Reporting and Analysing the Environmental Impact of Language Models on the Example of Commonsense Question Answering with External Knowledge Infusion

Anonymous ACL submission

Abstract

Human-produced emissions are growing at an alarming rate, causing already observable changes in the climate and environment in general. Each year global carbon dioxide emissions hit a new record, and it is reported that 0.5% of total US greenhouse gas emissions are attributed to data centres as of 2021 ((Siddik et al., 2021)). The release of ChatGPT in late 2022 sparked social interest in Large Language Models (LLMs), the new generation of Language Models with numerous parameters and trained on massive amounts of data. Currently, numerous companies are releasing products featuring various LLMs, with many more models in development and awaiting release. Deep Learning research is a competitive field, with only models that reach top performance attracting attention and being utilized. Hence, achieving better accuracy and results is often the first priority, while the model's efficiency and the environmental impact of the study are neglected. However, LLMs demand substantial computational resources and are very costly to train, both financially and environmentally. It becomes essential to raise awareness and promote conscious decisions about algorithmic and hardware choices. Providing information on training time, the approximate carbon dioxide emissions and power consumption would assist future studies in making necessary adjustments and determining the compatibility of available computational resources with model requirements. In this study, we infused T5 LLM with external knowledge and fine-tuned the model for Question-Answering task. Furthermore, we calculated and reported the approximate environmental impact for both steps. The findings demonstrate that the smaller models may not always be sustainable options, and increased training does not always imply better performance. The most optimal outcome is achieved by carefully considering both performance and efficiency factors.

Model	Parameters	Estimated emissions	Equivalence in # of flights
BERT	110M	1.59	1.9
BLOOM	176B	25	30
OPT	175B	75	90
Gopher	280B	380	456
GPT-3	175B	500	600

Table 1: The first column shows state-of-the-art LLMs, along with the corresponding number of parameters in the second column and emissions produced during training in metric tons CO_2eq in the third column. The last column represents the equivalence in the number of round flights between London and New York.

1 Introduction

With the growing problem of climate change, LLMs can potentially accelerate that process by contributing to greenhouse gas emissions. LLMs with billions of parameters may require several weeks of training time, and this duration is expected to increase further with the emergence of new models ((Scao et al., 2022; Radford et al., 2019; Brown et al., 2020)). Table 1 demonstrates the most recent LLMs released by famous research labs, the number of parameters of each model, the estimated emissions in net metric tons CO_2eq and the equivalence in flights. The amount of produced emissions doubles if taken into consideration the manufacturing of computers. Considering the computational expenses involved, it is only essential to prevent executing identical experiments and adopt a sustainability mindset in research endeavours. This means that researchers have to report not only performance but also training time, energy consumption, pre-training and fine-tuning requirements, and any other metrics that demonstrate the model's efficiency. Reporting training time and energy consumption can help to identify resource-intensive approaches to avoid or optimize them later. Carbon dioxide equivalence helps to assess the environmental impact of the research holistically. Ultimately, understanding and minimizing resource consumption and reducing carbon emissions promote the development of more sustainable practices and making informed decisions towards effective and efficient solutions.

In this study, we focus on Commonsense Reasoning and Question-Answering NLP tasks. Commonsense is a set of implicit pre-knowledge about the everyday world. For example, it is common knowledge that a refrigerator can be found in the kitchen and that summer comes after spring. Commonsense reasoning requires human experience, together with social, physical, temporal and spatial information of everyday life. Learning and using implicit knowledge for humans is an easy everyday task, which makes their language concise yet precise. However, machines do not possess common sense and are not able to learn such knowledge by interacting with the environment. That makes the Commonsense Question Answering (CSQA) task one of the major goals in the Artificial Intelligence (AI) community.

A way to teach models common sense and reasoning is through training them on a commonsense data. The study conducted by (Lal et al., 2021) introduced a new dataset for CSQA and fine-tuned three LLMs to showcase the TellMeWhy dataset. The authors fine-tuned and tested the performances of T5 ((Raffel et al., 2020)), GPT 2 ((Radford et al., 2019)) and UnifiedQA ((Khashabi et al., 2020)). To assess the effect of data size, model parameters and points mentioned above, our study similarly explores the T5 model and fine-tunes it on the TellMeWhy dataset.

In this study, we focus on two aspects: 1) how long does it take to train the T5 model and how we can calculate and report the environmental impact; 2) how does knowledge infusion from Knowledge Graphs (KG) influence the T5 model's ability to perform on CSQA task. Both goals are supposed to be achieved by injecting the commonsense knowledge from KG and fine-tuning the model on the CSQA dataset.

2 Related Work

Many advocate making efficiency reports a routine practice in deep learning research. Yet, when diving deeper into the problem, it is clear that part of the reason why very few researchers report efficiency results is because of the absence of a standard of measurement. There are numerous metrics available to assess the quality of the model, and often times improved performance means a better prediction ability. Some even argue that modern AI does not actually learn and is just a result of utilizing massive amounts of data and large computation power. Although sustainability in AI is still in its infancy, there are already great studies being held to bring awareness to the research community. In this section, we mention works that have been held to quantify and measure the carbon footprint of LLMs. By the end of the section, we will also briefly mention studies in knowledge infusion, which is also part of our study.

Multiple studies have already focused on energy consumption and carbon emissions accounting; some even propose methods for mitigating the problem. Prioritizing the model's efficiency over performance is becoming more relevant as more powerful machines are being developed. Several factors contribute to the increase in training time directly or indirectly, the development of more robust and powerful hardware, more complex machine learning algorithms and approaches, data growth, and social demand.

The study of Strubell et al. draws attention to the potentially hazardous impacts of training large models on our environment and proposes solutions to mitigate the problem. As an example, they trained a few state-of-the-art LLMs and put them into perspective by quantifying carbon emissions produced during training. Later they compared the results with the emissions produced during a flight and cloud computing prices. The work has concluded that training BERT emits roughly the same amount of carbon into the atmosphere as a trans-American flight. The paper was one of the first papers to draw attention to the environmental impact done by LLMs. Additionally, the authors provided standardized a reporting metrics for the emissions produced during training by comparing them to a more common metrics, like price and flights. Such comparison is still being used fore reporting in the recent papers.

The work of Wu et al. goes beyond measuring carbon emissions during training. The study also includes model development and inference phases. The authors encourage not only to look at the training phase but to consider the machine learning pipeline end-to-end, starting from data collection until inference. They examine the ML development cycle across the industry scale. Operational and manufacturing carbon footprint is also taken into account, by the end of the study the authors discuss how hardware choices and optimization techniques can help to reduce the carbon footprint of an AI system.

Work conducted by Patterson et al. proved that most of the companies and research groups try to avoid pre-training and prefer executing fine-tuning and inference stages. The study suggests that such stages are as important as pre-training and should not be neglected when it comes to carbon footprint accounting. The study proved that the inference can produce a significant amount of emissions as well.

So far, we have looked into measures for energy consumption and studies conducted on green AI. Now we will inspect KG-infusion methods, as it is a promising approach for carbon footprint reduction by enabling hybrid or neuro-symbolic AI. According to Bauer et al. providing LLMs with external knowledge enhances its ability to reason on a downstream task, i.e. QA, summarization, etc. Knowledge infusion enriches the model's vocabulary and allows it to "think out of the box". Some works have already attempted to incorporate commonsense knowledge into the BERT model to enhance reading comprehension (Yang et al.), and relation classification (Zhang et al.).

A recently conducted study by Lal et al. extends their previous work on Commonsense QA. The authors utilize COMET KG as an external knowledge source and inject the knowledge into the LLM. As a result, they observed an increase in performance. However, none of the studies measure and report the environmental impact of their work.

3 Experimental Setup

3.1 Dataset

3.1.1 Knowledge infusion

As external knowledge for our LLM, we combined ConceptNet (Speer et al.) and ATOMIC (Sap et al.), which are both large-scale Knowledge Graphs containing information about events in everyday life. Both KGs have to be pre-processed prior to being fed into the LLM. In the scope of this study, we pre-processed and verbalised only ConceptNet KG and combine it with already pre-processed ATOMIC KG Guan et al..

ConceptNet is constructed of multiple triplets (Subject, Relation, Object) with corresponding relation weight. We start by iterating over subjects and sorting them based on their relationship weight. Then we select the top 100 triplets with respect to relation weight and transform them into sentences using simple verbalization templates (Levy et al.), see Figure 1. The combined pre-processed dataset contains 1,174,267 sentences in the train set and 66,856 in the validation set. The dataset contains physical, spatial, social, and temporal aspects of daily life.

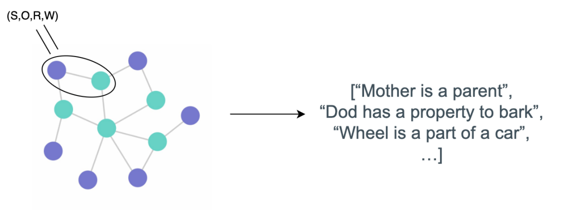

Figure 1: Transforming KG into a natural language sentence.

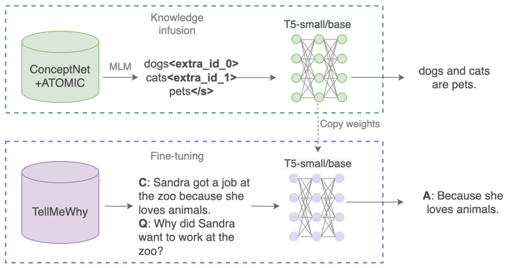

Figure 2: Study workflow. First step: Infusing T5-small/-base with knowledge from Knowledge Graphs. Second step: Fine-tuning model with injected knowledge for QA task.

3.1.2 Fine-tuning

Following Lal et al. example, models are fine-tuned on the TellMeWhy corpus, the largest Commonsense QA dataset. It incorporates various stories, 30K open-ended questions, and free-form answers. The provided short narratives describe why characters performed certain actions. The answers can be explicit and found in the narrative, as well as implicit, answers that require external knowledge, some intuitive knowledge about the world.

3.2 Methods

As mentioned earlier, injecting commonsense knowledge from KG beforehand should prepare a solid base for the fine-tuning step. Fine-tuning is performed on the TellMeWhy dataset for Commonsense Question-Answering task, which results in better rationalization and reasoning abilities.

Originally, T5 was pre-trained on the large unlabelled corpus, Colossal Clean Crawled Corpus (C4) corpus (Raffel et al.) cleaned information from the web. The web-crawled data consists of over 300M sentences for various topics; hence, further training the model on commonsense knowledge data might strengthen T5's ability to form constructive sequences and generate better reasoning. Step 1 is the continuation of T5's original unsupervised pre-training on the Masked Language Modelling task (MLM). The words in encoded sentences are masked with a 15% probability, together with the reversed masks as labels they are fed into the network. We maintained input and output in the same fashion as the original pre-training. Due to the size of this dataset, the knowledge infusion step runs only for one epoch, more extended training seems to lead to overfitting.

Once the knowledge infusion is completed, the model is fine-tuned for the QA task. Encoded context and question serve as input to produce the predicted answer, which is then compared to the reference answer. This step allows the model to only concentrate on a specific NLP task.

We maintained the same parameters for the fine-tuning phase as Lal et al.. We set the maximum number of epochs to 50, with a learning rate of 5-e5 and a batch size of 16. Our experiments also run until the validation loss does not improve for three iterations. However, we set the maximum source length to 255.

3.3 Models

In the scope of this study, we utilized T5-small and T5-base models (Raffel et al.). T5 is a transformer-based model that can be used for multiple NLP tasks without making any architectural changes in contrast to other language models, due to the unified text-to-text format. Such architecture enables the application of transfer learning techniques to reduce the training cost. Both, the input and the output, are string types. Task specifications are added in the beginning and separated by the colon from the input.

The T5-small version has 60 million, and T5-base has 220 million parameters. T5-large with 11B parameters was computationally too expensive for our servers; hence, it was not utilized in this study.

4 Results

In the scope of our study, we conducted 6 experiments, which will be further referred to as follows:

1. *T5s IK*: T5-small with injected knowledge from KG
2. *T5b IK*: T5-base with injected knowledge from KG
3. *T5s FT*: T5-small fine-tuned for QA task
4. *T5b FT*: T5-base fine-tuned for QA task
5. *T5s IK+FT*: T5-small with injected knowledge from KG and fine-tuned for QA task
6. *T5b IK+FT*: T5-base with injected knowledge from KG and fine-tuned for QA task

4.1 Performance Analysis

To evaluate the model's performance, we utilized the same metrics as Lal et al.. BLEURT (Sellam et al.) and BLEU (Papineni et al.) scores are both learned evaluation metrics for natural text generation based on BERT. Being trained on WMT human annotations for the machine translation task, they correlate well with human judgments. The scores are generated based on the precision of tokens of a candidate sentence to the reference. While BertScore (Zhang et al.) uses only pre-trained contextual embeddings from BERT and matches words between two sentences by cosine similarity.

We also measured cosine similarity between the generated and the target answers and analysed the number of unique words presented in answer vocabulary that does not exist in context vocabulary.

Table 2 presents the model performance in various setups, the automatic evaluation provided on the official TellMeWhy GitHub repository[1]. Based on the results, we cannot prove that infusing T5 with commonsense knowledge from ConceptNet and ATOMIC influences the model's ability to reason. This could be due to the large size of the C4 corpus, and, thus, KGs ConceptNet, and ATOMIC failing to provide enough knowledge to teach the network. However, we can conclude that the T5 model is inherently bad at commonsense reasoning, due to the type of data it has been pre-trained on.

[1] `https://github.com/StonyBrookNLP/tellmewhy`

Experiment	Convergence epoch	BLEU	RG-L F1	BLEURT	BERTscore
Full Test Set					
T5s FT	12	21.93	0.25	-0.412	0.501
T5s IK+FT	13	22.94	0.25	-0.374	0.513
T5b FT	6	24.43	0.26	-0.359	**0.535**
T5b IK+FT	6	**24.57**	**0.26**	-0.3514	0.5338
Lal et al. T5-base	30-50	24.53	0.24	**-0.28**	0.48
Implicit-Answer Questions					
T5s FT	12	15.2	0.19	-0.618	0.429
T5s IK+FT	13	15.32	0.19	-0.589	0.431
T5b FT	6	**16.92**	0.2	-0.58	**0.452**
T5b IK+FT	6	16.53	**0.2**	-0.577	0.4457
Lal et al. T5-base	30-50	16.31	0.17	**-0.51**	0.34

Table 2: Performance of models on the full test set and implicit answer questions in the test set using automatic evaluation provided by Lal et al..

Yet, there is an observable difference in results between *T5s FT* and *T5s IK+FT* across most of the metrics. *T5s IK+FT* performed slightly better than *T5s FT*. The difference in the BLEU score is 1.01 and in BertScore is 2.4%, both are noticeable differences, considering the evaluation is for similar models and on the same dataset. We assume that due to the smaller size of the T5-small model, the significance of commonsense knowledge from ConcentNet and ATOMIC was more prominent. Compared to T5-base, T5-small seems to gain more from the knowledge infusion step. While for T5-base, ConceptNet and ATOMIC KGs are too small to make a visible difference.

The BLEURT score demonstrates that for all experiments, there exists a negative correlation between predicted and reference answers. The BLEURT and BLEU scores were specifically designed to assess the quality of the machine translation; this could explain the insignificance of the results. However, since BertScore only uses BERT embeddings and calculates the cosine similarity between two sentences, we observe a higher correlation between the predicted and goal answers.

Similarly to Lal et al., models perform best when the answer is explicitly given in the context. We observed a slight performance increase for *T5b IK+FT* compared to the *T5 base* results of Lal et al.. We anticipate that the ROUGE F-1 and BertScores scores are higher in our experiments, compared to that of Lal et al., because we set *max_len_seq* to 200, as this was the size of the longest token in our case. However, we still came to the conclusion that the most influence comes from fine-tuning step, but it seems like knowledge infusion makes some difference for smaller models.

It is worth noting that the comparable results in our experiments were achieved with fewer epochs. Lal et al. suggests that T5-base reaches the best performance between epochs 30 and 50. However, we could see that longer training does not add much to the performance and incorporating EarlyStopping is necessary to prevent not only overfitting but also resource over usage.

The semantic similarity between the answers increases as the training time and size of a model also increase, but the difference is not significant. Surprisingly, infusing models with commonsense knowledge and fine-tuning on QA resulted in the model using more TellMeWhy context vocabulary rather than the model that was just fine-tuned on QA.

4.2 Efficiency analysis

Some studies provide great solutions to facilitate carbon emissions and energy consumption calculation. Anthony et al. developed a library that accesses information about hardware and calculates the estimates after the first epoch. Their Carbontracker gives information about approximate carbon emissions in grams, energy consumption (KW/h), and an equivalent number of kilometres the car would have driven producing the same amount of emissions. Alternatively, Lacoste et al. developed a tool that can be used after executing experiments. By providing training time, location, and hardware type, you can estimate produced CO_2 emissions and also how much would

Experiment	Overall time (hr)	Energy use (KWh)	$CO_2eq.$ (kg)	Travel by car (km)
Step 1 (Knowledge Infusion)				
T5s IK	4.438	3.53	1.04	8.62
T5b IK	13.969	10.72	3.15	26.18
Step 2 (Fine-tuning)				
T5s FT	1.981	1.52	0.45	3.72
T5s IK+FT	6.612	5.19	1.53	12.68
T5b FT	3.793	2.74	0.81	6.7
T5b IK+FT	17.718	13.42	3.95	32.80

Table 3: Energy and emissions calculated by Carbontracker (Anthony et al.).

have been emitted if the experiment was held in a different datacenter.

In our study, we embedded Carbontracker (Anthony et al.) into the training loop, which approximates carbon emissions and power usage for the whole training after 1 epoch. Eq. (1) is used to calculate the power usage of an experiment p_t. The average GPU power draw p_g is usually obtained by querying the NVIDIA System Management Interface throughout the run. The value is then multiplied by the number of GPUs g and the Power Usage Effectiveness Coefficient (PUE) (1.55 for Germany[2]).

$$p_t = \frac{1.55 * t * g * p_g}{1000} \quad (1)$$

The number of emissions and power consumption depends on the data centre location and the local power grid it is connected to. The same experiments executed in two different locations may have different environmental impacts. As of 2022, the power sector emissions in Germany were approximately 380 grams of carbon dioxide produced per kilowatt-hours (gCO_2/KWh) for generated electricity[3]. To get the carbon emissions equivalence estimation (in kg per kilowatt-hour), emissions per hour are multiplied by the experiment's power usage, as shown in Eq. (2).

$$CO_2e = 0.380 * p_t \quad (2)$$

We report the overall time it took to execute one experiment, the energy use, carbon dioxide emissions equivalence, and the equivalence in travel by car. All experiments in this study were performed on 2 NVIDIA RTX A5000 GPU blocks with 24GB memory each.

[2]https://www.statista.com/statistics/1229367/data-center-average-annual-pue-worldwide/

[3]https://www.nowtricity.com/country/germany/

Table 3 presents the training time of each experiment and the corresponding efficiency metrics calculated by Carbontracker. Since we set an early stopping during the fine-tuning step, none of the experiments reached the maximum number of epochs. *T5s FT* ran until 12, while *T5s IK+FT* until 13, both *T5b FT* and *T5b IK+FT* stopped after 6 epochs. This fact demonstrates the importance of early stopping in research to prevent unnecessary resource waste and energy consumption when the models do not need long training.

Clearly, the knowledge infusion phase required a much longer training time, the difference between minor and base variants is also significant, with T5-base requiring 3 times more hours to complete 1 epoch. Having a look at the fine-tuning stage, we can see that the difference between *T5s FT* and *T5b FT* is around 2 hours, but *T5b FT* outperforms the former. In our case, infusing knowledge and fine-tuning T5-small did not give a desirable performance, hence T5-base is preferred even with a longer training time.

Looking at *T5b FT* and *T5b IK+FT*, we noticed that the latter outperforms the first one only by a mere percentage based on BLEU, F1 and BLEURT scores. On the contrary, BERTscore for *T5b FT* has been consistently higher than that for *T5b IK+FT*.

5 Conclusion

Numerous factors could have influenced the outcome of our study. We assume that among these factors, the nature of the data that we infused into our model influenced the most. While sorting the KG based on relation weight and extracting top N triples seems like a straightforward approach, it yields suboptimal results. The main limitation lies in the lack of diversity within the dataset, with many sentences being semantically close and having limited number of relationship types. The two main conclusions from our study include:

- More sophisticated approaches to linearize KGs in a meaningful way are required. The resulting dataset should be rich in terms of semantics and relationship types.
- When looking for a balance between performance and efficiency, *T5b FT* seems like a more reasonable choice.

Our study showed that it is important to consider a model not solely based on one parameter. Focusing only on performance could lead to uncontrollable energy waste, while trying to reduce energy consumption too much can lead to a weak model that is less sustainable in the long run. The balance between the two is the key to the most optimal solution.

Tracking carbon footprint at every stage of the study is an extremely challenging task and has much more room for improvement regarding the report standards. To get the full picture, one might also need to know how much it takes to build hardware, transport them to the data centre, as well as consider the lighting in the room, etc. Nevertheless, it is important to be aware of the factors that impact the quantity of carbon emissions produced by research. As we have seen, stages like fine-tuning can also produce an observable amount of emissions. Such a step towards a positive change can also greatly help follow-up studies in the field.

Limitations

Our study includes several limitations that couldn't been addressed in this study and could be an idea for future work. Firstly, the Knowledge Infusion part of our study did not yield desirable results due to the poor KG linearization strategy. This stage also took the most time to be executed and consumed the most computational power. Secondly, Due to server limitations, we couldn't perform any experiments on T5-large model, which restricts us from making bolder statements on LLM performance on the CSQA task. In this work, we wanted to draw attention to the importance of considering efficiency results together with the performance results of the study.

Ethics Statement

Our research focuses on accounting and reporting the environmental impact of LLMs. Such studies raise concerns about transparency and accountability of Deep Learning approaches. It is crucial that the processes and algorithms used in any study are transparent and open to scrutiny. We commit to making our methods and data publicly available for review and validation by the broader community.

While the benefits of this research field are clear, it is essential to acknowledge and address potential ethical considerations. Calculating the exact amount of emitted carbon into the atmosphere presents a challenging task that requires acquiring server production and transportation information, as well as considering local energy grid and its fuel type. Furthermore, we should also scrutinise the Deep Learning model exploitation and life-cycle periods to get a clearer picture of its environmental impact. Hence, this field of study still requires extensive research with its potential positive impact on the research community.

References

Lasse F. Wolff Anthony, Benjamin Kanding, and Raghavendra Selvan. 2020. Carbontracker: Tracking and predicting the carbon footprint of training deep learning models. *CoRR*, abs/2007.03051.

Lisa Bauer, Yicheng Wang, and Mohit Bansal. 2018. Commonsense for generative multi-hop question answering tasks. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 4220–4230. Association for Computational Linguistics.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Jian Guan, Fei Huang, Minlie Huang, Zhihao Zhao, and Xiaoyan Zhu. 2020. A knowledge-enhanced pretraining model for commonsense story generation. *Trans. Assoc. Comput. Linguistics*, 8:93–108.

Daniel Khashabi, Sewon Min, Tushar Khot, Ashish Sabharwal, Oyvind Tafjord, Peter Clark, and Hannaneh Hajishirzi. 2020. Unifiedqa: Crossing format boundaries with a single QA system. In *Findings of the*

Association for Computational Linguistics: EMNLP 2020, Online Event, 16-20 November 2020, volume EMNLP 2020 of *Findings of ACL*, pages 1896–1907. Association for Computational Linguistics.

Alexandre Lacoste, Alexandra Luccioni, Victor Schmidt, and Thomas Dandres. 2019. Quantifying the carbon emissions of machine learning. *CoRR*, abs/1910.09700.

Yash Kumar Lal, Nathanael Chambers, Raymond J. Mooney, and Niranjan Balasubramanian. 2021. Tellmewhy: A dataset for answering why-questions in narratives. In *Findings of the Association for Computational Linguistics: ACL/IJCNLP 2021, Online Event, August 1-6, 2021*, volume ACL/IJCNLP 2021 of *Findings of ACL*, pages 596–610. Association for Computational Linguistics.

Yash Kumar Lal, Niket Tandon, Tanvi Aggarwal, Horace Liu, Nathanael Chambers, Raymond J. Mooney, and Niranjan Balasubramanian. 2022. Using commonsense knowledge to answer why-questions. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 1204–1219. Association for Computational Linguistics.

Omer Levy, Minjoon Seo, Eunsol Choi, and Luke Zettlemoyer. 2017. Zero-shot relation extraction via reading comprehension. In *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*, pages 333–342, Vancouver, Canada. Association for Computational Linguistics.

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, July 6-12, 2002, Philadelphia, PA, USA*, pages 311–318. ACL.

David A. Patterson, Joseph Gonzalez, Quoc V. Le, Chen Liang, Lluis-Miquel Munguia, Daniel Rothchild, David R. So, Maud Texier, and Jeff Dean. 2021. Carbon emissions and large neural network training. *CoRR*, abs/2104.10350.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

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

Maarten Sap, Ronan Le Bras, Emily Allaway, Chandra Bhagavatula, Nicholas Lourie, Hannah Rashkin, Brendan Roof, Noah A. Smith, and Yejin Choi. 2019. ATOMIC: an atlas of machine commonsense for if-then reasoning. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*, pages 3027–3035. AAAI Press.

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilic, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, Jonathan Tow, Alexander M. Rush, Stella Biderman, Albert Webson, Pawan Sasanka Ammanamanchi, Thomas Wang, Benoît Sagot, Niklas Muennighoff, Albert Villanova del Moral, Olatunji Ruwase, Rachel Bawden, Stas Bekman, Angelina McMillan-Major, Iz Beltagy, Huu Nguyen, Lucile Saulnier, Samson Tan, Pedro Ortiz Suarez, Victor Sanh, Hugo Laurençon, Yacine Jernite, Julien Launay, Margaret Mitchell, Colin Raffel, Aaron Gokaslan, Adi Simhi, Aitor Soroa, Alham Fikri Aji, Amit Alfassy, Anna Rogers, Ariel Kreisberg Nitzav, Canwen Xu, Chenghao Mou, Chris Emezue, Christopher Klamm, Colin Leong, Daniel van Strien, David Ifeoluwa Adelani, and et al. 2022. BLOOM: A 176b-parameter open-access multilingual language model. *CoRR*, abs/2211.05100.

Thibault Sellam, Dipanjan Das, and Ankur P. Parikh. 2020. BLEURT: learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 7881–7892. Association for Computational Linguistics.

Md. Abu Bakar Siddik, Arman Shehabi, and Landon T. Marston. 2021. The environmental footprint of data centers in the united states. *Environmental Research Letters*, 16.

Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, February 4-9, 2017, San Francisco, California, USA*, pages 4444–4451. AAAI Press.

Emma Strubell, Ananya Ganesh, and Andrew McCallum. 2019. Energy and policy considerations for deep learning in NLP. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 3645–3650. Association for Computational Linguistics.

Carole-Jean Wu, Ramya Raghavendra, Udit Gupta, Bilge Acun, Newsha Ardalani, Kiwan Maeng, Gloria Chang, Fiona Aga Behram, Jinshi Huang, Charles Bai, Michael Gschwind, Anurag Gupta, Myle Ott, Anastasia Melnikov, Salvatore Candido, David Brooks, Geeta Chauhan, Benjamin Lee, Hsien-Hsin S. Lee, Bugra Akyildiz, Maximilian Balandat, Joe Spisak, Ravi Jain, Mike Rabbat, and Kim M. Hazelwood. 2022. Sustainable AI: environmental implications, challenges and opportunities. In *Proceedings of Machine Learning and Systems 2022, MLSys*

2022, Santa Clara, CA, USA, August 29 - September 1, 2022. mlsys.org.

An Yang, Quan Wang, Jing Liu, Kai Liu, Yajuan Lyu, Hua Wu, Qiaoqiao She, and Sujian Li. 2019. Enhancing pre-trained language representations with rich knowledge for machine reading comprehension. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 2346–2357. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with BERT. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Zhengyan Zhang, Xu Han, Zhiyuan Liu, Xin Jiang, Maosong Sun, and Qun Liu. 2019. ERNIE: enhanced language representation with informative entities. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 1441–1451. Association for Computational Linguistics.

A Example Appendix

This is a section in the appendix.