Personalized Reading and Writing Tutor: Improving Literacy Skills and Assessment Accuracy

Moksh Kopikar, Council Rock High School South; Naman Mandloi, Holland Middle School

Abstract—**This paper presents the development and evaluation of the Reading Writing Tutor (RWT), a personalized learning assistant powered by Large Language Models (LLMs), designed to enhance students' reading and writing skills according to state education standards. Recognizing the national decline in student literacy and limited writing practice opportunities, the RWT introduces a novel architecture employing distinct Generator and Evaluator LLMs. The Generator creates custom reading passages and assessment questions tailored to student interests and state guidelines, while the Evaluator provides accurate, standards-aligned assessment scoring and personalized feedback. This two-LLM approach, inspired by external work [5], leverages specialized AI capabilities for separate tasks: creative content generation and objective assessment. A key research component evaluated the accuracy of the RWT's LLM-based assessment grading compared to human experts, demonstrating high agreement across question types, including complex text dependent analysis. Qualitative feedback from students and teachers further supports the tutor's perceived helpfulness, engagement, and potential for providing personalized learning and saving teacher time. The successful implementation and evaluation of this novel two-LLM architecture demonstrate a promising direction for developing advanced, accurate, and personalized AI tutors.**

Index Terms—**AI in Education, Large Language Models, Personalized Learning, Reading Comprehension, Writing Skills, Automated Assessment, AI Tutors, Novel Architecture, K-12 Education.**

I. INTRODUCTION

Reading and writing are foundational literacy skills critical for academic success and future opportunities. Recent data indicates a significant decline in U.S. students' performance in these areas [1], exacerbated by factors including insufficient personalized instruction and limited opportunities for in-depth writing practice beyond standardized test formats. Traditional classroom settings often struggle to provide individualized attention and tailored materials necessary to address each student's specific learning needs and interests.

The advent of advanced Artificial Intelligence, particularly Large Language Models (LLMs) [2], presents a unique opportunity to create intelligent tutoring systems that can offer personalized learning experiences [3]. Research suggests that aligning educational content with personal interests can significantly increase student engagement and comprehension [4]. Leveraging this, we developed the Reading Writing Tutor (RWT), an AI-powered application designed to provide students with personalized reading content and practice assessment questions aligned with state educational standards. A **key novel contribution** of the RWT's design is its **two-LLM architecture,** utilizing one LLM specifically as a **Generator** for personalized content and another distinct LLM as an **Evaluator** for assessment. This separation of roles, inspired by approaches in other domains like opinion summary evaluation [5], allows each AI component to be optimized for its specific complex task: creative text and question generation aligned with pedagogical standards, and accurate, rubric-based assessment scoring.

This paper details the design and implementation of the RWT, focusing on this novel two-LLM structure and its ability to generate personalized content and accurately assess student responses according to standardized testing guidelines, using the Pennsylvania PSSA ELA exam as a case study. We investigate the research question: "How does the use of a personalized and interactive AI-based reading writing tutor influence students' reading and writing performance and engagement compared to traditional teaching methods?" We hypothesize that personalization will lead to better understanding and retention and expect outcomes including increased student engagement, positive feedback from users regarding personalization and test preparation, and positive feedback from teachers regarding time-saving and potential impact on test scores. Furthermore, a primary goal was to evaluate the accuracy of the Evaluator LLM's grading compared to human expert assessment, thereby validating a core function of our novel architecture.

I. SYSTEM & METHODOLOGY

The Reading Writing Tutor (RWT) is a web-based application developed using Python, leveraging the capabilities of LLMs (specifically the GPT-4o API or similar) for content generation and response evaluation. The system comprises two main components: a Generator LLM and an Evaluator LLM, working in conjunction to provide a personalized and adaptive learning experience. Google Forms was used for collecting user feedback surveys.

A. System Design

The RWT functions as a chatbot interface where students interact with the AI tutor. The process unfolds as follows:

Interest Capture: The Generator LLM initiates the interaction by probing the student to identify topics of interest (e.g., "Ronaldo's latest wins in soccer"). Input filtering is applied to ensure topic appropriateness.

Passage Generation: Based on the student's interest, the Generator LLM creates a custom reading passage. This passage is generated to match the specific grade's reading level (e.g., Pennsylvania 8th grade ELA level) and aligns with relevant state education standards and eligible content anchors (e.g., PSSA ELA Test Design [6], Common Core Appendix A Lexile bands [7], Achieve the Core Rubric [8], PA Assessment Anchors [9]). The generation process follows a detailed internal algorithm to ensure compliance with complexity and structure requirements.

Question Generation (Generator LLM): Following the passage, the Generator LLM generates assessment questions in formats typical of the state test, including Multiple Choice (MC), Evidence Based Selected Response (EBSR), and Text-Dependent Analysis (TDA) prompts [6]. These questions are created based on an analysis of past years' test samplers to match difficulty and style [11], [12].

Student Response and Evaluation (Evaluator LLM): The student reads the passage and answers the generated questions. The Evaluator LLM then scores the responses and provides personalized feedback based on state scoring guidelines and rubrics (e.g., PSSA TDA Scoring Guidelines [10], past Item and Scoring Samplers [11], [12]). This dedicated evaluation model is designed for accurate, objective scoring.

Adaptive Personalization through Human Feedback: A key feature for adaptation is the ability for students to upload their completed PSSA test papers that have been graded by their human teacher. This teacher-provided grading and feedback serves as crucial training data for both the Generator and Evaluator LLMs. It allows the Generator to better understand the nuances of content and question creation that aligns with human grading standards and helps the Evaluator refine its scoring accuracy by learning from expert human evaluations. This integration of human-graded data provides a vital feedback loop for continuous improvement of the AI models.

Screenshots of the user interface are provided in the appendix

B. Evaluation Methodology

To evaluate the effectiveness and accuracy of the RWT, particularly its grading capabilities, we conducted a study with 8th-grade students (n=16) and teachers (n=5) from Council Rock School District. Informed consent was obtained from participants or guardians.

Evaluator LLM Accuracy Study: We compared the scoring of the Evaluator LLM against an expert human teacher using responses from past PSSA ELA Grade 8 Item and Scoring Samplers (2023 and 2024) [11], [12]. The human expert scores served as the gold standard. We analyzed agreement across MC, EBSR, and TDA question types using the following metrics:

1. Percent Agreement (MC): Proportion of exact score matches.
2. Mean Absolute Error (MAE) (EBSR, TDA): Average absolute difference between LLM and human scores.
3. Quadratic Weighted Kappa (TDA): Agreement measure for ordinal scales, accounting for chance agreement, with quadratic weighting.

Qualitative and System Analysis: Students used the RWT to practice, reading at least one personalized passage and answering questions. Teachers also tried the application. Both groups provided anonymous feedback via Google Forms surveys. Survey questions assessed the perceived helpfulness, engagement, and comparison to human grading..

III. RESULTS

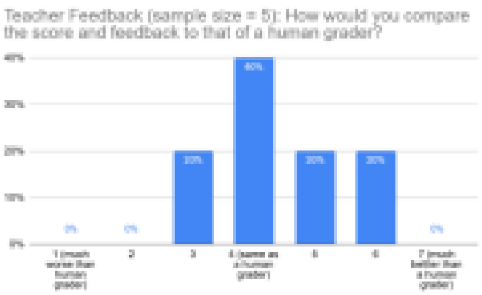
A. Evaluator LLM Accuracy: The comparison between the Evaluator LLM and the expert human grader showed high levels of agreement across all question types:

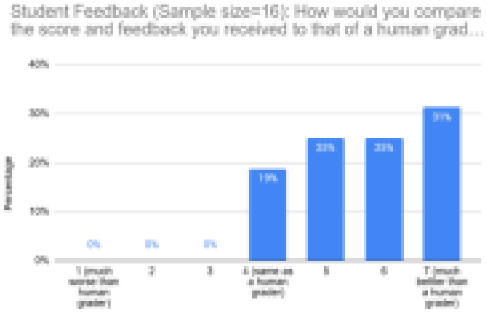
- **Multiple Choice (MC):** Achieved 100.00% Percent Agreement. This confirms the LLM's ability to accurately identify correct responses for this question type.
- **Evidence-Based Selected Response (EBSR):** Showed a Mean Absolute Error (MAE) of 0.10. This demonstrates a very high level of accuracy and close agreement with the expert's scoring, with the average difference being only 0.10 points.
- **Text-Dependent Analysis (TDA):** Demonstrated an MAE of 0.35 and a Quadratic Weighted Kappa of 0.85. The MAE of 0.35 indicates the average difference in holistic scores (on the 1-4 scale) was less than half a point. The Kappa value of 0.85 falls within the "Near Perfect Agreement" range, suggesting a very strong consistency between the Evaluator LLM and the expert in applying the TDA rubric holistically.

These results indicate that the Evaluator LLM can score student responses, including complex written analysis (TDA), with a high degree of accuracy comparable to a human expert.

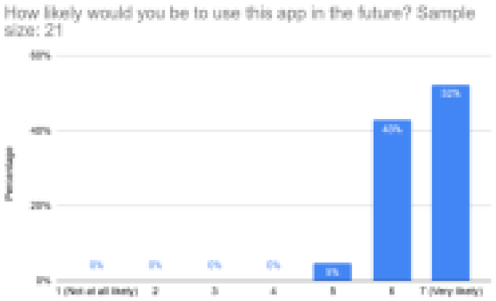
B. User Feedback and System Impact

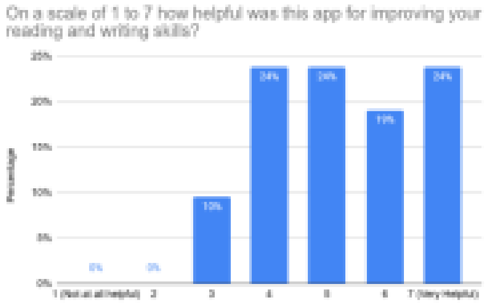
Qualitative feedback from students and teachers highlighted the perceived value of the RWT.
Comparison to Human Grader: Students (n=16) gave an average score of 5.69 (on a 1-7 scale, 7 being much better) when comparing the RWT's scoring and feedback to a human grader, suggesting they found the AI tutor's feedback beneficial, possibly due to its detailed nature based on specific rubrics. Teachers (n=5) gave an average score of 4.4, indicating they found the AI grader as good as a human.





Helpfulness and Future Use: The app received an average helpfulness score of 6.47 (on a 1-7 scale, 7 being very helpful) across both students and teachers. The high likelihood of future use score also supports its perceived value as a practice tool.





The project's impact was further demonstrated by winning the District Science Competition and advancing to the State competition. We are discussing piloting the app in the school district.

IV. CONCLUSIONS AND FUTURE WORK

The evaluation of the Reading Writing Tutor demonstrates its potential as an effective, personalized AI driven tool for improving student literacy skills and providing accurate, standards-aligned assessment feedback. A **key conclusion** is the successful implementation and validation of the **novel two-LLM architecture**, where separate Generator and Evaluator LLMs collaboratively provide personalized content and accurate grading. The high agreement rates between the dedicated Evaluator LLM and the human expert for scoring complex responses like TDAs validates this architectural approach and is a significant finding.

The RWT offers a novel solution to addressing challenges in reading and writing education by increasing student engagement through personalization, enabled by the Generator, and potentially alleviating teacher workload by providing automated, individualized practice and feedback, powered by the accurate Evaluator. This **distinct two-LLM design**, particularly when enhanced by a feedback mechanism incorporating human-graded student work, is a core innovation with promising applications in AI-driven education and adaptive learning systems.

Future work includes scaling the application for broader use across the school district.

Bibliography/References:

1. *du Plooy, E., Casteleijn D, and FranzsenD. Personalized adaptive learning in higher education: A scoping review of key characteristics and impact on academic performance and engagement ,Heliyon, Volume 10, Issue 21, 2024, https://doi.org/10.1016/j.heliyon.2024.e39630.* 2. *Open AI Platform*. platform.openai.com/docs/api-reference/introduction.

3. *PA Dept of Education. Pennsylvania Writing Assessment Domain Scoring Guide*. drive.google.com/file/d/1828iJfUU3856Q0u-_Ih0B399R0YIJzMr/view.

4. *Read-aloud and Scribing Guidelines for Operational Assessments*. www.pa.gov/content/dam/copapwp-pagov/en/education/documents/instruction/assessment-and accountability/pssa/accommodations/read%20aloud%20and%20scribing%20guidelines.pdf. Accessed 2025.

5. *Reber R, Canning EA, Harackiewicz JM. Personalized Education to Increase Interest. Curr Dir Psychol Sci. 2018 Dec;27(6):449-454. doi: 10.1177/0963721418793140. Epub 2018 Oct 31. PMID: 31467466; PMCID: PMC6715310.*

6. *US children fall further behind in reading, make little improvement in math on national exam* https://www.cnn.com/2025/01/29/us/education-standardized-test-scores/index.html 7. *Reading and mathematics scores decline during COVID-19 pandemic* https://www.nationsreportcard.gov/highlights/ltt/2022/

8. *Why Are Reading Scores Still Falling on the Nation's Report Card?* https://www.edweek.org/leadership/why-are-reading-scores-still-falling-on-the-nations-report

card/2025/01

9. *PSSA ELA Test Design https://www.pa.gov/content/dam/copapwp
pagov/en/education/documents/instruction/assessment-and-accountability/pssa/test
designs/pcs%20pssa%20english%20language%20arts%20test%20design.pdf*

10. *One Prompt To Rule Them All: LLMs for Opinion Summary Evaluation*
https://arxiv.org/pdf/2402.11683