Optimizing The Literature Review Process: Evaluating Generative AI Models on Summarizing Undergraduate Data Science Research Papers

Tse Saunders*
tsesaunders@gwu.edu
George Washington University
Washington, DC, USA

Nawwaf Aleisa*
nawwaf.aleisa@gwu.edu
George Washington University
Washington, DC, USA

Juliet Wield*
julietwield@gwu.edu
George Washington University
Washington, DC, USA

Joshua Sherwood*
jsherwood@gwu.edu
George Washington University
Washington, DC, USA

Xiaodong Qu
x.qu@gwu.edu
George Washington University
Washington, DC, USA

ABSTRACT

This analysis provides a preliminary exploration of the potential use of Generative AI (GenAI) by evaluating the summarization ability of what we consider to be some of the most popular GenAI models during May 2024. The models chosen were ChatGPT-4o, Google Gemini, Microsoft Copilot, and Claude 3 Sonnet. We assessed this ability by providing the models with a link or pdf to 5 undergraduate KDD submissions from 2023 with the prompt "Concisely summarize this paper in no more than 3 paragraphs". This process was performed 3 times per model. Their performance was assessed in four different categories: brevity, accuracy, readability, and content relevance. Although none of the models obtained perfect scores, we determined that ChatGPT-4o performed the best overall. The model produced three-paragraph summaries containing no irrelevant content and, on a scale from 0 to 5, it had an average accuracy of 3.05 and an average readability of 4.98. We therefore concluded that, with further advancements, GenAI will likely be able to provide high-quality summaries of research papers. This has the potential to aid in optimizing the literature review process since such summaries may allow researchers to more quickly understand the contents of different papers. Additionally, we hypothesize that future GenAI models may become advanced enough to simulate conversations regarding papers that would normally require the author, which would benefit both the research process and the development of GenAI models. The full data for this analysis can be found at:
https://github.com/jsherwood00/KDD2024-SAWS.git.

*The four undergraduate students are the primary authors and contributed equally to this research. Professor Xiaodong Qu served as the research advisor and mentor.

KDD-UC '24, August 25–29, 2024, Barcelona, Spain

ACM ISBN 978-1-4503-XXXX-X/23/08…$15.00
https://doi.org/XXXXXXX.XXXXXXX

CCS CONCEPTS

• **Human-centered computing** → **Human computer interaction (HCI)**; • **Information systems** → **Evaluation of retrieval results**; • **Computing methodologies** → **Information extraction**.

KEYWORDS

Review, Artificial Intelligence, Accuracy, Content, Readability, University Students, Retrieval, Generative AI.

ACM Reference Format:
Tse Saunders, Nawwaf Aleisa, Juliet Wield, Joshua Sherwood, and Xiaodong Qu. 2024. Optimizing The Literature Review Process: Evaluating Generative AI Models on Summarizing Undergraduate Data Science Research Papers. In *Proceedings of KDD Undergraduate Consortium (KDD-UC '24).* ACM, New York, NY, USA, 7 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

Generative AI (GenAI) has become popular since the release of ChatGPT in late 2022. In broad terms, it refers to models that accept an input and output a response based on a vast amount of training data. This output is nondeterministic, meaning the same input may produce different outputs with different trials. The content of the response varies according to the algorithm used. Some algorithms, such as General Adversarial Networks (GANs)[12] and diffusion models[4] are used to generate images from text descriptions. Others, such as Recurrent Neural Networks (RNNs)[13] and Transformers[35] are used to accomplish state-of-the-art text generation.

Large Language Models (LLMs), such as GPT-3.5, are a group of recently developed GenAI algorithms and have been generally successful at generating human-like responses to any readable prompt provided, within the size and data type limits of the specific model. However, the accuracy can vary greatly, with models sometimes providing inaccurate information in responses[3] and when citing works[45]. LLMs are trained using a version of the transformer model presented by Vaswani et al[35] in 2017 and require vast amounts of data. They also make use of existing techniques, such as comparing produced output to expected output by allowing users to like or dislike the response in order to constantly improve the

model. This along with pre-training on a large dataset of text allows them to better recognize language patterns and semantics and produce better outputs. This skill can be applied to many different tasks, one of which is summarizing the content of an input.

By reading the abstract, conclusion, or a summary provided by the author, readers can quickly understand the main points of a paper. This could potentially be improved by GenAI. Instead of only providing a summary of the paper, as done by authors, models may eventually be able to simulate a conversation with the author about their paper while explaining confusing concepts. This could prove useful to other authors, students, and members of the scientific community, especially considering the drastic increase in the number of research papers in recent years[17, 18]. The first step in achieving this is evaluating models on their ability to provide an adequate summary of research papers.

Achieving this goal presents some problems because some loss in representing a paper's ideas is inevitable. Different fields of work and areas of research may demand different details. As a result, training a GenAI model to perform a summarization task well such that all authors and readers generally agree with the evaluation of the summary provided by the model may prove difficult. In addition, determining objective evaluation criteria for what a summary should include is difficult because summary quality is subjective. Previous research provides some guidelines to resolve this, which will be further discussed in the related work section.

1.1 Research Questions

The goal of this analysis is to address whether GenAI can effectively summarize undergraduate data science research papers, focusing on the following questions:

(1) What criteria should be used to evaluate the performance of GenAI models in summarizing academic research papers?
(2) How do the summarization performances of ChatGPT-4, Google Gemini, Microsoft Copilot, and Claude 3 Sonnet compare? In which specific areas does each model excel or need improvement?
(3) What are the broader implications of using GenAI models for summarizing research papers on data collection and the field of data science?

2 RELATED WORK

2.1 Literature Review

Previous research has extensively surveyed the topic of large language models. Notable examples include Zhao et al.'s comprehensive survey on large language models, and Chang et al.'s detailed evaluation of these models [6, 14, 44]. Additionally, there is a growing body of work focusing on the application of large language models in specific domains, such as education, healthcare, music, and coding[5, 11, 28, 37, 40].

Chang et al.'s work[6] is particularly relevant to the content of this analysis due to the detailed exploration of how to evaluate GenAI model performance. The paper discusses the various applications of GenAI including summarization, identifies possible evaluation benchmarks, and provides two methods for evaluating GenAI model quality. The first is automatic evaluation, which involves the use of standard metrics and evaluation tools to determine the accuracy, calibrations, fairness, and robustness of a model's performance. They define accuracy as "how correct a model is on a given task", calibration as "the degree of agreement between the confidence level of the model output and the actual prediction accuracy", fairness as bias against different groups, and robustness as success rate "in the face of various challenging inputs"[6]. The second approach to evaluation is human evaluation, which is noted to sometimes have high variance and instability, but is more reliable when evaluating open-generation tasks since humans can provide qualitative analyses that are more useful than simple statistics. No paper summarization task is included in its list of existing LLM evaluation benchmarks, suggesting the need for research in this area, especially since the paper's summary of current LLM weaknesses includes information fabrication.

2.2 Generative AI in Research Settings

There have been several papers that mention the potential use of generative AI for research, writing, or data analytics[7, 33, 34]. They note the potential use of GenAI as a writing aid due to its ability to generate or correct text input. They also propose that it could serve as a guide, providing students with relevant research papers and organizing relevant data for easy reference. Despite identifying the potential of GenAI, there exists a gap in the existing literature when it comes to experiments that test the limits of such potential. In addition, most papers concentrated on the various iterations of ChatGPT, and few studies compared performance between different GenAI models.

2.3 Automatic Text Summarization

The focus of this analysis is automatic text summarization which is defined as "the task of producing a concise and fluent summary while preserving key information content and overall meaning"[2]. Initial research in the 1950s focused on using the most popular words and phrases in scientific papers to provide a summary[2]. Popular modern approaches can be classified into two categories: abstractive and extractive. The extractive approach chooses sentences from the input, whereas the abstractive is not limited to sentences from the input. Because of this, the abstractive approach has more output possibilities but is generally more complex to implement[36]. For abstractive summaries, guidelines suggest evaluating based on several factors, including only having content discussed in the input document, covering the most significant information in the input document, minimizing redundant information, and having coherent text, as per [15]. Recent research uses various machine learning approaches and focuses on these criteria.

3 METHODS

The results of this analysis are based on the performance of four GenAI models: ChatGPT-4o, Google Gemini, Copilot, and Claude 3 Sonnet. The summarizing capabilities of each model were tested on five undergraduate papers accepted at the KDD 2023 Undergraduate Consortium[8, 10, 26, 27, 38]. When selecting our papers from this conference, we used two criteria:

Rating	Accuracy	Readability
0	Contains less than 10% of significant points	Incoherent/Repetitive, poor writing or language
1	Covers between 10% and 24% of significant points in the paper	Difficult to follow, many confusing or unclear sentences
2	Covers between 25% and 49% of significant points in the paper	Somewhat difficult to follow, requires more effort to understand main points
3	Covers between 50% and 74% of significant points in the paper	Generally clear, may contain complex or convoluted sentence
4	Covers between 75% and 99% of significant points in the paper	Summarizes information logically with few confusing/unclear sentences
5	Covers 100% of significant points in the paper	Clear and concise in wording and content, could be easily understood by someone with understanding in the field

**If the model's response was greater than 3 paragraphs or provided incorrect information, the response was rated as 0 in all categories.

Table 1: Rating system

Model	Average Word Count	Average Accuracy	Average Readability	Average Relevant Content
ChatGPT-4o	278.4	3.05 (.67)	4.98 (.13)	1
Google Gemini	85.13	1.43 (.74)	4.98 (.13)	1
Claude 3 Sonnet	309.53	2.1 (1.55)	3.25 (2.08)	.73
Microsoft Copilot	188.47	2.1 (.92)	4.18 (.93)	1

Table 2: Model Breakdown of Average Ratings

- **Diversity:** We selected 5 papers that discussed different topics but were all within the realm of data science. This was done to examine whether the models could create summaries of similar quality over a variety of topics.
- **Quality and Recent Publication:** To ensure both quality and recency, we only selected papers from the 2023 KDD Undergraduate Consortium. These papers were reviewed and accepted by the consortium, guaranteeing their quality and rigor.

To determine the quality of the model's response, we each read every article and worked together to identify significant points we thought the article's authors would want in a summary. The same set of significant points was evaluated by each of our undergraduate authors. Paper [27] had 10 significant points, paper [10] had 9, paper [26] had 6, paper [38] had 12, and paper [8] had 5. The list of significant points can be found in our GitHub repository.

We prompted each model: "Concisely summarize this paper in no more than 3 paragraphs". This prompt was designed to enable the GenAI models' response variability by including only 2 requirements. For ChatGPT-4o and Claude 3 Sonnet, we provided a PDF of the article by uploading it. Gemini and Copilot did not have this capability, therefore we instead provided them with a link to the PDF.

We evaluated their performance in four categories: brevity, accuracy, readability, and relevant content. Brevity was measured in word count. Microsoft Copilot included a source list for their summaries when copied, but as this was not explicitly shown in the summaries on the Copilot website[1], this did not contribute to the word count. Accuracy and readability were rated along the six-point scales outlined in Table 1.

For relevant content, papers received either a rating of 0 or 1. A rating of 0 indicated the existence of any irrelevant information in the summary, while a rating of 1 indicated that the summary was completely on topic. Responses that solely consisted of irrelevant information or were not summaries were excluded, and the trial was restarted to maintain an equal number of responses per model.

Misinformation was distinguished from irrelevant information, with misinformation being defined as misrepresenting a point in the paper resulting in a 0 for all categories, as described in Table 1.

The models were prompted to summarize each of the 5 papers 3 times, resulting in 15 summaries per model. The reviewers then rated each summary privately to ensure no bias occurred in the rating process. In total, 5 papers x 3 summaries per paper x 4 reviewers per summary = 60 ratings collected for each model's accuracy, readability, and relevant content.

4 RESULTS

The results provide a quantifiable comparison between the performance of four popular GenAI models: ChatGPT-4o, Google Gemini, Claude 3 Sonnet, and Microsoft Copilot. Table 2 presents the average word count of each model's summaries, as well as the average rating they were given based on accuracy, readability, and relevant content. Standard deviation was included for accuracy and readability.

4.1 Word Count

ChatGPT produced summaries with an average word count of 278.4 words. The longest summary generated had a length of 322 words while the shortest was 246 words. The Google Gemini summaries averaged 85.13 words, with the longest summary containing 106 words. Microsoft Copilot generated summaries averaging 188.47 words, and the largest summary it generated was 230 words. Claude 3 Sonnet's summaries had the largest word count, at 309.53. However, this was caused by an outlier summary that ignored the given

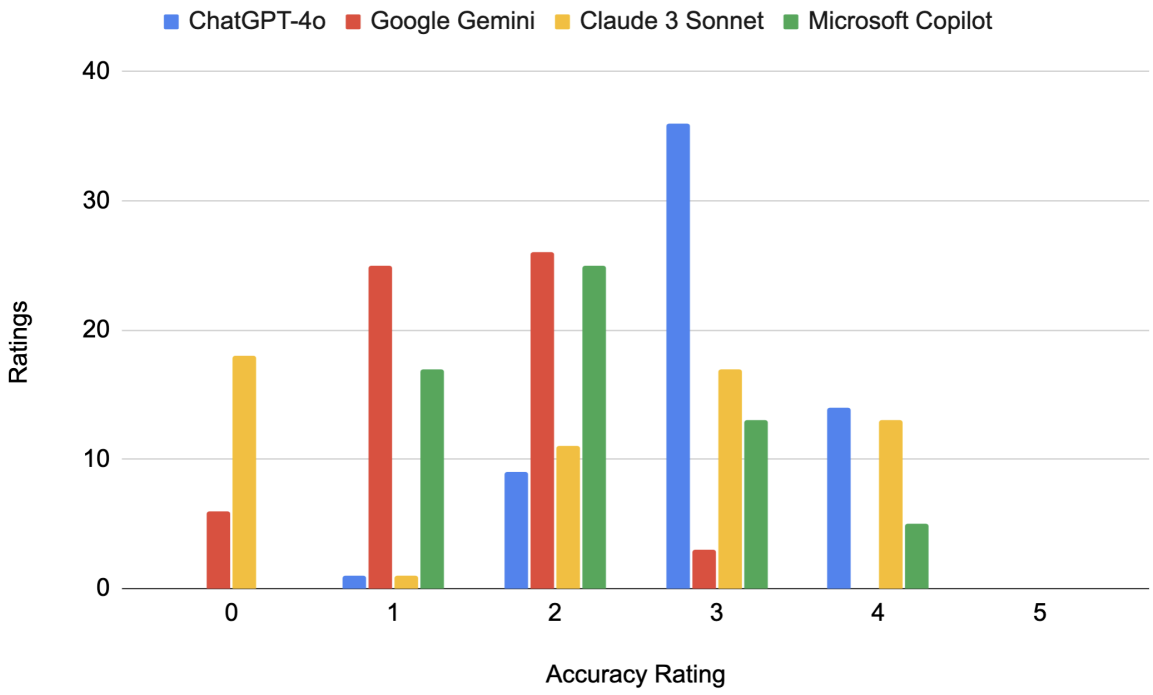

Figure 1: Total Accuracy Ratings by Model

prompt and produced a 1,346-word response. This outlier raised the average word count by 74.03 words.

4.2 Relevant Content

Claude 3 Sonnet was the only model to produce false or irrelevant information within the included trials. This is reflected in its average relevant content rating. While other models were rated to have relevant content every time and therefore had an average of 1, Claude 3 Sonnet had an average of .73, indicating that it only produced a relevant and misinformation-free summary of the input article 73% of the time.

4.3 Accuracy

ChatGPT-4o ranked first for accuracy with an average rating of 3.05 (sd=.67). The highest rating it received was 4 while the lowest was 1. Both Claude 3 Sonnet and Copilot had an average accuracy rating of 2.1 (sd=1.55, .92, respectively). Claude 3 Sonnet had ratings ranging from 0 to 4 while Copilot had ratings from 1 to 4. Additionally, Google Gemini had the lowest average accuracy rating, receiving a range of ratings from 0 to 2 which resulted in an average rating of just 1.43 (sd=.74). However, because it took several trials to produce a response that included a summary from Copilot, we gave it an overall rank of 4th in accuracy. Claude 3 Sonnet's accuracy rating was affected by the 0s it received as a result of failures regarding correctness and summary length. For example, for all trials for paper[26], it produced an incorrect number describing the size of a dataset the paper created. It is worth noting that while no model achieved a perfect rating, there were multiple instances of models only missing one of the significant points that we created. For example, all reviewers found 8 of 9 significant points were mentioned in ChatGPT-4o's second and third responses for paper[10].

4.4 Readability

Summaries produced by ChatGPT-4o and Google Gemini were rated as the most readable with both models receiving an average readability rating of 4.98 (sd=.13, .13). The ratings had little variation, with each receiving a rating of 4 only once while the remainder of the trials resulted in ratings of 5. Copilot received a lower average rating of 4.18 (sd=.93). The lowest readability rating it received was a 2 while the highest was a 5. Claude 3 Sonnet had the lowest average readability rating at 3.25 (sd=2.08). Its score was affected by the 0s it received as a result of incorrect information and format in the summaries.

5 DISCUSSION

5.1 Evaluation Criteria

As detailed by Chang et al., there are two common ways to evaluate the performance of GenAI models: automatic and human evaluation[6]. We used the human evaluation approach for this analysis, relying on a standardized evaluation rubric to quantitatively represent the qualitative results of the human evaluation. While this led to more variation and subjectivity in the resulting metrics, the averages generally reflected our perceptions of each model's performance.

The criteria used in this evaluation were not entirely novel. Previous research has measured the accuracy and correctness of GenAI or stated that they would be important to measure[6, 44]. The readability of responses, while not explicitly measured, was also a recurring theme in evaluative papers. Our analysis introduced another criterion, word count, which we found to be informative for results.

For example, though Google Gemini was found to be highly readable, it was also highly inaccurate. When defining accuracy as the percentage of significant points covered, a low word count with the same accuracy is more impressive. However, when the most accurate model has a higher word count, it could indicate the need for longer responses from the lower-performing models. Overall, we found that these criteria, while important, could not individually indicate the performance of a GenAI model. Rather, they are best understood in the context of the other qualitative and quantitative analyses that are conducted during a performance evaluation.

5.2 Model Comparison

- ChatGPT-4o: This model developed by OpenAI consistently showed the highest accuracy, with an average rating of 3.05. It also had a similar word count to Claude 3 Sonnet: with on average, a 31-word difference. This means that ChatGPT-4o was the best at summarizing the significant points from the different papers. While it had good performance, it was not perfect, as there were instances where it could have included more information and detail to raise its accuracy.
- Google Gemini: Google's GenAI model produced the shortest results with an average word count of 85.13 and with one paragraph in every trial for 4 of the 5 papers. The last paper had 2 paragraphs for each trial. While the responses were concise, they failed to go in-depth into the significant points of the paper, resulting in an average accuracy score of 1.43. However, it was rated as the most readable model. This calls to point out that the most readable AI responses may not be the most accurate, as the small summary may suggest less significant points covered. Areas of improvement include distinguishing between significant and insignificant content and providing more information in responses. Additionally, Gemini showed low variability in response content, with 4 of the 5 papers having the same response for each trial, and the 5th having only slightly different trials. Better evaluating randomness in Gemini response may benefit Google's model, as GenAI is supposed to be able to consistently generate a variety of outputs for the same input, which is not observed for Gemini.
- Claude 3 Sonnet: The AI model developed by Anthropic produces an accuracy rating of 2.1. It was also the only AI model to provide incorrect/wrong information, specifically when testing paper [26], which impacted its average score. Claude 3 Sonnet exhibited the greatest average word count at 309.53 words per response. For trial 2 on paper [38], the model gave an incoherent 1,346-word response, which was included in the average. This shows us that even when an AI model produces the most words in a response, they may not necessarily be significant to mention or even relevant to the paper. This model needs to better prevent the inclusion of both irrelevant and confusing information in its responses. For example, responses often included confusing annotations about paragraph numbers, such as </Paragraph 2>. Some form of this was found in 12 of the 15 responses evaluated. Also, adding quotes to parts of its response, as observed in trials 1 and 3 on paper [38] made these responses more difficult to read.
- Microsoft Copilot: Microsoft's AI model demonstrated moderate performance, with an average accuracy rating of 2.1 and an average word count of 188.46. As shown in Figure 2, it is in the middle of the graph, which shows its balanced performance across readability and word count. It tended to miss significant points, which may limit its validity for in-depth literature review and understanding. Copilot also repeatedly included confusing citations and phrases. It cited in a nonsensical way, often referencing the wrong source. In some cases, the model was able to generate proper responses, but in others, the model generated a message along the lines of "Unfortunately, I am unable to directly access external websites or download files"[1]. These were not included in responses as they did not include any summary of the paper, unlike Claude's outlier. Because each of the authors experienced this during data collection, if the responses were included the accuracy would be at least halved, causing Copilot to rank 4th in accuracy among the 4 models tested. However, further evaluation is needed to determine the extent of this issue.

Based on the results of this analysis, we recommend evaluating GenAI models by prioritizing accuracy and relevant content over brevity and readability. While brevity and readability are important for a summary, accuracy and relevant content determine its validity. Readability is also a completely subjective criterion, making it difficult to quantify.

5.3 Limitations

The results of this analysis provide only a preliminary evaluation of GenAI's capacity to aid in the literature review process due to several limiting factors. The two most prominent are finances and time. Due to a lack of funding, we were unable to conduct our experiment on the paid versions of these GenAI models, so our results are only indicative of the performance of the free versions. As of May 2024, the free versions of ChatGPT-4o and Claude 3 Sonnet limit the number of prompts per user. We also lacked the time to conduct a more extensive experiment. Given more time and greater finances, we could have analyzed potential barriers to entry provided by the paid models while comparing their summaries to those of their free counterparts.

As mentioned in section 3, trials that resulted in responses completely irrelevant to the paper were excluded from this experiment. The only model to be impacted by this was Copilot.

In designing the evaluative assessment, we faced challenges due to the need for unanimity, as defining a good summary is subjective. For example, we disagreed on whether model responses needed to include specific statistical differences [10], which highlighted the model's improvement over the baseline. Ultimately, we decided not

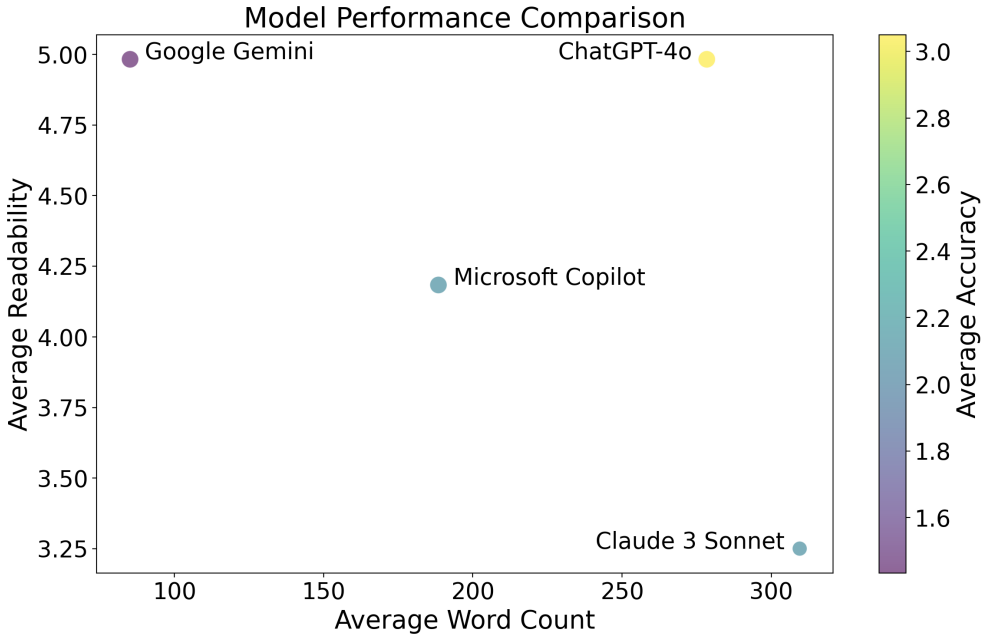

Figure 2: Average Data for Each Model

to include the statistic as a required significant point, and this choice affected the accuracy ratings of all models except Claude 3 Sonnet. A single evaluation rubric led to data loss, as it couldn't fully capture our differing opinions. Similarly, each summary's readability score was completely subjective. Each person has their own writing style, making it difficult to create objective criteria for readability. These limitations are inherent to the subjective nature of a qualitative performance evaluation and contribute to the difficulties involved in generalizing results.

5.4 Future Work

Future research could expand the scope by conducting more extensive experiments, including testing paid versions of GenAI models, generating summaries for a broader range of articles, or increasing the number of trials to improve reliability and generalizability. Another promising avenue for future research is exploring interdisciplinary applications, such as the intersection of AI and brain-computer interfaces[9, 16, 19, 29–32, 39], and others, [20–25, 41–43], to uncover novel insights and innovations. Additionally, investigating potential applications of GenAI in other stages of the research process, such as experiment design and data analysis, could reveal broader utility beyond data science.

Future studies should also consider diversifying the evaluation metrics and audiences involved. As the results of this study present an undergraduate level of model output evaluations, involving researchers with expertise in the subject matter could provide more nuanced and critical evaluations of model outputs. This would help in assessing the quality and applicability of AI-generated content more accurately.

6 CONCLUSION

The results of this analysis indicate that the best criteria to evaluate GenAI models are accuracy and relevant content. While brevity is important for a summary, accuracy and relevant content determine its validity. In addition, readability is a more subjective criterion, making it difficult to quantify.

For the purposes of a literature review, ChatGPT-4o set the highest standard out of the models evaluated. However, its performance, along with that of the other three models, leaves room for further improvement. Regarding accuracy, the models ranked from best to worst: ChatGPT-4o, Claude 3 Sonnet, Google Gemini, and Microsoft Copilot.[1] In readability, the models ranked from best to worst: ChatGPT-4o = Google Gemini, Microsoft Copilot, Claude 3 Sonnet. With relevant content: ChatGPT-4o, Google Gemini, and Microsoft Copilot ranked equally, and Claude 3 Sonnet ranked last.

Overall, using GenAI to generate summaries may help researchers filter out papers not conducive to their research goals, streamlining the literature review process.

[1]Note that Microsoft Copilot ranks differently than shown in previous figures because of the non-summary responses that were not included in the trials. If included, they would lower the accuracy score to below what Google Gemini achieved.

REFERENCES

[1] [n. d.]. Microsoft Copilot: Your everyday AI companion. https://ceto.westus2.binguxlivesite.net/

[2] Mehdi Allahyari, Seyedamin Pouriyeh, Mehdi Assefi, Saeid Safaei, Elizabeth D. Trippe, Juan B. Gutierrez, and Krys Kochut. 2017. Text Summarization Techniques: A Brief Survey. arXiv:1707.02268 [cs.CL]

[3] M. Bhattacharyya, V. M. Miller, D. Bhattacharyya, and L. E. Miller. 2023. High Rates of Fabricated and Inaccurate References in ChatGPT-Generated Medical Content. *Cureus* 15, 5 (2023), e39238. https://doi.org/10.7759/cureus.39238

[4] Hanqun Cao, Cheng Tan, Zhangyang Gao, Yilun Xu, Guangyong Chen, Pheng-Ann Heng, and Stan Z. Li. 2024. A Survey on Generative Diffusion Models. *IEEE Transactions on Knowledge and Data Engineering* (2024), 1–20. https://doi.org/10.1109/TKDE.2024.3361474

[5] Cecilia Ka Yuk Chan and Wenjie Hu. 2023. Students' voices on generative AI: Perceptions, benefits, and challenges in higher education. *International Journal of Educational Technology in Higher Education* 20, 1 (2023), 43.

[6] Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, et al. 2024. A survey on evaluation of large language models. *ACM Transactions on Intelligent Systems and Technology* 15, 3 (2024), 1–45.

[7] Joshua Ebere Chukwuere. 2024. The future of generative AI chatbots in higher education. arXiv:2403.13487 [cs.CY]

[8] Param Damle, Jo Watts, Glen Bull, and N Rich Nguyen. 2023. Extracting Features for Computational Thinking from Block-Based Code. (2023).

[9] Guangyao Dou, Zheng Zhou, and Xiaodong Qu. 2022. Time majority voting, a PC-based EEG classifier for non-expert users. In *International Conference on Human-Computer Interaction*. Springer, 415–428.

[10] Daniil Filienko, Yudong Lin, Kyler Robinson, Trevor Tomlin, and Martine De Cock. 2023. Predicting Time to Pushback of Flights in US Airports. (2023).

[11] Francisco García-Peñalvo and Andrea Vázquez-Ingelmo. 2023. What do we mean by GenAI? A systematic mapping of the evolution, trends, and techniques involved in Generative AI. (2023).

[12] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative Adversarial Networks. arXiv:1406.2661 [stat.ML]

[13] S. Grossberg. 2013. Recurrent neural networks. *Scholarpedia* 8, 2 (2013), 1888. https://doi.org/10.4249/scholarpedia.1888 revision #138057.

[14] Muhammad Usman Hadi, Rizwan Qureshi, Abbas Shah, Muhammad Irfan, Anas Zafar, Muhammad Bilal Shaikh, Naveed Akhtar, Jia Wu, Seyedali Mirjalili, et al. 2023. Large language models: a comprehensive survey of its applications, challenges, limitations, and future prospects. *Authorea Preprints* (2023).

[15] Lei Huang, Yanxiang He, Furu Wei, and Wenjie Li. 2010. Modeling document summarization as multi-objective optimization. In *3rd International Symposium on Intelligent Information Technology and Security Informatics, IITSI 2010*. 382–386. https://doi.org/10.1109/IITSI.2010.80 2010 International Symposium on Intelligent Information Technology and Security Informatics, IITSI 2010 ; Conference date: 02-04-2010 Through 04-04-2010.

[16] Matthew L Key, Tural Mehtiyev, and Xiaodong Qu. 2024. Advancing EEG-based gaze prediction using depthwise separable convolution and enhanced preprocessing. In *International Conference on Human-Computer Interaction*. Springer, 3–17.

[17] Sascha Kraus, Matthias Breier, Weng Marc Lim, Marina Dabić, Satish Kumar, Dominik Kanbach, Debmalya Mukherjee, Vincenzo Corvello, Juan Piñeiro-Chousa, Eric Liguori, Daniel Palacios-Marqués, Francesco Schiavone, Alberto Ferraris, Cristina Fernandes, and João J. Ferreira. 2022. Literature reviews as independent studies: guidelines for academic practice. *Review of Managerial Science* 16, 8 (2022). https://doi.org/10.1007/s11846-022-00588-8

[18] Peder Olesen Larsen and Markus von Ins. 2010. The rate of growth in scientific publication and the decline in coverage provided by Science Citation Index. *Scientometrics* 84, 3 (2010), 575–603. https://doi.org/10.1007/s11192-010-0202-z

[19] Weigeng Li, Neng Zhou, and Xiaodong Qu. 2024. Enhancing Eye-Tracking Performance Through Multi-task Learning Transformer. In *International Conference on Human-Computer Interaction*. Springer, 31–46.

[20] Yingzhou Lu, Tianyi Chen, Nan Hao, Capucine Van Rechem, Jintai Chen, and Tianfan Fu. 2024. Uncertainty quantification and interpretability for clinical trial approval prediction. *Health Data Science* 4 (2024), 0126.

[21] Yingzhou Lu, Kosaku Sato, and Jialu Wang. 2023. Deep Learning based Multi-Label Image Classification of Protest Activities. *arXiv preprint arXiv:2301.04212* (2023).

[22] Yingzhou Lu, Minjie Shen, Huazheng Wang, Xiao Wang, Capucine van Rechem, and Wenqi Wei. 2023. Machine learning for synthetic data generation: a review. *arXiv preprint arXiv:2302.04062* (2023).

[23] Xiaobo Ma. 2022. *Traffic performance evaluation using statistical and machine learning methods*. Ph. D. Dissertation. The University of Arizona.

[24] Xiaobo Ma, Adrian Cottam, Mohammad Razaur Rahman Shaon, and Yao-Jan Wu. 2023. A transfer learning framework for proactive ramp metering performance assessment. *arXiv preprint arXiv:2308.03542* (2023).

[25] Xiaobo Ma, Abolfazl Karimpour, and Yao-Jan Wu. 2024. Data-driven transfer learning framework for estimating on-ramp and off-ramp traffic flows. *Journal of Intelligent Transportation Systems* (2024), 1–14.

[26] Miguel Monares, Yuan Tang, Ritik Raina, and Virginia R de Sa. 2023. Analyzing Biases in AU Activation Estimation Toward Fairer Facial Expression Recognition. (2023).

[27] Nathan Koome Murungi, Michael Vinh Pham, Xufeng Caesar Dai, and Xiaodong Qu. 2023. Empowering Computer Science Students in Electroencephalography (EEG) Analysis: A Review of Machine Learning Algorithms for EEG Datasets. (2023).

[28] James Prather, Paul Denny, Juho Leinonen, Brett A Becker, Ibrahim Albluwi, Michelle Craig, Hieke Keuning, Natalie Kiesler, Tobias Kohn, Andrew Luxton-Reilly, et al. 2023. The robots are here: Navigating the generative ai revolution in computing education. In *Proceedings of the 2023 Working Group Reports on Innovation and Technology in Computer Science Education*. 108–159.

[29] Xiaodong Qu. 2022. *Time Continuity Voting for Electroencephalography (EEG) Classification*. Ph. D. Dissertation. Brandeis University.

[30] Xiaodong Qu, Peiyan Liu, Zhaonan Li, and Timothy Hickey. 2020. Multi-class time continuity voting for EEG classification. In *Brain Function Assessment in Learning: Second International Conference, BFAL 2020, Heraklion, Crete, Greece, October 9–11, 2020, Proceedings 2*. Springer, 24–33.

[31] Xiaodong Qu, Saran Liukasemsarn, Jingxuan Tu, Amy Higgins, Timothy J Hickey, and Mei-Hua Hall. 2020. Identifying clinically and functionally distinct groups among healthy controls and first episode psychosis patients by clustering on EEG patterns. *Frontiers in psychiatry* 11 (2020), 541659.

[32] Xiaodong Qu, Qingtian Mei, Peiyan Liu, and Timothy Hickey. 2020. Using EEG to distinguish between writing and typing for the same cognitive task. In *Brain Function Assessment in Learning: Second International Conference, BFAL 2020, Heraklion, Crete, Greece, October 9–11, 2020, Proceedings 2*. Springer, 66–74.

[33] Md Mostafizer Rahman and Yutaka Watanobe. 2023. ChatGPT for education and research: Opportunities, threats, and strategies. *Applied Sciences* 13, 9 (2023), 5783.

[34] Tareq Rasul, Sumesh Nair, Diane Kalendra, Mulyadi Robin, Fernando de Oliveira Santini, Wagner Junior Ladeira, Mingwei Sun, Ingrid Day, Raouf Ahmad Rather, and Liz Heathcote. 2023. The role of ChatGPT in higher education: Benefits, challenges, and future research directions. *Journal of Applied Learning and Teaching* 6, 1 (2023).

[35] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention Is All You Need. arXiv:1706.03762 [cs.CL]

[36] Adhika Pramita Widyassari, Supriadi Rustad, Guruh Fajar Shidik, Edi Noersasongko, Abdul Syukur, Affandy Affandy, and De Rosal Ignatius Moses Setiadi. 2022. Review of automatic text summarization techniques and methods. *Journal of King Saud University - Computer and Information Sciences* 34, 4 (2022), 1029–1046. https://doi.org/10.1016/j.jksuci.2020.05.006

[37] Frank F Xu, Uri Alon, Graham Neubig, and Vincent Josua Hellendoorn. 2022. A systematic evaluation of large language models of code. In *Proceedings of the 6th ACM SIGPLAN International Symposium on Machine Programming*. 1–10.

[38] Catherine Yang, Yuying Zhao, and Tyler Derr. 2023. The Friendship Paradox: An Analysis on Signed Social Networks with Positive and Negative Links. (2023).

[39] Long Yi and Xiaodong Qu. 2022. Attention-based CNN capturing EEG recording's average voltage and local change. In *International Conference on Human-Computer Interaction*. Springer, 448–459.

[40] Isshin Yunoki, Guy Berreby, Nicholas D'Andrea, Yuhua Lu, and Xiaodong Qu. 2023. Exploring AI Music Generation: A Review of Deep Learning Algorithms and Datasets for Undergraduate Researchers. In *International Conference on Human-Computer Interaction*. Springer, 102–116.

[41] Zhengming Zhang, Renran Tian, and Zhengming Ding. 2023. Trep: Transformer-based evidential prediction for pedestrian intention with uncertainty. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 37. 3534–3542.

[42] Zhengming Zhang, Renran Tian, Vincent G Duffy, and Lingxi Li. 2024. The comfort of the soft-safety driver alerts: Measurements and evaluation. *International Journal of Human–Computer Interaction* 40, 4 (2024), 904–914.

[43] Zhengming Zhang, Renran Tian, Rini Sherony, Joshua Domeyer, and Zhengming Ding. 2022. Attention-based interrelation modeling for explainable automated driving. *IEEE Transactions on Intelligent Vehicles* 8, 2 (2022), 1564–1573.

[44] Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. 2023. A survey of large language models. *arXiv preprint arXiv:2303.18223* (2023).

[45] Guido Zuccon, Bevan Koopman, and Razia Shaik. 2023. ChatGPT Hallucinates when Attributing Answers. In *Proceedings of the Annual International ACM SIGIR Conference on Research and Development in Information Retrieval in the Asia Pacific Region* (, Beijing, China,) *(SIGIR-AP '23)*. Association for Computing Machinery, New York, NY, USA, 46–51. https://doi.org/10.1145/3624918.3625329