LitLLM: A Toolkit for Scientific Literature Review

Shubham Agarwal[1,2,3], **Issam H. Laradji**[1,4], **Laurent Charlin**[2,3,5], **Christopher Pal**[1,2,5]
[1]ServiceNow Research, [2]Mila - Quebec AI Institute, [3]HEC Montreal, Canada
[4]UBC, Vancouver, Canada, [5]Canada CIFAR AI Chair
`Correspondence: shubham.agarwal@mila.quebec`

Abstract

Conducting literature reviews for scientific papers is essential for understanding research, its limitations, and building on existing work. It is a tedious task which makes an automatic literature review generator appealing. Unfortunately, many existing works that generate such reviews using Large Language Models (LLMs) have significant limitations. They tend to hallucinate—generate non-factual information—and ignore the latest research they have not been trained on. To address these limitations, we propose a toolkit that operates on Retrieval Augmented Generation (RAG) principles, specialized prompting and instructing techniques with the help of LLMs. Our system first initiates a web search to retrieve relevant papers by summarizing user-provided abstracts into keywords using an off-the-shelf LLM. Authors can enhance the search by supplementing it with relevant papers or keywords, contributing to a tailored retrieval process. Second, the system re-ranks the retrieved papers based on the user-provided abstract. Finally, the related work section is generated based on the re-ranked results and the abstract. There is a substantial reduction in time and effort for literature review compared to traditional methods, establishing our toolkit as an efficient alternative. Our open-source toolkit is accessible at `https://github.com/shubhamagarwal92/LitLLM` and Huggingface space (`https://huggingface.co/spaces/shubhamagarwal92/LitLLM`) with the video demo at `https://youtu.be/E2ggOZBAFw0`

1 Introduction

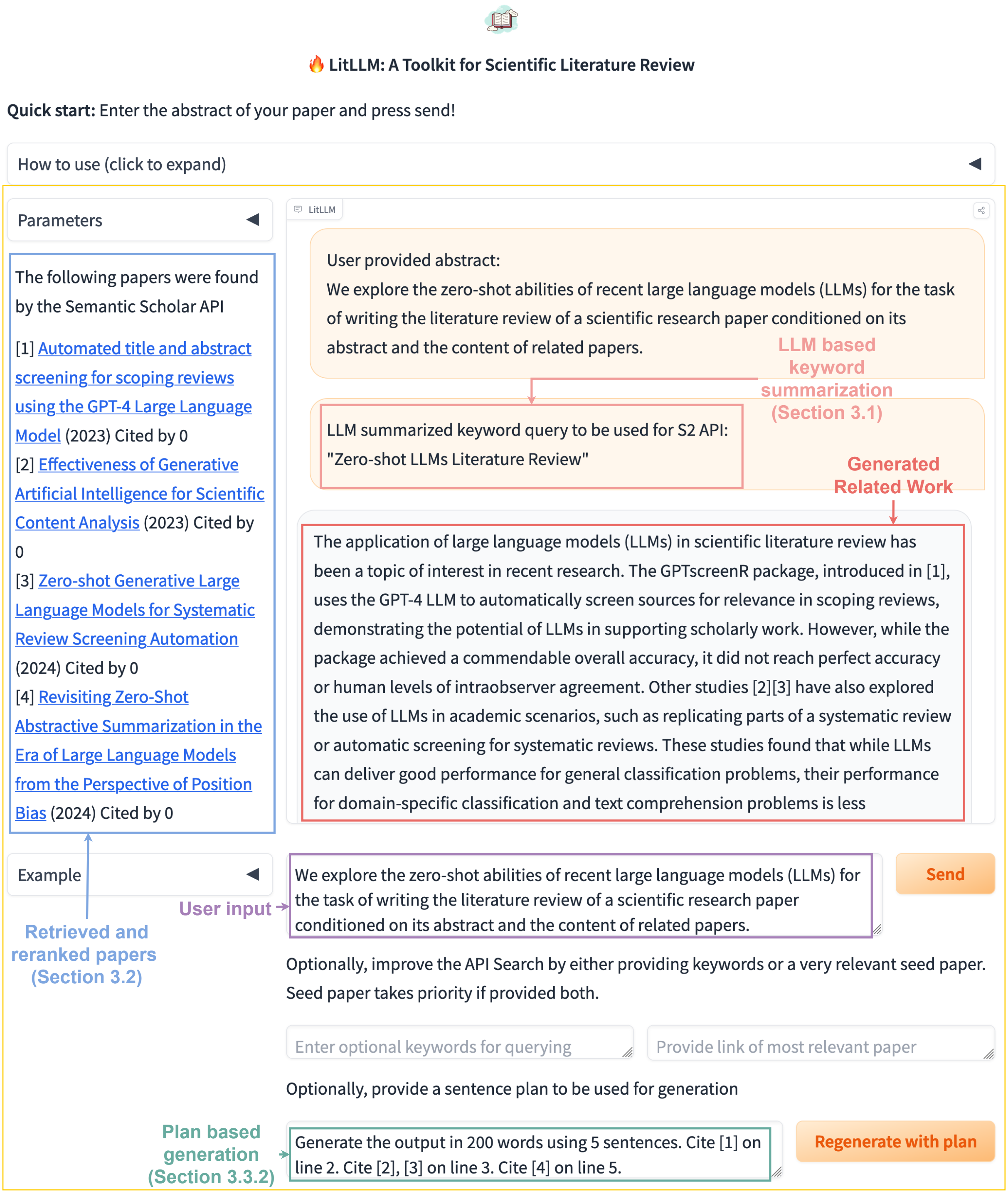
Scientists have long used NLP systems like search engines to find and retrieve relevant papers. Scholarly engines, including Google Scholar, Microsoft Academic Graph, and Semantic Scholar, provide additional tools and structure to help researchers further. Following recent advances in large language models (LLMs), a new set of systems provides even more advanced features. For example, Explainpaper[1] helps explain the contents of papers, and Writefull[2] helps with several writing tasks, including abstract and title generation. There are, of course, many other tasks where similar technologies could be helpful.

Systems that help researchers with literature reviews hold promising prospects. The literature review is a difficult task that can be decomposed into several sub-tasks, including retrieving relevant papers and generating a related works section that contextualizes the proposed work compared to the existing literature. It is also a task where factual correctness is essential. In that sense, it is a challenging task for current LLMs, which are known to hallucinate. Overall, creating tools to help researchers more rapidly identify, summarize and contextualize relevant prior work could significantly help the research community.

Recent works explore the task of literature review in parts or in full. For example, Lu et al. (2020) proposes generating the related works section of a paper using its abstract and a list of (relevant) references. Researchers also look at the whole task and build systems using LLMs like ChatGPT for literature review (Haman and Školník, 2023; Huang and Tan, 2023). While these LLMs tend to generate high-quality text, they are prone to hallucinations (Athaluri et al., 2023). For example, the Galactica system was developed to reason about scientific knowledge (Taylor et al., 2022). While it outperforms contemporary models on various scientific tasks, it generates made-up content like inaccurate citations and imaginary papers.[3]

As a step forward, we explore retrieval-augmented-generation (RAG) to improve factual correctness (Lewis et al., 2020). The idea is to use the retrieval mechanism to obtain a relevant list of

[1] `https://www.explainpaper.com/`
[2] `https://x.writefull.com/`
[3] see e.g., What Meta Learned from Galactica



🔥 LitLLM: A Toolkit for Scientific Literature Review

Quick start: Enter the abstract of your paper and press send!

The following papers were found by the Semantic Scholar API

[1] Automated title and abstract screening for scoping reviews using the GPT-4 Large Language Model (2023) Cited by 0
[2] Effectiveness of Generative Artificial Intelligence for Scientific Content Analysis (2023) Cited by 0
[3] Zero-shot Generative Large Language Models for Systematic Review Screening Automation (2024) Cited by 0
[4] Revisiting Zero-Shot Abstractive Summarization in the Era of Large Language Models from the Perspective of Position Bias (2024) Cited by 0

User provided abstract:
We explore the zero-shot abilities of recent large language models (LLMs) for the task of writing the literature review of a scientific research paper conditioned on its abstract and the content of related papers.

LLM summarized keyword query to be used for S2 API:
"Zero-shot LLMs Literature Review"

The application of large language models (LLMs) in scientific literature review has been a topic of interest in recent research. The GPTscreenR package, introduced in [1], uses the GPT-4 LLM to automatically screen sources for relevance in scoping reviews, demonstrating the potential of LLMs in supporting scholarly work. However, while the package achieved a commendable overall accuracy, it did not reach perfect accuracy or human levels of intraobserver agreement. Other studies [2][3] have also explored the use of LLMs in academic scenarios, such as replicating parts of a systematic review or automatic screening for systematic reviews. These studies found that while LLMs can deliver good performance for general classification problems, their performance for domain-specific classification and text comprehension problems is less

We explore the zero-shot abilities of recent large language models (LLMs) for the task of writing the literature review of a scientific research paper conditioned on its abstract and the content of related papers.

Optionally, improve the API Search by either providing keywords or a very relevant seed paper. Seed paper takes priority if provided both.

Optionally, provide a sentence plan to be used for generation

Generate the output in 200 words using 5 sentences. Cite [1] on line 2. Cite [2], [3] on line 3. Cite [4] on line 5.

How to use (click to expand)
Parameters
LitLLM
Example
Send
Enter optional keywords for querying
Provide link of most relevant paper
Regenerate with plan
LLM based keyword summarization (Section 3.1)
Generated Related Work
User input
Retrieved and reranked papers (Section 3.2)
Plan based generation (Section 3.3.2)

Figure 1: LitLLM on Huggingface Space. Our system works on the Retrieval Augmented Generation (RAG) principle to generate the literature review grounded in retrieved relevant papers. User needs to provide the abstract in the textbox (in purple) and press send to get the generated related work (in red). First, the abstract is summarized into keywords (Section 3.1), which are used to query a search engine. Retrieved results are re-ranked (in blue) using the Paper Re-Ranking module (Section 3.2), which is then used as context to generate the related work (Section 3.3). Users could also provide a sentence plan (in green) according to their preference to generate a concise, readily usable literature review (See Section 3.3.2).

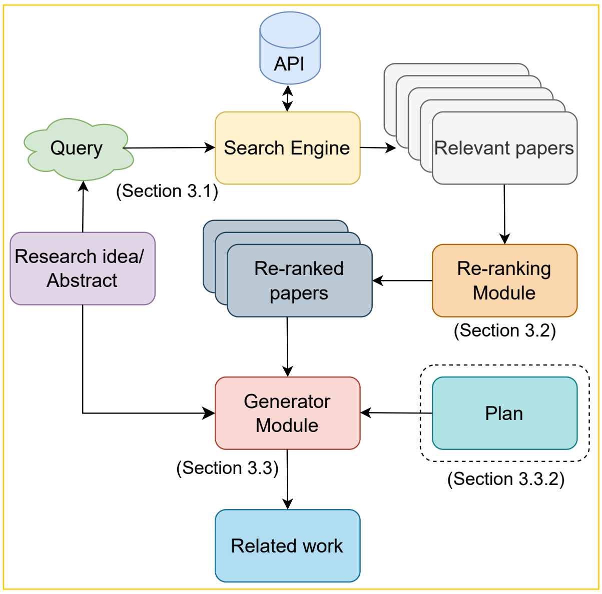


Figure 2: Schematic diagram of the modular pipeline used in our system. In the default setup, we summarize the research abstract into a keyword query, which is used to retrieve relevant papers from an academic search engine. We use an LLM-based reranker to select the most relevant paper relative to the provided abstract. Based on the re-ranked results and the user-provided summary of their work, we use an LLM-based generative model to generate the literature review, optionally controlled by a sentence plan.

existing papers to be cited which provides relevant contextual knowledge for LLM based generation.

LitLLM is an interactive tool to help scientists write the literature review or related work section of a scientific paper starting from a user-provided abstract (see Figure 1). The specific objectives of this work are to create a system to help users navigate through research papers and write a literature review for a given paper or project. Our main contributions are:

- We provide a system based on a modular pipeline that conducts a literature review based on a user-proposed abstract.
- We use Retrieval Augmented Generation (RAG) techniques to condition the generated related work on factual content and avoid hallucinations using multiple search techniques.
- We incorporate sentence-based planning to promote controllable generation.

2 Related Work

LLMs have demonstrated significant capabilities in storing factual knowledge and achieving state-of-the-art results when fine-tuned on downstream Natural Language Processing (NLP) tasks (Lewis et al., 2020).

However, they also face challenges such as hallucination, outdated knowledge, and non-transparent, untraceable reasoning processes (Huang et al., 2023; Gao et al., 2023; Li et al., 2024). These limitations have motivated the development of RAG (Retrieval Augmented Generation), which incorporates knowledge from external databases to enhance the accuracy and credibility of the models, particularly for knowledge-intensive tasks (Gao et al., 2023). RAG has emerged as a promising solution to the challenges faced by LLMs. It synergistically merges LLMs' intrinsic knowledge with the vast, dynamic repositories of external databases (Gao et al., 2023). This approach allows for continuous knowledge updates and integration of domain-specific information in an attempt to limit the effect of outdated knowledge. The proposed work builds upon the advancements around RAG to provide a more efficient solution for academic writing.

On the other hand, there has been a notable emphasis on utilizing Large Language Models (LLMs) for tasks related to information retrieval and ranking (Zhu et al., 2023). The work by Sun et al. (2023) leverages generative LLMs such as ChatGPT and GPT-4 for relevance ranking in information retrieval, demonstrating that these models can deliver competitive results to state-of-the-art supervised methods. Pradeep et al. (2023b,a) introduce different open-source LLM for listwise zero-shot reranking, further motivating the proposed approach of using LLMs for reranking in our work.

The exploration of large language models (LLMs) and their zero-shot abilities has been a significant focus in recent research. For instance, one study investigated using LLMs in recommender systems, demonstrating their promising zero-shot ranking abilities, although they struggled with the order of historical interactions and position bias (Hou et al., 2023). Another study improved the zero-shot learning abilities of LLMs through instruction tuning, which led to substantial improvements in performance on unseen tasks (Wei et al., 2021). A similar approach was taken to enhance the zero-shot reasoning abilities of LLMs, with the introduction of an autonomous agent to instruct the reasoning process, resulting in significant performance boosts (Crispino et al., 2023). The application of LLMs has also been explored in the context of natural language generation (NLG) assessment, with comparative assessment found to be superior to prompt scoring (Liusie et al., 2023). In the domain of Open-Domain Question Answering (ODQA), a Self-Prompting framework was proposed to utilize the massive knowledge stored in LLMs, leading to significant improvements over previous methods (Li et al., 2022). Prompt engineering has been identified as a key technique for enhancing the abilities of LLMs, with various strategies being explored (Shi et al., 2023).[4]

3 Pipeline

Figure 2 provides an overview of the pipeline. The user provides a draft of the abstract or a research idea. We use LLM to first summarize the abstract in keywords that can be used as a query for search engines. Optionally, the users could provide relevant keywords to improve search results. This query is passed to the search engine, which retrieves relevant papers with the corresponding information, such as abstracts and open-access PDF URLs. These retrieved abstracts with the original query abstract are used as input to the other LLM Re-ranker, which provides a listwise ranking of the papers based on the relevance to the query abstract. These re-ranked abstracts with the original query are finally passed to the LLM generator, which generates the related work section of the paper. Recently, Agarwal et al. (2024) showed that prompting the LLMs with the sentence plans results in reduced hallucinations in the generation outputs. These plans contain information about the number of sentences and the citation description on each line, providing control to meet author preferences. We include this sentence-based planning in the LLM generator as part of this system. In the following, we provide more details about each of the modules.

3.1 Paper Retrieval Module

In our toolkit, we retrieve relevant papers using the Semantic Scholar API. Other platforms could be used, but the S2 Platform is well-adapted to this use case. It is a large-scale academic corpus comprising 200M+ metadata records across multiple research areas, providing information about papers' metadata, authors, paper embedding, etc. The Recommendations API also provides relevant papers similar to any seed paper. Figure 3 shows our system's different strategies. We describe these three settings that we use to search for references:

- User provides an abstract or a research idea (roughly the length of the abstract). We prompt an LLM (see Figure 4) to summarize this abstract in keywords which can be used as a search query with most APIs.
- Users can optionally also provide keywords that can improve search results. This is similar (in spirit) to how researchers search for related work with a search engine. This is particularly useful in interdisciplinary research, and authors would like to include the latest research from a particular domain, which could not be captured much in the abstract.
- Lastly, any seed paper the user finds relevant enough to their idea could be used with the Recommendations API from search engines to provide other closely related papers.

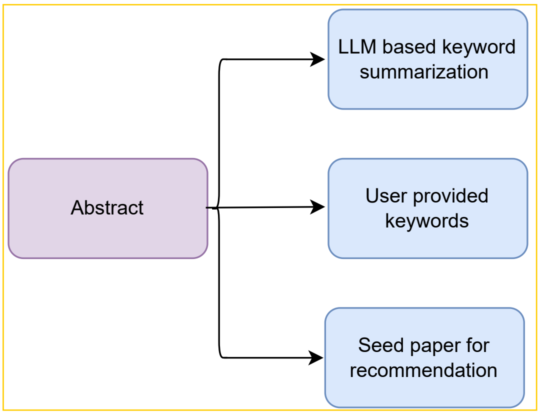

Abstract
LLM based keyword summarization
User provided keywords
Seed paper for recommendation

Figure 3: Different retrieval strategies as discussed in Section 3.1

Keyword summarization prompt

You are a helpful research assistant who is helping with literature review of a research idea. You will be provided with an abstract of a scientific document. Your task is to summarize the abstract in max 5 keywords to search for related papers using API of academic search engine.

```Abstract: {abstract}```

Figure 4: Prompt used to summarize the research idea by LLM to search an academic engine

3.2 Paper Re-Ranking Module

Recent efforts have explored the application of proprietary LLMs for ranking (Sun et al., 2023; Ma et al., 2023) as well as open-source models like

[4] This paragraph was generated using our platform with some minor modifications based on a slightly different version of our abstract.

(Pradeep et al., 2023a,b). These approaches provide a combined list of passages directly as input to the model and retrieve the re-ordered ranking list (Zhang et al., 2023). Typically, a retriever first filters top-k potential candidates, which are then re-ranked by an LLM to provide the final output list. In our work, we use the instructional *permutation generation* approach (Sun et al., 2023) where the model is prompted to generate a permutation of the different papers in descending order based on the relevance to the user-provided abstract, thus producing an ordered list of preferences against providing intermediate scores. Figure 5 showcases the prompt we used for LLM-based re-ranking.

Ranking prompt

~~~
You are a helpful research assistant who is helping with literature review of a research idea. You will be provided with an abstract or an idea of a scientific document and abstracts of some other relevant papers. Your task is to rank the papers based on the relevance to the query abstract. Provide only the ranks as [] > [] > []. Do not output anything apart from the ranks.
```Abstract: {abstract}
References:
[1]: {text}
[2]: {text}
...
```
~~~

Figure 5: Ranking prompt based on the permutation generation method

3.3 Summary Generation Module

We explore two strategies for generation: (1) Zero-shot generation and (2) Plan-based generation, which relies on sentence plans for controllable generation, described in the following

3.3.1 Zero-shot generation

While LLMs can potentially search and generate relevant papers from their parametric memory and trained data, they are prone to hallucinating and generating non-factual content. Retrieval augmented generation, first introduced in Parvez et al. (2021) for knowledge tasks, addresses this by augmenting the generation model with an information retrieval module. The RAG principles have been subsequently used for dialogue generation in task-oriented settings (Thulke et al., 2021), code generation (Liu et al., 2020; Parvez et al., 2021) and product review generation (Kim et al., 2020). RAG drastically reduces hallucinations in the generated output (Gao et al., 2023; Tonmoy et al., 2024).

RAG prompt

~~~
You will be provided with an abstract of a scientific document and other references papers in triple quotes. Your task is to write the related work section of the document using only the provided abstracts and other references papers. Please write the related work section creating a cohesive storyline by doing a critical analysis of prior work comparing the strengths and weaknesses while also motivating the proposed approach. You should cite the other related documents as [#] whenever you are referring it in the related work. Do not write it as Reference #. Do not cite abstract. Do not include any extra notes or newline characters at the end. Do not copy the abstracts of reference papers directly but compare and contrast to the main work concisely. Do not provide the output in bullet points. Do not provide references at the end. Please cite all the provided reference papers.
```Abstract: {abstract}
References:
[1]: {text}
[2]: {text}
```
~~~

Figure 6: Prompt for Retrieval Augmented Generation

Our work builds upon the principles of RAG, where we retrieve the relevant papers based on the query and augment them as context for generating the literature review. This also allows the system to be grounded in the retrieved information and be updated with the latest research where the training data limits the parametric knowledge of the LLM. Figure 6 shows our system's prompt for effective Retrieval Augmented Generation (RAG).

3.3.2 Plan based generation

To get the best results from LLM, recent research shifts focus on designing better prompts (Prompt Engineering) including 0-shot chain-of-thought prompting (Kojima et al., 2022; Zhou et al., 2022), few-shot prompting (Brown et al., 2020) techniques, few-shot Chain-of-thought prompting (Wei et al., 2022) and in-context prompting (Li and Liang, 2021; Qin and Eisner, 2021). However, the longer context of our problem statement (query paper and multiple relevant papers) hinders the application of these techniques for response generation.

We utilized sentence plan-based prompting techniques drawing upon insights from the literature of traditional modular Natural Language Generation (NLG) pipelines with intermediary steps of sentence planning and surface realization (Reiter and Dale, 1997; Stent et al., 2004). These plans provide a sentence structure of the expected output, which efficiently guides the LLM in generating the literature review in a controllable fashion as demonstrated in concurrent work (Agarwal et al., 2024). Figure 7 (in Appendix) shows the prompt for plan-based generation with an example template as:

```
Please generate {num_sentences} sentences in
{num_words} words. Cite {cite_x} at line {line_x}.
Cite {cite_y} at line {line_y}.
```

4 Implementation Details

We build our system using Gradio (Abid et al., 2019), which provides a nice interface to quickly and efficiently build system demos. Our user interface is also available at HuggingFace Space[5]. We query the Semantic Scholar API available through the Semantic Scholar Open Data Platform (Lo et al., 2020; Kinney et al., 2023) to search for the relevant papers. Specifically, we use the Academic Graph[6] and Recommendations[7] API endpoint. In this work, we use OpenAI API[8] to generate results for LLM using GPT-3.5-turbo and GPT-4 model. At the same time, our modular pipeline allows using any LLM (proprietary or open-sourced) for different components. We also allow the end-user to sort the retrieved papers by relevance (default S2 results), citation count, or year.

5 User Experience

As a preliminary study, we provided access to our user interface to 5 different researchers who worked through the demo to write literature reviews and validate the system's efficacy. We also provide an example in the demo with an abstract for a quick start. Particularly, the users found the 0-shot generation to be more informative about the literature in general while the plan-based generation to be more accessible and tailored for their research paper, as also evident in our demo video.[9]. Table 1 (in Appendix) shows the output-related work for a recent paper (Li et al., 2023) that was randomly chosen with a number of cited papers as 4. Our system generated an informative query *Multimodal Research: Image-Text Model Interaction* and retrieved relevant papers where the top recommended paper was also cited in the original paper. While zero-shot generation provides valuable insights into existing literature, plan-based generation produces a more succinct and readily usable literature review.

6 Conclusion and Future Work

In this work, we introduce and describe LitLLM, a system which can generate literature reviews in a few clicks from an abstract using off-the-shelf LLMs. This LLM-powered toolkit relies on the RAG with a re-ranking strategy to generate a literature review with attribution. Our auxiliary tool allows researchers to actively search for related work based on a preliminary research idea, research proposal or even a full abstract. We present a modular pipeline that can be easily adapted to include the next generation of LLMs and other domains, such as news, by changing the source of retrieval information.

Given the growing impact of different LLM-based writing assistants, we are optimistic that our system may aid researchers in searching relevant papers and improve the quality of automatically generated related work sections of a paper. While our system shows promise as a helpful research assistant, we believe that their usage should be disclosed to the readers, and authors should also observe caution in eliminating any possible hallucinations.

In the future, we would also like to explore academic search through multiple APIs, such as Google Scholar. This work only considered abstracts of the query paper and the retrieved papers, which creates a bottleneck in effective literature review generation. With the advent of longer context LLMs, we envision our system ingesting the whole paper (potentially leveraging an efficient LLM-based PDF parser) to provide a more relevant background of the related research. We consider our approach as an initial step for building intelligent research assistants which could help academicians through an interactive setting (Dwivedi-Yu et al., 2022).

[5] https://huggingface.co/spaces/shubhamagarwal92/LitLLM

[6] https://api.semanticscholar.org/api-docs/graph

[7] https://api.semanticscholar.org/api-docs/recommendations

[8] https://platform.openai.com/docs/guides/gpt

[9] https://youtu.be/E2ggOZBAFw0

References

Abubakar Abid, Ali Abdalla, Ali Abid, Dawood Khan, Abdulrahman Alfozan, and James Zou. 2019. Gradio: Hassle-free sharing and testing of ml models in the wild. *arXiv preprint arXiv:1906.02569*.

Shubham Agarwal, Issam Laradji, Laurent Charlin, and Christopher Pal. 2024. LLMs for Literature Review generation: Are we there yet? *Under submission*.

Sai Anirudh Athaluri, Sandeep Varma Manthena, V S R Krishna Manoj Kesapragada, Vineel Yarlagadda, Tirth Dave, and Rama Tulasi Siri Duddumpudi. 2023. Exploring the boundaries of reality: Investigating the phenomenon of artificial intelligence hallucination in scientific writing through chatgpt references. *Cureus*, 15.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners.

Nicholas Crispino, Kyle Montgomery, Fankun Zeng, Dawn Song, and Chenguang Wang. 2023. Agent instructs large language models to be general zero-shot reasoners. *ArXiv*, abs/2310.03710.

Jane Dwivedi-Yu, Timo Schick, Zhengbao Jiang, Maria Lomeli, Patrick Lewis, Gautier Izacard, Edouard Grave, Sebastian Riedel, and Fabio Petroni. 2022. Editeval: An instruction-based benchmark for text improvements. *arXiv preprint arXiv:2209.13331*.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, and Haofen Wang. 2023. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*.

Michael Haman and Milan Školník. 2023. Using chat-gpt to conduct a literature review. *Accountability in Research*, pages 1–3.

Yupeng Hou, Junjie Zhang, Zihan Lin, Hongyu Lu, Ruobing Xie, Julian McAuley, and Wayne Xin Zhao. 2023. Large language models are zero-shot rankers for recommender systems. *ArXiv*, abs/2305.08845.

Jingshan Huang and Ming Tan. 2023. The role of chat-gpt in scientific communication: writing better scientific review articles. *American Journal of Cancer Research*, 13(4):1148.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, et al. 2023. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *arXiv preprint arXiv:2311.05232*.

Jihyeok Kim, Seungtaek Choi, Reinald Kim Amplayo, and Seung-won Hwang. 2020. Retrieval-augmented controllable review generation. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 2284–2295.

Rodney Kinney, Chloe Anastasiades, Russell Authur, Iz Beltagy, Jonathan Bragg, Alexandra Buraczynski, Isabel Cachola, Stefan Candra, Yoganand Chandrasekhar, Arman Cohan, et al. 2023. The semantic scholar open data platform. *arXiv preprint arXiv:2301.10140*.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474.

Hang Li, Jindong Gu, Rajat Koner, Sahand Sharifzadeh, and Volker Tresp. 2023. Do dall-e and flamingo understand each other? In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1999–2010.

Junlong Li, Zhuosheng Zhang, and Hai Zhao. 2022. Self-prompting large language models for zero-shot open-domain qa.

Junyi Li, Jie Chen, Ruiyang Ren, Xiaoxue Cheng, Wayne Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. 2024. The dawn after the dark: An empirical study on factuality hallucination in large language models. *arXiv preprint arXiv:2401.03205*.

Xiang Lisa Li and Percy Liang. 2021. Prefix-tuning: Optimizing continuous prompts for generation. *arXiv preprint arXiv:2101.00190*.

Shangqing Liu, Yu Chen, Xiaofei Xie, Jingkai Siow, and Yang Liu. 2020. Retrieval-augmented generation for code summarization via hybrid gnn. *arXiv preprint arXiv:2006.05405*.

Adian Liusie, Potsawee Manakul, and Mark John Francis Gales. 2023. Llm comparative assessment: Zero-shot nlg evaluation through pairwise comparisons using large language models.

Kyle Lo, Lucy Lu Wang, Mark Neumann, Rodney Kinney, and Daniel Weld. 2020. S2ORC: The semantic scholar open research corpus. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4969–4983, Online. Association for Computational Linguistics.

Yao Lu, Yue Dong, and Laurent Charlin. 2020. Multi-XScience: A large-scale dataset for extreme multi-document summarization of scientific articles. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8068–8074. Association for Computational Linguistics.

Xueguang Ma, Xinyu Zhang, Ronak Pradeep, and Jimmy Lin. 2023. Zero-shot listwise document reranking with a large language model. *arXiv preprint arXiv:2305.02156*.

Md Rizwan Parvez, Wasi Uddin Ahmad, Saikat Chakraborty, Baishakhi Ray, and Kai-Wei Chang. 2021. Retrieval augmented code generation and summarization. *arXiv preprint arXiv:2108.11601*.

Ronak Pradeep, Sahel Sharifymoghaddam, and Jimmy Lin. 2023a. Rankvicuna: Zero-shot listwise document reranking with open-source large language models. *arXiv preprint arXiv:2309.15088*.

Ronak Pradeep, Sahel Sharifymoghaddam, and Jimmy Lin. 2023b. Rankzephyr: Effective and robust zero-shot listwise reranking is a breeze! *arXiv preprint arXiv:2312.02724*.

Guanghui Qin and Jason Eisner. 2021. Learning how to ask: Querying LMs with mixtures of soft prompts. *arXiv preprint arXiv:2104.06599*.

Leigang Qu, Meng Liu, Jianlong Wu, Zan Gao, and Liqiang Nie. 2021. Dynamic modality interaction modeling for image-text retrieval. *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*.

Ehud Reiter and Robert Dale. 1997. Building applied natural language generation systems. *Natural Language Engineering*, 3(1):57–87.

Fobo Shi, Peijun Qing, D. Yang, Nan Wang, Youbo Lei, H. Lu, and Xiaodong Lin. 2023. Prompt space optimizing few-shot reasoning success with large language models. *ArXiv*, abs/2306.03799.

Krishna Srinivasan, Karthik Raman, Jiecao Chen, Michael Bendersky, and Marc Najork. 2021. Wit: Wikipedia-based image text dataset for multimodal multilingual machine learning. *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*.

Amanda Stent, Rashmi Prasad, and Marilyn Walker. 2004. Trainable sentence planning for complex information presentations in spoken dialog systems. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL-04)*, pages 79–86, Barcelona, Spain.

Weiwei Sun, Lingyong Yan, Xinyu Ma, Pengjie Ren, Dawei Yin, and Zhaochun Ren. 2023. Is chatgpt good at search? investigating large language models as re-ranking agent. *arXiv preprint arXiv:2304.09542*.

Ross Taylor, Marcin Kardas, Guillem Cucurull, Thomas Scialom, Anthony Hartshorn, Elvis Saravia, Andrew Poulton, Viktor Kerkez, and Robert Stojnic. 2022. Galactica: A large language model for science. *arXiv preprint arXiv:2211.09085*.

David Thulke, Nico Daheim, Christian Dugast, and Hermann Ney. 2021. Efficient retrieval augmented generation from unstructured knowledge for task-oriented dialog. *arXiv preprint arXiv:2102.04643*.

SM Tonmoy, SM Zaman, Vinija Jain, Anku Rani, Vipula Rawte, Aman Chadha, and Amitava Das. 2024. A comprehensive survey of hallucination mitigation techniques in large language models. *arXiv preprint arXiv:2401.01313*.

Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V. Le. 2021. Finetuned language models are zero-shot learners. *ArXiv*, abs/2109.01652.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed H Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*.

Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. 2022. Coca: Contrastive captioners are image-text foundation models. *Trans. Mach. Learn. Res.*, 2022.

Xinyu Zhang, Sebastian Hofstätter, Patrick Lewis, Raphael Tang, and Jimmy Lin. 2023. Rank-without-gpt: Building gpt-independent listwise rerankers on open-source large language models. *arXiv preprint arXiv:2312.02969*.

Zijia Zhao, Longteng Guo, Xingjian He, Shuai Shao, Zehuan Yuan, and Jing Liu. 2022. Mamo: Fine-grained vision-language representations learning with masked multimodal modeling. *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*.

Yongchao Zhou, Andrei Ioan Muresanu, Ziwen Han, Keiran Paster, Silviu Pitis, Harris Chan, and Jimmy Ba. 2022. Large language models are human-level prompt engineers. *arXiv preprint arXiv:2211.01910*.

Yutao Zhu, Huaying Yuan, Shuting Wang, Jiongnan Liu, Wenhan Liu, Chenlong Deng, Zhicheng Dou, and Ji-Rong Wen. 2023. Large language models for information retrieval: A survey. *arXiv preprint arXiv:2308.07107*.

Appendix

Plan based generation prompt

You are a helpful research assistant who is helping with literature review of a research idea. You will be provided with an abstract of a scientific document and other references papers in triple quotes. Your task is to write the related work section of the document using only the provided abstracts and other references papers. Please write the related work section creating a cohesive storyline by doing a critical analysis of prior work comparing the strengths and weaknesses while also motivating the proposed approach. You are also provided a sentence plan mentioning the total number of lines and the citations to refer in different lines. You should cite all the other related documents as [#] whenever you are referring it in the related work. Do not cite abstract. Do not include any extra notes or newline characters at the end. Do not copy the abstracts of reference papers directly but compare and contrast to the main work concisely. Do not provide the output in bullet points. Do not provide references at the end. Please cite all the provided reference papers. Please follow the plan when generating sentences, especially the number of lines and citations to generate.

```Abstract: {abstract}

References:

[1]: {text}

[2]: {text}

Plan: {plan}

```

Figure 7: Prompt for sentence plan-based generation

In the following, we provide snippets of code to retrieve results from the Semantic Scholar API for both recommendation and query-based search:

```
# QUERY BASED SEARCH

def query_search_s2(query: str,
                    num_papers_api: int,
                    fields: str):
    rsp = requests.get("https://api.
    semanticscholar.org/graph/v1/paper/
    search",
    headers={"X-API-KEY": S2_API_KEY},
    params={"query": query, "limit":
    num_papers_api, "fields": fields})
    rsp.raise_for_status()
    results = rsp.json()
    # Total papers found
    total = results["total"]
    papers = results["data"]
    return papers
```

```
# PAPER SEARCH

def get_paper_data(paper_url: str,
                   fields: str):
    """
    Retrieves data of one paper based on
     URL
    """
    rsp = requests.get(f"https://api.
    semanticscholar.org/graph/v1/paper/
    URL:{paper_url}",
    headers={"X-API-KEY": S2_API_KEY},
    params={"fields": fields})
    results = rsp.json()
    return rewsults
```

```
# RECOMMENDATION API

def get_recommendations_from_s2(
    arxiv_id: str, num_papers_api: int,
    fields: str):
    """
    Get recommendations from S2 API
    """
    query_id = f"ArXiv:{arxiv_id}"
    rsp = requests.post("https://api.
    semanticscholar.org/recommendations/
    v1/papers/",
    json={
        "positivePaperIds": [query_id]},
    params={"fields": fields,
            "limit": num_papers_api})
    results = rsp.json()
    papers = results["recommendedPapers"
    ]
    return papers
```

Abstract for paper: ***Do DALL-E and Flamingo Understand Each Other?*** ***(Li et al., 2023)***

The field of multimodal research focusing on the comprehension and creation of both images and text has witnessed significant strides. This progress is exemplified by the emergence of sophisticated models dedicated to image captioning at scale, such as the notable Flamingo model and text-to-image generative models, with DALL-E serving as a prominent example. An interesting question worth exploring in this domain is whether Flamingo and DALL-E understand each other. To study this question, we propose a reconstruction task where Flamingo generates a description for a given image and DALL-E uses this description as input to synthesize a new image. We argue that these models understand each other if the generated image is similar to the given image. Specifically, we study the relationship between the quality of the image reconstruction and that of the text generation. We find that an optimal description of an image is one that gives rise to a generated image similar to the original one. The finding motivates us to propose a unified framework to finetune the text-to-image and image-to-text models. Concretely, the reconstruction part forms a regularization loss to guide the tuning of the models. Extensive experiments on multiple datasets with different image captioning and image generation models validate our findings and demonstrate the effectiveness of our proposed unified framework. As DALL-E and Flamingo are not publicly available, we use Stable Diffusion and BLIP in the remaining work. Project website: `https://dalleflamingo.github.io.`

LLM summarized query: Multimodal Research: Image-Text Model Interaction

Retrieved papers: (User Input: 4)

[1] CoCa: Contrastive Captioners are Image-Text Foundation Models (2022) Cited by 702 (Yu et al., 2022)

[2] MAMO: Fine-Grained Vision-Language Representations Learning with Masked Multimodal Modeling (2022) Cited by 0 (Zhao et al., 2022)

[3] Dynamic Modality Interaction Modeling for Image-Text Retrieval (2021) Cited by 88 (Qu et al., 2021)

[4] WIT: Wikipedia-based Image Text Dataset for Multimodal Multilingual Machine Learning (2021) Cited by 185 (Srinivasan et al., 2021)

Zero shot Generation (Section 3.3.1)

The field of multimodal research has seen significant advancements in recent years, with the development of models such as Flamingo and DALL-E that focus on image captioning and text-to-image generation respectively. However, the question of whether these models can understand each other and work in harmony is a topic of interest. In this context, the work of [1] presents the Contrastive Captioner (CoCa), a model that combines contrastive loss and captioning loss to pretrain an image-text encoder-decoder foundation model. This model, while efficient, does not fully address the interaction between image and text modalities at a fine-grained level.

The work of [2] addresses this gap by proposing a jointly masked multimodal modeling method that focuses on fine-grained multimodal representations. This method performs joint masking on image-text input and integrates both implicit and explicit targets for the masked signals to recover. However, this approach, while effective, does not fully address the challenges of intra-modal reasoning and cross-modal alignment that are inherent in image-text retrieval.

The work of [3] presents a novel modality interaction modeling network based on the routing mechanism, which is the first unified and dynamic multimodal interaction framework towards image-text retrieval. This model can dynamically learn different activated paths for different data, providing a more flexible approach to modality interaction. However, the design of interaction patterns in this model still relies heavily on expert experience and empirical feedback, which may limit its applicability in different contexts. The work of [4] introduces the Wikipedia-based Image Text (WIT) Dataset, a large-scale dataset for multimodal, multilingual learning. This dataset, while providing a rich resource for multimodal learning, does not directly address the question of how different models can understand each other and work together.

In light of these previous works, our study proposes a reconstruction task where Flamingo generates a description for a given image and DALL-E uses this description as input to synthesize a new image. We argue that these models understand each other if the generated image is similar to the given image. This approach allows us to study the relationship between the quality of the image reconstruction and that of the text generation, and to propose a unified framework to finetune the text-to-image and image-to-text models. Our extensive experiments validate our findings and demonstrate the effectiveness of our proposed unified framework.

Plan based generation (Section 3.3.2): Generate the output using 5 sentences. Cite [1] on line 2. Cite [2], [3] on line 3. Cite [4] on line 5.

The field of multimodal research has seen significant advancements in the comprehension and creation of both images and text, with models like Flamingo and DALL-E leading the way. In a similar vein, the Contrastive Captioner (CoCa) model presented in [1] pretrains an image-text encoder-decoder foundation model with contrastive loss and captioning loss, achieving state-of-the-art performance on a broad range of downstream tasks. Other works have also focused on improving the fine-grained image-text interaction, with [2] proposing a jointly masked multimodal modeling method and [3] developing a dynamic multimodal interaction framework for image-text retrieval. These models, while effective, often rely on expert experience and empirical feedback, which may limit their flexibility. The introduction of the Wikipedia-based Image Text (WIT) Dataset in [4] provides a large, multilingual dataset for pretraining multimodal models, offering a more diverse set of concepts and real-world entities for model training. This diversity and scale could potentially enhance the performance of models like Flamingo and DALL-E, further advancing the field of multimodal research.

Table 1: We show an example generated related work for a randomly chosen recent paper (Li et al., 2023) with LLM summarized query and retrieved papers. We show the generated related work from our system using both zero-shot and plan-based generation, producing a more succinct and readily usable literature review. Note: The number of citations is retrieved by Semantic Scholar at the date of submission of this work.