Infusing Knowledge into Large Language Models with Contextual Prompts

Kinshuk Vasisht
University of Delhi
New Delhi, India
kinshuk.mcs21@cs.du.ac.in

Balaji Ganesan
IBM Research India
Bengaluru, India
bganesa1@in.ibm.com

Vikas Kumar
University of Delhi
New Delhi, India
vikas@cs.du.ac.in

Vasudha Bhatnagar
University of Delhi
New Delhi, India
vbhatnagar@cs.du.ac.in

Abstract

Knowledge infusion is a promising method for enhancing Large Language Models for domain-specific NLP tasks rather than pre-training models over large data from scratch. These augmented LLMs typically depend on additional pre-training or knowledge prompts from an existing knowledge graph, which is impractical in many applications. In contrast, knowledge infusion directly from relevant documents is more generalisable and alleviates the need for structured knowledge graphs while also being useful for entities that are usually not found in any knowledge graph. With this motivation, we propose a simple yet generalisable approach for knowledge infusion by generating prompts from the context in the input text. Our experiments show the effectiveness of our approach which we evaluate by probing the fine-tuned LLMs.

1 Introduction

Unifying Large Language Models (LLMs) and Knowledge Graphs (KGs) is an active area of research for several reasons. Pan et al. (2023) presents a survey of different approaches, including knowledge infusion into Large Language Models from knowledge bases. Common techniques for infusing knowledge involve pre-training over a factually-rich corpus prepared from structured knowledge bases or learning soft knowledge prompts pertaining to entities using factual triples from a knowledge base for improving entity-specific inference (dos Santos et al., 2022).

Infusing knowledge directly into the model from knowledge bases, though more efficient than re-training LLMs from scratch, is unrealistic in many real-world applications. Maintaining knowledge graphs about customers, organizations and events mentioned in documents is cumbersome and incurs an overhead of maintaining privacy. Further, all entities are not equal in knowledge bases, with

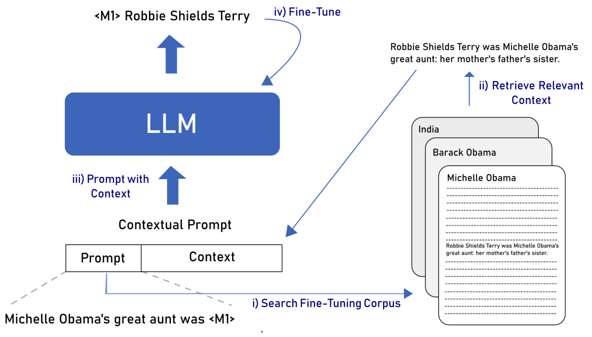

Figure 1: Contextual prompts to infuse knowledge about entities into Large Language Models

some entities being highly under-represented. This degrades the performance over downstream tasks for such entities, post-knowledge infusion, due to lack of sufficient knowledge. Such entities may be more frequent in domain-specific corpora, which can provide the relevant context for knowledge infusion. As an example, while entities such as *Barack Obama* and *Michelle Obama* are well known and represented in structured knowledge bases such as Wikidata, entities such as *Michelle Obama*'s great aunt *Robbie Shields Terry*, being relatively lesser known, exist only in the Wikipedia page of the subject and not as an entity in Wikidata.

Motivated by the above, we propose to exploit contextual text from a relevant domain-specific corpus to infuse knowledge into Large Language Models. Infusing knowledge directly from documents without having to create knowledge graphs is not only efficient, but also more general. Figure 1 outlines the four steps of our proposed approach. Given an input prompt composed of a task instance and accompanying instructions, we retrieve relevant context from an indexed corpus by identifying sentences mentioning involved entities. Knowledge is then infused into a pre-trained model by fine-tuning over input prompts augmented with the identified context. The knowledge-infused model obtained after fine-tuning can be leveraged for knowledge-intensive downstream tasks. We compare and contrast our knowledge-infusion method against other infusion techniques involving prompting with factual triples (Moiseev et al., 2022a) and natural sentences from triples (Agarwal et al., 2021).

Our approach offers significant advantages over existing knowledge infusion methods. The proposal alleviates the need for structured knowledge sources by relying solely on domain corpora. Using fine-tuning makes our approach simple, scalable, and employable in low-resource settings where excessive compute resources and data sources are unavailable. The method permits seamless integration of structured knowledge graphs, if available, to further enhance the identification of relevant context from corpora via entity linking and disambiguation. Our method is also extensible to other modalities like tabular data and graphs.

2 Related Works

Petroni et al. (2019) introduced the idea of language models being knowledge bases. Since then there have been continuous efforts to both extract facts from LLMs and also to infuse language models with facts from knowledge bases.

Agarwal et al. (2020) generated natural language sentences from triples and additionally pre-trained large language models with the generated sentences. Moiseev et al. (2022b) directly infuse triples into LLMs without generating sentences from the triples. Agarwal et al. (2023) infuse triples from domain-specific knowledge graphs into T5 models. These approaches show the interchangeable nature of knowledge infused from both triples and sentences containing the triples.

dos Santos et al. (2022) proposed *Knowledge Prompts*, where prompts are learnt for the most frequently occurring entities in Wikidata. For every triple, a prompt is initialised with random values and then updated via gradient descent on the triples mask prediction task. Other methods (Wang et al., 2021; Diao et al., 2023) explore the use of adapters for knowledge infusion with parameter efficient fine-tuning. De Cao et al. (2021) & Zhong et al. (2023) discuss related ideas in knowledge editing.

Existing approaches assume the existence of a well-populated KG, and hence suffer from limitations concerning practicality in a real-world setting. For example, a new customer entity, a new product or new terms in a news article or a court judgment may not exist in the KG.

3 Knowledge Infusion with context

Our motivation is to create a knowledge-prompting approach that draws on documents and large language models rather than only KGs. Accordingly, our approach leverages relevant context retrieved from a domain-specific fine-tuning corpus to infuse knowledge into a pre-trained language model. This is in contrast to approaches utilizing factual triples (Moiseev et al., 2022a) or corresponding natural sentences (Agarwal et al., 2021) from knowledge bases, or soft knowledge prompts prepared in an entity-specific manner (dos Santos et al., 2022) as the source for knowledge to be infused.

We prepare a pre-trained model by infusing knowledge using full fine-tuning over a specific downstream task, such as tail prediction in triples, question answering, or translation. For this purpose, along with a suitable dataset comprising of downstream task instances, we also identify a domain-specific fine-tuning corpus composed of relevant documents that contain information pertaining to involved entities, which serves as our source of knowledge. Primarily, we formulate instances from a training dataset into prompts, by prepending brief task-specific instructions alongside the instance data. For example, for a tail-prediction task, alongside a factual triple without the tail entity, an instruction describing the task of tail-prediction is prepended. Next, for a given prompt, we identify named entities present in the prompt and retrieve relevant context from the domain-specific fine-tuning corpus. This context is composed of useful information that could provide knowledge about the involved entities to aid in enhancing performance for the task. We infuse knowledge from the context surrounding the entity by augmenting the context alongside the text prompt. For each entity, our contextual information is the phrase, sentence, or paragraph surrounding the entity in a document index. Since we are doing full fine-tuning, the model parameters are updated, and the error is propagated based on the task.

During inference, task-specific prompts comprising of task instructions and the query to process are given, for which responses are generated based on the knowledge infused. Unlike the fine-tuning

Dataset	Model	Hits@1 ↑	Hits@5 ↑	Hits@10 ↑	AED ↓	MRR ↑
KELM-TEKGEN	google/flan-t5-small	0.019	0.036	0.045	18.75	0.024
	google/flan-t5-base	0.047	0.063	0.095	85.5	0.055
	google/flan-t5-large	0.082	0.102	0.138	109.5	0.088
	flan-t5-small-fine-tuned	0.528	0.535	0.541	96.75	0.538
	flan-t5-base-fine-tuned	0.514	0.520	0.539	83.25	0.525
	flan-t5-small-fine-tuned-w-context	0.800	0.801	0.804	2.75	0.805
	flan-t5-base-fine-tuned-w-context	**0.825**	**0.825**	**0.833**	**0.75**	**0.827**
TACRED	google/flan-t5-small	0.004	0.006	0.006	84.75	0.005
	google/flan-t5-base	0.004	0.014	0.018	9.75	0.008
	google/flan-t5-large	0.034	0.044	0.060	22.50	0.039
	flan-t5-small-fine-tuned	0.366	0.368	0.39	50.25	0.376
	flan-t5-small-fine-tuned-w-context	0.782	0.782	0.784	**3.75**	0.788
	flan-t5-base-fine-tuned-w-context	**0.818**	**0.820**	**0.824**	5.25	**0.823**
Re-TACRED	google/flan-t5-small	0.000	0.010	0.016	66.00	0.005
	google/flan-t5-base	0.006	0.016	0.028	28.50	0.010
	google/flan-t5-large	0.052	0.070	0.084	5.25	0.060
	flan-t5-small-fine-tuned	0.352	0.366	0.406	15.75	0.370
	flan-t5-small-fine-tuned-w-context	0.798	0.798	0.800	6.00	0.805
	flan-t5-base-fine-tuned-w-context	**0.846**	**0.846**	**0.850**	**0.00**	**0.852**

Table 1: Flan-T5 performance on relation prediction task on KELM-TEKGEN, TACRED and Re-TACRED datasets.

stage, relevant context is not retrieved and utilized for inference.

As an example, for the tail and link prediction task, a dataset for fine-tuning comprising triples is identified, alongside which we identify a relevant domain-specific corpus for the source of knowledge. We formulate the factual triple as a prompt for the LLM, post-masking one of the entities or the relation at random. For the unmasked entities and relation, we retrieve relevant context from the corpus and append it to the generated input prompt to obtain the contextual prompt finally used to fine-tune the LLM. Post fine-tuning, at inference the model is prompted with factual triples formulated as a prompt in the same format but without retrieved context.

Unlike the approach in dos Santos et al. (2022), neither entity linking via an external knowledge graph nor inference based on information/prompts derived from structured knowledge bases is performed. Further, the prompt formulation contrasts from Agarwal et al. (2022) and Saxena et al. (2022), where for entity and link prediction explicit mentions of the masked positions is present and additional context is not leveraged during fine-tuning.

In this work, we directly utilize retrieved-context alongside the input prompt, which is comprised of tokens and is therefore discrete in nature as opposed to soft prompts in a continuous latent space. To include larger amounts of information and bypass the context length restrictions of LLMs, contextual prompts can be prepared based on context embedded as soft prompts, allowing for further improvements to efficiency and performance. Further, for tasks where structured knowledge sources or information in other modalities, such as tabular data is available, our approach could be extended to leverage the information from these sources and create appropriate contextual prompts for knowledge infusion.

4 Experiments

We conduct experiments with pre-trained Flan-T5 models (Chung et al., 2022) to demonstrate the role of contextual text as discrete text prompts. We fine-tune the model on two downstream tasks, tail prediction in triples and question answering. Tail prediction, as the name indicates, predicts the object in subject, predicate, and object triples.

Our experiments leverage the following variants of Flan-T5 models: Flan-T5-Small with 80M parameters and Flan-T5-base with 250M parameters. Fine-tuning and inference were performed in FP-16 mixed-precision mode (Micikevicius et al., 2017) over the task-specific datasets using a single Nvidia A4000 GPU with 16GBs of VRAM available. Additionally, gradient accumulation and gradient check-pointing (Chen et al., 2016) techniques were employed to reduce memory consumption.

Relation Prediction

We perform the relation extraction experiment on four datasets. KELM-TEKGEN dataset from Agarwal et al. (2020) has over 15 million Wikidata triples and sentences generated with those triples, from which we leverage a sample of 1 million triples partitioned in a ratio of 9:1 for fine-tuning and evaluation. TACRED Zhang et al. (2017) is a well-known relation extraction dataset comprising of about 60,000 factual triples with associated text. Our experiments leverage the subset of 18,000 triples for which relations are known paritioned as before for fine-tuning and evaluation. Re-TACRED is a variant of TACRED introduced by Stoica et al. (2021) with a larger number of triples with known relations. We compare the results using metrics commonly utilized for entity and link prediction in Knowledge Graph literature, notably Mean Reciprocal Rank (MRR) and Hits@K (Voorhees and Tice (2000), Ali et al. (2022)) with varying values of K: 1,5 and 10. We also use an approximation of the Graph Edit Distance metric as in Swamy et al. (2021) for comparing the generated knowledge graphs against the ground truth, abbreviated as AED.

Table 1 summarizes the results of our experiments. We observe that all the fine-tuning methods perform better the base Flan-T5 models. Although we observe this across different sizes of Flan-T5 models we compare against (small, base and large), we believe the knowledge available to the model is more relevant in these experiments rather than the size of the models. This is discussed in Agarwal et al. (2022), which investigates the role of model size on the model's capacity to retain knowledge.

Among the fine-tuning methods, our introduction of contextual text as discrete text prompts significantly improves the performance across all datasets and metrics.

Soft Prompts

While the use of context significantly improves performance, to justify the efficacy of the method towards knowledge infusion, notably building of knowledge regarding entities present in the text and/or triples, we conduct additional investigatory experiments evaluating entity knowledge developed by the model as a consequence of the infusion process. We learn differentiable 'contextual' prompts termed as soft prompts, specific to each entity of the training corpus to aid inference in downstream tasks. We compare the performance of adapter models utilizing these contextual prompts over tail prediction and relation extraction against full fine-tuning with and without context in Table 2.

Model	Hits@1	AED	MRR
google/flan-t5-base	0.050	11.250	0.056
flan-t5-base-fine-tuned	0.441	17.250	0.452
flan-t5-base-fine-tuned-w-context	0.710	10.500	0.717
flan-t5-base-fine-tuned-soft-prompt	0.176	6.750	0.174

Table 2: Flan-T5 performance on relation prediction on a subset of KELM-TEKGEN test data with entities known from the training phase

Question Answering Task Evaluation

To evaluate our approach on tasks other than relation extraction, we use a closed-book open-domain question-answering task on the Trivia QA dataset Joshi et al. (2017). We use the rc-wikipedia partition of the dataset and evaluate the models fine-tuned over triples from the KELM corpus. The models are fine-tuned again for the question-answering task for a maximum of 10K steps and a batch size of 128.

For each question, we identify entities and perform a lookup in the KELM corpus to retrieve aligned sentences for the entity in question. The retrieved context is then used alongside for prompting. We evaluate the previously selected models after fine-tuning them over the question-answering task as described above using the same settings as described in the previous experiment. Table 3 summarizes the exact match scores observed.

Model	Exact Match (%)
google/flan-t5-small	27.1
flan-t5-small-fine-tuned	16.7
flan-t5-small-fine-tuned-w-context	20.0

Table 3: Flan-T5 performance on question-answering over the TriviaQA dataset

Impact of Variation in Context Size

In order to investigate the role of the context towards knowledge infusion and prediction, we conduct experiments to examine the impact of change in context length (in terms of either a change in the number of sentences or the length in terms of the number of tokens). This provides an insight into whether the models are able to produce satisfactory results only in the presence of a context that directly describes the entity to be predicted or whether the predictions are a result of the understanding and insight provided by the context. Results as summarized in table 4 demonstrate improvements with the use of larger contexts with more information.

Model	Hits@1	AED	MRR
flan-t5-small	0.798	6.00	0.805
flan-t5-small-longer-context	0.836	2.25	0.842
flan-t5-base	0.846	0.00	0.852
flan-t5-base-longer-context	0.914	3.75	0.917

Table 4: Impact of context lengths over entity and relation prediction performance

Legal Corpus

While the experiments in Table 1 are predominantly on news or web knowledge, we conducted a similar experiment on a legal knowledge graph and related documents from Singh Dhani et al. (2021). This is one of the example applications where we expect the domain knowledge graphs to be non-existent or incomplete. As shown in Table 5, our contextual prompts seem to work in this scenario, though the improvement is modest.

Model	Hits@10	AED	MRR
flan-t5-base	0.245	47.25	0.200
flan-t5-base-w-context	0.290	45.00	0.217

Table 5: Flan-T5 performance on entity and link prediction over the Legal KG, with corresponding judgments used as the context source

Limitations

In this work, we have limited our experiments to full fine tuning and examining the role of contextual text in helping LLMs surface the knowledge about entities. We have not compared our method with other knowledge infusion methods like additional pre-training, adapters and soft prompts. Further, an evaluation of knowledge propagation, such as in Onoe et al. (2023) has not been performed, which could be an interesting direction to explore in future research for understanding the extent of knowledge infused in models. Our work is more relevant to applications where an external knowledge graph is unavailable or is not well populated. We show that contextual text can be used in lieu of knowledge graphs in such applications.

Conclusion

We propose an alternative method to infuse knowledge into Large Language Models (LLMs) that does not assume the existence of a Knowledge Graph (KG). We use a search index to provide relevant sentences to be used as context alongside input prompts during fine-tuning for knowledge infusion. Results over relation extraction and tail predictions tasks demonstrate improved extents of knowledge infusion with the use of context.

Acknowledgements

This work was done by the first author as part of his graduate coursework, being a part of the Legal Text Analytics group at the University of Delhi. We thank the Legal Text Analytics group and the University of Delhi for providing this opportunity. The second author participated in this work as part of the Global Remote Mentorship program of IBM. We thank IBM University Relations for their support. We thank the reviewers for their valuable suggestions.

References

Ankush Agarwal, Sakharam Gawade, Sachin Channabasavarajendra, and Pushpak Bhattacharya. 2022. There is no big brother or small brother:knowledge infusion in language models for link prediction and question answering. In *Proceedings of the 19th International Conference on Natural Language Processing (ICON)*, pages 204–211, New Delhi, India. Association for Computational Linguistics.

Ankush Agarwal, Sakharam Gawade, Sachin Channabasavarajendra, and Pushpak Bhattacharyya. 2023. There is no big brother or small brother: Knowledge infusion in language models for link prediction and question answering. *arXiv preprint arXiv:2301.04013*.

Oshin Agarwal, Heming Ge, Siamak Shakeri, and Rami Al-Rfou. 2020. Knowledge graph based synthetic corpus generation for knowledge-enhanced language model pre-training.

Oshin Agarwal, Heming Ge, Siamak Shakeri, and Rami Al-Rfou. 2021. Knowledge graph based synthetic corpus generation for knowledge-enhanced language model pre-training. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3554–3565, Online. Association for Computational Linguistics.

Mehdi Ali, Max Berrendorf, Charles Tapley Hoyt, Laurent Vermue, Mikhail Galkin, Sahand Sharifzadeh, Asja Fischer, Volker Tresp, and Jens Lehmann. 2022. Bringing light into the dark: A large-scale evaluation of knowledge graph embedding models under a unified framework. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(12):8825–8845.

Tianqi Chen, Bing Xu, Chiyuan Zhang, and Carlos Guestrin. 2016. Training deep nets with sublinear memory cost. *CoRR*, abs/1604.06174.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022. Scaling instruction-finetuned language models.

Nicola De Cao, Wilker Aziz, and Ivan Titov. 2021. Editing factual knowledge in language models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6491–6506.

Shizhe Diao, Tianyang Xu, Ruijia Xu, Jiawei Wang, and Tong Zhang. 2023. Mixture-of-domain-adapters: Decoupling and injecting domain knowledge to pre-trained language models memories. *arXiv preprint arXiv:2306.05406*.

Cicero Nogueira dos Santos, Zhe Dong, Daniel Cer, John Nham, Siamak Shakeri, Jianmo Ni, and Yun hsuan Sung. 2022. Knowledge prompts: Injecting world knowledge into language models through soft prompts.

Mandar Joshi, Eunsol Choi, Daniel Weld, and Luke Zettlemoyer. 2017. triviaqa: A Large Scale Distantly Supervised Challenge Dataset for Reading Comprehension. *arXiv e-prints*, page arXiv:1705.03551.

Paulius Micikevicius, Sharan Narang, Jonah Alben, Gregory F. Diamos, Erich Elsen, David García, Boris Ginsburg, Michael Houston, Oleksii Kuchaiev, Ganesh Venkatesh, and Hao Wu. 2017. Mixed precision training. *CoRR*, abs/1710.03740.

Fedor Moiseev, Zhe Dong, Enrique Alfonseca, and Martin Jaggi. 2022a. SKILL: Structured knowledge infusion for large language models. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1581–1588, Seattle, United States. Association for Computational Linguistics.

Fedor Moiseev, Zhe Dong, Enrique Alfonseca, and Martin Jaggi. 2022b. Skill: Structured knowledge infusion for large language models. *arXiv preprint arXiv:2205.08184*.

Yasumasa Onoe, Michael J. Q. Zhang, Shankar Padmanabhan, Greg Durrett, and Eunsol Choi. 2023. Can lms learn new entities from descriptions? challenges in propagating injected knowledge.

Shirui Pan, Linhao Luo, Yufei Wang, Chen Chen, Jiapu Wang, and Xindong Wu. 2023. Unifying large language models and knowledge graphs: A roadmap. *arXiv preprint arXiv:2306.08302*.

Fabio Petroni, Tim Rocktäschel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, Alexander H Miller, and Sebastian Riedel. 2019. Language models as knowledge bases? *arXiv preprint arXiv:1909.01066*.

Apoorv Saxena, Adrian Kochsiek, and Rainer Gemulla. 2022. Sequence-to-sequence knowledge graph completion and question answering. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2814–2828, Dublin, Ireland. Association for Computational Linguistics.

Jaspreet Singh Dhani, Ruchika Bhatt, Balaji Ganesan, Parikshet Sirohi, and Vasudha Bhatnagar. 2021. Similar Cases Recommendation using Legal Knowledge Graphs. *arXiv e-prints*, page arXiv:2107.04771.

George Stoica, Emmanouil Antonios Platanios, and Barnabas Poczos. 2021. Re-TACRED: Addressing shortcomings of the TACRED dataset. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(15):13843–13850.

Vinitra Swamy, Angelika Romanou, and Martin Jaggi. 2021. Interpreting language models through knowledge graph extraction. In *Advances in Neural Information Processing Systems (NeurIPS), 1st Workshop on eXplainable AI Approaches for Debugging and Diagnosis*.

Ellen M. Voorhees and Dawn M. Tice. 2000. The TREC-8 question answering track. In *Proceedings of the Second International Conference on Language Resources and Evaluation (LREC'00)*, Athens, Greece. European Language Resources Association (ELRA).

Ruize Wang, Duyu Tang, Nan Duan, Zhongyu Wei, Xuan-Jing Huang, Jianshu Ji, Guihong Cao, Daxin Jiang, and Ming Zhou. 2021. K-adapter: Infusing knowledge into pre-trained models with adapters. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1405–1418.

Yuhao Zhang, Victor Zhong, Danqi Chen, Gabor Angeli, and Christopher D. Manning. 2017. Position-aware attention and supervised data improve slot filling. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP 2017)*, pages 35–45.

Zexuan Zhong, Zhengxuan Wu, Christopher D Manning, Christopher Potts, and Danqi Chen. 2023. MQuAKE: Assessing knowledge editing in language models via multi-hop questions. *arXiv preprint arXiv:2305.14795*.