Sematic Search Augmented Conversation for Enhanced Dialogue Generation

Yibo Yao[1a], Azlan Mohd Zain[2b,*], Kai-Qing Zhou[3c]

[1,2] Faculty of Computing, Universiti Teknologi Malaysia, 80310 Skudai, Johor, Malaysia
[3] School of Communication and Electronic Engineering, Jishou University, Jishou, China

[a]yibo.y@graduate.utm.my,[b]azlanmz@utm.my,[c]kqzhou@jsu.edu.cn

Abstract. Although advanced conversational models like ChatGPT are capable of generating rich and coherent content, the generated responses often contain fictional facts and knowledge hallucinations. A mainstream approach to addressing this problem in the past has been fine-tuning or retraining models by injecting external knowledge into pre-trained language models. However, given the enormous scale of current state-of-the-art language models, these methods require continuous retraining to update the knowledge embedded in the model parameters, which is undeniably challenging, slow, expensive, and the updated models lack scalability. In this work, we explore the use of semantic search based on user input and local knowledge to prompt language models for enhanced dialogue generation. We experiment with different domains of dialogue on four popular large language models (LLMs), and the results show that our approach, compared to the method of injecting knowledge into LLMs, can effectively improve the utilization efficiency of knowledge, significantly reduce knowledge hallucination problems, and has almost unlimited scalability.

Keywords: LLMs, semantic search, knowledge, prompt tuning, conversation generation

1 Introduction

Recently, universal large language models represented by ChatGPT and GPT-4[1] have achieved unprecedented success. These language models, which contain billions or even trillions of parameters, require large-scale pretraining on text data to acquire general capabilities for solving various tasks. In order to adapt these large language models to more specific task objectives, especially for unseen tasks, it is necessary to further utilize a large amount of text data in natural language format.

Currently, the commonly used methods for adaptation are Instruction Tuning [2, 3, 4, 5] and Alignment Tuning [2, 6, 7]. The former utilizes template-based examples (prompts) consisting of task descriptions, input-output pairs, and a small number of demonstrations to guide the model on what actions to take or what outputs to generate for specific tasks. With this guidance, the potential of large language models is further unlocked, leading to stronger generalization capabilities. The latter aims to avoid the

G. Guan et al. (eds.), *Proceedings of the 2023 3rd International Conference on Education, Information Management and Service Science (EIMSS 2023)*, Atlantis Highlights in Computer Sciences 16,
https://doi.org/10.2991/978-94-6463-264-4_84

generation of inaccurate, biased, or discriminatory statements [2, 7], in order to align with universal human values or preferences. The specific approach involves collecting human feedback, setting up reward models, and using reinforcement learning algorithms such as PPO [8] to learn from human feedback.

Although the adjusted dialogue models have performed very close to human-like, there are still three main challenges:

- **Slow and expensive**. Training even small-scale fine-tuning for large models is time-consuming and computationally expensive. Additionally, before supervised training using labelled instructions, a significant amount of human effort is required to design high-quality prompts. While methods have emerged to allow models to generate prompts for themselves or other large language models, some degree of manual inspection or filtering is still needed.
- **"Hallucination" problem** [9]. While the model's weights implicitly contain a vast amount of knowledge, the lack of collection and training on specific domains leads to seemingly reasonable yet inaccurate statements. For example, providing non-existent URLs or fabricating a work that does not belong to a specific director.
- **Data recency**. ChatGPT is only aware of events up until 2021, and even the most powerful GPT-4 cannot grasp information that occurred after the moment of its training without being connected to the internet or using plugins. This limitation is unfriendly for tasks relying on up-to-date information.

After problems above, introducing external knowledge has been proven to be an effective approach. Inspired by Jina AI's semantic search (or called neural search) method proposed in 2020, we attempt to improve the responses generated by large language models by utilizing semantic search. Neural search refers to the use of deep learning techniques to search unstructured data with unstructured data. The unstructured data can be diverse documents, images, or videos. This aligns well with the desired scenario for dialogue models, where users ask questions in natural language, and the model directly searches vast amounts of unstructured knowledge existing in the network or the real world based on the query. After referring to the search results, the model provides responses that are as reasonable and accurate as possible without requiring additional model training. The knowledge is up-to-date, and it possesses strong domain adaptability.

To evaluate the performance of LLMs with semantic search, we handpicked 2 state-of-the-art lightweight LLMs. By comparing the generated content of these models before and after applying semantic search, we evaluate and analyse the knowledgeable and factually incorrect using human evaluation of conversations.

2 Related Works

In the past few years, many researchers have been devoted to incorporating external knowledge, such as knowledge graphs, annotations, or unstructured texts, into language representation models to enhance input representations [10, 11, 12, 13] or dialogue generation [14, 15, 16]. The typical approach is to jointly train the text and knowledge base, enabling the language model to acquire prior knowledge. One important step is to establish accurate links between the entities mentioned in the input text and the knowledge in the knowledge base. K-ADAPTER [17] takes a different approach from jointly training a large model with text and knowledge base representations. It keeps the large model unchanged and adds adapters for specific knowledge behind the model. This approach ensures that the usage of the large model remains unaffected and prevents any deviation caused by injecting knowledge. Considering the differences between knowledge graphs and text structures, KELM [18] first transforms the knowledge graph into natural language text before integrating it into the pretrained language model.

In addition to the methods based on incorporation, there has been a recent resurgence of methods based on combining large models with retrievers. These methods are primarily used in Open-Book Question Answering (QA) tasks, where the LLM can query useful information or evidence from external knowledge bases or document collections and then use the extracted content to answer questions [19, 20, 21]. To select relevant knowledge from external resources, a retriever needs to be separately trained or jointly trained with the LLM. Alternatively, it can be combined with a search engine, where the generated text or the user's query serves as the search query, and the retrieval results are passed as additional context to the LLM [22, 23].

3 Approach

When users interact with a conversational agent, they typically have a specific purpose, such as obtaining information or completing a task, rather than engaging in purely casual conversation. Therefore, the core idea of our approach is to encode the user's question and unstructured data containing relevant knowledge (such as PDF documents) into vectors separately. These vectors are then matched and ranked using semantic search, and the sorted results are sent to the LLM in the form of prompt templates. Figure 1 illustrates the complete framework.

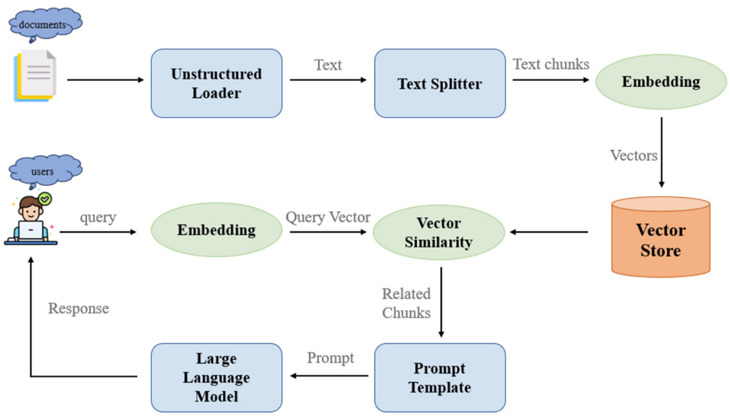

Fig. 1. Illustration of sematic search augmented conversation via prompted LLM.

3.1 Text embedding

The prerequisite for applying the semantic search method is that any form of data can be represented as vectors. Therefore, we use the same embedding technique to vectorize both the query and the documents to be retrieved. Since the documents to be retrieved may contain a large amount of textual content, it is necessary to split all the loaded documents into multiple text chunks to narrow down the retrieval results and improve retrieval accuracy. Each text chunk is encoded as a vector and stored in vector databases like FAISS [24], which are capable of handling millions of documents.

3.2 Calculation of search similarity

The similarity score is the cosine similarity between the query embedding and the embedding of each text chunk. After calculating the similarity scores, the text chunks are then sorted in descending order based on the scores. The top-K (K=3 in this paper) text chunks with the highest similarity scores are selected as the best matches.

3.3 Construct prompt

The construction of prompt templates aims to incorporate all the relevant text chunks and the query matched in the previous step into a unified prompt template, which is then sent to the LLMs for response generation. For multi-turn conversations, the historical dialogue information also needs to be considered when constructing the prompt templates.

4 Experiment and analysis

4.1 Dataset

To reduce the reliance of the large model on internal knowledge obtained from large-scale training corpora, we strictly limited the knowledge domain and recency of the knowledge documents. Ultimately, we manually collected 40 news articles or reports as external knowledge sources, forming a small dataset for testing purposes. This dataset covers topics in healthcare, science & technology, society, and movies & entertainment.

4.2 Evaluation

To assess the quality of knowledge generation, we evaluated the informativeness and factual correctness of the generated responses.

- Knowledgeable: Does the response contains relevant knowledge?
- Factual incorrect: Dose the response contains incorrect factual information?

The experiments were conducted in the form of single-turn conversations, where the queries were pre-set, and corresponding ground truth responses were provided as evaluation references. Table 1 provides a simple example.

Table 1. An example of evaluation

Query	What is the Hayao Miyazaki's latest film?		
Ground-Truth	The Boy and the Heron		
Answer	I'm sorry, as I am an AI language model, I cannot directly access the internet to obtain information. There is no information on Hayao Miyazaki's latest movies in my knowledge base. Hayao Miyazaki is a famous Anime director. His latest works may need to find other materials to obtain	**Answer with sematic search augmentation**	According to known information, Hayao Miyazaki's latest film, **The Boy and the Heron**, was released in Japan on July 14, 2023. It is an original story directed by Miyazaki and produced by Studio Ghibli, and tells the story of a boy named Ashuro and his friend, the Princess of Ashuro. The film follows GKids's usual policy of not releasing any further details or marketing materials.
Knowledgeable	NO	**Knowledgeable**	Yes
Factually incorrect	Incorrect	**Factually incorrect**	Correct

4.3 Result

In the experiments, we used Vicuna [25] and ChatGLM2 [26] to generate baseline answers and answers enhanced through semantic search. Table 2 presents the statistical results of the human evaluation.

Table 2. Comparation of generated response between LLM and LLM with sematic search

	Base answer		Augmented answer	
LLM	Knowledgeable	Factually incorrect	Knowledgeable	Factually incorrect
Vicuna-7B	78.2%	23.2%	82.8%	63.0%
ChatGLM2-6B	71.6%	18.7%	80.7%	50.5%

5 Conclusion and discussion

This study explores the issues of knowledge recency and knowledge illusion in large language models. By developing methods that incorporate external knowledge into the generation process by enhancing the original query, we demonstrate that such approaches can showcase more knowledge and generate fewer factual errors when interacting with humans. Future work should focus on how to incorporate multimodal data, such as images and videos, as well as structured data, such as knowledge graphs and tables, into the search framework. The biggest challenge in this process is how to unify heterogeneous information in a common semantic space for representation.

Acknowledgments

This work was supported in part by the Ministry of Higher Education Malaysia (MOHE) under Fundamental Research Grant Scheme (FRGS) R. J130000.7809.5F524.

References

1. OpenAI 2023 Gpt-4 technical report OpenAI. https://cdn.openai.com/papers/gpt-4.pdf
2. Ouyang L et al 2022 Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, **35**, 27730-27744.
3. Taori R, Gulrajani I, Zhang T, Dubois Y, Li X, Guestrin C, Liang P, and Hashimoto T B, Stanford alpaca: An instruction-following llama model
4. Wei J, Bosma M, Zhao V, Guu K, Yu A W, Lester B, Du N, M Dai A, Le Q V 2022 Finetuned Language Models are Zero-Shot Learners *in The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*
5. Chung H W et al 2022 Scaling instruction-finetuned language models *arXiv preprint arXiv*: 2210.11416
6. Glaese A et al 2022 Improving alignment of dialogue agents via targeted human judgements *arXiv preprint arXiv:2209.14375*
7. Kenton Z, Everitt T, Weidinger L, Gabriel I, Mikulik V and Irving G 2021 Alignment of language agents *arXiv preprint arXiv:2103.14659*
8. J Schulman, F Wolski, P Dhariwal, A. Radford and O Klimov 2017 Proximal policy optimization algorithms *arXiv preprint arXiv:*1707.06347, 2017
9. Maynez J, Narayan S, Bohnet B and McDonald R 2020 On faithfulness and factuality in abstractive summarization *In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1906–1919, Online. Association for Computational Linguistics
10. Zhang Z, Han X, Liu Z, Jiang X, Sun M and Liu Q 2019 ERNIE: Enhanced Language Representation with Informative Entities *In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (pp. 1441-1451)
11. Liu W, Zhou P, Zhao Z, Wang Z, Ju Q, Deng H and Wang P 2020 K-bert: Enabling language representation with knowledge graph *In Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. **34**, No. 03, pp. 2901-2908)

12. Sun Y, Wang S, Li Y, Feng S, Tian H, Wu H and Wang H 2020 Ernie 2.0: A continual pre-training framework for language understanding *In Proceedings of the AAAI conference on artificial intelligence* (Vol. **34**, No. 05, pp. 8968-8975)
13. Wang X, Gao T, Zhu Z, Zhang Z, Liu Z, Li J and Tang J 2021 KEPLER: A unified model for knowledge embedding and pre-trained language representation. *Transactions of the Association for Computational Linguistics*, 9, 176-194
14. Zhao X, Wu W, Xu C, Tao C, Zhao D and Yan R 2020 Knowledge-grounded dialogue generation with pre-trained language models *arXiv preprint arXiv*:2010.08824.
15. Chen X, Meng F, Li P, Chen F, Xu S, Xu B and Zhou J 2020 Bridging the gap between prior and posterior knowledge selection for knowledge-grounded dialogue generation *In Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP)* (pp. 3426-3437)
16. Li Y, Peng B, Shen Y, Mao Y, Liden L, Yu Z and Gao J 2021 Knowledge-grounded dialogue generation with a unified knowledge representation *arXiv preprint arXiv*:2112.07924
17. Wang R, Tang D, Duan N, Wei Z, Huang X, Cao G, Jiang D Zhou M 2020. K-adapter: Infusing knowledge into pre-trained models with adapters *arXiv preprint arXiv*:2002.01808
18. Agarwal O, Ge H, Shakeri S and Al-Rfou R 2020 Knowledge graph based synthetic corpus generation for knowledge-enhanced language model pre-training. *arXiv preprint arXiv*:2010.12688.
19. Izacard G et al 2022 Few-shot learning with retrieval augmented language models. *arXiv preprint arXiv*:2208.03299
20. Guu K, Lee K, Tung Z, Pasupat P and Chang M 2020 November Retrieval augmented language model pre-training. *In Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event,* 2020, pp. 3929–3938
21. Lewis P et al 2020 Retrieval-augmented generation for knowledge-intensive NLP tasks. *Advances in Neural Information Processing Systems*, 33, 9459-9474
22. Nakano R et al 2021 Webgpt: Browser-assisted question-answering with human feedback *arXiv preprint arXiv*:2112.09332
23. Lazaridou A, Gribovskaya E, Stokowiec W and Grigorev N 2022 Internet-augmented language models through few-shot prompting for open-domain question answering. *arXiv preprint arXiv*:2203.05115
24. Johnson J, Douze M and Jégou H 2019 Billion-scale similarity search with gpus *IEEE Transactions on Big Data*, 7(3), 535-547.
25. Chiang W L et al 2023 Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality *See https://vicuna. lmsys. org (accessed 14 April 2023)*
26. Du Z, Qian Y, Liu X, Ding M, Qiu J, Yang, Z and Tang J 2021 Glm: General language model pretraining with autoregressive blank infilling *arXiv preprint arXiv*:2103.10360