Large Language Model for Knowledge Extraction from Scientific Literature

By Dr. Haihua Chen

May 5, 2024

Group No - 1

Krinalben Monpara – 11616965, Harshavardhan Solingaram - 11617963

Tharun Sai V T – 11617958, Sreekar Vangimalla - 11642356

Contents:

Abstract:

Knowledge extraction from scientific literature is a crucial task in facilitating research progress and innovation. In this study, we explore the effectiveness of leveraging Large Language Models (LLMs) for extracting semantic information from research articles in the field of Library and Information Science (LIS). Specifically, we focus on extracting research contributions from a diverse set of articles using state-of-the-art LLMs, including ChatGPT-3 and Claude AI. Our methodology involves data collection, preprocessing, information extraction using LLMs, and evaluation using BERT Score and Cosine similarity metrics. We present a detailed analysis of the experiments conducted to assess the performance of LLMs in extracting research contributions compared to human Insights . Our findings demonstrate promising results, indicating that LLMs can effectively capture the essence of research contributions from scientific literature. However, challenges such as domain-specific terminology, precise evaluation metrics and context ambiguity require further investigation and refinement of LLM-based methods.

Background:

- In today's information age, the exponential growth of scientific literature presents both opportunities and challenges for researchers. While the abundance of research articles enhances knowledge dissemination, it also exacerbates the difficulty of efficiently extracting and synthesizing relevant information.
- Traditional methods of literature review and data extraction are time-consuming and often yield incomplete results. Consequently, there is a growing need for automated tools to assist researchers in extracting semantic information from scientific literature.
- The emergence of Large Language Models (LLMs) represents a significant breakthrough in natural language processing (NLP). These models, such as GPT-3 and BERT, have demonstrated remarkable capabilities in understanding and generating human-like text. Leveraging LLMs for knowledge extraction from scientific literature holds promise in revolutionizing the research process by automating tasks such as identifying research problems, objectives, methodologies, and contributions.

Related Work:

- DATA Extraction:
  - "Web scrapping: Data extraction from websites." (Almaqbali, Iqtibas Salim Hilal, et al. 2019).
  - "Web Scraping for Data Analytics: A BeautifulSoup Implementation." (Abodayeh, Ayat, et al. WiDS PSU IEEE, 2023).
- LLM's Data Extraction:
  - Borji, Ali, and Mehrdad Mohammadian. "Battle of the Wordsmiths: Comparing ChatGPT, GPT-4, Claude, and Bard." GPT-4, Claude, and Bard (June 12, 2023) (2023).Lozić, E., & Štular, B. (2023).
  - "Data-efficient information extraction from documents with pre-trained language models." (Sage, Clément, et al. September 5–10, 2021).
- Evaluation:
  - "Bertscore: Evaluating text generation with bert." (Zhang, Tianyi, et al. arXiv preprint arXiv:1904.09675 - 2019).
  - "Semantic cosine similarity." (Rahutomo, F., Kitasuka, T., & Aritsugi, M. 2012, October)
  - "Semantic similarity measures between words: A brief survey." (Ali, Ashraf, Fayez Alfayez, and Hani Alquhayz. 2018)

Methodology:

- In our project, we aim to evaluate the effectiveness of using LLMs for knowledge extraction from scientific literature, with a focus on research contributions. Our methodology consists of the following steps:
- Data Collection: We collect a large dataset of scientific research articles about 912 papers from https://aclanthology.org/events/acl-2023/#2023acl-long.
- Data Preprocessing: The collected articles are pre-processed to remove noise, standardize formatting, and extract relevant metadata.
- Information Extraction: We employ state-of-the-art LLMs, such as ChatGPT-3 and Claude AI, to extract semantic information from the pre-processed text. Specifically, we focus on identifying the research contributions of each paper, including novel findings, theoretical advancements, and practical implications.
- Evaluation: We compare the contributions extracted by LLMs with those manually annotated by human experts using metrics such as BERTScore. This allows us to assess the accuracy and robustness of LLMs in capturing the essence of research contributions from scientific literature.

Collection & Preprocessing:

1. Collecting all the text data from every paper from the website using BeautifulSoup and webscarpping.
2. The website : https://aclanthology.org/events/acl-2023/#2023acl-long.
3. We have cleaned the text – double spaces, hyphenated words, join lines… etc.
4. Now we have divided the text of every paper using header into sections.

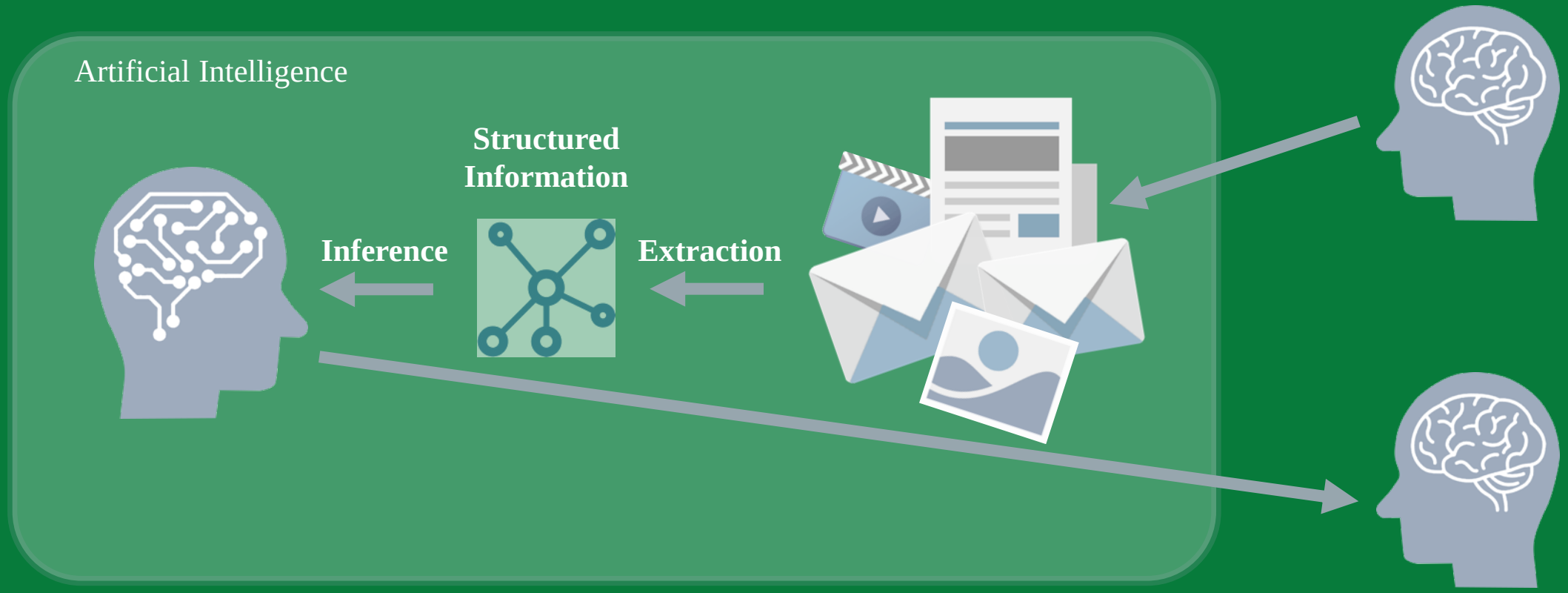
   headers = ['Abstract', 'Background', 'Introduction', 'Method', 'Methodology', 'Approach’, 'Analysis', 'Experiment Results', 'Experiments', 'Experimental Setup’, 'Evaluation', 'Related Work’, 'Related Works', 'Conclusion’, 'Conclusions']
5. We clubbed the similar columns to create a uniform dataset.
6. Now the text data we have is around 18,000 in some of the columns,

Continue…

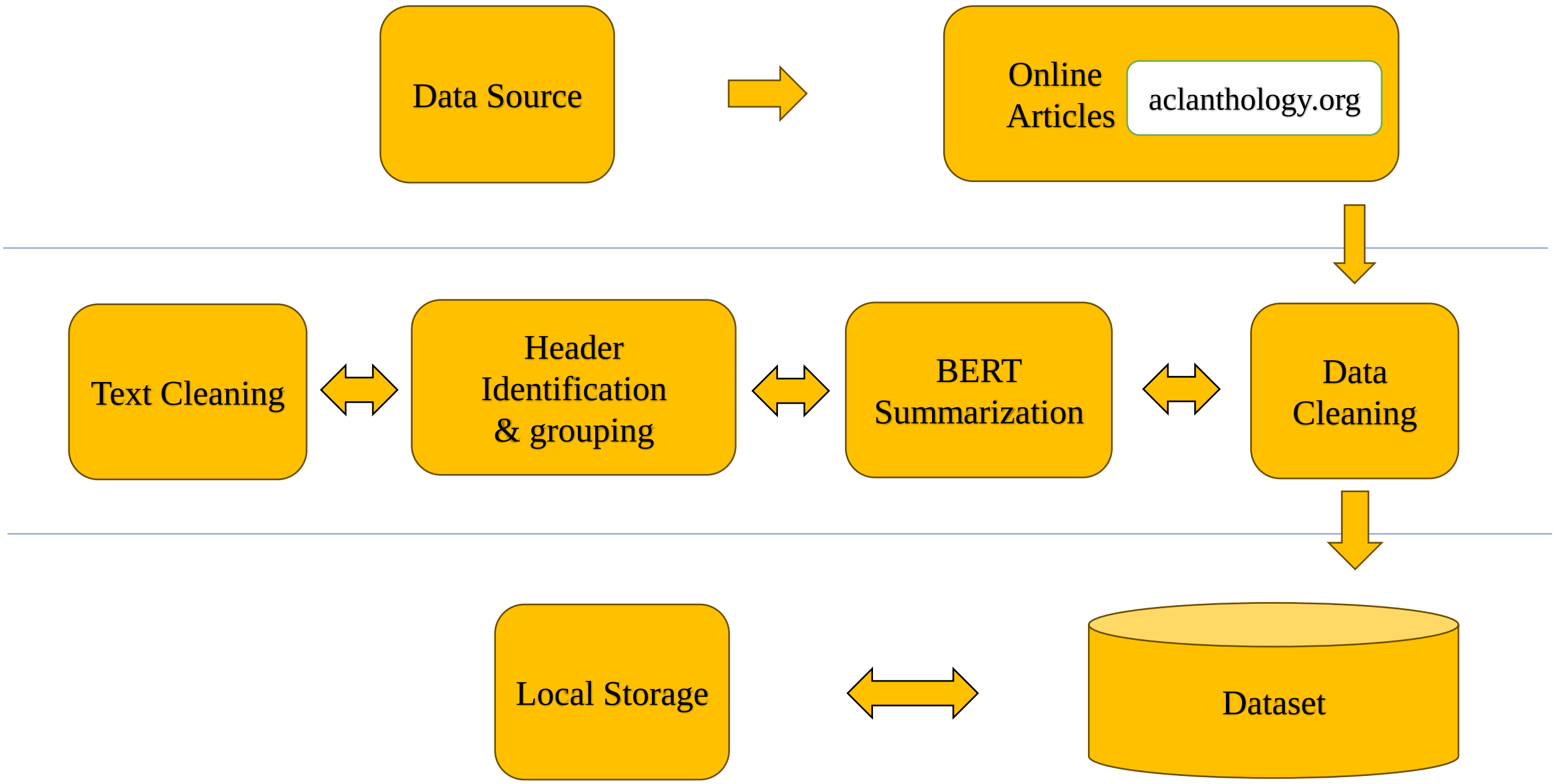
6. To reduce large tokens data, We have used Bert summarizer to reduce the text without losing its actual meaning.
7. This is the final Dataset we created.

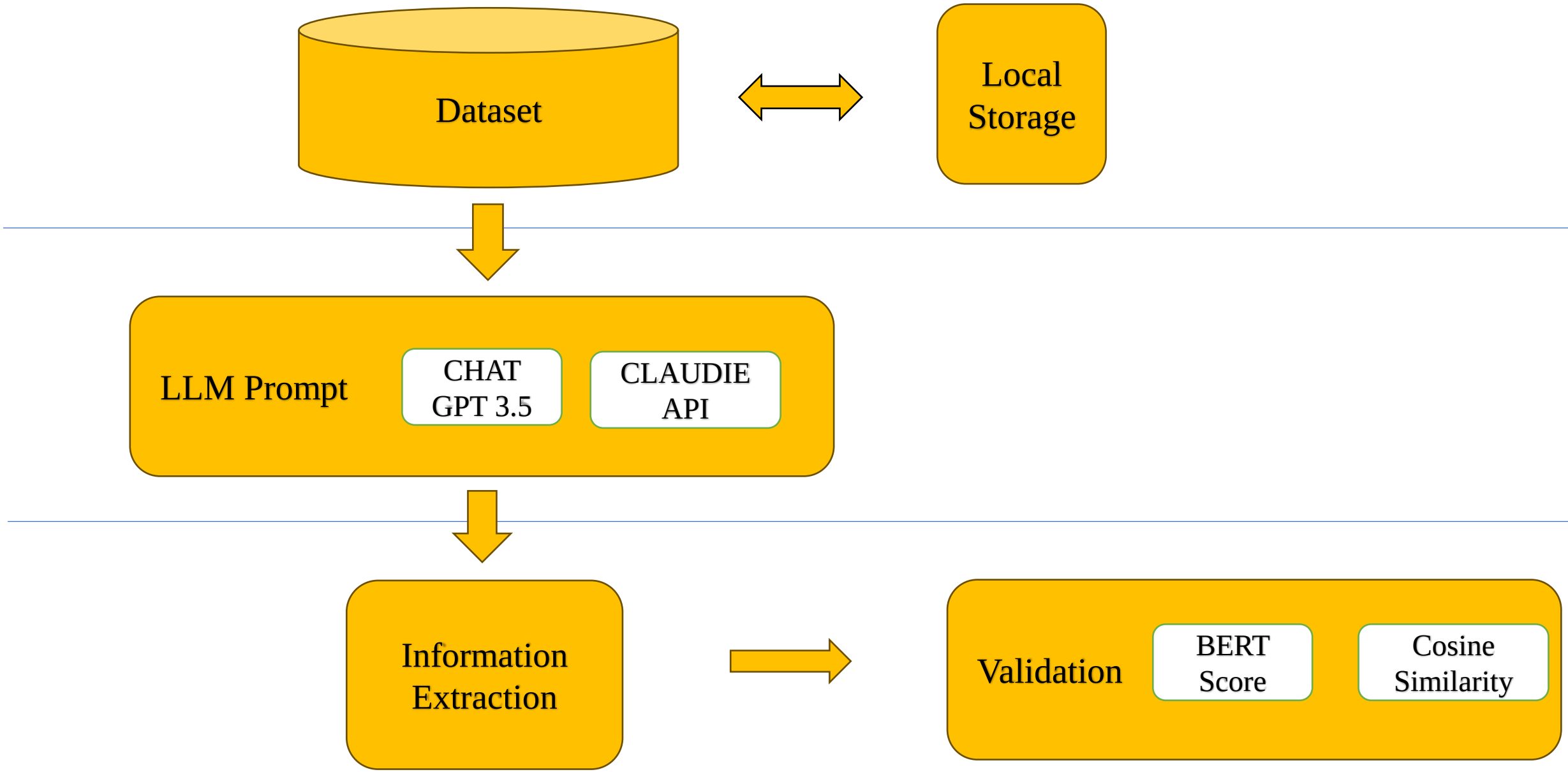
	Name	Abstract	Background	Introduction	Methodology	Experiments	Related Work	Conclusion
0	One Cannot Stand for Everyone! Leveraging Mult...	Abstract User simulators are agents designed t...	Background Dialogue system. Taskoriented dialo...	Introduction Taskoriented dialogue systems aim...	method called MUST adaptive that balances i) t...	experiments, we observed that the dialogue sys...		Conclusion In this paper, we propose a framewo...
1	SafeConv: Explaining and Correcting Conversati...	Abstract One of the main challenges opendomain...		Introduction Safety of artificial intelligence...	method is onetime checking and rewriting—direc...	Experiments show that the detected unsafe beha...	Related Work Dialogue Safety Datasets Datasets...	Conclusion In this paper, we study how to expl...
2	Detecting and Mitigating Hallucinations in Mac...	Abstract While the problem of hallucinations i...	Background and Setting In this section, we des...	Introduction Hallucinations in machine transla...	method that evaluates the percentage of the so...	experiments.1 1 Introduction Hallucinations in...		Conclusions We start by asking how far we can...
3	Explainable Recommendation with Personalized R...	Abstract Explainable recommendation is a techn...		Introduction Recent years have witnessed a gro...	method has been employed to identify and selec...	experiments on three datasets show that our mo...	Related Work 2.1 Explainable Recommendation wi...	Conclusion In this paper, we propose a novel m...

A Quick Overview to Information Extraction:



With the exponential growth of data from various sources especially the Internet, there is an increasing need for **Information Extraction** technology that extracts **machine-readable structured information** to support downstream applications.


Data Source
Online Articles
aclanthology.org
Text Cleaning
Header Identification & grouping
BERT Summarization
Data Cleaning
Local Storage
Dataset


Dataset
Local Storage
LLM Prompt
CHAT GPT 3.5
CLAUDIE API
Information Extraction
Validation
BERT Score
Cosine Similarity

Information Extraction:

1. We have used two of the top LLM’s. CHATGPT-3 and Claude.
2. Imported the dataset created,
3. Imported the required version of Openai == 0.28.0 for GPT and anthropic for Claude.
4. Accessing the chat prompt using API key of both the LLM’s.
5. Then mapping the column for the desired category to extract information.
6. Specify the category to extraction using LLM’s. here I used '**contribution**’.

```
category_mappings = {   'purpose': ['Abstract', 'Introduction', 'Background’],
                        'contribution': ['Abstract', 'Conclusion’],
                        'method': ['Methodology’],
                        'dataset': ['Experiments’],
                        'findings': ['Conclusion’],
                        'future_work': ['Conclusion’],
                        'conclusion': ['Conclusion’],          }
```

Claude:

```
    prompt = f"Extract the {category} of the paper:\n\n{text}"
    message = client.messages.create(
        model="claude-3-opus-20240229",
        max_tokens=4000,
        temperature=0,
        system=f"i want to know {category} of the text\n",
        messages=[
            {
                "role": "user",
                "content": [
                    {
                        "type": "text",
                        "text": f"{text}"
                    }
                ]
            }
        ]
    )
    response_message = message.content
```

GPT-3:

```
# Use ChatGPT to extract information based on the combined text
def chatgpt_completion(text):
    response = openai.ChatCompletion.create(
        model="gpt-3.5-turbo",
        messages=[
            {
            "role": "system",
            "content": f"Extract the {category} of the paper"
        },
        {
            "role": "user",
            "content": text
        }
        ]
    )
    return response.choices[0].message['content'].strip()
```

	Manual_ExtractedInfo	GPT_ExtractedInfo	Claude_ExtractedInfo
0	a new framework called MUST (Multi-user Simula...	The paper introduces a framework called MUST (...	Based on the abstract, the key contributions a...
1	The authors constructed a new dataset called S...	The paper introduces a new dataset called SAFE...	Based on the abstract, this paper makes the fo...
2	The authors propose a method that evaluates th...	The contribution of the paper can be summarize...	Based on the conclusions section, the key poin...
3	The authors propose a novel model, ERRA (Expla...	The paper proposes a novel model called ERRA (...	Based on the abstract, the key contributions o...

Results:

We conduct experiments to evaluate the performance of LLMs in extracting research contributions from scientific literature. The experiments involve comparing the contributions identified by LLMs with those manually annotated by human experts. Additionally, we explore the impact of different LLM architectures, training strategies, and fine-tuning approaches on the extraction performance.

Preliminary results indicate that LLMs exhibit promising performance in extracting research contributions from scientific literature, achieving competitive accuracy compared to human annotations. However, further experimentation and refinement are needed to address challenges such as domain-specific terminology, context ambiguity, and noise in the data.

Results (Continue…):

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

GPT_ExtractedInfo:
The paper introduces a framework called MUST (Multiple UserSimulaTors) to optimize Task-oriented Dialogue (ToD) systems by leveraging multiple user simulators simultaneously. The main motivation behind this framework is to address the limitations of training ToD systems with a single ad hoc user simulator, which may not accurately represent the diverse behaviors of human 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 training, 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 simultaneously.
2. It designs various strategies for implementing MUST, formulating it as a MAB problem and introducing the novel MUST adaptive 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 improved generalization ability when tested on unseen user simulators, highlighting its potential for diverse real-world applications.

Results (Continue…):

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 leveraging multiple user simulators simultaneously during training. This is to address the issue that optimizing towards a single 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 optimization 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 called 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.

Evaluation:

1. **Cosine similarity**: It is a metric used to measure how similar two vectors are, often used in text mining and document analysis .Cosine similarity calculates the cosine angle between two vectors, irrespective of their magnitudes. Cosine similarity is computationally efficient, especially for sparse high-dimensional data like text documents represented as term frequency-inverse document frequency (TF-IDF) vectors. It involves simple arithmetic operations, such as dot products and vector norms, making it suitable for large-scale text analysis tasks.

2. **Bert Score**: It is metric used to evaluate the quality of text generated by language models, especially in tasks like Text summarization, text generation and machine translation. BERT Score leverages pre-trained BERT (Bidirectional Encoder Representations from Transformers) models to compute the similarity between the reference (ground truth) text and the generated text. BERT Score considers the contextual information present in both the reference and generated text. Unlike simple matching-based metrics, BERT Score evaluates the similarity at the token level while considering the surrounding context.

	Manual_ExtractedInfo	GPT_ExtractedInfo	Claude_ExtractedInfo	Claude_BERTScore	Claude_CosineSimilarity	GPT_BERTScore	GPT_CosineSimilarity
0	a new framework called MUST (Multi-user Simula...	The paper introduces a framework called MUST (...	Based on the abstract, the key contributions a...	0.856249	0.522594	0.861111	0.530671
1	The authors constructed a new dataset called S...	The paper introduces a new dataset called SAFE...	Based on the abstract, this paper makes the fo...	0.871570	0.693517	0.858482	0.627444
2	The authors propose a method that evaluates th...	The contribution of the paper can be summarize...	Based on the conclusions section, the key poin...	0.853713	0.431445	0.869151	0.407156

```
Shape of dataset: (40, 7)
'GPT' column is better than 'Claude' column using bert score: 21
'Claude' column is better than 'GPT' column using bert score: 19
```

Discussion and Implications:

- The findings of our study have several implications for both researchers and practitioners in the field of Library and Information Science. Automated knowledge extraction using LLMs has the potential to significantly streamline the literature review process, enabling researchers to quickly identify relevant papers and extract key insights. By automating mundane tasks such as data extraction and summarization, LLMs empower researchers to focus on higher-level cognitive activities, such as hypothesis generation, theory development, and data analysis.
- Furthermore, the successful application of LLMs in knowledge extraction from scientific literature opens up new avenues for research in NLP and information retrieval.
- Future work may involve exploring advanced techniques, developing interactive interfaces for researchers to interact with LLM-based systems, and investigating the ethical implications of automated literature analysis, advance LLMs & advance evaluation techniques .

Conclusion:

In this paper, we focused on reviewing existing studies that utilize LLM's for knowledge extraction from scientific papers. We first extracted the text data from all the scientific papers and later manually annotated the text data and evaluated the information extracted using LLM models like Chat GPT and Claudie by Cosine Similarity and Bert Score. Potentially, The performance of Chat GPT 3 is slightly better when compared to Claudie AI. This approach implies that without NLP training we can utilize existing models such as Chat GPT 3 or Claudie to extract knowledge for highly scientific papers. We hope this kind of approach enables us to rapidly extract related information for the advancement of scientific knowledge. This approach offers a promising avenue for rapidly accessing and synthesizing valuable insights from a vast array of scientific literature. By harnessing the capabilities of LLMs, researchers and practitioners can streamline the process of knowledge extraction, facilitating the dissemination and advancement of scientific knowledge. This not only accelerates the pace of research but also fosters collaboration and innovation across various scientific disciplines.

References:

1. Almaqbali, Iqtibas Salim Hilal, et al. "Web scrapping: Data extraction from websites." Journal of Student Research (2019).
2. Abodayeh, Ayat, et al. "Web Scraping for Data Analytics: A BeautifulSoup Implementation." 2023 Sixth International Conference of Women in Data Science at Prince Sultan University (WiDS PSU). IEEE, 2023.
3. Gartlehner, Gerald, et al. "Data extraction for evidence synthesis using a large language model: A proof-of-concept study." Research Synthesis Methods (2024).
4. Borji, Ali, and Mehrdad Mohammadian. "Battle of the Wordsmiths: Comparing ChatGPT, GPT-4, Claude, and Bard." GPT-4, Claude, and Bard (June 12, 2023) (2023).
5. Lozić, E., & Štular, B. (2023). ChatGPT v Bard v Bing v Claude 2 v Aria v human-expert. How good are AI chatbots at scientific writing?(ver. 23Q3). arXiv preprint arXiv:2309.08636.
6. Zhang, Tianyi, et al. "Bertscore: Evaluating text generation with bert." arXiv preprint arXiv:1904.09675 (2019).
7. Gunawan, Dani, C. A. Sembiring, and Mohammad Andri Budiman. "The implementation of cosine similarity to calculate text relevance between two documents." Journal of physics: conference series. Vol. 978. IOP Publishing, 2018.
8. Sage, Clément, et al. "Data-efficient information extraction from documents with pre-trained language models." Document Analysis and Recognition–ICDAR 2021 Workshops: Lausanne, Switzerland, September 5–10, 2021, Proceedings, Part II 16. Springer International Publishing, 2021.

References:

1. Rahutomo, F., Kitasuka, T., & Aritsugi, M. (2012, October). Semantic cosine similarity. In *The 7th international student conference on advanced science and technology ICAST* (Vol. 4, No. 1, p. 1). South Korea: University of Seoul.
2. An article in 2024 [1] - This paper, they propose SciEval, a comprehensive and multi-disciplinary evaluation benchmark toAn article in 2023 [2] - This paper presents a comprehensive survey of ChatGPT-related (GPT-3.5 and GPT-4) research, state-of-the-art large language models (LLM) from the GPT series, and their prospective applications across diverse domains.
3. An article in 2024 [3] - Large language models (LLMs) improve evidence synthesis data extraction efficiency and accuracy. LLM (Claude 2) data extraction from published papers was compared to systematic review human data extraction in this proof-of-concept project. Our convenience sample included 10 English-language; open-access randomized controlled trial manuscripts from a systematic review.
4. Sun, L., Han, Y., Zhao, Z., Ma, D., Shen, Z., Chen, B., ... & Yu, K. (2024, March). Scieval: A multi-level large language model evaluation benchmark for scientific research. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 38, No. 17, pp. 19053-19061).
5. Liu, Y., Han, T., Ma, S., Zhang, J., Yang, Y., Tian, J., ... & Ge, B. (2023). Summary of chatgpt-related research and perspective towards the future of large language models. *Meta-Radiology*, 100017.
6. Gartlehner, G., Kahwati, L., Hilscher, R., Thomas, I., Kugley, S., Crotty, K., ... & Chew, R. (2024). Data extraction for evidence synthesis using a large language model: A proof-of-concept study. *Research Synthesis Methods*.

Contributions:

Harshavardhan	Tharun Sai	Krinalben	Sreekar
Data extraction & preprocessing using web scraping.	Drafting Project proposal	Drafting Project proposal	Drafting Project proposal, flow charts.
Methodology: model selection, insights extraction using GPT-3, Claude including coding.	Data preprocessing and Tried various Model selection insights for extraction like Gemini , LLAMA 2.	Manual knowledge extraction from papers for evaluation	Manual knowledge extraction from papers for evaluation
Tried using Llama.	Flow graphs in Presentation and proposal .	preprocessing. Until headers.	Data extraction
Evaluation metrics selection and results.	Manual knowledge extraction from papers for evaluation	Participated in online meetings.	Participated in online, offline meetings.
Participated in online, offline meetings. Drafting Project presentation. Related work and references.	Participated in online, offline meetings.	Drafting Project presentation	Contributed for selecting Related work.
Manual insights Information about enhances and Future work. Going to work on Final report.	Drafting Project presentation. Related work and references.	Going to work on Final report	Going to work on Final report
Coding from scratch to final. And git push.	Going to work on Final report	Contributed to selecting Related work and references.	

Source Link:

Git link:
Every file is uploaded in IPYNB and Datasets in Json and included a zip file.

https://github.com/HarshaSolingaram/INFO_5731_Project_Group1

Thank You!!!

Open to questions…