Augmenting Large Language Models with External Knowledge

Jinheon Baek
Graduate School of AI, KAIST
jinheon.baek@kaist.ac.kr

1 Introduction

Large Language Models (LLMs) have shown impressive capabilities in processing and generating text with remarkable accuracy, even outperforming human experts across diverse tasks and domains [14, 9]. However, the knowledge acquired from extensive training on a vast amount of corpora and internalized in their parameters can sometimes be inaccurate, incomplete, or outdated, leading to outputs that, while seemingly plausible, are factually incorrect. This problem is widely known as hallucination or confabulation [15], and it poses a substantial risk of spreading misinformation, potentially misleading users who rely on machine-generated information, especially in real-world application settings.

In my Ph.D. study, I strive to tackle the significant limitation of LLMs in generating factually incorrect responses by developing machine learning models and algorithms that 1) augment LLMs with external knowledge sources, 2) represent this supplemental knowledge (used for augmentation) holistically, and 3) apply these resulting knowledge-augmented LLMs to the challenging real-world applications. Specifically, as LLMs cannot memorize all the world knowledge into their parameters, the key to handling this limitation is to provide them with relevant knowledge retrieved from external knowledge sources, which enables LLMs to stay accurate and up-to-date. To operationalize this, we should answer the important question of how to enhance (or augment) LLMs with external knowledge, and I detail my efforts [11, 12, 4, 5, 7] to tackle this question in Section 2. In addition, the nature follow-up question to answer is how this knowledge is structured, represented, and retrieved, which I discuss based on my previous work [3, 10, 6, 1, 13, 2] in Section 3. Lastly, I am committed to making my approaches to knowledge-augmented LLMs useful and impactful to real-world applications (such as internet search or scientific research), and I showcase them [7, 16, 8] in Section 4. Broadly, I can define that my research area is in the field of machine learning for languages, knowledge, and their intersections at scale, and have published papers (as a first author) in both ML and NLP conferences, such as NeurIPS, ICML, ICLR, ACL, EMNLP, NAACL, and WWW. In the near future, I am eager to explore how LLMs can become more accurate, reliable, and trustworthy by leveraging the diverse knowledge existing in a variety of sources and formats, including figures, tables, and videos within a relational database system but not limited to them, which I describe in Section 5.

2 LLM Augmentation with (Un)structured Knowledge Sources

In this section, I describe my previous approaches to handle the limitation of (large) language models in generating factually incorrect outputs by augmenting them with external knowledge sources. It is worth noting that the knowledge that I utilize is ideally represented in a structured format (e.g., graph), since the knowledge gains more depth and meaning when it is interconnected with each other; yet, my previous approaches are flexible, which can utilize the knowledge in an unstructured format.

Knowledge-Augmented LLM Prompting I proposed a simple yet effective fundamental approach to augment LLMs with knowledge from external (graph-structured) knowledge sources, where this knowledge is injected into LLMs via prompting (Figure 1). More specifically, given a query from a user, the proposed Knowledge-Augmented language model PromptING (KAPING) method first retrieves the relevant knowledge to the given query from external knowledge sources (such as knowledge graphs). After that, the KAPING directly prepends the retrieved knowledge in the input of LLMs along with the query, from which LLMs can generate the correct answer grounded in the injected knowledge. I showed that KAPING can impressively improve the performance of various LLMs on multiple question answering tasks, which means it contributes to generating more factually correct answers.

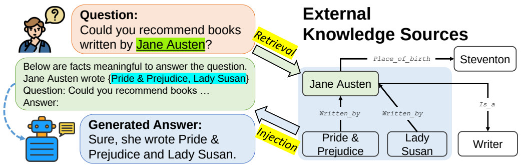
Question:
Could you recommend books written by Jane Austen?
Retrieval
External Knowledge Sources
Below are facts meaningful to answer the question.
Jane Austen wrote {Pride & Prejudice, Lady Susan}
Question: Could you recommend books …
Answer:
Generated Answer:
Sure, she wrote Pride & Prejudice and Lady Susan.
Injection
Jane Austen
Place_of_birth
Steventon
Written_by
Written_by
Is_a
Pride & Prejudice
Lady Susan
Writer

Figure 1: I proposed KAPING, that retrieves relevant facts to a query and augments LLMs by injecting them as a prompt, to improve factuality [4].

Knowledge-Augmented LLM Verification I observed that, despite a remarkable success of knowledge-augmented LLMs in generating factually correct outputs, they are still not reliable and largely problematic to deploy in prediction. This is because the model may fail to retrieve the knowledge relevant to the given query or the model may not faithfully reflect the retrieved knowledge in the generated output, which leads to generating misinformation (See Figure 2, Left). To overcome these substantial challenges, I built a verifier (KALMV) that can detect those two types of errors based on a combination of question, knowledge, and answer (Figure 2). Also, if errors are detected from the proposed verifier, I developed strategies that can rectify them by retrieving the knowledge and generating the output iteratively until making the correct prediction. I showed that KALMV can significantly reduce hallucinations, improving reliability of LLM-based systems.

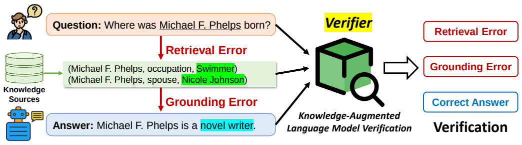
Question: Where was Michael F. Phelps born?
Retrieval Error
Knowledge Sources
(Michael F. Phelps, occupation, Swimmer)
(Michael F. Phelps, spouse, Nicole Johnson)
Grounding Error
Answer: Michael F. Phelps is a novel writer.
Verifier
Knowledge-Augmented Language Model Verification
Retrieval Error
Grounding Error
Correct Answer
Verification

Figure 2: I proposed Knowledge-Augmented Language Model Verification (KALMV), which identifies knowledge retrieval and grounding errors.

Knowledge-Augmented LLM Personalization Based on aforementioned two approaches [4, 5] that augment LLMs with external knowledge sources but also verify and rectify their outputs, the responses of LLMs can be more accurate, reliable, and trustworthy. However, one particularly important challenge of knowledge-augmented LLMs is that they do not reflect the personal knowledge and interests of individual users when handling their queries. To this end, I developed Knowledge-augmented Language Models for Personalization (K-LaMP) (Figure 3). Specifically, K-LaMP first constructs the personal knowledge store of each individual user by extracting and aggregating entities from their interaction histories with electronic devices (e.g., search and browsing logs), which can capture different views of user's knowledge (e.g., familiar or unfamiliar entities). After that, given a query from the specific user, K-LaMP retrieves their relevant knowledge from the personal knowledge store by matching entities between the query and store. Then, based on this retrieved knowledge, K-LaMP augments LLMs, which enables generating responses that reflect user's personal knowledge and interests.

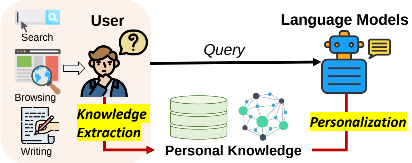

Figure 3: I proposed K-LaMP that personalizes outputs of LLMs based on the personal knowledge store of each user.

Knowledge-Augmented Language Model Modulation Before the era of LLMs, the most common way to direct Language Models (LMs) in generating desired outputs is to fine-tune them with appropriate techniques. In line of this paradigm, my previous work (namely KALA [11] and SURGE [12]) includes language model augmentation approach that modulates intermediate representations of LMs with respect to representations of external knowledge sources, to inject external knowledge into LMs. It is worth noting that, even though my previous modulation approach requires fine-tuning to align representations of LMs with external knowledge, in the test-time, it does not increase the number of tokens to integrate the external knowledge with LMs (unlike LLM prompting approaches). This framework thus offers significant efficiency benefits, and extending it to LLMs would be fascinating.

3 Representation Learning for Utilizing Structured Knowledge

To enhance the efficacy of knowledge-augmented LLMs, it is crucial to provide them with appropriate sources of knowledge (and oftentimes its representation). However, despite the fact that knowledge is most impactful when modeled as a graph (since its interconnected nature adds depth and insight), the methods to handle the graph-structured knowledge remain relatively underexplored. This section discusses my previous efforts to improve the retrieval strategy for graph-structured knowledge, along with several methods to advance the representation learning that supports more meaningful retrieval.

Retrieving Graph-Structured Knowledge Each fact on a graph-structured knowledge can be represented as a triplet consisting of two entities and one relation between them. To retrieve this fact, existing approaches typically perform three sequential steps: entity detection, entity disambiguation, and relation classification. However, this pipeline approach is complex and suboptimal, which is prone to error propagation and lacks generalizability. To this end, I developed a paradigm-shifting approach (Figure 4) that verbalizes facts and utilizes LMs to represent them, which enables direct retrieval of facts based on their representational similarities with input queries [6].

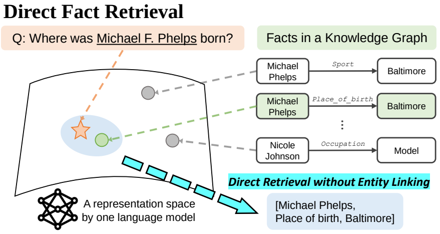
Direct Fact Retrieval
Q: Where was Michael F. Phelps born?
Facts in a Knowledge Graph
Michael Phelps
Sport
Baltimore
Michael Phelps
Place_of_birth
Baltimore
⋮
Nicole Johnson
Occupation
Model
Direct Retrieval without Entity Linking
A representation space by one language model
[Michael Phelps, Place of birth, Baltimore]

Figure 4: I proposed DiFaR [6], which represents facts with LMs and then retrieves them based on their similarities with input queries.

Holistic Representation Learning of Graph-Structured Knowledge To improve the retrieval of graph-structured knowledge (to ultimately improve the knowledge-augmented LLM performance), it is crucial to comprehensively represent all components of the graph, such as nodes, edges, and entire (sub)graphs. Yet, existing approaches for graph-structured data tend to focus predominantly on static nodes. To address these challenges, I developed novel methods to holistically represent graph-structured knowledge, which include representation learning for nodes not seen during training by simulating them with meta-learning for training [3], edges by transforming them as nodes with dual hypergraph transformation [10], and entire (sub)graphs by compressing them with graph multiset transformers [1]. Notably, they contribute to more accurate retrieval of structured knowledge [11, 12].

Learning Strategies for Real-World Graph-Structured Knowledge In some real-world scenarios, graph-structured knowledge might lack training labels or be distributed across various clients. To address them, I introduced two innovative learning strategies: for unlabeled data, I developed a self-supervised learning objective based on graph edit distances [13]; for decentralized data, I proposed a personalized federated learning framework that utilizes community structures of graphs [2]. These two approaches contribute to allowing for the utilization of structure knowledge in extreme contexts.

4 Real-World Applications of Knowledge-Augmented LLMs

In this section, I showcase how the knowledge-augmented LLMs developed and improved through my previous work (detailed in Sections 2 and 3) can be effectively utilized to solve real-world challenges.

Personalized Contextual Query Suggestion for Web Search The goal of query suggestion is to recommend new, relevant queries to users, which is highly practical having been shipped in web-scale search engines (Google). In my recent work [7], I further improved this process by suggesting the next queries that are additionally conditioned on what a user is currently viewing and knowing. This is achieved by augmenting LLMs with contextual and personalized knowledge relevant to each user.

Generating Research Ideas for Accelerating Scientific Research While scientific research is vital for improving human life, it is hindered by its inherent complexity, slow pace, and the need for specialized experts. To revolutionize this process, I developed a scientific knowledge-augmented LLM that automatically generates problems, methods, and experiment designs with scientific literature [8].

Data Augmentation for Low-Resource Domain Tasks Despite the notable successes of LLMs, their performance significantly deteriorates in low-resource settings, where the data available for training is very scarce (such as in the case of emerging events like novel viruses) or, in certain cases, completely unavailable (such as in privacy-sensitive enterprise contexts). To address this challenge, I proposed to augment LLMs with samples from an external database, to generate diverse samples [16].

5 Future Research

Due to the static nature of LLM parameters, the issues that their knowledge is incomplete and outdated continue to be significant challenges, resulting in hallucinations. While my prior work on developing knowledge-augmented LLMs has made strides towards addressing this issue, I further aim to enhance this augmentation framework by enabling it to consume a broader range of multimodal knowledge sources. Also, I aspire to empower real-world applications with my knowledge-augmented LLMs.

Augmenting Multimodal LLMs with Multimodal Knowledge One compelling avenue for LLM augmentation involves enhancing it with diverse types of knowledge beyond textual contexts, such as images, videos, and audio. This strategy allows LLMs to gain a more holistic understanding of multifaceted, real-world problems, ultimately broadening their functionality. For example, augmenting multimodal LLMs with video-embedded web pages can yield more comprehensive answer generation to the given query. I envision pursuing this line of research built upon my prior work in a few years.

Database-Augmented Multimodal LLMs In practice, a vast amount of enterprise data is stored in a structured database, which consists of various types of data such as texts, numbers, and multimedia, and is often organized into tables to support efficient data manipulation, which far differs from usual benchmark data that we have used for LLM augmentation. In light of this, I aim to further extend my approaches to augmenting LLMs, sitting them on top of databases to handle richer and dynamically updated data samples. This will potentially transform how businesses utilize LLMs with their data.

Augmenting (Multimodal) LLMs in Real World To tackle the problem of LLMs that are limited by their static knowledge (a significant barrier in production), I have actively applied my expertise in knowledge-augmented LLMs to practical scenarios (Section 4). Moving forward, I intend to expand their capabilities by integrating more robust, multimodal knowledge sources, and aligning them closely with real-world contexts and user-specific needs, ensuring that LLMs are not only powerful in their computational abilities but also truly useful and adaptable in practical, everyday applications.

References

[1] Jinheon Baek*, Minki Kang*, et al. Accurate learning of graph representations with graph multiset pooling. 2021.

[2] Jinheon Baek*, Wonyong Jeong*, et al. Personalized subgraph federated learning. *ICML*, 2023.

[3] Jinheon Baek et al. Learning to extrapolate knowledge: Transductive few-shot out-of-graph link prediction. *NeurIPS*, 2020.

[4] Jinheon Baek et al. Knowledge-augmented language model prompting for zero-shot knowledge graph question answering. *NRLSE @ ACL*, 2023.

[5] Jinheon Baek et al. Knowledge-augmented language model verification. *EMNLP*, 2023.

[6] Jinheon Baek et al. Direct knowledge graph retrieval without entity linking. *ACL*, 2023.

[7] Jinheon Baek et al. Knowledge-augmented large language models for personalized contextual query suggestion. *WWW*, 2024.

[8] Jinheon Baek et al. Researchagent: Iterative research idea generation over scientific literature with large language models. *arXiv preprint arXiv:2404.07738*, 2024.

[9] Gemini Team Google. Gemini: A family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

[10] Jaehyeong Jo*, Jinheon Baek*, Seul Lee*, et al. Edge representation learning with hypergraphs. *NeurIPS*, 2021.

[11] Minki Kang*, Jinheon Baek*, et al. Kala: Knowledge-augmented language model adaptation. *NAACL*, 2022.

[12] Minki Kang*, Jin Myung Kwak*, Jinheon Baek*, et al. Knowledge-consistent dialogue generation with knowledge graphs. *KRLM @ ICML*, 2022.

[13] Dongki Kim*, Jinheon Baek*, et al. Graph self-supervised learning with accurate discrepancy learning. *NeurIPS*, 2022.

[14] OpenAI. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

[15] Anna Rohrbach et al. Object hallucination in image captioning. 2018.

[16] Minju Seo*, Jinheon Baek*, et al. Retrieval-augmented data augmentation for low-resource domain tasks. *arXiv preprint arXiv:2402.13482*, 2024.