SYMBOLIC KNOWLEDGE DISTILLATION

Yuchen Cao, Xinyi Xie, Mason Ma, Nathan Nguyen, Dennis Tang
Duke University

ABSTRACT

Our project follows the *from-machine-to-corpus-to-machine* workflow outlined in West et al. (2021) for Symbolic Knowledge Distillation. We explore the design space and limitations of the student model, including various sizes, architectures, and data-augmentation methods in order to determine the optimal design for capturing commonsense knowledge. In addition, we train a critic model to automatically score the output of the student model, and to serve as a heuristic for beam search in order to improve performance without the need for a larger model. The effectiveness of the proposed method is demonstrated through experimentation and examples of rejected output sentences.

1 INTRODUCTION

We explore using symbolic knowledge distillation to solve commonsense reasoning tasks with small student models. A large pre-trained language model contains commonsense, which is knowledge about the world that all humans know. For example, if "Jack was disappointed to not see Santa Claus" humans can infer that it is sometime near Christmas. Recent research has employed a new workflow for symbolic knowledge distillation - the *from-machine-to-corpus-to-machine* (West et al., 2021) approach, where commonsense knowledge that is stored implicitly in a large model can be extracted explicitly to a knowledge graph and further used to train a student model with significantly fewer parameters. This method is different from traditional knowledge distillation.

However, there are still a lot of issues that need to be addressed. This paper adopted GPT-2 XL model as the student model and trained with the generated knowledge graph on the ATOMIC$^{20}_{20}$ dataset. While GPT-2 XL has "only" 1.5 billion parameters, which is significantly smaller compared to the GPT-3 Curie model with 6.7B and GPT-3 Davinci model with 175B parameters, it is still considered to be "too large" compared to models used on edge devices like MobileBERT (Sun et al., 2020) with only 25M parameters. Moreover, for symbolic knowledge distillation, no optimization techniques of the student model have been evaluated carefully. In other words, the design space of the student model has not been explored extensively as it is still unknown what structures of student models would prompt the most effective capture of commonsense.

In this project, we aim to explore the design space and limitations of the student model and find out the optimal design by experimenting with various sizes and structures of student models (e.g. GPT2-Large, fine-tuning). We show that the student model can learn common sense from the teacher model, and the smaller model such as GPT-1 can capture commonsense effectively after fine-tuning with knowledge graphs. Additionally, fine-tuning using ATOMIC data is an effective technique for improving the acceptance rate of inference corpora. It is also shown that fine-tuning on machine-generated ATOMIC data gives a significant improvement in the acceptance rate.

Furthermore, we attempt to use the critic model to automatically score the output sentences, that is, whether we should accept or reject the inferred sentence without requiring onerous manual labeling. We also explored the different usages of the critic model by using it as a heuristic for the beam search to improve the performance of the student model without the requirement of larger models.

2 RELATED WORK

One research stream that is related to our project is called Knowledge Distillation. Knowledge distillation refers to the process of transferring knowledge from a large model to a small model. Normally, deploying a large deep neural network is challenging and takes many resources. To tackle

this problem, knowledge distillation was proposed to transfer the knowledge from a large model (known as the teacher model) into training a much smaller model (known as the student model) without any significant loss in performance. This notion was firstly proposed by Hinton et al. (2015), by using the class probabilities of large models as soft targets. Recent work about Distilled BiLSTM (Tang et al., 2019) showed that by distilling BERT-large into a single-layer BiLSTM, the number of parameters was reduced by 100 times, and the speed was increased by 15 times. MiniLM (Wang et al., 2020b) introduced the role of teaching assistant. When the number of layers and dimensions of the student model is much smaller, we should first distill out a teaching assistant model with a small dimension but the same number of layers as the teacher's model, and then transfer the knowledge of the teaching assistant to the students. With the TA mechanism, the value relation transfer can make the student model imitate the teacher model in deep. All these previous work has shown that we could employ knowledge distillation to train the small-size student model by learning the knowledge from the larger teacher model, which is the foundation of our project.

Furthermore, it is a very promising and valuable idea to treat a pre-trained language model as an open, readily available knowledge base. This knowledge base has several advantages over artificially constructed knowledge bases, including being more flexible, easily accessible, and containing a wide variety of knowledge. There appears to be a lot of work being done on the Language Models as a Knowledge Bases paradigm right now. Petroni et al. (2019) proposed the idea of language models as a knowledge base (LM as KB) and creates a LAMA dataset to assess the model's capacity to store knowledge. He explored whether the language model has learned and stored any factual knowledge (subject-relation-object triad or question-answer pairs) that can be determined by pre-training on large text corpora. Jiang et al. (2020) suggested two types of methods for building prompts to further explore the ability of models to extract knowledge. For example, mining candidate prompts from large text corpus (Wikipedia). Hence, according to the extant work, we have chosen the pre-trained language model such as GPT-2 as our knowledge base.

The ultimate objective of our project is to address commonsense reasoning tasks with a small-size student model. Commonsense reasoning refers to a natural language processing (NLP) model's ability to capture information that is not explicitly written in the text. Wang et al. (2020a) showed how to build knowledge maps from pre-trained language models (such as BERT, GPT-2/3) without manual supervision. Papanikolaou & Pierleoni (2020) employed a method of data enhancement relation extraction by fine-tuning GPT-2. Chen et al. (2020) used the pre-trained language model RoBERTA-large to encode the question, and an answer selection awareness mechanism is proposed to integrate all implicit representations of previous modules. Besides, Moghimifar et al. (2020) considered the zero-shot sample learning, which used ATOMIC, which is a short text description containing 300,000 events and an if-then relationship. Large pre-trained language model(like GPT-3 (Brown et al., 2020), GPT-2(Radford et al., 2019)) has been shown to contain implicit commonsense representations, which can be extracted to build new knowledge graph (Bosselut et al., 2019), by finetuning through manually constructed CSKG(common sense knowledge graph) seed dataset like ATOMIC$^{20}_{20}$ (Hwang et al., 2021). Therefore, based on the above findings, we have used pre-trained language models such as GPT-2, the datasets such as ATOMIC$^{20}_{20}$, and several techniques such as fine-tuning to improve the commonsense reasoning performances with student model.

3 APPROACH

Our work is based on previous work of Symbolic Knowledge Distillation (West et al., 2021), in which the workflow of distilling commonsense knowledge from teacher model to student model was proposed

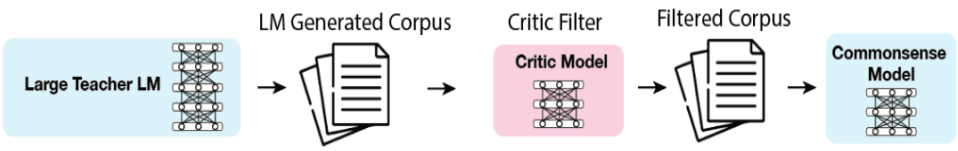

Figure 1: Workflow of Symbolic Knowledge Distillation

However, the authors didn't do a detailed analysis related to student models. In this project, we focus on exploring the design space and limitations related to different student model selections. We also tested the possibility of using a critic model to evaluate whether the generated commonsense knowledge is correct or not. Based on these, we proposed using the critic model to help the student model learn better.

3.1 BACKGROUND

In the previously mentioned work, the authors proposed a new workflow for symbolic knowledge distillation - *from-machine-to-corpus-to-machine*, where commonsense knowledge that is stored implicitly in a large model can be extracted explicitly to a knowledge graph and further used to train a student model with significantly fewer parameters. The method is different from traditional knowledge distillation.

Given a dataset that uses *events* as nodes and *relations* as edges to build a knowledge graph to represent commonsense knowledge, the distillation workflow is as follows as figure 1: **1** Query the teacher model with a list of events that have the same relations in a manually built commonsense knowledge graph to generate new events. **2.** Combine the new events, relations, and responses from the teacher model to the existing events to build prompts, as the input to the teacher model, in order to generate new responses to the new events and assemble new events and new responses with their relations as triplets. **3.** Train a critic model to filter out the new triplets with low scores and use the triplets with high scores to build a much bigger commonsense knowledge graph with higher quality. **4.** Train a student model with the new-generated larger commonsense knowledge graph.

3.2 METHOD

In this project, we evaluated how well different small-size student models capture commonsense knowledge, given different architectural design features such as model sizes, training data for fine-tuning, architectures, etc.

We generate different variants of GPT models as student models, including differences in size, pretrained-corpus, and structures. We perform fine-tuning on these models using ATOMIC$^{\mathbf{10x}}$ and ATOMIC$^{20}_{20}$ data, as well as augmentation variants of ATOMIC$^{\mathbf{10x}}$. For each student model after fine-tuning, we generated 500 inference corpora from 500 initially chosen event/relation prompts.

To examine the *quality* of knowledge in each generated corpus of the student models. We conduct a human evaluation with assisted code, labeling each inference corpus with one of five labels: "always/often", "sometimes/likely", "farfetched/never", "invalid", and "too unfamiliar to judge". Then we treat "always/often" and "sometines/likely" as acceptable, and other options as unacceptable. By hand-labeling these generated sentences with commonsense, we could compute the acceptance rate of different models, and make an evaluation based on that.

By finetuning different student models, we would like to answer the following questions:

1. Does the student model learn commonsense through finetuning?
2. Does ATOMIC$^{\mathbf{10x}}$ dataset help student models learn more commonsense compared with ATOMIC$^{20}_{20}$?
3. Does pretrained corpus affect student models?
4. What are the results like for student models with a different number of parameters?

To answer Question 1, we compared the acceptance rate of GPT-2 with and without finetuning on the commonsense dataset. To answer Question 2, we compared the acceptance rate of GPT-2 finetuned on ATOMIC$^{\mathbf{10x}}$ and ATOMIC$^{20}_{20}$. To answer Question 3, we compared the acceptance rate of original GPT-2 and GPT-2 finetuned on ArXiv dataset before. To answer Question 4, we compared the acceptance rate of GPT-2 model of different sizes, including GPT-2 small, GPT-2 medium, GPT-2 large and distilled GPT-2.

Besides, We tested two types of substitution on the training data: *changename* and *changerelation*. Originally, a corpus of ATOMIC$^{\mathbf{10x}}$ has a format of using subjects with the names `PersonX`, `PersonY`. For *changename* augmentation, we change `PersonX, PersonY` to `Alice, Bob`,

respectively. For *changerelation*, we replace relations with *explicit phrases*, e.g. **HinderedBy** $\rightarrow$ "is hindered by". We wanted to explore whether such a change could help student models learn more commonsense.

We also did experiments related to GPT-1 and GPTNeo, to explore the possible influence of GPT model design.

3.3 Extensions

3.3.1 Critic Model

Although human labeling can help us get the accurate acceptance rate of a model, it is highly time-consuming, and the results will be influenced by the different scorers. In this project, we would like to train a critic model, which can help us evaluate the generated sentence, whether it's acceptable (at least happen sometimes) or unacceptable (wrong or seldom happens). We used our dataset which was labeled to evaluate the acceptance rate of student models. We also compared the performance of different text-classification pre-trained models.

To help the critic model learn more difference between sentences and their connections with accepted or not. We generated 3000 more sentences using GPT-2, and human-labeled them. We combine them with previously labeled data to finetune the critic model.

3.3.2 Critic Model-Assisted Student Model Improvement

As mentioned above, we have manually evaluated the quality of the student models. With thousands of labeled corpora, we have also trained a critic model as mentioned in West et al. (2021). While we could follow the original method to use the critic model as a filter to remove the low-quality results generated by the teacher model, we also use the critic model to improve the quality of the student model to further unleash its potential.

All of the previously mentioned student models are using a greedy approach to generate texts. During generation, a student model starts with a prompt and then generates the next word based on the probability distribution of words that it learned during training. It then uses this predicted word as part of the context for generating the next word, and so on. This process continues until the model reaches the desired length of text or is stopped by the user. The greedy approach is fast and efficient, but it can lead to repetitive or nonsensical output if the model's predictions are not accurate.

We apply a novel approach to improve the performance of the student model without the need to increase its parameter size or re-training them as Figure 2. To improve the quality of the student model, we use the critic model to assist the student model during generation and apply beam search to generate multiple (k) outputs. Beam search is a method for generating multiple outputs from a language model, such as a GPT model, by considering multiple possible options at each step and selecting the top-k most likely outcomes. After generating k results, the results are then scored by the critic model. The critic model then determines which result is more likely to be favored by humans and returns that result to the user. Using this approach, we have tested the naive GPT2-medium student model and the critic model-assisted GPT2-medium student model with an unseen test set of 220 inferences and found that the student model now can generate more coherent and meaningful inferences.

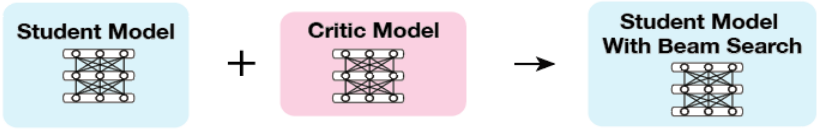

Figure 2: Critic model assisted student model improvement

4 EXPERIMENTS

4.1 DATASETS AND EVALUATION

Our project is based on ATOMIC$^{20}_{20}$ and ATOMIC$^{\textbf{10x}}$ dataset.

ATOMIC$^{20}_{20}$ (Hwang et al., 2021), is a human-authored commonsense knowledge graph. It exhibits a total of 23 relations, which the original author West et al. (2021) limited to only 7 relations. This should also be our main interest in using this commonsense knowledge graph for fine-tuning GPT-2 and obtaining a variety of student models. The 7 relations are **xAttr**: how X is perceived after *event*, **xReact**: how X reacts in response to *event*, **xEffect**: what X does after *event*, **xIntent**: X's intent in *event*, **xWant**: what X wants after *event*, **xNeeded**: what X needed for *event* to take place, **HinderedBy**: what X does that might hinder *event*.

ATOMIC$^{\textbf{10x}}$ is a model-generated commonsense knowledge graph built in West et al. (2021). Firstly, 100 high-quality events are selected from ATOMIC$^{20}_{20}$ that avoid grammatical or logical errors and minimize semantic overlaps. For each generation batch, 10 of these events are randomly sampled. For each event and each one of the 7 relations above, 10 resulting inferences are generated by Curie GPT-3 model, which yields 6.46M ATOMIC-style data triples. This is called ATOMIC$^{\textbf{10x}}$ since it contains more triples than ATOMIC$^{20}_{20}$ given the consideration of only 7 relations above.

We cleaned the ATOMIC$^{\textbf{10x}}$ dataset based on the score of sentences, by which we get a dataset of 2512720 lines. We used 500 for the basic test of student models, and 3000 as a test set for training critic models.

Each element in the training data has the following format: EventA **Relation** [GEN] EventB. An example of a corpus is as shown,

PersonX learns to use a PC **HinderedBy** [GEN] PersonX can't find a PC.

The way we calculate the acceptance rate for each model is by the formula

$$\text{Acceptance rate} = \frac{\#\text{corpus labeled with always/often or sometimes/likely}}{\#\text{all corpora}}$$

4.2 MODEL AND TRAINING DETAILS

The models of different structures, checkpoints we used as student models, and critic models are provided on Hugging Face. To train the model, we used Trainer provided by hugging face and finetuning the models in batched form. We used the default parameters for the trainer, the optimizer is AdamW, and the learning rate is 5e-5. For each model with the dataset, we finetune the model on the whole dataset for one time. For generating sentences, we used the pipeline provided by Hugging Face. The text generated by basic student models is selected with greedy search, for the critic model-assisted student model, we used beam search. For finetuning the critic model, we finetuned the model on a human-labeled dataset for 20 epochs. Most of our training work is done on Google Colab.

4.3 RESULTS

4.3.1 STUDENT MODELS' PERFORMANCE

After hand-labeling outcomes of student models, we obtained the resulting acceptance rate (as in table 1).

Model	Dataset	Acceptance rate
GPT-2 (124M)	no fine-tuning	4%
GPT-2 (124M)	ATOMIC^{10x}	78%
GPT-2 (124M)	ATOMIC$^{20}_{20}$	57.8%
GPT-2 medium (354M)	ATOMIC^{10x}	86%
GPT-2 large (774M)	**ATOMIC^{10x}**	**89.3%**
GPT-1	ATOMIC^{10x}	85.6%
GPT-2 (124M)	ATOMIC^{10x} changename	80.4%
GPT-2 (124M)	ATOMIC^{10x} changerelation	84.6%

Table 1: Acceptance rate for 8 different student models obtained from fine-tuning different GPT-2 models and different commonsense knowledge graphs.

4.3.2 CRITIC MODEL'S RESULTS

We finetuned different pretrained text-classification models provided by Hugging-face. The results are as Tabel 2. We can find the results are close.

Model	Accuracy	F1-socre	Precision	Recall
RobeRTa-base	0.84	0.91	0.84	0.99
Bert-base-uncased	0.83	0.91	0.85	0.98
Bert-base-cased	0.83	0.90	0.85	0.96

Table 2: Critic model results

Based on the critic model, we evaluate more student models and get the acceptance rate as table 3. For this part, the human-labeled result of GPT-2 is based on 3000 human-labeled data, the test set is different from before. Therefore, the result is different. We also compare the result of human labeling and critic model on the basic GPT-2 model. But for other models here, the acceptance rate was got only by the critic model, so we don't have a human-labeled acceptance rate for these models.

Model	Dataset	Acceptance rate (Human)	Acceptance rate (Critic)
new GPT-2 (124M)	ATOMIC^{10x}	84.3%	86.425%
Distill GPT-2 (82M)	ATOMIC^{10x}	-	88.425%
ArXiv GPT-2 (124M)	ATOMIC^{10x}	-	90.675%
GPT-Neo (125M)	ATOMIC^{10x}	-	91.575%

Table 3: Acceptance rate from critic model for other student models. new GPT-2 is the basic GPT-2 model on another test set, ArXiv GPT-2 is the basic GPT-2 model pretrained on ArXiv dataset.

4.3.3 CRITIC MODEL-ASSISTED STUDENT MODEL

We trained another separate critic model based on the `BERT-base-uncased` checkpoint. With the critic model, we build a pipeline where the critic model selects the best text from the candidates generated by the beam search. For comparison, we used the same GPT2-medium model with exactly the same pipeline and recorded the outputs from both models. We manually evaluated 220 pairs of outputs and grouped each pair into three categories: model 1 is better, model 2 is better, or roughly the same. The result shows that the GPT-2 model is able to generate outputs that are more favorable to human evaluators. The results are shown in Table 4.

	Counts	Acceptance rate
GPT2+Critic is better	158	72%
GPT2 is better	13	6%
Roughly the same	49	22%

Table 4: Acceptance rate for vanilla GPT-2 and critic model-assisted GPT-2.

5 ANALYSIS

5.1 ANALYSIS OF STUDENT MODELS & CRITIC MODELS

According to the results in table 1 and 3, we can answer our questions proposed in section 3.2 about student models. For question 1, we can find that GPT-2 alone does not capture enough commonsense knowledge in our task, since without fine-tuning with any commonsense knowledge graphs, the commonsense inference power is significantly small (4%). For question 2, we can find that fine-tuning small GPT-2 using human-authored commonsense knowledge graphs is shown to be less effective than fine-tuning using machine-generated commonsense graphs (57.8% versus 78%), which means the ATOMIC$^{\textbf{10x}}$ is shown to contain more commonsense knowledge and can help student models learn that. For question 3, we found that the GPT-2 pretrained on ArXiv performed better than the original GPT-2. It seems that pretrained model on commonsense corpus helps it learn knowledge. For question 4, it seems like in general the bigger the GPT-2 for the original model, the better the student model after fine-tuning, which indicated the target to minimize the student model is really hard. Besides, we can find that the results of GPT model with different structures are different.

Furthermore, we can find that performing augmentation on ATOMIC$^{\textbf{10x}}$ improves the acceptance rate when training on small GPT-2 (from 78% without any augmentation for ATOMIC$^{\textbf{10x}}$ to 80.4% for *changename* augmentation and 84.6% for *changerelation* augmentation). This might be a pathway to help student models learn.

For the critic models, we can find that the result is close to human labeling, it may be a way to evaluate the correctness of commonsense sentences.

5.2 ANALYSIS OF THE PERFORMANCE OF CRITIC MODEL-ASSISTED STUDENT MODEL

In this part, we analyze the outputs generated by both models.

One noticeable difference is the lengths from both outputs. The critic model tends to choose longer sentences and this is expected since a lot of the manually-labeled rejections are one-word outputs that do not fit into the context of the preceding event or are essentially meaningless. Longer sentences are more likely to be coherent and meaningful. For example, the improved GPT-2 generates "PersonX is pleased with his money, unlike a lot of people who don't have enough to pay their bills" for the event "PersonX withdraws $800 from the bank" whereas the vanilla model gives "surprised". Here the longer sentence is more meaningful and applies to broader cases.

In many cases, the critic model is able to improve the quality of the generated sentence and appeals to commonsense knowledge. For example, the improved model generates "PersonX is happy with his purchase" for the prompt "buys CDs at a record store xEffect", while the vanilla model just gives "PersonX becomes a musician" which was labeled as rejected by the human evaluators.

However, in some cases, the critic model is not able to improve the quality of the generated sentence. For example, the improved model gives "PersonX's family talks to her HinderedBy [GEN] PersonY is too busy with school and doesn't have time for them" while the vanilla model gives "PersonX's family talks to her HinderedBy [GEN] PersonX's family is afraid to talk to her". The vanilla model is able to capture the semantics of the preceding event and the critic model is not able to improve the quality of the sentence. This is likely due to the fact that the critic model is trained on a relatively small dataset that is generated by human beings. Besides, the dataset has duplicated entries because the same inputs are used to compare different student models. This might lead to overfitting of the critic model.

5.3 Analysis of accepted and rejected sentences

Acceptances fell into two primary categories - "blanket truths" and "context-specific". "Blanket truths" refer to generating phrases that could be likely for various events. For example, "PersonX organizes others to work is hindered by [GEN] PersonX is sick". In this instance, while being sick would indeed hinder one's ability to "organize others to work", the generated phrase "PersonX is sick " would hinder almost any event that precedes it. The second category of generated labels we denote as "context-specific" because the generated event includes some knowledge specific to the event that precedes it. For example "PersonX moves a towel out of the way, is hindered by [GEN] PersonX can't find the towel." About 20% of acceptances are "Blanket Truths".

For rejected sentences, the two primary reasons for rejection were grammatical and logical incoherence. A large majority of the rejections were due to logical incoherence such as in the phrase, "PersonX has an urge to run so [GEN] PersonX is seen as pushy." In a few instances, the generated phrase was able to partially capture the semantics of the preceding event such as "PersonX flies a kite so, as a result, [GEN] PersonX wants to be a pilot". However, in this case, while flying a kite is somewhat conceptually similar to flying an airplane as a pilot, most people who fly kites do not necessarily want to be pilots. Overall, 95% of rejections were due to logical incoherence.

The portion of blanket truth sentences is a non-trivial amount (around 20%). This indicates that while small-size student models have a great commonsense inference power, the resulting generations are not often meaningful. This might suggest some "naive" learning pattern of commonsense, for example, a large proportion of sentences with relation **HinderBy** are "blanket truths", since if EventA is of the form "PersonX do something", then the inference of EventB will be "PersonX don't do that thing", which makes sense, but it's trivial. It seems like the models are still on the side of capturing the semantics of the corpus to perform inference rather than using commonsense knowledge learned from the teacher model. It is unclear for now if this behavior is less observable in GPT-2 XL models described in West et al. (2021).

5.4 Limitation and Future Works

Our work is limited to the few relations we have in ATOMIC data, which could be generalized better if we extend to more relations. Further work is needed to obtain a better comparison between student models. For example, we can better explore the effect of human-authored ATOMIC$^{20}_{20}$ versus the machine-generated ATOMIC$^{\mathbf{10x}}$ over fine-tuning GPT-2 models if we train more baseline models with ATOMIC$^{20}_{20}$. Additionally, due to the limitation of our available computing resources, we are unable to perform a comprehensive hyperparameter search on the critic model to make it performs better in the hybrid model. We would also expect it to perform better by having a larger manually labeled dataset that indicates the flavor of human evaluators. Finally, the effect of data augmentation on ATOMIC$^{\mathbf{10x}}$ is limited to only training on small GPT-2, so it would be better if we explore this using other models.

6 Conclusion

We have investigated knowledge distillation from the perspective of student and critic models. Our work extends how symbolic knowledge distillation works with different small-size student models, and how well they can learn commonsense from the teacher model. Our results provide a greater perspective on optimization techniques and architectural designs that would give a well-performed student model that is less space-costly, which can have many practical impacts on small-scale devices. Our findings further extend from West et al. (2021) how different commonsense knowledge graphs impact the learning of student models and emphasize the significance of using symbolic distillation as an alternative to human-authored knowledge distillation.

Author Contributions

All team members contributed partly to the coding and writing part of the project. Mason and Yuchen mostly took charge of the design and workflow of the project, especially in coding experiments. Yuchen and Xinyi worked on data preprocessing. Nathan created a data-labeling platform. Dennis created the presentation and data visualizations.

REFERENCES

Antoine Bosselut, Hannah Rashkin, Maarten Sap, Chaitanya Malaviya, Asli Celikyilmaz, and Yejin Choi. Comet: Commonsense transformers for automatic knowledge graph construction. *arXiv preprint arXiv:1906.05317*, 2019.

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

Qianglong Chen, Feng Ji, Haiqing Chen, and Yin Zhang. Improving commonsense question answering by graph-based iterative retrieval over multiple knowledge sources. *arXiv preprint arXiv:2011.02705*, 2020.

Geoffrey Hinton, Oriol Vinyals, Jeff Dean, et al. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2(7), 2015.

Jena D Hwang, Chandra Bhagavatula, Ronan Le Bras, Jeff Da, Keisuke Sakaguchi, Antoine Bosselut, and Yejin Choi. (comet-) atomic 2020: On symbolic and neural commonsense knowledge graphs. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 6384–6392, 2021.

Zhengbao Jiang, Frank F Xu, Jun Araki, and Graham Neubig. How can we know what language models know? *Transactions of the Association for Computational Linguistics*, 8:423–438, 2020.

Farhad Moghimifar, Lizhen Qu, Yue Zhuo, Mahsa Baktashmotlagh, and Gholamreza Haffari. Cosmo: Conditional seq2seq-based mixture model for zero-shot commonsense question answering. *arXiv preprint arXiv:2011.00777*, 2020.

Yannis Papanikolaou and Andrea Pierleoni. Dare: Data augmented relation extraction with gpt-2. *arXiv preprint arXiv:2004.13845*, 2020.

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

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

Zhiqing Sun, Hongkun Yu, Xiaodan Song, Renjie Liu, Yiming Yang, and Denny Zhou. Mobilebert: a compact task-agnostic bert for resource-limited devices, 2020. URL https://arxiv.org/abs/2004.02984.

Raphael Tang, Yao Lu, Linqing Liu, Lili Mou, Olga Vechtomova, and Jimmy Lin. Distilling task-specific knowledge from bert into simple neural networks. *arXiv preprint arXiv:1903.12136*, 2019.

Chenguang Wang, Xiao Liu, and Dawn Song. Language models are open knowledge graphs. *arXiv preprint arXiv:2010.11967*, 2020a.

Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. *Advances in Neural Information Processing Systems*, 33:5776–5788, 2020b.

Peter West, Chandra Bhagavatula, Jack Hessel, Jena D Hwang, Liwei Jiang, Ronan Le Bras, Ximing Lu, Sean Welleck, and Yejin Choi. Symbolic knowledge distillation: from general language models to commonsense models. *arXiv preprint arXiv:2110.07178*, 2021.