← Back to **Author Console** (/group?id=aclweb.org/ACL/ARR/2024/February/Authors#your-submissions)

# TOP-Training: Target-Oriented Pretraining for Medical Extractive Question Answering

**(/pdf?id=jWlYUpK7-n)**

*Anonymous*

16 Feb 2024 ACL ARR 2024 February Blind Submission Readers: February, Paper2804 Senior Area Chairs, Paper2804 Area Chairs, Paper2804 Reviewers, Paper2804 Authors Show Revisions (/revisions?id=jWlYUpK7-n)

**Abstract:** This work studies extractive question answering in the medical domain (Medical EQA). This problem casts two main challenges: i) domain specificity; most AI models lack the necessary knowledge; ii) extraction-based answering style, which restricts most autoregressive LLMs due to the potential hallucination concern. To handle those challenges, this work proposes TOP-Training, a target-oriented pretraining paradigm that stands out among all domain adaptation techniques with two desirable features: i)TOP-Training moves one step further than popular domain-oriented fine-tuning since it not only moves closer to the target domain, but also familiarizes itself with the target dateset, ii) it does not assume the existence of large set of unlabeled instances towards the target domain. Specifically, for a target medical EQA dataset, we extract its entities and leverage large language models (LLMs) to generate synthetic texts containing those entities; pretraining on this synthetic text data is shown better performance on the target medical EQA benchmarks. Overall, our contributions are threefold: i) TOP-Training, a new pretraining technique to effectively adapt LLMs to better solve a target problem; ii) TOP-Training has a wide application scope because it doesn't require the target problem to have a large set of unlabeled data; iii) Our experiments highlight the limitations of autoregressive LLMs, emphasizing TOP-Training as a means to unlock the true potential of bidirectional LLMs.

**Paper Type:** long

**Research Area:** Question Answering

**Contribution Types:** Approaches to low-resource settings, Publicly available software and/or pre-trained models

**Languages Studied:** English

*Revealed to Saptarshi Sengupta, Connor Heaton, Shreya Ghosh, Wenpeng Yin, Preslav Nakov, Prasenjit Mitra*

15 Feb 2024 (modified: 16 Feb 2024) ACL ARR 2024 February Submission

**Authors:** *Saptarshi Sengupta (/profile?id=~Saptarshi_Sengupta1), Connor Heaton (/profile?id=~Connor_Heaton1), Shreya Ghosh (/profile?id=~Shreya_Ghosh3), Wenpeng Yin (/profile?id=~Wenpeng_Yin1), Preslav Nakov (/profile?id=~Preslav_Nakov2), Prasenjit Mitra (/profile?id=~Prasenjit_Mitra1)*

**TL;DR:** PreTraining on corpus from generative text models is helpful for in-domain pre-training for extractive question answering.

**Reassignment Request Action Editor:** This is not a resubmission

**Reassignment Request Reviewers:** This is not a resubmission

**Software:** zip (/attachment?id=1mfqFlKjr2j&name=software)

**Preprint:** no

**Preprint Status:** We are considering releasing a non-anonymous preprint in the next two months (i.e., during the reviewing process).

**Existing Preprints:** https://arxiv.org/abs/2310.16995 (https://arxiv.org/abs/2310.16995)

**Preferred Venue:** ACL 2024

**Consent To Share Data:** yes

**Consent To Review:** yes

**Consent To Share Submission Details:** On behalf of all authors, we agree to the terms above to share our submission details.

**A1:** yes

**A1 Elaboration For Yes Or No:** After section 7 (conclusion)

**A2:** yes

**A2 Elaboration For Yes Or No:** We discuss the ethical implications of our research after the limitations section.

**A3:** yes

**A3 Elaboration For Yes Or No:** Abstract & Introduction - Section 1

**B:** yes

**B1:** yes

**B1 Elaboration For Yes Or No:** We cite the required models & datasets wherever they were used.

**B2:** yes

**B2 Elaboration For Yes Or No:** All models used are openly available.

**B3:** yes

**B3 Elaboration For Yes Or No:** We explain the intended uses of the models and also explain how we used them creatively for our purposes.

**B4:** yes

**B4 Elaboration For Yes Or No:** Ethics Statement & Appendix E - We take measures to not reveal any identifying information in the generated corpus.

**B5:** yes

**B5 Elaboration For Yes Or No:** We describe the pre-training data of the foundational models that were used.

**B6:** yes

**B6 Elaboration For Yes Or No:** For the datasets on which we tested our models, we report the statistics in Section 5 (Experiments) & Tables 3,4 contain numeric details of our generated corpus.

**C:** yes

**C1:** yes

**C1 Elaboration For Yes Or No:** Limitations, Appendix D, Table 3 & 4

**C2:** yes

**C2 Elaboration For Yes Or No:** Appendix F

**C3:** yes

**C3 Elaboration For Yes Or No:** Yes, we provide mean and standard deviation scores from all our trials across the random seeds which were used.

**C4:** yes

**C4 Elaboration For Yes Or No:** Yes, all models & packages ex. Pytorch & HuggingFace were mentioned in the paper, Appendix H

**D:** no

**E:** no

**E1:** n/a

Add **Official Comment** **Author-Editors Confidential Comment** **Withdraw**

Reply Type: all Author: everybody Visible To: all readers **5 Replies**

Hidden From: nobody

[−]

## Official Review of Paper2804 by Reviewer 5Qxh

*ACL ARR 2024 February Paper2804 Reviewer 5Qxh*

27 Mar 2024 ACL ARR 2024 February Paper2804 Official Review Readers: Program Chairs, Paper2804 Senior Area Chairs, Paper2804 Area Chairs, Paper2804 Reviewers Submitted, Paper2804 Authors Show Revisions (/revisions?id=1LOxbwRfAGh)

**Recommended Process Of Reviewing:** I have read the instructions above

**Paper Summary:**

The author argues that decoder-only LLMs still struggle with some certain specialized language, in their case medical domain, and advocate the idea that encoder-only model is a better fit for closed-domain extraction-based QA tasks. The author proposes TOP-training: using an off-the-shelf LLM to generate target-domain-specific synthetic corpora, upon which an MLM model is fine-tuned to excel on the particular target-domain. The author conducts extensive baseline experiments on two benchmarks (COVID-QA and RadQA). They show that their proposed TOP-training leads to slightly better performance (0.5 absolute points on COVID-QA)

**Summary Of Strengths:**

The author conducts extensive baseline experiments on two benchmarks.

The proposed TOP-training leads to (slightly) better performance while eliminating the need of large amount of annotated data.

**Summary Of Weaknesses:**

The novelty of proposed approach is limited since the key idea is centered around **using an LLM to generate target-domain-specific synthetic corpora, upon which an MLM model is fine-tuned to excel on the particular target-domain**, which has been widely explored in the literature (i.a., [1]; see Comments, Suggestions And Typos).

The performance gain of TOP-Training is really limited (comparing table 1 and 3), which is just around 0.5 absolute points. Also, the author only showcased their proposed TOP-Training on top of two relatively weak LMs (BERT, RoBERTa), which significantly limits the impact of this work. If encoder is really vital, I would suggest experimenting with encoder-decoder models such as T5 as well, and you may choose to only use their encoder part for your study.

Section 6 (results and analysis) mainly lists results without providing in-depth analysis. Considering entity extraction and synthetic data generation are two major components in your design, it's advised to conduct more ablation study on these two aspects. For example, you may study how different LLMs' generated data would have different impact for the continual training, and how truthful/factual the generated data is.

The presentation needs to be improved and I spot out some typos, e.g., "dateset" in line 16, and line 167 doesn't read natural to me. The inconsistent usage of "§" makes it hard for me to read (line 079 and 083). Also, the caption of Figure 1 is not well-informed and should have included more sentences to explain modules presented in Figure 1 in a more detailed way. See more in Comments, Suggestions And Typos.

I don't see any discussion on whether the synthetic dataset or the training script would be released.

**Comments, Suggestions And Typos:**

What is FM in line 61?

What do you mean by saying "the former is more suitable due to its inherent capabilities" (line 078-079)

Why is line 240 split into two lines?

Line250-251 are hard to read.

Missing space between TOP-training and i.e., in line 289.

It's in my understanding that you are not doing model fine-tuning on top of the synthetic data but continual pretraining (line 296-298). Basically, I believe you are using MLM loss to train your models on the synthetic data with no EQA-specific supervision signals. Correct me if I am wrong here.

The following paper also studies domain adaptation and synthetic data generation for medical QA, and they follow the same idea as yours, i.e., generate synthetic data on target domain, which is used to fine-tune a downstream QA model. It's expected that you can discuss the difference and similarities in your work: CliniQG4QA: Generating Diverse Questions for Domain Adaptation of Clinical Question Answering

**Soundness:** 2 = Poor: Some of the main claims/arguments are not sufficiently supported. There are major technical/methodological problems.

**Overall Assessment:** 2 = Revisions Needed: This paper has some merit, but also significant flaws, and needs work before it would be of interest to the community.

**Confidence:** 4 = Quite sure. I tried to check the important points carefully. It's unlikely, though conceivable, that I missed something that should affect my ratings.

**Best Paper:** No

**Ethical Concerns:**
N/A

**Needs Ethics Review:** No

**Reproducibility:** 4 = They could mostly reproduce the results, but there may be some variation because of sample variance or minor variations in their interpretation of the protocol or method.

**Datasets:** 1 = No usable datasets submitted.

**Software:** 1 = No usable software released.

**Knowledge Of Or Educated Guess At Author Identity:** No

**Knowledge Of Paper:** N/A, I do not know anything about the paper from outside sources

**Knowledge Of Paper Source:** N/A, I do not know anything about the paper from outside sources

**Impact Of Knowledge Of Paper:** N/A, I do not know anything about the paper from outside sources

**Reviewer Certification:** 5Qxh

Add **Official Comment** **Author-Editors Confidential Comment**

[−]

## Official Review of Paper2804 by Reviewer im5p

*ACL ARR 2024 February Paper2804 Reviewer im5p*

26 Mar 2024 ACL ARR 2024 February Paper2804 Official Review Readers:
Program Chairs, Paper2804 Senior Area Chairs, Paper2804 Area Chairs, Paper2804 Reviewers Submitted, Paper2804 Authors Show Revisions (/revisions?id=gwDbnUxvCt)

**Recommended Process Of Reviewing:** I have read the instructions above

**Paper Summary:**
This paper introduced a pre-training approach to improving the performance of extractive question-answering (EQA) tasks in the medical domain in resource-scarce settings. They generated the synthetic dataset using the language model and continued the pre-training to introduce the target domain knowledge into the pre-trained language models.

**Summary Of Strengths:**
The proposed method outperforms the baseline models Extensive experiments and analysis on two benchmark datasets.

**Summary Of Weaknesses:**
Why the first stage of fine-tuning with SQuAD is performed? The authors did not discuss this. More importantly what was its impact on the final performance of the model? I did not see any results and analysis on this.

Why did the authors show the usefulness of the LLM-generated dataset on BERT and Roberta? Regarding this, no results were reported where only BERT and RoBERTa (without using the synthetic data) were used to compute the performance, like the results reported in Table 1 and Table 2.

Ideally, the authors should show the usefulness of the LLM-generated dataset on the best-performing PubMedBERT and LUKE for the COVID-QA dataset and, similarly, the best-performing PubMedBERT and CODER for the RadQA dataset. These results are missing in the paper.

Can the authors provide some details about the zero-shot experiments on Galactica, MedLLaMA, and MedAlpaca? The reported results are a bit skeptical.

More detail on the “Filtration - 50k corpus” setting is required. What did the authors mean by ill-formed entities here?

Since the datasets are generated from the language models, which are eventually used for training, it is quite possible that the generated dataset contains errors and incorrect facts in the generated abstract. How did the authors deal with this issue? Also, there needs to be a human evaluation of the model-generated dataset reported in the paper. It should be discussed in the paper.

Direct quantitative comparisons between Human-Generated Contexts using Wikipedia and model-generated contexts are not possible as the authors themselves state, “....entities available for this baseline are much smaller....” (lines 381-386). In this case, a fair comparison would be to report the results using the same subset of the dataset generated by models and for which the Wikipedia pages exist.

The paper does not explain why the authors use the Galactica model for corpus generation. In lines 329-333, they said they explored three decoder models: Galactica, MedLLaMA, and MedAlpaca.

There is no comparison or discussion with the existing approach on the COVID-QA and RedMAD datasets. Without providing a comparison to the existing approaches, how can the authors claim that their approach set SOAT on the COVID-QA dataset?

**Comments, Suggestions And Typos:**
The acronym FMs is not defined in the paper. In many places, texts are overlapping, e.g., lines 289, 291

**Soundness:** 2.5

**Overall Assessment:** 2.5

**Confidence:** 5 = Positive that my evaluation is correct. I read the paper very carefully and am familiar with related work.

**Best Paper:** No

**Ethical Concerns:**
None

**Needs Ethics Review:** No

**Reproducibility:** 3 = They could reproduce the results with some difficulty. The settings of parameters are underspecified or subjectively determined, and/or the training/evaluation data are not widely available.

**Datasets:** 1 = No usable datasets submitted.

**Software:** 3 = Potentially useful: Someone might find the new software useful for their work.

**Knowledge Of Or Educated Guess At Author Identity:** No

**Knowledge Of Paper:** N/A, I do not know anything about the paper from outside sources

**Knowledge Of Paper Source:** N/A, I do not know anything about the paper from outside sources

**Impact Of Knowledge Of Paper:** N/A, I do not know anything about the paper from outside sources

**Reviewer Certification:** im5p

Add **Official Comment** **Author-Editors Confidential Comment**

[−]
## Official Review of Paper2804 by Reviewer QBbA

*ACL ARR 2024 February Paper2804 Reviewer QBbA*

20 Mar 2024 ACL ARR 2024 February Paper2804 Official Review Readers:

Program Chairs, Paper2804 Senior Area Chairs, Paper2804 Area Chairs, Paper2804 Reviewers Submitted, Paper2804 Authors Show Revisions (/revisions?id=xYLQQPmMfSh)

**Recommended Process Of Reviewing:** I have read the instructions above

**Paper Summary:**

- This paper proposes a data augmentation method for extractive QA in the medical domain. They first extracted entities using spaCy and applied rule-based filtering, then used an LLM to generate synthetic texts usable for training a domain-specific model. The experimental results showed that some of their models outperformed existing BERT-like models.
- They conducted experiments based on the assumption that generative models (when without striding) do not handle long contexts for medical extractive QA tasks and showed that the EM scores were actually almost close to zero.

**Summary Of Strengths:**

- This paper comprehensively compares classifier-based models for medical extractive QA tasks. The authors also showed some performance gains when the training dataset was augmented using the output from generative models.

**Summary Of Weaknesses:**

1. To be honest, I do not find the considered superiority of their proposed approach using synthetic data, as it underperforms domain-adapted BERT models that require less computation cost (e.g., the best BERT-based baseline in Table 1 achieved 68.47, while their best BERT-based model in Table 3 achieved only 63.84).
2. The paper does not provide implementation details (e.g., filtering rules in Appendix A and regex in L473), making it hard for readers to reproduce. Furthermore, their choice of baseline does not necessarily allow concluding the superiority of their proposed method. Some examples are (1) regarding prompt design, "normal prompt" in L370 only feeds the words to generate pseudo-reports, but its inferiority is almost obvious as it is very different from the pretraining task of next token prediction (and same can be said for other trials in Appendix E), and (2) human-generated contexts (L371), whose lack of necessary knowledge is natural when humans only refer to Wikipedia, whereas the generative models are trained on other corpora.
3. Sec 6.1.1 only lists the results of various models and comments on each of them but does not provide any clear conclusion. No insights from these comparisons are considered in the proposed methods. This section should be reorganized, possibly in a separate paper, if it is not related to the proposed method.

**Comments, Suggestions And Typos:**

1. According to L220 and L278, Galactica also seems to lack the necessary knowledge. Please explain why the text generated by such a model helps the QA model acquire such knowledge. I feel like a BERT/RoBERTa baseline trained on a similar corpus to the Galactica model (i.e., L255-256) should be introduced if the richness of these corpora is the key.
2. L242 What does the train split of COVID-QA mean? (It is inconsistent with L301)
3. L245 Do these entities contain duplications? (I do not think 47k and 11k unique entities are in the corpus)
4. L478 Please explain why this performance gap comes from the writing style. (I suspect it came from the smaller number of training steps when the corpus is filtered)
5. It is better to explain why the proposed method is named "TOP" if it has any meaning.
6. Section 5.1 (Baselines and influencing factors) should be divided into two sections.
7. L016 dateset -> dataset
8. L061 What is an "FM"? It could be a fine-tuned model, but I do not think it is a common abbreviation.
9. L167 A period is missing after "span".
10. L265 Red texts should not be italicized when they are fixed string
11. Tables 1-4 Captions should be placed below tables. See paper formatting guidelines.

**Soundness:** 2 = Poor: Some of the main claims/arguments are not sufficiently supported. There are major technical/methodological problems.

**Overall Assessment:** 2 = Revisions Needed: This paper has some merit, but also significant flaws, and needs work before it would be of interest to the community.

**Confidence:** 3 = Pretty sure, but there's a chance I missed something. Although I have a good feel for this area in general, I did not carefully check the paper's details, e.g., the math or experimental design.

**Best Paper:** No

**Ethical Concerns:**
None

**Needs Ethics Review:** No

**Reproducibility:** 3 = They could reproduce the results with some difficulty. The settings of parameters are underspecified or subjectively determined, and/or the training/evaluation data are not widely available.

**Datasets:** 1 = No usable datasets submitted.

**Software:** 2 = Documentary: The new software will be useful to study or replicate the reported research, although for other purposes it may have limited interest or limited usability. (Still a positive rating)

**Knowledge Of Or Educated Guess At Author Identity:** No

**Knowledge Of Paper:** N/A, I do not know anything about the paper from outside sources

**Knowledge Of Paper Source:** N/A, I do not know anything about the paper from outside sources

**Impact Of Knowledge Of Paper:** N/A, I do not know anything about the paper from outside sources

**Reviewer Certification:** QBbA

Add **Official Comment** **Author-Editors Confidential Comment**

## [−] Official Review of Paper2804 by Reviewer Qpec

*ACL ARR 2024 February Paper2804 Reviewer Qpec*

19 Mar 2024 ACL ARR 2024 February Paper2804 Official Review Readers: Program Chairs, Paper2804 Senior Area Chairs, Paper2804 Area Chairs, Paper2804 Reviewers Submitted, Paper2804 Authors Show Revisions (/revisions?id=NUncjcYilg)

**Recommended Process Of Reviewing:** I have read the instructions above

**Paper Summary:**

The paper proposes a new pre-training framework, TOP-Training, that automatically extracts named entities from a target domain dataset (biomedical) and synthetically creates a pre-training dataset using these entities. The pre-training is followed by two additional rounds of fine-tuning, first on SQuAD and then on the target domain training dataset (COVID-QA and RadQA). The framework is evaluated against a series of baseline transformer models pre-trained on different general and biomedical domain datasets. The results are promising and show little improvement over the baselines on the COVID-QA dataset. On the RadQA dataset, the framework outperforms the models pre-trained on Wikipedia articles that were selected based on the extracted set of entities, however, did not surpass the best performance of other baseline transformer models.

**Summary Of Strengths:**

- The evaluation results show that the proposed framework is effective and can be used to improve performance in domain-specific contexts.
- The proposed framework can help reduce the pre-training costs by enabling similar performance gains (if not more) than that from pre-training on massive datasets.
- The described approach is simple and intuitive and seems easily applicable to specific domains such as biomedicine.

**Summary Of Weaknesses:**

1. The approach is shown to be effective as compared to the Wikipedia baseline for both datasets, however, it could not perform better as compared to other domain-specific models for RadQA. The paper is missing a discussion on the performance gains vs pre-training costs in such scenarios.
2. Since the focus of the paper is on the proposed pre-training framework, further exploration of the entity extraction and dataset generation step should be conducted. E.g., What are the similarity measures of generated data when compared to the actual datasets? What kinds of entities were extracted for each dataset?

Are certain kinds of entities more useful for pre-training data generation?

**Comments, Suggestions And Typos:**

1. Including some examples in the paper will greatly improve the understanding of the datasets.
2. Please define the "inherent capabilities" mentioned on line 79.

**Typos:**

There are multiple typos and convoluted sentences in the paper (e.g., on lines 16, 151-4, 289).

The abbreviations are not defined for FM, MLM, and NLP.

Line 392: “respectively” – the order of tables is reversed.

**Soundness:** 3 = Acceptable: This study provides sufficient support for its major claims/arguments. Some minor points may need extra support or details.

**Overall Assessment:** 4 = This paper represents solid work, and is of significant interest for the (broad or narrow) sub-communities that might build on it.

**Confidence:** 4 = Quite sure. I tried to check the important points carefully. It's unlikely, though conceivable, that I missed something that should affect my ratings.

**Best Paper:** No

**Ethical Concerns:**
None

**Needs Ethics Review:** No

**Reproducibility:** 4 = They could mostly reproduce the results, but there may be some variation because of sample variance or minor variations in their interpretation of the protocol or method.

**Datasets:** 1 = No usable datasets submitted.

**Software:** 1 = No usable software released.

**Knowledge Of Or Educated Guess At Author Identity:** No

**Knowledge Of Paper:** N/A, I do not know anything about the paper from outside sources

**Knowledge Of Paper Source:** N/A, I do not know anything about the paper from outside sources

**Impact Of Knowledge Of Paper:** N/A, I do not know anything about the paper from outside sources

**Reviewer Certification:** Qpec

Add **Official Comment** **Author-Editors Confidential Comment**

[−]

## Supplementary Materials by Program Chairs

*ACL ARR 2024 February Program Chairs*

16 Feb 2024 ACL ARR 2024 February Paper2804 Supplementary Materials Readers: Program Chairs, Paper2804 Reviewers, Paper2804 Authors, Paper2804 Area Chairs, Paper2804 Senior Area Chairs Show Revisions (/revisions?id=Cdy35TPECdD)

**Software:** zip (/attachment?id=Cdy35TPECdD&name=software)

**Reassignment Request Action Editor:** This is not a resubmission

**Reassignment Request Reviewers:** This is not a resubmission

**A1:** yes

**A1 Elaboration For Yes Or No:** After section 7 (conclusion)

**A2:** yes

**A2 Elaboration For Yes Or No:** We discuss the ethical implications of our research after the limitations section.

**A3:** yes

**A3 Elaboration For Yes Or No:** Abstract & Introduction - Section 1

**B:** yes

**B1:** yes

**B1 Elaboration For Yes Or No:** We cite the required models & datasets wherever they were used.

**B2:** yes

**B2 Elaboration For Yes Or No:** All models used are openly available.

**B3:** yes

**B3 Elaboration For Yes Or No:** We explain the intended uses of the models and also explain how we used them creatively for our purposes.

**B4:** yes

**B4 Elaboration For Yes Or No:** Ethics Statement & Appendix E - We take measures to not reveal any identifying information in the generated corpus.

**B5:** yes

**B5 Elaboration For Yes Or No:** We describe the pre-training data of the foundational models that were used.

**B6:** yes

**B6 Elaboration For Yes Or No:** For the datasets on which we tested our models, we report the statistics in Section 5 (Experiments) & Tables 3,4 contain numeric details of our generated corpus.

**C:** yes

**C1:** yes

**C1 Elaboration For Yes Or No:** Limitations, Appendix D, Table 3 & 4

**C2:** yes

**C2 Elaboration For Yes Or No:** Appendix F

**C3:** yes

**C3 Elaboration For Yes Or No:** Yes, we provide mean and standard deviation scores from all our trials across the random seeds which were used.

**C4:** yes

**C4 Elaboration For Yes Or No:** Yes, all models & packages ex. Pytorch & HuggingFace were mentioned in the paper, Appendix H

**D:** no

**E:** no

**E1:** n/a

**Note From EiCs:** These are the confidential supplementary materials of the submission. If you see no entries in this comment, this means there haven't been submitted any.

Add **Official Comment** **Author-Editors Confidential Comment**