Knowledge Injection to Counter Large Language Model (LLM) Hallucination*

Ariana Martino[0009−0005−0747−8984], Michael Iannelli[0000−0002−5967−2026], and Coleen Truong[0009−0009−7317−7492]

Yext, New York NY 10011, USA {amartino,miannelli,ctruong}@yext.com
https://www.yext.com/

Abstract. A shortfall of Large Language Model (LLM) content generation is hallucination, i.e., including false information in the output. This is especially risky for enterprise use cases that require reliable, fact-based, controllable text generation at scale. To mitigate this, we utilize a technique called Knowledge Injection (KI), where contextual data about the entities relevant to a text-generation task is mapped from a knowledge graph to text space for inclusion in an LLM prompt. Using the task of responding to online customer reviews of retail locations as an example, we have found that KI increases the count of correct assertions included in generated text. In a qualitative review, fine-tuned bloom-560m with KI outperformed a non-fine-tuned text-davinci-003 model from OpenAI, though text-davinci-003 has 300 times more parameters. Thus, the KI method can increase enterprise users' confidence leveraging LLMs to replace tedious manual text generation and enable better performance from smaller, cheaper models.

Keywords: large language model · knowledge graph · prompt engineering · hallucination · bloom · gpt-3.

1 Introduction

One limitation of Large Language Model (LLM) content generation is hallucination, or false assertions in the generated text [2]. Enterprise use cases require reliable, fact-based text generation at scale, making investment into LLM-generated text risky. To mitigate hallucination, we utilize a technique called Knowledge Injection (KI) where contextual data about entities relevant to a task is mapped from a knowledge graph to text space for inclusion in an LLM prompt. In our use case of responding to online customer reviews of retail locations, KI increases the rate at which assertions are correct while improving overall text quality.

While LLM parameters encode knowledge [7], they are still susceptible to hallucination because: (1) not all current data can be present during training of the model (e.g., updates to business information made post-training) and (2) it is difficult to encode all knowledge into the model's parameters [6].

* All work in this paper was supported by and conducted at Yext.

KI begins with a knowledge graph that includes the entity relevant to the task and connections to other entities from which context can be derived. KI aims to generate controllable text with business information from a knowledge graph that is not general knowledge (e.g., the business' phone number will not likely be common knowledge that the LLM knows from base training). Controllable Text Generation (CTG) is subject to controlled constraints such as sentiment or, in our use case, alignment with source-of-truth business information [8].

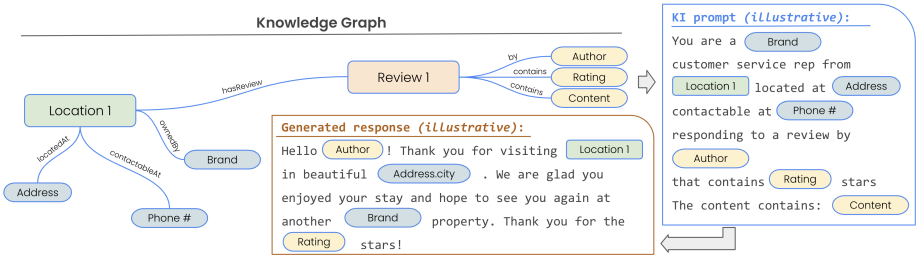

Fig. 1. A templated text prompt with KI is compiled by navigating the entity's neighborhood and inserting relevant contextual fields. In this example, the KI prompt requests the model generate a text response to an online customer review based on the relevant review, location, and brand entities. In contrast, a review-only prompt would contain only the yellow fields (author, rating, and content).

Text fields from the knowledge graph are inserted into a templated prompt to map the graph-based context to text space, forming the input to the LLM. This is demonstrated in Fig. 1, where an LLM-generated response to an online customer review is requested. The relevant entity, Review 1, and its neighbors, e.g., Location 1, in the knowledge graph are mapped to a templated prompt.

2 Problem Setup and Experiments

2.1 Hallucination

We set out to determine if KI reduces hallucination in LLM-generated responses to online customer reviews. LLMs using bloom-560m [4] were fine-tuned using reviews and responses written by human customer service agents. Generated responses from a review-only model fine-tuned with only information from the review (i.e., author, rating, and content) vs. a KI-prompted model fine-tuned with added context about the linked entities were evaluated. The models were fine-tuned on a dataset of ∼35K review-response pairs.

Domain experts counted correct and incorrect assertions in each generated response. Assertions included specification of a location name, contactable at phone number or web address, owned by brand name, and located at location address. Incorrect (i.e., hallucinated) assertions contained untrue information

contradicted by the knowledge graph, like directing customers to call a fictitious phone number. Factual assertions were those not otherwise marked as incorrect.

2.2 Generated Response Quality

In addition to testing KI's impact on hallucination, we also tested its impact on overall quality of generated review responses. Subject matter experts graded generated responses from non-KI prompted OpenAI's text-davinci-003 text generation model, aka GPT-3 [1], and KI prompted bloom-560m on the overall quality based on a 3-point scale (Table 1).

Table 1. Scoring rubric used in qualitative response quality analysis

Score	Quality	Criteria
1	Bad	Unusable generated response with potential negative business brand reputation impact.
2	Good	Usable generated response with potential for human-intervention to refine using business brand standards.
3	Great	Usable generated response with minimal to no requirement for human- intervention and aligns with business brand standards.

3 Results and Discussion

3.1 Hallucination

The KI increased the count of correct assertions while decreasing the count of incorrect assertions (Table 2), suggesting it is useful for enterprise tasks like review response, which are manual and costly when done by humans, but require factual context about the business to produce trustworthy generated text.

Table 2. Assertions in generated text from review-only vs. KI LLMs (bloom-560m)

Avg. # of assertions per inference	Review-only prompt *n = 64*	KI prompt *n = 78*	Δ
Correct	**0.61**	**1.86**	**+205%**
Incorrect	0.23	0.19	−18%
Total	0.84	2.05	+143%

3.2 Generated Response Quality

The KI model received higher quality scores for generated responses, suggesting KI is useful for helping models to align with business brand standards (Table 3). Though text-davinci-003 has ∼300 times as many parameters as bloom-560m, the smaller model fine-tuned with KI outperformed the larger OpenAI model. Thus, fine-tuning with KI could help businesses save on cost by training and hosting a smaller model while producing higher quality generated responses [5]. Furthermore, using smaller models could improve inference speed [3].

Table 3. Quality of Generated Responses

Model	**Params.**	**Avg. Score**
OpenAI text-davinci-003 *(n=94)*	175b	1.80
bloom-560m fine-tuned with KI *(n=94)*	**0.56b**	**2.14**

4 Conclusions and Future Work

Experiments on both hallucination and generated response quality highlighted how KI can help businesses generate more reliable, fact-based, and higher quality text from LLMs. In order to take advantage of this, businesses would require a factual and robust knowledge graph of entities relevant to their business, like locations, reviews, products, documents, etc.

To help mitigate this limitation, in future experimentation, we intend to continue researching methods to build out robust knowledge graphs for businesses through entity and edge extraction leveraging LLMs.

References

1. Brown, T.B., et al.: Language models are few-shot learners. CoRR **abs/2005.14165** (2020), https://arxiv.org/abs/2005.14165
2. Ji, Z., et al.: Survey of hallucination in natural language generation. ACM Computing Surveys **55**(12), 1–38 (mar 2023). https://doi.org/10.1145/3571730
3. Menghani, G.: Efficient deep learning: A survey on making deep learning models smaller, faster, and better. ACM Computing Surveys **55**(12), 1–37 (2023)
4. Scao, T.L., et al.: Bloom: A 176b-parameter open-access multilingual language model (2022). https://doi.org/10.48550/ARXIV.2211.05100
5. Sharir, O., Peleg, B., Shoham, Y.: The cost of training nlp models: A concise overview. arXiv preprint arXiv:2004.08900 (2020)
6. Singhal, K., et al.: Large language models encode clinical knowledge (2022)
7. Wang, C., Liu, X., Song, D.: Language models are open knowledge graphs (2020), https://arxiv.org/abs/2010.11967
8. Zhang, H., et al.: A survey of controllable text generation using transformer-based pre-trained language models. ArXiv **abs/2201.05337** (2022)