There is No Big Brother or Small Brother: Knowledge Infusion in Language Models for Link Prediction and Question Answering

Ankush Agarwal[1*]**, Sakharam Gawade**[1*]**,**
Sachin Channabasavarajendra[2]**, Pushpak Bhattacharyya**[1]
[1]IIT Bombay,
[2]Honeywell Technology Solutions Pvt Ltd
{ankushagrawal, sakharamg, pb}@cse.iitb.ac.in,
sachin.channabasavarajendra@honeywell.com

Abstract

The integration of knowledge graphs with deep learning is thriving in improving the performance of various natural language processing (NLP) tasks. In this paper, we focus on knowledge-infused link prediction and question answering using language models, T5, and BLOOM across three domains: Aviation, Movie, and Web. In this context, we infuse knowledge in large and small language models and study their performance, and find the performance to be similar. For the link prediction task on the Aviation Knowledge Graph, we obtain a 0.2 hits@1 score using T5-small, T5-base, T5-large, and BLOOM. Using template-based scripts, we create a set of 1 million synthetic factoid QA pairs in the aviation domain from National Transportation Safety Board (NTSB) reports. On our curated QA pairs, the three models of T5 achieve a 0.7 hits@1 score. We validate our findings with the paired student t-test and Cohen's kappa scores. For link prediction on Aviation Knowledge Graph using T5-small and T5-large, we obtain a Cohen's kappa score of 0.76, showing substantial agreement between the models. Thus, we infer that small language models perform similar to large language models with the infusion of knowledge.

1 Introduction

A large number of pre-trained language models (LMs) are used for downstream tasks, such as Question Answering (QA). Generally, these language models are trained on generic domain data, such as Web data and News Forums. Recently, LMs are used for downstream tasks in domain-specific fields, namely, healthcare (Michalopoulos et al., 2021), radiology (Kale et al., 2022), and aviation (Agarwal et al., 2022). For tasks such as Information Extraction (IE) and Question Answering (QA), Knowledge Graphs (KGs) are used as a source of external knowledge to boost the performance of models. To a great extent, researchers focus on the synergy of Knowledge Graph and Deep Learning (Miller et al., 2016a; Saxena et al., 2020, 2022). With the increase in data, it is observed that larger models are preferred for different tasks across various domains.

The Large Language Models (LLMs) are preferred to obtain better results than small or non-pre-trained models as they have a vast number of parameters and have been trained on a large amount of data. But, the larger model increases the need for computation power and training time. In this paper, we show that small and large models perform likewise with the infusion of knowledge. We can use non-pre-trained models for different tasks across domains that require less computation power and time and still attain the same performance as pre-trained models.

We validate our hypothesis with the LLMs, *i.e.*, T5 (Raffel et al., 2020) & BLOOM[1]. We perform two tasks: a) Link Prediction, and b) Question Answering on different datasets: a) Aviation Knowledge Graph (AviationKG) (Agarwal et al., 2022), and Aviation QA pairs (section 4.4), b) Movie Knowledge Base (MovieKB) & MetaQA (a set of QA pairs), both present in the MetaQA dataset (Zhang et al., 2018), and c) Complex Web Questions (CWQ) (Talmor and Berant, 2018), which uses subsets of Freebase (Chah, 2017). We perform hypothesis testing to validate our hypothesis. We use paired Student T-test and attempt to reject our hypothesis that models have a negligible difference in performance. But, we were not able to repudiate our hypothesis. To strengthen our findings, we use Cohen's kappa measure and show significant agreement between models.

Our **contributions** are as follows:

---

*Equal contribution

[1]https://huggingface.co/bigscience/bloom

1. We create a synthetic dataset, AviationQA [2], a set of 1 million factoid QA pairs from 12,000 National Transportation Safety Board (NTSB) reports using templates explained in section 4.4. These QA pairs contain questions such that answers to them are entities occurring in the AviationKG (Agarwal et al., 2022). AviationQA will be helpful to researchers in finding insights into aircraft accidents and their prevention.

2. We show that the size of a language model is inconsequential when knowledge is infused from the knowledge graphs. With AviationKG, we obtain 0.22, 0.23, and 0.23 hits@1 scores for link prediction using T5-small, T5-base, and T5-large, respectively. On AviationQA, we get a 0.70 hits@1 score on the three sizes of the T5 model. We validate our hypothesis with paired student t-test, and Cohen's kappa explained in section 6. We obtain a substantial Cohen's kappa score of 0.76 for link prediction on AviationKG using T5-small and T5-large. For Question Answering using T5-small and T5-large, we get a Cohen's kappa score of 0.53 on the MetaQA dataset. Hence, we provide evidence that we can substitute larger models with smaller ones and achieve the same performance with less computational cost and power.

2 Motivation

As stated earlier, in Section 1, LMs are trained on generic datasets. So, knowledge from different sources, *i.e.*, KGs, are used to perform downstream tasks in specific domain areas. LLMs infused with knowledge are required to perform such tasks, namely, QA and link prediction, which increases the need for computation power and time. We show that computational resources can be saved by using smaller language models for tasks.

It is rare to obtain datasets related to the aviation domain, which is in increased demand. We scrape NTSB reports from NTSB's website [3] and create QA pairs that can be used by the aviation industry and researchers for Information Retrieval (IR) and QA purposes. The created dataset will help find insights into aircraft accidents and develop solutions

to prevent accidents.

3 Background & Related Work

A Knowledge Graph is a collection of entities and relations represented in the form of triplets (subject, relation, object). Querying KG in Natural Language (NL) is a long-standing work. Early work focused on rule-based and pattern-based systems (Affolter et al., 2019). Recently, the work is shifted to seq2seq architecture (Zhong et al., 2017) and pre-trained models with the advent of neural networks. Querying KGs remains a challenge because of the conversion of NL to the graph query language, namely, SPARQL, Cypher, etc.

With the value increase of knowledge in the world, the popularity of the KG has escalated. Researchers are keenly interested in the synergy of knowledge graphs and deep learning. Several methods are exploited considering synergy: a) Integrating triplets of KG into the neural network (Liu et al., 2020; Saxena et al., 2022), b) Computing the relevance of entity and relations in a KG using a neural network (Sun et al., 2019; Yasunaga et al., 2021).

Deep Learning models use representations of entities and relations to integrate triplets of KG. Knowledge Graph Embeddings are widely used to obtain representations (Dai et al., 2020). The KG embedding models are trained on link prediction over triplets to obtain representations (Wang et al., 2021). Recent work has focused on using fine-tuned language models over KGE models for link prediction to reduce the number of parameters required to obtain the representations (Saxena et al., 2022).

LMs and KGs are extensively used to improve task-specific performance. Still, no study has been done to understand the characteristics of a language model during the synergy of KG and DL. In this paper, we observe the behavior of language models after knowledge infusion with different domain datasets.

4 Methodology and Experimental Design

This section presents our approach (flow diagram in figure 1), discusses the experiment datasets, creation of AviationQA, describes the model configurations, and explains the evaluation technique.

4.1 Approach

We observe the performance of small and large language models with the infusion of knowledge for link prediction and QA. Experiments are performed with the following models (detailed in section 4.6): a) T5-small non-pre-trained, b) T5-base pre-trained, c) T5-large pre-trained, and SOTA d) BLOOM 1b7. We make use of different domain datasets for our approach, explained in section 4.2. Figure 1 demonstrates link prediction and question answering on the data after pre-processing.

We inject knowledge into the LMs. The knowledge is injected by the process of fine-tuning the pre-trained LM. Fine-tuning requires a learning objective and training data. In our case, the training data is triplets from the KG (table 1), and the learning objective is triple completion. Triple completion involves getting tail entity given head entity and relation. Triple completion is also called link prediction. Thus, the LM absorbs the knowledge. The link prediction results with triplets are shown in table 3.

After fine-tuning on triplets for link prediction, the language model learns representations of entities and relations. The checkpoint with the best result on link prediction is used for the question-answering task. We again fine-tune the selected checkpoint with QA pairs (table 2) and obtain the QA results shown in table 4.

4.2 Experiment Data

We are using three datasets: a) Aviation Knowledge Graph (AviationKG) (Agarwal et al., 2022) & Aviation QA pairs (section 4.4), b) MetaQA (Zhang et al., 2018), which consists of a KB constructed from WikiMovies dataset (Miller et al., 2016b) and question-answer pairs, and c) Complex Web Questions (CWQ) (Talmor and Berant, 2018), which uses subsets of Freebase (Chah, 2017). The statistic of these datasets is shown in table 1 & 2. We chose these datasets because they belong to different domains and vary in size.

MetaQA KB & AviationKG are from the movie and aviation domains, respectively, which is useful to represent the diversity of datasets and validate our hypothesis. CWQ is based on Freebase, a huge KG, which is crowd-sourced. We require a knowledge base and the corresponding QA pairs for our experimentation, described in section 4.5. MetaQA and CWQ are openly available datasets. But, there is no available QA pairs dataset for the aviation

domain. We create a set of QA pairs in the aviation domain and contribute to the research community, detailed in section 4.4. The datasets used in the paper are pre-processed and split before running experiments, as explained in section 4.3 and 4.5.

Dataset	Train	Validation	Test
AviationKG	173,372	10,000	10,000
MovieKB	249,482	10,000	10,000
CWQ	27,590,648	10,000	10,000

Table 1: Statistics of triplets (subject, relation, object) for three knowledge bases: AviationKG (Agarwal et al., 2022), MetaKB (Zhang et al., 2018), and Complex Web Question (CWQ) (Talmor and Berant, 2018). Subsets of Freebase (Chah, 2017) are used for CWQ.

Dataset	Train	Validation	Test
AviationQA	367,304	10,000	10,000
MetaQA	184,230	10,000	10,000
CWQ	61,619	3,519	3,531

Table 2: Statistics of Question Answer pairs from three domains: Aviation, Movie, and Web. For MetaQA, we use 1-hop questions. For more details, refer to section 4.5.

4.3 Data Pre-processing

We make use of KG and QA pairs (section 4.2) from 3 domains, Aviation, Movie, and General domain. These datasets are cleaned and structured for our experiments. For the link prediction task, the dataset is created similar to Saxena et al. (2022), described below:

predict head: subject | relation | object
predict tail: object | relation | subject
The triplets {subject, relation, object} are extracted from the AviationKG, MovieKB, and Freebase individually.

All these knowledge bases are associated with the corresponding QA pairs. As explained in section 4.4, we construct the AviationQA pairs and use MetaQA 1-hop and CWQ for question answering. For QA fine-tuning, the dataset is in the given format:

predict answer: question | answer.
E.g., **predict answer**: What is the capital of India? | New Delhi.
Multiple answers exist for a question in AviationQA, MetaQA, and CWQ. These collective instances are separated as individual QA pairs.
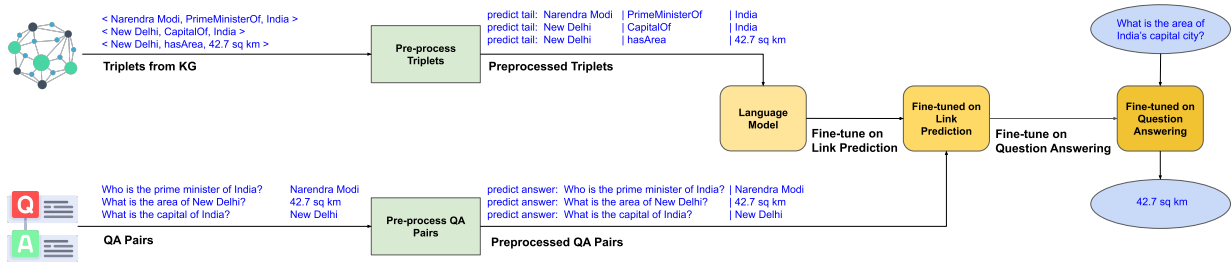
Figure 1: Flow diagram of the approach adopted in our paper. The model is first fine-tuned on KG triplets for Link Prediction. Next, the fine-tuned model is again fine-tuned on question answering. Because of the link-prediction task, the model learns KG completion and can answer multi-hop questions. E.g., If the model knows India's capital is New Delhi and New Delhi's area size, then the model should predict the area of India's capital correctly without explicitly mentioning New Delhi in the question

E.g., What countries did Narendra Modi visit in the year 2021? Answers: United States, Italy. Every QA pair is segregated in the current layout: a) What countries did Narendra Modi visit in the year 2021? | United States. b) What countries did Narendra Modi visit in the year 2021? | Italy.

With small KGs, *i.e.*, AviationKG, and MovieKB, triplet samples are added during QA fine-tuning to avoid overfitting. The added triplets are in the same format as mentioned for the link prediction task. The pre-processing of triplets and QA pairs is shown in figure 1.

4.4 Creation of AviationQA

We web scrape the National Transportation Safety Board (NTSB) website and download 12k reports from 2009-2022. A set of 90 question templates is prepared using the common structure of documents in the format:

- Where did the accident [ ] take place?

- What is the model/series of the aircraft bearing accident number [ ]?

- Was there fire on the aircraft of the accident number [ ]?

The template of questions is created, and answers to those questions are extracted from every NTSB report. Because every report is associated with an accident number, we place [ ] in the template to indicate which report the question pertains to, e.g., CHI07LA273, LAX07LA148. NTSB reports are semi-structured, containing unstructured data in paragraphs and structured data in tabular format. We extract answers from each report w.r.t the template using the regular expression method. Later,

QA pairs are scrutinized. As some reports' structure varies, different scripts are written to fetch answers for those reports.

We successfully created 1 million factoid QA pairs in the aviation domain using the template-based method. The dataset will contribute to research and development in the aviation industry.

4.5 Dataset Description

After pre-processing the data (section 4.3), we split it to train, validate, and test for link prediction and question answering. Table 1 shows the split of triplets from AviationKG, MovieKB, and subsets of Freebase. CWQ uses subsets of Freebase, which is of size 27 million. AviationKG and MovieKB are domain-specific datasets of sizes 170k and 250k. Valid and test splits are equal in size to 10k each.

Our motive for considering different sizes and domain datasets is to strengthen our intuition that the performance of varying size models remains the same with an infusion of knowledge in language models. Table 3 shows the correctness of our intuition with the link prediction task.

Table 2 shows the split of QA pairs for question-answering. We use 387,304 instances for AviationQA from 1 million QA pairs (section 4.4). The scrutinization is based on reports used to create AviationKG (Agarwal et al., 2022) from 1962 to 2015. We use QA pairs that have information available in the AviationKG. Moreover, we ensured that an answer to a question is an entity in the AviationKG.

For comparison between the movie and the aviation data, the split of valid and test set is the same in both, *i.e.*, 10k. CWQ dataset is smaller than AviationQA and MetaQA, so we use the same validation and test split, as mentioned in Saxena et al. (2022).

4.6 Model Configuration

In this paper, we are using four models: T5-small non-pretrained (60 million parameters), T5-base pre-trained (220 million parameters), T5-large pre-trained (770 million parameters), and BLOOM (1.72 billion parameters). These models are considered to validate our statement that with the injection of knowledge, small and large model performs the same. Both tasks, link prediction and question answering, are performed using these models. The T5 model is considered in our experiment as it is trained to perform multiple downstream tasks, i.e., translation, classification, and question answering. We use BLOOM as it is similar to the SOTA model GPT-3 (Brown et al., 2020), which has outperformed other language models on tasks such as QA and summarization.

4.7 Evaluation Technique

We evaluate the performance of our models using the hits@1 score for link prediction and question answering. Table 3 and 4 show the hits@1 score for link prediction and question answering, respectively, on different datasets. We choose the hits@1 score for evaluation as it is more precise than other hits@k scores. If the first predicted value matches the actual answer, then the score is 1; otherwise, 0. We are using the hits@1 metric and not other metrics such as BLEU score (Papineni et al., 2002) and semantic similarity (Miller and Charles, 1991) to validate the correctness of our hypothesis (introduced in section 1). BLEU score is generally used for comparing sentences, whereas, for link prediction and QA tasks, the answer is a compound noun, *i.e.*, an entity in the knowledge graph. Since the entities are ranked for tasks, the hits@1 score is the best metric. As the answers to link prediction and QA are entities of KG, the semantic similarity would not be able to distinguish between 2 different entities with semantically the same meaning. After considering all drawbacks of other metrics, we adapted the hits@1 score for the evaluation.

5 Results and Analysis

This section analyzes the performance of two models: T5 and BLOOM. Table 3 & 4 show the hits@1 score for link prediction and QA tasks, respectively. With table 3, we can clearly observe that the hits@1 score for three variations of the T5 model & BLOOM is proximate for three different datasets (section 4.5). The three T5 models score 0.22 &

Model	AviationKG	MetaKB	CWQ
T5-small	0.2258	0.0257	0.2153
T5-base	0.2387	0.0286	0.2273
T5-large	0.2323	0.0301	0.2207
BLOOM 1b7	0.2163	0.0365	0.2155

Table 3: Link Prediction results on three knowledge bases: Aviation Knowledge Graph (KG) (Agarwal et al., 2022), Meta Knowledge Base (Zhang et al., 2018), and subsets of Freebase (Chah, 2017) for Complex Web Questions (CWQ) (Talmor and Berant, 2018).

Model	AviationQA	MetaQA	CWQ
T5-small	0.7031	0.2144	0.2225
T5-base	0.7093	0.2158	0.2736
T5-large	0.7013	0.2371	0.2632
BLOOM 1b7	0.5507	0.2386	0.1517

Table 4: Question Answering (QA) results in three QA datasets: AviationQA (4.4), MetaQA (Zhang et al., 2018), and Complex Web Questions (CWQ) (Talmor and Berant, 2018).

0.23 hits@1 for link prediction on AviationKG. Similarly, scores with MetaKB and CWQ have very less differences among models. LMs on MetaKB perform poorly for link prediction compared to other datasets; 0.02 & 0.03 are the hits@1 scores on the T5 model & BLOOM. The reason is the extensiveness of triplets in the MetaKB and the presence of noise in the dataset. We chose MetaKB to have a diversity of datasets and justify our claim (explained in section 1).

The main observation with the link prediction task is that the T5-small non-pre-trained model performs alike to pre-trained models. The T5-base with 220 million parameters shows results like T5-large & BLOOM, which comprises 770 million & 1.7 billion parameters, respectively. Link prediction results (in table 3) infers our claim that small and large models perform the same with the infusion of knowledge.

To support our claim, we also performed QA with the same set of models as used for the link prediction task. With the AviationQA dataset, we achieved 0.7 hits@1 scores on T5-small, T5-base, and T5-large. LLMs such as T5-large & BLOOM are expected to perform better for QA than small models as they are trained with a large amount of data and vice-versa, T5-small non-pre-trained, and T5-base are expected to perform direly. But, we observe that the performance of all three T5 models

Hypothesis Testing	AviationKG			MetaQA		
	T5-small T5-large	T5-base T5-large	T5-large Bloom	T5-small T5-large	T5-base T5-large	T5-large Bloom
Paired Student T-test	Cannot Reject	Cannot Reject	Cannot Reject	Cannot Reject	Cannot Reject	Cannot Reject
Cohen's kappa Score	0.76	0.75	0.68	0.49	0.53	0.33
Agreement (%)	91.77	91.36	89.16	82.50	83.62	75.73

Table 5: Hypothesis Testing on link prediction with 'AviationKG' and question-answering with 'MetaQA' datasets. We choose two measures for the test: a) paired Student T-test (Hsu and Lachenbruch, 2014), and b) Cohen's kappa Score (Cohen, 1968), to prove our hypothesis- after injection of knowledge, small and large models perform the same. Student T-test with 0.1 significance value is done on 2000 instances of the test set selected randomly, and our hypothesis is not rejected 7 out of 10 times. We use the entire test set of 10,000 instances for the kappa score. Cohen's kappa scores on link prediction for AviationKG are between 0.6 and 0.8, and on question-answering for MetaQA, between 0.4 and 0.6. With these scores, we are able to prove that our claim is correct.

is the same for QA with the AviationQA dataset. Similarly, we observe that MetaQA achieves 0.2 hits@1 scores for non-pre-trained T5, pre-trained T5-base, T5-large, and BLOOM.

Through our experiments, we have shown how different model sizes perform on QA after infusion of knowledge using link prediction. Pre-trained and non-pre-trained models of different sizes have shown similar results on different domain datasets for link prediction and QA tasks. This contribution to the research community is pivotal as high accuracy can be achieved efficiently with less computation power, time, and cost.

The source code for our paper is publicly available on GitHub[4].

6 Hypothesis Testing

We attempt to contradict our hypothesis (1) that the difference in scores for the two models is negligible. We choose paired student t-test (Hsu and Lachenbruch, 2014) to refute our hypothesis. In our testing, the significance level (p-value) is 0.1, and the sample size is 20% of the test set selected randomly. In comparing the pair of models (section 4.6), we predicted T5-large to perform better than T5-base & T5-small and Bloom to perform better than all three models of T5 because of its large size. But, 7 out of 10 times student t-test was unable to reject our hypothesis, and the significance level among the pair of models was greater than 0.1. Table 5 clearly shows the paired student t-test on AviationKG (table 1) and MetaQA (table 2) for different pairs of models, and the result is the same,

our hypothesis cannot be rejected.

After not being able to reject the hypothesis, our next step was to strengthen it, so, we calculate Cohen's kappa (Cohen, 1968) score of the pair of models with different datasets (table 1 & 2). We consider a pair of models as two annotators and the hits@1 score corresponding to each sample in the test set as their annotations. Since our evaluation technique (section 4.7) uses hits@1 score and the score is binary for each sample, Cohen's kappa score is used to measure the reliability between the two models. The kappa score is calculated for all instances of the test set. Table 5 shows the Cohen's kappa score and % agreement for AviationKG and MetaQA datasets between pair of models. For link prediction on AviationKG, the kappa score is between 0.6 and 0.8, and agreement is near 90%. These results clearly denote the substantiality of our claim with high scores. We extend the test for question-answering with MetaQA. The pair of T5 models score 0.4-0.6, denoting moderate agreement as more than 80% of agreement. T5-large and Bloom pair scores 0.33 with 75.7% agreement, which is fair.

Thus, the testing supports our hypothesis, and we prove that the level of performance of different models with the infusion of knowledge remains the same.

7 Conclusion and Future Work

We have successfully created a million factoid QA pairs from the NTSB aircraft accident reports. The QA pairs are used in our experiments with AviationKG. We have validated our claim that with the infusion of knowledge to language models, the per-

---

[4]https://github.com/ankush9812/
Knowledge-Infusion-in-LM-for-QA

formance of the small language model is similar to the large language model. We substantiate with different language models and a diversity of datasets. Our investigation will benefit researchers in selecting the appropriate language model when working with knowledge and save computation power and time.

The future line of work is to investigate the performance of models with incomplete and noisy knowledge graphs and study the extent to which the models can outright the domain knowledge.

Acknowledgements

References

Katrin Affolter, Kurt Stockinger, and Abraham Bernstein. 2019. A comparative survey of recent natural language interfaces for databases. *The VLDB Journal*, 28(5):793–819.

Ankush Agarwal, Raj Gite, Shreya Laddha, Pushpak Bhattacharyya, Satyanarayan Kar, Asif Ekbal, Prabhjit Thind, Rajesh Zele, and Ravi Shankar. 2022. Knowledge graph–deep learning: A case study in question answering in aviation safety domain. *arXiv preprint arXiv:2205.15952*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Niel Chah. 2017. Freebase-triples: A methodology for processing the freebase data dumps. *arXiv preprint arXiv:1712.08707*.

Jacob Cohen. 1968. Weighted kappa: nominal scale agreement provision for scaled disagreement or partial credit. *Psychological bulletin*, 70(4):213.

Yuanfei Dai, Shiping Wang, Neal N. Xiong, and Wenzhong Guo. 2020. A survey on knowledge graph embedding: Approaches, applications and benchmarks. *Electronics*, 9(5).

Henry Hsu and Peter A Lachenbruch. 2014. Paired t test. *Wiley StatsRef: statistics reference online*.

Kaveri Kale, Pushpak Bhattacharyya, Aditya Shetty, Milind Gune, Kush Shrivastava, Rustom Lawyer, and

Spriha Biswas. 2022. Knowledge graph construction and its application in automatic radiology report generation from radiologist's dictation.

Weijie Liu, Peng Zhou, Zhe Zhao, Zhiruo Wang, Qi Ju, Haotang Deng, and Ping Wang. 2020. K-bert: Enabling language representation with knowledge graph. In *AAAI*.

George Michalopoulos, Yuanxin Wang, Hussam Kaka, Helen Chen, and Alexander Wong. 2021. UmlsBERT: Clinical domain knowledge augmentation of contextual embeddings using the Unified Medical Language System Metathesaurus. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1744–1753, Online. Association for Computational Linguistics.

Alexander Miller, Adam Fisch, Jesse Dodge, Amir-Hossein Karimi, Antoine Bordes, and Jason Weston. 2016a. Key-value memory networks for directly reading documents. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1400–1409, Austin, Texas. Association for Computational Linguistics.

Alexander Miller, Adam Fisch, Jesse Dodge, Amir-Hossein Karimi, Antoine Bordes, and Jason Weston. 2016b. Key-value memory networks for directly reading documents. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1400–1409, Austin, Texas. Association for Computational Linguistics.

George A Miller and Walter G Charles. 1991. Contextual correlates of semantic similarity. *Language and cognitive processes*, 6(1):1–28.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, Peter J Liu, et al. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21(140):1–67.

Adam Roberts, Colin Raffel, and Noam Shazeer. 2020. How much knowledge can you pack into the parameters of a language model? In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5418–5426, Online. Association for Computational Linguistics.

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.

Apoorv Saxena, Aditay Tripathi, and Partha Talukdar. 2020. Improving multi-hop question answering over knowledge graphs using knowledge base embeddings. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4498–4507, Online. Association for Computational Linguistics.

Haitian Sun, Tania Bedrax-Weiss, and William Cohen. 2019. PullNet: Open domain question answering with iterative retrieval on knowledge bases and text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2380–2390, Hong Kong, China. Association for Computational Linguistics.

Alon Talmor and Jonathan Berant. 2018. The web as a knowledge-base for answering complex questions. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 641–651, New Orleans, Louisiana. Association for Computational Linguistics.

Meihong Wang, Linling Qiu, and Xiaoli Wang. 2021. A survey on knowledge graph embeddings for link prediction. *Symmetry*, 13(3).

Quan Wang, Zhendong Mao, Bin Wang, and Li Guo. 2017. Knowledge graph embedding: A survey of approaches and applications. *IEEE Transactions on Knowledge and Data Engineering*, 29(12):2724–2743.

Michihiro Yasunaga, Hongyu Ren, Antoine Bosselut, Percy Liang, and Jure Leskovec. 2021. QA-GNN: Reasoning with language models and knowledge graphs for question answering. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 535–546, Online. Association for Computational Linguistics.

Yuyu Zhang, Hanjun Dai, Zornitsa Kozareva, Alexander J Smola, and Le Song. 2018. Variational reasoning for question answering with knowledge graph. In *Thirty-second AAAI conference on artificial intelligence*.

Victor Zhong, Caiming Xiong, and Richard Socher. 2017. Seq2sql: Generating structured queries from natural language using reinforcement learning. *arXiv preprint arXiv:1709.00103*.

A   Appendix

A.1   Examples of AviationQA

Below, we mention some examples from our created Aviation question-answering dataset (section 4.4):

- **Q:** Which seat was occupied by the pilot responsible for accident no. CEN18LA272?
  A: Left

- **Q:** Are there other Aircraft Rating(s) for the pilot of accident no. GAA18CA489?
  A: None

- **Q:** What is the make of the aircraft bearing accident no. CEN18LA272?
  A: Cessna

- **Q:** What is the category of the aircraft involved in accident no. GAA18CA489?
  A: Gyroplane

- **Q:** What is the Airworthiness Certificate of accident no. GAA18CA297?
  A: Normal