Board 62: Work in progress: A Comparative Analysis of Large Language Models and NLP Algorithms to Enhance Student Reflection Summaries

Dr. Ahmed Ashraf Butt, Carnegie Mellon University

Ahmed Ashraf Butt has recently completed his Ph.D. in the School of Engineering Education at Purdue University, where he cultivated a multidisciplinary research portfolio bridging learning science, Human-Computer Interaction (HCI), and engineering education. His primary research focuses on designing and developing educational technologies that facilitate different student learning aspects (e.g., engagement). Further, he is interested in designing instructional interventions and exploring their relationship with different aspects of first-year engineering (FYE) students' learning (e.g., motivation and learning strategies). Before Purdue University, he worked as a lecturer at the University of Lahore, Pakistan. Additionally, he has been associated with the software industry in various capacities, from developer to consultant.

Eesha tur razia babar, University of California, Irvine

Eesha Tur Razia Babar holds a master's degree in Electrical and Computer Engineering from the University of California, Irvine. She completed her undergraduate studies in Electrical Engineering at the University of Engineering and Technology in Lahore, Pakistan. Her primary research interests include educational technology, educational data mining, and educational data science.

Dr. Muhsin Menekse, Purdue University, West Lafayette

Muhsin Menekse is an Associate Professor at Purdue University with a joint appointment in the School of Engineering Education and the Department of Curriculum & Instruction. Dr. Menekse's primary research focuses on exploring K-16 students' engagement and learning of engineering and science concepts by creating innovative instructional resources and conducting interdisciplinary quasi-experimental research studies in and out of classroom environments. Dr. Menekse is the recipient of the 2014 William Elgin Wickenden Award by the American Society for Engineering Education. He is also selected as an NSF SIARM fellow for the advanced research methods for STEM education research. Dr. Menekse received four Seed-for-Success Awards (in 2017, 2018, 2019, and 2021) from Purdue University's Excellence in Research Awards programs in recognition of obtaining four external grants of $1 million or more during each year. His research has been generously funded by grants from the Institute of Education Sciences (IES), the U.S. Department of Defense (DoD), Purdue Research Foundation (PRF), and the National Science Foundation (NSF).

Ali Alhaddad, Purdue University, West Lafayette

Work in progress: A Comparative Analysis of Large Language Models and NLP Algorithms to enhance Student Reflection Summaries

Abstract

The advent of state-of-the-art large language models has led to remarkable progress in condensing enormous amounts of information into concise and coherent summaries, benefiting fields like education, health, and public policy, etc. This study contributes to the current effort by investigating two NLP approaches' effectiveness in summarizing students' reflection text. This approach includes Natural Language Processing (NLP) algorithms customized for summarizing students' reflections and ChatGPT, a state-of-the-art large language model. To conduct the study, we used the CourseMIRROR application to collect students' reflections from s sections of the engineering course at a large Midwestern university. Over the semester, students were asked to reflect after each lecture on two aspects of their learning experience, i.e., what they found 1) interesting and 2) confusing in the lecture? In total, we collected reflections from 42 lectures, and the average class size was 80 students in each section. To inform the study, we generated a reflection summary for all reflection submissions in each lecture using both NLP approaches and human annotators. Furthermore, we evaluated the quality of reflection summaries by assessing the ROUGE-N measure for each lecture's reflection summary generated by all three approaches. These summaries were then aggregated for each approach by averaging different metrics of ROUGE scores. Subsequently, we see the differences between the average ROUGE scores of the two NLP approaches and human-generated reflection summaries. Preliminary findings suggest that NLP algorithms outperformed ChatGPT in creating human-like reflection summaries. This finding implies that, despite being trained on a large corpus of textual data, the prominent large language model ChatGPT still requires improvements to surpass or match the performance of NLP algorithms tailored for solving custom problems.

Introduction:

In the field of education, there has been a consistent generation of qualitative datasets such as students' reflections, assessments, discussions, feedback, and more. This dataset can potentially enhance educational outcomes by providing insights into the students' learning process and teachers' pedagogical practices [1]. Traditionally, analyzing this qualitative data has been labor-intensive, relying on manual methods using human experts [2]. This reliance on manual analysis often limits the ability to process this data on time and fully utilize its potential to inform our educational practices. However, technological advancements, particularly in Natural Language Processing (NLP), have revolutionized how we handle and process data by introducing efficient, automated ways to extract insights and provide valuable insights from this qualitative educational dataset [3]. As a result, there has been huge interest by educators and researchers to fully utilize these powerful tools to extract the insights in these datasets timely, and then use them to make evidence-based decisions in education.

In this regard, one of the aspects that has seen an enormous shift is the summarization of the educational data e.g., [4 - 5]. Prior studies have predominantly used two summarization approaches such as extractive (selecting and rearranging existing text[6]) and abstractive (generating new condensed sentences[7]) to summarize and extract meaningful insights from

diverse academic data sources [8]. In line with this, different education data sources have been used to create summaries, such as student reflections (e.g., [9]), discussion forum transcripts (e.g., [10]), and course materials (e.g., [11]). This diverse source has helped students to better self-regulate their learning, access the study material (e.g., [12]), and help teachers to make timely changes in their pedagogical practices(e.g., [13]).

In recent years, Generative Pre-trained Transformer (ChatGPT), a large language model, has revolutionized the NLP field and emerged as a powerful tool for summarizing, sentiment analysis, and language translation [14]. This model has been trained with a large corpus of data, enabling it to generate human-like responses based on learned patterns and contexts. Even though it has shown promising results in the general summarization of the qualitative dataset [15], there is still a question on its ability to generate summaries within a context [16]. In other words, ChatGPT's ability to capture the nuanced and subjective nature of a task compared to summarization algorithms specifically designed to achieve a particular task still needs further investigation. In this regard, this study aims to bridge the gap by assessing the effectiveness of the traditional NLP approach and the capabilities of ChatGPT in performing a similar summarization task. In our study, we used ChatGPT and an NLP algorithm trained on previous students' reflection datasets to generate summaries from students' reflections submitted for each lecture in the classroom. Then, we assess to what extent the generated summaries are similar to human-annotated reflection summaries. More specifically, our research study is guided by the research question: To what extent do ChatGPT and traditional NLP algorithms generate reflection summaries similar to human-annotated summaries?

Literature Review

In education research, qualitative data is everywhere. This data can be students' responses to open-ended survey questions about pre-requisite preparation for a course, students' discussions on an online learning platform, students' assignments, or their reflections about learning experiences in a particular course. Analysis of this qualitative data can provide valuable insights into teaching and learning practices in schools and help improve these practices for effective teaching and better learning [1], [17], [18]. However, often, this qualitative data is present at scale. The manual analysis process of this data without the help of any automated process can be complicated and slow. Natural language processing (NLP) comes to the rescue in this situation and helps condense this large amount of educational data to find critical information and patterns efficiently [5]. The process of reducing the text while retaining the essential information is known as text summarization.

In the education literature, different NLP algorithms for text summarization have been employed that can be categorized into two broad approaches, i.e., the Extractive and abstractive approaches. In the extraction summarization approach, algorithms (e.g., TextRank, LexRank, and Latent Semantic Analysis [19]) create summaries by identifying and extracting important sentences based on the Bayesian statistic approaches and language modeling. For instance, andhin et al. utilized the extractive summarization algorithm to generate summaries of articles to facilitate the students reading difficulties [20]. In the abstractive summarization approach, algorithms (e.g., Transformer-based models such as BERT [12]) generate new summaries by understanding the original content and then paraphrasing it to create a coherent summary. For instance, Benedetto et al. [21] used the

abstractive approach to summarize video lectures to aid learners, particularly in challenging contexts or for those with special needs. Their results showed the efficacy of this approach in producing fluent summaries. Similarly, studies [6] have compared both approaches in the text summarization task and have discussed that the selection of one over the other depends on various factors, such as the nature of the task, abstraction level, or the specific task objectives. Furthermore a significant challenge has been discussed for these approaches is the requirement for large training corpora of human-written summaries, which may limit their effectiveness [22].

With the advent of ChatGPT, there is a notable shift in the field of NLP. Researchers have started exploring its potential in academia and industry due to their notable performance in various applications [23]. Educational literature is still novice regarding ChatGPT, whereas, in the border literature, the potential of ChatGPT has been investigated in NLP tasks such as sentiment analysis (e.g., [24]), question-answering(e.g., [25]), and language translation[26] and has shown promising results. Similarly, there have been promising results for the text summarization task, but it has has yet to be explored rigorously [24]. For instance, in a study by Soni and Wade [27], they conducted a human assessment and found reviewers had difficulties differentiating handwritten and ChatGPT-generated summaries. However, in the context of educational literature, Katz et al.[28] showed how both the LLMs and NLP could be used for analyzing unstructured text data to identify themes and patterns in student writing with accuracy. However, they also mentioned that researchers should remain cautious while using LLMs for analyzing text data as LLMs can exhibit unfavorable traits, for example, bias and toxicity behaviors, inherited from their training data. Another key focus of this study is on developing effective methods to generate real-time, human-like reflective summaries. The importance of real-time processing is emphasized in other domains, such as augmented reality in healthcare[29], [30].

Research Method

This study compares an NLP algorithm and ChatGPT's ability to generate a reflection summary close to the human-annotated summary. The evaluation consists of four components: data collection, reflection summary generation, pre-processing system, and evaluation metrics. The reflection is collected through the CourseMIRROR application and fed into two algorithms individually to produce the reflection summaries. Subsequently, their reflections are pre-processed before being evaluated using the human-annotated summary.

Data collection instrument

We used the CourseMIRROR application to collect students' reflections in first-year engineering courses over a semester. Each section had almost 80 students who used this app to reflect on their learning experiences. Students were enrolled in the application at the start of the semester and were prompted to reflect on two open-ended questions after each lecture throughout the semester. The prompts for these questions were: 'What did you find interesting in the lecture? (POI)' and 'What did you find confusing in the lecture? (MP).' Furthermore, the application employs NLP algorithms to scaffold students' reflection writing and create reflection summaries by combining reflections based on common themes [31].

Dataset

The dataset used in the study consists of two sets of reflections, where students reflect on the interesting (POI) and confusing (MP) aspects of the lecture over a semester. Using the CourseMIRROR application, these reflections were collected from seven sections of the first-year engineering course over a semester. The course covered fundamental concepts in computer programming, visualizing data, and formulating solutions for engineering challenges. Overall, the dataset had three data sources: reflection summaries generated by an NLP algorithm, ChatGPT, and a human-annotated algorithm for both POI and MP of the lecture.

Evaluation Metric

The study used the commonly used evaluation measure, i.e., ROUGE (Recall-Oriented Understudy for Gisting Evaluation), to evaluate the ability of NLP algorithms to generate human-like reflection summaries [19]. Within the ROUGE, we used ROUGE 1 (evaluates the overlap of unigrams), 2 (assesses the overlap of bi-grams), and Rouge L (considers the overlap of most extended common sequence). These metrics provide a comprehensive understanding of the similarity between the GPT and NLP algorithm-generated reflection summaries and the human-annotated reflection summaries. The ROGUE score for each metric gives three key results, i.e., Recall Score (the ability of the summary to include all the relevant information present in the reference documents), Precision Score (measures the accuracy of a system by calculating the ratio of correctly identified relevant items to the total number of items specified by the system), and F1 Score (harmonic mean of precision and recall).

NLP Approaches

Following are the two approaches used to generate a reflection summary of students' reflections on both aspects of each lecture in the dataset.

ChatGPT

The advent of the (LLMs) has been at the forefront of the recent AI revolution in the NLP fields. The powerful model (e.g., GPT-3) has shown an unprecedented ability to understand, generate, and context-aware responses in the various language-related task (e.g., chatbot). Seeing its potential, we utilized ChatGPT API with the GPT-3.5 turbo model to generate summaries of the student's reflections. In the ChatGPT API, we send HTTP requests containing students' reflections and prompts with different parameters (e.g., temperature controls the randomness of the responses), and receive a reflection summary as output.

In this experiment, we used multiple versions of prompts and varied the temperature parameter of the API, choosing the one that produced the most sensible reflection summary, as determined by the research team and ROUGE score. The prompts were created through discussions within the research team to produce human-like reflection summaries. Also, the research team refined the prompts until the reflection summary consistently focused on the major topics discussed in across all reflections in a lecture. To control the ChatGPT's diverse response, the temperature parameter of the API was controlled, and different experiments were conducted to determine which one was closer to human-annotated reflection. The parameter's value ranges from 0 to 1,

where 0 produces a more focused response, and 1 produces a more diverse response. Table 1 displays the few tested temperatures:

Table 1. Rouge score to create reflection summary with different temperature score(randomness)

Temperature	Rouge 1			Rouge 2			Rouge LCS		
	Precision	**Recall**	**F1**	**Precision**	**Recall**	**F1**	**Precision**	**Recall**	**F1**
Point of interest – Reflection summary									
0	0.48	0.28	0.34	0.15	0.09	0.11	0.32	0.18	0.22
0.5	0.47	0.27	0.33	0.15	0.09	0.1	0.32	0.18	0.22
1	0.45	0.26	0.31	0.14	0.08	0.1	0.3	0.17	0.2
Muddiest Point– Reflection summary									
0	0.46	0.26	0.32	0.16	0.09	0.11	0.3	0.17	0.21
0.5	0.47	0.26	0.32	0.15	0.08	0.1	0.31	0.17	0.21
1	0.44	0.26	0.31	0.15	0.09	0.1	0.29	0.17	0.21

NLP Algorithm

In our approach, we used an extension of the BERT (Bidirectional Encoder Representations from Transformers) model that is specifically developed for extractive summarization tasks. We used both the original based and fine-tuned version. For training on the CNN/DM news dataset, we used a checkpoint named "bertext_cnndm_transformer" and employed the original codebase to select five reflections. Also, we fine-tuned extension of BERT models on our reflection dataset and the FEWSUMM AMAZON dataset to assess the advantages of our data for summarization tasks. We utilized an off-the-shelf BERT-EXT model to form the candidate sets to condense the original documents into eight reflections. Finally, we crafted a summary for each lecture, consisting of five sentences extracted from these reflections.

System Architecture

The system for evaluation is composed of four unit as mentioned in the beginning of this section.

Reflection collection: The initial phase involved gathering the reflections of students through the CourseMIRROR application.

Generation of reflection summaries: The three datasets of students' reflection summaries were created using the NLP algorithm technique, ChatGPT, and human annotator. The process to create a reflection summary set is explained in the NLP approaches section. For the human-annotated reflection summaries, we employed 11 undergraduate students to summarize the reflections for each lecture. These summaries served as the reference to evaluate the ability of NLP approaches to produce human-like reflection summaries. The students were trained in three batches to achieve interrater agreement before working individually to generate reflection summaries. We measure their agreement using the ROUGE scores, which are ROGUE 1=48.31,

ROGUE 2 = 43.52, and ROGUE L= 43.52, which is an acceptable agreement as discussed in the [7].

Pre-processing: In this step, we used the commonly used NLTK (Natural Language Toolkit) library to clean and filter the stopping words and punctuation from the dataset.

Evaluation: Different metrics of the ROGUE score were used to evaluate the effectiveness of the NLP approaches in producing human-generated summaries.

Analysis & Results

To inform the study, we calculate the ROGUE 1, 2, and L for both the MP and POI reflection summary datasets. The result is shown in the table 2 for POI and for MP in the table 3.

Table 2. POI - ROGUE score performance by NLP algorithms

Algorithms	Evaluation Metrics	Precision	Recall	F1
NLP	Rouge 1	0.55	0.5	0.51
	Rouge 2	0.38	0.36	0.36
	Rouge L	0.4	0.36	0.37
ChatGPT	Rouge 1	0.47	0.27	0.33
	Rouge 2	0.15	0.09	0.10
	Rouge L	0.32	0.18	0.22

Table 3. MP - ROGUE score performance by NLP algorithms

Algorithms	Evaluation Metrics	**Precision**	**Recall**	**F1**
NLP	Rouge 1	0.55	0.47	0.49
	Rouge 2	0.39	0.34	0.36
	Rouge L	0.4	0.34	0.36
ChatGPT	Rouge 1	0.47	0.26	0.32
	Rouge 2	0.15	0.08	0.1
	Rouge L	0.31	0.17	0.21

The result clearly shows that precision, recall, and F1 scores for both POI and MP refection summary datasets of NLP algorithms are higher as compared to the ChatGPT API. In other words, this shows that reflection summaries generated by the NLP algorithm were more similar to the human-annotated reflection summaries compared to the ChatGPT.

Discussion & Conclusion:

This study investigated the ability of ChatGPT and an NLP algorithm to generate reflection summaries similar to human-annotated summaries. To inform our study, we generated reflection summaries using both NLP approaches from the collected student reflections on their learning

experiences in the classroom. To this end, we calculated different metrics of the ROGUE score, including precision, recall, and F1 score. We found that the NLP algorithm consistently outperformed ChatGPT in producing human-like annotated reflection summaries. The robustness of the NLP algorithm may stem from its tailored design for this summarization task, enabling it to better understand and represent the complexities of student reflections, consistent with the literature [32]. Additionally, the NLP algorithm was trained on the ways students write reflections, which captured the students' writing style. In contrast, ChatGPT is mostly trained on writing from the web, which is usually formal.

Overall, these findings contribute to our understanding of summarization methods for students' reflection and also highlight the potential of customized design and development of NLP algorithms to perform a specific task. Also, there is a need to further explore the potential of technological advancement to enhance educational outcomes.

Limitations & Future Directions

As this is an exploratory study, one limitation was that the analysis relied on a single dataset from a single course. This could limit the generalizability of the findings. Therefore, further studies could test this on varied educational datasets. Another limitation is the use of one ROGUE metric, which has been discussed as ineffective when it comes to assessing factual inconsistency [33]. Therefore, other metrics (e.g., SummaC, FactCC and DAE) might be explored to compare the efficiency of both NLP approaches. Furthermore, our follow-up study would be to run a similar experiment with a customized ChatGPT, where we would train the model with the same reflection dataset used for the NLP Algorithm. This approach will allow for a balanced comparison between both NLP approaches.

References:

[1] Innovare, "Using Qualitative Data in Education For Better Student Outcomes," Innovare | Social Innovation Partners. Accessed: Feb. 07, 2024. [Online]. Available: https://innovaresip.com/resources/blog/qualitative-data-in-education-student-outcomes/

[2] J. W. Neal, Z. P. Neal, E. VanDyke, and M. Kornbluh, "Expediting the Analysis of Qualitative Data in Evaluation: A Procedure for the Rapid Identification of Themes From Audio Recordings (RITA)," *American Journal of Evaluation*, vol. 36, no. 1, pp. 118–132, Mar. 2015, doi: 10.1177/1098214014536601.

[3] K. Crowston, E. E. Allen, and R. Heckman, "Using natural language processing technology for qualitative data analysis," *International Journal of Social Research Methodology*, vol. 15, no. 6, pp. 523–543, Nov. 2012, doi: 10.1080/13645579.2011.625764.

[4] P. Pandiaraja, K. B. Boopesh, T. Deepthi, M. Laksmi Priya, and R. Noodhana, "An Analysis of Document Summarization for Educational Data Classification Using NLP with Machine Learning Techniques," in *Applied Computational Technologies*, B. Iyer, T. Crick, and S.-L. Peng, Eds., Singapore: Springer Nature Singapore, 2022, pp. 127–143.

[5] G. Yang, N.-S. Chen, Kinshuk, E. Sutinen, T. Anderson, and D. Wen, "The effectiveness of automatic text summarization in mobile learning contexts," *Computers & Education*, vol. 68, pp. 233–243, Oct. 2013, doi: 10.1016/j.compedu.2013.05.012.

[6] Y. Chen, Y. Ma, X. Mao, and Q. Li, "Multi-Task Learning for Abstractive and Extractive Summarization," *Data Science and Engineering*, vol. 4, no. 1, pp. 14–23, Mar. 2019, doi: 10.1007/s41019-019-0087-7.
[7] A. Magooda, M. Elaraby, and D. Litman, "Exploring Multitask Learning for Low-Resource AbstractiveSummarization," *arXiv preprint arXiv:2109.08565*, 2021.
[8] J. P. Verma and A. Patel, "An Extractive Text Summarization approach for Analyzing Educational Institution's Review and Feedback Data," *International Journal of Computer Applications*, vol. 143, pp. 51–55, 2016.
[9] W. Luo and D. Litman, "Summarizing student responses to reflection prompts," in *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, 2015, pp. 1955–1960.
[10] O. Almatrafi and A. Johri, "Improving MOOCs Using Information From Discussion Forums: An Opinion Summarization and Suggestion Mining Approach," *IEEE Access*, vol. 10, pp. 15565–15573, 2022, doi: 10.1109/ACCESS.2022.3149271.
[11] L. Cagliero, L. Farinetti, and E. Baralis, "Recommending Personalized Summaries of Teaching Materials," *IEEE Access*, vol. 7, pp. 22729–22739, 2019, doi: 10.1109/ACCESS.2019.2899655.
[12] Y. AlRoshdi, M. AlBadawi, A. Alhamadani, and M. Sarrab, "An Extractive Summarization for utilizing Learning Content using Deep Learning algorithm: Proposed Framework and Implementation," *International Journal of Computing and Digital Systems*, vol. 13, no. 1, pp. 461–474, 2023.
[13] X. Fan, W. Luo, M. Menekse, D. Litman, and J. Wang, "CourseMIRROR: Enhancing Large Classroom Instructor-Student Interactions via Mobile Interfaces and Natural Language Processing," in *Proceedings of the 33rd Annual ACM Conference Extended Abstracts on Human Factors in Computing Systems*, in CHI EA '15. New York, NY, USA: Association for Computing Machinery, 2015, pp. 1473–1478. doi: 10.1145/2702613.2732853.
[14] J. Kocoń *et al.*, "ChatGPT: Jack of all trades, master of none," *Information Fusion*, vol. 99, p. 101861, Nov. 2023, doi: 10.1016/j.inffus.2023.101861.
[15] S. Tian *et al.*, "Opportunities and challenges for ChatGPT and large language models in biomedicine and health," *Briefings in Bioinformatics*, vol. 25, no. 1, p. bbad493, Jan. 2024, doi: 10.1093/bib/bbad493.
[16] Z. Luo, Q. Xie, and S. Ananiadou, "Chatgpt as a factual inconsistency evaluator for text summarization." ArXiv, 2023.
[17] A. Magooda, D. Litman, A. A. Butt, and M. Menekse, "Improving the quality of students' written reflections using natural language processing: Model design and classroom evaluation," in *International Conference on Artificial Intelligence in Education*, Springer, 2022, pp. 519–525.
[18] A. A. Butt, S. Anwar, A. Magooda, and M. Menekse, "Comparative analysis of the rule-based and machine learning approach for assessing student reflections," in *Proceeding of International Society of the Learning Sciences (ISLS)*, 2022, pp. 1577–1580.
[19] A. Kumar, A. Sharma, S. Sharma, and S. Kashyap, "Performance analysis of keyword extraction algorithms assessing extractive text summarization," in *2017 International Conference on Computer, Communications and Electronics (Comptelix)*, Jul. 2017, pp. 408–414. doi: 10.1109/COMPTELIX.2017.8004004.

[20] K. Nandhini and S. R. Balasundaram, “Improving readability through extractive summarization for learners with reading difficulties,” *Egyptian Informatics Journal*, vol. 14, no. 3, pp. 195–204, Nov. 2013, doi: 10.1016/j.eij.2013.09.001.
[21] I. Benedetto, M. La Quatra, L. Cagliero, L. Canale, and L. Farinetti, “Abstractive video lecture summarization: applications and future prospects,” *Education and Information Technologies*, Jun. 2023, doi: 10.1007/s10639-023-11855-w.
[22] A. Nenkova and K. McKeown, “Automatic Summarization,” *Foundations and Trends® in Information Retrieval*, vol. 5, no. 2–3, pp. 103–233, 2011, doi: 10.1561/1500000015.
[23] M. U. Hadi *et al.*, “Large language models: a comprehensive survey of its applications, challenges, limitations, and future prospects,” *Authorea Preprints*, 2023.
[24] K. Bu, Y. Liu, and X. Ju, “Efficient utilization of pre-trained models: A review of sentiment analysis via prompt learning,” *Knowledge-Based Systems*, vol. 283, p. 111148, Jan. 2024, doi: 10.1016/j.knosys.2023.111148.
[25] K. Nassiri and M. Akhloufi, “Transformer models used for text-based question answering systems,” *Applied Intelligence*, vol. 53, no. 9, pp. 10602–10635, May 2023, doi: 10.1007/s10489-022-04052-8.
[26] B. D. Lund and T. Wang, “Chatting about ChatGPT: how may AI and GPT impact academia and libraries?,” *Library Hi Tech News*, vol. 40, no. 3, pp. 26–29, 2023.
[27] M. Soni and V. Wade, “Comparing Abstractive Summaries Generated by ChatGPT to Real Summaries Through Blinded Reviewers and Text Classification Algorithms,” *arXiv preprint arXiv:2303.17650*, 2023.
[28] A. Katz, M. Norris, A. M. Alsharif, M. D. Klopfer, D. B. Knight, and J. R. Grohs, “Using natural language processing to facilitate student feedback analysis,” in *2021 ASEE Virtual Annual Conference Content Access*, 2021.
[29] T. Khan *et al.*, “AR in the OR: exploring use of augmented reality to support endoscopic surgery,” in *Proceedings of the 2022 ACM International Conference on Interactive Media Experiences*, in IMX ’22. New York, NY, USA: Association for Computing Machinery, 2022, pp. 267–270. doi: 10.1145/3505284.3532970.
[30] T. Khan *et al.*, “Understanding Effects of Visual Feedback Delay in AR on Fine Motor Surgical Tasks,” *IEEE Transactions on Visualization and Computer Graphics*, vol. 29, no. 11, pp. 4697–4707, Nov. 2023, doi: 10.1109/TVCG.2023.3320214.
[31] M. Menekse, S. Anwar, and S. Purzer, “Self-Efficacy and Mobile Learning Technologies: A Case Study of CourseMIRROR,” in *Self-Efficacy in Instructional Technology Contexts*, C. B. Hodges, Ed., Cham: Springer International Publishing, 2018, pp. 57–74. doi: 10.1007/978-3-319-99858-9_4.
[32] E. H. Park, H. I. Watson, F. V. Mehendale, and A. Q. O’Neil, “Evaluating the Impact on Clinical Task Efficiency of a Natural Language Processing Algorithm for Searching Medical Documents: Prospective Crossover Study,” *JMIR Med Inform*, vol. 10, no. 10, p. e39616, Oct. 2022, doi: 10.2196/39616.
[33] J. Maynez, S. Narayan, B. Bohnet, and R. McDonald, “On Faithfulness and Factuality in Abstractive Summarization,” in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, D. Jurafsky, J. Chai, N. Schluter, and J. Tetreault, Eds., Online: Association for Computational Linguistics, Jul. 2020, pp. 1906–1919. doi: 10.18653/v1/2020.acl-main.173.