

Distribution Agreement

In presenting this thesis as a partial fulfillment of the requirements for an advanced degree from Emory University, I hereby grant to Emory University and its agents the non-exclusive license to archive, make accessible, and display my thesis in whole or in part in all forms of media, now or hereafter known, including display on the world wide web. I understand that I may select some access restrictions as part of the online submission of this thesis. I retain all ownership rights to the copyright of the thesis. I also retain the right to use in future works (such as articles or books) all or part of this thesis.

Catherine Baker

Date

Read with Emora: Revolutionizing Children's Reading Comprehension Through An
LLM-Powered Personalized Intelligent Tutoring System for Vocabulary, Grammar,
and Context-Based Learning

By

Catherine Baker
Master of Science
Computer Science

Dr. Jinho D. Choi
Advisor

Dr. Li Xiong
Committee Member

Dr. Emily Wall
Committee Member

Kimberly Jacob Arriola, Ph.D, MPH
Dean, James T. Laney Graduate School

Date

Read with Emora: Revolutionizing Children's Reading Comprehension Through An
LLM-Powered Personalized Intelligent Tutoring System for Vocabulary, Grammar,
and Context-Based Learning

By

Catherine Baker
Computer Science and Mathematics BS

Advisor: Dr. Jinho D. Choi

An abstract of
A thesis submitted to the Faculty of the
James T. Laney School of Graduate Studies of Emory University
in partial fulfillment of the requirements for the degree of
Master of Science
in Computer Science
2024

Abstract

Read with Emora: Revolutionizing Children’s Reading Comprehension Through An LLM-Powered Personalized Intelligent Tutoring System for Vocabulary, Grammar, and Context-Based Learning
By Catherine Baker

Reading comprehension is critical for educational success, yet many students in the United States struggle with low proficiency. Decades of research on reading comprehension techniques show depth-based in-context approaches for practical learning components like grammar and vocabulary, spaced repetition and practice for long-term retention, and personalized learning to ensure all students learn at the same rate. However, many educators have yet to incorporate these findings into classroom practice opting instead for core or standardized curricula, which can contribute to suboptimal student reading progression. This thesis presents the development of Read with Emora, an Intelligent Tutoring System designed to enhance current reading comprehension-focused tutoring systems by incorporating LLM-powered personalized learning. Read with Emora integrates large language models to generate AI-driven reading passages, context-based questions, and adaptive feedback tailored to each student’s grade level, vocabulary list, grammar rules, and chosen topics. By integrating LLMs for personalized instruction, Read with Emora aims to improve content generation and personalized learning features present in current reading comprehension-focused ITSs.

This project demonstrates the potential for LLM-driven tutoring systems to provide accessible and free reading comprehension tools for parents, teachers, and students. Read with Emora generates personalized reading materials and assessments that would otherwise require purchasing, making reading practice more available to students regardless of background or resources. The evaluation of Read with Emora focuses on three key areas: (1) the integration of critical components such as learning objectives, grammar rules, and vocabulary words into generated passages through assigned integration scores, (2) the accuracy of the system’s grading for context-based comprehension questions compared to human evaluators (analyzed through loss between human and LLM-assigned scores), and (3) the variety of questions generated by assigning questions and analyzing their distribution across six categories. The results show that while the system excels in integrating personalized topics and generating factual recall questions, it faces challenges in more context-dependent areas such as complex grammar rules, factual content integration, and Cause and Effect questions. The system also exhibited leniency in grading, often assigning partial credit for incomplete or incorrect responses, highlighting the need for refinement in the grading prompt.

This project offers a promising innovation for improving reading comprehension systems through LLMs. Future directions include refining Read with Emora’s question generation and grading systems, particularly in generating nuanced question types and improving grading accuracy, incorporating second language learning benefits, and expanding the system’s feature set with tools like speech-to-text functionality and image

generation. Ultimately, this thesis underscores the importance of integrating LLMs into comprehension-focused educational tools to provide children with personalized, effective learning experiences.

Read with Emora: Revolutionizing Children's Reading Comprehension Through An
LLM-Powered Personalized Intelligent Tutoring System for Vocabulary, Grammar,
and Context-Based Learning

By

Catherine Baker
Computer Science and Mathematics BS

Advisor: Dr. Jinho D. Choi

A thesis submitted to the Faculty of the
Emory College of Arts and Sciences of Emory University
in partial fulfillment of the requirements for the degree of
Master of Science
in Computer Science
2024

Acknowledgments

I would like to express my gratitude to Dr. Jinho Choi for his invaluable guidance and support throughout the development of this thesis. His expertise and insights were instrumental in shaping this project, and his encouragement has been a source of constant motivation. I would like to thank my thesis committee members Drs. Li Xiong and Emily Wall as well for their time, feedback, and edits. The suggestions made have helped to greatly improve this work in both content and writing quality. I would also like to sincerely thank Hyoung Gyoo Seo for his exceptional work on this project. While I focused on background research, frontend development, UI, data review and acquisition, project management, and program logic and algorithms, Hyoung Seo focused on the backend development, database management and maintenance, and web hosting, which was critical in deploying our program on an external platform.

I would also like to thank my parents, Todd and Diane, and my grandparents, Ron and Rita, who helped me get to Emory University and allowed me to attend college without extensive financial burden. Their support and encouragement allowed me to be in the position to write this thesis, and without them, I would have never been able to achieve the level of academic success I have so far.

Along with them, I thank my sisters Julianne and Caroline, who have always looked out for me and helped me through any mistake or tribulation. Their friendship has guided and saved me many times, and I cannot be grateful enough to have them in my life.

Contents

1	Introduction	1
1.1	Reading Comprehension in Children	1
1.2	LLMs in Personalized Learning Intelligent Tutoring Systems	2
1.3	Research Questions	4
2	Background	6
2.1	Theoretical Approaches to Reading Comprehension and Personalized Learning	6
2.1.1	Current Classroom Limitations	6
2.1.2	Learning Theories Supporting ITS Development	8
2.2	Existing Intelligent Tutoring Systems for Reading Comprehension	9
2.2.1	My Turn To Read	10
2.2.2	AutoTutor	11
2.2.3	The Reading Tutor	12
2.3	Advances in ITS with LLM and Spaced Repetition Features	13
2.3.1	ITSs and Spaced Repetition Techniques	13
2.3.2	ITSs Utilizing LLMs for Personalized Learning	16
2.4	Novelty of Proposed Personalized Intelligent Tutoring System: Read with Emora	19
3	Read with Emora (RwE)	21

3.1	Materials and Data Collection	21
3.1.1	Learning Objectives	22
3.1.2	Grammar Rules	28
3.1.3	Vocabulary Words	31
3.2	Theoretical Approach and Core Algorithms	33
3.2.1	Adaptive Scoring System	33
3.2.2	Spaced Repetition and Learning Algorithms	34
3.2.3	Grammar and Vocabulary Integration	37
3.2.4	RwE Personalization	38
3.3	Implementation of Key Algorithms and Features	41
3.3.1	Program Structure	42
3.3.2	Passage Generation	42
3.3.3	Purpose Generation	49
3.3.4	Important Sentence Identification	53
3.3.5	Question Generation	57
3.3.6	Gold Standard Answer Generation	60
3.3.7	Iterative Feedback and Grader AI	64
3.3.8	Scoring and Ease Factor Adjustments	68
3.4	Interface	68
3.4.1	Page Layout and Navigation	69
3.4.2	Data Communication and Storage	78
4	Evaluation	79
4.1	Experimental Setup	79
4.1.1	Research Questions	82
4.1.2	Research Question 1: Evaluating Passage Generation	83
4.1.3	Research Question 2: Evaluating Question Generation and Grading Accuracy	85

4.1.4	Considerations for Experiment Design	86
5	Analysis	87
5.1	Results for RQ1: Passage Generation	87
5.1.1	Feature Integration by Grade	88
5.1.2	Feature Integration by Feature	88
5.2	Results for RQ2: Question Generation and Grading	91
5.2.1	Question Distribution and Variety	92
5.2.2	Grading Discrepancies by Category	92
5.2.3	Additional Findings	93
6	Conclusion	95
6.1	Evaluation Results	96
6.1.1	Passage Evaluation	96
6.1.2	Question Evaluation	97
6.2	Limitations	97
6.3	Future Directions	98
6.3.1	Vocabulary and Grammar	98
6.3.2	Feature Expansion	99
6.3.3	Higher Fidelity Implementations	100
6.3.4	Second Language Learning Benefits	100
6.4	Final Thoughts	100
A	Appendix	102
A.1	Learning Objectives by Grade	102
A.2	Excluded Learning Objectives by Grade	103
A.3	Grammar Rules by Grade	103
A.4	GitHub and Data	103

List of Figures

3.1	General Literature Learning Objective Structure Based on the Massachusetts DOE Guide	23
3.2	General Informational Learning Objective Structure Based on the Massachusetts DOE Guide	24
3.3	Program Structure Outline through our Designated AI Models	43
3.4	Read with Emora Sign-In Page	69
3.5	Read with Emora Sign-Up Page	70
3.6	The Session Page before passage generation.	71
3.7	The Session Page After Passage and Question Generation.	72
3.8	The Session Page After Finishing a Session: The results card represents the student responses to the questions.	74
3.9	The Progress Page After Completing Five or More Passages.	75
3.10	The Profile Page After Signing In.	77

List of Tables

2.1	User performance grading for ease-factor in SM-2 algorithm	14
3.1	First Through Third Grade Learning Objectives Based on the MDOE Guide	25
3.2	Objectives Excluded from Second-Grade Learning per the MDOE . .	28
3.3	Second Grade Grammar Rules per Big Learners and Night Zookeeper	30
3.4	Number of Vocabulary Words by Grade as Sourced from Flocabulary and k12Reader.	32
3.5	First Ten Vocabulary Words per Grade.	32
3.6	Words per minute by grade level	39
3.7	Recommended Passage Length by Grade Level	39
4.1	Grading Scale for Passage Features	84
5.1	Average Integration Score for Each Feature by Grade	87
5.2	Question count across grade and categories	92
5.3	Average loss between GPT and human grades by grade Level	93
A.1	Fourth Grade Learning Objectives Based on the MDOE Guide	102
A.2	Fifth Grade Learning Objectives Based on the MDOE Guide	103
A.3	First Grade Objectives Excluded from Learning Based on the MDOE Guide	104

A.4	Third Grade Objectives Excluded from Learning Based on the MDOE Guide	105
A.5	Fourth Grade Objectives Excluded from Learning Based on the MDOE Guide	106
A.6	Fifth Grade Objectives Excluded from Learning Based on the MDOE Guide	107
A.7	First Grade Grammar Rules per Big Learners and Night Zookeeper .	107
A.8	Third Grade Grammar Rules per Big Learners and Night Zookeeper .	108
A.9	Fourth Grade Grammar Rules per Big Learners and Night Zookeeper	109
A.10	Fifth Grade Grammar Rules per Big Learners and Night Zookeeper .	110

Chapter 1

Introduction

1.1 Reading Comprehension in Children

In 2022, the Program for International Student Assessment (PISA) found that 20% of adolescent American 15-year-olds showed low reading proficiency and only 14% showed high reading proficiency [55]. The PISA findings also highlighted a substantial gap in reading literacy between students in the 90th and 10th percentiles, one of the largest disparities among participating countries, which points to a wide range of reading abilities within U.S. reading comprehension [55]. Similarly, the National Assessment of Educational Progress's (NAEP) 2019 report for mathematics and reading showed that only 35% of fourth graders and 34% of eighth graders can read proficiently [1]. Additionally readers who are at basic proficiency have poor prosody and make more errors than proficient readers [32]. have With the majority of fourth and eighth graders reading below NAEP proficiency levels and a the bulk of PISA students only reaching average proficiency, current classroom reading comprehension instruction must be in some way lacking.

The reading comprehension challenges many lower-proficiency students face are often due to inadequate classroom instruction rather than a lack of ability. Comprehen-

sion is influenced by factors such as language understanding, background knowledge, vocabulary, and inference skills, all of which require time and focused practice to develop [18]. The highest predictor of reading comprehension is vocabulary, including understanding definitions and word usage [36]. Because focused practice on these constituent skills are often neglected, many lower-proficiency students remain behind their peers, and American classrooms continue to see a wide gap in reading comprehension performance [18, 42].

1.2 LLMs in Personalized Learning Intelligent Tutoring Systems

The majority of American classrooms adopt traditional learning methods using a "one-size-fits-all" method that does not account for individual learning paces [3]. There is a significant reliance on mass learning techniques like massed practice in American classrooms, prioritizing cramming large amounts of information into single sessions without adequate reinforcement [28]. Often this is implemented in standardized classroom projects, curriculum, and instruction with no individualism in learning for students [3]. This standardized instruction can make teaching simpler, but does not account for many of the innate differences in student learning needs that lead to an uneven distribution of skill acquisition in the classroom [42].

Personalized learning is a method of instruction that tailors educational content and learning to each student's needs [3]. It is currently considered one of the most efficient methods of teaching and commonly is incorporated into Intelligent Tutoring Systems (ITSs), technology tailored to student instruction and tutoring, through features like a feedback mechanism, unbiased assessment and grading, and analysis of performance [3]. Studies on ITSs show that students benefit more from systems that provide iterative feedback than systems that display the correct answers alone [62, 6].

Furthermore, this feedback can increase student engagement in educational exercises [10]. We aim to incorporate these personalized features in our proposed platform.

Although human tutors offer effective personalized learning, they are often expensive and time-consuming, making them inaccessible for many students [62]. Technology-based solutions, particularly Intelligent Tutoring Systems (ITS), can provide tailored instruction at a lower cost that improves student understanding and performance more than just classroom instruction alone [62, 6, 37]. LLMs can generate reliable tailored content [52]. The controllability of LLMs allows LLM-powered ITSs to generate text that meets specific criteria as dictated by the user [52]. This control allows LLM-powered ITSs to incorporate features like learning objectives, vocabulary, and grammar into reading passages. Further, LLMs have been shown to provide context-sensitive corrections and suggestions to writing in ITSs like Tsai et al. [54], which can enhance error correction in response to student answers. By incorporating corrective feedback in RC-ITSs, we can further reduce educator workload allowing comprehensive out-of-classroom guidance [54].

Current reading comprehension-focused ITS (RC-ITS) implementations have limitations in offering personalized instruction, focusing more on reading comprehension proficiency assessment, cognitive skills, and phonology [36, 47, 38]. While cognitive skills like phonics are foundational for early reading literacy and acquisition, vocabulary and grammar instruction are critical for ongoing reading development and literacy, as they enable deeper comprehension, text analysis, and academic success [42]. Additionally these sources rarely implement language models for content generation despite evidence showing that LLMs are effective for generating coherent, understandable narratives and diverse and context-based questions, making them a promising choice for producing reading passages for children’s tutoring [57, 51, 52]. Logical clarity and flow in generated stories is essential for readability and comprehension in educational materials and question answering on reading passages is fundamental for enhancing

reading comprehension ability [61, 27]. LLMs can provide an advantage over existing reading comprehension platforms by generating dynamic context-based content.

Advances in large language models (LLMs) like with generative pre-trained transformers (GPTs) open the door for more sophisticated, individualized tutoring experiences than currently implemented. Integrating LLMs into ITSs has made it possible to generate study-focused conversational agents [15, 12, 45]. Research has also shown that reading comprehension can be boosted with automatically-generated context-based questions [34, 51]. We expand on previous comprehension research by creating an RC-ITS using LLMs to generate context-appropriate reading passages and questions that cater to a student’s unique grade and learning. We further expand on RC-ITSs by using LLMs to adapt common personalized learning features like corrective feedback, adaptive scoring, and performance analysis into an RC-ITS.

1.3 Research Questions

Our project focuses on the integration of LLMs for personalized learning into existing reading comprehension research in the form of an ITS. We aim to address two research questions in this thesis that focus on our novel LLM integration’s performance in generated and assessing these personalized features:

1. **RQ1: Can an LLM generate a reading passage that integrates learning objectives, grammar rules, vocabulary words, and personalized topics from the user?**

This question investigates whether LLM-generated content can be tailored to include specific educational components. Some current systems like The Reading Tutor demonstrate the potential for AI-driven tools to enhance reading comprehension [38]. Still, they often lack the integration of personalized learning objectives and complex grammar instruction necessary for holistic reading

education [49]. We propose a passage generation system that integrates practical learning components like learning objectives, vocabulary, and grammar with student-provided topics to personalize the tutoring experience.

2. **RQ2: Given a passage, can an LLM generate a categorical variety of context-based questions and accurately grade their responses?**

This question explores whether an LLM can generate context-based comprehension questions on a given reading passage, and respond appropriately to student answers. Implementing such features in our proposed system provides a practice platform for comprehension study.

Thesis Statement: Through corrective feedback, content generation, adaptive scoring, and performance analysis, we aim to create an LLM-powered, personalized Intelligent Tutoring System that narrows the gap between technological advancements in ITSs and LLMs for online reading comprehension study.

Chapter 2

Background

2.1 Theoretical Approaches to Reading Comprehension and Personalized Learning

Classrooms face a range of instructional limitations, such as a lack of extended comprehension instruction, reliance on massed practice, depth-based grammar instruction, and varying student learning paces, that can hinder students' understanding if not addressed comprehensively [48, 28, 41, 42].

2.1.1 Current Classroom Limitations

American classrooms often do not dedicate enough time to focused comprehension instruction. This limits the development of essential skills like vocabulary, background knowledge, and inference ability [48]. Reading comprehension instruction ideally involves building context and integrating knowledge across different subjects and domains through diverse texts. American classrooms, however, often prioritize language arts concepts and fictional texts over informational texts like science articles or biographies [18], limiting students' contextual knowledge and vocabulary and reducing the relevance of these skills to real-life contexts.

Massed practice (often denoted "cramming") attempts to amass all necessary learning into one session to maximize learning from a single session [28]. It also works to maintain associations between topics by introducing them together [28]. This technique can be more convenient to teach when working with a larger group of students, with less time to go over corrections and mistakes, but has not been shown to significantly increase student performance compared to spaced practice [28]. It also does not include robust misunderstanding-handling for students who make mistakes and removes personalization from instruction. However, massed practice is currently used in most American school systems and curricula, despite evidence supporting techniques like spaced repetition for studying [28].

Traditional grammar instruction follows a depth-based approach. This technique uses step-by-step grammar mastery (learning one rule at a time and building on those rules) [41]. Additionally, teaching techniques often involve standardized worksheets and lessons which excludes direct addressing of common errors, individual error handling, and context-based grammar learning, thereby removing a practical understanding of grammar from the lesson [11]. A depth-based approach can be beneficial, but in practice, students tend to prefer a more sporadic, breadth-based approach [41].

In classrooms there is often an uneven learning plane where some students acquire skills faster than others. Many classroom assessments may be too easy or difficult for learners, resulting in scores that do not accurately represent their comprehension level [42, 47]. The unknown variability in student progress complicates the recommended teaching instruction techniques, which heavily emphasize the inter-relatedness of rules and structures. With some students ready to move on to new topics much faster than others, low proficiency students are often left behind to make up their understanding themselves [42].

2.1.2 Learning Theories Supporting ITS Development

Spaced Repetition and Practice

Spaced repetition is a proven technique to reinforce specific learning items, such as vocabulary or flashcards, by spacing out review intervals based on previous recall success [21, 53, 4]. Correctly answered items appear less frequently than those answered incorrectly, which optimizes recall by strengthening both short- and long-term memory. Spaced repetition is especially effective in vocabulary and grammar acquisition, with applications in second-language learning [21].

Spaced practice, which spaces review sessions over time, complements this by extending reinforcement across sessions. It has worked similarly to improve long-term recall in adults and children as young as four [53]. This practice improves term recall more than massed practice in the same amount of time and reduces study redundancies. Spaced practice offers efficient study time and cost benefits, especially when applied within ITS environments for learning reinforcement [28, 53].

Grammar Teaching Techniques

Breadth-based grammar approaches fluctuate the mastery levels of various rules at any time [41]. By varying the current grammar topic through different contexts, students can better understand the inter-relatedness of grammar rules [41]. This approach should be combined with connections between grammar, and its usual context, communicative purpose, and error correction [41, 11]. This contextualized grammar instruction allows students to comprehensively understand grammar in its standard setting, and to recognize and use it more easily. Personalized learning platforms can adapt grammar instruction to these findings by offering context-based grammar lessons that introduce grammar rules in reading material without the stress of direct testing on them.

Student Skill Development Constraints

Students behind in classrooms cannot always simply "catch up" to their peers. In most cases, individual instruction and monitoring are proven to be most beneficial, but they also require the most time and resources for parents and educators [42]. One successful technique seen has been reading "interventions," in which classrooms or students at home use a specialized reading comprehension tool involving individual instruction and improvement for a period of time [36]. These interventions can improve classroom techniques by providing more practice and instruction in a shorter time period [36]. By creating well-rounded personalized ITS platforms for reading comprehension, we can make technological personalized learning for interventions or study more accessible.

2.2 Existing Intelligent Tutoring Systems for Reading Comprehension

The lack of application of researched instruction techniques in the classroom leads to a lack of this practice by students outside of the classroom [53]. By enabling students to learn at their own pace with personalized learning features like corrective feedback, adaptive scoring, and spaced repetition, ITSs can offer an alternative educational experience to traditional classroom methods for comprehension study [37].

ITSs remain one of the most affordable and accessible options for personalized learning [6, 3]. ITS effects can vary student-to-student with human tutoring traditionally providing more impact on student reading proficiency than ITSs themselves [62, 6]. Despite this, ITSs have consistently improved student reading proficiency compared to just classroom instruction [62, 6, 37], and though students of all proficiency levels see improvements [50], reading comprehension-focused ITSs (RC-ITSs) tend to benefit students with lower reading proficiency most significantly [6].

By acting as a secondary educational tool, a tutoring system can supplement

classroom shortcomings from teacher skills, lesson plans, or developmental constraints [50]. Increasing the duration and intensity of tutored sessions has been shown to benefit student performance, but is not necessary to see improvements [62, 59]. This makes ITS programs for reading comprehension adaptable as students only need to practice when they need proficiency improvements.

Many existing ITSs for reading comprehension focus on specific, narrow aspects of the skill, such as cognitive skills [36], reading level assessment [47], or phonological skills like pronunciation and fluency [38, 32]. Often, these systems prioritize cognitive or phonological skills over practical vocabulary and grammar instruction, despite research showing that vocabulary breadth and grammar understanding are essential to reading proficiency [48]. Furthermore, while comprehension-focused ITSs frequently incorporate reading passages and or reading assessments, they rarely use LLMs for dynamic passage generation. Instead, they rely on fixed datasets of pre-written content like with Mostow et al. [38], which can make instruction repetitive and reduce opportunities for personalized engagement. Most current RC-ITSs rely on basic accuracy scores and rote assessments ([36, 47, 38], which limits their ability to adapt to individual learning needs and provide meaningful, context-aware feedback.

2.2.1 My Turn To Read

My Turn To Read (MTTR) is a reading support application designed to assist low-proficiency readers in building fluency through an interleaved reading model [32]. MTTR combines audiobook narration with user-read turns, allowing children to alternate between listening and reading to reduce reading fatigue and maintain focus [32]. This interactive, turn-based approach aims to help young readers transition from word-by-word reading to fluent, enjoyable reading over separated sections of text. By pairing a narrated portion with each user turn, MTTR encourages sustained reading practice, making the experience less daunting and more engaging for children [32].

MTTR’s effectiveness was assessed in two summer camp programs, where 36 children ages 8-11 used the app consistently over multiple weeks. The evaluation demonstrated a high level of engagement, with children spending over 61 hours listening to stories and 9.5 hours answering 2,104 comprehension questions, achieving a 65% accuracy rate [32]. Surveys indicated positive user feedback, with most children believing the app helped them to become better readers. Although MTTR shows promise for enhancing fluency and motivation, it primarily focuses on reading engagement rather than targeted vocabulary or grammar instruction, areas that are foundational for comprehension proficiency [32]. MTTR also forgoes dynamic content generation in favor of pre-existing e-books and audiobooks and human-generated questions. This focus on fluency support through static data without broader comprehension assessment limits MTTR’s applicability for comprehensive reading skill development and practice.

2.2.2 AutoTutor

AutoTutor ([47]) is a dynamic, intelligent tutoring system that adaptively responds to user input. It facilitates learning through one-on-one conversations with adult users. The AI tool gains an understanding of the user’s reading comprehension level in these conversations. AutoTutor’s conversation facilitation allows the user to engage in discussions of their choice with the tool without using LLMs, opting for a guided-dialogue approach instead [47]. Conversational study may be friendlier to children and encourage them to interact with/use the tutoring system. Additionally, the dynamic features of AutoTutor are impressive and indicate the importance of the tool conforming to the user to allow for improved instruction [47].

Through the Woodcock-Johnson III Tests, researchers assessed participants reading proficiency level and then conducted sessions with AutoTutor. This tool then used user question-answering accuracy to assess user reading proficiency and was found

to fairly accurately predict proficiency in adults [47]. Despite its accuracy in user assessment, many aspects of AutoTutor went largely untested. While it performs well in assessing user proficiency, the authors do not report extensive results relating to direct improvements or improvement methods in comprehension proficiency [47]. While promising in determining reading levels in adults, this tool needs many more evolutions before being beneficial to comprehension performance in young children.

2.2.3 The Reading Tutor

The Reading Tutor ([38]) is a computer-based tool that improves children’s oral reading through speech recognition technology. This tool was developed under Project LISTEN and, although not always highly accurate, allows the program to monitor and respond to reading miscues and provide assistance for difficult words [38]. It incorporates automated skill leveling based on correct-words-read-per-minute and interactive tutorials to guide users, offering phonological support through prompts like ”rhymes with” hints or grapheme pronunciation. The program’s automation, including skill leveling and scaffolding, enables individualized pacing, reducing the need for present educators [38]. In evaluations comparing the Reading Tutor to sustained silent reading (SSR), students using the tool demonstrated significant gains in Word Identification and Basic Skills, outperforming SSR participants, particularly in word recognition and reading comprehension [38].

Despite the Reading Tutor’s strong focus on phonological support, a noted difference between it and our proposed system is the lack of grammatical and vocabulary instruction, which are fundamental for reading comprehension [48]. While the Reading Tutor aids reading fluency by responding to real-time pronunciation challenges it removes all assessment and contextual testing from reading comprehension instruction [38]. The Reading Tutor then increases student skill levels (grades) by the number of words read correctly per minute [38] despite research showing that time spent

reading is not a significant measure of a students reading comprehension ability [47]. The Reading Tutor was shown to improve student word recognition and reading comprehension as compared to SSR students [38], however the lack of domain-specific study and invalidated skill progression system limit the effect of this tool.

2.3 Advances in ITS with LLM and Spaced Repetition Features

Many classroom shortcomings can be made up for with personalized learning in ITSs [37]. LLMs and spaced repetition can enhance ITSs by providing text generation and reasoning skills as well as systematic practice intervals to programs. Spaced repetition has often been used in various flashcard programs like Anki [5] to aid term retention. LLMs appear more frequently in ITSs focused on communicative assessment or learning like Collins-Thompson and Callan [15], Chen et al. [12], Ruan et al. [45]. These conversational agents use language to promote engagement and practice in learning, mostly with factually-based information.

2.3.1 ITSs and Spaced Repetition Techniques

SuperMemo-2: Spaced Repetition Scoring

The SuperMemo-2 (SM-2) algorithm is the most widely adopted and accepted spaced repetition algorithm [60]. The SM-2 algorithm optimizes learning by adjusting repetition intervals based on item difficulty, similar to the Anki algorithm (a derivation of the SM-2 algorithm). Initially, each item is assigned an ease factor, a decimal number representing the ease with which a student can correctly answer for that item. A numerical grade is assigned to a user's responses based on their hesitation and time spent answering, as shown in Table 2.1.

Grade	Description
5	Perfect response
4	Correct response after a hesitation
3	Correct response recalled with serious difficulty
2	Incorrect response; where the correct one seemed easy to recall
1	Incorrect response; the correct one remembered
0	Complete blackout

Table 2.1: User performance grading for ease-factor in SM-2 algorithm

The ease-factor value is then adjusted based on a user’s performance grade after each repetition. The ease factor is adjusted higher for better performance and lower for worse, ranging from 1.3 to 4.0. Repetition intervals start at one day, then six days, and increase by multiplying the previous interval by the ease factor [60]. The ease factor is updated using a complex exponential formula (found at the SM-2 website) based on student performance.

When compared to other spaced repetition algorithms (like the Leitner algorithm), the SM-2 was found to result in a 10.11% improvement in test performance [14]. This performance gap indicates the algorithm’s effectiveness over others in scenarios requiring memory recall, where spaced repetition plays a critical role [14]. It also highlights the benefits of spaced repetition for term retention in academic settings.

Anki

The Anki platform is an open-source flashcard-learning program that uses a derivation of the SM-2 spaced repetition to drill terms. It is a perfect example of the numerous ITSs using spaced repetition for flashcard practice. The Anki program, specifically, is famous for its versatile flashcard mechanism and its classic UI. The Anki interface features a simple design with minimal graphics and a focus on function over form. It resembles a clean but somewhat dated look with basic icons and straightforward menus. Many paid versions have been adopted and specialized for different education

programs, including popular ones for medical students [5]. In Anki, user response is used in combination with user performance to space terms. When new words are introduced to a user, they enter the learning phase where their intervals are largely pre-determined (first a 1-day interval, then a 6-day interval) [5]. Once seen a few times, cards transition to review mode, where the user is prompted to identify the ease with which they correctly answered a question: Easy, Good, Hard, Again. If you answer 'easy', 'good', or 'hard', the ease factor is adjusted and the card interval is re-calculated. If you answer 'again,' the card re-enters the learning phase with determined intervals [5].

In 2022 a reported 70% of first-year medical students were using Anki flashcards in conjunction with medical lectures for study [20]. Anki's widespread use by medical students is a testament to its effectiveness in term retention, especially in fields with high memorization and cognitive demands. In one orthopedic surgery study, Anki was used to review terms for the OPBS exam resulting in a 92% pass rate among users [20]. This compared to the national pass rate of 67% reinforces Anki's utility in studying for exams and improving student test scores [20]. Anki has been shown to reduce burnout and improve student performance due to its reliance on spaced repetition for term retention.

Other Spaced Repetition Platforms

Other platforms like the flashcard program Quizlet and the language learning program Duolingo have implemented spaced repetition features for better spaced term retention and skill acquisition. Quizlet's Learning Assistant Platform facilitates spaced repetition for provided study terms by varying seen terms based on previous study performance and scope [35]. Duolingo implements spaced repetition in their recently updated learning path organization which varies aspects of language learning such that skills and vocabulary are tested over time in a spaced repetition format [39]. These platforms'

spaced repetition implementations have largely been untested in ITS and spaced repetition research, with most research being done on second language learning independent from the spaced repetition process. Other work like that by Amiri [4] highlights how spaced repetition can facilitate data sampling in neural network training and other background machine processes. To see the complete and more detailed implementations of the Quizlet, Duolingo, Anki, and SM-2 spaced repetition algorithms, please visit the linked websites or the sources in Section A.4.

2.3.2 ITSs Utilizing LLMs for Personalized Learning

Web Corpus Grade Prediction Model

The Web Corpus Grade Prediction Model leverages a Smoothed Unigram variation of a multinomial Naïve Bayes classifier to assess readability by grade level [15]. The authors created a custom corpus of 550 documents across 12 grade levels, given that no comprehensive dataset labeled by reading difficulty as grades existed [15]. Each grade level was modeled with a unique language model based on unigrams, where words are understood independently, allowing effective training with limited data. This approach aimed to discern unique vocabulary across grades, predicting reading difficulty by grade by determining the likelihood of word distribution in a text. The unigram model yielded accurate classifications despite the model’s simplicity showing a language model’s ability to discern texts by grade level through word use [15].

Compared to traditional measures such as the Flesch-Kincaid readability score and the type/token ratio, the Smoothed Unigram Model achieved stronger performance across web data and was particularly robust for shorter passages, making it suitable for fragmented or brief web content. Its evaluation showed 90% vocabulary coverage on test sets, demonstrating reliable performance even with minimal training data [15]. By emphasizing relative word frequencies across grades and smoothing techniques, this model provides a scalable, cross-linguistic approach to reading level assessment

that could complement traditional readability tools in digital applications [15].

ChatTutor

ChatTutor is a chat-based ITS that provides an information on a selected topic through user conversations. The system is structured around three core processes: Interaction, Reflection, and Reaction [12]. Interaction facilitates learning through dialogue, maintaining user engagement by tailoring responses to user responses or requests for information, Reflection monitors user learning progress and preferences, and Reaction adjusts the course plan and generates quizzes based on the insights from Reflection and Interaction [12]. System interaction is controlled using a Meta Agent and memories are used to maintain long-term context [12]. Together, these processes facilitate user conversations to deliver an interactive and personal learning experience.

The ChatTutor system was evaluated using pre-defined topics for conversation, ranging in complexity to test the system’s robustness. Results indicate that ChatTutor’s memory-based Reflection and Reaction processes improve instructional consistency and quiz relevance. Participants report experiencing consistent guidance and fine-grained quiz questions tailored to their progress with models using Reflection and Reaction models. Comparisons of ChatTutor with versions without the Reflection or Reaction processes show ChatTutor’s superior stability, with fewer repetitive or omitted objectives. However, it was noted that ChatTutor’s responsiveness may lag due to the intensive backend processing required for memory updates, hinting at the trade-off between adaptiveness and prompt interaction in LLM-based ITS applications. This structured yet adaptive framework from ChatTutor offers a potential guideline to inform future ITSs that incorporate LLMs for personalized learning.

QuizBot

QuizBot is a chat-based learning system that uses LLMs, semantic similarity, and adaptive question-sequencing to facilitate user dialogue. This allows QuizBot to ask the user questions on a given topic to foster factual knowledge retention [45]. Using a structured quiz mode and casual chat mode, QuizBot engages students with questions via both typing and selection-based responses. The adaptive question-sequencing algorithm, selects questions based on prior student performance, while a supervised semantic similarity model compares student answers to target responses, offering hints and explanations when needed [45]. Through features like "Frosty" the penguin and corrective feedback, QuizBot aims to create a conversational, low-pressure study environment, making learning more immersive and less draining to students [45].

Two studies compared QuizBot with a traditional flashcard system, focusing on both effectiveness and engagement. In controlled studies, QuizBot enabled a higher rate of accurate recall, with students correctly answering 21.4% more multiple-choice and 21.1% more fill-in-the-blank questions compared to flashcards [45]. Further analysis indicated that QuizBot was over twice as engaging; students used it longer and more frequently. This increased engagement led to better long-term retention, demonstrated by students' improved performance on delayed post-study tests [45]. These results underline QuizBot's potential to support long-term learning for factual content, suggesting that conversational agents might be particularly beneficial for students motivated by interactive, human-like study experiences over time-bound test preparation.

2.4 Novelty of Proposed Personalized Intelligent Tutoring System: Read with Emora

Our proposed RC-ITS, Read with Emora (RwE), uniquely integrates structured LLM interaction for guided, but personalized comprehension learning. Our LLMs facilitate passage generation and question generation, allowing for a greater variety in student review and study content. This content generation process is specific to different grade levels through hybrid prompting and maintains variety in reading material, reducing the repetitiveness of conventional ITSs that rely on static datasets like Mostow et al. [38] or Madnani et al. [32].

Question generation is a natural language processing task for reading comprehension, intelligent tutoring systems, and virtual assistants. Most current question generation platforms use complex language models trained on question-answer datasets to generate questions from a given single sentences [29, 61, 10]. Our platform implements novel question generation techniques by basing questions off of pivotal sentences given the context of the whole passage. We combine current available research techniques into this novel approach by relying on the context of both the individual sentence and the passage as a whole using GPTs.

In addition to its LLM-driven content generation, RwE maintains novelty by providing personalized learning features like corrective feedback specifically designed for reading comprehension. Drawing from systems like Ruan et al. [45]’s QuizBot and [12]’s ChatTutor and Tsai et al. [54]’s Linggle that offer corrective feedback and structured LLM interactions, RwE’s feedback is embedded within reading tasks, guiding students to improve comprehension skills without shifting focus away from the passage, a feature not incorporated in modern reading comprehension-focused ITSs. By adapting these principles to reading comprehension specifically, RwE ensures that RC-ITSs receive the same attention in technological advancement as ITSs for general

study and factually-based content like QuizBot, ChatTutor and the Web Corpus Grade Prediction Model.

Unlike typical reading comprehension systems, which often focus on cognitive, phonological or other niche skills, RWE facilitates reading sessions incorporating practical, in-context vocabulary and grammar integration to passages and questions. This approach addresses an essential gap found in other ITSs, which tend to either focus on isolated vocabulary development [5] or lack grammar integration altogether [38, 47, 36, 32]. By contextualizing vocabulary and grammar learning within reading comprehension tasks, RWE avoids creating a stressful rule and memorization-focused environment for learners.

RWE incorporates adaptive scoring and spaced repetition. Our system uses LLMs for student response scoring and a unique derivation of the SM-2 algorithm to vary content review intervals based on individual performance and amount of feedback needed. This further varies the context in which vocabulary and grammar is seen by ensuring students are tested first on contexts that are most challenging to them.

Ultimately, RWE aims to create an advanced reading comprehension tutoring experience by addressing classroom shortcomings with LLM-powered personalized learning. With its integration of LLMs for content generation, contextualized vocabulary and grammar instruction, and adaptive scoring techniques, RWE is positioned to meet diverse reading comprehension needs in ways that existing RC-ITS platforms do not. This system offers varied contexts and promotes a balanced development of reading skills, tailored to the student’s unique learning pace and proficiency level.

Chapter 3

Read with Emora (RwE)

3.1 Materials and Data Collection

Before gathering our materials, we reviewed existing research on reading comprehension and identified three essential areas for data collection: learning objectives, grammar, and vocabulary. Learning objectives are the foundation of a child’s reading comprehension skills, focusing on concepts such as ”understanding key details” or ”comparing and contrasting stories.” By creating a well-researched and organized set of these objectives, we can generate prompts that align with specific goals and assess students across various contexts through reading passages. Generating these passages also requires a carefully curated list of grammar rules and vocabulary words, each tailored to the appropriate grade level. Learning objectives, grammar rules, and vocabulary are distinct for each grade to ensure age-appropriate content and challenges. We focus on reading comprehension for grades one through five as our goal is to aim this tool toward children in elementary school.

3.1.1 Learning Objectives

Each state produces a set of Department of Education approved guidelines on reading and literacy for education by grade which outline reading objectives for students in that state. These guidelines' exact details, structure, and content differ by state, but many follow a similar structure to the Massachusetts DOE (MDOE) guidelines for reading, comprehension, and literacy due to Massachusetts' education's national acclaim. Massachusetts public schools produce some of the highest-performing students in the country, and they continuously rank among the highest proficiencies on the NAEP report [1]. We choose to adopt the MDOE guidelines for children's reading proficiency due to their national standing in education and effortless objective organization.

Structure of Learning Objectives

For each grade, the Massachusetts Department of Education (MDOE) outlines three primary categories: 'Reading Standards for Literature,' 'Reading Standards for Informational Texts,' and 'Reading Standards for Foundational Skills.' The first two categories consist of a comprehensive list of learning objectives and detailed explanations of how these objectives can be assessed through reading comprehension and textual analysis of literature and informational texts [33]. The third category, 'Reading Standards for Foundational Skills,' emphasizes classroom techniques such as phonics and fluency development, typically assessed through direct teacher observation rather than standardized testing [33].

In our work, we opted to incorporate the 'Reading Standards for Literature' (RSL) and 'Reading Standards for Informational Texts'(RSI) as our learning objectives since the skills covered by these sections were more easily translatable to automated assessments. As shown in Figure 3.1, the RSL category is divided into three main categories (orange) for first through fifth graders: 'Key Ideas and Details,' 'Integration of Knowledge and Ideas,' and 'Craft and Structure. Within each of the main, orange

categories, there are 1-3 (green) learning objectives. The MDOE lists explanations of how the objective relates to and can be incorporated into student learning.

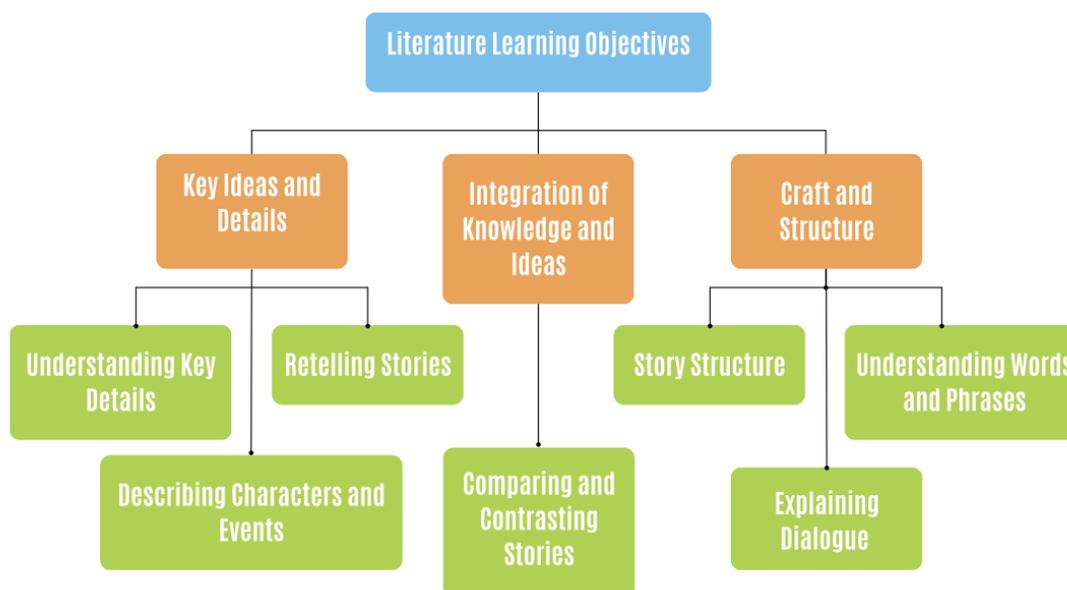


Figure 3.1: General Literature Learning Objective Structure Based on the Massachusetts DOE Guide

For RSI, Figure 3.2 shows the two main (orange) categories of objectives for first through third graders: 'Key Ideas and Details' and 'Integration of Knowledge and Ideas', each with 1-3 objectives as well. The exact number of main and subcategories per RSL and RSI vary by grade, but follow the general structure in Figures 3.1 and Figure 3.2.

The Massachusetts reading standard for literature and informational texts helps students develop critical skills like understanding rhetoric, analyzing centralized themes and ideas, evaluating arguments, and interpreting complex texts [33]. These objectives serve as a framework for the prompts fed into our passage generation model. These prompts outline specially curated story narratives that provide varied contexts for our passage generation model.

There is an additional fourth category in each Literature and Informational section

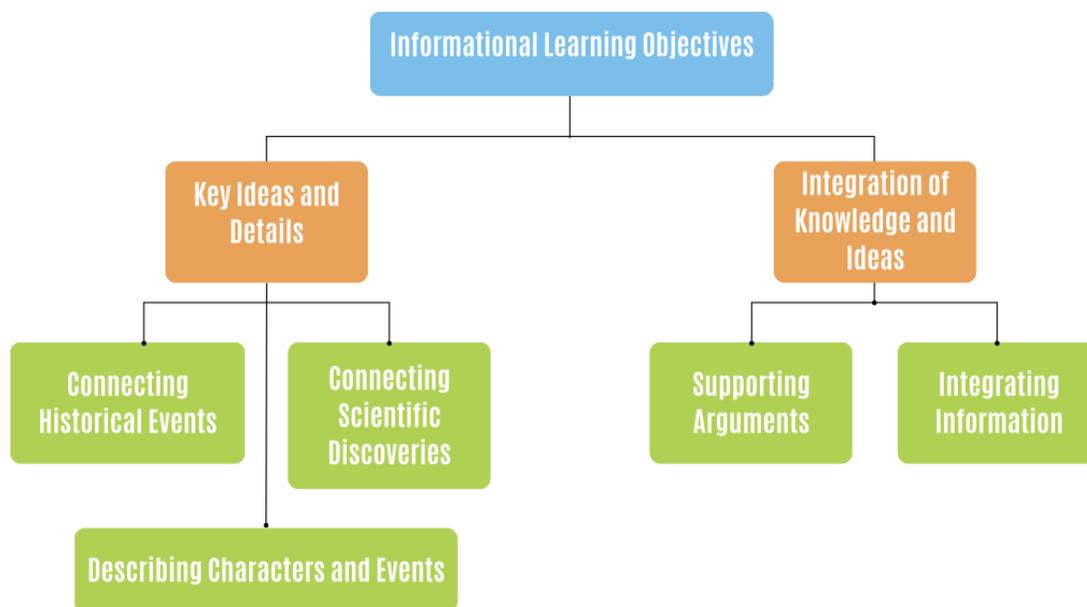


Figure 3.2: General Informational Learning Objective Structure Based on the Massachusetts DOE Guide

titled 'Range of Reading and Level of Text Complexity.' This category contains a single description per grade, such as, "Independently and proficiently read and comprehend literary texts representing a variety of genres, cultures, and perspectives and exhibiting complexity appropriate for at least grade 2" for grade 2 [33]. We elected not to incorporate this category to prevent redundancy, as it consistently summarizes the goals and intentions of the previous three categories.

Learning Objectives as Data

We provide the complete learning objectives applicable to grades first through third in Table 3.1.1. Learning objectives for all other grades can be found in Chapter A in Table A.1 and Table A.2. Fourth Grade objectives expand on the sets provided for previous grades by adding a 'Craft and Structure' section under the RSI designation through which the 'Article Structure' objective is found. Fifth grade expands on this further by adding the 'Comparing and Contrasting Perspectives' objective to the RSI

'Craft and Structure' section.

RSL	Key Ideas and Details	Understanding Key Details
		Retelling Stories
		Describing Characters and Events
	Craft and Structure	Understanding Words and Phrases
		Story Structure
		Explaining Dialogue
Integration of Knowledge and Ideas	Comparing and Contrasting Stories	
RSI	Key Ideas and Details	Connecting Historical Events
		Connecting Scientific Discoveries
		Describing People and Events
	Integration of Knowledge and Ideas	Supporting Arguments
		Integrating Information

Table 3.1: First Through Third Grade Learning Objectives Based on the MDOE Guide

Although the learning objectives are consistent across grades, the way each grade is assessed on these objectives varies. For every objective, we include both a prompt and a purpose in our data. The objective prompt guides passage generation, ensuring that specific elements, such as plot structure, character development, or moral lessons, are appropriately integrated into the passage. Story premises provide a clear framework for generating engaging, logical narratives for LLMs [31]. The objective purpose is used in the purpose generation phase, where it helps articulate how the passage aligns with the assigned learning goals.

As an example, the first-grade learning objective "Understanding Key Details" includes the following prompt:

Ensure that there is a main character in the story generation.

- ↪ Make sure that the main character faces certain challenges
- ↪ during the story, and detail the main character's
- ↪ procedure for tackling the challenges.

and the following purpose:

This learning objective aims to help students focus on the

- essential details of a story, ensuring they grasp the
- fundamental elements such as characters, settings, and
- major events. The passage is intended to have a main
- character who experiences and tackles challenges through
- the narrative. Students will be challenged to remember and
- analyze key details from the story that affect plot
- progression and character development.

By varying the prompts for each objective, we ensure that the passages generated remain diverse and closely aligned with MDOE standards. These prompts are adapted to reflect grade-specific requirements, ensuring that passage content is appropriate for the student’s educational level. Objective prompts are consistent across grades where the MDOE sets similar goals across grade levels. Occasionally higher grades may see a learning objective description increase in complexity or skill demand. In these cases we upgrade the objective prompt to reflect any upgrade in objective description by the MDOE. We are confident this approach does not detract from grade-level personalization, as the prompts are enriched with grade-specific details such as vocabulary, grammar rules, customizable topics, and word count. The LLM-models (OpenAI’s gpt-4o-mini) are explicitly provided with the student’s grade level to ensure that the generated passage adheres to the appropriate scope and complexity. For more details on this process, refer to Section 3.3.2 on passage generation and Section 3.3.3 on purpose generation.

Excluded Learning Objectives

We chose not to adopt the ‘Reading Standards for Foundational Skills’ because many of these competencies—such as giving presentations or demonstrating oral fluency—rely heavily on teacher observation and in-person feedback, making them difficult to assess

through an AI tutoring program designed for comprehension [33]. Additionally, some learning objectives or student standards were excluded from our data due to their inability to be translated into our online format. Most of these objectives were excluded because of our inability to track or test the information requested. In all cases we either already test the information through our website, cannot test the information due to development constraints (like image incorporation), or exclude the objective because testing the information would be redundant to other incorporated objectives.

One example of objective redundancy is the excluded Craft and Structure objective for RSI: "Determine the meaning of words and phrases in a text relevant to a grade 2 topic or subject area" [33]. This objective was determined redundant because of its focus on two aspects, "determine the meaning of words and phrases in a text" and "words and phrases in a text relevant to a grade 2 topic or subject area". All passages in RWE are generated with prompts that rely on grade-specific learning component data (vocabulary, grammar, objectives) and specify the grade in the prompt, ensuring all content, words, and phrases are relevant to the grade of the student. Additionally, the corresponding included RSL Craft and Structure objectives contains the objective "Understanding Words and Phrases" which is defined in the MDOE as: "Describe how words and phrases (e.g., regular beats, alliteration, rhymes, repeated lines) supply rhythm and meaning in a story, poem, or song" [33], which directly tests a student's understanding of the meaning of words and phrases relevant to a text. Since we test the targeted skill (understanding words and phrases) this excluded objective presents in other RSL included objectives, and ensure grade relevant content across our program, we consider the excluded objective redundant and remove it from our corpus of objectives. Similar analysis processes were done across all learning objectives for all grades.

Table 3.2 shows the complete list of excluded objectives for our second grade learning objective data. Excluded objectives for grades one, three, four, and five can

be found in Chapter A in Tables A.3, A.4, A.5, and A.6.

Objective	Description	Exclusion Reason
Integration of Knowledge and Ideas	Use information gained from the visual aids and words in a print or digital text to demonstrate understanding of its characters, setting, or plot	Untestable - requires visual aids
Craft and structure	Determine the meaning of words and phrases in a text relevant to a grade 2 topic or subject area	Tested - through other objectives
Craft and structure	Know and use various text features (e.g., captions, bold print, subheadings, glossaries, indexes, electronic menus, icons) to locate key facts or information in a text efficiently	Tested - through website use
Craft and structure	Identify the main purpose of a text, including what the author wants to answer, explain, or describe	Tested - through other objectives
Integration of Knowledge and Ideas	Explain how specific visual aids (e.g., a diagram showing how a machine works) contribute to and clarify a text	Untestable - requires visual aids

Table 3.2: Objectives Excluded from Second-Grade Learning per the MDOE

3.1.2 Grammar Rules

We took time to reach out to authors and paid for some academically sourced grammar and vocabulary literature out of pocket to ensure we reviewed all possible avenues of academically backed sources. Unfortunately even proprietary resources were not grade-specific and no authors responded to our inquiries for access or more information. We found numerous English Language Arts textbooks providing grammar information with no indication of one being more reliable than another. All textbooks found were also proprietary, limiting our access. For these reasons we opted to source grammar rules and vocabulary from less academically-robust sources like online educational

websites. While we recognize the issues with using non-academically-robust sources for our data, this project focuses more on the integration of learning components through advanced LLMs rather than the components themselves. We hope in the future grammar and vocabulary can become a customizable feature of our program, allowing students to upload, change, and delete rules and words. This would add further personalization to our platform.

Sources and Preprocessing

For this project, grammar rules were gathered primarily from two educational websites: Big Learners [7] and Night Zookeeper [40]. The rules from these sources were straightforward, and minimal preprocessing was needed to adapt them to our data structures. After finding these sources, we removed duplicates and ensured consistency in rule progression across grades. While these sources provided a basis of understanding for general grammar rules, more robust, academically backed sources specific to grammar rule labeling per grade were not found or were not available to the authors, highlighting a significant gap in grade-specific elementary grammar rules and their translation to RC-ITSs.

Most available academic grammar sources offered rules as general guidelines rather than grade-specific content. For example, resources like Cowan [17] and Conklin [16] present grammar rules in a comprehensive list with great detail in the instruction but do not distinguish between individual grades. Similarly, other websites like Learning at the Primary Pond provide grammar conventions for K–second-grade students, but lack the detailed breakdown across all five grades we aimed to implement in our project [30]. Table 3.3 lists the grammar rules compiled for second-grade data. Full lists of the grammar rules for each grade can be found in Chapter A in Tables A.7, A.8, A.9, A.10,

Grammar Rule
Sentence Structure
Complete and Incomplete Sentences
Subject and Predicate / Expanding Sentences
Statements, Questions, Commands, Exclamations
Parts of speech
Nouns/Pronouns
Collective and Plural Nouns
Reflexive Pronouns
Verbs / Adjectives / Adverbs
Punctuation
Commas / Contractions
Quotation Marks / Apostrophes
Capitalization
Spelling and Spelling Patterns
Short Vowel Sounds
Long Vowel Sounds
Blends, Digraphs, and Diphthongs
R-Controlled Patterns
Hard and Soft Sound of C and G
Variant Vowels / Y as a Vowel
Final E Sound / Silent Letters

Table 3.3: Second Grade Grammar Rules per Big Learners and Night Zookeeper

3.1.3 Vocabulary Words

Sources and Preprocessing

In a process similar to our approach with grammar rules, vocabulary words were collected from Flocabulary [19] and k12Reader [26] that serve as online educational data sources for teachers. While other websites, such as Vocabulary provided additional lists, they mostly contained duplicate entries from the original two sources [56]. Preprocessing the vocabulary data involved converting all words to lowercase, alphabetizing the lists, and removing duplicates.

As with grammar rules, we encountered difficulties locating academic journal articles containing specific vocabulary lists. Although we searched for peer-reviewed sources, available materials containing grade-specific english vocabulary were rare and, when found, proprietary or focused on general vocabulary instruction without a detailed breakdown of word suggestions by grade. This lack of accessible, grade-specific vocabulary resources further complicated our efforts to source vocabulary from academic literature, as with grammar. In the future, this problem should not persist as we hope to develop our program to allow teachers or parents to input custom vocabulary words and grammar rules from their children's or students' classrooms, enabling a more personalized and adaptable learning experience tailored to the specific curriculum needs of each student.

Vocabulary Words by Grade

As students progress through grades, their exposure to vocabulary expands, reflecting their growing ability to comprehend language in more complex contexts. This requires a broader and richer vocabulary with more complex words. Table 3.4 displays the total number of vocabulary words assigned to each grade level, showcasing how the volume of vocabulary words increases as students advance.

Grade	Number of Vocabulary Words
First Grade	437
Second Grade	676
Third Grade	701
Fourth Grade	847
Fifth Grade	847

Table 3.4: Number of Vocabulary Words by Grade as Sourced from Flocabulary and k12Reader.

Due to the extensive vocabulary lists, we have included Table 3.5, which displays the first ten alphabetically ordered vocabulary words for each grade. This table makes apparent the increasing complexity of vocabulary words for higher grades through word length and structure. Vocabulary words for higher grades are less likely to consist of high-frequency words commonly encountered in everyday English. For access to the complete list of vocabulary words, please refer to the data available on our GitHub repository, linked in Chapter A.

First Grade	Second Grade	Third Grade	Fourth Grade	Fifth Grade
about	able	ability	absolute	abacus
above	acceptable	absorb	abuse	abolish
ache	act	account	academic	absent
adjust	add	accuse	accurate	absurd
affordable	addition	act	achieve	abuse
alarm	adorable	action	across	access
all	affect	active	acrylic	accompany
alone	after	actual	actor	accomplish
an	again	actually	actress	according
and	age	adapt	acute	achievement

Table 3.5: First Ten Vocabulary Words per Grade.

3.2 Theoretical Approach and Core Algorithms

3.2.1 Adaptive Scoring System

The adaptive scoring system in our program plays a crucial role in providing immediate feedback and adjusting learning objectives based on a student's performance. The system is designed to offer corrective feedback through our grading LLM-model, allow for multiple attempts on each question, and track progress over time. The adaptive nature of feedback and grading allows students to understand and correct their mistakes. We also include a results card and a Progress page so teachers or parents can quickly review session results and track user progress over time.

LLM-Generated Corrective Feedback

When a user submits a response, our AI model grades the answer and provides feedback. In the case the student does not achieve full credit, the model generates a comment to guide the student toward the correct answer, and the question is presented again for another attempt. The student may answer up to three times, receiving LLM-generated feedback at each attempt. The final score is recorded after the last attempt and displayed to the user after all questions are answered.

Session Review with Results Card

Once the student has answered all questions for a passage, the system generates a results card. This card is displayed on the session page in place of the question card and provides detailed feedback on each question. It lists:

- The original question
- The student's response
- A gold standard answer

- The grade received
- The number of attempts taken

The gold standard answers are LLM-generated responses to questions given the passage context. These allow students a point of comparison for questions they may not have understood or answered incorrectly. This allows students to review their answers immediately with a teacher or parent, even if they were not monitored during the reading and question-answering process.

Tracking Performance Over Time

In addition to immediate feedback, the system tracks performance over time through a dedicated results page. This page displays the ease factor scores of each learning objective as percentages. A bar chart showing performance on the last five passages helps users analyze student learning progress trends through average attempts and average score values. Finally, three percentages represent a student's average attempts, average score, and average learning objective score over all time. This page is shown and discussed more in Section 3.4.1.

3.2.2 Spaced Repetition and Learning Algorithms

Adaptation to Reading Comprehension

Our algorithm adapts the SM-2 ease factor technique to grade learning objectives without directly relying on student responses. As derived from the MDOE guidelines, each learning objective is assigned an initial ease factor of 2.5 when a user first instantiates passage generation. After reading the passage, students are presented with context-based questions, answered one at a time.

Our grading AI model provides comments to the student if an answer does not receive full credit as explained in Section 3.2.1.

RwE Ease Factor Calculation

Our model grades all questions on a 0 to 2 scale. 0 represents a completely incorrect answer, 1 represents a partially incorrect answer, and 2 represents a fully correct answer. After answering all questions for a passage (of which there are at least four and up to ten), we calculate the average score and average attempts over all questions for that passage. First, the scoring average is adjusted by a penalty factor based on the number of attempts, with higher attempts resulting in a more significant penalty. We normalize the average attempts (which is originally on a 0 to 2 scale) by dividing it by 2. We then multiply that by 0.25 to apply a penalty of 0 to 0.25, ensuring that students are not overly punished for multiple tries while still accounting for number of attempts in our scoring. This adjustment is designed to slightly penalize students for taking multiple attempts without drastically lowering their score to the next level. In working with average scores and attempts rather than individual scores and attempts, we avoid overly penalizing or rewarding student performance based off of only one or two student responses, generalizing performance to the entire passage. This is shown in Equation 3.1 where 'score' represents the average score and 'attempts' represents the average attempts over all questions for a passage:

$$s = \text{score} - \frac{\text{attempts}}{2} \times 0.25 \quad (3.1)$$

This adjustment ensures that students are encouraged to reach the correct answer with minimal attempts while allowing for learning through retries. It balances feedback and progression without discouraging students from guessing or trying their best. The ease factor is updated based on the following Equation 3.2:

$$\text{update_ease}(s, e) = \begin{cases} \max(1.3, e \times 0.8), & \text{if } s \leq 0.5 \\ \max(1.3, e - 0.15), & \text{if } 0.5 < s \leq 1 \\ e, & \text{if } 1 < s \leq 1.5 \\ \min(4.0, e + 0.15), & \text{if } s > 1.5 \end{cases} \quad (3.2)$$

where s is the updated score and e is the current ease factor. If a student's average score after adjustment is below a 1, the ease factor is penalized representing the student's difficulty in attaining full credit. If the average score is higher than 1.5, the ease factor increases, representing a student's mastery of that objective. The ease factor remains the same for intermediate performance.

Learning Objectives Importance in Ease Factor Calculation

The learning objectives are the foundation of the passage generation process. Since the context-based questions are derived directly from the passage content, the student's performance on these questions reflects their understanding of the underlying learning objectives. By adjusting the ease factor based on question performance, the program tracks the student's progress over time, ensuring that the content is tested more or less depending on student performance.

It is important to note that the equations used in this system, such as the ease factor and score adjustments, are well-considered but ultimately flexible. The equations' exact values and ratios can be fine-tuned in the future based on user performance and response data to optimize learning progression.

3.2.3 Grammar and Vocabulary Integration

Teaching Grammar in Context

Research indicates that grammar instruction for young learners is most effective when introduced through engaging activities, such as storytelling, songs, and chants, rather than direct rule-based instruction. Consciousness-raising activities that bring grammatical forms to students' attention without relying on complex terminology are preferred by students over traditional methods [24]. We implement this approach by embedding grammar rules in our passages and questions, ensuring that students are naturally exposed to these rules through LLMprompting and prompt engineering rather than being directly tested.

To align with recent research advocating for a breadth-based approach to grammar instruction [41], we randomize the grammar rules applied in each passage, selecting them from grade-specific sets. This approach increases the contextual understanding of grammar with a more varied approach than introducing one at a time. It avoids isolating individual rules, promoting the inter-relatedness of grammatical concepts—a critical aspect often missing from traditional American classrooms [41]. Additionally, our corrective feedback mechanism is designed to identify and correct significant grammar or spelling errors, which addresses the need for handling and correcting student errors, a facet of instruction that is well-supported by research but underutilized in classrooms [11].

Each passage is generated with two grammar rules incorporated. The rules used are displayed on the interface, allowing teachers and parents to adapt the program further to their needs. Educators can use this information to create exercises that help students identify where grammar rules were applied in the passage or track concepts that students have studied.

Teaching Vocabulary in Context

Our vocabulary integration follows a similar structure to grammar. A random set of vocabulary words drawn from grade-specific word lists is included for each passage to promote exposure and contextual learning. Context-building and exposure are critical factors in effective vocabulary acquisition, as they are in grammar study. The vocabulary words are also incorporated into the questions through prompt engineering, which is further discussed in Section 3.3.4.

In addition, we provide a list of vocabulary words alongside the displayed grammar rules for each passage, enabling teachers and parents to tailor supplemental exercises or activities to reinforce vocabulary or grammar. This flexibility allows for greater adaptability and customization of the learning process based on individual student needs with minimal effort from the parent or teacher.

3.2.4 RWE Personalization

Our program incorporates personalization at multiple levels to enhance the student learning experience. This includes grade-specific personalization, which tailors content to a student’s grade level, and user-specific personalization, which adapts the learning process to each student’s individual needs and preferences. Further, the integration of existing samples from data (e.g. vocabulary words or grammar rules used in a passage) with LLM content generation to drive story narrative creates combinatorial creativity that ensures diversity and originality between passages [31].

Grade-Specific Personalization

Grade-specific personalization is driven by variations in story length, vocabulary, grammar rules, and learning objectives that align with the student’s grade level. Each grade has a set of recommended vocabulary lists, grammar rules, and learning objectives, with some overlap in vocabulary across grades for more common words.

Passage length is determined based on the typical reading speed for each grade, as reported by Brysbaert [9], outlined in Table 3.6.

Grade	Words Per Minute (WPM)
First Grade	80
Second Grade	115
Third Grade	138
Fourth Grade	158
Fifth Grade	173

Table 3.6: Words per minute by grade level

Research shows that reading time is not necessarily an indicator of reading proficiency [47]. We aim to ensure that students are neither rushed nor stressed by timing constraints. Although research on attention span in younger children is limited, one study suggests that attention begins to wane after 15-30 minutes in university students [8]. To maintain engagement, we designed passages to have 3-5 minutes reading duration, depending on the grade level. This is followed by our question answering section which we usually expect to take 5-15 minutes.

Grade	Recommended Passage Length
First Grade	100-200
Second Grade	200-300
Third Grade	300-400
Fourth Grade	400-500
Fifth Grade	500-600

Table 3.7: Recommended Passage Length by Grade Level

We incorporated the length recommendations in Table 3.7 into the AI prompts, allowing the model to generate passages that match the appropriate reading times. We found that by providing a range instead of a single number as a length suggestion, the passage generator was more likely to output complete stories meeting the given requirements which is analogous to findings that GPT models perform poorly when given specific numerical constraints [52]. Additionally, through piloting we found that the passage lengths tended to go just over the recommended cap which encouraged us

to provide passage ranges that were slightly shorter than our intended passage length. In cases where students find passages too simple, they are encouraged to increase their grade level to encounter more complex prompts, vocabulary, grammar rules, and longer passage lengths. Older grade levels include more advanced language arts concepts, such as figurative language and narrative structure.

User-Specific Personalization

User-specific personalization is at the core of our intelligence system, enabling customization based on the individual student's needs. Students can choose the topics for their passages, allowing them to engage with content that interests them while reinforcing vocabulary and grammar in varied contexts. Prompt engineering provides a method of moderation for content generation as the LLM model is instructed to only generate grade-appropriate context and to ignore overly sensitive, dangerous, or illegal topics. Furthermore, students can adjust their assigned grade level through the Profile Page to ensure the learning experience is neither too challenging nor too simple.

One feature of our system is its adaptability through dynamic logic. Parameters such as the number of questions, vocabulary words, grammar rules, and passage length are randomized within appropriate grade-level ranges with every generation, ensuring variety while maintaining alignment with learning objectives. These learning objectives are selected based on the individual student's ease factor score—the system dynamically chooses objectives with the lowest ease factor, allowing students to focus on areas where they need the most improvement.

Our program's most significant element of user-specific customization is the corrective feedback mechanism, as detailed in Section 3.2.2. Our corrective feedback provides a response to a student answer submission. If a student answer is graded less than perfect (less than 2) corrective feedback is displayed to the UI noting what the

student answer missed or why it was incorrect, and the student is allowed to resubmit their answer. Feedback prompting encourages the LLM to flag and correct significant grammar or spelling errors, reinforcing proper in-context usage. Photos and examples of these personalized features are available in Section 3.3.

3.3 Implementation of Key Algorithms and Features

Read with Emora’s key personalization features are implemented in a set structure that drives our AI model communication. We facilitate passage generation, passage-purpose generation, important sentence identification, question generation, student response grading, and gold standard answer generation through six different AI models based on GPT technologies, specifically OpenAI’s gpt-4o-mini LLM. These six models are fine-tuned to their independent functions, with each user prompt serving as a structured template for dynamic data, increasing our work’s replicability [22]. We create a separate API call for each generation based on separate system and user prompts. For clarity, we will refer to these six models as Passage AI, Purpose AI, Sentence AI, Question AI, Grader AI, and Answer AI, respectively. We ensure our user prompts consistently include appropriate previous context (such as the original reading passage) to ensure models like Question AI and Answer AI can provide long-form context-based generated texts as recommended by LLM content generation research [63]. Additionally, our user prompts are structured in a template format where data sent to our LLMs is changed depending on the details of a user’s session (like vocab or passage viewed most recently). Using prompt templates increases the replicability of our work and supports a guided learning experience [22].

3.3.1 Program Structure

RC-ITSs generally focus on providing a length-appropriate passage to be read and analyzing the student’s proficiency based on that reading [38, 47, 36]. This is sometimes followed by a set of context-based questions on that passage, and a scoring section where student responses are graded or proficiency is assessed [47, 34]. Context-based questions, even when generated automatically like by LLMs, have been shown to boost reading proficiency [34] and we incorporate this testing on reading passage comprehension in our system. To facilitate this process, we defined a distinct program structure mirroring the structured LLM-interactions of previous LLM-based ITSs [45, 12]. In following a similar outline to previous work, we hope students can find familiarity in experiencing our program’s structure and so that this program could help students practically in future assessments that might use that same structure.

We updated this with corrective feedback to follow a conversational structure more closely, as we wanted to model a helpful teacher guiding a student. This structure is outlined in Figure 3.3. First our program generates a passage with Passage AI. Using that passage a passage-purpose is generated too. After reading the passage, Sentence AI generates important sentences which are fed to Question AI to generate questions. Once questions are generated we also generate a set of gold standard answers through Answer AI, based on those questions. Lastly Grader AI grades and provides iterative feedback to students as they submit answers.

3.3.2 Passage Generation

The passage generation algorithm is the first step in our program, responsible for generating reading passages tailored to the student’s grade level, learning objective, topic, length, vocabulary, and grammar knowledge. The process involves several steps, including prompt crafting, data retrieval, and communication with an AI model (Passage AI), and is triggered by pressing the ”Generate Passage” button on our

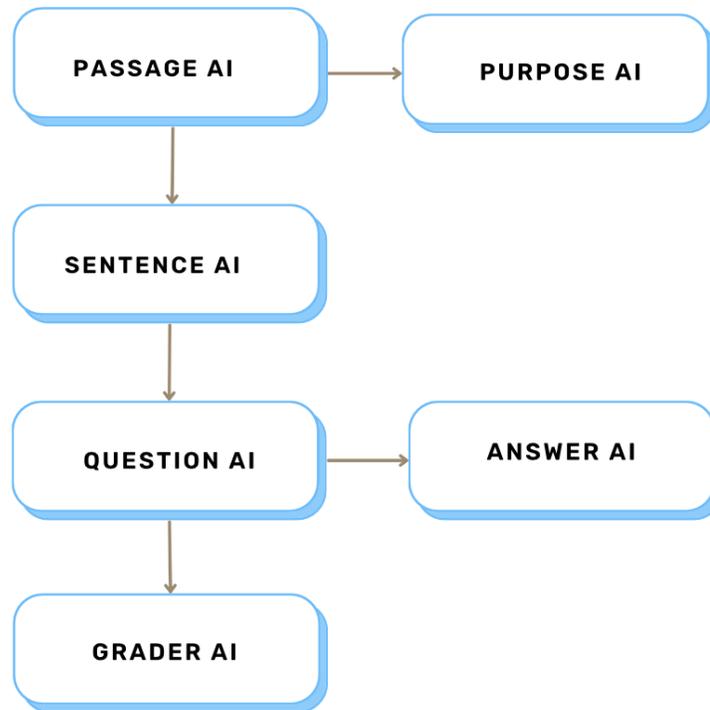


Figure 3.3: Program Structure Outline through our Designated AI Models

Session Page. Below is an in-depth look at this process.

Prompt Crafting:

The passage generation system creates a prompt based on the student's grade, an identified learning objective, and specific vocabulary and grammar features as learning components to ensure a generated passage incorporates educational standards. We also integrate a user-specified topic into passage generation so that reading material remains specific to user preferences which has the potential to increase student engagement. The prompt is sent to the Passage AI model, which generates a passage for reading comprehension. The system selects these features as follows:

- **Vocabulary:** Based on the student's grade, a random set of vocabulary words is selected from the grade-specific JSON file. Higher grades allow for more vocabulary incorporation.

- **Grammar:** Two grammar rules are chosen based on the student's grade level.
- **Topic:** The user provides a topic for the passage, or the system defaults to a predefined one: "something fun and interesting."
- **Learning Objective:** The LO is selected from our database based on the student's ease factor and tracked objectives. LOs with lower ease factors and those not recently tested are prioritized. From the LO we receive a 'literature_type' based on the chosen prompt. This may be a story, poem, biography, or article.
- **Length:** The student's grade level determines the passage length as per Figure 3.7.

The prompt is then formatted as follows:

```
Write a ${literature_type} for ${grade} graders to read on
→ ${topic}. It should try to include most of the vocabulary
→ words in this list: '${selected_vocab}' and at least one
→ example from this list of rules: '${selected_grammar}.'
→ The story should be about ${length} words long.
```

Through this process:

1. Select a random set of vocabulary words based on the
 - student's grade.
2. Select two grammar rules based on the student's grade.
3. Format the prompt using:

```
Vocabulary words: {selected_vocab}
```

```
Grammar rules: {selected_grammar}
```

```
Topic: {topic}
```

```
Grade: {grade}
```

```
Literature Type:
```

```
{objective_details.Prompts[literature_type]}
```

```
Length: "{grade * 100} - {(grade + 1) * 100}"
```

4. Concatenate user prompt with learning objective prompt
→ containing passage and narrative-specific information.
5. Return the formatted prompt.

Explanation: Here, we describe our hybrid prompt crafting, dynamically incorporating features such as the student’s grade, topic, and specific learning objectives. The ”Length” feature prints as a string so for a grade 2 student, a length of ”200 - 300” would be included. This process ensures that the generated passage is appropriate for the student’s learning level while adding variability in data and covering a particular educational objective. This is essential for tailoring the passage to the individual student’s grade-specific needs and user-specific preferences.

Passage Generation from Prompt:

Once the above user prompt is generated, the system communicates with the Passage AI model to generate the passage. First a system prompt is assigned in the message history to the model, and then a user prompt is sent. The system prompt defines the models main purpose in all future messages and outlines any rules that the model should follow. For passage generation we use the following system and above user prompt to conduct passage generation:

System Prompt:

Play the role of a passage generator for children's reading
 ↪ comprehension books. You will be given vocabulary words,
 ↪ grammar rules, and a topic to incorporate into the story.
 ↪ Return only the story as your response without a title.
 ↪ You should not generate any stories about anything
 ↪ inappropriate for an elementary school student. If an
 ↪ inappropriate topic is requested, choose a random, more
 ↪ appropriate topic to write on.

The system prompt provides a simple method of moderation by instructing the LLM model to avoid inappropriate topics. Originally we defined "inappropriate topics" to the model as things "illegal, sensitive, or dangerous." Through piloting we found our LLM model was able to self-moderate topics without the need for specific examples. When removing this definition of "inappropriate" from the system prompt, the LLM generated more diverse stories with acceptable moderation. This moderation system may need future refinement as more sensitive topics are able to be tested. The pseudocode for this process is:

1. Pass the user prompt created in the previous step with the
 ↪ system prompt in the message history.
2. Await the AI model's response.
3. If a passage is generated successfully, check if the
 ↪ passage contains the required vocabulary and grammar.
 (b) Otherwise, try passage generation again up to three
 ↪ times.
 (c) If an error persists, return the error.
4. Return the generated passage.

Explanation: This pseudocode shows how the prompt is passed to the model, the response is processed, and the output is either displayed or handled as an error. If

an error is encountered we retry passage generation up to three times before quitting and displaying the error message.

Example

Here is an example of a prompt used for generating a passage:

Write a story for third graders to read about something fun

- and exciting. It should try to include most of the
- vocabulary words in this list: 'wrong, engine' and at
- least one example from this list of rules: 'Orthographic
- Patterns, Sentence Combining and Compound Sentences'. The
- story should be about 300 - 400 words long.

Ensure that the generated story has a main character that

- learns a lesson or moral. You may retell stories,
- including fables and folktales from diverse cultures, with
- a central message, lesson, or moral.

The first line is generated through our prompt creation in Section 3.3.2, and the second line is the concatenated objective prompt. The passage that was generated from this prompt is:

In a small town called Maplewood, there lived a curious boy

- named Timmy. Timmy loved to explore and learn about
- everything around him. One sunny Saturday, he decided to
- visit the local science fair that was held in the park. He
- had heard that there would be amazing projects, and he was
- eager to see them.

When Timmy arrived, he saw colorful tents and excited children
→ running around. There was a giant robot powered by a
→ special engine, a volcano that erupted with bubbly lava,
→ and even a booth dedicated to rocks from outer space!
→ Timmy was amazed by how creative everyone was.

As he walked around, he met Emily, a girl who was building a
→ small rocket. "Hi, Timmy! I'm trying to make the best
→ rocket to launch into the sky!" she exclaimed. Timmy
→ thought her idea was fantastic. He offered to help her,
→ but they quickly ran into a problem. They needed a special
→ piece that Emily didn't have.

"I guess it's wrong to think we can build a rocket without all
→ the pieces," Emily said, looking disappointed.

Timmy felt sad for Emily, but then he remembered something.
→ "What if we combine our ideas? I have some old toy parts
→ at home. We could use those!"

Emily's eyes lit up. "That's a great idea! Let's see what we
→ can find!" They rushed home to gather supplies. Timmy
→ brought a few toy cars, and Emily had her rocket parts.
→ Together, they spent the afternoon building, laughing, and
→ sharing ideas.

Finally, their rocket was ready! They painted it bright colors
→ and added stickers. It looked amazing! They set it up in
→ the park, and with a little help from a mentor at the
→ fair, they prepared for the launch. They counted down,
→ "Three, two, one... blast off!" The rocket soared high
→ into the air, leaving behind a trail of glittery smoke.

When it came back down, they were the talk of the fair! Timmy
→ and Emily learned an important lesson that day: things
→ might go wrong, but by working together, they can create
→ something amazing. They celebrated their success with
→ smiles, knowing that teamwork makes dreams work. And from
→ then on, they were the best of friends, ready for more
→ adventures!

This passage contains 371 words, which matches the goal of a 3-5 minute reading time for third graders as per Figure 3.7. The story contains a moral as instructed by the learning objective section of the prompt and the vocabulary words "engine" and "wrong." The passage demonstrates orthographic patterns through proper capitalization of names like "Timmy" and "Emily" and using punctuation such as commas, quotation marks for dialogue, and exclamation marks for excitement. It also features compound sentences, such as "He had heard that there would be amazing projects, and he was eager to see them," where two independent clauses are joined by the conjunction 'and'. This passage meets the length and complexity criteria for third graders and aligns with the educational goal of providing a moral by promoting cooperation and critical thinking through storytelling as instructed.

3.3.3 Purpose Generation

Once a passage is generated and read, Read with Emora generates a passage-purpose which explains how the passage content relates to the assigned learning objective. This explanation, generated by Purpose AI, is not displayed until after the "Finished Reading" button is pressed so as not to distract students during the reading process. This explanation aims to inform parents or educators about how the passage supports the student's reading comprehension skills through a specific learning objective and allows them to further test the assigned objective.

System and User Prompts

To generate this explanation, Read with Emora uses the Purpose AI model with prompts designed to connect the passage's content to the learning objective. The goal is to produce a concise paragraph that outlines how the passage's essential details align with the assigned learning objective and its overall purpose.

System Prompt:

```
You are generating explanations of how reading comprehension
→ passages for children connect to learning objectives. You
→ will be provided with a passage, general purpose, and
→ learning objective. You must use these to formally explain
→ in a short paragraph how reading and analyzing this
→ specific passage helps develop the given learning
→ objective for improving reading comprehension in young
→ kids. You should base your response on the given purpose
→ and use details or elements from the passage to add
→ specifics.
```

User Prompt:

```
Given
Passage: ${passage}
General Purpose: ${purpose}
Learning Objective: ${name}
```

```
Generate a few sentences on how this passage's key details
→ connect to the overall purpose and learning objective.
```

The system prompt directs the Purpose AI model to generate a formal explanation that uses details from the passage to demonstrate how the passage supports the

assigned learning objective. The output is typically a short paragraph that connects the critical information to the broader educational goals.

Pseudocode for Purpose Generation

1. The system selects the appropriate learning objective for
→ passage generation.
2. After the passage is generated, the objective-specific
→ purpose is pulled.
3. The passage, purpose, and learning objective are sent to
→ the Purpose AI model as a user prompt.
4. The GPT model generates a paragraph that:
 - (a) Explains how the passage supports the assigned
→ learning objective.
 - (b) Highlights critical details from the passage that
→ relate to the purpose.
5. The generated explanation is stored and later displayed
→ alongside the reading comprehension results for the parent
→ or educator to review.
6. If an error occurs, the system returns an error message.

Explanation: After generation, the Purpose AI model produces a formal explanation, displayed at the same time as the questions.

Example

Here is an example of the purpose generation process based on the passage about Timmy from Section 3.3.2. The Purpose AI input would appear as follows:

Input:

Given

Passage: "In a small town called Maplewood, there lived a
→ curious boy named Timmy. Timmy loved to explore and learn
→ about everything..."

General Purpose: "This learning objective aims to introduce
→ morals and lessons into stories for readers. Some of these
→ passages may be new stories based on familiar morals or
→ retellings of older, more well-known fables and tales.
→ Retelling familiar stories helps students internalize the
→ narrative, improve their memory, and enhance their
→ understanding of the story structure and themes."

Learning Objective: "Retelling Stories"

Generate a few sentences on how this passage's key details
→ connect to the overall purpose and learning objective.

Output:

The passage about Timmy's adventure at the science fair
→ effectively connects to the learning objective of
→ retelling stories by exemplifying the classic moral of
→ teamwork and problem-solving. By focusing on Timmy and
→ Emily's journey from encountering obstacles to
→ collaboratively building a rocket, the narrative
→ underscores the importance of working together to overcome
→ challenges.

This aligns with the overall purpose of introducing morals
→ into stories, as young readers can internalize the lesson
→ that collaboration can lead to success. The vivid details
→ of their interactions and the moment of triumph during the
→ rocket launch provide a structure that helps students
→ grasp the narrative, enhancing their memory and
→ understanding of the moral theme that teamwork makes
→ dreams work.

In this example, the generated explanation emphasizes how the passage about Timmy and Emily connects to the learning objective titled "Retelling Stories", which focuses on generating new stories that use morals and lessons from classic tales and fables. The purpose generation shows this through specific passage examples and details.

3.3.4 Important Sentence Identification

After a passage is generated, users select the "Finished Reading" button, which triggers the question-generation process, starting with important sentence identification. We adopt this process as outlined initially by He et al. [23] for AI-generated context-based comprehension questions on reading material for children and students. We simplify their techniques into two steps: important sentence identification and question generation. In important sentence identification, the goal is to identify sentences in the passage that are critical to the narrative or subject matter and can be used as the basis for reading comprehension questions.

System and User Prompts

To identify important sentences, we employ a GPT model (Sentence AI) with system and user prompts that emphasize both the contextual significance of sentences and

their relevance to the vocabulary introduced in the passage. The prompts are as follows:

System Prompt:

You are to act as an important sentence identifier. You will
 → be provided with a reading passage (which may be a story,
 → poem, article, or biography) written for children and an
 → array of vocab words. You should identify and return only
 → the most contextually important sentences from the passage
 → to the overall story (please try to find at least 4 but as
 → many as possible). These sentences or lines will be turned
 → into reading comprehension questions so they must contain
 → enough material to be contextually relevant. Please ensure
 → at least one sentence focuses on one or more of the
 → vocabulary words included in the list.

User Prompt:

Passage: $\{\text{passage}\}$

Vocab: $\{\text{vocab}\}$

The system prompt instructs the GPT model to analyze the passage for meaningful sentences, ensuring the output is relevant for question generation. It also emphasizes that at least one of the selected sentences must contain a vocabulary word from the provided list to increase further the context through which vocabulary is seen.

Pseudocode for Important Sentence Identification

1. The user clicks the "Finished Reading" button.
2. The system sends the passage and vocabulary words to the
 → Sentence AI model.

3. Sentence AI analyzes the passage, selecting meaningful
 - sentences based on contextual relevance to the passage's
 - overall story or message and the presence of vocabulary
 - words in at least one sentence.
4. The selected sentences are returned in a numbered list.
5. If an error occurs, we return it to the front end.
6. The sentences are stored for the next step in question
 - generation.

Explanation: This pseudocode outlines identifying essential sentences from the passage. After the user submits the passage, the system sends the passage and vocabulary words to Sentence AI, which returns a list of key sentences to understanding the passage. These sentences form the basis for the subsequent question-generation process.

Example

Here, we provide an example of the important sentence identification process performed on our example passage from Section 3.3.2. The Sentence AI input would appear as:

Input:

Passage: "In a small town called Maplewood, there lived a
→ curious boy named Timmy. Timmy loved to explore and learn
→ about everything..."
Vocab: ['engine', 'wrong']

Note that the passage is truncated here only for clarity. The actual user prompt always contains the entirety of the passage to give the Sentence AI the full context.

Output:

1. "One sunny Saturday, he decided to visit the local science
→ fair that was held in the park."
2. "There was a giant robot powered by a special engine, a
→ volcano that erupted with bubbly lava, and even a booth
→ dedicated to rocks from outer space!"
3. "As he walked around, he met Emily, a girl who was building
→ a small rocket."
4. "Timmy felt sad for Emily, but then he remembered
→ something."
5. "Together, they spent the afternoon building, laughing, and
→ sharing ideas."
6. "Timmy and Emily learned an important lesson that day:
→ things might go wrong, but by working together, they can
→ create something amazing."
7. "And from then on, they were the best of friends, ready for
→ more adventures!"

The important sentences identified here are key to the narrative of the original passage. Our LLM model does not pick up every single sentence that may define the story's narrative, instead focusing on standalone sentences with testable information. Sentence AI also successfully ensured that at least one of the sentences included a vocabulary word ("wrong" and "engine"). These sentences will be used to generate

context-based questions in the next phase of the system.

3.3.5 Question Generation

Once the important sentences from a passage are identified, the next step in our system is question generation. This process creates a set of reading comprehension questions based on the identified essential sentences through Question AI. The goal is to create simple, focused questions that help students reflect on the key aspects of the passage.

System and User Prompts

To generate reading comprehension questions, the system uses the Question AI model with prompts designed to create questions directly related to the important sentences from the passage. The prompts are crafted to ensure the inclusion of at least one question containing a vocabulary word to increase vocabulary contextual awareness.

System Prompt:

You are to act as a question generator for reading
→ comprehension passages for $\{\text{grade}\}$ grade children. You
→ will be provided with the passage, a numbered list of
→ important sentences from the passage, and a list of vocab
→ words. You are to return a numbered list of reading
→ comprehension questions based on the important sentences
→ of the same number. At least one question should include
→ or be centered around at least one vocabulary word and its
→ context in the story. If a sentence is not included with
→ one of these words, please create your own question based
→ on a vocabulary word from those provided. If there are any
→ important sentences that would not translate well into a
→ question, feel free to omit them. These questions should
→ be simple and should only ask about one thing.

User Prompt:

Passage: $\{\text{passage}\}$

Important Sentences: $\{\text{sentences}\}$

Vocab: $\{\text{vocab}\}$

Pseudocode for Question Generation

The process for question generation is as outline here:

1. After identifying important sentences, the system sends
→ them, along with the passage and vocab, to the Question AI
→ model.
2. The GPT model generates questions, ensuring that:
 - (a) Each question corresponds to an important sentence.

- (b) At least one question is vocabulary-based (if
→ possible).
- 3. The system receives the questions.
- 4. If no questions are generated, an error is returned.
- 5. The questions are stored for the next phase in the program.

Explanation: The pseudocode outlines the simple process of sending important sentences, the passage, and vocabulary to the Question AI model to generate reading comprehension questions. Each question corresponds to one of the important sentences, with the system ensuring that the questions are contextually relevant to the story.

Example

Here is an example of the question generation process based on the passage from Section 3.3.2 and important sentences from Section 3.3.4. The Question AI input would appear as follows:

Input:

Passage: "In a small town called Maplewood, there lived a
→ curious boy named Timmy. Timmy loved to explore and learn
→ about everything..."

Important Sentences:

1. "One sunny Saturday, he decided to visit the local science
→ fair that was held in the park."
2. "There was a giant robot powered by a special engine, a
→ volcano that erupted with bubbly lava, and even a booth
→ dedicated to rocks from outer space!"

3. "As he walked around...

Vocab: ['engine', 'wrong']

Note that the passage and sentences are truncated here only for clarity. The actual user prompt always contains the entirety of the passage and sentence list to provide the Question AI with full context.

Output:

1. 'Where did Timmy decide to go on the sunny Saturday?'
2. 'Who did Timmy meet while he was exploring the science
→ fair?'
3. 'What did Timmy feel when he saw Emily was disappointed?'
4. 'What did Timmy and Emily do together in the afternoon?'
5. 'What lesson did Timmy and Emily learn about teamwork?'

The generated questions focus on the critical moments in the story, following the important sentences. The Question AI is instructed to omit sentences to ensure well-contextualized questions. Though it is encouraged for Question AI to incorporate sentences with vocabulary words, the questions generated do not explicitly test vocabulary knowledge. The vocabulary words' presence in the questions helps to increase students' contextual awareness of the vocabulary. We discuss this more in Section 6.3.

3.3.6 Gold Standard Answer Generation

After students complete and submit their answers to the reading comprehension questions, the system generates a list of gold-standard answers. These answers serve as points of comparison for the students' responses and are displayed alongside the students' answers and scores after the answer submission and grading are complete.

Presenting these gold-standard answers allows students to review their work, understand the correct responses, and address any common mistakes they may have made.

System and User Prompts

To generate gold-standard answers, the system uses the Answer AI model with specific prompts designed to generate concise, accurate answers to the reading comprehension questions. These answers are based solely on the passage and are typically limited to one sentence for reviewers. The system prompt is:

System Prompt:

```
You are to act as a gold standard answer generator for reading
→ comprehension passages for children. You will be provided
→ with a passage and a list of questions on that passage.
→ Please return a numbered list of gold standard answers
→ corresponding to each question in the given list. Your
→ answers should be based on the passage information and
→ short, at most one sentence. Your return should be in this
→ format: 1. 'answer text here'\n2. 'answer text here'\n3.
→ 'answer text here', and so on.
```

where “\n” represents a newline character. And the user prompt is:

User Prompt:

```
Passage: ${passage}
Questions: ${questions}
```

The system prompt instructs Answer AI to provide concise answers to each question in a numbered list format. These answers are based directly on the passage’s content and serve as the gold standard for comparison.

Pseudocode for Gold Standard Answer Generation

1. After question generation, the system sends the passage and
→ questions to the Answer AI model.
2. The GPT model generates a numbered list of gold standard
→ answers corresponding to the question list order.
3. If the gold standard answers are successfully generated the
→ system stores these answers for comparison with student
→ answers.
4. The answers are displayed in the results section alongside
→ the student's answers after final answer submission and
→ grading.
5. If an error occurs, the system returns an error message.

Explanation: After the student submits their responses, the system sends the passage and questions to the Answer AI model, which returns a list of short, 'gold-standard' answers. These answers are stored and displayed in the results section for the student to review. Students can use these to understand where they missed passage information. These answers can also serve as a basis for understanding grading as our grading system is refined.

Example

Here is an example of the gold standard answer generation process based on the passage from Section 3.3.2 and questions from Section 3.3.5. The Answer AI input would appear as:

Input:

Passage: "In a small town called Maplewood, there lived a
→ curious boy named Timmy. Timmy loved to explore and learn
→ about everything..."

Questions:

1. 'Where did Timmy decide to go on the sunny Saturday?'
2. 'Who did Timmy meet while he was exploring the science
→ fair?'
3. 'What did Timmy feel when he saw Emily was disappointed?'
4. 'What did Timmy and Emily do together in the afternoon?'
5. 'What lesson did Timmy and Emily learn about teamwork?'

Output:

1. 'Timmy decided to go to the local science fair that was
→ held in the park.'
2. 'Timmy met Emily, a girl who was building a small rocket.'
3. 'Timmy felt sad for Emily when he saw she was
→ disappointed.'
4. 'Timmy and Emily spent the afternoon building, laughing,
→ and sharing ideas.'
5. 'Timmy and Emily learned that by working together, they can
→ create something amazing.'

The generated gold standard answers provide accurate, one-sentence responses based on the passage, giving students a reference for incorrect or partially incorrect answers. The gold standard answers are displayed on the Session Page alongside the student's responses on the results card after question answering.

3.3.7 Iterative Feedback and Grader AI

Grader AI assigns scores and provides iterative feedback to help students improve their responses. This feedback system encourages students to reflect on their mistakes and try again, fostering more holistic understanding and learning retention.

System and User Prompts

The Grader AI model generates a grade and provides a comment as feedback. The prompts provide the passage, a specific question, and the student's response, allowing the model to analyze the answer and generate feedback. The system prompt is designed to encourage leniency in grading, ensuring students receive partial credit for mostly correct but slightly unclear answers.

System Prompt:

```
You are to act as an answer grader for reading comprehension
→ passages for children. You will be provided with a
→ passage, a question on that passage, and a student answer.
→ Please grade the student response on a ternary scale (2
→ for correct, 1 for partially correct, and 0 for incorrect)
→ based on the passage information. You should give partial
→ credit if a student answer was almost right or partially
→ right, but still incorrect in some way. Be lenient with
→ full credit, if the student seems to understand the answer
→ but, say for example, is slightly unclear, use common
→ sense to infer if that counts for a full score, but be
→ generous.
```

If the student has incorrect grammar or makes a spelling
 → mistake, you should mention it in the comment, but not let
 → it affect their grade unless it's very bad or illegible.
 → If you scored the student less than 2, then the comment
 → should be encouraging and guide them to the correct answer
 → in the passage, but never directly mention the answer to
 → the question. Please keep comments short at 1-2 sentences
 → maximum. You will return a comment and a grade in this
 → format: Comment: 'comment text here'\nGrade: 2."

User Prompt:

Passage: \${passage}

Question: \${question}

Student Answer: \${answer}

This setup allows the system to assess the answer in the passage's context, provide supportive feedback, and assign a score.

Pseudocode for Comment and Grade Generation

The general pseudocode for this process is as follows:

1. The user submits an answer to the current question.
2. The system sends the passage, question, and answer to the
 → Grader AI model.
3. Grader AI analyzes the student's response and:
 - (a) Assigns a grade (2, 1, or 0) based on the accuracy of
 → the answer.
 - (b) Provides a comment if the grade is less than 2.
4. The comment and grade are stored in the system.

5. If the student has not reached the maximum number of
 - attempts and the score is less than 2:
 - (a) Display the AI feedback.
 - (b) Allow the student to resubmit the answer.
 - (c) Otherwise, move on to the next question
7. After all questions are answered, display the results card
 - for review.

Explanation: Students can resubmit their answers for up to three attempts. The system moves to the next question if a perfect score is achieved or the maximum attempts are reached.

Example

Here's an example of how Grader AI works:

Input:

Passage: "In a small town called Maplewood, there lived a
→ curious boy named Timmy. Timmy loved to explore and learn
→ about everything..."

Question: "Where did Timmy decide to go on the sunny
→ Saturday?"

Student Answer: "He went on a walk"

Output:

Comment: Timmy went somewhere specific, not just a walk, to
→ see something interesting on that sunny Saturday. Try
→ looking at the beginning of the passage for a clearer
→ answer.

Grade: 0

Here the grade provided applies to the student answer and the comment address the student answer's shortcomings. The student is awarded no credit in this example because the answer is incorrect. While Timmy could have walked to the science fair, the passage does not specify any mode of transportation. Grader AI provides an encouraging comment, allowing the student to reflect and resubmit a more accurate answer. Let's see how the comment and grade change after updating our answer to something partially correct:

Input:

Passage: "In a small town called Maplewood, there lived a
→ curious boy named Timmy. Timmy loved to explore and learn
→ about everything..."

Question: "Where did Timmy decide to go on the sunny
→ Saturday?"

Student Answer: "He went to the park"

Output:

Comment: You're close! Remember to mention the event that
→ Timmy went to see at the park. Keep improving!

Grade: 1

In this example, Grader AI can recognize the science fair as a higher priority detail than the park and grades the students accordingly. Their score improves as there is more correct information in the student's answer than previously, but full credit is not awarded due to the exclusion of essential details. When we update our response on the third attempt to "He went to the science fair at the park", Grader AI returns a score of 2 and moves on to the next question.

3.3.8 Scoring and Ease Factor Adjustments

Once all questions have been answered and graded, the system adjusts the ease factor of the learning objective associated with the passage. The ease factor represents how well the student has mastered the learning objective, and it is updated based on the student's average score and number of attempts.

First, we calculate the student's average score and average attempts. We updated the score based on the attempts using Equation 3.1. Next, the ease factor is updated using Equation 3.2, which uses the adjusted score to account for both student performance and assistance needed to reach that performance. If the student performs poorly, the ease factor decreases, ensuring the objective is tested sooner than others next time (which results in a different passage based off of different vocab, grammar, and a topic with the same objective prompt). If the student did well, the ease factor increases, decreasing the priority of testing that objective next time. Once the ease factor is recalculated, the system updates the database with the new value, ensuring the adaptive nature of the program is maintained.

3.4 Interface

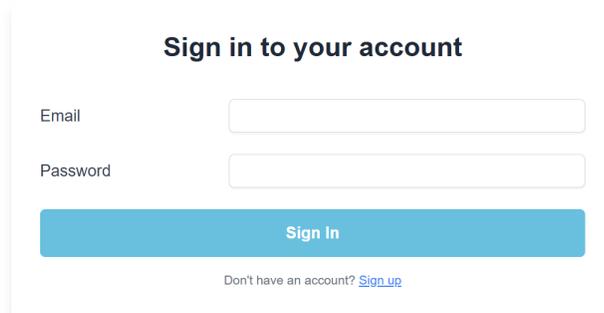
Our website interface is streamlined with a light blue color scheme and the Poppins font to ensure child-friendliness and ease of navigation. Each page features large, clearly labeled buttons, making the site intuitive for young users. Navigation is done through a navigation bar at the top of the website with tabs to all pages. We limited the number of tabs in the navigation bar to keep the focus on the logic and intelligence behind the project and to reduce potential confusion in site navigation for younger users.

3.4.1 Page Layout and Navigation

Our website has five distinct pages designed with simplicity and usability in mind. The Read with Emora link routes to the Sign In page, allowing users to sign in or navigate to the Sign Up page if they need to create an account. After logging in, users can access three main pages: Session, Progress, and Profile.

Sign In/Sign Up

Our sign-in page (Figure 3.4) allows users to enter their email and password to access the website. If they are not yet registered, the 'Sign Up' button directs them to the Sign Up page (Figure 3.5), where they can create a new account. After signing up, a user's account information is stored in our database. The design of these pages is minimalistic to ensure ease of use. The form fields are clearly labeled, and error messages are shown immediately if the input is invalid or incomplete.



The image shows a sign-in form with the following elements:

- Title: **Sign in to your account**
- Input field: Email
- Input field: Password
- Button: Sign In (blue)
- Link: Don't have an account? [Sign up](#)

Figure 3.4: Read with Emora Sign-In Page

Initially, we implemented email verification for account creation. However, due to

Create a new account

First Name Last Name

Email Grade

Password Password Confirmation

By creating an account, you agree to our [terms and conditions](#) and [privacy policy](#)

Already have an account? [Log in](#)

Figure 3.5: Read with Emora Sign-Up Page

recurring issues with users not receiving the verification emails, we decided to remove this step, as it is not a necessary security measure at this project stage. This decision improved the user experience and eliminated unnecessary friction for parents accessing the platform.

Session Page

The Session Page’s user interface guides students through the passage generation, question-answering, and feedback process. The interface adapts dynamically depending on the student’s interaction, allowing for step-by-step learning, by displaying the passage card on session instantiation, displaying the objective purpose and vocabulary and grammar cards after passage reading, displaying or removing corrective feedback by student question performance, and displaying the results card on session completion.

Initially, the user is shown a clean interface with an input box to specify a topic and a “Generate Passage” button. This is displayed in Figure 3.6. If no topic is

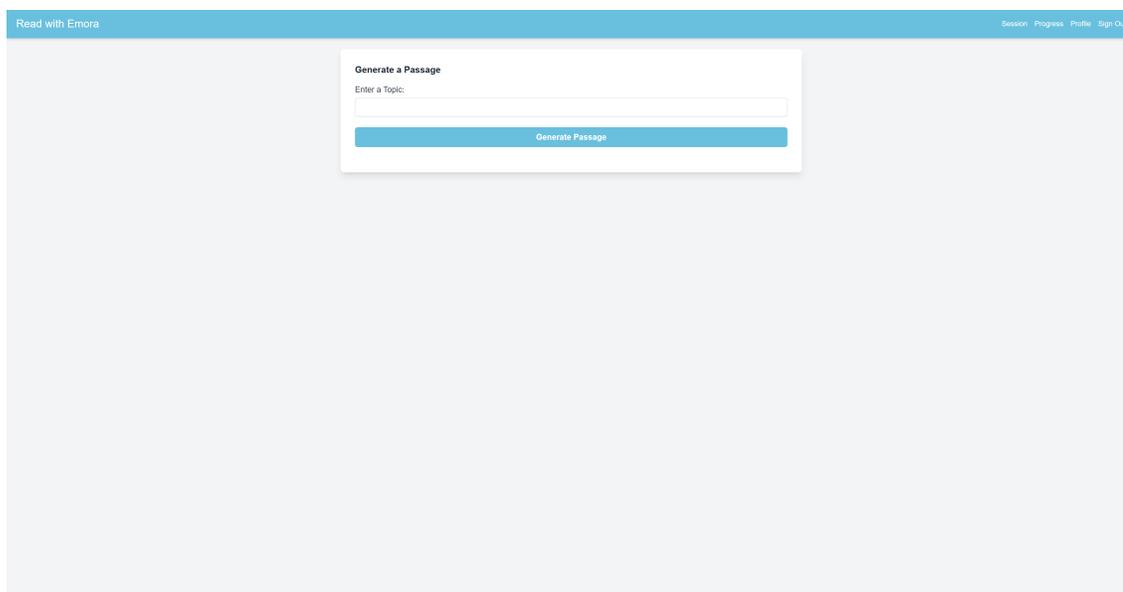


Figure 3.6: The Session Page before passage generation.

provided, the system defaults to generating a passage on a random, grade-appropriate topic. No passage or question cards are visible at this point.

Once the "Generate Passage" button is pressed, the system triggers the passage generation process using the student's input or default topic. After that, the passage appears on the page. The student can read through the passage, and once they are ready, they can press the "Finished Reading" button to display and begin answering questions.

After the question generation process, the page displays the first question card, the learning objective information, and vocabulary and grammar information, as shown in Figure 3.7.

In this state, the interface presents questions individually, starting with the first question related to the passage. Each question is in a card format, with an input field for students to submit their answer. Once the student submits their answer, the system evaluates it through Grader AI and provides feedback, displayed between the open text box and the "Submit Answer" button. The system then advances to the next question if the answer is perfect or if the attempts are exhausted.

Generate Passage

Passage

In a small town filled with visual delight, Lived a young girl named Lily, full of light. With her sketchbook in hand and colors so bright, She dreamed of adventures from morning till night. One day, while walking beneath the ambient trees, She heard a loud noise that danced with the breeze. "Wow! What's that?" she exclaimed with a cheer, A magical creature had suddenly appeared! It was a quirky old dragon, bright with three tails, "I'm looking for friends! Let's set sail on the gales!" But alas! The dragon's map was inaccurate and torn, "Without it," the dragon sighed, "I feel so forlorn." "Don't worry!" said Lily, her courage ablaze, "We'll solve this together, we'll find our way!" With adjectives flying, she described the scene, "Look at the river, so shiny and clean!" Through valleys and hills, they started to roam, Lily knew in her heart that with friends they'd find home. They danced on the clouds and floated on air, "I'll help you, dear dragon, I promise!" she declared. With each little step, they paved a new road, Finding clues in the nature, lightening their load. "Maybe we can follow the sound of the stream, And find our way back to your beautiful dream." "At each twist and turn," said the dragon with glee, "We'll see all the wonders that nature can be!" The hours turned to days as they traveled afar, Lily and the dragon, under the same star. "Look, over yonder," said the dragon one night, "There's a village so cozy, with twinkling light!" "Let's ask for some hints," Lily gently suggested, As they arrived at the place, both feeling quite rested. The villagers, friendly, gathered around, "Have you seen this map?" asked the dragon astound. With a positive smile, they pointed to the hill, "Just climb to the top, and you'll find what you will!" So up they both climbed, with hearts full of glee, Determined and brave, their spirits flew free. And then from the top of the hill with bright hues, They spotted the treasure – three colorful clues! With joy in their hearts, they followed the signs, Beneath the soft sky, through bushes and vines. Eventually, the dragon could finally see, The path to his home, where he could be free. "Thank you, dear Lily, for being so kind, In every adventure, true friends you will find!" And as the sun set in a blaze of bright gold, Lily smiled wide, her heart feeling bold. "In this great big world, so thrilling and wide, There's magic around us, with friends by our side. So here's a reminder, in times tough and long, That together we conquer, together we're strong!"

Question 1 of 5

What did Lily have in her hand while she dreamed of adventures?

Submit Answer

Objective: Understanding Key Details

The passage centers around Lily, a young girl who embarks on an adventure with a quirky dragon, providing students with a vivid exploration of essential story elements such as characters, settings, and major events. As they read, students can identify key details, like Lily's courage in resolving the dragon's challenge of an inaccurate map and the supportive interactions with the villagers, which highlight character development and propel the plot forward. By analyzing these details, students enhance their ability to discern the connections between characters and their challenges, thereby deepening their understanding of key story elements and developing their skills in reading comprehension.

Vocabulary Words Used:

- visual
- ambient
- inaccurate

Grammar Syntax Included:

- Interjection / Prepositions
- Verb Tenses

Figure 3.7: The Session Page After Passage and Question Generation.

The bottom half of the page also displays two important cards: the objective card and the vocabulary and grammar card. The objective card shows the specific learning objective for the passage along with the generated passage-purpose, while the vocabulary and grammar card highlights key vocabulary words and grammar rules featured in the passage.

Once the student has answered all the questions, the question card transitions to the results card, as seen in Figure 3.8, this section provides a comprehensive overview of the student's performance and is described further in section 3.2.1.

The interface is designed to facilitate student study sessions with little teacher maintenance or monitoring. Then after session completion students and teachers can review performance immediately, making it easier to identify areas for improvement. Additionally, the learning objectives and vocabulary and grammar cards remain visible in the results view so educators may expand on student sessions.

Progress Page

The Progress Page (Figure 3.9) provides a comprehensive overview of the user's performance across reading comprehension sessions, helping students, teachers, and parents track progress over time. The layout is designed to be user-friendly, featuring a combination of charts, tables, and informational tips that guide the user in interpreting the displayed data.

At the top of the page, three doughnut charts provide visual summaries of calculated percentages:

- **Average Assistance Needed:** This chart represents the average number of attempts a student requires to answer questions across all passages, shown as a percentage. A higher percentage indicates more assistance was needed (i.e., multiple attempts were made), while a lower percentage reflects a student's ability to answer correctly with fewer attempts.

Generate Passage

Passage

In a small town filled with visual delight, Lived a young girl named Lily, full of light. With her sketchbook in hand and colors so bright, She dreamed of adventures from morning till night. One day, while walking beneath the ambient trees, She heard a loud noise that danced with the breeze. "Wow! What's that?" she exclaimed with a cheer, A magical creature had suddenly appeared! It was a quirky old dragon, bright with three tails, "I'm looking for friends! Let's set sail on the gales!" But alas! The dragon's map was inaccurate and torn, "Without it," the dragon sighed, "I feel so forlorn." "Don't worry!" said Lily, her courage ablaze, "We'll solve this together, we'll find our way!" With adjectives flying, she described the scene, "Look at the river, so shiny and clean!" Through valleys and hills, they started to roam, Lily knew in her heart that with friends they'd find home. They danced on the clouds and floated on air, "I'll help you, dear dragon, I promise!" she declared. With each little step, they paved a new road, Finding clues in the nature, lightening their load. "Maybe we can follow the sound of the stream, And find our way back to your beautiful dream." "At each twist and turn," said the dragon with glee, "We'll see all the wonders that nature can be!" The hours turned to days as they traveled afar, Lily and the dragon, under the same star. "Look, over yonder," said the dragon one night, "There's a village so cozy, with twinkling light!" "Let's ask for some hints," Lily gently suggested, As they arrived at the place, both feeling quite rested. The villagers, friendly, gathered around, "Have you seen this map?" asked the dragon astound. With a positive smile, they pointed to the hill, "Just climb to the top, and you'll find what you will!" So up they both climbed, with hearts full of glee, Determined and brave, their spirits flew free. And then from the top of the hill with bright hues, They spotted the treasure – three colorful clues! With joy in their hearts, they followed the signs, Beneath the soft sky, through bushes and vines. Eventually, the dragon could finally see, The path to his home, where he could be free. "Thank you, dear Lily, for being so kind, In every adventure, true friends you will find!" And as the sun set in a blaze of bright gold, Lily smiled wide, her heart feeling bold. "In this great big world, so thrilling and wide, There's magic around us, with friends by our side. So here's a reminder, in times tough and long, That together we conquer, together we're strong!"

Your Results

- 1. What did Lily have in her hand while she dreamed of adventures?**
Your Answer: her sketchbook
Gold Standard Answer: Lily had a sketchbook in her hand while she dreamed of adventures.
Score: 2
Attempts: 0
- 2. How did the dragon feel about his map being inaccurate and torn?**
Your Answer: He sighed and felt forlorn
Gold Standard Answer: The dragon felt forlorn about his map being inaccurate and torn.
Score: 2
Attempts: 0
- 3. What did Lily promise the dragon they would do together?**
Your Answer: That she would help him
Gold Standard Answer: Lily promised the dragon they would solve the map issue together.
Score: 1
Attempts: 2
- 4. What did the dragon say they would see at each twist and turn during their adventure?**
Your Answer: all the wonders that nature can be
Gold Standard Answer: The dragon said they would see all the wonders that nature can be at each twist and turn during their adventure.
Score: 2
Attempts: 2
- 5. What important thing did the dragon find at the end of their journey?**
Your Answer: I'm not sure
Gold Standard Answer: The important thing the dragon found at the end of their journey was the path to his home.
Score: 0
Attempts: 2

Want more practice? Generate a new passage up top!

Objective: Understanding Key Details

The passage centers around Lily, a young girl who embarks on an adventure with a quirky dragon, providing students with a vivid exploration of essential story elements such as characters, settings, and major events. As they read, students can identify key details, like Lily's courage in resolving the dragon's challenge of an inaccurate map and the supportive interactions with the villagers, which highlight character development and propel the plot forward. By analyzing these details, students enhance their ability to discern the connections between characters and their challenges, thereby deepening their understanding of key story elements and developing their skills in reading comprehension.

Vocabulary Words Used:

- visual
- ambient
- inaccurate

Grammar Syntax Included:

- Interjection / Prepositions
- Verb Tenses

Figure 3.8: The Session Page After Finishing a Session: The results card represents the student responses to the questions.

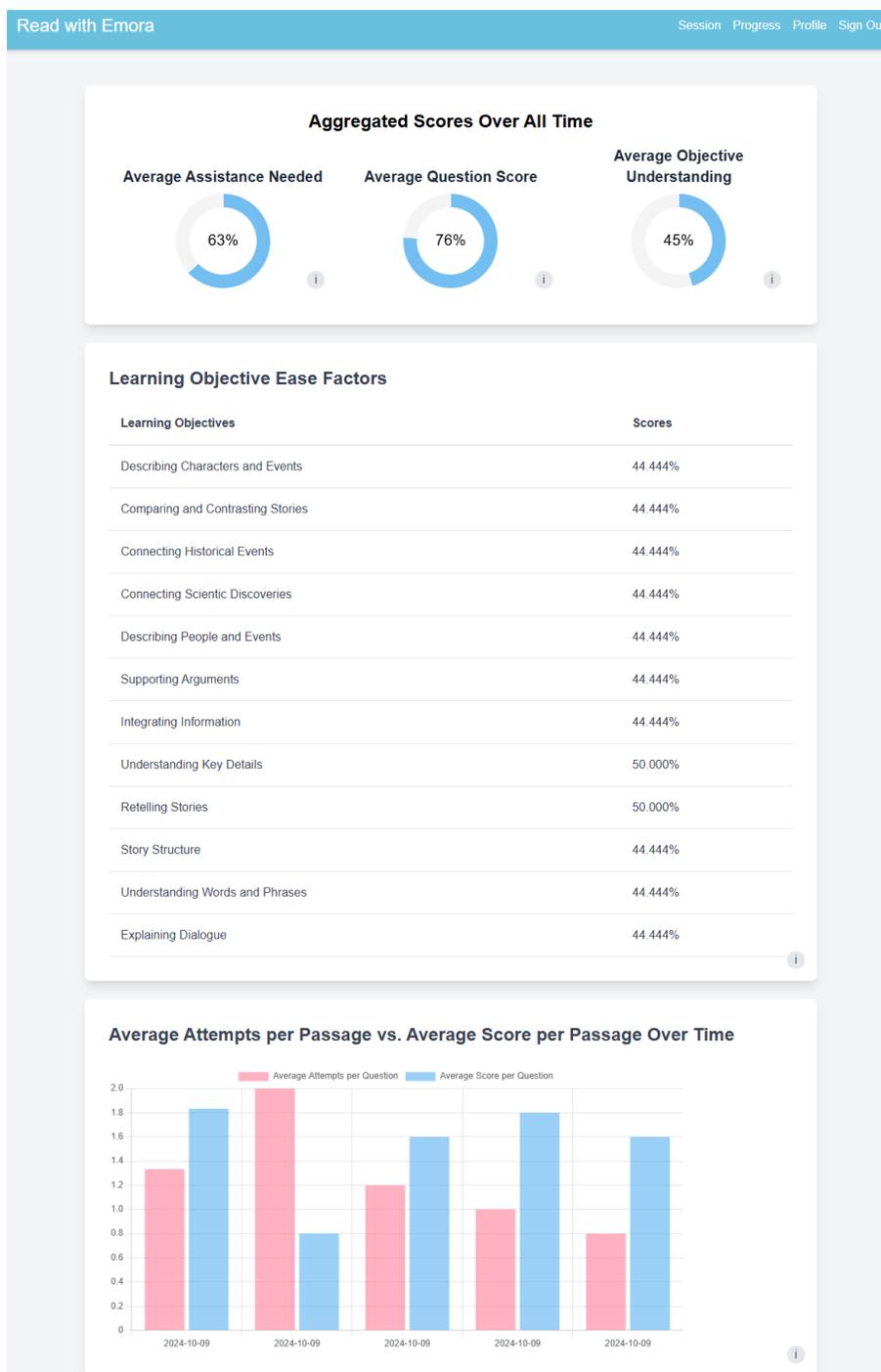


Figure 3.9: The Progress Page After Completing Five or More Passages.

- **Average Question Score:** The second doughnut chart illustrates a student's average score per question, also displayed as a percentage. This score is based on the system's ternary grading system, where each answer is scored 0, 1, or 2. This percentage should increase as students improve, reflecting better comprehension and more correct answers on the first attempt.
- **Average Objective Understanding:** The final doughnut chart shows the mastery of learning objectives, averaged from the ease factors assigned to each objective. A higher percentage reflects a better understanding and mastery of these objectives over time.

Directly below these charts, helpful informational tips are included to assist users in interpreting the data. By hovering over the info icons next to each chart, users can access explanations that clarify what the percentages represent and how they should expect these values to change as the student improves. For example, the tooltip for the "Average Assistance Needed" chart explains how more attempts correspond to a higher percentage, guiding users on how to interpret the chart in relation to student improvement.

Further down the page, a table labeled "Learning Objective Ease Factors" lists the learning objectives the student has been working on alongside their ease factor (as a percentage). These ease factors reflect how easily the student has mastered the material. The data here is dynamically updated to show the latest scores for each objective, allowing the student's progress in understanding each learning goal to be tracked over time.

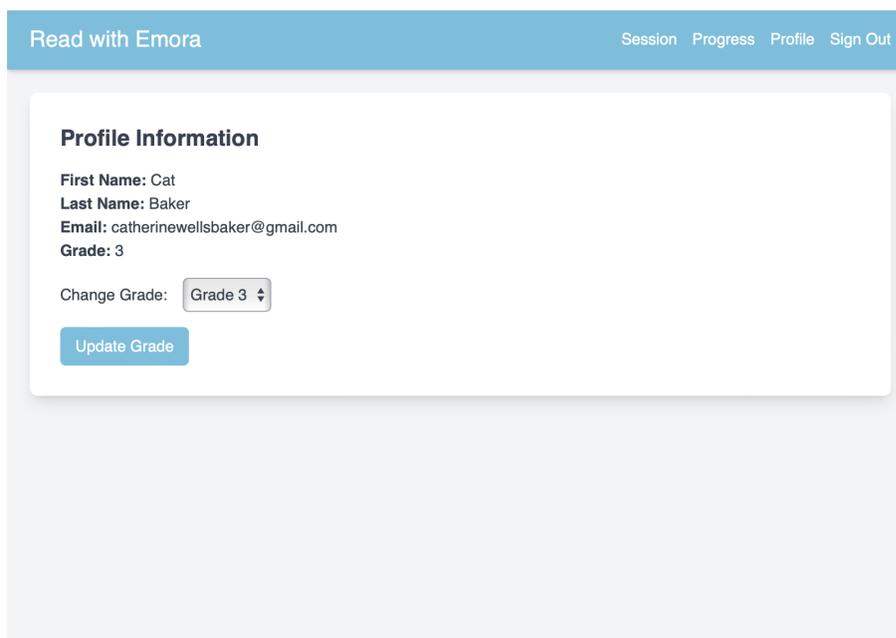
A bar chart at the bottom of the page compares the average number of attempts and average scores for recent passages. Each pair of bars represents one passage, showing the average attempts made and the average score achieved across all questions for the last five passages for a user. Ideally, over time, this chart will show a decrease in attempts (as students answer more questions correctly on the first try) and an

increase in scores, indicating improved comprehension. The side-by-side nature of the bars helps users quickly compare performance between passages and spot trends in the student's progress.

Including info tips for the doughnut charts, objective list, and the bar chart ensures that users, regardless of technical expertise, can understand the significance of the displayed data. This design helps foster a supportive learning environment, providing clear, accessible insights into the student's progress. The charts and tables offer an intuitive way for parents and teachers to monitor performance and guide future learning.

Profile Page

The Profile Page provides students with an interface to view their basic information and update their grade level. The profile page is essential for personalized learning as it adjusts content based on the user's grade, ensuring appropriate difficulty levels in passages, questions, and feedback.



The screenshot shows the 'Profile Page' in the 'Read with Emora' application. The page has a blue header with the text 'Read with Emora' on the left and navigation links 'Session', 'Progress', 'Profile', and 'Sign Out' on the right. The main content area is white and contains the following information:

- Profile Information**
- First Name:** Cat
- Last Name:** Baker
- Email:** catherinewellsbaker@gmail.com
- Grade:** 3
- Change Grade:** Grade 3 (with a dropdown arrow)
- Update Grade** (button)

Figure 3.10: The Profile Page After Signing In.

As seen in Figure 3.10, the current Profile Page displays personal user information: the user's first name, last name, email, and current grade. Users can update their grade level using a dropdown menu and 'Update Grade' button. This is particularly useful for students progressing through grades or when adjustments are needed for educational content. This change is reflected across the platform, ensuring that future passages and questions are tailored to the selected grade. As mentioned in Section 6.3, we hope to see the Profile Page eventually be expanded to include additional customization features, such as the ability to upload or change personal vocabulary words and grammar rules.

3.4.2 Data Communication and Storage

For this project, we use a Supabase database to manage and track data due to its intuitive user interface and generous free tier, which allows up to 500GB of storage. During the development phase, this solution provides easy access and seamless integration with our system for handling the necessary data.

The database primarily tracks the ease factors associated with learning objectives and stores user-specific information. On the Progress page, we display the updated ease factors of the learning objectives, alongside additional data pulled from the passages table. This table stores relevant information after each session, including the questions asked, the student's responses, and the grades awarded. This stored passage information is visualized on the Progress page.

Chapter 4

Evaluation

This chapter evaluates Read with Emora based on two core research questions (RQ1 and RQ2 introduced in Section 1.3) through passage generation and question analysis. The experiments assess Read with Emora’s effectiveness in generating educational reading materials tailored to specific learning objectives, generating context-based questions that test students’ material understanding, and grading student responses accurately.

4.1 Experimental Setup

Most of our research into existing ITS and LLM programs used user feedback [12, 45, 34, 57] or expert generated responses and grading He et al. [23] to evaluate their systems. RC-ITSs, specifically, often use question answering accuracy [38, 47, 32] or cognitive ability testing through tests like Woodcock-Johnson III Tests [36, 47] to analyze how their programs improve reading proficiency in students. This project assesses the impact of our contribution by evaluating our llm-based content generation. Automatic content generation is not unique to reading comprehension or ITSs and many systems have been created to facilitate this process [34, 23, 57, 52, 31]. In our project, we evaluate 2 facets of our content generation: passage generation and

question generation.

One system by Wang and Kreminski [57] evaluates their LLM-generated passages through a user feedback survey assessing the passage over a 7-point scale in different categories. Wang and Kreminski [57] then observe this data and draw conclusions off of averaged scores. Our approach adopts analagous techniques by conducting a rated review on a 3-point scale of learning component integration in individual passages. Sun et al. [52] performs an analysis of content integration in generated text by assigning a binary 0/1 success or fail score to integrated content like topics or keywords. Additionally, they analyze story generation quality through measures of repetition, diversity, and coherence across various passages [52]. Similarly, Ma et al. [31] evaluated content quality and integration in story premises using GPT-4-turbo, instead of human evaluators, to assess fascination, completeness, and originality. Their system assigned a completeness score based on how well each story premise integrated essential narrative components [31]. Other systems like UltraEval by [22] replicate this by using GPT-4 to simulate human evaluation by assigning a numerical score to passages over metrics like coherence, completeness, and relevance. Some additional papers such as Wang et al. [58], Prabhumoye et al. [43] evaluate content integration and quality of generated texts through empirical metrics like Rouge-L F1 scores, test-BLEU, and perplexity or through comparison with human annotations. These tests provide great insight to systems using or comparing various models' performance, however automated evaluations for generation systems often fall short of capturing the nuances of a model's capacity for generation or comprehension [27, 61, 51], which is likely why so many text generation systems opt for unique or score-based evaluations instead. We build on these approaches, focusing on content integration specifically, by assigning integration scores for various incorporated features on a ternary scale to account for partial integration.

In one question generation evaluation by Mazidi and Nielsen [34] workers were

asked to rate automatically-generated questions on a 1 to 3 scale for grammaticality and clarity and then again for how well the question helps them to understand the meaning of a sentence. The authors then analyze the mean agreement scores over all workers and questions and provide the Pearson Correlation Coefficient to analyze statistical significance [34]. Question generation systems like Xu et al. [61], Kalbaliyev and Sirts [27] or ?] alternately analyze the impact of their question generation by comparing machine-generated questions to the question-answer pairs from the training dataset using empirical analysis like Rouge-L F1 scores. Further, Xu et al. [61] and ?] analyze generated question variety by recording the distribution of who, what, when, where, why, and how questions. Other work like that by Kalbaliyev and Sirts [27] uses similar techniques by categorizing their questions by the skill being tested instead of the leading word. However, studies such as Chan et al. [10] rely on expert human evaluators by analyzing the accuracy of their answer choices across questions rather than dataset comparison or skill categorization. Other question generation systems, like that proposed by Jouault and Seta [25] or Kumar et al. [29], contain little to no evaluation of their question generation. These systems focus more on their program’s development and structure and rely on example outputs from their program to provide evidence to its performance and abilities [25, 29]. Most papers presenting generated questions from language models also recognize the limitations in evaluating generation systems as common and automatic evaluation measures often fail to accurately represent a model’s comprehension or generation abilities [27, 61, 51].

We reviewed evaluation methods in Automatic Text Scoring (ATS) for short answer questions and essays. These systems often have varied evaluation methods given the variety of approaches taken to this task. Some systems grade using numerical methods [46, 64, 2] while others grade in a preference-ranking system[44]. Schneider et al. [46] and Zeng et al. [64] perform analyses of their numerically-based grading methods on exam responses and short answers by having expert human educators model grade

the same assessments as a language model (a GPT LLM and a deep learning model, respectively). The authors measured discrepancies in human and LLM scoring and identify instances where the LLM grades diverged significantly from human grades [46]. Other studies adopt similar methods using pre-annotated data. These LLMs are trained on specific datasets of questions with expert-responses and then rely on those human expert annotations to guide the LLMs ability to grade similar responses [13, 44]. Chu et al. [13], specifically, took an advanced approach to LLM grading assessment by iteratively refining grading guidelines with additional LLM agents that collectively update LLM-assigned grades given human grading data. Many question grading LLM systems employ various empirical methods like answer accuracy, Cohen’s Kappa, Mean Reciprocal Rank (MRR), Normalized Discounted Cumulative Gain, the Jaccard coefficient, and Fleiss’ Kappa to measure alignment between human and model grading [13, 44, 2, 64]. The exact empirical tests used are varied and unique across papers. Because approaches to ATS are so varied, these systems require unique evaluation methods to compliment their exact scoring techniques. Our approach adopts similar techniques to these systems by implementing human and LLM evaluations of question and answer pairs, and comparing the discrepancies between those.

This thesis explores whether the system functions correctly as a proof-of-concept implementation. Our evaluations address the question of how and if LLMs work in reading comprehension-based applications, however, it does not fully address how well the LLM meets the these criteria or how it impacts reading comprehension proficiency. Additionally we do not rely on external datasets to train our models. Therefore, our evaluation methods rely on analysis of content integration, question variety, and answer grading.

4.1.1 Research Questions

Our experiments focus on the following two research questions:

- **RQ1:** Can an LLM generate an engaging story that integrates learning objectives, grammar rules, vocabulary words, and personalized topics from the user?
- **RQ2:** Given a passage, can an LLM generate a categorical variety of context-based questions and accurately grade their responses?

We choose to adopt evaluation methods most similar to Wang and Kreminski [57], Sun et al. [52], Mazidi and Nielsen [34], and Ma et al. [31] for LLM passage generation assessment by applying integration scores on a 0-2 scale (matching Mazidi and Nielsen [34]’s 1-3 scale and simplifying Chen et al. [12]’s 1-5 scale) for our intended integrated learning components over 75 generated passages. This will address research question one providing evidence that provided learning components can effectively be integrated into LLM-generated passages. To address research question two we conduct two evaluations across 402 LLM-generated questions. To assess question variety, as Xu et al. [61], Sugawara et al. [51], and Kalbaliyev and Sirts [27] did in their evaluations, we categorize questions into one of six general types and analyze the distribution of questions across grades. To assess LLM grading accuracy we conduct human grading on the same 402 context-based questions and compare results to LLM grading similar to Schneider et al. [46] and Zeng et al. [64].

4.1.2 Research Question 1: Evaluating Passage Generation

This experiment determines whether the LLM can successfully generate engaging stories that integrate four key components: learning objectives, grammar rules, vocabulary words, and personalized topics.

Methodology

- **Passage Generation:** We generate 15 passages per grade level (grades 1–5) based on varied topics. We ensure that each passage integrates the following

four key features: learning objectives, grammar rules, vocabulary words, and the user-specified topic.

- **Evaluation Criteria:** Each passage will be manually reviewed and assigned an integration score for each incorporated feature on a 0–2 scale as shown in Table 4.1.
- **Data Aggregation and Analysis:** We aggregate the scores for all passages per grade level. We calculate the average score for each feature across all passages and grade levels. We will use these calculations to determine which features (vocab, grammar, LO, or topic) tend to be most effectively integrated and which require improvement.

Score	Criteria	Description
0	Not Integrated	Feature is absent or irrelevant in the passage.
1	Partially Integrated	Feature is present but weakly integrated (e.g., vocabulary word mentioned only once and not further emphasized).
2	Well-Integrated	Feature is fully and meaningfully incorporated into the passage (e.g., vocab is relevant to context, LO is central to passage).

Table 4.1: Grading Scale for Passage Features

Discussion

This experiment evaluates how well the LLM balances multiple educational features within a passage. We expect Read with Emora’s Passage AI to excel with simpler components such as vocabulary integration and struggle with more abstract elements like learning objectives. We identify areas where prompt refinement may be necessary to enhance the relevance of specific features through averaged scores over all passages within each grade.

4.1.3 Research Question 2: Evaluating Question Generation and Grading Accuracy

This experiment evaluates Read with Emora’s ability to generate a variety of context-based questions from passages and its accuracy in grading student responses. We aim to determine whether the LLM can effectively guide students through comprehension tasks, providing effective questions and accurate grading.

Methodology

- **Question Generation:** For consistency, we generate 5–10 context-based questions for each passage used in RQ1. These questions are categorized by type (e.g., Factual Recall, Inference, Cause and Effect).
- **Evaluation Criteria:** Each question is categorized and evaluated using the same 0–2 grading scale as Grader AI. This categorization helps assess the variety and depth of questions generated by Question AI.
- **Grading Accuracy:** Each generated question requires a student answer. Grader AI and a human separately evaluated student responses on a 0–2 scale.
- **Data Aggregation and Analysis:** We compare the LLM’s grades with human grades to assess alignment and consistency. We analyze loss by taking the absolute value of the difference between the human grade and the AI grade. We then analyze this loss across question type and grade. Additionally, we analyze which question types (e.g., factual, inferential) the LLM grades most accurately and identify areas requiring further refinement.

Discussion

We evaluate Grader AI’s ability to accurately assess student responses. To ensure a fair evaluation, we test the model with various input types, including partial answers,

paraphrased responses, typographical errors, and synonyms. This allows us to assess the model’s precision and ability to handle real-world variations in student responses.

By incorporating both ideal and challenging inputs, we ensure a balanced assessment that highlights both the strengths and limitations of the LLM. The analysis also provides insights into how closely Grader AI aligns with human grading and where further improvements are needed.

4.1.4 Considerations for Experiment Design

Using the same passages for RQ1 and RQ2 ensures consistency in evaluating the LLM’s performance across different tasks. This approach simplifies the experimental design, however, we ensure that the passage content does not overly bias Question AI toward certain question types.

Additionally, to focus solely on the LLM’s grading performance, we limit our evaluation to the first answer submission from each question answerer. While the LLM provides corrective feedback, this feature is not evaluated here, as this experiment focuses on initial grading accuracy rather than iterative feedback.

Chapter 5

Analysis

This section presents the results for both research questions, analyzing the system’s performance by grade level and feature. For each research question, we summarize Read with Emora’s strengths and weaknesses, compare its performance across different categories (e.g., vocabulary vs. grammar, factual vs. inferential questions), and explore the impact of grade level on the overall results.

5.1 Results for RQ1: Passage Generation

The results of our passage analysis can be found in Table 5.1. Further break downs of our data can be found at the link to our google sheet in Chapter A.

Grade	Learning Objectives	Grammar	Vocabulary	Topic
Grade 1	1.667	2.000	1.733	1.733
Grade 2	1.533	2.000	1.600	2.000
Grade 3	1.733	1.800	1.667	1.933
Grade 4	1.733	2.000	1.667	1.933
Grade 5	1.800	1.733	1.467	2.000
Total Average	1.693	1.907	1.627	1.920

Table 5.1: Average Integration Score for Each Feature by Grade

5.1.1 Feature Integration by Grade

The scores for each feature were consistent across grades, with no clear trend of improvement or decline as the grade level increased. This suggests that the program was able to manage the increased complexity of higher-grade content—including greater vocabulary, length, and detail—without compromising the integration of key features such as learning objectives and grammar rules. This stability in performance across grades highlights the robustness of the passage generation algorithm.

The variation in scores across different features also provides insight into the program’s strengths and areas for improvement. Below, we explore the findings for each feature in detail.

5.1.2 Feature Integration by Feature

Learning Objectives

Learning objective integration, with an overall average score of 1.693, was among the lower-performing features. While many objectives were easily incorporated into fictional story contexts, such as ‘Describing people and Events’, or ‘Understanding Key Details’, more specialized objectives posed challenges. In particular, objectives like ‘Connecting Historical Events’ and ‘Connecting Scientific Discoveries’ were difficult to integrate effectively, as they require fact-based narratives, biographies, or other non-fictional content. The Passage AI model demonstrated a preference for generating fictional stories, which made it harder to incorporate these fact-based objectives seamlessly.

In performing well on fictional objectives and poorly on fact-based objectives, our findings suggest that Passage AI excels in creative storytelling but struggles when tasked with integrating factual information into narrative contexts. This limitation points to the need for more precise prompt engineering or future model adjustments

to better incorporate historical or scientific topics where needed.

Grammar

Grammar integration performed well overall, with an average score of 1.907. However, there was some variation across grade levels. Higher grades, which required more complex grammatical structures—such as correlative conjunctions and homographs—showed slightly lower grammar integration scores. In these cases, these advanced grammar rules were either underrepresented or used only sparingly, typically with only one or two examples.

In slight contrast, grammar rules for younger grades were consistently well-integrated. This discrepancy indicates that the LLM handles basic grammar competently but may need further fine-tuning to incorporate more complex grammar rules for higher-grade content.

Vocabulary

Vocabulary integration, with an average score of 1.627, presented challenges when the passages required the inclusion of more than five vocabulary words. In such cases, the vocabulary was often included only once or in a supplemental manner, rather than being central to the story. This pattern was particularly noticeable in Grade 5, which had the lowest vocabulary score (1.467) and the highest suggested range for pulling vocabulary words. While all required vocabulary words were mentioned, the focus of these passages was more centered around the topic than the vocabulary, suggesting a need for improvements in prompt design to ensure more seamless vocabulary integration.

Overall, the program handled vocabulary effectively when fewer words were required, with all words being mentioned at least once and often central to the story. However, as the number of required words increased, maintaining meaningful integration became

more challenging, pointing to an area for future refinement.

Topic

Topic integration was the strongest-performing feature, with an average score of 1.920. Across all grades, the passages consistently incorporated the given topics, with little variation or decline in performance. The system was able to handle a wide range of topics, including more abstract ones assigned to higher grades, such as 'war' and 'non-violent protests' sensitively. However, it is important to note that we deliberately avoided testing for highly sensitive or inappropriate topics to prevent conflicts with the GPT model's content moderation prompting. This decision ensured that the topic integration scores remained unaffected by any filtering or blocking behavior from the model.

The high scores for topic integration reflect the model's strength in adapting to user-specified topics, indicating that it can reliably maintain topical relevance across varying grade levels and content types.

Additional Observations

In addition to feature-level analysis, two notable patterns emerged during the passage generation process: repetitive story intros and outros and character naming. The LLM tended to generate repetitive or formulaic introductions and conclusions for many stories. This was often paired with traditional names like "Mary", "Bob", or "Timmy" which showed low cultural diversity. While this did not significantly affect the integration scores, it may reduce the overall engagement of the passages, particularly when multiple stories are read consecutively. Across multiple passages, the LLM frequently reused character names, which could reduce the diversity and uniqueness of the stories especially in regards to diversity, equity, and inclusion in narratives. Although this did not impact the core feature scores, it highlights a

potential limitation in narrative variety that may warrant further exploration.

5.2 Results for RQ2: Question Generation and Grading

This section presents the results of our evaluation of the system’s ability to generate questions and grade student responses accurately. We analyzed five main question categories: Factual Recall (FR), Character Actions or Feelings (CAF), Inferences (I), Cause and Effect (CE), and Lessons/Moral of the Story (MS). Our analysis focused not only on the variety of questions generated but also on the discrepancies between human and GPT grading.

The authors performed non-expert human grading to evaluate responses generated by the system and compared these with GPT’s automated grading. While the human graders were not reading comprehension experts, their evaluations help reveal how well the model handles common mistakes, nuanced errors, and partially correct answers. The simulated student responses, also provided by the authors, included varied answer formats to test the grading system’s robustness. The bias involved in using the authors for both the evaluation question-answering and question-grading was unavoidable due to resource constraints that prevented hiring of expert graders or student testers. This is discussed more in Section 6.2 and we hope future evaluations can include expert evaluators. These responses intentionally incorporated common issues such as:

- Providing partial answers with only some of the required elements.
- Using synonyms or paraphrased content instead of exact phrasing.
- Including both minor and major spelling mistakes.
- Responding with incorrect answers (e.g., "I don't know").

This setup allowed us to assess the Grader AI model’s ability to handle nuanced answers, spelling errors, and context-based grading.

5.2.1 Question Distribution and Variety

The total number of questions generally increased with grade level, as shown in Table 5.2. However, certain question categories, particularly Inferences (I) and Cause and Effect (CE), were generated less frequently compared to others like Factual Recall (FR) and Character Actions/Feelings (CAF).

Grade	FR	CAF	I	CE	MS	Total
Grade 1	29	20	1	7	19	76
Grade 2	26	22	2	7	21	78
Grade 3	25	27	4	4	20	80
Grade 4	25	23	9	6	18	81
Grade 5	25	27	12	11	12	87
Total	130	119	28	35	90	402

Table 5.2: Question count across grade and categories

This discrepancy reflects the nature of these question types: Inference and Cause and Effect questions require deeper understanding of the story’s context, making them more challenging for the question-generation model to create consistently. The lower frequency of these questions suggests that the system may benefit from enhancements to capture more nuanced aspects of the passage content. This could most likely be encouraged through prompt engineering of specific question types or goals in question generation (such as mandating certain question types like inference or cause and effect).

5.2.2 Grading Discrepancies by Category

Table 5.3 highlights the average loss (discrepancy) between GPT and human grades across categories. For example, if Grader AI scored a 2 and the human grader scored

a 1, this was a discrepancy of 1. Likewise, if Grader AI scored a 0 and the human grader scored a 2, this was a discrepancy of 2. The largest discrepancies were observed in the Cause and Effect (CE) category, with an average loss of 0.514. This suggests that the model struggled with context-based reasoning, often awarding full credit to responses that were only partially correct.

Grade	FR	CAF	I	CE	MS
Grade 1	0.276	0.150	0.000	0.857	0.158
Grade 2	0.231	0.273	0.000	0.429	0.190
Grade 3	0.120	0.259	0.000	0.500	0.150
Grade 4	0.040	0.130	0.222	0.333	0.167
Grade 5	0.280	0.222	0.167	0.455	0.250
Total Averages	0.192	0.210	0.143	0.514	0.178

Table 5.3: Average loss between GPT and human grades by grade Level

In contrast, categories like Factual Recall (FR) and Inferences (I) showed lower discrepancies, indicating that the system was more reliable in grading straightforward factual content or identifying correct inferences. However, Inference questions were rarely generated, making it difficult to draw broad conclusions about the system’s performance in this category.

5.2.3 Additional Findings

Several noteworthy grading patterns emerged during the evaluation. After identifying the categories with more and less loss, we reviewed the corresponding question and answer pairs to identify specific human and LLM grading discrepancy scenarios. While discrepancies sometimes had niche explanations, common trends across loss scores include:

- **Reluctance to Assign Zeros:** In many instances where students provided incorrect or incomplete answers, Grader AI assigned a score of 1 instead of 0, demonstrating a tendency to be lenient in its grading if the student answer was

at all related to the question.

- **Inconsistent Treatment of Spelling and Grammar Errors:** The system's handling of spelling and grammar errors was inconsistent. Minor errors, such as a single-letter typo, sometimes resulted in a reduced score, while more significant errors, such as every word in a sentence being misspelled, occasionally received full credit.
- **Challenges with Complex Answers:** When students provided partially correct answers, such as including one out of two required elements, Grader AI sometimes assigned full credit despite the incompleteness of the response. This indicates that the model struggles with nuanced grading that requires precise assessment of multiple elements.

While these limitations were consistent enough to mention, ultimately, they don't take away from our discrepancy scores in Table 5.3 which show a relatively low rate of discrepancy. Additionally we originally intended to include vocabulary-based questions in the question generation process, but we found that the system often struggled to test students on the vocabulary's meaning consistently. This work provides evidence that LLM's can be used to generate a variety of context questions and perform semi-accurate grading, however care needs to be taken to ensure prompts direct models as needed.

Chapter 6

Conclusion

In this thesis, we explored the potential of using an LLM to enhance personalized reading comprehension learning. Our goal was to integrate learning objectives, grammar rules, and vocabulary words into an intelligent tutoring system for elementary students to engage students while addressing critical areas of reading comprehension instruction. We aim to create a platform that fosters independent learning and engages elementary students across various grade levels by providing personalized topics, adaptive scoring, and immediate corrective feedback.

Unlike traditional resources that often require the purchase of finite, supplementary materials, our program offers an accessible solution with varied, infinite content generation. With user-specific input such as grade levels and customized topics, grade-specific content such as LOs, vocabulary, grammar, and length, Read with Emora facilitates targeted sessions that align with researched educational standards and RC-ITS and ITS personalized learning features. By allowing students to learn independently and providing immediate feedback, we create a platform that makes personalized reading comprehension teaching more accessible.

6.1 Evaluation Results

In our evaluation, we focused on two key research questions. First, we assessed the system’s ability to generate educationally relevant reading passages incorporating learning objectives, grammar rules, vocabulary words, and personalized topics. Second, we evaluated the system’s capability to create practical, context-based questions from these passages and grade student responses accurately.

6.1.1 Passage Evaluation

The results from the passage generation experiment indicate that the LLM effectively generates engaging, educationally relevant content across grade levels. The system integrated topics exceptionally well, demonstrating its ability to adapt specific topics to full narratives. Grammar rules were generally well-incorporated, though the system struggled with more advanced grammatical constructs in higher-grade passages. Vocabulary integration was effective when fewer than five words were required but became more challenging with larger sets, often resulting in vocabulary being supplemental rather than central to the story.

Learning objectives, particularly those requiring factual integration (e.g., historical events or scientific discoveries), posed the most significant challenge for the system over each grade. This suggests that while excelling in creative storytelling, the LLM may require additional prompt refinement to incorporate non-fictional content better. Furthermore, repetitive narrative elements and character names indicate areas where content generation could be improved through prompt engineering to enhance engagement and variety.

These findings highlight the system’s strengths in storytelling and topical relevance, with opportunities for refinement in higher-grade content and specialized educational objectives.

6.1.2 Question Evaluation

The question generation and grading evaluation revealed strengths and limitations in Read with Emora's performance. Question AI excelled in generating factual and character-based questions, which align well with straightforward, context-independent content. It struggled with more complex question types such as Cause and Effect and Inference, which require a deeper contextual understanding of the passages.

Grading discrepancies were most pronounced in the Cause and Effect category. Grader AI often awarded full credit for partially correct answers, reflecting a limitation in the model's context-based reasoning abilities. Additionally, the system demonstrated inconsistent handling of spelling and grammar errors, sometimes penalizing minor mistakes while ignoring more severe ones. Additionally Grader AI presented a reluctance to assign zero scores, emphasizing our encouragement toward lenient grading in prompts which could impact the reliability of its assessments.

Despite these limitations, Read with Emora demonstrates potential as a tool for automated grading. Aligning its grading more closely with human evaluators, particularly for nuanced, context-based questions-remains an area for future improvement.

6.2 Limitations

While this project achieved its primary objectives, several limitations impacted the development and evaluation process. One of the main challenges was the need for academically sourced grade-specific vocabulary and grammar lists. We took efforts to reach out to educators and authors of robust sources, but were unable to make connections or access their expertise. Without these sources, it was more difficult to ensure consistency and rigor of the educational content we provided across grade levels. Developing or accessing a standardized set of grammar rules and vocabulary words for future iterations would enhance the system's reliability.

The same individual who developed the system also provided the student answers and human-graded scores for the evaluation. While every effort was made to maintain objectivity, this introduces a potential source of bias. Future studies could involve multiple graders to ensure more consistent and unbiased evaluations. Furthermore, we originally aimed to include user feedback in this thesis for further analysis on the corrective feedback system and the spaced repetition algorithm's impact on learning objective variety. Due to technical issues (CORS and database difficulties), we were unable to gather enough feedback in time for evaluation and inclusion in this thesis.

6.3 Future Directions

The current implementation of our ITS demonstrates potential for LLMs in personalized reading comprehension, but several areas for further development and exploration remain.

6.3.1 Vocabulary and Grammar

Academic journal sources on elementary grammar by grade are extremely limited. One key area for future work is creating more robust and academically sourced vocabulary lists and grammar rules that students can adhere to by grade. Having academically sourced sets of vocabulary and grammar data would enhance our system's ability to provide more targeted and reliable instruction. This access would ensure that the most relevant grammar rules are introduced at the appropriate grade levels.

We were unable to generate vocabulary-based questions that tested a student's knowledge of the meaning of new words. A more refined question generating system prompt could help to implement a vocabulary scoring system that could better assess a student's knowledge of new words. A vocabulary scoring system would allow for a spaced repetition grading system for vocabulary. We believe this would significantly

improve our program scoring as term recall is one aspect of comprehension most benefited by spaced repetition [53]. Instead, spaced repetition works in our program to vary passage context through learning objectives.

6.3.2 Feature Expansion

We planned to implement several features that were not completed due to time and scope constraints. These feature implementations remain promising avenues for future work in Read with Emora and reading comprehension. These include:

- **Image generation:** Adding images to passages could enhance student engagement and aid comprehension, particularly for younger learners. This would also help us to be able to incorporate more objectives from the MDOE into our project, as many were excluded based on the necessity of visual aids.
- **Speech diarization and accents:** Integrating speech recognition and dialect customization would improve accessibility for students with diverse language backgrounds. This could also help students learning English as a second language.
- **Speech-to-text and text-to-speech:** These features could help younger users interact with the system by allowing them to read or listen to passages aloud, reinforcing comprehension and expanding our program's targeted age range. It would also allow parents and teachers a more hands-off approach when navigating students through sessions as the website could use audible directives to guide students through the site itself.

These features could significantly enhance the learning experience and accessibility of Read with Emora by further personalizing the tutoring system to student needs.

6.3.3 Higher Fidelity Implementations

A higher-fidelity implementation of Read with Emora could lead to more impactful results [59]. Testing the system in real-world classroom environments or at home and monitoring how learning objective ease factors, question scores, and number of attempts change over time would provide valuable insights into how the system performs in practice. This evaluation would also provide concrete numerical results as to the effect of our tutoring system on reading comprehension directly and user feedback on our corrective feedback mechanism. We can better understand how the system supports student learning over time by conducting more controlled studies and gathering feedback from teachers, students, and parents.

6.3.4 Second Language Learning Benefits

Reading comprehension and spaced repetition activities have been shown to support second language learning by increasing contextual awareness of vocabulary and grammar rules [21]. Our system’s focus on contextualized learning objectives and language instruction positions it to benefit both elementary and second-language learners. In future iterations, we aim to explore how our ITS can be tailored to meet student needs in English second-language learning, leveraging its capacity for personalized instruction to support diverse learning goals.

6.4 Final Thoughts

In conclusion, the development and evaluation of our intelligent tutoring system demonstrates that LLMs can indeed enhance personalized reading comprehension instruction. While the system excels in creative storytelling and topic customization, challenges remain in integrating fact-based content and aligning automated grading with human evaluations. These findings highlight the potential of AI-driven learning

platforms to provide accessible, engaging, and practical education for elementary students and beyond.

With further development, including improved vocabulary and grammar integration, more nuanced question generation, and real-world testing, our system could offer a powerful tool for both general education and second-language learning. As AI technologies continue to evolve, personalized learning platforms like Read with Emora have the potential to revolutionize how students learn, making high-quality education more accessible to all.

Appendix A

Appendix

A.1 Learning Objectives by Grade

See Tables A.1 and A.2.

Literature	Key Ideas and Details	Understanding Key Details
		Retelling Stories
		Describing Characters and Events
	Craft and Structure	Understanding Words and Phrases
		Story Structure
		Explaining Dialogue
Integration of Knowledge and Ideas	Comparing and Contrasting Stories	
Informational	Key Ideas and Details	Connecting Historical Events
		Connecting Scientific Discoveries
		Describing People and Events
	Craft and Structure	Comparing and Contrasting Perspectives
		Supporting Arguments
	Integration of Knowledge and Ideas	Integrating Information

Table A.1: Fourth Grade Learning Objectives Based on the MDOE Guide

Literature	Key Ideas and Details	Understanding Key Details
		Retelling Stories
		Describing Characters and Events
	Craft and Structure	Understanding Words and Phrases
		Story Structure
		Explaining Dialogue
Integration of Knowledge and Ideas	Comparing and Contrasting Stories	
Informational	Key Ideas and Details	Connecting Historical Events
		Connecting Scientific Discoveries
		Describing People and Events
	Craft and Structure	Article Structure
		Comparing and Contrasting Perspectives
	Integration of Knowledge and Ideas	Supporting Arguments
Integrating Information		

Table A.2: Fifth Grade Learning Objectives Based on the MDOE Guide

A.2 Excluded Learning Objectives by Grade

See Tables A.3, A.4, A.5, A.6.

A.3 Grammar Rules by Grade

See Tables A.7, A.8, A.9, A.10

A.4 GitHub and Data

Our full datasets, including full vocabulary lists, grammar rules, and learning objective sets per grade, can be found at our public Read with Emora Github repository (<https://github.com/catherinewbaker/Read-with-Emora>). Data from our experiments and evaluation can be found at our public google sheet (https://docs.google.com/spreadsheets/d/12SflWH-2N0sLYC6vTJQaVprB5jkU_ah6BLwLMPicU/edit?usp=sharing).

Objective	Description	Exclusion Reason
Integration of Knowledge and Ideas	Use visual aids and details in a story to describe its characters, setting, or events	Untestable - requires visual aids
Craft and structure	Ask and answer questions to help determine or clarify the meaning of words and phrases in a text	Untestable - requires human monitoring
Craft and structure	Know and use various text features (e.g., headings, tables of contents, glossaries, electronic menus, icons) to locate key facts or information in a text	Tested - through website use
Craft and structure	Distinguish between information provided by pictures or other visual aids and information provided by the words in a text	Untestable - requires visual aids
Integration of Knowledge and Ideas	Use the visual aids and details in a text to describe its key ideas	Untestable - requires visual aids

Table A.3: First Grade Objectives Excluded from Learning Based on the MDOE Guide

Objective	Description	Exclusion Reason
Integration of Knowledge and Ideas	Explain how specific aspects of a text’s visual aids contribute to what is conveyed by the words in a story (e.g., create mood, emphasize aspects of a character or setting)	Untestable - requires visual aids
Craft and structure	Determine the meaning of general academic and domain-specific words and phrases in a text relevant to a grade 3 topic or subject area	Tested - through other objectives
Craft and structure	Use text features and search tools (e.g., key words, sidebars, hyperlinks) to locate information relevant to a given topic efficiently	Tested - through website use
Craft and structure	Distinguish their own point of view from that of the author of a text	Tested - through other objectives
Integration of Knowledge and Ideas	Use information gained from visual aids (e.g., maps, photographs) and the words, numbers, and symbols in a text to demonstrate understanding of the text (e.g., where, when, why, and how key events occur)	Untestable - requires visual aids

Table A.4: Third Grade Objectives Excluded from Learning Based on the MDOE Guide

Objective	Description	Exclusion Reason
Integration of Knowledge and Ideas	Make connections between a written story or drama and its visual or oral presentation, identifying where the presentation reflects specific descriptions and directions in the written text	Untestable - requires visual aids
Craft and structure	Determine the meaning of general academic and domain-specific words or phrases in a text relevant to a grade 4 topic or subject area	Tested - through other objectives
Craft and structure	Describe the overall structure (e.g., chronology, comparison, cause/effect, problem/solution) of events, ideas, concepts, or information in a text or part of a text	Tested - through other objectives
Integration of Knowledge and Ideas	Interpret information presented visually, orally, or quantitatively (e.g., in charts, graphs, diagrams, timelines, animations, or interactive elements on webpages) and explain how the information contributes to an understanding of the text in which it appears	Untestable - requires visual aids

Table A.5: Fourth Grade Objectives Excluded from Learning Based on the MDOE Guide

Objective	Description	Exclusion Reason
Key Ideas and Details	Quote or paraphrase a text accurately when explaining what the text states explicitly and when drawing inferences from the text	Tested - through question-answering
Integration of Knowledge and Ideas	Analyze how visual and multimedia elements contribute to the meaning, tone, or beauty of a text (e.g., graphic novel; multimedia presentation of fiction, folktale, myth, poem)	Untestable - requires visual aids
Craft and structure	Determine the meaning of general academic and domain-specific words and phrases in a text relevant to a grade 5 topic or subject area.	Tested - through other objectives
Integration of Knowledge and Ideas	Draw on information from multiple print or digital sources, demonstrating the ability to locate an answer to a question quickly or to solve a problem efficiently	Tested - through platform use

Table A.6: Fifth Grade Objectives Excluded from Learning Based on the MDOE Guide

Grammar Rule
Nouns and Pronouns
Singular and Plural
Verbs
Adjectives
Subjects and Predicate
Use of Is and Are/Contractions With Not
Conjunctions
Determiners and Articles
Using Prepositions
Prefix and Suffix
Statements, Questions, and Commands
Capitalization, Period, and Exclamation Mark
Use basic subject-verb agreement
Use common, proper, and possessive nouns
Use the articles "a", "an", and "the" in complete simple sentences
Use action verbs to apply past, present & future tense in sentences
Sentence Structure

Table A.7: First Grade Grammar Rules per Big Learners and Night Zookeeper

Grammar Rule
Parts of speech
Common and Proper Nouns
Singular, Plural, and Irregular Nouns
Abstract and Concrete Nouns
Pronouns
Adjectives and Articles
Verbs / Verb Tenses
Adverbs
Prepositions
Sentence Structure
Subject and Predicate
Subject and Verb Agreement
Pronoun and Antecedent Agreement
Declarative and Interrogative Sentences
Imperative and Exclamatory Sentences
Statements, Questions, Commands, and Exclamations
Coordinating and Subordinating Conjunctions
Sentence Combining and Compound Sentences
Punctuation and Capitalization
Use of Comma
Commas in Dates, Locations, and Addresses
Capitalization
Spelling and Spelling Patterns
Blends and Contractions
Orthographic Patterns
Homophones and Homographs
Words in Alphabetical Order
Vocabulary and Concept Development
High Frequency Words
Use Dictionary
Compound Words
Antonyms and Synonyms
Vocabulary Acquisition and Use
Literal and Figurative Meaning of Words/Phrases
Use abstract nouns
Use possessive nouns correctly
Write in simple, compound and complex sentences

Table A.8: Third Grade Grammar Rules per Big Learners and Night Zookeeper

Grammar Rule
Conventions of Standard English
Relative Pronouns and Adverbs
Form and Use of Progressive Verb Tenses
Use Modal Auxiliaries
Order Adjectives within Sentences
Form and Use Prepositional Phrases
Complete Sentences
Using Confusing Words Correctly
Capitalization
Use Commas and Quotation Marks
Using Commas with Coordinating Conjunctions
Spelling
Knowledge of Language
Conveying Ideas with Words and Phrases
Choose Punctuation for Effect
Use of Formal and Informal English
Vocabulary Acquisition and Use
Define Words and Phrases From Context Clues
Greek and Latin Affixes and Roots
Using Dictionaries, Glossaries, and Thesauruses
Similes and Metaphors
Idioms, Adages, and Proverbs
Antonyms and Synonyms

Table A.9: Fourth Grade Grammar Rules per Big Learners and Night Zookeeper

Grammar Rule
Conventions of Standard English
Conjunctions
Interjection / Prepositions
Perfect Verb Tenses
Verb Tenses
Inappropriate Shifts in Verb Tense
Correlative Conjunctions
Commas: Items in a Series
Using Commas to Separate the Introductory Elements
Comma Usage Within Sentences
Using Titles in Text
Grade-Appropriate Spelling Words
Knowledge of Language
Combine, Expand, and Reduce Sentences
Define Words and Phrases From Context Clues
Greek and Latin Affixes and Roots
Pronunciation and Meaning: Using Reference Materials
Similes and Metaphors
Idioms, Adages, and Proverbs
Antonyms, Synonyms, and Homographs
Use verb tense to display time, sequence and conditions

Table A.10: Fifth Grade Grammar Rules per Big Learners and Night Zookeeper

Bibliography

- [1] Results from the 2019 mathematics and reading assessments: National average scores and percentiles. Technical report, National Center for Education Statistics (NCES), October 2019. URL <https://www.nationsreportcard.gov>. NAEP Report.

- [2] Rajat Agarwal, Varun Khurana, Karish Grover, Mukesh Mohania, and Vikram Goyal. Multi-relational graph transformer for automatic short answer grading. In Marine Carpuat, Marie-Catherine de Marneffe, and Ivan Vladimir Meza Ruiz, editors, *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2001–2012, Seattle, United States, July 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.naacl-main.146. URL <https://aclanthology.org/2022.naacl-main.146>.

- [3] Yasar Akyuz. Effects of intelligent tutoring systems (its) on personalized learning (pl). *Creative Education*, 11(06):26, 2020. doi: 10.4236/ce.2020.116069. URL <https://www.scirp.org/journal/paperinformation?paperid=101287>.

- [4] Hadi Amiri. Neural self-training through spaced repetition. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 21–31,

- Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1003. URL <https://aclanthology.org/N19-1003>.
- [5] *Anki Manual*. Anki, 2024. URL <https://docs.ankiweb.net/intro.html>. Accessed: 2024-06-21.
- [6] Handan Atun. Intelligent tutoring systems (its) to improve reading comprehension: A systematic review. *Journal of Teacher Education and Lifelong Learning*, 2(2): 77–89, 2020.
- [7] Big Learners. Big learners: Grammar rules. <https://www.biglearners.com>, 2024. URL <https://www.biglearners.com>. Accessed: 2024-6-15.
- [8] Neil A. Bradbury. Attention span during lectures: 8 seconds, 10 minutes, or more? *Advances in Physiology Education*, 40(4):509–513, 2016. doi: 10.1152/advan.00109.2016. URL <https://doi.org/10.1152/advan.00109.2016>. PMID: 28145268.
- [9] Marc Brysbaert. How many words do we read per minute? a review and meta-analysis of reading rate. *Journal of Memory and Language*, 109:104047, 2019. ISSN 0749-596X. doi: <https://doi.org/10.1016/j.jml.2019.104047>. URL <https://www.sciencedirect.com/science/article/pii/S0749596X19300786>.
- [10] Sophia Chan, Swapna Somasundaran, Debanjan Ghosh, and Mengxuan Zhao. AGReE: A system for generating automated grammar reading exercises. In Wanxiang Che and Ekaterina Shutova, editors, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 169–177, Abu Dhabi, UAE, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-demos.17. URL <https://aclanthology.org/2022.emnlp-demos.17>.

- [11] W. W. Charters. Constructing a language and grammar scale. *The Journal of Educational Research*, 1(4):249–257, 1920. doi: 10.1080/00220671.1920.10879050. URL <https://doi.org/10.1080/00220671.1920.10879050>.
- [12] Yulin Chen, Ning Ding, Hai-Tao Zheng, Zhiyuan Liu, Maosong Sun, and Bowen Zhou. Empowering private tutoring by chaining large language models. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management (CIKM '24)*, page 11, New York, NY, USA, October 21–25 2024. ACM. doi: 10.1145/3627673.3679665. URL <https://dl.acm.org/doi/pdf/10.1145/3627673.3679665>.
- [13] Yucheng Chu, Hang Li, Kaiqi Yang, Harry Shomer, Hui Liu, Yasemin Copur-Gencturk, and Jiliang Tang. A llm-powered automatic grading framework with human-level guidelines optimization, 2024. URL <https://arxiv.org/abs/2410.02165>.
- [14] Stephen Colbran, Wayne Jones, and John Milburn. Comparing spaced repetition algorithms for legal digital flashcards. *Open oceans: Learning without borders*, 92: 92–102, 2018.
- [15] Kevyn Collins-Thompson and James P Callan. A language modeling approach to predicting reading difficulty. In *Proceedings of the human language technology conference of the North American chapter of the association for computational linguistics: HLT-NAACL 2004*, pages 193–200, 2004.
- [16] Benjamin Y. Conklin. *A Complete Graded Course in English Grammar and Composition*. American Book Co., 1888. URL https://www.google.com/books/edition/_/nUQ5AQAAMAAJ?hl=en&gbpv=1&pg=PA1&dq=complete+english+grammar+rules+by+grade.
- [17] R. Cowan. *The Teacher's Grammar of English with Answers: A Course Book*

- and Reference Guide*. The Teacher's Grammar of English. Cambridge University Press, 2008. ISBN 9780521809733. URL <https://books.google.com/books?id=WCCk829jmzUC>.
- [18] Amy M. Elleman and Eric L. Oslund. Reading comprehension research: Implications for practice and policy. *Policy Insights from the Behavioral and Brain Sciences*, 6(1):3–11, 2019. doi: 10.1177/2372732218816339. URL <https://doi.org/10.1177/2372732218816339>.
- [19] Flocabulary. Word lists. <https://www.flocabulary.com/wordlists/>, 2024. Accessed: 2024-10-09.
- [20] Matthew Goldman, Jaimie Bryan, and Brandon Lucke-Wold. Evidence-based algorithm “anki” for optimization of medical education: The evolution of knowledge retention. *Preprints*, July 2023. doi: 10.20944/preprints202307.1955.v1. URL <https://doi.org/10.20944/preprints202307.1955.v1>.
- [21] Aroline E. Seibert Hanson and Christina M. Brown. Enhancing l2 learning through a mobile assisted spaced-repetition tool: an effective but bitter pill? *Computer Assisted Language Learning*, 33(1-2):133–155, 2020. doi: 10.1080/09588221.2018.1552975. URL <https://doi.org/10.1080/09588221.2018.1552975>.
- [22] Chaoqun He, Renjie Luo, Shengding Hu, Ranchi Zhao, Jie Zhou, Hanghao Wu, Jiajie Zhang, Xu Han, Zhiyuan Liu, and Maosong Sun. UltraEval: A lightweight platform for flexible and comprehensive evaluation for LLMs. In Yixin Cao, Yang Feng, and Deyi Xiong, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 247–257, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-demos.23. URL <https://aclanthology.org/2024.acl-demos.23>.

- [23] Guoxiu He, Jiacheng Yao, and Xin Xu. A self-questioning framework towards knowledge self-organization in children’s readings via hybrid prompt learning, 07 2024. URL https://www.researchgate.net/publication/381994924_A_Self-Questioning_Framework_Towards_Knowledge_Self-Organization_in_Children’s_Readings_via_Hybrid_Prompt_Learning. Preprint, not peer-reviewed.
- [24] Lesley Ito. How should grammar be taught to children? In *Japan Associate for Language Teaching*, Japan, May 2019. JALT Publications. URL <https://jalt-publications.org/jj>.
- [25] Corentin Jouault and Kazuhisa Seta. Content-dependent question generation for history learning in semantic open learning space. In Stefan Trausan-Matu, Kristy Elizabeth Boyer, Martha Crosby, and Kitty Panourgia, editors, *Intelligent Tutoring Systems*, pages 300–305, Cham, 2014. Springer International Publishing. ISBN 978-3-319-07221-0. URL https://link.springer.com/chapter/10.1007/978-3-319-07221-0_37#citeas.
- [26] K12 Reader. Fifth grade spelling words. <https://www.k12reader.com/fifth-grade-spelling-words/>, 2024. Accessed: 2024-10-09.
- [27] Emil Kalbaliyev and Kairit Sirts. On narrative question answering skills. In Kevin Duh, Helena Gomez, and Steven Bethard, editors, *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 814–820, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.naacl-short.73. URL <https://aclanthology.org/2024.naacl-short.73>.
- [28] Sean H. K. Kang. Spaced repetition promotes efficient and effective learning:

- Policy implications for instruction. *Policy Insights from the Behavioral and Brain Sciences*, 3(1):12–19, 2016. doi: 10.1177/2372732215624708. URL <https://doi.org/10.1177/2372732215624708>.
- [29] Vishwajeet Kumar, Sivaanandh Muneeswaran, Ganesh Ramakrishnan, and Yuanfang Li. ParaQG: A system for generating questions and answers from paragraphs. In Sebastian Padó and Ruihong Huang, editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP): System Demonstrations*, pages 175–180, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-3030. URL <https://aclanthology.org/D19-3030>.
- [30] Learning at the Primary Pond. Learning at the primary pond. <https://learningatthepriarypond.com>, 2024. Accessed: 2024-9-26.
- [31] Yan Ma, Yu Qiao, and Pengfei Liu. MoPS: Modular story premise synthesis for open-ended automatic story generation. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2135–2169, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.117. URL <https://aclanthology.org/2024.acl-long.117>.
- [32] Nitin Madnani, Beata Beigman Klebanov, Anastassia Loukina, Binod Gyawali, Patrick Lange, John Sabatini, and Michael Flor. My turn to read: An interleaved E-book reading tool for developing and struggling readers. In Marta R. Costajussà and Enrique Alfonseca, editors, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 141–146, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-3024. URL <https://aclanthology.org/P19-3024>.

- [33] Massachusetts Department of Elementary and Secondary Education. *English Language Arts and Literacy: Grades Pre-Kindergarten to 12, Massachusetts Curriculum Framework*. Massachusetts Department of Elementary and Secondary Education, Malden, MA, 2017. URL <https://www.doe.mass.edu/frameworks/ela/2017-06.pdf>. Board of Elementary and Secondary Education Members: Paul Sagan (Chair), James Morton (Vice Chair), et al.
- [34] Karen Mazidi and Rodney D. Nielsen. Pedagogical evaluation of automatically generated questions. In Stefan Trausan-Matu, Kristy Elizabeth Boyer, Martha Crosby, and Kitty Panourgia, editors, *Intelligent Tutoring Systems*, pages 294–299, Cham, 2014. Springer International Publishing. ISBN 978-3-319-07221-0.
- [35] Shane Mooney. Spaced repetition for all: Cognitive science meets big data in a procrastinating world, March 2017. URL <https://quizlet.com/blog/spaced-repetition-for-all-cognitive-science-meets-big-data-in-a-procrastinating-world>. Published on the Quizlet Blog.
- [36] Amy L Moore. Reading comprehension: Cognitive skills, strategies, and interventions, 2014. URL <https://api.semanticscholar.org/CorpusID:12862819>. Published online on Semantic Scholar.
- [37] Brent Morgan, Marie Hogan, Andrew Hampton, Anne Lippert, and Arthur Graesser. *The Need for Personalized Learning and the Potential of Intelligent Tutoring Systems*, pages 1–38. 03 2020.
- [38] Jack Mostow, Jessica Nelson-Taylor, and Joseph E. Beck. Computer-guided oral reading versus independent practice: Comparison of sustained silent reading to an automated reading tutor that listens. In *Journal of Educational Computing Research*, volume 49, pages 249–276. Baywood Publishing Co., Inc., 2013. doi: 10.2190/EC.49.2.g. URL

<https://www.istation.com/hubfs/6.%20Research/1b.%20Full%20Studies/Amira%20Research%20Study%20-%20Comparison%20of%20Sustained%20Silent%20Reading%20to%20Amiras%20Automated%20Tutor.pdf>.

- [39] Holly Munson, Anton Yu, Ananya Rajgarhia, and AJ Noh. Introducing the new duolingo learning path: Learn more about the science behind our exciting home screen redesign!, May 2023. URL <https://blog.duolingo.com/new-duolingo-home-screen-design/>. Published on the Duolingo Blog.
- [40] Night Zookeeper. Night zookeeper: Grammar. <https://www.nightzookeeper.com/language-arts/grammar>, 2024. Accessed: 2024-6-15.
- [41] David Nunan. Teaching grammar in context. *English Language Teaching Journal*, 52(2):101–109, 04 1998. ISSN 0951-0893. doi: 10.1093/elt/52.2.101. URL <https://doi.org/10.1093/elt/52.2.101>.
- [42] Scott G. Paris. Reinterpreting the development of reading skills. *Reading Research Quarterly*, 40(2):184–202, April/May/June 2005. ISSN 0034-0553. doi: 10.1598/RRQ.40.2.3. URL https://lchc.ucsd.edu/MCA/Mail/xmcamail.2013_09.dir/pdfy5hhXiK7kw.pdf.
- [43] Shrimai Prabhumoye, Chris Quirk, and Michel Galley. Towards content transfer through grounded text generation. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2622–2632, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1269. URL <https://aclanthology.org/N19-1269>.
- [44] Corby Rosset, Guoqing Zheng, Victor Dibia, Ahmed Awadallah, and Paul Bennett. Axiomatic preference modeling for longform question answering. In Houda

- Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 11445–11475, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.702. URL <https://aclanthology.org/2023.emnlp-main.702>.
- [45] Sherry Ruan, Liwei Jiang, Justin Xu, Bryce Joe-Kun Tham, Zhengneng Qiu, Yeshuang Zhu, Elizabeth L. Murnane, Emma Brunskill, and James A. Landay. Quizbot: A dialogue-based adaptive learning system for factual knowledge. In *CHI Conference on Human Factors in Computing Systems Proceedings (CHI 2019)*, pages 1–13, New York, NY, USA, 2019. ACM. doi: 10.1145/3290605.3300587. URL <https://doi.org/10.1145/3290605.3300587>.
- [46] Johannes Schneider, Bernd Schenk, and Christina Niklaus. Towards llm-based autograding for short textual answers, 2024. URL <https://arxiv.org/abs/2309.11508>.
- [47] G. Shi, A.M. Lippert, and K. et al Shubeck. Exploring an intelligent tutoring system as a conversation-based assessment tool for reading comprehension. *Behaviormetrika*, 45:615–633, 2018. doi: 10.1007/s41237-018-0065-9. URL <https://doi.org/10.1007/s41237-018-0065-9>.
- [48] Catherine Snow. *Reading for Understanding: Toward an RD Program in Reading Comprehension*. RAND Corporation, Santa Monica, CA, 2002. URL https://www.rand.org/pubs/monograph_reports/MR1465.html.
- [49] Venkat Srinivasan and Hemavathi Murthy. Improving reading and comprehension in k-12: Evidence from a large-scale ai technology intervention in india. *Computers and Education: Artificial Intelligence*, 2:100019, 2021. ISSN 2666-920X. doi: <https://doi.org/10.1016/j.caeai.2021.100019>.

[//doi.org/10.1016/j.caeai.2021.100019](https://doi.org/10.1016/j.caeai.2021.100019). URL <https://www.sciencedirect.com/science/article/pii/S2666920X21000138>.

- [50] Venkat Srinivasan and Hemavathi Murthy. Improving reading and comprehension in k-12: Evidence from a large-scale ai technology intervention in india. *Computers and Education: Artificial Intelligence*, 2:100019, 2021. ISSN 2666-920X. doi: <https://doi.org/10.1016/j.caeai.2021.100019>. URL <https://www.sciencedirect.com/science/article/pii/S2666920X21000138>.
- [51] Saku Sugawara, Nikita Nangia, Alex Warstadt, and Samuel Bowman. What makes reading comprehension questions difficult? In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6951–6971, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.479. URL <https://aclanthology.org/2022.acl-long.479>.
- [52] Jiao Sun, Yufei Tian, Wangchunshu Zhou, Nan Xu, Qian Hu, Rahul Gupta, John Wieting, Nanyun Peng, and Xuezhe Ma. Evaluating large language models on controlled generation tasks. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 3155–3168, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.190. URL <https://aclanthology.org/2023.emnlp-main.190>.
- [53] Thomas C. Toppino, Jane E. Kasserman, and Wayne A. Mracek. The effect of spacing repetitions on the recognition memory of young children and adults. *Journal of Experimental Child Psychology*, 51(1):123–138, 1991. ISSN 0022-0965. doi: [https://doi.org/10.1016/0022-0965\(91\)90079-8](https://doi.org/10.1016/0022-0965(91)90079-8). URL <https://www.sciencedirect.com/science/article/pii/0022096591900798>.

- [54] Chung-Ting Tsai, Jhih-Jie Chen, Ching-Yu Yang, and Jason S. Chang. LinggleWrite: a coaching system for essay writing. In Asli Celikyilmaz and Tsung-Hsien Wen, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 127–133, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-demos.17. URL <https://aclanthology.org/2020.acl-demos.17>.
- [55] U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics. Highlights of U.S. PISA 2022 results web report (nces 2023-115 and 2024-113), 2023. URL <https://nces.ed.gov/surveys/pisa/pisa2022/>. Accessed: October 28, 2024.
- [56] Vocabulary.com. Word list: 172808. <https://www.vocabulary.com/lists/172808#view=list>, 2024. Accessed: 2024-10-09.
- [57] Phoebe J. Wang and Max Kreminski. Guiding and diversifying llm-based story generation via answer set programming, 2024. URL <https://arxiv.org/abs/2406.00554>.
- [58] Wenlin Wang, Zhe Gan, Hongteng Xu, Ruiyi Zhang, Guoyin Wang, Dinghan Shen, Changyou Chen, and Lawrence Carin. Topic-guided variational auto-encoder for text generation. In Jill Burstein, Christy Doran, and Tamar Solorio, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 166–177, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1015. URL <https://aclanthology.org/N19-1015>.
- [59] Kausalai (Kay) Wijekumar, Bonnie J.F. Meyer, and Puiwa Lei. High-fidelity implementation of web-based intelligent tutoring system improves fourth and

- fifth graders content area reading comprehension. *Computers Education*, 68:366–379, 2013. ISSN 0360-1315. doi: <https://doi.org/10.1016/j.compedu.2013.05.021>. URL <https://www.sciencedirect.com/science/article/pii/S0360131513001498>.
- [60] Piotr A. Wozniak. Optimization of learning: A new approach and computer application. Master’s thesis, University of Technology in Poznan, Computer Science Center, June 1990. Specialization: Programming Techniques and Information Systems, Individual Specialization: Application of Computers in Human Biology.
- [61] Ying Xu, Dakuo Wang, Mo Yu, Daniel Ritchie, Bingsheng Yao, Tongshuang Wu, Zheng Zhang, Toby Li, Nora Bradford, Branda Sun, Tran Hoang, Yisi Sang, Yufang Hou, Xiaojuan Ma, Diyi Yang, Nanyun Peng, Zhou Yu, and Mark Warschauer. Fantastic questions and where to find them: FairytaleQA – an authentic dataset for narrative comprehension. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 447–460, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.34. URL <https://aclanthology.org/2022.acl-long.34>.
- [62] Zhihong Xu, Kausalai Wijekumar, Gilbert Ramirez, Xueyan Hu, and Robin Irey. The effectiveness of intelligent tutoring systems on k-12 students’ reading comprehension: A meta-analysis. *British Journal of Educational Technology*, 50, 03 2019. doi: 10.1111/bjet.12758. URL <https://bera-journals.onlinelibrary.wiley.com/doi/10.1111/bjet.12758>.
- [63] Mo Yu, Jiangnan Li, Shunyu Yao, Wenjie Pang, Xiaochen Zhou, Zhou Xiao, Fandong Meng, and Jie Zhou. Personality understanding of fictional characters during book reading. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors,

Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 14784–14802, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.826. URL <https://aclanthology.org/2023.acl-long.826>.

- [64] Zijie Zeng, Xinyu Li, Dragan Gasevic, and Guanliang Chen. Do deep neural nets display human-like attention in short answer scoring? In Marine Carpuat, Marie-Catherine de Marneffe, and Ivan Vladimir Meza Ruiz, editors, *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 191–205, Seattle, United States, July 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.naacl-main.14. URL <https://aclanthology.org/2022.naacl-main.14>.