Knowledge-Augmented Large Language Models with Personalized Knowledge Representation, Retrieval, Injection, and Verification

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

1 Introduction

Large Language Models (LMs) have shown numerous successes in a variety of natural language tasks, such as question answering and dialogue [6, 12]. However, the knowledge internalized in their parameters can sometimes be inaccurate, incomplete, or outdated, which leads them to generate outputs that are factually incorrect although they appear plausible. This problem is widely known as hallucination or confabulation [13], and it poses a substantial risk of spreading misinformation, potentially misleading users who rely on the information, especially in real-world production settings.

Alternatively, there exists plenty of external knowledge sources that we can use in order to complement the knowledge of LMs. However, this knowledge rarely stands alone; rather it gains more depth and meaning when it is interconnected. For example, considering two entities "Steve Jobs" and "Tim Cook" in isolation offers limited insight. Yet, when they are linked with the context of Apple Inc., such as their roles as CEOs of Apple Inc., we can gain a more comprehensive and profound understanding. To model such a large-scale interconnected knowledge, a graph is a commonly used data structure that can denote how everything is connected to everything else with nodes and edges. Also, to leverage this structured graph data, we typically use Graph Neural Networks (GNNs) [7], which represent each node by iteratively aggregating the features from its neighboring nodes. However, existing GNNs are still suboptimal to represent diverse real-world graph structures, which sometimes hinders us from directly utilizing the external graph-structured knowledge to complement the knowledge of LMs.

In my Ph.D. study, in order to tackle the critical limitation of LMs in generating factually incorrect outputs, I propose several methods [9, 10, 5] that augment LMs by injecting the knowledge from the graph-structured knowledge; and, in order to inject the right source of the knowledge effectively, I propose a series of fundamental works [4, 8, 2, 1, 11, 3] that can represent and retrieve the graph-structured knowledge and data. In particular, in my previous works on learning graph-structured data, I devise novel methods that can represent edges and entire graphs as well as unseen entities that evolve over time and facts that consist of two entities and one relation, to overcome challenges of existing GNNs that are limited to representing mostly the static nodes. Also, I propose new learning strategies for GNNs, which can either represent the graph-structured data without any labels on it based on the self-supervision from graph distances, or train the distributed subgraphs over multiple clients without sharing their data based on the graph-structure-aware federated learning. After that, by leveraging my previous methods on representing graph-structured knowledge, I further contribute to developing knowledge-augmented LMs. Specifically, I propose multiple techniques that inject not only the represented knowledge from Knowledge Graphs (KGs) into the input or intermediate layers of LMs, but also the verbalized knowledge retrieved from KGs into large LMs via prompting.

While I identify and address the challenging problems of representing the graph-structured knowledge and data with novel GNNs, and then augmenting the LMs based on the representation, retrieval, and injection of the knowledge from KGs, there still exist remaining challenges to be tackled. Specifically, while LMs can contain a vast amount of general knowledge in their parameters, the text generated from LMs may not be specific and relevant to the individual users since they cannot capture their personal knowledge. Yet, the users may have particular experiences or interests in some concepts, topics, or subjects (e.g., Apple Inc.), which are substantially worthwhile to be reflected in the generated texts. Therefore, I am planing to work on the personalization of LMs by augmenting them with the user's personal knowledge. Also, despite the successes of knowledge-augmented LMs in generating more factually correct outputs, they are still suboptimal since they may fail to retrieve the relevant knowledge to the given query or may not faithfully reflect the injected knowledge in the output texts. In order to overcome these, I am planning to develop a verification method that can detect errors in both knowledge retrieval and answer generation but also can rectify them if identified. The comprehensive overview of my previous and potential research projects is visualized in Figure 1.

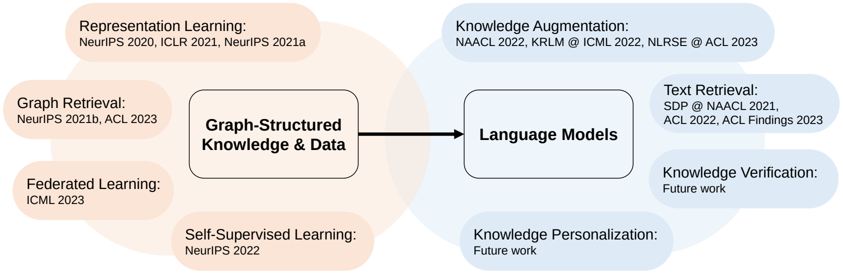


Figure 1: **Overview of previous and future research**, which includes learning the graph-structured knowledge and data, and then retrieving and injecting them to augment large language models with the knowledge verification and personalization methods for real-world natural language applications.

2 Previous Research Experience

In this section, I explain my previous works on representation learning for graph-structured knowledge and data, which aim to tackle several limitations of GNNs in handling real-world graph structures, and then introduce my previous efforts to overcome challenges of LMs in memorizing facts by retrieving and injecting the pertinent knowledge to queries based on my structured knowledge learning methods.

2.1 Learning on Graph-Structured Knowledge and Data

I first discuss a series of my previous methods on representing graph-structured knowledge and data.

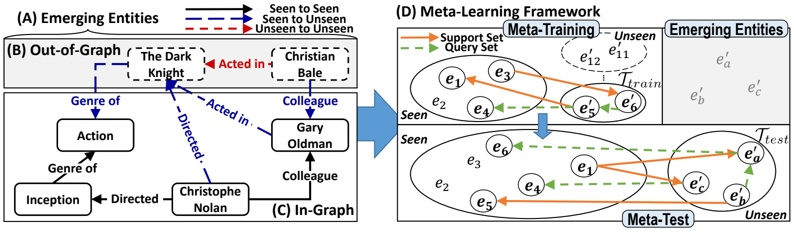


Figure 2: **Graph Extrapolation Network (GEN)** [4]. (Left) An illustration of unseen entities and their out-of-graph link prediction problem with other seen and unseen entities. (Right) An illustration of a meta-learning framework to represent unseen entities first and then predict links for them.

Representation Learning on Unseen Entities In many real-world knowledge graphs, unseen entities continuously emerge while having relationships with other entities (Figure 2, Left). However, existing GNNs, which are trained and evaluated over the same set of seen entities, are suboptimal to handle such emerging entities. To this end, I first introduce an out-of-graph link prediction problem [4], whose goal is to represent unseen nodes to incorporate them onto the existing graph with seen nodes, by predicting links between seen and unseen as well as unseen entities themselves. Then, I tackle this problem with a novel meta-learning framework [4], which meta-learns any GNNs to extrapolate the knowledge from seen to unseen entities (Figure 2, Right). I validate the effectiveness of the proposed GEN on multiple knowledge graphs, showing its efficacy in handling unseen entities.

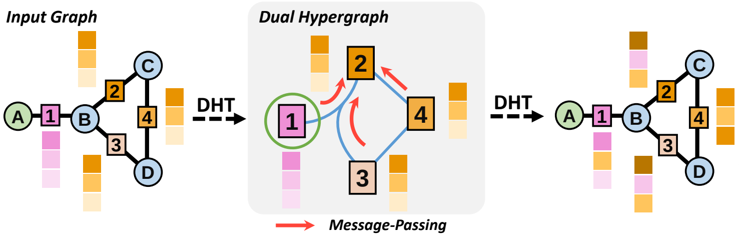


Figure 3: **Edge HyperGraph Neural Network (EHGNN)** [8]. Dual Hypergraph Transformation (DHT) transforms the edges of the original graph as the nodes of the dual hypergraph, which allows us to represent edges as nodes over the dual hypergraph with any existing graph neural networks.

Representation Learning on Edges In graph-structured knowledge and data, edges play a pivotal role. For example, in knowledge graphs, edges themselves carry semantic meaning that provides a more profound understanding of two linked entities. However, existing GNNs focus mostly on representing nodes, and consider edges as mere paths for sharing information between nodes. To tackle this limitation, I propose to transform the edges of the original graph to nodes of the hypergraph and nodes to hyperedges with dual hypergraph transformation [8] (Figure 3). After that, we can use any existing GNNs to represent nodes of the dual hypergraph, for learning the representations of edges of the original graph. The proposed EHGNN is applicable to any graph structure, and I validate it on node and graph classification tasks as well as graph reconstruction and generation tasks.

Representation Learning for Fact Retrieval To retrieve facts from KGs, which consist of two entities and one relation, for downstream tasks (e.g., question answering), existing works typically perform three sequential steps: entity detection, entity disambiguation, and relation classification, which is yet complex and suboptimal. Thus, I propose to represent each verbalized fact with LMs [2], which enables us to retrieve facts directly by calculating their representational similarities with input queries (Figure 4). I demonstrate the efficacy of the proposed DiFaR on fact retrieval.

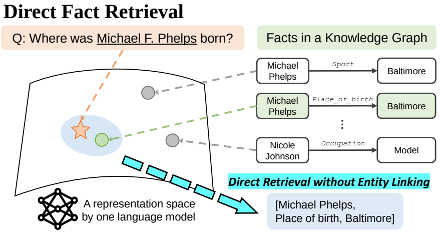

Direct Fact Retrieval
Q: Where was Michael F. Phelps born?
Facts in a Knowledge Graph
Michael Phelps
Sport
Baltimore
Michael Phelps
Place_of_birth
Baltimore
Nicole Johnson
Occupation
Model
Direct Retrieval without Entity Linking
A representation space by one language model
[Michael Phelps, Place of birth,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

Figure 4: **Direct Fact Retrieval (DiFaR)** [2], which represents and retrieves facts based on their similarities to input queries, using LMs.

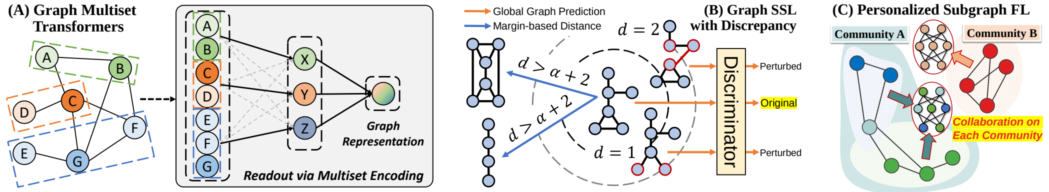


Figure 5: **(A) Graph Multiset Transformers (GMT)** [1], which hierarchically represents the entire graphs with multiset-based transformers. **(B) Discrepancy-based Self-supervised LeArning (D-SLA)** [11], which learns discrepancy between graphs based on their edit distances to represent graphs without labels, during self-supervised learning. (C) **FEDerated Personalized sUBgraph learning (FED-PUB)** [3], which trains locally accessible distributed subgraphs by considering their community structures, in order to jointly improve the personalized models working on interconnected subgraphs.

Representation Learning with Graph Pooling Some practical scenarios (e.g., graph classification and retrieval) require obtaining the hierarchical or global representations of sub- and entire graphs by summarizing representations of their nodes and edges. To this end, I propose the graph pooling operator that can hierarchically learn the representations of entire graphs based on a novel multiset-based transformer architecture [1], illustrated in Figure 5 (A). Further, I theoretically prove that the proposed graph pooling, namely GMT, is as powerful as the Weisfeiler-Lehman graph isomorphism test [14] in distinguishing graphs, and empirically show that GMT outperforms existing graph pooling on graph classification, reconstruction, and generation tasks with high memory and time efficiency.

Strategies for Graph Representation Learning In some real-world scenarios, graph-structured knowledge and data may not have labels for training or may be distributed over multiple local clients. However, existing graph representation learning methods are suboptimal to handle such challenging scenarios, and I propose two novel strategies to tackle them. In particular, to train GNNs without any labeled data, I devise the pre-text task for graph-structured data, which learns the discrepancy between graphs not only by discriminating real graphs from their perturbed graphs but also by directly learning the graph edit distances between them over self-supervised learning [11], visualized in Figure 5 (B). On the other hand, in order to train with multiple subgraphs that are locally accessible due to privacy restrictions, I propose the personalized federated learning framework, which aims at promoting the joint improvement of local subgraphs that belong to the same community by selectively utilizing the knowledge across subgraphs from diverse communities [3], illustrated in Figure 5 (C). I demonstrate the efficacy of my self-supervised learning (D-SLA) and federated learning (FED-PUB) methods on node classification, link prediction, and graph classification tasks of various real-world graphs.

2.2 Knowledge-Augmented (Large) Language Models

In this subsection, I describe my previous methods on knowledge-augmented LMs to complement their limited knowledge, where the knowledge is modeled by my proposed works in Subsection 2.1.

Knowledge-Augmented Language Model Modulation To inject the relevant graph-structured knowledge into LMs, I first propose to modulate the intermediate representations of LMs with respect to the KG representations [9]. Specifically, I first represent entities and relations in KGs, which are associated with the input query (e.g., entities appear in the input text), based on GNNs. After that, I propose to linearly scale and shift the intermediate representations of entities in LMs with the entity

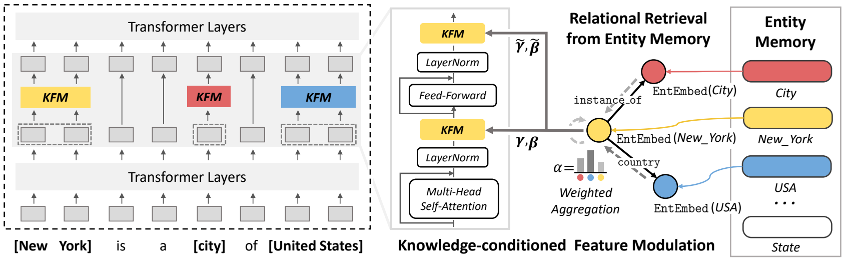
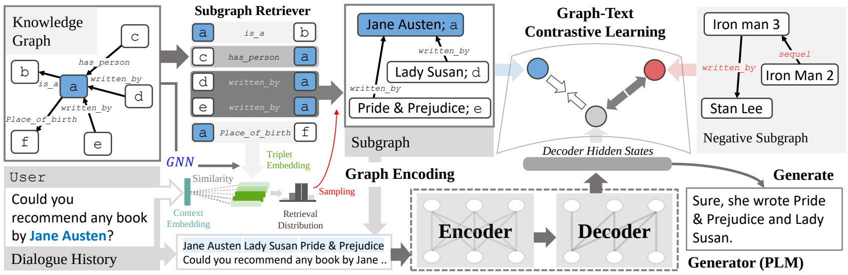


Figure 6: **Knowledge-Augmented Language model Adaption (KALA)** [9], which injects the representations of entities from KGs into LMs by modulating the intermediate entity representations of LMs via the Knowledge-conditioned Feature Modulation layer (KFM) that is interleaved in LMs.

representations from KGs, which I refer to as the Knowledge-conditioned Feature Modulation (KFM) layer that is interleaved in LMs (See Figure 6). Then, I validate the proposed method, namely KALA, on question answering and named entity recognition tasks, showing its effectiveness.



Could you recommend any book by Jane Austen?

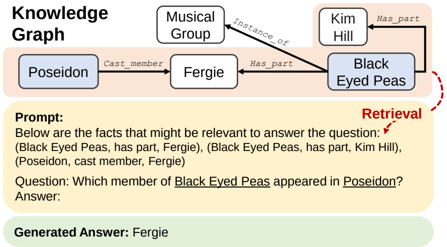
Jane Austen Lady Susan Pride & Prejudice Could you recommend any book by Jane ..

Sure, she wrote Pride & Prejudice and Lady Susan.

Figure 7: **SUbgraph Retrieval-augmented Generation (SURGE)** [10], which consists of three components. It first retrieves the subgraph relevant to the dialogue from KGs. Then, it encodes the graph-structured information of the retrieved subgraph into the input text of the generator. Lastly, it uses contrastive learning to enforce the model to generate a response consistent with the subgraph.

Knowledge-Augmented Language Models for Text Generation In contrast to the aforementioned work that injects the knowledge implicitly into the representations of LMs, I also propose to explicitly incorporate the graph-structured knowledge from KGs into the input text of LMs [10]. Specifically, as shown in Figure 7, I retrieve the subgraph relevant to the input query by calculating similarities between the representations of the input text from LMs and its associated facts from GNNs, and then generate the response by injecting the retrieved facts in the LM input, where the retriever and generator are jointly trained. Further, I propose the graph encoder to reflect the graph-structured information in the text representation, but also the contrastive learning strategy to ensure similarities between the text and its relevant subgraphs. I validate the proposed SURGE in generating responses on dialogue tasks, showing they are more informative thanks to leveraging relevant facts from KGs.

Knowledge-Augmented Large Language Model Prompting Unlike smaller LMs, large LMs are capable of answering questions without fine-tuning. Yet, they are still vulnerable to generating factually incorrect answers since they cannot store all the knowledge in their parameters. To overcome this without further training, I propose to augment large LMs with the knowledge from KGs via prompting [5]. Specifically, similar to the aforementioned two works, I first retrieve the relevant facts to the given query from KGs and then directly append them in the input of LMs along with the query, which is then forwarded to the LMs to generate the correct answer (See Figure 8). I validate the proposed Knowledge-Augmented language model PromptING (KAPING) on knowledge graph question answering, showing it impressively improves the performance of various large LMs.



Prompt:
Below are the facts that might be relevant to answer the question:
(Black Eyed Peas, has part, Fergie), (Black Eyed Peas, has part, Kim Hill), (Poseidon, cast member, Fergie)
Question: Which member of Black Eyed Peas appeared in Poseidon?
Answer:

Generated Answer: Fergie

Figure 8: **Knowledge-Augmented language model PromptING (KAPING)** [5], which retrieves relevant facts to the query and augments LMs by injecting them as the prompt.

3 Ongoing and Future Research

My ongoing and future research is tackling more realistic and practical yet challenging problems that arise when deploying large LMs in real-world scenarios, e.g., output validation and personalization.

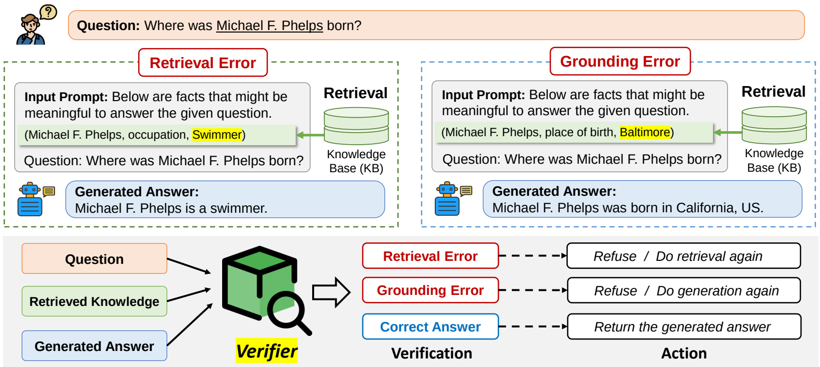

Question: Where was Michael F. Phelps born?
Retrieval Error
Input Prompt: Below are facts that might be meaningful to answer the given question.
(Michael F. Phelps, occupation, Swimmer)
Question: Where was Michael F. Phelps born?
Retrieval
Knowledge Base (KB)
Generated Answer: Michael F. Phelps is a swimmer.
Grounding Error
Input Prompt: Below are facts that might be meaningful to answer the given question.
(Michael F. Phelps, place of birth, Baltimore)
Question: Where was Michael F. Phelps born?
Retrieval
Knowledge Base (KB)
Generated Answer: Michael F. Phelps was born in California, US.
Question
Retrieved Knowledge
Generated Answer
Verifier
Retrieval Error
Grounding Error
Correct Answer
Verification
Refuse / Do retrieval again
Refuse / Do generation again
Return the generated answer
Action

Figure 9: **Concept of Knowledge-Augmented Language Model Verification**, which not only identifies knowledge retrieval and grounding errors but also rectifies the outputs if errors are detected.

Knowledge-Augmented LM Verification Knowledge-augmented LMs have achieved remarkable success in generating factually correct outputs. Yet, they are still not reliable and rather problematic to deploy in production, since 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 text, which leads to generating misinformation. To overcome these substantial challenges, I plan to build a verifier that can detect such two types of errors based on the combination of question, knowledge, and answer (See Figure 9). Moreover, I plan to further develop strategies that can rectify such two types of errors if detected, by retrieving the knowledge and generating the output iteratively until making the correct.

Personalized, Knowledge-Augmented LMs One practical challenge of knowledge-augmented LMs is that they do not reflect the personal knowledge and interests of individual users when handling their queries. To this end, as illustrated in Figure 10, I plan to extract the personal knowledge of users from their interactions with devices (e.g., search and browsing history) and store it as unstructured memory and structured knowledge graphs. After that, I augment LMs with the extracted knowledge of users, which leads to generating outputs that reflect the user's personal knowledge and interests (e.g., providing information about Tim Cook, if the user is interested in Apple Inc. and familiar with Steve Jobs but not with Tim Cook).

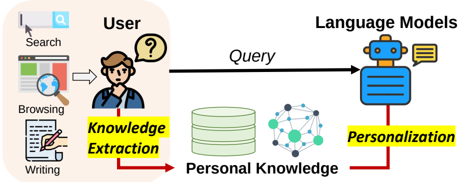


Figure 10: **Concept of Personalized Knowledge Augmented LMs**, which extracts the personal knowledge of individual users from their interactions with devices, and augments LMs with it for personalized text generation.

4 Conclusion

In this research statement, I introduced a series of my previous research on augmenting LMs with graph-structured knowledge and data by representing, retrieving, and injecting them in innovative and effective ways. As I move forward, my focus will be on improving knowledge-augmented LMs by addressing more practical problems, which include the knowledge personalization of LMs for individual users and the post hoc verification of LM outputs, at scale. I strongly hope that my previous and ongoing researches will bring substantial advancements and practical benefits to the community.

References

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

[2] Jinheon Baek, Alham Fikri Aji, Jens Lehmann, and Sung Ju Hwang. Direct knowledge graph retrieval without entity linking. *ACL*, 2023.

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

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

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

[6] Tom Brown et al. Language models are few-shot learners. *NeurIPS*, 2020.

[7] William L. Hamilton. Graph representation learning. *Synthesis Lectures on Artificial Intelligence and Machine Learning*, 2020.

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

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

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

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

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

[13] Anna Rohrbach et al. Object hallucination in image captioning. In *EMNLP*, 2018.

[14] B. Yu. Weisfeiler and A. A. Leman. Reduction of a graph to a canonical form and an algebra arising during this reduction. 1968.