AI Insights: A Case Study on Utilizing ChatGPT Intelligence for Research Paper Analysis

Anjalee De Silva[1,*,†], Janaka L. Wijekoon[2,3,*,†], Rashini Liyanarachchi[4,*,†], Rrubaa Panchendrarajan[5,*,†] and Weranga Rajapaksha[6,*,†]

[1]University of Queensland, Brisbane, Australia
[2]Victorian Institute of Technology, Adelaide, South Australia
[3]Keio University, Yokohama, Japan
[4]University of New South Wales(UNSW), Sydney, Australia
[5]Queen Mary University of London, United Kingdom
[6]University of South Australia, Adelaide, Australia

Abstract

This paper discusses the effectiveness of leveraging Chatbot: Generative Pre-trained Transformer (ChatGPT) versions 3.5 and 4 for analyzing research papers for effective writing of scientific literature surveys. The study selected the *Application of Artificial Intelligence in Breast Cancer Treatment* as the research topic. Research papers related to this topic were collected from three major publication databases Google Scholar, Pubmed, and Scopus. ChatGPT models were used to identify the category, scope, and relevant information from the research papers for automatic identification of relevant papers related to Breast Cancer Treatment (BCT), organization of papers according to scope, and identification of key information for survey paper writing. Evaluations performed using ground truth data annotated using subject experts reveal, that GPT-4 achieves 77.3% accuracy in identifying the research paper categories and 50% of the papers were correctly identified by GPT-4 for their scopes. Further, the results demonstrate that GPT-4 can generate reasons for its decisions with an average of 27% new words, and 67% of the reasons given by the model were completely agreeable to the subject experts.

Keywords

ChatGPT-3.5, ChatGPT-4, Research Paper, Academic Writing, Artificial Intelligence, Research Paper Classification,

1. Introduction

Artificial Intelligence (AI), the term coined by John McCarthy in 1956 [1] and discussed by prominent scientists such as Nikola Tesla in 1890 [2], Vannevar Bush in 1945 [3], and Alan Turing in 1950 [4], is a concept of machines thinking and applying knowledge. The “Imitation Game" is a

14th International Workshop on Bibliometric-enhanced Information Retrieval, 24–28 March, 2024, Glasgow, Scotland
*Corresponding author.
†Author names appear alphabetically, and all authors contributed equally.
anjalee98@gmail.com (A. D. Silva); janaka.wijekoon@gmail.com (J. L. Wijekoon); rashinikavindya@gmail.com (R. Liyanarachchi); r.panchendrarajan@qmul.ac.uk (R. Panchendrarajan); weranga.rajapaksha_gedara@mymail.unisa.edu.au (W. Rajapaksha)
0009-0001-5881-4408 (A. D. Silva); 0000-0001-6092-3366 (J. L. Wijekoon); 0000-0003-0228-6658 (R. Liyanarachchi); 0000-0002-1403-2236 (R. Panchendrarajan); 0000-0001-8361-884X (W. Rajapaksha)

significant historical example of machines demonstrating the ability to think and use knowledge to solve problems [4]. Ever since then, the research in AI has seen tremendous developments including the introduction of Neural Networks in 1990 [5]. Notably, the momentum in AI was gained in various industries since 2017 [6], following the introduction of the transformer model: a parallel multi-head attention mechanism, by Ashish Vaswani et al. [7]. The Transformer model led to the development of a Chatbot: Generative Pre-trained Transformer 3.5 (ChatGPT-3.5) in late 2022 [8], marking a revolutionary change in the AI domain [9]. Subsequently, various industries have employed [10, 6, 11, 12, 13, 14], debated [15, 16], and disputed [17, 18] the use of GPT models.

Background studies revealed some studies embracing GPT [19, 20, 14, 21] and focused on using GPT to automate academic writing [22, 23, 24]. Whereas, some approached GPT cautiously [25, 26, 27, 28] and some emphasised the need for regulation and introducing new guidelines for using generative AI [29]. Consequently, as a part of our ongoing research project, this paper discusses the effectiveness of using GPT models to analyze research papers for writing scientific literature surveys.

The project is carried out to produce a review paper revealing how AI applications are used in Breast Cancer Treatment (BCT). The overall study comprises five stages (Refer to Figure 1) starting from the construction of a taxonomy depicting the branches of BCT, followed by the research paper collection, and automatic analysis of research papers for drafting the survey paper. Subsequently, this paper presents the effectiveness of using ChatGPT to automatically analyze the gathered research papers.

We used both GPT-3.5 (2022 January update) and GPT-4 (2023 April update) models to automatically analyze the research papers. Various information presented in the research paper including the title, abstract, and textual content was used at different stages of the study. Research papers gathered from three major publication databases Google Scholar, Pubmed, and Scopus were merged without duplicates to form a unified corpus related to *AI in BCT*. ChatGPT models were employed in this corpus to identify the research paper categories and scope automatically and to retrieve information required for survey paper writing. The performance of the models in identifying the category and scope of a research paper was evaluated against ground truth data annotated by subject experts. Further, the capability of the models to generate reasons for their own decisions was analyzed with the help of the same subject experts. Results of our experiments are presented in Section 3.

2. Methodology

Figure 1 illustrates the flow of the methodology carried out in this study via various stages. Stage 1 comprises the construction of a taxonomy depicting the branches of BCT. Following this, in stage 2, the search queries were developed based on the taxonomy, and the research papers were gathered by querying Google Scholar, Pubmed, and Scopus, followed by duplication removal to form a unified corpus of research articles related to *AI in BCT* [1].

Stage 3 focused on identifying the research paper category to filter out the relevant papers related to BCT. We compared the performance of both GPT-3.5 and GPT-4 in identifying research

[1]Automating scripts can be found at https://github.com/janakawest/ScholrlyDatagathering

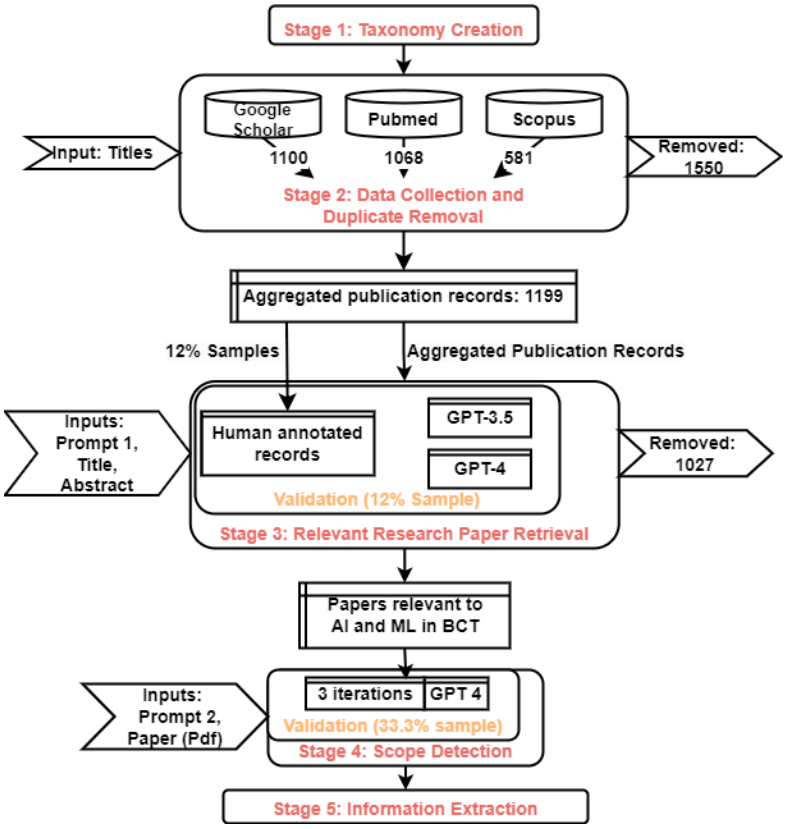

Figure 1: Overall Methodology

paper categories using an human-annotated sample of 12% research papers from the unified corpus. Then, the best GPT model was used to identify research paper categories, and research papers related to *AI in BCT* proceeded to the next stage. In Stage 4, we identified the BCT scope of the research papers to organize them in the survey paper according to their area of study. We also validated the performance of GPT model in identifying the scope of a research paper using a human-annotated sample of 33% relevant research papers. Finally, in Stage 5, we extracted the relevant information from the research paper required to write the survey article.

2.1. Taxonomy Construction

Our study began with constructing a taxonomy depicting all the branches of BCT. This taxonomy was used to retrieve research papers from publication databases and to categorize and organize various treatment methods into a structured format for easier understanding and analysis. Figure 2 depicts the taxonomy used in this study.

Specifically, the BCT options were primarily organized under three broader oncological categories namely medical oncology, surgical oncology, and radiation oncology. Further, each category is divided into sub-categories. We included the type *Other* in each branch to indicate the treatment types excluded in the taxonomy. For example, the *Other* type under *Medical*

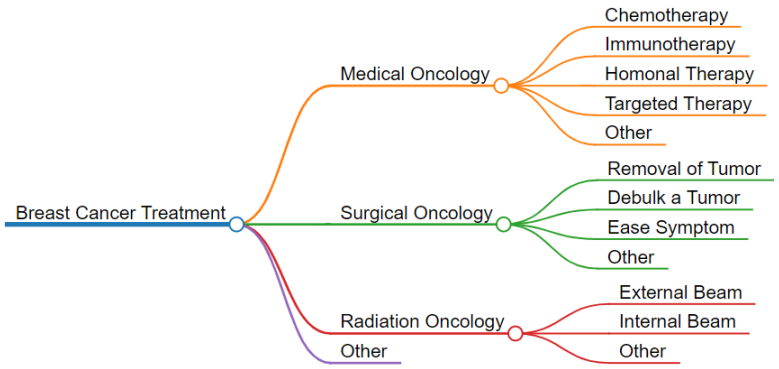

Figure 2: Taxonomy of *BCT* used for the Study

Oncology was included to cover all the sub-branches of medical oncology excluded in our taxonomy. As this is an ongoing project, we hope to extend our taxonomy to cover all the BCT types, and this is discussed as a future work in Section 4.

2.2. Data Collection

We collected research articles from three major publication databases, Google Scholar [2], Pubmed [3], and Scopus [4]. Each database was queried independently using the keywords related to the topic *AI for BCT*. The search keywords were defined based on the taxonomy (Refer Figure 2), such that each keyword represents a node or a branch of the taxonomy. For example, the search keyword *Radiology* represents all three sub-branches of *Radiation Oncology*, whereas the search keyword *Chemotherapy* represents a node in the taxonomy. Together with the keywords, the search queries included *"AI", "Artificial Intelligence", "Breast cancer", " «breast cancer treatment type»"* to target the search to BCT. For example, *"AI", "Artificial Intelligence", "Breast cancer", "Radiology"* is an example search query used to retrieve research papers related to radiology in BCT. Altogether, we used 13 keywords resulting in 13 search queries, and the research papers were retrieved by querying the publication database using these search queries.

We used the built-in API of Google Scholar and Pubmed to query the research papers automatically, and Scopus was queried using the user interface manually. For Google Scholar API, we had to mention how many publication records the scraper should retrieve manually. For this study, we limited each query to get 110 records per search to avoid the increase in noise. Figure 3 presents the number of research papers retrieved for each source for the 13 search keywords.

Once the research papers were collected, we removed the duplicate papers within the source as well as across the sources to generate a unified corpus of research papers related to the topic *AI in BCT*. We used the title of the research papers to identify the duplicates in this stage of the study. Table 1 shows the statistics of the unified corpus. Referring to the table, it can be observed, that Pubmed and Google Scholar APIs resulted in a higher number of duplicates

[2]https://scholar.google.com/
[3]https://pubmed.ncbi.nlm.nih.gov/
[4]https://www.scopus.com/

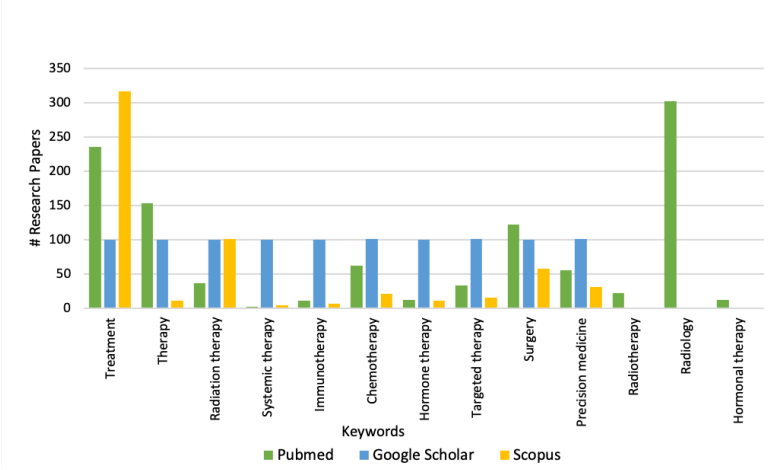

Figure 3: Distribution of Research Papers Retrieved using Keyword related to *BCT*

Table 1
Statistics of the Dataset

Source	# of Papers Collected	# of Unique Papers
Pubmed	1,068	516
Google Scholar	1,100	511
Scopus	581	462
# of Unique Papers across Sources		**1,199**

(roughly 50% of duplicate records) due to their nature in indexing research papers from multiple other repositories and their retrieval mechanism. After removing the duplicate research articles within each source, we ended up with 462 - 516 records per each source. Further, we merged the three repositories and removed the duplicates across the sources resulting in a final corpus of size 1199. Figure 1 illustrates the number of records removed as duplicates.

2.3. Relevant Research Paper Retrieval

The research papers retrieved through the search keywords in stage 1 did not ensure the retrieval of relevant papers related to *AI in BCT*. A manual analysis of the unified corpus revealed that, in addition to the papers related to AI in BCT, it comprised a mix of research papers related to AI for breast cancer diagnosis, some review papers in breast cancer research, research papers focusing on both BCT and diagnosis, along with completely irrelevant papers. Therefore, this step focused on automatically identifying the research paper categories to filter out articles related to BCT from the unified corpus.

Given the title, and abstract of the research paper, the ChatGPT model was instructed to identify the objective of the paper from the aforementioned options. For this purpose, we designed the Prompt 1 as a single-option multiple-choice question, allowing the model to choose the right answer from the given options.

Prompt 1 Category Identification Prompt

```
Title:
Abstract:
Analyze the scientific paper's title and abstract to determine the
most accurate description of its objectives
from the following options:

A: A Review or survey paper summarizing research related to breast cancer
B: A Research study on breast cancer diagnosis
C: A Research study on breast cancer treatment
D: A Research study on breast cancer diagnosis and treatment
E: None of the above
F: Not sure

Share your response in the format:
Answer
Reason
```

We included option E indicating the research articles that are not related to BCT, and option F was included in Prompt 1 to enable the model to indicate that it is unsure about the decision. We noticed that the model may choose different options for the same article's title and abstract during different prompt executions as it is a generative model [30]. Therefore, for each GPT model, we executed Prompt 1 three times and chose the majority option selected by the model. If no majority was found, then we marked the ChatGPT answer as *F: Not sure*, to indicate the confusion of the model in determining the answer. Once the majority option was selected, papers indicated as either option *C* or *D* were filtered out as relevant papers for the next stage of the study. We ended up with 143 relevant papers focusing on BCT as their key objective.

We used both GPT-3.5 and GPT-4 for this stage of the study and the performance of both models on categorizing the research papers (Refer Section 3.1.1) was evaluated using 12% human-annotated sample of the unified corpus. Due to the unsurpassed performance of the GPT-4 model, compared to GPT-3.5, we carried out the remaining study using the GPT-4 model.

2.4. Scope Detection

Each relevant paper focusing on BCT may discuss the application of AI in different BCT. Therefore, the next stage of our study was focused on automatically identifying the *scope* of the research paper, indicating the BCT type(s) experimented by the authors. In this stage, we defined the *scope* as a path in the taxonomy (Figure 2). For example, *Medical Oncology - Chemotherapy* is a possible scope of BCT. Similar to the relevant paper detection, we designed Prompt 2 for ChatGPT as a multiple-choice question with all the possible scopes (all the paths in the taxonomy) as the answers. This resulted in 15 possible answers for the scope detection prompt. Here, options A to M indicate the paths in the taxonomy.

We executed Prompt 2 by uploading the PDF file of each relevant paper to GPT-4. The model was instructed to read the entire research article and identify the main focus of the research paper from the list of scope options. The model was instructed to select more than one answer among options A-M if the authors had contributed to more than one BCT type. We ran the

Prompt 2 Scope Detection Prompt

```
Make sure you thoroughly read through the entire research paper and classify the content of
the paper from the following options:

A: Chemotherapy of Medical Oncology
B: Immunotherapy of Medical Oncology
C: Hormonal therapy of Medical Oncology
...
L: Other treatments of Radiation Oncology
M: Other treatment categories directly related to breast cancer
N: Reviews or meta-analyses on breast cancer treatments
O: Studies not directly related to the treatment of breast cancer

In case, if the authors contributed on many areas of treatments, select all the necessary
options from the list A to M as the answer. Else, If the paper provides a detailed review of
existing breast cancer treatments, then choose Option "N" as the answer. Otherwise, If the
content of the paper does NOT contribute to breast cancer treatment at all, then choose
option "O" as the answer.

PLEASE Do not provide any explanations. Just mention the suitable options.

NOTE: Make sure you read the entire research paper to select the options.
```

prompt three times for each research paper, and the options that appeared in more than one execution were chosen as the scope of the research paper.

Similar to the previous stage, we evaluated the performance of the GPT-4 model in correctly identifying the scope of the research paper using a human-annotated sample of size 33% of the relevant papers. Further details on this experiment are discussed in Section 3.2.

2.5. Information extraction

After utilizing GPT-4 to identify the scope of the paper, this stage of the study was aimed at extracting the information required for survey paper writing. We uploaded the paper to GPT-4 and provided Prompt 3 to extract specific details, including background & objective, methods, key findings, conclusions, and limitations of the study. Subject experts verified the extracted information for a sample research paper[5] [31] and the observations are discussed in Section 3.3. As this is ongoing research, Prompt 3 will be further enhanced to retrieve scope-specific information as one of the main future works.

3. Results and Discussion

In this section, we analyze and present the performance of GPT in 1.) identifying research paper categories using the title and abstract, 2.) analyzing the reasons provided by the GPT for category detection, and 3.) identifying the scope of a research paper by reading the entire

[5]Data of a large sample is not presented due to page limitation

Prompt 3 Information Extraction Prompt

```
Analyze the research article thoroughly and generate a concise summary in tabular form that
captures the fundamental aspects of the study. Include the following key points:

1. Background and Objective:
   - Context: Briefly describe the background and context of the research.
   - Objective: Specify the goals or discoveries the authors aimed to achieve.
2. Methods:
  - Methodologies: Summarize the methods, experiments, or analysis employed by the authors
   in conducting their research.
3. Key Findings:
  - Main Results: Provide a clear overview of the primary results or discoveries presented
   in the paper.
4. Conclusion:
   - Final Takeaways: Summarize the concluding remarks of the research.
  - Future Directions: Highlight any recommendations or future directions suggested by
   the authors.
5. Limitations:
  - Constraints: Identify and summarize any limitations or constraints mentioned in the study.

Ensure that the summary is comprehensive, and coherent, and avoids unnecessary jargon.
Present the information in a well-organized tabular format for easy comprehension.
```

paper. The temperature of both GPT models was not altered assuming their default temperature values, i.e., 1 for both models [32].

3.1. Relevant Research Paper Retrieval

We randomly sampled around 12% of the research papers from the initial corpus of size 1199. However, there were some research papers in the sample without abstracts and we had to drop those papers for a fair evaluation of the model in identifying the research paper category using title and abstract. This resulted in a final sample size of 132. The sampled data were used to analyze the performance of GPT-3.5 and GPT-4, and the evaluation process and the performance of the model are discussed in the following sections.

3.1.1. Performance on Category Identification

The selected sample papers were first given to two subject experts for annotation based on the title and abstract, and the human-annotated data were used to evaluate the performance of the model in identifying the research paper category. The subject experts annotated the sample using the six options listed in Section 2.3, and any disagreements between the subject experts were resolved through discussion. The accuracy of the models on the sample data is presented in Table 2. The table indicates the average accuracy of three iterations and the accuracy of the majority option selection by the models during the three iterations (i.e., two selections out of three iterations). It is noteworthy that, in both cases, the GPT-4 model significantly outperformed the GPT-3.5 model in identifying research paper categories.

Table 2
Performance of ChatGPT Models in Category Identifation

Model	Accur. (Avg.)	Accur. (Maj.)	Avg. Response Length	New Words Response (Avg.%)
GPT-3.5	65.15% (±6.8)	65.15%	45	18%
GPT-4	**76.21%** (±5)	**77.3%**	68.5	27%

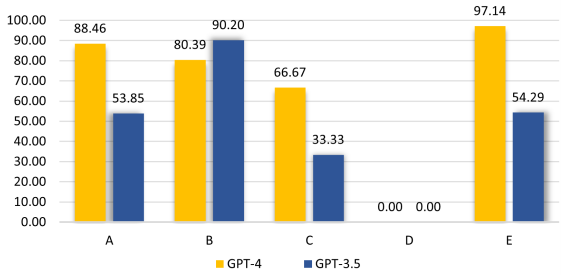

Figure 4: Category-wise Accuracy of ChatGPT Models (Refer to Prompt 1 for Categories)

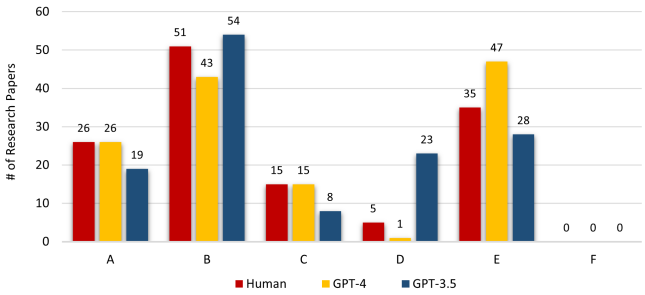

Figure 5: Statistics of the Research Papers Categorized by Human and ChatGPT Models (Refer to Prompt 1 for Categories)

Furthermore, we analysed the accuracy of identifying individual research paper categories. Figures 4 and 5 show the category-wise performance of the models and the number of research articles classified under each category by subject experts Vs. ChatGPT models. It can be observed that the performance of the GPT-4 model is significantly higher than GPT-3.5 across all the categories, except the research articles focusing on breast cancer diagnosis (*Option B*). This could be possibly due to the GPT-4 model classifying some of the breast cancer detection papers as *Option E - None of the above*. Further, both models show drastically different behavior in identifying research articles focusing on both diagnosis and treatment (*Option D*) and fail to correctly identify any of the papers belonging to this category. Interestingly, none of the models choose the *Option F - Not sure* within the sample data.

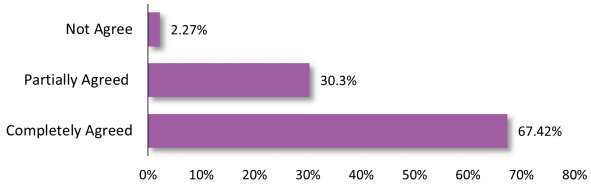

Figure 6: Distribution of Agreement Level of Reasoning

3.1.2. Reasoning Analysis

In addition to the capability of the GPT model to identify the category of a research paper, we analyzed their reasoning capability while making a decision [33]. We instructed the GPT model to give a reason for choosing the option among the six categories in Prompt 1. Our domain experts read the research article's title, abstract, and GPT's choice and the reasons generated by the models, and marked it as either *Completely Agreed*, *Partially Agreed*, or *Not Agree* indicating the level of accurate reasoning the model produces with respect to its choice. Figure 6 presents the distribution of agreement level of the reasoning of the GPT-4 model. It is evident that the model produces completely agreeable reasoning most of the time (67.42%), and very rarely it generates invalid reasoning (2.27%). Table 2 shows the average length of the reasons generated by both GPT-3.5 and GPT-4 models. It can be observed, that GPT-3.5 generates relatively shorter reasons compared to GPT-4.

We further analyze the capability of the model to generate reasons with its wording instead of copying the content from the abstract and the title. We computed the percentage of new words appearing in the reason compared to the abstract. Here, the stop words[6] were not considered as words. The observed results are presented in Table 2. The table presents the average number of new words produced by both models, and interestingly GPT-4 model produces relatively more new words in the reason compared to GPT-3.5, though the former generates lengthier reasons. The distribution of the percentage of new words generated by the GPT-4 model during reasoning is presented in Figure 7. It can be noticed that the GPT-4 model generates 25-30% of new words when reasoning and the highest value of 78% is also observed in the sample data.

3.2. Scope Detection

To effectively evaluate the model's ability to identify the scope of selected papers in BCT, a detailed annotation process was employed using 15 predefined scope options mentioned in Prompt 2. Similar to the category identification, initially, two subject experts were presented with the full text of approximately 33% of the relevant papers related to BCT and asked to independently determine its treatment scope from the above-mentioned options. Noting that a single paper could potentially align with multiple scopes. In instances of disagreement between the annotators, resolution was achieved through discussion. Subsequently, GPT-4 was utilized to process each paper, providing both Prompt 2 and the PDF file over three separate iterations for thorough analysis.

[6]https://www.nltk.org/search.html?q=stopwords

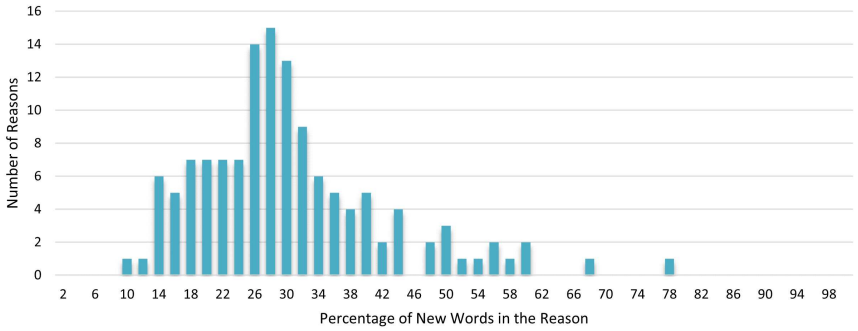

Figure 7: Distribution of Percentage of New Words Used by GPT-4 for Reasoning

Table 3
Scope Detection Percentage Distribution of GPT-4

Match Level	Percentage
Complete Match	50%
Intermediate Match	22%
No Match	28%

The yielded results are presented in Table 3. Some of the papers have multiple scopes whereas some of the papers focus on a single scope. GPT-4 correctly identified the scope of 50% of the research papers, and for 22%, it identified a portion of the scope annotated by subject experts. Within the 22% intermediate matched, there were 4% instances where GPT identified a smaller scope than the experts and 16% times it provided a broader range of scopes compared to what subject experts had identified.

3.3. Information extraction

This stage of the research is still under development, hence we present only the information retrieved by GPT-4 for a sample research paper [31] for the information extraction Prompt 3. The extracted information is presented in Table 4. According to the table, the GPT-4 successfully extracted the important information.

3.4. Limitations

This section details the limitations we encountered during the entire process of our study.

- **Noisy data retrieval** - Given the nature of Google Scholar and Pubmed, they retrieve and compile data from various publication databases. Consequently, when we automatically queried using APIs, the search results had several flaws, including missing abstracts, incomplete titles, abstracts and titles extracted from the body of texts, partial abstracts, or a high noise level in the retrieved data. For instance, when utilizing the scholarly API[7],

[7]https://pypi.org/project/scholarly/

Table 4
Information Extracted from a Sample Research Paper [31]

Aspect	Details
Background and Objective	**Context:** Focuses on improving radiotherapy planning through automated contouring tools. **Objective:** Assessing the performance of AI-Rad, a machine-learning automated contouring tool, compared to manual and another automated tool.
Methods	**Methodologies:** AI-Rad used for automated contouring in 28 patients, compared with manual and SS contours. Evaluation metrics included Dice similarity coefficient, sensitivity, precision, and Hausdorff distance.
Key Findings	**Main Results:** AI-Rad produced clinically acceptable contours, often superior to SS, with higher efficiency and minimal editing requirements. Some structures, like the larynx, were challenging, but AI-Rad showed promise in improving radiotherapy planning efficiency.

it was observed that requesting large numbers, such as 200, led to a significant portion of these results containing irrelevant or incorrect data. Therefore, we limited the Google Scholar search to the first 110 research papers.

- **Inconsistent Chat-GPT response** - As we already mentioned, Both GPT models were inconsistent in generating the response and provided different options in various iterations. Furthermore, when executing the prompt in bulk (e.g. 5 research papers together), we received responses for more than the number of research articles queried; maybe GPT is now lazy as mentioned in [34].
- **Limited Chat-GPT functionality** - We noticed that the inherent performance limitations of GPT models, specifically when using GPT-4, the message limit of 50 messages for every 3 hours slow down our analysis significantly [35]. Moreover, even with the paid version, additional charges for API querying limited us to manual execution of all the prompts hindering the efficiency of the automation.
- **Iterative prompt creation** - One observation we made during the prompt creation was that the prompts had to undergo several iterations of edits to refine them and achieve the optimal results reported in this study. We included only the prompts that achieved optimal performance due to space limitations.

4. Conclusion

This paper presents a preliminary study of ongoing research focusing a survey on the topic *AI for breast cancer treatment (BCT)*. We analyzed the effectiveness of using ChatGPT models for automating the research paper analysis for survey paper writing. Specifically, we evaluated GPT-3.5 and GPT-4 for automatic paper category identification, scope detection, and information extraction. Experiment results compared to ground truth data reveal GPT-4 model can be used to identify the category of the research papers for automating the research paper analysis. However, the model seems to struggle when accurately identifying the scope of a research study. Further, we detailed the limitations that could potentially hinder the adoption of GPT models for scholarly work. As a future work, we will extend this work to come up with a comprehensive taxonomy of BCT and compile a survey article on *AI for BCT*.

References

[1] R. Anyoha, The history of artificial intelligence, 2017. URL: https://sitn.hms.harvard.edu/flash/2017/history-artificial-intelligence/.

[2] P. Ratner, 10 of nikola tesla's most amazing predictions, 2016. URL: https://bigthink.com/the-future/10-nikola-teslas-most-amazing-predictions/.

[3] V. BUSH, As we may think, 1945. URL: https://cdn.theatlantic.com/media/archives/1945/07/176-1/132407932.pdf.

[4] A. M. TURING, I.—COMPUTING MACHINERY AND INTELLIGENCE, Mind LIX (1950) 433–460. doi:10.1093/mind/LIX.236.433. arXiv:https://academic.oup.com/mind/article-pdf/LIX/236/433/30123314/lix-236-433.p

[5] J. L. Elman, Finding structure in time, Cognitive Science 14 (1990) 179–211. URL: https://www.sciencedirect.com/science/article/pii/036402139090002E. doi:https://doi.org/10.1016/0364-0213(90)90002-E.

[6] K. Hu, Chatgpt sets record for fastest-growing user base - analyst note, 2024. URL: https://rb.gy/v32gzt.

[7] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, I. Polosukhin, Attention is all you need, 2023. arXiv:1706.03762.

[8] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. L. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, J. Schulman, J. Hilton, F. Kelton, L. Miller, M. Simens, A. Askell, P. Welinder, P. Christiano, J. Leike, R. Lowe, Training language models to follow instructions with human feedback, 2022. URL: https://openai.com/research/instruction-following. arXiv:2203.02155.

[9] P. P. Ray, Chatgpt: A comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope, Internet of Things and Cyber-Physical Systems 3 (2023) 121–154. URL: https://www.sciencedirect.com/science/article/pii/S266734522300024X. doi:https://doi.org/10.1016/j.iotcps.2023.04.003.

[10] R. Shewale, Chatgpt statistics for 2024 (users, trends & more), 2024. URL: https://rb.gy/9e5yvi.

[11] M. Abdullah, A. Madain, Y. Jararweh, Chatgpt: Fundamentals, applications and social impacts, in: 2022 Ninth International Conference on Social Networks Analysis, Management and Security (SNAMS), IEEE, 2022, pp. 1–8.

[12] J. Deng, Y. Lin, The benefits and challenges of chatgpt: An overview, Frontiers in Computing and Intelligent Systems 2 (2022) 81–83.

[13] Microsft, Microsoft and openai extend partnership, 2023. URL: https://blogs.microsoft.com/blog/2023/01/23/microsoftandopenaiextendpartnership/.

[14] Z. Bahroun, C. Anane, V. Ahmed, A. Zacca, Transforming education: A comprehensive review of generative artificial intelligence in educational settings through bibliometric and content analysis, Sustainability 15 (2023). URL: https://www.mdpi.com/2071-1050/15/17/12983. doi:10.3390/su151712983.

[15] H. Yu, Reflection on whether chat gpt should be banned by academia from the perspective of education and teaching, Frontiers in Psychology 14 (2023). URL: https://www.frontiersin.org/articles/10.3389/fpsyg.2023.1181712. doi:10.3389/fpsyg.2023.1181712.

[16] M. U. Haque, I. Dharmadasa, Z. T. Sworna, R. N. Rajapakse, H. Ahmad, " i think this is the

most disruptive technology": Exploring sentiments of chatgpt early adopters using twitter data, arXiv preprint arXiv:2212.05856 (2022).
[17] K. Hines, Exploring italy's chatgpt ban and its potential impact, 2023. URL: https://www.searchenginejournal.com/chatgpt-ban-italy/484157/.
[18] M. Farina, A. Lavazza, Chatgpt in society: emerging issues, Frontiers in artificial intelligence 6 (2023) 1130913. doi:10.3389/frai.2023.1130913.
[19] D. Baidoo-Anu, L. Owusu Ansah, Education in the era of generative artificial intelligence (ai): Understanding the potential benefits of chatgpt in promoting teaching and learning, SSRN Electronic Journal (2023). doi:10.2139/ssrn.4337484.
[20] O. Aydın, E. Karaarslan, OpenAI ChatGPT Generated Literature Review: Digital Twin in Healthcare, 2022, pp. 22–31. doi:10.2139/ssrn.4308687.
[21] G. C. A. Pividori M, publishing infrastructure for ai-assisted academic authoring., bioRxiv [Preprint] (2023). doi:10.1101/2023.01.21.525030.
[22] R. Golan, R. Reddy, A. Muthigi, R. Ramasamy, Artificial intelligence in academic writing: a paradigm-shifting technological advance, Nature Reviews Urology (2023) 1–2.
[23] H. M., Could ai help you to write your next paper?, Nature 5 (2023). doi:https://doi.org/10.1038/d41586-022-03479-w.
[24] J. de la Torre-López, A. Ramírez, J. R. Romero, Artificial intelligence to automate the systematic review of scientific literature, Computing (2023) 1–24.
[25] I. L. Alberts, L. Mercolli, T. Pyka, G. Prenosil, K. Shi, A. Rominger, A. Afshar-Oromieh, Large language models (llm) and chatgpt: what will the impact on nuclear medicine be?, European journal of nuclear medicine and molecular imaging 50 (2023) 1549–1552.
[26] H. Yu, Reflection on whether chat gpt should be banned by academia from the perspective of education and teaching, Frontiers in psychology 14 (2023) 1181712. URL: https://europepmc.org/articles/PMC10267436. doi:10.3389/fpsyg.2023.1181712.
[27] G. Hu, Challenges for enforcing editorial policies on ai-generated papers, Accountability in Research 0 (2023) 1–3. URL: https://doi.org/10.1080/08989621.2023.2184262. doi:10.1080/08989621.2023.2184262.
[28] C. A. Gao, F. M. Howard, N. S. Markov, E. C. Dyer, S. Ramesh, Y. Luo, A. T. Pearson, Comparing scientific abstracts generated by chatgpt to real abstracts with detectors and blinded human reviewers, NPJ Digital Medicine 6 (2023) 75.
[29] The ai writing on the wall, Nature Machine Intelligence 5 (2023). doi:https://doi.org/10.1038/s42256-023-00613-9.
[30] A. Everett, Unlocking the potential of gpt: Why the same prompt can lead to different results, 2023. URL: https://www.promptprodigy.blog/post/unlocking-the-potential-of-gpt-why-the-same-prompt-can-lead-to-different-results.
[31] Y. Hu, H. Nguyen, C. Smith, T. Chen, M. Byrne, B. Archibald-Heeren, J. Rijken, T. Aland, Clinical assessment of a novel machine-learning automated contouring tool for radiotherapy planning, Journal of Applied Clinical Medical Physics 24 (2023). doi:10.1002/acm2.13949.
[32] OpenAI, Api reference - openai api, 2023. URL: https://platform.openai.com/docs/api-reference/introduction.
[33] A. Everett, Chatting with machines: The turing test and the rise of chatgpt, 2023. URL: https://www.promptprodigy.blog/post/

chatting-with-machines-the-turing-test-and-the-rise-of-chatgpt.
[34] E. Price, Openai acknowledges gpt-4 is getting 'lazy', 2023. URL: https://www.pcmag.com/news/openai-acknowledges-gpt-4-is-getting-lazy.
[35] Natalie, Higher message limits for gpt-4 (july 19, 2023), 2023. URL: https://help.openai.com/en/articles/6825453-chatgpt-release-notes.