# Retrieval-Augmented Data Augmentation for Low-Resource Domain Tasks

**Minju Seo*** **Jinheon Baek*** **James Thorne** **Sung Ju Hwang**
KAIST
{minjuseo, jinheon.baek, thorne, sjhwang82}@kaist.ac.kr

## Abstract

Despite large successes of recent language models, they suffer from severe performance degeneration in low-resource settings with limited training data available. Many existing works tackle this problem by generating synthetic data from the training data and then training models on them, recently using Large Language Models (LLMs). However, in low-resource settings, the amount of seed data samples to use for data augmentation is very small, which makes generated samples suboptimal and less diverse. To tackle this challenge, we propose a novel method that augments training data by incorporating a wealth of examples from other datasets, along with the given training data. Specifically, we first retrieve relevant instances from other datasets, such as their input-output pairs or contexts, based on their similarities with the given seed data, and prompt LLMs to generate new samples with the contextual information within and across the original and retrieved samples. This approach can ensure that the generated data is not only relevant but also more diverse than what could be achieved using the limited seed data alone. We validate our Retrieval-Augmented Data Augmentation (RADA) framework on multiple datasets under low-resource settings of training and test-time data augmentation scenarios, on which it outperforms existing data augmentation baselines.

## 1 Introduction

Recent advances in language models [7, 58, 46, 3] have achieved numerous successes across various natural language tasks. The common practice to further enhance their performances is to perform fine-tuning on task-specific datasets, which has been proven substantially effective regardless of model sizes [23, 41]. However, the efficacy of this fine-tuning is closely tied to the volume and quality of the data available for training. Meanwhile, in real-world scenarios, there is often a scarcity of training instances, and the manual annotation of additional training samples is costly and time-consuming.

To address this challenge, various approaches have been proposed to augment the training data automatically, which range from altering the texts of existing training samples [54, 64], to leveraging generative models to produce new instances based on initial seed samples [68, 2, 34] with LLMs that eliminates the burden of performing task-specific training [28, 65, 35]. However, in low-resource environments where only a limited number of training instances are available, generating new data from these minimal seed samples results in poor diversity and variation (See Figure 1, (B)). We note that, while a recent approach attempts to overcome this by iteratively including generated samples as seed data for further data generation [60], it is still ill-suited, which is not only constrained by the limited diversity of the initial seed data but also vulnerable to recursively diminishing the quality of subsequent augmentations due to the potential low-quality of prior augmentations.

Despite the limited seed data in low-resource settings, there is an abundance of examples and resources accumulated in existing data pools, which can be utilized for data augmentation. Moreover,

*Equal contribution

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

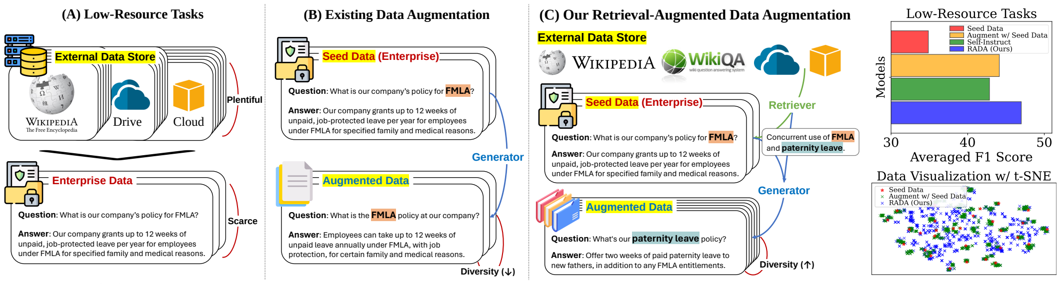


Figure 1: **(A) Low-Resource Tasks** refer to problems (usually on the specific domains) where there is a limited amount of data available. **(B) Existing Data Augmentation** approaches expand the seed data with itself, which results in the limited diversity of the generated data samples. **(C) Our Retrieval-Augmented Data Augmentation (RADA)** framework generates the new data with the external context, retrieved from the external datasets, along with the seed data, yielding more diverse and useful samples. **(Upper Right:)** Our RADA outperforms existing methods, demonstrating the quality of generated samples. **(Lower Right:)** The generated data samples from RADA are more diverse than existing data augmentation, based on the t-SNE visualization.

by leveraging the contextual understanding capabilities of LLMs, we can effectively utilize a mixture of samples drawn from the initial seed data, other datasets, or a combination of both. This can enable the synthesis of new samples, which mirror the characteristics of the seed data while being diverse.

However, not all samples from external datasets are useful for data augmentation, as most of them may not align with the characteristics of the seed data. Thus, inspired by the motivation to use external data instances while overcoming the problem of many of their irrelevancies, in this work, we propose a novel LLM-powered Retrieval-Augmented Data Augmentation (RADA) framework (See Figure 1, (C)). Specifically, the input of our data augmentation approach consists of in-context examples containing example instances, along with a target context that elicits a new sample generation. Then, our RADA flexibly employs multiple retrieval strategies to construct these in-context and target-context with samples from both original and external datasets, enabling diverse data augmentation.

We validate the effectiveness of RADA in augmenting low-resource datasets on multiple domain-specific datasets, where we consider both the training and test-time data augmentation scenarios. Then, the experimental results show that RADA consistently surpasses several LLM-powered data augmentation baselines. In addition, a key finding from our analyses is the dual benefit offered by our RADA: the incorporation of external data sources enhances the diversity of the generated instances, while the retrieval mechanism ensures maintaining their semantic alignment with the initial seed data.

# 2 Methodology

## 2.1 Problem Statement

**Low-Resource Domain-Specific Tasks** Before explaining the low-resource tasks that we focus on, we define conventional natural language tasks. Formally, their goal is to predict a label $\boldsymbol{y}$ given an input $\boldsymbol{x}$, where $\boldsymbol{x}$ and $\boldsymbol{y}$ are comprised of a sequence of tokens: $\boldsymbol{x} = [x_1, x_2, ..., x_{|\boldsymbol{x}|}]$ and $\boldsymbol{y} = [y_1, y_2, ..., y_{|\boldsymbol{y}|}]$. Then, the training data $\mathcal{D}$ can be represented as an aggregation of input-output pairs: $\mathcal{D} = \{(\boldsymbol{x}_i, \boldsymbol{y}_i)\}_{i=1}^{N}$ where its size $N$ can vary widely from just a few dozens to several millions. In this work, we target handling challenging scenarios where $N$ is notably small, referred to as low-resource settings. These settings are particularly prevalent in domain-specific tasks (for example, within legal, medical, or technical fields), where the availability of labeled data is inherently limited due to the specialized nature of the domain or the scarcity of domain experts for annotation.

**LLMs for Data Augmentation** A typical way to handle the low-resource domain tasks is to expand the training data $\mathcal{D}$ with data augmentation techniques, which has been recently powered by LLMs due to their strong text-generation capabilities. Formally, let us first describe the LLM as a model parameterized by $\theta$, which takes the input $\boldsymbol{x}$ and generates the output $\boldsymbol{y}$, as follows: $\boldsymbol{y} = \texttt{LLM}_\theta(\boldsymbol{x})$. Here, $\theta$ is trained with massive text corpora with several training strategies and, after that, it usually remains fixed due to the costs of further training. Also, $\boldsymbol{x}$ can be any form of text, referred to as a prompt, which includes task-dependent instructions and contexts (such as demonstrations), to guide LLMs in generating outputs that align with the user's intent, which is data augmentation in our work.

The primary goal of data augmentation is to expand the diversity and amount of data $\mathcal{D}$ available for model training (and for testing in certain use cases such as test-time adaption), without manually collecting the new data, for tackling specific tasks especially on low-resource domains. Formally, this data augmentation process can be represented as follows: $\mathcal{D}' = f(\mathcal{D})$, where $f$ is the model (or technique) designed to generate new input-output pairs $(\boldsymbol{x}', \boldsymbol{y}')$ for the augmented dataset $\mathcal{D}'$, which is achieved by leveraging the underlying patterns, contexts, and knowledge existing in seed data $\mathcal{D}$. However, unlike existing works that mainly focus on expanding the original data $\mathcal{D}$ with itself, we can potentially incorporate any external sources of information easily available at hand, which could introduce greater diversity and quality in generating the samples for data augmentation. In addition, especially in low-resource settings, the available data to use as a source for expansion is largely scarce, which poses a challenge as the augmentation method $f$ is operationalized with only limited samples, leading to the generation of samples that may lack the desired diversity and quality.

### 2.2 Retrieval-Augmented Data Augmentation

To tackle the aforementioned drawbacks of existing data augmentation approaches, we propose a novel data augmentation method (from a different angle), that leverages available external datasets.

**Data Generation with External Resources** We redefine the concept of previous data augmentation to incorporate samples from external resources, as follows: $\mathcal{D}' = f(\mathcal{D}, \mathcal{C})$ where $\mathcal{C}$ is an external data store that is composed of input-output pairs $(\boldsymbol{x}, \boldsymbol{y})$ aggregated from all available datasets. However, not all the external data samples can be accommodated within the context length of LLMs, but also many of these samples may not be pertinent for generating valuable augmentations for $\mathcal{D}$.

**Retrieving Relevant Instances** To tackle the aforementioned challenges, we propose to retrieve contextually relevant instances from the data store $\mathcal{C}$, which is critical as it ensures that the data produced by LLMs is not only diverse and high-quality but also contextually coherent and aligned with the nuances of the target dataset $\mathcal{D}$. In the following, we first provide the general formulation of the retrieval and then propose our two specific instantiations of the retrieval for data augmentation.

Formally, for a given input instance $\boldsymbol{q}$, the goal of a retriever is to identify and fetch a ranked list of $k$ entries from a large corpus, which are deemed most relevant to the input, represented as follows: $\{\boldsymbol{c}_i\}_{i=1}^{k} = \texttt{Retriever}(\boldsymbol{q}, \mathcal{C})$ where $\boldsymbol{c}_i \in \mathcal{C}$. Here, $\boldsymbol{q}$ can be a textual query; $\mathcal{C}$ is the corpus (which is typically a large collection of documents) from which the relevant information is to be retrieved; `Retriever` is designed with keyword-based algorithms or neural embedding-based models [52, 32].

#### 2.2.1 Retrieval for Data Augmentation

The input to LLMs can be viewed from two different perspectives: in-context learning which refers to their ability to learn from the input demonstrations; task-solving where the model executes specific tasks requested by users (e.g., data augmentation). According to them, we propose two instantiations of retrieval for LLM-powered data augmentation (See Figure 2).

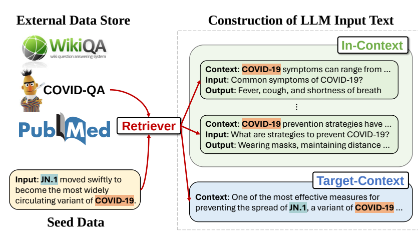


Figure 2: **RADA Framework Overview**. We first retrieve the external instances (relevant to the seed data) from the external data store, and construct in-context and target-context of LLM prompts with the retrieved samples along with the seed data.

**Retrieval for In-Context Learning** In-context learning plays a crucial role in enabling LLMs to align their outputs with the contextual cues provided in the input examples. Similarly, in data augmentation, it may enable LLMs to learn from examples (e.g., input-output pairs) in the seed data, to generate new input-output pairs. Yet, in low-resource settings, the combination of data samples to provide as the examples in the input prompt is largely limited. This limitation highlights the advantage of our retrieval-augmented data augmentation framework, which can fill the input demonstrations with samples from external datasets. Yet, as not all the samples are relevant, we retrieve only the relevant samples based on the similarity between the sample in seed data $\mathcal{D}$ and the external sample in data store $\mathcal{C}$, as follows: $\{\boldsymbol{c}_i\}_{i=1}^{k} = \texttt{Retriever}(\boldsymbol{q}, \mathcal{C})$ where $\boldsymbol{q} \in \mathcal{D}$. Mathematically, the combination of demonstrations to use as the LLM input is expanded to $O((k \times |\mathcal{D}|)^3)$ from $O(|\mathcal{D}|^3)$, where $|\mathcal{D}|$ is typically small in the low-resource setting.

**Retrieval for Target Sample Generation** Unlike in-context examples providing background information for data augmentation, the context to be retrieved and used here has a different goal, which should serve as a source for generating a complete input-output pair or one among them when given the other, depending on specific use cases. Specifically, a certain document can be used as a context to derive a query-answer pair, along with their in-context examples. Another example is to provide a question as a context and then generate its answers, or vice versa to augment queries.

Table 1: **Training data augmentation results**, where 10, 30, and 100 denote the number of initial seed data.

| | Covid QA | | | Policy QA | | | Tech QA | | | Average | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Methods** | 10 | 30 | 100 | 10 | 30 | 100 | 10 | 30 | 100 | 10 | 30 | 100 |
| Seed Data | 57.07 | 66.93 | 68.97 | 6.25 | 16.26 | 28.09 | 12.28 | 17.59 | 33.90 | 25.20 | 33.59 | 43.65 |
| PAQ (non-LLM) | 65.23 | 66.55 | 66.72 | 24.37 | 25.87 | 27.48 | 24.03 | 25.65 | 29.89 | 37.88 | 39.36 | 41.36 |
| Augment w/ Seed Data | 62.74 | 64.69 | 65.01 | 28.08 | 27.49 | 25.89 | 40.20 | 42.07 | 42.42 | 43.67 | 44.75 | 44.44 |
| Self-Instruct | 63.34 | 61.90 | 64.20 | 27.48 | 27.50 | 27.53 | 33.20 | 39.13 | 37.55 | 41.34 | 42.84 | 43.09 |
| QA Generation | 51.72 | 48.98 | 39.05 | 20.04 | 20.46 | 20.95 | 30.01 | 30.99 | 32.80 | 33.92 | 33.48 | 30.93 |
| CQA Generation | 67.00 | 67.01 | 67.80 | 27.30 | 24.96 | 25.94 | 28.08 | 30.94 | 31.88 | 40.79 | 40.97 | 41.87 |
| Seed + External Data | 62.30 | 62.81 | 63.50 | 25.72 | 25.60 | 29.34 | 34.82 | 35.46 | 37.06 | 40.95 | 41.29 | 43.30 |
| **RADA (Ours)** | 67.55 | 67.95 | 68.36 | 28.83 | 28.25 | 28.88 | 40.44 | 44.41 | 45.81 | **45.61** | **46.87** | **47.68** |

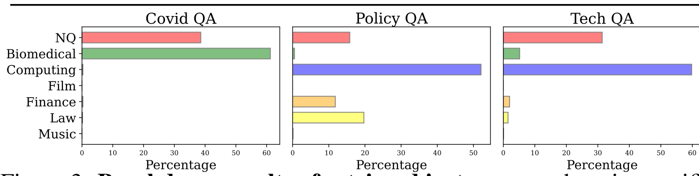


Figure 3: **Breakdown results of retrieved instances** on domain-specific data, where samples in the retrieval pool are one of Biomedical, Computing, Film, Finance, Law, and Music, as well as NQ (covering general domains).

| Domains | Covid QA | Tech QA |
|---|---|---|
| All | 67.55 | 40.44 |
| Biomedical | **67.75** | 40.09 |
| Computing | 66.70 | **42.67** |

Table 2: **Results with the hand-crafted data store**, selectively using only the most suitable external domain as the retrieval pool.

Formally, $\{\boldsymbol{c}_i\}_{i=1}^{k} = \texttt{Retriever}(\boldsymbol{q}, \mathcal{C})$ where $\boldsymbol{q}$ can be either the document or the question from $\mathcal{D}$. Also, the augmented samples generated directly from the retrieved instances are similar in nature to the original samples, as we consider relevant top-$k$ instances, ensuring a high degree of contextual coherence with seed samples while being more diverse against the generation with seed.

## 3 Experimental Setups and Results

**Experimental Setups** We validate our RADA on training data augmentation in Covid [44], Policy [1] and Tech [8] datasets and test-time data augmentation scenarios in MMLU [26]. For external resources for retrieval, we use Natural Questions (NQ) [33] and labeled subset [67] of MS MARCO [45], as well as MMLU's auxiliary data from similar datasets. For data augmentation, we use Llama2-7B-Chat [58] for all methods. For fine-tuning, we use T5-base [48] or Llama2-7B, to measure the effectiveness of different approaches without worrying about data contamination as they are not trained on any downstream tasks/datasets. We provide additional details in Appendix A.

**Main Results** We conduct experiments on two different data augmentation scenarios and report the results of training data augmentation in Table 1 and the test-time augmentation results in Table 3 (See Table 8 and Table 9 for standard deviations). As shown in them, RADA substantially outperforms all baselines, demonstrating its effectiveness. We note that the average score of the non-LLM-based PAQ approach is low, compared to LLM-based methods, which confirms the effectiveness of using LLMs for data augmentation perhaps thanks to their prior knowledge (See Appendix B for more results and discussion). Moreover, as shown in Table 3, RADA is highly effective in the challenging test-time data augmentation scenario (where no data is available for training), outperforming the model trained with all the external data instances. This may be due to our retrieval strategy, which results in generating samples that are relevant to the test data.

Table 3: **Test-time data augmentation results** on sub-domains of MMLU and domain-specific QA datasets.

| **MMLU** | **CS** | **Biology** | **Law** | **Average** |
|---|---|---|---|---|
| 5-Shots w/ Training | 32.00 | 47.74 | 64.46 | 48.07 |
| External Data | 48.00 | 54.52 | 66.12 | 56.21 |
| **RADA (Ours)** | **49.00** | **55.48** | **70.25** | **58.24** |
| **Domain-Specific QA** | **Covid** | **Policy** | **Tech** | **Average** |
| External Data | 54.02 | 19.32 | 12.97 | 28.77 |
| PAQ (non-LLM) | 61.22 | 25.03 | 19.83 | 35.36 |
| **RADA (Ours)** | **66.03** | **29.14** | **29.17** | **41.45** |

**Analysis of Retrieval** To understand which data instances are retrieved for data augmentation and what are their effectiveness, we conduct a comprehensive analysis. Firstly, we visualize the categories of retrieved instances for domain-specific QA in Figure 3, which shows that (mostly) only the relevant instances are retrieved and used for data augmentation for each specific task. For example, the Biomedical domain is the dominant field of retrieval source for Covid QA; meanwhile, the Computing domain is for Tech QA. In addition, to see the contribution of relevant retrieval, we restrict the retrieval domain to the one that is the most relevant to the given specific dataset. For example, we use only the Biomedical domain for Covid QA and the Computing domain for Tech QA. As shown in Table 2, we observe that when manipulating the retrieval pool, the performance further increases (as instances from irrelevant domains are not retrieved), which reaffirms the effectiveness of retrieval and its room for improvement for data augmentation.

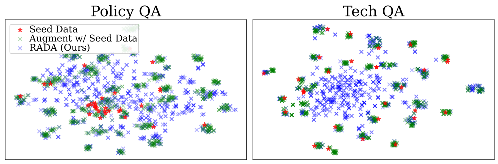


Figure 4: **Embedding-space visualization of samples** including the seed data and augmented data.


Figure 5: **ROUGE-L score distributions** measured between the seed and generated data on Tech QA.

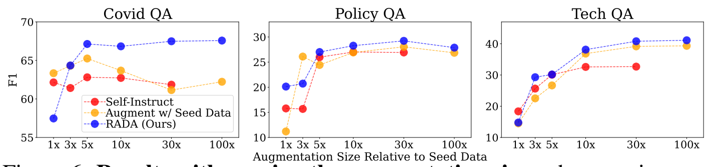


Figure 6: **Results with varying the augmentation size**, where we increase the size by factors of 1, 3, 5, 10, 30, and 100 relative to the seed data size.

| **Methods** | **Tech QA** |
|---|---|
| **RADA (Ours)** | **44.41** |
| w/o In-context Retriever | 41.24 |
| w/o Target-context Retriever | 34.42 |
| w/o All Retrievers | 30.38 |

Table 4: **Ablation study** of RADA on the Tech QA dataset.

**Analysis of Augmented Data Diversity** A notable advantage of RADA is that it intuitively can generate more diverse samples than what could be achieved by existing data augmentation approaches that use the seed data alone, by augmenting this process with the retrieval from external data samples. To measure this ability, we visualize the embedding space of the augmented samples in Figure 4 and report their lexical overlaps in Figure 5. Specifically, for the visualization, we first embed the generated instances with Sentence-BERT [50] into the latent space and project them with t-SNE [59]. From this, we observe that, unlike Augment w/ Seed Data whose generated samples are close to the seed data, the samples generated from RADA are broadly dispersed. Further, we measure the max ROUGE-L scores between the seed and generated instances where lower scores indicate higher diversity. As shown in Figure 5, RADA generates distinct samples to the seed data thanks to retrieving and utilizing the external contexts beyond the seed data, unlike baselines that rely solely on it.

**Analysis of Augmented Data Size** To see how the performance changes as a function of the size of augmented data samples, we vary the augmentation size relative to the seed data size and report the results in Figure 6[2]. Firstly, when the amount of augmented data is very small, baseline performances are comparable with RADA since the data samples that can be generated from the seed data alone can have a certain diversity level as we augment only a small amount. Yet, as the augmentation size expands, RADA consistently outperforms baselines, showcasing its ability to generate broader and richer samples through retrieval augmentation.

**Ablation Study** To see how each component of RADA affects the overall performance, we conduct an ablation study where we replace our in-context and target-context retrieval modules with random retrievals. As shown in Table 4, we observe that, without retrieving relevant instances, the performances drop substantially since irrelevant samples are used to construct the in-context examples and target context, leading to generating the samples not useful. Also, the target-context retriever is particularly important for data augmentation, as this is used to directly derive instances for training.

## 4 Conclusion

In this work, we pointed out the limitation of existing data augmentation approaches that use the seed data alone, leading to generating suboptimal and less diverse instances, despite the existence of plenty of external samples available. Inspired by this, we proposed the LLM-powered Retrieval-Augmented Data Augmentation (RADA) framework, which augments the seed data by leveraging samples retrieved from the external data store based on their relevance with the seed data. Specifically, the input to LLMs for data augmentation can be viewed from two different angles of in-context examples and task-solving context, and we constructed them through samples from within and across the seed data and the retrieved data. Through extensive evaluation results on multiple datasets with training and test-time data augmentation scenarios, we showed that RADA outperforms strong LLM-powered data augmentation baselines substantially. Also, our findings reveal that the data samples generated from our approach are much more diverse against baselines while being relevant to the seed data, due to leveraging retrieval for data augmentation. We believe that RADA will pave the way for enhancing the model performances on realistic low-resource domain-specific tasks, which have arisen as very important problems recently due to the limited availability and privacy concerns of data.

[2]Due to the cost of Self-Instruct, we are not able to generate its samples for the 100 times augmentation-level.

## References

[1] Wasi Uddin Ahmad, Jianfeng Chi, Yuan Tian, and Kai-Wei Chang. Policyqa: A reading comprehension dataset for privacy policies. In Trevor Cohn, Yulan He, and Yang Liu, editors, *Findings of the Association for Computational Linguistics: EMNLP 2020, Online Event, 16-20 November 2020*, volume EMNLP 2020 of *Findings of ACL*, pages 743–749. Association for Computational Linguistics, 2020. doi: 10.18653/V1/2020.FINDINGS-EMNLP.66. URL https://doi.org/10.18653/v1/2020.findings-emnlp.66.

[2] Ateret Anaby-Tavor, Boaz Carmeli, Esther Goldbraich, Amir Kantor, George Kour, Segev Shlomov, Naama Tepper, and Naama Zwerdling. Do not have enough data? deep learning to the rescue! In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 7383–7390. AAAI Press, 2020. URL https://doi.org/10.1609/aaai.v34i05.6233.

[3] Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M. Dai, Anja Hauth, Katie Millican, David Silver, Slav Petrov, Melvin Johnson, Ioannis Antonoglou, Julian Schrittwieser, Amelia Glaese, Jilin Chen, Emily Pitler, Timothy P. Lillicrap, Angeliki Lazaridou, Orhan Firat, James Molloy, Michael Isard, Paul Ronald Barham, Tom Hennigan, Benjamin Lee, Fabio Viola, Malcolm Reynolds, Yuanzhong Xu, Ryan Doherty, Eli Collins, Clemens Meyer, Eliza Rutherford, Erica Moreira, Kareem Ayoub, Megha Goel, George Tucker, Enrique Piqueras, Maxim Krikun, Iain Barr, Nikolay Savinov, Ivo Danihelka, Becca Roelofs, Anaïs White, Anders Andreassen, Tamara von Glehn, Lakshman Yagati, Mehran Kazemi, Lucas Gonzalez, Misha Khalman, Jakub Sygnowski, and et al. Gemini: A family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023. URL https://doi.org/10.48550/arXiv.2312.11805.

[4] Jinheon Baek, Alham Fikri Aji, and Amir Saffari. Knowledge-augmented language model prompting for zero-shot knowledge graph question answering. *arXiv preprint arXiv:2306.04136*, 2023. URL https://doi.org/10.48550/arXiv.2306.04136.

[5] Teodoro Baldazzi, Luigi Bellomarini, Stefano Ceri, Andrea Colombo, Andrea Gentili, and Emanuel Sallinger. Fine-tuning large enterprise language models via ontological reasoning. In Anna Fensel, Ana Ozaki, Dumitru Roman, and Ahmet Soylu, editors, *Rules and Reasoning - 7th International Joint Conference, RuleML+RR 2023, Oslo, Norway, September 18-20, 2023, Proceedings*, volume 14244 of *Lecture Notes in Computer Science*, pages 86–94. Springer, 2023. URL https://doi.org/10.1007/978-3-031-45072-3_6.

[6] Luiz Henrique Bonifacio, Hugo Queiroz Abonizio, Marzieh Fadaee, and Rodrigo Frassetto Nogueira. Inpars: Unsupervised dataset generation for information retrieval. In Enrique Amigó, Pablo Castells, Julio Gonzalo, Ben Carterette, J. Shane Culpepper, and Gabriella Kazai, editors, *SIGIR '22: The 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, Madrid, Spain, July 11 - 15, 2022*, pages 2387–2392. ACM, 2022. doi: 10.1145/3477495.3531863. URL https://doi.org/10.1145/3477495.3531863.

[7] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin, editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/1457c0d6bfcb4967418bfb8ac142f64a-Abstract.html.

[8] Vittorio Castelli, Rishav Chakravarti, Saswati Dana, Anthony Ferritto, Radu Florian, Martin Franz, Dinesh Garg, Dinesh Khandelwal, J. Scott McCarley, Mike McCawley, Mohamed Nasr, Lin Pan, Cezar Pendus, John F. Pitrelli, Saurabh Pujar, Salim Roukos, Andrzej Sakrajda, Avirup

Sil, Rosario Uceda-Sosa, Todd Ward, and Rong Zhang. The techqa dataset. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel R. Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 1269–1278. Association for Computational Linguistics, 2020. doi: 10.18653/V1/2020.ACL-MAIN.117. URL `https://doi.org/10.18653/v1/2020.acl-main.117`.

[9] Jiaao Chen, Zichao Yang, and Diyi Yang. Mixtext: Linguistically-informed interpolation of hidden space for semi-supervised text classification. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel R. Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 2147–2157. Association for Computational Linguistics, 2020. URL `https://doi.org/10.18653/v1/2020.acl-main.194`.

[10] Meng Chen, Hongyu Zhang, Chengcheng Wan, Zhao Wei, Yong Xu, Juhong Wang, and Xiaodong Gu. On the effectiveness of large language models in domain-specific code generation. *arXiv preprint arXiv:2312.01639*, 2023. URL `https://doi.org/10.48550/arXiv.2312.01639`.

[11] Mingda Chen, Jingfei Du, Ramakanth Pasunuru, Todor Mihaylov, Srini Iyer, Veselin Stoyanov, and Zornitsa Kozareva. Improving in-context few-shot learning via self-supervised training. In Marine Carpuat, Marie-Catherine de Marneffe, and Iván Vladimir Meza Ruíz, editors, *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2022, Seattle, WA, United States, July 10-15, 2022*, pages 3558–3573. Association for Computational Linguistics, 2022. URL `https://doi.org/10.18653/v1/2022.naacl-main.260`.

[12] Mingda Chen, Xilun Chen, and Wen-tau Yih. Efficient open domain multi-hop question answering with few-shot data synthesis. *arXiv preprint arXiv:2305.13691*, abs/2305.13691, 2023. doi: 10.48550/ARXIV.2305.13691. URL `https://doi.org/10.48550/arXiv.2305.13691`.

[13] Haixing Dai, Zhengliang Liu, Wenxiong Liao, Xiaoke Huang, Yihan Cao, Zihao Wu, Lin Zhao, Shaochen Xu, Wei Liu, Ninghao Liu, et al. Auggpt: Leveraging chatgpt for text data augmentation. *arXiv preprint arXiv:2302.13007*, 2023. URL `https://api.semanticscholar.org/CorpusID:257631936`.

[14] Zhuyun Dai, Vincent Y. Zhao, Ji Ma, Yi Luan, Jianmo Ni, Jing Lu, Anton Bakalov, Kelvin Guu, Keith B. Hall, and Ming-Wei Chang. Promptagator: Few-shot dense retrieval from 8 examples. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023. URL `https://openreview.net/pdf?id=gmL46YMpu2J`.

[15] Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. Qlora: Efficient finetuning of quantized llms. *arXiv preprint arXiv:2305.14314*, abs/2305.14314, 2023. doi: 10.48550/ARXIV.2305.14314. URL `https://doi.org/10.48550/arXiv.2305.14314`.

[16] Bosheng Ding, Chengwei Qin, Linlin Liu, Lidong Bing, Shafiq R. Joty, and Boyang Albert Li. Is gpt-3 a good data annotator? In *Annual Meeting of the Association for Computational Linguistics*, 2022. URL `https://api.semanticscholar.org/CorpusID:254877171`.

[17] Bosheng Ding, Chengwei Qin, Ruochen Zhao, Tianze Luo, Xinze Li, Guizhen Chen, Wenhan Xia, Junjie Hu, Anh Tuan Luu, and Shafiq R. Joty. Data augmentation using large language models: Data perspectives, learning paradigms and challenges. 2024. URL `https://api.semanticscholar.org/CorpusID:268249187`.

[18] Chandra Kiran Reddy Evuru, Sreyan Ghosh, Sonal Kumar, S. Ramaneswaran, Utkarsh Tyagi, and Dinesh Manocha. Coda: Constrained generation based data augmentation for low-resource nlp. *ArXiv*, abs/2404.00415, 2024. URL `https://api.semanticscholar.org/CorpusID:268819699`.

[19] Steven Y. Feng, Varun Gangal, Jason Wei, Sarath Chandar, Soroush Vosoughi, Teruko Mitamura, and Eduard H. Hovy. A survey of data augmentation approaches for NLP. In Chengqing Zong,

Fei Xia, Wenjie Li, and Roberto Navigli, editors, *Findings of the Association for Computational Linguistics: ACL/IJCNLP 2021, Online Event, August 1-6, 2021*, volume ACL/IJCNLP 2021 of *Findings of ACL*, pages 968–988. Association for Computational Linguistics, 2021. URL `https://doi.org/10.18653/v1/2021.findings-acl.84`.

[20] Saumya Gandhi, Ritu Gala, Vijay Viswanathan, Tongshuang Wu, and Graham Neubig. Better synthetic data by retrieving and transforming existing datasets. *ArXiv*, abs/2404.14361, 2024. URL `https://api.semanticscholar.org/CorpusID:269292964`.

[21] Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Qianyu Guo, Meng Wang, and Haofen Wang. Retrieval-augmented generation for large language models: A survey. *ArXiv*, abs/2312.10997, 2023. URL `https://api.semanticscholar.org/CorpusID:266359151`.

[22] Sreyan Ghosh, Utkarsh Tyagi, Sonal Kumar, Chandra Kiran, Reddy Evuru, S. Ramaneswaran, S Sakshi, Dinesh Manocha, Adams Wei Yu, David Dohan, Minh-Thang Luong, Rui Zhao, Kai Chen, and Mohammad Norouzi. Abex: Data augmentation for low-resource nlu via expanding abstract descriptions. *ArXiv*, abs/2406.04286, 2024. URL `https://api.semanticscholar.org/CorpusID:270285859`.

[23] Arnav Gudibande, Eric Wallace, Charlie Snell, Xinyang Geng, Hao Liu, Pieter Abbeel, Sergey Levine, and Dawn Song. The false promise of imitating proprietary llms. *arXiv perprint arXiv:2305.15717*, 2023. URL `https://doi.org/10.48550/arXiv.2305.15717`.

[24] Biyang Guo, Yeyun Gong, Yelong Shen, Songqiao Han, Hailiang Huang, Nan Duan, and Weizhu Chen. Genius: Sketch-based language model pre-training via extreme and selective masking for text generation and augmentation. *ArXiv*, abs/2211.10330, 2022. URL `https://api.semanticscholar.org/CorpusID:253708143`.

[25] Demi Guo, Yoon Kim, and Alexander M. Rush. Sequence-level mixed sample data augmentation. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 5547–5552. Association for Computational Linguistics, 2020. URL `https://doi.org/10.18653/v1/2020.emnlp-main.447`.

[26] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. URL `https://openreview.net/forum?id=d7KBjmI3GmQ`.

[27] Sebastian Hofstätter, Sheng-Chieh Lin, Jheng-Hong Yang, Jimmy Lin, and Allan Hanbury. Efficiently teaching an effective dense retriever with balanced topic aware sampling. In Fernando Diaz, Chirag Shah, Torsten Suel, Pablo Castells, Rosie Jones, and Tetsuya Sakai, editors, *SIGIR '21: The 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, Virtual Event, Canada, July 11-15, 2021*, pages 113–122. ACM, 2021. doi: 10.1145/3404835.3462891. URL `https://doi.org/10.1145/3404835.3462891`.

[28] Or Honovich, Thomas Scialom, Omer Levy, and Timo Schick. Unnatural instructions: Tuning language models with (almost) no human labor. In Anna Rogers, Jordan L. Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 14409–14428. Association for Computational Linguistics, 2023. URL `https://doi.org/10.18653/v1/2023.acl-long.806`.

[29] Or Honovich, Thomas Scialom, Omer Levy, and Timo Schick. Unnatural instructions: Tuning language models with (almost) no human labor. In Anna Rogers, Jordan L. Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 14409–14428. Association for Computational Linguistics, 2023. doi: 10.18653/V1/2023.ACL-LONG.806. URL `https://doi.org/10.18653/v1/2023.acl-long.806`.

[30] Soyeong Jeong, Jinheon Baek, Sukmin Cho, Sung Ju Hwang, and Jong C. Park. Test-time self-adaptive small language models for question answering. *ArXiv*, abs/2310.13307, 2023. URL `https://api.semanticscholar.org/CorpusID:264406056`.

[31] Akbar Karimi, L. Rossi, and Andrea Prati. Aeda: An easier data augmentation technique for text classification. In *Conference on Empirical Methods in Natural Language Processing*, 2021. URL `https://api.semanticscholar.org/CorpusID:237353017`.

[32] Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. Dense passage retrieval for open-domain question answering. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.550. URL `https://aclanthology.org/2020.emnlp-main.550`.

[33] Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur P. 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. Natural questions: a benchmark for question answering research. *Trans. Assoc. Comput. Linguistics*, 7:452–466, 2019. doi: 10.1162/TACL_A_00276. URL `https://doi.org/10.1162/tacl_a_00276`.

[34] Dong Bok Lee, Seanie Lee, Woo Tae Jeong, Donghwan Kim, and Sung Ju Hwang. Generating diverse and consistent QA pairs from contexts with information-maximizing hierarchical conditional vaes. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel R. Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 208–224. Association for Computational Linguistics, 2020. URL `https://doi.org/10.18653/v1/2020.acl-main.20`.

[35] Dong-Ho Lee, Jay Pujara, Mohit Sewak, Ryen White, and Sujay Kumar Jauhar. Making large language models better data creators. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 15349–15360. Association for Computational Linguistics, 2023. URL `https://aclanthology.org/2023.emnlp-main.948`.

[36] Patrick Lewis, Yuxiang Wu, Linqing Liu, Pasquale Minervini, Heinrich Kuttler, Aleksandra Piktus, Pontus Stenetorp, and Sebastian Riedel. Paq: 65 million probably-asked questions and what you can do with them. *Transactions of the Association for Computational Linguistics*, 9: 1098–1115, 2021. URL `https://api.semanticscholar.org/CorpusID:231924957`.

[37] Bohan Li, Yutai Hou, and Wanxiang Che. Data augmentation approaches in natural language processing: A survey. *AI Open*, 3:71–90, 2022. URL `https://doi.org/10.1016/j.aiopen.2022.03.001`.

[38] Xian Li, Ping Yu, Chunting Zhou, Timo Schick, Luke Zettlemoyer, Omer Levy, Jason Weston, and Mike Lewis. Self-alignment with instruction backtranslation. *arXiv preprint arXiv:2308.06259*, abs/2308.06259, 2023. doi: 10.48550/ARXIV.2308.06259. URL `https://doi.org/10.48550/arXiv.2308.06259`.

[39] Chen Ling, Xujiang Zhao, Jiaying Lu, Chengyuan Deng, Can Zheng, Junxiang Wang, Tanmoy Chowdhury, Yun-Qing Li, Hejie Cui, Xuchao Zhang, Tian yu Zhao, Amit Panalkar, Wei Cheng, Haoyu Wang, Yanchi Liu, Zhengzhang Chen, Haifeng Chen, Chris White, Quanquan Gu, Jian Pei, Carl Yang, and Liang Zhao. Domain specialization as the key to make large language models disruptive: A comprehensive survey. 2023. URL `https://api.semanticscholar.org/CorpusID:259502302`.

[40] Quanyu Long, Wenya Wang, and Sinno Jialin Pan. Adapt in contexts: Retrieval-augmented domain adaptation via in-context learning. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 6525–6542. Association for Computational Linguistics, 2023. URL `https://aclanthology.org/2023.emnlp-main.402`.

[41] Kai Lv, Yuqing Yang, Tengxiao Liu, Qinghui Gao, Qipeng Guo, and Xipeng Qiu. Full parameter fine-tuning for large language models with limited resources. *arXiv preprint arXiv:2306.09782*, 2023. URL `https://doi.org/10.48550/arXiv.2306.09782`.

[42] Xinxi Lyu, Sewon Min, Iz Beltagy, Luke Zettlemoyer, and Hannaneh Hajishirzi. Z-ICL: zero-shot in-context learning with pseudo-demonstrations. In Anna Rogers, Jordan L. Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 2304–2317. Association for Computational Linguistics, 2023. doi: 10.18653/V1/2023.ACL-LONG.129. URL `https://doi.org/10.18653/v1/2023.acl-long.129`.

[43] Sewon Min, Mike Lewis, Luke Zettlemoyer, and Hannaneh Hajishirzi. Metaicl: Learning to learn in context. In Marine Carpuat, Marie-Catherine de Marneffe, and Iván Vladimir Meza Ruíz, editors, *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2022, Seattle, WA, United States, July 10-15, 2022*, pages 2791–2809. Association for Computational Linguistics, 2022. doi: 10.18653/V1/2022.NAACL-MAIN.201. URL `https://doi.org/10.18653/v1/2022.naacl-main.201`.

[44] Timo Möller, Anthony Reina, Raghavan Jayakumar, and Malte Pietsch. Covid-qa: A question answering dataset for covid-19. 2020. URL `https://api.semanticscholar.org/CorpusID:229079829`.

[45] Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. MS MARCO: A human generated machine reading comprehension dataset. In Tarek Richard Besold, Antoine Bordes, Artur S. d'Avila Garcez, and Greg Wayne, editors, *Proceedings of the Workshop on Cognitive Computation: Integrating neural and symbolic approaches 2016 co-located with the 30th Annual Conference on Neural Information Processing Systems (NIPS 2016), Barcelona, Spain, December 9, 2016*, volume 1773 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2016. URL `https://ceur-ws.org/Vol-1773/CoCoNIPS_2016_paper9.pdf`.

[46] OpenAI. GPT-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. doi: 10.48550/ARXIV.2303.08774. URL `https://doi.org/10.48550/arXiv.2303.08774`.

[47] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Z. Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 8024–8035, 2019. URL `https://proceedings.neurips.cc/paper/2019/hash/bdbca288fee7f92f2bfa9f7012727740-Abstract.html`.

[48] 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. *J. Mach. Learn. Res.*, 2020. URL `http://jmlr.org/papers/v21/20-074.html`.

[49] Ori Ram, Yoav Levine, Itay Dalmedigos, Dor Muhlgay, Amnon Shashua, Kevin Leyton-Brown, and Yoav Shoham. In-context retrieval-augmented language models. *arXiv preprint arXiv:2302.00083*, abs/2302.00083, 2023. doi: 10.48550/ARXIV.2302.00083. URL `https://doi.org/10.48550/arXiv.2302.00083`.

[50] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 11 2019. URL `http://arxiv.org/abs/1908.10084`.

[51] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan, editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 3980–3990. Association for Computational Linguistics, 2019. doi: 10.18653/V1/D19-1410. URL `https://doi.org/10.18653/v1/D19-1410`.

[52] Stephen E. Robertson, Steve Walker, Susan Jones, Micheline Hancock-Beaulieu, and Mike Gatford. Okapi at TREC-3. In Donna K. Harman, editor, *Proceedings of The Third Text REtrieval Conference, TREC 1994, Gaithersburg, Maryland, USA, November 2-4, 1994*, volume 500-225 of *NIST Special Publication*, pages 109–126. National Institute of Standards and Technology (NIST), 1994. URL `http://trec.nist.gov/pubs/trec3/papers/city.ps.gz`.

[53] Jon Saad-Falcon, Omar Khattab, Keshav Santhanam, Radu Florian, Martin Franz, Salim Roukos, Avirup Sil, Md. Arafat Sultan, and Christopher Potts. UDAPDR: unsupervised domain adaptation via LLM prompting and distillation of rerankers. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 11265–11279. Association for Computational Linguistics, 2023. URL `https://aclanthology.org/2023.emnlp-main.693`.

[54] Gözde Gül Sahin and Mark Steedman. Data augmentation via dependency tree morphing for low-resource languages. In Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii, editors, *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 5004–5009. Association for Computational Linguistics, 2018. URL `https://doi.org/10.18653/v1/d18-1545`.

[55] Gaurav Sahu, Olga Vechtomova, Dzmitry Bahdanau, and Issam Hadj Laradji. Promptmix: A class boundary augmentation method for large language model distillation. *arXiv preprint arXiv:2310.14192*, 2023. URL `https://api.semanticscholar.org/CorpusID:264426761`.

[56] Vinay Samuel, Houda Aynaou, Arijit Ghosh Chowdhury, Karthik Venkat Ramanan, and Aman Chadha. Can llms augment low-resource reading comprehension datasets? opportunities and challenges. *arXiv preprint arXiv:2309.12426*, abs/2309.12426, 2023. doi: 10.48550/ARXIV.2309.12426. URL `https://doi.org/10.48550/arXiv.2309.12426`.

[57] Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. BEIR: A heterogenous benchmark for zero-shot evaluation of information retrieval models. *arXiv preprint arXiv:2104.08663*, abs/2104.08663, 2021. URL `https://arxiv.org/abs/2104.08663`.

[58] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton-Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurélien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023. URL `https://doi.org/10.48550/arXiv.2307.09288`.

[59] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9:2579–2605, 2008. URL `http://www.jmlr.org/papers/v9/vandermaaten08a.html`.

[60] Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A. Smith, Daniel Khashabi, and Hannaneh Hajishirzi. Self-instruct: Aligning language models with self-generated instructions. In Anna Rogers, Jordan L. Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 13484–13508. Association for Computational Linguistics, 2023. doi: 10.18653/V1/2023.ACL-LONG.754. URL `https://doi.org/10.18653/v1/2023.acl-long.754`.

[61] Yue Wang, Xinrui Wang, Juntao Li, Jinxiong Chang, Qishen Zhang, Zhongyi Liu, Guannan Zhang, and Min Zhang. Harnessing the power of david against goliath: Exploring instruction data generation without using closed-source models. *arXiv preprint arXiv:2308.12711*, abs/2308.12711, 2023. doi: 10.48550/ARXIV.2308.12711. URL `https://doi.org/10.48550/arXiv.2308.12711`.

[62] Jason Wei and Kai Zou. Eda: Easy data augmentation techniques for boosting performance on text classification tasks. In *Conference on Empirical Methods in Natural Language Processing*, 2019. URL `https://api.semanticscholar.org/CorpusID:59523656`.

[63] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In Sanmi Koyejo, S. Mohamed, A. Agarwal, Danielle Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, 2022. URL `http://papers.nips.cc/paper_files/paper/2022/hash/9d5609613524ecf4f15af0f7b31abca4-Abstract-Conference.html`.

[64] Jason W. Wei and Kai Zou. EDA: easy data augmentation techniques for boosting performance on text classification tasks. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan, editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 6381–6387. Association for Computational Linguistics, 2019. URL `https://doi.org/10.18653/v1/D19-1670`.

[65] Chenxi Whitehouse, Monojit Choudhury, and Alham Fikri Aji. Llm-powered data augmentation for enhanced crosslingual performance. *arXiv preprint arXiv:2305.14288*, 2023. URL `https://doi.org/10.48550/arXiv.2305.14288`.

[66] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. Transformers: State-of-the-art natural language processing. In Qun Liu and David Schlangen, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, EMNLP 2020 - Demos, Online, November 16-20, 2020*, pages 38–45. Association for Computational Linguistics, 2020. doi: 10.18653/V1/2020.EMNLP-DEMOS.6. URL `https://doi.org/10.18653/v1/2020.emnlp-demos.6`.

[67] Ying Xu, Xu Zhong, Antonio José Jimeno-Yepes, and Jey Han Lau. Forget me not: Reducing catastrophic forgetting for domain adaptation in reading comprehension. In *2020 International Joint Conference on Neural Networks, IJCNN 2020, Glasgow, United Kingdom, July 19-24, 2020*, pages 1–8. IEEE, 2020. doi: 10.1109/IJCNN48605.2020.9206891. URL `https://doi.org/10.1109/IJCNN48605.2020.9206891`.

[68] Kaichun Yao, Libo Zhang, Tiejian Luo, Lili Tao, and Yanjun Wu. Teaching machines to ask questions. In Jérôme Lang, editor, *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI 2018, July 13-19, 2018, Stockholm, Sweden*, pages 4546–4552. ijcai.org, 2018. URL `https://doi.org/10.24963/ijcai.2018/632`.

[69] Jiacheng Ye, Jiahui Gao, Qintong Li, Hang Xu, Jiangtao Feng, Zhiyong Wu, Tao Yu, and Lingpeng Kong. Zerogen: Efficient zero-shot learning via dataset generation. *arXiv preprint arXiv:2202.07922*, 2022. URL `https://api.semanticscholar.org/CorpusID:246867045`.

# A Additional Experimental Setups

## A.1 Tasks and Datasets

We validate our RADA on training data augmentation and test-time data augmentation scenarios.

**Training Data Augmentation** The goal of training data augmentation is to expand the samples available, which is useful when new events occur that the model needs to adapt to, while having only limited data for training. To test RADA with this scenario, we use three low-resource domain-specific datasets: Covid QA [44] that is annotated by medical doctors for tackling the COVID-19 pandemic; Policy QA [1] that is designed with specialized policies about website privacy; and Tech QA [8] that is constructed with questions on technical public forums for the IT domain. Additionally, to simulate the low-resource settings, we sample 10, 30, and 100 instances from the training dataset.

**Test-Time Data Augmentation** The assumption of test-time data augmentation is more challenging, considering the case where there is no data available for training due to strict privacy concerns (e.g., users or institutions may not want to share their private data to train models) [30]. For this scenario, we select and use three specific domains from the MMLU dataset [26] as it does not have direct training instances (aligned with our validation purpose), as well as using previous Covid QA, Policy QA, and Tech QA without any training samples for this setup.

**External Resources for Retrieval** We construct the external data store serving as a retrieval source by aggregating samples from other datasets. Specifically, for Covid QA, Policy QA, and Tech QA designed for open-domain Question Answering (QA), we use the Natural Questions (NQ) [33] and the labeled subset [67] of MS MARCO [45], covering broad domains with questions asked on web search. For MMLU that targets multi-choice QA, we use its official auxiliary data collected from similar datasets.

## A.2 Baselines and Our Model

We compare our approach to several LLM-powered data augmentation baselines to ensure a fair evaluation. Also, we include non-LLM-based approaches for reference purposes, contrasting them with LLM-based methods (see Appendix B for further discussion and results on them).

1. **Seed Data** – It uses only the seed data for training models without extra data augmentation steps.
2. **PAQ (non-LLM)** – It [36] is a state-of-the-art non-LLM-based method, which selects passages, extracts answers, generates questions, and filters some of them, with conventional NER tools.
3. **Augment w/ Seed Data** – It expands the seed data by generating new data instances from the seed data, where samples for in-context learning and target-context selection are randomly picked.
4. **Self-Instruct** – It [60] aims to bootstrap new tasks only with limited seed examples, by incorporating the generated data instances in the data pool and leveraging them along with the seed data iteratively, where the samples in the pool are used to construct the in-context and target samples.
5. **CQA Generation** – It [56] generates a context and then, based on it, subsequently generates a question-answer pair, where existing seed data samples are used for in-context learning. Its variant (**QA Generation**) generates a question-answer pair with in-context learning [69].
6. **Seed + External Data** – It trains the models with the seed data instances as well as all the instances available in the external data pool.
7. **RADA** – This is our model that generates samples by retrieving samples (relevant to the seed data) from the external corpus and using them for in-context and target context.

We note that, for the test-time data augmentation scenario, since the samples having complete input-output pairs are unavailable, we cannot compare against the baselines requiring in-context examples; yet, RADA can run with only the target context.

## A.3 Implementation Details

**Models** We use Llama2-7B-Chat [58] as the basis for data augmentation across all methods. For fine-tuning, we use either T5-base [48] or Llama2-7B, to measure the effectiveness of different approaches directly without worrying about data contamination as they are not trained on any

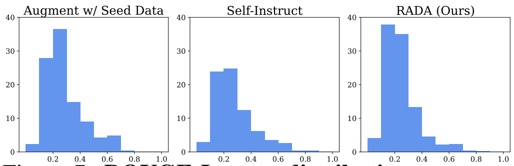


Figure 7: **ROUGE-L score distributions** measured between the seed and generated data on Covid QA.

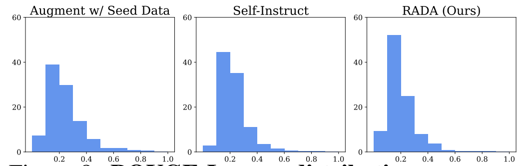


Figure 8: **ROUGE-L score distributions** measured between the seed and generated data on Policy QA.

downstream tasks/datasets. For the number of data augmented, unless otherwise stated, we produce samples amounting to 30 times that of the seed data and train models with the seed and generated data. A retriever used to retrieve instances is DistilBert TAS-B [27]. We report results with the F1 score for Covid QA, Policy QA, and Tech QA datasets, and the accuracy for MMLU, following standard evaluation protocols.

**Fine-tuning Details** Here, we provide more details on how to fine-tune models on the seed and augmented data samples. Firstly, for T5-base, we train it over 5 epochs with a batch size of 8 and a learning rate of $3\times10^{-5}$, selecting the best epoch to report the performance with inference. For Llama-7B, to train it with our computational resources available, we use the QLORA [15] technique, on which we use the epoch size of 30, the batch size of 1, and the learning rate of $2\times10^{-4}$. Lastly, we report the fine-tuning results with three runs.

**Prompts** The prompt used to elicit the data augmentation is provided in Table 12. For the domain-specific datasets including Covid QA, Policy QA, and Tech QA, we use the following prompt to generate the answer: "Context: { } Question: { } Answer: ". For the MMLU dataset, we use the following prompt: "Question: { } Answer Options: { } Answer:" where 5-shot examples prepended are the same as the one in the official code repository[3].

**Computational Resources and Time** We train and inference all baselines and our model by using one of the TITAN RTX, NVIDIA GeForce RTX 3080, NVIDIA GeForce RTX 3090, NVIDIA RTX A4000, NVIDIA RTX A5000, and Quadro RTX 8000 GPUs, depending on their availability at the time of run. The time required for training RADA ranges from a few minutes to about one and half day, which also depends on the number of the augmented data used for model fine-tuning.

**Deep Learning Libraries** In our experiments, we utilize the deep learning libraries as follows: PyTorch [47], Transformers [66], SentenceTransformers [51], and BEIR [57].

## B Additional Experimental Results

**More Analysis of Data Diversity** In addition to the result of ROUGE-L score distributions on Tech QA in Figure 5, we provide results on Covid QA and Policy QA in Figure 7 and Figure 8, respectively. Additionally, for their actual ROUGE-L scores, please see Table 5.

Table 5: The average ROUGE-scores between the original data samples and the augmented data samples.

| | Covid | Policy | Tech |
|---|---|---|---|
| Augment w/ Seed Data | 0.34 | 0.29 | 0.39 |
| Self-Instruct | 0.33 | 0.28 | 0.32 |
| **RADA (Ours)** | **0.30** | **0.25** | **0.24** |

To compare the diversity of augmented samples between other baselines and our method, we have provided further visualizations using t-SNE embeddings for Covid QA, Policy QA and Tech QA in Figure 9, Figure 10 and Figure 11, respectively.

**Results of Llama on Domain-Specific QA** Here we discuss the training data augmentation results of Llama on domain-specific QA data (such as Covid QA). Specifically, in Table 6, we report its 0-shot and 5-shot performances, as well as its fine-tuning performances on seed data and augmented data. As shown in Table 6,

Table 6: Training time augmentation results on Covid QA with T5 and Llama as the base for fine-tuning.

| # of seed | Bases | 0-shot | 5-shot | Seed | RADA (Ours) |
|---|---|---|---|---|---|
| 10 | T5 | N/A | N/A | **53.94** | **67.49** |
| | Llama2 | 12.79 | 16.43 | 50.62 | 56.50 |
| 30 | T5 | N/A | N/A | **66.50** | **68.15** |
| | Llama2 | 12.79 | 16.43 | 55.48 | 53.62 |

[3]https://github.com/hendrycks/test

Table 7: **Results with filtering mechanisms** on domain-specific QA with training data augmentation settings.

| Methods | Covid QA | | | Policy QA | | | Tech QA | | | Average | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 10 | 30 | 100 | 10 | 30 | 100 | 10 | 30 | 100 | 10 | 30 | 100 |
| **RADA (Ours)** | **67.49** | **68.15** | **68.57** | **29.23** | **28.49** | **29.18** | **40.81** | 44.37 | **46.93** | **45.84** | **47.00** | **48.23** |
| w/ ROUGE-based Filtering | 66.21 | 67.25 | 66.84 | 28.35 | 28.09 | 28.31 | 37.75 | **44.64** | 46.74 | 44.10 | 46.66 | 47.30 |
| w/ Embedding-based Filtering | 67.19 | 67.67 | 67.27 | 28.62 | 28.13 | 28.65 | 40.02 | **44.64** | 46.74 | 45.27 | 46.82 | 47.55 |
| w/o Answer Filtering | 66.78 | 66.65 | 67.09 | 28.78 | 28.44 | 29.12 | 40.55 | 42.43 | 42.56 | 45.37 | 45.84 | 46.26 |

Table 8: Training data augmentation results where we report the standard deviations in parentheses and the statistically significant results (under the t-test of p-value < 0.05) in bold.

| Methods | Covid QA | | | Policy QA | | | Tech QA | | |
|---|---|---|---|---|---|---|---|---|---|
| | 10 | 30 | 100 | 10 | 30 | 100 | 10 | 30 | 100 |
| Seed Data | 57.07 (2.76) | 66.93 (0.38) | 68.97 (0.46) | 6.25 (1.21) | 16.26 (3.46) | 28.09 (0.49) | 12.28 (2.37) | 17.59 (0.48) | 33.90 (2.34) |
| PAQ (non-LLM) | 65.23 (0.66) | 66.55 (0.24) | 66.72 (0.47) | 24.37 (0.18) | 25.87 (0.60) | 27.48 (0.46) | 24.03 (0.48) | 25.65 (1.39) | 29.89 (0.35) |
| Augment w/ Seed Data | 62.74 (1.41) | 64.69 (0.01) | 65.01 (0.51) | 28.08 (0.41) | 27.49 (0.47) | 25.89 (0.16) | 40.20 (0.92) | 42.07 (1.52) | 42.42 (1.01) |
| Self-Instruct | 63.34 (1.58) | 61.90 (0.18) | 64.20 (0.24) | 27.48 (0.53) | 27.50 (0.13) | 27.53 (0.27) | 33.20 (0.75) | 39.13 (0.76) | 37.55 (0.53) |
| QA Generation | 51.72 (1.15) | 48.98 (1.82) | 39.05 (1.91) | 20.04 (0.77) | 20.46 (0.55) | 20.95 (0.22) | 30.01 (0.13) | 30.99 (0.23) | 32.80 (0.78) |
| CQA Generation | 67.00 (0.32) | 67.01 (0.18) | 67.80 (0.17) | 27.30 (0.26) | 24.96 (0.17) | 25.94 (0.70) | 28.08 (0.92) | 30.94 (0.68) | 31.88 (0.95) |
| Seed + External Data | 62.30 (0.44) | 62.81 (0.28) | 63.50 (0.55) | 25.72 (0.41) | 25.60 (1.07) | 29.34 (0.12) | 34.82 (0.21) | 35.46 (0.94) | 37.06 (0.02) |
| **RADA (Ours)** | **67.55 (0.15)** | **67.95 (0.20)** | 68.36 (0.25) | **28.83 (0.37)** | **28.25 (0.21)** | 28.88 (0.50) | 40.44 (0.53) | **44.41 (0.45)** | **45.81 (0.97)** |

despite the large number of parameters that Llama2-7B has (which is ten times larger than T5), we observe that Llama2 is inferior to T5. We conjecture that this may be because the general massive corpus used to pre-train Llama2 has little (to no) overlap or relevance with instances in domain-specific tasks. In other words, eliciting the domain-specific ability of Llama2 with fine-tuning may be largely suboptimal, when it does not have internalized knowledge about its corresponding domain-specific tasks. In addition, this result may further highlight the fact that not all the larger models perform always better than the smaller models in low-resource settings, which gives us a promise to take advantage of computational efficiency, especially when dealing with extreme domain-specific tasks, or that specific LLMs may be required to handle each specific domain.

**Results with Filtering Strategies** We try various filtering approaches on the augmented data to fine-tune models with only the samples of high quality. Specifically, to further promote diversity in the generated samples from our RADA, we filter samples if they are similar to the already generated samples, based on their ROUGE scores or their embedding-level distances. Then, as shown in Table 7, these filtering techniques do not improve the model performance. This may further strengthen our claim that the augmented instances from RADA are already very diverse but also relevant to the seed data, which does not necessitate additional filtering mechanisms. On the other hand, if we relax the assumption that the passage should include the answer to the question for domain-specific QA, and subsequently do not apply the filtering strategy, the performance drops slightly in Table 7.

**Results with Standard Deviation** We report the average performance of three different runs and their standard deviation on training-time data augmentation and test-time data augmentation scenarios in Table 8 and Table 9, respectively. These results show that our proposed RADA achieves the statistically significant results over baselines on the most cases.

**More Results of Non-LLM-based Baselines** It is worth noting that making a comparison of LLM-based approaches (including our RADA) over non-LLM-based methods is unfair since different LMs have different capabilities in generating outputs, which leads to far different quality of augmented samples. Therefore, to ensure a fair comparison across all data augmentation approaches, we set Llama2 as the basis for data augmentation. Nevertheless, to see the efficacy of non-LLM-based approaches, we compare our

Table 9: Test-time data augmentation results where the standard deviations are in parentheses and the statistically significant results (p-value < 0.05) are in bold.

| Domain-Specific QA | Covid | Policy | Tech |
|---|---|---|---|
| External Data | 54.02 (0.42) | 19.32 (0.11) | 12.97 (0.52) |
| PAQ (non-LLM) | 61.22 (0.22) | 25.03 (0.34) | 19.83 (0.83) |
| **RADA (Ours)** | **66.03 (0.15)** | **29.14 (0.18)** | **29.17 (0.98)** |

Table 10: Comparison results of RADA against non-LLM-based methods on the challenging TechQA dataset, with the training time augmentation scenario. We report the standard deviations in parentheses and the statistically significant results (under the t-test) in bold.

| | 10 | 30 | 100 |
|---|---|---|---|
| PAQ | 24.03 (0.48) | 25.65 (1.39) | 29.89 (0.35) |
| GENIUS | 12.28 (2.37) | 26.90 (0.50) | 43.55 (0.45) |
| EDA | 38.27 (0.53) | 41.93 (0.26) | 45.21 (0.64) |
| AEDA | 38.86 (0.30) | 41.98 (0.30) | 45.24 (0.16) |
| **RADA (Ours)** | **40.44 (0.53)** | **44.41 (0.45)** | **45.81 (0.97)** |

RADA against several recent and popular (non-LLM-based) methods, namely PAQ [36], GENIUS [24], EDA [62], and AEDA [31], on the most challenging dataset (TechQA) that we observe in Table 1. Then, we report the results in Table 10. From this, we observe that RADA significantly outperforms previous non-LLM-based methods, demonstrating the effectiveness of using the LLM-based approach for data augmentation under low-resource settings, thanks to LLM's prior knowledge.

**Analysis of Using Different LLMs** We conduct an auxiliary analysis to see whether the superiority of RADA is consistent across different LLMs, compared to existing baselines. In particular, we use ChatGPT 3.5 (released on June 13, 2023) as the basis model for data augmentation, and report the results in Table 11. From this, we observe that RADA significantly outperforms baselines with another data augmentation LLM, demonstrating its robustness across different LLMs for data augmentation.

Table 11: Results of another LLM (ChatGPT) for data augmentation with seed examples of 10.

| | Covid | Policy | Tech | Average |
|---|---|---|---|---|
| Self-Instruct | 57.86 | 26.20 | 33.42 | 39.16 |
| CQA Generation | 65.64 | 27.20 | 34.16 | 42.33 |
| **RADA (Ours)** | **67.19** | **28.59** | **36.17** | **43.98** |

**Quantitative Analysis** In Table 13, 14, 15, we provide examples of the augmented instances across different methods on Covid QA, Policy QA, and Tech QA. A key finding from these results is that the existing approach that uses only the seed data results in a limited diversity of generated samples, unlike our RADA which generates distinct yet contextually coherent samples with the seed data, thanks to the retrieval of relevant external samples.

## C Related Work

In this section, we provide detailed discussions about the relevant literature.

**Large Language Models** Large Language Models (LLMs), which are trained on vast amounts of textual corpora with multiple training strategies along with a large number of parameters, have demonstrated remarkable capability of handling diverse tasks [7, 58, 46, 3]. A notable feature of these models is their ability to perform in-context learning, which means they can understand and learn from examples or instructions provided in the input and then adapt their responses based on this information, without requiring retraining for each specific task [7, 63, 43, 11, 21]. Due to its simplicity yet effectiveness and versatileness, several approaches have been introduced to improve the quality of the LLM context. In particular, Lyu et al. [42] constructs pseudo-demonstrations, for the case where examples in the context are unavailable, by retrieving relevant instances from the external corpus based on their similarities with the input query. Similarly, Ram et al. [49] and Baek et al. [4] augment LLMs by prepending relevant documents or facts retrieved from the external corpus in their input context, to improve the factuality of LLM responses. Lastly, Long et al. [40] targets adapting LLMs with in-context examples (which are adaptively retrieved) for domain adaptation. However, existing works do not focus on augmenting the data based on the retrieval of its relevant samples from other datasets, through in-context learning of LLMs.

**Data Augmentation** Despite the notable successes of LLMs, their performance significantly deteriorates in low-resource settings, particularly for domain-specific environments where the data available for training is very scarce (for instance, in the case of emerging events like novel viruses) or, in certain cases, completely unavailable (such as in privacy-sensitive enterprise contexts) [39, 10, 5]. Further, they are less likely to be trained with ones similar to these specialized data, leading to constrained capability in handling them. To address this challenge, numerous studies have proposed to expand the original seed data with various data augmentation techniques [19, 37]. Early works utilized token-level perturbation approaches, which either alter texts [54, 64] or interpolate them [9, 25]. Recent studies have shifted the focus towards utilizing the capability of generative language models, since they may internalize the useful knowledge to generate samples relevant to the seed data. Previous works on this line trained relatively smaller language models, based on the input-output pairs of the seed data to generate new outputs from the input variants [68, 2, 34]. Also, more recent works have used LLMs, which have much greater capability in generating high-quality data (sometimes surpassing human-level performances) without requiring task-specific training [16, 28, 65, 35]. Specifically, in information retrieval, some studies have generated synthetic queries with LLMs, to match the unlabeled documents with them [6, 14, 53]. Similarly, some other studies have proposed

LLM-powered methods for specific down-stream tasks, such as text classification [13, 55], reading comprehension [56], natural language understanding [18, 22] or multi-hop question answering [12]. This trend also goes to empowering the collection of instruction-tuning and alignment datasets for LLM training, which expands actual data samples with synthetic samples generated from LLMs themselves [29, 60, 61, 38, 17, 20]. However, in the low-resource setting, the seed data samples available to use for data augmentation are extremely scarce, which may result in suboptimal quality and limited diversity of the generated data. In this work, we propose to overcome this limitation by augmenting the data generation process with retrieval from larger external samples.

# D Limitations

We faithfully discuss some remaining room for improvements to our RADA framework. First of all, the effectiveness of our retrieval-augmentation approach (by its nature) depends on the quality and relevance of the external data store. Thus, the performance of RADA may degenerate if the retrieval source is not truly aligned with our seed data, and we leave exploring this new setting as future work. Also, investigating the scenario of continuously updating the retrieval pool over time would be interesting for future work as well. On the other hand, due to the heavy cost of fine-tuning LLMs, data sample efficiency (i.e., reducing the amount of samples to train while maintaining the model performance) becomes an important agenda. While we do have some preliminary results on filtering augmented samples in Appendix B, it would be interesting to developing more on this direction.

# E Broader Impacts

While RADA is superior in generating more diverse and high-quality samples (compared to existing data augmentation approaches), its performance is not flawless: the retriever might retrieve offensive or harmful instances for data augmentation, and the generator might produce plausible yet factually incorrect instances. Therefore, it may be carefully used for mission-critical domains, such as biomedical or legal fields, (perhaps with the help of domain-experts during the augmentation process).

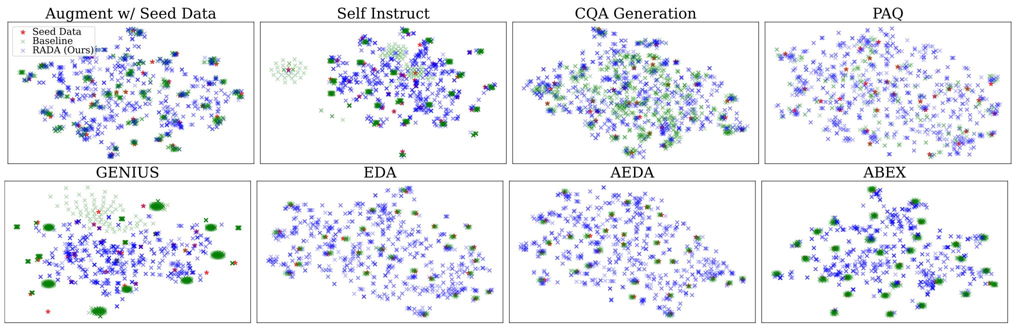

Figure 9: **Embedding-space visualization results** using t-SNE on Covid QA.

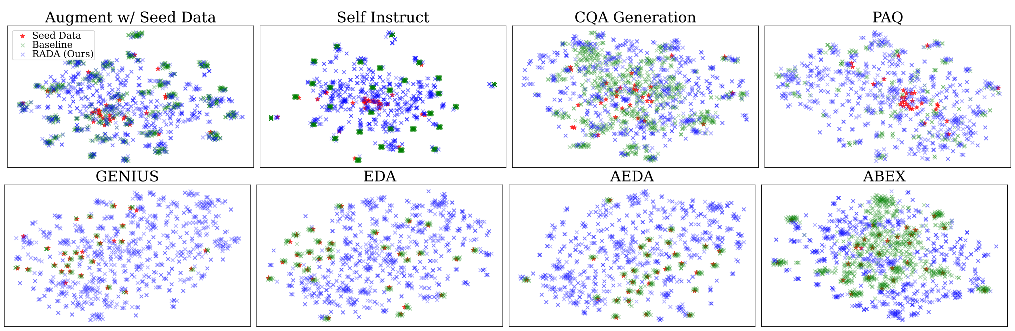

Figure 10: **Embedding-space visualization results** using t-SNE on Policy QA.

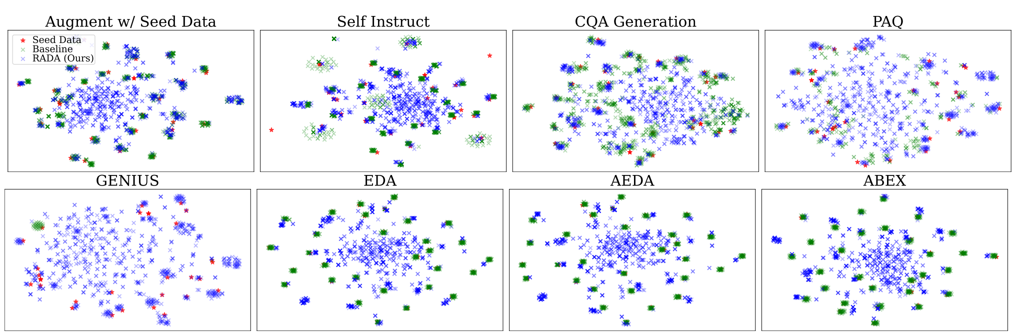

Figure 11: **Embedding-space visualization results** using t-SNE on Tech QA.

Table 12: A list of prompts that we use for data augmentation with the proposed RADA framework. It is worth noting that the variable inside the parentheses {} is replaced with its actual string (e.g., context, question, answer options, and answer). Also, the last sentence of the prompt represents the target context, which is used as the main source of information to generate the augmented instance. For MMLU, we use the combinations of Version 1 and Version 2 for data augmentation.

| Types | Prompts |
|---|---|
| **Domain-specific QA** | I want you to act as a question and answer generator. Your goal is to create an extractive question-answer pair based on a given context. The answer to the question must be a specific span from the given context.<br>Context: {context 1}<br>Question: {question 1}<br>Answer: {answer 1}<br>Context: {context 2}<br>Question: {question 2}<br>Answer: {answer 2}<br>Context: {context 3}<br>Question: {question 3}<br>Answer: {answer 3}<br>Context: {context} |
| **MMLU (Version 1)** | I want you to act as an answer options and answer generator. Your goal is to create four answer options and the answer pair based on a given question. The answer must be one of the generated answer options.<br>Question: {question 1}<br>Answer Options: {answer options 1}<br>Answer: {answer 1}<br>Question: {question 2}<br>Answer Options: {answer options 2}<br>Answer: {answer 2}<br>Question: {question 3}<br>Answer Options: {answer options 3}<br>Answer: {answer 3}<br>Question: {question} |
| **MMLU (Version 2)** | I want you to act as a question and answer generator. Your goal is to create an extractive question-answer pair based on the given answer options. The answer to the question must be selected from the given answer options.<br>Answer Options: {answer options 1}<br>Question: {question 1}<br>Answer: {answer 1}<br>Answer Options: {answer options 2}<br>Question: {question 2}<br>Answer: {answer 2}<br>Answer Options: {answer options 3}<br>Question: {question 3}<br>Answer: {answer 3}<br>Answer Options: {answer options} |

Table 13: The example question-answer pairs generated from different models on Covid QA.

| Types | Samples |
|---|---|
| **Augment w/ Seed Data** | **Context**: polymerase chain reaction testing, the time lag between hospitalization and reporting was longer for early cases compared with that of more recent cases. Among the seven locations reporting importation, the total volume of inbound passengers from China was m = 63.1 million per year in 2017 [9] , of which 100q = 2.1% were from Wuhan [10] , a home of n = 19.0 million people as the catchment population of Wuhan airport.<br>**Generated Question**: What was the total volume of inbound passengers from China to the seven locations reporting importation in 2017?<br>**Generated Answer**: 63.1 million per year |
| **Self-Instruct** | **Context**: polymerase chain reaction testing, the time lag between hospitalization and reporting was longer for early cases compared with that of more recent cases. Among the seven locations reporting importation, the total volume of inbound passengers from China was m = 63.1 million per year in 2017 [9] , of which 100q = 2.1% were from Wuhan [10] , a home of n = 19.0 million people as the catchment population of Wuhan airport.<br>**Generated Question**: What is the total volume of inbound passengers from China?<br>**Generated Answer**: m = 63.1 million per year in 2017 |
| **RADA (Ours)** | **Context**: MYTH: The flu is just a bad cold. Influenza may cause bad cold symptoms, like sore throat, runny nose, sneezing, hoarseness, and cough. But in the United States alone, 36,000 people die and more than 200,000 are hospitalized each year because of the flu. During the 2017/18 flu season, flu activity has significantly increased throughout the majority of the country with the A(H3N2) viruses predominating so far, according to the Centers for Disease Control (CDC).<br>**Generated Question**: What are the consequences of getting the flu?<br>**Generated Answer**: 36,000 people die and more than 200,000 are hospitalized each year because of the flu. |

Table 14: The example question-answer pairs generated from different models on Policy QA.

| Types | Samples |
|---|---|
| **Augment w/ Seed Data** | **Context**: IP Addresses. We use your IP address to help diagnose problems with our server, and to administer our Web site. Your IP address is used to help identify you and, in the case of the Coffee Review Marketplace, your shopping cart and to gather broad demographic information in the interest of improving our site and service.<br>**Generated Question**: How does the company use IP addresses?<br>**Generated Answer**: to help diagnose problems with our server, and to administer our Web site. |
| **Self-Instruct** | **Context**: IP Addresses. We use your IP address to help diagnose problems with our server, and to administer our Web site. Your IP address is used to help identify you and, in the case of the Coffee Review Marketplace, your shopping cart and to gather broad demographic information in the interest of improving our site and service.<br>**Generated Question**: What information does the company use IP addresses for?<br>**Generated Answer**: to help diagnose problems with our server, and to administer our Web site. Your IP address is used to help identify you and, in the case of the Coffee Review Marketplace, your shopping cart and to gather broad demographic information in the interest of improving our site and service. |
| **RADA (Ours)** | **Context**: The 1998 Data Protection Act sets the rules for how UK businesses can gather, store and use people confidential data. The Act is organised around eight principles of data protection. They include the idea that personal data should be obtained and processed fairly and lawfully, that only the right data should be used, that data should be accurate and so on. Principle Seven covers data security. It states:<br>**Generated Question**: What is the main purpose of the 1998 Data Protection Act?<br>**Generated Answer**: The 1998 Data Protection Act sets the rules for how UK businesses can gather, store and use people confidential data. |

Table 15: The example question-answer pairs generated from different models on Tech QA.

| Types | Samples |
|---|---|
| **Augment w/ Seed Data** | **Context**: VULNERABILITY DETAILS<br><br>CVEID: CVE-2015-1920 [http://cve.mitre.org/cgi-bin/cvename.cgi?name=CVE-2015-1920]<br>DESCRIPTION: WebSphere Application Server could allow a remote attacker to execute arbitrary code by connecting to a management port and executing a specific sequence of instructions.<br>CVSS Base Score: 9.3<br>CVSS Temporal Score: See https://exchange.xforce.ibmcloud.com/vulnerabilities/102404 [https://exchange.xforce.ibmcloud.com/vulnerabilities/102404] for the current score<br>CVSS Environmental Score*: Undefined<br>CVSS Vector: (AV:N/AC:M/Au:N/C:C/I:C/A:C)<br>**Generated Question**: What is the CVSS Base Score of the vulnerability in IBM WebSphere Application Server?<br>**Generated Answer**: The CVSS Base Score of the vulnerability in IBM WebSphere Application Server is 9.3. |
| **Self-Instruct** | **Context**: VULNERABILITY DETAILS<br><br>CVEID: CVE-2015-1920 [http://cve.mitre.org/cgi-bin/cvename.cgi?name=CVE-2015-1920]<br>DESCRIPTION: WebSphere Application Server could allow a remote attacker to execute arbitrary code by connecting to a management port and executing a specific sequence of instructions.<br>CVSS Base Score: 9.3<br>CVSS Temporal Score: See https://exchange.xforce.ibmcloud.com/vulnerabilities/102404 [https://exchange.xforce.ibmcloud.com/vulnerabilities/102404] for the current score<br>CVSS Environmental Score*: Undefined<br>CVSS Vector: (AV:N/AC:M/Au:N/C:C/I:C/A:C)<br>**Generated Question**: What is the CVSS score for the vulnerability in WebSphere Application Server?<br>**Generated Answer**: CVSS Base Score 9.3. |
| **RADA (Ours)** | **Context**: The Wii Homebrew Applications (such as The Homebrew Channel, Emulators, Games, Media players, etc.), USB Loader GX's NAND emulation feature and DIOS MIOS (Lite) require a FAT32 partition format. You can use a FAT32 SD/SDHC card for that purpose and set your USB hard drive as NTFS or Ext partition format.<br>**Generated Question**: What partition format is required for certain Wii Homebrew applications?<br>**Generated Answer**: FAT32 |

## NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: See Abstract and Introduction, as well as all other sections followed by them.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: See Section D.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

Justification: Our paper does not have any theoretical results.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: See Section 3 and Appendix A.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [No]

Justification: We will provide the data and code, with instructions to reproduce the results.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyper-parameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: See Appendix A.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: See Appendix B.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.

- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: See Appendix A.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We follow the NeurIPS Code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: See Appendix E.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to

generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.

- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: We do not release new data or models.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We cite all the assets and follow their terms.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [NA]

Justification: We do not release the new assets.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA]

Justification: We do not have crowdsourcing experiments and research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: We do not have human study and its participants.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.