Is Prompt Transfer Always Effective? An Empirical Study of Prompt Transfer for Question Answering

Minji Jung[1,3†§], Soyeon Park[2†], Jeewoo Sul[2,3†§], and Yong Suk Choi[1,2*]

[1]Department of Intelligence and Convergence Hanyang University, Seoul, Korea
[2]Department of Computer Science Hanyang University, Seoul, Korea
[3]LG Electronics
{jminji98, ssoyaavv, jeewoo25, cys}@hanyang.ac.kr

Abstract

Prompt tuning, which freezes all parameters of a pre-trained model and only trains a soft prompt, has emerged as a parameter-efficient approach. For the reason that the prompt initialization becomes sensitive when the model size is small, the prompt transfer that uses the trained prompt as an initialization for the target task has recently been introduced. Since previous works have compared tasks in large categories (e.g., summarization, sentiment analysis), the factors that influence prompt transfer have not been sufficiently explored. In this paper, we characterize the question answering task based on features such as answer format and empirically investigate the transferability of soft prompts for the first time. We analyze the impact of initialization during prompt transfer and find that the train dataset size of source and target tasks have the influence significantly. Furthermore, we propose a novel approach for measuring catastrophic forgetting and investigate how it occurs in terms of the amount of evidence. Our findings can help deeply understand transfer learning in prompt tuning[1].

1 Introduction

Advances in large language models (LLMs) (Devlin et al., 2018; Brown et al., 2020; Raffel et al., 2020) have continued to be made since the advent of the Transformer (Vaswani et al., 2017). As LLMs grow larger and larger, prompt tuning (Lester et al., 2021) is introduced to reduce the computational costs. This approach, which freezes all parameters of a pre-trained model and only trains a soft prompt, requires updating fewer parameters than fine-tuning while achieving comparable performance in many natural language processing (NLP) systems.

However, especially in model sizes below 11B parameters, prompt initialization causes performance differences. Recently, prompt transfer (PoT) was proposed in SPoT (Vu et al., 2022) as a way to better initialize prompts, in which a prompt embedding trained for a source task is used for initialization before training the target prompt. TPT_{TASK} (Su et al., 2022) claims that the performance is effective when initialized with the best zero-shot prompt. Several studies modified the prompt learning process to improve performance (Li et al., 2022; Asai et al., 2022; Zhong et al., 2022; Wang et al., 2023; Xie et al., 2023). The effectiveness of these approaches achieves better or comparable performance with prompt tuning and fine-tuning.

Nevertheless, previous studies unexplored the factors influencing transferability and only focused on large categories of tasks. Therefore, our goal is not only to refine the categorization of Question Answering (QA) tasks but also to investigate the impact on prompt transferability.

Our study is the first to examine PoT across QA datasets, and we report four important findings: (1) Transferability has different trends for each target task. (2) Initialization with the prompt that has high cosine similarity or high zero-shot performance does not always guarantee positive transferability. (3) Transferability is related to the difference in the train dataset size between the source and target tasks. (4) We identify the conditions for catastrophic forgetting to occur from an amount of evidence perspective and propose a new method to measure it.

2 Preliminary

2.1 Formulation

Similar to T5 (Raffel et al., 2020), we applied a text-to-text approach to the QA task. Given N train data,

† Equal contributions. Alphabetical order.
* Corresponding author.
§ This work was conducted while the authors were at Hanyang University. Currently, Minji Jung and Jeewoo Sul are working at LG Electronics.
[1]We release our code and prompt checkpoints at https://github.com/ailab-prompt-transfer/qa_prompt_transfer.

Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers), pages 528–539
June 16-21, 2024 ©2024 Association for Computational Linguistics

Dataset	Answer format	Amount of evidence	Train	Valid	Test
DuoRC (Saha et al., 2018)	Freeform	Partial	60,094	12,845	12,415
NQ-Open (Lee et al., 2019)	Freeform	**No**	79,132	8,793	3,610
WQ (Berant et al., 2013)	Freeform	**No**	3,400	378	2,032
MRQA-NewsQA (Trischler et al., 2017)	Extractive	*Single*	66,744	7,416	4,212
SQuAD (Rajpurkar et al., 2016)	Extractive	*Single*	78,839	8,760	10,570
BoolQ (Clark et al., 2019)	Categorical	*Single*	8,484	943	3,270
MultiRC (Khashabi et al., 2018)	Categorical	*Single*	24,518	2,725	4,848
TQA (Joshi et al., 2017)	Freeform	Partial	78,859	8,763	11,313
CosmosQA (Huang et al., 2019)	Multi-choice	Partial	22,735	2,527	2,985
SIQA (Sap et al., 2019)	Multi-choice	Partial	30,069	3,341	1,954
SQuAD w/o ctx	Freeform	**No**	78,839	8,760	10,570
BoolQ w/o ctx	Categorical	**No**	8,484	943	3,270
MultiRC w/o ctx	Categorical	**No**	24,518	2,725	4,848
TQA w/o ctx	Freeform	**No**	78,859	8,763	11,313
CosmosQA w/o ctx	Multi-choice	**No**	22,735	2,527	2,985
SIQA w/o ctx	Multi-choice	**No**	30,069	3,341	1,954

Table 1: The details of QA datasets. "w/o ctx" refers to the removal of context from the original dataset to evaluate the influence of the amount of evidence.

we performed gradient updates to the following log-likelihood objective: $\max_{\Theta} \sum_i^N \log p_{\Theta}(y_i|x_i)$ where x_i is the input text, and y_i is the output sequence.

$$\max_{\theta_{\mathbf{P}}} \sum_i^N \log p_{\theta,\theta_{\mathbf{P}}}(y_i|[\mathbf{P}; x_i]) \tag{1}$$

The prompt tuning method proposed in Lester et al. (2021) is represented by Equation 1. The parameter of a pre-trained language model θ is fixed, and only the prompt parameter $\theta_{\mathbf{P}}$ of the soft prompt $\mathbf{P} = [p_1, p_2, \ldots, p_l] \in \mathbb{R}^{l \times d}$ is learnable. We use the prompt length $l = 100$, and d is the input dimension of the model.

2.2 Datasets

Following the two classification systems from Rogers et al. (2023), we show 16 QA datasets[2] used in our analysis in Table 1. Detailed descriptions of each dataset are provided in Appendix A.

First, the amount of evidence is how much evidence is provided to answer the question. *Single Source* indicates that the information required to answer the question is explicitly contained within a context. Partial Source means that although some evidence is available, it needs to be integrated with external knowledge to answer the question. **No Source** needs to find answers solely from implicit knowledge. The more evidence available to answer a question, the more explicit knowledge exists; conversely, the less evidence, the more implicit knowledge exists.

Second, the answer format is divided into four types. Extractive format refers to when the answer span can be found within the provided context. Categorical format denotes that the correct answer is in a pre-defined option, exclusively employing yes or no formats in our dataset. Multi-choice format indicates that answer options are given, and the answer is to be chosen from among them. Lastly, Freeform format refers to cases where the model generates answers without following a specific format.

3 Results and Analysis

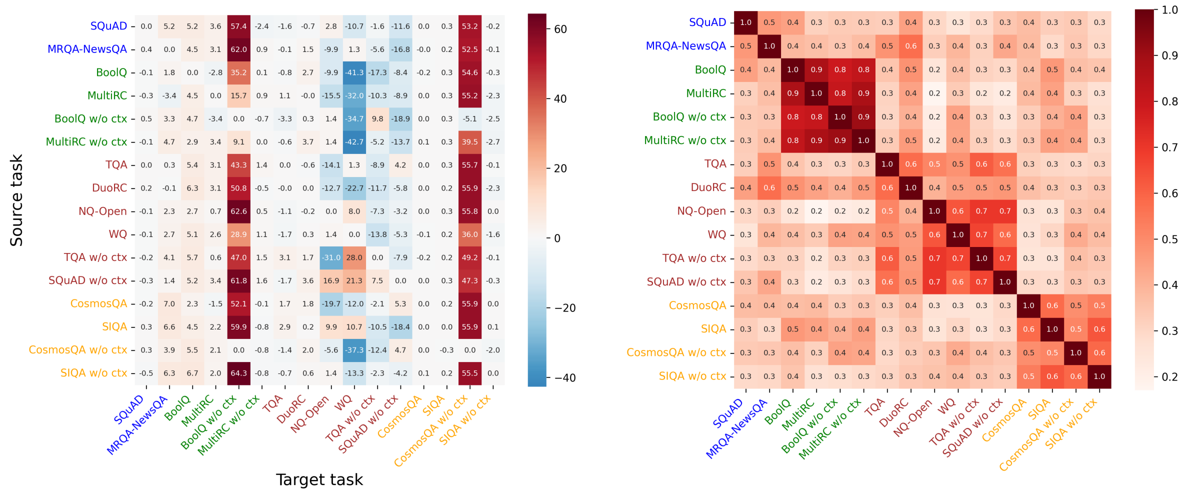
To study the transferability of soft prompts, we used 16 QA datasets as the source and target tasks. The main terms referred to in this section are as follows: (1) vanilla prompt tuning (Vanilla PT), the result of training the prompt in Equation 1 after random initializing; (2) zero-shot performance, the result of solving the target task using the source prompt without additional training; and (3) prompt transfer (PoT), the result of initializing the target prompt with the selected source prompt and training it as shown in Equation 1. For our experiments, we used the T5_{BASE}[3] as our base LM. Further experimental details are in Appendix B.

3.1 Transferability with Initialization

Can transferability be interpreted as cosine similarity? As shown in Figure 1, we investigated the prompt transferability with cosine-similarity. We can observe that prompt embeddings with the

[2] In cases where only one of valid or test datasets was available such as Rajpurkar et al. (2016), we used it in the testing process. Additionally, we split the train datasets into a 9:1 ratio, and used it in the train and valid process, respectively. The number of datasets we used is shown in Table 1.

[3] https://huggingface.co/t5-base



(a) (b)

Figure 1: (a) Heatmap of our task transferability results. (b) Heatmap of the cosine similarities between the source prompt embeddings. The colors of the task names indicate the answer format type: Blue, Extractive; Green, Categorical; Brown, Freeform; Yellow, Multi-choice.

Target Task	Random	Best Source Task	Zero-shot	PoT	Worst Source Task	Zero-shot	PoT	Δ
DuoRC	2.14	SQuAD	32.86	35.56	BoolQ	0.77	36.79	-1.23
NQ-Open	0.00	SQuAD w/o ctx	1.66	2.30	MultiRC w/o ctx	0.00	1.99	+0.31
WQ	0.00	NQ-Open	3.69	3.99	MultiRC	0.00	2.51	+1.48
MRQA-NewsQA	4.80	SQuAD	38.39	41.90	MultiRC	1.16	38.49	+3.41
SQuAD	13.96	DuoRC	78.90	81.57	CosmosQA	1.07	81.28	+0.29
BoolQ	0.00	MultiRC	67.37	76.70	SIQA w/o ctx	0.00	78.38	-1.68
MultiRC	0.06	BoolQ	69.68	74.05	TQA	0.00	78.57	-4.52
TQA	13.21	DuoRC	39.51	43.58	MultiRC	1.87	44.06	-0.48
CosmosQA	2.91	SIQA	78.22	82.81	MultiRC	0.00	82.81	0.00
SIQA	0.61	CosmosQA	99.28	99.59	BoolQ	0.00	99.64	-0.05
SQuAD w/o ctx	0.00	NQ-Open	0.96	1.74	BoolQ	0.00	1.65	+0.09
BoolQ w/o ctx	19.27	BoolQ	47.83	51.13	SIQA w/o ctx	0.00	62.17	-11.04
MultiRC w/o ctx	43.05	MultiRC	57.86	58.15	SQuAD w/o ctx	0.00	58.54	-0.39
TQA w/o ctx	0.15	SQuAD w/o ctx	5.09	4.06	BoolQ w/o ctx	0.02	4.15	-0.09
CosmosQA w/o ctx	0.20	SIQA w/o ctx	74.64	82.65	MultiRC	0.00	82.45	+0.20
SIQA w/o ctx	0.46	SQuAD	26.46	99.39	BoolQ	0.00	99.33	+0.06

Table 2: Relativeness of zero-shot and PoT performance. **Random** indicates the performance after random initialization. **Best Source Task** represents the best performance task in a zero-shot setting. **Worst Source Task** represents the worst performance task in a zero-shot setting. Each score is EM. The difference in the PoT scores between **Best Source Task** and **Worst Source Task** is denoted by Δ. When the **Zero-shot** scores are equal, we chose the source task with the higher **PoT** score.

same answer formats are clustered together in Figure 1(b). However, Figure 1(a) demonstrates that the high similarity score between the source and target task does not necessarily result in positive transferability. For example, even though the transfer BoolQ (Clark et al., 2019) → MultiRC (Khashabi et al., 2018) has the highest similarity score of 0.9, it yields a negative transferability of -2.8%. We note that the PoT performance varies significantly depending on the target task. Therefore, prompt initialization with high cosine-similarity does not guarantee performance improvement. As a result, we find that it is not suitable to interpret transferability through cosine-similarity in the QA task.

Can transferability be interpreted as zero-shot performance? To verify the effectiveness of selecting the best zero-shot prompt when used for initialization, we compare PoT performance between the best and worst zero-shot prompts in Table 2. When initialized with the best zero-shot prompt, it only outperforms the worst one in 7 out of 16

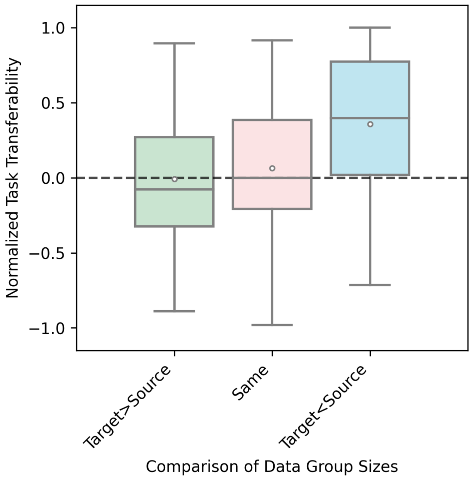


Figure 2: Normalized transferability difference based on the train datasets size. The X-axis refers to each case where the target dataset size is bigger, equal, and smaller than the source dataset size. The Y-axis denotes the difference between the PoT performance and the vanilla prompt performance after normalization.

cases. The mean absolute error was 1.58, indicating that the performance difference is approximate. Additionally, Figure 5 and Figure 6 indicate that most cases converge to similar values as the epoch progresses, regardless of which source prompt is selected. It can therefore be seen that the method proposed in Su et al. (2022) cannot assure better or comparable transfer performance in the QA task.

Effect of Dataset Size In Table 4, the PoT performance varies considerably depending on the target task. Therefore, we applied min-max normalization[4] to each target task to compare the correlation between the source and target tasks. We classified the QA datasets based on the number of train datasets into small, medium, and large (see Appendix D). Subsequently, we divided into three groups [5] founded on the difference in size between the source task and target task as follows: *Target* > *Source*, *Same* and *Target* < *Source*.

As shown in Figure 2, the normalized task transferability results are based on the difference in the dataset group size between the source task and the target task. Regarding the *Target* < *Source* group, most cases show positive transferability. The median (Q2) of each box plot indicates a tendency to drop in the sequence of *Target* < *Source*, *Same*, and *Target* > *Source*. We demonstrate that the dataset size of the source and target tasks in the QA task is a key factor in transferability.

3.2 Investigating Catastrophic Forgetting

Catastrophic Forgetting Formula Catastrophic forgetting (Kirkpatrick et al., 2017) is the tendency for previously learned task knowledge to be abruptly lost as information relevant to the current task is incorporated. However, there is still no clear method for measuring this phenomenon.

Therefore, we propose a novel metric for evaluating catastrophic forgetting:

$$\frac{(Zero\text{-}shot\ correct) \cap (PoT\ incorrect)}{Zero\text{-}shot\ correct} \tag{2}$$

where *Zero-shot correct* is the case of correct responses in a zero-shot setting, and *PoT incorrect* is the case of incorrect answers after prompt transfer in the target task. In a zero-shot setting, correct responses indicates that the trained prompt from the source task retains valuable information for the target task. On the other hand, incorrect answers after additional learning with the target task indicate forgetting of source task knowledge relevant to the target task.

We analyse catastrophic forgetting in terms of the amount of evidence in the QA datasets. *Single Source* use the most explicit knowledge, followed by Partial Source, and **No Source**. The relationship between explicit and implicit knowledge is a trade-off. When comparing the quantity of explicit and implicit knowledge with the amount of evidence, Equation 2 is used for cases where the target task has a bigger, equal, or smaller amount of explicit knowledge than the source task.

Analyzing Catastrophic Forgetting As illustrated in Figure 3, the results compare the extent of catastrophic forgetting based on the levels of explicit and implicit knowledge in each dataset. If the source task has more explicit knowledge or less implicit knowledge than the target task, catastrophic forgetting tends to occur. In the right side[6] of Figure 3, *Partial-Single*, *No-Single*, and *No-Partial* are displayed mixed together and the left also shows a similar trend. As a result, the existence of a knowledge gap between the source and target task is more influential in catastrophic forgetting than the extent of the knowledge gap.

[4] See the formula in Appendix C.

[5] For example, *Target* > *Source*, indicating the train dataset group of the target task is larger than the source task (*e.g.*, target task: large, source task: small).

[6] **Explicit: Target < Source** indicates that the amount of explicit knowledge of the target task is less than the amount of explicit knowledge of the source tasks.

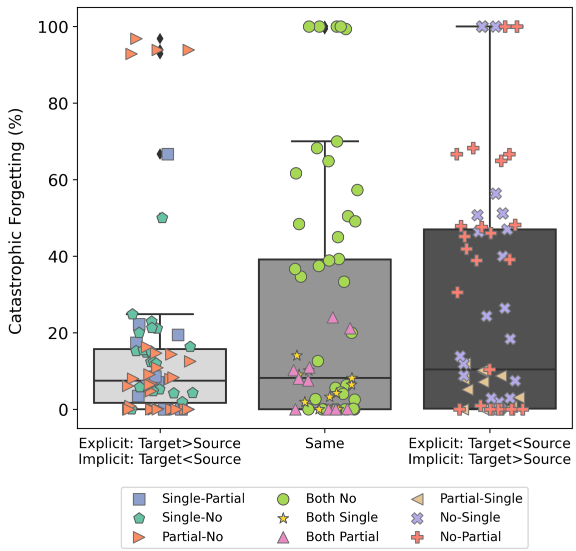


Figure 3: Percentage of Catastrophic Forgetting based on the amount of evidence. The X-axis shows the differences between the target and source task according to the amount of evidence. The Y-axis represents the percentage of catastrophic forgetting. Each label indicates the amount of evidence type in Target-Source order.

4 Conclusion

In this paper, we study PoT in the QA task. In particular, we empirically investigate prompt initialization, demonstrating that the difference of train dataset size between source and target tasks is affecting the transferability. We also define a novel method to measure catastrophic forgetting and show that there is a relationship between the amount of evidence in QA datasets and the tendency of catastrophic forgetting. Finally, our fine-grained analyses provide meaningful insights to help improve the performance of PoT.

Limitations

Our paper has two limitations as follows: First, we only perform all experiments on the T5_{BASE} model. We cannot serve results on various models and model sizes because of our limited computational resources. Secondly, although we show the type of occurrence for catastrophic forgetting by our proposed evaluation metric, we do not present an approach to mitigate them.

In further experiments, we observe the possibility that prompt transferability could be influenced by different model architectures, prefixes, or other factors. Therefore, in the future work, we will explore strategies to save the knowledge of source tasks related to target tasks and investigate the use of various backbone models.

Acknowledgements

We thank anonymous reviewers for valuable feedback and helpful suggestions. This work was supported in part by the National Research Foundation of Korea (NRF) under Grant 2018R1A5A7059549 and Grant 2020R1A2C1014037; and in part by the Institute of Information and Communications Technology Planning and Evaluation (IITP), funded by the Korean Government [Ministry of Science and Information and Communication Technology (MSIT)] under Grant 2020-0-01373.

References

Akari Asai, Mohammadreza Salehi, Matthew Peters, and Hannaneh Hajishirzi. 2022. ATTEMPT: Parameter-efficient multi-task tuning via attentional mixtures of soft prompts. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 6655–6672, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Jonathan Berant, Andrew Chou, Roy Frostig, and Percy Liang. 2013. Semantic parsing on Freebase from question-answer pairs. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1533–1544, Seattle, Washington, USA. Association for Computational Linguistics.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D 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 Ziegler, Jeffrey Wu, Clemens Winter, Chris 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*, volume 33, pages 1877–1901. Curran Associates, Inc.

Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. 2019. BoolQ: Exploring the surprising difficulty of natural yes/no questions. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2924–2936, Minneapolis, Minnesota. Association for Computational Linguistics.

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

Adam Fisch, Alon Talmor, Robin Jia, Minjoon Seo, Eunsol Choi, and Danqi Chen. 2019. MRQA 2019 shared task: Evaluating generalization in reading comprehension. In *Proceedings of the 2nd Workshop on Machine Reading for Question Answering*, pages 1–13, Hong Kong, China. Association for Computational Linguistics.

Lifu Huang, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2019. Cosmos QA: Machine reading comprehension with contextual commonsense reasoning. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2391–2401, Hong Kong, China. Association for Computational Linguistics.

Mandar Joshi, Eunsol Choi, Daniel Weld, and Luke Zettlemoyer. 2017. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1601–1611, Vancouver, Canada. Association for Computational Linguistics.

Daniel Khashabi, Snigdha Chaturvedi, Michael Roth, Shyam Upadhyay, and Dan Roth. 2018. Looking beyond the surface: A challenge set for reading comprehension over multiple sentences. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 252–262, New Orleans, Louisiana. Association for Computational Linguistics.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. 2017. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 114(13):3521–3526.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:452–466.

Kenton Lee, Ming-Wei Chang, and Kristina Toutanova. 2019. Latent retrieval for weakly supervised open domain question answering. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6086–6096, Florence, Italy. Association for Computational Linguistics.

Brian Lester, Rami Al-Rfou, and Noah Constant. 2021. The power of scale for parameter-efficient prompt tuning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3045–3059, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Junyi Li, Tianyi Tang, Jian-Yun Nie, Ji-Rong Wen, and Xin Zhao. 2022. Learning to transfer prompts for text generation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3506–3518, Seattle, United States. Association for Computational Linguistics.

SGOPAL Patro and Kishore Kumar Sahu. 2015. Normalization: A preprocessing stage. *arXiv preprint arXiv:1503.06462*.

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. *The Journal of Machine Learning Research*, 21(1):5485–5551.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas. Association for Computational Linguistics.

Anna Rogers, Matt Gardner, and Isabelle Augenstein. 2023. Qa dataset explosion: A taxonomy of nlp resources for question answering and reading comprehension. *ACM Computing Surveys*, 55(10):1–45.

Amrita Saha, Rahul Aralikatte, Mitesh M. Khapra, and Karthik Sankaranarayanan. 2018. DuoRC: Towards complex language understanding with paraphrased reading comprehension. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1683–1693, Melbourne, Australia. Association for Computational Linguistics.

Maarten Sap, Hannah Rashkin, Derek Chen, Ronan Le Bras, and Yejin Choi. 2019. Social IQa: Commonsense reasoning about social interactions. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4463–4473, Hong Kong, China. Association for Computational Linguistics.

Yusheng Su, Xiaozhi Wang, Yujia Qin, Chi-Min Chan, Yankai Lin, Huadong Wang, Kaiyue Wen, Zhiyuan Liu, Peng Li, Juanzi Li, Lei Hou, Maosong Sun, and Jie Zhou. 2022. On transferability of prompt tuning for natural language processing. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3949–3969, Seattle, United States. Association for Computational Linguistics.

Adam Trischler, Tong Wang, Xingdi Yuan, Justin Harris, Alessandro Sordoni, Philip Bachman, and Kaheer Suleman. 2017. NewsQA: A machine comprehension dataset. In *Proceedings of the 2nd Workshop on Representation Learning for NLP*, pages 191–200, Vancouver, Canada. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Tu Vu, Brian Lester, Noah Constant, Rami Al-Rfou', and Daniel Cer. 2022. SPoT: Better frozen model adaptation through soft prompt transfer. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5039–5059, Dublin, Ireland. Association for Computational Linguistics.

Zhen Wang, Rameswar Panda, Leonid Karlinsky, Rogerio Feris, Huan Sun, and Yoon Kim. 2023. Multitask prompt tuning enables parameter-efficient transfer learning. In *The Eleventh International Conference on Learning Representations*.

Kaige Xie, Tong Yu, Haoliang Wang, Junda Wu, Handong Zhao, Ruiyi Zhang, Kanak Mahadik, Ani Nenkova, and Mark Riedl. 2023. Few-shot dialogue summarization via skeleton-assisted prompt transfer. *arXiv preprint arXiv:2305.12077*.

Qihuang Zhong, Liang Ding, Juhua Liu, Bo Du, and Dacheng Tao. 2022. Panda: Prompt transfer meets knowledge distillation for efficient model adaptation. *arXiv preprint arXiv:2208.10160*.

A Datasets

Table 1 displays the datasets used in our experiments. More descriptions of each dataset are as follows:

- **DuoRC** (Saha et al., 2018) is a reading comprehension dataset with low lexical overlap between questions and context. It has unique question-answer pairs generated from a movie plots collection. The original dataset included *no answer*, but we only used the data that has an answer. Background and common sense are required to derive the answer, surpassing the context's explicit knowledge. We used the dataset from `https://huggingface.co/datasets/duorc/viewer/SelfRC`.

- The original Natural Questions (NQ) dataset was introduced by Kwiatkowski et al. (2019). The **NQ-Open**, which removes the context from NQ, was introduced by Lee et al. (2019). We used the dataset from `https://huggingface.co/datasets/nq_open`.

- The WebQuestions (**WQ**, Berant et al., 2013) dataset consists of questions that can be answered via Freebase. The dataset link we used, `https://huggingface.co/datasets/web_questions`, only provides a freebase link, so we did not use a separate context.

- **MRQA-NewsQA**. NewsQA (Trischler et al., 2017) is a machine comprehension dataset composed of CNN news articles. The answers consist of spans of texts from the article. We used the NewsQA dataset from `https://huggingface.co/datasets/mrqa`, within the MRQA (Fisch et al., 2019) benchmark.

- The Stanford Question Answering Dataset (**SQuAD**, Rajpurkar et al., 2016) is a benchmark dataset in machine reading comprehension. It comprises Wikipedia articles accompanied by question-answer pairs formulated by human annotators. Answers in SQuAD are spans of text directly extracted from the provided context. We used the dataset from `https://huggingface.co/datasets/squad`.

- **BoolQ** (Clark et al., 2019) is a format of yes/no questions. A context is given along with question-answer pairs. We used the dataset from `https://huggingface.co/datasets/boolq`.

- The **MultiRC** (Khashabi et al., 2018) dataset consists of a paragraph (context), question, and answer as well as a label to determine whether the answer to the question was correct. We used this dataset to solve a categorical answer format problem that determines whether the answer to a question is correct. We used the dataset from `https://huggingface.co/datasets/super_glue/viewer/multirc`.

- TriviaQA (**TQA**, Joshi et al., 2017) is a reading comprehension dataset, which is more challenging than other QA datasets because the questions cannot be answered directly by span prediction and the context is much longer than other benchmarks. We used two versions of TQA: the *unfiltered* version (TQA) and the *unfiltered.nocontext* version (TQA w/o ctx). We used the dataset from https://huggingface.co/datasets/trivia_qa.

- **CosmosQA** (Huang et al., 2019) requires commonsense-based reading comprehension and consists of questions that require additional knowledge rather than extracting spans from the context. The answer is in the form of choosing one of four options. We used the dataset from https://huggingface.co/datasets/cosmos_qa.

- Social Interaction QA (**SIQA**, Sap et al., 2019) is a dataset for testing commonsense reasoning about social situations. The answer is to choose one of three options. We used the dataset from https://huggingface.co/datasets/social_i_qa.

B Training Details

In prompt tuning, we trained a soft prompt using a NVIDIA RTX A5000 single GPU with 23GB memory. We applied the AdamW optimizer with a learning rate 0.005, set batch size of 16, and used early stopping in three steps. We set the soft prompt length l=100, which is the same as most prompt transfer settings (Vu et al., 2022; Asai et al., 2022; Su et al., 2022; Wang et al., 2023).

C Min-Max Normalization

The PoT performance for each target task is normalized using a formula derived with reference to (Patro and Sahu, 2015). We remove the denominator from the formula because sometimes it becomes zero. The formula we used is as follows :

$$\textit{Normalized PoT score} = \frac{(PoT\ score) - (Vanilla\ PT\ score)}{\max(PoT\ score) - \min(PoT\ score)} \quad (3)$$

D Comparing Train Dataset Size

Figure 4 illustrates the categorization of QA datasets based on the size of train datasets.

E Prompt Transfer Performance in Each Epoch

Figure 5 and Figure 6 show that as the epoch progresses, the influence of initialization gradually decreases. The red and blue lines denote the scores

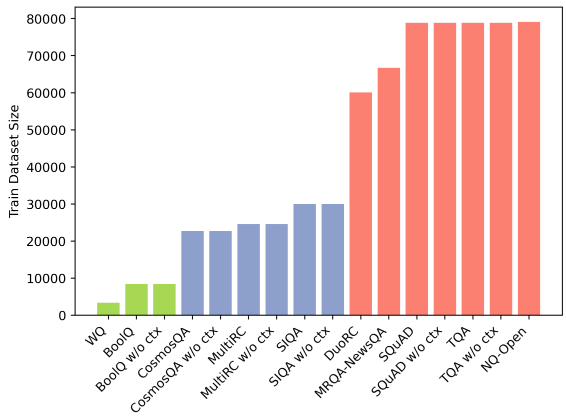


Figure 4: QA dataset size. Each color indicates the train dataset group: Green, small; Blue, medium; Red, large.

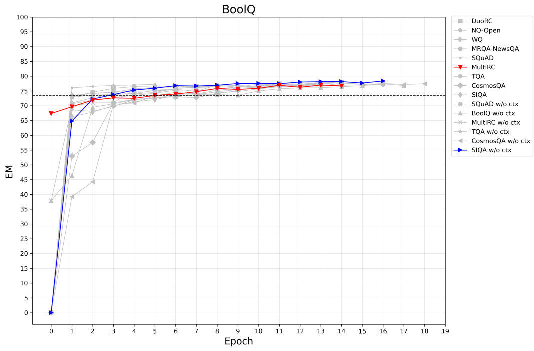
per epoch for the **Best Source Task** and **Worst Source Task** shown in Table 2. Specifically, even though some prompts have EM score of 0 in the zero-shot setting, they achieve better or comparable PoT performance than prompts with the best zero-shot performance.

F Zero Shot Performance

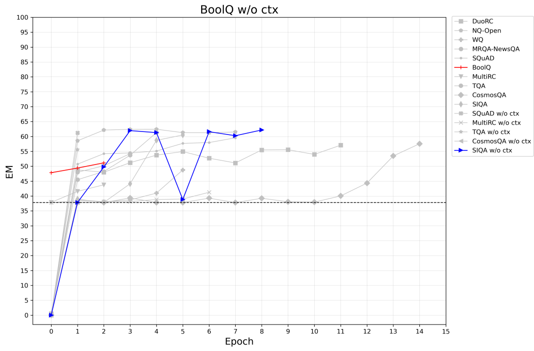
The full results of zero-shot performance are shown in Table 3.

G Prompt Transfer Performance

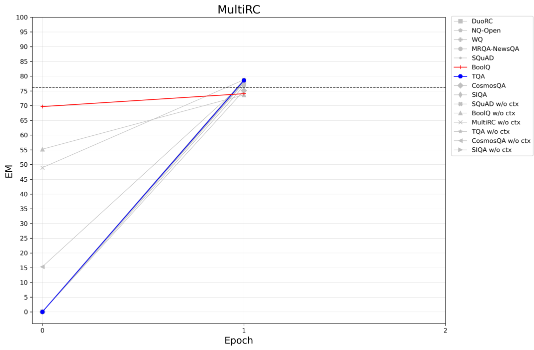
The full results in our experiments are shown in Table 4.



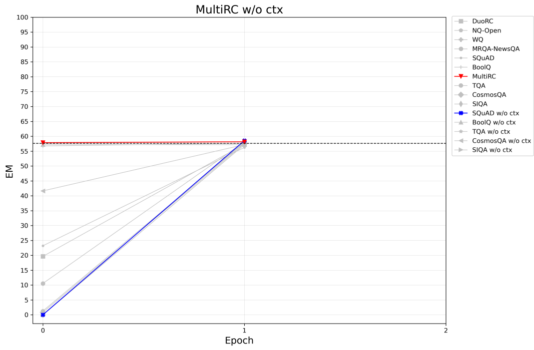
(a) BoolQ



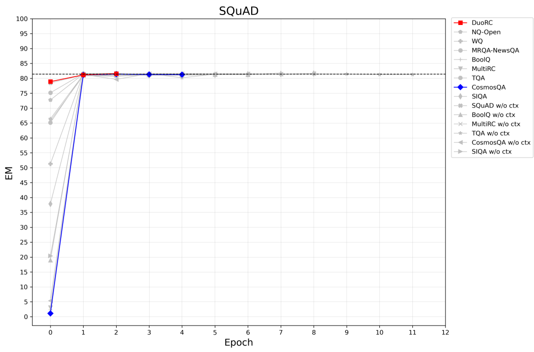
(b) BoolQ w/o ctx



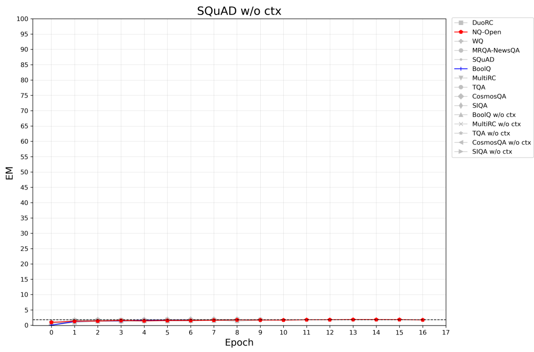
(c) MultiRC



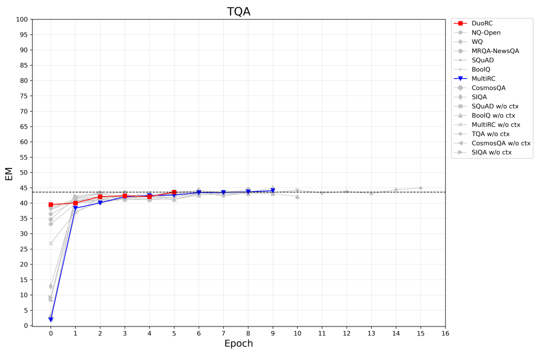
(d) MultiRC w/o ctx



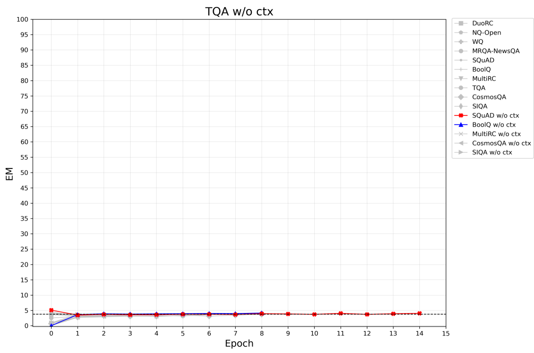
(e) SQuAD



(f) SQuAD w/o ctx

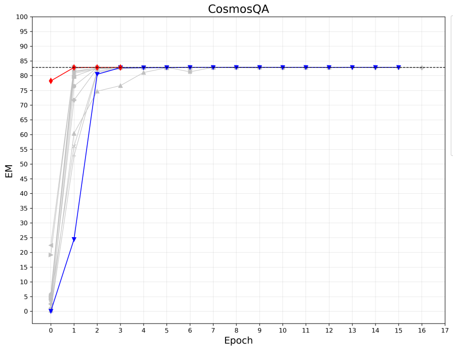


(g) TQA

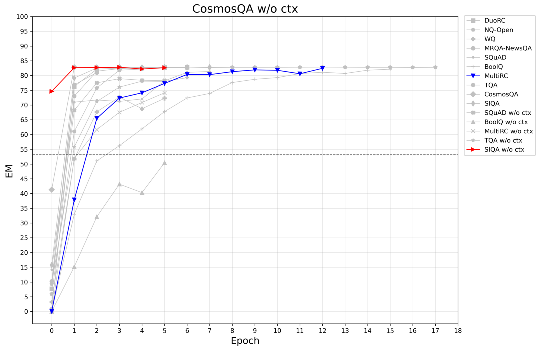


(h) TQA w/o ctx

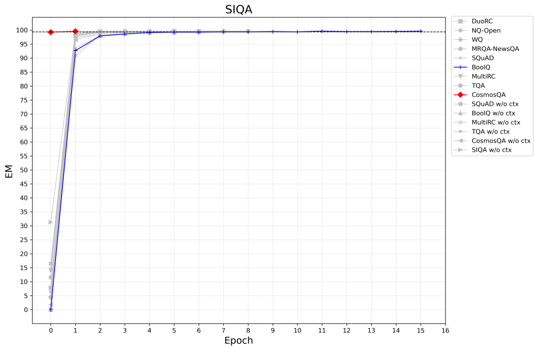
Figure 5: Prompt Transfer Performance in Each Epoch.



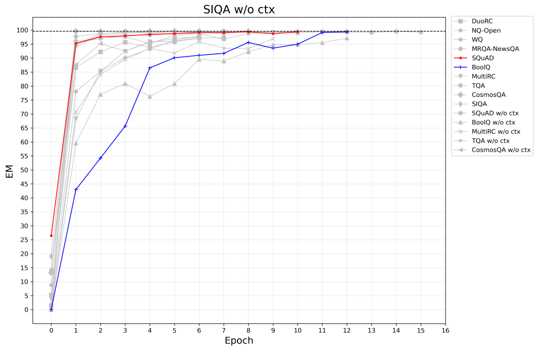
(a) CosmosQA



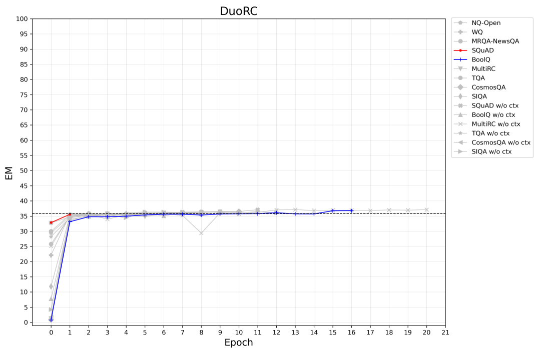
(b) CosmosQA w/o ctx



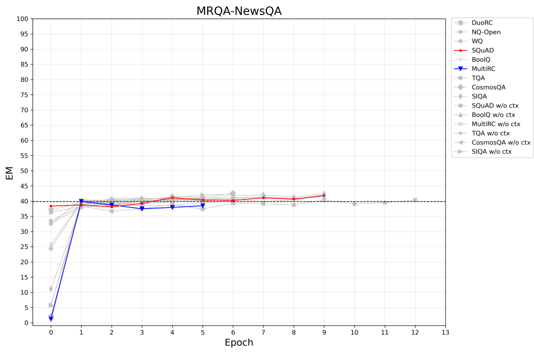
(c) SIQA



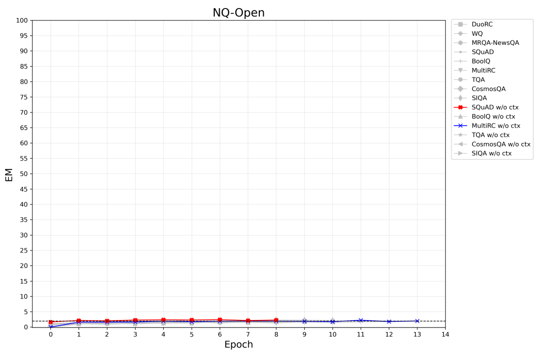
(d) SIQA w/o ctx



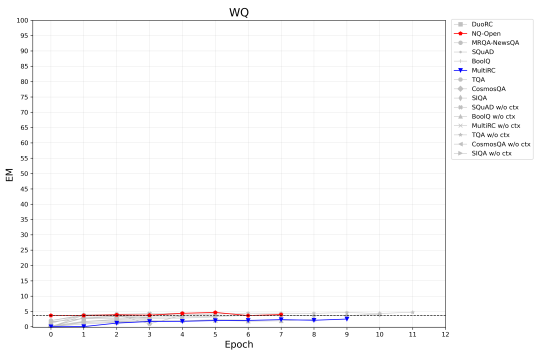
(e) DuoRC



(f) MRQA-NewsQA



(g) NQ-Open



(h) WQ

Figure 6: Prompt Transfer Performance in Each Epoch.

		Target Task							
		DuoRC	**NQ-Open**	**WQ**	**MRQA-NewsQA**	**SQuAD**	**BoolQ**	**MultiRC**	**TQA**
	Random	2.14 / 8.88	0.00 / 0.20	0.00 / 0.45	4.80 / 17.16	13.96 / 30.59	0.00 / 0.03	0.06 / 0.06	13.21 / 22.62
Source Task	DuoRC	-	0.64 / <u>4.38</u>	1.53 / 10.52	<u>37.16</u> / <u>53.91</u>	**78.90** / **87.37**	0.18 / 0.24	0.08 / 0.14	**39.51** / **47.31**
	NQ-Open	29.46 / 39.07	-	**3.69** / **12.14**	33.57 / 47.68	72.72 / 80.79	0.00 / 0.00	0.00 / 0.07	36.39 / 43.93
	WQ	22.13 / 30.11	1.02 / 4.06	-	24.41 / 36.76	51.29 / 60.77	0.00 / 0.02	0.00 / 0.08	33.12 / 40.83
	MRQA-NewsQA	25.83 / 39.97	<u>1.19</u> / 4.21	2.02 / 9.28	-	75.16 / <u>86.68</u>	0.00 / 0.02	0.04 / 0.09	34.69 / 44.89
	SQuAD	**32.86** / **44.07**	1.14 / 4.05	1.67 / 9.95	**38.39** / **55.64**	-	0.00 / 0.03	0.08 / 0.12	37.87 / 46.28
	BoolQ	0.77 / 0.96	0.00 / 0.00	0.00 / 0.00	1.40 / 1.88	2.34 / 2.54	-	**69.68** / **69.68**	3.35 / 3.84
	MultiRC	1.26 / 1.43	0.00 / 0.00	0.00 / 0.00	1.16 / 1.54	3.02 / 3.26	**67.37** / **67.37**	-	1.87 / 2.08
	TQA	29.97 / 39.32	0.58 / 2.85	1.13 / 6.07	32.69 / 47.37	65.10 / 75.47	0.00 / 0.20	0.00 / 0.06	-
	CosmosQA	0.84 / 6.83	0.08 / 0.88	0.05 / 1.17	2.18 / 7.83	1.07 / 8.56	0.00 / 0.22	0.00 / 0.02	2.54 / 11.65
	SIQA	11.82 / 23.13	0.03 / 2.27	0.20 / 4.08	11.11 / 25.80	37.86 / 57.14	0.00 / 0.20	0.00 / 0.01	12.83 / 24.79
	SQuAD w/o ctx	<u>32.77</u> / <u>43.26</u>	**1.66** / **5.02**	<u>2.66</u> / <u>11.53</u>	36.35 / 51.90	<u>78.45</u> / 86.20	0.00 / 0.02	0.04 / 0.07	<u>38.75</u> / <u>46.41</u>
	BoolQ w/o ctx	7.76 / 9.67	0.00 / 0.00	0.00 / 0.00	5.96 / 8.58	18.94 / 21.00	<u>37.83</u> / <u>37.83</u>	<u>55.22</u> / <u>55.26</u>	8.56 / 10.82
	MultiRC w/o ctx	25.16 / 33.89	0.00 / 0.00	0.00 / 0.00	25.78 / 37.66	65.79 / 74.47	<u>37.83</u> / <u>37.83</u>	48.95 / 49.04	26.82 / 33.31
	TQA w/o ctx	28.22 / 37.89	0.86 / 3.23	2.02 / 7.91	32.53 / 46.69	66.41 / 75.48	0.00 / 0.01	0.00 / 0.08	38.06 / 45.05
	CosmosQA w/o ctx	1.76 / 6.53	0.08 / 2.11	0.20 / 4.62	1.88 / 8.31	5.37 / 18.49	0.09 / 0.11	15.33 / 15.38	2.06 / 9.55
	SIQA w/o ctx	4.16 / 13.28	0.00 / 2.99	0.25 / 6.07	5.75 / 18.43	20.37 / 37.78	0.00 / 0.11	0.04 / 0.14	9.14 / 18.52

		Target Task							
		CosmosQA	**SIQA**	**SQuAD w/o ctx**	**BoolQ w/o ctx**	**MultiRC w/o ctx**	**TQA w/o ctx**	**CosmosQA w/o ctx**	**SIQA w/o ctx**
	Random	2.91 / 22.39	0.61 / 8.79	0.00 / 0.27	19.27 / 19.27	43.05 / 43.05	0.15 / 1.53	0.20 / 3.54	0.46 / 6.76
Source Task	DuoRC	4.46 / 21.02	7.83 / 20.55	0.70 / **7.86**	0.55 / 0.85	19.64 / 19.64	3.70 / **12.57**	7.74 / 29.03	5.27 / 16.85
	NQ-Open	3.85 / 17.74	14.48 / 27.74	**0.96** / 6.22	0.00 / 0.23	1.30 / 1.30	<u>4.53</u> / 10.20	5.96 / 25.31	8.75 / 26.37
	WQ	2.48 / 15.79	4.30 / 11.35	0.61 / 5.22	0.00 / 0.09	0.12 / 0.17	3.60 / 9.65	3.18 / 16.77	1.89 / 8.23
	MRQA-NewsQA	5.26 / 22.04	11.57 / 24.27	<u>0.95</u> / 6.67	0.46 / 0.79	10.54 / 10.54	4.11 / 9.98	10.25 / 31.47	14.07 / 28.56
	SQuAD	6.13 / 23.13	13.82 / 24.82	0.90 / <u>6.72</u>	0.00 / 0.33	23.21 / 23.21	4.13 / 10.88	14.20 / 36.87	**26.46** / **43.90**
	BoolQ	0.00 / 0.02	0.00 / 0.00	0.00 / 0.01	**47.83** / **47.83**	56.75 / 56.75	0.02 / 0.10	0.00 / 0.00	0.00 / 0.00
	MultiRC	0.00 / 0.00	0.00 / 0.00	0.00 / 0.01	<u>37.86</u> / <u>37.86</u>	**57.86** / **57.86**	0.02 / 0.06	0.00 / 0.00	0.00 / 0.00
	TQA	0.84 / 8.66	4.45 / 15.30	0.60 / 4.51	0.00 / 0.00	0.00 / 0.00	2.52 / 7.24	0.77 / 6.56	1.28 / 11.25
	CosmosQA	-	**99.28** / **99.91**	0.02 / 1.01	0.03 / 0.24	0.85 / 0.89	0.72 / 3.20	<u>41.31</u> / <u>59.34</u>	13.15 / <u>31.99</u>
	SIQA	**78.22** / **92.91**	-	0.03 / 2.65	0.00 / 0.24	0.00 / 0.04	0.86 / 5.66	15.71 / 35.10	<u>19.09</u> / 28.84
	SQuAD w/o ctx	5.36 / 21.12	16.48 / 29.65	-	0.00 / 0.17	0.00 / 0.01	**5.09** / <u>11.35</u>	9.72 / 28.17	13.72 / 27.54
	BoolQ w/o ctx	0.34 / 1.93	0.00 / 0.00	0.00 / 0.01	-	<u>57.22</u> / <u>57.22</u>	0.02 / 0.06	0.00 / 0.00	0.00 / 0.00
	MultiRC w/o ctx	3.79 / 16.71	0.00 / 0.00	0.00 / 0.01	37.80 / 37.80	-	0.02 / 0.08	0.00 / 0.41	0.00 / 0.00
	TQA w/o ctx	5.83 / 16.46	6.35 / 18.41	0.89 / 4.96	0.00 / 0.05	0.08 / 0.08	-	9.15 / 18.83	3.94 / 13.94
	CosmosQA w/o ctx	<u>22.41</u> / <u>44.05</u>	1.79 / 10.78	0.01 / 3.53	0.00 / 0.11	41.65 / 41.65	0.95 / 6.92	-	14.02 / 30.83
	SIQA w/o ctx	19.13 / 33.23	<u>31.22</u> / <u>44.86</u>	0.03 / 4.18	0.00 / 0.21	0.00 / 0.03	0.97 / 7.23	**74.64** / **89.94**	-

Table 3: Zero shot Performance (EM / F1). **Bold** and <u>underline</u> fonts denote the best and the second best score. Zero-shot performance is not measurable when the source task and target task are the same.

Source Task	DuoRC	NQ-Open	WQ	MRQA-NewsQA	SQuAD	BoolQ	MultiRC	TQA
DuoRC	35.80 / 45.71	1.72 / 5.99	2.85 / 13.05	39.81 / 57.96	81.57 / 89.75	78.04 / 78.04	78.57 / 78.57	43.58 / 49.28
NQ-Open	35.74 / 45.85	1.97 / 6.51	3.99 / 14.28	40.74 / 58.47	81.32 / 89.72	75.38 / 75.38	76.71 / 76.71	43.12 / 48.78
WQ	35.90 / 45.78	1.99 / 6.45	3.69 / 14.36	40.93 / 58.57	81.32 / 89.72	77.19 / 77.19	78.20 / 78.20	42.85 / 48.79
MRQA-NewsQA	36.35 / 46.33	1.77 / 6.25	3.74 / 13.61	39.84 / 58.12	81.70 / 90.03	76.76 / 76.76	78.55 / 78.55	43.53 / 49.18
SQuAD	35.56 / 45.44	2.02 / 6.48	3.30 / 12.70	41.90 / 59.31	81.40 / 89.80	77.25 / 77.25	**78.92** / **78.92**	42.91 / 48.66
BoolQ	36.79 / 46.72	1.77 / 6.09	2.17 / 11.66	40.55 / 57.96	81.32 / 89.74	73.43 / 73.43	74.05 / 74.05	43.26 / 49.13
MultiRC	35.80 / 45.72	1.66 / 6.25	2.51 / 13.30	38.49 / 57.05	81.19 / 89.71	76.70 / 76.70	76.20 / 76.20	44.06 / 49.76
TQA	35.59 / 45.87	1.69 / 5.90	3.74 / 12.62	39.96 / 57.48	81.41 / 89.68	77.37 / 77.37	78.57 / 78.57	43.59 / 49.35
CosmosQA	36.46 / 46.23	1.58 / 6.06	3.25 / 13.48	**42.62** / 59.60	81.28 / 89.68	75.11 / 75.11	75.08 / 75.08	44.31 / 49.96
SIQA	35.89 / 45.80	2.16 / 6.49	4.08 / 14.90	42.45 / 59.53	81.63 / 89.87	76.73 / 76.73	77.87 / 77.87	44.83 / 50.45
SQuAD w/o ctx	37.10 / 46.83	**2.30** / 6.55	4.48 / 14.96	40.38 / 58.43	81.12 / 89.66	77.22 / 77.22	78.77 / 78.77	42.85 / 48.79
BoolQ w/o ctx	35.91 / 45.92	1.99 / **6.67**	2.41 / 12.45	41.14 / 58.82	**81.82** / **90.07**	76.85 / 76.85	73.64 / 73.64	42.16 / 47.88
MultiRC w/o ctx	**37.12** / **47.06**	1.99 / 6.54	2.12 / 12.10	41.71 / 59.19	81.32 / 89.69	75.57 / 75.57	78.77 / 78.77	43.32 / 48.94
TQA w/o ctx	36.40 / 46.20	1.36 / 6.24	**4.72** / **15.55**	41.45 / 59.11	81.27 / 89.69	77.61 / 77.61	76.69 / 76.69	**44.93** / **50.50**
CosmosQA w/o ctx	36.51 / 46.55	1.86 / 6.00	2.31 / 12.18	41.41 / 58.96	81.62 / 89.80	77.43 / 77.43	77.81 / 77.81	42.99 / 48.72
SIQA w/o ctx	36.00 / 46.51	1.99 / 6.28	3.20 / 14.08	42.33 / **59.62**	81.01 / 89.58	**78.38** / **78.38**	77.70 / 77.70	43.28 / 49.30

Source Task	CosmosQA	SIQA	SQuAD w/o ctx	BoolQ w/o ctx	MultiRC w/o ctx	TQA w/o ctx	CosmosQA w/o ctx	SIQA w/o ctx
DuoRC	82.81 / **96.53**	99.59 / 99.95	1.69 / 8.16	57.06 / 57.06	57.34 / 57.34	3.33 / 7.99	82.81 / 96.40	97.34 / 98.72
NQ-Open	82.85 / 96.47	**99.64** / **99.97**	1.74 / 7.95	61.50 / 61.50	57.90 / 57.90	3.50 / 8.19	82.78 / **96.54**	99.59 / 99.87
WQ	82.71 / 96.48	99.59 / 99.92	1.70 / 7.82	48.75 / 48.75	58.25 / 58.25	3.25 / 8.17	72.26 / 87.36	98.00 / 98.93
MRQA-NewsQA	82.78 / 96.49	99.54 / 99.89	1.49 / 7.73	61.28 / 61.28	58.13 / 58.13	3.56 / 8.35	81.01 / 95.12	99.54 / 99.89
SQuAD	82.85 / 96.51	**99.64** / **99.97**	1.59 / 7.83	59.54 / 59.54	56.25 / 56.25	3.71 / 8.68	81.37 / 95.37	99.39 / 99.78
BoolQ	82.68 / 96.43	**99.64** / 99.94	1.65 / 7.84	51.13 / 51.13	57.67 / 57.67	3.12 / 7.84	82.14 / 96.28	99.33 / 99.81
MultiRC	82.81 / 96.50	**99.64** / 99.93	1.64 / **8.25**	43.76 / 43.76	58.15 / 58.15	3.39 / 8.83	82.45 / 96.19	97.34 / 98.34
TQA	82.81 / 96.51	**99.64** / 99.93	1.87 / 8.24	54.22 / 54.22	58.44 / 58.44	3.44 / 8.30	82.71 / 96.41	99.49 / 99.89
CosmosQA	82.81 / 96.50	99.59 / 99.96	**1.89** / 8.20	57.55 / 57.55	57.55 / 57.55	3.69 / 8.67	82.81 / 96.49	99.59 / 99.96
SIQA	82.81 / 96.47	99.39 / 99.86	1.47 / 7.73	60.49 / 60.49	57.20 / 57.20	3.38 / 8.02	**82.85** / 96.51	**99.64** / **99.97**
SQuAD w/o ctx	82.85 / 96.49	**99.64** / 99.93	1.80 / 8.06	61.19 / 61.19	**58.54** / **58.54**	4.06 / 9.06	78.26 / 92.22	99.28 / 99.81
BoolQ w/o ctx	82.85 / 96.49	**99.64** / **99.97**	1.46 / 7.79	37.83 / 37.83	57.22 / 57.22	**4.15** / **9.36**	50.42 / 67.32	97.13 / 98.52
MultiRC w/o ctx	**82.88** / 96.51	**99.64** / 99.92	1.55 / 7.69	41.25 / 41.25	57.63 / 57.63	3.58 / 8.59	74.14 / 88.43	96.88 / 97.87
TQA w/o ctx	82.65 / 96.47	99.59 / 99.91	1.66 / 7.99	55.63 / 55.63	58.50 / 58.50	3.77 / 8.73	79.26 / 93.57	99.54 / 99.89
CosmosQA w/o ctx	82.85 / 96.47	99.13 / 99.65	1.88 / 8.07	37.83 / 37.83	57.20 / 57.20	3.31 / 8.16	53.13 / 71.82	97.59 / 98.42
SIQA w/o ctx	**82.88** / **96.53**	99.59 / 99.96	1.72 / 8.06	**62.17** / **62.17**	57.20 / 57.20	3.69 / 8.86	82.65 / 96.44	99.59 / 99.90

Table 4: Prompt Transfer Performance (EM / F1). **Bold** and underline fonts denote the best and the second best score.