Dear ARR Reviewer,

We write to kindly request you to consider the following where we address the issues raised in the previous round of reviews to this paper. Text in blue are the additions to the paper. We are also attaching a color-coded version of our paper. Text in blue are either new text or new sections.

Reviewer KnHk

1. We respectfully disagree. The innovation in the paper is the fact that we show that foundational models even those specifically trained in the domain do not work very well for some closed-domain question answering in the medical area. We then suggest how one can extract entities, use them to create prompts, generate synthetic text and retrain existing models in order to improve these models' performance. We demonstrate their efficacy and superiority on two closed-domain medical question-answering benchmarks, COVID-QA and RadQA. The method is not obvious and the experiments generate knowledge about our proposed method working better than any existing method. There lies the innovation.

   With respect to generality, the point is taken that it would be good if we can run these on other datasets. However, that argument can be made against any paper. We have a limited number of resources both with respect to human time and computational time. We also have limitations on the space we are allowed to write a paper. In this space and with these resources we have addressed the issues to the best of our ability.

   Is this the end of the story? No. More work is needed and our group is already working on extending this. Does this paper advance science? A resounding yes. We show that at least in these two domains using entity-based corpora generation works improve the state-of-the-art. We strongly believe the scientific community deserves to know this result. And, we are willing to make whatever specific edits, suggestions etc. your committee may have to improve the paper but a generic "not enough" as raised by the previous review is hard to address.

   We highlight our contributions at the end of the Introduction section,

   Line 111: In summary, our contributions are,

   Proposing a pipeline for generating customized pre-training data for closed (in our case medical) domains.

   Demonstrating the effectiveness of synthetic data for achieving sizable gains with reduced memory footprint.

   Showing the benefits of creative prompting and dataset awareness.

   Benchmarking numerous pre-trained biomedical FMs on COVID-QA and RadQA and contrasting our models with them by showing their superiority.

2. We have put in a dedicated Related Work section and addressed all known works, to the best of our knowledge, in the area directly related to our efforts.

   Line 124: 2 Related Work

   In this section, we discuss prior work on decoder-based FMs for QA, highlight their limitations and describe recent knowledge distillation pipelines.

3. Our paper does not require customized prompts. We can use the same basic template where we use the entity name to instruct a FM to generate synthetic texts related to that prompt. If we customize the prompt, we may get slightly better results. I asked practitioners whether they would be interested in manually creating a 2-3 line prompt template for the fancy template and

the work is so minimal that all practitioners were ready to do so. We can also automate this process with a heuristic solution. This should not be an impediment in the real-world to using our work nor would it take that much manual effort to set up a "fancy" prompt template (as defined in our paper).

4. We regret the unprofessional writing in the last draft. We have removed colloquial language, archaic phrasing, unnecessary unprofessional abbreviations and worked to the best of our ability given the time we had. We will do another full pass with a native speaker to edit the paper for the final camera-ready version should this paper be accepted.

5. We have tried to meticulously check every claim made in the paper in every line and either give citations or tone down the claims to make it clear that that is a probable conjecture. We still keep the conjecture in the hope that the intuition is useful for a follow-up work to verify them more rigorously. But, we agreed with the previous review and marked our conjectures more carefully.

6. We have tried to explain the methods and settings as much as we can given the space provided. We include the rest in the appendix due to space limitations. We will have a full-version of a technical report where we will report all parameters, etc. carefully to enable full reproduction.

Reviewer cmuw

1. Novelty – we claim that no prior work has done entity-based synthetic corpora generation and shown that it improves domain-specific foundation models on our closed-domain benchmarks. We would welcome counterexamples that disprove our claim; our best literature review did not find a paper that does that.
2. We have added real-life examples in the whole presentation.

   Line 932: Appendix B – We refer readers to Figure 6 (top) to see an example of such kind of report.

3. We show improvements even in a non data-leak setting at the end. Note here that a "data-leak setting" is nuanced here. For example, if I have a student who does badly in a test I give, I can give the student a set of entities to read up on and then give another test. In this case, although I have given the student the list of entities to read up on, the second test is not really a seen test.

   In fact, real practitioners may be fine giving us a set of diseases and entities of interest on which they want to ask questions and then we can generate our synthetic documents based on them and prepare for the unseen exam. Then, the practitioners can use the fine-tuned model to ask questions they seek to ask.

   Nevertheless, as we should in our cross-validation experiments on COVID-QA, our method still performs better. Also, note that in RadQA, there is no data leak. So, the claim that our method works only in data-leak settings is flat out wrong.

   Reviewer: rTgr

   1. This may be an interesting solution but we do not claim that we have tried out all possible solutions. It would be another project or part of a new project. We do not know if that would work. Our claim is that our method of augmenting FMs work on our datasets that

we establish by our experiments. We have tried to make the claims we make in the paper more tight and accurate though.

2. We have tried to address the presentation issues.