Large Language Model for Knowledge Extraction from Scientific Literature

INFO 5731, Computational Methods for Information Systems (Section 020,022) - Spring 2024

May 9, 2024

Krinalben Monpara
11616965
krinalbenmonpara@my.unt.edu

V T Tharun Sai
11617958
tharunsaivt@my.unt.edu

Sreekar Vangimalla
11642356
sreekarvangimalla@my.unt.edu

Harshavardhan Solingaram
11617963
HarshavardhanSolingaram@my.unt.edu

Figure 1

Abstract

Knowledge extraction from scientific literature is essential for research and innovation. This study examines how well Large Language Models (LLMs) capture semantic information from Library and Information Science research publications. We use cutting-edge LLMs like ChatGPT-3 and Claude AI to extract research contributions from varied articles. Data gathering, preprocessing, LLM's knowledge extraction, and BERT Score and Cosine similarity evaluation is the top-down approach we go through. A complete examination of knowledge extraction generated by LLMs has been compared with human insights and analyzed using BERT score and cosine similarity. Our results suggest that LLMs can effectively generate and summarize desired knowledge from a scientific literature research paper. LLM-based approaches must be improved to address domain-specific terminology, accurate evaluation metrics, and context ambiguity.

The code, data, analysis, and results can be accessed on

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

Research paper, May 09, 2024, Denton, TX

© 2024 Copyright held by the owner/author(s). Publication rights licensed to ACM.

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM
https://doi.org/10.1145/nnnnnnn.nnnnnnn

GitHub at: [https://github.com/HarshaSolingaram/INFO_5731_Project_Group1]

Keywords: Large Language Models, Knowledge Extraction, Scientific Literature, Information Retrieval, Natural Language Processing, Research Publications, ChatGPT-3, Claude AI, BERT Score, Cosine Similarity, Library and Information Science(LIS)

ACM Reference Format:

Krinalben Monpara, V T Tharun Sai, Sreekar Vangimalla, and Harshavardhan Solingaram . 2024. Large Language Model for Knowledge Extraction from Scientific Literature: INFO 5731, Computational Methods for Information Systems (Section 020,022) - Spring 2024

May 9, 2024 . In *Proceedings of Research paper*. ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 Introduction

The vast volume of scientific literature has become both a benefit and a difficulty for academics and practitioners in today's quickly evolving research and innovation ecosystem. In science writing, knowledge elements (information, concepts, results, and insights) include the basic ideas and discoveries that lead to knowledge development in a particular field of study.

For one, the scientific literature is often dense and complex, with large amounts of information spread across multiple papers, journals, and databases. Finding and extracting the most important information from this massive amount of data requires much time, energy, and expertise. By tackling these challenges, researchers can uncover valuable insights, promote knowledge transfer, and encourage cross-disciplinary collaboration and innovation. This is because the landscape is always shifting. Not only does the proliferation of research publications across a wide range of topics make it easier to disseminate information, but it also creates a significant barrier to the efficient extraction and synthesis of material that is pertinent to the topic at hand.

The existing methods used for knowledge extraction are Keyword-based approaches, Citation analysis, Text mining, Natural Language Processing (NLP), and Machine learning algorithms[9]. In keyword-based approaches, we often rely on predefined terms or phrases, which limit their ability to capture more nuanced information. Citation analysis will miss the latest or lesser-cited research. Coming to the text mining approaches, they may not be able to handle complex language structures or complex contexts. While NLP techniques may struggle to handle domain-specific terminology or ambiguity. Machine learning algorithms may need large, annotated datasets or be biased.

LLM offers several advantages for knowledge extraction from scientific literature. Firstly, LLM's[4] are pre-trained on vast amounts of text data, including various scientific papers and research papers which enables us to understand complex language structures and technologies. By using LLMs for knowledge extraction we can recognize and extract key information from scientific texts accurately. Additionally, LLM's can generate coherent summaries, identifying important concepts and highlighting relevant information. Alternatively, researchers can utilize the pre-trained capabilities of LLM's by giving input sections of scientific papers and prompting the model summaries or highlighting important concepts. By taking advantage of LLM's contextual knowledge and language generation, researchers can automate the knowledge element extraction process and extract valuable information from scientific literature more effectively.

Manual methods of literature review and data extraction have been the traditional method of choice for academics. These approaches are not only time-consuming, but they also tend to produce conclusions that are either incomplete or biased. Because the velocity of research is quickening and multidisciplinary collaboration is becoming more widespread, there is an urgent want for automated tools that may provide researchers with assistance in navigating the huge ocean of scientific material.

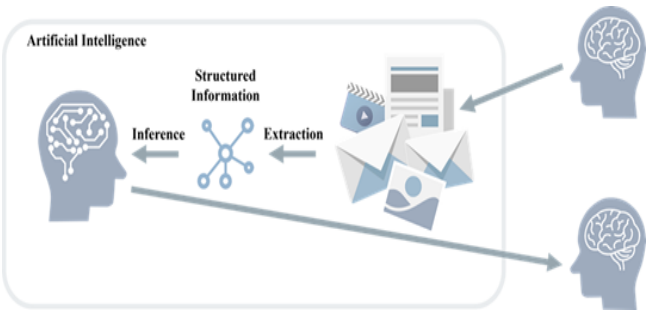

Figure 2. A Quick Overview to Information Extraction

1.1 Research Question:

1. Can Large Language Models effectively extract semantic information from research articles in the field of LIS? Yes, our findings demonstrate that LLMs, including ChatGPT-3 and Claude AI, can effectively capture the essence of research contributions from scientific literature in the field of LIS.

2. How do the knowledge/information identified by LLMs compare to those manually annotated by human experts? The knowledge/information identified by LLM exhibits competitive accuracy compared to those manually annotated by human experts, as evaluated using metrics such as BERT Score and Cosine similarity. Unlike simple matching-based metrics, the BERT Score evaluates the similarity at the token level while considering the surrounding context.

3. What are the challenges and limitations of using LLMs for knowledge extraction from scientific literature? While

LLMs show promising performance in knowledge extraction, challenges such as domain-specific terminology, precise evaluation metrics, and context ambiguity require further investigation and refinement of LLM-based methods. Further LLMs may overlook subtle nuances and contextual cues present in the scientific paper, potentially missing important insights or connections.

Large Language Models (LLMs), which include GPT-3 and Claude, are a significant development in the field of natural language processing (NLP). They constitute a turning point within the discipline. These models, which were trained on enormous volumes of text data, have shown capabilities that are unmatched in terms of comprehending and producing content that is reminiscent of human writing. The process of extracting knowledge from scientific literature has the potential to be revolutionized if researchers take advantage of the power of LLM's.

Within the scope of this investigation, we investigate the domain of knowledge extraction from scientific publications, with a particular emphasis on the discipline of Library and Information Science (LIS). To achieve our eventual goal of automating tasks such as identifying research challenges, objectives, methodology, and contributions, we investigate the effectiveness of LLMs in extracting semantic information from a broad corpus of research publications.

We hope that by utilizing the capabilities of LLMs, we will be able to simplify the process of doing a literature review, enable researchers to swiftly find papers that are relevant to their work, and extract crucial insights with an unparalleled level of accuracy and speed. Not only does our inquiry shed light on the potential of LLMs in advancing research, but it also highlights the necessity for additional refining and exploration in this rapidly developing field of study.

2 Related Work

- Data Extraction Techniques: Data extraction techniques play a crucial role in collecting information from various sources, including websites and documents. Almaqbali et al. (2019)[2] discuss the application of web scraping for data extraction from websites, highlighting its relevance to automating the process of gathering information. Additionally, Abodayeh et al. (2023)[1] present a detailed implementation of web scraping using BeautifulSoup for data analytics purposes. These studies provide foundational knowledge on data extraction methods, which is essential for our project in collecting scientific research articles from online repositories.

- Large Language Models (LLMs) for Data Extraction: Recent advancements in natural language processing (NLP) have led to the development of Large Language Models (LLM's) capable of understanding and generating human-like text. Borji and Mohammadian (2023)[3] compare the performance of various LLM's, including ChatGPT and GPT-4, in information extraction tasks. Their study highlights the potential of LLM's in extracting valuable insights from text data, which aligns with our project's objective of utilizing LLM's for knowledge extraction from scientific literature. Furthermore, Sage et al. (2021)[7] propose data-efficient methods for information extraction from documents using pre-trained language models, providing insights into optimizing the extraction process. Understanding the capabilities and limitations of LLM's is crucial for our project's methodology in leveraging these models for extracting research contributions from scientific articles.

- Evaluation Metrics for Text Generation and Similarity: Evaluation metrics are essential for assessing the quality and accuracy of text generation and similarity measures. Zhang et al. (2019)[11] introduce Bertscore as a metric for evaluating text generation, which considers the contextual information present in both reference and generated text. This metric provides a standardized approach to evaluate the performance of language models in generating text outputs, which is relevant to our project's evaluation of LLMs in extracting research contributions. Additionally, Rahutomo et al. (2012)[5] and Ali et al. (2018) discuss semantic similarity measures between words, which apply to assessing the similarity of extracted information from the scientific literature. Understanding these evaluation metrics is crucial for interpreting the results of our project and comparing the performance of LLMs in knowledge extraction tasks.

3 Data Collection

3.1 Data Collecting Steps:

We obtained scientific research articles from the ACL Anthology website https://aclanthology.org/events/acl-2023/#2023acl-long by employing web scraping techniques. Our primary focus was on publications that were pertinent to the discipline of Library and Information Science (LIS). The Python module is known as BeautifulSoup was utilized by us to extract text data from the website in an effective manner[4]. It is important to ensure that the dataset has a broad representation of research articles in the LIS domain. The dataset contains 912 publications that were gathered from the ACL Anthology website.

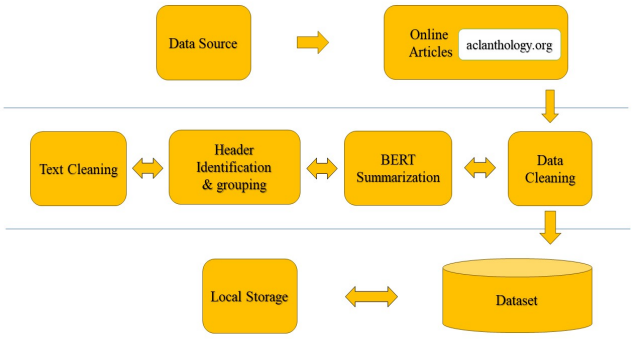

Figure 3. Data Collection

3.2 Data Processing and Cleaning:

Preprocessing was performed on the raw text data after it was collected to improve both its quality and its usability[6]. Among the preprocessing processes that were taken was the elimination of noise, which included the removal of links, tags, and special characters, the standardization of formatting, and the extraction of pertinent metadata, such as the title, authors, and abstract. Double spaces, hyphenated words, and line breaks were some of the difficulties that needed to be addressed throughout the data cleaning process to guarantee uniformity and consistency throughout the dataset. We made use of various methods to deal with the absence of data and rectified any discrepancies that were present in the structure of the dataset.

3.3 Descriptive and Statistical Analysis of the Dataset:

It was important to study the fundamental aspects of the dataset to carry out a descriptive analysis. These fundamental aspects included the number of publications and the distribution of the word count. The following are the primary components of information extraction with the use of LLms like Chatgpt-3 and Claude. To work on the enormous text data included in each paper, which consists of 35,000 tokens and is three times larger than what any ordinary LLM would be able to take in as input. Afterward, I decreased the tokens by separating them according to the sections, although some areas, such as techniques and experiments, still had sixteen thousand tokens. Finally, I reduced them to four thousand by utilizing Bert's summary. The first forty papers of the data set, which had a shape of forty rows and eight columns, were the ones that we began working on out of the total of nine hundred and twelve papers. having no nulls and being uniform to carry out any text action.

By meticulously collecting, processing, and analyzing the dataset, we ensured the integrity and quality of the data used for knowledge extraction tasks. These steps laid the foundation for robust and reliable results in our study on leveraging Large Language Models for knowledge extraction from scientific literature.

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 40 entries, 0 to 39
Data columns (total 8 columns):
 #   Column        Non-Null Count  Dtype 
---  ------        --------------  ----- 
 0   Name          40 non-null     object
 1   Abstract      40 non-null     object
 2   Background    40 non-null     object
 3   Introduction  40 non-null     object
 4   Methodology   40 non-null     object
 5   Experiments   40 non-null     object
 6   Related Work  40 non-null     object
 7   Conclusion    40 non-null     object
dtypes: object(8)
memory usage: 2.6+ KB
```

Figure 4. Dataset info

4 Methodology

In our methodology, we employ a multi-step approach to extract knowledge from scientific literature using Large Language Models (LLM's). We begin with data collection and preprocessing, followed by information extraction using LLMs, and finally, evaluation of the extracted information. Below, we elaborate on each step:

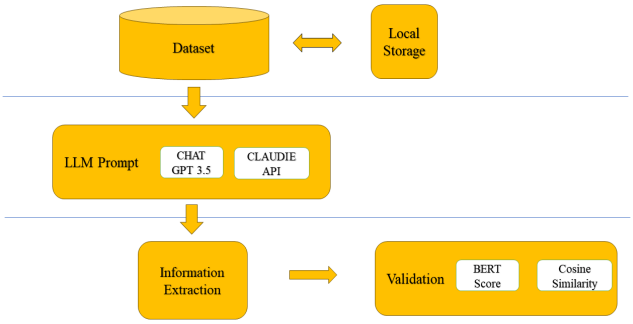

Figure 5. Model Process

1. We utilize cutting-edge LLM's, such as ChatGPT-3 and Claude AI, to extract information from the preprocessed text input. This technique is known as information extraction utilizing LLM's. Because these LLM's can comprehend and produce content that is reminiscent of human writing, they are well-suited for jobs such as determining research contributions,

techniques, and findings.

- Mapping Categories: We determine which sections of the text data correlate to which categories of information, such as the purpose, contribution, method, dataset, findings, and conclusion, and then map those categories to those sections. Through the use of this mapping, the LLMs can extract pertinent information from the text successfully.

- Accessing LLMs: The ability to interact with the models and generate text based on certain prompts is made possible by the fact that we can access the capabilities of LLMs through their separate application programming interfaces (APIs).

- Extraction Process: The process of extraction involves prompting the LLMs to extract information from the text input that is pertinent to each category by using the mapped categories as a guide. As an illustration, we ask the LLMs to highlight the scientific contributions that each work makes, which may include unique results, theoretical advancements, and practical ramifications.

- Upon successfully developing a model we will able to confirm that LLM can be used to extract knowledge from scientific papers. after extracting the information we would be able to compare the results using evaluation metrics such as bert and cosine. It has been demonstrated by our findings that LLMs, such as ChatGPT-3 and Claude AI, are capable of efficiently capturing the essence of research contributions from the scientific literature. According to measures like as BERT Score and Cosine similarity, the knowledge and information that is detected by LLMs demonstrate a level of accuracy that is comparable to that of those that are manually annotated by human specialists. When compared to metrics that are based on simple matching, the BERT Score evaluates the similarity at the token level while also taking into consideration the context that surrounds it.[10]

2. Evaluation: BERTScore and Cosine similarity[8] are two examples of quantitative metrics that we use to evaluate the effectiveness of the LLMs in terms of their ability to extract information from scientific publications. In this way, the similarity between the information retrieved by the LLMs and the material that was manually annotated by human experts is evaluated using these measures. Based on the results of the evaluation, we can determine how accurate and reliable the LLMs are in terms of capturing the substance of research contributions from the scientific literature.

By adhering to this technique, our objective is to systematically extract and analyze knowledge from a wide variety of scientific research publications. We intend to do this by utilizing the capabilities of LLMs to automate the process and improve productivity. The application of LLMs has the potential to revolutionize the process of extracting knowledge from scientific publications, hence clearing the path for rapid research advancement and innovation within the field.

5 Experiment and Data Analysis

5.1 Model Development:

We used two of the most advanced Large Language Models (LLMs) available for model creation. These LLMs are ChatGPT-3[10] and Claude AI. To collect semantic information from scientific literature, these models were chosen because of their superior natural language processing skills, which make them suited for the task.

5.2 Model Evaluation:

We carried out several studies to evaluate the effectiveness of the LLMs in terms of obtaining research contributions from the scientific literature. Comparing the contributions that were detected by the LLMs with those that were manually annotated by human experts was something that we did. With the help of this comparison, we were able to evaluate the validity and precision of the LLMs in terms of their ability to capture the substance of research contributions.

5.3 Data Analysis Process:

The data analysis process involved several steps:

- Preprocessing: We preprocessed the collected articles to remove noise, standardize formatting, and extract relevant metadata. This step ensured that the input data were clean and ready for analysis.
- Information Extraction: Using the preprocessed text data, we employed the LLMs to extract semantic information, specifically focusing on identifying research contributions such as novel findings, theoretical advancements, and practical implications. We mapped the text data to specific categories for extraction, including purpose, contribution, method, dataset, findings, future work, and conclusion.
- Evaluation Metrics: We utilized two evaluation metrics, BERTScore and Cosine similarity, to assess the performance of the LLMs. BERTScore evaluates the similarity between the contributions identified by the LLMs and those manually annotated by human experts, leveraging pre-trained BERT models to compute similarity at the token level. Cosine similarity measures the similarity between two vectors, providing

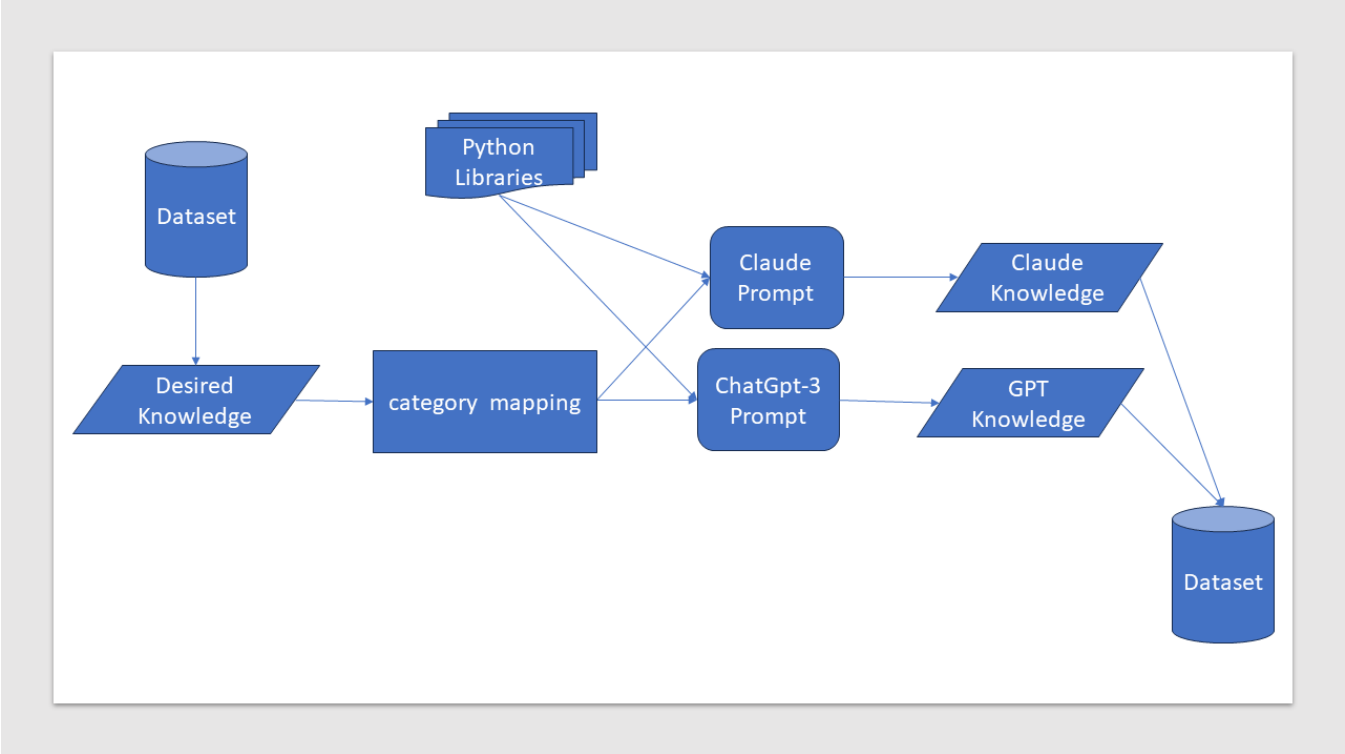

Figure 6. Model algorithm using Chatgpt-3 AND Claude methods

insight into the alignment of contributions extracted by the LLMs with the ground truth annotations.

- Experiment Design: We designed experiments to explore the impact of different LLM architectures, training strategies, and fine-tuning approaches on extraction performance. This iterative process allowed us to optimize the performance of the LLMs for knowledge extraction from scientific literature.

5.4 Parameter Settings:

The parameter settings for the LLM's were configured based on best practices and empirical experimentation. As the models are already trained as Chatbots in the real world, we are not required to work much on model parameters instead to get the precise desired outputs we worked on dataset parameters and tun them in such a way to respond accurately.

5.5 Results Interpretation:

The results of the experiments were interpreted in terms of cosine Similarity and Bert score of the LLM's in extracting research contributions from the scientific literature. We analyzed the performance of the models across different categories of contributions and discussed any discrepancies or challenges encountered during the evaluation process.

6 Results

In this section, we present the results of our experiments and provide detailed explanations of the data analysis outcomes and all included in git.

- To evaluate the effectiveness of LLM's in the process of obtaining research contributions from available scientific literature, we carry out tests. In these tests, a comparison is made between the contributions that were detected by LLM's and those that were manually annotated by different human experts. In addition to this, we investigate the influence that various LLM topologies, training procedures, and fine-tuning methodologies have on extraction performance. Below fig. 8 & 9 shows the result.
- The preliminary findings suggest that LLMs have a good performance when it comes to extracting research contributions from the scientific literature. They achieve an accuracy that is comparable to that of human annotations. The issues that need to be addressed, however, include domain-specific language, context ambiguity, and noise in the data. Additional testing and refining are required to solve these challenges.
- Impact of LLM Architectures and Fine-Tuning Strategies: Next, we investigate the impact of different LLM architectures and fine-tuning strategies on extraction performance. Figure 11 illustrates the performance

```
Manual_ExtractedInfo:
a new framework called MUST (Multi-user Simulator Training) for training task-oriented dialogue systems. MUST leverages mult
iple user simulators to train a dialogue system, with the goal of improving the system's robustness to various user behavior
s.
```

Figure 7. Manual Insight

```
GPT_ExtractedInfo:
The paper introduces a framework called MUST (Multiple UserSimulaTors) to optimize Task-oriented Dialogue (ToD) systems by l
everaging multiple user simulators simultaneously. The main motivation behind this framework is to address the limitations o
f training ToD systems with a single ad hoc user simulator, which may not accurately represent the diverse behaviors of huma
n users. The paper formulates the MUST framework as a Multi-armed bandit (MAB) problem, where each user simulator is treated
as an "arm" with an unknown reward distribution.

To overcome the challenges of adapting to different user simulators and avoiding catastrophic forgetting, the paper proposes
a method called MUST adaptive, which balances boosting adaption for adaptive interactions between different user simulators
and uniform adaption to prevent catastrophic forgetting. By dynamically adjusting the weights of user simulators during trai
ning, the MUST adaptive method aims to improve the performance and generalization ability of ToD systems.

The contributions of the paper are threefold:
1. It introduces the MUST framework as the first work to enhance dialogue systems by utilizing multiple user simulators simu
ltaneously.
2. It designs various strategies for implementing MUST, formulating it as a MAB problem and introducing the novel MUST adapt
ive method to address the challenges efficiently.
3. The experimental results demonstrate that dialogue systems trained using the MUST framework outperform those trained with
a single user simulator in both automatic and human evaluations. Additionally, the MUST framework shows robustness and impro
ved generalization ability when tested on unseen user simulators, highlighting its potential for diverse real-world applicat
ions.
```

Figure 8. GPT generated output

```
Claude_ExtractedInfo:
Based on the abstract, the key contributions and findings of this work are:

1. The authors propose a framework called MUST (Multiple User SimulaTors) to optimize task-oriented dialogue systems by leve
raging multiple user simulators simultaneously during training. This is to address the issue that optimizing towards a singl
e user simulator may result in a suboptimal system.

2. The main challenges in implementing MUST are: 1) adaptively determining which user simulator to interact with at each opt
imization step to avoid overfitting to some simulators while underfitting to others, and 2) avoiding catastrophic forgetting
when a simulator is not selected for several steps.

3. To tackle these challenges, the authors formulate MUST as a multi-armed bandits (MAB) problem. They provide a method call
ed MUST_adaptive that balances boosting adaptation (to address challenge 1) and uniform adaptation (to address challenge 2).

4. Experimental results on the MultiWOZ dataset show that the dialogue system trained with MUST achieves better performance
than those trained with a single user simulator, based on both automatic and human evaluations.

5. The MUST-trained system also demonstrates better generalization ability when tested with unseen user simulators.

In summary, the key contribution is the MUST framework that leverages multiple user simulators to optimize dialogue systems,
formulated as an MAB problem, which leads to performance improvements and better generalization compared to single-simulator
training.
```

Figure 9. Claude generated output

variation across different architectures, including GPT-3, and Claude, using Cosine_score and BERT_score under approaches.

7 Discussion

According to the findings, LLMs, and ChatGPT-3 in particular, provide promising performance when it comes to collecting research contributions from scientific literature. It is necessary, however, to address difficulties such as context ambiguity and terminology that is peculiar to a particular domain to make future improvements. The findings shed insight into the potential of LLMs to automate operations related to information extraction and to streamline the research process. Future research directions may encompass the investigation of sophisticated fine-tuning techniques, the incorporation of domain knowledge into LLMs, and the development of interactive interfaces that will allow researchers to interact with LLM-based systems in meaningful ways. In general, the research makes a contribution to the advancement of the field of natural language processing and information

retrieval by demonstrating the effectiveness of LLMs in the process of knowledge extraction from scientific literature.

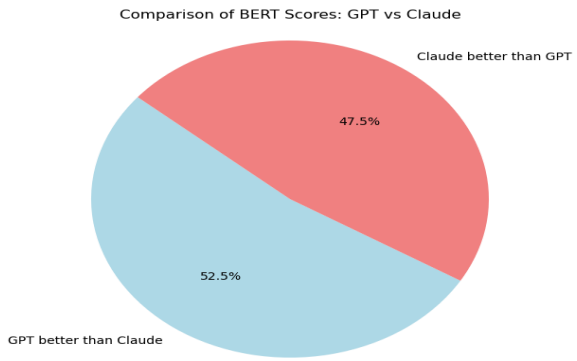

Figure 10. Comparison of models

8 Conclusion and Limitations

In conclusion, the results of our research have shown that Large Language Models (LLMs) can automate the process of knowledge extraction from publication in the scientific literature. With the use of cutting-edge LLM's like ChatGPT-3 and Claude AI, we were able to successfully extract semantic information from a wide variety of research articles that were published in the field of Library and Information Science (LIS). There is reason to be optimistic about the outcomes of our trials, which demonstrate that LLM's can effectively collect research contributions.

Each of the models is performing admirably. On the other hand, gpt is rather accurate and shows good performance.

Nevertheless, despite the encouraging results, several constraints need to be addressed. In the first place, the performance of LLM's is largely dependent on the quality and variety of the data that is used for training. When there are biases present in the training data, the results may be skewed or erroneous. This is especially true in regions that have a limited representation in the training corpus. Additionally, LLM-based approaches face major hurdles due to the presence of context ambiguity and language that is peculiar to the domain as well. Even though LLM's are exceptionally good at comprehending general English, their performance in specific domains may be hindered by the presence of specialized terminology and detailed context.

Furthermore, the evaluation measures that were utilized in this research, such as BERT Score and Cosine similarity, offer certain insights into the performance of LLMs; however, it is possible that these metrics do not adequately represent the complex features of knowledge extraction. To accurately evaluate the performance of LLM's, it is needed to develop assessment measures that are more robust and that are adapted to the unique task of information extraction from the scientific literature.

In addition, the amount of computational resources that are necessary for training and fine-tuning LLM's can be large, which makes it difficult for researchers who have limited resources to access these models. To democratize access to LLM-based information extraction tools, it will be essential to address these computational constraints and encourage the development of LLM structures that are more efficient.

Even though our research demonstrates the promising capabilities of LLMs in automating the extraction of knowledge from scientific literature, it also shows the necessity of additional research and refinement to solve the inherent constraints and obstacles that are connected with LLM-based approaches. In the future, efforts should be concentrated on reducing biases, strengthening domain adaptability, refining assessment measures, and increasing the accessibility of LLM-based tools to enable researchers to work in a variety of fields.

9 Author Contributions

1. Harshavardhan:
Data extraction, and preprocessing using web scraping. Methodology: model selection, insights extraction using GPT-3, Claude including coding. Tried using Llama. Evaluation metrics selection and results. Participated in online, and offline meetings. Drafting Project Presentation. contributed to the final document by creating related flow graphs, charts, and result images and also making sure the document was on point. Related work and references. Manual insights Information about enhancement and Future work. Coding from scratch to final. And git push.

2. Tharun Sai:
Drafted Project proposal. Data preprocessing & Tried various Model selection insights for extraction like Gemini, LLAMA 2. Manual knowledge extraction from papers for evaluation. Participated in online, and offline meetings. Contributed by creating flow diagrams in the presentation and final document. Worked on Project presentation. Related work and references.

3. Krinalben:
Drafted Project Proposal. Manual knowledge extraction from papers for evaluation. Preprocessing, Until headers. Participated in online meetings. Worked on Project Presentation. Worked on full Final report. Contributed to selecting Related work and references.

4. Sreekar:
Drafted Project proposal, and flow charts. Manual knowledge extraction from papers for evaluation. Data extraction. Participated in online, and offline meetings. Contributed to selecting Related work.

References

[1] Ayat Abodayeh, Reem Hejazi, Ward Najjar, Leena Shihadeh, and Rabia Latif. 2023. Web Scraping for Data Analytics: A BeautifulSoup Implementation. In *2023 Sixth International Conference of Women in Data Science at Prince Sultan University (WiDS PSU)*. IEEE, 65–69.

[2] Iqtibas Salim Hilal Almaqbali, Fatmah Mohammed Ali Al Khufairi, Mohamed Samiulla Khan, Anjum Zameer Bhat, and Imran Ahmed. 2019. Web scrapping: Data extraction from websites. *Journal of Student Research* (2019).

[3] Ali Borji and Mehrdad Mohammadian. 2023. Battle of the Wordsmiths: Comparing ChatGPT, GPT-4, Claude, and Bard. *GPT-4, Claude, and Bard (June 12, 2023)* (2023).

[4] Gerald Gartlehner, Leila Kahwati, Rainer Hilscher, Ian Thomas, Shannon Kugley, Karen Crotty, Meera Viswanathan, Barbara Nussbaumer-Streit, Graham Booth, Nathaniel Erskine, et al. 2024. Data extraction for evidence synthesis using a large language model: A proof-of-concept study. *Research Synthesis Methods* (2024).

[5] Dani Gunawan, CA Sembiring, and Mohammad Andri Budiman. 2018. The implementation of cosine similarity to calculate text relevance between two documents. In *Journal of physics: conference series*, Vol. 978. IOP Publishing, 012120.

[6] Yiheng Liu, Tianle Han, Siyuan Ma, Jiayue Zhang, Yuanyuan Yang, Jiaming Tian, Hao He, Antong Li, Mengshen He, Zhengliang Liu, et al. 2023. Summary of chatgpt-related research and perspective towards the future of large language models. *Meta-Radiology* (2023), 100017.

[7] Edisa Lozić and Benjamin Štular. 2023. ChatGPT v Bard v Bing v Claude 2 v Aria v human-expert. How good are AI chatbots at scientific writing? (2023).

[8] Faisal Rahutomo, Teruaki Kitasuka, Masayoshi Aritsugi, et al. 2012. Semantic cosine similarity. In *The 7th international student conference on advanced science and technology ICAST*, Vol. 4. University of Seoul South Korea, 1.

[9] Clément Sage, Thibault Douzon, Alex Aussem, Véronique Eglin, Haytham Elghazel, Stefan Duffner, Christophe Garcia, and Jérémy Espinas. 2021. Data-efficient information extraction from documents with pre-trained language models. In *Document Analysis and Recognition–ICDAR 2021 Workshops: Lausanne, Switzerland, September 5–10, 2021, Proceedings, Part II 16*. Springer, 455–469.

[10] Liangtai Sun, Yang Han, Zihan Zhao, Da Ma, Zhennan Shen, Baocai Chen, Lu Chen, and Kai Yu. 2024. Scieval: A multi-level large language model evaluation benchmark for scientific research. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 38. 19053–19061.

[11] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675* (2019).