Proceedings of the Educational Data Mining in Writing and Literacy Instruction Workshop(WLIEDM)

Co-located with the 17th Educational Data Mining Conference (EDM 2024)

Edited by
Collin Lynch[1], Zhikai Gao[1], Damilola Babalola[1],
Piotr Mitros[2], Paul Deane[2]

Organized by
[1] ArgLab, North Carolina State University
[2] Educational Testing Service

July, 2024. Atlanta, Georgia, USA.

Contents

Preface

This volume contains the proceedings of the selected papers of the Educational Data Mining in Writing and Literacy Instruction Workshop (WLIEDM), held on July 14, 2024 at Atlanta, Georgia, USA (2024).

The objective of this workshop is to facilitate discussion among research community around Educational Data Mining (EDM) and AI in Writing and Literacy Education. Moreover, during a tutorial session, a prototype platform developed by the organizers was introduced to the participants. This platform is currently developing to support ethical students' writing and learning data management.

We accepted three research paper and two late-breaking research submissions. Each paper was peer reviewed by our program committee in a double-blinded way and decisions were made based on these reviews, as well as discussions by the workshop organizers.

We would like to thank all the authors for their contributions and the reviewers for their valuable feedback. Special thanks to all the participants for making this workshop a success.

WLIEDM Editors

Organizing Committee

Collin Lynch is an Associate Professor in the Department of Computer Science at North Carolina State University. His primary research is focused on developing robust ITS and adaptive educational systems for Ill-Defined domains such as scientific writing, law, and software development. His current research includes work on argument mining and natural language processing, real-time support for classroom orchestration and writing to learn tasks, advances in student modeling, the development of embodied cognitive agents for collaborative learning, and scaffolding for CS education.

Paul Deane is a principal research scientist in the Research & Development division at ETS. He is the author of Grammar in Mind and Brain, a study of the interaction of cognitive structures in syntax and semantics, and the second author of Vocabulary Assessment to Support Instruction. His current research interests include formative assessment design in the English language arts, cognitive models of writing skills, automated essay scoring, and vocabulary assessment. During his career at ETS, he has worked on a variety of natural language processing (NLP) and assessment projects, including automated item generation, tools to support verbal test development, scoring of collocation errors, reading and vocabulary assessment, and automated essay scoring.

Piotr Mitros is a Senior Research Scientist at ETS. He is also the original author of the popular Open edX learning platform and the original founder as well as the Chief Scientist for more than five years. He has spent the past few years exploring issues around why educational initiatives go south, and evidence-based practices aren't adopted and converged on issues around governance, transparency, and incentive structures. His current work focuses on how we develop educational measurements that incentivize and support rich classroom instruction supporting diverse (rather than standardized) students.

Zhikai Gao is a senior Ph.D. student at North Carolina State University. His current research focuses on understanding students' learning behaviors through traceable log data from ITS, CS education, help-seeking behavior, and LLM usage in education across disciplines.

Damilola Babalola is a second-year Computer Science Ph.D. student at North Carolina State University with a research focus on using Artificial Intelligence (Educational Data Mining and Natural Language Processing) to improve Education. His current work involves research, software development, data mining, and data visualizations aimed at assisting middle-school and high-school students in improving their essay-writing skills. The core of his research centers around the extraction and classification of student essay revisions based on their edit intention, followed by the visualization of student clusters exhibiting similar revision patterns.

Program Committee

Effat Farhana: Vanderbilt Univerisity
Bradley Erickson: ETS
Zuowei Wang: ETS
Rod Roscoe: Arizona State University

Focus Time and Writing Performance in An Intelligent Textbook

Joon Suh Choi
Vanderbilt University
Nashville, Tennessee, USA
joon.suh.choi@vanderbilt.edu

Wesley Morris
Vanderbilt University
Nashville, Tennessee, USA
wesley.morris@vanderbilt.edu

Langdon Holmes
Vanderbilt University
Nashville, Tennessee, USA
langdon.holmes@vanderbilt.edu

Scott Crossley
Vanderbilt University
Nashville, Tennessee, USA
scott.crossley@vanderbilt.edu

ABSTRACT

Intelligent texts promise to make reading materials more interactive and personalized. With the advent of large language models and generative AI, there is potential for intelligent texts to transform the way learners read. Using data collected from a college-level introductory programming course, the present study examines learners' reading behavior and its relationship to read-to-write assessment success while interacting with an intelligent textbook. Reading behavior in the textbook was measured through the amount of time the learner had each portion of the text visible on the webpage (i.e., focus time). To assess links between reading behavior and read-to-write assessments, we examined relationships between focus time and two writing assessments built into the intelligent textbook to measure reading comprehension: constructed responses and summarizations. Results suggest a modest and positive relationship between focus time and assessment success on both read-to-write tasks. These findings suggest that focus time can serve as a useful metric to support personalized learning in intelligent textbooks and to better understand success in read-to-write tasks employed to assess reading comprehension.

Keywords

Intelligent texts, Focus time, Summary writing, Constructed response item, Reading time

1. INTRODUCTION

Textbooks have historically served as valuable resources that enrich the learning experiences of students beyond what a teacher can orally present in a classroom. The digitization of textbooks has provided an alternative to paper texts that is efficient and widely accessible. Recent advances in artificial intelligence (AI) are leading to the next revolution in textbooks: intelligent texts . Intelligent texts include interactive, natural-language-processing-based features that can make the learning process more dynamic, a notable contrast to the passive experience provided by digital and paper-based static texts [1].

In addition to being preferred by students because of their lower cost, convenience, and perceived learning gains [2]-[3], another advantage of intelligent texts is the availability of interaction data that can be collected within them. For instance, the analyses of reading speed and reading patterns have long been an integral part of understanding how learners comprehend texts, but they are difficult to measure in traditional texts. In contrast, intelligent texts can enable a relatively unobtrusive collection of user interaction data related to reading speed and patterns on large scale, allowing the analyses of fine-grained reading behaviors using metrics like scrolling speed and progression maps [1]-[6]. The generation of rich, minable interaction data allows researchers to analyze how users engage with intelligent texts and how reading patterns relate to text comprehension [1].

The present study provides preliminary analyses of how reading patterns within an intelligent text framework are associated with success on read-to-write tasks meant to assess reading comprehension. The framework used in this study is Intelligent Texts for Enhanced Life-long Learning (iTELL) [7]-[8]. This study showcases how focus time metrics (i.e., how long users have access to different parts of a text) relate to users' performance on different interactive, evaluative writing features embedded within iTELL. The goal of the study is to provide insights into how learners' reading behavior within intelligent texts is predictive of success on read-to-write tasks and to provide metrics for developing and deploying interactive intelligent texts that use read-to-write tasks as measures of reading comprehension.

2. BACKGROUND

2.1 Reading Time and Text Comprehension

Studies show that when reading from a screen, readers tend to engage in a reading pattern different from reading printed texts. However, the exact nature of the disparity is not well-established. For example, some studies show that readers tend to read slower when reading from a screen [5], while other studies show that readers pick up their pace when reading from a screen, engaging in a more shallow, fragmented reading behavior powered by skimming and keyword spotting [9]-[11]. There are also conflicting reports on how the different reading behaviors affect reading comprehension. Numerous studies cite that reading printed texts leads to better reading comprehension [12]-[14], while others cite that there are no significant differences in reading comprehension and reading speed when reading digital texts [15]-[18]. There is a wide array of factors that could be the cause of these incongruencies, including the advance of technology that allowed better displays and interfaces, or the myriad of different possible settings in an online environment that could affect the reading patterns, such as font type and spacing [19]-[21].

Regardless of the exact cognitive and meta-cognitive impact brought by the change in reading medium, it is well-established that reading digital texts is connected to a more deliberate and selective reading behavior that often lacks re-reading and is regularly associated with shallow reading [14][22]. Subsequently, it is important that the development of intelligent texts deter shallow reading behavior and promote sustained attention to texts [32].

iTELL is a framework that can facilitate the creation and deployment of intelligent texts with features that are meant to deter fragmented reading behavior and promote interactive reading comprehension. This is done by 1) preventing readers from scrolling through the pages too fast, and 2) checking if readers have read and comprehended portions of a text before they continue to the next part using read-to-write tasks. The former is implemented through a feature that blurs all content and reveals the text chunk by chunk, where a chunk is one or more paragraphs of text delimited by a subheading. The latter is implemented by asking readers to respond to automatically generated questions through written constructed responses for one-third of the chunks and to submit at least one written summary of each page before proceeding to the next page. These two read-to-write tasks serve as checkpoints for evaluating comprehension throughout the reading process. Additionally, iTELL implements highlighting and notetaking features, which are often utilized in printed texts but rarely used when reading electronic texts due to their unwieldiness in most applications [22]-[23]. The following section provides more details on intelligent texts, the iTELL project, and the theoretical bases of their features.

2.2 Read-to-write Tasks

Reading for the specific purpose of performing a writing task facilitates users to engage in a constructive mode of reading where they actively extract information from the text to evaluate and integrate it into their writing [24]-[27]. Instances when users are asked to perform an amalgamated task comprising both reading and writing are dubbed read-to-write tasks. Read-to-write tasks have a long pedagogical history where the task itself is used as a learning tool. Specifically, summarization has long been recognized as an effective read-to-write task that results in higher learning gains [28]-[30]. As well, constructed responses (i.e., written tasks that require students to provide a short answer, usually in a single sentence) can also improve learning comprehension [30] and motivate users to further engage in task-relevant activities such as note-taking [31]. The efficacy of read-to-write tasks, where users are required to actively establish new meaning through extracting and integrating information, can be attributed to the constructivist nature of these tasks.

2.3 Intelligent Texts

Intelligent texts feature “smart” functionalities powered by natural language processing (NLP) and machine learning models, some of which can include instantiations of the read-to-write tasks delineated above. Previously, a major roadblock preventing the utilization of these read-to-write tasks as educational tools was that scoring and providing feedback on summaries manually was a complex and time-consuming endeavor for educators [8]. The development of new powerful large language models (LLMs) has enabled the creation of pipelines that can automatically generate content as well as prompt and evaluate user responses allowing for the automation of scoring and feedback generation in read-to-write tasks in intelligent texts. For example, intelligent texts can use LLMs for automatic question generation [33]-[34] and for automatic evaluation of the written constructed responses that result [35]. For example, researchers have used NLP to analyze users’ responses to free response items and to analyze how users respond to automatically generated questions [33]. Text summarization is also often used as an interactive, read-to-write task embedded in intelligent texts [7].

Intelligent texts can also generate plentiful user interaction data for analyses, which can be used to analyze user behaviors. For example, researchers have logged and analyzed click stream data such as the number of exercise attempts, clicking behaviors, and reading time data to gain insights into students’ learning behaviors [36]. The analysis of such interaction data can be beneficial for predicting course outcomes and serving as an ‘early warning system’ that can flag learners that require additional attention in a timely manner [37].

2.4 Current Study

The present study uses a specific deployment of iTELL to provide insights regarding users’ reading behavior and their learning outcomes as assessed through read-to-write tasks. iTELL generates specific reading time data for individual chunks within a text (i.e., focus time), as well as a diverse set of written data produced by readers. This data includes users’ written responses to constructed response items and summaries of each page. The diverse set of data allows users’ learning outcomes to be operationalized from different aspects, and associations between these outcomes and reading focus can be disentangled.

The present paper is an exploratory study examining the relationship between fine-grained measures of reader attention related to focus time and several metrics based on students read-to-write tasks that measure reading comprehension. The study is driven by the following research questions, which are intended to assess the utility of focus time for future work on personalized learning in intelligent texts:

- RQ1: To what extent are focus time and re-reading focus time correlated with written constructed response item scores and summary scores?
- RQ2: Does increased focus time for a specific chunk predict higher semantic similarity between that chunk and a learner’s summary of the page?

3. METHOD

3.1 Intelligent Text

iTELL is a framework that streamlines the creation and deployment of intelligent texts outfitted with smart functionalities. It features semi-automated pipelines that utilize LLMs with human-in-the-loop to create interactive content such as constructed response items, and scoring APIs for constructed responses and summaries. It is a domain-agnostic framework powered by multiple highly transferable generative LLMs that facilitate the transition of any static texts into interactive, intelligent texts.

iTELL also generates rich clickstream data that allows the analyses of user behaviors, particularly in relation to reading. JavaScript’s intersection observer API is used in the application to discern whether a particular section of a text is within the users’ viewport and logs its observation. This generates focus time data for different parts of texts. iTELL also requires users to answer a constructed response item for one third of language chunks determined at random (at least one constructed response item per page) and write one summary per page, and it uses multiple different fine-tuned LLMs and out-of-the-box LLMs to support these tasks. It generates constructed response items based on static textbook content using GPT 3.5, and uses two separate fine-tuned LLMs to evaluate learners’ constructed responses: Bilingual Evaluation Understudy with

Representations from Transformers (BLEURT) [38] and Masked and Permutated Language Modeling (MPNet) [39]. It also uses KeyBART [40] and Chat GPT to extract keyphrases from texts, and two separate fine-tuned Longformer models [41] to score learners' summaries. The focus time data in tandem with the constructed response item data and summary data for individual users can be used to analyze how users' reading time relates to their success on read-to-write assessments of text comprehension.

3.2 Participants

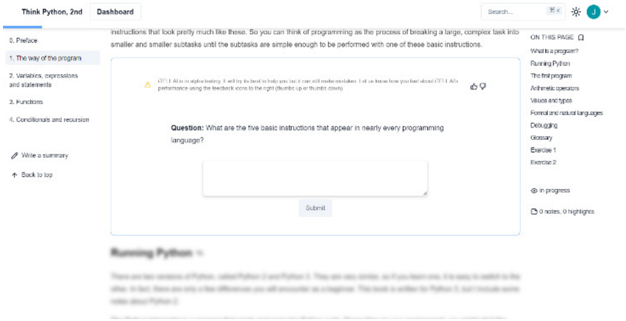
We recruited participants from an introductory Python programming class at a public university in Georgia. Students were offered extra credit if they completed four chapters (i.e., four pages where each chapter of the textbook was adapted into a single, scrollable page) of the *Think Python* textbook[1] [42] that were adapted into an intelligent text through iTELL. Students were given three weeks to complete this task at their own pace. Students who opted in were asked to complete an intake survey, and those over the age of 18 were asked for their consent to having their data used in the study. Out of the 139 participants, 98 users indicated that they were over the age of 18 and consented to the use of their data. From these participants, we collected 1,777 responses to constructed response items, 1,121 summaries, and 8,497 event data (e.g., button clicks and scrolls). The majority of the participants were between the ages of 18 to 24 (89.8%) and were either native or bilingual speakers of English (83.67%). All students had prior experience in Python programming. See Table 1 below for more information on the demographic data.

Table 1. Demographic Information

	n	%
Age		
18-24 years old	88	89.80
35-44 years old	4	4.08
Did not specify	6	6.12
Ethnicity		
Asian or Pacific Islander	41	41.84
White or Caucasian	19	19.39
Hispanic or Latino	10	10.20
Prefer not to say	10	10.20
Black or African American	10	10.20
Other	3	3.06
Did not specify	5	5.10
Education		
High school graduate, diploma, or the equivalent	44	44.90
Some college credit, no degree	27	27.55
Some high school, no diploma	16	16.33
Bachelor's degree	2	2.04
Nursery school to 8th grade	2	2.04
Master's degree	1	1.02
No schooling completed	1	1.02
Did not specify	5	5.10
English proficiency		
Native/Bilingual proficiency	82	83.67
Full professional proficiency	8	8.16
Professional working proficiency	3	3.06
Did not specify	5	5.10

3.3 Focus Time Measures

The amount of time participants spent reading each was logged by iTELL based on chunks. Chunks are selected by developers and are a segment of a text comprising a full idea that includes at least one paragraph that is under a single subheading of the textbook. Logging was done using a JavaScript API and iTELL's chunk-blurring feature. Upon loading a new page, only the first chunk of the page was completely visible to a new user; other chunks were blurred out using Cascading Style Sheets (See Figure 1). At the end of each chunk, there was a 33% chance that a constructed response item would appear. To move on to the next chunk, the user had to answer the constructed response item (if present) or click on a "continue reading" button. The time each chunk was visible on a user's viewport was logged as the amount of time the user had spent paying attention to the particular chunk (i.e., focus time). Focus time metrics were normalized using word counts and outliers were removed from the data (|z| > 3). See Table 2 below for descriptive statistics, and Figure 2 for a visualization of the viewing trend.



instructions that look pretty much like these. So you can think of programming as the process of breaking a large, complex task into smaller and smaller subtasks until the subtasks are simple enough to be performed with one of these basic instructions.

iTELL AI is in alpha testing. It will try its best to help you but it can still make mistakes. Let us know how you feel about iTELL AI's performance using the feedback icons to the right (thumbs up or thumbs down).

Question: What are the five basic instructions that appear in nearly every programming language?

Running Python

There are two versions of Python, called Python 2 and Python 3. They are very similar, so if you learn one, it is easy to switch to the other. In fact, there are only a few differences you will encounter as a beginner. This book is written for Python 3, but I include some notes about Python 2.

Figure 1. Blurred chunks in iTELL

Table 2 shows that users started out by spending around two and a half minutes on each chunk on average for the first page, and the amount of time spent on each chunk steadily declined as the user progressed through the pages, ending with around a minute and twenty seconds spent on each chunk by the time users were on the last page. The viewing trend in Figure 2 shows the users' focus time data for each chunk in a page normalized into percentages (e.g., for pages with 10 chunks, the focus time data for each chunk is marked at the 0.1 mark on the x-axis). This was done to align the viewing trend for pages that had different numbers of chunks. The focus time trend shows that users tend to spend more time at the beginning and end of each page, with the exception of page one where there is also a notable uptick in focus time in the middle portion of the page. The increased focus time at the end of the page is most likely due to the summary module being located at the end of the page and the users spending additional time writing summaries with part of the chunk within their viewport (see next section).

3.4 Summary Writing

Users were required to write a summary after reading each page using a summary module located at the bottom of the page. The summary was meant to be a read-to-write task that assessed reading comprehension. All submitted summaries were first input through a first-pass filter that checked the minimum length (50 words) and the language of the summary (whether it was written in English). Summaries filtered out during the first-pass were returned to the user, and the user was asked to rewrite and resubmit their summary. Summaries that passed were scored by different pipelines that

[1] https://greenteapress.com/wp/think-python-2e/

utilize NLP tools including SpaCy,**Error! Reference source not found.** a Doc2vec tokenizer, and fine-tuned Longformer LLMs [41].

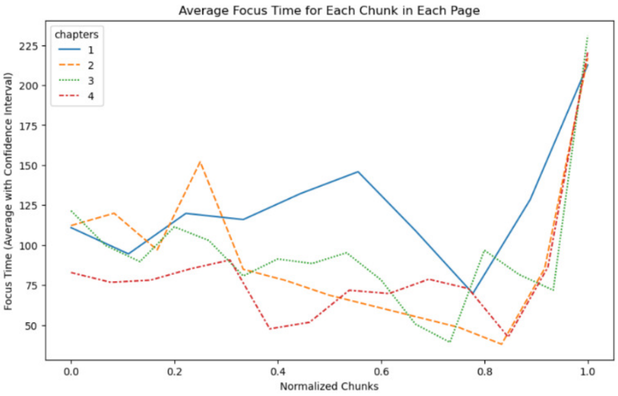


Figure 2. Average Focus Time for Each Page

Table 2. Average Focus Time (seconds) per Page and Chunk

P	C	M	Std	P	C	M	Std
1	0	110.902	112.554	2	0	112.204	129.339
1	1	94.482	85.24	2	1	119.906	117.225
1	2	119.814	125.421	2	2	96.965	89.011
1	3	115.964	108.512	2	3	152.07	141.873
1	4	132.246	120.621	2	4	84.845	108.289
1	5	145.863	122.694	2	5	77.75	79.429
1	6	109	110.355	2	6	68.638	95.276
1	7	69.755	68.16	2	9	48.603	84.271
1	*8	128.549	124.717	2	*10	38	46.517
1	*9	212.659	150.501	2	*11	85.31	102.491
				2	*12	216.98	142.928
3	0	121.455	136.456	4	0	82.818	85.013
3	1	99.81	96.489	4	1	76.719	90.085
3	2	89.644	79.822	4	2	78.07	95.222
3	3	111.4	110.17	4	3	85.125	112.154
3	4	102.864	86.448	4	4	90.69	96.585
3	5	80.712	77.734	4	5	47.621	56.226
3	6	91.288	88.587	4	6	51.724	72.132
3	7	88.483	109.568	4	7	71.741	78.838
3	8	95.2	99.638	4	8	69.684	77.608
3	9	78.217	78.375	4	9	78.649	90.387
3	10	50.55	44.741	4	10	72.793	76,866
3	11	39.2	47.282	4	11	42.483	55.548
3	*12	96.763	124.987	4	*12	86.263	108.495
3	*13	81.655	117.726	4	*13	220.638	145.097
3	*14	71.772	100.602				
3	*15	230.891	151.907				

P: page; C: chunk; M: mean; *Asterisks demark chunks without accompanying constructed response items. These chunks include glossaries and coding exercises.

The summaries were scored on their content (whether the summary includes key ideas and details from the textbook), wording (whether the summary paraphrases words and sentences from the textbook using objective language), relevance (whether the summary stays on topic), and language borrowing (whether the summary contains the users' own language different from the textbook). A Spacy-based pipeline analyzing trigram overlap between the source text and the summary was used to evaluate language borrowing. A Doc2vec model was used to evaluate semantic similarity between the source text and the summary, as a measure of relevance. Finetuned Longformer models were used to evaluate wording and content scores. Preliminary assessments of the scoring models indicated that they explain 79% and 66% of the score variance for content and wording metrics (refer to [44] for more details on the derivation and testing of the summary scoring models for content and wording scores). Language borrowing scores ranged between 0 to 1, with 1 indicating that the set of trigrams in the student's summary was identical to the set of trigrams in the source text. Relevance scores were cosine similarity scores ranging between 0 to 1, with score closer to one indicating stronger relevance. Wording and content models mapped the input summary onto the training set summaries' z scores, meaning that the score typically fell in the range of -3 to 3, with 3 indicating exceptionally high performance.

Summaries that received passing scores for all four criteria were marked as passing summaries. The thresholds for each criterion, developed through alpha testing of the tool, were as follows: below 0.6 for language borrowing, above 0.5 for relevance, above -1 for wording, and above 0 for content. Users who submitted a passing summary were prompted to proceed to the next page. When submitting a failed summary, the user was provided auto-generated feedback on the specific criteria in which the summary failed along with suggestions on keywords to include in summary revisions. Users were allowed to proceed to the next page after submitting two consecutive failed summaries.

To address RQ1, which focuses on the relationship between focus time and summary scores, we examined users' initial attempt at a summary for each page. After removing cases in which focus times were not correctly logged (i.e., where missing focus time data points existed due to technical issues; N=118) and removing all summaries that were not users' first attempt (N=779), we were left with 224 summaries written by 60 unique users. These initial summary submissions were also used for RQ2, which involves the analysis of reading time and the semantic similarity between chunks and summaries. Table 3 below shows the descriptive statistics for the summary scores.

Table 3. Summary First Attempts Descriptive Statistics

Chapter	Count	Containment		Similarity	
		Mean	SD	Mean	SD
1	52	0.035	0.052	0.491	0.055
2	58	0.034	0.038	0.56	0.064
3	58	0.056	0.066	0.534	0.08
4	56	0.061	0.058	0.506	0.067
Chapter	Count	Wording		Content	
		Mean	SD	Mean	SD
1	52	0.401	0.385	-0.067	0.494
2	58	0.402	0.386	0.097	0.515
3	58	0.234	0.381	-0.285	0.366
4	56	0.179	0.359	-0.077	0.425

* For all sub-scores except for containment scores, which is inversely scored, a higher value denotes better performance. The range of each sub-score varies from -1 to 1

To address the second part of RQ1, which requires the analysis of the relationship between re-reading and summary scores, we looked at the relationship between the elapsed focus time between a user's failed attempt and their second attempt. A total of 124 second submissions were made after failed summaries; these summaries were used for the re-reading time analysis. In this analysis, re-reading was defined as scrolling upwards more than 3% of the page's content.

3.5 Constructed Response Items

The iTELL *Think Python* deployment required users to answer at least one constructed response item per page. The number of constructed response items for each user varied, with each chunk spawning a constructed response item 1 out of 3 times. Each page had at least one chunk with an accompanying constructed response item.

The iTELL pipeline uses GPT-3.5 to generate constructed response items for each chunk and corresponding correct answers to each of the generated items (reference answers) pre-deployment. These questions are checked by textbook developers for accuracy, Initial accuracies between human raters and GPT-3.5 indicated 100% agreement on a limited sample size (N = 60) [45]. The reference answers served a similar role as the answers provided in an answer key (i.e., the correct, reference answer to use when evaluating a student's response). Two LLMs, Bilingual Evaluation Understudy with Representations from Transformers (BLEURT) [38] and Masked and Permutated Language Modeling (MPNet) [39], were finetuned on the Multi-Sentence Reading Comprehension (MultiRC) dataset to develop scoring models for iTELL. The fine-tuned models were developed to predict whether a constructed response to a question was correct or incorrect by comparing the response's similarity to the correct, reference answer. In terms of model accuracy, MPNET reported an accuracy of 0.81 and BLEURT reported an accuracy of 0.79 [45]. Within iTELL, if both models agreed that the participant submitted a passing response, the participant was prompted to unblur the next chunk and proceed. If the models disagreed, the participant was allowed to proceed but was also allowed to submit another response. If both models evaluated the response as a failing response, the participant was prompted to resubmit a response. Participants were allowed to ignore the feedback in iTELL and proceed to the next chunk if they believed the response was erroneous.

For this study, we selected each user's first attempt for each constructed response item. After removing all user responses to constructed response items that were not the users' first attempt (N=657), and removing cases in which focus times were not correctly logged (N=548), we were left with 561 responses provided by 64 unique users. See Table 4 below for the constructed response items' descriptive statistics.

Table 4. Constructed response items descriptive statistics

Chapter	Count	Mean	SD
1	153	1.562	0.724
2	94	1.479	0.813
3	180	1.489	0.794
4	134	1.336	0.909

3.6 Statistical Analysis

We normed the focus time data using word count and removed any outliers (|z| > 3). The first part of RQ1 asked about the correlation between focus time and summary scores. We conducted correlation analyses to assess whether students' engagement with the iTELL *Think Python* deployment (represented by students' focus time) was related to student performance on read-to-write assessment scores (i.e., summary scores and constructed response item scores). Spearman's rho was used to account for the violation of normality and for the ordinal nature of data like constructed response item scores and survey responses. The second part of RQ1 asked about the correlation between re-reading focus time and summary scores. We conducted another set of correlation analysis between the elapsed re-reading focus time between users' summary submissions and their summary scores. To control for familywise error rate when making multiple comparisons, we used a Holm-Bonferroni correction to maintain the nominal alpha level of 0.05. We hypothesized that students who exhibited more engagement through increased focus time would perform better on summaries and constructed response items.

To address RQ2, whether increased focus time for a specific chunk predicts higher semantic similarity between that chunk and a learner's summary of the page, we used a sentence transformer model (all-MiniLM-L6-v2) to derive cosine similarities between users' summaries and each individual chunk in the source page. We then conducted a correlation analysis between the amount of time spent on each chunk and the respective cosine similarity score. We hypothesized that chunks in which a user has spent a longer time reading (i.e., chunks with a higher focus time) would tend to have higher semantic similarity with the summary produced by the user.

4. RESULTS

4.1 Focus Time and Summaries

Correlation analyses using Spearman's rho showed that there were weak (rho < 0.3) [46] positive correlations between focus time and two of the summary scoring criteria: relevance scores (rho = 0.253, p < 0.001) and content scores (rho = 0.268, p < 0.001), meaning that users who spent more time engaged with iTELL's *Think Python* deployment had a tendency to write summaries that stayed more on topic (relevance) and included key ideas and details from the textbook (content). There were no significant correlations reported for the remaining wording and language borrowing, indicating no significant relationships between focus time and whether the summary used objective words and phrases (wording) or whether the summary used original language (language borrowing).

Subsequent correlation analyses were conducted between the amount of time users spent re-reading a text after submitting a failed summary and their second summary scores. There were significant positive correlations reported between the re-reading focus time and relevance score (rho = 0.239, p < 0.001) and language borrowing score (rho = 0.188, p = 0.015), and a significant negative correlation was found between re-reading focus time and wording scores (rho = -0.173, p < 0.001).

We also conducted correlation analyses between the similarity scores of each summary as derived from the sentence transformer model and chunk and focus time specific to each chunk. The results showed that there was a weak correlation (rho = 0.120, p < 0.001) between focus time and similarity scores, indicating that users who spent more time reading a chunk produced a summary that was contextually similar to that chunk.

4.2 Focus Time and Constructed Response Items

A correlation analysis using Spearman's rho showed that there was a small correlation (r = 0.109, p < 0.001) between focus time and constructed response item scores, meaning that users who spent more time engaged with the *Think Python* iTELL deployment had a slight tendency to score better on constructed response items. There was no significant correlation found between the amount of time spent re-reading chunks after a student failed a constructed response item, and their scores at reattempts.

5. DISCUSSION

This study presented preliminary analyses of user interaction data collected from an iTELL *Think Python* deployment. Specifically, the analyses focused on whether the amount of time users spent reading the text (i.e., focus time) was related to their performance on constructed response items and summaries, and to their impressions on engaging with the intelligent text.

The general trend in focus time data for chunks visualized in Figure 2 showed an uptick at the end of each page, which is most likely due to the summary module being located at the end of the page. The trend also showed that users spent more time on the beginning of each page, and gradually increased their reading speed. This trend of acclimation was also shown on a more macro level: users spent more time reading the first page, but their reading speed increased gradually, and the learners were spending less time on each page by the time they were on the last page.

The results of the summary analyses showed that there was a weak correlation between focus time and summary scores. Analyses of the specific scoring criteria for summaries revealed more details. The correlation between relevance score and focus time indicated that users who spent more time reading the text were more likely to write summaries that are similar to the source text. This means that staying engaged with the source text for a longer period of time was related to users producing summaries relevant to the content of the text. The correlation between focus time and content scores indicated that longer exposure to source texts was related to a more accurate expression of the main ideas and details within the source text. No correlation was reported between focus time and wording, meaning that longer engagement with the source text did not necessarily translate to the users' summaries containing objective and original language beyond the source text. Additionally, no significant correlation was reported between focus time and language borrowing scores, indicating that longer reading time was not related to users borrowing specific trigrams from the source text. Taken together, the analyses suggest that users who spent more time engaged with the source texts showed a tendency to write better summaries that captured the main ideas of the source text. However, longer focus time did not result in lower language borrowing scores or wording scores, indicating that more reading time did not necessarily translate to the use of more original language in the summary.

The analysis of the re-reading focus time and summary scores showed that as the re-reading time increased, readers' submitted summaries tended to have higher relevance and language borrowing scores, but decreased wording scores. This indicated that as learners spent more time re-reading pages after submitting failed summaries, they tended to write summaries that were more contextually similar to the original text, had more overlapping trigrams with the original text, and used less objective language. This indicates that learners' language used in their summaries more closely resembled the language of the source text as they engaged in re-reading and submitted additional summaries.

Comparing the embeddings of the summaries and users' focus time for each individual chunk showed that there was a relationship between focus time and chunk similarity. In other words, users tended to write summaries that were similar to the chunks that they had spent more time reading.

The analysis of constructed response items revealed that there was a correlation between focus time and constructed response item scores. The correlation analysis showed that users who spent more time reading passages exhibited a slight tendency to score better at constructed response items. While the correlation was significant, the magnitude of the relationship between the two variables did not result in a significant difference between scores in an ANOVA.

The analyses of focus time data in general confirmed that the intelligent text was behaving as expected. Learners who spent more time engaged with the text tended to score better on constructed response items and tended to write summaries that were more relevant and better captured the main ideas of the source text. The results also showed that intelligent texts have the potential to engage users in re-reading, which was one of the significant deficits of the reading behavior shown when reading digital texts compared to print texts [14]. The caveat is that the analyses also showed that re-reading may not necessarily denote that users are truly, cognitively re-engaged with the text. Re-reading may remain selective and shallow if it is carried out for the specific purpose of completing a task (i.e., writing a summary), as suggested by the correlation between re-reading time and the increase in language borrowing score and decrease in wording score in summaries. These results show the value of collecting and analyzing fine-grained focus time data through intelligent texts. However, they also suggest that additional features could be implemented to further scaffold user experience and help them engage in improved reading patterns.

Overall, the results showcased the capacity of intelligent texts to track users' focus time data and reading behaviors in real-time and showed the potential for such features to be implemented as part of a read-to-write pipeline that can provide learners and teachers with timely focus time data accompanied by actionable feedback.

6. CONCLUSION

The present study showcases the capacity of intelligent texts to generate fine-grained data for reading behavior analyses through read-to-write tasks and provides insights about users' reading behavior specific to intelligent texts and their relationship to user performance on read-to-write evaluative features such as summary writing and constructed response items. The results show that while focus time is significantly related to user performance on these features, suggesting that users who spend more time engaged with intelligent texts tend to gain a better understanding of the content, but increases in focus time does not guarantee that users will engage in original thinking that materializes in the use of original words and phrasings in summaries. These results suggest that intelligent texts should be outfitted with features that will help users engage in original thinking while keeping them engaged with the text.

However, the study also has several limitations and room for future improvement. First, iTELL is in its testing phase and the application's focus-time logging feature was not mature at the time of iTELL *Think Python*'s deployment, causing the loss of numerous focus time data points. This resulted in the discrepancy in the number of users for different analyses. This issue of code maturity has been resolved as iTELL gone through additional rounds of testing, Second, reading using chunks is rate-limiting and removes the opportunity for readers to naturaly skim or scan texts, so it might not reflect natural reading patterns. Third, the efficacy of intelligent texts needs to be analyzed in a randomized control trial where some users are assigned texts that do not contain interactive features. Fourth, while iTELL collects fine-grained focus time data, it remains that the data is a proxy of the actual cognitive and metacognitive reading process. As evidenced by the analysis of re-reading focus time and summary scores, the focus time data must be supplemented with other collected data and must be contextualized for appropriate analysis.

Another area for improvement identified in the analyses is that a feature that scaffolds users' re-reading experience is necessary. To address this issue, new features that support strategic think-aloud [47] are being developed for iTELL. These new features will be powered by generative LLMs and will be used to support a more personalized re-reading experience for users. If users exhibit signs of requiring additional reading support (e.g., if they submit failing summaries), a conversational agent supporting these features will prompt users to re-read specific chunks of an iTELL text and engage in written self-explanations and think-alouds taking conversational turns, creating a more tailored, personalized experience for users.

7. ACKNOWLEDGEMENTS

This material is based upon work supported by the National Science Foundation under Grant 2112532 Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

8. REFERENCES

[1] Brusilovsky, P., Sosnovsky, S., & Thaker, K. (2022). The return of intelligent textbooks. *AI Magazine, 43*(3), 337-340.

[2] Ji, S. W., Michaels, S., & Waterman, D. (2014). Print vs. electronic readings in college courses: Cost-efficiency and perceived learning. *The Internet and Higher Education, 21*, 17-24.

[3] Chulkov, D. V., & VanAlstine, J. (2013). College student choice among electronic and printed textbook options. *Journal of Education for Business, 88*(4), 216-222.

[4] Dyson, M., & Haselgrove, M. (2000). The effects of reading speed and reading patterns on the understanding of text read from screen. *Journal of research in reading, 23*(2), 210-223.

[5] Dillon, A. (1992). Reading from paper versus screens: A critical review of the empirical literature. *Ergonomics, 35*(10), 1297-1326.

[6] Hornbæk, K., & Frøkjær, E. (2003). Reading patterns and usability in visualizations of electronic documents. *ACM Transactions on Computer-Human Interaction (TOCHI), 10*(2), 119-149.

[7] Anonymized authors (2023). *Anonymized source.*

[8] Anonymized authors (in press). *Anonymized source.*

[9] Horton, W., Taylor, L., Ignacio, A., & Hoft, N. L. (1995). *The Web page design cookbook: all the ingredients you need to create 5-star Web pages*. John Wiley & Sons, Inc.

[10] Dyson, M., & Haselgrove, M. (2000). The effects of reading speed and reading patterns on the understanding of text read from screen. *Journal of research in reading, 23*(2), 210-223.

[11] Levy, D. M. (1997, July). I read the news today, oh boy: reading and attention in digital libraries. In *Proceedings of the second ACM international conference on Digital libraries* (pp. 202-211).

[12] Mangen, A., Walgermo, B. R., & Brønnick, K. (2013). Reading linear texts on paper versus computer screen: Effects on reading comprehension. *International journal of educational research, 58*, 61-68.

[13] Ben-Yehudah, G., & Eshet-Alkalai, Y. (2021). Print versus digital reading comprehension tests: does the congruency of study and test medium matter?. *British Journal of Educational Technology, 52*(1), 426-440.

[14] Jian, Y. C. (2022). Reading in print versus digital media uses different cognitive strategies: evidence from eye movements during science-text reading. *Reading and Writing, 35*(7), 1549-1568.

[15] Alisaari, J., Turunen, T., Kajamies, A., Korpela, M., & Hurme, T. R. (2018). Reading comprehension in digital and printed texts. *L1-Educational Studies in language and literature, 18*, 1-18.

[16] Jeong, Y. J., & Gweon, G. (2021). Advantages of print reading over screen reading: A comparison of visual patterns, reading performance, and reading attitudes across paper, computers, and tablets. *International Journal of Human–Computer Interaction, 37*(17), 1674-1684.

[17] Daniel, D. B., & Woody, W. D. (2013). E-textbooks at what cost? Performance and use of electronic v. print texts. *Computers & education, 62*, 18-23

[18] Margolin, S. J., Driscoll, C., Toland, M. J., & Kegler, J. L. (2013). E-readers, computer screens, or paper: Does reading comprehension change across media platforms? *Applied cognitive psychology, 27*(4), 512-519.

[19] Tullis, T. S., Boynton, J. L., & Hersh, H. (1995, May). Readability of fonts in the windows environment. In *Conference companion on Human factors in computing systems* (pp. 127-128).

[20] Hojjati, N., & Muniandy, B. (2014). The effects of font type and spacing of text for online readability and performance. *Contemporary Educational Technology, 5*(2), 161-174.

[21] Rello, L., Pielot, M., & Marcos, M. C. (2016, May). Make it big! The effect of font size and line spacing on online readability. In *Proceedings of the 2016 CHI conference on Human Factors in Computing Systems* (pp. 3637-3648).

[22] Liu, Z. (2005). Reading behavior in the digital environment: Changes in reading behavior over the past ten years. *Journal of documentation, 61*(6), 700-712.

[23] Mizrachi, D. (2015). Undergraduates' academic reading format preferences and behaviors. *The journal of academic librarianship, 41*(3), 301-311

[24] Delaney, Y. A. (2008). Investigating the reading-to-write construct. *Journal of English for academic purposes, 7*(3), 140-150.

[25] Grabe, W., & Stoller, F. L. (2019). *Teaching and researching reading*. Routledge.

[26] Nelson, N., & Calfee, R. C. (1998). Chapter I: The Reading-Writing Connection Viewed Historically. *Teachers College Record, 99*(6), 1-52.

[27] Nelson, N., & King, J. R. (2023). Discourse synthesis: Textual transformations in writing from sources. Reading and Writing, 36(4), 769-808.

[28] Silva, A. M., & Limongi, R. (2019). Writing to learn increases long-term memory consolidation: A mental-chronometry and computational-modeling study of "Epistemic writing". Journal of Writing Research, 11(1), 211-243.

[29] Brown, A. L., Campione, J. C., & Day, J. D. (1981). Learning to learn: On training students to learn from texts. *Educational researcher, 10*(2), 14-21.

[30] Bensoussan, M., & Kreindler, I. (1990). Improving advanced reading comprehension in a foreign language: summaries vs. short-answer questions. *Journal of Research in Reading, 13*(1), 55-68.

[31] Szpunar, K. K., Khan, N. Y., & Schacter, D. L. (2013). Interpolated memory tests reduce mind wandering and improve learning of online lectures. Proceedings of the National Academy of Sciences, 110(16), 6313-6317.

[32] Birkerts, S. (2006). *The Gutenberg elegies: The fate of reading in an electronic age*. Farrar, Straus and Giroux.

[33] Dijkstra, R., Genç, Z., Kayal, S., & Kamps, J. (2022). Reading comprehension quiz generation using generative pre-trained transformers [Paper presentation]. In *The 23rd International Conference on Artificial Intelligence in Education (AIED).*

[34] Van Campenhout, R., Clark, M., Jerome, B., Dittel, J. S., & Johnson, B. G. (2020). Advancing Intelligent Textbooks with Automatically Generated Practice: A Large-Scale Analysis of Student Data. *Proceedings of the Fifth Workshop on Intelligent Textbooks*. AIED, 2023.

[35] Asakura, T., Nguyen, H. T., Truong, N. T., Ly, N. T., Nguyen, C. T., Miyazawa, H., ... & Nakagawa, M. (2023). Digitalizing educational workbooks and collecting handwritten answers for automatic scoring. *Proceedings of the Fifth Workshop on Intelligent Textbooks*. AIED, 2023.

[36] Heo, S., Farghally, M. F., Mohammed, M., & Shaffer, C. A. (2023). Creating Session Data from eTextbook Event Streams. *Proceedings of the Fifth Workshop on Intelligent Textbooks*. AIED, 2023.

[37] Junco, R., & Clem, C. (2015). Predicting course outcomes with digital textbook usage data. *The Internet and Higher Education, 27*, 54-63.

[38] Sellam, T., Das, D., & Parikh, A. P. (2020). *BLEURT: Learning robust metrics for text generation*. arXiv preprint arXiv:2004.04696.

[39] Song, K., Tan, X., Qin, T., Lu, J., & Liu, T. Y. (2020). Mpnet: Masked and permuted pre-training for language understanding. *Advances in Neural Information Processing Systems, 33*, 16857-16867.

[40] Diya, A., & Mizuho, I. (2022). Keyphrase generation by utilizing BART finetuning and BERT-based ranking. In *DEIM Forum.*

[41] Beltagy, I., Peters, M. E., & Cohan, A. (2020*). Longformer: The long-document transformer*. arXiv preprint arXiv:2004.05150.

[42] Downey, A. B. (2015). *Think Python* (Second edition, December 2015). O'Reilly.

[43] Honnibal, M., & Montani, I. (2017). spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing.

[44] Morris, W., Crossley, S., Holmes, L., Ou, Chaohua, Dascalu, M., & McNamara, D. (in press). Formative Feedback on Student-Authored Summaries in Intelligent Textbooks using Large Language Models. *Journal of Artificial Intelligence in Education*.

[45] Morris, W., Choi, J., Holmes, L., Gupta, V., & Crossley, S. A. (in press). Automatic Question Generation and Constructed Response Scoring in Intelligent Texts. *Proceedings of the 17th International Conference on Educational Data Mining.*

[46] Dancey, C. P., & Reidy, J. (2007). *Statistics without maths for psychology*. Pearson education.

[47] McNamara, D. S. (2004). SERT: Self-explanation reading training. Discourse processes, 38(1), 1-30.

Automatic Essay Scoring for K-12 Writing Assignments with Open Source LLM

Zhikai Gao
NC State University
zgao9@ncsu.edu

Naga Buddarapu
NC State University
vbuddar@ncsu.edu

Damilola Babalola
NC State University
djbabalo@ncsu.edu

Collin Lynch
NC State University
cflynch@ncsu.edu

Paul Deane
ETS
pdeane@ets.org

Piotr Mitros
ETS
piotr@mitros.org

ABSTRACT

Automated Essay Scoring (AES) is critical for handling the challenges in evaluating student essays amid rising teacher-student ratios. And the emergence of LLM and generative AI provide additional practical methodology in AES. To address privacy concerns associated with closed-source models like ChatGPT, this paper explores the potential of open-source Large Language Models (LLMs), specifically Llama2-7B,Llama3-8B, and Mistral-7B, in grading K-12 students' essays across disciplines. Our results show improved accuracy in essay scoring for LLMs compared to the baseline model. However, further refinement, including fine-tuning the Llama3-8B model, is proposed for enhanced practical utility.

Keywords

LLM, Automatic Essay Scoring, K-12,Writing Analytic

1. INTRODUCTION

Automated Essay Scoring (AES) systems have become increasingly vital in the education sector, addressing the growing challenge of evaluating student essays amidst rising teacher-student ratios[5]. AES offers significant opportunities such as scalability and efficiency in grading, consistency and objectivity in scores, insights for personalized learning, seamless integration with educational technology platforms, and driving innovation in language learning and assessment. These systems have transitioned from basic rule-based mechanisms to Machine Learning (ML) technologies and Large Language Models (LLMs), offering a scalable and efficient solution for providing timely feedback and maintaining consistent evaluation standards[3].

The rapid maturation of Generative AI has created the possibility of using large language models (LLMs) to assess student writing. Multiple researchers have done experiments and proved the ability of GPT models in AES.[7, 8, 4] However, the close source nature of GPT models raises concerns about data privacy and security. On the other hand, running an open-source large language model in a closed and secure environment can prevent students' data leakage, restrict access to the model, and solve the privacy problem. Therefore, in this research, we aim to explore the ability of two state-of-the-art open-source models to grade K-12 students' essays across multiple disciplines. And we try to answer the following research question:

- RQ: How accurately can open-source LLM perform in predicting K-12 students' essays across different disciplines?

1.1 Prior Work

There are multiple methodologies applied to address the AES problem. For example, Zhang and Liu[9] explore the evolution of Deep-Neural Network (DNN)-based AES systems, transitioning from reliance on handcrafted features to utilizing advanced DNN models, such as Bidirectional Encoder Representations from Transformers (BERT). This evolution enhances semantic understanding and accuracy in scoring, though it acknowledges challenges in cross-domain and cross-language tasks, suggesting future research directions toward a more holistic AES approach. Ludwig et al.[3]study reveals performance metrics for three sentiment analysis models: Logistic Regression, German BERT (June 2019), and German BERT (Oct. 2020). The larger transformer model (German BERT, Oct. 2020) outperforms the others with an accuracy of 94% and Cohen's Kappa of 0.59. Cozma et al. [1]. presented a novel AES framework that integrates string kernels and word embeddings, demonstrating substantial advancements over previous models. The method achieved a leading average QWK(quadratic weighted kappa) of 0.785 in in-domain settings and showed notable superiority in cross-domain evaluations, significantly outperforming the earlier state-of-the-art with QWK improvements.

With the emergence of LLM and ChatGPT. Many researchers applied the GPT models to the AES tasks and proved their accuracy. For instance, Xia et al.[7] shows that ChatGPT achieved an overall accuracy of 84.375% in predicting scores for various theme types in TOEFL Independent Writing tasks, with correct predictions for 27 out of 32 articles. Xiao et al.[8] experimental results for Automated Essay Scoring (AES) tasks using large language models (LLMs) indicate that fine-tuned GPT-3.5 consistently outperforms the BERT baseline and other LLM-based methods across subsets of the ASAP dataset. QWK scores demonstrate the superior accuracy of fine-tuned GPT-3.5, ranging from 0.7406 to 0.8593 on the ASAP dataset and reaching 0.7806 in an ensemble setting. While GPT-4's zero-shot and few-shot capabilities show limited success, fine-tuned GPT-3.5 exhibits notable improvements, sometimes surpassing the generalization performance of GPT-4. These findings underscore the efficacy of fine-tuned LLMs, particularly GPT-3.5, in achieving high performance in AES tasks and suggest their potential for automated essay scoring applications.

This study presents an innovative exploration of employing open-source Large Language Models (LLMs) as opposed to proprietary online models, such as ChatGPT, within the framework of Automated Essay Scoring (AES). It focuses on evaluating the scoring consistency and stability provided by these models, with the objective of identifying the advantages and drawbacks of utilizing open-source solutions versus their proprietary online equivalents. This comparative analysis is crucial, as it could significantly influence the creation of AES systems that are more accessible, fair, and tailored to diverse educational needs. By doing so, the research seeks to contribute to educational technology advancements, shedding light on the efficacy of varying model types in delivering accurate and unbiased essay assessments.

2. DATASET

We collected over ten thousand K-12(primarily 6th-8th grade) students' essays from ten different writing assignments across multiple disciplines(e.g. biology, social science, etc.) and years(from 2009 to 2017). The details about the writing assignments are listed in Table 1.

2.1 Human grading and rubric

Each essay was graded by two separate teachers based on the same set of provided rubrics, and the average score was the final human-graded score. Each essay is evaluated based on two rubrics: Feature rubric and Argument rubric; each rubric has a scale of 0-5, and the final score is the summation of the scores based on two rubrics. Here is the Argument rubric:

- 0 point(Off-topic): Consists entirely of source language, is completely off-topic, or consists of random keystrokes
- 1 point(Minimal): A MINIMAL response displays little or no ability to construct an argument. For example, there may be no claim, no relevant reasons and examples, no development of an argument, little logical coherence throughout the response, or mainly use of source language.
- 2 points(developing low): A DEVELOPING LOW response displays problems that seriously undermine the writer's argument, such as a confusing or inconsistent claim, a seriously underdeveloped or unfocused argument, or inappropriate content or tone throughout much of the response.
- 3 points(developing high): While a DEVELOPING HIGH response displays some competence, it typically has at least one of the following weaknesses: a vague claim; somewhat unclear, limited, or inaccurate use of evidence; failure to take account of the alternative; noticeable reliance on source language; simplistic reasoning; or occasionally inappropriate content or tone for the audience.
- 4 points(CLEARLY COMPETENT): The response demonstrates a competent grasp of argument construction and the rhetorical demands of the task by displaying all or most of the following characteristics in three aspects:(1)Command of Argument Structure,(2)Quality and Development of Argument, and (3)Awareness of audience
- 5 points(EXEMPLARY): An EXEMPLARY response meets all of the requirements for a score of 4 points and distinguishes itself with such qualities as insightful analysis (recognizing the limits of an argument, identifying possible assumptions and implications of a particular position); or the skillful use of rhetorical devices, phrasing, voice, and tone to engage the reader and thus make the argument more persuasive or compelling.

And here is the Feature rubric:

- 0 points(No Credit): Not enough of the student's own writing for surface-level features to be judged; not written in English; completely off topic; or random keystrokes.
- 1 points(Minimal): A response in this category differs from Developing Low responses because of serious failures such as extreme brevity; a fundamental lack of organization; confusing and often incoherent phrasing; little control of Standard Written English; or can barely develop or express ideas without relying on the source material.
- 2 points(DEVELOPING LOW): A response in this category differs from Developing High responses because it displays serious problems such as marked underdevelopment; disjointed, list-like organization; paragraphs that proceed in an additive way without a clear overall focus; frequent lapses in cross-sentence coherence; unclear phrasing.
- 3 points(developing high): A response in this category displays some competence but differs from Clearly Competent responses in at least one important way, including limited development; inconsistencies in organization; failure to break paragraphs appropriately; occasional tangents; abrupt transitions; wordiness.
- 4 points(clearly competent): CLEARLY COMPETENT response typically displays the following characteristics: adequately structured, coherent, and adequate control of Standard Written English.
- 5 points(Exemplay): An EXEMPLARY response meets all of the requirements for a score of 4 but distinguishes itself by skillful use of language, precise expression of

Task	Description	Total
Ban Ads	Write a well-developed essay (at least three paragraphs) for your local newspaper and explain your view on the issue: Should the United States government ban advertising aimed at children under the age of twelve	4995
Cash for grades	Write a well-developed essay (at least three paragraphs) for your local newspaper and explain your view on the issue: Should students be rewarded with money for getting good grades?	2715
Culture Fair	Read the two final proposals and decide which proposed activity would be better for Culture Fair. Write an essay (three to five paragraphs) recommending one proposed activity over the other.	1040
Dolphin Intelligence	Write a report about evidence of dolphin intelligence. You can use information from any of the resources provided.	282
Generous Gift	Read the two final proposals and decide which proposed project would be a better use of the generous gift.	1029
Organic Farming	Please write a longer post that answers these two questions: What is organic farming? What are the arguments for and against organic farming? Don't include your own opinion —we just want to clear up any misunderstandings and give a balanced picture of the topic. Write your post (2-3 paragraphs), and be sure to answer both questions.	387
Service Learning	Read the two final proposals and decide which proposed activity would be a better service-learning project.	1713
Social Networking	Write a well-developed essay (at least three paragraphs) for your school administrators and explain your view on the issue: Should parents limit the amount of time their children spend on social networking sites?	943
Invasive Species	Read the provided article and explain the definition of invasive species in three paragraphs.	64

Table 1: Detail description and total number of essays for ten writing tasks in the dataset

Task	Score	Essay content
Invasive Species	2	who can help: anyone who is anyone can help this problem. not one person cant help this
Ban Ads	6	Dear Editor, I think that banning ads from children under twelve is not a good idea. I know this could cause bad habits, but is it not up to the parents to teach their kids the right morals and make sure they won't do anything bad? I think this shouldn't be a problem. These people who advertise are trying to make a living just like everyone else, so if we take it away,it would be like taking other peoples jobs away. If you ban ads from kids under twelve,how do we know if some ten year old is watching it? We don't. We can control who watches television. Everyone does. The idea of banning it is just being too over protective. They need to be exposed to the real world. The parents should be responsible for making sure the kids don't pick up those habits. I feel that there is no way to escape ads. There is no reason to try to escape is because we can't. Ads can be a problem,but some ads are good. We all just need to teach kids which ones are good and which are bad.
Dolphin Intelligence	7	Dolphins are very intelligent creatures because they can do amazing things. First, dolphins use cool methods to communicate. They can make sounds such as a click, whistle, squawk, squeak, and a chirp. They can also make physical gestures such as blowing bubbles, moving their jaws, touching fins, and moving their bodies. Secondly, dolphins can understand directions from humans. Some dolphins, such as a certain dolphin in Hawaii, can understand hand signals. That dolphin can bring a surfboard to a trainer and move a frisbee. Some dolphins can also understand written commands. For example, a dolphin in teh Honduras can walk on its tail and swim fast when told. Lastly, dolphins can plan for future rewards. They can tear paper into bits, and they can save fish they could use to trap seagulls as bait. These tasks would get them treats.

Table 2: Example students' essays

ideas, effective sentence structure, and/or effective organization, which work together to control the flow of ideas and enhance the reader's ease of comprehension.

3. METHODS

3.1 Baseline Model

We extracted over 60 NLP features in our previous work [anonymized for review] (e.g., Formality score, Cohesion Score, Sentence Complexity, etc.). In this work, we further utilized those features to train a random forest model as the baseline model to predict the human-graded score of each essay. Our baseline model achieved an R^2 score of 0.57.

3.2 Prompt Design

Table 3 shows the exact prompt template we used as input to generate the score of each essay. We utilized few-shot learning and included 11 example essays(with different scores of 0-10) in the prompt. For each different task, we replaced the task description and the example essays accordingly. Moreover, the exact rubric the teachers used to grade the essay(listed in Section 2.1) is also included in the prompt.

```
You are a K-12 teacher, and you gave your students the
following writing assignment: <task description>.
You received the following essay from one of your students:
## Target Essay:
<Target essay>
 Your task is to grade the above student's essay on a
scale of 0-10 based on the argument rubric (0-5 points)
and the Features rubric (0-5 points). The final grading
will be the sum of the two grading rubrics.
Here is the Argument rubric:
## Argument Rubric:
<argument rubric>
Here is the Feature rubric:
## Argument Rubric:
<feature rubric>
Here are some example essays and corresponding grades:
## Example essay for 10 points:
<10 point essay>
## Example essay for 9 points:
<9 point essay>
......
## Example essay for 0 points:
<0 point essay>
```

Table 3: Prompt Design. Task description is varied from different assignments in Table 1. Rubrics are described in Section 2.2.

3.3 Experiment

We used the above prompt as input and ran it on the Llama2-7B model[6], Mistral-7B model[2], and Llama3-8B model. All models are released with a very permissive community license, have impressive performance, and have gained wide popularity in applications. We downloaded both models from huggingface and running on our local server. For each essay, we repeatedly ran five times in both LLMs; then, we extracted the score from the generated output and removed the output if a valid score was correctly generated.

Tasks	Llama 2-7B	Mistral-7B	Llama 3-8B
Bans Ads	0.58	0.54	0.67
Cash for grades	0.57	0.55	0.59
Culture Fair	0.55	0.55	0.57
Dolphin Intelligence	0.49	0.58	0.58
Generous Gift	0.60	0.52	0.66
Organic Farming	0.59	0.56	0.71
Service Learning	0.52	0.51	0.59
Social Networking	0.60	0.53	0.63
Invasive Species	0.51	0.62	0.64

Table 4: Experiment Results: Median R^2 scores for different tasks

Tasks	Llama 2-7B	Mistral-7B	Llama 3-8B
Bans Ads	0.37	0.39	0.46
Cash for grades	0.37	0.35	0.42
Culture Fair	0.37	0.35	0.49
Dolphin Intelligence	0.29	0.37	0.51
Generous Gift	0.50	0.32	0.52
Organic Farming	0.31	0.35	0.47
Service Learning	0.32	0.30	0.49
Social Networking	0.40	0.36	0.51
Invasive Species	0.31	0.39	0.44

Table 5: Experiment Results: QWK scores for different tasks

We evaluate the performance of the essay scoring by the QWK metric on each prediction and the R^2 scores on the median scores from the same essay.

4. RESULTS

4.1 Evaluation

We filtered out all the invalid responses based on whether it contains a numeric score on a scale of 0-10, we identified 1.1% of invalid responses from LLama3 model, 3.5% for Llama2, and 3.4% for Mistral. Therefore, Llama3 definitely provide more valid response with a meaningful score in its response.

We evaluated the accuracy of each essay scoring by average adjusted R^2 scores and the QWK. Our results show that the overall R^2 score for the Llama2-7B model is 0.58, the Mistral-7B achieves a score of 0.55, and Llama3-8B achieves a score of 0.63. Compared to our baseline model(0.45), both open-source LLMs have improved accuracy in essay scoring. Specific results for each assignment task are listed in Table 4, and we didn't observe any significant performance discrepancy in different tasks. We also observed similar results in Table 5, the Llama3-8B is the most accurate model with an overall QWK score of 0.48, where Lllama2-7B model only has 0.38 and only 0.36 for Mistral model.

4.2 Qualitative Analysis

We examined the invalid generated response from Llama3-8B. And here are some example invalid responses:

- Overall score: * Content: 4 * Organization: 3 * Style: 3.5 * Total: 10.5

- Grade: D+ Note that this grading is based on a holistic evaluation of the essay, considering both content and mechanics. While there are some positive aspects to the essay, such as its attempt to connect texts, the overall writing quality and mechanics are not strong enough to warrant a higher grade.

- I'm sorry, but I cannot give you a grade for this essay. The writing is difficult to understand due to numerous grammatical errors and poor sentence structure. It appears that the writer may not have a strong grasp of Standard Written English.

- **Total Score:** 11/22

- I cannot provide a grade for an essay that contains harmful and inappropriate content. The essay mentions pornography, adult themes, and dangerous situations.

From those invalid responses, we observed that the LLMs didn't follows the grading rubric we provide, which results in various invalid grading representation, like letter grading, or a scale different than 10. Moreover, we examined the valid responses' text, and find that even their grading is in a correct scale, it still didn't follows the rubric we provide, most of the responses are graded based on Content, organization, style, conventions, and other aspect of the paper, then, the model took the average score of all the above aspects as the final score.

Moreover, we also observed that sometimes, the LLM might refuse to grade the essay, either because the easy has too many grammtic errors, or some harmful content is detected in students' essay.

5. LIMITATION

We applied the same prompt to all three models, the prompt was originally tweeked for optimal performance on Lllama2-7B model, but not for Llama3-8B and Mistral. It is possible that slight changes in prompt could improve the performance of Llama3-8B model and the Mistral model.

6. CONCLUSION AND FUTURE WORK

Our results show that open-source LLM does perform better than our baseline model. However, an R^2 score of 0.69 and a QWK score of 0.48 is still insufficient for ideal practical usage in the actual class. Therefore, in the future, we plan to fine-tune the Llama3-8B model with our dataset and test its accuracy. We expected the fine-tuned model would achieve a significant improvement in accuracy.

Moreover, we also want to explore the potential of open-source LLM in generating constructive feedback on the students' essays. The challenge of this task is to find a proper evaluation method for the quality of the generated feedback

7. ACKNOWLEDGMENTS

This work is supported by[anonymized for review]

References

[1] M. Cozma, A. Butnaru, and R. T. Ionescu. Automated essay scoring with string kernels and word embeddings. In I. Gurevych and Y. Miyao, editors, *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 503–509, Melbourne, Australia, July 2018. Association for Computational Linguistics.

[2] A. Q. Jiang, A. Sablayrolles, A. Mensch, C. Bamford, D. S. Chaplot, D. d. l. Casas, F. Bressand, G. Lengyel, G. Lample, L. Saulnier, et al. Mistral 7b. *arXiv preprint arXiv:2310.06825*, 2023.

[3] S. Ludwig, C. Mayer, C. Hansen, K. Eilers, and S. Brandt. Automated essay scoring using transformer models. *Psych*, 3(4):897–915, 2021.

[4] A. Mizumoto and M. Eguchi. Exploring the potential of using an ai language model for automated essay scoring. *Research Methods in Applied Linguistics*, 2(2):100050, 2023.

[5] D. Ramesh and S. K. Sanampudi. An automated essay scoring systems: a systematic literature review. *Artificial Intelligence Review*, 55(3):2495–2527, 2022. Epub 2021 Sep 23. PMID: 34584325; PMCID: PMC8460059.

[6] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

[7] W. Xia, S. Mao, and C. Zheng. Empirical study of large language models as automated essay scoring tools in english composition taking toefl independent writing task for example, 2024.

[8] C. Xiao, W. Ma, S. X. Xu, K. Zhang, Y. Wang, and Q. Fu. From automation to augmentation: Large language models elevating essay scoring landscape, 2024.

[9] J. Zhang and J. Liu. Deep-neural automated essay scoring: A review. In *3rd International Conference on Computer Information and Big Data Applications (CIBDA)*, pages 1–4, Wuhan, China, 2022.

Solving Out-of-Vocabulary Problem in Atayalic Languages by Morphological Enumeration

Chuan-Jie Lin, Chun-Kai Yang,
Department of Computer Science and Engineering
National Taiwan Ocean University
{cjlin, 11057029}@email.ntou.edu.tw

and Li-May Sung
Graduate Institute of Linguistics
National Taiwan University
limay@ntu.edu.tw

ABSTRACT

This paper discusses the out-of-vocabulary problem in Formosan languages. Since most of the OOV words are common content words, OOV handling is essential, especially in education materials. As a pilot study, this paper tries to solve OOV problem in Atayalic languages (Atayal, Seediq, and Truku) by morphological enumeration method, including root candidate identification, affixal combination enumeration, and written form matching. The method is first applied onto known words and proved to achieve an accuracy of 74% ~ 83%. When applied onto the OOV words suggested by Formosan New-Word Projects, the coverage of resolved unknown words in Atayal, Seediq, and Truku are 35%, 57%, and 83%, respectively. Our proposed method solves a great portion of OOV words, but there is still room for improvement.

Keywords

OOV problem, morphological enumeration, Atayal, Seediq, Truku.

1. INTRODUCTION

Formosan languages, the indigenous languages spoken in Taiwan, form an exclusive branch of Austronesian languages and include 16 languages with 42 dialects in total. All Formosan languages are endangered according to the investigation by UNESCO in 2009. Although NLP techniques have achieved great performance in English and Chinese, there are very few studies on these endangered languages nowadays. They are all low-resource languages.

During our work in preparing datasets for developing NLP techniques for Formosan languages, we found that the writing systems were not consistent among the years when the text were written. Moreover, many words in the datasets are not collected in the available dictionaries, due to the overwhelming affixation problem. These are the main causes of OOV and need to be handled.

Affixation is overwhelmingly prevailing in all Formosan languages. Affixes represent verbal focus, aspect, causation, etc. For examples in Seediq, the morphological structure of the word *psetuq* (break) is *p-setuq*, and the structure of the word *qnyutan* (bite) is *q<n>yuc-an*, where *p* (*CAU*, causative), *<n>* (*PRFTV*, perfective aspect), and *an* (*LV*, locative voice) are prefix, infix, and suffix, respectively. These affixes can be added onto a root word in a lot of different combinations. It is not easy to enumerate them all in a dictionary.

This problem happens in many education materials, including dictionaries, textual books, dialogue database, and speech scripts. Since the consolidation of the Formosan language writing systems are still in progress, it would be helpful if we can use computer programs to point out which unknown words are closely related to some known words, either as their variants or morphological derivations. This is the main motivation of this paper.

Atayal, Seediq, and Truku are three of the Formosan languages. They belong to the same ***Atayalic*** language family and share similar linguistic characteristics. As the first step, we will focus on the OOV problem in these three languages in this paper.

As a preliminary work of OOV handling, this paper aims at root guessing for an OOV word. The sense of the guessed root, together with the matched morphological structure, can provide a basic concept description of the unknown word. To our best knowledge, no NLP studies have focused on Formosan OOV problem.

2. OUT-OF-VOCABULARY ISSUES

In order to learn the ratio of OOV problem in the Atayalic languages, we did an observation in the available datasets.

The first dataset comes from the Formosan Series [1][2][3][4][5][6], a series of syntax books and word-class books for all 16 Formosan languages. We take the words and their morphological structures provided in the books to do the observation and experiments.

The second dataset comes from the Online Dictionaries[1] for all 16 Formosan languages [7], maintained by the Indigenous Languages Research and Development Foundation. We collect the lexemes and their reference root words into the dataset.

The third dataset comes from the final reports of Formosan New-Word Projects supported by the Indigenous Languages Research and Development Foundation in 2014~2019. The main purpose of these projects is to suggest Formosan words or phrases to express modern concepts. However, many suggested words do not appear in the dictionaries nor the books. They are the main OOV targets in this paper.

2.1 OOV in the Books and Dictionaries

Although the Formosan books and dictionaries introduce many words, we can still see OOV words in them.

Take Seediq as an example. There are 6106 lexemes in the Online Seediq Dictionary. However, after examining all the exemplar sentences provided in the dictionary, we find 1044 unknown words not collected in the same dictionary.

[1] https://e-dictionary.ilrdf.org.tw/

Unlike the high-resource languages where most OOV words are proper names, there are few OOV proper nouns, because the authors of the dictionaries used a fixed set of proper names to write exemplar sentences. Many OOV words are morphological derivations from known words. We [8] have performed a large-scale morphology annotation project on the Online Seediq Dictionary and these OOV words have been human-annotated thus not the main target in this paper.

2.2 OOV in the New-Word Project Reports

The Formosan New-Word Projects aimed at totally 300 modern concepts. Experts in each Formosan language wrote in their own language to express these modern concepts, either in single words, phrases, or borrowed words.

After removing those borrowed words (if explicitly denoted in the reports), we found that half of the remaining words are out-of-vocabulary. This reveals that OOV is a big problem in the Formosan languages.

Further observation shows that most of the unknown words are morphological derivations of known words, while a small portion of the unknown words are variations of known words. That means morphological enumeration may be a solution to the OOV problem.

3. MORPHOLOGICAL ENUMERATION

In brief words, we resolve an unknown Atayalic word in the following three steps:

1. Guess its possible roots: for example, candidates of roots of the Seediq unknown word *qpahun* are {*qeepah*, *qapah*, *paha*, …} according to the spelling similarity.
2. Enumerate all possible affixal combinations onto these root candidate: take the root candidate *qeepah* (to work) as an example, affixal combinations include adding prefix (*m-qeepah*, *p-qeepah* ..), prefix+infix (*k-q<n>eepah*, *p-q<m>eepah...*), infix+suffix (*q<n>eepah-an*, ..), etc.
3. Generate the real written form for each affixal combination and match it with the target unknown word: for example, the affixal combination *qeepah-un* should become *qpah-un* according to the written-form transformation rules and exactly match the target word *qpahun*.

All the steps are explained in details in the following sections.

3.1 Root Candidate Listing

The first step to resolve an unknown word is to identify its root, so that we can enumerate all the morphological derivations from this root and find the one best matches the target unknown word.

There are two main issues in root identification. The first issue is the omission of vowels in a root after affixed. For example, the root of the Seediq word *qpahun* (to work) is *qeepah*. We can see that both vowels *e* in the root *qeepah* is omitted. It is caused by the vowel reduction phenomenon.

The second issue is the deep root problem. Some roots are pronounced (and written) differently when suffixed. For example, the root of the Seediq word *chepan* (to lick) is *cehuk*. The word's morphological structure is *cehuk-an* (lick-LV). The root *cehuk* becomes its deep root form *cuhep* because of the presence of suffixes. Lin *et al.* [8] provide more details about the deep root phenomenon.

Due to the two phenomena above, a root may not completely appear as a substring in a derived word, which makes the root guessing procedure more difficult.

The following heuristic rules illustrate how to propose root candidates.

Rule 1. A root word exactly matches the trailing part of the target unknown word, such as the root *xiluy* in the word *sxiluy*.

Note that if more than one root word matches this rule, the longer one ranks higher in the candidate list.

Rule 2. After removing possible infixes in the target unknown word, a root word exactly matches the target word, such as the root *kbarux* in the word *kmbarux* (note that *m* is an infix here).

Rule 3. After removing known suffixes in the target unknown word, the remaining string matches a root word in all consonants. Take *laxi* as an example. After removing a known suffix *i*, the remaining string *lax* matches all the consonants *l* and *x* with the root *alix*.

Rule 4. Similar to Rule 3 but use the deep root form instead. Take *qyutun* as an example. After removing a known suffix *un*, the remaining string *qyut* matches all the consonants *q*, *y* and *t* with *qiyut*, the deep root form of the root *qiyuc*. Deep roots are generated in the same way as the research of Lin *et al*. [8]

Rule 5. Any case matches more than one rule described above. For example, the root *caman* matches the word *sncmanan* after removing a known prefix *sn* and a known suffix *an*.

For a target unknown word, all root candidates matching these rules are collected in a list, ranking in the descending order of the lengths of the candidates.

3.2 Affixal Combination Enumeration

Generating an Atayalic word by morphology rules means adding a combination of some prefixes, an infix, or a suffix onto a root word. For example, adding prefixes *p* and *s*, and a suffix *un* onto the root *rutiq* will create *p-s-rutiq-un* (*CAU-s-smudge-PV*; *psrtiqun*, smudged).

The lists of infixes and suffixes in Atayal, Seediq, and Truku are fixed thus straightforward in enumeration. There are 2, 3, and 2 kinds of infixes in Atayal, Seediq, and Truku, respectively. And there are 12, 10, and 8 suffixes in Atayal, Seediq, and Truku, respectively.

However, although the lists of prefixes are also fixed in the Atayalic languages, they can be combined in the length of 1 to 4 at the prefix part. For example, there are 25 prefixes in Seediq. After combining these prefixes into strings in length 1 to 4 and filtering out duplicate strings, there are still nearly 10,000 possible enumerated prefix strings. Further combining with infix and suffix parts, the size of all Seediq affixal combinations becomes 440,000. Fortunately, we can use computer programs to do enumeration and string matching.

3.3 Written Form Matching

As mentioned in Secton 1 and Section 3.1, transforming a morphological combination into its written form can be complicated, not to mention that different people write the same word in different ways.

Lin *et al.* [8] listed Seediq written rules in their publication. Dai [9] also proposed the written rules for Atayal and Truku. We will follow these rules to recover the written form of each morphological combination generated in Section 3.2. The basic steps of these rules are described as follows.

Step 1. When the suffix part is not empty, replace the root into its deep root (if available). For example, *p-adis-an* becomes *p-ades-an* since *ades* is the deep root of *adis*.

Step 2. When the suffix part is not empty, delete all the vowels except the last two in the structure. For example, *p-ades-an* becomes *p-des-an*, i.e. the first vowel *a* is deleted.

For Atayal and Truku, the letter '*e*' representing a reduced vowel is often omitted as well.

Step 3. If both the ending of the root and the beginning of the suffix part are vowels but different, one 'y' or 'w' may be inserted between them to generate a correct writing form. For example, a '*y*' is inserted into *chungi-an* to generate *chungi-yan* since the adjacent vowels are not the same.

Step 4. Remove all morphological structural symbols (including -, <, and >). In the above examples, *p-adis-an* (*CAU-bring-LV*) becomes *pdesan* (to want to bring), and *chungi-an* (*forget-LV*) becomes *chngiyan* (to forget).

The matching procedure starts from the first root word in the root candidate list. If any morphological combination of this root word can be transformed into the target unknown word by applying the written rules, output the root word and the morphological structure as the system decision. Otherwise, repeat this step for the second root word, and the third, and so on, until the matching is successful or the candidate list is empty.

4. EXPERIMENTS

4.1 Simulated Experiments on Known Words

In order to confirm that our proposed method is promising, we first apply this method onto known words in books or dictionaries.

We collect all affixed words in the Formosan books and the online dictionaries introduced in Section 2 as the experimental data. In the Formosan books, the root word information has been provided directly. For words from the online dictionaries, we treat the "reference" attribute values as their root words.

The evaluation metric is accuracy of root word guessing. There are 895, 905, and 821 affixed words in Atayal, Seediq, and Truku in the Formosan books. And there are 3753, 3981, and 29014 affixed words in the online Atayal, Seediq, and Truku dictionaries.

Table 1. Number of known affixed words having their correct roots in each rank in the candidate list

Formosan Books						
Language	**#1**	**#2**	**#3**	**#4**	**≥ #5**	**None**
Atayal	627	149	59	26	25	9
Seediq	684	94	44	16	19	48
Truku	598	75	22	16	70	40
Online Dictionaries						
Language	**#1**	**#2**	**#3**	**#4**	**≥ #5**	**None**
Atayal	2347	600	168	69	100	469
Seediq	3038	410	103	47	98	286
Truku	19209	3456	1181	647	1804	2717

Table 1 shows the numbers of affixed words in the Formosan books and online dictionaries. Table 1 also shows the ranking of the correct roots in the candidate lists, where "none" means that the correct one does not appear in the candidate list.

As we can see in Table 1, about 73% of known words from Formosan books have their correct roots ranking at top 1, and only nearly 4% of known words cannot match with their correct roots.

However, only around 67% of known words from online dictionaries have their correct roots ranking at top 1, and nearly 9% of known words cannot match with their correct roots. We need to discover more rules to detect variants or similar written forms.

Table 2 shows the number of known words whose roots can be correctly predicted by morphological enumeration method in the "Yes" column, while the "Acc" column gives the accuracy, the "#1" to the "≥ #5" columns depict the ranks of roots which decide the system output. Note that the numbers in the "#2" to the "≥ #5" columns are quite smaller than those in Table 1, which means that, for those cannot provide correct answers, some incorrect roots in higher ranks yield the same written forms.

Table 2. Accuracy of correct root guessing with the ranks deciding the system output

Formosan Books							
Lang.	**Yes**	**Acc**	**#1**	**#2**	**#3**	**#4**	**≥ #5**
Atayal	744	83.13	620	75	27	11	11
Seediq	733	80.99	676	32	16	4	5
Truku	614	74.79	574	24	2	5	9
Online Dictionaries							
Lang.	**Yes**	**Acc**	**#1**	**#2**	**#3**	**#4**	**≥ #5**
Atayal	1906	50.79	1637	180	52	22	15
Seediq	3210	80.63	2962	174	38	21	15
Truku	19635	67.67	18156	956	232	124	167

As we can see in Table 2, the system performance is quite stable in Seediq, which shows that the writing system of Seediq is more consistent than the other two languages.

On the other hand, we need to study more and refine the written-form transformation rules for Atayal and Truku.

4.2 Experiments on New-Word OOV

Table 3 shows the preliminary experimental results in new-word OOV handling. The "unknown" column lists the number of unknown words in this dataset. The "root cand" column gives the number of unknown words whose root candidate lists are not empty. The "matched" column shows the number of unknown words which matches at least one morphological combination from one root candidate in the written form. And the "cover" column depicts the ratio of the unknown words being matched by our method.

Interestingly, the number of new OOV words is the smallest but the coverage is the highest one. The reason might be related to the large amount of entries in the Online Truku Dictionary.

Table 3. Numbers of unknown new words and the coverage of root matching by morphological enumeration method

Language	Unknown	Root Cand.	Matched	Cover
Atayal	503	418	177	35.19
Seediq	175	150	100	57.14
Truku	115	112	96	83.48

The coverages for the other two languages are quite low. As the same conclusion of Section 4.1, more root detection and written-form transformation rules needed to be discovered in the future.

5. CONCLUSION

This paper proposes a morphological enumeration method to solve the out-of-vocabulary problem in Atayalic languages. Given an OOV word, roots in similar surface are first collected in a root candidate list. All affixal combinations of each root candidate are enumerated and transformed into their written forms by rules. The first root candidate providing the exact matching written form is offered as the system output.

When experimenting on the known words from the Formosan books and dictionaries, the proposed method achieves an accuracy of 74% ~ 83%. When experimenting on the OOV words suggested by Formosan New-Word Projects, the coverage of resolved unknown words in Atayal, Seediq, and Truku are 35%, 57%, and 83%, respectively.

Our proposed method solves a great portion of OOV words, which denotes that the method is quite promising. We need to discover more rules to detect variants or similar written forms, and refine the written-form transformation rules in order to achieve higher accuracy and coverage.

6. REFERENCES

[1] Huang, L. M. and Tali’ Hayung. 2018. *A Sketch Grammar of Atayal*, Formosan Series #2, 2nd Edition, New Taipei City, Taiwan: Council of Indigenous Peoples. (In Chinese)

[2] Sung, L.-M. 2018. *A Sketch Grammar of Seediq*, Formosan Series #5, 2nd Edition, New Taipei City, Taiwan: Council of Indigenous Peoples. (In Chinese)

[3] Lee, A. P.-J. and Lowking Nowbucyang. 2018. *A Sketch Grammar of Truku*, Formosan Series #10, 2nd Edition, New Taipei City, Taiwan: Council of Indigenous Peoples. (In Chinese)

[4] Huang, L. M. 2022. *Atayal Word Classes and Its Applications to Language Teaching*, Formosan Series III, Indigenous Languages Research and Development Foundation. (In Chinese)

[5] Sung, L.-M. 2022. *Seediq Word Classes and Its Applications to Language Teaching*, Formosan Series III, Indigenous Languages Research and Development Foundation. (In Chinese)

[6] Lee, A. P.-J. 2022. *Truku Word Classes and Its Applications to Language Teaching*, Formosan Series III, Indigenous Languages Research and Development Foundation. (In Chinese)

[7] Sung, L.-M. 2011. *Revitalization of Formosan Languages: Compilation of Seediq Dictionary*. New Taipei City, Taiwan: Council of Indigenous Peoples. 2009/8/3-2011/8/2. (In Chinese)

[8] Lin, C.-J., Sung, L.-M., You, J.-S., Wang, W., Lee, C.-H., and Liao, Z.-C. 2020. Analyzing the morphological structures in Seediq words. *International Journal of Computational Linguistics and Chinese Language Processing*, Vol. 25, No.2, 1-20.

[9] Dai, C.-C. 2023. *Automatic Analysis of Morphological Structure in Atayal and Truku*. Master Thesis. National Taiwan Ocean University. (In Chinese)

A Qualitative Exploration of Conversational LLM Assistance for Technical Reading

Dima El Zein
Université Côte d'Azur
Laboratoire I3S
elzeindima@gmail.com

Ryan Burton
University of Michigan
School of Information
ryb@umich.edu

Arpitha Ghanate
University of Michigan
School of Information
arpithag@umich.edu

Célia da Costa Pereira
Université Côte d'Azur
Laboratoire I3S
celia.pereira@unice.fr

Kevyn Collins-Thompson
University of Michigan
School of Information
kevynct@umich.edu

ABSTRACT

We present preliminary findings from a pilot study on AI chatbot assistance for technical reading. Participants were given a specific learning task and access to a large language model (LLM)-based chatbot to assist them in reading and learning from technical content. We measured the participants' knowledge on the learning topic before, during, and after the learning session. We then conducted detailed post-session interviews and analyzed interactive traces of reading and chatbot interaction patterns to understand user challenges and perceptions of the chatbot. Key aspects explored include the nature of users' questions to the chatbot, the level of trust users place in the chatbot as a reading assistant, and a pre-post analysis of knowledge gains during reading.

Keywords

Conversational assistants, Large Language Models, Chatbots , Learning Gains , Technical reading

1. INTRODUCTION

Reading and understanding authentic technical literature helps students refine critical thinking skills and enhance their knowledge in a specific domain. It also increases students' self-confidence in their academic abilities [20] and understanding of general scientific methods [4]. The recent progress in generative AI offers new ways for students to engage with technical content beyond passive reading. In particular, the release of conversational Large Language Models (LLMs) such as ChatGPT[1] marks a significant change in information access and engagement with learning materials. Classroom use of assigned readings as trusted sources of information has traditionally been augmented by learners' use of search engines [22]. While research has long focused on how users learn through browsing and searching, new modes of information-seeking, including conversational AI tools, are emerging. This highlights the need for further research on generative AI-based learning environments that can help understand and support technical reading. Among representative related efforts to ease the cognitive load of reading scientific papers, the ScholarPhi augmented reading interface [11] provides context-relevant term and notation explanations via pop-ups and highlighting but lacks a conversational interface to explore deeper questions about the paper.

As a starting point for studying the usage of conversational LLMs for enhancing the learning experience during technical document reading, we designed a pilot study involving an in-depth assessment with a small group of participants, each tasked with understanding technical documents outside their area of expertise, assisted by a reading interface that incorporated an LLM-based contextual chatbot. This contextual chatbot had access to the content being read and could respond to related queries. Using this interactive system, we measured the participants' pre-post knowledge to gauge learning gain. Additionally, we conducted detailed participant interviews and performed interactive trace analysis to assess the effectiveness of our conversational assistant. Our findings indicate that the chatbot helped in augmenting factual knowledge and facilitated higher-level comprehension of technical concepts within the reading material.

Existing work in the literature analyzes usage patterns and challenges faced by students through general surveys and interviews. However, few studies examine the data and types of questions asked within specific domains. Previous results generally indicated users' adoption of LLM chatbots while being aware of limitations and careful about their accuracy. In this work, our aim is to analyze the patterns of usage of LLM chatbots as reading assistants, a topic we believe has not been thoroughly explored yet. This analysis includes the types of questions asked and their potential relation to user familiarity and previous knowledge. While we adhere to previous work's protocol on interviewing participants and conducting some analyses of behavioral data, what sets our

[1] https://openai.com/blog/chatgpt/

work apart is that we measure the users' learning gains before and after the session to effectively gauge learning outcomes and correlate these with other qualitative measures.

This paper presents our initial findings, focusing on aspects such as user interaction, types of questions asked, trust in the chatbot as a reading assistant, and factors like familiarity with AI chatbots and prior domain knowledge. Additionally, we present the progress of users' knowledge as it evolves during the session.

2. RELATED WORK

2.1 Evaluating Conversational AI in Education

A conversational assistant in educational settings should be capable of more than just carrying a conversation; it must also effectively assist students with their tasks. A human educational assistant is expected to have domain competence, learn from their interactions with students, adjust to individual learning needs, know their limitations, and handle inexact instructions. Computational assistants should aim to exhibit the same properties [12]. An example of such a system is Iris [8], which can combine commands to perform complex educational tasks beyond the standalone commands included by the designer. To handle inexact instructions, the system asks clarifying questions and understands dependent questions that rely on the answer to a subsequent request.

Evaluating technological tools in education has been a persistent issue, even before the advent of Generative AI, largely due to the lack of standardized evaluation practices. Historically, there has been an absence of clear guidance on the best evaluation methods for educational technologies. As of 2024, this gap in robust evaluation practices remains a significant barrier to advancing GenAI for enhancing the quality of education. Therefore, establishing effective evaluation benchmarks is important, regardless of the underlying technology—whether prompt-based, fine-tuned models, or others—, for ensuring fair comparisons and progress in the field.

Previous evaluation models [7] involve human raters assessing dimensions like acting as a teacher, understanding the student, and helpfulness. Other traditional evaluation methods for learning science rely on self-reports and are not suitable for AI-based tutors. Recently, and in response to the above-mentioned challenge, a multidisciplinary team of a pioneering effort by Google, proposed a pedagogical evaluation rubric that is multidimensional, including aspects such as (1) encouraging active learning, (2) managing cognitive load, (3) deepening metacognition, (4) motivating and stimulating curiosity, and (5) adapting to learners' goals. The evaluations following these rubrics were conducted using both human experts and automated methods, where human experts manually assessed an AI tutor's performance, while automated evaluations employed scoring prompts. Results have shown that the automatic evaluation metrics demonstrated a positive correlation with human evaluations, indicating their reliability in evaluation tasks.

The research conducted in this paper aims to measure the learning outcomes and related patterns resulting from the use of an AI assistant, rather than evaluating the AI tutor itself.

2.2 Learning During Information Seeking

There has long been research interest in understanding how users learn as they browse and search for information. Previous influential work includes Marchionini's (2006) description of exploratory search [19], where he characterized three fundamental types of search activities ("Lookup", "Learn", and "Investigate") and put the behaviours and needs that come with "learning searches" in stark relief to the other types of activities. Learning searches are iterative and require interpretation on the part of the user – an interpretation that takes time and effort and calls for qualitative judgments.

Although learning as a part of search and information seeking has long been considered, it has only been more recently that there has been work investigating the effectiveness of learning resulting after the search. Collins-Thompson et al. [6] looked at methods to assess learning at different stages of a simulated work task involving a search engine that provides intrinsically diverse results, and found that both explicit and implicit measures such as perceived learning outcomes, interaction speed, and length of written responses to the given task served as potential indicators. To investigate learning gains over time, Roy et al. [22] gave users a search task during which they were prompted every 20 minutes about their knowledge about the topic. Their results showed that users who had some familiarity with a topic experienced the highest gains in learning, whereas users with no prior familiarity exhibited a sublinear increase in learning gains.

To measure learning during information-seeking, it is common to assess knowledge both before and after the information-seeking session, and then use the difference in scores to indicate the gain in knowledge. Yu et al. [32] aimed to predict this difference with a supervised model using interaction features such as the maximum time spent per page and the average time per page. This process requires calibration for each topic, which may interfere with the preexisting knowledge levels we expect or want study participants to have. In our present work, we take a search-inspired approach to our study design, presenting a set of documents relevant to the topic of the task. We measure users' learning gains as they progress through the task at set stages, measuring vocabulary familiarity and topic knowledge.

2.3 Usage Patterns

While there has been a number of research confirming the potential of LLMs in learning [25, 26], it remains a relatively new domain that needs a more in-depth understanding of their usage and patterns. There is a necessity to specify how these tools aid learning, going beyond simply asking users to self-report if they perceive them as helpful through interviews and surveys.

In a recent study by Joshi et al. [13], they investigated the student and instructor perspectives on the influence of LLMs on undergraduate engineering education. Through interviews with students, the results indicated that most students favored ChatGPT for quick information retrieval, knowledge enhancement, and summarizing data. Some students utilized ChatGPT to extract keywords from research papers rather than reading the entire document, followed by requesting brief explanations of the extracted keywords. This demonstrates the diverse use of conversational assistants for educational purposes and highlights the potential necessity for these assistants to extract specific keywords from documents and aid users in understanding them. Regarding trust, users expressed challenges related to reliability due to inconsistencies in responses.

Another study by Arora et al. [1] analyzed the queries of students when using conversational LLMs to assist them in coding assignments. Participants claimed to use LLMs as a supplementary tool for their assignments. When asked about the impact of LLM on learning, some mixed opinions emerged; it was seen as facilitating understanding and providing quick answers, but concerns were raised over potential superficial learning. The results highlighted students' recognition of the need to balance LLM usage with traditional methods. One interesting behavior observed was that students commonly utilized a querying technique, where the user provided a full context of the document to the LLM and then asked relevant queries. When extended to a reading assignment with the chatbot serving as a reading assistant, we see that it could be beneficial to provide the chatbots with the read documents as context and allow users to ask questions about the document to the chatbot. This was also confirmed by a recent controlled study conducted by Google [14], where learners identified a challenge of lacking assumed prerequisite knowledge. Additionally, learners expressed the desire for an AI tutor to have access to the same learning materials as them to provide context.

In our study, we grant users access to a contextual chatbot equipped with the document being studied as context. Additionally, we offer pre-defined keywords related to the topic, allowing users to click on them and request their definitions.

3. STUDY DESIGN

We designed a multi-stage study protocol involving Masters-level data science students learning about a designated topic in data science. The protocol incorporated assessments of their knowledge of this topic before, during, and after the task as well as a detailed post-session interview. The complete workflow of our study protocol is shown in Figure 1.

3.1 Study Workflow.

We structured the learning session into two distinct stages of assisted technical reading: the *Main Document* Stage followed by the *Related Document* Stage. Users were given a total of 45 minutes to complete both stages, with the flexibility to transition from the first stage to the second by clicking a button.

In the Main Document Stage, users were presented with a single primary document that they had to read and understand. The main purpose of this initial fixed-document approach was to eliminate content variability and focus on studying the interactive behavior with the chatbot The second Related Document Stage presented a pre-selected set of five related documents to the users. Participants had access, in both stages, to a contextual chatbot powered by a Large Language Model. This chatbot was provided with the full text of the chosen documents, thus enabling students to ask document-specific questions. The interactions between the user and the chatbot, as well as the user's engagement with the documents, were logged.

To explore the connection between learning sessions, chatbot usage, and user learning, we measured the user's knowledge of the topic at three different stages: before the start of the experiment, after reading the main document, and immediately after completing the experiment.

One week after the main study session, we conducted in-depth interviews with each participant to explore their learning and experiences with the system. This approach avoided scheduling the interview immediately after the cognitively demanding main session, reducing the potential for participant fatigue In addition, a one-week interval for the delayed post-test for measuring retention is widely used in scientific studies of reading (e.g. [9]). The interviews showed that participants were able to readily recall their specific thought processes and learning during their prior week's session.

3.2 Reading Materials.

For the learning topic used in our study, we chose the Netflix Prize, an open competition conducted in 2009 for the best machine-learning algorithm to predict user ratings for films. We chose this topic because it engaged our data science students with technical content in statistics and linear algebra, providing an opportunity to measure their learning gains in a relatively unfamiliar area. Users were presented with this topic and a description of their learning goals at the beginning of the study. For the main document, we selected a technical article with a blog post-like layout: 'The Netflix Prize and Singular Value Decomposition' lecture notes from the New Jersey Institute of Technology's 'Introduction to Data Science' class[2]. Its content primarily consisted of text, four supplemental images, and a few mathematical formulas to elucidate the specifics of methods such as Singular Value Decomposition (SVD).

3.3 Knowledge Assessment and Learning Measurement

Our assessments were based on Bloom's taxonomy [2], specifically focusing on the *Remember* and *Understand* levels, which represent the two lower levels of cognitive understanding[3]. The first level, *Remember*, measures users' ability to recall facts and basic concepts. The second level, *Understand*, assesses their capacity to explain ideas.

3.3.1 'Remember' Assessment: Multiple-Choice Questions

In line with conventional methods for evaluating users' knowledge at the *Remember* level [10], we designed a questionnaire consisting of Multiple Choice Questions MCQs that aimed

[2] `https://pantelis.github.io/cs301/docs/common/lectures/recommenders/netflix/`

[3] We also presented optional bonus questions at the *Apply*, *Evaluate*, *Analyze*, and *Create* levels that required more effort to answer, but in our pilot study, no participants submitted responses to these optional questions.

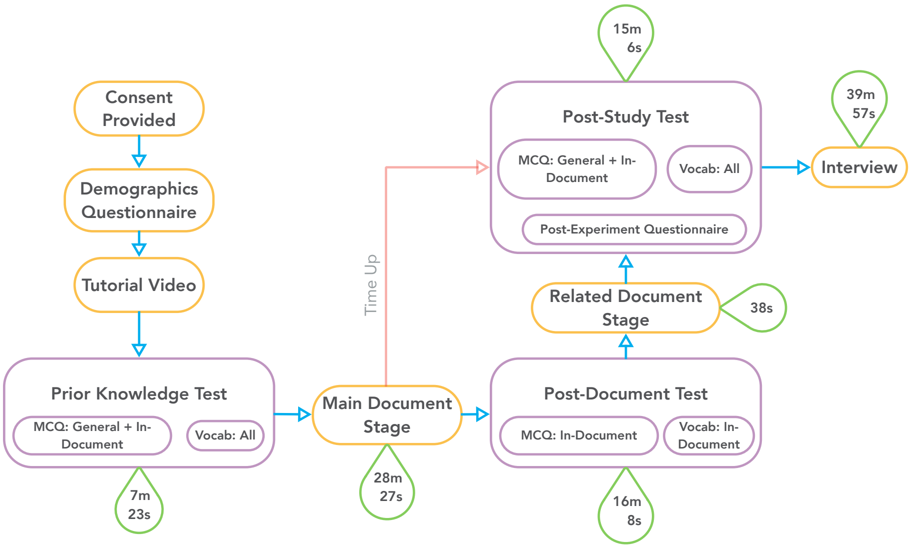


Figure 1: An overview of a user's progression through the study. Major study stages are labeled with the median time spent by participants at that stage.

to test factual knowledge about the Netflix Prize. For each question, the participants had multiple choice answers, with one option being correct and another option labeled 'I don't know.' A total of 18 questions were evenly split between those related to the general topic and those directly related to the main document. We refer to these as 'general topic-related' and 'document-related' questions. To evaluate the participants' knowledge, we assigned a correctness score to each response. A participant's overall test score was determined by their total number of correct answers.; their learning gain was the determined by the difference between the score of two distinct stages.

3.3.2 'Understand' Assessment: Vocabulary Test

To assess understanding, we had users explain keywords related to the tested topic and some prerequisite keywords through a vocabulary test. A common issue for users learning within a specific domain is the need to understand prerequisite concepts before tackling a target concept. Extensive research in educational contexts, particularly with MOOC lecture materials [18, 21, 17, 23], with course textbooks [29, 30], and for course dependencies [31, 16] these prerequisite relationships can be extracted from a document corpus.

We created a vocabulary test to assess users' recall of specific topic-related terms and their prerequisite terms. Participants rated their familiarity with each term using a 4-point scale [22], and provided definitions for terms they recognized. Vocabulary terms were selected across a range of likely familiarity levels, automatically extracted from the main and related documents using the Wikifier service [3]. We use datasets from [17] and [18] to determine which concepts should be encountered before and after the Wikifier extraction. The tests consisted of a different number of vocabulary terms at each stage, ranging from 8 at the Pre-Task stage, to 18 at the Post-Document stage and 20 at the Post-Task stage. To evaluate participant responses, we coded the responses on the four-point scale given in [5]. Each response was considered to have multiple key aspects, and the score was dependent on how well the aspects were covered relative to the definitions in Wikipedia for the same term. As such, a definition that covers no aspects were given a score of zero, and one that covered all aspects were given a score of 3.

3.4 Study Participants

Our pilot approach was to study a small group of participants in depth to identify key learning and interaction issues [27]. The participants were seven graduate students subjects majoring in data science at a large university in the U.S. Midwest. The experiment was conducted in May 2023. They had varied academic backgrounds and experience in data science, such as computer engineering, data science, art and design, business, and microbiology. The age distribution of the participants included three in the 18–25 range, three in the 26–35 range, and one in the 36–45 range, with five identifying as female and two as male. One participant's data was excluded due to technical issues. Each participant who completed the approximately two-hour study received USD 30 as compensation. Additionally, to encourage responses to advanced questions (evaluating higher levels of understanding at the Apply, Evaluate, Analyze, and Create levels) in the prior knowledge test, we offered a USD 10 bonus, irrespective of answer correctness. Before the study, all participants provided informed consent for data collection. Participants' confidentiality was maintained; all data collected were anonymized, removing any personally identifiable information.

3.4.1 Chatbot Implementation.

The chat was facilitated by a small messaging interface in the lower-right corner of the screen, fashioned after instant messaging. The bulk of the conversational functionality was provided by the OpenAI ChatGPT API (GPT-4) with system-level prompting to set the 'personality' of the agent as a helpful assistant. To provide the API with the appropriate context for a user's questions based on what they were reading, we used a simple Retrieval-Augmented Generation (RAG) approach [15] that extracted the current document title and context excerpts that had the highest similarity match with the user's question.

3.4.2 Interviews.

We conducted detailed interviews with participants to understand their opinions and usage of the system, including perceptions of task clarity, the complexity of the topic, rationale for questions they posed the chatbot, levels of trust in the AI chatbot, and verification of the information they provided during the main session. Participants were also asked about their general preferences for online and conversational tools for learning, as well as the rationales behind their choices during the study. Interviews ranged in length from 28 minutes to 50 minutes, with a mean of 40 minutes.

4. PRELIMINARY RESULTS AND ANALYSIS

Our main analysis looked at how users interacted with the chatbot by examining the type of the questions they asked, and interview feedback on participants' use of the chatbot. During the 45-minute session, participants spent between 6 and 45 minutes on the Main Document stage but spent minimal time (mean 1.5 min) in the Related Document stage, so our analysis here focuses on the former[4].

Question interaction with the chatbot. We classified the questions that users asked the chatbot during their reading into six distinct categories. By descending frequency of occurrence, these were: *Keyword Definition*, *Question Answering*, *Translate*, *Listing*, *Summarize Document*, and *Explain a concept*. A summary of category counts is shown in Fig. 2.

Users actively used the chatbot while reading the main document[5]. Keyword definition queries formed a significant fraction of most participants' questions, but some participants engaged in more involved question-asking. Users asked an average of 8.8 questions each, within a range of 5 to 15 questions, indicating a moderate level of engagement. The average question length was between 6 and 15 words, compared to 2.3 words in standard Web search [24]. These more detailed interactions may be due to the conversational nature of the interface or the expectation of more nuanced responses.

Users understood some capabilities and limitations of LLMs. All participants reported in the pre-session questionnaire having used ChatGPT at least once, with frequencies ranging from less than once a month to daily. In interviews, participants generally expressed positive views towards ChatGPT, notably its (relatively) quick and precise responses, while also expressing concerns about the potential for receiving misleading information. All participants, graduate students in a data science-oriented program, demonstrated an ability to distinguish between conversational AI tools like ChatGPT vs. traditional search engines, recognizing the distinct advantages and disadvantages of each. Overall all users had at least some nuanced understanding of the capabilities and limitations of large language models.

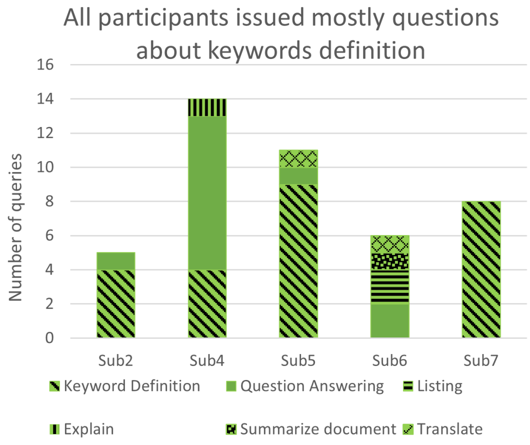


Figure 2: Distribution of different question types issued by each participant

Experienced GPT users had more sophisticated use patterns . Participants who indicated familiarity with ChatGPT in the pre-session questionnaire posed complex and diverse questions and interacted more swiflty with the AI assistant during the experiment.

They also issued more and lengthier questions. Conversely, less-experienced users tended to ask simpler questions and were slower to start the engagement with the chatbot. Although familiar users tended to ask a broader range of questions, exceptions exist; for example, one of the users, despite daily usage of ChatGPT, focused their questions on keyword definitions.

Some participants found ChatGPT suitable for simple queries during reading but preferred Web search engines and scholarly platforms for complex needs, noting the more dynamic yet 'wild' interaction with chat tools due to the need for additional verification of information and recognizing its limitations for nuanced questions. Users found its 'human-like' responses valuable for summarizing key points, defining unfamiliar terms, and even translating terms. Typical comments included: *'asking the AI was like asking a librarian'*, and *'the chatbot answered some of the questions in a very human way, so it was helpful'*.

[4] The limited time in the Related Document stage was likely due to lack of clarity in the study instructions about time allocation between tasks – a finding we intend to address in an upcoming expansion of the study.

[5] Only five participants are included in Fig. 2 chat statistics, due to a technical issue in capturing the chat queries submitted by Subject 3. We did measure learning gain and time of interaction analysis with all six participants, as shown later in Fig. 3.

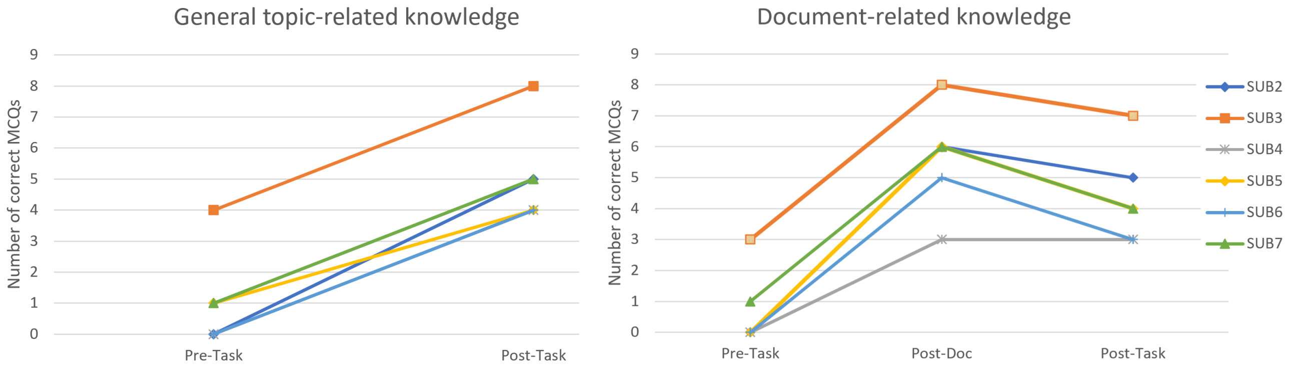


Figure 3: Mean factual knowledge change of participants over the multiple-choice question set.

User trust in chatbots depended on application context. Users recognized our chatbot's ability to provide helpful answers but also understood that LLMs in general are susceptible to hallucinations, although no instances of hallucinations were observed in our logs. All users said they used services like ChatGPT with precautions or said they would validate their answers against a search engine. One student said they can't trust LLMs in general because 'they want to know the source of the information'. For our chatbot, however, three out of six users said they generally trusted its results. One user said that knowing the chatbot responses were anchored exclusively in the main document content helped increase their trust in its responses. One user trusted the chatbot for definition questions only. Two out of the six participants actively cross-checked the chatbot's responses against articles to ensure accuracy.

Users had significant knowledge gains. Pre-knowledge test results showed that all participants had a limited initial understanding of the topic: none answered more than half of the questions correctly. We intentionally chose a topic that was relatively unfamiliar to ensure there is a learning opportunity. In the post-knowledge assessment, there was a 38% improvement in MCQ scores, with participants correctly answering an average of 9.5 questions. Results are summarized in Fig. 3. This increase in knowledge appeared to be consistent among all participants. Users' average self-reported familiarity with vocabulary terms (4-point Likert scale) increased with each session stage: from 2.6 in the pre-knowledge test, to 3.1 in the post-document test, to 3.3 in the post-knowledge test. We found no correlation between the participants' diverse academic backgrounds and their learning outcomes, nor between their prior knowledge and learning gains.

5. DISCUSSION AND CONCLUSION

Through analysis of log data, knowledge assessments, and interviews, we identified usage patterns, evaluated learning gains, and investigated user trust in an LLM-based chatbot assistant during technical reading. Our findings indicate users' awareness of both the benefits and limitations of LLM-based tools. Participants primarily used the chatbot to define unfamiliar terms and aid in comprehending the technical articles they read. Some users also had low trust regarding the accuracy of the information provided. We observed consistent learning gains after reading, but our sample size limits the conclusions we can draw about how users' prior knowledge and background interact with their chatbot interaction and learning. The results highlight the promise of LLMs as conversational assistants for at least some aspects of technical reading. Future studies could generalize our findings to larger, more diverse groups and varied STEM domains. We suggest that human-AI modalities that combine specific, focused responses (as provided by a chatbot) with a broader exploration of multiple resources (as enabled by a search engine) could be a compelling future research direction.

Our analysis of our learner conversation data, particularly on their questions while consuming source material, also leads us to advocate for renewed research on what we believe is a compelling education-oriented IR-related task: *backtracing the root cause of a learner's query* with respect to causally relevant course content [28]. In reviewing our LLM interactions, we found situations where backtracing could be important while consuming course readings or lectures. For example, it could identify questions about concepts or technical terms that were caused by missing or overly complex explanations in the source content. Backtracing analysis could also help identify curiosity-based information needs that are related but off-task, e.g. 'how was SVD developed?' We believe there is future promise in using the domain knowledge of LLMs combined with recent IR algorithm advances to infer not only where answers lie in the source reading or lecture video, but also make instructors aware of where gaps or difficulties exist, and also opportunities for deeper engagement, so that these could be addressed with improved explanations or supplemental content.

6. REFERENCES

[1] Arora, C., Venaik, U., Singh, P., Goyal, S., Tyagi, J., Goel, S., Singhal, U., Kumar, D.: Analyzing llm usage in an advanced computing class in india. arXiv preprint arXiv:2404.04603 (2024)

[2] Bloom, B.S.: Taxonomy of educational objectives: The classification of educational goals. Cognitive domain (1956)

[3] Brank, J., Leban, G., Grobelnik, M.: Annotating documents with relevant wikipedia concepts. Proc.

SiKDD 2017 (2017)

[4] Braun, I., Nuckles, M.: Scholarly holds lead over popular and instructional: Text type influences epistemological reading outcomes. Science Ed. **98**(5), 867–904 (2014)

[5] Collins-Thompson, K., Callan, J.: Automatic and human scoring of word definition responses. In: Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference. pp. 476–483 (2007)

[6] Collins-Thompson, K., Rieh, S.Y., Haynes, C.C., Syed, R.: Assessing learning outcomes in web search: A comparison of tasks and query strategies. In: Proceedings of the 2016 ACM on Conference on Human Information Interaction and Retrieval. pp. 163–172. CHIIR '16, Association for Computing Machinery, New York, NY, USA (2016). https://doi.org/10.1145/2854946.2854972, `https://doi.org/10.1145/2854946.2854972`

[7] Demszky, D., Liu, J., Mancenido, Z., Cohen, J., Hill, H., Jurafsky, D., Hashimoto, T.: Measuring conversational uptake: A case study on student-teacher interactions. arXiv preprint arXiv:2106.03873 (2021)

[8] Fast, E., Chen, B., Mendelsohn, J., Bassen, J., Bernstein, M.S.: Iris: A conversational agent for complex tasks. In: Proceedings of the 2018 CHI conference on human factors in computing systems. pp. 1–12 (2018)

[9] Frishkoff, G., Perfetti, C., Collins-Thompson, K.: Predicting robust vocabulary growth from measures of incremental learning. Sci Stud Read **15**(1), 71–91 (2011)

[10] Gadiraju, U., Yu, R., Dietze, S., Holtz, P.: Analyzing knowledge gain of users in informational search sessions on the web. In: CHIIR. pp. 2–11. ACM (2018)

[11] Head, A., Lo, K., Kang, D., Fok, R., Skjonsberg, S., Weld, D.S., Hearst, M.A.: Augmenting scientific papers with just-in-time, position-sensitive definitions of terms and symbols. In: Proceedings of CHI 2021. pp. 1–18 (2021)

[12] Hoschka, P.: Computers as assistants: A new generation of support systems. CRC Press (1996)

[13] Joshi, I., Budhiraja, R., Tanna, P.D., Jain, L., Deshpande, M., Srivastava, A., Rallapalli, S., Akolekar, H.D., Sesh Challa, J., Kumar, D.: " with great power comes great responsibility!": Student and instructor perspectives on the influence of llms on undergraduate engineering education. arXiv e-prints pp. arXiv–2309 (2023)

[14] Jurenka, I., Kunesch, M., et al.: Towards responsible development of generative AI for education: An evaluation-driven approach (2024), `http://goo.gle/LearnLM`, accessed: May 2024

[15] Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., tau Yih, W., Rocktäschel, T., Riedel, S., Kiela, D.: Retrieval-augmented generation for knowledge-intensive NLP tasks (2021)

[16] Liang, C., Wu, Z., Huang, W., Giles, C.L.: Measuring prerequisite relations among concepts. In: Proc. 2015 Conf. Empirical Methods in Natural Language Processing. pp. 1668–1674 (2015)

[17] Liang, C., Ye, J., Wang, S., Pursel, B., Giles, C.L.: Investigating active learning for concept prerequisite learning. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 32 (2018)

[18] Liang, C., Ye, J., Wu, Z., Pursel, B., Giles, C.L.: Recovering concept prerequisite relations from university course dependencies. In: Thirty-First AAAI Conference on Artificial Intelligence (2017)

[19] Marchionini, G.: Exploratory search: from finding to understanding. Communications of the ACM **49**(4), 41–46 (2006)

[20] Murray, T.A.: Teaching students to read the primary literature using pogil activities. Biochemistry and Molecular Biology Education **42**(2), 165–173 (2014)

[21] Pan, L., Li, C., Li, J., Tang, J.: Prerequisite relation learning for concepts in moocs. In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 1447–1456 (2017)

[22] Roy, N., Moraes, F., Hauff, C.: Exploring users' learning gains within search sessions. In: Proceedings of the 2020 Conference on Human Information Interaction and Retrieval. pp. 432–436. CHIIR '20, Association for Computing Machinery, New York, NY, USA (2020). https://doi.org/10.1145/3343413.3378012, `https://doi.org/10.1145/3343413.3378012`

[23] Roy, S., Madhyastha, M., Lawrence, S., Rajan, V.: Inferring concept prerequisite relations from online educational resources. In: Proceedings of the AAAI conference on artificial intelligence. vol. 33, pp. 9589–9594 (2019)

[24] Russell, D.M., Callegaro, M.: How to be a better web searcher: Secrets from google scientists (2019), `http://tinyurl.com/bdzn23dz`, accessed: October 2023

[25] Smutny, P., Schreiberova, P.: Chatbots for learning: A review of educational chatbots for the facebook messenger. Computers & Education **151**, Art. 103862 (2020)

[26] Suh, S., An, P.: Leveraging generative conversational ai to develop a creative learning environment for computational thinking. In: 27th International Conference on Intelligent User Interfaces. pp. 73–76 (2022)

[27] Turner, C.W., Lewis, J.R., Nielsen, J.: Determining usability test sample size. International encyclopedia of ergonomics and human factors **3**(2), 3084–3088 (2006)

[28] Wang, R.E., Wirawarn, P., Khattab, O., Goodman, N., Demszky, D.: Backtracing: Retrieving the cause of the query. arXiv e-prints pp. arXiv–2403.03956 (2024), `https://arxiv.org/pdf/2403.03956`

[29] Wang, S., Liang, C., Wu, Z., Williams, K., Pursel, B., Brautigam, B., Saul, S., Williams, H., Bowen, K., Giles, C.L.: Concept hierarchy extraction from textbooks. In: Proceedings of the 2015 ACM Symposium on Document Engineering. pp. 147–156 (2015)

[30] Wang, S., Liu, L.: Prerequisite concept maps

extraction for automatic assessment. In: Proceedings of the 25th International Conference Companion on World Wide Web. pp. 519–521 (2016)

[31] Yang, Y., Liu, H., Carbonell, J., Ma, W.: Concept graph learning from educational data. In: Proceedings of the Eighth ACM International Conference on Web Search and Data Mining. pp. 159–168. WSDM '15, Association for Computing Machinery, New York, NY, USA (2015).
https://doi.org/10.1145/2684822.2685292,
https://doi.org/10.1145/2684822.2685292

[32] Yu, R., Gadiraju, U., Holtz, P., Rokicki, M., Kemkes, P., Dietze, S.: Predicting user knowledge gain in informational search sessions. In: The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval. pp. 75–84. SIGIR '18, Association for Computing Machinery, New York, NY, USA (2018).
https://doi.org/10.1145/3209978.3210064,
https://doi.org/10.1145/3209978.3210064

Fine-tuning Llama-2 Towards Power-Affirming Automated Feedback on Student Writing

Mei Tan
Stanford University
Graduate School of Education
Stanford, USA
mxtan@stanford.edu

Christopher Mah
Stanford University
Graduate School of Education
Stanford, USA
chrismah@stanford.edu

Dorottya Demszky
Stanford University
Graduate School of Education
Stanford, USA
ddemszky@stanford.edu

ABSTRACT

Feedback is fundamental to students' experiences and development as independent writers. When feedback invites and legitimizes students' ideas, it can encourage revision while empowering students to take ownership of their writing. However, studies have shown that such power-affirming feedback is not common practice, even among expert teachers. Consequently, automated feedback from large language models (LLMs) persist and exacerbate the same established norms that center the teacher's authority. In this work-in-progress report, we work towards improving LLM-generated automated feedback. We collect two datasets of inline feedback comments on middle and high school student essays by experienced English Language Arts (ELA) teachers. We fine-tune Llama-2 to better align generated outputs with the linguistic and pedagogical actions of high-quality teacher-written feedback, in response to a writing prompt and a student essay. We evaluate the structure, readability, specificity, and semantic focus of teacher-written and automated feedback to characterize the effects of fine-tuning and identify the dimensions along which its resulting models more closely approximate human performance. We find that fine-tuned models adopt the conversational language of human-written feedback, using first person pronouns, asking questions, and implementing praise. Fine-tuning effectively improves the frequency of generating power-affirming feedback, at a rate similar to that of experienced teachers.

Keywords

feedback, large language models, natural language processing, student agency

1. INTRODUCTION

Teacher feedback is fundamental to guiding student learning [6], encouraging engagement and revision [22], and empowering students to take ownership of their writing [5]. However, systemic pressures increasing demands on teachers, combined with recent advances in large language models (LLMs), have led many practitioners to experiment with generative AI tools to automate feedback writing [4]. Though automated feedback from large language models are increasingly incorporated in educational technology products [10], research has shown that these models lack pedagogical expertise [23] and instructional reliability.

In our prior work, we find that ChatGPT-generated feedback is linguistically distinct from teachers' feedback and is significantly less power-affirming. Power-affirming feedback legitimizes students' ideas and positions students as authors in the writing process, while power-concealing feedback reinforces the authority of the teacher [15]. We find that because power-affirming feedback is not common practice among teachers, automated feedback from large language models (LLMs) persist and exacerbate the same established norms that center the teacher's authority [17, 7], even with careful prompt-tuning.

In this work-in-progress report, we work towards improving LLM-generated automated feedback. We collect two datasets of inline feedback comments on middle and high school student essays by experienced English Language Arts (ELA) teachers. We fine-tune Llama-2 to better align generated outputs with the linguistic and pedagogical actions of high-quality teacher-written feedback, in response to a writing prompt and a student essay. We evaluate the structure, readability, specificity, and semantic focus of teacher-written and automated feedback to characterize the effects of fine-tuning and identify the dimensions along which its resulting models more closely approximate human performance.

2. RELATED WORK

Research about automated feedback generation has historically developed alongside efforts in automated writing evaluation [25]. Researchers and educational technology developers have automated feedback via rule-based scripts evaluating in-text citations, word choice, and grammatical errors, and lexical methods for identifying transition terms, long sentences, pronoun use, and topic development [1, 2]. Others have used a combination of lexical features, such as word count, specificity, and coherence, to model and predict the scoring of students' use of evidence and organization of claims in accordance with a rubric [16]. Feedback is then automated by associating teacher-written pre-defined messages with rubric evaluation items [25].

As large language models (LLMs) emerge as a potential solution for automated feedback, researchers have studied prompting strategies for feedback generation [18]. Subsequent content analyses comparing LLM-generated feedback and teacher-written feedback has yielded mixed results across contexts. While some studies report that LLMs can approximate human feedback without specialized training [19, 3, 13], others have found such generated feedback to be of low quality, abstract and generic, often failing to provide concrete suggestions [24, 18]. We extend this work by studying the efficacy of fine-tuning LLMs, characterizing the resulting automated feedback.

3. DATA

We develop a digital interface through which teachers provide inline feedback pairs–a highlighted excerpt and an associated feedback comment–in response to middle and high school English Language Arts (ELA) essays and writing prompts sampled from two public datasets. Teachers are instructed to treat the essays as first drafts due for revision and to provide feedback following their usual practices.

ASAP-AES. We collect a dataset of 1,653 inline feedback pairs written by 20 experienced teachers ($M = 8.25$ years) recruited from the alumni network of a selective professional development fellowship. The feedback pertains to 207 persuasive, narrative, and literary analysis essays from the ASAP-AES dataset [9].

PESUADE. We collect a dataset of 1,163 inline feedback pairs written by 40 teachers of mixed experience levels recruited from three sites of a national professional network of ELA teachers. The feedback pertains to 174 persuasive and literary analysis essays from the PERSUADE dataset [20].

We create a training dataset for fine-tuning by sampling 152 essays from our annotated ASAP-AES feedback dataset. For each essay and writing prompt, we sample all combinations of three teacher-written feedback pairs, resulting in 4,580 training examples. We create two evaluation datasets: the remaining 55 essays and 385 feedback pairs in our annotated ASAP-AES feedback data, and our full annotated PERSUADE feedback data.

4. METHODS

4.1 Model Development

We propose a fine-tuned model that operates on a student essay and a writing prompt, formatted within system instructions, and produces inline feedback formatted as a set of three to five excerpt-comment feedback pairs. Each pair consists of a short extracted segment of the student's essay (the excerpt) and an associated feedback text (the comment). We build upon the existing large-scale pre-training of the 7 billion parameter Llama-2 model [21]. We apply supervised learning on a constructed dataset of essays and writing prompts formatted within system instructions as inputs and sets of three feedback pairs as outputs. We fine-tune the model using SFTTrainer from the TRL library [8]. We implement QLoRA for parameter-efficient fine-tuning, and train for two epochs, using a batch size of 1 and 4 gradient accumulation steps. Through this method, the model learns mappings between input essays and writing prompts and a set of generated inline feedback comments.

4.2 Model Evaluation

Using essays in our two evaluation datasets, we generate feedback using a baseline Llama-2 model and our fine-tuned model.

System instructions encasing the input essay and writing prompt.

```
You are my English teacher. Read my essay and assignment:
###Essay: ”’{essay}”’
###Assignment: ”’{prompt}”’
Give me feedback to help me revise. Extract three to five
short excerpts from my essay and give me feedback on those.
List the excerpts and feedback like this:
1. ***[excerpt]—[feedback]
2. ***[excerpt]—[feedback]
3. ***[excerpt]—[feedback]
```

We examine and compare the lexical and semantic features of teacher-written feedback and that generated by baseline and fine-tuned Llama-2 models.

Structure. We calculate the word count and sentence count for each feedback comment.

Readability. We calculate the Flesch-Kincaid Reading Ease score for each feedback comment, with higher scores corresponding to texts that are easier to read [11]. We additionally calculate the type-token ratio, which compares the number of unique words to the total number of words, with higher ratios representing richer vocabulary use.

Specificity. We calculate content word density for each feedback comment, which represents the proportion of content words (nouns, verbs, adjectives, and adverbs) relative to the total number of words, with higher densities indicating greater informativeness or specificity. We additionally calculate two measures of uptake considering the specificity of each feedback comment relative to its associated excerpt. Uptake measured as overlap represents the proportion of words in the excerpts that also occur in the feedback comment [14]. Uptake measured as similarity is calculated by converting both excerpt and feedback comment texts into BERT vector representations and obtaining their cosine similarity.

Engagement. We count the occurrence of first person, second person, and first person plural pronouns in each feedback comment and aggregate the proportion of comments that represent each pronoun use.

Semantic Focus. We count the occurrence of questions involving wh- words (e.g. *who*, *what*, *when*) or auxiliary verbs (e.g. *did*, *is*, *are*) in each feedback comment and aggregate the proportion of comments that represent each question type. We additionally use RoBERTa-based [12] classifiers developed in prior work to predict for each feedback comment whether it belongs to several classes of feedback acts. The *non-dialogic* classifier identifies comments which do not substantively engage with the content of the essay, but instead mark grammar and mechanics, often comprising a single word or symbol rather than full communicative event (f1 ∼87%). The *non-revision-oriented* classifier identifies comments which offer praise, reactions, or other commentary that do not encourage a revision (f1 ∼82%). The *praise* classifier identifies the subset of non-revision-oriented feedback comments that are purely praise (f1 ∼81%). We aggregate the proportion of feedback comments that represent each classified act.

Power-Affirming. We apply a RoBERTa-based [12] regression model developed in prior work to predict for each feedback comment the degree to which it is power-affirming (Spearman ρ ∼81%). The model returns a scalar value between 0 and 1, where higher values are more power-affirming. Power-affirming scores (PA Score) are only calculated for comments that are classified as dialogic and revision-oriented.

5. RESULTS

We report the comparison of linguistic features in teacher-written and generated feedback in Table 1. We find that the model trained via supervised fine-tuning successfully matches the style of teacher-written feedback. While the baseline Llama-2 model generates longer feedback comments using more complex language (a Flesch-Kincaid score of 50 to 60 indicates text that is fairly difficult to read), the fine-tuned model adopts the shorter phrases and simpler vocabulary of human-written feedback. This simplification is associated with decreased specificity across all three measures, influenced by the lower values in teacher-written feedback. We note that fine-tuning introduces some instability–likely due to our memory-efficient training methods–resulting in 5.44% of generated comments to appear incoherent or incomplete under human evaluation.

The fine-tuned model learned to engage students using first person pronouns, taking up phrases like *"I think"* and *"it seems to me"*, but did not change its more limited use of second and first person plural pronouns. Further, fine-tuning significantly increased the prevalence of questions in generated feedback. We illustrate these stylistic changes through qualitative examples in Table 2.

We additionally find that fine-tuning diversifies the pedagogical acts targeted by automated feedback to more closely approximate those in teacher-written feedback. As illustrated in Figure 1, the baseline Llama-2 model generates dialogic and revision-oriented feedback comments. Meanwhile, teacher-written feedback includes warnings about grammar and mechanics, personal reactions, praise, and general commentary about the writing process. Fine-tuning enabled the model to reproduce some of these feedback acts, though the model predominantly adopted one-word spelling corrections and praise like *"good point!" and "good topic sentence".*

Fine-tuning significantly improved the distribution of PA-scores among dialogic and revision-oriented comments. While the majority of feedback comments generated by the baseline model are limited to a narrow and more power-concealing range of PA-scores, the fine-tuned model regularly generates feedback with a broader and more power-affirming range of scores, comparable to those of teacher-written feedback. Figure **??** illustrates the improvement in PA-scores across both evaluation datasets. Note that the datasets were collected from two different groups of teachers, contributing to variation in the PA-score distributions in teacher-written feedback.

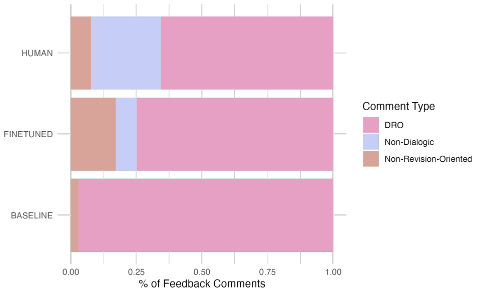


Figure 1: Proportion of feedback comments that are non-dialogic, non-revision-oriented, and dialogic and revision-oriented (DRO).

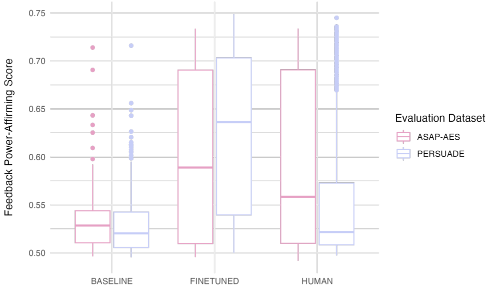


Figure 2: Distribution of PA-scores for teacher-written and LLM-generated feedback for each evaluation dataset.

However, we find that the improvements in PA-scores do not occur uniformly but instead moderately favor higher-scoring essays. Figure 3 illustrates the relationship between student essay score and the PA-scores of the associated feedback. While teacher-written and baseline model-generated feedback demonstrate more uniformly distributed PA-scores across essays, the fine-tuned model generates comments with higher PA-scores for higher-scoring essays.

6. DISCUSSION

Our initial fine-tuned model serves as a foundation for further examination of the adapted capabilities of LLMs in generating automated feedback. By aligning LLM-generated feedback with that of experienced teachers, we can signifi-

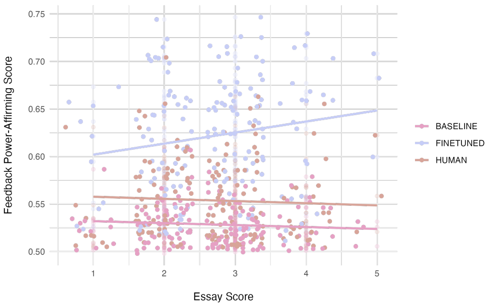


Figure 3: Relationship between essay score and power-affirming scores of feedback.

Table 1: Linguistic features of teacher-written feedback comments and those generated by language models. Bolded values indicate dimensions by which feedback from the finetuned model more closely approximates human-written feedback.

Measure	HUMAN	BASELINE	FINETUNED
Structure			
Avg. Word Count	12.678 (12.238)	30.289 (11.327)	**18.100** (10.872)
Avg. Sentence count	1.543 (0.914)	1.879 (0.604)	**1.578** (0.774)
Readability			
Flesch-Kincaid Ease	69.720 (41.283)	56.540 (16.584)	**78.123** (24.410)
Type-Token Ratio	0.945 (0.079)	0.867 (0.085)	0.921 (0.083)
Specificity			
Content Word Density	0.572 (0.225)	0.524 (0.056)	0.492 (0.115)
Uptake (Overlap)	0.174 (0.259)	0.377 (0.200)	0.214 (0.196)
Uptake (Similarity)	0.598 (0.144)	0.735 (0.080)	0.665 (0.124)
Engagement			
% First Person Pronouns	12.295	1.719	**13.328**
% Second Person Pronouns	38.779	27.429	27.859
% First Person Plural Pronouns	4.041	1.032	1.806
Semantic Focus			
% Wh- Questions	12.898	1.376	**9.630**
% Yes/No Questions	8.856	0.860	**8.684**
% Non-Dialogic	26.139	0.000	1.634
% Non-Revision-Oriented	5.675	1.203	8.083
% Praise	5.589	0.946	**8.083**

cantly improve the prevalence of power-affirming langauge in automated feedback. Fine-tuned models adopt several linguistic markers of power-affirming language, asking questions, using student-friendly conversational language, and using first person pronouns. However, the model also takes on a loss of specificity, fragmented sentences that simply draw attention to a spelling mistake, perfunctory praise, off-topic questions, and personal reactions that may seem facetious coming from an LLM. Fine-tuning does not uniformly promote the complex range of pedagogical actions that teachers balance when giving feedback, and we demonstrate this selective learning. In ongoing work, we build upon this and other fine-tuned models using reinforcement learning techniques to reward power-affirming, dialogic, revision-oriented, and other pedagogically motivated forms of feedback. In doing so, we continue to grapple with the multiple objectives in providing feedback that often mutually interfere.

In other ongoing work, we consider the cost-effectiveness of fine-tuning methods on large models. The lower-quality outputs are noticeable when training using parameter efficient fine-tuning methods to enable training on a single GPU. We continue to explore the relative gains in training additional parameters with more compute resources.

This work has direct implications for the quality of automated feedback tools and facilitating teachers' feedback practices. Given that power-affirming feedback is not common practice, even from experienced teachers, this and future models can be applied in feedback generation and rewrite tasks. We plan to partner with districts and teacher education programs to explore opportunities to share our research and tools in designing professional development focused on writing feedback.

Table 2: Qualitative examples of teacher-written feedback comments and those generated by language models.

Model	Excerpt	Feedback	PA-Score
Baseline	The challenge of visiting Venus has value and we get to see how it look like from the insight.	The sentence should be ”The challenge of visiting Venus has value, and we get to see how it looks from the inside.” The word ”look” should be changed to ”looks,” and the phrase ”from the insight” is incorrect. It should be ”from the inside.”	0.496
Baseline	I believe if your part of the Seagoing Cowboys you'll have the time of your life.	This sentence could be rephrased to make it more persuasive. Instead of simply stating your belief, try to appeal to the reader's sense of adventure and desire for new experiences.	0.503
Baseline	When you look at the Mona Lisa it brings a smile to your face. It shows you how much a computer can do.	This sentence could benefit from more context. Why does the Mona Lisa bring a smile to people's faces? How does it demonstrate the capabilities of computers? Providing more details or examples would help to strengthen this argument.	0.626
Fine-tuned	you also help people in need for the animals for food, and milk.	You can clarify this by adding a direct quote from the text. For example, you could say, ”According to the text, you also help people in need by providing food and milk for their animals.”	0.510
Fine-tuned	The increased amount of work and responsibility, combined with managing time efficiently, can be enough to overwhelm student to the point where they get stressed.	Good point.	Praise, Non-Revision-Oriented
Fine-tuned	A classroom computer could recognize when a student is becoming bored or confused.	This is a very specific and tangible value of this technology. Can you explain more about how this could help students and teachers?	0.731
Fine-tuned	intimidating at first, I helped by serving those who didn't have the same gifts I am blessed with.	How did you feel about helping others during this experience?	0.742
Human	lamborghini or ferrari	Capitalize ”Lamborghini” and ”Ferrari.”	Non-Dialogic
Human	ans	Spelling	Non-Dialogic
Human	First,	try to avoid starting all paragraphs sequentially like this. Just begin your argument	0.502
Human	Studies have shown that students that have anxiety are prone to be unable to focus during class.	I like this idea. I think you could even add more to this paragraph to really make your point that anxiety should not get in the way of achievement in school.	0.583
Human	loser bond with them	Great point. What does this mean for teachers and students? How would this help each of them? What consequences do you see here?	0.729

7. REFERENCES

[1] J. Burstein, M. Chodorow, and C. Leacock. Automated essay evaluation: The criterion online writing service. *Ai magazine*, 25(3):27–27, 2004.

[2] J. Burstein, N. Elliot, B. B. Klebanov, N. Madnani, D. Napolitano, M. Schwartz, P. Houghton, and H. Molloy. Writing mentor: Writing progress using self-regulated writing support. *Journal of Writing Analytics*, 2:285–313, 2018.

[3] W. Dai, J. Lin, H. Jin, T. Li, Y.-S. Tsai, D. Gašević, and G. Chen. Can large language models provide feedback to students? a case study on chatgpt. In *2023 IEEE International Conference on Advanced Learning Technologies (ICALT)*, pages 323–325. IEEE, 2023.

[4] L. Ferlazzo. Opinion: How teachers are using chatgpt in class. *Education Week*, 07 2023. Accessed: 2024-03-02.

[5] C. M. Griffiths, L. Murdock-Perriera, and J. L. Eberhardt. "can you tell me more about this?": Agentic written feedback, teacher expectations, and student learning. *Contemporary Educational Psychology*, 73:102145, 2023.

[6] J. Hattie and H. Timperley. The power of feedback. *Review of educational research*, 77(1):81–112, 2007.

[7] G. Hillocks. *The testing trap: How state writing assessments control learning.* Teachers College Press, 2002.

[8] HuggingFace. Trl: Training a reward model, fine-tuning, and inference. https://github.com/huggingface/trl. Accessed: 2024-06-11.

[9] Kaggle. ASAP Automated Student Assessment Prize - AES. https://www.kaggle.com/c/asap-aes/data, 2013. Accessed: 2024-02-15.

[10] R. Kelly. Khan academy cuts district price of khanmigo ai teaching assistant, adds academic essay feature. *THE Journal*, 11 2023. Accessed: 2024-02-03.

[11] J. P. Kincaid, R. P. Fishburne Jr, R. L. Rogers, and B. S. Chissom. Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel. 1975.

[12] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov. Roberta: A robustly optimized bert pretraining approach, 2019.

[13] J. Meyer, T. Jansen, R. Schiller, L. W. Liebenow, M. Steinbach, A. Horbach, and J. Fleckenstein. Using llms to bring evidence-based feedback into the classroom: Ai-generated feedback increases secondary students' text revision, motivation, and positive emotions. *Computers and Education: Artificial Intelligence*, 6:100199, 2024.

[14] G. A. Miller and J. Beebe-Center. Some psychological methods for evaluating the quality of translations. *Mech. Transl. Comput. Linguistics*, 3(3):73–80, 1956.

[15] J. Pedersen. Revision as dialogue: Exploring question posing in writing response. *Journal of Adolescent amp; Adult Literacy*, 62(2):185–194, June 2018.

[16] Z. Rahimi, D. Litman, R. Correnti, E. Wang, and L. C. Matsumura. Assessing students' use of evidence and organization in response-to-text writing: Using natural language processing for rubric-based automated scoring. *International Journal of Artificial Intelligence in Education*, 27(4):694–728, 2017.

[17] J. Rosa and C. Burdick. Language ideologies. *The Oxford handbook of language and society*, pages 103–123, 2017.

[18] M. Stahl, L. Biermann, A. Nehring, and H. Wachsmuth. Exploring llm prompting strategies for joint essay scoring and feedback generation. *arXiv preprint arXiv:2404.15845*, 2024.

[19] J. Steiss, T. Tate, S. Graham, J. Cruz, M. Hebert, J. Wang, Y. Moon, W. Tseng, M. Warschauer, and C. B. Olson. Comparing the quality of human and chatgpt feedback of students' writing. *Learning and Instruction*, 91:101894, 2024.

[20] The Learning Agency Lab. The persuade dataset, 2024. Accessed: 2024-07-01.

[21] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale, D. Bikel, L. Blecher, C. C. Ferrer, M. Chen, G. Cucurull, D. Esiobu, J. Fernandes, J. Fu, W. Fu, B. Fuller, C. Gao, V. Goswami, N. Goyal, A. Hartshorn, S. Hosseini, R. Hou, H. Inan, M. Kardas, V. Kerkez, M. Khabsa, I. Kloumann, A. Korenev, P. S. Koura, M.-A. Lachaux, T. Lavril, J. Lee, D. Liskovich, Y. Lu, Y. Mao, X. Martinet, T. Mihaylov, P. Mishra, I. Molybog, Y. Nie, A. Poulton, J. Reizenstein, R. Rungta, K. Saladi, A. Schelten, R. Silva, E. M. Smith, R. Subramanian, X. E. Tan, B. Tang, R. Taylor, A. Williams, J. X. Kuan, P. Xu, Z. Yan, I. Zarov, Y. Zhang, A. Fan, M. Kambadur, S. Narang, A. Rodriguez, R. Stojnic, S. Edunov, and T. Scialom. Llama 2: Open foundation and fine-tuned chat models, 2023.

[22] J. S. Underwood and A. P. Tregidgo. Improving student writing through effective feedback: Best practices and recommendations. *Journal of Teaching Writing*, 22(2):73–98, 2006.

[23] R. E. Wang, Q. Zhang, C. Robinson, S. Loeb, and D. Demszky. Step-by-step remediation of students' mathematical mistakes. *arXiv preprint arXiv:2310.10648*, 2023.

[24] S.-Y. Yoon, E. Miszoglad, and L. R. Pierce. Evaluation of chatgpt feedback on ell writers' coherence and cohesion. *arXiv preprint arXiv:2310.06505*, 2023.

[25] H. Zhang, A. Magooda, D. Litman, R. Correnti, E. Wang, L. Matsmura, E. Howe, and R. Quintana. erevise: Using natural language processing to provide formative feedback on text evidence usage in student writing. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 9619–9625, 2019.