grading them limit widespread usage. While closed-response formats like multiple choice can be scored automatically, the automated scoring of open-ended responses presents significant technical challenges. One common open-ended question format for assessing reading comprehension is short-answer. This section examines the evolution of approaches to automatic short answer grading (ASAG), with particular attention to reading comprehension assessment. Early research in ASAG faced two major limitations. First, developing these systems required large collections of human-graded

student responses to train the models - datasets that were both expensive and time-consuming to create. Second, the models needed to understand domain-specific vocabulary and concepts, making it difficult to adapt them for different subjects or types of questions.

Prior to the emergence of the most recent generation of LLMs (e.g., GPT-4, Bard, Claude), there had been promising research pointing to the potential for LLMs to assist with literacy assessment. For instance, Sung et al. (2019) experimented with fine-tuning BERT, a widely used pre-trained transformer-based language model, to classify pairs of student answers as correct or incorrect. They found that this approach produced superior results across multiple domains, including a 10% improvement over previous state-of-the-art results on the SemEval-2013 task, and was able to classify student responses almost at human-level rates of agreement.

A significant benchmark in reading comprehension assessment came from the NAEP automated scoring challenge. The winning approach, developed by Fernandez et al., (2023), used a specialized language model to achieve agreement scores of 0.84 with expert human raters (compared to human-to-human agreement of 0.88). However, this system, like its predecessors, required extensive technical expertise to implement and could only be used for the specific types of questions it was trained on, and struggled to generalize with similar but distinct tasks, for instance, reading comprehension for a different topic, question type, or level of difficulty.

The newest generation of Large Language Models has demonstrated significant advances in language understanding and generation. Through new training approaches that incorporate human feedback and preferences (called Reinforcement Learning from Human Feedback, or RLHF), these models have developed broader capabilities than their predecessors. Unlike earlier models that required extensive task-specific training, these newer models can often handle complex language tasks with just a few examples of the desired behavior and clear instructions about the task - an approach known as few-shot learning.

These capabilities suggest these models could be particularly well-suited for evaluating reading comprehension responses. When grading a student's answer, the model can consider both the original reading passage and the question, which provides all the context needed to assess the response's accuracy. This approach differs from earlier automated scoring systems that needed to

be extensively trained on large sets of pre-graded student answers. While early results appear promising, rigorous empirical research on using these models for reading comprehension assessment remains limited. The following chapters address this gap, particularly Chapter 5, which presents the first systematic study of how well these models can grade short answer reading comprehension questions compared to human raters.

2.4.2 Automatically Assessing Reading Fluency

Because of the strong empirical and theoretical links between oral reading fluency and reading comprehension, ORF tasks are widely used for both diagnostic and formative assessment purposes. These assessments are typically conducted one-on-one between a student and a trained evaluator, making them time-consuming and costly to administer. This creates potential benefits for using NLP and ASR to partially or fully automate ORF assessments (Duong et al., 2011). Ideally, an automated version of this assessment would record the student reading, analyse the errors, calculate the total, and report scores that are as reliable as, if not more reliable than, human ratings. Therefore, automatic scoring would involve two main tasks: (1) transcribing a recording of the student's oral reading, and (2) calculating the accuracy rate by comparing the transcript of the student's reading to the original document they were reading from. This challenge is non-trivial because in addition to having to deal with unique aspects of children's speech, these models must accurately distinguish between correct and incorrectly read words, whereas most speech recognition models would use the context and surrounding words to replace incorrectly read words with the most plausible alternative.

Transcribing children's speech has been a long-standing challenge in ASR due to limited high-quality child speech datasets available for training, different linguistic and acoustic features, and different pronunciation abilities (Gray et al., 2014). Children also tend to use imaginative words, ungrammatical phrases, and incorrect pronunciations, which pose challenges for language modelling (Claus et al., 2013; Gray et al., 2014). As a result, implementing ASR in classrooms with young students was previously considered complex, time-consuming, and expensive (Cincarek et al., 2009; Gerosa et al., 2009). However, over the past five years, research has made significant progress in ASR for adult speech, achieving human-level performance and researchers are actively exploring the adaptation of these models to enhance ASR for children (Baevski et al., 2020; Radford et al., 2022).

Recent developments in self-supervised learning have shown improvements in child speech recognition (Jain et al., 2023a; Thienpondt & Demuynck, 2022). In one set of experiments, Jain et al. (2023b) explored using the ASR model wav2vec 2.0 with different pre-training and fine-tuning approaches. Their models outperformed the unmodified wav2vec 2.0 on child speech, achieving a Word Error Rate (WER) below 10% using as little as 10 hours of child speech data for fine-tuning (Jain et al., 2023b). In a subsequent study, they compared Whisper, a large ASR model released by OpenAI, with fine-tuned, self-supervised models like wav2vec2.0 (Fan & Alwan, 2022). Whisper, trained with significantly more data and multilingual resources, was expected to improve child speech recognition through fine-tuning. The results showed that fine-tuning Whisper on child speech led to significant improvements in ASR performance compared to non–fine-tuned Whisper models, but fine-tuned wav2vec 2.0 models still outperformed the fine-tuned Whisper model.

One of the most widely used measures of ORF—WCPM—has the benefits of simplicity and strong correlations with other measures of comprehension but it excludes a component of fluency—prosody. Due to the technical complexity, only recently has prosody been attempted to be measured in automated assessments of fluency. Some researchers formally define prosodic features, such as duration, pitch, or filled-pauses and measure these characteristics directly via spectrographic analysis, but these approaches could not satisfactorily classify sentences relative to human judgments (Duong et al., 2011). Other researchers attempted to create computer programs to simulate expert raters, training a classification algorithm on examples of stronger and weaker prosody in oral reading as scored by humans using a rubric. Bolanos et al. (2013) used 783 recordings of one-minute text passages read aloud by first- through fourth-grade students, to train an algorithm to classify these recordings on the NAEP 4-point fluency scale. The algorithm would first assign a score of 1-4, then would classify each recording as “not-fluent” (levels 1 or 2) or “fluent” (levels 3 or 4). Based on judgments from two expert human judges for each text passage read aloud, the proposed system had an accuracy of 76% on the 4-point NAEP scale and over 90% on the binary scale (“fluent” versus “non-fluent”). Additionally, the agreement between the algorithm and each scorer was significantly higher than the agreement between scorers.

These early attempts to automatically assess prosody, while showing promise, highlighted a crucial challenge in automated assessment: ensuring the system performs consistently across different groups of students. This challenge of fair and unbiased assessment becomes even more critical as automated systems take on greater roles in educational measurement. While automated scoring of various reading measures has shown increasing accuracy, careful attention must be paid to potential biases that could disadvantage certain groups of students.

2.4.3 Potential for Algorithmic Bias

The issue of algorithmic bias in education 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). Harms resulting from algorithmic bias are typically categorized as allocative, involving the withholding of opportunities or resources from certain groups, or representational, where a system diminishes the identity of a particular group (Mitchell et al., 2021). 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).

Some sources of bias are related to insufficient and inaccurate datasets used to train models. 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 large 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, it is 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).

Importantly, biases in educational algorithms not only stem from the training data but also from biases inherent in societal contexts (Tempelaar et al., 2020). 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). To address algorithmic bias in education, various strategies can be employed at different stages of the creation of these models, including data collection, preparation, model development, evaluation, post-processing, and deployment.

2.5 Conclusion

The emergence of large pre-trained language models and improved speech recognition systems has created new possibilities for automating reading assessment. While some skeptics raise valid concerns about NLP-based approaches missing deeper aspects of student writing, these critiques may be less relevant for basic reading comprehension assessment, where the goal is primarily to evaluate understanding rather than complex reasoning. Given the near human-level performance of these models on a variety of benchmark tasks, these findings suggest that it may now be technically viable to automatically administer and score both ORF and open-ended reading comprehension tasks, making it a promising area for further research (McNamara et al., 2010).

However, several crucial questions remain unanswered. Can ASR models accurately transcribe and score oral reading fluency for students in diverse linguistic contexts? How well can large language models evaluate open-ended reading comprehension responses without extensive task-specific training? And critically, can these automated approaches maintain fairness and reliability across different student populations? This dissertation addresses these questions through three interrelated studies examining the potential of AI technologies to support reading assessment in Ghana. By investigating both the technical capabilities and practical

implications of these tools, this research aims to understand how automated assessment might help make regular reading assessment more feasible in low-resource educational settings.

Chapter 3 – Methodology

This chapter outlines the methodological approaches used across the empirical studies in this dissertation, with particular attention to measuring accuracy and agreement between human raters and automated systems. I first explain key statistical measures and approaches used to evaluate assessment reliability, and then detail the development of two novel datasets collected in collaboration with Rising Academies in Ghana. These datasets, which include audio recordings of oral reading, written and spoken responses to comprehension questions, and story retells from 130 Ghanaian students, provide a foundation for investigating how artificial intelligence might support reading assessment in diverse educational contexts.

3.1 Measuring Accuracy and Agreement

Measures of accuracy and agreement are closely related but distinct ways of measuring the performance of ML models and human raters. Accuracy is simply a measure of how often a model's or rater's judgment matches that of a true score (Belur et al., 2018). If accuracy is low, it means the model or rater does not match the predetermined value. Agreement, on the other hand, is a measure of the degree to which two or more raters (or models) concur in their judgment. If agreement is low, it may suggest an inherently ambiguous task, a poorly specified task, or inadequately trained raters (Landis & Koch, 1977). Inter-rater reliability (IRR) is a measure of agreement between the results of judging tasks performed by different raters, also commonly referred to as annotators, judges, observers, or coders (Belur et al., 2018; El Dehaibi & MacDonald, 2020). Intra-rater agreement evaluates to what extent a single rater (or model) agrees with themselves when presented with the same item multiple times.

There are several statistical metrics for quantifying accuracy and agreement. The simplest measure is exact agreement, which merely calculates the proportion of ratings in which the judgements concur. However, this approach does not account for the level of accuracy/agreement that would be expected by chance. A preferred measure of accuracy is the F1 score, which partially accounts for chance by considering both false positives and false negatives (Banerjee et al., 2008). However, in many real-world datasets, the classes are not evenly distributed. For example, in a medical diagnosis test, positive cases might be much rarer than negative cases.

When classes are unbalanced, traditional accuracy metrics can be misleading. A model could achieve high accuracy simply by always predicting the majority class, even though it fails to identify the minority class correctly. F1 score attempts to address this by balancing precision and recall, but it still may not fully capture the nuanced performance across different class distributions. (Banerjee et al., 2008; Belur et al., 2018). In contrast, chance-adjusted metrics of agreement, such as kappa scores, can be used to evaluate both a model's performance against a reference set and the level of agreement between two or more raters. These metrics are particularly useful in handling datasets with uneven class distributions. Referred to as measures of inter-rater reliability, these scores range from -1 to 1. Like correlation scores, values of 1 indicate perfect agreement, values of 0 suggest that the agreement is no better than chance, and values less than 0 indicate that the agreement is worse than chance.

3.1.1 Determining Ground-Truth Scores

To evaluate a model's ability to score student answers to open-ended questions, it is necessary to have a human-generated reference score, commonly referred to as the "ground truth" score. For straightforward tasks with more objective criteria, items can usually be evaluated by a single rater. However, tasks that are more inherently ambiguous or dependent on individual judgment require a more nuanced approach. In such contexts, researchers typically employ multiple raters to mitigate the challenges of subjective assessment. This method involves developing comprehensive instructions, creating detailed rubrics, and providing carefully selected example responses (Nangia et al., 2021). Rater training and calibration are critical steps in ensuring consistency and reliability in the scoring process (Pavlick & Kwiatkowski, 2019).

Despite these methodological precautions, less-than-perfect agreement among raters remains a common challenge in many assessment contexts. The process of establishing ground truth reflects the complex nature of human judgment in educational assessment. When raters disagree, researchers have developed strategies for aggregating individual judgments. These methods typically involve using majority opinion or consensus techniques to generate a definitive score. Such approaches recognize that no single rater's judgment is entirely definitive, and that collective assessment can provide a more robust evaluation (Nangia et al., 2021).

Even in seemingly more objective tasks, such as audio transcription, establishing a ground truth requires careful consideration. Researchers often rely on multiple transcription attempts, with a single rater making repeated careful transcriptions or multiple raters independently producing transcripts. This method has been demonstrated to improve accuracy and can be understood as a form of judgment pooling that draws on the strengths of multiple perspectives. This method has been demonstrated in previous studies to improve accuracy and can be considered a rough approximation of pooling judgements, utilizing the wisdom of crowds (Belur et al., 2018).

However, in educational assessment, the goal is not merely to aggregate opinions, but to develop a systematic method for generating a reliable reference score that captures the nuanced nature of student responses. Importantly, the ground truth is not an absolute or immutable construct. The careful construction of ground truth scores is particularly critical in contexts where assessment involves complex cognitive skills, such as reading comprehension. By employing multiple raters, developing comprehensive scoring guidelines, and using sophisticated aggregation techniques, researchers can generate reference scores that provide meaningful insights into student performance.

3.1.2 Choosing Appropriate Measures

While Cohen's kappa is well-suited for situations dealing with questions of interrater agreement, it relies on the assumption that all the items were evaluated by only two raters (Hallgren, 2012). There are several different measures of chance-adjusted agreement, including Cohen's kappa (for two raters), Fleiss' kappa (for 3 or more raters), the Pearson r correlation coefficient, and Krippendorff's Alpha. When possible, Cohen's kappa is reported, as it is a more widely used and understood metric, and is well suited for scenarios when predictions are being compared against the ground-truth values (Banerjee et al., 2008). When evaluating two-class ratings (incorrect/correct) I used Linear Weighted kappa (LWK), and when evaluating rubric-based scores where the number of categories is 3 or greater, I used Quadratic Weighted kappa (QWK), as the three-classes, incorrect, partially correct and correct, are best understood as an ordinal scale (De Raadt et al., 2021). Krippendorff's Alpha is more a versatile measure because it can be used with both nominal or ordinal data and can handle situations with more than two raters; it can also be used when if the raters do not all rate the exact same same set of items (Krippendorff,

2011). For evaluating agreement between human raters on large and complex datasets, Krippendorff's Alpha is usually a better solution.

3.2 Creation of Novel Datasets

A key part of my dissertation is the collection of two novel datasets comprised of Ghanaian students' answers to a series of literacy assessment tasks. Facilitated by a long-standing professional relationship and potential benefits for their own teaching and assessment practices, Rising Academies, a network of schools in Ghana, collaborated with me in this endeavour. In summer 2022, I administered a series of short literacy assessments to 162 of their students between the ages of 9 and 18. By making the dataset publicly available, I hope to facilitate further research on literacy assessment in this context.

All the assessment items and passages administered were taken from released items from the 2016 PrePIRLS assessment. PrePIRLS is an international comparative assessment that measures the reading ability of fourth graders across 60 countries and is explicitly designed for use in low- and middle-income countries and countries with lower reading levels (*Methods and Procedures in PIRLS 2016*, 2018). Its reading comprehension section uses shorter reading passages and focuses more on direct straightforward comprehension and less on higher-order reading skills compared to the standard PIRLS assessment (see Appendix B for an example).

Before beginning the full assessment, students were asked to start reading a short passage aloud, both as a warm-up/icebreaker, and to identify students who were non-readers. Students who struggled significantly with reading aloud were permitted to stop the assessment to avoid potential negative emotional impact. All 162 students were first given screening tests and the 32 students determined to be non-readers did not proceed to subsequent tasks. The remaining 130 students were asked to read two short passages aloud, each approximately 100 words long (see Appendix A). The oral reading sessions were audio-recorded and later transcribed and marked for reading errors by expert raters.

The remaining 130 students were then asked to silently read one of two 400-word fictional stories and answer six short answer questions, which were primarily related to

information retrieval and direct comprehension (see Appendix B). Of those 130 students, 48 students were asked to read a similar short story silently and answered the 6 reading comprehension questions aloud rather than in writing, with spoken responses audio-recorded and then transcribed by humans. Some students did the oral reading first, while others started with the silent reading comprehension, and a subset of 48 students read a second short story silently and then answered the reading comprehension questions aloud rather than in writing.

Figure 3.1
Assessment Types and Scoring Procedures

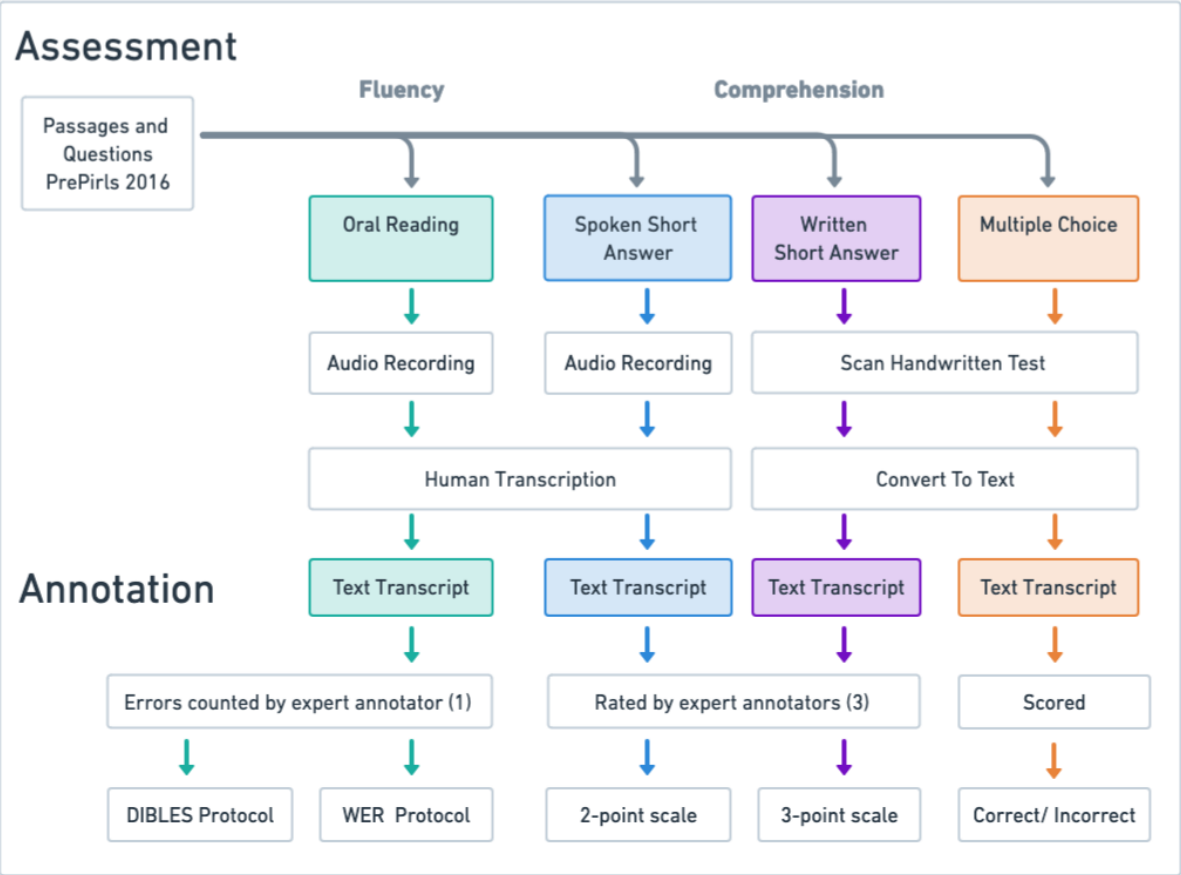


Figure 3.1, above, illustrates items completed by the 130 students, as well as the transcription and grading process. The results of this process were then turned into two, publicly available linked datasets, described below.

Table 3.1
Key Characteristics of ROARS and Ghana ORF datasets

Category	Count
Total Students	162
Student Age: Average, min, max	13, 9, 18
Student English at Home: count (percent)	16 (12%)
Student Female vs Male: count (percent)	86 (65%) / 46 (35%)
Pre-readers vs Readers	32 / 130
Students Attempting Oral Reading Assessment	162
Students Attempting Written Assessment	130
Students Attempting Additional Spoken Assessment	48
Student Completing Retell Task	95
Total Short Answer Responses	1,068
Written SA Responses vs Spoken SA Responses	768 / 300
Total Multiple-Choice Responses	534
Total Retell Responses	95

3.3 ROARS Dataset

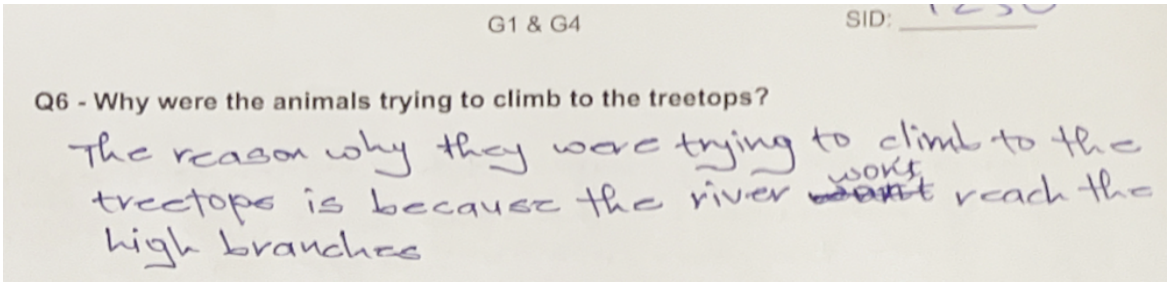
The Rising & Oxford African Reading Comprehension **S**hort Answer Dataset (subsequently referred to as the ROARS dataset). This dataset contains the results of various reading assessment tasks administered to 130 students between the ages of 9 and 18 in Ghana. The ROARS dataset contains the original tests, questions, and over 1,000 student answers to multiple-choice, short answer, and story retell questions. It also contains expert human scores for each student answer.

3.3.1 Scoring Procedure

All the answers were scanned and transcribed by human raters who were all Master of Education students. The raters then independently evaluated students' short answers and student retells, according to various rubrics (see chapters 6 and 7 for more detailed descriptions of the rubrics used). Blank or empty responses were marked incorrect. In cases where all three raters didn't agree, the ground truth scores were determined by averaging human judgments and rounding to the nearest class value. For instance, in the case of the three-class rubric, if one judge rated a response as 0 and two judges rated it as 2, the average score would be 1.67, which would be

rounded to 2, resulting in a ground truth score of 2. For the short answer grading, raters first read the relevant short story and then had access to the relevant passage, the student answer, and some examples of correct and incorrect answers. They were also instructed to ignore small spelling and grammar issues and to determine whether the student's answer demonstrated that they had understood the question and answered it correctly.

Figure 3.2
Example of Short Answer Scoring

<table>
<tr><th colspan="2">Passage</th></tr>
<tr><td colspan="2">“The river is flooding,” said the giraffe. “A wall of water is racing down the valley and will soon be here.” “What can we do?” asked the gazelle. “It’s too late to run away.” “Climb up here,” called the monkey from the treetops. “The river won’t reach the high branches.” The animals raced to the trees. But some of them could not climb up the slippery tree trunks. Their hooves and tails were not made for climbing.</td></tr>
<tr><th colspan="2">Question</th></tr>
<tr><td colspan="2">Why were the animals trying to climb to the treetops?</td></tr>
<tr><th>Accepted Answers</th><th>Incorrect Answers</th></tr>
<tr><td>to stay out of the water; because it was flooding; to escape the river</td><td>to see the valley; to reach the sweet leaves; to hide from the leopard</td></tr>
<tr><th colspan="2">Actual Student Response Score = Correct</th></tr>
<tr><td colspan="2"> 
</td></tr>
</table>

For the retell tasks, raters used three different rubrics: a binary rubric, a three-class rubric, and a five-class rubric. These rubrics were adapted from previous measures in the literature (see more detail in chapter 6). I then examined each rater's individual ratings and calculated both a ground truth score (using a majority vote approach) and an average score for each student response.

Figure 3.3
Example of Retell Scoring

Student Retell			
lucy was a girl who like learning around she round through the country stole that night allways when everyone including the sheeps and lambs were as sleep. lucy helped to save the shepherd when the shepherd got a broke in his leg.			
	2 - class	**3 – class**	**5 - class**
Rater 1	**0**	**1**	**1**
Rater 2	**0**	**1**	**2**
Rater 3	**0**	**1**	**2**
Ground Truth	**0**	**1**	**2**

3.4 Ghanaian Student Oral Reading Fluency Dataset

The University of Oxford & Rising Academies Ghanaian Student Oral Reading Fluency Dataset (subsequently referred to as the Ghana ORF Dataset) consists of the results from an oral reading fluency task, administered to 130 students between the ages of 9 and 18 in Ghana. Students were asked to read a series of short passages aloud, which were audio-recorded and later transcribed by human annotators. To generate the reference transcriptions, the expert annotator listened to the recording while reading along on the source story, monitoring for added or repeated words. During this process, they were able to pause and relisten to audio files. I estimate that this careful transcription took approximately 5 minutes per story. Figure 3.4 shows an example transcript. The Ghana ORF dataset contains the original passages, the student audio clips, and high-quality human transcriptions.

Figure 3.4
Example of Passage and Human Scoring

Original 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	

3.5 CU Dataset

In addition to the novel datasets that I created for this dissertation, I also made use of a pre-existing research dataset. The CU Kids Read and Summarized Story Corpus contains speech recordings and transcripts from 106 children summarizing stories they had read. The data were collected in 2003 by the University of Colorado Boulder researchers. The participants were native English-speaking elementary students in grades 3–5, approximately 8–12 years old. Each child was asked to read a short story and then provide a spoken retelling summarizing what happened in their own words. No specific guidance was given regarding the length or content of the summaries. The children's spoken summaries were audio-recorded and manually transcribed by graduate students. The transcripts intentionally preserved natural speech phenomena like filler words, repetitions, and lack of punctuation. Access to the dataset has been granted by the researchers who collected the data. As per the ethics protocols, no personally identifying information is included in the transcripts.

Figure 3.5
Example of Human Transcriptions

Text Being Read	Human Transcript of Audio
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.	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 fluffy clouds come on sunny days," Tim said.
Audio file of student reading	

3.6 Disclosure Statement

I acknowledge my ongoing professional relationship with Rising Academies. The fieldwork and data collection for this study were completed in the summer of 2022. In 2023, I began working as a consultant with Rising Academies on a separate initiative, helping to design a foundational mathematics tool. It is important to note that I am not being compensated in any way by Rising Academies or other organizations for the work presented in this dissertation. Furthermore, this dissertation does not report on projects for which I am consulting with Rising Academies, nor does it discuss Rising Academies' overall quality. Rising Academies is mentioned only in the context of describing the student population and acknowledging their role in facilitating the original research collaboration. I do not believe there is a conflict of interest, nor any way for the current relationship to affect my judgment or interpretation of the results presented in this work. However, given the ongoing professional relationship, it is important to acknowledge this connection in the interest of full transparency.

Chapter 4 – Can Speech Models Assess Oral Reading Fluency?

4.1 Introduction

Frequent diagnostic and 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). 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 countries, conducting regular literacy assessments can be challenging due to large rural populations, unreliable infrastructure, and weak organizational capacity (Dubeck & Gove, 2015). Given 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.

Previous research has investigated whether Automatic Speech Recognition (ASR) models are able to help automate ORF assessment; however they generally failed to perform satisfactorily on child voice datasets (Jain et al., 2023a). 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 many countries in Sub-Saharan Africa, there are added complications, including 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 speech and multi-modal models (e.g., GPT-4, Whisper v2, and wav2vec 2.0) have demonstrated human-level performance across a range of speech recognition tasks. The growth in the size of training datasets and the number of model parameters appears to have dramatically improved their overall performance. This suggests that large-scale speech models could be deployed to support a variety of educational activities, including helping with assessment of 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.

This study addresses this gap in the literature in three ways. First, I use 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 contains student responses to an ORF task, which was administered to approximately 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 Words Correct Per Minute (WCPM) scores to those produced by expert humans. The main finding — that the best-performing model produces ORF scores similar to those of human raters—has important implications for educational assessment, and particularly for the diagnostic and formative assessment of reading ability. 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—this study aims to better understand the feasibility and limitations of using large-scale speech models for automatic ORF assessment in the anglophone Global South.

4.2 Prior Work

4.2.1 Assessing Oral Reading Fluency

While Oral Reading Fluency can be assessed in various ways, one of the most widely used measures is WCPM (Roehrig et al., 2008). To calculate WCPM, the student is asked to read a passage aloud for a specified period of time, while the assessor follows along, counting reading errors — specifically word deletions or substitutions—but not word insertions or repetitions. There is no theoretical rationale for not counting insertions or repetitions and this decision may have been made pragmatically due to the challenges of scoring by humans in real time (Riedel,

2007; Roehrig et al., 2008). At the end of the allocated time, the student is told to stop, the assessor tallies all the errors, and subtracts that number from the total number of words read to calculate the final score.

WCPM = (total words in story - total errors) x (60 / duration of audio clip in seconds)

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 between a student and a trained rater, making it time-consuming and costly to administer. Therefore, there are potential benefits to using ASR to partially or fully automate ORF assessments (Duong et al., 2011).

4.2.2 ASR in Educational Settings

While ASR had some limited applications in educational contexts 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. 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., 2013; Haller et al., 2022). However, their implementation in real-world educational environments were generally hindered by two main challenges. First, fine-tuning can be a complicated process that demands advanced technical skills, and the acquisition and labelling of the fine-tuning dataset remains an expensive and labour-intensive undertaking (Jain et al., 2023a; Radford et al., 2022). Secondly, models that demonstrate human-level performance when trained on a specific dataset struggled to perform well when tested on a new dataset (Geirhos et al., 2020). 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, recent progress in ASR means that models now can transcribe adult speech at human levels, and there has been increased application in non-adult educational settings (Baevski et al., 2020; Radford et al., 2022).

Research in fine-tuning self-supervised speech model learning has shown improvements in child speech recognition. In one set of experiments, Jain et al. (2023b) explored wav2vec 2.0 with different pre-training and fine-tuning approaches, finding that their models outperformed the unmodified wav2vec 2.0 on child speech. They reported that, with as little as 10 hours of child speech data for fine-tuning, they were able to achieve a WER below 15%, on par with adult speech on well-established student speech corpora, such as MyST Corpus, PFSTAR dataset, and CMU Kids dataset. In a subsequent study, they compared Whisper v2, a large ASR model released by OpenAI, with fine-tuned self-supervised models like wav2vec 2.0. Whisper v2, trained with significantly more data and multilingual resources, was expected to improve child speech recognition through fine-tuning. The results showed significant improvements in performance compared to the non-fine-tuned model, reaching parity with human raters. I consider these results, which are summarized in Table 4.1 below, to be the best results to-date in child speech transcription. However, it remains unclear how ASR models like Whisper would perform in scenarios which combined children's speech, regional accents, and environmental noise (e.g., background conversations).

Table 4.1
Performance of ASR Model on Child-Specific Datasets Measured with WER

	MyST Corpus	**PFSTAR dataset**	**CMU Kids dataset**
wav2vec 2.0 large (10 hours finetuning)	27.17	3.50	21.35
wav2vec 2.0 large (65 hours finetuning)	7.42	2.99	14.18
Whisper v2 Large (unmodified)	25.00	73.68	12.69
Whisper v2 Large (10 hours finetuning)	15.79	2.88	15.22
Whisper v2 Large (65 hours finetuning)	13.34	4.17	17.11

Note: Adapted from Jain et. al. 2023b

4.2.3 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 the MyST Corpus, PFSTAR dataset, and CMU Kids dataset. However, these datasets are typically focused on children's natural speech rather than on oral reading. When assessing ORF, it is critical for transcripts to accurately preserve reading errors, but in normal ASR, the goal of the task is typically to produce representative transcripts. The models themselves are trained on combined audio and transcripts, which leads them to often correct for verbal slips, repetitions, and mispronunciations. While this is a reasonable decision for most ASR applications, in the context of ORF, it is precisely these small discrepancies between the source text and recording 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., H_2O) with a soft "t" in the middle and a 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 my 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.

4.3 Current Study

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, I evaluated the performance of two state-of-the-art speech models, Whisper v2 and wav2vec 2.0, on the Ghana ORF Dataset. By testing the models' ability to accurately transcribe recordings containing regional accents and background classroom noise without any fine-tuning, I 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 for a better understanding of 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, I used the transcriptions generated by Whisper v2 in the first experiment to automatically calculate students' ORF scores using the WCPM metric. I 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. Additionally, because this dataset has demographic information linked to each response, this allows for investigation of whether the models grade students' answers differently from human raters based on students' background attributes such as age, gender, and home language ability.

4.3.1 Rising Academies and Oxford University Oral Reading Fluency Dataset

This dataset comprises a series of literacy assessments conducted with students from Rising Academies, a school network in Ghana. Students aged between 9 and 18 were asked to complete various tasks aimed at assessing their reading abilities. They were first given a quick screening which determined that 32 students were non-readers. 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 (*Methods and Procedures in PIRLS 2016*, 2018). The ORF tasks were audio-recorded and later transcribed by a rater listening to the audio recording while reading along with the source story–paying careful attention to include added or repeated words During this process they were able to pause and relisten to audio files. This careful transcription took approximately five minutes per story. As a quality control measure, I conducted the same process independently on a subset of the audio files, but I found no systematic differences in transcripts.

4.3.2 CU Kids Read and Summarized Story Corpus

The dataset was compiled in 2003 by researchers at the University of Colorado Boulder, who granted me access to the original dataset. 106 elementary students in grades 3–5 were asked to read a short story and then provide a spoken retelling summarizing what happened in their own words. Both the student oral reading and spoken summaries were audio-recorded and manually transcribed by graduate students. The transcripts were over 99% accurate but contain natural speech phenomena like filler words, repetitions, and lack of punctuation.

4.3.3 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 multiplied by 100 for ease of interpretation. Hence, WER scores have a minimum value 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. 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. As discussed above, WCPM is a widely used measure of ORF that combines aspects of reading accuracy and reading rate. 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.

4.3.4 Relationship Between WCPM and WER

From a theoretical standpoint, it is also plausible that there would be a stable and linear relationship between errors calculated using WCPM and WER measurement. This implies that the measure would not differentially impact stronger or weaker readers, making it a fair and objective assessment tool. To evaluate this, I conducted an experiment using a subset of the ORF recordings. A total of 50 audio files were selected across the 5 different stories (See Appendix A). Stories 1, 2, and 3 had 10 examples each; story 4 had 7 examples due to fewer students reading this story and poorer audio quality, and story 5 had 23 examples. The audio files were first transcribed to serve as the gold standard for comparison, with leniency given for accent-affected pronunciation, stuttering, and tester speech.

Next, a rater listened to each of the 50 audio files and calculated the number of errors using both the WER and DIBELS protocols. To ensure consistency, the rater cross-checked their grading against the transcriptions. Surveys were created on Qualtrics with the chosen audio recordings and the correct story scripts, which had two versions for each story: one with DIBELS instructions and another with WER. A minimum of five individuals participated per survey with a goal of obtaining at least three 'useful' grades per recording. Submissions were rejected if participants went too fast or provided ratings inconsistent with those of other participants.

Upon analysis, both crowdworkers and the expert rater found more errors using the WER rules compared to the DIBELS rules, which was expected. On average, the expert rater identified more errors per story than the crowd workers (7 and 4.8 errors, respectively). A t-test revealed that the difference between my grading and the average ratings of the raters was statistically significant. When examining the crowdworkers' results specifically, a higher total variance in error counts was observed when using the WER protocol compared to DIBELS. This is not surprising, as the WER rules allow for more errors to be counted, resulting in more opportunities for disagreement.

4.4 Experiment 1 - Evaluating Speech Models' Performance on the Ghana ORF Dataset

To better evaluate the performance of two state-of-the-art open-source ASR models, Whisper v2 large and wav2vec 2.0, I conducted a comparative analysis using a subset of the Ghana ORF Dataset. I selected 60 audio files based on their having higher audio quality and the student having completed reading the story (approximately 30 students were allowed to stop the task early when it became apparent they were functionally non-readers). This curated subset represented a best-case scenario for evaluating the underlying model performance, as the audio clips generally had better audio quality and more fluent student readers. It is important to note that I did not fine-tune or modify these models, using them "out-of-the-box" for this evaluation. Figure 4.1 below shows a comparison of the original text, the human transcript, and the model generated transcript. Highlighted areas denote discrepancies between the original text and what the student read.

As can be seen in Figure 4.1, the human and model transcriptions—while largely similar—have some differences. Because the expert-human transcript is considered the ground truth in this case, I calculated the WER for Whisper v2 and wav2vec 2.0 in reference to the human transcript of what the student read (rather than text of the story). To calculate WER, I standardized all three transcriptions by removing punctuation and converting the text to lowercase. I then calculated the WER for the Whisper v2 and wav2vec 2.0 transcriptions relative to the human transcription using JiWER, a simple and fast Python package designed to evaluate ASR systems (*JiWER*, n.d.). Results of this analysis are summarized in Table 4.2 below.

Whisper v2 achieved a WER of 10.3, which is typically considered strong performance when transcribing real-world speech. 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. 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 (Radford et al., 2022).

Figure 4.1

Example of Passage and Human Scoring

Original Text (122 words)	
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	
Audio file of student reading	
Human generated transcript (differences in yellow)	**wav2vec 2.0 generated transcript** (differences in blue)
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	a farmer went out one day to search for a lost scarf he went to the valley and serched by the riverbed among the reeds behind the rocks and in the russian water he climbed the slopes of the high mountain with its rocky cliffs he locked he looked behind a large rog in a case the calf had head headed there to escape the storm and that was the and that was where he stopped there, on a ladge ledge of rog, was a most usual sit 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
Human - 15 Student reading errors	Model - 17 Student reading errors

Table 4.2

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

Manually inspecting the differences between how the models 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 similar to how an average American might say the word "Russian". Other examples include "rog" for "rock" and the pause between "river" and "bed".

Figure 4.2
Alternate Transcripts of Same Audio Clip

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 rock in 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 and serched by the river bed among the reeds behind the rocks and in the russian water he climbed the slopes of the high mountain with its rocky cliffs he locked he looked behind a large rog in a case the calf had head headed there to escape the storm and that
Audio file of student reading	

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 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.

These results are more notable given the historical challenges of transcribing children's speech and regional accents and suggest that ASR models may have reached a level of accuracy sufficient for use in educational applications. To better understand this potential, I next investigated whether the model's performance varies depending on students' accents, and

subsequently explore whether the WER is low enough to generate reliable estimates of students' ORF.

4.5 Experiment 2 - Evaluating Speech Models' Performance on CU Dataset

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). The recordings in the Ghana ORF Dataset differ from many other children-specific ASR datasets in that the recordings were made in classrooms, students spoke English as a second (or third) language, and the local accent is distinct from many North American accents. Consequently, I wanted to better understand whether the accuracy of model transcriptions was worse than for North American students' ORF recordings. To do so, I 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 data were collected in 2003 by researchers at the University of Colorado Boulder. The participants were English-speaking elementary students in grades 3–5, approximately 8–12 years old. This dataset includes high-quality transcripts of students' oral reading that critically preserve reading errors, and most of the students lacked strong regional accents (example of student reading).

I selected 45 audio recordings, representing a mix of age and ability levels, and used Whisper V2—the strongest performing ASR model from Experiment 1—to generate transcriptions. I then calculated the WER of the Whisper v2-generated transcript relative to the high-quality human transcripts, which I considered the ground truth. Comparing the WER of Whisper v2 on the CU Dataset with the Ghana ORF dataset, it is possible to understand the potential impact of accent on model performance.

Table 4.3
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

As shown above in Table 4.3 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. This contrasts with previous studies that reported large gaps in WER, such as Koenecke et al. (2020) who reported a gap in WER of over 20% between Black and White speakers. This 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, I consider these results provisional and believe that this area can benefit from continued research.

4.6 Experiment 3 - Automatically Calculating WCPM From a Transcript

Experiments 1 and 2 demonstrate the rapid improvement of ASR performance, however, this might not translate directly to improved performance in automatic ORF assessment. To explore the real-world viability of this approach, in this experiment, I generated transcripts for a larger portion of the Ghana ORF Dataset, and then take these transcripts and use an edit-distance algorithm to automatically calculate WCPM scores.

To expand the range of recordings transcribed, I used both the original set of 60 audio files used in experiments 1 and 2 and an additional 95 recording, with more background noise and weaker readers, which allowed me to evaluate the model's robustness in more challenging scenarios. I took the 165 transcriptions generated by Whisper v2 and used the JiWER Python package to calculate the WER between the transcript and the original story text. To get a sense of the performance, I compared 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 should not be seen as transcription errors, but instead as accurately capturing student reading mistakes. Hence, for this question, I am looking to see how similar the respective WERs are between the human transcripts and the model transcripts.

Table 4.4
Human vs Model Generated Estimates of Student Reading Errors

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

Table 4.4 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. These results indicated that Whisper v2 is producing broadly similar but not identical transcriptions to those 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 automatically generated WCPM scores, I divided the number of correct words in the model transcripts by the length of the audio files, according to the following formula:

*Model WCPM = (total words in story * (1-WER)) * (60 / duration of audio clip in seconds)*

To calculate human generated WCPM scores, I took the high-quality human transcriptions of these audio files and manually counted the student's reading errors according to the standard WER protocol (insertions, deletions, and substitutions), subtracted the number of errors from the

total story length, which was then divided by the duration of the audio clip in seconds and multiplied by 60.

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

Because the other values in WCPM (length of story and time spent reading) are objective, any discrepancies between expert human raters and the model estimates will be attributable to either the transcription itself or the error calculation rate. This allowed me to compare the automatically generated scores to those made by expert human raters. Table 4.5 reports the agreement between the automated and human-generated scores.

Table 4.5
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	

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 typically considered small in the context of ORF tests (Fuchs et al., 2001). Further, automated scores and the expert human scores were highly correlated (.98), suggesting that the same students were receiving similar scores relative to one another, which is arguably the most important measure in terms of practical implications.

Taken as a whole, I 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 exactly match expert-generated scores, this discrepancy may not significantly impact its

practical utility. WCPM is inherently a proxy measure of reading ability that typically shows natural variation even when assessing the same student multiple times. Additionally, since ORF is primarily used as a formative or diagnostic tool rather than for summative assessment, small measurement errors are less likely to have negative consequences for students. 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.

4.7 Analysing Potential Bias

While bias and fairness in AI models in educational contexts have been actively debated (see Baker & Hawn, 2022 for a comprehensive review), their definitions remain complex and often contested. I operationalize bias as a situation where the model exhibits a higher error rate when transcribing speech of students from specific socioeconomic or demographic groups compared to human raters. 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. In this analysis, I examine the differences between model estimates and human estimates of both WER and WCPM across two salient demographic groups: gender and native language status.

Table 4.6

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

In both scenarios, I 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, I treat these results with caution and believe that these types of checks should be conducted regularly and more extensively.

It is important to note that this 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, this is at least partially mitigated by the relatively objective nature of the task (identifying oral reading errors from audio recordings) and rigorous coding procedures, which involved listening to each clip multiple times. Nevertheless, a thorough examination of potential rater bias is an important and complex question that merits further research but falls outside the scope of this study.

4.8 Discussion

4.8.1 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—a task that 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 as 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 these findings are provisional, they suggest that newer ASR models may better handle diverse accents and point 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 to those of expert human raters. The study found a high correlation (0.98) between automated WCPM scores and those generated by expert human raters, as well as a small absolute average difference in the scores.

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.

4.8.2 Limitations and Further Research

While this study provides valuable insights into the potential of ASR models for the automated assessment of ORF, it is not without limitations. First, the study was conducted using only two specific ASR models, Whisper v2 and wav2vec 2.0. While these models are considered state-of-the-art, the findings may not be generalizable to other ASR models. Second, 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.

One key direction for further investigation is the development of datasets that represent the wide range of accents and languages found in many educational contexts. By training and testing speech models on these datasets, their generalizability can be improved, ensuring that they can accurately assess ORF across various populations. Another important area for continued exploration is 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. By making this dataset available for research purposes, other researchers can contribute to this important topic.

4.9 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 focused particularly 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 utilized two state-of-the-art ASR models, 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 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 larger context of diagnostic and formative assessment, these findings indicate that ASR could assist in the assessment of student work across various contexts. These models' demonstrated ability to accurately transcribe student speech 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.

Chapter 5 - Can LLMs Grade Short Answer Reading Comprehension Questions?

5.1 Introduction

The potential applications of Large Language Models (LLMs) in education are myriad, and grading student work in low-stakes situations, such as diagnostic and formative assessment, is a particularly promising area. Diagnostic and formative assessments can help students monitor their progress and provide teachers with a deeper understanding of student knowledge and reasoning (Black & Wiliam, 2009; Shute, 2008). Closed-response question formats such as multiple-choice and true/false, while efficient to grade, can be time-consuming to create and have other limitations, including being more indicative of test-taking strategies rather than actual comprehension, and in some cases, may lack face validity (Bellinger & DiPerna, 2011; Landauer et al., 2009). In contrast, open-ended formats such as short answer questions require students to answer in their own words and can evaluate more integrative skills such as communication, reasoning, and analysis (Morjaria et al., 2024; Shapiro et al., 2014). However, students' responses to short answer questions can be challenging to score consistently and are more time-consuming for teachers to mark, which limits their widespread usage, particularly in diagnostic formative assessment (Magliano & Graesser, 2012).

Automatic short answer grading (ASAG) is a longstanding research area, which aims to develop computational methods for evaluating student responses to open-ended questions. Early ASAG models employed statistical or domain-specific neural network approaches but struggled to generalize across different domains and contexts, requiring painstaking construction for each specific task (Burrows et al., 2015). In recent years, transfer learning, a technique involving adapting a model trained on one Natural Language Processing (NLP) task for use on another related task, has gained traction and demonstrated dramatic performance across various NLP tasks including ASAG (Brown et al., 2020; Haller et al., 2022). By leveraging the transfer learning paradigm, where a pre-trained model is fine-tuned on a specific task using a labelled dataset, researchers have been able to achieve high levels of accuracy across a wide range of ASAG tasks (Fernandez et al., 2023; Sung et al., 2019). For instance, Fernandez et al. (2023) achieved an agreement score with expert raters of 0.84 (Cohen's kappa) on evaluating open-

response reading comprehension questions, which was close to the human-to-human agreement of 0.88.

Despite this progress, there remain two major hurdles limiting the usage of ASAG in real-world education settings. First, the process of adapting these models requires substantial technical knowledge, and second, assembling and annotating a sufficiently large training dataset. Additionally, these transfer learning approaches frequently encounter difficulties maintaining consistent performance when applied to contexts beyond their original training scope, even for seemingly related tasks (Bommasani et al., 2022; Raffel et al., 2020; Sultan et al., 2022). This lack of generalizability limits the practical benefits of using transfer learning models, as they cannot be easily adapted and deployed across different educational contexts.

More recently, a new generation of LLMs (e.g., GPT-4, Claude, Gemini) commonly referred to as "generative LLMs" have exhibited improved accessibility. These models enable users without technical backgrounds to interact through natural language instructions (commonly referred to as prompting), while demonstrating remarkable flexibility in tackling diverse tasks, frequently requiring only minimal example demonstrations (Kojima et al., 2022; Ouyang et al., 2022). These generative models have been designed to generate text based on the input provided, rather than being specifically trained for classification tasks. However, these LLMs have shown evidence of being able to perform classification tasks exceptionally well through the use of few-shot prompting, where the model is given a few examples as part of the system prompt (Cohn et al., 2024; Kortemeyer, 2023; Ouyang et al., 2022). The ability to adapt a model to new classification tasks (i.e., grading) with minimal training data and technical overhead could provide a more accessible and efficient way to evaluate student understanding and provide targeted feedback.

However, there are also important concerns about the use of LLMs to assist in the marking of educational assessments. First, there are questions as to whether the LLMs' judgments align with those of human experts who typically mark student responses (Caines et al., 2023; Weidinger et al., 2022). Secondly, the non-deterministic nature of these models means that they may not always produce the same output for a given input, which can potentially

undermine the reliability of the assessment process, making it difficult to compare student performance across different time points or cohorts. Thirdly, LLMs are trained on vast amounts of text data that may contain inherent biases and stereotypes, which could potentially influence the model's grading decisions and lead to unfair or discriminatory assessment of student work (Baker & Hawn, 2022). To date, there has been very little empirical research on the use of generative LLMs to mark short answer questions, highlighting the need for further investigation to better understand their potential impact and limitations.

The current study addresses the limited amount of empirical research on the use of generative LLMs for educational assessment. To do so, I use the **R**ising & **O**xford **A**frican **R**eading Comprehension **S**hort Answer Dataset (subsequently referred to as the ROARS dataset). This dataset contains the results of various reading assessment tasks administered to 130 students between the ages of 9 and 18 in Ghana. The ROARS dataset contains the original test, over 1,000 student answers to reading comprehension questions, and scores generated by expert human raters for each answer. Importantly, the fact that this dataset was developed in collaboration with an education organization in Ghana helps to broaden the scope of research on AI in education to include underrepresented groups. This dataset can serve as a resource for future researchers investigating the potential of LLMs in educational contexts in low and middle-income countries (LMICs). Additionally, this paper provides the first empirical research on the ability of the most recent generation of generative LLMs to grade short answer reading comprehension questions. The main finding, that the newest generation of LLMs performed extremely well, exceeding previous benchmarks and approaching human-level performance, has implications for educational assessment in general and the assessment of reading comprehension in particular.

5.2 Prior Work

The key rationale for ASAG is that while instructors often prefer the type of information they can get from open-ended questions, grading student responses individually is a laborious process, potentially resulting in overburdened educators and compromised feedback quality, negatively affecting students' comprehension and capacity for critical engagement with the subject matter (Matelsky et al., 2023). As a result, some instructors opt for simpler formats like multiple-choice questions, which provide instant feedback but lack personalized and insightful remarks (Burrows

et al., 2015; Magliano & Graesser, 2012). If, however, short answer questions could be graded automatically, it would be possible for teachers to get the rich feedback they need to best inform instruction without creating an unsustainable workload.

5.2.1 Prior Approaches for ASAG

Automatic short answer grading (ASAG) has been an active area of research for over a decade, and it is beyond the scope of this study to cover it in detail. Burrows et al. (2015) provide a comprehensive overview of approaches up until 2015, while Haller et al., (2022) discuss more recent developments up to - but not including - the debut of the latest generation of LLMs. As noted in Haller et al. (2022), over the last few years there has been a move from models based on handcrafted features to models using word-embedding and/or representation-learning approaches. This approach enables the model to adapt to the specifics of the target task and domain using only a few hundred labelled examples, as opposed to the thousands of examples that would have been required using previous approaches (Brown et al., 2020). However, dependence on prompt-specific training data often meant that it was necessary to re-train the model for each individual short answer prompt, which was costly, time-consuming, and in most cases simply infeasible. Even in situations where a school may have access to such expertise and can develop such a model, a labelled dataset of a few hundred examples of correct and incorrect student answers is required. Furthermore, the model must be fine-tuned each time the question changes (Mishra et al., 2022; Ye et al., 2021).

Prior research on ASAG and Automatic Essay Scoring (AES) has shown that prompt-specific fine-tuning, while effective for individual prompts, performs poorly in cross-prompt settings. In other words, a system trained using only responses from one prompt may not generalize well to answers written in response to another prompt, even when the prompts are similar in nature (Alikaniotis et al., 2016; Ridley et al., 2020). For example, in the context of the ASAP dataset, which contains short student essays (150-500 words) in response to eight different questions, models achieved high performance grading responses to one prompt - reading an excerpt and explaining why the author chose a specific conclusion. However, these models were not able to successfully grade responses to a similar but distinct prompt - reading an excerpt and describing the mood created by the author (Ridley et al., 2020). This further indicates that

transfer learning-based approaches to ASAG are unlikely to be practical for classroom-based assessment, including diagnostic and formative assessment.

5.2.2 Potential of Generative LLMs for ASAG

The more recent generation of Large Language Models (LLMs) such as GPT-4, Claude, Llama, Mistral, Gemini can interpret human-written natural language instructions, enabling users to adapt a model to new tasks simply by modifying the prompts they pass to the language model, rather than requiring further training or fine-tuning. Generative LLMs have exhibited the ability to perform various linguistic tasks that previously required the use of task-specific, fine-tuned LLMs (Kojima et al., 2022; Wei et al., 2022). In the specific case of using generative LLMs for grading, instead of fine-tuning a custom head for a pre-trained LLM using a task-specific dataset, a user can now simply write an explanation of how they wish the model to classify (i.e., grade) student answers, typically accompanied by a few examples.

While there has been a fair amount of work on using LLMs for scoring essays, there is relatively little research on using generative LLMs to evaluate short answer student responses (Kortemeyer, 2023; Mizumoto & Eguchi, 2023; Schneider et al., 2023). Morjaria et al. (2024) found that ChatGPT performed similarly to a single expert rater when marking short answer assessments in an undergraduate medical program, but their study only marked 6 distinct questions with 10 student answers each. A review by Schneider et al. (2023) concluded that "while 'out-of-the-box' LLMs provide a valuable tool to offer a complementary perspective, their readiness for independent automated grading remains a work in progress." Cohn et al. (2024) explored short answer responses to formative assessments in high school science and found that GPT-4 successfully scored student answers using a human-in-the-loop approach combining few-shot and active learning with chain-of-thought reasoning. Kortemeyer (2023) found that while LLMs could be useful for preliminary grading of introductory physics assignments, they fell short in certain aspects. However, in all the above cases, the studies were conducted with a small sample size of student answers, typically in the range of dozens. Additionally, the focus of these studies was primarily on high school and university students, and there was much more limited exploration of short answer responses for students in elementary or middle school. To help fill

this gap in the literature, this study presents a novel dataset and my own empirical analysis of LLMs' ability to grade student short answers.

5.2.3 Pre-Existing Short Answer Datasets

There are many publicly available datasets commonly used for evaluating ASAG models as well as for studying students' reading ability. However, there are far fewer that contain reading comprehension short answer questions, especially ones that contain data from students from lower-income and/or underrepresented populations. These limitations are the main motivation behind the proposed dataset. Some of the most well-known datasets include ASAP, which provide a broad range of subjects, from the sciences to reading comprehension, but are tailored to high school students in the US. The CREG and CREE datasets concentrate on language learning data rather than literacy. Other datasets, such as CS and CSSAG, focus on niche subjects like computer science at the university level.

As can be seen in Table 5.1 below, the datasets remain predominantly focused on North American and European contexts, and lack data that focus on foundational educational skills, such as basic reading comprehension. This research attempts to fill the gap by creating a dataset containing short answer responses to reading comprehension questions answered by students in Ghana. Additionally, the new dataset stands out by providing a detailed demographic breakdown, which is often absent in other datasets.

Table 5.1
Summary of Publicly Available Short-Answer Datasets

Corpus	Subject	Language	Question Type	Answers	Additional Info
ASAP	Sciences, biology, reading comprehension	English	Single-prompt essay questions	1,300	Collected from Grades 7-10 in the USA
ASAP-DE	Sciences	German	3 prompt essay questions	903	Answers crowdsourced
Beetle	Physics	English	3 prompts given for short-answer questions	3,000	Completed by native English speakers
CESA	Physics and computer science	Chinese	3 short answer questions	1,800	Collected from university students
CREE	Reading comprehension for language learning	English	62 prompts given for short-answer questions	566	Participants are language learners, age unknow.
CREG	Reading comprehension for language learning	German	177 reading comprehension questions	1,032	Collected with Universities in the US. Participants are language learners, age unknown
CS (Mohler)	Computer science	English	21 short-answer questions	630	Collected from computer science students in the US
CSSAG	Computer science	German	21 short-answer questions	1,926	Collected from mostly native German-speaking university students.
Mindreading	Explaining behavior	English	10 short answer questions	11,311	Collected from 7-11 year old students
Powergrading	US Immigration exams	English	10 prompts to short-answer questions	6,980	Collected by the US for immigration purposes. Population unclear
PT_ASAG	Biology	Portuguese	15 prompts for short-answer questions	3,675	Collected from students aged 12-14 years
SciEntsBank	Science questions	English	197 questions in 15 science domains.	10,000	Collected as part of the Student Response Analysis corpus

Note: ASAP: Automated Student Assessment Prize; ASAP-DE: Automated Student Assessment Prize - Deutsche Edition, CESA: Chinese Evaluation for Short Answers, CREE: Content-based Reading comprehension Exercise in English; CREG: Content-based Reading comprehension Exercise in German;. CSSAG: Computer Science Short Answer Grading; PT_ASAG: Portuguese Automatic Short Answer Grading; SciEntsBank: Science Entailment Bank

5.3 Current Study

The goal of this study is to evaluate the performance of generative LLMs and various prompting strategies for grading on a novel dataset, the **R**ising & **O**xford **A**frican **R**eading Comprehension **S**hort Answer Dataset (subsequently referred to as the ROARS dataset), which contains the results of a reading comprehension assessment taken by approximately 150 students between the ages of 9 and 18 in Ghana. As discussed above, there is limited empirical work evaluating the use of generative LLMs for ASAG. While the models appear to be extremely capable with classification tasks, they were not explicitly trained for them, making rigorous testing of their capabilities both under-researched and practically important. If generative LLMs demonstrate high accuracy and reliability, and do not introduce unintentional bias into the marking process, this would have important practical implications. This is primarily because generative LLMs are relatively straightforward to use and do not require the technical expertise and medium-sized datasets necessary for task-specific fine-tuning approaches, such as transfer learning, meaning they could realistically be used at the school and/or classroom level.

5.3.1 Human Grading of Student Responses

In total, there are 1,068 student short answer responses to reading comprehension questions. The passages, stories, and questions were all released items from the 2016 PrePIRLS (see Appendix A). Of the 1,068 short answer questions, 87 were left blank by students and were automatically scored as incorrect; they have been left in the dataset so that student-level accuracy rates can be calculated. The remaining 981 are scored on a binary rubric (correct | incorrect) and a three-class rubric (correct | partially correct | incorrect), by two human raters, both of whom were graduate students at the University of Oxford.

Figure 5.1

Examples of Student Answers and Ratings

<table>
<tr><th colspan="3">Passage</th></tr>
<tr><td colspan="3">“The river is flooding,” said the giraffe. “A wall of water is racing down the valley and will soon be here.” “What can we do?” asked the gazelle. “It’s too late to run away.” “Climb up here,” called the monkey from the treetops. “The river won’t reach the high branches.” The animals raced to the trees. But some of them could not climb up the slippery tree trunks. Their hooves and tails were not made for climbing.</td></tr>
<tr><th colspan="3">Question</th></tr>
<tr><td colspan="3">Why were the animals trying to climb to the treetops?</td></tr>
<tr><th colspan="3">Student Answer 1</th></tr>
<tr><td colspan="3">The reason why they were trying to climb to the treetops is because the river ~~wasn't~~ won't reach the high branches</td></tr>
<tr><th colspan="3">Rater Assigned Scores</th></tr>
<tr><td></td><th>2 - class</th><th>3 - class</th></tr>
<tr><td>Rater 1</td><td>correct</td><td>partially correct</td></tr>
<tr><td>Rater 2</td><td>incorrect</td><td>incorrect</td></tr>
<tr><td>Rater 3</td><td>correct</td><td>partially correct</td></tr>
<tr><td>Ground Truth</td><td>correct</td><td>partially correct</td></tr>
<tr><th colspan="3">Student Answer 2</th></tr>
<tr><td colspan="3">* the flooded river was splashing around the animals.</td></tr>
<tr><th colspan="3">Rater Assigned Scores</th></tr>
<tr><td></td><th>2 - class</th><th>3 - class</th></tr>
<tr><td>Rater 1</td><td>correct</td><td>correct</td></tr>
<tr><td>Rater 2</td><td>correct</td><td>correct</td></tr>
<tr><td>Rater 3</td><td>correct</td><td>correct</td></tr>
<tr><td>Ground Truth</td><td>correct</td><td>correct</td></tr>
</table>

The raters first read the relevant short story and then had access to the relevant passage, the student answer, and some examples of correct and incorrect answers. They were also

instructed to ignore small spelling and grammar issues and to determine whether the student's answer demonstrated that they had understood the question and answered it correctly. In the case of the three-class rating, they were told the following: "*An answer is wrong if it definitely does not answer the question correctly. An answer is partially correct if it gets part of the answer but leaves out important information. An answer is correct if it answers the question and includes the key information.*"

To establish ground truth scores against which the model's performance would be evaluated, I also rated the questions. While most answers had unanimous agreement (approximately 90%), in cases where the first two raters disagreed, the third rater's score was used to cast a tie-breaking vote, which was used as the ground truth for whether a student answer was correct or incorrect. These ratings were paired with the students' answers, and background demographic information to produce the final dataset.

5.3.2 Model Selection and Task Design

The experimental part of this paper focuses on five interconnected research questions, including: (1) How well does the best-performing LLM mark student short answer responses relative to human raters? (2) What types of straightforward prompting strategies produce the best results? (3) How large is the performance drop between the top-performing model and the previous generation of models? (4) Do the models demonstrate any converging behaviours in relation to intra-rater reliability (i.e., does the model grade a given response the same each time)? (5) Does the model demonstrate any differential performance based on students' gender, age, or home language?

At the time of writing, the strongest performing model publicly available across a variety of benchmarks was GPT-4-0613. Hence, this model was selected as the upper benchmark. To test the performance of a previous generation model and better understand recent performance improvements, GPT-3.5 turbo was also used. To evaluate both the broader effectiveness of generative LLMs relative to transfer learning approaches and to examine how model scale and prompt design choices influence performance, comparative testing was conducted across four GPT variants. I configured all models with a temperature setting of zero to minimize output variance, following best practices for classification tasks. The evaluation protocol involved

presenting each model with a combination of passage, question, and potential answer, requiring it to determine the correctness of the candidate response. The model was shown only the passage, question, and student answers; it was never shown the ground-truth answer or any student demographic information.

5.3.3 Prompting Strategy

A relatively simple prompting strategy was employed to simulate what might be realistic for a teacher to create to help them mark a couple hundred assessment questions. The prompting strategy included a system prompt (see figure 5.2 below) that gave explicit instructions about the general task and a user prompt that provided the specific passage, question, and answer to evaluate. The system prompt tasked the model with deciding if the student's answer was correct or incorrect. In the case of the three-class evaluation, I provided the model with a slightly more detailed rubric. I also experimented with few-shot prompting, by providing the model with two examples of correct answers and two examples of incorrect answers within the prompt itself. These examples were subsequently removed from the dataset when calculating QWK.

Figure 5.2

System Prompts Used in Experiment

System Prompt - 2 Class	System Prompt - 3 Class
You are evaluating students' reading comprehension. Students have been asked to read a short story and answer reading comprehension questions. You will be given a story, a question, and a student response. Your task is to classify student responses as correct or incorrect. You will be given a story, a question, and a student response. If the student's response is correct you will respond “1”. If the student's response is wrong you will respond “0”.	You are evaluating students' reading comprehension. Students have been asked to read a short story and answer reading comprehension questions. You will be given a story, a question, and a student response. Your task is to classify student responses as wrong, partially correct, or correct. An answer is wrong if it does not answer the question correctly. An answer is partially correct if it gets part of the answer, but leaves out important information. An answer is correct if it answers the question and includes the key information. If the student's response is wrong you will respond “0”. If the student's response is partially correct you will respond “1”. If the student's response is correct you will respond “2”.

5.4 Results

5.4.1 Model Performance

Tables 5.2 and 5.3 below report the model performance across the eight pertinent combinations. As mentioned above, 87 of the 1,068 answers were left blank by students. While these blank responses are considered incorrect answers in the context of the assessment, and all version of the LLMs correctly identified all 87 of them as incorrect, I have removed them from the below analysis, as well as from calculations of human agreement. This is because identifying a completely blank answer as incorrect is a rule-based decision, and leaving them in the evaluation set would inflate estimates of agreement by 0.1-0.3. The results reported below are relative to the human-generated ground truth answer, so a score of 1.0 would indicate that the model gave the same answer as the human grader for all the questions. As discussed above in Section 3.3, while precision, recall, and F1 scores offer valuable insights into model performance, the preferred holistic measure of model performance is the kappa score. In Table 5.2, which reports on the two-class score, I use Linear Weighted kappa (LWK), whereas in Table 5.3, which reports on the three-class score, I use Quadratic Weighted kappa (QWK). In both tables, the precision, recall, F1, and kappa scores represent model performance relative to the human-generated, ground truth labels.

GPT-4 consistently outperformed GPT-3.5 on kappa and F1 scores in both the two-class and three-class conditions. I also found that few-shot prompting improves performance across conditions, which is consistent with prior research on prompting strategies across a variety of domains (Ye et al., 2021; Perez et al., 2020). The top-performing combination was few-shot GPT-4, achieving a LWK of 0.87 in the two-class condition and a QWK of 0.91 in the three-class condition. Another noteworthy trend is that the models' performance with partial credit scores was consistently worse than their performance on fully correct or fully incorrect answers, suggesting that the models struggle more with identifying partially correct responses compared to the other two categories.

Table 5.2
Performance of Four Varieties of GPT Models on two-class Task

	Model Prediction	Precision	Recall	F1	LWK
	Incorrect	0.82	0.76	0.79	
basic-zeroshot-3.5	*Correct*	0.91	0.93	0.92	0.71
	Average	0.88	0.88	0.88	
	Incorrect	0.79	0.80	0.80	
basic-fewshot-3.5	*Correct*	0.92	0.92	0.92	0.72
	Average	0.88	0.88	0.88	
	Incorrect	0.80	0.99	0.89	
basic-zeroshot-4	*Correct*	1.00	0.90	0.95	0.83
	Average	0.94	0.93	0.93	
	Incorrect	0.83	0.99	0.91	
basic-fewshot-4	*Correct*	1.00	0.92	0.96	0.87
	Average	0.95	0.94	0.94	

Table 5.3
Performance of Four Varieties of GPT Models on three-class Task

	Prediction	Precision	Recall	F1	QWK
basic-zeroshot-3.5	*Incorrect*	0.87	0.43	0.57	0.69
	Partial	0.22	0.71	0.33	
	Correct	0.93	0.80	0.86	
	Average	0.84	0.69	0.73	
basic-fewshot-3.5	*Incorrect*	0.79	0.76	0.77	0.73
	Partial	0.20	0.10	0.14	
	Correct	0.85	0.93	0.89	
	Average	0.77	0.80	0.78	
basic-zeroshot-4	*Incorrect*	0.84	0.98	0.91	0.86
	Partial	0.33	0.49	0.39	
	Correct	0.96	0.83	0.89	
	Average	0.87	0.84	0.85	
basic-fewshot-4	*Incorrect*	0.89	0.98	0.93	0.91
	Partial	0.39	0.39	0.39	
	Correct	0.95	0.91	0.93	
	Average	0.88	0.87	0.87	

Finally, the improvement in performance from GPT-3.5 to GPT-4 is particularly striking, with GPT-4 consistently outperforming its predecessor across all metrics and classification tasks. A closer examination of the results reveals that the performance jump from GPT-3.5 to GPT-4 seems to have mainly come from more accurately grading incorrect answers, as evidenced by the much larger increases in F1 scores for the incorrect class between the two model versions. This suggests that GPT-4 has developed a better representation of what constitutes an incorrect response, leading to improved overall performance.

5.4.2 Interpretation

In terms of interpreting the strength of these results, there are a few potential approaches. The first and perhaps most general approach is to use established ranges considered to indicate a satisfactory level of agreement. While there are several limitations with this approach, kappa scores above 0.5 are considered meaningful, scores above 0.6 are considered substantial, and some consider scores between 0.67 and 0.80 suitable for drawing provisional conclusions (Landis & Koch, 1977). A slightly better approach is to compare model performance against similar models' performance on tasks that are similar in type (i.e., short answer), or domain (i.e., reading ability). This has the benefit of being able to make a direct comparison but has the shortcoming of not accounting for the fact that different tasks may have different levels of complexity, and that QWK, while it attempts to adjust for the number of classes, typically produces higher scores for tasks with fewer classes. One close comparison would be the results on the NAEP reading comprehension challenge discussed in Section 2.3.

Another potential point of comparison is the degree of inter-rater agreement between expert raters when evaluating student reading fluency for other items on NAEP. Various studies have reported inter-rater reliability ranging between 0.74 and 0.82 on this task (Smith & Paige, 2019). Using either approach, the QWK of 0.91 in the three-class condition and LWK of 0.87 in the two-class condition would be considered strong results. Another way to evaluate model performance is to compare it to the rate of agreement among human raters. As discussed more extensively in section 3.3, the intuition is that the human inter-rater agreement effectively sets a ceiling on the accuracy I could expect from a model.

Table 5.4
Performance Across Different Approaches to Scoring Responses

	GPT-4 Few-Shot	**Expert Human Raters**
***3 - Class** (QWK)*	*0.91*	*0.90*
***2 - Class** (LWK)*	*0.87*	*0.93*

The model's performance is strong here as well; its agreement level with the ground truth answers exceeded the agreement level of expert human raters in the case of a three-class judgment. To my knowledge, this is the first reported result of an ASAG model matching or

exceeding human agreement levels when evaluating open-response reading comprehension questions. In sum, these results can be considered as strong model performance. Furthermore, because the ROARS dataset was recently compiled and was not previously publicly available, the results would not have been part of prior GPT training runs, which reduces the potential of the LLM having "learned" the correct answers.

5.4.3 Evaluating Model Reliability

Feature-based models and fine-tuned LLMs are deterministic, always producing the same output for a given input. In contrast, Generative LLMs can produce varying outputs even when given identical inputs. While setting the model temperature to zero should theoretically ensure consistent outputs, the reality is more complex. Since commercial LLM providers regularly adjust their models' parameters and users lack full control over the automated grading process, there remains a possibility that these models will produce different results across multiple runs. Interestingly, this is similar to human graders, who occasionally will award different ratings to the same student response, when asked to re-rate it after a period of time. Measures of intra-rater reliability are intended to evaluate the extent to which a single rater agrees with their own judgment over time.

To investigate the intra-rater reliability of these models, I tested the best-performing model configurations by re-marking the entire dataset and compared the results on the two-class task. I observed high but not perfect rates of agreement, with seven students' answers having different labels applied to them; resulting in a Cohen's kappa score of 0.98. While this level of agreement would be considered exceptionally high for human raters, the lack of perfect agreement was an intriguing finding, given the model had the exact same inputs, and had its temperature set to zero. An in-depth examination of responses where classification outcomes differed revealed no distinguishing patterns or characteristics. The observed variability likely stems from OpenAI's reported implementation of a "mixture of experts" architecture, where intermediary models first analyze and direct incoming API requests to specialized task-optimized models, potentially introducing subtle inconsistencies into the processing pipeline.

5.4.4 Evaluating Potential Bias

Bias and fairness in AI models, particularly in the context of education, have been extensively studied in the literature (see Baker & Hawn 2022 for a comprehensive review). However, the definition of fairness and bias in the AI literature is complex and often contested. In this context, I consider a model to be biased if it exhibits judgments that are significantly different from human raters, based on a student's membership in a socio-economic group. For example, if the model systematically rated responses from female students lower than those of male students, in comparison to the scores given by human raters, one might suspect the model was exhibiting a form of bias. The underlying mechanisms by which a generative LLM might operate in this way are complex and open for debate, especially because I provide the same prompt for all students and the model has no access to student demographic information. However, it is possible, in principle, that due to the corpus on which the model was trained, it might "prefer" certain types of student responses, such as those of native English speakers, over those of non-native speakers. If this preference differed from that of expert human raters, it would be a cause for concern. It is also possible that such a discrepancy could point to implicit biases or prejudices held by the human raters themselves, although a proper treatment of this question is outside the scope of this article. Nonetheless, an initial check for potential bias was conducted by investigating if the model misclassified students' answers (relative to human raters) at a higher rate based on the students' gender and whether English was spoken at home..

Table 5.5
Misclassification Rates by Demographic Group

	Percent of Entire Dataset	**Percent of Misclassifications**
Male	33% | (354)	35% | (9)
Female	67% | (714)	65 % | (17)
English at Home	13% | (144)	8 % | (2)
No English at Home	87% | (924)	92 % | (24)

As can be seen in Table 5.5, the rates at which misclassifications occurred by group, closely corresponds with the overall makeup of the dataset, and I found no statistically significant difference due to gender, or home language. However, given the small number of total misclassifications (26) these results should be treated with caution.

5.5 Discussion

The topline finding, that GPT-4 can reach equivalent performance levels as expert human raters on ASAG tasks, has important implications. If LLMs can perform as well as human raters, as my research suggests, it indicates that they could form the basis of assessment systems that could be used in schools. Notably, these results were obtained with no fine-tuning and limited prompt engineering, indicating that generative LLMs could effectively support informal assessment activities in educational practice. This capability holds particular promise for regions where assessment resources are scarce or where technical constraints prevent model customization.

Implementing an assessment system that leverages LLMs for diagnostic and formative assessment purposes in real-world situations would require some technical infrastructure. In high-income countries, I believe there is significant potential in the short term for using LLMs to support more frequent and effective assessments. This could potentially have a disproportionate benefit for lower-resourced schools, where it can be challenging to ensure the ideal quantity and quality of diagnostic and formative assessment. While these findings suggest that such approaches could be effective with student work from anglophone LMICs, implementing them in practice would still face challenges. For example, schools would need appropriate devices, infrastructure, and connectivity, and teachers and school staff would need training. In the near term, a more likely use case in LMIC contexts would be for large-scale assessments meant to screen for student ability and mastery of basic skills. Encouragingly, the costs of using these models are decreasing, while the tools being developed for non-expert users are advancing rapidly. A better understanding of these practical limitations will be crucial for realizing the potential benefits of generative LLMs in automated short answer grading and ensuring that these advancements are accessible to a wide range of educational institutions and learners.

5.5.1 Limitations

This study provides valuable insights into the potential of generative LLMs for automated short answer grading; however, it is important to acknowledge its limitations. Though the ROARS dataset allows for the evaluation of an underrepresented group, it ultimately consists of student responses, in English, from a single country. It is unclear whether model's performance would

generalize to other languages or dialects. Future research could examine the performance of generative LLMs across a wider range of cultural settings, languages, and dialects to help better understand their potential application; it may also limit the generalizability of the findings to other educational contexts and student populations. Additionally, while the ROARS dataset provides some demographic information about the students, such as age, gender, and whether they speak English at home, it may not capture the full range of socioeconomic and cultural factors that could influence student performance.

In terms of the experimental design, a key limitation is the type and difficulty of the reading comprehension questions used. The questions were primarily direct comprehension questions, assessing students' ability to retrieve explicit information from the given passages. It remains uncertain how the model's performance would change as the complexity of the topics and the difficulty of the questions increase. Furthermore, the study did not explore the model's performance across different subject areas, which could have varying levels of complexity and require different types of reasoning skills. Additionally, the study employed simple binary and three-class rubrics that primarily focus on overall correctness. It is unclear how the model would perform if the scoring criteria were more complex or had more potential classes, such as those that assess the quality of the student's writing or the depth of their understanding. Finally, this study evaluated only GPT-4 rather than multiple generative LLMs. Given the rapid pace of development in the field of AI, it is likely that the performance of these models will continue to improve over time. As such, the results of this study should be interpreted as a snapshot of the current state of the technology rather than a definitive assessment of its potential.

5.5.2 Further Research

Further research used to evaluate generative LLMs' performance in automated short answer grading should explore datasets that cover a wider range of grade levels, subject areas, and question types, as well as varying levels of difficulty. By testing the models on a more diverse set of educational contexts, researchers can better understand the generalizability and limitations of these technologies. This study focused solely on English-language responses from Ghanaian students, and future research should explore the performance of LLMs in non-English languages, as this is key a consideration for determining their potential for global education.

Another important area for further research is the ethical implications of using generative LLMs for automated grading. While the current study briefly addresses the topic of bias in the model's performance, it does not fully explore the potential concerns regarding privacy, fairness, and the perpetuation or amplification of existing educational inequities. Future research should delve deeper into these ethical considerations and develop guidelines for the responsible use of generative LLMs in educational assessment. Moreover, further investigation into the intra-rater reliability of model results is necessary, as ensuring the consistency and trustworthiness of the grading outcomes is critical for educators to have confidence in the technology. Finally, while this study focused on GPT-4, it is essential to evaluate the performance of other generative LLMs, including open-source models that could potentially be more affordable and transparent. By exploring a wider range of models and their capabilities, researchers can better inform the development and implementation of these technologies in real-world educational settings.

5.6 Conclusion

This study aimed to address the lack of empirical research on the use of generative Large Language Models (LLMs) for educational assessment, particularly in the context of low and middle-income countries (LMICs). To achieve this goal, the study introduced a novel dataset, the ROARS dataset, containing over 1,000 student answers to reading comprehension questions from 130 students aged 9 to 18 in Ghana. This dataset provides a unique opportunity to evaluate the performance and potential of generative LLMs in a diverse educational context, while also offering insights into the educational challenges faced by students in LMICs.

The experimental design focused on evaluating the performance of GPT-4, a state-of-the-art generative LLM, in grading short answer reading comprehension questions from the ROARS dataset. The study employed various prompting strategies, including zero-shot and few-shot learning, and compared the model's performance to that of expert human raters. The results showed that GPT-4, with minimal prompt engineering, achieved near-parity with expert human raters, attaining a Quadratic Weighted kappa (QWK) score of 0.91 in the three-class condition and a Linear Weighted kappa (LWK) score of 0.87 in the two-class condition. These findings demonstrate the remarkable ability of generative LLMs to evaluate student responses accurately,

even when dealing with a novel dataset from an underrepresented educational context. The implications of these findings are significant for both high-income countries and LMICs. In high-income countries, generative LLMs could be used to support more frequent and effective diagnostic and formative assessments, particularly in lower-resourced schools where the quantity and quality of assessment is typically lower. In LMICs, while implementing these approaches may face challenges at the individual school level, they could be valuable for large-scale diagnostic assessment. However, it is important to note that implementing these approaches in real-world situations would require some technical infrastructure and support.

This study contributes to the growing body of research on the application of generative LLMs in education by providing empirical evidence of their potential for automated short answer grading. The introduction of the ROARS dataset aims to help fill the gap in available datasets of student short answer questions, particularly from LMICs. As the field of AI in education continues to evolve, it is crucial to ensure that the benefits of these technologies are accessible to students from all backgrounds and educational contexts. Future research should explore the potential of generative LLMs in various educational settings, while also addressing the challenges and ethical considerations associated with their implementation.

Chapter 6 - Can LLMs Grade Story Retell

6.1 Introduction

The multifaceted nature of reading comprehension has led to the development of various assessment approaches, including open-response questions, which typically require a student to read (or listen) to a passage and then answer questions in their own words (Shapiro et al., 2014). Some researchers claim that open-ended questions are better suited to evaluate reading comprehension than closed-response questions because they require students to actively reconstruct and communicate about what they have just read, which more closely corresponds with real-world applications (Bishop & Adams, 1990; Cain & Oakhill, 2006). The most widely used types of open-ended questions are short answer, story retell, and summarization. For story retell tasks, the focus of this study, students are asked to read a passage and write what they remember about the story in their own words (Reed & Vaughn, 2012). This typically involves a more comprehensive recounting of the text, often maintaining the original sequence of events. Story retell can be scored in a variety of ways including counting the total number of words in a retell, analysing the presence of specific story elements, or providing a holistic score (Reed et al., 2012; Shapiro et al., 2014). However, it can be challenging for teachers to score story retell tasks time-effectively and reliably, making them difficult to grade at scale.

Large Language Models (LLMs) present an intriguing possibility for addressing the challenges of grading retell assessments; however, there has been limited research on this question to date. This paper contributes to the existing literature through three interrelated studies using data from the Rising & Oxford African Reading Comprehension Short Answer Dataset (ROARS) to investigate: (1) the reliability of different measures of story retell, (2) the convergent validity of story retell as a measure of reading comprehension, and (3) the capacity of LLMs to grade story retell in a manner similar to human graders.

In the first experiment, I examine the impact of rubric choice on inter-rater reliability among expert human raters when evaluating student story retells. Specifically, I consider how the use of binary, three-class, and five-class rubrics affects the level of agreement between raters. In the second experiment, I explore the convergent validity of different measures of story retell.

This involves analysing the relationship between fluency-based and rubric-based measures of retell, as well as examining how these measures correlate with other indicators of reading ability such as oral reading fluency and responses to short answer and multiple-choice reading comprehension questions. In the third experiment, I investigate the potential of LLMs in grading student story retells and how closely different configurations of LLMs and prompting strategies can replicate expert human judgments.

6.2 Prior Work

6.2.1 Assessing Reading Comprehension with Retell

Put simply, the underlying rationale for story retell as a measure of reading comprehension is that it is impossible for a student to retell a story they have not comprehended, so being able to do so indicates comprehension. A more nuanced justification for using retell as an assessment task is rooted in its ability to engage multiple cognitive processes involved in comprehension: retelling a story challenges a student's capacity to draw upon memory for factual details, generate inferences to fill gaps in their recall, and finally to reconstruct and sequence events and major concepts presented in the text (Reed et al., 2012; Wilson et al., 1985). Furthermore, retelling requires students to recall—rather than simply recognize—information from the text and express it either orally or in writing without referring to the passage (Spooner et al., 2004). This process of active recall and reconstruction is thought to more closely mirror the cognitive demands of reading comprehension.

Prior findings on the effectiveness of retell measures have been mixed, with some studies showing moderate to strong correlations between retell scores and other measures of reading comprehension (Reed & Vaughn, 2012), and others finding inconsistent evidence of a unique relationship to students' reading abilities (Hagtvet, 2003; Marcotte & Hintze, 2009; Riedel, 2007). A literature review by Reed & Vaughn (2012) looked at 54 studies involving retell measures and found moderate correlations with standardized measures of reading comprehension across grades. Nonetheless, many argue that retelling can provide teachers with valuable insights into student comprehension, and serves as a valuable tool for comprehensively assessing and

supporting students' reading comprehension skills. However, the process requires individual administration and can be time-intensive to both administer and evaluate.

6.2.2 Different Measures of Story Retell

There are multiple different approaches for scoring story retell, and all approaches have their own shortcomings (Bellinger & DiPerna, 2011; Fuchs et al., 1988). One common approach involves using a text analysis system to divide passages into idea units and assign these units to levels of importance (Maria, 1990). While this method enables a granular analysis of a student's retell comprehensiveness and accuracy, the process of meticulously categorizing and weighting narrative elements is inherently time-intensive and requires substantial analytical effort. Moreover, because such rubrics necessitate individual development for each specific text, comparative analysis across different stories becomes methodologically challenging. Researchers encounter significant obstacles in establishing standardized evaluation criteria that can be consistently applied across varied narrative contexts.

A related approach is component-based scoring, which focuses on assessing the presence of specific narrative elements such as main ideas, characters, and setting. Although this method relies on more generalizable story components compared to the idea-unit approach, it remains substantially complex and labor-intensive. Component-based scoring frequently encounters reliability challenges, with researchers observing inconsistent inter-rater agreements that complicate systematic assessment (Klesius & Homan, 1985). Yet another approach, holistic scoring, involves assigning an overall score based on the quality and completeness of the retell, offers a middle ground between the efficiency of fluency-based measures and the detail of component-based scoring. However, holistic scoring can be subjective and can struggle to achieve adequate interrater reliability if the categories are not carefully defined and if the raters are not given a chance to calibrate their scores.

Finally, a more objective measure, used by Fuchs et al. (1988), involves simply counting the total number of words included in a participants' retell. They argue that this provides an objective unit of measure and is related to ORF which has a strong relationship with reading comprehension. The approach offers several apparent advantages: it is straightforward to implement, provides a quantifiable metric, and can be easily automated. However, this method

raises significant methodological concerns. Reed (2011) has argued that only considering the volume of the retell, rather than the quality, can be misleading. Students could potentially manipulate the metric through several strategic behaviors that do not reflect genuine reading comprehension.

As can be seen, each of these approaches has its strengths and limitations, and the choice of scoring method depends on the specific goals of the assessment, the age of the students, and the resources available for administration and scoring. As research in this area continues to evolve, it is likely that new methods and technologies will emerge to further refine and improve the assessment of reading comprehension through story retell measures. Given the challenges associated with traditional methods of scoring story retell assessments, it is worth exploring how emerging technologies, particularly generative Large Language Models (LLMs), might offer new approaches to this assessment challenge.

6.2.3 Potential of Generative LLMs for grading Story Retell

In the context of story retell assessment, LLMs present several distinctive capabilities that set them apart from traditional scoring methods. Unlike previous assessment methods that rely on rigid, predefined scoring rubrics, LLMs have demonstrated an impressive capacity to generalize to new tasks, often with minimal exposure to examples (Ouyang et al., 2022). While previous approaches have been constrained by manual, time-intensive evaluation processes, these models can complete new tasks simply by modifying their prompts, rather than requiring additional training or fine-tuning.

However, Schneider et al. (2023) note that while these models provide valuable insights, their readiness for completely independent automated grading is still uncertain. The complexity of human narrative interpretation means that LLMs must navigate subtle contextual and interpretive challenges that go beyond straightforward computational approaches. This chapter explores whether these models can be systematically applied to story retell assessment, specifically within the context of the ROARS dataset. By examining the models' performance across different rubric structures, it is possible to better understand their potential and limitations in educational assessment. This study aims to systematically evaluate LLMs' effectiveness in scoring story retell assessments, potentially offering a reliable and efficient alternative to

traditional scoring methods while maintaining the pedagogical value of this important comprehension measure.

6.3 Dataset and Human Grading

In the ROARS 130 students read one of two 400-word fictional stories and answered a series of comprehensions questions including a story retell task. Out of the 130 students, only 95 completed the retell task, as many responses were left blank. These 95 story retells were transcribed, and a count of the total number of words was taken. Next, three human raters, all graduates of Master of Education programs and former classroom teachers, evaluated the same set of student story retells using three separate rubrics: a binary rubric, a three-class rubric, and a five-class rubric. These rubrics were drawn from previous measures in the literature (see Figure 6.1) and then lightly adapted to standardise and streamline the rating process (see Figure 6.2).

Figure 6.1
Original Rubrics

Number of Categories	Rubric Name	Definition	Qualitative Elements	Score Range
2 – class	BADER Reading and Language Inventory Bader and Pearce (2009)	Number of unprompted plus prompted ideas	Judgement of whether the retell was organized	Yes / No
3 – class	Analytical Reading Inventory Wood & Moe (2007)	N/A	Retells judged for completeness, organizations and sentence structure	All, Some, None
5 - class	Informal Reading Inventory Roe & Burns (2007)	N/A	Guiding questions about completeness, accuracy, and organization.	1 “poorly” to 5 “very well”

Figure 6.2
Adapted Rubrics

Number of Categories	Rubric Categories and Definitions
2 – class	**0 - Bad retelling** (student answer has few accurate story elements, but also lack of narrative flow or reasonable relevant details) **1 - Good retelling** (student answer includes multiple accurate story elements, a sense of narrative flow, and a strong focus on the story, emphasizing the central character and their experience)
3 – class	**0 - Bad retelling** (student answer has few accurate story elements, but also lack of narrative flow or reasonable relevant details) **1 - Mediocre retelling** (student answer is neither bad nor good, it has elements of both) **2 - Good Retelling** (student answer includes multiple accurate story elements, a sense of narrative flow, and a strong focus on the story, emphasizing the central character and their experience)
5 – class	**0 - Bad retelling (**No retelling / no accuracy / no detail / lacks coherence) **1 - Poor retelling** – (Contains one or two brief elements of the story or one developed element. No / very little narrative flow / recounts one isolated story element) **2 - Mediocre retelling** (Contains three or four brief elements or two or three developed elements. Mostly accurate. May contain some narrative flow.) **3 - Acceptable retelling** (Contains three or four brief elements or two or three developed elements. Accurate. Some narrative flow.) **4 - Good retelling (**Contains multiple brief and developed story elements. Fully accurate. Clear evidence of narrative flow.)

In cases where all three raters did not agree, the ground truth scores were determined by averaging human judgments and rounding to the nearest whole class. For instance, in the case of the three-class rubric, if one judge rated a response as 0 and two judges rated it as 2, the average score would be 1.67, which would be rounded to 2, resulting in a human ground truth score of 2. The very need for a process to determine ground truth scores emphasizes that human agreement is not unanimous in most cases, introducing an element of uncertainty even within the data used to generate the ground truth score.

Figure 6.3
Examples of different rater's scoring by rubric

Retell 1			
lucy was a girl who like learning around she round through the country stole that night allways when everyone including the sheeps and lambs were as sleep. lucy helped to save the shepherd when the shepherd got a broke in his leg.			
	2 - class	**3 – class**	**5 - class**
Rater 1	**0**	**1**	**1**
Rater 2	**0**	**1**	**2**
Rater 3	**0**	**1**	**2**
Ground Truth	**0**	**1**	**2**
Retell 2			
Lucy was different from all the other sheep right from the start. One day something terrible happened. The shepherd fell over and broke his leg. The shepherd lay in the grass, not knwoing what to do. one day a truck driver noticed Lucy and stopped in the road. "Going to the valley?" he asked Lucy nodded. Lucy and the farmer drove on the tractor to find the other sheep. The farmer took his to the hospital right away. The shepherd had his leg in a cast for six weeks, before he could go back to the sheep. when he returned he gace Lucy a big smile. Lucy was free to roam as she pleased.			
	2 - class	**3 – class**	**5 - class**
Rater 1	**1**	**2**	**4**
Rater 2	**1**	**2**	**3**
Rater 3	**1**	**1**	**4**
Ground Truth	**1**	**2**	**4**

6.4 Experiment 1 - Inter-Rater Agreement on Story Retell

In the first of the three interrelated experiments, I examine the level of inter-rater agreement for the story retell tasks across three different rubrics. This analysis aims to provide insights into the adequacy of each rubric for assessing story retell performance, as well as to investigate how inter-rater agreement might vary depending on the assessment tool used. I report the average rating awarded by each rater to better understand if there were any systematic differences between the raters. For context, in all three rubrics, zero was the lowest score, so a higher overall average number indicates that a rater judged student retells to be of higher quality. As can be seen in Table 6.1, it appears that Rater 3 was more lenient than Raters 1 and 2 for the two-class and three-class rubric, with more convergence on the five-class rubric. Perhaps, when presented

with an edge case, Rater 3 might have been more likely to award the student the higher of the two categories. However, this is only one of many possible interpretations, and this type of difference between raters is precisely the kind of discrepancy that would typically be addressed during rater calibration.

Table 6.1
Comparison of Average Score by Rubric and Rater

	Two-class Rubric	**Three-class Rubric**	**Five-class Rubric**
Rater 1	0.17	0.37	1.41
Rater 2	0.21	0.36	1.14
Rater 3	0.37	0.59	1.22
(Min Score | Max Score)	(0 | 1)	(0 | 2)	(0 | 4)

The selection of statistical measures for quantifying inter-rater agreement requires careful consideration of the underlying data structure and research objectives. For the binary rubric, I employed Fleiss' kappa, a robust statistical measure specifically designed for nominal categorical data with three or more raters. This approach recognizes the fundamental nature of binary classifications, where each rating represents a discrete, non-ordered category.

In contrast, the three-class and five-class rubrics introduce an ordinal dimension, where the scoring categories represent meaningful hierarchical levels of performance. Kendall's W was selected because it can capture the inherent hierarchical relationships between different score levels. Unlike nominal measures, Kendall's W accounts for the fact that a disagreement between a score of 0 and 4 represents a more significant deviation than a disagreement between adjacent categories. As an additional check, I also calculated Cohen's kappa (LWK for two-class rating and QWK for three-class and five-class) for each pairwise combination of raters and took the average of the results. While averaging pairwise Cohen's kappa values serves as an intuitive supplementary check but has statistical limitations. Unlike Fleiss's kappa, which is designed for multiple raters, this approach doesn't properly account for rater interdependence and may inflate agreement estimates (Zapf et al., 2016; Hallgren, 2012). Nevertheless, it provides a useful

comparison point that confirms the pattern of improving agreement across increasingly granular rubrics..

Table 6.2 presents the results of the inter-rater reliability analysis, illustrating the levels of agreement achieved across the different rubrics and measurement approaches. For comparison, I have included the average Cohen's kappa scores; however, I consider Fleiss's kappa for the two-class rubric, and Kendall's W for the three-class and five-class rubrics to be the most appropriate measures based on the literature.

Table 6.2
Comparison of Interrater Agreement by Rubric and Measure

Measure	two-class	three-class	five-class
Fleiss's kappa	0.56		
Kendall's W	0.74	0.81	0.85
Cohen's kappa Pairwise Average	0.60	0.74	0.78

While agreement was moderate in the two-class condition, inter-rater agreement improved substantially in the three-class condition and improved slightly once again for the five-class condition. Additionally, the overall trend moving from lower to higher agreement is also seen in the unconventional, “sense-check”, average Cohen’s kappa metric. On the whole, these results suggest there is a relatively high level of inter-rater agreement between human raters on the retell task for the three-class and five-class rubrics, even with relatively simple rubrics and little-to-no rater calibration. In the following section, I investigate how closely these rubric-based ratings of retell quality correspond with fluency-based measures of retell and the extent to which either measure corresponds to other indicators of reading ability.

6.5 Experiment 2 – Convergent Validity of Story Retell

The second experiment aims to evaluate the validity of story retell as a measure of reading comprehension. First, I compare the correlation between scores generated by three distinct rubrics (two-class, three-class, and five-class) and a fluency-based measure. This comparison allows for analysis of the consistency and potential interchangeability of these scoring methods.

Because the same student responses are being used, it would be expected that these different scoring methods would have a relatively strong correspondence between them. If the relationship is too weak, it could be inferred that one or more of these measures is unreliable, or that they are evaluating different underlying constructs.

Second, I examine the correlation between story retell measures of reading ability and other established indicators of reading proficiency, specifically oral reading fluency, multiple-choice and short answer reading comprehension questions. Because the same students performed the same tasks, and all the tasks are intended to measure reading ability, one would expect there to be a correlation between these different measures. Relatively high levels of correlations between these different measures, could be taken as evidence of convergent validity, whereas low levels of correlations could suggest that story retell was a relatively unreliable measure of reading ability, or that it was in fact measuring a distinct construct.

I took the expert human ratings for story retell across the three distinct rubrics (two-class, three-class, and five-class) and calculated the mean score for each item on each rubric. I then calculated a fluency-based measure of retell, in this case the total number of characters in the student's retell, for each student response. This resulted in each of the 95 student retells having four scores: the average score on the two-class rubric, average score on the three-class rubric, average score on the five-class rubric, and number of total characters in the retell.

To standardize the different scales, I calculated each student's percentile rank relative to other students on each measure. Because the different scores are for the same student response, if the different measures were fully equivalent and interchangeable, one would expect students' percentile ranks to remain unchanged. If, however, student percentile ranks change based on retell measure, one could conclude that the change in the rubric had an impact on how students were being evaluated. Next, I calculated Pearson correlations between these normalized scores to quantify the relationships between different measures. The results, presented in Table 6.3, show the following: high correlations (above 0.7) were observed among the three-class rubric, five-class rubric, and fluency-based measure.

Table 6.3
Correlation Between Different Measures of Story Retell

	Character Count	Rubric 2 Class	Rubric 3 Class	Rubric 5 Class
Character Count	1.00	-	-	-
Rubric 2 Class	0.57	1.00	-	-
Rubric 3 Class	0.76	0.72	1.00	-
Rubric 5 Class	0.84	0.78	0.86	1.00

The high correlations between character count, the three-class rubric, and the five-class rubric suggest that any of these approaches could potentially serve as effective methods for story retell assessment. In contrast, the two-class (binary) rubric consistently showed lower correlations with other measures, indicating it may be less effective at capturing the nuances in students' retell performance.

I then examined how students' retell performance related to other established measures of reading ability: oral reading fluency (ORF) and reading comprehension. As detailed in Section 3, students completed a comprehensive assessment battery where they read a short story aloud (scored for ORF) and then completed silent reading comprehension assessments consisting of 3 multiple-choice questions, 6 short-answer questions, and 1 story retell task.

For this analysis, I standardized the various measures by converting the raw scores to percentile ranks. The original ORF scores (measured as words correct per minute) and reading comprehension scores (percentage of correct answers out of nine questions) were transformed into percentile ranks to facilitate comparison across different measurement scales. I then calculated Pearson correlations between these normalized scores to quantify the relationships between the different aspects of reading ability. The results of this analysis are presented in Table 6.4.

Table 6.4
Correlation Between Measures of Retell, Oral Reading Fluency and Comprehension

	Character Count	Rubric 2 Class	Rubric 3 Class	Rubric 5 Class	Oral Reading Fluency	Reading Comprehension
Character Count	1.00	-	-	-	-	-
Rubric 2 Class	0.57	1.00	-	-	-	-
Rubric 3 Class	0.76	0.72	1.00	-	-	-
Rubric 5 Class	0.84	0.78	0.86	1.00	-	-
Oral Reading Fluency	0.27	0.25	0.26	0.32	1.00	-
Reading Comp	0.25	0.58	0.41	0.50	0.43	1.00

While correlations between these percentile ranks revealed moderate relationships (between 0.3 and 0.4) between story retell measures, ORF, and the composite reading comprehension score, the levels of correlation are relatively similar across the three rubrics. These moderate correlations may also be reflective of developmental differences, with ORF and story retell serving as precursors to reading comprehension and potentially functioning as good indicators for lower-level readers.

The moderate correlations observed between different reading measures provide important insights into their relationships. If ORF and reading comprehension had shown very high correlation while retell measures demonstrated substantially lower correlations, this might have indicated either assessment implementation issues or that story retell was measuring a fundamentally different construct. Instead, the pattern of moderate correlations (ranging from 0.25 to 0.58) across all measures suggests a convergent relationship between these different aspects of reading ability. However, the magnitude of these correlations also indicates that each measure likely captures distinct components of reading proficiency rather than reflecting a single underlying construct. For example, the three-class rubric's stronger correlation with reading comprehension (0.41) compared to ORF (0.26) suggests it may be more closely aligned with

comprehension skills than fluency. These findings underscore the multidimensional nature of reading development and reinforce the value of employing multiple assessment approaches when evaluating students' reading abilities, particularly in developmental contexts where different skills may emerge at varying rates.

6.6 Experiment 3 - Evaluating LLMs' Ability to Score Student Story Retells

The prior two experiments established that there was a relatively high level or correlation between different methods for scoring story retell, and that story retell appears to be related to but independent from other measures of reading ability. In this third experiment, I examine the capability of LLMs to score student story retells in a way that mimics expert human raters. To explore this, I prompted a LLM to score student retells using the same rubrics as were used by human raters: a two-class rubric (bad/good retell), a three-class rubric (bad/mediocre/good retell), and a five-class rubric (ranging from "bad retell" to "good retell" with intermediate levels). The model was provided with a simple prompt that explained the task and the relevant rubric, one example for each rubric category, the original story, and the student response. It was then instructed to grade each student response according to the relevant rubric. I intentionally began with this simpler prompting strategy so that I could assess the model's inherent capabilities with the aim of establishing a baseline for understanding how well LLMs could mimic human rating.

To evaluate the LLM's performance, I compared the scores it assigned to the student retells against the ground truth scores determined by human raters. I employed traditional machine learning performance metrics—including precision, recall, and F1 score—to quantify the model's accuracy in classifying the responses. Additionally, I reported the kappa measures to assess the level of agreement between the LLM and human raters beyond what would be expected by chance alone. Specifically, Cohen's kappa was utilized as it is appropriate for comparing categorical data between two raters—in this case, the LLM and the aggregate human ratings. For the two-class ratings, I used Linear Weighted kappa (LWK), and for the three-class and five-class ratings, I used Quadratic Weighted kappa (QWK). The results of the baseline approach are summarized below in Table 6.5.

Table 6.5
Performance of baseline prompt Using GPT4o-2024-05-13

	Model Prediction	Precision	Recall	F1	Support	LWK/QWK
2 class	*Bad Retell (0)*	0.86	1.00	0.93	74	0.35
	Good Retell (1)	1.00	0.25	0.40	16	
	Average	0.89	0.87	0.83	90	
3 class	*Bad Retell (0)*	0.95	0.94	0.94	64	0.78
	Mediocre Retell (1)	0.57	0.81	0.67	16	
	Good Retell (2)	1.00	0.25	0.40	8	
	Average	0.89	0.85	0.84	88	
5 class	*Bad Retell (0)*	0.68	0.91	0.78	23	0.82
	Poor Retell (1)	0.77	0.68	0.72	34	
	Mediocre Retell (2)	0.65	0.58	0.61	19	
	Acceptable Retell (3)	0.50	0.38	0.43	7	
	Good Retell (4)	1.00	1.00	1.00	1	
	Average	0.69	0.69	0.68	84	

In the case of the two-class rubric, the LLM achieved a precision of 0.86 for the "Bad Retell" category and 1.00 for the "Good Retell" category, but with a recall of 1.00 and 0.25 respectively. The LWK for this rubric was 0.35, indicating low agreement with human raters. For the three-class rubric, the model showed improved performance, particularly in the "Bad Retell" category, with a precision and recall of 0.95 and 0.94 respectively, and a QWK of 0.78. Similarly, with the five-class rubric, the LLM demonstrated an even higher QWK of 0.82.

To better understand these results, I compare the model-to-human agreement to the human-to-human agreement. As shown in Table 6.6, the kappa values of 0.78 for the three-class

rubric and 0.82 for the five-class rubric, nearly reach the level of human-to-human agreement levels of 0.81 and 0.85. In contrast, the model's performance with the two-class rubric was notably lower, with an LWK of 0.35 compared to the human-to-human agreement of 0.56.

Table 6.6
Model Human Agreement compared to Human-to-Human Agreement

	Model to Human	**Human to Human**
2 – Class	*0.35*	*0.56*
3 - Class	*0.78*	*0.81*
5 - Class	*0.82*	*0.85*

These results suggest that LLMs can closely approximate human scoring in retell tasks when the rubric provides sufficient granularity to capture the nuances in student responses but seems to struggle with the simplistic classification of responses into just two categories. This is intriguing, as one might expect that fewer categories would simplify the task for the model. However, the lower performance aligns with my earlier findings that the two-class rubric itself had lower inter-rater reliability among human raters. The binary rubric may not capture the complexity of the students retells, leading both humans and the model to make more variable judgments. Taken together, these results present evidence that LLMs can effectively mimic human scoring using only relatively simple prompting approaches, if provided well-designed rubrics.

6.7 Discussion

The findings of this study contribute to our understanding of both the measurement properties of retell tasks and the potential role of Large Language Models in grading these assessments. The first experiment revealed relatively high levels of inter-rater agreement across different rubrics and scoring approaches, with particularly strong agreement when using the three-class and five-class rubrics compared to the binary rubric. This suggests that story retell tasks are capturing a relatively stable construct, and that various scoring approaches could be justified depending on the specific educational context. This experiment also indicated that more granular rubrics might enable more reliable scoring while still maintaining sufficient ease of use for regular classroom assessment.

The second experiment examined the relationship between retell assessments and other established measures of reading ability, such as ORF and reading comprehension questions, finding they were low to moderate. While this finding aligns with prior research suggesting that retell tasks assess different components of reading proficiency, it warrants careful interpretation given the developmental stage of the participants. Story retell represents a relatively basic form of reading comprehension assessment compared to tasks requiring inference, analysis, or evaluation. For adolescent readers, as in this study, one would expect stronger correlations with more complex comprehension tasks that better match their developmental level. The low to moderate correlations observed might reflect this mismatch between task complexity and student capability rather than indicating problems with the assessment itself. The relationship between retell performance and student effort also merits consideration. Since the retell task was untimed, variations in student engagement likely influenced both the fluency-based and rubric-based scores. Students who invested more time and effort would naturally produce longer and more detailed responses. This suggests that retell tasks might capture a combination of reading comprehension and non-cognitive skills like persistence and willingness to elaborate. While this could be seen as a limitation, it might also represent an opportunity to assess important academic behaviors alongside reading comprehension. Future research could explore this relationship by implementing both timed and untimed retell tasks to better understand how time constraints affect performance and the relationship with other measures of reading ability.

The third experiment examined the performance of LLMs in grading story retell tasks and found that they effectively mimic human raters with minimal prompting, achieving agreement levels comparable to inter-rater reliability among experts. This suggests potential for supporting more regular use of retell assessment in classroom settings.

First, investigation is needed into how these models perform across different types of texts and different levels of text complexity. Second, research should examine whether model performance varies across different student populations and age groups, particularly given the developmental considerations discussed above. Third, studies should explore how automated scoring might be integrated into existing assessment practices in ways that support rather than replace teacher judgment. Additional areas for future research include investigating how different prompting strategies might affect model performance, examining the potential for

providing automated feedback beyond simple scoring, and exploring how these technologies might support assessment in languages other than English. There are also important questions about how automated scoring systems might affect student motivation and engagement with retell tasks, and whether knowing their responses will be machine-scored influences how students approach the task.

The practical implications of these findings extend beyond just the technical feasibility of automated scoring. The ability to reliably score retell tasks without extensive time investment could enable more regular use of this assessment format, providing teachers with richer information about student comprehension while maintaining reasonable workload demands. However, successful implementation would require careful attention to teacher training, assessment protocols, and the integration of automated scoring with human judgment. These considerations suggest the need for implementation research that examines how automated scoring systems can best be incorporated into educational practice.

6.8 Conclusion

In conclusion, this study contributes to the literature on the assessment of reading comprehension through story retell tasks and the potential application of LLMs in grading such assessments. The findings indicate that story retell is moderately associated with other measures of reading ability, reinforcing its utility as a diagnostic and formative assessment tool. The relatively high correspondence between fluency-based and rubric-based measures of retell suggests that these approaches capture similar constructs. However, the lower correlations between retell measures and other reading assessments like ORF and comprehension questions suggest that retell tasks may assess distinct subcomponents of reading ability, possibly influenced by factors such as student effort and engagement. This highlights the importance of considering multiple measures when evaluating reading comprehension to obtain a holistic understanding of a student's abilities. The exploration of LLMs for grading story retells revealed that even with relatively simple prompting strategies, LLMs can mimic human grad with accuracy and agreement levels comparable to human raters.

However, this study has limitations that warrant careful consideration when interpreting the findings. First, the sample size was relatively small, comprising only 94 student retell responses from a specific educational context in Ghana, with participants within a 5-year age range. Such a limited and specific sample may affect the generalizability of the results to other populations or educational settings. Secondly, the rubrics themselves, while adapted from prior studies, may not have fully captured all relevant aspects of retell quality, potentially limiting the comprehensiveness of the evaluation.

Chapter 7 - Discussion

7.1 Summary of Findings

This dissertation investigated the potential of AI to support the assessment of English reading ability in West Africa through three interrelated empirical studies. Each study examined a different aspect of reading assessment: oral reading fluency, short answer reading comprehension questions, and story retell tasks. Together, these studies provide evidence that recent advances in AI, particularly in Natural Language Processing and Automatic Speech Recognition, could enable more frequent and flexible diagnostic and formative assessment of reading ability.

The first study examined the feasibility of using ASR models to assess oral reading fluency. Using a novel dataset of Ghanaian students' oral reading, I found that Whisper v2 could transcribe student reading with an accuracy rate of 10.3 WER without any fine-tuning. Notably, the model's performance showed similar accuracy when transcribing recordings from both Ghanaian and American students. While a WER of 10.3 still has room for improvement, when the ASR generated transcripts were used to calculate Words Correct Per Minute scores, the model generated score had high agreement with scores from expert human raters (correlation of 0.98). This suggests that automated ORF assessment could be viable in diverse educational contexts.

The second study investigated the potential of Large Language Models to grade short answer reading comprehension questions. Using a novel dataset which contained over 1,000 student responses, the study found that GPT-4 could achieve performance levels comparable to expert human raters with minimal prompt engineering. The model achieved a Quadratic Weighted kappa of 0.91 on a three-class rubric and a Linear Weighted kappa of 0.87 when using a binary rubric. These results exceeded previous benchmarks in automated short answer grading and approached human-level performance (Fernandez et al., 2023).

The third study examined the use of LLMs for grading story retell tasks, while also investigating the reliability and validity of different retell measures. When evaluating LLM performance on grading student retells, I found that GPT-4 could closely approximate human scoring, achieving agreement levels (QWK of 0.82) nearly matching human-to-human agreement

(.85) when using well-designed rubrics. I also found high inter-rater reliability among human raters when using three-class and five-class rubrics (Kendall's W of 0.81 and 0.85 respectively). I also found only moderate correlations between measures of retell and other indicators of reading ability, suggesting that retell tasks may capture distinct aspects of reading comprehension.

Several cross-cutting themes emerged from these studies. First, in all three studies, the new generation of LLMs and ASR models achieved levels of agreement with human raters that were comparable to inter-rater agreement among experts, suggesting that these technologies could potentially provide reliable scoring for diagnostic and formative assessment purposes. Second, the models achieved these results without requiring task-specific fine-tuning or extensive prompt engineering, obviating the need for substantial technical expertise and task-specific training data. Third, there was limited evidence of systematic bias in model performance based on student demographics - though this finding should be interpreted cautiously given the relatively small sample sizes.

These findings contribute to our understanding of how AI might support literacy assessment in diverse educational contexts. While previous research has demonstrated the potential of AI for educational assessment in high-resource settings (e.g., Burrows et al., 2015), this dissertation provides some of the first empirical evidence for the feasibility of using these technologies to support reading assessment in anglophone LMICs. The strong performance across different assessment types and student populations suggests that recent advances in AI could enable more regular and comprehensive assessment of reading ability, particularly in contexts where traditional assessment methods are constrained by resource limitations.

7.2 Implications

The findings from these studies have significant implications for both theory and practice in educational assessment, particularly in the context of reading instruction in low and middle-income countries.

7.2.1 Theoretical Implications

This research contributes to our understanding of the capabilities of using NLP and ASR for reading assessment in several ways. First, the strong performance of AI models across different

assessment types challenges previous assumptions about the necessity of task-specific training for automated assessment systems (Mayfield & Black, 2020). Earlier approaches to automated assessment required extensive fine-tuning on task-specific datasets, with models trained separately for different question types, domains, and student populations. The ability of current generation models to perform well across varied assessment tasks with minimal prompt engineering suggests a fundamental shift in how we should understand machine learning capabilities in educational assessment.

Second, these findings contribute to our understanding of different approaches to reading assessment. The research examined three distinct assessment types - oral reading fluency, short answer responses, and story retell tasks - each designed to measure different aspects of reading ability. The varying patterns of correlation between these measures support theoretical models that conceptualize reading comprehension as a multifaceted construct (Nation & Snowling, 1997). More specifically, the strong correlation between automated and human scoring of oral reading fluency supports theories about the objective nature of fluency assessment, while raising new questions about how prosodic elements might be captured through automated means (Fuchs et al., 2001; Kuhn et al., 2010). Additionally, the moderate correlations between story retell measures and other indicators of reading ability provide empirical support for theoretical frameworks that position retell as capturing distinct aspects of comprehension, such as organization and synthesis of information (Shapiro et al., 2014).

Third, the studies highlighted the importance of assessment design in successful automation. In both the short answer and story retell studies, more granular rubrics (three-class and five-class) led to higher agreement levels among both human raters and automated systems compared to binary rubrics. This suggests that well-designed rubrics that capture meaningful gradations in student performance may be particularly amenable to automated scoring.

These findings contribute to broader theoretical discussions about the nature of reading comprehension assessment and how different measurement approaches might complement each other. They suggest that a combination of assessment types, potentially supported by automated scoring, might complement each other in providing a complete picture of student ability (Bellinger & DiPerna, 2011; Reed & Vaughn, 2012).

7.2.2 Practical Implications

The finding that widely available NLP and ASR models can effectively assess different aspects of reading ability without requiring extensive technical expertise or task-specific training data suggests that automated assessment could be more feasible than previously thought. In the context of oral reading fluency assessment, the strong performance of ASR models suggests that automated systems could help scale up regular ORF assessment, which is currently limited by the time-intensive nature of one-on-one administration (Roehrig et al., 2008). Similarly, the ability of LLMs to grade short answer and story retell responses could enable more regular use of open-ended assessment formats, which are often avoided due to grading burden despite their pedagogical benefits (Magliano & Graesser, 2012).

However, implementing these approaches in practice, particularly in lower resource contexts, would require careful consideration of several factors. While the models themselves may be accessible, schools would need appropriate devices and infrastructure for recording and processing student responses, which would be a major impediment in many schools in West Africa. Second, teachers and school staff would need training in how to effectively integrate these tools into their assessment practices (Abadzi, 2008). Additionally, the cost implications of implementing these systems also warrant careful consideration. While the per-response cost of using commercial LLMs has been rapidly decreasing, it could still be a limiting factor when scaled across many students and frequent assessments. Open-source alternatives might offer more cost-effective solutions, though they would require more technical infrastructure to implement (Matelsky et al., 2023). These factors suggest that initial implementation might be most feasible for annual diagnostic assessments, rather than ongoing formative assessment, particularly in low-resource contexts.

7.3 Limitations

7.3.1 Exclusive Focus on English

A fundamental limitation of this research is its exclusive focus on English language assessment. While English is the primary language of instruction in many West African schools, it is often not the first language of students, teachers, or their communities (Spaull et al., 2020). The

linguistic landscape in West Africa is diverse, with many countries having dozens of widely spoken languages, some of which lack standardized written forms, which presents challenges for both assessment design and the application of AI technologies.

From an assessment perspective, evaluating reading comprehension in English may not fully capture students' literacy capabilities, particularly for those who are more fluent in other languages. Research has shown that literacy skills can transfer across languages, but the nature and extent of this transfer varies depending on the linguistic features of the languages involved (Nag, 2017). The validity of reading assessments becomes particularly complex when students are simultaneously developing proficiency in both basic literacy and the language of instruction (Klaas & Trudell, 2011).

From a technological perspective, the strong performance of AI models observed in this research may not generalize to other languages. LLMs and ASR systems have historically shown weaker performance on languages with smaller digital footprints, which typifies many African languages. While recent advances in multilingual models show promise, significant gaps remain (Nouza et al., 2023).

7.3.2 Size and Composition of Datasets

While the ROARS dataset represents a useful contribution to the field, particularly in its inclusion of students from Ghana, the overall number of participants remains relatively small. This limitation potentially impacts the statistical power of the analyses and the generalizability of the findings to broader and different populations, particularly regarding how well the findings might apply to other linguistic, cultural and economic contexts, both within and beyond Ghana. Another related question is how these models would perform with student populations that have fewer low-level readers. The distribution of answers in the ROARS dataset included a substantial number of clearly incorrect answers. It remains unclear how the models would perform in scenarios where most answers are correct or partially correct, and the main task is distinguishing between more subtle differences in response quality.

Another limitation is the limited range of AI models evaluated. While the studies demonstrate strong performance of LLMs and ASR models on specific tasks, it remains uncertain how well these results would translate to other types of reading assessments, different educational levels, or languages other than English. This is a particularly serious concern in linguistically diverse regions such as West-Africa, where model improvement may be slower due to the lack of extensive data and limited incentives for large tech companies to focus on these languages (Thienpondt & Demuynck, 2022). Another important limitation is related to the quality of the audio recordings. The recordings in my datasets were single-speaker, high-quality recordings – however ASR models can have their performance seriously degraded when in interpreting noisy classroom environments and low-quality recordings, which are often the reality in many educational settings (Fan et al., 2022).

7.3.3 Assessment Design

The format of reading comprehension assessments I administered to create the datasets used in the studies, particularly the distinction between multiple-choice (MC) and short answer (SA) items, is a critical consideration as it potentially blurs the distinction between reading and writing abilities. While reading and writing are related skills that draw on shared knowledge, they are separable constructs, and SA reading assessment tasks introduce writing ability as a possible source of performance variability in reading comprehension (Jenkins et al., 2003). The lower absolute levels of learning mean that many students may not be proficient writers, meaning it would be possible for a student with some emergent reading abilities to not demonstrate any basic reading skills that they have or have not attained, making it impossible to identify existing foundational skills that can be built upon to promote learning (Dowd & Pisani, 2013).

In the case of oral reading fluency, based on this prior research and the complications of reliably scoring prosody, I chose to adopt one of the most widely used of measure oral reading fluency or Words Correct Per Minute (WCPM), which is a combined measure of reading accuracy and reading speed, which ignore prosody. As discussed in Chapters 3 and 6, WCPM can be calculated using DIBELS or WER-style protocols, but in both cases the underlying formula subtracts the number of errors from the text length and divides by the time spent

reading. When conducting these assessments, I found the qualitative difference between a student making 1-2 errors and one making 8-10 errors over a 150-word passage was pronounced. However, when calculating WCPM a mere 2-3 second difference in reading rate can offset the impact of these additional errors. This raises the question of whether misreading words should perhaps be given more weight relative to reading rates in fluency measures.

A separate but equally important topic is the perceived face-validity of the results. The structure of LLMs make it difficult to understand why a particular judgment was made, and this lack of interpretability is potentially problematic in educational contexts where fairness, student motivation, and instructional usefulness are paramount concerns. If a student feels a result from a LLM grader is unfair, it is unclear how this could be reviewed or challenged. There is also a possibility that students might take assessments less seriously if they know they are being graded by an AI system, or try to find ways to game the assessment (Baker, 2008).

7.3.4 Specific Educational Context

During the fieldwork for this research, I also noticed several patterns of student behaviour that might have influenced the results. First, students appeared reluctant to ask clarifying questions, even when encouraged to do so. This hesitancy was also observed between students and local teachers involved in the assessment process (i.e., it wasn't simply an issue of students being nervous of foreign researchers). However, students seemed more comfortable seeking clarification from their peers rather than from the researchers or teachers administering the tests. This behaviour raises questions about the cultural and educational norms that might influence test-taking strategies (Kanjee, 2009).

A second common behaviour was direct copying from the passage in students' written responses. While this practice was not explicitly prohibited, it raises questions regarding whether certain patterns of copying correlate with different levels of reading proficiency. For instance, whether stronger readers create their own answers, while weaker readers copy correct or incorrect parts of the story. Understanding these patterns could provide insights into reading comprehension strategies (Nation et al., 2010).

7.4 Directions for Further Research

7.4.1 Investigating Regional Languages

Future research must address the complexities of applying AI-supported assessment in multilingual contexts. A first step would be systematic evaluation of how automated assessment systems perform across different languages, particularly those with distinct linguistic features from English. This research should examine not only model performance but also how different orthographic systems and linguistic structures might require modifications to assessment approaches. For instance, studies might investigate how tonal languages affect the reliability of ASR-based fluency measures.

Additionally, research could investigate how the models' performance varies when assessing students with different language backgrounds. The current study included some students who spoke English at home, but this group was too small for detailed statistical analysis. A focused study comparing model performance across students with different levels of English exposure could help understand how language background influences automated assessment accuracy.

A related line of investigation could examine recordings of students reading English texts while occasionally switching to local languages for specific words or phrases, a pattern I also observed in classroom settings. Understanding how ASR models handle such transitions, and how LLMs interpret responses containing mixed language use, would provide valuable insights for developing more robust assessment systems.

7.4.2 Developing More Comprehensive Datasets

Building on the ROARS dataset, one priority would be collecting additional data from stronger readers to evaluate how well the models distinguish between different levels of proficiency. This is particularly important given that the current dataset includes a substantial number of clearly incorrect answers, making it difficult to assess model performance on more nuanced distinctions between partial and full understanding.

Another valuable extension would involve gathering recordings under varying classroom conditions to understand how environmental factors affect ASR performance. This could be accomplished through a systematic study varying recording conditions (e.g., background noise levels, microphone quality, distance from speaker) while keeping other factors constant. Such research would help establish practical guidelines for implementing automated oral reading assessment in typical classroom environments.

The dataset could also be extended to include more examples of different reading assessment tasks, particularly those requiring longer responses or more complex comprehension tasks. This would help evaluate whether the strong model performance observed on relatively straightforward comprehension questions generalizes to more sophisticated assessment tasks. Such work could begin with existing assessment items but gradually introduce more complex elements to test the boundaries of current model capabilities.

7.4.3 Refining Assessment Design

The preliminary observations about oral versus written responses warrant systematic investigation. A focused study comparing response modalities could help identify whether certain types of comprehension questions are better suited to oral or written responses, and how this might vary based on student proficiency level. This research could also examine whether the mode of response affects the reliability of automated scoring.

Research could also explore methods for improving the interpretability of automated assessments. This might involve having models generate explanatory rationales for their scores, or developing hybrid scoring approaches that combine statistical measures with rule-based components. Such work could help address concerns about the "black box" nature of current automated assessment while potentially providing more actionable feedback for teachers and students.

Further research is also needed to refine measures of oral reading fluency, specifically the relationship between WCPM calculations and reading error patterns. This could be investigated by reanalysing the existing audio recordings with different weighting schemes for accuracy versus speed. This research could explore whether certain types of reading errors (e.g.,

substitutions versus omissions) could be weighted differently when calculating fluency scores, particularly for more advanced readers where current measures may lack sensitivity.

7.4.4 Investigating Cultural Context

Cultural factors affecting assessment performance and interpretation also warrant further investigation. The observed patterns of student behavior in this study—such as reluctance to ask questions and tendencies to copy directly from texts—suggest the need for more detailed cross-cultural studies of assessment-taking strategies. Understanding how different cultural and educational contexts influence student responses could inform both the design of assessment tasks and the development of more culturally responsive automated scoring systems.

The reluctance of students to seek clarification from teachers, despite showing comfort with peer assistance, suggests the need for research into how assessment conditions affect student behavior. A focused study comparing student behavior under different assessment conditions (e.g., with peers versus teachers, or with different types of pre-assessment preparation) could provide insights into how assessment environments might be optimized to better capture student abilities.

7.5 Conclusions

This dissertation demonstrates that recent advances in AI, particularly in Natural Language Processing and Automatic Speech Recognition, have created new possibilities for automating various aspects of reading assessment. The strong performance of these models across different assessment types and student populations suggests that AI-supported assessment could help address some of the practical challenges that have historically limited regular diagnostic and formative assessment of reading ability.

The key contribution of this work is providing empirical evidence for the feasibility of using AI to support reading assessment in a diverse educational context. While previous research has demonstrated the potential of AI for educational assessment in high-resource settings, this dissertation offers some of the first evidence that these technologies could be viable for

supporting reading assessment in lower-resource contexts, more specifically in anglophone West Africa. However, realizing this potential will require careful attention to both technical and practical considerations. While the models demonstrated strong performance, successful implementation will depend on appropriate infrastructure, clear protocols for data handling, and effective training for educators. Furthermore, ongoing research will be crucial for understanding how these technologies perform across different linguistic, educational, and economic contexts, and how they can best be integrated into existing assessment practices.

Looking forward, the rapid pace of advancement in AI technology suggests that capabilities will continue to improve, potentially enabling even more sophisticated approaches to automated assessment. If successful, this can enable more regular and comprehensive assessment while freeing up educator time for the complex work of interpreting results and adjusting instruction. By demonstrating the current capabilities of AI in supporting reading assessment while highlighting areas requiring further investigation, this research contributes to our understanding of how artificial intelligence might help make regular reading assessment more feasible in low-resource educational settings.

Chapter 8 - References

Alderson Abadzi, H. (2008). Efficient learning for the poor: New insights into literacy acquisition for children. *International Review of Education, 54*(5-6), 581-604. https://doi.org/10.1007/s11159-008-9102-3

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.

Alderson, J. C., Huhta, A., & Nieminen, L. (2016). Characteristics of weak and strong readers in a foreign language. *The Modern Language Journal, 100*(4), 853-879. https://doi.org/10.1111/modl.12367

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

Alikaniotis, D., Yannakoudakis, H., & Rei, M. (2016). Automatic text scoring using neural networks. *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 715-725. https://doi.org/10.18653/v1/P16-1068

August, D., Francis, D. J., Hsu, H. A., & Snow, C. E. (2006). Assessing reading comprehension in bilinguals. *The Elementary School Journal, 107*(2), 221-238. https://doi.org/10.1086/510656

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. (2008). Why students engage in "gaming the system" behavior in interactive learning environments. *Journal of Interactive Learning Research, 19*(2), 185-224.

Baker, R. S., & Hawn, A. (2022). Algorithmic bias in education. *International Journal of Artificial Intelligence in Education, 32*(4), 1052-1092. https://doi.org/10.1007/s40593-021-00285-9

Banerjee, M., Capozzoli, M., Mcsweeney, L., & Sinha, D. (2008). Beyond kappa: A review of interrater agreement measures. *Canadian Journal of Statistics, 27*, 3-23. https://doi.org/10.2307/3315487

Banerji, R. (2016). Annual Status of Education Report (ASER) assessment in India: Fast, rigorous and frugal. In *Understanding what works in oral reading assessments: Recommendations from donors, implementers and practitioners* (pp. 187-201). UNESCO Institute for Statistics.

Banerji, R. (2020). Interview with Rukmini Banerji regarding oral reading fluency and the ASER assessment [In Person].

Bartlett, L., Dowd, A. J., & Jonason, C. (2015). Problematizing early grade reading: Should the post-2015 agenda treasure what is measured? *International Journal of Educational Development, 40*, 308-314. https://doi.org/10.1016/j.ijedudev.2014.10.002

Bejar, I. I. (1984). Educational diagnostic assessment. *Journal of Educational Measurement, 21*(2), 175-189. https://doi.org/10.1111/j.1745-3984.1984.tb00228.x

Bellinger, J. M., & DiPerna, J. C. (2011). Is fluency-based story retell a good indicator of reading comprehension? *Psychology in the Schools, 48*(4), 416-426. https://doi.org/10.1002/pits.20563

Belur, J., Tompson, L., Thornton, A., & Simon, M. (2018). Interrater reliability in systematic review methodology: Exploring variation in coder decision-making. *Sociological Methods & Research, 50*, 004912411879937. https://doi.org/10.1177/0049124118799372

Benjamin, R. G., & Schwanenflugel, P. J. (2010). Text complexity and oral reading prosody in young readers. *Reading Research Quarterly, 45*(4), 388-404. https://doi.org/10.1598/RRQ.45.4.2

Bennett, R. E. (2011). Formative assessment: A critical review. *Assessment in Education: Principles, Policy & Practice, 18*(1), 5-25. https://doi.org/10.1080/0969594X.2010.513678

Bishop, & Adams, C. (1990). A prospective study of the relationship between specific language impairment, phonological disorders and reading retardation. *Journal of Child Psychology and Psychiatry, 31*(7), 1027-1050. https://doi.org/10.1111/j.1469-7610.1990.tb00844.x

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

Blanchard, G., Lee, G., & Scott, C. (2011). Generalizing from several related classification tasks to a new unlabeled sample. *Advances in Neural Information Processing Systems 24 (NIPS 2011)*.

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

Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M. S., Bohg, J., Bosselut, A., Brunskill, E., Brynjolfsson, E., Buch, S., Card, D., Castellon, R., Chatterji, N., Chen, A., Creel, K., Davis, J. Q., Demszky, D., … Liang, P. (2022). On the opportunities and risks of foundation models (arXiv:2108.07258). *arXiv*. https://doi.org/10.48550/arXiv.2108.07258

Braun, H., Kanjee, A., Bettinger, E., & Kremer, M. (2006). *Improving education through assessment, innovation, and evaluation*. American Academy of Arts and Sciences.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., … Amodei, D. (2020).

Language models are few-shot learners. *arXiv:2005.14165 [Cs]*. http://arxiv.org/abs/2005.14165

Burns, P. C., & Roe, B. D. (2007). *Informal reading inventory: Preprimer to twelfth grade* (6th ed.). Houghton Mifflin.

Burrows, S., Gurevych, I., & Stein, B. (2015). The eras and trends of automatic short answer grading. *International Journal of Artificial Intelligence in Education, 25*(1), 60-117. https://doi.org/10.1007/s40593-014-0026-8

Cain, K., & Oakhill, J. (2006). Assessment matters: Issues in the measurement of reading comprehension. *British Journal of Educational Psychology, 76*(4), 697-708. https://doi.org/10.1348/000709905X69807

Caines, A., Benedetto, L., Taslimipoor, S., Davis, C., Gao, Y., Andersen, O., Yuan, Z., Elliott, M., Moore, R., Bryant, C., Rei, M., Yannakoudakis, H., Mullooly, A., Nicholls, D., & Buttery, P. (2023). On the application of Large Language Models for language teaching and assessment technology (arXiv:2307.08393). *arXiv*. http://arxiv.org/abs/2307.08393

Camus, L., & Filighera, A. (2020). Investigating transformers for automatic short answer grading. In I. I. Bittencourt, M. Cukurova, K. Muldner, R. Luckin, & E. Millán (Eds.), *Artificial Intelligence in Education* (Vol. 12164, pp. 43-48). Springer International Publishing. https://doi.org/10.1007/978-3-030-52240-7_8

Carver, R. P. (1998). Predicting reading level in grades 1 to 6 from listening level and decoding level: Testing theory relevant to the simple view of reading. *Reading and Writing: An Interdisciplinary Journal, 10*, 35.

Castles, A., & Coltheart, M. (2004). Is there a causal link from phonological awareness to success in learning to read? *Cognition, 91*(1), 77-111. https://doi.org/10.1016/S0010-0277(03)00164-1

Castles, A., Rastle, K., & Nation, K. (2018). Ending the reading wars: Reading acquisition from novice to expert. *Psychological Science in the Public Interest, 19*(1), 5-51. https://doi.org/10.1177/1529100618772271

Catts, H. W., Adlof, S. M., & Weismer, S. E. (2006). Language deficits in poor comprehenders: A case for the simple view of reading. *Journal of Speech, Language, and Hearing Research, 49*(2), 278-293. https://doi.org/10.1044/1092-4388(2006/023)

Chall, J. (1983). *Stages of reading development*. McGraw-Hill.

Cheng, J., Chen, X., & Metallinou, A. (2015). Deep neural network acoustic models for spoken assessment applications. *Speech Communication, 73*, 14-27. https://doi.org/10.1016/j.specom.2015.07.006

Chowdhery, A., Narang, S., Devlin, J., Bosma, M., Mishra, G., Roberts, A., Barham, P., Chung, H. W., Sutton, C., Gehrmann, S., Schuh, P., Shi, K., Tsvyashchenko, S., Maynez, J., Rao, A., Barnes, P., Tay, Y., Shazeer, N., Prabhakaran, V., … Salakhutdinov, R. (2023). PaLM: Scaling language modeling with pathways.

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), 65-88. https://doi.org/10.1016/j.csl.2008.03.001

Claus, F., Gamboa-Rosales, H., Petrick, R., Hain, H.-U., & Hoffmann, R. (2013, August 25). A survey about databases of children's speech. https://doi.org/10.21437/Interspeech.2013-561

Cohn, C., Hutchins, N., Le, T., & Biswas, G. (2024). A chain-of-thought prompting approach with LLMs for evaluating students' formative assessment responses in science (arXiv:2403.14565). *arXiv*. http://arxiv.org/abs/2403.14565

Collins, A. A., Compton, D. L., Lindström, E. R., & Gilbert, J. K. (2020). Performance variations across reading comprehension assessments: Examining the unique

contributions of text, activity, and reader. *Reading and Writing, 33*(3), 605-634. https://doi.org/10.1007/s11145-019-09972-5

Cowie, R., Douglas-Cowie, E., & Wichmann, A. (2002). Prosodic characteristics of skilled reading- Fluency and expressiveness in 8-10-year-old readers. *Language and Speech, 45*(1), 47-82.

Cunningham, A. E., & Stanovich, K. E. (1997). Early reading acquisition and its relation to reading experience and ability 10 years later. *Developmental Psychology, No. 33*, 12. https://doi.org/10.1037//0012-1649.33.6.934.

Daane, M. C., Campbell, J. R., Grigg, W. S., Goodman, M. J., & Oranje, A. (2002). *The Nation's Report Card: Fourth-grade students reading aloud: NAEP 2002 special study of oral reading* (p. 84).

De Raadt, A., Warrens, M. J., Bosker, R. J., & Kiers, H. A. L. (2021). A comparison of reliability coefficients for ordinal rating scales. *Journal of Classification, 38*(3), 509-528. https://doi.org/10.1007/s00357-021-09386-5

Deno, S. L., Fuchs, L. S., Marston, D., & Shin, J. (2001). Using curriculum-based measurement to establish growth standards for students with learning disabilities. *School Psychology Review, 30*.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv:1810.04805 [Cs]*. http://arxiv.org/abs/1810.04805

DiCerbo, K. D., Lai, E., & Ventura, M. (2020). Assessment design with automated scoring in mind. In *Handbook of automated scoring: Theory into practice*.

Dowd, A. J., & Pisani, L. (2013). Two wheels are better than one: The importance of capturing the home literacy environment in large-scale assessments of reading. *Research in Comparative and International Education, 8*(3), 359-372. https://doi.org/10.2304/rcie.2013.8.3.359

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

Dumais, S. T. (2005). Latent semantic analysis. *Annual Review of Information Science and Technology, 38*(1), 188-230. https://doi.org/10.1002/aris.1440380105

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

Ehri, L. C. (1995). Phases of development in learning to read words by sight. *Journal of Research in Reading, 18*(2), 116-125. https://doi.org/10.1111/j.1467-9817.1995.tb00077.x

El Dehaibi, N., & MacDonald, E. F. (2020). Investigating inter-rater reliability of qualitative text annotations in machine learning datasets. *Proceedings of the Design Society: DESIGN Conference, 1*, 21-30. https://doi.org/10.1017/dsd.2020.153

Elsahar, H., & Gallé, M. (2019). To annotate or not? Predicting performance drop under domain shift. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 2163-2173. https://doi.org/10.18653/v1/D19-1222

Fan, R., & Alwan, A. (2022). DRAFT: A novel framework to reduce domain shifting in self-supervised learning and its application to children's ASR. *Interspeech 2022*, 4900-4904. https://doi.org/10.21437/Interspeech.2022-11128

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), 1281-1293. 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

Fernandez, N., Ghosh, A., Liu, N., Wang, Z., Choffin, B., Baraniuk, R., & Lan, A. (2023). Automated scoring for reading comprehension via in-context BERT tuning (arXiv:2205.09864). *arXiv*. http://arxiv.org/abs/2205.09864

Foltz, P. W., Gilliam, S., & Kendall, S. (2000). Supporting content-based feedback in on-line writing evaluation with LSA. *Interactive Learning Environments, 8*(2), 111-127. https://doi.org/10.1076/1049-4820(200008)8:2;1-B;FT111

Foltz, P., Rupp, A. A., & Yan, D. (2020). The past, present, and future of automated scoring. In *Handbook of automated scoring*. CRC Press.

Foy, P., & Drucker, K. T. (Eds.). (2013). *PIRLS 2011 user guide for the international database: prePIRLS released passages and items*. TIMSS & PIRLS International Study Center, Lynch School of Education, Boston College; International Association for the Evaluation of Educational Achievement (IEA).

Frith, U. (2017). Beneath the surface of developmental dyslexia. In K. E. Patterson, J. C. Marshall, & M. Coltheart (Eds.), *Surface dyslexia* (1st ed., pp. 301-330). Routledge. https://doi.org/10.4324/9781315108346-18

Fuchs, L. S., Fuchs, D., & Maxwell, L. (1988). The validity of informal reading comprehension measures. *Remedial and Special Education, 9*(2), 20-28. https://doi.org/10.1177/074193258800900206

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*, 1-8. https://doi.org/10.1145/1640377.1640384

Geva, E., & Massey-Garrison, A. (2013). A comparison of the language skills of ELLs and monolinguals who are poor decoders, poor comprehenders, or normal readers. *Journal of Learning Disabilities, 46*(5), 387-401. https://doi.org/10.1177/0022219412466651

Geva, E., & Siegel, L. S. (2000). Orthographic and cognitive factors in the concurrent development of basic reading skills in two languages. *Reading and Writing: An Interdisciplinary Journal*.

Gilardi, F., Alizadeh, M., & Kubli, M. (2023). ChatGPT outperforms crowd workers for text-annotation tasks. *Proceedings of the National Academy of Sciences, 120*(30), e2305016120. https://doi.org/10.1073/pnas.2305016120

Good, R. H., Simmons, D. C., & Kame'enui, E. J. (2001). The importance and decision-making utility of a continuum of fluency-based indicators of foundational reading skills for third-grade high-stakes outcomes. *Scientific Studies of Reading, 5*(3), 257-288. https://doi.org/10.1207/S1532799XSSR0503_4

Gough, P. B., & Tunmer, W. E. (1986). Decoding, reading, and reading disability. *Remedial and Special Education, 7*(1), 6-10. https://doi.org/10.1177/074193258600700104

Gove, A., & Cvelich, P. (2011). *Early reading: Igniting education for all: A report by the Early Grade Learning Community of Practice*. RTI International.

Graesser, A. C., Wiemer-Hastings, P., Wiemer-Hastings, K., Harter, D., Tutoring Research Group, T. R. G., & Person, N. (2000). Using latent semantic analysis to evaluate the

contributions of students in AutoTutor. *Interactive Learning Environments, 8*(2), 129-147. https://doi.org/10.1076/1049-4820(200008)8:2;1-B;FT129

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

Gregory, R. J. (2007). *Psychology testing: History, principles, and applications* (5th Edition). Ally & Bacon.

Hagtvet, B. E. (2003). Listening comprehension and reading comprehension in poor decoders: Evidence for the importance of syntactic and semantic skills as well as phonological skills. *Reading and Writing: An Interdisciplinary Journal, 16*(6), 505-539. https://doi.org/10.1023/A:1025521722900

Haller, S., Aldea, A., Seifert, C., & Strisciuglio, N. (2022). Survey on automated short answer grading with deep learning: From word embeddings to transformers (arXiv:2204.03503). *arXiv*. http://arxiv.org/abs/2204.03503

Hallgren, K. (2012). Computing inter-rater reliability for observational data: An overview and tutorial. *Tutorials in Quantitative Methods for Psychology, 8*, 23-34. https://doi.org/10.20982/tqmp.08.1.p023

Hanley, R., Masterson, J., Spencer, L., & Evans, D. (2004). How long do the advantages of learning to read a transparent orthography last? An investigation of the reading skills and reading impairment of Welsh children at 10 years of age. *The Quarterly Journal of Experimental Psychology Section A, 57*(8), 1393-1410. https://doi.org/10.1080/02724980343000819

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

Head, M. H., Readence, J. E., & Buss, R. R. (1989). An examination of summary writing as a measure of reading comprehension. *Reading Research and Instruction, 28*(4), 1-11. https://doi.org/10.1080/19388078909557982

Henkel, O., Hills, L., Roberts, B., & McGrane, J. (2024). Can LLMs Grade Open Response reading comprehension questions? An empirical study using the ROARs dataset. *International journal of artificial intelligence in education*, 1-26

Henkel, O., Horne-Robinson, H., Hills, L., Roberts, B., & McGrane, J. (2025). Supporting Literacy Assessment in West Africa: Using State-of-the-Art Speech Models to Assess Oral Reading Fluency. *International Journal of Artificial Intelligence in Education*, 1-22.

Holmlund, T. B., Chandler, C., Foltz, P. W., Cohen, A. S., Cheng, J., Bernstein, J. C., Rosenfeld, E. P., & Elvevåg, B. (2020). Applying speech technologies to assess verbal memory in patients with serious mental illness. *Npj Digital Medicine, 3*(1), 1-8. https://doi.org/10.1038/s41746-020-0241-7

Hoover, A., & Gough, B. (1990). The simple view of reading. *Reading and Writing, Vol 2*(2), 34.

Hulme, C., & Snowling, M. J. (2011). Children's reading comprehension difficulties: Nature, causes, and treatments. *Current Directions in Psychological Science, 20*(3), 139-142. https://doi.org/10.1177/0963721411408673

Jain, R., Barcovschi, A., Yiwere, M., Corcoran, P., & Cucu, H. (2023a). Adaptation of Whisper models to child speech recognition. *INTERSPEECH 2023*, 5242-5246. https://doi.org/10.21437/Interspeech.2023-935

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

Jenkins, J. R., Fuchs, L. S., van den Broek, P., Espin, C., & Deno, S. L. (2003). Sources of individual differences in reading comprehension and reading fluency. *Journal of Educational Psychology, 95*(4), 719-729. https://doi.org/10.1037/0022-0663.95.4.719

Jenkins, J. R., Johnson, E., & Hileman, J. (2004). When is reading also writing: Sources of individual differences on the new reading performance assessments. *Scientific Studies of Reading, 8*(2), 125-151. https://doi.org/10.1207/s1532799xssr0802_2

Jiang, W., & Pardos, Z. A. (2021). Towards equity and algorithmic fairness in student grade prediction. *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, 608-617. https://doi.org/10.1145/3461702.3462623

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

Joulin, A., Grave, E., Bojanowski, P., & Mikolov, T. (2017). Bag of tricks for efficient text classification. *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, 5.

Kanjee, A. (2009). Enhancing teacher assessment practices in South African schools: Evaluation of the assessment resource banks. *Education as Change, 13*(1), 73-89. https://doi.org/10.1080/16823200902940599

Keenan, J. M., Betjemann, R. S., & Olson, R. K. (2008). Reading comprehension tests vary in the skills they assess: Differential dependence on decoding and oral comprehension. *Scientific Studies of Reading, 12*(3), 281-300. https://doi.org/10.1080/10888430802132279

Kim, Y.-S., Petscher, Y., Schatschneider, C., & Foorman, B. (2010). Does growth rate in oral reading fluency matter in predicting reading comprehension achievement? *Journal of Educational Psychology, 102*(3), 652-667. https://doi.org/10.1037/a0019643

Kintsch, W. (1988). The role of knowledge in discourse comprehension: a construction-integration model. *Psychological review*, *95*(2), 163.

Kintsch, W., & van Dijk, T. A. (1978). Toward a model of text comprehension and production. *Psychological Review*.

Kirsch, I., & Guthrie, J. T. (1977). The concept and measurement of functional literacy. *Reading Research Quarterly, 13*(4), 485-507. https://doi.org/10.2307/747509

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

Klaas, A., & Trudell, B. (2011). Effective literacy programmes and independent reading in African contexts. *Language Matters, 42*(1), 22-38. https://doi.org/10.1080/10228195.2011.569739

Klesius, J. P., & Homan, S. P. (1985). A validity and reliability update on the informal reading inventory with suggestions for improvement. *Journal of Learning Disabilities, 18*(2), 71-76. https://doi.org/10.1177/002221948501800202

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

Kojima, T., Gu, S. S., Reid, M., Matsuo, Y., & Iwasawa, Y. (2022). Large language models are zero-shot reasoners.

Kormos, J., Košak Babuder, M., & Pižorn, K. (2019). The role of low-level first language skills in second language reading, reading-while-listening and listening performance: A study of young dyslexic and non-dyslexic language learners. *Applied Linguistics, 40*(5), 834-858. https://doi.org/10.1093/applin/amy028

Kortemeyer, G. (2023). Performance of the pre-trained large language model GPT-4 on automated short answer grading (arXiv:2309.09338). *arXiv*. http://arxiv.org/abs/2309.09338

Krippendorff, K. (2011). Computing Krippendorff's alpha-reliability.

Kuhn, M. R., Schwanenflugel, P. J., Meisinger, E. B., Levy, B. A., & Rasinski, T. V. (2010). Aligning theory and assessment of reading fluency: Automaticity, prosody, and definitions of fluency. *Reading Research Quarterly, 45*(2), 230-251. https://doi.org/10.1598/RRQ.45.2.4

Kuzman, T., Mozetič, I., & Ljubešić, N. (2023). ChatGPT: Beginning of an end of manual linguistic data annotation? Use case of automatic genre identification (arXiv:2303.03953). *arXiv*. http://arxiv.org/abs/2303.03953

LaBerge, D., & Samuels, J. (1974). Toward a theory of automatic information processing in reading. *Cognitive Psychology, 6*.

Landauer, T. K., Foltz, P. W., & Laham, D. (1998). An introduction to latent semantic analysis. *Discourse Processes, 25*(2-3), 259-284. https://doi.org/10.1080/01638539809545028

Landauer, T. K., Lochbaum, K. E., & Dooley, S. (2009). A new formative assessment technology for reading and writing. *Theory Into Practice, 48*(1), 44-52. https://doi.org/10.1080/00405840802577593

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33*(1), 159-174. https://doi.org/10.2307/2529310

Lervåg, A., Hulme, C., & Melby-Lervåg, M. (2018). Unpicking the developmental relationship between oral language skills and reading comprehension: It's simple, but complex. *Child Development, 89*(5), 1821-1838. https://doi.org/10.1111/cdev.12861

Lockheed, M. (2008). Measuring progress with tests of learning: Pros and cons for "cash on delivery aid" in education. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.1213162

Loukina, A., Zechner, K., & Chen, L. (2014). Automatic evaluation of spoken summaries: The case of language assessment. *Proceedings of the Ninth Workshop on Innovative Use of NLP for Building Educational Applications*, 68-78. https://doi.org/10.3115/v1/W14-1809

Ma, W., Adesope, O. O., Nesbit, J. C., & Liu, Q. (2014). Intelligent tutoring systems and learning outcomes: A meta-analysis. *Journal of Educational Psychology, 106*(4), 901-918. https://doi.org/10.1037/a0037123

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

Magliano, J. P., & Millis, K. K. (2003). Assessing reading skill with a think-aloud procedure and latent semantic analysis. *Cognition and Instruction, 21*(3), 251-283. https://doi.org/10.1207/S1532690XCI2103_02

Marcotte, A. M., & Hintze, J. M. (2009). Incremental and predictive utility of formative assessment methods of reading comprehension. *Journal of School Psychology, 47*(5), 315-335. https://doi.org/10.1016/j.jsp.2009.04.003

Maria, K. (1990). *Reading comprehension instruction: Issues and strategies*. York Press.

Matelsky, J. K., Parodi, F., Liu, T., Lange, R. D., & Kording, K. P. (2023). A large language model-assisted education tool to provide feedback on open-ended responses (arXiv:2308.02439). *arXiv*. http://arxiv.org/abs/2308.02439

Mayfield, E., & Black, A. W. (2020). Should you fine-tune BERT for automated essay scoring? *Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications*, 151-162. https://doi.org/10.18653/v1/2020.bea-1.15

McArthur, D., Lewis, M., & Bishay, M. (2005). The roles of artificial intelligence in education. *Journal of Educational Technology, 1*.

McNamara, D. S., Allen, L. K., Crossley, S. A., Dascalu, M., & Perret, C. A. (2017). Chapter 8: Natural language processing and learning analytics. In *Handbook of Learning Analytics*.

McNamara, D. S., Louwerse, M. M., McCarthy, P. M., & Graesser, A. C. (2010). Coh-Metrix: Capturing linguistic features of cohesion. *Discourse Processes, 47*(4), 292-330. https://doi.org/10.1080/01638530902959943

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.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv:1301.3781 [Cs]*. http://arxiv.org/abs/1301.3781

Miller, J., & Schwanenflugel, P. J. (2008). A longitudinal study of the development of reading prosody as a dimension of oral reading fluency in early elementary school children. *Reading Research Quarterly, 43*(4), 336-354. https://doi.org/10.1598/RRQ.43.4.2

Mishra, S., Khashabi, D., Baral, C., & Hajishirzi, H. (2022). Cross-task generalization via natural language crowdsourcing instructions (arXiv:2104.08773). *arXiv*. http://arxiv.org/abs/2104.08773

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

Mizumoto, A., & Eguchi, M. (2023). Exploring the potential of using an AI language model for automated essay scoring. *Research Methods in Applied Linguistics, 2*(2), 100050. https://doi.org/10.1016/j.rmal.2023.100050

Morjaria, L., Burns, L., Bracken, K., Levinson, A. J., Ngo, Q. N., Lee, M., & Sibbald, M. (2024). Examining the efficacy of ChatGPT in marking short-answer assessments in an

undergraduate medical program. *International Medical Education, 3*(1), 32-43. https://doi.org/10.3390/ime3010004

Mueller, J., & Thyagarajan, A. (2016). Siamese recurrent architectures for learning sentence similarity. 7.

Nag, S. (2017). *Assessment of literacy and foundational learning in developing countries*. HEART (Health & Education Advice & Resource Team).

Nangia, N., Sugawara, S., Trivedi, H., Warstadt, A., Vania, C., & Bowman, S. R. (2021). What ingredients make for an effective crowdsourcing protocol for difficult NLU data collection tasks? (arXiv:2106.00794). *arXiv*. http://arxiv.org/abs/2106.00794

Nation, K., & Snowling, M. (1997). Assessing reading difficulties: The validity and utility of current measures of reading skill. *British Journal of Educational Psychology, 67*(3), 359-370. https://doi.org/10.1111/j.2044-8279.1997.tb01250.x

Nation, K., Allen, R., & Hulme, C. (2001). The limitations of orthographic analogy in early reading development: Performance on the clue-word task depends on phonological priming and elementary decoding skill, not the use of orthographic analogy. *Journal of Experimental Child Psychology, 80*(1), 75-94. https://doi.org/10.1006/jecp.2000.2614

Nation, K., Cocksey, J., Taylor, J. S. H., & Bishop, D. V. M. (2010). A longitudinal investigation of early reading and language skills in children with poor reading comprehension: A longitudinal investigation of early reading and language skills. *Journal of Child Psychology and Psychiatry, 51*(9), 1031-1039. https://doi.org/10.1111/j.1469-7610.2010.02254.x

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

OECD. (2018). PISA for Development Reading Framework. In OECD, *PISA for Development Assessment and Analytical Framework* (pp. 25-48). OECD. https://doi.org/10.1787/9789264305274-4-en

Ouellette, G. P. (2006). What's meaning got to do with it: The role of vocabulary in word reading and reading comprehension. *Journal of Educational Psychology, 98*(3), 554-566. https://doi.org/10.1037/0022-0663.98.3.554

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P., Leike, J., & Lowe, R. (2022). Training language models to follow instructions with human feedback (arXiv:2203.02155). *arXiv*. https://doi.org/10.48550/arXiv.2203.02155

Page, E. (1966). The imminence of... grading essays by computer. *The Phi Delta Kappan, 7*.

Pavlick, E., & Kwiatkowski, T. (2019). Inherent disagreements in human textual inferences. *Transactions of the Association for Computational Linguistics, 7*, 677-694. https://doi.org/10.1162/tacl_a_00293

Pearson, P. D., & Hamm, D. N. (2006). The assessment of reading comprehension: A review of practices—Past, present, and future. In *Children's reading comprehension and assessment*. Lawrence Erlbaum Associates.

Pennington, J., Socher, R., & Manning, C. (2014). Glove: Global vectors for word representation. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1532-1543. https://doi.org/10.3115/v1/D14-1162

Perez, E., Kiela, D., & Cho, K. (2021). True few-shot learning with language models.

Perfetti, C. A., & Hogaboam, T. (1975). Relationship between single word decoding and reading comprehension skill. *Journal of Educational Psychology, 67*(4), 461-469. https://doi.org/10.1037/h0077013

Perfetti, C., & Adlof, S. M. (2012). Reading comprehension: A conceptual framework from word meaning to text meaning. In *Measuring Up: Advances in How We Assess Reading Ability*.

Perin, D., & Lauterbach, M. (2018). Assessing text-based writing of low-skilled college students. *International Journal of Artificial Intelligence in Education, 28*(1), 56-78. https://doi.org/10.1007/s40593-016-0122-z

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. (2018). Deep contextualized word representations. *arXiv:1802.05365 [Cs]*. http://arxiv.org/abs/1802.05365

Piper, B., Schroeder, L., & Trudell, B. (2016). Oral reading fluency and comprehension in Kenya: Reading acquisition in a multilingual environment: Fluency and comprehension in Kenya. *Journal of Research in Reading, 39*(2), 133-152. https://doi.org/10.1111/1467-9817.12052

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

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., & Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv:1910.10683 [Cs, Stat]*. http://arxiv.org/abs/1910.10683

Rasinki, T. (2004). What research says about reading. *Educational Leadership, 61*(6), 46-51.

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

Reed, D. K. (2011). A review of the psychometric properties of retell instruments. *Educational Assessment, 16*(3), 123-144. https://doi.org/10.1080/10627197.2011.604238

Reed, D. K., & Vaughn, S. (2012). Retell as an indicator of reading comprehension. *Scientific Studies of Reading, 16*(3), 187-217. https://doi.org/10.1080/10888438.2010.538780

Reed, D. K., Vaughn, S., & Petscher, Y. (2012). The validity of a holistically scored retell protocol for determining the reading comprehension of middle school students. *Learning Disability Quarterly, 35*(2), 76-89. https://doi.org/10.1177/0731948711432509

Ridley, R., He, L., Dai, X., Huang, S., & Chen, J. (2020). Prompt agnostic essay scorer: A domain generalization approach to cross-prompt automated essay scoring (arXiv:2008.01441). *arXiv*. http://arxiv.org/abs/2008.01441

Riedel, B. W. (2007). The relation between DIBELS, reading comprehension, and vocabulary in urban first-grade students. *Reading Research Quarterly, 42*(4), 546-567. https://doi.org/10.1598/RRQ.42.4.5

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

Schneider, J., Schenk, B., Niklaus, C., & Vlachos, M. (2023). Towards LLM-based autograding for short textual answers.

Schwanenflugel, P. J., Hamilton, A. M., Kuhn, M. R., Wisenbaker, J. M., & Stahl, S. A. (2004). Becoming a fluent reader: Reading skill and prosodic features in the oral reading of young readers. *Journal of Educational Psychology, 96*(1), 119-129. https://doi.org/10.1037/0022-0663.96.1.119

Shapiro, E. S., Fritschmann, N. S., Thomas, L. B., Hughes, C. L., & McDougal, J. (2014). Concurrent and predictive validity of reading retell as a brief measure of reading

comprehension for narrative text. *Reading Psychology, 35*(7), 644-665. https://doi.org/10.1080/02702711.2013.790328

Share, D. L. (2008). On the Anglocentricities of current reading research and practice: The perils of overreliance on an "outlier" orthography. *Psychological Bulletin, 134*(4), 584-615. https://doi.org/10.1037/0033-2909.134.4.584

Shermis, M., Burstein, J., Higgins, D., & Zechner, K. (2010). Automated essay scoring: Writing assessment and instruction. In *International Encyclopedia of Education* (pp. 20-26). Elsevier. https://doi.org/10.1016/B978-0-08-044894-7.00233-5

Shute, V. J. (2008). Focus on formative feedback. *Review of Educational Research, 78*(1), 153-189. https://doi.org/10.3102/0034654307313795

Smith, G., & Paige, D. (2019). A study of interrater reliability across multiple raters when using the NAEP and MDFS rubrics to measure oral reading fluency. *Reading Psychology, 40*, 1-33. https://doi.org/10.1080/02702711.2018.1555361

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), 1-14. 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

Spirgel, A. S., & Delaney, P. F. (2016). Does writing summaries improve memory for text? *Educational Psychology Review*. https://doi.org/10.1007/s10648-014-9290-2

Spooner, A. L. R., Baddeley, A. D., & Gathercole, S. E. (2004). Can reading accuracy and comprehension be separated in the Neale Analysis of Reading Ability? *British Journal of Educational Psychology, 74*(2), 187-204. https://doi.org/10.1348/000709904773839833

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

Sultan, M. A., Sil, A., & Florian, R. (2022). Not to overfit or underfit the source domains? An empirical study of domain generalization in question answering. *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 3752-3761. https://doi.org/10.18653/v1/2022.emnlp-main.247

Sung, C., Dhamecha, T., Saha, S., Ma, T., Reddy, V., & Arora, R. (2019). Pre-training BERT on domain resources for short answer grading. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 6070-6074. https://doi.org/10.18653/v1/D19-1628

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*, 1-9. https://doi.org/10.1145/3465416.3483305

Taghipour, K., & Ng, H. T. (2016). A neural approach to automated essay scoring. *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 1882-1891. https://doi.org/10.18653/v1/D16-1193

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

Thienpondt, J., & Demuynck, K. (2022). Transfer learning for robust low-resource children's speech ASR with transformers and source-filter warping. *Interspeech 2022*, 2213-2217. https://doi.org/10.21437/Interspeech.2022-10964

Torfi, A., Shirvani, R. A., Keneshloo, Y., Tavaf, N., & Fox, E. A. (2020). Natural language processing advancements by deep learning: A survey. *arXiv:2003.01200 [Cs]*. http://arxiv.org/abs/2003.01200

Ullmann, T. (2019). Automated analysis of reflection in writing: Validating machine learning approaches. *International Journal of Artificial Intelligence in Education, 29*. https://doi.org/10.1007/s40593-019-00174-2

UNESCO Institute for Statistics. (2016). *Understanding what works in oral reading assessments*. UNESCO Institute for Statistics.

Valencia, S. W., Smith, A. T., Reece, A. M., Li, M., Wixson, K. K., & Newman, H. (2010). Oral reading fluency assessment: Issues of construct, criterion, and consequential validity. *Reading Research Quarterly, 45*(3), 270-291. https://doi.org/10.1598/RRQ.45.3.1

Van den Bergh, H. (1990). On the construct validity of multiple-choice items for reading comprehension. *Applied Psychological Measurement, 14*(1), 1-12. https://doi.org/10.1177/014662169001400101

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. arXiv:1706.03762 [Cs]. 31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach California. http://arxiv.org/abs/1706.03762

Vellutino, F. R., Tunmer, W. E., Jaccard, J. J., & Chen, R. (2007). Components of reading ability: Multivariate evidence for a convergent skills model of reading development. *Scientific Studies of Reading, 11*(1), 3-32. https://doi.org/10.1080/10888430709336632

Wagner, D. A., Lockheed, M., Mullis, I., Martin, M. O., Kanjee, A., Gove, A., & Dowd, A. J. (2012). The debate on learning assessments in developing countries. *Compare: A Journal of Comparative and International Education, 42*(3), 509-545. https://doi.org/10.1080/03057925.2012.670480

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

Weidinger, L., Uesato, J., Rauh, M., Griffin, C., Huang, P.-S., Mellor, J., Glaese, A., Cheng, M., Balle, B., Kasirzadeh, A., Biles, C., Brown, S., Kenton, Z., Hawkins, W., Stepleton, T., Birhane, A., Hendricks, L. A., Rimell, L., Isaac, W., … Gabriel, I. (2022). Taxonomy of risks posed by language models. *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, 214-229. https://doi.org/10.1145/3531146.3533088

Wilson, R. M., Gambrell, L. B., & Pfeiffer, W. R. (1985). The effects of retelling upon reading comprehension and recall of text information. *The Journal of Educational Research, 78*(4), 216-220. https://doi.org/10.1080/00220671.1985.10885604

Wimmer, H., & Goswami, U. (1994). The influence of orthographic consistency on reading development: Word recognition in English and German children. *Cognition, 51*(1), 91-103. https://doi.org/10.1016/0010-0277(94)90010-8

Winograd, P. N. (1984). Strategic difficulties in summarizing texts. *Reading Research Quarterly, 19*(4), 404-425. https://doi.org/10.2307/747913

Wolf, M., & Katzir-Cohen, T. (2001). Reading fluency and its intervention. *Scientific Studies of Reading, 5*(3), 211-239. https://doi.org/10.1207/S1532799XSSR0503_2

Wolf, M., Bally, H., & Morris, R. D. (1986). Automaticity, retrieval processes, and reading: A longitudinal study in average and impaired readers. *Child Development*, 14.

Woods, M. L., & Moe, A. J. (2007). *Analytical reading inventory: Comprehensive standards-based assessment for all students including gifted and remedial*. Pearson/Merrill Prentice-Hall.

Ye, Q., Lin, B. Y., & Ren, X. (2021). CrossFit: A few-shot learning challenge for cross-task generalization in NLP (arXiv:2104.08835). arXiv. http://arxiv.org/abs/2104.08835

Zapf, A., Castell, S., Morawietz, L., & Karch, A. (2016). Measuring inter-rater reliability for nominal data—Which coefficients and confidence intervals are appropriate? *BMC Medical Research Methodology, 16*. https://doi.org/10.1186/s12874-016-0200-9

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), 1280-1294. https://doi.org/10.1109/JSTSP.2022.3182537

Zhao, S., Li, B., Reed, C., Xu, P., & Keutzer, K. (2020). Multi-source domain adaptation in the deep learning era: A systematic survey (arXiv:2002.12169). arXiv. http://arxiv.org/abs/2002.12

Appendices

Appendix A: Oral Reading Fluency Passages

Passages excerpted from 2011 PIRLS released passages and items.

PIRLS 2011 Assessment. Copyright © 2013 International Association for the Evaluation of Educational Achievement (IEA). Publisher: TIMSS & PIRLS International Study Center, Lynch School of Education, Boston College, Chestnut Hill, MA and International Association for the Evaluation of Educational Achievement (IEA), IEA Secretariat, Amsterdam, the Netherlands

---

Passage 1 | Excerpt from *Fly Eagle Fly* | word count 122

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.

Passage 2 | Excerpt from *The Giant Tooth Mystery* | word count 110

A fossil is the remains of any creature or plant that lived on the Earth many, many years ago. People have been finding fossils for thousands of years in rocks and cliffs and beside lakes. I now know that some of these fossils were from dinosaurs. Long ago, people who found huge fossils did not know what they were. Some thought the big bones came from large animals that they had seen or read about, such as hippos or elephants. But some of the bones people found were too big to have come from even the biggest hippo or elephant. These enormous bones led some people to believe in giants.

Passage 3 | Excerpt from *Enemy Pie* | word count 122

I went outside to play. All the while, I listened to the sounds of my dad in the kitchen. This could be a great summer after all. I tried to imagine how horrible Enemy Pie must smell. But I smelled something really good. As far as I could tell, it was coming from our kitchen. I was confused. I went inside to ask Dad what was wrong. The pie shouldn't smell this good. But Dad was smart. "If it smelled bad, your enemy would never eat it," he said. I could tell he'd made this pie before. The oven buzzer rang. Dad put on oven mitts and pulled out the pie. It looked good enough to eat! I was beginning to understand.

Passage 4 | Excerpt from *The Giant Tooth Mystery* | word count 119

Hundreds of years ago in France, a man named Bernard was a famous pottery maker. When he went to make his pots, he found many tiny fossils in the clay. He studied the fossils and wrote that they were the remains of living creatures. This was not a new idea. But Bernard also wrote that some of these creatures no longer lived on earth. They had completely disappeared. They were extinct. Was he rewarded for his discovery? No! He was put in prison for his ideas. As time

went by, some people became more open to new ideas about how the world might have been long ago. Then, in the 1820s, a huge fossil tooth was found in England.

Passage 5 | Excerpt from *Fly Eagle Fly* | word count 131

The farmer smiled at him and said, “Of course it’s a chicken. Look— it walks like a chicken, it eats like a chicken. It thinks like a chicken. Of course it’s a chicken.” But the friend was not convinced. “I will show you that it is an eagle,” he said. The farmer’s children helped his friend catch the bird. It was fairly heavy, but the farmer’s friend lifted it above his head and said, “You are not a chicken but an eagle. You belong not to the earth but to the sky. Fly, Eagle, fly!” The bird stretched out its wings, looked about, saw the chickens feeding, and jumped down to scratch with them for food. “I told you it was a chicken,” the farmer said, and he roared with laughter.

Appendix B: Reading Comprehension Passages

Stories and questions taken from 2011 prePIRLS released passages and items (Foy & Drucker, 2013).

Story 1: Brave Charlotte

Charlotte was different from all the other sheep right from the start. When all the other lambs just stood shyly by their mothers, Charlotte was leaping around, ready for adventure.

Charlotte lived with all the other sheep on a hillside far from the farm. They had a shepherd to look after them and he had an old dog named Jack. Jack tried to keep Charlotte under control, but she wasn't scared of him.

Q1 - What did Jack try to do with Charlotte? (Short Answer)
- Process: Focus on and Retrieve Explicitly Stated Information
- Acceptable Response - indicates that Jack tries to keep Charlotte under control.
    - *Examples: keep Charlotte under control, control, keep her calm*
    - *Unacceptable Examples: put Charlotte to sleep*

One time, Charlotte jumped over the side of a riverbank and went for a swim in the fast-running stream. "Tut, tut," said the older sheep, shaking their heads. What would they say if they knew that at night Charlotte secretly roamed through the countryside?

When all the other sheep were sleeping, she would slip away to her special place and gaze at the moon. Even Jack didn't notice. But he didn't have very good ears these days.

One day something terrible happened. The shepherd fell over and broke his leg. Jack barked and circled around him, but that didn't help one bit. The shepherd lay in the grass, not knowing what to do.

"Oh dear, oh dear," said the older sheep. "Somebody must go to the farmer's house in the valley and get help."

"Jack should go. He is the only one who knows the way."

"But it is too far. He hardly manages with the herd these days."

"Yes, that's true," said the others, shaking their heads in despair.

Q2 - Why was the shepherd lying in the grass? (Multiple-choice)
- A) He didn't want to visit the farmer.
- B) He had broken his leg.

- C) He wanted to go to sleep.
- D) He was gazing at the moon.

Q3 - Where did the animals need to go to get help? (Short Answer)
- Process: Focus on and Retrieve Explicitly Stated Information
- Acceptable Response identifies the farmer's house in the valley as the place to go for help
    - *Examples: the farmer's house to the valley; to the farmer*
    - *Unacceptable Examples: a farm; to the road; to the stream*

Then Charlotte said, "I'll do it. I'll go." "Charlotte?" muttered the older sheep. "Out of the question! A sheep has never gone to the valley alone."

The older sheep were beside themselves with worry. But Charlotte couldn't hear them. She had already left to find the right way to the valley. She bounded over fields, through the stream, and over the mountain.

Q4 - Why couldn't Charlotte hear the older sheep? (Short Answer)
- Process: Focus on and Retrieve Explicitly Stated Information
- Acceptable Response indicates that Charlotte could not hear the sheep because she had already left to go to the farmer's house.
    - *Examples: She had already left; She already was on her way to the farmer's house; She was too far away.*
    - *Unacceptable Examples: She had bad ears*.

When Charlotte reached the busy road, it was the middle of the night. She stood and watched the traffic. A truck driver noticed Charlotte and stopped in the road. "Going to the valley?" he asked. Charlotte nodded.

Q5 - When did Charlotte reach the road? (Short Answer)
- Process: Focus on and Retrieve Explicitly Stated Information
- Acceptable Response identifies the time (middle of the night) that Charlotte reached the
    - *Examples: in the middle of the night; at night; night*
    - *Unacceptable Examples: when the farmer was sleeping; sunset, late*

It was so nice to speed along in the truck that Charlotte was almost sorry when they reached the farmer's house. The farmer was asleep when Charlotte tapped on his window with her nose.

"It's Charlotte," said the farmer, "and she is all alone. Something must be wrong."

Q6 - Why was Charlotte almost sorry to reach the farmer's house? (Multiple-choice)
- A) She missed the other sheep.
- B) She would have to wake up the farmer.
- C) She liked the ride in the truck.
- D) She was worried about the shepherd.

Q7 - How did the farmer know something was wrong? (Multiple-choice)

- A) Charlotte was all alone.
- B) Charlotte told him.
- C) Charlotte came in a truck.
- D) Charlotte looked sorry.

Charlotte and the farmer drove on the tractor to find the other sheep. When they arrived, the poor shepherd was still lying in the grass. The farmer took him to the hospital right away.
The shepherd had his leg in a cast for six weeks before he could go back to the sheep. When he returned, he gave Charlotte a big smile. From then on, Charlotte was free to roam as she pleased.

Q8 - How long did the shepherd have a cast on his leg? (Short Answer)

- Process: Focus on and Retrieve Explicitly Stated Information
- Acceptable Response indicates that the shepherd wore the cast for six (6) weeks.
    - *Examples: six weeks; 6 weeks*
    - *Unacceptable Examples: a long time; forever*

Q9 - Why did the shepherd give Charlotte a big smile? (Short Answer)

- Process: Interpret and Integrate Ideas and Information
- Acceptable Response indicates that the shepherd smiled at Charlotte because he was *grateful to her, or a more general indication that the shepherd was proud of Charlotte.*
    - *Examples: She had gone to get the farmer; She had gone to get help: He was thankful: She saved him: She helped him: He was proud of her: because she was brave*
    - *Unacceptable Examples: He liked her; He was happy: From then on, Charlotte was free to roam as she pleased.*

Q10 – In your own words, please describe what happened in the story. (Story Retell)

Story 2: The Lonely Giraffe

The jungle animals were a friendly bunch. All the animals met every morning to talk about the jungle news. Everyone took their turn to speak, but no one listened to the giraffe.
The giraffe was just too tall. By the time he lowered his head to the ground to speak the other animals had lost interest.

Q1 - Why didn't anyone listen to the giraffe? (Short Answer)

- Process: Focus on and Retrieve Explicitly Stated Information
- Acceptable Response indicates that it was because he was too tall or that it took too long to wait for him.
    - *Examples: He was too tall, it took too long*
    - *Unacceptable Examples: He did not wait his turn to speak: He spoke too quietly: He was not friendly.*

So the giraffe would lift his long neck and wander off. He spent all day with his head in the trees eating the sweetest leaves. He didn't know that the birds were frightened of his large head suddenly appearing in the treetops. Or that the small animals on the ground ran away because they were scared of being stepped on. After a while, the lonely giraffe stopped trying to speak to anyone. This went on for the rest of the long dry summer.

Q2 - Why were the animals on the ground afraid of the giraffe? (Multiple-choice)

- A) he would suddenly appear
- B) he had a large head
- C) he might step on them
- D) he ate all their leaves

Q3 - What did the giraffe stop doing over the summer? (Short Answer)

- Process: Focus on and Retrieve Explicitly Stated Information
- Acceptable Response indicates that the Giraffe had stopped speaking to anyone else
    - Examples: *speaking to anyone, talking with the other animals*
    - Unacceptable Examples*:* wandering off; frightening the birds; appearing in the treetops

Then the rainy season came. The rain poured down for days. The animals huddled together beneath the bushes. Then the leopard heard a distant roar. But nobody could think what it was. The giraffe looked over the heads of the animals on the ground. His big eyes widened like saucers and he slowly bent down until the worried animals could hear him.

"The river is flooding," said the giraffe. "A wall of water is racing down the valley and will soon be here."

"What can we do?" asked the gazelle. "It's too late to run away."

Q4 - Why did the animals huddle together beneath the bushes? (Short Answer)

- Process: Focus on and Retrieve Explicitly Stated Information
- Acceptable Response indicates that the animals were avoiding the rain
    - Examples: It was raining, they were trying to stay dry, they didn't want to get wet
    - Unacceptable Examples*:* They were scared of the giraffe; They heard a roar; It was hard to climb the trees.

Q5 - What made the roaring sound in the distance? (Multiple-choice)

- A) a leopard
- B) a flood
- C) a monkey
- D) a tree falling

“Climb up here,” called the monkey from the treetops. “The river won’t reach the high branches.”

The animals raced to the trees. But some of them could not climb up the slippery tree trunks. Their hooves and tails were not made for climbing.

Q6 - Why were the animals trying to climb to the treetops? (Short Answer)

- Process: Focus on and Retrieve Explicitly Stated Information
- Acceptable Response indicates that the animals were trying to stay out of the rising water
    - Examples: to stay out of the water; because it was flooding, to escape the river
    - Unacceptable Examples*:* to see the valley; to reach the sweet leaves: to hide from the leopard

Then the giraffe had an idea. He bent his knees and spoke to the animals. “Climb on to my back. The water is almost here.” The flooded river was splashing around the animals. The monkey jumped onto the giraffe’s neck and called to the others. The hairy wart hog was next to carefully climb on. One by one the animals helped each other to safety.

Q7 - What was the giraffe’s idea for the animals? (Multiple-choice)

- A) to climb the trees
- B) to swim in the river
- C) to hide behind his legs
- D) to climb on his back

The giraffe straightened his knees as the water flooded the jungle. He stretched up his long neck and the last few animals hurried into the branches. The water washed around the giraffe’s strong legs and sprayed the animals in the trees.

Then the flood rushed on. The water slowly sank back to the ground and the sun came out from behind the clouds. The giraffe poked his head up into the high branches and the animals slid down his back to the ground.

From that day on the giraffe was never lonely again.

Q8 - What did the animals do when the sun came out again? (Short Answer)

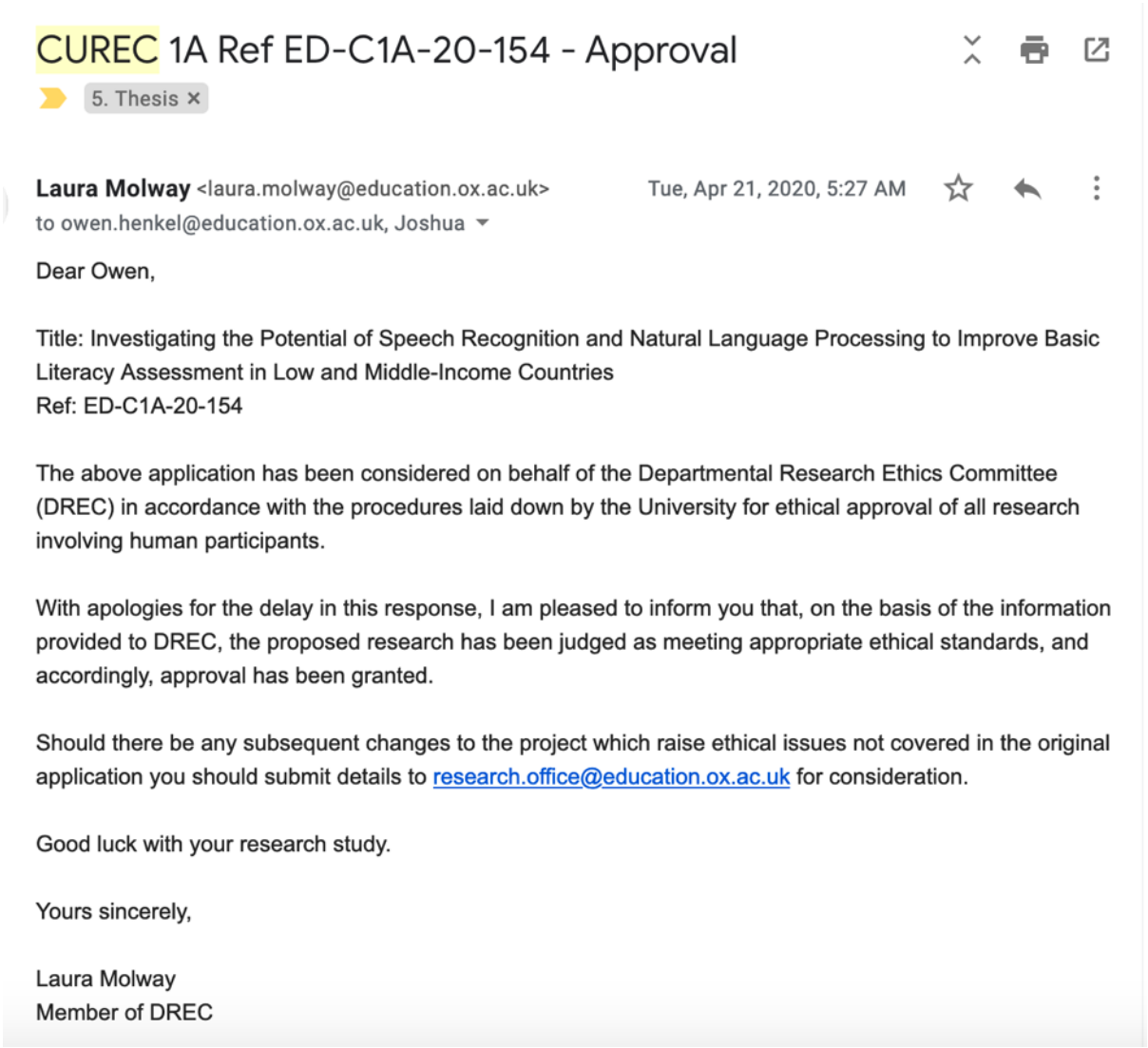
- Process: Focus on and Retrieve Explicitly Stated Information
- Acceptable Response indicates that the animals came down from the trees or slid down the giraffe’s back.
    - Examples: They came down from the trees; they came down; They climbed down his neck.
    - Unacceptable Examples: played; went swimming

Q9 - How was the giraffe’s life different after the flood? (Short Answer)

- **Process**: Interpret and Integrate Ideas and Information
- Acceptable Response provides an appropriate description of how the giraffe's life was different (i.e., improved) after the flood.
    - Examples: He wasn't lonely; He had friends; He was happier.
    - Unacceptable Response: He was lonely, He saved the other animals.

Q10 - Please describe what happened in the story. (Story Retell)

Appendix C: CUREC Application



CUREC 1A Ref ED-C1A-20-154 - Approval

5. Thesis ×

Laura Molway <laura.molway@education.ox.ac.uk> Tue, Apr 21, 2020, 5:27 AM

to owen.henkel@education.ox.ac.uk, Joshua

Dear Owen,

Title: Investigating the Potential of Speech Recognition and Natural Language Processing to Improve Basic Literacy Assessment in Low and Middle-Income Countries
Ref: ED-C1A-20-154

The above application has been considered on behalf of the Departmental Research Ethics Committee (DREC) in accordance with the procedures laid down by the University for ethical approval of all research involving human participants.

With apologies for the delay in this response, I am pleased to inform you that, on the basis of the information provided to DREC, the proposed research has been judged as meeting appropriate ethical standards, and accordingly, approval has been granted.

Should there be any subsequent changes to the project which raise ethical issues not covered in the original application you should submit details to research.office@education.ox.ac.uk for consideration.

Good luck with your research study.

Yours sincerely,

Laura Molway
Member of DREC

SECTION A: Filter for CUREC 2 application This section determines whether your study raises more complex issues requiring the completion of a full application for ethical review, known as the CUREC 2 application. **(Please mark 'X' in the Yes/No column.)**		
1. Are research participants classed as ? (This may include under 18s (although see "competent youths"), prisoners, or adults "at risk".) Your attention is drawn to the University's **Safeguarding Code of Practice** and its implications for researchers involving children or adults at risk. This includes the need for the work to be risk assessed and for researchers to undertake related training. (**Note**: If any of your participants are aged 16 or under, answer 'Yes' here **and** also answer question 5 below.)	Yes ☒	No ☐
2. By taking part in the research, will participants be at risk of criminal prosecution (e.g., by providing information on drug abuse or child abuse)?	Yes ☐	No ☒
3. Does the research involve the **deception** of participants?	Yes ☐	No ☒
4. Does your research raise issues relevant to the **Counter-Terrorism and Security Act (the Prevent duty)**, which seeks to prevent people from being drawn into terrorism? Please see advice on this on our **Best Practice Guidance web page**.	Yes ☐	No ☒
If you answered 'No' to all the questions above, go to **Section B**. If you answered 'Yes' to any question above, continue to question 5 below.		
5. Is your project covered by a CUREC **Approved Procedure** (formerly known as "CUREC Protocols")?	Yes ☒	No ☐
If yes, give the specific Approved Procedure number(s): **IDREC 25**		
If you answered '**Yes'** to **ANY** of questions **1-4**, **and** answered '**No'** to question **5**, **stop completing this checklist and do not submit it for ethical review**. Instead, complete the CUREC 2 application form from the CUREC website, then submit that for ethical review. If you answered '**Yes'** to ANY of questions 1-3, **and** answered '**Yes'** to question 5, go on to **Section B**.		

SECTION B: Contact details and project description	
Contact details:	
1. **Principal investigator** OR supervisor (if student research) (give title and full name)	Associate Professor Joshua McGrane
2. Name of student (if student research)	Owen Henkel
3. Degree programme (if student research), e.g., BA, BSc, MSc, MPhil, DPhil	DPhil

4. Department or Institute name	Oxford University Centre for Educational Assessment, Department of Education
5. Address for correspondence (if different from above)	
6. University (**not** private) e-mail address and telephone number	owen.henkel@education.ox.ac.uk
7. Name and status of others taking part in the project (e.g., third year undergraduate; postdoctoral research assistant)	Professor Therese Hopfenbeck (co-supervisor)

Project description:	
8. Title of research project	Investigating the Potential of Speech Recognition and Natural Language Processing to Improve Basic Literacy Assessment in Low and Middle-Income Countries
9. List of location(s) where project will be conducted	Johannesburg, South Africa
10. If your research involves overseas fieldwork or travel and your department requires a travel risk assessment, will you have completed and returned a risk assessment form beforehand? (This must be approved by your department before you travel. If you are travelling overseas, you are strongly advised to take out University travel insurance.) Please also address any physical or psychological risks for Oxford researchers and local fieldworkers in Section 16 below and discuss with your safety officer.	Yes ☒ No ☐ Not required in this instance ☐
11. Anticipated duration of overall research project	4 years, 8 months
12.a) Anticipated start and end dates of the part of the research project involving human participants and/or personal data	From: (30/03/2020) To: (01/01/2025) **Note**: You will need ethics approval **before** you start your research. CUREC 1As may take up to 30 days to process. **Retrospective ethics approval cannot be granted.**
12. b) In the case of international or collaborative research, will you submit or have you submitted this project for ethical review or consideration elsewhere	Yes ☒ No ☐

(e.g., collaborator's/local ethics committee, or other local approval)?	If 'Yes', please attach ethics or other approvals and give more details below. If 'No', please explain your reasons below. Please also refer to the Best Practice Guidance on Ethical Review of social-sciences based research conducted outside the UK (BPG 16), which includes an Ethics Issues Checklist for International Research (Appendix A)
13. **External** organisation funding the research (if applicable)	

14. a) Title and brief description of **research** (about 150 words) in lay language.
When describing the research, include your methodology, how you are applying professional guidelines, and the use to which results/data will be put. **Please also declare any conflicts of interest here.**

Investigating the Potential of Speech Recognition and Natural Language Processing to Improve Basic Literacy Assessment in Low and Middle-Income Countries

Facilitated by a long-standing professional relationship and potential benefits for their own teaching and assessment practices, SPARK Schools in South Africa has agreed to allow the researchers to conduct a variety of short (5-10 min) oral literacy assessment tasks with their early grade students (**please see attached documents**). These short assessments, which are similar to ones already conducted on a regular basis, will be conducted by during a normal class period. The. researcher will record the oral literacy assessment in addition to scoring in traditional manner (i.e., running record of errors). After removing student names as well and other uniquely identifiable information, SPARK Schools will share student demographic data (age, gender, home language) and previous academic data (reading scores or formative assessments) that it has already collected. Researchers will use UID numbers to match secondary data with audio files. This will create a corpus of recordings of oral literacy assessment tasks, cross-referenced with reliable, longitudinal measures of "true" reading ability. I plan to use this data set to investigate three key questions: a) the feasibility of partially automating oral assessment using automated speech processing and machine transcription, b) the potential for semantic analysis of short passages to be used to grade story summarization tasks, c) training a speech recognition algorithm to "identify" fluent readers.

Potential conflicts of interest: I have worked with SPARK Schools for over four years, have visited all their schools and conducted over 100 classroom observations, and I currently serve on their board. Given the longstanding relationship I see two potential risks: one is SPARK agreeing to a research collaboration despite reservations, and the second would be my relationship with SPARK impacting my ability to be impartial about the results (positive or negative).

In order to ensure my research proposal was treated like any other, I requested that I pass through their normal research approval processes, which involves a committee of senior academic staff evaluating proposals for any potential impact on student safety or student learning. The committee approved my proposal in November. As this research will not be making any claims about the overall efficacy or quality of student learning, I do not anticipate any potential reputational benefit or risk to SPARK, and hence any risk to my ability to be impartial

14.b) Description of participants and how you will **obtain informed consent** to take part in the research (about 200 words in total)

1. Description of participants **and** your criteria for inclusion/exclusion

> Participants will be students in Grade 1-3 across the 20 SPARK schools in greater Johannesburg. As my research does not require a specific sampling approach, I will work with SPARK management to select schools that are highest capacity and best able to participate in. the study without affecting the learning environment. Students will further be excluded from the study if: (a) parental consent is not secured; (b) the classroom teacher judges that the assessment would be in anyway detrimental to a specific student; (c) the student declines to participate when asked at the time of assessment or indicates lack of assent through "avoidance of the testing situation".

2. Your method(s) of recruitment

> Recruitment will be done via each school's main office, who will circulate an announcement/flyer/email to parents of children in the included classrooms. This document, drafted by the researchers, will explain the research in simple but non-patronizing language both the aims of the research and how the data will be used and protected (draft letter attached).

3. Your processes for obtaining consent from participants

> SPARK Schools regularly obtain signed parental-consent from parents for a variety of reasons. I will work with school operations staff to use mix of approaches they deem appropriate such as sending flyers home with students, or having staff share them at parent teacher meetings. The consent form would be opt-in and would request specific consent for the recording of audio. A signed form from a parent/guardian will be necessary to deem that consent has been given. Additionally, before commencing the assessment, I will ask for the teacher confirmation the student is able to participate, and for assent from the student (assessment script attached).

Please **attach separate supporting documents (in Word)** if appropriate for your research (English language versions only). Tick those you are submitting below. If appropriate supporting documents are not submitted, you will be asked to provide these separately, which may delay the ethical review process.

- ☐ Recruitment and advertisement material (e.g., a poster, social media recruitment text, or brief invitation letter/ email)
- ☒ Information for participants to read (or hear) before they agree to take part (e.g., written information or, if applicable, an outline oral information script).
- ☒ A document to record informed consent. Templates for written consent forms and/or oral information scripts (in case of an oral consent process) are available from the CUREC website
- ☒ Questions to be asked of participants (e.g., interview questions, or a preliminary scope of questions, or a sample questionnaire)
- ☐ (If relevant) debriefing document after participants have taken part
- ☐ If you feel the above approaches are not appropriate for your study, provide details on how you will obtain consent from participants
- ☐ Please complete section 15 if you cannot obtain informed consent

Attached to application:

- Information and Consent Form
- Pre-assessment script and indicative reading passages
- SPARK research request and approval

15. If you cannot obtain informed consent from participants according to CUREC guidelines and good practice in your discipline, please give a brief explanation and justification of this decision below.

I will be obtaining informed consent from parents of guardians of student participants, in line with IDREC 25.

16. What are the ethical issues connected with your research and what steps have you taken to address them? **Please do not answer 'none'.** We need to see evidence that you have identified potential ethical issues with respect to your research and have taken steps to address them. If applicable, please address:

The research does not involve topics that are “ethically, emotionally or politically sensitive”, and students already participate in regular formative reading assessment. Further, both students and teacher are accustomed to visitors, researchers, and central office staff visiting and interacting with classrooms. Having conducted numerous class visits in the course of my work with SPARK, during which I interacted with both teacher and students, I am confident that neither my presence in classroom nor the short 1-on-1 assessment I will be conducting will disrupt the learning environment, or be stressful to students in any systematic manner. However, it is entirely possible that particular students might be uncomfortable conducting the assessment either because of ongoing issues or simply because they are having a bad day. I plan to mitigate this risk by asking the classroom teacher if there are students who they believe are not in the right emotional state to be assessed. I will also ask each student if they are comfortable and assent before beginning the assessment.

Regarding child-safeguarding issues, I am well acquainted with the associated risks, as I lead two board-level child-safeguarding committees. The assessments will only be conducted by myself and members of SPARK Schools staff, who themselves have passed background checks and safeguarding training. All assessment/recording will be done in open classrooms, with a minimum of a second adult in the room. Should the other adult need to leave the room for a moment, I will postpone testing and wait at the front of the classroom, leaving the door to the room open.

Finally, in my professional opinion, SPARK takes child safeguarding and welfare extremely seriously and implements procedures and standards that are far above local norms and are line with international best practice. I firmly believe that if they saw potential for this study to negatively impact students’ learning or in any way compromise their safety, they would not have agreed to participate.

- ☐ Your own physical and psychological safety as a researcher or of fieldworkers you may employ (see the University’s and Social Science Division’s Safety in Fieldwork guidance)

I previously worked as a remedial reading teacher at a low-performing high school in New Orleans, where I conducted similar types of literacy assessments with over-aged, functionally illiterate, special needs students. SPARK is a far more positive learning environment and students are much younger, so I am confident of researcher safety within the school/classroom.
While violent crime is an issue in greater Johannesburg, SPARK Schools have full time security guards, detailed safety procedures and to-date have never had intruders on any campus during school hours. Hence the only safety risk would be when commuting around, for which I plan to use taxis rather than public transportation to avoid unnecessary risks. Having already visited South Africa over 10 times, I have felt safe and secure my entire time there and have never experienced any issues where I felt unsafe.

- Data protection/ confidentiality (also see Section 18).

Audio recordings of students reading and summarizing short stories will be the only new data being collected. The **content** of the audio will **not** be personally identifiable, nor will it contain sensitive information. However, because voice patterns are distinctive and theoretically could be used to identify individuals, audio files will be treated with an appropriate level of security.

The audio recording will NOT include student name, rather I will start the audio recording by stating student UID, and then asking them for assent to continue. Audio will be recorded on an Olympus 9500 (or equivalent) voice recorder, which offer 256-bit encryption, pin protection, read only mode USB connectivity. Audio files will be uploaded directly from audio recorder to researcher computer (i.e., not via wi-fi). Files will be located on dedicated, password protected account, using Firevault encryption of the home directory and all the files in it. Files will be identified by UID, not student name. In case of damage to researcher computer, a backup of the original files will be kept on the researcher's Nexus365 University Account.

After removing student names as well and other uniquely identifiable information, SPARK Schools will share participant demographic data (age, gender, home language) and previous academic data (reading scores or formative assessments) which it has already collected. Researchers will use UID numbers to match this secondary data with audio files. Researcher will not have access to list matching UIDs and student names. Only SPARK will retain this matching list, which will be destroyed at the conclusion of the research.

17. Will your research involve discussing sensitive issues? This could be information relating to race or ethnic origin, political opinions, religious beliefs, physical/mental health, trade union membership, sexual life or criminal activities.	Yes ☐	No ☒

18. Management and handling of personal and other research data

For the purpose of completing this section, all information provided by participants is considered **research data**. Any research data from which participants can be identified is known as *personal data*; any personal data which is sensitive is considered *special category data*.

Management of **personal data** and **special category data** of human participants, either directly or via a third party, must comply with the requirements of the General Data Protection Regulation (GDPR) and the Data Protection Act 2018, as set out in the University's Guidance on Data Protection and Research. In answering the questions below, please also consider the points raised in the Data Protection Checklist. For advice on research data management and security, please consult with the University's Research Data Team (researchdata@ox.ac.uk) and/or your local IT department, and the University's web pages on research data management.

a.) **Please mark 'X' against the data you will collect for your research**

Consent records (written consent forms, audio-recorded consent, assent forms (for research involving minors) including participant name	☒
Online consent (may be anonymous)	☐
Opt-out forms	☐
Contact details for research purposes only (destroyed when no longer needed for this research)	☐
Contact details kept for future studies	☐
Audio recordings (preferably using PIN-protected audio recorder and stored on device's hard drive)	☒
Video recordings	☐
Transcript of audio/video recordings	☒
Photographs	☐
Task results (e.g., paper/online tasks, diary completion)	☐
Questionnaire answers	☐
Field notes	☒
Other (please specify below)	☒
Running-records of student reading accuracy Teacher and researcher assigned fluency scores	

b.) For **each** of the types of data selected above, state how this will be physically transferred from where it is collected to a local secure storage site (and backed up as necessary). This includes paper records and data captured electronically.

All documents and date mentioned below (will be kept by researchers will be located on dedicated, password protected account, using Firevault encryption of the home directory and all the files in it.) Files will be identified by UID, not student name. In case of damage to researcher computer, a backup of the original files will be kept on the researcher's Nexus365 University account.

Audio Recording (personally identifiable) – Audio will be recorded on an Olympus 9500 (or equivalent) voice recorder, which offer 256-bit encryption, pin protection, read only mode USB connectivity. Audio files will be uploaded directly from audio recorder to researcher computer (i.e., not via wi-fi). The audio recording will NOT include student name, only UID. Researcher will keep audio files on dedicated, password protected computer with a back-up on University Nexus365 account.

Consent records (not personally identifiable) - Collected at school offices and scanned at the central office. SPARK school will keep electronic copies of consent forms along with other student information in its internal data warehouse. Researcher will keep electronic copies on dedicated, password protected computer with a back-up on University Nexus365 account

Transcripts (not personally identifiable) **–** generated from audio files. Kept on dedicated, password protected computer with a back-up on University Nexus365 account.

Running records and ratings (not personally identifiable) **-** will be captured on paper at time of assessment. They will be scanned at SPARK head office. Researcher will keep electronic copies on dedicated, password protected computer with a back-up on University Nexus365 account.

c.) How and where will **each type of data** be stored during the research (until the end of all participant involvement)? Describe the arrangements for ensuring confidentiality, i.e., location of storage (e.g., Nexus 365 OneDrive for Business, SharePoint), security arrangements and de-identification of such data. Do not store unencrypted data in freely available cloud services or unprotected USB drives.

All documents and date mentioned below (will be kept by researchers will be located on dedicated, password protected account, using Firevault encryption of the home directory and all the files in it.) Files will be identified by UID, not student name. In case of damage to researcher computer, a backup of the original files will be kept on the researcher's Nexus365 University account.
Some computational tasks will require the use of cloud computing. To avoid doubt, this will be used only for the calculations required in training of statistical model, the audio files themselves will not be uploaded nor "live" on the

cloud. I plan to use either Amazon AWS, Microsoft Asure, or Google Cloud, all of which are GDPR compliant and have robust security protocols, including default encryption.		
d.) Will you use a unique participant number on research data instead of a participant name? If **yes,** state whether or not you will retain a list of participant names against numbers (i.e., pseudonymisation via a linkage list). Where will the list be stored, and when will it be destroyed?		
Yes. The researcher will receive aforementioned demographic and academic data matched to UIDs, which will be used to link them to audio files. SPARK Schools will be the only party to retain the list linking UIDs to student names. At the conclusion of the research, SPARK will destroy this list.		
e.) Who will have access to the research data?		
Researcher and two academic supervisors		
f.) If research data is to be shared with another organisation, how will it be transferred / disclosed securely?		
g.) When and how will identifiable data (including audio/video recordings & photos) be destroyed or deleted? **Note:** Records of consent should be retained for a **minimum of three years after publication or public release**. Some funders may require longer periods (see http://www.dcc.ac.uk/resources/policy-and-legal/overview-funders-data-policies). If you wish to retain contact details in order to re-approach participants about future studies, you must detail this in information provided to them and obtain specific consent for this.		
Electronic versions of consent forms, and audio files will be kept for 5 years after data collection. Audio recordings will be kept on secure computer (as described above) and will be destroyed by permanently deleting them from the computer and hard drive on which they are stored and reformatting the dedicated computer. There will be no need to re-approach participants as this is not a longitudinal study.		
h.) Please confirm that you will store **other research data** safely for at least 3 years after final publication or public release and adhere to any additional research funder policies. For more information about the University policies, please see the University's web pages on research data management.	Yes ☒	No ☐
If '**Yes'**, please give details of who will store the data and on storage format, location and security. Note that open science is encouraged. If **'No',** please provide further details below.		
Audio files and student demographic and academic data cannot be shared and will be destroyed 5 years after collection as outline above. Transcription, prosody, and summarization models that are produced as part of the research will continue to be the intellectual property of the researcher and able to be used for further research.		
i.) Does your research involve the use of secondary (i.e., previously collected) data? Common sources of secondary data include censuses, information collected by government departments, organisational records and data that was originally collected for other research purposes	Yes ☒	No ☐

(If "**No**", please go to section **19**.)		
j.) Do you have data access agreements for the use of this secondary data? (If so, please attach these.)	Yes ☒	No ☐
k.) Is your use of this secondary data compatible with what data subjects/participants agreed that their data should be used for?	Yes ☒	No ☐
l.) Could this data be linked back to an individual or individuals? If yes, address how securely any personally identifiable data will be transferred to you, and where and for how long it will be stored during or after the research. Who will have access to it?	Yes ☒	No ☐
After removing student names as well and other uniquely identifiable information, SPARK Schools will share participant demographic data (age, gender, home language) and previous academic data (reading scores or formative assessments) which it has already collected. Researchers will use UID numbers to match this secondary data with audio files. Researcher will not have access to list matching UIDs and student names, only SPARK will retain this matching list, which will be destroyed at the conclusion of the research. While this information is not uniquely identifiable, it will still be stored with the same protocols as above (fully encrypted, dedicated account on research computer with back up on University Nexus365 account).		

19. Publication and dissemination of research data	
How will you disseminate and feedback project outcomes at the end of the research?	As this research project requires close collaboration of SPARK schools, any actionable or potentially beneficial information will be shared with them as and when it is discovered. SPARK Schools will decide how best to communicate any information to parents and students. The results from the study will be used for publications and conferences. However, any information shared, will be at an aggreagate level, and will not be traceable to individual students.

SECTION C: Methods and procedures to be used	
Method used: Please ensure you have addressed any potential ethical issues related to these methods in Section 14 and in your Participant Information Sheet	**Please mark 'X'**
1. Analysis of existing records	☐
2. Snowball sampling (recruiting through contacts of existing participants)	☐
3. Use of casual or local workers e.g., interpreters	☐
4. Participant observation	☐
5. Covert observation	☐
6. Observation of specific organisational practices	☐

7. Participant completes questionnaire in hard copy	☐
8. Participant completes online questionnaire or other online task	☐
9. Using social media	☐
10. Participant performs paper and pencil task	☐
11. Participant performs verbal or aural task (e.g., for linguistic study)	☐
12. Focus group	☐
13. Interview	☐
14. Audio recording of participant (you will generally need specific consent from participants for this)	☒
15. Video recording of participant (you will generally need specific consent from participants for this)	☐
16. Photography of participant (you will generally need specific consent from participants for this)	☐
17. Others (please specify below)	☐

SECTION D: Professional guidelines and training		
1. In this section, please mark 'X' against at least one of the following professional guidelines you aim to adhere to. You should use the principles listed in your chosen guideline(s) in conducting your own research. **Note**: this is not an exhaustive list.		**Please mark 'X'**
Research specialism/ methodology	**Association and guidance document**	
Anthropology	Association of Social Anthropologists of the UK and Commonwealth	☐
Computer Sciences	ACM Code of Ethics and Professional Conduct	☒
Criminology	http://www.britsoccrim.org/ethics/	☐
Education	British Educational Research Association Ethical Guidelines for Educational Research	☒
Geography	Association of American Geographers Statement on Professional Ethics	☐
History	Oral History Society of the UK Ethical Guidelines	☐
Internet-based Research	British Psychological Society: Conducting Research on the Internet Association of Internet Researchers Ethics Guide ACM Code of Ethics and Professional Conduct Also see our Best Practice Guidance on internet-based research	☐
Law (Socio-Legal)	Socio-Legal Studies Association: Statement of Principles of Ethical Research	☐
Management	Academy of Management's Professional Code of Ethics	☐
Political Science	American Political Science Association (APSA) Guide to Professional Ethics in Political Science	☐
Politics	Political Studies Association. Guidelines for Good Professional Conduct	☐
Psychology	British Psychological Society Code of Ethics and Conduct	☐
Social Research	Social Research Association: Ethical Guidelines	☐
Sociology	The British Sociological Association: Statement of Ethical Practice	☐
Visual Research	ESRC National Centre for Research Methods Review Paper: Visual Ethics: Ethical Issues in Visual Research	☐
Other professional guidelines. Please specify the other guidelines used here*:*		☐
2. Please indicate what training in research ethics (or research methodology) the researchers involved with this study have received, e.g., the title of the course and date completed (online training available at http://researchsupport.admin.ox.ac.uk/support/training/ethics), or discussions between researchers and supervisors, if applicable.		

SECTION E: Signatures or email endorsements (The SSH IDREC Secretariat accepts either option below. If you have a DREC, check which signature option it prefers.) ☐ **Option 1:** Email confirmations from a University of Oxford email address can be accepted. Separate emails should come from each of the relevant signatories as outlined below, indicating acceptance of the relevant responsibilities. **Pasted images of signatures cannot be accepted.** ☐ **Option 2:** Handwritten (wet-ink) signatures. Please scan them and the rest of the checklist pages to create a single PDF document and email to us.	
Please ensure this checklist is signed by:	
For staff research:	**For student research**:
1. Principal investigator	1. Principal investigator **(project supervisor)**
2. **Head of Department (or nominee)**	2. **Head of Department (or nominee)**
	3. **Student researcher**

1. Principal Investigator signature/supervisor signature (if student research)

I understand my responsibilities as **principal investigator** as outlined in the CUREC glossary and guidance on the CUREC website.

I declare that the answers above accurately describe the research as presently designed, and that a new checklist will be submitted should the research design change in a way which would alter any of the above responses so as to require completion of CUREC 2 (involving full scrutiny by an IDREC). I will inform the relevant IDREC if I cease to be the principal investigator on this project and supply the name and contact details of my successor if appropriate.

Signature (or email endorsement using the above declaration): ……..

Print name (block capitals): …….. **Date**: ……...

2. Departmental endorsement signature

I have read the research project application named above. On the basis of the information available to me, I:

(i) consider the principal investigator to be aware of her/his ethical responsibilities in regard to this research;
(ii) consider that any ethical issues raised have been satisfactorily resolved or are covered by relevant professional guidelines and/or CUREC approved procedures, and that it is appropriate for the research to proceed (noting the principal investigator's obligation to report should the design of the research change in a way which would alter any of the above responses so as to require completion of a CUREC 2 full application);
(iii) am satisfied that: the proposed project design and scientific methodology is sound; the project has been/will be subject to appropriate **peer review**; and is likely to contribute to existing knowledge and/or to the education and training of the researcher(s) and that it is in the **public interest**.

Signed by Head of Department or nominee *(example nominees for student research include the Director of Graduate Studies/ Director of Undergraduate Studies)*:

Signature (or email endorsement using the above declaration): ……

Print name (block capitals): ……..... **Date**: ……...

3. **Student signature (if student research)**

I understand the questions and answers that have been entered above describing the research, and I will ensure that my practice in this research complies with these answers, subject to any modifications made by the principal investigator properly authorised by the CUREC system.

Signature by student (or email endorsement using the above declaration):

Appendix D: Abstract of Published Article 1

International Journal of Artificial Intelligence in Education
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
© The Author(s) 2024

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

Published online: 28 January 2025

Appendix E: Abstract of Published Article 2

International Journal of Artificial Intelligence in Education
https://doi.org/10.1007/s40593-024-00431-z

ARTICLE

Can LLMs Grade Open Response Reading Comprehension Questions? An Empirical Study Using the ROARs Dataset

Owen Henkel[1] · **Libby Hills**[2] · **Bill Roberts**[3] · **Joshua McGrane**[4]

Accepted: 19 September 2024
© International Artificial Intelligence in Education Society 2024

Abstract
Formative assessment plays a critical role in improving learning outcomes by providing feedback on student mastery. Open-ended questions, which require students to produce multi-word, nontrivial responses, are a popular tool for formative assessment as they provide more specific insights into what students do and do not know. However, grading open-ended questions can be time-consuming and susceptible to errors, leading teachers to resort to simpler question formats or conduct fewer formative assessments. While there has been a longstanding interest in automating the grading of short answer questions, previous approaches have been technically complex, limiting their use in formative assessment contexts. The newest generation of Large Language Models (LLMs) potentially makes grading short answer questions more feasible, as the models are lexible and easier to use. This paper addresses the lack of empirical research on the potential of the newest generation LLMs for grading of short answer questions in two ways. First, it introduces a novel dataset of short answer reading comprehension questions, drawn from a battery of reading assessments conducted with over 150 students in Ghana. This dataset allows for the evaluation of LLMs in a new context, as they are predominantly designed and trained on data from high-income countries. Second, the paper empirically evaluates how well various configurations of generative LLMs can grade student short answer responses compared to expert human raters. The findings show that GPT-4, with minimal prompt engineering, performed extremely well on evaluating the novel dataset (QWK 0.91, F1 0.87), reaching near-parity with expert human raters. To our knowledge, this work is the first to empirically evaluate the performance of generative LLMs on short answer reading comprehension questions using real student data. These findings suggest that generative LLMs could be used to grade formative assessment tasks, potentially benefiting real-world educational settings.

Keywords Large language models · Automatic short answer grading · Reading comprehension

Extended author information available on the last page of the article

Published online: 18 October 2024