

# Quality > Quantity: Synthetic Corpora from Foundation Models for Closed-Domain Extractive Question Answering

Anonymous ACL submission

## Abstract

Domain adaptation, the process of training a model in one domain and applying it to another, has been extensively explored in machine learning. While training a domain-specific foundation model (FM) from scratch is an option, recent methods have focused on *adapting* pre-trained FMs for domain-specific tasks. However, our experiments reveal that either approach does not consistently achieve state-of-the-art (SOTA) results in the target domain. In this work, we study extractive question answering (EQA) in closed domains and introduce the concept of *targeted pre-training*. This involves determining and generating relevant data for further pre-training, as opposed to the conventional philosophy of using domain-specific FMs trained on an overarching collection of domain data thus avoiding concepts orthogonal to the downstream dataset and focusing only on relevant ones. Our proposed framework uses Galactica to *generate* synthetic, “targeted” corpora that align with specific writing styles and topics, such as *research papers* and *radiology reports*. This process can be viewed as a form of *knowledge distillation*. We apply our method to two biomedical EQA datasets, COVID-QA and RadQA, achieving a new benchmark on the former and demonstrating overall improvements on the latter.

## 1 Introduction

Our work revolves around three key pillars: **EQA**, **Domain Adaptation**, and **knowledge distillation** through *prompting* generative foundation models (FMs). EQA, a long-standing problem in natural language processing (NLP) involves identifying a token span in a text passage to answer a given question. The task is typically evaluated using datasets like SQuAD (Rajpurkar et al., 2016) and DuoRC (Saha et al., 2018). While recent architectures like BERT (Devlin et al., 2019), T5 (Raffel et al., 2020), and GPT-3 (Brown et al., 2020) have made remarkable advancements in this task, their performance

suffers when applied to domain-specific data, especially in the medical domain (Moradi et al., 2021).

The performance discrepancy in models is linked to the definition of a *domain*, i.e., the loose NLP equivalency of *domain* = *genre* or *thematic content* of a dataset. This definition is quite restrictive (Plank, 2016). Ideally, a model pre-trained on a specific theme should excel in tasks related to that subject matter. However, not all domain-specific models are equal as illustrated by the differing performances of BioBERT (Lee et al., 2020) and PubMedBERT (Gu et al., 2021), even though both trained on PubMed data. We suggest redefining *domain* = [*genre* + *dataset*], emphasizing the importance of tailoring the training data to the subject matter of the task. This approach acknowledges that a *one-domain-model-to-rule-them-all* is not universally applicable, and the learning should focus on concepts relevant to specific tasks. We define “closed-domains” as *datasets* related to highly specialized subjects like medicine, law, or finance.

The third pillar supporting our work is the recent progress in generative FMs (Ye et al., 2023; OpenAI, 2023). While ChatGPT performs well on the USMLE (Kung et al., 2023), our experiments demonstrate that large, general-domain (and even closed) FMs struggle with tasks involving highly specialized language, such as COVID-QA (Möller et al., 2020) and RadQA (Soni et al., 2022). Additionally, their autoregressive architecture is not well-suited for EQA as they are designed to *synthesize* new text rather than *extract spans* from given text (c.f. sec. 3.1). Also, when presented with sequences exceeding the model’s context length, they need to be divided into overlapping segments. Although this challenge applies to both bi-directional and generative models, bi-directional models are more suitable due to their inherent capabilities. While generative models can generate an answer for each segment, it lacks the ability to indicate their confidence in each answer, a feature provided

by bi-directional models.

To overcome these limitations, we propose distilling the knowledge from causal FMs into smaller MLM-style models better suited for EQA. Our approach involves using a causal FM **to generate a synthetic corpus** tailored to a specific application and fine-tuning open-domain MLM models on this corpus. We **adapt** open-domain models instead of using domain-centric ones to emphasize the resource benefits from our pipeline i.e. without relying on gigabytes of domain-specific pre-training data, we can achieve respectable gains by honing our models on modest-sized corpora. Our results demonstrate the efficacy and running time improvements w.r.t existing domain-specific models.

Seminal to knowledge distillation, West et al. (2022) demonstrated how GPT3 could be utilized to create high-quality knowledge graphs via prompting. He et al. (2022) show how a GPT model could be used as a “teacher” to distil knowledge into a “student.” Similarly, Peris et al. (2022) used unlabelled task-relevant data and trained multilingual students with varying proportions of general/task-specific data and report the most gains using “only the downstream task’s unlabelled data”.

Gururangan et al. (2020) introduced DAPT (Domain-Adaptive Pretraining) and TAPT (Task-Adaptive Pretraining), which share similarities with our work. DAPT involves extended pretraining on domain-specific corpora without labels, while TAPT focuses on pretraining on the unlabelled training set of the downstream task. Although they demonstrate the effectiveness of TAPT compared to DAPT, closed-domain datasets like COVID-QA typically lack a separate unlabelled training set and may not even have train/dev/test splits. Further, DAPT considers knowledge beyond what is necessary to the task data, whereas our approach confines training to only required concepts.

In summary, **our contributions** are (a) proposing a pipeline for generating customized pre-training data for closed domains, (b) demonstrating the effectiveness of synthetic data in achieving respectable gains with reduced memory footprint, (c) showcasing the benefits of creative prompting and dataset awareness, (d) setting a new benchmark on COVID-QA & overall improvements on RadQA.

## 2 Methodology

In Figure 1, we present our method and compare it to existing pre-training paradigms. The current ap-

proaches involve training a randomly initialized architecture from scratch (top) on either open-domain data (e.g., BERT/RoBERTa (Liu et al., 2019)) or closed-domain data (e.g., SciBERT (Beltagy et al., 2019), PubMedBERT (Gu et al., 2021)), or extended pre-training (middle), where the model is initially trained on open-domain data and then further pre-trained on unlabelled domain-specific text (e.g., BioBERT) to adapt it to the closed-domain. The former emphasizes stronger domain representations, while the latter prioritizes computational efficiency, by not requiring the model to relearn general token representations. After training, few rounds of fine-tuning are done to learn the task-specific head and to be adapted to the final dataset.

While the above techniques have achieved much success, they typically rely on high quantities of unlabelled corpora to yield useful results, thus raising the question: *What happens when we do not have enough “relevant” domain data, either in style or volume?* Inspired by the works of Gunasekar et al. (2023) and Zhou et al. (2023), who hearken the use of smaller yet *better quality* corpora, we introduce the notion of **targeted pre-training**, which focuses on a specific subset of the domain, tailor-made for the ultimate downstream dataset.

Our method works as follows. First, we **combine all the questions and contexts** from the training split of the EQA dataset. Unfortunately, COVID-QA does not have a train-dev-test split. In such a situation, we consider the entire dataset for the next step (we test for cheating/information leakage in this case as described in sec. 4.3). Next, we **extract entities** through Named Entity Recognition (NER) using scispaCy (Neumann et al., 2019). Comparing the *small* and *large* versions of the NER models, we found the former (en\_core\_sci\_sm) yields qualitatively better & quantitatively more, entities.

Next, we **create prompts** for the identified entities to generate contexts mimicking the respective datasets. This required studying the characteristics of the datasets such as the style of contexts (full research articles in COVID-QA & radiology reports in RadQA), their lengths and relevant keywords. The collection of prompts were then supplied to Galactica (Taylor et al., 2022), to **generate the corpora** ( $\cup$  generated contexts) for pre-training.

Galactica is a decoder-based FM pre-trained on a collection of text encompassing research articles, knowledge bases, code and even L<sup>A</sup>T<sub>E</sub>X markup. Interestingly, Galactica is able to generate research papers by being prompted as “**Title:**

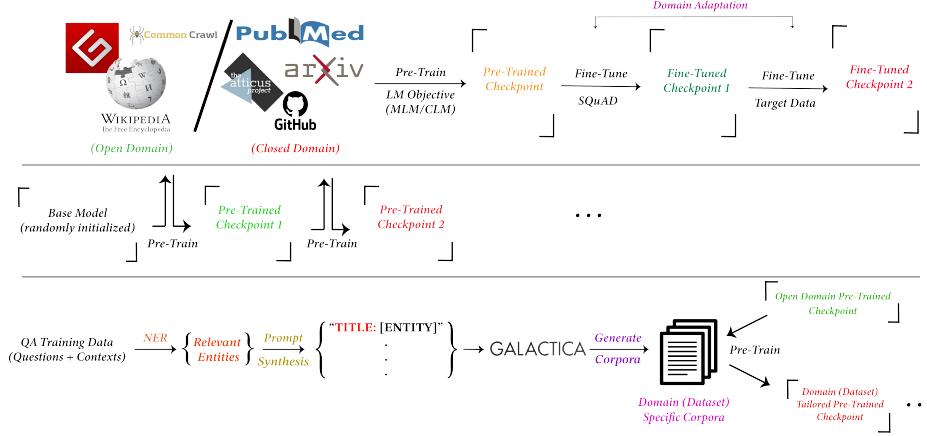


Figure 1: Pre-Training Pathways: From scratch (top); Extended (middle); **Targeted** (bottom; ours) | Note: We only show Fine-Tuning on EQA as it is the task of interest | The prompt handle is written in CAPITAL for emphasis.

[*entity*]]” (where **Title:** is the prompt handle/keyword and *entity* is the entity for which we require generated content). We considered other generative models such as BLOOM (Scao et al., 2022) and PubMedGPT. However, due to multilingual (BLOOM) or qualitatively inferior (both) generations to Galactica, they were not chosen.

The choice of prompt for COVID-QA is straightforward (as above) seeing as its contexts are research articles themselves. RadQA, on the other hand, presents a bigger challenge. Its contexts are redacted radiology reports without any consistent format (Hartung et al., 2020). This is challenging since we do not have a template to synthesize prompts for. However, after going through the samples in the dataset, we realized that the Findings and Impressions section are the most vital in a patient’s report (akin to the experiment and results section in a research paper). Such clues led us to construct our prompt for RadQA as, “Patient has [*entity*]. FINDINGS AND IMPRESSION”. This was very interesting for us since Galactica had never seen radiology reports during training and we found a way to get it to synthesize *pseudo-reports* in this manner bypassing any privacy concern. We specifically wrote our prompt in this way so as to acquire text for both sections in a single go (for computational efficiency) and, to avoid *chain-of-thought-reasoning* (Wei et al., 2022) since we were using the base variant of Galactica (1.3B) which according to Wei et al. (2022), would not be able to keep track of logic seeing as it’s  $\ll \sim 100B$  parameters. After generating contexts, we perform targeted pre-training i.e., extended pre-training of BERT/roBERTa on our generated corpus followed

by two rounds of fine-tuning (SQuAD  $\rightarrow$  COVID-QA/RadQA).

A natural question to ask is why we *generated* a corpus instead of using existing text. We do this for 3 reasons, (a) *flexibility* to create content of a certain style (b) some corpora can be *unavailable* for privacy reasons, ex. clinical diagnostic reports or “internal” versions of corpora such as that used by Gururangan et al. (2020) and (c) our tests can be used to determine if the content produced by such FMs is *factually grounded* & able to teach student models specific writing styles.

### 3 Experiments

Our study focuses on two datasets: COVID-QA, comprising 2,019 answerable QA pairs (no train/dev/test splits) sourced from CORD-19 (Wang et al., 2020), and RadQA, consisting of 6,148 QA pairs from radiology reports, with a train/dev/test split of 4,878/656/614. We conduct experiments in two areas: benchmarking and targeted pre-training and provide mean and standard deviation scores over three random seeds for each.

#### 3.1 Benchmarking

We identify thirteen encoder models in total to apply to each dataset taking into account which model made most sense to apply to a dataset. The models applied to COVID-QA were checkpoints fine-tuned on SQuAD v1 while RadQA, containing unanswerable questions, was tackled with those fine-tuned on SQuAD v2 (Rajpurkar et al., 2018). For consistency, we utilized the *cased, base* version of each architecture when available. Models applied to COVID-QA were fine-tuned using five-fold cross-validation and the resulting average performance



across folds is presented in Table 1. Results of models applied to the prescribed splits are presented in Table 3. The metrics used are exact match (EM), binary measure of whether the prediction & gold spans are identical & F1, the harmonic mean of the number of shared words in the two spans w.r.t number of words in the prediction (precision) and w.r.t number of words in the gold span (recall).

To assess the **zero-shot** performance of three decoder models, namely Galactica-base (1.3B), MedLLaMA, and MedAlpaca (both 13B), we measure their ability to generate answers without further fine-tuning on our datasets, considering that decoders do not extract spans, but generate answers for comparison to gold spans. We selected Galactica-1.3B for consistency with our corpus generation experiments, MedLLaMA as a strong open-source medical checkpoint, and MedAlpaca as a medical QA-specific LLaMA checkpoint. Following Yue et al. (2021), each sample is formatted as `Question:<question_text>` `Context:<part_of_context>` `Answer:.` The text generated after `Answer:` is considered as the predicted span. Due to the large size of COVID-QA contexts, they were segmented as they exceeded the maximum sequence length of each model (2,048 tokens). We report overall EM/F1 on each dataset and average best EM/F1 (parenthesis in Table 1) from each Q+C+A chunk for COVID-QA (N/A for RadQA since the context size was << models maximum input length).

## 3.2 Targeted Pre-training

Targeted pre-training begins by identifying entities in the final datasets. scispaCy (en\_core\_sci\_sm) identifies roughly 47k and 11k entities in COVID-QA and RadQA, resp. Next, Galactica is used to generate contexts for the identified entities, constituting the synthetic dataset for targeted pre-training. To maintain size-parity, five contexts are generated for each entity identified in RadQA, yielding around 55k (11k\*5) total contexts. Galactica is allowed to use its full context size of 2,048 tokens to generate the synthetic data for each entity.

### 3.2.1 Corpus Size

When training models for COVID-QA, we investigated the impact of synthetic dataset size on downstream performance. We examined the effects of generating one context per entity and also explored generating ten contexts per entity, resulting in a

dataset that was 10 times larger than the baseline. This analysis allowed us to assess the scalability of the proposed approach.

### 3.2.2 Context Length

The average context length for COVID-QA is 6k tokens, and Galactica has a maximum context size of 2k, resulting in a misalignment between the synthetic corpus and the target dataset. Increasing the context size of Galactica would mean training it from scratch with architectural changes which is infeasible. Thus, we explore the impact of sequence length in the synthetic corpus by limiting the records to only 1k tokens. While we cannot determine if *longer* sequences are *beneficial*, we can evaluate if *shorter* ones are *detrimental*.

### 3.2.3 Token Filtering

We performed entity filtering as a common ablation technique for both datasets. We used regular expressions to remove entities with special characters such as \*, !, etc., as well as specific text patterns like `https*` and `baby`. We implemented a length-based filter, retaining only entities longer than a certain number of characters. Additionally, for COVID-QA, we applied a second round of filtration using TF-IDF, considering the questions + context as the corpus and retaining the top 25k entities with the highest IDF scores. However, as this approach did not lead to substantial gains, we decided not to use it for RadQA. Due to the large number of possible combinations, we did not extensively explore these settings in our experiments.

### 3.2.4 Prompting Style

We explore the use of two different prompts when encouraging Galactica to generate *pseudo* radiology reports - “**Patient has [entity]. FINDINGS AND IMPRESSION**” (referred to as *fancy prompt*), and simply “[entity]” (referred to as *normal prompt*). We had to experiment with prompt styles since Galactica was **not** pretrained on radiology reports & as such, the ideal prompt is not immediately apparent.

### 3.2.5 Human-Generated Contexts

We establish a *Wikipedia* baseline alongside our domain-specific models to assess the influence of content and text structure during domain adaptation. Wikipedia was chosen as it has been extensively used in the training data of varied models offering reliable information. Additionally, as reported by Micallef et al. (2022), mBERTu Wiki, which

was pre-trained on Maltese Wikipedia data, surpassed the performance of mBERT, thereby proving to be a competitive baseline. Further, we see that Wikipedia is often used (Borkakoty and Espinosa Anke, 2023; Jurczyk et al., 2018) as it provides a common ground for comparison across different models and studies.

For each entity, we query Wikipedia and retrieve the complete page associated with the top search result. The number of entities available for this baseline is  $\ll$  than our approach since most of them do not exist in Wikipedia due to either being extremely esoteric, e.g., `pulmonary parenchymal infiltrate` or improperly formed, e.g., `Bao &`.

## 4 Discussion

Here we discuss the results of benchmarking existing models (Tables 1 & 3) as well as results for our targeted pre-training (Tables 2 & 4). The last 3 rows of Tables 1 & 3 provide zero-shot performance of our chosen decoder models on RadQA & COVID-QA resp. We do not perform multiple trials here as the performance was seen to be extremely poor and would not benefit from additional runs. Granted they were not fine-tuned on our datasets, their size, pre-training data coverage & reported zero-shot performance on related datasets, should have allowed them to at least perform on par or better than open-domain BERT/RobERTa. **Overall, we see that MedAlpaca seems to be the “best” among the three for RadQA and only marginally poorer in terms of F1 for COVID-QA.** This makes sense since MedAlpaca is an instruction-tuned version of MedLLaMA. In terms of EM (for COVID-QA), none of the models generated text in line with the gold standard ( $\sim 0$  EM) & only showed positive F1.

### 4.1 COVID-QA

In this subsection, we present the benchmarking trials and contrast them with the proposed method, highlighting the latter’s benefits, as observed.

#### 4.1.1 Benchmarking

**Baselining our chosen models demonstrates that a one-size-fits-all approach does not work for domain adaptation.** BioBERT and PubMedBERT were trained on similar corpora and yet score in the same range, indicating no clear winner. That said, PubMedBERT does perform better as it is trained from scratch using a custom vocabulary covering a range of medical jargon.

Surprisingly, the SciBERT (+CORD-19) checkpoint, trained on CORD-19 articles, performs worse than regular SciBERT, suggesting potential issues in training choices or noisiness in the data. **Notably, LUKE, trained solely on Wikipedia data, emerges as the best baseline model**, possibly due to its entity-recognition pre-training objective, which aids in identifying relevant entities for QA tasks (Van Aken et al., 2019) and highlighting the need for entity representations in closed-domains. Models capable of handling longer context i.e. Longformer, BigBird and XLNet do not show marked improvements. In fact, XLNet, degrades completely on COVID-QA, potentially due to token permutation hindering its reasoning across large and conceptually dense contexts.

#### 4.1.2 Targeted Pre-Training

→ **Wiki Baseline:** Fine-Tuning BERT on Wikipedia yields marginal improvement of 0.9% EM and 1.2% F1, while RoBERTa shows a 3.7% increase in EM and a 1.3% increase in F1. This makes sense as we hypothesize that our Wikipedia corpus, while small, contains relevant information pertaining to COVID literature which in turn aids in answering related questions.

→ **47k corpus:** With targeted pre-training both models demonstrate even further improvements. BERT achieves a 4.01% increase in EM and a 3.5% increase in F1, while RoBERTa shows a 6.7% increase in EM and a 2.5% increase in F1, setting a new SOTA on COVID-QA. Remarkably, RoBERTa even outperforms the previous SOTA model (LUKE) by 1.2% in EM and 1.3% in F1, despite using a training corpus significantly smaller (67.4 MB/0.032B words) than LUKE’s 3.5B-word corpus (0.9% of the size). Moreover, we find that **any variation of our approach, ablation or otherwise, yields better performance for both models over the Wikipedia baseline. This confirms our hypothesis that having the right content alone is insufficient without proper structure/style.**

→ **470k [10x] corpus:** Training with a 10x corpus (10 contexts per entity) led to the most improvements for BERT with EM increasing by 8.2% and F1 by 6.4%. This is consistent with Liu et al. (2019) who argue that BERT was *significantly undertrained*. Though this improvement does not achieve RoBERTa-esque performance, it does **demonstrates the scalability of our approach**. Additionally, while RoBERTa does improve, its gains are not as great as using the base

Table 1: Benchmarking Bio Models (COVID-QA). <sup>1</sup>(Peng et al., 2019); <sup>2</sup>(Yuan et al., 2022); <sup>3</sup>(Beltagy et al., 2020); <sup>4</sup>(Zaheer et al., 2020); <sup>5</sup>(Yamada et al., 2020); **Blue = best/red = worst** scores overall; **bold = best decoder**

Model	Pre-Training Corpus	Corpus Size	EM	F1
BioBERT	PubMed	4.5B words	37.62 ± 0.18	65.73 ± 0.35
SciBERT	Semantic Scholar	3.2B words	37.52 ± 0.23	65.58 ± 0.18
SciBERT(+CORD-19)	Semantic Scholar + CORD-19	3.2B words + 20GB	35.61 ± 0.30	63.60 ± 0.59
PubMedBERT	PubMed	3.1B words / 21GB	39.87 ± 0.74	<b>68.47 ± 0.13</b>
BlueBERT <sup>1</sup>	PubMed + MIMIC	4.5B words	27.35 ± 0.30	52.18 ± 0.40
CODER <sup>2</sup>	Unified Medical Language System	NA	39.33 ± 0.47	67.01 ± 0.30
Longformer <sup>3</sup>	Books + Wiki + RealNews + Stories	6.5B tokens	37.79 ± 0.39	66.58 ± 0.22
BigBird <sup>4</sup>	Books + CC-News + Stories + Wiki	160GB (same as RoBERTa)	32.79 ± 0.13	60.06 ± 0.41
LUKE <sup>5</sup>	Wikipedia	3.5B words	<b>41.01 ± 0.30</b>	68.23 ± 0.19
XLNET	BooksCorpus + Wikipedia + Giga5 + ClueWeb 2012-B + Common Crawl	32.89B words	2.45 ± 0.08	8.64 ± 0.19
Galactica	c.f. section 2	106B tokens	0 (0)	5.01 (11.11)
MedLLaMA	Medical Corpora	NA	0 (0)	<b>5.81 (12.79)</b>
MedAlpaca	Medical Meadow	NA	<b>0.03 (0.2)</b>	5.21 (12.73)

Table 2: Targeted Pre-training (COVID-QA). Time#: to generate corpus; ♣: entity filter; Gal = Galactica; max\_length = Context Max Length. **Blue = best/red = worst** scores overall; **bold = best BERT/RoBERTa setup**.

Model	Train Dataset	Time#	Corpus Size	EM	F1
BERT	NA [Vanilla Fine-Tuning]	NA	NA	33.62 ± 0.59	60.01 ± 0.36
RoBERTa	NA [Vanilla Fine-Tuning]	NA	NA	38.89 ± 0.52	67.44 ± 0.47
BERT	Wikipedia	≈2.5 hrs	139.6 MB	33.95 ± 0.13	60.76 ± 0.78
RoBERTa	Wikipedia	≈2.5 hrs	139.6 MB	40.33 ± 0.60	68.30 ± 0.54
BERT	Gal(47k)	≈ 6.5 hrs	67.4 MB	34.97 ± 0.18	62.11 ± 0.32
RoBERTa	Gal(47k)	≈ 6.5 hrs	67.4 MB	<b>41.51 ± 0.48</b>	<b>69.10 ± 0.27</b>
BERT	Gal(470k) [10x]	≈ 2.5 days	558.2 MB	<b>36.39 ± 0.27</b>	<b>63.84 ± 1.16</b>
RoBERTa	Gal(470k) [10x]	≈ 2.5 days	558.2 MB	41.31 ± 0.22	68.84 ± 0.28
BERT	Gal(25k*2 = 50k)♣	≈ 6.5 hrs	64.0 MB	35.03 ± 0.38	62.14 ± 0.48
RoBERTa	Gal(25k*2 = 50k)♣	≈ 6.5 hrs	64.0 MB	41.36 ± 0.35	69.00 ± 0.52
BERT	Gal(47k) [max_length = 1k] ≈ 2.5 hrs	≈ 2.5 hrs	44.8 MB	34.90 ± 0.14	62.02 ± 0.95
RoBERTa	Gal(47k) [max_length = 1k] ≈ 2.5 hrs	≈ 2.5 hrs	44.8 MB	<b>41.57 ± 0.33</b>	68.98 ± 0.31

47k corpus. We attribute this to the *undertraining* of BERT making it more *malleable* to learn new concepts as opposed to RoBERTa which we reason, from our trials, to have hit its *ceiling for learning* in this domain i.e. it gathered everything it needs to know from the 47k corpus and does not benefit from additional data.

→ **Filtration - 50k corpus:** We expected that removing ill-formed entities would improve the results, the fifth row of Table 2 shows that performance actually declined w.r.t the best BERT (10x corpus - row 4) and best RoBERTa (base 47k corpus - row 3) model. We reason that our regex-based filtering rules mistakenly removed entities relevant to research articles such as author names or URLs, leading to the decline in performance.

→ **Reduced context length - 47k (at most 1k context tokens) corpus:** When we decrease the context length due to limitations in Galactica’s token generation i.e. at most 2048 tokens (last row of Table 2), we again observe a **decline in performance** for both metrics and both models (over their best versions). This outcome is expected as Galactica is unable to generate content matching the style

of COVID-QA’s contexts (research papers), under-scoring the importance of domain aware writing styles for adaptation pipelines.

## 4.2 RadQA

Results on RadQA are presented for both its validation and test splits. We analyze results by considering various combinations of contexts (prompts) and entity filtration. Overall, higher test scores on average are observed as compared to validation scores, which we attribute to fewer unanswerable questions in the test set (154 vs. 231) and slightly shorter contexts (73.82 vs. 78.1 tokens). We check for information leakage and find no irregularities. Thus, while we report scores for both, our analysis mainly focuses on the validation set, serving as the first point of evaluation in the RadQA domain.

### 4.2.1 Benchmarking

**RadQA benchmarks were a bit less unanimous.** On the dev set, **CODER has the best overall EM & F1 but suffers slightly v/s PubMedBERT on only answerable questions.** This is not surprising as CODER is a PubMedBERT checkpoint

Table 3: Benchmarking Bio Models (RadQA). H(F1): HasAns\_F1, H(EM): HasAns\_EM; **Blue = best/red = worst** scores overall; **bold = best decoder**; <sup>1</sup>(Yan et al., 2022); <sup>2</sup>(Alsentzer et al., 2019); <sup>3</sup>(Gururangan et al., 2020)

Model	Corpus	Corpus Size	Dev				Test			
			EM	F1	H(EM)	H(F1)	EM	F1	H(EM)	H(F1)
BioBERT	PubMed	4.5B words	26.42 ± 0.49	44.26 ± 0.09	40.79 ± 0.76	68.31 ± 0.14	49.95 ± 1.08	63.32 ± 0.40	45.65 ± 1.21	63.50 ± 0.57
SciBERT	Semantic Scholar	3.2B words	27.03 ± 0.32	44.40 ± 0.06	41.65 ± 0.62	68.45 ± 0.22	53.04 ± 0.38	67.17 ± 0.73	48.62 ± 0.70	67.49 ± 1.02
PubMedBERT	PubMed	3.1B words / 21GB	31.45 ± 0.17	47.89 ± 0.46	48.40 ± 0.27	73.77 ± 0.62	54.07 ± 0.71	68.76 ± 0.22	49.49 ± 0.87	69.09 ± 0.80
BlueBERT	PubMed + MIMIC	4.5B words	30.08 ± 1.33	47.14 ± 0.81	46.12 ± 2.01	73.44 ± 2.78	54.99 ± 1.91	68.11 ± 1.41	48.55 ± 1.66	66.06 ± 1.40
CODER	UMLS	N/A†	40.50 ± 1.31	57.32 ± 1.74	47.37 ± 1.70	73.34 ± 1.29	53.74 ± 0.71	68.36 ± 0.36	49.86 ± 0.50	69.36 ± 1.00
LUKE	Wikipedia	3.5 billion words	27.44 ± 0.70	44.77 ± 0.40	42.35 ± 1.08	69.10 ± 0.62	50.92 ± 1.26	64.47 ± 1.75	46.16 ± 0.25	64.25 ± 1.28
RadBERT <sup>1</sup>	Radiology reports	2.6 GB	30.34 ± 1.50	48.00 ± 1.43	45.73 ± 0.68	73.00 ± 0.73	54.40 ± 2.84	67.34 ± 1.74	51.52 ± 0.87	68.80 ± 0.80
ClinicalBERT <sup>2</sup>	MIMIC	0.5B words / 3.7GB	27.18 ± 1.89	44.69 ± 0.54	41.88 ± 2.86	68.90 ± 0.71	50.27 ± 1.63	63.40 ± 1.52	46.89 ± 0.13	64.41 ± 0.16
BioMed-RoBERTa <sup>3</sup>	S2ORC	7.55B tokens / 47GB	27.44 ± 1.10	45.44 ± 0.64	42.35 ± 1.70	70.14 ± 0.99	52.82 ± 0.57	66.52 ± 0.32	48.62 ± 0.33	66.91 ± 0.82
Galectica	c.f. section 2	106B tokens	1.37	8.5	1.37	8.5	0.49	10.23	0.49	10.23
MedLLaMA	Medical Corpora	NA	0.3	10.63	0.3	10.63	0.16	12.14	0.16	12.14
MedAlpaca	Medical Meadow	NA	<b>1.68</b>	<b>15.18</b>	<b>1.68</b>	<b>15.18</b>	<b>1.3</b>	<b>16.95</b>	<b>1.3</b>	<b>16.95</b>

extended to learn clinical embeddings from the UMLS knowledge graph which covers several terms found in radiology reports.

Interestingly, PubMed/Blue-BERT perform similarly on both the dev and test sets. This is surprising as theoretically, BlueBERT should have performed better being pre-trained on MIMIC clinical notes. Even more surprising is RadBERT’s performance, which not only is a superior RoBERTa architecture, but is specifically trained on radiology reports. Although this led to marginal improvements over PubMed/Blue-BERT, it comes at the cost of a fraction of the training data. This again indicates the importance of proper domain alignment i.e. *what* data the models are trained on.

Unfortunately, LUKE performed poorly here as compared to Bio/Sci-BERT, showing little, v/s both on dev, to no gain, v/s SciBERT on test, during evaluation. Next, the impact of writing styles in the training corpora is clearly evident in the performance gap between Clinical/Rad-BERT. While the former was trained on more clinical data, it was not the *right* type of data i.e. radiology reports, leading RadBERT to outperform it on both splits.

#### 4.2.2 Targeted Pre-Training

→ **Wiki Baseline:** Training on the Wikipedia corpus, BERT shows an increase in EM only (2.8%) on the dev set, but an overall improvement in all measures on the test set compared to vanilla fine-tuning. RoBERTa, on the other hand, showed overall improvement on the dev set while test scores suffered w.r.t the vanilla baseline.

→ **Normal Prompting:** Training on the unfiltered corpus with normal prompts (row 3) led to an improvement of 3.7% EM for BERT and RoBERTa, and 1.4% F1 for RoBERTa, over the vanilla baseline, which is **greater than our Wiki baseline.**

However, **when the filter is applied (row 4), we observe even more improvements.** BERT shows 7% EM & 0.5% F1 increase while RoBERTa shows 6% EM & 3% F1 increase over regular fine-tuning. These scores indicate the **necessity of using even a simple filter for selecting a subset of training corpora.** That said, the drawback is that care needs to be taken when designing the filter since poor choices can hamper performance (§Table 2).

→ **Best BERT setting:** The most notable improvement for BERT occurs when both filtered entities and the corpus from the fancy prompt is used (row 6), resulting in enhancements over basic fine-tuning (8.1% EM, 0.4% F1 on dev) and the Wikipedia baseline (5.2% EM, 1.1%F1 on dev). We note two things here a) the benefit of studying RadQA and designing targeted prompts and b) BERT reaches these scores with a modest 34.3MB corpus which is << than the benchmarked models.

→ **Best RoBERTa setting:** RoBERTa also demonstrates improvements across different combinations of filtration methods and prompt styles, as well as when trained on the Wikipedia corpus. However, the improvements are inconsistent w.r.t a specific approach. That is, excluding the combined corpora settings, we see it achieve the best performance, on validation, using the corpus obtained from the filtered entities and normal prompting (row 4) and on test, using filtered entities and fancy prompting (row 6).

→ **RoBERTa v/s Benchmarks:** In row 4, RoBERTa, outperforms Bio/Sci-BERT and LUKE on all metrics along with long-context models Big-Bird & Longformer. Interestingly, this **RoBERTa even beats out BioClinicalBERT which was trained using much higher quality clinical notes** scoring 1.9% EM & 1.1% F1 more than it.



Table 4: Targeted Pre-training Results (RadQA). H(F1): HasAns\_F1, H(EM): HasAns\_EM; \*: [Vanilla Fine-Tuning]. [†: normal, ‡: fancy] prompt, ♣: entity filter. **Blue** = **best/red** = **worst** overall; **bold** = **best BERT/RoBERTa setup**.

Model	Time	Corpus Size / Train dataset	Dev				Test			
			EM	F1	H(EM)	H(F1)	EM	F1	H(EM)	H(F1)
BERT	NA*	NA	23.83 ± 0.49	42.91 ± 0.46	36.79 ± 0.76	66.23 ± 0.70	46.20 ± 1.96	59.42 ± 1.10	40.65 ± 1.15	58.30 ± 0.74
RoBERTa			26.12 ± 0.69	43.83 ± 0.44	40.31 ± 1.06	67.65 ± 0.68	51.68 ± 0.73	64.94 ± 0.62	46.45 ± 0.66	64.14 ± 0.42
BERT	≈30 mins	18.4 MB/Wikipedia	24.49 ± 0.38	42.62 ± 0.08	37.80 ± 0.59	65.79 ± 0.13	47.40 ± 1.85	60.11 ± 1.59	41.95 ± 1.73	58.92 ± 1.56
RoBERTa			27.19 ± 0.32	44.49 ± 0.31	41.96 ± 0.49	68.68 ± 0.47	50.54 ± 0.62	63.43 ± 0.75	45.72 ± 0.50	62.92 ± 0.67
BERT	≈11 hrs	81.6 MB/Galactica(≈55k) †	24.70 ± 0.46	42.88 ± 0.30	38.12 ± 0.71	66.19 ± 0.46	46.74 ± 1.84	59.69 ± 0.62	41.60 ± 2.31	58.88 ± 1.64
RoBERTa			27.09 ± 0.70	44.42 ± 0.55	41.81 ± 1.09	68.56 ± 0.86	51.25 ± 0.41	64.41 ± 0.95	46.45 ± 0.66	64.01 ± 0.71
BERT	≈11 hrs	80.3 MB/Galactica(≈55k) †♣	25.51 ± 0.38	43.11 ± 0.23	40.38 ± 2.31	<b>64.76 ± 2.75</b>	46.36 ± 1.79	59.41 ± 1.67	46.59 ± 8.06	57.83 ± 1.81
RoBERTa			27.69 ± 0.23	<b>45.16 ± 0.35</b>	42.74 ± 0.36	69.70 ± 0.54	51.14 ± 0.49	64.29 ± 0.38	46.31 ± 1.36	63.85 ± 0.76
BERT	≈11 hrs	38.1 MB/Galactica(≈55k) ‡	25.10 ± 0.32	42.78 ± 0.55	38.74 ± 0.49	66.03 ± 0.84	46.85 ± 1.74	59.54 ± 1.19	41.66 ± 0.45	58.60 ± 0.37
RoBERTa			27.64 ± 0.49	44.99 ± 0.11	42.67 ± 0.76	69.45 ± 0.16	<b>52.39 ± 0.80</b>	<b>65.57 ± 0.93</b>	<b>47.76 ± 1.85</b>	<b>65.34 ± 1.97</b>
BERT	≈11 hrs	34.3 MB/Galactica(≈55k) ‡♣	<b>25.76 ± 0.66</b>	<b>43.10 ± 0.27</b>	<b>39.68 ± 1.11</b>	<b>66.44 ± 0.55</b>	46.52 ± 1.20	<b>58.98 ± 1.09</b>	<b>40.44 ± 0.95</b>	<b>57.06 ± 0.94</b>
RoBERTa			27.08 ± 0.46	44.67 ± 0.23	41.81 ± 0.72	68.95 ± 0.35	51.30 ± 1.17	64.14 ± 0.41	47.39 ± 1.64	64.53 ± 0.85
BERT	≈22 hrs	120.8 MB/Galactica(≈100k) ‡‡	25.10 ± 0.54	42.74 ± 0.06	38.74 ± 0.83	65.97 ± 0.09	47.07 ± 1.23	60.04 ± 1.19	42.46 ± 0.13	59.78 ± 0.41
RoBERTa			27.49 ± 0.49	44.82 ± 0.38	42.43 ± 0.75	69.17 ± 0.58	51.79 ± 1.98	64.99 ± 1.80	47.03 ± 1.45	64.45 ± 1.21
BERT	≈22 hrs	115.6 MB/Galactica(≈100k) ‡‡♣	24.75 ± 0.23	42.96 ± 0.10	38.20 ± 0.36	66.31 ± 0.16	47.50 ± 1.06	60.38 ± 0.48	41.96 ± 1.31	59.14 ± 1.97
RoBERTa			<b>27.85 ± 0.09</b>	<b>45.16 ± 0.11</b>	<b>42.98 ± 0.14</b>	<b>69.71 ± 0.18</b>	51.79 ± 0.59	64.77 ± 0.40	46.88 ± 1.28	64.21 ± 0.75

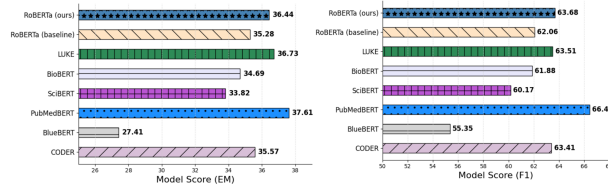


Figure 2: Information Leakage Validation Trials (Left - EM | Right - F1): RoBERTa (ours) was trained on a subset of the 47k corpus with entities only from the 80% train set. All of the models were fine-tuned in the usual manner i.e. SQuAD→COVID-QA (80% train set) and evaluated on the 20% test set.

→ **Combined Prompting Styles:** We examine the effectiveness of combining different context styles (rows 7 and 8) for our approach. The corpora was created by merging the contexts from both prompt styles for the filtered and the unfiltered entities separately. In this setup, BERT shows better performance in row 7 i.e. using the unfiltered corpus (1.7% EM increase over the filtered variant and roughly same F1) and RoBERTa in row 8 i.e. using the filtered corpus (1.3% EM and 0.8%F1 over the unfiltered variant). Overall, this version of RoBERTa resulted in our best overall model beating out the same collection of benchmarks as mentioned above. Such findings suggest incorporating a mixture of prompt styles will create more diverse corpora, enhancing domain alignment.

### 4.3 Investigating Information Leakage

Given that the synthetic corpus generated for COVID-QA in §3 contains entities identified in the *entire* COVID-QA dataset - not from the *train* split within each *fold* - we explore if the performance gains from targeted pre-training are a result of in-

formation leak. To this end, we construct a roughly 80%/20% train/test split (1,676/343 records), ensuring no context overlap, and apply a suite of models to this new split. When applying our targeted pre-training, a synthetic corpus is generated *only* from entities identified in the train split. The results from a brief parameter search for this assay are presented in Figure 2.

As we can see, the RoBERTa model subjected to targeted pre-training still yields strong performance in this restricted scenario, only surpassed by PubMedBERT (& marginally by LUKE in EM), demonstrating that the **improved performance on COVID-QA cannot be attributed to information leak from the test set**. Although the scores are lower than those in Table 2, the relative scores produced by each model leads to a similar conclusion that targeted pre-training yields optimal results.

## 5 Conclusion & Future Work

In this study, we demonstrate the effectiveness of bootstrapping corpora for domain adaptation using FMs, prompting & dataset awareness. We achieve SOTA on COVID-QA and observe notable improvements on RadQA by using combinations of corpora, occasionally surpassing prior benchmarks. However, this work is nascent, and there is much scope for exploration. Our future endeavors involve using larger versions of Galactica to generate even more extensive contexts & fact-checking to eliminate inaccuracies, potentially enhancing the performance of our 10x COVID-QA corpus (§3). Lastly, we aspire to explore complete EQA dataset generation to be used for additional fine-tuning instead of solely relying on pre-training.



## 631 Limitations

632 We identify two limitations of our work. First,  
633 we use a number of GPUs to generate our corpus.  
634 While we were fortunate to have access to powerful  
635 computing clusters, this could form a bottleneck  
636 when being deployed on low-end hardware. How-  
637 ever, with cloud services being made more and  
638 more affordable, we feel that this point can only  
639 be a deal-breaker in severely budget-constrained  
640 settings. And second, in this study, we have only  
641 shown how to generate corpora for the biomedical  
642 domain. For an even wider applicability, we need  
643 to study generation techniques for other closed do-  
644 mains such as Finance, Law, Aviation, etc.

## 645 Ethics Statement

646 As our work relied on publicly available datasets,  
647 we believe that the ethical ramifications here are  
648 limited. That being said, we recognize that to use  
649 RadQA, we had to acquire certifications to access  
650 it. This shows that even though the data in it is  
651 redacted, loosely disseminated patient reports are a  
652 threat to their privacy. Moreover, we had to make  
653 sure that when generating our synthetic reports, we  
654 were not mentioning any patient names, which even  
655 with a small probability might bear resemblance to  
656 an actual person.

## 657 References

658 Emily Alsentzer, John Murphy, William Boag, Wei-  
659 Hung Weng, Di Jindi, Tristan Naumann, and  
660 Matthew McDermott. 2019. [Publicly available clinical BERT embeddings](#). In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 72–78, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

665 Helena Balabin, Charles Tapley Hoyt, Colin Birkenbihl, Benjamin M Gyori, John Bachman, Alpha Tom Kodamullil, Paul G Plöger, Martin Hofmann-Apitius, and Daniel Domingo-Fernández. 2022. Stonkgs: a sophisticated transformer trained on biomedical text and knowledge graphs. *Bioinformatics*, 38(6):1648–1656.

672 Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. [SciBERT: A pretrained language model for scientific text](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3615–3620, Hong Kong, China. Association for Computational Linguistics.

Iz Beltagy, Matthew E Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*.

Hsuvas Borkakoty and Luis Espinosa Anke. 2023. [WIKITIDE: A Wikipedia-based timestamped definition pairs dataset](#). In *Proceedings of the 14th International Conference on Recent Advances in Natural Language Processing*, pages 207–216, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. [BERT: Pre-training of deep bidirectional transformers for language understanding](#). In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2021. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare (HEALTH)*, 3(1):1–23.

Suriya Gunasekar, Yi Zhang, Jyoti Aneja, Caio César Teodoro Mendes, Allie Del Giorno, Sivakanth Gopi, Mojan Javaheripi, Piero Kauffmann, Gustavo de Rosa, Olli Saarikivi, et al. 2023. Textbooks are all you need. *arXiv preprint arXiv:2306.11644*.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. [Don’t stop pretraining: Adapt language models to domains and tasks](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8342–8360, Online. Association for Computational Linguistics.

Michael P Hartung, Ian C Bickle, Frank Gaillard, and Jeffrey P Kanne. 2020. How to create a great radiology report. *RadioGraphics*, 40(6):1658–1670.

Xuanli He, Islam Nassar, Jamie Kiros, Gholamreza Haffari, and Mohammad Norouzi. 2022. [Generate, annotate, and learn: NLP with synthetic text](#). *Transactions of the Association for Computational Linguistics*, 10:826–842.

Tomasz Jurczyk, Amit Deshmene, and Jinho D Choi. 2018. Analysis of wikipedia-based corpora for question answering. *arXiv preprint arXiv:1801.02073*.

Tiffany H Kung, Morgan Cheatham, Arielle Medenilla, Czarina Sillos, Lorie De Leon, Camille Elepaño,

736	Maria Madriaga, Rimel Aggabao, Giezel Diaz-Candido, James Maningo, et al. 2023. Performance of chatgpt on usmle: Potential for ai-assisted medical education using large language models. <i>PLoS digital health</i> , 2(2):e0000198.	791
737		792
738		793
739		794
740		
741	Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. Biobert: a pre-trained biomedical language representation model for biomedical text mining. <i>Bioinformatics</i> , 36(4):1234–1240.	795
742		796
743		797
744		798
745		799
746	Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. <i>arXiv preprint arXiv:1907.11692</i> .	800
747		
748		
749		
750		
751	Kurt Micallef, Albert Gatt, Marc Tanti, Lonneke van der Plas, and Claudia Borg. 2022. <a href="#">Pre-training data quality and quantity for a low-resource language: New corpus and BERT models for Maltese</a> . In <i>Proceedings of the Third Workshop on Deep Learning for Low-Resource Natural Language Processing</i> , pages 90–101, Hybrid. Association for Computational Linguistics.	801
752		802
753		803
754		804
755		805
756		806
757		807
758		
759	Timo Möller, Anthony Reina, Raghavan Jayakumar, and Malte Pietsch. 2020. <a href="#">COVID-QA: A question answering dataset for COVID-19</a> . In <i>Proceedings of the 1st Workshop on NLP for COVID-19 at ACL 2020</i> , Online. Association for Computational Linguistics.	808
760		809
761		810
762		811
763		812
764	Milad Moradi, Kathrin Blagec, Florian Haberl, and Matthias Samwald. 2021. Gpt-3 models are poor few-shot learners in the biomedical domain. <i>arXiv preprint arXiv:2109.02555</i> .	813
765		814
766		815
767		816
768	Mark Neumann, Daniel King, Iz Beltagy, and Waleed Ammar. 2019. <a href="#">ScispaCy: Fast and robust models for biomedical natural language processing</a> . In <i>Proceedings of the 18th BioNLP Workshop and Shared Task</i> , pages 319–327, Florence, Italy. Association for Computational Linguistics.	817
769		818
770		819
771		
772		
773		
774	OpenAI. 2023. Gpt-4 technical report. <i>arXiv</i> .	820
775	Yifan Peng, Shankai Yan, and Zhiyong Lu. 2019. <a href="#">Transfer learning in biomedical natural language processing: An evaluation of BERT and ELMo on ten benchmarking datasets</a> . In <i>Proceedings of the 18th BioNLP Workshop and Shared Task</i> , pages 58–65, Florence, Italy. Association for Computational Linguistics.	821
776		822
777		823
778		824
779		825
780		
781	Charith Peris, Lizhen Tan, Thomas Gueudre, Turan Gokayev, Pan Wei, and Gokmen Oz. 2022. <a href="#">Knowledge distillation transfer sets and their impact on downstream NLU tasks</a> . In <i>Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: Industry Track</i> , pages 128–137, Abu Dhabi, UAE. Association for Computational Linguistics.	826
782		827
783		828
784		829
785		830
786		831
787		832
788		833
789	Barbara Plank. 2016. <a href="#">What to do about non-standard (or non-canonical) language in NLP</a> . In <i>Proceedings of the 13th Conference on Natural Language Processing, KONVENS 2016, Bochum, Germany, September 19-21, 2016</i> , volume 16 of <i>Bochumer Linguistische Arbeitsberichte</i> .	834
790		835
		836
		837
		838
		839
		840
		841
		842
		843
		844
		845
		846
		847

Paul Mooney, Dewey A. Murdick, Devvret Rishi, Jerry Sheehan, Zhihong Shen, Brandon Stilson, Alex D. Wade, Kuansan Wang, Nancy Xin Ru Wang, Christopher Wilhelm, Boya Xie, Douglas M. Raymond, Daniel S. Weld, Oren Etzioni, and Sebastian Kohlmeier. 2020. [CORD-19: The COVID-19 open research dataset](#). In *Proceedings of the 1st Workshop on NLP for COVID-19 at ACL 2020*, Online. Association for Computational Linguistics.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837.

Peter West, Chandra Bhagavatula, Jack Hessel, Jena Hwang, Liwei Jiang, Ronan Le Bras, Ximing Lu, Sean Welleck, and Yejin Choi. 2022. Symbolic knowledge distillation: from general language models to commonsense models. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4602–4625.

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. 2020. [Transformers: State-of-the-art natural language processing](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Ikuya Yamada, Akari Asai, Hiroyuki Shindo, Hideaki Takeda, and Yuji Matsumoto. 2020. [LUKE: Deep contextualized entity representations with entity-aware self-attention](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6442–6454, Online. Association for Computational Linguistics.

An Yan, Julian McAuley, Xing Lu, Jiang Du, Eric Y Chang, Amilcare Gentili, and Chun-Nan Hsu. 2022. Radbert: Adapting transformer-based language models to radiology. *Radiology: Artificial Intelligence*, 4(4):e210258.

Junjie Ye, Xuanting Chen, Nuo Xu, Can Zu, Zekai Shao, Shichun Liu, Yuhan Cui, Zeyang Zhou, Chao Gong, Yang Shen, et al. 2023. A comprehensive capability analysis of gpt-3 and gpt-3.5 series models. *arXiv preprint arXiv:2303.10420*.

Zheng Yuan, Zhengyun Zhao, Haixia Sun, Jiao Li, Fei Wang, and Sheng Yu. 2022. Coder: Knowledge-infused cross-lingual medical term embedding for term normalization. *Journal of biomedical informatics*, 126:103983.

Zhenrui Yue, Bernhard Kratzwald, and Stefan Feuerriegel. 2021. [Contrastive domain adaptation for question answering using limited text corpora](#). In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9575–9593, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Manzil Zaheer, Guru Guruganesh, Kumar Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontanon, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, et al. 2020. Big bird: Transformers for longer sequences. *Advances in neural information processing systems*, 33:17283–17297.

Chunting Zhou, Pengfei Liu, Puxin Xu, Srinu Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, et al. 2023. Lima: Less is more for alignment. *arXiv preprint arXiv:2305.11206*.

## A Synthetic Corpora Samples

We provide samples from our generated corpus. In Figure 3 we show two positive results for our COVID-QA directed corpus. The top one is in the style of a research paper while the bottom one, though shorter in length, details useful information on the required entity. We show negative examples of the same in Figure 4. The top one completely degrades into noise while the bottom one although coherent is talking about an unrelated topic.

For RadQA, we show similar examples. In Figure 5 we show positive (top) and negative (bottom) samples from normal prompting. Note that since programming languages were a part of Galactica’s training corpora, it sometimes generates the same during prompting. Finally, Figure 6 shows positive (top) and negative (bottom) examples of fancy prompting. While the positive sample shows a strong correlation with an actual radiology report, the bottom one completely breaks down.

## B ChatGPT trials

We were curious to see if ChatGPT (with GPT3.5) would be able to answer a few questions from COVID-QA. Only the questions were provided to it without the context since they were much longer than its input capacity. We did not test RadQA since its questions were heavily linked to the patient report on hand and were not as *general* as the former’s. Figure 7 shows a few examples from our trials. As we can see, it seems to answer the questions correctly. However, as mentioned before, these answers need validation from a professional in the field. Also, we see that sometimes it **cannot** answer questions on topics not found in its training



<p><b>Title: chest pain</b>  <b>Title:</b> an unusual cause of chest pain in an elderly patient  <b>Authors:</b> Sushila Shrivastava, Vinay Kumar, Pramod Kumar  <b>Keywords:</b> Chest pain, Eosinophilic pneumonia, Chest pain</p> <p># Abstract</p> <p># Background</p> <p>Pulmonary eosinophilic infiltration is one of the rare causes of non-specific chest pain in the elderly patients. We present a case of elderly patient who presented with a non-specific chest pain of 1 day duration with eosinophilia and chest CT scan showing ground glass opacities.</p> <p># Case presentation</p> <p>A 74-year-old lady presented with a 1-day history of chest pain, nausea and vomiting. The patient had a history of cough for 1 year, and she was on regular medication for cough. The patient had a significant history of hypertension, diabetes mellitus and COPD. On examination, the patient was afebrile with an oxygen saturation of 98% on room air. She had a heart rate of 90 beats per minute and blood pressure of 115/70 mmHg. Chest auscultation revealed a crackles in the lower lung fields. Her chest X-ray showed bilateral pulmonary infiltrates. A diagnosis of eosinophilic pneumonia was made and she was started on oral corticosteroids. She recovered without any further complications after two weeks of treatment.</p> <p># Conclusion</p> <p>Chest pain in the elderly patient is not always indicative of a life-threatening condition. A high index of suspicion for this rare entity should be kept in mind when a non-specific chest pain in elderly patients is seen.</p> <p># Background</p> <p>Pulmonary eosinophilic infiltration is a rare cause of non-specific chest pain in elderly patients. The clinical presentation of the disease is similar to other lung diseases with non-specific symptoms, and a high index of suspicion for this rare entity is required in elderly patients presenting with non-specific chest pain. We present a case of a 74-year-old lady with a history of cough who presented with a non-specific chest pain of 1 day duration with eosinophilia.</p> <p># Case presentation</p> <p>A 74-year-old lady was admitted in a state of mild depression. The patient had a significant history of hypertension, diabetes mellitus and COPD. She had a history of chronic cough for 1 year and was on regular medication for the cough. On examination, the patient was afebrile with an oxygen saturation of 98% on room air. She had a heart rate of 90 beats per minute and blood pressure of 115/70 mmHg. Chest auscultation revealed a crackles in the lower lung fields. She was investigated with a chest X-ray which showed bilateral pulmonary infiltrates. An ECG showed no sign of ST-T segment elevation. Her complete blood count (CBC) was as follows: haemoglobin 12.4 gm%, leucocytes 10,300/cu mm, platelets 253,000/cu mm, red blood cells (RBCs) 2.28 million/cu mm, and haemoglobin percentage (Hb%) 24.9%. Her electrolytes were normal. Her C-reactive protein (CRP) was 37.1 mg/L. D-dimers 0.24 µg/ml, and erythrocyte sedimentation rate (ESR) was 66 mm/1st hour. Her sputum was clear. The chest X-ray showed bilateral pulmonary infiltrates (Fig. 1). A diagnosis of eosinophilic pneumonia was made, and she was started on oral corticosteroids. She recovered without any further complications after two weeks of treatment. Fig. 1: Chest X-ray shows bilateral pulmonary infiltrates</p> <p># Discussion</p> <p>Eosinophilic pneumonia is an uncommon cause of non-specific chest pain. The incidence of this condition is approximately 0.1% of all pulmonary infiltrations [ Idiopathic eosinophilic pneumonia: clinical features, radiological appearances and long-term outcome., Brown, Eosinophilic pneumonia: imaging findings., Bongiovanni]. The most common age group affected is between 40 and 60 years [ Idiopathic eosinophilic pneumonia: clinical features, radiological appearances and long-term outcome., Brown]. However, there have been several case reports of eosinophilic pneumonia in the elderly, with the peak age between 60 and 70 years [ Eosinophilic pneumonia in elderly patients: a case series and review of the literature, Kara, Eosinophilic pneumonia in the elderly, De Luca]. Most of the patients with eosinophilic pneumonia are asymptomatic at presentation, but it can also present with fever, dyspnea, cough, and chest pain [ Idiopathic eosinophilic pneumonia: clinical features, radiological appearances and long-term outcome., Brown, Eosinophilic pneumonia in elderly patients: a case series and review of the literature, Kara, Eosinophilic pneumonia in the elderly, De Luca]. The chest pain in patients with eosinophilic pneumonia can be localized or generalized and can be of mild, moderate or severe intensity. It can be persistent or acute in nature [ Idiopathic eosinophilic pneumonia: clinical features, radiological appearances and long-term outcome., Brown, Eosinophilic pneumonia in the elderly, De Luca]. The patient can be admitted to hospital with a cough, dyspnea, chest pain and fever. However, the patient may not show any of these symptoms at presentation [ Idiopathic eosinophilic pneumonia: clinical features, radiological appearances and long-term outcome., Brown, Eosinophilic pneumonia in the elderly, De Luca].</p> <p>The most common clinical feature of eosinophilic pneumonia is cough, which is present in 85% of patients [ Idiopathic eosinophilic pneumonia: clinical features, radiological appearances and long-term outcome., Brown]. Fever is present in 50% to 75% of patients [ Idiopathic eosinophilic pneumonia: clinical features, radiological appearances and long-term outcome., Brown, Eosinophilic pneumonia in the elderly, De Luca]. Dyspnea and chest pain are seen in less than 10% of patients [ Idiopathic eosinophilic pneumonia: clinical features, radiological appearances and long-term outcome., Brown, Eosinophilic pneumonia in the elderly, De Luca]. The most common respiratory symptoms are cough and dyspnea [ Eosinophilic pneumonia in the elderly, De Luca]. The majority of the patients with eosinophilic pneumonia are asymptomatic, while a minority of patients present with non-specific symptoms of cough, fever, dyspnea and chest pain [ Idiopathic eosinophilic pneumonia: clinical features, radiological appearances and long-term outcome., Brown, Eosinophilic pneumonia in the elderly, De Luca]. Patients with pulmonary eosinophilia and elevated ESR or CRP have been reported to have a poorer prognosis [ Idiopathic eosinophilic pneumonia: clinical features, radiological appearances and long-term outcome., Brown]. Chest radiography can be nonspecific and shows bilateral or unilateral infiltrates [ Idiopathic eosinophilic pneumonia: clinical features, radiological appearances and long-term outcome., Brown, Eosinophilic pneumonia in the elderly, De Luca]. A chest CT scan is often required to diagnose eosinophilic pneumonia and show a variety of appearances. These include ground glass opacities (GGO) and reticular opacities [ Idiopathic eosinophilic pneumonia: clinical features, radiological appearances and long-term outcome., Brown, Eosinophilic pneumonia in the elderly, De Luca]. In our case, the patient had a history of chronic cough and had no fever on presentation. A chest X-ray showed bilateral infiltrates. A chest CT scan showed GGO and reticular opacities.</p> <p>The etiology of eosinophilic pneumonia is unknown, but it has been proposed that it is a T-helper 2-mediated inflammatory disorder [ Idiopathic eosinophilic pneumonia: clinical features, radiological appearances and long-term outcome., Brown]. The diagnosis of eosinophilic pneumonia can be difficult because of the absence of typical radiological findings. It can be difficult to distinguish eosinophilic pneumonia from other inflammatory lung diseases. Treatment consists of corticosteroids, and the use of immunosuppressive drugs, such as azathioprine, hydroxychloroquine, and methotrexate, has been reported to improve the symptoms of eosinophilic pneumonia [ Idiopathic eosinophilic pneumonia: clinical features, radiological appearances and long-term outcome., Brown, Eosinophilic pneumonia in the elderly, De Luca].</p> <p># Conclusion</p> <p>Eosinophilic pneumonia is a rare cause of non-specific chest pain in elderly patients. In these patients, the diagnosis can be difficult and a high index of suspicion should be kept in mind in such patients. Early diagnosis and prompt treatment can improve the prognosis.</p> <p># Consent</p> <p>Written informed consent was obtained from the patient for publication of this case report and any accompanying images. A copy of the written consent is available for review by the Editor-in-Chief of this journal.</p> <p>Footnote: Competing interests</p> <p>The authors declare that they have no competing interests.</p> <p>Footnote: Authors' contributions</p> <p>SS, KK and PK were involved in the management of the patient. SS drafted the manuscript. All authors read and approved the final manuscript.</p>	
<p><b>Title: Polymerase chain reaction</b>  <b>#1:</b> a new diagnostic tool in the management of tuberculosis.</p> <p>Abstract: Polymerase chain reaction (PCR) is an advanced molecular diagnostic technique. The potential of PCR in the diagnosis of tuberculosis (TB) has been evaluated by several investigators. PCR techniques have been used in the diagnosis of extra-pulmonary TB as well as the diagnosis of drug resistance. This article reviews the current status of PCR in the diagnosis of TB.</p>	

Figure 3: COVID-QA positive examples (underline = entity | **red** = prompt) | **olive** = generation





Figure 4: COVID-QA negative examples (underline = entity | **red** = prompt) | **olive** = generation

The role of FNA in the diagnosis of metastases to non-breast related lymph nodes has not been extensively studied. It has been reported that FNA was helpful in the diagnosis of metastases in patients with primary cancers other than breast cancer, such as small cell lung cancer [1]. The diagnostic value of fine-needle aspiration cytology for the diagnosis of nodal metastases in patients with non-small cell lung cancer, Kilm, melanoma [ Cytodiologic diagnosis of metastases to nonbreast-related lymph nodes in breast cancer patients, Park], esophageal cancer [ Fine needle aspiration cytology of lymph node metastases in esophageal carcinoma, Yamagishi], and non-small cell lung cancer [ Fine-needle aspiration cytology of metastatic lung cancer to non-breast-related lymph nodes, Kang]. In our study, 44 lymph nodes were diagnosed as metastatic by FNA (Table 4). Table 4. The accuracy of FNA in staging of metastatic breast cancer to non-breast-related lymph nodes

*In conclusion, FNA of the cervical lymph nodes was accurate in diagnosing metastatic breast cancer. In cases of lymph node enlargement in patients with breast cancer, FNA should be performed. In patients with negative cytomorphology on FNA, clinical, imaging, and biopsy findings should be considered. In these cases, additional procedures should be considered.*

```
TEST_P(ShapeOpTest, ShapeNotSameRank_3D){
  Shape shape;
  shape.set_rank(1);
  shape.set_dims({1, 2, 3});
  shape.set_strides({1, 2, 3});
  shape.set_sizes({1, 2, 3});
  Shape shape_copy = *shape;
  shape.set_rank(0);
  shape_copy.set_rank(1);
  shape_copy.set_dims({1, 2, 3});
  shape_copy.set_strides({1, 2, 3});
  shape_copy.set_sizes({1, 2, 3});
  Shape expected_shape_copy;

  TF_EXPECT_OK(
    TF_EXPECT_OK(shape.AddScalar(2.0, "add_shape=*nullptr,*/is_dynamic=*false");
    TF_EXPECT_OK(shape.AddScalar(3.0, "add_shape=*nullptr,*/is_dynamic=*false");
    TF_EXPECT_OK(shape.AddScalar(4.0, "add_shape=*nullptr,*/is_dynamic=*false");
    TF_EXPECT_OK(
      shape.AddScalar(2.0, "add_shape=*nullptr,*/is_dynamic=*true");
      TF_EXPECT_OK(shape.AddScalar(3.0, "add_shape=*nullptr,*/is_dynamic=*true");
      TF_EXPECT_OK(shape.AddScalar(4.0, "add_shape=*nullptr,*/is_dynamic=*true");
      TF_EXPECT_OK(shape.AddScalar(2.0, "add_shape=*nullptr,*/is_dynamic=*true");
      TF_EXPECT_OK(shape.AddScalar(3.0,
```

14

[illegible]

Figure 6: RadQA - fancy prompt (underline = entity | red = prompt) | olive = generation

**Question:** What is DC-GENR and where is it expressed?

**ChatGPT:** As an AI language model, I do not have access to up-to-date information or research beyond my training data. Therefore, I am not aware of any current information on "DC-GENR." It is possible that this refers to a specific term or concept within a particular field or context. If you could provide me with more context or information about what you are referring to, I may be able to assist you further.

**Question:** What is the main cause of HIV-1 infection in children?

**ChatGPT:** The main cause of HIV-1 infection in children is mother-to-child transmission of the virus during pregnancy, childbirth, or breastfeeding. This is also known as vertical transmission. If the mother is HIV-positive and not receiving treatment, there is a high risk of transmitting the virus to her baby. However, with proper medical care and treatment, the risk of mother-to-child transmission can be significantly reduced.

**Question:** What plays the crucial role in the Mother to Child Transmission of HIV-1 and what increases the risk?

**ChatGPT:** *The level of the mother's HIV viral load plays a crucial role in mother-to-child transmission of HIV-1. A higher viral load in the mother's blood and genital secretions increases the risk of transmission to the baby. Factors that can increase the risk of mother-to-child transmission include lack of antiretroviral therapy during pregnancy, vaginal delivery, and breastfeeding. However, with proper medical care and treatment, the risk of mother-to-child transmission can be greatly reduced. Pregnant women living with HIV should seek medical care and treatment to help protect their babies*

Figure 7: Testing ChatGPT

data, a problem not faced by our encoder models which rely only on the provided context.

## C Hardware Details

To run our corpus generation code, we used a total of 8 40GB NVIDIA A100 or 8 32GB NVIDIA Tesla V100 cards, depending on our institutes job scheduler and availability. Although the former is a superior card, we preferred the latter since it gives relatively quick turnaround and is a general purpose GPU unlike the A100 which is required for even heavier compute. In fact we were requested by our system administrator to use the V100 over the A100, if it was possible, to avoid blocking off even more compute-heavy experiments, which we did not have an issue obeying. That being said, if the A100 cluster was free, we did take advantage of it to run our code. For pre-training and fine-tuning BERT/RobERTa, we used only 1 A100 80GB GPU as RoBERTa was unable to fit on the 40/32 GB cards owing to the large batch size.

## D Prompts Tried

As mentioned before, we knew the prompt to be used for COVID-QA since it was a feature of Galactica. However, we had to experiment with a few prompt styles for RadQA before landing on our final one. We describe the prompts we tried below for the entity `small-bowel injury` and the reasons for not using them. The text in `courier` is the prompt and in *italics* is the generated text.

- IMPRESSION: *small-bowel injury (SBI) was defined as injury to the small bowel wall, or transmural disruption of the bowel wall, in the absence of penetrating abdominal trauma ( The management of small-bowel injuries in the emergency room., Brown). Patients were excluded if the SBIs occurred after admission to the emergency department or were complicated by perforation. ... (truncated)*

– Issue - Did not include the FINDINGS section.

- FINDINGS: *small-bowel injury was observed in 11.1% and 21.9% of patients, respectively. Small-bowel injury associated with liver injury occurred in 10.1%. The mortality rate was significantly higher in the small-bowel injury group than in the*

*non-small-bowel injury group (11.6% vs. 2.1%,  $P = 0.005$ ). This study shows that small-bowel injury is associated with a high mortality rate, and that this complication is an important cause of morbidity and mortality in patients with severe acute pancreatitis.*

– Issue - Did not include the IMPRESSIONS section.

- Patient has small-bowel injury. FINAL REPORT *This patient was treated with conservative methods. The small-bowel injury was treated surgically with resection of the bowel, primary anastomosis, and end-to-end anastomosis between the ends of the resected bowel. She did not have recurrence of intestinal obstruction after 3 years.*

– Issue - Did not include either IMPRESSIONS or FINDINGS section.

- John Doe has small-bowel injury. IMPRESSIONS : *This case study was made possible by his support by a research grant from the National Institutes of Health (CA 10644).*

– Issue - Had to make sure that the pronouns used were same throughout the report otherwise it would not be consistent.

- Jane Doe has small-bowel injury. IMPRESSIONS : *This case study was made possible by the support of a research grant from the Department of Veterans Affairs.*

– Issue - Same as above.

- Patient I has small-bowel injury *He has no history of abdominal trauma or surgery. At first, he has only mild abdominal pain. But with the progress of disease, he has become distressed. At the initial examination, the abdominal pain was alleviated, but the patient still has slight distention of abdomen. He was admitted to our hospital on the third day of admission. His vital signs were: body temperature 37.5 °C, blood pressure 110/70 mmHg, pulse 105 beats/min, respiratory rate 26 breaths/min and oxygen saturation 97% on room air.*



*He had mild abdominal distention. The results of routine blood tests, urinalysis and abdominal ultrasound examination were normal. X-ray showed free air in the abdomen. ... (truncated)*

- Issue - Galactica could get confused between I (alphabet) and I (roman numeral).

- Patient A has small-bowel injury (Fig. 1). He has no history of abdominal trauma or surgery. At first, an abdominal CT scan was performed to evaluate abdominal pain. Abdominal CT showed a partial small-bowel obstruction at the splenic flexure of the colon (Fig. 2). An oral contrast medium was then administered via a nasogastric tube and an abdominal CT scan was performed. ... (truncated)

- Issue - Galactica could get confused misinterpret “A” for the beginning of a sentence (it was observed for a few cases)

- Patient alpha has small-bowel injury requiring an open operation and was diagnosed with *C. difficile colitis*. The patient had a history of severe anemia and leukopenia, which were also observed in patient beta. Patient alpha was treated with a single course of intravenous moxifloxacin 400 mg/day, which resulted in the resolution of the diarrhea. He was followed closely after discharge. In patient beta, he had a similar episode of diarrhea after hospital admission, but the diarrhea did not resolve. He was prescribed with an oral vancomycin 1000 mg twice a day, and the diarrhea was resolved. ... (truncated)

- Issue - This would have been a good choice. However, as we see above, patient qualifiers such as A, I and alpha, might bias the model towards more *male patients* (He). As such, we decided to drop the qualifier altogether.

## E Hyperparameters Used

Hyperparameters for each experiment is detailed in Table 5. These were selected mostly from preexisting implementations or through minimal exploration of known settings.

Experiment	Hyperparameters
Corpus Generation	random seed: 42 renormalize_logits: True do_sample: True max_length (prompt + generated tokens): 2,048 top_p: 0.9 temperature: 0.9
Pre-Training	batch_size: 40 learning_rate: 5e-5 epochs: 3
Fine-Tuning (SQuAD)	batch_size: 16 max_input_length (question + context): 384 stride: 128 learning_rate: 2e-5 epochs: 3 n_best (top n answer spans): 20 max_answer_length: 30 optimizer_type: AdamW
Fine-Tuning (COVID-QA)	batch_size: 40 max_input_length (question + context): 384 stride: 128 learning_rate: 2e-5 epochs: 1 n_best (top n answer spans): 20 max_answer_length: 1000 optimizer_type: AdamW
Fine-Tuning (RadQA)	batch_size: 16 max_length: 384 stride: 128 learning_rate: 3e-5 epochs: 1 n_best (top n answer spans): 20 max_answer_length: 1000 optimizer_type: AdamW

Table 5: Hyperparameters for each experiment. We use three random seeds during pre-training and fine-tuning, 41, 42, 43 but only 42 when generating the corpus. This is done since otherwise to run the entire pipeline from generation to training across all ablations would take an infeasible amount of time.

## F Model Cards

All models used in this study were downloaded from the HuggingFace library (Wolf et al., 2020). Each model, along with its model card (name as it appears in the HuggingFace model hub) and URL is listed in Table 6.

## G Note on Stability

All of our experiments were run using PyTorch 1.13.1 and Huggingface 4.26.1. However, we have noticed fluctuations in results when training with other versions of these libraries. Thus, in order to replicate our scores to the best extent, we recommend installing the aforementioned versions of the packages.

Model	Model Card (URL)
BERT-Base, Cased	<a href="#">bert-base-cased</a>
BERT-Base, Cased, SQuAD v1	<a href="#">batterydata/bert-base-cased-squad-v1</a>
BERT-Base, Cased, SQuAD v2	<a href="#">deepset/bert-base-cased-squad2</a>
RoBERTa-Base	<a href="#">roberta-base</a>
RoBERTa-Base, SQuAD v1	<a href="#">csarron/roberta-base-squad-v1</a>
RoBERTa-Base, SQuAD v2	<a href="#">deepset/roberta-base-squad2</a>
BioBERT	<a href="#">dmis-lab/biobert-base-cased-v1.2</a>
SciBERT	<a href="#">allenai/scibert_scivocab_uncased</a>
SciBERT (+CORD-19)	<a href="#">lordtt13/COVID-SciBERT</a>
PubMedBERT	<a href="#">microsoft/BiomedNLP-PubMedBERT-base-uncased-abstract-fulltext</a>
BlueBERT	<a href="#">bionlp/bluebert_pubmed_mimic_uncased_L-12_H-768_A-12</a>
CODER	<a href="#">GanjinZero/UMLSBert_ENG</a>
LUKE	<a href="#">studio-ousia/luke-base</a>
XLNet, SQuAD v1	<a href="#">arrafmousa/xlnet-base-cased-finetuned-squad</a>
<b>STonKGs *</b>	<a href="#">stonkgs/stonkgs-150k</a>
RadBERT	<a href="#">zzxslp/RadBERT-RoBERTa-4m</a>
Clinical BERT	<a href="#">emilyalsentzer/Bio_ClinicalBERT</a>
BioMed-RoBERTa	<a href="#">allenai/biomed_roberta_base</a>
MedLLaMA	<a href="#">chaoyi-wu/MedLLaMA_13B</a>
MedAlpaca	<a href="#">medalpaca/medalpaca-13b</a>
Galactica	<a href="#">facebook/galactica-1.3b</a>
Longformer, SQuAD v1	<a href="#">valhalla/longformer-base-4096-finetuned-squadv1</a>
Longformer, SQuAD v2	<a href="#">mrm8488/longformer-base-4096-finetuned-squadv2</a>
BigBird, SQuAD v1	<a href="#">FredNajjar/NF-bigbird-squad</a>
BigBird, SQuAD v2	<a href="#">FredNajjar/bigbird-QA-squad_v2</a>

Table 6: Model cards and URLs for all models used in our paper. \* We wanted to use STonKGs (Balabin et al., 2022). However, there was no vocabulary file for the model which resulted in errors.