Knowledge Enhancement in Language Models: Retrieval Augmentation or Adapter-Based Integration?

Dong Lu
Northeastern University
lu.dong1@northeastern.edu

Abstract

Large language models (LLMs) have demonstrated strong linguistic capabilities across a wide range of NLP tasks. However, they still struggle to reliably access and apply precise factual knowledge, making them prone to hallucination. A common strategy to address this limitation is to enhance language models with external knowledge sources. Since existing knowledge enhancement methods are often studied in isolation, their relative effectiveness remains unclear. In this project, we conduct a comparative study of two representative knowledge enhancement paradigms: Retrieval-Augmented Generation (RAG), which incorporates external knowledge at inference time, and K-Adapter, which injects structured knowledge into model parameters via adapter modules. We evaluate model performance using a factual recall task and report accuracy, mean reciprocal rank (MRR), and cosine similarity. Experimental results show that RAG achieves the most consistent and strongest performance, while K-Adapter improves over the baseline but remains sensitive to data scale and knowledge quality. These findings suggest that retrieval-based knowledge enhancement is more effective under limited domain-specific data settings, whereas adapter-based knowledge integration is a viable but more data-dependent alternative. Future work includes exploring hybrid approaches that combine retrieval and adapter-based mechanisms to further improve factual reliability.

1 Introduction

Large language models (LLMs) and pre-trained language models (PLMs) have reshaped modern natural language processing (NLP), particularly with the introduction of Transformer-based architectures (Vaswani et al., 2017). Models such as BERT (Devlin et al., 2019) and its successors (e.g., RoBERTa; Liu et al., 2019) learn strong linguistic representations from large-scale self-supervised pretraining and can be adapted to a wide range of downstream tasks with minimal supervision. Despite these advances, LLMs still struggle to reliably access and apply precise factual knowledge during inference, especially in domain-specific settings. This limitation often leads to hallucination, where models generate fluent but incorrect statements or fail to recall relevant facts.

Hallucination arises from how LLMs are trained and what they are optimized for. These models primarily learn statistical regularities from large text corpora, where factual knowledge is often long-tailed and not consistently recalled at inference time (Kandpal et al., 2023). As a result, LLMs can produce fluent and confident outputs without a built-in mechanism to verify the correctness. This becomes a serious concern in real-world applications such as healthcare, law, and education, making hallucination mitigation a critical challenge for improving the reliability and trustworthiness of LLMs.

A common direction for mitigating hallucination is to enhance language models with external knowledge (Logan et al., 2019). Existing approaches can be broadly categorized into two paradigms. The first consists of retrieval-based methods, which retrieve relevant factual information at inference time and use it as additional context to guide generation. Retrieval-Augmented Generation (RAG) is a representative framework in this category, where external evidence is retrieved and incorporated into the model input to improve factual grounding (Lewis et al., 2020; Guu et al., 2020). The second paradigm includes parametric knowledge integration methods, which aim to encode structured knowledge directly into model parameters during training. Examples include entity-aware models such as KnowBERT (Peters et al., 2019) and LUKE (Yamada et al., 2020), as well as adapter-based approaches such as K-Adapter (Wang et al., 2021), which inject knowledge through lightweight adapter modules without fully fine-tuning the backbone model.

While both retrieval-based and parametric approaches have been studied for mitigating hallucination, existing research often focuses on one method in isolation, making it difficult to directly compare their effectiveness or understand the conditions under which one approach may be preferable to the other. In this project, we conduct a controlled comparison between these two methods under an identical experimental setting. We construct a domain-specific knowledge graph (KG) and use it in two different ways: for RAG, factual information is retrieved from the KG at inference time and appended to the model input; for K-Adapter, the same KG is used to train a knowledge adapter that injects structured knowledge into the model parameters. By comparing these approaches using a common domain-adapted BERT backbone without knowledge enhancement, we aim to better understand their relative effectiveness and limitations in mitigating hallucination.

Our experimental results show that knowledge enhancement leads to clear improvements in factual recall and semantic understanding over text-only baselines. Under small-scale data settings, retrieval-based methods are more stable and effective, while adapter-based knowledge integration remains a promising direction that may benefit from larger corpora, stronger supervision, or improved training objectives.

2 Related Work

In recent years, there has been extensive research on knowledge enhancement for language models, which has played an important role in advancing natural language processing. Several influential works in this area form the foundation of our study.

2.1 BERT

Bidirectional Encoder Representations from Transformers (BERT) is a pretrained language model introduced by Google AI Research that significantly advanced natural language understanding at the time of its release (Devlin et al., 2019). Unlike earlier unidirectional models, BERT employs a Transformer encoder to model both left and right context simultaneously, enabling richer contextual representations. BERT is trained using masked language modeling on large-scale general-purpose text corpora, followed by task-specific fine-tuning, which allows it to generalize effectively across a wide range of downstream NLP tasks. However, because this pretraining focuses on broad, open-domain data, BERT has relatively weak domain-specific factual knowledge, making it a suitable baseline for this study.

2.2 Retrieval-Augmented Generation

Retrieval-Augmented Generation (RAG) is a framework that enhances language models by combining neural retrieval with generation or prediction (**?**). Rather than relying solely on knowledge stored in model parameters, RAG retrieves relevant documents or factual evidence from an external knowledge source at inference time and incorporates this information into the model input. This retrieval-based design grounds model predictions in explicit evidence, making it particularly effective for knowledge-intensive tasks. By decoupling knowledge storage from model parameters, RAG also allows language models to leverage large or frequently updated knowledge sources without retraining, and has been shown to reduce hallucination and improve factual reliability.

2.3 K-Adapter

K-Adapter is a lightweight and flexible approach for injecting structured knowledge into pretrained language models through adapter modules (Wang et al., 2021). Instead of modifying the backbone model, K-Adapter attaches knowledge-specific adapters to intermediate Transformer layers, allowing different types of knowledge to be learned independently. These adapters can be trained using supervision from knowledge graph triples or other structured signals, while keeping the pretrained backbone largely frozen. This design enables memory-efficient training and supports the integration of multiple knowledge sources without disrupting the original language representations learned during pretraining.

3 Methodology

3.1 RAG-Based Knowledge Augmentation

Our first approach adopts Retrieval-Augmented Generation (RAG) with a domain-specific knowledge graph (KG), where relevant factual information is retrieved at inference time and incorporated into the model input to improve factual grounding.

The domain-specific KG is constructed through the following steps:

1. Data Collection: We scrape data from a curated set of psychology Wikipedia articles.
2. Triple Extraction: We use REBEL-large (Huguet Cabot and Navigli, 2021) to extract subject–relation–object (SRO) triples from the text, followed by a light post-processing step to clean and filter noisy or low-quality triples.
3. Knowledge Storage: The extracted triples (head, relation, and tail) are organized in a graph-structured format.

Then, for a given input prompt, we implement RAG as follows:

1. Tokenization: The prompt is first tokenized to extract keywords.
2. Knowledge Retrieval: Relevant SRO triples are retrieved from the knowledge graph using a BM25-based retriever.
3. Response Generation: The top retrieved triples are appended to the original prompt as additional context. The augmented prompt is then passed to the language model to generate the final response.

3.2 Adapter-Based Knowledge Integration (K-Adapter)

In addition to retrieval-based augmentation, we implement the K-Adapter architecture as an alternative method that directly injects structured knowledge into a pretrained large language model. Following the original K-Adapter methodology (Wang et al., 2021), we use a domain-adaptive BERT-base model (BERT-DAPT) as the shared backbone, which is also used in the RAG setting for a controlled comparison. The model includes two types of adapters: a knowledge graph adapter trained with supervision from knowledge graph triples, and a text adapter trained using masked language modeling (MLM).

Training is performed in two stages. In the first stage, the backbone model is frozen and only the adapter parameters are updated using a joint objective that combines KG-based prediction and MLM. This stage injects structured knowledge into the adapters while preserving the backbone representations. In the second stage, both the backbone and adapters are further trained using standard MLM on the domain corpus to better align the injected knowledge with natural language usage.

3.3 Baseline

Our baseline for comparison is a BERT-base model without any knowledge graph assistance. To account for domain-specific language usage, we perform domain-adaptive pretraining (DAPT) on BERT-base using the crawled Wikipedia corpus (after cleaning) with an MLM objective, resulting in a BERT-DAPT model. This model captures domain-specific contextual semantics and terminology and serves as the shared backbone for both the RAG and K-Adapter approaches.

The original BERT-base model is retained as a text-only baseline. Comparing BERT-base, BERT-DAPT, and their knowledge-enhanced variants allows us to evaluate whether augmenting a domain-adapted language model with structured knowledge provides additional benefits beyond learning domain-specific semantics from text alone.

4 Experiments

4.1 Data

Our dataset consists of 1,243 Wikipedia articles collected from three closely related subcategories: decision-making, heuristics, and cognitive biases. The articles are crawled via a Python-based interface to the Wikipedia API, following the hierarchical organization of Wikipedia categories. This hierarchical structure helps maintain strong domain specificity and topical coherence across the corpus. Table 1 presents a selection of Wikipedia articles included in our dataset.

The collected articles are used to construct a domain-specific knowledge graph (KG), which serves as a shared knowledge source for both RAG-based retrieval and K-Adapter knowledge injection. In addition, the factual recall evaluation is derived from the same KG, ensuring consistency between knowledge construction and downstream assessment.

4.2 Evaluation Method

We design a cloze-style factual recall test to evaluate the ability of different models to recall domain-specific factual knowledge. This evaluation setup leverages the fact that all models are BERT-based, where prediction fundamentally relies on a masked language modeling objective.

The factual recall test is conducted as follows:

1. Question Construction: We use GPT-4 to generate 200 fill-in-the-blank questions from

Cognitive bias	Confirmation bias	Availability heuristic	Attribution bias
Bandwagon effect	Belief bias	Authority bias	Attentional bias
Bias blind spot	Choice-supportive bias	Congruence bias	Actor–observer asymmetry
Attribute substitution	Anchoring effect	Automation bias	Affinity bias
Framing effect	Hindsight bias	Overconfidence bias	Status quo bias

Table 1: Assortment of some of the psychology Wikipedia articles used in our dataset.

high-quality triples extracted from the constructed knowledge graph. For each question, the tail entity is replaced with a [MASK] token, and the relation is paraphrased into a natural-language form to reduce reliance on surface patterns.

2. Prediction: Models are asked to predict the masked token, and the full probability distribution over the vocabulary is examined.

3. Ranking-Based Evaluation: Model performance is evaluated based on the rank of the ground-truth entity in the predicted probability distribution.

4. Metrics: We report Absolute Accuracy, Mean Reciprocal Rank (MRR), and Average Cosine Similarity.

Absolute Accuracy measures the percentage of queries where the ground-truth word is exactly returned. Mean Reciprocal Rank (MRR) measures the average of the reciprocal ranks of the ground-truth word in the probability distribution for each query. Average Cosine Similarity measures the average of the cosine similarities between the returned word vector and the ground-truth word vector for all queries. These three metrics are defined below for evaluation on N queries, where r_i and g_i denote the predicted word and the ground-truth word for query i, respectively:

$$\text{Accuracy} = \frac{\text{correct predictions}}{N} \tag{1}$$

$$\text{MRR} = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{\text{rank}_i} \tag{2}$$

$$\text{Cosine Similarity} = \frac{1}{N}\sum_{i=1}^{N}\frac{\mathbf{p}_i \cdot \mathbf{g}_i}{|\mathbf{p}_i||\mathbf{g}_i|} \tag{3}$$

4.3 Experimental Details

All experiments adopt **BERT-base** as the backbone model. We first perform *domain-adaptive pretraining* (DAPT) by further training BERT-base on the cleaned domain corpus using a masked language modeling (MLM) objective. This step adapts the model to domain-specific language usage and terminology. The pretraining is conducted for three epochs on NVIDIA H200 GPUs, resulting in a BERT-DAPT model that serves as the shared backbone for both the RAG and K-Adapter approaches.

K-Adapter training follows a two-stage procedure. In the first stage, the BERT-DAPT backbone is frozen and only the adapter parameters are updated. Two adapters are trained jointly: a knowledge adapter, supervised by structured knowledge graph triples, and a text adapter, trained with MLM on verbalized sentences derived from those triples. The training objective jointly optimizes knowledge prediction and language modeling, defined as:

$$L = \lambda_{MLM} \cdot L_{MLM} + \lambda_{KG} \cdot L_{KG} \tag{4}$$

where L_{MLM} denotes the masked language modeling loss and L_{KG} represents the knowledge graph prediction loss. We set $\lambda_{\text{MLM}} = 1.0$ and $\lambda_{\text{KG}} = 6.0$ to emphasize structured knowledge supervision during this stage.

In the second stage, both the backbone and adapter parameters are unfrozen and further trained using standard MLM on the domain corpus. This stage aims to better align the knowledge encoded in the adapters with natural language usage in real text, stabilizing the adapter representations and improving their integration with the backbone model. Both training stages are conducted on NVIDIA H200 GPUs.

For retrieval-based augmentation, we implement RAG with a BM25 retriever. For each query, BM25 retrieves the top-3 most relevant textual facts from the knowledge graph, which are appended to the input as additional context before being passed to the BERT-DAPT model for prediction.

Model	Accuracy	MRR	Cosine Similarity
BERT-base	0.23	0.309	0.024
BERT-DAPT	0.26	0.332	0.026
BERT-RAG (BM25 + Top-3)	**0.52**	**0.569**	**0.054**
K-Adapter	0.37	0.436	0.021

Table 2: Factual recall performance across different models on the full dataset of 1,243 Wikipedia articles.

Model	Accuracy	MRR	Cosine Similarity
BERT-base	0.20	0.273	0.025
BERT-DAPT	0.22	0.288	0.026
BERT-RAG (BM25 + Top-3)	**0.484**	**0.533**	**0.041**
K-Adapter	0.225	0.301	0.026

Table 3: Factual recall performance across different models on an initial experiment using a smaller corpus of 160 Wikipedia articles and 100 factual recall questions.

4.4 Results

Table 2 summarizes the results of the factual recall evaluation across different models. Both knowledge-enhanced approaches outperform the baselines, with BERT-RAG achieving the strongest performance across all metrics. The K-Adapter model also improves over BERT and BERT-DAPT, though the gains are more modest compared to retrieval-based augmentation.

Table 3 reports results from an initial experiment conducted using a smaller corpus of 160 Wikipedia articles and a factual recall test consisting of 100 questions, which motivated the expansion of the dataset and further refinement of the knowledge graph in subsequent experiments.

5 Analysis and Discussion

The experimental results show that incorporating external knowledge substantially improves factual recall performance compared to models trained solely on text. Both retrieval-based and adapter-based knowledge enhancement strategies outperform the BERT-base and BERT-DAPT baselines, indicating that explicit access to structured knowledge helps reduce hallucination, particularly in scenarios requiring precise factual grounding.

Among the evaluated approaches, RAG demonstrates the most stable behavior and achieves the strongest performance across all metrics. By explicitly retrieving relevant facts from the knowledge graph and providing them as additional context, RAG supplies concrete evidence to the model during prediction. This makes retrieval-based augmentation an effective and reliable strategy for domain-specific factual recall, especially under limited data conditions.

The performance of the K-Adapter model improves after increasing the training data size and refining the constructed knowledge graph, suggesting that adapter-based knowledge injection is sensitive to both data scale and knowledge quality. Given the strong performance of K-Adapter reported in the original work (Wang et al., 2021), this trend indicates that K-Adapter has the potential to benefit from larger corpora, stronger supervision, or improved training objectives.

In contrast, domain-adaptive pretraining (DAPT) yields only marginal gains over the original BERT-base model. One possible reason is the mismatch between the DAPT objective and the evaluation task. While DAPT relies on masked language modeling with randomly masked tokens, the factual recall evaluation specifically masks knowledge entities derived from the knowledge graph. As a result, although both involve MLM, they emphasize different types of information, which may limit the effectiveness of text-only domain adaptation for improving factual recall.

Overall, these findings suggest that knowledge enhancement is an effective strategy for improving factual recall and mitigating hallucination. Retrieval-based methods provide a more direct and robust solution under constrained data settings, while adapter-based approaches remain a promising direction for future work, particularly with larger datasets, stronger supervision, or hybrid retrieval–adaptor designs.

6 Conclusion

In this project, we studied knowledge enhancement strategies for improving domain-specific factual recall and reducing hallucination in large language models. We compared retrieval-based augmentation with RAG and adapter-based knowledge injection with K-Adapter under a unified experimental setting. Our results show that incorporating structured knowledge leads to clear improvements over text-only baselines, with retrieval-based methods providing the most stable and effective performance, while adapter-based approaches show promising gains as data scale increases.

The primary limitation of this study is that we do not evaluate other existing methods for knowledge enhancement. Our initial study design included embedding-based models such as LUKE; however, we found that this class of models faces practical challenges in domain adaptation. Specifically, they rely on a fixed entity vocabulary and an entity linker, which makes adaptation difficult when domain-specific entities are not present in the original knowledge graph (Yamada et al., 2020).. This limitation prevents effective learning in new domains and motivated our focus on RAG and K-Adapter instead. At the same time, this observation highlights the flexibility of K-Adapter, which learns structured knowledge through adapter modules without modifying the backbone model.

Acknowledgments

We thank Northeastern University for providing access to high-performance computing resources, which were essential for conducting the experiments in this study. We also thank Professor Feifei Pan for her valuable feedback and guidance.

References

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics*, pages 4171–4186.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Ming-Wei Chang. 2020. Retrieval-augmented language model pre-training. In *Proceedings of the 37th International Conference on Machine Learning*.

Pere-Lluís Huguet Cabot and Roberto Navigli. 2021. Rebel: Relation extraction by end-to-end language generation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2370–2381.

Nikhil Kandpal, Hao Deng, Adam Roberts, Eric Wallace, and Colin Raffel. 2023. Large language models memorize, but do not generalize: The case of factual knowledge. *arXiv preprint arXiv:2309.12288*.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, and Sebastian Riedel. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, pages 5296–5305.

Robert Logan, Eric Wallace, Matt Gardner, Sameer Singh, and Hannaneh Hajishirzi. 2019. Barack's wife hillary: Using knowledge graphs for fact-aware language modeling. In *Proceedings of ACL*.

Matthew E. Peters, Mark Neumann, Robert Logan, Roy Schwartz, Vidur Joshi, Sameer Singh, and Noah A. Smith. 2019. Knowledge enhanced contextual word representations. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, pages 43–54.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pages 6000–6010.

Ruize Wang, Wenhao Yu, Changyou Zhu, Tong Zhang, and Heng Ji. 2021. K-adapter: Infusing knowledge into pre-trained models with adapters. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics*, pages 567–578.

Ikuya Yamada, Akari Asai, Hiroyuki Shindo, Hideaki Takeda, and Yuji Matsumoto. 2020. Luke: Deep contextualized entity representations with entity-aware self-attention. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 6442–6454.