Enhancing Knowledge Graph Consistency Through Open Large Language Models: A Case Study

Ankur Padia, Francis Ferraro, Tim Finin

University of Maryland, Baltimore County (UMBC), MD, USA
pankur1@umbc.edu, ferraro@umbc.edu, finin@umbc.edu

Abstract

High-quality knowledge graphs (KGs) play a crucial role in many applications. However, KGs created by automated information extraction systems can suffer from erroneous extractions or be inconsistent with provenance/source text. It is important to identify and correct such problems. In this paper, we study leveraging the emergent reasoning capabilities of large language models (LLMs) to detect inconsistencies between extracted facts and their provenance. With a focus on "open" LLMs that can be run and trained locally, we find that few-shot approaches can yield an absolute performance gain of 2.5-3.4% over the state-of-the-art method with only 9% of training data. We examine the LLM architectures' effect and show that Decoder-Only models underperform Encoder-Decoder approaches. We also explore how model size impacts performance and counterintuitively find that larger models do not result in consistent performance gains. Our detailed analyses suggest that while LLMs can improve KG consistency, the different LLM models learn different aspects of KG consistency and are sensitive to the number of entities involved.

Introduction

Knowledge graphs (KGs) represent knowledge using nodes and edges, where nodes denote entities (e.g., Washington and United States) and edges signify relationships between entity pairs (e.g., "located_at"). KGs can be either domain-specific or encompass general world knowledge and are useful in diverse downstream tasks, including question answering (Zhu et al. 2021), semantic search (Wang et al. 2020), and guided conversations (Liu et al. 2019b).

Automatic construction of these extensive KGs often relies on Information Extraction (IE) systems. However, automatic construction can suffer from noise and contain semantic inconsistencies, which hinder downstream applications. In Fig. 1, we show an example of a fact that is inconsistent with the provenance text from which it was extracted. Thus, an automated approach is necessary to identify inconsistencies within a KG.

Previous approaches have used feature-based ensembles (Viswanathan et al. 2015), KG-embedding-based approaches (Pan et al. 2018), and neural graph-based approaches (Fung et al. 2021) to detect inconsistent facts.

Copyright © 2024, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

Extracted Fact: Mauritania; org:alternate_names; CPPCC

Provenance Text: China thanked Mauritania for supporting China on Taiwan, Tibet, Xinjiang, human rights and other issues concerning the country's core interests, Yu said. Yu said the CPPCC would like to work with the ESC of Mauritania to carry out exchanges and promote bilateral relations.

Figure 1: An example of an incorrectly extracted fact and associated provenance text. Here, the provenance text does not support the extracted fact. Notice that "facts" can be extracted across multiple sentences.

However, these approaches do not fully consider the linguistic signals present in the provenance text: recently, Padia, Ferraro, and Finin (2022) proposed a synchronous neural encoding approach to identify inconsistencies and introduced new relations to reconcile the extracted facts with provenance text. However, all of these models require substantial training data, making it challenging to apply in more specialized domains or when increasing computational demands of larger neural models makes training on larger datasets difficult and time-consuming.

Pretrained large language models (LLMs) have gained significant attention in downstream applications due to their robust understanding and reasoning capabilities. In this paper, we explore the capabilities and limitations of LLMs in reasoning about inconsistencies within KGs. However, we place particular emphasis on *open* models (Rogers et al. 2023): those where static versions can be downloaded, checkpointed, and replicated by researchers on their own machines. This focus is important from a scientific perspective and potential broader use perspective: not every downstream user may be able to use a proprietary cloud-based LLM, and so it is critical to understand the strengths and limitations of open, locally runnable models. To our knowledge, we are the first to evaluate the quality of the open-source LLMs on the knowledge graph consistency evaluation task.

Through our experiments, we found that Encoder-Decoder models outperform the previous baseline (Padia, Ferraro, and Finin 2022) using only around 9% of the entire training data. Our research indicates that Few-shot Encoder-

Decoder models better detect inconsistencies between facts and their corresponding provenance text than Decoder-Only models. Moreover, we found that increasing the number of parameters does not necessarily increase the consistency of KG. We also found that these larger models are sensitive to the number of entities mentioned in the provenance information. We have made our code and models available for further exploration at https://github.com/Ebiquity/kgc_llm.

Method

Problem Statement. We assume we have a provenance-endowed KG $\mathcal{G} = (\mathcal{S}, \mathcal{R}, \mathcal{O}, \mathcal{P})$, such as automatically created from an IE system. Here, $\mathcal{S}$ denotes the subject entity, $\mathcal{O}$ denotes the object entity, and $\mathcal{R}$ is the relation between them. The provenance $\mathcal{P}$ is the set of sentence(s) from which the information was extracted. The task of **KG consistency** is to determine if the extracted fact $(\mathcal{S}, \mathcal{R}, \mathcal{O})$ is supported with associated provenance $\mathcal{P}$ or not. While previous approaches (Pan et al. 2018; Fung et al. 2021; Padia, Ferraro, and Finin 2022) formulated it as a classification task, here we model it as a multi-choice question-answering one.

KG Consistency as Multi-Choice Question-Answering. We convert the fact and associated provenance information using the prompt template shown below to align our task with this LLM objective.

```
Context: China thanked Mauritania
for supporting China on Taiwan,
Tibet, Xinjiang, human rights and
other issues concerning the country's
core interests, Yu said. The CPPCC
would like to work with the ESC of
Mauritania to carry out exchanges and
promote bilateral relations, Yu said.

Question: Which of the following
answers is most applicable for
"Mauritania;org:alternate_names;CPPCC"
(a) True, or (b) False?
```

We found that converting KG consistency to a Multi-Choice Question-Answering entailment task better aligned it with the LLM pre-trained objective of predicting the next token. However, LLMs can generate text that is not controllable and can repeat itself (Zhang et al. 2022). To control the generation process, we added multiple-choice options. We tried other templates but found that most derailed the text-generation process or failed to generate precise answers.

Datasets

We considered two challenging datasets from the previous baseline (Padia, Ferraro, and Finin 2022), TAC 2015 and TAC 2017. Both contain facts extracted from more than 70 English-based IE systems.We considered these two datasets for several reasons: (i) they enable a fair comparison with the previous state-of-the-art baseline (Padia, Ferraro, and Finin 2022), (ii) subject matter experts manually annotated the datasets following the rigorous guidelines provided by Ellis (2015), and (iii) existing LLM benchmark datasets do not contain data for the knowledge inconsistency task.

	Train	Valid	Test
TAC-2015	626	6859	6856
TAC-2017	552	5734	5729

Table 1: Statistics of the dataset used in the experiments. These datasets are originally from Padia, Ferraro, and Finin (2022): while the validation and test splits are identical, note that we have downsampled the training set by randomly sampling five facts per relation from the training dataset, which results in using only 9% of training data. Each number represents the number of annotated $(\mathcal{S}, \mathcal{R}, \mathcal{O}, \mathcal{P})$ present in the data split. Each example is manually annotated as either True/False and as consistent/inconsistent.

Table 1 shows their statistics. Both datasets contain relations from three domains: (i) Geopolitical (e.g., `subsidiaries`), (ii) Person (e.g., `birthstate_or_provience`), and (iii) Organization (`top_member_or_employee`).

Evaluation

We use this multi-choice question-answering formulation in two ways: first, we use it for few-shot in-context learning (ICL). Second, we use it for fine-tuning but in a limited training data regime. Both approaches allow the models to perform the underlying task without requiring considerable training data. We convert each fact and its associated provenance text into the above prompt template with provenance text as context and extracted fact as part of the question.

Few-shot In-Context Learning. In-Context Learning (ICL) (Dong et al. 2023) allows the model to predict unseen instances based on the demonstration instances added as part of the prompt. In the ICL approach, we sample two additional instances for every evaluation instance to provide the few-shot examples. To avoid sample bias, each evaluation instance has different examples. We used GPT-3.5 for zero-shot in-context learning, i.e., no demonstration instances are added as part of the prompt.

Few-shot Fine-tuning. Compared to few-shot in-context learning, where we give examples as part of the evaluation-time prompt, in few-shot fine-tuning, we randomly selected a maximum of five facts supported by the corresponding provenance information and five facts not supported by the provenance information for each relation from TAC-2015 and TAC-2017 to create a dataset for fine-tuning. We used 9% of the training dataset as less than 9% resulted in lower performance compared to Padia, Ferraro, and Finin (2022).

Metric. We compared the ground truth inconsistency with the predicted inconsistency generated from the LLM's answer as an output from the model. We considered the LLM response to be valid only if the output was "(a)", "True", "(b)" or "False" as shown below. Each fact is associated with

Learning Approach	Model	TAC 2017			TAC 2015		
		P	R	F1	P	R	F1
Baseline (Padia, Ferraro, and Finin 2022)		48.1	**98.0**	63.2	**50.8**	65.2	57.1
ZS.	GPT-3.5	41.6	46.7	43.9	40.4	41.6	41.0
	Flan-T5 (large)	50.9	37.4	43.1	63.0	29.0	39.7
ICL.	Flan-T5 (large)	39.3	64.8	48.9	41.2	44.9	43.0
FT. Dec.	Galactica	34.8	40.2	37.3	29.1	64.0	40.0
	OPT	37.3	45.7	41.1	31.7	61.4	41.8
	Vicuna	35.9	95.1	52.2	27.0	**83.3**	40.8
FT. Enc.-Dec.	BART	34.3	65.1	44.9	29.9	79.9	43.6
	Flan-T5 (large)	**65.3**	66.5	**65.9**	49.5	77.5	**60.5**

Table 2: Performance of recent language models on KG consistency tasks. Padia, Ferraro, and Finin (2022) uses the full training dataset to train the model while the others use significantly less, 9%, of the training data. Here, FT stands for Fine Tuning, ICL for In-context Learning, ZS for zero-shot, Dec for Decoder-only architecture, and Enc-Dec for the encoder-decoder architecture.

a context and evaluated independently. In cases where multiple facts can be extracted from the same context, each fact is considered independently with the same context. We compare model output with ground truth to calculate the F1 score for the decoder model based on the predicted output and evaluate the Encoder-Decoder model using an exact match between the generated response and ground truth.

Context: China thanked Mauritania for supporting China on Taiwan, Tibet, Xinjiang, human rights and other issues concerning the country's core interests, Yu said. Yu said that the CPPCC would like to work with the ESC of Mauritania to carry out exchanges and promote bilateral relations.

Question: Which of the following answers is most applicable for "Mauritania;org:alternate_names;CPPCC" (a) True, or (b) False?

Response from LLM: (b)

Architectures and Implementation. We broadly consider two popular LLM architectures: (i) Encoder-Decoder and (ii) Decoder-Only. The encoder-decoder can be used for either ICL or fine-tuning, where in fine-tuning, the encoder-decoder is trained to continue the input text (Context + Question) using a sequence-to-sequence objective. Similarly, for the Decoder-only model, we used the final token representation to determine the likelihood of the response and fine-tuned it to maximize the likelihood of correct response.

With the exception of GPT-3.5, we evaluated using publicly available models. We considered six encoder-decoder models, Flan-T5 and its variants (base, large, x-large, and xx-large) (Chung et al. 2022) and BART (base and large) (Liu et al. 2019c) ranging from 0.3B parameters to 11B parameters. We evaluated two decoder-only models: (i) Galactica (Taylor et al. 2022), and (ii) OPT (Zhang et al. 2022) ranging from 0.3B parameters to 30B parameters, and (iii) Vicuna (Chiang et al. 2023) descendent of LLaMA model.

We used publicly available HuggingFace models (Wolf et al. 2019). We set the learning rate to 2e-5 and the L2 penalty (λ) to zero. We used early stopping and trained models for three epochs. We used the Adam optimizer (Kingma and Ba 2014) and set the maximum token length to 1024. We use an Ubuntu 18.04 with four 48GB RTX 8000 Nvidia GPUs, 200GB RAM, and 10 TB hard disk space.

Results

Table 2 shows the results. We considered two popular LLM architectures: (i) Encoder-Decoder and (ii) Decoder-only.

Encoder-Decoder: Overall, the Flan-T5 model performed better in identifying inconsistencies between extracted facts and their provenance. Comparing the fine-tuned Flan-T5 and BART models, Flan-T5 is more coherent in generating answers using the options available from the prompt, resulting in better performance. On the other hand, the fine-tuned BART performed poorly in generating completion text and occasionally repeated the input prompt without producing final answers. Comparing the performance of ICL Flan-T5 and FT Flan-T5 clearly indicates the benefit of fine-tuning.

Decoder-only: Among the Decoder-only models, Vicuna outperformed OPT and Galactica as Vicuna is instruct-finetuned with LLaMA (Touvron et al. 2023) as the base model. On the other hand, OPT performed better compared to Galactica and was on par compared to BART. We believe this is due to the nature of the pre-training corpus. OPT is pre-trained on a dataset used in RoBERTa (Liu et al. 2019a), the Pile (Gao et al. 2020), and PushShift.io Reddit (Baumgartner et al. 2020), which includes diverse data sources improving general cross-domain knowledge of the model.

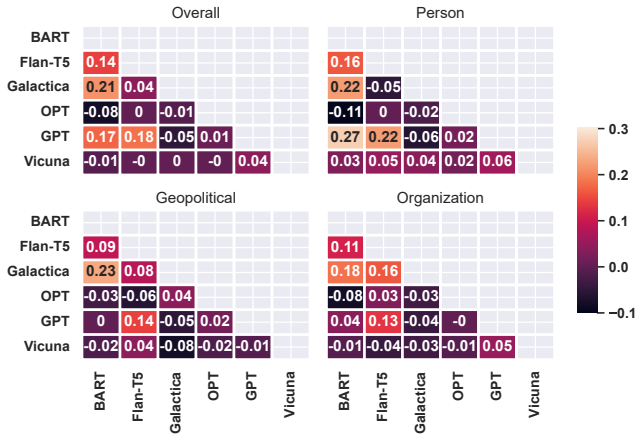

Figure 2: Pearson correlation between predictions from different Large Language Models. The higher/lower the correlation, the brighter/darker the cell, which indicates whether the models made the same/different predictions.

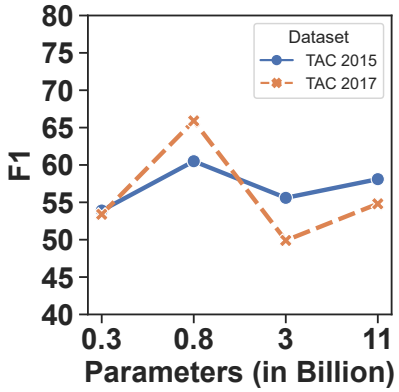

Figure 3: Change in performance vs. the number of parameters (in billions) in Flan-T5 models for the KG consistency task. Note that increasing the number of parameters does not uniformly lead to higher performance.

Discussion and Findings

Finding 1: Generic Models do not Outperform Fine-tuned Models to Identify Inconsistencies

We used the production-level LLM GPT-3.5-Turbo to identify inconsistencies in the knowledge graph, applying the same prompt as other models. As indicated in Table 2, GPT-3.5 performs poorly compared to the Flan-T5 model. We attribute this to the fact that the GPT-3.5 model is used in a zero-shot setting without fine-tuning. On the other hand, when comparing the in-context learning-based model with GPT-3.5, performance is slightly increased, mainly due to additional sample examples provided during in-context learning.

Finding 2: Different Architectures Learn Different Consistency Aspects

We consider the output generated from the above models to understand the effect of different architectures, the pre-training procedures of the LLM models, and the ability to understand and extract facts with provenance. We followed existing annotation guidelines (Ellis 2015) and used the relation prefix to group relations into three categories: Person, Geopolitical, and Organization. Figure 2 displays the Pearson correlation between predictions of different models. In the Overall heatmap (top-left), the Encoder-Decoder architectures BART and Flan-T5 show relatively higher correlations between them. However, there is little or a negative correlation among the Decoder architecture models, indicating that the different architectures produce different predictions, even with similar performance as shown in (Table 2).

When examining the heatmaps by domains (top-left and bottom row), all models generally encounter challenges in generating concise predictions for relations within the "Geopolitical" and "Organization" domains. Moreover, the correlation between BART and Flan-T5 also decreases. This difficulty stems from the rigorous and intricate annotation scheme utilized by the annotators (Ellis 2015), which requires that each extracted fact and its provenance adhere to a set of rules to be considered correct. Deviation from these annotation rules leads to the extracted fact and its provenance being labeled as incorrect.

Finding 3: Increasing Model Size does not Increase KG Consistency

Figure 3 illustrates the change in performance with an increase in the size of the Flan-T5 models (ranging from 0.3 to 11 billion parameters). A performance gain occurs when the model size is increased to a certain number of parameters. However, beyond that point, performance either decreases or remains constant. This suggests that better techniques are required to maximize the benefits of larger models. We hypothesize that this phenomenon results from quantization, the type of prompt used for fine-tuning the models, and the limited number of examples used in training.

Finding 4: Large Model Variants Perform Differently Based on the Relation Domain

Figure 4 shows the dependency of the F1-metric on the model size and training data domain. Overall, the Flan-T5-Large (0.8B) model performed best across all three categories. The rest of the model variants (base, x-large, and xx-large) performed similarly for the Person domain, while xx-large performed relatively better compared to x-large for the Geopolitical domain and Organization.

Finding 5: All Models Sensitive to Number of Entities

Depending on the design of the information extraction system, it can have multiple sentences as part of the provenance, each containing a varying number of entities. Figure 5 shows the number of entities' effect on the models' performance. For each fact-provenance pair, we calculate the number of entities present in the associated provenance and then take the average across all the pairs in the dataset

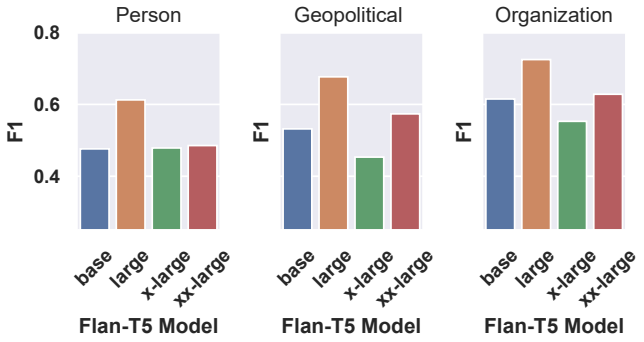

Figure 4: The performance of the Flan-T5 models on relations across domains varies. Compared to the base model, Flan-T5-base, the larger Flan-T5-large excels at identifying consistency in person, organization, and relations. However, even larger models, Flan-T5-x-large and xx-large, do slightly worse than smaller ones. As the number of parameters increases from Flan-T5-small to Flan-T5-base, the performance gap for GPE decreases.

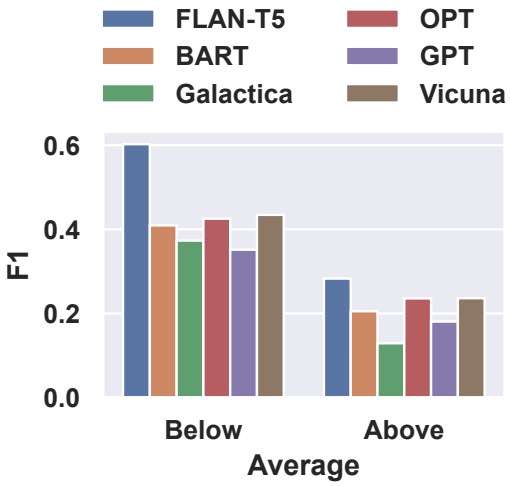

Figure 5: Drop in performance as the number of entities in the provenance increases above an average number of entities in the data.

to calculate the average number of entities. The term 'Below' ('Above') indicates the model's performance on examples with fewer (greater) than the average number of entities across the dataset. As it is evident, all the models perform better when fewer entities are present in the example compared to more entities, which results in a drop in performance.

Finding 6: LLMs Initially Learn Faster with more Datapoints, then Slower

To understand the effect of LLM on the availability of data, we sampled at five sample sizes n = {5,10,20,40,80} facts per relation from TAC-2017 and independently fine-tuned the five models. Overall, the performance of the LLM increases as the sample size increases. Initially, as n increases,

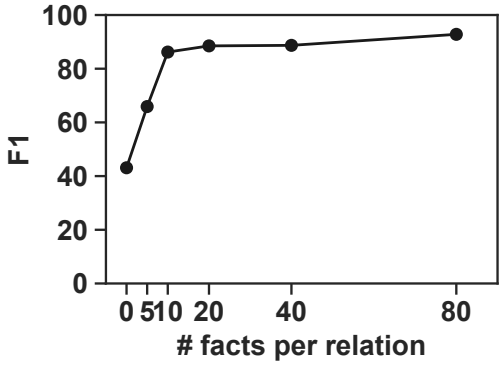

Figure 6: F1 score of Flan-T5-large as the number of training facts increases. As more training data is available, performance increases quickly until 20 facts per relation, after which adding more training data does not result in a significant gain. Here "0" indicates zero-shot.

there are significant performance gains until n=20, after which there are minor performance gains.

Conclusion and Future Work

Recently, several large language models like BART, Flan-T5, and Vicuna have been developed and pre-trained on a large scale and have been shown to improve downstream applications. We explored the limitations and capabilities of such large language models on the Knowledge Graph Consistency task to enhance the quality of the knowledge graph by identifying incorrectly extracted facts inconsistent with provenance information. We investigated the effects of architecture, such as Encoder-Decoder and Decoder, size, and the impact of entities on the identification capabilities of large language models. In the future, we plan to determine the consistency of domain-specific knowledge graphs.

Acknowledgements

We thank the anonymous reviewers for their comments, questions, and suggestions. This material is partly based on work supported by the National Science Foundation under Grant Nos. IIS-2024878 and DGE-2114892. This material is also based on research that is in part supported by the Army Research Laboratory, Grant No. W911NF2120076, and by the Air Force Research Laboratory (AFRL), DARPA, for the KAIROS program under agreement number FA8750-19-2-1003. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes, notwithstanding any copyright notation thereon. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either express or implied, of the Air Force Research Laboratory (AFRL), DARPA, or the U.S. Government.

References

Baumgartner, J.; Zannettou, S.; Keegan, B.; Squire, M.; and Blackburn, J. 2020. The pushshift reddit dataset. In *Pro-*

ceedings of the international AAAI conference on web and social media, volume 14, 830–839.

Chiang, W.-L.; Li, Z.; Lin, Z.; Sheng, Y.; Wu, Z.; Zhang, H.; Zheng, L.; Zhuang, S.; Zhuang, Y.; Gonzalez, J. E.; Stoica, I.; and Xing, E. P. 2023. Vicuna: An Open-Source Chatbot Impressing GPT-4 with 90%* ChatGPT Quality.

Chung, H. W.; Hou, L.; Longpre, S.; Zoph, B.; Tay, Y.; Fedus, W.; Li, E.; Wang, X.; Dehghani, M.; Brahma, S.; et al. 2022. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*.

Dong, Q.; Li, L.; Dai, D.; Zheng, C.; Wu, Z.; Chang, B.; Sun, X.; Xu, J.; Li, L.; and Sui, Z. 2023. A Survey on In-context Learning. arXiv:2301.00234.

Ellis, J. 2015. TAC KBP 2015 assessment guidelines. Technical report, Linguistic Data Consortium.

Ellis, J.; Getman, J.; Fore, D.; Kuster, N.; Song, Z.; Bies, A.; and Strassel, S. M. 2015. Overview of Linguistic Resources for the TAC KBP 2015 Evaluations: Methodologies and Results. In *Proceedings of the Eighth Text Analysis Conference*. NIST.

Fung, Y.; Thomas, C.; Reddy, R. G.; Polisetty, S.; Ji, H.; Chang, S.-F.; McKeown, K.; Bansal, M.; and Sil, A. 2021. Infosurgeon: Cross-media fine-grained information consistency checking for fake news detection. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 1683–1698.

Gao, L.; Biderman, S.; Black, S.; Golding, L.; Hoppe, T.; Foster, C.; Phang, J.; He, H.; Thite, A.; Nabeshima, N.; Presser, S.; and Leahy, C. 2020. The Pile: An 800GB Dataset of Diverse Text for Language Modeling. *arXiv preprint arXiv:2101.00027*. https://pile.eleuther.ai/.

Getman, J.; Ellis, J.; Song, Z.; Tracey, J.; and Strassel, S. M. 2017. Overview of Linguistic Resources for the TAC KBP 2017 Evaluations: Methodologies and Results. In *Proceedings of the 2017 Text Analysis Conference*. NIST.

Kingma, D. P.; and Ba, J. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; and Stoyanov, V. 2019a. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv preprint arXiv:1907.11692*.

Liu, Z.; Niu, Z.-Y.; Wu, H.; and Wang, H. 2019b. Knowledge Aware Conversation Generation with Explainable Reasoning over Augmented Graphs. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, 1782–1792. Association for Computational Linguistics.

Liu, Z.; Niu, Z.-Y.; Wu, H.; and Wang, H. 2019c. Knowledge Aware Conversation Generation with Explainable Reasoning over Augmented Graphs. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing*, 1782–1792. Association for Computational Linguistics.

Padia, A.; Ferraro, F.; and Finin, T. 2022. Jointly Identifying and Fixing Inconsistent Readings from Information Extraction Systems. In *Proceedings of Deep Learning Inside Out (DeeLIO 2022): The 3rd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures*, 42–52. Association for Computational Linguistics.

Pan, J. Z.; Pavlova, S.; Li, C.; Li, N.; Li, Y.; and Liu, J. 2018. Content Based Fake News Detection Using Knowledge Graphs. In *The Semantic Web – ISWC 2018*, 669–683. Cham: Springer.

Rogers, A.; Balasubramanian, N.; Derczynski, L.; Dodge, J.; Koller, A.; Luccioni, S.; Sap, M.; Schwartz, R.; Smith, N. A.; and Strubell, E. 2023. Closed AI Models Make Bad Baselines. https://hackingsemantics.xyz/2023/closed-baselines/.

Taylor, R.; Kardas, M.; Cucurull, G.; Scialom, T.; Hartshorn, A.; Saravia, E.; Poulton, A.; Kerkez, V.; and Stojnic, R. 2022. GALACTICA: A Large Language Model for Science. *arXiv preprint arXiv:2211.09085*.

Touvron, H.; Lavril, T.; Izacard, G.; Martinet, X.; Lachaux, M.-A.; Lacroix, T.; Rozière, B.; Goyal, N.; Hambro, E.; Azhar, F.; Rodriguez, A.; Joulin, A.; Grave, E.; and Lample, G. 2023. LLaMA: Open and Efficient Foundation Language Models. arXiv:2302.13971.

Viswanathan, V.; Rajani, N. F.; Bentor, Y.; and Mooney, R. 2015. Stacked ensembles of information extractors for knowledge-base population. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 177–187.

Wang, K.; Shen, Z.; Huang, C.; Wu, C.-H.; Dong, Y.; and Kanakia, A. 2020. Microsoft academic graph: When experts are not enough. *Quantitative Science Studies*, 1(1): 396–413.

Wolf, T.; Debut, L.; Sanh, V.; Chaumond, J.; Delangue, C.; Moi, A.; Cistac, P.; Rault, T.; Louf, R.; Funtowicz, M.; et al. 2019. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

Zhang, S.; Roller, S.; Goyal, N.; Artetxe, M.; Chen, M.; Chen, S.; Dewan, C.; Diab, M.; Li, X.; Lin, X. V.; Mihaylov, T.; Ott, M.; Shleifer, S.; Shuster, K.; Simig, D.; Koura, P. S.; Sridhar, A.; Wang, T.; and Zettlemoyer, L. 2022. OPT: Open Pre-trained Transformer Language Models. *2205.01068*.

Zhu, F.; Lei, W.; Wang, C.; Zheng, J.; Poria, S.; and Chua, T.-S. 2021. Retrieving and reading: A comprehensive survey on open-domain question answering. *arXiv preprint arXiv:2101.00774*.