

Do We Still Need Clinical Language Models?

Eric Lehman^{1,2}
 Evan Hernandez^{1,2}
 Diwakar Mahajan³
 Jonas Wulff²
 Micah J. Smith²
 Zachary Ziegler²
 Daniel Nadler²
 Peter Szolovits¹
 Alistair Johnson⁴
 Emily Alsentzer^{5,6}

LEHMER16@MIT.EDU
 DEZ@MIT.EDU
 DMAHAJA@US.IBM.COM
 JONAS@XYLA.COM
 MICAH@XYLA.COM
 ZACK@XYLA.COM
 DANIEL@XYLA.COM
 PSZ@MIT.EDU
 ALISTAIR.JOHNSON@SICKKIDS.CA
 EALSENTZER@BWH.HARVARD.EDU

¹MIT ²Xyla ³IBM Research ⁴The Hospital for Sick Children
⁵Brigham and Women’s Hospital ⁶Harvard Medical School

Abstract

Although recent advances in scaling large language models (LLMs) have resulted in improvements on many NLP tasks, it remains unclear whether these models trained primarily with general web text are the right tool in highly specialized, safety critical domains such as *clinical text*. Recent results have suggested that LLMs encode a surprising amount of medical knowledge. This raises an important question regarding the utility of smaller domain-specific language models. With the success of general-domain LLMs, is there still a need for specialized clinical models? To investigate this question, we conduct an extensive empirical analysis of 12 language models, ranging from 220M to 175B parameters, measuring their performance on 3 different clinical tasks that test their ability to parse and reason over electronic health records. As part of our experiments, we train T5-Base and T5-Large models from scratch on clinical notes from MIMIC III and IV to directly investigate the efficiency of clinical tokens. We show that relatively small specialized clinical models substantially outperform all in-context learning approaches, even when finetuned on limited annotated data. Further, we find that pretraining on clinical tokens allows for smaller, more parameter-efficient models that either match or outperform much larger language models trained on general text. We release the code and the models used under the PhysioNet Credentialed Health Data license and data use agreement.¹

1. Introduction

Large language models (LLMs) have shown strong performance on a wide variety of natural language processing (NLP) tasks. State-of-the-art LLMs are pretrained on billions of tokens scraped from a mixture of general sources, varying widely in both subject matter and quality. With relatively little task-specific training data, these models can be adapted to new tasks by **finetuning** the model’s weights on labeled data (Devlin et al., 2019) or by including examples of the task **in-context** (Kaplan et al., 2020; Wei et al., 2022). This has made them a promising tool for many different applications.

Recent findings have shown that LLMs with over 100B+ parameters contain embedded clinical knowledge (Singhal et al., 2022). For example, Agrawal et al. (2022) found that GPT-3 competes with or outperforms smaller models on a small set of clinical tasks including acronym disambiguation, co-reference resolution, and medication extraction. Similarly, ChatGPT achieved passing scores on the US Medical Licensing Exam (Kung et al., 2022). From a performance standpoint, these findings raise an important question about the role of smaller models that are *specifically* tailored for clinical text (Alsentzer et al., 2019; Li et al., 2022). With the success of LLMs, **is there still a need for specialized clinical models?**

To answer this question, we take the perspective of a reasonably equipped healthcare system that is attempting to automate a clinical task involving electronic health record (EHR) notes. For example, suppose a hospital wishes to implement semantic search of clinical notes. Without automation, a doctor at the hospital would have to

1. https://github.com/elehman16/clinical_llm

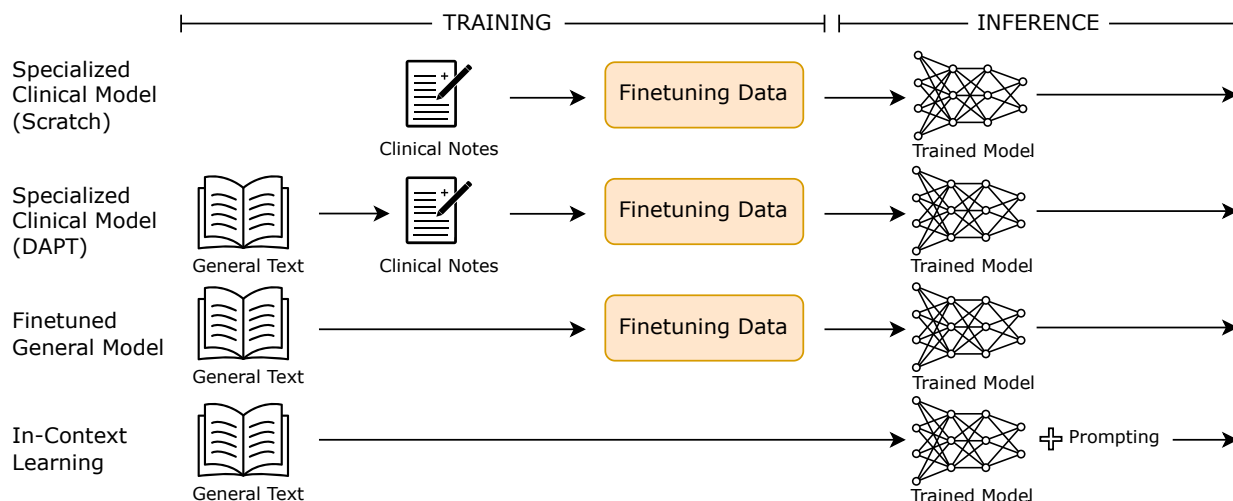


Figure 1: We consider three options for how a healthcare system with access to clinical notes might approach a clinical problem. First, the healthcare system could use a specialized language model pretrained on clinical notes. This model could be pretrained from scratch (Row 1) or from a publicly available checkpoint of a LM pretrained on general text (Row 2). Alternatively, the healthcare system could directly finetune a publicly available general-purpose language model to perform the clinical task (Row 3). Finally, the healthcare system could use a state-of-the-art LLM such as GPT-3, without any additional finetuning, by prompting the LLM to perform the clinical task (Row 4).

manually review all of a patient’s previous notes to understand their patient’s medical history. A language model, however, would allow the hospital to automatically extract answers to questions about a patient’s medical history, using hundreds of past clinical notes as source material. A hospital would have three reasonable options for applying a language model to address this type of clinical problem (Figure 1):

1. Create a **specialized clinical model** by pretraining a language model on in-house clinical notes and finetuning it for a specific downstream task² (Figure 1, first and second rows).
2. Finetune a publicly available pretrained language model, which has largely been pretrained on non-clinical text (Figure 1, third row).
3. Use a state-of-the-art LLM, such as GPT-3, which is made available through an API, and adapt the model to the task using in-context learning (Figure 1, last row).

In this paper, we ask whether there is still a need for **specialized clinical language models**, even with the

availability of impressive domain-agnostic LLMs. To answer this question, we perform an extensive experimental evaluation of 12 different LMs on 3 different clinical tasks that use EHR notes. In addition, we train T5-Base and T5-Large from scratch on clinical notes written primarily in English from the Medical Information Mart for Intensive Care (MIMIC)-III and MIMIC-IV databases (Johnson et al., 2016, 2023). Our results show that relatively small specialized clinical models (345M parameters) substantially outperform all in-context learning approaches that we evaluate, even when finetuned on limited annotated data. We further find that pretraining on clinical tokens allows for smaller, more parameter-efficient models that either match or outperform much larger LMs trained on general text. We will release the code and models from our experiments under the PhysioNet Credentialed Health Data license and data use agreement.

2. Background & Related Work

We specifically focus on clinical tasks that use EHR notes. These notes, which are written by clinicians, contain important information about a patient’s past medical history, lab results, medications, and current clinical presentation. The text in clinical notes differs substantially

2. Hospitals could also use a model pretrained on MIMIC.

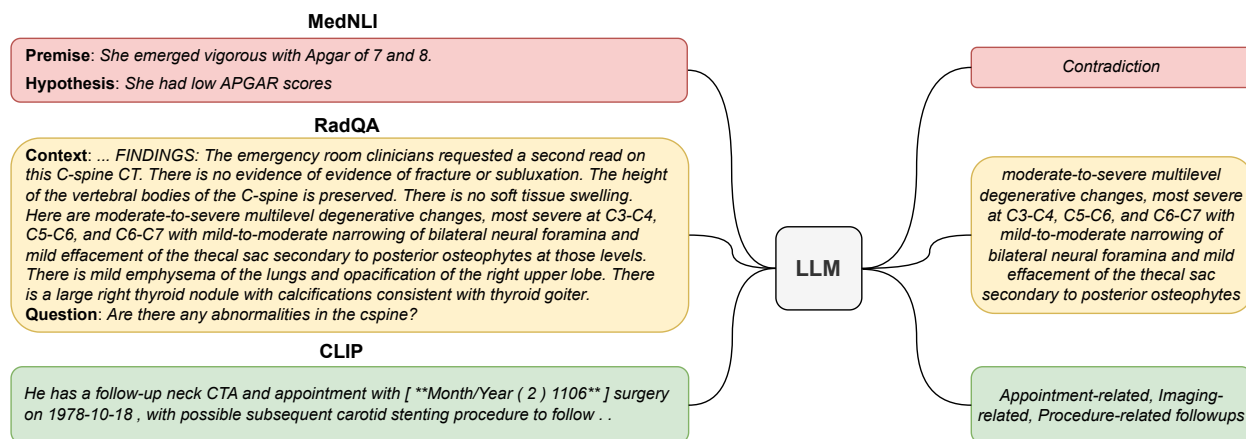


Figure 2: An example of the tasks we consider in this paper. In MedNLI, the goal is determine if the two sentences entail, contradict or are neutral to each other. RadQA is an extractive question answering task over radiology reports. In CLIP, the goal is to identify the different types of patient follow-up information in each sentence of a discharge summary (if any). These examples illustrate the difficulty of parsing clinical text.

from the general-domain text found in LM training corpora. Some of these differences are highlighted in Figure 2: EHR notes often contain grammatical errors (“no evidence of evidence of fracture”), include abbreviations not defined in the context (APGAR, CTA), and reference domain-specific terminology (carotid stenting, subluxation). These peculiarities also lead to substantial differences between clinical text and biomedical text (such as PubMed). Despite the overall shared domain of medicine, biomedical text is otherwise fluent, edited, and polished. This makes clinical tasks that involve these notes particularly challenging. In this section, we briefly describe the three different approaches that one could use for applying a LM to a clinical task (Figure 1).

2.1. Specialized Clinical Models

We define a *specialized clinical model* to be a model pretrained over clinical notes, and refer to models trained on mostly open-domain web text as *general-purpose models*. A specialized clinical model can be trained from scratch, or it can be initialized from a previous checkpoint of a biomedical or general-domain model and pretrained further on clinical data in a process known as domain-adaptive pretraining (DAPT, Gururangan et al. 2020). Models pretrained on clinical notes have shown improved performance compared to their domain-agnostic equivalents (Alsentzer et al., 2019; Lewis et al., 2020; Li et al., 2022; Yang et al., 2022). The semi-structured and abbreviated text found in clinical notes may negatively im-

pact the performance of models pretrained on grammatical biomedical and general text. Further pretraining on clinical text may help these more general models adapt to this domain-shift.

However, pretraining a LM on clinical notes incurs a high upfront cost. This expense may not be justified if it results in only minimal improvements on downstream clinical tasks. Additionally, there is a concern that specialized clinical models pretrained on hospital records may retain sensitive patient information (Carlini et al., 2018; Lehman et al., 2021). For example, Yang et al. (2022) train but *do not release* multi-billion parameter models using notes from the University of Florida Health system, likely due to the unknown risk of the models emitting previously seen protected health information.

2.2. Finetuning General Purpose LLMs for Clinical Tasks

As an alternative to pretraining a specialized clinical model, ML practitioners can finetune a general-purpose LM such as the GPT family of models (Radford and Narasimhan, 2018) or T5 (Raffel et al., 2020), on the clinical task. The capabilities of these models have been well established in the literature: finetuned general-purpose models are effective at clinical question-answering (Pampari et al., 2018), protected health information (PHI) de-identification (Alsentzer et al., 2019), and relation-extraction (Wei et al., 2020). Using a finetuned domain-agnostic model may be necessary in settings where pre-

training a language model from scratch is too costly. While finetuning a general-purpose LM eliminates the cost of pretraining altogether, it may lead to more expensive *inference*-time costs compared to specialized models if the general model must be larger to obtain the same performance. Furthermore, these models may still require regular re-finetuning if the data distribution of the EHR shifts, which may happen if, for example, the hospital system changes how medical personnel write notes (Payne et al., 2010; Blease et al., 2020). This requires substantially more infrastructure and technical expertise to maintain as model sizes grow. There is ongoing research into methods for parameter efficient training (Li and Liang, 2021; Singhal et al., 2022), which reduce the computational cost of finetuning. However, in this work, we only consider finetuning of the *entire* model and leave exploration of these techniques to future work. Finally, in addition to these concerns, models pretrained on text from the general web likely contain additional unexpected and harmful biases towards protected classes and other groups (Bender et al., 2021).

2.3. Using In-Context Learning

A cheaper alternative to finetuning a LM is to use **in-context learning** (ICL). In this setting, examples of the task are included in the input prompt to the model, and no weights are modified. ICL has many potential advantages for the clinical domain because there is often a limited set of labeled data due to the high level of expertise needed for annotation.

In-context learning, paired with LLMs like GPT-3, has shown strong performance on a number of tasks (Brown et al., 2020). Agrawal et al. (2022) found that GPT-3 competes with or outperforms smaller models on several clinical tasks, including acronym disambiguation, coreference resolution, and medication extraction. Due to OpenAI’s data policies,³ Agrawal et al. (2022) are only able to directly test GPT-3’s ability on a restricted set of tasks. Similarly, Kung et al. (2022) found that ChatGPT was able to achieve passing scores on all three stages of the US Medical Licensing Exam (USMLE). However, it is unclear whether such performance on tasks requiring clinical knowledge translates to tasks that require parsing semi-structured, abbreviation-laden clinical notes.

In practice, ICL performs best in very large models (Singhal et al., 2022) or in models explicitly trained for

ICL (Wei et al., 2021a). These models perform as well as — or better than — many finetuned models on several language tasks, which makes ICL a quick and easy option for many NLP problems. However, GPT-scale models are typically accessible only through APIs hosted by private companies, which may add additional concerns about security and data privacy. Additionally, these models have a tendency to generate realistic, but factually incorrect content, which may be especially problematic in the safety-critical medical domain. We exclude ChatGPT and GPT-4 from our analysis due to the ambiguity surrounding model size, architecture and details about which datasets were included in their training procedure. For these reasons, it is difficult to draw any conclusions about the performance of these models.

3. Experimental Setup

We examine the performance of 12 different LMs on three different clinical tasks derived from MIMIC (Figure 2).

3.1. Tasks

We select tasks that test the ability to parse and reason over clinical notes. We describe these tasks below:

- **MedNLI** (Romanov and Shivade, 2018) is a natural language inference task in which the goal is to determine whether a hypothesis written by a doctor can be inferred from a premise taken directly from a clinical note (multi-class classification with labels *entailment*, *neutral*, or *contradiction*). We measure performance using accuracy.
- **RadQA** (Soni et al., 2022) is a question-answering (QA) task on radiology reports. Doctors were provided text describing the clinical reason for the imaging and were instructed to ask questions about the radiology report. The answers, if available, were extracted from the report. We measure performance using token-level F1 and exact string match metrics.
- **CLIP** (Mullenbach et al., 2021) is a multi-label classification task in which the goal is to identify key-sentences that contain some follow-up information in discharge summaries. Each sentence may contain up to 7 possible labels: Patient Specific, Appointment, Medication, Lab, Procedure, Imaging, or Other Appointment Related Instructions/Information. We measure performance using micro and macro F1-Score.

3. When Agrawal et al. (2022) was released, OpenAI stored all inputs to be used as training data, which violated MIMIC’s data use agreement. As of January 16, 2023, it is now possible to use OpenAI models via Microsoft Azure’s HIPAA certified platform (Boyd, 2023).

Model	Size	Architecture	General PTT	BioMed PTT	Clinical PTT
T5-Base	220M	Encoder-Decoder	34B	0.5B	–
Clinical-T5-Base-Ckpt	220M	Encoder-Decoder	34B	0.5B	13B
Clinical-T5-Base	220M	Encoder-Decoder	–	–	40B
RoBERTa-Large	345M	Encoder Only	2200B	–	–
BioClinRoBERTa	345M	Encoder Only	–	2037B	65B
GatorTron	345M	Encoder Only	40B	92B	1570B
T5-Large	770M	Encoder-Decoder	34B	0.5B	–
Clinical-T5-Large	770M	Encoder-Decoder	–	–	38B
PubMedGPT	2.7B	Decoder Only	–	300B	–
T5-XL	3B	Encoder-Decoder	34B	0.5B	–
Flan-T5-XXL	11B	Encoder-Decoder	34B	0.5B	–
GPT-3	175B	Decoder Only	?	?	?

Table 1: We show all the models used in this paper, as well as their size, architecture and make up of pretraining data. We are unable to provide any information on GPT-3. We focus only on pretraining data, and ignore any finetuning data. PTT stands for pretraining tokens. We use ? for GPT-3, as details about its training have not been released.

3.2. Models

We experiment with two existing clinical models, BioClinRoBERTa⁴ (Lewis et al., 2020) and GatorTron (Yang et al., 2022), which are both 345M parameter encoder-only models based on the BERT-Large architecture (Devlin et al., 2019). GatorTron was trained on a combination of Wikipedia, PubMed, MIMIC-III, and notes from the University of Florida Health system, whereas BioClinRoBERTa was trained exclusively over PubMed and MIMIC-III. One additional difference between these two models is that GatorTron is trained using the objective function presented in Lan et al. (2019), while BioClinRoBERTa is trained using the techniques described in Liu et al. (2019).

Relative to the general and biomedical domains, there are only a small number of available clinical LMs, primarily due to the paucity of publicly available clinical notes. To supplement our experiments using *specialized clinical models*, we train three different clinical T5 models on MIMIC III and MIMIC IV, which total ≈ 1.2 B words (2B tokens). The T5 models are encoder-decoder LMs that are trained with a generative masked language modeling loss (Devlin et al., 2019). Raffel et al. (2020) pretrain several T5 models of varying size (T5-Base, T5-Large, T5-XL, etc.) on text from the general web. We describe our pre-trained models below and provide an extensive detail on training method, data preprocessing, and model hyperparameters in Appendix A:

- **Clinical-T5-Base-Ckpt:** We initialize from the T5-Base (220M) checkpoint and train on MIMIC for 13B tokens. This would classify as a Specialized Clinical Model (DAPT) in row two of Figure 1.
- **Clinical-T5-Base:** We randomly initialize T5-Base from scratch and train on MIMIC for 40B tokens. This would classify as a Specialized Clinical Model (Scratch) in row one of Figure 1.
- **Clinical-T5-Large:** We randomly initialize T5-Large (770M) from scratch and train on MIMIC for 38B tokens. This would classify as a Specialized Clinical Model (Scratch) in row one of Figure 1.

To ground the results of the specialized clinical models, we compare to several different general domain models (Table 1), including RoBERTa (Liu et al., 2019), T5-Base, and T5-Large. RoBERTa shares the same architecture as GatorTron and BioClinRoBERTa, while T5-Base and T5-Large share the same architecture as Clinical-T5-Base and Clinical-T5-Large, respectively. However, RoBERTa, T5-Base and T5-Large are trained exclusively on general-domain text.

In order to examine how specialized clinical models compare to significantly larger, non-clinical models, we compare to PubMedGPT (Bolton et al., 2022) and T5-XL, as these are the largest models that we are able to fully finetune. All finetuning hyperparameters are reported in Appendix B. Additionally, we examine how these specialized clinical models compare to LLMs used with ICL. For these experiments, we use GPT-3 (text-davinci-003, Ouyang et al. 2022) and T5-

4. We rename the model (RoBERTa-large-PM-M3-Voc) from Lewis et al. (2020) to be BioClinRoBERTa.

Flan-XXL (Chung et al., 2022). We explore using a number of different prompts (~ 10 -20) and report additional details in Appendix D.

4. Clinical Models Are Parameter Efficient

In this section, we study how smaller specialized clinical models compare to larger models trained on the general domain. We fix the *model architecture* and compare models pretrained on general data (T5-Base, T5-Large, T5-XL) versus clinical data (Clinical-T5-Base-Ckpt, Clinical-T5-Base, Clinical-T5 Large). We find that Clinical-T5-Base-Ckpt and Clinical-T5-Base outperform their general domain counterpart, T5-Base, while Clinical-T5-Large outperforms T5-Large (Table 2). This is despite the fact that we pretrain for several epochs (15+) on the relatively small set of tokens present in MIMIC, which Raffel et al. (2020) shows negatively impacts performance relative to pretraining on unique text for less than one epoch. Furthermore, we find that pretraining from scratch on clinical data yields the largest performance gains. While domain adaptive pretraining of T5-Base on clinical data improves performance over T5-Base, training from scratch is more effective, leading to +3% and +5% gains over Clinical-T5-Base-Ckpt on RadQA and CLIP, respectively. The weaker performance of Clinical-T5-Base-Ckpt could be explained by a sub-optimal learning rate. Selecting a continuation learning rate is a known challenge of domain-adaptive pretraining (Hoffmann et al., 2022).

While there is substantial evidence that specialized clinical models can outperform their similarly sized general domain equivalents (Lewis et al., 2020; Liu et al., 2019; Alsentzer et al., 2019), it is less clear whether specialized clinical models can outperform *larger* general-domain models. We investigate this by comparing T5 models of varying sizes. We find that Clinical-T5-Base slightly outperforms T5-Large ($3.5\times$ larger) on all three tasks, but fails to outperform T5-XL ($13.5\times$ larger). Similarly, Clinical-T5-Large slightly outperforms or performs similarly to T5-XL ($3.5\times$ larger). This comparison between models trained on in-domain data and larger domain-agnostic models demonstrates that **specialized clinical models can achieve comparable or better performance with significantly fewer computational resources**. This is particularly important for hospital systems, which often lack the infrastructure necessary to run computationally intensive models. By training models specifically on in-domain data, hospitals can still benefit from state-of-the-art LLMs, but with a smaller, more man-

ageable model that can operate in computationally constrained environments.

4.1. When Is Pretraining From Scratch More Efficient?

Pretraining a specialized clinical model from scratch has a high initial one-time cost. However, performing this pretraining, as our results above suggest, enables the model to be significantly smaller than a general-purpose model while still exhibiting similar downstream performance. This means that despite a high initial cost, the cost of both finetuning and running inference on a specialized clinical model greatly decreases. In this section, we determine **at what point** it is more computationally expensive to use a *larger* domain-agnostic model versus pretraining a *smaller* specialized model from scratch. We measure the cost of a model in terms of FLOPs (Kaplan et al., 2020), which is a function of model size and number of pretraining tokens. We compare the costs of pretraining, finetuning, and performing inference on specialized clinical models versus finetuning and performing inference on an existing general domain model. We assume here that the entire model is updated during the finetuning process.

The training cost C_t and inference cost C_i of a model are a function of the number of parameters P in the model and the number of tokens T that are processed (Kaplan et al., 2020):

$$C_t(P, T) = 6PT \quad (1)$$

$$C_i(P, T) = 2PT \quad (2)$$

The number of tokens T in the above cost functions depend on the vocabulary and tokenization process. One additional benefit of training from scratch is that it enables use of an in-domain vocabulary: words previously broken up into word-pieces by a general tokenizer may now be treated as a single token. We find that for every 1 clinical token, there are ≈ 1.12 general tokens.⁵ We model this using an additional token cost weight w , with $w_c = 1.0$, $w_g = 1.12$ for clinical and general-domain tokenizers, respectively. Using T_{pt} pretraining tokens, T_{ft} finetuning tokens (both fixed), and T_i inference tokens, we can write the total cost required to pretrain, finetune, and perform inference as follows:

5. We calculate this by running the T5-Base tokenizer over all of MIMIC, as compared to Clinical-T5-Base (same vocabulary size). There is roughly a 65% overlap between the two vocabularies.

Size	Model	MedNLI	RadQA		CLIP	
		Acc.	EM	F1	Micro F1	Macro F1
220M	T5-Base	0.818	0.479	0.662	0.767	0.594
	Clinical-T5-Base-Ckpt	0.852	0.507	0.689	0.772	0.605
	Clinical-T5-Base	0.855	0.531	0.710	0.793	0.652
770M	T5-Large	0.849	0.537	0.700	0.779	0.629
	Clinical-T5-Large	0.872	0.550	0.745	0.800	0.663
3B	T5-XL	0.869	0.568	0.729	0.780	0.640

Table 2: We compare the performance of T5-models with varying pretraining setups. Performance is based on the mean of 3 seeds. Specialized clinical models can outperform larger, general-purpose models like T5-XL.

$$\begin{aligned}
C_m(P, T_i, T_{pt}, T_{ft}, w) & \\
&= C_t(P, wT_{pt}) + C_t(P, wT_{ft}) + C_i(P, wT_i) \\
&= 6Pw(T_{pt} + T_{ft}) + 2PwT_i \quad (3)
\end{aligned}$$

We can now compare the cost of a small, specialized clinical model of size P_c with a larger, general-domain, previously pretrained (i.e. $T_{pt} = 0$) model of size P_g , with $P_c < P_g$. Assuming the same amount of finetuning tokens, T_{ft} , the costs of both models (C_c and C_g) to run inference over T_i tokens becomes:

$$\begin{aligned}
C_c(P_c, T_{pt}, T_{ft}, T_i, w_c) &= 6P_cw_c(T_{pt} + T_{ft}) \\
&\quad + 2P_cw_cT_i \quad (4)
\end{aligned}$$

$$\begin{aligned}
C_g(P_g, T_{pt} = 0, T_{ft}, T_i, w_g) &= 6P_gw_gT_{ft} \\
&\quad + 2P_gw_gT_i \quad (5)
\end{aligned}$$

Equating (4) and (5) and solving for the number of inference tokens, T_i , we find the point at which the costs of running inference with the clinical and the general model become equal:

$$T_{i, breakeven} = \frac{3[w_cP_c(T_{pt} + T_{ft}) - w_gP_gT_{ft}]}{w_gP_g - w_cP_c} \quad (6)$$

Ignoring finetuning costs and using Clinical-T5-Large and T5-XL as our comparison models, it would take $\sim 40B$ tokens of inference to recover the costs of pretraining from scratch on clinical data. For reference, we estimate that University of Florida Health, which is a large health system with over 1000 beds, records $\sim 15B$ tokens per year (Yang et al., 2022). While it would take ~ 2.5

years to recover the cost of a specialized clinical model for a single task that runs over each note once, in practice, such a model would be used for numerous tasks and potentially operate over multiple years of clinical notes. Given that the two models perform similarly, these results suggest that training a smaller specialized clinical model would allow hospitals to leverage the benefits of LMs, without the higher inference-time and environmental costs of running significantly larger models.

5. In-Domain Tokens Are More Valuable

In Section 4, we examine performance based on a *fixed* model architecture. In this section, we expand the models we consider to include two more specialized clinical models (GatorTron, BioClinRoBERTa), as well non-clinical models that were trained for a similar number of FLOPs (RoBERTa, PubMedGPT). We aim to explore how performance changes as a function of the amount of general, biomedical and clinical FLOPs used during pretraining.

BioClinRoBERTa and GatorTron achieve the highest performance on all tasks (Table 3). This is despite the fact that both of these models are less than 12% of the size of T5-XL, suggesting that model size alone does not guarantee state-of-the-art performance. Another hypothesis is that the total number of FLOPs drives performance; notably, both BioClinRoBERTa and GatorTron were trained for significantly more FLOPs than T5-XL. However, we find that RoBERTa, which is trained for more total FLOPs than GatorTron and BioClinRoBERTa and shares the same BERT-Large architecture, fails to outperform both of these models. This suggests that the high performance of GatorTron and BioClinRoBERTa stems from the makeup of their training data, rather than the total number of FLOPs.

Size	Model	Compute FLOPs			MedNLI	RadQA		CLIP	
		General	BioMed	Clinical	Acc.	EM	F1	Micro	Macro
220M	T5-Base	4.5E+19	6.6E+17	–	0.818	0.479	0.662	0.767	0.594
	Clinical-T5-Base	–	–	5.3E+19	0.855	0.531	0.710	0.793	0.652
345M	RoBERTa	4.6E+21	–	–	0.852	0.521	0.684	0.793	0.677
	BioClinRoBERTa	–	4.2E+21	1.4E+20	0.900	0.604	0.759	0.805	0.707
	GatorTron	1.4E+19	1.9E+20	3.3E+21	0.883	0.583	0.759	0.791	0.690
770M	T5-Large	2.6E+19	2.3E+18	–	0.849	0.537	0.700	0.779	0.629
	Clinical-T5-Large	–	–	1.8E+20	0.872	0.550	0.745	0.800	0.663
2.7B	PubMedGPT	–	4.9E+21	–	0.870	0.512	0.698	0.819	0.666
3B	T5-XL	1E+20	9E+18	–	0.869	0.568	0.729	0.780	0.640
11B	Flan-T5-XXL	3.7E+20	5.5E+18	–	0.808	0.300	0.602	0.164	0.178
175B	GPT-3	?	?	?	0.805	0.362	0.619	0.154	0.146

Table 3: A comparison of clinical and general models trained with varying FLOPs on the three clinical tasks. We only evaluate the ICL methods on 25% of the test set for CLIP due to the time required for inference on the dataset. We report the mean performance over 3 random seeds. GatorTron and BioClinRoBERTa obtain the highest performance on all metrics except Micro F1 on CLIP. EM stands for exact-match. Macro and Micro stand for Macro and Micro F1 respectively.

Similarly, we find that PubMedGPT, which is trained on PubMed for the largest number of total FLOPs, fails to outperform significantly smaller clinical models. This is especially striking considering that PubMedGPT achieves a high performance on the United States Medical Licensing Exam (USMLE), a set of standardized tests required for medical licensure in the United States (Bolton et al., 2022). In fact, we find that GatorTron scores 10 points *worse* than PubMedGPT on the USMLE, suggesting that there is a difference between the ability to leverage conventional medical knowledge and parse a clinical note.

As we saw in Section 4, clinical models outperform their domain-agnostic equivalents. Figure 3 additionally highlights that domain-agnostic models do so with fewer parameters. Furthermore, given a fixed level of performance, we see that clinical models are more computationally efficient than general-domain models. For example, Clinical-T5-Large and T5-XL achieve comparable performance on MedNLI, yet T5-XL requires 3.5 times as many FLOPs. While model architecture differences make a direct comparison difficult, we see that these trends hold for the non-T5 models as well. These results suggest that **increasing the number of biomedical and clinical FLOPs, as opposed to the number of parameters or**

total FLOPs, is the most promising approach for improving performance on tasks based on clinical text.

6. In-Context Learning Underperforms Task Specific Models

Recent works have shown that LLMs can be adapted to new domains simply through ICL (Wei et al., 2022; Li’evin et al., 2022; Agrawal et al., 2022; Sanh et al., 2021). This type of approach is especially appealing in settings where there is a limited amount of labeled data. To properly compare ICL to specialized clinical models and general-purpose models, we simulate a setting in which we have access to very limited data, even as low as < 100 samples. Concretely, we finetune RoBERTa, BioClinRoBERTa, GatorTron, Clinical-T5-Large and PubMedGPT on 1%, 5%, 10%, 25% and 100% of the available finetuning data for each task and compare the finetuned models to ICL with GPT-3 and Flan-T5-XXL.

We find that **models finetuned on all available data significantly outperform any ICL approach** for all of our tasks (Figure 4). This is consistent with prior results, which compared ICL with parameter-efficient finetuning (Liu et al., 2022). These findings are particularly relevant to the safety critical clinical domain, where ML practi-

DO WE STILL NEED CLINICAL LANGUAGE MODELS?

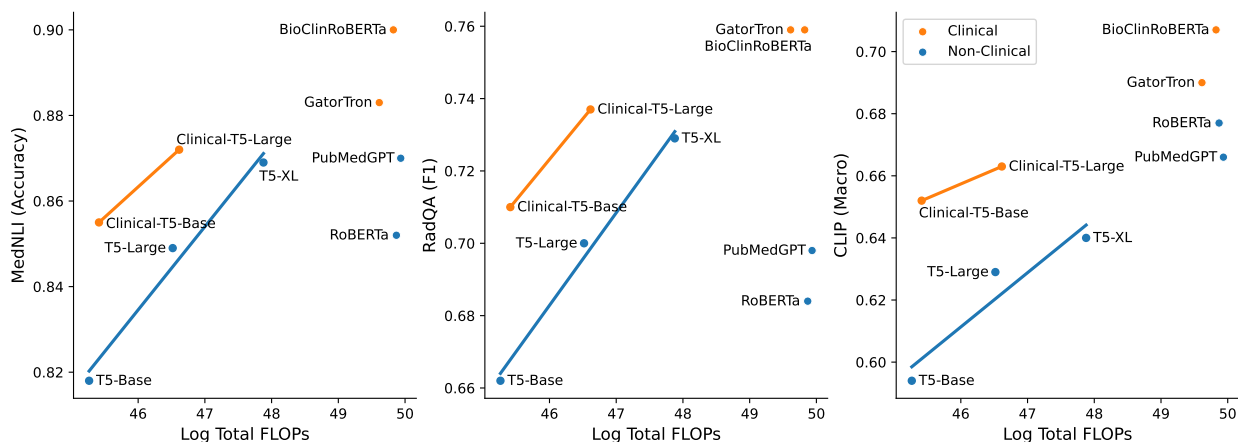


Figure 3: Log total pretraining FLOPs by performance for MedNLI, RadQA, and CLIP. When comparing models with a similar number of FLOPs or performance, clinical models outperform general models. We add regression curves for all T5 models, which are comparable in architecture and training process and differ only in model size and pretraining domain. The T5 models demonstrate the effectiveness of clinical tokens relative to tokens taken from the general web.

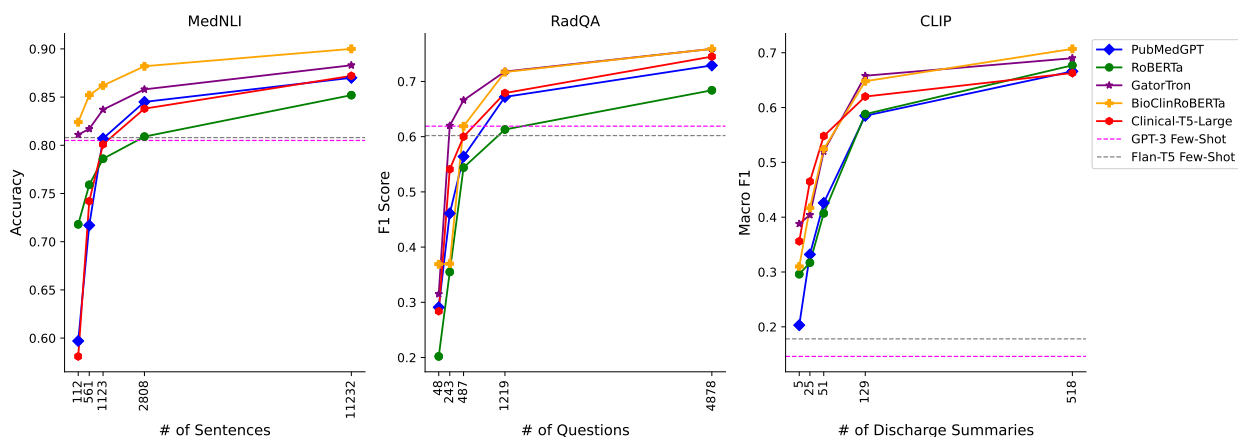


Figure 4: An ablation study in which we compare models trained with 1%, 5%, 10%, 25%, and 100% of available training data for each task. Except for RadQA at 1%, GPT-3 and T5-Flan-XXL perform worse than GatorTron at all ablation points. We report mean performance over three random seeds.

tioners may be willing to gather additional finetuning data for improved performance in high-risk settings.

The utility of specialized clinical models in the few-shot setting varies across datasets. On MedNLI, both BioClinRoBERTa and GatorTron outperform GPT-3 in all resource-restricted settings. On RadQA, GPT-3 and Flan-T5-XXL outperform the smaller specialized clinical models, but only when the specialized models are trained on 1% (49 question-answer pairs) of training data. It is worth noting that GPT-3 and Flan-T5-XXL are finetuned on question-answering style tasks (Ouyang et al., 2022;

Chung et al., 2022), albeit it is unlikely that these tasks are from the clinical domain.

We find that all models outperform GPT-3 and Flan-T5-XXL on CLIP, even when only 5 discharge summaries are used for training data. We believe that this can be attributed to the aggressive sentence-segmentation of the discharge summaries in the CLIP dataset, as well as the lack of specificity of the task labels.⁶ For example, GPT-3 struggles to categorize

6. The aggressive sentence-segmentation leads to sentences like “Discharge Instructions:”. If important follow-up information follows a

size labels of type `Other Appointment Related Instructions`, which significantly lowers its overall performance on CLIP. Further, unlike RadQA and MedNLI, the label space of this task is different from the type of tasks that GPT-3 and Flan-T5-XXL were finetuned on.

On two of the three datasets, the 11B Flan-T5-XXL model outperforms the much larger 175B GPT-3 model. Flan-T5-XXL is publicly available and can be run with ICL locally on a single GPU, particularly with the aid of libraries such as DeepSpeed (Rajbhandari et al., 2019), making it a promising option for ICL when compute is limited.

We can also examine the gap in performance between clinical (GatorTron, BioClinRoBERTa, Clinical-T5-Large) and non-clinical (RoBERTa, PubMedGPT) pretrained models. For RadQA and CLIP in particular, there is a clear gap in performance between clinical and non-clinical models. This gap is largest in limited data settings (5% and 10%), and slowly diminishes as the amount of finetuning data increases. This suggests that pretraining on in-domain data can be especially advantageous when there is a low amount of text available for finetuning.

7. Limitations & Future Work

In this paper, we test 12 different LMs on 3 different clinical tasks. We specifically select tasks that test the ability to reason over and parse clinical notes. However, we do not test the ability of these models to reason over *long text*, which is a considerable challenge when working with clinical notes. We also do not consider tasks that require generating clinical text (e.g. summarization), which would likely be challenging for encoder-only models. Further, this work does not consider the various techniques that can be used to reduce model size (e.g., distillation (Hinton et al., 2015), pruning (Janowsky, 1989)) or perform parameter-efficient training (e.g., prompt-tuning (Li and Liang, 2021)). Another limitation is that we make some comparisons *across* different architectures. While this is still a valuable comparison, we cannot attribute improvements in performance to the pretraining data distribution versus the model architecture. Lastly, we do not use any instruction-tuned models (Wei et al., 2021b), which are finetuned on a collection of tasks described via instructions, in our finetuning experiments, and we do not compare against ChatGPT or GPT-4 because the model

header sentence, then the header is also marked with the label of the following sentence. This makes it particularly challenging to do in an ICL setting; however, it is possible that extensive heuristics may help alleviate this issue.

size, architecture, and training data are unknown. Future work is needed to evaluate ChatGPT and GPT-4 on private EHR datasets that are unlikely to be included in the training data. In the future, we would also like to compare to these models and develop instruction-tuned models tailored to the clinical domain.

8. Conclusion

In this paper, we explore whether there is still a need for smaller specialized clinical language models. To answer this question, we conduct an extensive experimental analysis of 12 models, ranging from 220M to 175B parameters, on 3 different clinical tasks that test the ability to parse and reason over electronic health records. Our results suggest that smaller models, specifically tailored for clinical text, are more parameter efficient than larger domain-agnostic models. Further, we find that using in-context learning with extremely large language models, like GPT-3, is not a sufficient replacement for finetuned specialized clinical models. These findings highlight the importance of developing models for highly specialized domains such as clinical text.

Data and Code Availability We use the MIMIC-III (Johnson et al., 2016) and MIMIC-IV (Johnson et al., 2023) datasets in this paper for training and evaluating specialized clinical models. We experiment with several tasks, including MedNLI (Romanov and Shivade, 2018), RadQA (Soni et al., 2022), and CLIP (Mullenbach et al., 2021). All of these datasets are available on PhysioNet (Goldberger et al., 2000 (June 13)). In the appendix, we also report results on the n2c2 2010 (Uzuner et al., 2011), 2012 (Sun et al., 2013), and 2014 (Stubbs et al., 2015) challenges. We plan to make our code and models available upon paper acceptance.

Institutional Review Board (IRB) This work solely leverages de-identified clinical notes from PhysioNet and the n2c2 challenges and therefore does not require IRB approval. All authors who interacted with the data were credentialed PhysioNet and n2c2 users.

References

- Monica Agrawal, Stefan Heggelmann, Hunter Lang, Yoon Kim, and David A. Sontag. Large language models are zero-shot clinical information extractors. *ArXiv*, abs/2205.12689, 2022.
- Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jindi, Tristan Naumann, and Matthew

- McDermott. Publicly available clinical BERT embeddings. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 72–78, Minneapolis, Minnesota, USA, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/W19-1909. URL <https://aclanthology.org/W19-1909>.
- Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. On the dangers of stochastic parrots: Can language models be too big? *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 2021.
- Sid Black, Stella Biderman, Eric Hallahan, Quentin Anthony, Leo Gao, Laurence Golding, Horace He, Connor Leahy, Kyle McDonell, Jason Phang, Michael Pieler, USVSN Sai Prashanth, Shivanshu Purohit, Laria Reynolds, Jonathan Tow, Ben Wang, and Samuel Weinbach. GPT-NeoX-20B: An open-source autoregressive language model. In *Proceedings of the ACL Workshop on Challenges & Perspectives in Creating Large Language Models*, 2022. URL <https://arxiv.org/abs/2204.06745>.
- Charlotte Blease, John Torous, and Maria Hägglund. Does patient access to clinical notes change documentation? *Front. Public Health*, 8:577896, November 2020.
- Elliot Bolton, David Hall, Michihiro Yasunaga, Percy Liang, Michael Carbin, Jonathan Frankle, Abhinav Venigalla, Chris Manning, and Tony Lee. Pubmed gpt: a domain-specific large language model for biomedical text, Dec 2022. URL <https://crfm.stanford.edu/2022/12/15/pubmedgpt.html>.
- Eric Boyd. General availability of azure openai service expands access to large, advanced ai models with added enterprise benefits, Jan 2023. URL <https://azure.microsoft.com/en-us/blog/general-availability-of-azure-openai-service-expands-access-to-large-advanced-ai-models-with-added-enterprise-benefits/>.
- Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, T. J. Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeff Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. *ArXiv*, abs/2005.14165, 2020.
- Nicholas Carlini, Chang Liu, Úlfar Erlingsson, Jernej Kos, and Dawn Xiaodong Song. The secret sharer: Evaluating and testing unintended memorization in neural networks. In *USENIX Security Symposium*, 2018.
- Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. Scaling instruction-finetuned language models, 2022. URL <https://arxiv.org/abs/2210.11416>.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *ArXiv*, abs/1810.04805, 2019.
- A. L. Goldberger, L. A. N. Amaral, L. Glass, J. M. Hausdorff, P. Ch. Ivanov, R. G. Mark, J. E. Mietus, G. B. Moody, C.-K. Peng, and H. E. Stanley. PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals. *Circulation*, 101(23):e215–e220, 2000 (June 13). Circulation Electronic Pages: <http://circ.ahajournals.org/content/101/23/e215.full> PMID:1085218; doi: 10.1161/01.CIR.101.23.e215.
- Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. Don’t stop pretraining: Adapt language models to domains and tasks. *ArXiv*, abs/2004.10964, 2020.
- Geoffrey E. Hinton, Oriol Vinyals, and Jeffrey Dean. Distilling the knowledge in a neural network. *ArXiv*, abs/1503.02531, 2015.
- Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katie Millican, George van den Driessche, Bogdan Damoc,

- Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, Jack W. Rae, Oriol Vinyals, and L. Sifre. Training compute-optimal large language models. *ArXiv*, abs/2203.15556, 2022.
- S A Janowsky. Pruning versus clipping in neural networks. *Phys. Rev. A Gen. Phys.*, 39(12):6600–6603, June 1989.
- Alistair E. Johnson, Lucas Bulgarelli, Lu Shen, Alvin Gayles, Ayad Shammout, Steven Horng, Tom J. Pollard, Benjamin Moody, Brian Gow, Li-wei H. Lehman, and et al. Author correction: Mimic-iv, a freely accessible electronic health record dataset. *Scientific Data*, 10(1), 2023. doi: 10.1038/s41597-023-01945-2.
- Alistair EW Johnson, Tom J Pollard, Lu Shen, H Lehman Li-wei, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3:160035, 2016.
- Jared Kaplan, Sam McCandlish, T. J. Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeff Wu, and Dario Amodei. Scaling laws for neural language models. *ArXiv*, abs/2001.08361, 2020.
- Tiffany Kung, Morgan Cheatham, Arielle Medenilla, Czarina Sillos, Lorie De Leon, Camille Elepaño, Maria Madriaga, Rimel Aggabao, Giezel Diaz-Candido, James Maningo, Victor Tseng, and ChatGPT. Performance of chatgpt on usmle: Potential for ai-assisted medical education using large language models, 2022. URL <https://doi.org/10.1101/2022.12.19.22283643>.
- Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. Albert: A lite bert for self-supervised learning of language representations. *ArXiv*, abs/1909.11942, 2019.
- Eric P. Lehman, Sarthak Jain, Karl Pichotta, Yoav Goldberg, and Byron C. Wallace. Does bert pretrained on clinical notes reveal sensitive data? *ArXiv*, abs/2104.07762, 2021.
- Patrick Lewis, Myle Ott, Jingfei Du, and Veselin Stoyanov. Pretrained language models for biomedical and clinical tasks: Understanding and extending the state-of-the-art. In *Proceedings of the 3rd Clinical Natural Language Processing Workshop*, pages 146–157, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.clinicalnlp-1.17. URL <https://aclanthology.org/2020.clinicalnlp-1.17>.
- Xiang Lisa Li and Percy Liang. Prefix-tuning: Optimizing continuous prompts for generation. *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, abs/2101.00190, 2021.
- Yikuan Li, Ramsey M. Wehbe, Faraz S. Ahmad, Hanyin Wang, and Yuan Luo. Clinical-longformer and clinical-bigbird: Transformers for long clinical sequences. *ArXiv*, abs/2201.11838, 2022.
- Valentin Li’evin, Christoffer Egeberg Hother, and Ole Winther. Can large language models reason about medical questions? *ArXiv*, abs/2207.08143, 2022.
- Haokun Liu, Derek Tam, Mohammed Muqeeth, Jay Mohata, Tenghao Huang, Mohit Bansal, and Colin Raffel. Few-shot parameter-efficient fine-tuning is better and cheaper than in-context learning. *ArXiv*, abs/2205.05638, 2022.
- Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *ArXiv*, abs/1907.11692, 2019.
- Ilya Loshchilov and Frank Hutter. Fixing weight decay regularization in adam. *ArXiv*, abs/1711.05101, 2017.
- J. Mullenbach, Yada Pruksachatkun, Sean Adler, Jennifer M. Seale, Jordan Swartz, T. Greg McKelvey, Hui Dai, Yi Yang, and David A. Sontag. Clip: A dataset for extracting action items for physicians from hospital discharge notes. *ArXiv*, abs/2106.02524, 2021.
- Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke E. Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Francis Christiano, Jan Leike, and Ryan J. Lowe. Training language models to follow instructions with human feedback. *ArXiv*, abs/2203.02155, 2022.
- Anusri Pampari, Preethi Raghavan, Jennifer J. Liang, and Jian Peng. emrqa: A large corpus for question answering on electronic medical records. In *Conference on Empirical Methods in Natural Language Processing*, 2018.

- Giovanni Paolini, Ben Athiwaratkun, Jason Krone, Jie Ma, Alessandro Achille, Rishita Anubhai, Cicero Nogueira dos Santos, Bing Xiang, and Stefano Soatto. Structured prediction as translation between augmented natural languages. In *9th International Conference on Learning Representations, ICLR 2021*, 2021.
- Thomas H Payne, Aharon E tenBroek, Grant S Fletcher, and Mardi C Labuguen. Transition from paper to electronic inpatient physician notes. *J. Am. Med. Inform. Assoc.*, 17(1):108–111, January 2010.
- Long Phan, James T. Anibal, Hieu Tran, Shaurya Chanana, Erol Bahadroglu, Alec Peltekian, and Grégoire Altan-Bonnet. Scifive: a text-to-text transformer model for biomedical literature. *ArXiv*, abs/2106.03598, 2021.
- Alec Radford and Karthik Narasimhan. Improving language understanding by generative pre-training. 2018.
- Colin Raffel, Noam M. Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *ArXiv*, abs/1910.10683, 2020.
- Samyam Rajbhandari, Jeff Rasley, Olatunji Ruwase, and Yuxiong He. Zero: Memory optimizations toward training trillion parameter models, 2019. URL <https://arxiv.org/abs/1910.02054>.
- Alexey Romanov and Chaitanya Shivade. Lessons from natural language inference in the clinical domain. August 2018. URL <http://arxiv.org/abs/1808.06752>.
- Victor Sanh, Albert Webson, Colin Raffel, Stephen H. Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Teven Le Scao, Arun Raja, Manan Dey, M Saiful Bari, Canwen Xu, Urmish Thakker, Shanya Sharma Sharma, Eliza Szczechla, Taewoon Kim, Gunjan Chhablani, Nihal Nayak, Debajyoti Datta, Jonathan Chang, Mike Tian-Jian Jiang, Han Wang, Matteo Manica, Sheng Shen, Zheng Xin Yong, Harshit Pandey, Rachel Bawden, Thomas Wang, Trishala Neeraj, Jos Rozen, Abheesht Sharma, Andrea Santilli, Thibault Fevry, Jason Alan Fries, Ryan Teehan, Stella Biderman, Leo Gao, Tali Bers, Thomas Wolf, and Alexander M. Rush. Multitask prompted training enables zero-shot task generalization, 2021.
- K. Singhal, Shekoofeh Azizi, Tao Tu, Said Mahdavi, Jason Lee Kai Wei, Hyung Won Chung, Nathan Scales, Ajay Kumar Tanwani, Heather J. Cole-Lewis, Stephen J. Pfohl, P A Payne, Martin G. Seneviratne, Paul Gamble, Chris Kelly, Nathaneal Scharli, Aakanksha Chowdhery, P. D. Mansfield, Blaise Agueray Arcas, Dale R. Webster, Greg S. Corrado, Y. Matias, Katherine Hui-Ling Chou, Juraj Gottweis, Nenad Tomaev, Yun Liu, Alvin Rajkumar, Joelle K. Barral, Christopher Semturs, Alan Karthikesalingam, and Vivek Natarajan. Large language models encode clinical knowledge. *ArXiv*, abs/2212.13138, 2022.
- Sarvesh Soni, Meghana Gudala, Atieh Pajouhi, and Kirk Roberts. RadQA: A question answering dataset to improve comprehension of radiology reports. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 6250–6259, Marseille, France, June 2022. European Language Resources Association. URL <https://aclanthology.org/2022.lrec-1.672>.
- Amber Stubbs, Christopher Kotfila, and Özlem Uzuner. Automated systems for the de-identification of longitudinal clinical narratives: Overview of 2014 i2b2/UTHealth shared task track 1. *J. Biomed. Inform.*, 58 Suppl:S11–S19, December 2015.
- Weiyi Sun, Anna Rumshisky, and Ozlem Uzuner. Annotating temporal information in clinical narratives. *Journal of Biomedical Informatics*, 46: S5–S12, 2013. ISSN 1532-0464. doi: <https://doi.org/10.1016/j.jbi.2013.07.004>. URL <https://www.sciencedirect.com/science/article/pii/S1532046413001032>. Supplement: 2012 i2b2 NLP Challenge on Temporal Relations in Clinical Data.
- Ozlem Uzuner, Brett R South, Shuying Shen, and Scott L DuVall. 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text. *Journal of the American Medical Informatics Association*, 18(5): 552–556, 06 2011. ISSN 1067-5027. doi: 10.1136/amiajnl-2011-000203. URL <https://doi.org/10.1136/amiajnl-2011-000203>.
- Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V. Le. Finetuned language models are zero-shot learners. *ArXiv*, abs/2109.01652, 2021a.
- Jason Wei, Maarten Bosma, Vincent Y. Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai,

and Quoc V. Le. Finetuned language models are zero-shot learners, 2021b. URL <https://arxiv.org/abs/2109.01652>.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed Huai hsin Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. Emergent abilities of large language models. *ArXiv*, abs/2206.07682, 2022.

Qiang Wei, Zongcheng Ji, Yuqi Si, Jingcheng Du, Jingqi Wang, Firat Tiryaki, Stephen Wu, Cui Tao, Kirk Roberts, and Wang Qi. Relation extraction from clinical narratives using pre-trained language models. *AMIA ... Annual Symposium proceedings. AMIA Symposium*, 2019:1236–1245, 03 2020.

Xi Yang, Aokun Chen, Nima PourNejatian, Hoo Chang Shin, Kaleb E. Smith, Christopher Parisien, Colin Compas, Cheryl Martin, Anthony B. Costa, Mona G. Flores, and et al. A large language model for electronic health records. *npj Digital Medicine*, 5(1), 2022. doi: 10.1038/s41746-022-00742-2.

Appendix A. MIMIC Preprocessing and Model Training

In this section, we walk through the steps required to pre-train the T5 specialized clinical models.

A.1. Data Preprocessing

We use notes from both MIMIC-III & MIMIC-IV for pre-training. These datasets are not entirely disjoint, as a portion of the notes that appear in MIMIC-III also appear in MIMIC-IV. However, MIMIC-IV only contains discharge summaries and radiology reports. We take the union of MIMIC-III and MIMIC-IV notes such that patient records are not repeated (Table 4). This includes notes from all CAREVUE patients and all notes that are not discharge summaries or radiology reports. We also remove patients that overlap with the tasks we consider in this paper (except for MedNLI). This is important because it is unlikely that models will be pretrained on the same data used at inference time in a realistic deployment scenario.

We remove duplicates of notes from MIMIC-III using `charttime`, `storetime` and `cgid`. Duplicate notes can occur when clinicians draft and later edit a note; these duplicates generally differ by 1-2 words. After this preprocessing, there are 430M words in MIMIC-III (Table 4).

A.2. Tokenization of DEID Tokens

All data in MIMIC is fully de-identified. In MIMIC-III, protected health information (PHI) is replaced with special deidentification tags (e.g. `[**First Name 123**]`), and in MIMIC-IV PHI is replaced with the generic placeholder `----`. While these de-identification tags can be informative, tokenizers typically break each tag into multiple subwords, dramatically increasing the number of tokens. We find that replacing all DEID tags with several special DEID tokens (e.g., `[NAME]`), which we add to the tokenizer vocabulary, reduces the size of MIMIC from 2,400,714,781 tokens to 2,335,573,220 tokens. To perform this replacement on MIMIC-IV, we were granted special access to a file that maps PHI locations to the type of PHI it is. Using this mapping, we add the appropriate DEID tokens to MIMIC-IV text so that the DEID information is stored in a similar manner across both datasets.

We experimented with 3 different tokenization methods prior to pretraining our specialized clinical models. To select the best tokenizer, we pretrained 3 different models for 10 epochs initializing from T5-Base. In the first model, which we use in the paper, we add special DEID tokens and replace the existing ones in MIMIC. For the second model, we do not modify the tokenizer at all. In the last model, we replace all DEID tags with realistic PHI. We frame the problem as a masked language modeling task and query a T5-Large model to generate realistic PHI (e.g. patient names, hospital names, etc.). We evaluated each model on the n2c2 2012 challenge (Sun et al., 2013), and we found that the performance of these models was comparable. Using the evaluation script provided by Paolini et al. (2021), we found that n2c2 2012 scores were 0.800, 0.803, 0.802, for the first, second, and third model, respectively. These models can be made available upon request.

A.3. Model Pretraining

We train and test three different T5 models, following the original T5 training pretraining scheme where possible. We describe the process for training each below.

1. Clinical-T5-Base: We pretrain the model from scratch on MIMIC notes for 310K steps, which is roughly 40B tokens worth of pretraining. The model was trained for 200K steps on a TPU before an error with the TPU caused us to switch training to a GPU cluster. The batch size was 32 per TPU/GPU. Due to an issue in the code, the model uses a lowercased vocabulary. All other models are cased.

Name	# Patients	#Notes	#Words
MIMIC-III	46K	2M	429M
MIMIC-IV	246K	2.6M	921M
MIMIC-III + MIMIC-IV	291K	4.1M	1.2B

Table 4: We break down the MIMIC-III and MIMIC-IV datasets. There is an overlap in notes between MIMIC-III & MIMIC-IV.

2. **Clinical-T5-Base-Ckpt:** We initialize the model with T5-Base and trained the model for an additional 100K steps on the MIMIC notes. The model was trained on 8xA6000 (48GB) GPUs with a batch-size of 32 per GPU. Each epoch took roughly 6 hours. We used 40K warm-up steps (compared to 10K in the original T5 paper) because we were training the model on a fewer number of tokens. We suspect that this was too many warm-up steps and may have negatively impacted performance.
3. **Clinical-T5-Large:** We train this model from scratch on MIMIC notes for 780K steps or approximately 38B tokens. We use a TPU v3.8 cluster with a batch size of 12 per TPU. The cost of training was approximately 1,800 USD, and the training process took approximately 220 hours.

Appendix B. Detailed Model Training and Performance

In the following section, we describe our process for finetuning language models on MedNLI, RadQA, and CLIP. Due to space limitations, we only show results for 12 models in the main body of the paper. However, in this expanded appendix, we report the performance of 16 different general, biomedical, and clinical language models, adding results for ClinicalBERT (Alsentzer et al., 2019), ClinicalLongformer (Li et al., 2022), SciFive (Phan et al., 2021), and SciFive Large. All of these models were trained use DAPT. ClinicalBERT was initialized from BioBERT and further pretrained over MIMIC-III. Similarly, ClinicalLongformer was initialized from the Longformer (?) and trained over MIMIC-III. Lastly, SciFive and SciFive-Large were initialized from T5-Base and T5-Large, respectively, and trained over PubMed.

B.1. Hyperparameter Tuning

We largely follow the guidance of Raffel et al. (2020) for finetuning all of the T5 models. Raffel et al. (2020) suggest using a constant learning rate of $1e-3$ for all finetuning experiments (with adafactor optimizer). We found that

this was too large and that $1e-4$ performed significantly better across all tasks.

For PubMedGPT, we follow Bolton et al. (2022) and train using AdamW with a learning rate of $2e-6$. We experimented with $2e-5$, but found that $2e-6$ performed much better. For ClinicalBERT, GatorTron, and Clinical-Longformer, we do a hyperparameter search over learning rates of $2e-5$, $3e-5$ and $5e-5$. For RoBERTa and BioClin-RoBERTa, we follow the guidance of Lewis et al. (2020), and use a learning rate of $1e-5$. We select whichever learning rate works best on the validation set. The optimal learning rate varies for each task. We use the AdamW optimizer (Loshchilov and Hutter, 2017).

To train T5-XL and PubMedGPT with limited GPU resources, we leverage the DeepSpeed library (Rajbhandari et al., 2019). This enables the models to be trained on 32GB GPUs by using CPU offloading at the expense of increasing train run time.

We train until convergence for all tasks. The time to convergence differs across tasks. Generally, we find that T5-XL converges much faster than the other T5 models. On MedNLI, for example, T5-XL converges within 15 epochs whereas Clinical-T5-Large needs roughly 30-40 epochs to converge. We ran all experiments with an effective batch size of 64. We select the optimal hyperparameters according to the performance on the validation set for each task (accuracy for MedNLI, F1 for RadQA, and Macro F1 for CLIP).

B.2. Computational Resources and Run-Time

We used a wide-range of GPUs for our experiments, including 80GB V100s, 48GB A6000, 32GB V100, and 12GB 2080Tis. The encoder-only models take around 20-40 minutes to run on MedNLI and RadQA and 3 hours to run on CLIP. We find that the T5-Base models take around an hour to run on MedNLI and RadQA and 4 hours on CLIP (these models are trained for additional epochs compared to the encoder-only models because they are slower to converge). The T5-Large models take around 1.5 hours to run on MedNLI and RadQA and roughly 10 hours to run on CLIP. PubMedGPT and T5-XL take around 6 hours to run on MedNLI and RadQA. For CLIP, this took roughly 40 hours to run (on 4x48GB GPUs). The

Model	Size	General PTT	BioMed PTT	Clinical PTT	Unique PTT
ClinicalBERT	110M	137B	46B	0.6B	3.4B / 32B / 0.6B
Clinical LongFormer	150M	2200B	–	15B	55B / – / 0.8B
T5-Base	220M	34B	0.5B	–	34B / 0.5B / –
Clinical-T5-Base-Ckpt	220M	34B	0.5B	13B	34B / 0.5B / 2.3B
Clinical-T5-Base	220M	–	–	40B	– / – / 2B
RoBERTa-Large	345M	2200B	–	–	55B / – / –
BioClinRoBERTa	345M	–	2037B	65B	– / 32B / 0.8B
GatorTron	345M	40B	92B	1570B	4B / 9B / 157B
T5-Large	770M	34B	0.5B	–	34B / 0.5B / –
Clinical-T5-Large	770M	–	–	38B	– / – / 2B
SciFive	220M	34B	27B	–	34B / 27B / –
SciFive-Large	770M	34B	14B	–	34B / 14B / –
PubMedGPT	2.7B	–	300B	–	– / 50B / –
T5-XL	3B	34B	0.5B	–	34B / 0.5B / –
Flan-T5-XXL	11B	34B	0.5B	–	34B / 0.5B / –
GPT-3	175B	–	–	–	–

Table 5: PTT stands for pretraining tokens. All of the models tested and considered for the project. We show the models, their size, what they were initialized from, and the make up of their pretraining data. We are, of course, unable to provide any information on GPT-3. We focus only on pretraining data, and ignore any instruction tuning data.

use of the DeepSpeed library increased the time required for finetuning PubMedGPT and T5-XL.

B.3. Task-Specific Details

We produce answers with the T5 models by generating the label or extracted text with beam search. For the encoder-only models and PubMedGPT, we add a task-specific linear layer on top of the base model. We next outline finetuning details that are specific to each task.

MedNLI. We train the encoder-only models and PubMedGPT for 20 epochs, and we train T5-XL for 15 epochs. All clinical and general-domain T5-Base and T5-Large models are trained for 40 epochs. For all T5 models, we use a beam search width of 3.

RadQA. As before, we train the encoder-only models and PubMedGPT for 20 epochs, and we train T5-XL for 15 epochs. We trained all T5-Base and T5-Large models for 50 epochs. For all T5 models, we use a beam search width of 1. We found that increasing the beam-search width did not consistently improve performance; we experimented with beam search widths of 3, 5, and 10, and found that it increased exact-match at the expense of F1-Score.

CLIP. Again, we train the encoder-only models and PubMedGPT for 20 epochs, and we train T5-XL for 15 epochs. We trained all T5-Base and T5-Large models

for 40 epochs. For all T5 models, we use a beam search width of 5. We did not experiment with different beam search widths for CLIP. To generate multiple labels for each sentence, we ask the T5 models to produce a comma-delimited list of labels, ordered alphabetically. We use a context window of 256 for all experiments with CLIP. This resulted in a slightly lower performance compared to the results presented in [Mullenbach et al. \(2021\)](#), which used a window of 512 tokens.

Appendix C. Additional Discussion of Model Performance

C.1. MedNLI

We report results for all models in Table 7. We find that ClinicalBERT performs similarly to T5-Base, while ClinicalLongFormer performs similarly to T5-Large. We additionally test SciFive and SciFive-Large ([Phan et al., 2021](#)), which outperform T5-Base and T5-Large, respectively. However, these models fail to outperform Clinical-T5-Base and Clinical-T5-Large. This may be because SciFive and SciFive-Large are trained via DAPT, while Clinical-T5-Base and Clinical-T5-Large are trained from scratch. Further, SciFive and SciFive-Large are trained on biomedical tokens, rather than clinical tokens.

Task	Type	Labels	Max Sequence Length	Train / Val / Test	Units
MedNLI	NLI	3	256	11K / 1K / 1K	Sentence Pairs
RadQA	QA	–	1024	4.8K / 1K / 1K	Question + Answer Pairs
CLIP	CLS	7	256	107K / 10K / 10K	Sentences

Table 6: We summarize some task statistics. CLS stands for classification.

Model	Size	BioMed PT	Clinical PT	Accuracy	Std.
ClinicalBERT	110M	✗	✓	0.815	0.008
ClinicalLongFormer	150M	✗	✓	0.846	0.003
T5-Base	220M	✗	✗	0.818	0.006
SciFive	220M	✗	✗	0.835	0.003
Clinical-T5-Base-Ckpt	220M	✗	✓	0.852	0.007
Clinical-T5-Base	220M	✗	✓	0.855	0.004
GatorTron	345M	✓	✓	0.883	0.002
RoBERTa	345M	✗	✗	0.852	0.002
BioClinical RoBERTa	345M	✓	✓	0.900	0.003
T5-Large	770M	✗	✗	0.849	0.008
SciFive Large	770M	✓	✗	0.857	0.005
Clinical-T5-Large	770M	✗	✓	0.872	0.008
PubmedGPT	2.7B	✓	✗	0.870	0.009
T5-XL	3B	✗	✗	0.869	0.004
Flan-T5-XL	11B	✗	✗	0.808	–
GPT-3	175B	–	–	0.807	–

Table 7: We show the performance of all models considered on MedNLI. Results are based on at least 3 seeds.

We also show how performance changes depending on the number of DAPT steps (Table 8). We find that training Clinical-T5-Base-Ckpt for 20K pretraining steps gives a reasonable boost in performance over T5-Base. Training from 20K to 80K steps does not seem to provide any additional performance gains. However, we find that training for 100K steps does improve performance versus training for 80K steps. This is likely due to the learning rate scheduler. It is possible that at 40K to 80K steps, the learning rate is too large.

C.2. RadQA

We report results for all models in Table 9. We find that ClinicalBERT performs extremely poorly on RadQA, while the ClinicalLongformer performs similar to Clinical-T5-Base-Ckpt. Similar to MedNLI, SciFive and SciFive-Large outperform T5-Base and T5-Large, respectively. However, both of these models fail to outperform their clinical equivalents.

C.3. CLIP

We report results for all models in Table 10. We find that ClinicalBERT and ClinicalLongformer perform very well on this task, performing comparably to or outperforming the much larger T5-XL model. This is likely due to the fact that the T5 models *generate* answers, which is challenging for a multi-label classification task. As we saw in other experiments, SciFive and SciFive-Large underperform their clinical-domain counterparts. PubMedGPT has the highest Micro F1 performance, outperforming both GatorTron and BioClinRoBERTa, which excelled across all other tasks.

Appendix D. Additional Details about In Context Learning Experiments

In this section, we provide additional information about our approach for performing in context learning with GPT-3 and Flan-T5-XXL.

We experiment with approximately 10-20 different prompts for each task, crafting prompts to reflect the prompts used during instruction tuning of Flan-T5 and GPT-3. We pair each prompt with one to three randomly sampled examples for in-context learning. We select the best prompt based on the performance on a random sample of 200 examples from the validation set. We use a temperature of 0 and a beam search width of 1.

For MedNLI and CLIP, we found that the performance of the models was largely stable, regardless of the input

prompt. For example, GPT-3 was achieving 70%+ accuracy on all of the MedNLI prompts given to it. For RadQA, there was more variation in its performance, as the model needed to be “pushed” towards correctly identifying unanswerable questions.

There are two options for generating labels for CLIP, which is a multi-label classification task. The model can either generate predictions for each label independently or all at once. We experiment with both options using Flan-T5-XXL and find that both approaches perform similarly. However, independently prompting the model for each label results in higher inference time costs. Therefore, we ask the model to generate predictions for all labels at once for GPT-3.

We list the prompts that were used on the test set below. Note that we only include the prompt itself and do not include the in-context examples.

- MedNLI - T5-Flan-XXL & GPT-3: Answer entailment, contradiction or neutral. Premise: {Premise} Hypothesis: {Hypothesis}
- RadQA - GPT-3 & GPT-3: Context: {Context}, {Question} Answer N/A if there is no answer or give a quote from the context:
- CLIP - T5-Flan-XXL:
 1. Context: {Context}. Does the above sentence contain information about current or future appointments? Options: -Yes -No
 2. Context: {Context}. Does the above sentence contain information about medications? Options: -Yes -No
 3. Context: {Context}. Does the above sentence contain any important actionable information? Options: -Yes -No
 4. Context: {Context}. Does the above sentence contain any information about laboratory tests? Options: -Yes -No
 5. Context: {Context}. Does the above sentence contain any information about what to do

Model	Clinical PTT	Accuracy	Std.
T5-Base	–	0.818	0.006
Clinical-T5-Base-Ckpt-20K	2B	0.831	0.001
Clinical-T5-Base-Ckpt-40K	5B	0.831	0.002
Clinical-T5-Base-Ckpt-60K	8B	0.836	0.007
Clinical-T5-Base-Ckpt-80K	10B	0.836	0.002
Clinical-T5-Base-Ckpt	13B	0.852	0.007

Table 8: We report the performance of Clinical-T5-Base-Ckpt on MedNLI when trained on an increasing number of tokens from MIMIC. We find that pretraining for a high warmup initially boosts performance by 1%.

Model	Size	BioMed PT	Clinical PT	Exact Match	F1
ClinicalBERT	110M	✗	✓	0.457 ± 0.002	0.626 ± 0.008
ClinicalLongformer	150M	✗	✓	0.518 ± 0.036	0.689 ± 0.018
T5-Base	220M	✗	✗	0.479 ± 0.014	0.662 ± 0.010
SciFive	220M	✓	✓	0.506 ± 0.010	0.697 ± 0.007
Clinical-T5-Base-Ckpt	220M	✗	✓	0.505 ± 0.014	0.684 ± 0.009
Clinical-T5-Base	220M	✗	✓	0.531 ± 0.013	0.710 ± 0.005
RoBERTa	345M	✗	✗	0.521 ± 0.014	0.684 ± 0.004
BioClinical RoBERTa	345M	✗	✗	0.604 ± 0.012	0.759 ± 0.029
GatorTron	345M	✓	✓	0.583 ± 0.008	0.759 ± 0.008
T5-Large	770M	✗	✗	0.537 ± 0.019	0.700 ± 0.012
SciFive-Large	770M	✓	✗	0.541 ± 0.016	0.704 ± 0.013
Clinical-T5-Large	770M	✗	✓	0.550 ± 0.018	0.745 ± 0.008
PubMedGPT	2.7B	✓	✗	0.512 ± 0.005	0.698 ± 0.004
T5-XL	3B	✗	✗	0.568 ± 0.007	0.729 ± 0.005
Flan-T5-XXL	11B	✗	✗	0.300	0.602
GPT-3	175B	✗	✗	0.362	0.620

Table 9: Performance of all models on RadQA. We report the mean performance and standard deviation of models trained with at least 3 random seeds.

Model	Size	BioMed PT	Clinical PT	Micro F1	Macro F1
ClinicalBERT	110M	✗	✓	0.777 ± 0.006	0.649 ± 0.007
ClinicalLongformer	150M	✗	✓	0.790 ± 0.003	0.659 ± 0.008
T5-Base	220M	✗	✗	0.767 ± 0.008	0.594 ± 0.011
SciFive	220M	✓	✓	0.769 ± 0.008	0.603 ± 0.004
Clinical-T5-Base-Ckpt	220M	✗	✓	0.772 ± 0.005	0.605 ± 0.009
Clinical-T5-Base	220M	✗	✓	0.793 ± 0.001	0.652 ± 0.009
RoBERTa	345M	✓	✗	0.793 ± 0.001	0.677 ± 0.008
BioClinRoBERTa	345M	✓	✗	0.805 ± 0.005	0.707 ± 0.007
GatorTron	345M	✓	✗	0.791 ± 0.003	0.690 ± 0.010
T5-Large	770M	✗	✗	0.779 ± 0.008	0.629 ± 0.011
SciFive-Large	770M	✓	✗	0.774 ± 0.008	0.630 ± 0.011
Clinical-T5-Large	770M	✗	✓	0.800 ± 0.008	0.663 ± 0.007
PubMedGPT	2.7B	✓	✗	0.819 ± 0.003	0.666 ± 0.003
T5-XL	3B	✗	✗	0.780 ± 0.021	0.640 ± 0.022
Flan-T5-XXL	11B	✗	✗	0.164	0.178
GPT-3	175B	✗	✗	0.154	0.146

Table 10: Performance of all models on CLIP. We report the mean performance and standard deviation of models trained with at least 3 random seeds. T5-Flan-XXL and GPT-3 are based on a sample of 25% of the test data.

post-discharge? Options: -Yes -No

6. Context: {Context}. Does the above sentence contain any information about procedures (e.g., surgeries)? Options: -Yes -No

7. Context: {Context}. Does the above sentence contain any information about an imaging followup? Options: -Yes -No

- CLIP - GPT-3: Context: {Context}. Label the above sentence as one or more of the following, delimited by comma: Options: -Appointment-related followup information -Medication-related followup information -Lab-related followup information -Case-specific instructions for the patient -Procedure-related followup information -Imaging-related followup information -None of the above

also report Exact Match on RadQA in Table 13 and Micro F1 on CLIP in Table 14. We initially experimented with GPT-Neo-X (Black et al., 2022) in addition to GPT-3 and T5-Flan-XXL. However, in our initial experiments, we found that its performance on MedNLI was less than 40%. Therefore, we dropped it from our remaining experiments.

We will make all of our prompts available, along with their validation set performance scores. Consistent with prior literature, we find that the performance of these models is extremely dependent on the prompt (Chung et al., 2022). For example, when evaluating Flan-T5-XXL on MedNLI, we find that using the following prompt leads to a drop in accuracy from 83.5% to 62% on the validation set: Answer entailment, neutral or contradiction. Premise: Premise Hypothesis: Hypothesis. Answer: '.

Post-processing was required to map the text generated by GPT-3 and Flan-T5-XXL to the label space. For MedNLI, we check if the string contains the word entailment, contradiction or neutral. If none of these three words appear, we predict neutral. For CLIP, we search the generated string for the label types. This allows for the models to generate predictions in any order. GPT-3 and Flan-T5-XXL sometimes produce answers to RadQA questions that cannot be extracted directly from the radiology report. In such cases, we calculate F1-score regardless. Had we enforced that the model produce a string directly from the text, the F1-score would have dropped to ~ 40 for both models.

Finally, we report the exact performance metrics shown in Figure 4 in Table 11, Table 12 and Table 15. We

Model	1%	5%	10%	25%	100%
PubMedGPT	0.597 +/- 0.011	0.717 +/- 0.011	0.807 +/- 0.011	0.845 +/- 0.006	0.870 +/- 0.009
GatorTron	0.811 +/- 0.001	0.817 +/- 0.005	0.837 +/- 0.023	0.858 +/- 0.001	0.883 +/- 0.002
RoBERTa	0.718 +/- 0.008	0.759 +/- 0.010	0.786 +/- 0.008	0.809 +/- 0.004	0.852 +/- 0.002
BioClinRoBERTa	0.824 +/- 0.025	0.852 +/- 0.004	0.862 +/- 0.004	0.882 +/- 0.006	0.900 +/- 0.003
Clinical-T5-Large	0.581 +/- 0.029	0.742 +/- 0.033	0.801 +/- 0.003	0.838 +/- 0.007	0.872 +/- 0.008

Table 11: Accuracy on MedNLI for models finetuned with varying amounts of annotated data. Percentages refer to fraction of the training set for the task. We report the mean and standard deviation over three random seeds. We always evaluate on the full test set.

Model	1% (F1)	5% (F1)	10% (F1)	25% (F1)	100% (F1)
PubMedGPT	0.291 +/- 0.017	0.461 +/- 0.002	0.564 +/- 0.012	0.672 +/- 0.014	0.729 +/- 0.005
GatorTron	0.315 +/- 0.027	0.620 +/- 0.011	0.666 +/- 0.001	0.718 +/- 0.008	0.759 +/- 0.008
RoBERTa	0.202 +/- 0.014	0.355 +/- 0.015	0.544 +/- 0.006	0.613 +/- 0.008	0.684 +/- 0.004
BioClinRoBERTa	0.369 +/- 0.001	0.370 +/- 0.011	0.619 +/- 0.021	0.717 +/- 0.011	0.759 +/- 0.029
Clinical-T5-Large	0.284 +/- 0.024	0.541 +/- 0.027	0.600 +/- 0.021	0.679 +/- 0.012	0.745 +/- 0.008

Table 12: F1 score on RadQA for models finetuned with varying amounts of annotated data. Percentages refer to fraction of the training set for the task. We report the mean and standard deviation over three random seeds. We always evaluate on the full test set.

Model	1% (EM)	5% (EM)	10% (EM)	25% (EM)	100% (EM)
PubMedGPT	0.231 +/- 0.004	0.332 +/- 0.012	0.362 +/- 0.009	0.476 +/- 0.013	0.512 +/- 0.005
GatorTron	0.263 +/- 0.022	0.482 +/- 0.010	0.507 +/- 0.004	0.554 +/- 0.012	0.583 +/- 0.008
RoBERTa	0.187 +/- 0.021	0.295 +/- 0.004	0.415 +/- 0.009	0.462 +/- 0.009	0.521 +/- 0.014
BioClinRoBERTa	0.322 +/- 0.009	0.322 +/- 0.009	0.479 +/- 0.016	0.561 +/- 0.019	0.604 +/- 0.012
Clinical-T5-Large	0.206 +/- 0.015	0.358 +/- 0.016	0.435 +/- 0.024	0.495 +/- 0.006	0.550 +/- 0.018

Table 13: Exact Match performance on RadQA for models finetuned with varying amounts of annotated data. Percentages refer to fraction of the training set for the task. We report the mean and standard deviation over three random seeds. We always evaluate on the full test set.

Model	1% (Micro)	5% (Micro)	10% (Micro)	25% (Micro)	100% (Micro)
PubMedGPT	0.580 +/- 0.006	0.706 +/- 0.010	0.740 +/- 0.006	0.789 +/- 0.003	0.819 +/- 0.003
GatorTron	0.686 +/- 0.010	0.725 +/- 0.009	0.759 +/- 0.006	0.785 +/- 0.002	0.793 +/- 0.001
RoBERTa	0.703 +/- 0.014	0.726 +/- 0.002	0.739 +/- 0.001	0.768 +/- 0.006	0.791 +/- 0.003
BioClinRoBERTa	0.692 +/- 0.007	0.714 +/- 0.003	0.739 +/- 0.003	0.770 +/- 0.001	0.805 +/- 0.005
Clinical-T5-Large	0.616 +/- 0.004	0.716 +/- 0.016	0.743 +/- 0.013	0.777 +/- 0.000	0.800 +/- 0.008

Table 14: Micro F1 score on CLIP for models finetuned with varying amounts of annotated data. Percentages refer to fraction of the training set for the task. We report the mean and standard deviation over three random seeds. We always evaluate on the full test set.

Model	1% (Macro)	5% (Macro)	10% (Macro)	25% (Macro)	100% (Macro)
PubMedGPT	0.203 +/- 0.010	0.332 +/- 0.014	0.426 +/- 0.001	0.585 +/- 0.020	0.666 +/- 0.003
GatorTron	0.296 +/- 0.006	0.317 +/- 0.007	0.407 +/- 0.015	0.588 +/- 0.014	0.677 +/- 0.008
RoBERTa	0.388 +/- 0.014	0.404 +/- 0.003	0.520 +/- 0.043	0.658 +/- 0.007	0.690 +/- 0.010
BioClinRoBERTa	0.310 +/- 0.004	0.417 +/- 0.015	0.524 +/- 0.018	0.648 +/- 0.006	0.707 +/- 0.007
Clinical-T5-Large	0.356 +/- 0.007	0.465 +/- 0.047	0.548 +/- 0.012	0.620 +/- 0.008	0.663 +/- 0.007

Table 15: Macro F1 score on CLIP for models finetuned with varying amounts of annotated data. Percentages refer to fraction of the training set for the task. We report the mean and standard deviation over three random seeds. We always evaluate on the full test set.