A Study of Zero-shot Adaptation with Commonsense Knowledge

Jiarui Zhang JRZHANG@ISI.EDU
Information Sciences Institute, University of Southern California

Filip Ilievski ILIEVSKI@ISI.EDU
Information Sciences Institute, University of Southern California

Kaixin Ma KAIXINM@CS.CMU.EDU
Language Technologies Institute, Carnegie Mellon University

Jonathan Francis JON.FRANCIS@US.BOSCH.COM
Human-Machine Collaboration, Bosch Research Pittsburgh

Alessandro Oltramari ALESSANDRO.OLTRAMARI@US.BOSCH.COM
Human-Machine Collaboration, Bosch Research Pittsburgh

Abstract

Self-supervision with synthetic training data built from knowledge graphs has been proven useful to enhance the language model accuracy in zero-shot evaluation on commonsense reasoning tasks. Yet, since these improvements are reported in aggregate, little is known about how to *select* the appropriate knowledge for generalizable performance across tasks, how to *combine* this knowledge with neural language models, and how these pairings affect *granular task performance*. In this paper, we study the sensitivity of language models to knowledge sampling strategies, modeling architecture choices, and task properties. We evaluate the accuracy overall and in relation to four task properties: domain and vocabulary overlap between the train and the test data, answer similarity, and answer length. Our experiments show that: (i) encoder-decoder models benefit from more data to learn from, (ii) sampling strategies that balance across different aspects or focus on knowledge dimensions yield best accuracy, (iii) synthetic data is most effective for tasks with low domain overlap, and questions with short answers and dissimilar answer candidates, and (iv) our best T5 model reaches state-of-the-art results on zero-shot commonsense reasoning, narrowing the gap with supervised models, which is a side effect of our overall study.

1. Introduction

Common sense is the human knowledge about the world and the methods for making inferences from this knowledge [Davis, 2014]. Commonsense knowledge includes facts about events (including actions) and their effects, about knowledge and how it is obtained, about beliefs and desires, and facts about material objects and their properties [McCarthy, 1989]. Building open-world AI agents equipped with common sense that possess a wide range of everyday knowledge about naive physics, folk psychology, and causality is still an open challenge [Ma et al., 2019, Francis et al., 2022].

In recent years, large pre-trained language models (LMs) have been shown to perform well on commonsense reasoning [Devlin et al., 2018, Liu et al., 2019b] tasks based on developing one fine-tuned model for each benchmark. Recognizing that the assumption of benchmark-specific training data is unrealistic for open-world commonsense reasoning, the community has proposed lightweight alternatives to fine-tuning [Shin et al., 2020, Li and Liang, 2021] and has increasingly focused on zero- and few-shot tasks and reasoning models [Shwartz et al., 2020, Ma et al., 2021a]. Zero-shot reasoning methods rely on careful self-supervision of LMs with external resources: commonsense

KGs [Banerjee and Baral, 2020, Ma et al., 2021a], elicitation of pre-existing knowledge in the LM [Shwartz et al., 2020, Paranjape et al., 2021], or instruction-prompted training with a diverse set of tasks [Sanh et al., 2021]. State-of-the-art zero-shot performance has been obtained by adapting LMs with synthetic data from commonsense KGs [Ma et al., 2021a, Dou and Peng, 2022]. Yet, as these improvements are reported in aggregate, little is known about (i) how to *select* the appropriate knowledge for solid performance across tasks, (ii) how to *combine* this knowledge with neural language models, and (iii) how these pairings affect *granular task performance*.

This paper conducts an empirical study of the adaptation (self-supervision) of state-of-the-art LMs with KGs for zero-shot commonsense reasoning. Our contributions are: 1. A research design that captures key dependencies between the selected knowledge as synthetic data,[1] the LM, and the properties of the task, through five questions that have not been answered so far in such zero-shot evaluation setting. 2. Formal framework for self-supervision of LMs with synthetic data from commonsense KGs, which answers the research questions by supporting a wide variety of sampling strategies and sizes, language model variants, and meaningful task properties. 3. Rigorous experimentation of seven sampling strategies with seven knowledge sizes, five language models belonging to two representative model architectures, in relation to four properties of five commonsense reasoning tasks. We observe that: (i) encoder-decoder models benefit from more data to learn from, (ii) sampling strategies that balance across different aspects or focus on knowledge dimensions yield best performance, (iii) synthetic data is most effective for tasks with low domain overlap, and questions with short answers and dissimilar answer candidates, and (iv) our best T5 model reaches new state-of-the-art results on zero-shot commonsense reasoning, narrowing the gap with supervised models, which is a side effect of our overall study.

2. Research Questions

Here, we motivate each question and indicate the novelty introduced by studying the question for zero-shot commonsense reasoning with KGs.

RQ1: What is the overall impact of model and knowledge choices on the generalizability of self-supervision of LMs with KGs? Prior work on zero-shot commonsense reasoning with KGs [Ma et al., 2021a, Dou and Peng, 2022] has reported large gains over vanilla LMs across benchmarks. Yet, the gap between these results and the performance of supervised models remains large. It is unclear how much this gap can be bridged by other knowledge sampling strategies or LM architectural choices.

RQ2: How much data is needed to adapt LMs to commonsense reasoning tasks? Finding a right number of QA pairs to adapt a model with is crucial to reach optimal performance, prevent overfitting, and optimize efficiency. Ma et al. [2021a] report accuracy gains with a hand-selected KG subset, whereas Ilievski et al. [2021a] show that adapting language models with questions from certain knowledge dimensions is much more beneficial than others, and may fare better than using the entire set of questions. No prior work has performed systematic analysis of the relation between knowledge sample size and the zero-shot commonsense reasoning performance.

RQ3: How to best sample data for model adaptation? It is unclear what is the effect of the sampling strategy on the model performance. Sampling can focus on specific knowledge types or their optimal combination [Ilievski et al., 2021a], training time uncertainty [Swayamdipta et al., 2020], or

1. Here, the synthetic data is formulated as multiple choice questions generated by leveraging the structure of commonsense knowledge graphs.

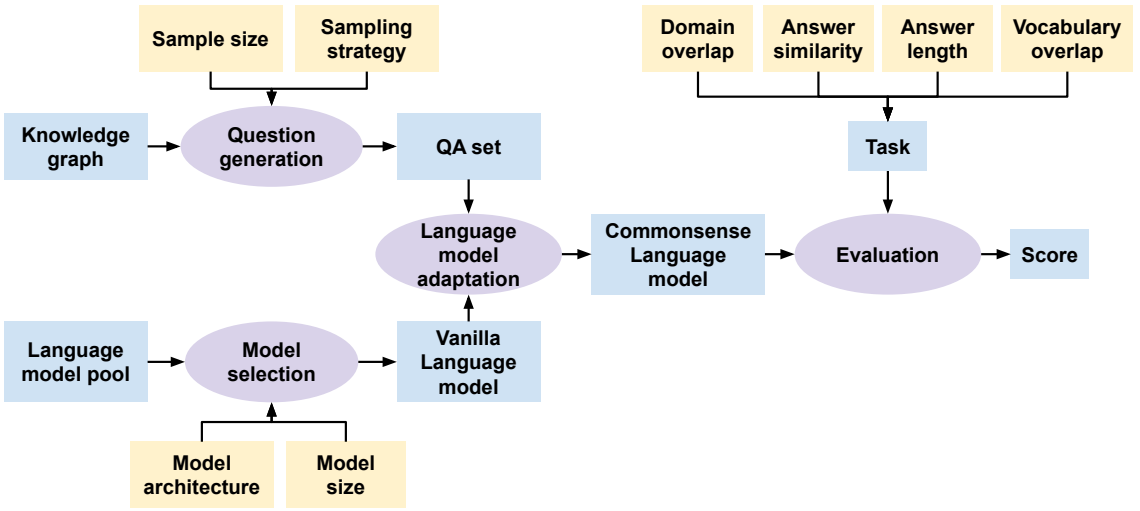

Figure 1: Overview of our study framework. The question generation step takes a KG as input, and yields a synthetic QA set. The QA set depends on the sampling size and strategy. In parallel, a LM is chosen out of a pool of models based on two factors: architecture and size. The selected vanilla LM is adapted based on the synthetic QA set, resulting in a commonsense LM. The accuracy score of the LM is evaluated on a task that is characterized by a degree of domain overlap, answer similarity, answer length, and vocabulary overlap.

confidence [Pleiss et al., 2020]. While corresponding strategies exist in the literature, they have not been applied to the task of zero-shot QA with KGs.

RQ4: Can models generalize well to tasks with low domain overlap? Knowledge-enhanced zero-shot models have higher impact on tasks that resemble the synthetic data, which is confirmed by the relatively larger gains obtained when using ConceptNet for CSQA and ATOMIC for SocialIQA, compared to using these sources on datasets like WinoGrande [Mitra et al., 2019, Ma et al., 2021a]. Whether models can generalize well to questions with low domain overlap is an open question.

RQ5: What is the connection between model's accuracy and properties of the task? Prior work has reported that models rely on spurious correlations, such as lexical properties, to answer questions [Gururangan et al., 2018, McCoy et al., 2019], and that fine-tuned models perform much better on questions that resemble the training data [Ma et al., 2021b]. Only considering the answer candidates may bring high performance on some tasks, indicating that the properties of the answer candidates have a large impact on the model [Li et al., 2021]. While it is intuitive that questions and answers with different properties (similarity, length, vocabulary, and overlap) may require different reasoning, these investigations have not been conducted for zero-shot commonsense evaluation.

3. Method

We follow the task formulation of *generalizable commonsense reasoning* proposed by Ma et al. [2021a]. The input consists of a natural language question Q and n candidate answers A_i, $|A_i| = n$. Exactly one of the candidate answers, marked with A, is correct. The remaining $(n-1)$ candidate

answers serve as distractors. As we assume a zero-shot setup, the models have no access to the benchmark-specific training data. Each model is adapted once, after which they are fixed, and directly applied on test partitions across benchmarks. We assume a knowledge-driven QA framework (Figure 1), where pre-trained LMs are adapted with artificial QA sets derived from KG data. This framework allows us to investigate the performance of LMs adapted with synthetic data from KGs as a function of the 1) size and architecture of the language model; 2) size and sampling strategies of the knowledge used for model adaptation; and 3) properties of the task, such as overlap with knowledge and answer length. Given a natural language question Q, and n candidate answers $\{A_1, ..., A_n\}$, the LM has to select the most probable answer A during training. Once the LM adaptation is done, the updated LM is applied across QA tasks in a zero-shot manner.

3.1 Language Models

Model architectures. We adopt two widely-used pre-trained models: RoBERTa [Liu et al., 2019a] and T5 [Raffel et al., 2019]. RoBERTa is an encoder-only masked language model (MLM), whereas T5 is an encoder-decoder model which converts tasks into text-to-text format. Following Ma et al. [2021a], for RoBERTa each input sequence is a concatenation of the question and one of its answer candidates. We mask one non-stop token in the sequence at a time, and compute the masked token's loss. We then take the averaged loss for the sequence and this is repeated for every answer candidates. We then train the model with the margin loss: $L = \frac{1}{n}\sum^{n}_{\substack{i=1 \\ i \neq y}} max(0, \eta - S_y + S_i)$, where S_y and S_i are the negative averaged loss for correct answer and distractor respectively. During inference, we take the candidate with highest score S as the answer. For T5, each input sequence is the same as RoBERTa except that we prefix it with a task-specific term "reasoning:" following the original paper [Raffel et al., 2019]. The model is pre-trained to generate "true" or "false" token. L_{true} represents the loss of the "true" token logits, while L_{false} represents the "false" token logits. For each answer candidate, we compute the score $S = L_{true} - L_{false}$ and use the same margin loss function as in RoBERTa to jointly predict the optimal candidate.[2]

Model sizes. We use RoBERTa's base and large models with 125M and 355M parameters, respectively. We experiment with three T5 models: small (60M parameters), large (740M), and 3b (2.85B).

3.2 Synthetic question generation

We sample knowledge from the CommonSense Knowledge Graph (CSKG) [Ilievski et al., 2021b], which combines seven commonsense knowledge resources under a shared representation, including ConceptNet [Speer et al., 2017], ATOMIC [Sap et al., 2019a], and Visual Genome [Krishna et al., 2017]. In total, CSKG contains over 7 million commonsense statements, consisting of head (h), relation (r), and tail (t). CSKG describes over 2 million nodes with 58 relations. Each knowledge statement in CSKG is categorized into one of 13 dimensions: lexical, similarity, distinctness, taxonomic, part-whole, creation, utility, comparative, quality, temporal, spatial, motivational, and relational-other [Ilievski et al., 2021a].

2. We also tried to score the answers individually, or to concatenate the question with all answer candidates, and teach the model to predict the position or make a copy of the right candidate, following [Khashabi et al., 2020]. These loss strategies performed consistently worse, and we leave them out of the paper.

Sample sizes. We use all relations of the CSKG subset that combines ATOMIC, ConceptNet, WordNet [Miller, 1995], Wikidata [Vrandečić and Krötzsch, 2014], and Visual Genome. We sample synthetic QA sets that correspond to $K\%$ from this knowledge set, $K \in \{1, 5, 10, 33, 50, 100\}$. In comparison, Ma et al. [2021a] use 100% of the data for 14 manually-selected semantic relations.

Sampling strategies. We experiment with seven strategies to sample the $K\%$ questions. (i) *Random* draws $K\%$ of the question pool by chance, without replacement. (ii) *Dimension* selects the questions that belong to a knowledge dimension. We evaluate the five most populous dimensions in CSKG: temporal, desire/goal, taxonomic, quality, and relational-other. For fair comparison, we limit the questions selected for a dimensions to the equivalent of $K\%$ of the entire question set. (iii) *Uniform* selects an equal number of questions from each of the thirteen dimensions, leading to a total of $K\%$ of the entire question set. (iv) *Vanilla-confidence* samples questions based on the confidence of the vanilla LM, i.e., before any adaptation. We experiment with selecting the questions with either lowest or highest confidence. (v) *Confidence* samples questions based on the mean LM confidence for the true label across the adaptation epochs. We first train a model on the entire QA set and record each question's training statistics as in [Swayamdipta et al., 2020]. We design two variants: confidence-low and confidence-high, analogous to the vanilla-confidence strategy. (vi) *Variability* detects the $K\%$ of the questions with extreme standard deviation for the true label across the adaptation epochs [Swayamdipta et al., 2020]. We experiment with variability-low and variability-high sampling. (vii) *Margin* selects the $K\%$ with the most extreme mean difference between the confidence of the correct answer and the incorrect ones [Pleiss et al., 2020]. We consider margin-low and margin-high sampling. For every strategy, $K = 100$ corresponds to the entire synthetic QA pool, while $K = 0$ is the vanilla pre-trained LM without adaptation.

3.3 Tasks

Task properties. We evaluate accuracy on five benchmarks for multiple-choice commonsense question answering, described in the Appendix of this paper. We compute granular model accuracy on task partitions based on four properties. (i) *Domain overlap* between the KG and the task. We refer to the necessary commonsense knowledge for solving a particular set of tasks as a *domain*. Two of the five benchmarks are known to have high domain overlap (HDO) with existing KGs [Mitra et al., 2019, Ma et al., 2021a]: CSQA has been devised based on knowledge in ConceptNet, while SocialIQA has been created based on the ATOMIC KG [Sap et al., 2019a]. The remaining three benchmarks have been created independently of the KGs, therefore, we consider them to have low domain overlap (LDO) with our KGs. We compare LM accuracy on the benchmarks with HDO and LDO. (ii) *Answer similarity (AS)* between the answer candidates. We partition the task into quartiles based on the Jaccard similarity between the tokens of the candidates A_i and A_j: $AS(q) = \frac{|T_{A_i} \cap T_{A_j}|}{|T_{A_i} \cup T_{A_j}|}$. Here, T_{A_i} and T_{A_j} are the set of tokens of candidates A_i and A_j, respectively. (iii) *Answer length (AL)*. We partition a task into quartiles based on the answer length, computed by summing the tokens T_{A_i} of the candidates A_i: $AL(q) = \sum_{i=1}^{n} |T_{A_i}|$. (iv) *Vocabulary overlap (VO)* between the task questions and the synthetic QA set. Given a task question, we compute the average frequency of the candidate tokens in the synthetic data. To increase the effect of the tokens with low frequency, we use the reciprocal value of the token frequencies: $VO(q) = \frac{1}{m} \sum_{k=1}^{m} \frac{1}{f(t_k)}$. Here, m is the number of tokens in the combination of the candidates ($|\bigcup_i^n T_{A_i}|$ for each question, t_k is the k-th token in the answer candidates, and $f(t_k)$ is its frequency in the synthetic data. When splitting the task based on *answer similarity*, *answer length*, and *vocabulary overlap*, we use RoBERTa's tokenizer, and

Table 1: Zero-shot results of our LMs with their optimal data size and sampling strategy, all the results are the average accuracy from 3 runs. We compare to the best versions of relevant baselines. '*' indicates that the average is computed on an incomplete set of benchmarks. Best results per column are shown in bold.

Model	LDO			HDO		Avg(LDO)	Avg(HDO)	Avg
	aNLI	WG	PIQA	SIQA	CSQA			
Majority [Ma et al., 2021a]	50.8	50.4	50.5	33.6	20.9	50.6	27.25	41.2
RoBERTa-large [Liu et al., 2019b]	65.5	57.5	67.6	47.3	45.0	63.5	46.1	56.6
COMET [Bosselut et al., 2019]	-	-	-	50.1	-	-	*50.1	*50.1
Self-Talk [Shwartz et al., 2020]	-	54.7	70.2	46.2	32.4	*62.5	39.3	50.9
SMLM [Banerjee and Baral, 2020]	65.3	-	-	48.5	38.8	*65.3	43.7	50.9
Ma et al. [Ma et al., 2021a]	70.5	60.9	72.4	63.2	67.4	67.9	65.3	66.8
Dou & Peng [Dou and Peng, 2022]	-	-	-	59.9	67.4	-	63.6	63.6
RoBERTa-base (ours)	59.9	53.1	65.7	54.6	53.6	59.6	54.1	57.4
RoBERTa-large (ours)	71.5	60.0	72.6	63.6	66.4	68.0	65.0	66.8
T5-small (ours)	50.6	51.6	56.2	42.3	36.4	52.8	39.4	47.4
T5-large (ours)	66.1	58.7	70.8	57.5	63.1	65.2	60.3	63.2
T5-3b (ours)	**76.6**	**71.0**	**76.7**	**65.3**	**69.9**	**74.7**	**67.6**	**71.9**
RoBERTa-large (supervised)	85.6	79.3	79.2	76.6	78.5	81.4	77.5	79.8
T5-3b (supervised)	87.5	84.4	76.3	78.6	81.5	82.7	80.1	81.7

we focus on the PIQA benchmark [Bisk et al., 2020] which has a variety of properties among its questions.

4. Results

Finding 1: Careful knowledge sampling and model design leads to state-of-the-art zero-shot accuracy across tasks. Table 1 shows the results obtained with the best performing knowledge sampling strategy (random) and the best data size per model architecture (5% for RoBERTa models, 33% for T5 models), all the results are the average accuracy from 3 runs. [3] We have observed that the variance across runs is within 1% accuracy points. The standard deviation statistics for all models is lower than 0.5 (see Table 5 in Appendix section). The relatively low variance means the results are consistent when we are doing random samplings for training data for different times. The results show that the best performance is clearly obtained with self-supervision of the model T5-3b with 33% of the data. The zero-shot result of this model is 71.9 on average over the five benchmarks, which is 15.3 points higher than the vanilla RoBERTa-large model and 5.1 points higher than the previous state-of-the-art result of Ma et al. [2021a]. This result is especially encouraging in comparison with the supervised RoBERTa-large and T5-3b LM, which is now only 7.9 and 9.8 points higher than our result, despite relying on benchmark-specific training data. Our second best model is RoBERTa-large, which is able to outperform the RoBERTa-large model in Ma et al. despite relying on only 5% of the training data. As expected, LM size correlates with accuracy, as T5-3b $>$ T5-large $>$ T5-small and RoBERTa-large $>$ RoBERTa-base.

Finding 2: The optimal synthetic data size depends on the LM size and architecture. Figure 2 shows the average accuracy of the LMs adapted with different synthetic data sizes. The encoder-only model, RoBERTa, benefits less from the increase of the synthetic data size. RoBERTa-large achieves

3. We did not post the average result of T5-3b due to the excessive time cost.

its optimum of 66.8% accuracy with only 5% of the data, whereas the accuracy of RoBERTa-base increases only marginally with more than 5% of the data.

The generative encoder-decoder LM, T5, benefits from more commonsense data. Especially the accuracy of our largest model, T5-3b, grows around 7% by increasing K from 5 to 33%. Yet, the performance of the T5 models generally peaks around 33%, and plateaus with the increase of the data size. We also plot the learning curves of RoBERTa and T5 models in Appendix, and we found that RoBERTa's adaptation loss to be consistently lower than T5. This is probably because RoBERTa is adapted with the similar objective as its pretraining (MLM), whereas T5 is adapted with a new prefix. This observation also helps explain why RoBERTa is more data efficient and achieves better results than T5 of similar size.

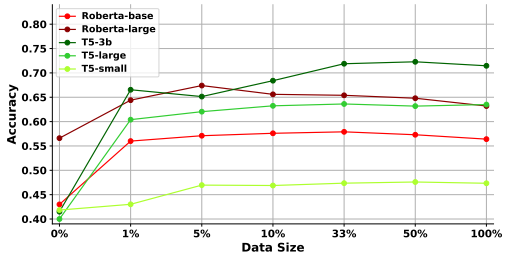

Figure 2: Evaluation of the LMs adapted with different sizes (K) of the synthetic data. We show the mean accuracy over the five benchmarks.

Table 2: Evaluation results on the five benchmarks of **T5-large** with different sampling strategies. All samples have equivalent sizes, corresponding to 5% of the training data. The best result per column is marked in bold. We focus on 5% of the data for computational reasons.

Strategy		LDO			HDO		Avg(LDO)	Avg(HDO)	Avg
		aNLI	WG	PIQA	SIQA	CSQA			
Random	5%	65.9	56.5	70.5	55.4	61.9	64.3	58.7	62.0
Dimension	temporal	66.6	56.4	**71.2**	54.9	**63.4**	**64.7**	**59.2**	**62.5**
	desire	64.4	57.9	69.6	55.9	62.2	64.0	59.1	62.0
	taxonomic	61.8	54.0	66.8	52.8	57.5	60.9	55.2	58.6
	quality	**66.8**	**58.4**	70.0	**56.4**	59.6	65.1	58.0	62.2
	rel-other	61.0	52.5	65.9	51.7	54.0	59.8	52.9	57.0
Uniform		65.3	57.5	69.2	56.6	62.7	64.0	59.7	62.3
Vanilla-conf	high	65.3	56.8	69.0	55.5	57.5	63.7	56.5	60.8
	low	64.0	56.0	68.1	52.0	59.6	62.7	55.8	59.9
Conf	high	62.9	53.8	66.5	53.9	57.0	61.1	55.5	58.8
	low	41.8	48.5	42.0	24.7	07.7	44.1	16.2	32.9
Varibility	high	64.0	54.6	65.1	51.1	54.5	61.2	52.8	57.9
	low	61.7	54.9	66.8	52.7	55.9	61.1	54.3	58.4
Margin	high	63.8	54.5	67.2	52.8	56.9	61.8	54.9	59.0
	low	41.5	45.0	43.7	24.1	09.1	43.4	16.6	32.7

Finding 3: Preserving the natural distribution of the data is the optimal sampling strategy. We study the impact of different sampling strategies on RoBERTa-large and T5-large, which have the same order of magnitude (hundreds of millions of parameters). The results in Table 2 show that random, uniform, and temporal sampling perform best.[4] The finding that random and uniform sampling lead to a strong and balanced model is consistent with the finding that random sampling of distractors is better than heuristic- and embedding-based strategies [Ma et al., 2021a], both emphasiz-

4. Similar result has been obtained for RoBERTa-large, see Appendix.

Table 3: Examples of benchmark questions that are correctly answered with only one model, which is adapted with dimension-based knowledge. (*) denotes the correct answer.

dimension: temporal Q:Jan went out with Quinn's friends and had a great time.What does Jan need to do before this? A1:get dressed(*); A2:cancel her plans; A3:see Quinn's Friends again
dimension: desire Q:Robert has no regret for punching Justin in the nose because _ was the victim of injustice. A1:Robert(*); A2:Justin
dimension: quality Q:What can machines do that humans cannot? A1:fail to work; A2:perform work; A3:answering questions; A4:see work; A5:fly(*)

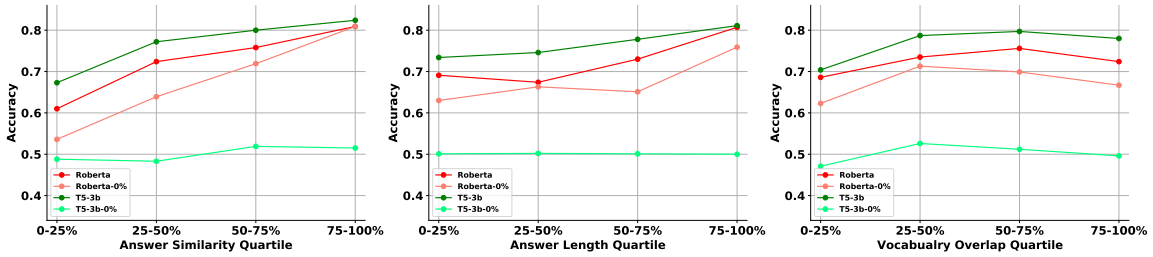

Figure 3: Accuracy of the best performing RoBERTa-large and T5-3b models in relation to the answer similarity, answer length, and vocabulary overlap between the data used for pretraining and testing.

ing the benefit of preserving the natural data distribution. Meanwhile, the strong performance of the dimension-based strategies, whose data samples are disjoint by design, indicates that LMs adapted on these dimensions capture complementary knowledge. As an illustration, Table 3 shows three benchmark questions which are only answered correctly with the most suitable dimension-based LM.

Finding 4: The superior accuracy of T5 owes mainly to better generalization to low domain overlap tasks. Table 1 shows that T5-3b's improvement over RoBERTa is on average 6.5% on the LDO benchmarks, but only 1.4% on the HDO benchmarks. This generalization ability of T5-3b can largely be attributed to the larger capacity of T5-3b, which allows it to represent additional knowledge and associations between terms. In addition, this Table shows that the HDO benchmarks have been much more popular in prior work, and much larger gains over the vanilla LM have been reported on them (up to 15.1 points on SIQA and 22.4 points on CSQA). Conversely, results on the LDO benchmark have rarely been reported in prior work on zero-shot commonsense reasoning, and the maximum improvement obtained in prior work is only 4.4 points on average across these benchmarks. Therefore, our accuracy improvement of 0.3 points for RoBERTa and 6.8 points for T5-3b is a notable leap towards robust performance on domains with low overlap.

Finding 5: Synthetic data is most effective for questions with short answers and dissimilar answer candidates. Figure 3 (left) shows that all models that use a discriminative loss (vanilla RoBERTa, and the adapted RoBERTa and T5) perform equally well on the questions with similar answers. The adaptation of T5 brings a large improvement on similar questions, which can be attributed to adapting its existing knowledge to the task at hand. For RoBERTa, the impact of

the synthetic data grows with the decrease of the answer similarity. On most dissimilar questions, RoBERTa benefits from the synthetic data by improving its accuracy by nearly 10 points, and T5 excels over RoBERTa by leveraging its higher capacity to learn more effectively from the synthetic data. Given that the data used for pre-training is designed to only include questions with non-overlapping answers, this finding is intuitive, and explains the source of improvement of performance reported in prior work [Ma et al., 2021a, Dou and Peng, 2022]. We observe an analogous outcome in terms of the answer length. Both RoBERTa and T5 perform best on questions with longer answers, while T5-3b is advantageous for short answers (Figure 3, middle). Notably, the synthetic QA data mostly consists of short answers, showing again that the performance gain of T5-3b owes to its capacity to extend original knowledge during the commonsense adaptation stage. Finally, we expect that questions with a higher vocabulary overlap will be easier for the models to learn with the synthetic data. Figure 3 (right) shows no clear correlation between vocabulary overlap and model accuracy. Further analysis should investigate whether this is an artifact of our method of computing vocabulary overlap, or the models are indeed insensitive to the task vocabulary.

5. Discussion

Our experiments show that the choice of LM size and architecture, as well as knowledge size and sampling strategy, affects the ability of models to answer commonsense questions across benchmarks. Encoder-decoder models benefit from more data to learn from, whereas sampling strategies that balance across different aspects yield best performance. Most of the accuracy gain with synthetic data adaptation came on tasks with low domain overlap, signifying strong generalization, and on questions with short answers and dissimilar answer candidates, owing to the synthetic data properties. Next, we revisit three key assumptions of this study, and provide an alternative inspired by the results of our experiments.

1. From a single model to mixture of models. Balanced sampling strategies (uniform and random) show very robust performance on these task, as they preserve the natural distribution between the different properties of the data. More specialized strategies, e.g., focusing on knowledge dimensions, perform well on subsets of the task, but under-perform on other subsets. This model specialization questions the assumption that a single zero-shot model is sufficient to perform optimally on different aspects of common sense, and suggests that model combinations, such as Mixture of models [Gururangan et al., 2021], might provide a more comprehensive and trustworthy commonsense model. This would entail, e.g., combining models from different dimensions, or models that capture complementary training dynamics.

2. From implicit to explainable zero-shot commonsense reasoning. In our current framework, the rich and diverse commonsense knowledge is taught to an LM through a large set of QA pairs. Given the simplicity of these questions, an implicit assumption of this study is that LMs can reverse engineer these questions to learn commonsense knowledge implicitly, and apply this newly acquired knowledge on unforeseen benchmarks, whose surface properties may be different, but their underlying commonsense knowledge may be largely shared. While this is a reasonable assumption, our commonsense models are black boxes, and they do not provide an explicit justification for their decisions. A natural extension of this work is to devise *explainable models*, i.e., models whose output includes the explicit reasoning steps associated with a predicted answer.

3. From textual-based question answering to more realistic tasks. A key aspect of our zero-shot framework is its generalization across QA tasks and knowledge domains. The gap between

zero-shot and fine-tuning performance is closing down, bringing a natural question: are zero-shot models, with an adaptation of neural techniques with background knowledge, able to generalize across knowledge domains and QA settings, or have they merely learned how to answer questions statistically? To address this question, we propose a shift towards more realistic tasks across QA settings and domain based on common sense, such as story understanding [Kalyanpur et al., 2020], dialogue modeling [Ghosal et al., 2021], and text-based games [Murugesan et al., 2021]. Recently, multi-modal question answering tasks also have progressed by contextualizing the questions in a visual setting [Zellers et al., 2019] or an embodied simulation [Das et al., 2018]. The rich prior work that focuses on these tasks has assumed the existence of benchmark-specific training data; zero-shot models have not been thoroughly explored.

6. Related Work

Generalizable Commonsense Reasoning. UNICORN [Lourie et al., 2021] investigates continual learning of commonsense knowledge from multiple benchmarks, ultimately aiming to perform well on all of the benchmarks. This work demonstrates that LMs can learn from commonsense benchmarks effectively and efficiently, reaching relatively high accuracy based on little training examples. Prefix-tuning [Li and Liang, 2021] adapts LMs, by keeping the model parameters intact, but extending them with a small set of additional parameters tuned separately for each benchmark. Rather than updating model parameters, Autoprompt [Shin et al., 2020] extends the model input with trigger tokens, which are updated during training. Such efforts share our vision to develop models that can generalize to multiple commonsense benchmarks simultaneously. However, they assume availability of training data, while we focus on zero-shot commonsense QA models.

Zero-shot Commonsense Reasoning. Zero-shot commonsense reasoning methods may elicit knowledge from pre-trained LMs, via self-talk clarification prompts [Shwartz et al., 2020] or by asking LMs to generate contrastive explanations [Paranjape et al., 2021]. As shown in prior work [Ma et al., 2021a, Dou and Peng, 2022], KG-based approaches achieve superior performance compared to pure LM-based methods for zero-shot commonsense QA. To use KGs for zero-shot pretraining and evaluation, Banerjee and Baral [2020] pre-train an LM to perform knowledge completion, whereas Bosselut et al. [2020] enhance the question based on knowledge completion models, and score an answer candidate in relation to the context, question, and generated knowledge. Our work is based on the framework of Ma et al. [2021a], which generates synthetic QA pairs from a consolidated KG to pre-train LMs. They investigate the impact of different loss functions and knowledge sources, showing that margin loss performs better than masked language modeling, and that more knowledge generally performs better, though this might change depending on the knowledge-task alignment. Dou and Peng [2022] extend this framework with several data transformation methods, out of which measuring consistency between different prompt versions performs best. Complementing these efforts, we perform a systematic study of model size and architecture, knowledge sampling and size, and task properties, which allows us to obtain new state-of-the-art results, clarify the contribution and interplay of different system components, and relate these contributions to task properties.

Model Generalization and Data Selection. Sen and Saffari [2020] analyzed LM's ability to generalize across five different QA datasets. Ma et al. [2021b] showed that models can have drastically different performances by fine-tuning on different subset of the data. Swayamdipta et al. [2020] proposed to select training instances based on models' confidence and variability, noting that training on less-confident examples is more beneficial for generalization. Follow-up

work [Ethayarajh et al., 2021] proposed an information-theoretic metric for estimating the difficult of a training example, treating as difficult the examples for which information is missing. Pleiss et al. [2020] propose to identify erroneous data points based on their rank in the area under the margin of a machine learning model. While prior work analyses model robustness by sub-sampling instances from the task's training set, we investigate the impact of knowledge selection, model selection, and task properties when models are adapted on large KGs for commonsense QA. Ilievski et al. [2021a] split synthetic data from KGs into 12 commonsense dimensions, revealing that some kinds of knowledge are much more useful for pre-training compared to others. Our study provides a comprehensive study framework that consolidates prior efforts on the task of zero-shot commonsense QA with KGs.

7. Conclusions

Building robust AI agents with common sense requires in-depth understanding of the strengths and weaknesses of current zero-shot adaptation methods. We designed a framework to study systematically the effect of synthetic knowledge, model choices, and task properties on the generalization of LMs across commonsense QA tasks. Our best model improved over prior best zero-shot performance by 5 points, obtaining new state-of-the-art results, and narrowing the gap with supervised models. Using the same LM as prior work still performs better despite using less synthetic data. Closer analysis revealed that optimal knowledge size and sampling strategy is model-dependent, with encoder-only models learning quicker from less data than encoder-decoder models. Interestingly, strategies that perform balanced knowledge sampling led to robust performance. Strategies that focus on semantic data dimensions also performed well as their questions are more challenging for the models. Most of the improvement with synthetic data adaptation came on tasks with low domain overlap, signifying strong generalization, and on questions with short answers and dissimilar answer candidates, owing to the synthetic data properties. These findings point to three key directions for future work that uses self-supervision with large KGs to create generalizable commonsense reasoning agents: creating mixtures of specialized commonsense models, explainable zero-shot reasoning, and shift in evaluation to more realistic tasks like story completion and embodied QA. All our code and data can be downloaded at https://github.com/saccharomycetes/commonsense-with-KG.

References

Pratyay Banerjee and Chitta Baral. Self-supervised knowledge triplet learning for zero-shot question answering. *ArXiv*, abs/2005.00316, 2020.

Chandra Bhagavatula, Ronan Le Bras, Chaitanya Malaviya, Keisuke Sakaguchi, Ari Holtzman, Hannah Rashkin, Doug Downey, Scott Wen-tau Yih, and Yejin Choi. Abductive commonsense reasoning. *arXiv preprint arXiv:1908.05739*, 2019.

Yonatan Bisk, Rowan Zellers, Ronan Le Bras, Jianfeng Gao, and Yejin Choi. PIQA: Reasoning about Physical Commonsense in Natural Language. In *Thirty-Fourth AAAI Conference on Artificial Intelligence*, pages 7432–7439, 2020.

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

Antoine Bosselut, Ronan Le Bras, and Yejin Choi. Dynamic neuro-symbolic knowledge graph construction for zero-shot commonsense question answering, 2020.

Abhishek Das, Samyak Datta, Georgia Gkioxari, Stefan Lee, Devi Parikh, and Dhruv Batra. Embodied question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

Ernest Davis. *Representations of commonsense knowledge*. Morgan Kaufmann, 2014.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

Zi-Yi Dou and Nanyun Peng. Zero-shot commonsense question answering with cloze translation and consistency optimization. *arXiv preprint arXiv:2201.00136*, 2022.

Kawin Ethayarajh, Yejin Choi, and Swabha Swayamdipta. Information-theoretic measures of dataset difficulty. *arXiv preprint arXiv:2110.08420*, 2021.

Jonathan Francis, Nariaki Kitamura, Felix Labelle, Xiaopeng Lu, Ingrid Navarro, and Jean Oh. Core challenges in embodied vision-language planning. *Journal of Artificial Intelligence Research*, 74: 459–515, 2022.

Deepanway Ghosal, Pengfei Hong, Siqi Shen, Navonil Majumder, Rada Mihalcea, and Soujanya Poria. Cider: Commonsense inference for dialogue explanation and reasoning. *arXiv preprint arXiv:2106.00510*, 2021.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel R Bowman, and Noah A Smith. Annotation artifacts in natural language inference data. *arXiv preprint arXiv:1803.02324*, 2018.

Suchin Gururangan, Mike Lewis, Ari Holtzman, Noah A Smith, and Luke Zettlemoyer. Demix layers: Disentangling domains for modular language modeling. *arXiv preprint arXiv:2108.05036*, 2021.

Filip Ilievski, Alessandro Oltramari, Kaixin Ma, Bin Zhang, Deborah L McGuinness, and Pedro Szekely. Dimensions of commonsense knowledge. *Knowledge-Based Systems*, 2021a.

Filip Ilievski, Pedro Szekely, and Bin Zhang. Cskg: The commonsense knowledge graph. In *Extended Semantic Web Conference (ESWC)*, 2021b.

Aditya Kalyanpur, Tom Breloff, David Ferrucci, Adam Lally, and John Jantos. Braid: Weaving symbolic and neural knowledge into coherent logical explanations. *arXiv preprint arXiv:2011.13354*, 2020.

Daniel Khashabi, Sewon Min, Tushar Khot, Ashish Sabharwal, Oyvind Tafjord, Peter Clark, and Hannaneh Hajishirzi. Unifiedqa: Crossing format boundaries with a single qa system. *arXiv preprint arXiv:2005.00700*, 2020.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123(1):32–73, 2017.

Xiang Lisa Li and Percy Liang. Prefix-tuning: Optimizing continuous prompts for generation, 2021.

Xiang Lorraine Li, Adhi Kuncoro, Cyprien de Masson d'Autume, Phil Blunsom, and Aida Nematzadeh. A systematic investigation of commonsense understanding in large language models. *arXiv preprint arXiv:2111.00607*, 2021.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *ArXiv*, abs/1907.11692, 2019a.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019b.

Nicholas Lourie, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. Unicorn on rainbow: A universal commonsense reasoning model on a new multitask benchmark. *arXiv preprint arXiv:2103.13009*, 2021.

Kaixin Ma, Jonathan Francis, Quanyang Lu, Eric Nyberg, and Alessandro Oltramari. Towards generalizable neuro-symbolic systems for commonsense question answering. In *Proceedings of the First Workshop on Commonsense Inference in Natural Language Processing*, pages 22–32, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-6003. URL https://aclanthology.org/D19-6003.

Kaixin Ma, Filip Ilievski, Jonathan Francis, Yonatan Bisk, Eric Nyberg, and Alessandro Oltramari. Knowledge-driven Data Construction for Zero-shot Evaluation in Commonsense Question Answering. In *AAAI*, 2021a.

Kaixin Ma, Filip Ilievski, Jonathan Francis, Satoru Ozaki, Eric Nyberg, and Alessandro Oltramari. Exploring strategies for generalizable commonsense reasoning with pre-trained models. *EMNLP 2021*, 2021b.

John McCarthy. Artificial intelligence, logic and formalizing common sense. In *Philosophical logic and artificial intelligence*, pages 161–190. Springer, 1989.

R Thomas McCoy, Ellie Pavlick, and Tal Linzen. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. *arXiv preprint arXiv:1902.01007*, 2019.

George A Miller. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11): 39–41, 1995.

Arindam Mitra, Pratyay Banerjee, Kuntal Kumar Pal, Swaroop Mishra, and Chitta Baral. Exploring ways to incorporate additional knowledge to improve natural language commonsense question answering. *arXiv preprint arXiv:1909.08855*, 2019.

Keerthiram Murugesan, Mattia Atzeni, Pavan Kapanipathi, Pushkar Shukla, Sadhana Kumaravel, Gerald Tesauro, Kartik Talamadupula, Mrinmaya Sachan, and Murray Campbell. Text-based rl agents with commonsense knowledge: New challenges, environments and baselines. In *AAAI*, pages 9018–9027, 2021.

Bhargavi Paranjape, Julian Michael, Marjan Ghazvininejad, Hannaneh Hajishirzi, and Luke Zettlemoyer. Prompting contrastive explanations for commonsense reasoning tasks. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4179–4192, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.findings-acl.366. URL https://aclanthology.org/2021.findings-acl.366.

Geoff Pleiss, Tianyi Zhang, Ethan R Elenberg, and Kilian Q Weinberger. Identifying mislabeled data using the area under the margin ranking. *arXiv preprint arXiv:2001.10528*, 2020.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv preprint arXiv:1910.10683*, 2019.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. Winogrande: An adversarial winograd schema challenge at scale. *arXiv preprint arXiv:1907.10641*, 2019.

Victor Sanh, Albert Webson, Colin Raffel, Stephen H. Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Teven Le Scao, Arun Raja, Manan Dey, M Saiful Bari, Canwen Xu, Urmish Thakker, Shanya Sharma Sharma, Eliza Szczechla, Taewoon Kim, Gunjan Chhablani, Nihal Nayak, Debajyoti Datta, Jonathan Chang, Mike Tian-Jian Jiang, Han Wang, Matteo Manica, Sheng Shen, Zheng Xin Yong, Harshit Pandey, Rachel Bawden, Thomas Wang, Trishala Neeraj, Jos Rozen, Abheesht Sharma, Andrea Santilli, Thibault Fevry, Jason Alan Fries, Ryan Teehan, Stella Biderman, Leo Gao, Tali Bers, Thomas Wolf, and Alexander M. Rush. Multitask prompted training enables zero-shot task generalization, 2021.

Maarten Sap, Ronan Le Bras, Emily Allaway, Chandra Bhagavatula, Nicholas Lourie, Hannah Rashkin, Brendan Roof, Noah A. Smith, and Yejin Choi. ATOMIC: an atlas of machine commonsense for if-then reasoning. In *Proc. of AAAI*, pages 3027–3035, 2019a.

Maarten Sap, Hannah Rashkin, Derek Chen, Ronan Le Bras, and Yejin Choi. Social IQa: Commonsense reasoning about social interactions. In *Proc. of EMNLP-IJCNLP*, pages 4463–4473, November 2019b.

Priyanka Sen and Amir Saffari. What do models learn from question answering datasets? In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2429–2438, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.190. URL https://www.aclweb.org/anthology/2020.emnlp-main.190.

Taylor Shin, Yasaman Razeghi, Robert L. Logan IV, Eric Wallace, and Sameer Singh. AutoPrompt: Eliciting knowledge from language models with automatically generated prompts. In *EMNLP*, 2020.

Vered Shwartz, Peter West, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. Unsupervised commonsense question answering with self-talk. *ArXiv*, abs/2004.05483, 2020.

Robyn Speer, Joshua Chin, and Catherine Havasi. Conceptnet 5.5: An open multilingual graph of general knowledge. In *Proc. of AAAI*, AAAI'17, page 4444–4451, 2017.

Swabha Swayamdipta, Roy Schwartz, Nicholas Lourie, Yizhong Wang, Hannaneh Hajishirzi, Noah A Smith, and Yejin Choi. Dataset cartography: Mapping and diagnosing datasets with training dynamics. *arXiv preprint arXiv:2009.10795*, 2020.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. CommonsenseQA: A question answering challenge targeting commonsense knowledge. In *Proc. of NAACL*, pages 4149–4158, June 2019.

Denny Vrandečić and Markus Krötzsch. Wikidata: a free collaborative knowledgebase. *Communications of the ACM*, 57(10):78–85, 2014.

Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. From recognition to cognition: Visual commonsense reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

Implementation

For T5 training, we add the prefix "reasoning:" in front of every concatenation of question and answer, then ask the model to predict "1" for true, and "2" for false.
Regarding libraries, we used python 3.7.10, pytorch 1.9.0 and transformers 4.11.3.
Among all the training sets, we are using learning rate of $1e^{-5}$, batch size of 32, weight decay 0.01, training epochs of 5, adam-epsilon of $1e^{-6}$, $\beta 1 = 0.9, \beta 2 = 0.98$, warm-up proportion of 0.05, margin of 1.0.
For CPUs, we used Intel(R) Xeon(R) Gold 5217 CPU @ 3.00GHz (32 CPUs, 8 cores per sockets, 263GB ram).
For GPUs, we used Nvidia Quadro RTX 8000, and Nvidia Geforce 2080Ti.

Benchmarks

CommonsenseQA (CSQA) [Talmor et al., 2019] is a five-choice question answering benchmark which evaluates a broad range of common sense aspects. *SocialIQA (SIQA)* [Sap et al., 2019b] is a three-choice QA benchmark that requires reasoning about social interactions. *Abductive NLI (aNLI)* [Bhagavatula et al., 2019] is formalized as natural language inference, where, given the beginning and the ending of a story, the task is to choose the more plausible hypothesis out of two options. *PhysicalIQA (PIQA)* [Bisk et al., 2020] is a binary choice task, which tests the ability of models to employ physical reasoning. *WinoGrande (WG)* [Sakaguchi et al., 2019] is a binary choice anaphora resolution task.

Training curves

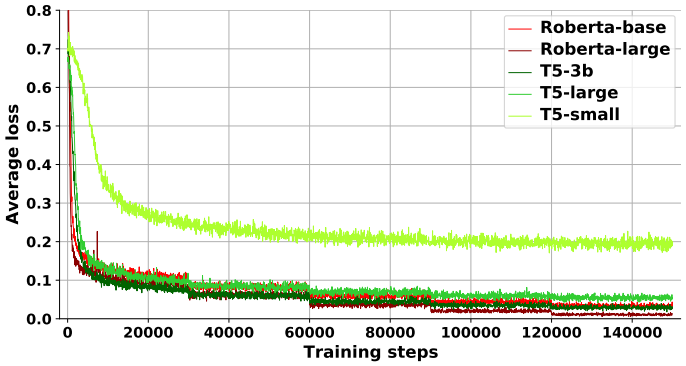

Figure 4: Training curves of the models: RoBERTa-base, RoBERTa-large, T5-small, T5-large, and T5-3b. We use all of the models with 100% of our training data.

Effect of sampling strategies on RoBERTa

The result in table 4 shows that the random, uniform and some dimensions perform best. It is worth noting that the temporal dimension perform better than random in aNLI task, which is highly related

to the time order, showing that specialized sampling strategies focusing on knowledge dimensions perform well on particular subset of tasks.

Table 4: Evaluation results on the five benchmarks of **RoBERTa-large** with different sampling strategies. All samples have equivalent sizes, corresponding to 5% of the training data. The best result per column is marked in bold. We focus on 5% of the data for computational reasons.

Strategy		**LDO**			**HDO**		**Avg(LDO)**	**Avg(HDO)**	**Avg**
		aNLI	**WG**	**PIQA**	**SIQA**	**CSQA**			
Random	5%	71.5	60.0	**72.6**	**63.6**	**66.4**	68.0	**65.0**	**66.8**
Dimension	temporal	**72.7**	61.1	72.1	62.3	65.8	**68.6**	64.1	**66.8**
	desire	70.2	59.5	72.4	60.9	64.3	67.4	62.6	65.5
	taxonomic	67.0	58.0	69.2	51.0	59.0	64.7	55.0	60.8
	quality	71.3	**61.8**	72.0	58.5	64.6	68.4	61.6	65.6
	rel-other	65.3	55.5	69.7	51.5	58.1	63.5	54.8	60.0
Uniform		69.6	58.0	72.4	61.7	64.3	66.7	63.0	65.2
Vanilla-conf	high	63.3	59.1	67.6	49.4	47.2	63.3	48.3	57.3
	low	57.9	51.9	55.6	33.1	21.7	55.1	27.4	44.0
Conf	high	66.2	58.9	70.3	59.4	62.2	65.1	60.8	63.4
	low	71.4	59.2	72.1	62.6	65.7	67.6	64.2	66.2
Varibility	high	67.4	56.8	65.5	48.2	44.0	63.2	46.1	56.4
	low	65.4	56.0	68.6	54.4	61.0	63.3	57.7	61.1
Margin	high	67.1	58.2	70.7	60.1	62.3	65.3	61.2	63.7
	low	72.3	60.5	71.2	62.7	65.0	68.0	63.9	66.3

Table 5: Standard deviations of 4 models training on random sampled data in our main experiment, all of them are relatively small. That means the results are consistent when we are doing random samplings for training data for different times. We did not run three trials for T5-3b due to the excessive time cost.

models	Roberta-base	Roberta-large	T5-base	T5-large
standard deviations	0.330	0.511	0.170	0.236

The effect of task properties in relation to the synthetic data size

Table 6 shows that both models perform better on the questions with dissimilar answers when they are trained with more data. The models perform optimal on the questions with similar answers with less data. This confirms our explanation that the synthetic data directs the models towards better performance on the questions with dissimilar answers. Furthermore, we see that T5 is able to exploit maximum amount of data for short answers, which is expected, given that most of the synthetic questions are relatively short. When it comes to longer answers, T5 performs best with less data, which indicates that the pre-training data has limited utility for this set of questions. Curiously, this pattern is not observed for RoBERTa - RoBERTa is unable to leverage more than 1% of the data to improve its performance on the questions with short answers. We hypothesize that this is due to the

Table 6: Evaluation results on the similarity, length, and vocabulary overlap quartiles of PIQA data for the models RoBERTa and T5-3b with different data sizes. Best results per model and similarity quartile are marked in bold.

Model	Data Size	Similarity				Length				Vocabulary overlap			
		25%	**50%**	**75%**	**100%**	**25%**	**50%**	**75%**	**100%**	**25%**	**50%**	**75%**	**100%**
Roberta	0%	53.6	63.9	71.9	**80.9**	63.0	66.3	65.1	75.9	62.3	71.3	69.9	66.7
	1%	56.6	**73.5**	75.6	78.7	**68.8**	**69.6**	68.0	78.0	**68.6**	71.3	74.3	70.2
	5%	**60.3**	72.4	**76.5**	80.4	66.7	68.3	**72.1**	**82.6**	**68.6**	**73.5**	**74.9**	**72.6**
	10%	58.2	71.1	73.2	79.8	67.1	65.9	70.2	79.1	68.0	72.4	70.6	71.3
	33%	58.8	72.0	74.7	78.0	68.0	68.7	70.6	76.3	66.9	71.7	72.8	72.2
	50%	57.1	70.4	73.9	80.2	66.4	66.1	71.2	77.8	65.8	70.0	**74.9**	70.9
	100%	55.6	68.3	69.3	74.8	62.7	65.2	66.9	73.0	63.4	65.9	71.9	66.7
T5-3b	0%	48.8	48.3	51.9	51.5	50.1	50.2	50.1	50.0	47.1	52.6	51.2	49.6
	1%	61.7	73.9	73.6	78.7	71.2	70.0	70.2	76.5	68.2	70.4	76.9	72.4
	5%	60.3	72.4	71.9	79.3	68.4	68.7	67.8	79.1	64.1	70.7	77.6	71.7
	10%	65.4	75.9	75.4	82.6	70.6	74.8	72.5	81.3	69.9	74.1	77.8	77.4
	33%	67.3	**77.2**	80.0	82.4	**73.4**	74.6	**77.8**	81.1	70.4	78.7	79.7	78.0
	50%	**68.8**	**77.2**	**80.2**	**84.3**	73.2	**76.3**	77.3	**83.7**	72.1	**79.1**	**80.2**	**79.1**
	100%	68.2	76.1	78.2	84.1	72.3	71.9	76.5	79.3	**78.9**	77.2	75.8	76.7

limited model capacity of RoBERTa, causing limited ability to store additional knowledge from the synthetic data. We do not see a clear correlation between vocabulary overlap and model accuracy across different data sizes.