Are Clinical T5 Models Better for Clinical Text?

Yahan Li* YAHANLI@USC.EDU
Department of Computer Science, University of Southern California

Keith Harrigian* KHARRIGIAN@JHU.EDU
Department of Computer Science, Johns Hopkins University

Ayah Zirikly AZIRIKLY@JHU.EDU
Center for Language and Speech Processing, Whiting School of Engineering, Johns Hopkins University
Malone Center for Engineering in Healthcare, Johns Hopkins University

Mark Dredze MDREDZE@CS.JHU.EDU
Department of Computer Science, Johns Hopkins University

Abstract

Large language models with a transformer-based encoder/decoder architecture, such as T5 (Raffel et al., 2023), have become standard platforms for supervised tasks. To bring these technologies to the clinical domain, recent work has trained new (Lehman et al., 2023) or adapted existing (Lu et al., 2022) models to clinical data. However, the evaluation of these clinical T5 models and comparison to other models has been limited. Are the clinical T5 models better choices than FLAN-tuned (Chung et al., 2022a) generic T5 models? Do they generalize better to new clinical domains that differ from the training sets? We comprehensively evaluate these models across several clinical tasks and domains. We find that clinical T5 models provide marginal improvements over existing models, and perform worse when evaluated on different domains. Our results inform future choices in developing clinical LLMs.

Keywords: Language Models, Domain Generalization, Pre-training

Data and Code Availability We do not release any new data or models as part of this study. The code in available on GitHub: https://github.com/yli-z/ml4h_are_clinical_t5_models_better_for_clinical_text.

Institutional Review Board (IRB) Our clinical datasets were obtained from the PhysioNet website under the PhysioNet Credentialed Health Data Use Agreement and from an anonymized Hospital System following IRB approval.

* These authors contributed equally

1. Introduction

There is an ongoing conversation in the community about the best strategy for developing Large Language Models (LLMs) for specialized domains. general-purpose LLMs, trained on massive amounts of diverse text, can generalize remarkably well to new domains and tasks (Brown et al., 2020; Touvron et al., 2023; Bubeck et al., 2023), especially with a small amount of domain-specific alignment data (Singhal et al., 2022). At the same time, specialized LLMs trained using domain-specific data can outperform their generic counterparts (Wu et al., 2023b; Taylor et al., 2022; Lehman et al., 2023; Lu et al., 2022). For supervised tasks, Text-to-Text Transfer Transformer (T5) models are a popular choice as they provide the benefits of both extensive pre-training and task generalization(Raffel et al., 2023); however, there are few studies on the benefits of domain-specialized T5 training.

Consider clinical text from Electronic Health Records (EHR). Access to this data is severely restricted and the data differs substantially from popular pre-training data sets. Recent work has explored developing T5 models for this domain. One approach adapts an existing T5 model first to the biomedical domain (journal articles) and then to the clinical domain using a small amount of clinical text (Lu et al., 2022). Another approach starts from scratch, developing a new vocabulary and model focused on clinical data (Lehman et al., 2023). As these works were completed concurrently, they do not include a direct comparison between the approaches, leaving the best strategy for T5 models an open question.

Furthermore, while a specialized clinical T5 model sounds attractive, it has several potential shortcomings. First, the severe limitation of available EHR pre-training text means specialized T5 models are not only trained on less data overall, but also that the pre-training text is unlikely to adequately represent the true diversity of the domain. In contrast, pre-trained T5 models benefit from massive amounts of diverse data, which may yield better generalization abilities. Second, existing general T5 models may make use of extensive supervised datasets for FLAN tuning (Chung et al., 2022a), potentially leading to better generalization to new supervised tasks. Specialized clinical T5 models do not have the same opportunity nor benefit. Does FLAN training mean better generalization, even without access to EHR pre-training?

We thereby present a series of evaluations of existing general-purpose and clinical T5 models to answer two questions – 1) What existing T5 model should clinical NLP practitioners utilize when training a supervised system, and 2) What training strategy is more effective in the clinical domain with severe data limitations nowadays?

We make the following contributions.

- We evaluate 2 clinical, 2 non-clinical, and 1 FLAN-tuned T5 variants across 7 clinical and biomedical tasks. We independently reproduce prior results (Lu et al., 2022; Lehman et al., 2023) and extend the evaluation to new clinical datasets. We find that general T5 models achieve similar results as the clinical variants.

- We investigate model performance under data-limited conditions and find that FLAN-T5 excels on clinical tasks in the low-data settings.

- We evaluate clinical T5 models on new clinical domains and show they perform worse than general T5 models, suggesting that existing clinical models are overfitted to the limited clinical datasets available.

Taken together with other studies on clinical LLMs (Alsentzer et al., 2019; Lewis et al., 2020a; Singhal et al., 2022; Lu et al., 2022; Lehman et al., 2023), our findings suggest that future work should focus on adapting general T5 models to the clinical domain through new supervised training sets rather than training new methods from scratch.

2. Clinical T5 Models

LLMs have become the standard base models upon which NLP systems are built. Instruction following tasks typically leverage general-purpose decode LLMs (e.g., GPT, LlaMa) (Radford et al., 2019; Brown et al., 2020; Touvron et al., 2023; Bubeck et al., 2023), whereas supervised tasks typically rely on encoder/decoder (Raffel et al., 2023) or encoder models (Devlin et al., 2019).

Historically, clinical NLP systems have benefited from specialized versions of these models due to differences in the topic, style, and vocabulary of clinical text relative to other language domains. However, these advantages have been tempered under various circumstances (Gutiérrez et al., 2023; Lewis et al., 2020a), and are not necessarily guaranteed to hold given the existence of much larger contemporary LLMs. This uncertainty raises several key questions for clinical NLP practitioners.

What existing models should be used to develop a supervised clinical NLP system?

T5 (Raffel et al., 2023) and Flan-T5 (Chung et al., 2022a) do exceedingly well on a wide variety of supervised tasks, benefited from large amounts of pre-training and diverse task-specific fine-tuning, respectively. These models balance out-of-the-box zero-shot performance and supervised fine-tuning when a moderate amount of supervised data is available. Within the clinical domain, two T5-based models (Lehman et al., 2023; Lu et al., 2022) have been trained on clinical text from MIMIC (Johnson et al., 2016, 2023). Which model should we prefer and in what setting? Unfortunately, the publication of these two models was concurrent, so a head-to-head comparison is unavailable. Therefore, we conduct a direct comparison with the same training and evaluation settings to determine which model should be preferred.

Should we train from scratch on clinical data or adapt an existing pre-trained model?

Beyond determining which model is preferred, these two models represent different strategies for developing domain-specific models:

Lu et al. (2022) utilize continued pre-training to tune an existing model for the clinical setting, which allows them to benefit from large amounts of previous training on general-purpose and related data. They use MIMIC-III (Johnson et al., 2016) to train SciFive

(Phan et al., 2021), which is itself adapted to the scientific domain from the base T5 model (Raffel et al., 2023). In contrast, Lehman et al. (2023) train a new T5 model from scratch on clinical text from MIMIC-III and MIMIC-IV (Johnson et al., 2016, 2023), which enables them to leverage a domain-specific vocabulary.

These models follow larger trends in developing domain-specific models, including training from scratch (Wu et al., 2023b; Taylor et al., 2022; Lewis et al., 2020a) and continued pre-training or aligning existing models (Alsentzer et al., 2019; Singhal et al., 2022). In the clinical domain specifically, encoder (masked) LLMs have been adapted from pre-trained models (e.g., BioBERT (Lee et al., 2019) and Clinical-BioBERT (Alsentzer et al., 2019) from BERT (Devlin et al., 2019)) and have also been trained from scratch (e.g., GatorTron was pre-trained on a combination of Wikipedia articles, PubMed publications, and de-identified clinical notes (Yang et al., 2022)). Decoder (GPT-style) LLMs have been adapted to the medical domain, but not necessarily the clinical domain (e.g., Med-PaLm (Singhal et al., 2022) from FLAN-PaLm (Chowdhery et al., 2022)).

Arguments have been made in support of both pre-training strategies. We are interested in understanding whether one strategy is preferable over another when a) there is a need to generalize to a new clinical domain, and b) when only a limited amount of supervised data in the target domain is available.

Should we prefer domain-specialized models over powerful, general-purpose models?

Domain-specificity might not be optimal: biomedical language models have been shown to be robust to domain-general tokenization (Gutiérrez et al., 2023), and pre-trained LLMs have been shown to possess significant knowledge about the medical domain despite not having been trained specifically on biomedical data (Agrawal et al., 2022; Singhal et al., 2023). In the clinical domain, pre-trained clinical BERT models do not consistently demonstrate a significant improvement over non-clinical variants of BERT (Lewis et al., 2020b; Harrigian et al., 2023a; Yue et al., 2020). This suggests that the benefits of in-domain pre-training may not always justify the additional complexity and computational cost.

Morover, in the case of T5 models, there exist high-quality FLAN-tuned versions that do well on a variety of supervised tasks. While they are not clinically tuned, they do incorporate a lot of supervision. Should we prefer these general-purpose models to clinical models, especially in new clinical settings or those with limited supervision? We compare these models to clinical T5 models to determine which should be preferred in clinical settings.

Overall, our analysis provides practical guidance to practitioners, provides new evidence on LLM domain specialization methods, and explores the tradeoff between domain specialized and general-purpose LLMs for supervised tasks.

3. Experiment Setup

3.1. Models

Text-To-Text Transfer Transformer (T5) (Raffel et al., 2023) and its variants are sequence-to-sequence models utilizing an encoder/decoder Transformer architecture (Vaswani et al., 2017). We evaluate several existing pre-trained models, all being the large variant (∼770M parameters), by fine-tuning them on (x, y) pairs, where x and y are the input and output of a task, formatted as text.

Additional information about each model is included in Table 4.

T5 The original T5 model was trained in two stages. First, it was trained with a self-supervised text denoising objective on web crawl data (Raffel et al., 2023), and then with a conditional generation objective on labeled data from a diverse mixture of downstream tasks. To separate the effects of pre-training using non-clinical web data and performing supervised multitask pre-training, we evaluate two versions of T5: **T5-Den** which is trained only using the denoising objective and **T5-Sup** which is further pre-trained on supervised multi-task tasks.[1]

FLAN-T5 FLAN-T5 (Chung et al., 2022a) took the LM-Adapted T5, which was trained for additional 100K steps on the LM objective from T5-Den (Raffel et al., 2023; Lester et al., 2021), and further trained with extensive supervised instruction fine-tuning. The instruction tuning process typically improves the model's generalization ability to new tasks (Longpre et al., 2023). We include FLAN-T5 to evaluate its generalization capabilities on clinical tasks.

1. Most studies only evaluate T5-Sup and call it a pre-trained only model.

Clinical-T5 We evaluate two concurrently developed Clinical-T5 models – similarly named, but pre-trained on different corpora and with different weight initialization strategies. We refer to the model introduced by Lehman et al. (2023) as MIMIC-T5 since their model is only pre-trained on the union of all notes from MIMIC-III and MIMIC-IV (Johnson et al., 2016, 2023), including discharge summaries and radiology reports, from random initialization. We refer to the model introduced by Lu et al. (2022) as SciFive+MIMIC-T5 since their model is trained on MIMIC-III (Johnson et al., 2016) but initializes from a biomedically pre-trained model (SciFive), which gives it additional exposure to biomedical data. SciFive was itself trained from the base T5 initialization.

3.2. Datasets

We evaluate models on the union of datasets considered by each of the clinical T5 studies (Lehman et al., 2023; Lu et al., 2022), as well as clinical data not from MIMIC-III.

MIMIC-III Datasets We evaluate on CLIP (Mullenbach et al., 2021), a multi-label classification dataset, RadQA (Soni et al., 2022), a question-answering dataset, and MedNLI (Romanov and Shivade, 2018), a natural language inference dataset. Each dataset is comprised of instances from MIMIC-III (Johnson et al., 2016). Note that these datasets annotate data that comes from the same corpus used for clinical model pre-training.

Biomedical Datasets We evaluate on HOC (Baker et al., 2015), a multi-label classification dataset, and two named entity recognition datasets, BC5CDR-disease (Li et al., 2016a) and NCBI-disease (Doğan et al., 2014). HOC and NCBI-disease are datasets derived from PubMed abstracts [2], and the BC5CDR-disease dataset is derived from full PubMed articles. Note that some of them annotate data that comes from the same corpus used for SciFive+MIMIC-T5 pre-training.

Clinical Stigmatizing Language Datasets Due to the extremely limited availability of clinical text data, public clinical NLP datasets are drawn from MIMIC (MIMIC-III (Johnson et al., 2016) in particular), the same dataset used to pre-train the clinical-T5 models. How well do these clinical models generalize to a novel clinical domain and task? We obtain data from Hospital System (anonymized) that includes five medical specialties: Internal Medicine, Emergency Medicine, Pediatrics, OB-GYN, and Surgery. We evaluate the task of detecting stigmatizing language about patients by using the setup as Harrigian et al. (2023b). We also use their annotations from MIMIC-IV.

4. Are Clinical T5 Models Better for Clinical Text?

Both MIMIC-T5 and SciFive+MIMIC-T5 have demonstrated improved performance on clinical EHR text compared to a general-purpose model T5. Our first step is to replicate these findings with a more extensive and direct evaluation.

Methods We began by reproducing the results reported in Lehman et al. (2023) and Lu et al. (2022), adhering closely to each study's original evaluation methodology. Both studies only leveraged a single train/dev/test split for each respective task, and only Lehman et al. (2023) estimated uncertainty around their performance point estimates.[3] To understand whether prior results were sensitive to the initial dataset split, we conducted an additional set of cross-validation experiments. We merged existing training and development splits and randomly split them into 5 subsets to facilitate 5-fold cross-validation. Additionally, the original held-out test split was used to evaluate generalization and compare to prior work. We considered the same evaluation metrics used in prior work (Table 8), using a bootstrap procedure to estimate confidence intervals. For uniformity across the metrics (e.g., exact match, F1 score), differences in performance between models were assessed using a paired t-test (paired by fold). We referred to MIMIC-T5 and SciFive-T5's evaluation metrics on these datasets. Training and evaluation details are included in Appendix D.

Results Cross-validation results are reported in Table 1, while reproducibility efforts are reported alongside the originally reported performance metrics in Appendix F. Our results for MIMIC-T5 perform slightly worse than previously reported on MedNLI and RadQA, but better on CLIP. SciFive+MIMIC-T5 performs slightly better than previously reported on the three biomedical tasks. The average perfor-

2. https://pubmed.ncbi.nlm.nih.gov/

3. Lehman et al. (2023) used at least three random seeds in their training procedure to estimate uncertainty.

Dataset	Metrics	T5-Den	T5-Sup	FLAN-T5	MIMIC-T5	SciFive+MIMIC-T5
MedNLI	Acc.	$85.9_{(85.3,86.2)}$	$84.9_{(84.4,85.5)}$	$86.0_{(85.7,86.4)}$	$86.8_{(86.3,87.3)}$	$85.6_{(85.1,86.4)}$
RadQA	EM	$52.6_{(51.9,53.3)}$	$52.0_{(51.0,53.0)}$	$53.0_{(51.8,54.0)}$	$54.5_{(54.0,55.1)}$	$53.4_{(52.7,54.2)}$
	F1	$68.9_{(68.0,69.8)}$	$68.7_{(67.8,69.8)}$	$70.6_{(69.8,71.3)}$	$73.4_{(72.8,74.0)}$	$70.4_{(69.4,71.5)}$
CLIP	Macro F1	$63.9_{(62.2,65.6)}$	$62.4_{(61.4,63.4)}$	$64.3_{(62.9,65.7)}$	$66.1_{(65.9,66.3)}$	$63.5_{(62.2,64.8)}$
	Micro F1	$78.6_{(77.5,79.8)}$	$78.4_{(78.0,78.8)}$	$79.4_{(78.9,79.9)}$	$79.3_{(78.7,79.9)}$	$78.2_{(77.9,78.5)}$
HOC	F1	$84.9_{(84.4,85.6)}$	$84.8_{(84.4,85.1)}$	$84.8_{(84.2,85.3)}$	$82.8_{(82.6,83.0)}$	$85.1_{(84.8,85.4)}$
	P	$84.7_{(84.2,85.6)}$	$84.8_{(84.4,85.1)}$	$84.6_{(83.9,85.2)}$	$82.9_{(82.8,83.1)}$	$85.0_{(84.7,85.5)}$
	R	$85.1_{(84.7,85.6)}$	$84.8_{(84.3,85.2)}$	$85.0_{(84.5,85.4)}$	$82.6_{(82.3,83.0)}$	$85.2_{(84.9,85.5)}$
BC5CDR	F1	$82.9_{(82.3,83.4)}$	$82.9_{(82.5,83.5)}$	$83.3_{(82.9,83.8)}$	$81.2_{(80.9,81.5)}$	$83.6_{(83.4,83.8)}$
	P	$81.7_{(80.8,82.5)}$	$81.6_{(81.3,82.0)}$	$82.1_{(81.5,82.7)}$	$80.6_{(80.2,80.9)}$	$83.1_{(82.5,83.6)}$
	R	$84.2_{(83.6,84.6)}$	$84.3_{(83.6,85.0)}$	$84.5_{(84.2,84.9)}$	$81.7_{(81.2,82.3)}$	$84.2_{(83.6,84.8)}$
NCBI	F1	$85.1_{(84.6,85.5)}$	$84.6_{(84.2,85.0)}$	$84.4_{(83.9,85.0)}$	$79.1_{(78.3,79.9)}$	$85.4_{(85.0,86.2)}$
	P	$83.5_{(82.8,84.2)}$	$83.6_{(83.1,84.1)}$	$82.2_{(81.6,83.0)}$	$78.4_{(77.6,78.9)}$	$83.8_{(83.0,84.7)}$
	R	$86.7_{(86.3,87.2)}$	$85.6_{(85.2,86.2)}$	$86.8_{(86.3,87.3)}$	$79.9_{(78.7,81.3)}$	$87.1_{(86.9,87.5)}$

Table 1: Mean test performance and 95% confidence intervals across 5-fold cross-validation. MIMIC-T5 outperforms alternative models on clinical tasks, but struggles on biomedical tasks. Inclusion of biomedical data during pre-training is useful for the biomedical tasks.

mance over the cross-validation procedure mirrored the single train/dev/test split results.

Despite minor differences from the originally-reported results, we confirm the findings of Lehman et al. (2023) that MIMIC-T5 outperforms T5-Sup across the three clinical tasks at a statistically significant level. In comparison, SciFive+MIMIC-T5 either slightly underperformed or slightly outperformed the non-clinical T5 variants across the clinical tasks, albeit not at a statistically significant level. While the former outcome seems to corroborate the efficacy of domain-specific clinical pre-training, the latter outcome suggests the effect may be moderated inconsistently by pre-training on out-of-domain text. Given that MIMIC-T5 was pre-trained on text from the same corpus used for curating the evaluation datasets, the observed improvement may be a consequence of overfitting, which we explore below. Moreover, we note that the absolute increase in performance ($\sim$1.0 to $\sim$1.5 points across clinical tasks) falls within a range that is reasonable to expect when using domain adaptive pre-training (i.e., continued pre-training) (Gururangan et al., 2020).

Results on biomedical (journal article) datasets highlight the differences between the adaptation and train from scratch strategies. SciFive+MIMIC-T5 does better than MIMIC-T5 on these tasks, sometimes by more than 4 points, presumably since it was adapted from SciFive which is trained on similar data, whereas MIMIC-T5 has only seen clinical data. However, SciFive+MIMIC-T5 only slightly outperforms the general-purpose models, and does not do so consistently at a statistically significant level. It is possible that the general-purpose models have already seen data similar enough to the biomedical data such that further pre-training has minimal effect.

Finally, we observe the benefits of supervised instruction fine-tuning. FLAN-T5 performs comparably to MIMIC-T5 on the three clinical datasets and to SCIFIVE+MIMIC-T5 on the three biomedical datasets. MIMIC-T5 only outperforms FLAN-T5 on RadQA dataset, and only under 1 metric (F1 score), with statistical significance. Recent work suggests that training on relevant target tasks, either through instruction tuning or multitask training, has broad downstream benefits (Chung et al., 2022b; Mueller et al., 2022). T5-Den and T5-Sup perform similarly on biomedical tasks, but not the clinical tasks, where T5-Den outperforms T5-Sup slightly, albeit not at a statistically significant level. Whether the remaining gap in performance between non-clinical and clinical T5 models can be decreased via supervised pre-training on related tasks or extensive pre-training in general is an important open area of study, especially

given that non-clinical task datasets are significantly more abundant than clinical ones.

5. Do Clinical T5 Models Generalize to New Clinical Text?

A key limitation in developing clinical NLP systems is the lack of publicly available data due to PHI concerns. Mostly public available clinical models, including MIMIC-T5 and SCIFIVE+MIMIC-T5, and evaluation sets utilize MIMIC, so we do not know if they generalize to new clinical text sources. The weaker performance of MIMIC-T5 on the biomedical datasets may suggest its limited ability to generalize, or perhaps its benefits remain when focused on clinical EHR text. Since the clinical pre-training and evaluation sets represent only a single medical domain from a single institution, we ask: do clinical T5 models generalize to new clinical text sources and tasks better than general-purpose models?

Methods To evaluate generalization, we consider a new task for which we can evaluate on both a MIMIC dataset and a Hospital System dataset. The task is to characterize stigmatizing language in medical records, which includes three multi-class classification tasks that operate similarly to word sense disambiguation. The two datasets were curated using the annotation guidelines set forth by Harrigian et al. (2023b). We perform 5-fold cross-validation experiments as above on these two datasets, and compare the in-domain (MIMIC-IV) and out-of-domain (Hospital System (anonymized)) performance.

Results Table 2 shows the performance of the T5 models on the two stigmatizing language datasets. On the MIMIC-IV dataset, T5-Sup, MIMIC-T5, and SciFive+MIMIC-T5 achieve similar levels of performance. For this task, the benefits of MIMIC-T5 do not materialize, despite the MIMIC-IV dataset being within the distribution of its pre-training data.[4]

When we consider the out-of-distribution Hospital System (anonymized) data, MIMIC-T5 falls behind these other two models, perhaps either because MIMIC-T5 has been overfit to MIMIC data, or because it lacks the large-scale pre-training of the other models. SciFive+MIMIC-T5 performs slightly better than MIMIC-T5 on these two clinical datasets. However, neither MIMIC-T5 or SciFive+MIMIC-T5 show improvement over T5-Sup or FLAN-T5.

Finally, these evaluations further highlight the generalization ability of T5-Sup and FLAN-T5, which do better than clinical T5 models on both datasets. Even without clinical tasks in hand, the additional supervised training yields a model that does better on a novel clinical task across both data sources.

6. Do Clinical T5 Models Perform Well in Low-Resource Settings?

One of the possible reasons we have observed only a small performance gap between general-purpose and clinical models may be the presence of sufficient task training data. Indeed, previous work has found that tuning general-purpose BERT models for a clinical task using supervised data can nullify a clinical model's advantage (Harrigian et al., 2023a). In this regard, we may expect to see larger benefits to clinical models in a low-resource setting. Importantly, such settings are common in clinical tasks, where annotation typically requires advanced expertise, or is nigh impossible due to data privacy constraints. We therefore explore if the advantages of clinical models emerge more prominently in low-resource settings.

Methods We selected MedNLI and the two stigmatizing language datasets to test a low-resource setting. For MedNLI, we used the same cross validation splits as before, but downsampled the training subset to 1% (∼99 examples on average per cross validation fold). For the stigmatizing language datasets, which has less data than MedNLI to begin with, we were able to explore multiple downsampled training size settings – 1% (∼2 examples on average per fold), 5% (∼12 examples on average per fold), and 25% (∼59 examples on average per fold) for every task. We note that 1% of training data represents an extreme scenario, albeit one that is not inconceivable given the complexities of annotating some clinical datasets (Spasic and Nenadic, 2020). Experimental details are described in the Appendix D.

Results Figure 1 and Figure 2 show results for the Hospital System (anonymized) and MIMIC-IV stigmatizing language datasets, respectively, while Table 3 shows results for MedNLI. Additional task-specific results are reported in the Appendix G.

In the 1% MedNLI experiment, MIMIC-T5 is no longer leading, nor is it even competitive with, the other T5 models. For the two stigmatizing language

4. The MIMIC-IV stigmatizing language dataset may have been seen verbatim during MIMIC-T5 pre-training, as they did not leverage the same evaluation splits.

Model	Hospital System (anonymized)				MIMIC-IV			
	Credibility & Obstinacy	Compliance	Descriptors	**Average**	Credibility & Obstinacy	Compliance	Descriptors	**Average**
T5-Sup	$88.2_{(86.1,90.6)}$	$86.7_{(85.5,88.2)}$	$90.6_{(89.3,91.4)}$	$\mathbf{88.5}_{(87.8,89.0)}$	$76.0_{(74.1,77.9)}$	$92.9_{(91.7,94.2)}$	$85.7_{(83.4,88.0)}$	$84.9_{(84.2,85.8)}$
MIMIC-T5	$88.1_{(83.1,94.1)}$	$84.7_{(81.0,88.7)}$	$88.9_{(88.2,89.5)}$	$87.2_{(85.1,89.7)}$	$75.0_{(72.1,77.8)}$	$91.6_{(90.4,92.7)}$	$84.8_{(81.9,88.2)}$	$83.8_{(82.5,84.5)}$
SciFive+MIMIC-T5	$87.3_{(82.1,91.8)}$	$87.2_{(86.8,87.6)}$	$90.0_{(89.2,90.7)}$	$88.2_{(86.6,89.2)}$	$74.6_{(70.6,78.8)}$	$91.0_{(89.9,92.1)}$	$86.3_{(85.0,87.8)}$	$84.0_{(82.0,85.8)}$
FLAN-T5	$90.4_{(88.3,92.3)}$	$88.0_{(86.8,88.9)}$	$86.8_{(80.0,90.7)}$	$88.4_{(86.6,90.0)}$	$77.4_{(75.0,79.3)}$	$92.9_{(91.3,94.5)}$	$86.3_{(85.3,87.0)}$	$\mathbf{85.6}_{(85.0,86.1)}$

Table 2: Macro F1 scores on stigmatizing language datasets from the Hospital System (anonymized) and MIMIC-IV.

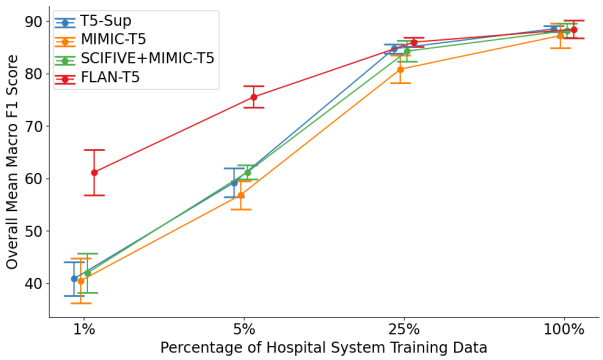

Figure 1: Mean macro-F1 scores across the three stigmatizing language classification tasks for the Hospital System (anonymized) dataset using a random 1%, 5%, and 25% sample of the training data.

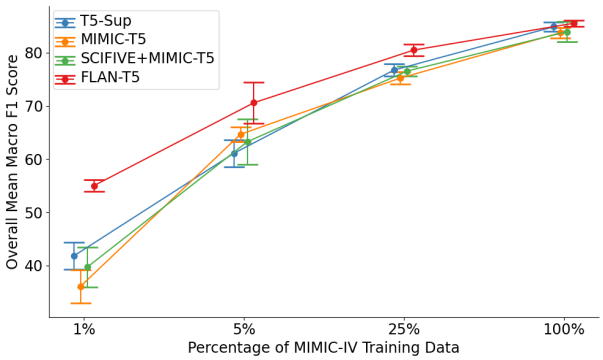

Figure 2: Mean macro-F1 scores across the three stigmatizing language classification tasks for the MIMIC-IV dataset using a random 1%, 5%, and 25% sample of the training data.

Model	Accuracy
T5-Sup	$78.80_{(77.69,79.90)}$
MIMIC-T5	$63.87_{(60.35,66.95)}$
SciFive+MIMIC-T5	$78.92_{(78.35,79.84)}$
FLAN-T5	$79.49_{(78.97,80.06)}$

Table 3: Performance with 1% training data on MedNLI dataset

datasets, T5-Sup, MIMIC-T5, and SciFive+MIMIC-T5 all perform roughly equivalently to one another in the low-resource settings as we gradually increase the sample size. That said, MIMIC-T5 finds itself at the bottom of relative performance rankings in more cases than not. Across all three tasks, FLAN-T5 excels in the low-resource settings compared to the three non-FLAN-tuned T5 models. Altogether, our results suggest that strong general-purpose language models are more appropriate for new distributions and those with limited amounts of task-specific training data.

7. Discussion

In this study, we independently replicated and extended prior work concerning specialized clinical T5 models (Lu et al., 2022; Lehman et al., 2023). By placing the models in an expanded context (i.e., additional non-clinical variants, evaluation datasets), we find ourselves in a position to make the following recommendations for clinical NLP practitioners.

Clinical models pre-trained from scratch using comparatively small datasets should not be used beyond their pre-training distributions.

Language models that are pre-trained on the target domain from scratch can achieve strong target-domain performance due to their in-domain pre-training (Gupta et al., 2023). For example, PubMedBERT (Gu et al., 2021), which was pre-trained from scratch on 3.2 billion words from PubMed Abstracts only, achieved remarkable performance in biomedical tasks drawn also from PubMed Abstracts. Unfortunately, large amounts of diverse and relevant pre-training data are generally hard to obtain for most real-world clinical applications.

Currently, MIMIC provides the largest public clinical text corpora, coming in at approximately 1.2 billion words across both MIMIC-III and MIMIC-IV (Lehman et al., 2023). However, practitioners must remember that MIMIC only represents a small subset of the broader clinical text landscape. Not only are there sample biases related to the demographics of patients in the corpora, but also topic biases (e.g., intensive care unit patients) and syntactic biases (e.g., hospital-specific documentation practices). LLMs trained on this comparatively narrow sample of data provide optimistic estimates of performance when evaluated on data also drawn from the MIMIC corpora (Lehman et al., 2023). As shown in our experiments, they struggle to adapt to the diverse needs of the clinical and biomedical space. So-called "clinical" language models may perhaps be better described as "MIMIC" language models.

Clinical models should only be used in downstream tasks if they have access to sufficient supervised training data.

In low-data settings, specificalized clinical models are more likely to underperform their generic counterparts that have been pre-trained on more data and/or using supervised fine-tuning. For example, in §6, we saw that T5-Sup, SciFive+MIMIC-T5, and FLAN-T5 were all more qualified to operate in at the 1% downsample regime than MIMIC-T5. Even larger generic LLMs such as ChatGPT and PaLM are likely to provide even more significant advantages in low-resource settings (Agrawal et al., 2022; Singhal et al., 2023). This leads us to our next recommendation.

The combination of task fine-tuning and FLAN instruction fine-tuning is hard to beat.

Recent research has shown that models pre-trained in a supervised manner are explicitly guided towards flat loss regions. Consequently, they are more robust to sequential fine-tuning in new domains (Mehta et al., 2023). It has also been shown that heterogeneity within pre-training data is important for promoting generalization in downstream tasks across diverse domains (Longpre et al., 2023). FLAN-T5's exceptional performance across all tasks in our study, especially in low-resource settings, affirms these prior findings.

Adapting general-purpose models rather than training from scratch, or leveraging a mix of both training strategies, may be the best option for clinical language models moving forward.

General-domain pre-training is beneficial given that clinical text has limited availability, and domain-adaptive or task-adaptive training can further improve performance on downstream tasks (Gururangan et al., 2020). For example, continued pre-training on small amounts of annotated clinical data (i.e., task-adaptive pre-training) has improved transfer to out-of-domain clinical data in a note-section classification task (Zhou et al., 2023), while adapting a generic language model, such as BERT, has been shown to be more efficient than pre-training a new model from scratch in unseen target domains (Lamproudis et al., 2022). Various contemporary works have presented ways to further optimize continued pre-training for adaptation. For example, by reverse engineering the effects of instruction-tuning (Fleshman and Durme, 2024), or by finding more efficient warm-up strategies (Gupta et al., 2023).

8. Conclusion

Our study aimed to determine whether existing clinical T5 models offer performance improvements over non-clinical T5 variants in clinical tasks. Our findings suggest that these clinical language models may outperform their general counterparts, but only under specific conditions: a) the task is on clinical EHR data sourced from MIMIC; b) sufficient annotated training data is available. Meanwhile, general-purpose language models, especially those leveraging supervised multi-task instructing tuning (e.g., FLAN-T5), excel when these criteria are not met.

More broadly, our results *do* provide support to previous claims that truly "domain-specific" language models, trained for specific data distributions and tasks, are ideal (Lehman et al., 2023). However, they also suggest that domain-specific models are not always practical. Data distributions may change over time (e.g., suddenly due to a pandemic, gradually as clinical practice evolves) (Khanday et al., 2020; Jeong et al., 2024) or across target populations (Harrigian et al., 2023a). Likewise, not all domains are data rich and diverse enough to support training a language model from scratch that endows broader linguistic abilities. In the clinical domain, non-clinical language models will almost certainly continue to have access to pre-training corpora that are orders of magnitude larger than clinical pre-training corpora, and that data scale is greatly advantageous for training high-performing LLMs (Kaplan et al., 2020).

Lastly, we would be remiss not to comment on the cost and sustainability implications of training domain-specific language models. While it is true that given the same inference budget, a clinical model trained from scratch is typically able to achieve better performance than a generic model (Lehman et al., 2023), such a perspective regarding model efficiency is inherently narrow. Using floating point operations (FLOPs) at inference time as a proxy for efficiency does not accurately reflect the additional cost complexities of training a language model from scratch – e.g., training compute, human labor (development and annotation), data access, and dealing with future distribution shift. Put another way, specialized models may achieve improvements in performance that are statistically significant, but not necessarily practically significant.

Limitations

T5 Model Family Our study focused on clinically pretrained T5 models, made possible by the released models from Lu et al. (2022) and Lehman et al. (2023). Comparing them allowed us to study the effect of pretraining strategies while controlling for the architecture. Future work could extend these investigations to other pretrained model architectures.

Data Availability Due to the scarcity of clinical text datasets, our evaluation was limited to a total of seven datasets. This constraint might not provide a comprehensive assessment for a wide range of clinical tasks. We have cited relevant work where applicable to supplement our findings.

Hyperparameter Search Due to computational costs associated with training each T5 model over multiple cross validation folds, we were only able to perform limited amounts of hyperparameter tuning at the onset of the study. Upon identifying a reasonable set of hyperparameters that worked well across the different datasets and models, we opted to fix them across all the cross validation runs. It is certainly possible that different hyperparameters may have yielded different outcomes. That said, the computational limitations faced here reflect what many practitioners face.

Statistical Measures We are aware that traditional statistical methods in NLP may not be sufficient to describe the divergence of performances in the clinical domain. While these statistical methods are effective in establishing the statistical significance of performance differences between models, they may fall short of providing a comprehensive understanding of how these differences manifest in real-world clinical settings. This limitation underscores the need for specialized evaluation frameworks in healthcare applications, thereby ensuring that the models not only perform well statistically but also deliver practical, reliable, and clinically relevant outcomes.

Future Directions The clinical domain presents unique challenges and opportunities. Future work could consider extending model architectures and training strategies (Gema et al., 2024; Labrak et al., 2024; Toma et al., 2023; Wu et al., 2023a), evaluation (Tang et al., 2023; Hager et al., 2024; Ness et al., 2024)), and techniques for addressing data scarcity and privacy issues (Lin et al., 2022).

Ethics Statement

Our clinical datasets were obtained from the PhysioNet website, accessed under the PhysioNet Credentialed Health Data Use Agreement, and from Hospital System (anonymized) after IRB approval. All data from MIMIC-III or MIMIC-IV has been de-identified to ensure privacy and confidentiality. Clinical-T5 models were pre-trained on MIMIC; we access and use them under the PhysioNet Credentialed Health Data license and data use agreement. All data processing and model training was done on remote servers secured using OS-level security protocols.

Acknowledgments

This work was supported by the National Institute on Minority Health and Health Disparities under grant number R01 MD017048. The content is solely the responsibility of the authors and does not necessarily represent the official views of NIMHD, NIH, or Johns Hopkins University.

We thank Eric Lehman, Qiuhao Lu, and Long Phan for clarifications about their code and experiments.

References

Monica Agrawal, Stefan Hegselmann, Hunter Lang, Yoon Kim, and David Sontag. Large language models are few-shot clinical information extractors. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, editors, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 1998–2022, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.130. URL https://aclanthology.org/2022.emnlp-main.130.

Emily Alsentzer, John R. Murphy, Willie Boag, Wei-Hung Weng, Di Jin, Tristan Naumann, and Matthew B. A. McDermott. Publicly available clinical bert embeddings, 2019.

Simon Baker, Ilona Silins, Yufan Guo, Imran Ali, Johan Högberg, Ulla Stenius, and Anna Korhonen. Automatic semantic classification of scientific literature according to the hallmarks of cancer. *Bioinformatics*, 32(3):432–440, October 2015. ISSN 1367-4803. doi: 10.1093/bioinformatics/btv585. URL http://dx.doi.org/10.1093/bioinformatics/btv585.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners, 2020.

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, Harsha Nori, Hamid Palangi, Marco Tulio Ribeiro, and Yi Zhang. Sparks of artificial general intelligence: Early experiments with gpt-4, 2023.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. Palm: Scaling language modeling with pathways, 2022.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, Dasha Valter, 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, 2022a.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*, 2022b.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep

bidirectional transformers for language understanding, 2019.

Rezarta Islamaj Doğan, Robert Leaman, and Zhiyong Lu. NCBI disease corpus: A resource for disease name recognition and concept normalization. *J. Biomed. Inform.*, 47:1–10, February 2014.

Rezarta Islamaj Doğan, Robert Leaman, and Zhiyong Lu. Ncbi disease corpus: A resource for disease name recognition and concept normalization. *Journal of Biomedical Informatics*, 47:1–10, Feb 2014. doi: 10.1016/j.jbi.2013.12.006.

William Fleshman and Benjamin Van Durme. Readapt: Reverse engineered adaptation of large language models, 2024.

Aryo Pradipta Gema, Pasquale Minervini, Luke Daines, Tom Hope, and Beatrice Alex. Parameter-efficient fine-tuning of llama for the clinical domain, 2024. URL https://arxiv.org/abs/2307.03042.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. Domain-specific language model pretraining for biomedical natural language processing. *ACM Trans. Comput. Healthcare*, 3(1), oct 2021. doi: 10.1145/3458754. URL https://doi.org/10.1145/3458754.

Kshitij Gupta, Benjamin Thérien, Adam Ibrahim, Mats L. Richter, Quentin Anthony, Eugene Belilovsky, Irina Rish, and Timothée Lesort. Continual pre-training of large language models: How to (re)warm your model?, 2023.

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, 2020.

Bernal Jiménez Gutiérrez, Huan Sun, and Yu Su. Biomedical language models are robust to sub-optimal tokenization, 2023.

Paul Hager, Friederike Jungmann, Robbie Holland, Kunal Bhagat, Inga Hubrecht, Manuel Knauer, Jakob Vielhauer, Marcus Makowski, Rickmer Braren, Georgios Kaissis, and et al. Evaluation and mitigation of the limitations of large language models in clinical decision-making. *Nature Medicine*, 30(9):2613–2622, Jul 2024. doi: 10.1038/s41591-024-03097-1.

Keith Harrigian, Tina Tang, Anthony Gonzales, Cindy X. Cai, and Mark Dredze. An eye on clinical bert: Investigating language model generalization for diabetic eye disease phenotyping, 2023a.

Keith Harrigian, Ayah Zirikly, Brant Chee, Alya Ahmad, Anne Links, Somnath Saha, Mary Catherine Beach, and Mark Dredze. Characterization of stigmatizing language in medical records. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 312–329, Toronto, Canada, July 2023b. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-short.28. URL https://aclanthology.org/2023.acl-short.28.

Hyewon Jeong, Sarah Jabbour, Yuzhe Yang, Rahul Thapta, Hussein Mozannar, William Jongwon Han, Nikita Mehandru, Michael Wornow, Vladislav Lialin, Xin Liu, et al. Recent advances, applications, and open challenges in machine learning for health: Reflections from research roundtables at ml4h 2023 symposium. *arXiv preprint arXiv:2403.01628*, 2024.

Alistair E. W. Johnson, Lucas Bulgarelli, Lu Shen, Alvin Gayles, Ayad Shammout, Steven Horng, Tom J. Pollard, Sicheng Hao, Benjamin Moody, Brian Gow, and et al. Mimic-iv, a freely accessible electronic health record dataset, Jan 2023. URL https://www.nature.com/articles/s41597-022-01899-x.

Alistair E.W. Johnson, Tom J. Pollard, Lu Shen, Li-wei H. Lehman, 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 (1), May 2016. doi: 10.1038/sdata.2016.35.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

Akib Mohi Ud Din Khanday, Syed Tanzeel Rabani, Qamar Rayees Khan, Nusrat Rouf, and Masarat Mohi Ud Din. Machine learning based approaches for detecting covid-19 using clinical text data. *International Journal of Information Technology*, 12: 731–739, 2020.

Yanis Labrak, Adrien Bazoge, Emmanuel Morin, Pierre-Antoine Gourraud, Mickael Rouvier, and Richard Dufour. Biomistral: A collection of open-source pretrained large language models for medical domains, 2024. URL https://arxiv.org/abs/2402.10373.

Anastasios Lamproudis, Aron Henriksson, and Hercules Dalianis. Evaluating pretraining strategies for clinical BERT models. In Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 410–416, Marseille, France, June 2022. European Language Resources Association. URL https://aclanthology.org/2022.lrec-1.43.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240, September 2019. ISSN 1367-4811. doi: 10.1093/bioinformatics/btz682. URL http://dx.doi.org/10.1093/bioinformatics/btz682.

Eric P. Lehman, Evan Hernandez, Diwakar Mahajan, Jonas Wulff, Micah J. Smith, Zachary M. Ziegler, Daniel Nadler, Peter Szolovits, Alistair E. W. Johnson, and Emily Alsentzer. Do we still need clinical language models? *ArXiv*, abs/2302.08091, 2023.

Brian Lester, Rami Al-Rfou, and Noah Constant. The power of scale for parameter-efficient prompt tuning. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih, editors, *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3045–3059, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.243. URL https://aclanthology.org/2021.emnlp-main.243.

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 Anna Rumshisky, Kirk Roberts, Steven Bethard, and Tristan Naumann, editors, *Proceedings of the 3rd Clinical Natural Language Processing Workshop*, pages 146–157, Online, November 2020a. Association for Computational Linguistics. doi: 10.18653/v1/2020.clinicalnlp-1.17. URL https://aclanthology.org/2020.clinicalnlp-1.17.

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, 2020b.

Jiao Li, Yueping Sun, Robin J Johnson, Daniela Sciaky, Chih-Hsuan Wei, Robert Leaman, Allan Peter Davis, Carolyn J Mattingly, Thomas C Wiegers, and Zhiyong Lu. BioCreative V CDR task corpus: a resource for chemical disease relation extraction. *Database (Oxford)*, 2016:baw068, May 2016a.

Jiao Li, Yueping Sun, Robin J. Johnson, Daniela Sciaky, Chih-Hsuan Wei, Robert Leaman, Allan Peter Davis, Carolyn J. Mattingly, Thomas C. Wiegers, and Zhiyong Lu. Biocreative v cdr task corpus: A resource for chemical disease relation extraction. *Database*, 2016, 2016b. doi: 10.1093/database/baw068.

Bill Yuchen Lin, Chaoyang He, Zihang Ze, Hulin Wang, Yufen Hua, Christophe Dupuy, Rahul Gupta, Mahdi Soltanolkotabi, Xiang Ren, and Salman Avestimehr. FedNLP: Benchmarking federated learning methods for natural language processing tasks. In Marine Carpuat, Marie-Catherine de Marneffe, and Ivan Vladimir Meza Ruiz, editors, *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 157–175, Seattle, United States, July 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-naacl.13. URL https://aclanthology.org/2022.findings-naacl.13.

Shayne Longpre, Gregory Yauney, Emily Reif, Katherine Lee, Adam Roberts, Barret Zoph, Denny Zhou, Jason Wei, Kevin Robinson, David Mimno, and Daphne Ippolito. A pretrainer's guide to training data: Measuring the effects of data age, domain coverage, quality, & toxicity, 2023.

Qiuhao Lu, Dejing Dou, and Thien Nguyen. ClinicalT5: A generative language model for clinical

text. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 5436–5443, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. URL https://aclanthology.org/2022.findings-emnlp.398.

Sanket Vaibhav Mehta, Darshan Patil, Sarath Chandar, and Emma Strubell. An empirical investigation of the role of pre-training in lifelong learning, 2023.

David Mueller, Nicholas Andrews, and Mark Dredze. Do text-to-text multi-task learners suffer from task conflict? *arXiv preprint arXiv:2212.06645*, 2022.

James Mullenbach, Yada Pruksachatkun, Sean Adler, Jennifer 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. *CoRR*, abs/2106.02524, 2021. URL https://arxiv.org/abs/2106.02524.

Hiroki Nakayama. seqeval: A python framework for sequence labeling evaluation, 2018. URL https://github.com/chakki-works/seqeval. Software available from https://github.com/chakki-works/seqeval.

Robert Osazuwa Ness, Katie Matton, Hayden Helm, Sheng Zhang, Junaid Bajwa, Carey E. Priebe, and Eric Horvitz. Medfuzz: Exploring the robustness of large language models in medical question answering, 2024. URL https://arxiv.org/abs/2406.06573.

Long N. 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 e-prints*, art. arXiv:2106.03598, May 2021. doi: 10.48550/arXiv.2106.03598.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. 2019. URL https://api.semanticscholar.org/CorpusID:160025533.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer, 2023.

Pranav Rajpurkar, Jian Zhang, and Percy Liang. Know what you don't know: Unanswerable questions for squad. In *ACL 2018*, 2018.

Lance Ramshaw and Mitch Marcus. Text chunking using transformation-based learning. In *Third Workshop on Very Large Corpora*, 1995. URL https://www.aclweb.org/anthology/W95-0107.

Alexey Romanov and Chaitanya Shivade. Lessons from natural language inference in the clinical domain, 2018.

Noam Shazeer and Mitchell Stern. Adafactor: Adaptive learning rates with sublinear memory cost, 2018.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S. Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, Perry Payne, Martin Seneviratne, Paul Gamble, Chris Kelly, Nathaneal Scharli, Aakanksha Chowdhery, Philip Mansfield, Blaise Aguera y Arcas, Dale Webster, Greg S. Corrado, Yossi Matias, Katherine Chou, Juraj Gottweis, Nenad Tomasev, Yun Liu, Alvin Rajkomar, Joelle Barral, Christopher Semturs, Alan Karthikesalingam, and Vivek Natarajan. Large language models encode clinical knowledge, 2022.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, Perry Payne, Martin Seneviratne, Paul Gamble, Chris Kelly, Abubakr Babiker, Nathanael Schärli, Aakanksha Chowdhery, Philip Mansfield, Dina Demner-Fushman, Blaise Agüera Y Arcas, Dale Webster, Greg S Corrado, Yossi Matias, Katherine Chou, Juraj Gottweis, Nenad Tomasev, Yun Liu, Alvin Rajkomar, Joelle Barral, Christopher Semturs, Alan Karthikesalingam, and Vivek Natarajan. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180, August 2023.

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.

Irena Spasic and Goran Nenadic. Clinical text data in machine learning: Systematic review. JMIR Medical Informatics, 03 2020. doi: 10.2196/17984.

Liyan Tang, Zhaoyi Sun, Betina Idnay, Jordan G. Nestor, Ali Soroush, Pierre A. Elias, Ziyang Xu, Ying Ding, Greg Durrett, Justin F. Rousseau, and et al. Evaluating large language models on medical evidence summarization. *npj Digital Medicine*, 6 (1), Aug 2023. doi: 10.1038/s41746-023-00896-7.

Ross Taylor, Marcin Kardas, Guillem Cucurull, Thomas Scialom, Anthony Hartshorn, Elvis Saravia, Andrew Poulton, Viktor Kerkez, and Robert Stojnic. Galactica: A large language model for science, 2022.

Augustin Toma, Patrick R. Lawler, Jimmy Ba, Rahul G. Krishnan, Barry B. Rubin, and Bo Wang. Clinical camel: An open expert-level medical language model with dialogue-based knowledge encoding, 2023. URL https://arxiv.org/abs/2305.12031.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. Llama: Open and efficient foundation language models, 2023.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2017.

Chaoyi Wu, Xiaoman Zhang, Ya Zhang, Yanfeng Wang, and Weidi Xie. Pmc-llama: Further finetuning llama on medical papers, 2023a.

Shijie Wu, Ozan Irsoy, Steven Lu, Vadim Dabravolski, Mark Dredze, Sebastian Gehrmann, Prabhanjan Kambadur, David Rosenberg, and Gideon Mann. Bloomberggpt: A large language model for finance, 2023b.

Xi Yang, Aokun Chen, Nima PourNejatian, Hoo Chang Shin, Kaleb E Smith, Christopher Parisien, Colin Compas, Cheryl Martin, Mona G Flores, Ying Zhang, Tanja Magoc, Christopher A Harle, Gloria Lipori, Duane A Mitchell, William R Hogan, Elizabeth A Shenkman, Jiang Bian, and Yonghui Wu. Gatortron: A large clinical language model to unlock patient information from unstructured electronic health records, 2022.

Michihiro Yasunaga, Jure Leskovec, and Percy Liang. Linkbert: Pretraining language models with document links, 2022.

Xiang Yue, Bernal Jimenez Gutierrez, and Huan Sun. Clinical reading comprehension: A thorough analysis of the emrQA dataset. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4474–4486, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.410. URL https://aclanthology.org/2020.acl-main.410.

Weipeng Zhou, Meliha Yetisgen, Majid Afshar, Yanjun Gao, Guergana Savova, and Timothy A Miller. Improving model transferability for clinical note section classification models using continued pretraining. *Journal of the American Medical Informatics Association*, 31(1):89–97, 09 2023. ISSN 1527-974X. doi: 10.1093/jamia/ocad190. URL https://doi.org/10.1093/jamia/ocad190.

Appendix A. Dataset Details

We adapted the data processing methods mostly from the original MIMIC-T5, SciFive+MIMIC-T5 and SciFive's paper and codebases.

MedNLI MedNLI (Romanov and Shivade, 2018) is a natural language inference dataset derived from MIMIC-III (Johnson et al., 2016). Given two sentences - premise and hypothesis - it labels the relation between these two sentences as "entailment", "natural", and "contradiction."

RadQA RadQA (Soni et al., 2022) is a question-answering dataset sourced from the radiology report in MIMIC-III (Johnson et al., 2016). It annotates the answer span to the questions; if no answers for the question, the answer text is empty.

CLIP CLIP (Mullenbach et al., 2021) is a multi-label classification dataset of clinical action items, such as "patient instructions" and "appointment." Due to the long context of clinical records of these two tasks, we referred to the methodology utilized by MIMIC-T5, which involves segmenting long records

and appropriately mapping labels to the segmented sequences.

BC5CDR-disease BC5CDR-disease (Li et al., 2016b) is a Disease Named Entity Recognition corpus, which annotates 1500 PubMed articles with disease entities labels. Suppose (x, y) is the input-output pair: y is the same as x except that disease entities are enclosed with "disease*" and "*disease".

NCBI-disease NCBI-disease (Doğan et al., 2014) is also a corpus for Disease Named Entity Recoginition from PubMed abtracts. Both BC5CDR-disease and NCBI-disease are preprocessed by SciFive and available in their repo.

HOC HOC dataset (Baker et al., 2015) is accessible on Huggingface. It is developed from PubMed publication abstracts. We use the sentence-level dataset, where each sentence is annotated for different hallmarks of cancer.

Clinical Stigmatizing Language Datasets Clinical Stigmatizing Language Datasets (Harrigian et al., 2023b) characterizes stigmatizing languages in three classification tasks:

1. Credibility & Obstinacy (Disbelief, Difficult, Exclude): expressions of doubt or resistance.
2. Compliance (Negative, Neutral, Positive): adherence to medical advice.
3. Descriptors (Negative, Neutral, Positive, Exclude): characterization of patient behavior and demeanor.

Appendix B. Prefix and Instructions

T5 Prefix We used the same prefix as MIMIC-T5 on MedNLI, RadQA, and CLIP datasets. We referred to SCFIVE-T5's github repository for NCBI, BC5CDR, NCBI's prefix. On the stigmatizing language datasets, we referred to its keyword category as prefix before inputs: "adamant" for the Credibility & Obstinacy task; "compliance" for the Compliance task; "other" for the Descriptors task.

FLAN-T5 Instructions We didn't exhaustively experiment with different instructions for FLAN-T5. We referred to some instructions from MIMIC-T5 and most instructions are straight-forward.

- MedNLI `Answer entailment, contradiction or neutral. Premise: {premise} Hypothesis:{hypothesis}`
- RadQA `Context: {context} Question: {question} If no answer is found in the context, do not reply; otherwise, give an answer from the context:`
- CLIP `Context: {context}. Label the above sentence as an empty string or as one or more of the following options, delimited by comma: Options: {labels}`
- HOC `Sentence: {input} Assign the above sentence as zero or more of the following class labels: {labels}`
- BC5CR-disease & NCBI-disease `Sentence: {input} Identify and label disease terms in the sentence:`
- Stigmatizing Language Dataset
  - Credibility & Obstinacy
    `Classify this sentence as difficult, disbelief, or exclude, regarding the credibility and obstinacy of the patient: {input}`
  - Compliance
    `Classify this sentence as negative, neutral, or positive, regarding the patient's compliance with medical advice: {input}`
  - Descriptors
    `Classify this sentence as exclude, negative, neutral, or positive, regarding the patient's behavior and demeanor: {input}`

Appendix C. Model Details

The details of the pre-training corpus for each model are listed in Table 4.

Appendix D. Training Details

D.1. Training settings

Training on Full Datasets We use code provided by prior work (Phan et al., 2021; Lehman et al., 2023) to fine-tune the various T5 variants with pytorch==2.1.2. Throughout our experiments, we used the adafactor optimizer (Shazeer and Stern, 2018) with constant $lr = 1e - 4$, batch size of 64, and set max sequence lengths that depend on the dataset.

	Weight Initialization	pre-training Data
T5-Sup (Raffel et al., 2023)	Random	Colossal Clean Crawled Corpus (C4) + Supervised Tasks (Raffel et al., 2023)
T5-Den (Raffel et al., 2023)	Random	Colossal Clean Crawled Corpus (C4) (Raffel et al., 2023)
Clinical-T5 (Lehman et al., 2023)	Random	MIMIC-III (Johnson et al., 2016) & MIMIC-IV (Johnson et al., 2023)
Clinical-T5 (Lu et al., 2022)	SciFive-PubMed-PMC (Phan et al., 2021)	MIMIC-III (Johnson et al., 2016)

Table 4: pre-training corpus for the models we compared. All models compared have 770 million parameters.

For all fine-tuning runs, we train for a fixed amount of epochs (30) and pick the best checkpoint based on validation set performance. We chose accuracy for MedNLI, and F-1 for other datasets to pick the best checkpoint during the validation and then evaluated on the complete test split per model and dataset.

Training on Downsampled Datasets Due to variations in the downsampled training data, we conducted a learning rate search for all models in the range of $[1e-3, 5e-4, 1e-4, 1e-5]$ using a batch size of 16 with 5% and 25% downsampling training data and using a batch size of 4 with 1% downsampling training data on stigmatizing language datasets and ran the learning rate search in the range of $[5e-3, 1e-3, 5e-4, 1e-4]$ with a batch size of 64 on MedNLI dataset. We trained each model for 30 epochs. We selected the best checkpoint for each model based on their average validation score across folds (i.e. macro-F1 for the stigmatizing task and accuracy for the MedNLI task) and subsequently evaluated on the test dataset. The best-found learning rates are reported in Table 5, Table 6, and Table 7.

D.2. Differences from Prior Work

We summarize a list of changes we made that possibly contribute to the difference in results compared to prior Clinical-T5s works (Lu et al., 2022; Lehman et al., 2023):

- We trained each model on each supervised dataset from scratch.
- We implemented the 5-fold cross-validation experiment, focusing on variance in data. This is a different evaluation structure from prior works.

1% Training data		
Model	Learning Rate	Batch Size
T5-Sup	1e-4	4
FLAN-T5	1e-4	4
MIMIC-T5	5e-4	4
SCIFIVE+MIMIC-T5	1e-4	4
5% Training data		
Model	Learning Rate	Batch Size
T5-Sup	1e-4	16
FLAN-T5	1e-4	16
MIMIC-T5	5e-4	16
SCIFIVE+MIMIC-T5	5e-4	16
25% Training data		
Model	Learning Rate	Batch Size
T5-Sup	1e-4	16
FLAN-T5	1e-4	16
MIMIC-T5	1e-4	16
SCIFIVE+MIMIC-T5	1e-4	16

Table 5: Best-found hyperparameter settings for down-sampling experiments on stigmatizing language dataset - MIMIC-IV

1% Training data		
Model	Learning Rate	Batch Size
T5-Sup	1e-3	4
FLAN-T5	1e-4	4
MIMIC-T5	1e-3	4
SCIFIVE+MIMIC-T5	5e-4	4
5% Training data		
Model	Learning Rate	Batch Size
T5-Sup	1e-4	16
FLAN-T5	1e-4	16
MIMIC-T5	1e-3	16
SCIFIVE+MIMIC-T5	5e-4	16
25 % Training data		
Model	Learning Rate	Batch Size
T5-Sup	1e-4	16
FLAN-T5	1e-4	16
MIMIC-T5	5e-4	16
SCIFIVE+MIMIC-T5	1e-4	16

Table 6: Best-found hyperparameter settings for down-sampling experiments on stigmatizing language dataset - Hospital System

1 % Training Data		
Model	Learning Rate	Batch Size
T5-Sup	1e-4	64
FLAN-T5	1e-4	64
MIMIC-T5	1e-3	64
SCIFIVE+MIMIC-T5	1e-4	64

Table 7: Best-found hyperparameter settings for down-sampling experiments on MedNLI dataset with 1% training data

- We added T5-Den and FLAN-T5 in addition to widely used T5-Sup into the comparison.

D.3. Computing Resources

We conducted our experiments on multiple 40G/80G A100 GPUs and Tesla M60 GPUs. Each run of the fine-tuning takes a few hours, except for CLIP dataset which could take approximately 10 hours for 30 epochs with multiple GPUs. However, it should be noted that these models are approaching a size that is not necessarily easy to accommodate in clinical settings.

Appendix E. Evaluation Details

We use the Python 3.9 packages sklearn==1.1.2, and evaluate==0.3.0 for evaluations.

Classification For the CLIP dataset, we followed the same evaluation strategy as Lehman2023DoWS by transforming predicted values into binary matrices, whose dimensions indicate the presence of class labels. And then used the package sklearn.metrics to calculate macro-F1 and micro-F1. On HOC dataset, we evaluate the outputs in the same manner as Phan et al., 2021 and Yasunaga et al., 2022. For stigmatizing datasets, we also used sklearn.metrics for macro-F1 scores.

Natural Language Inference For the MedNLI dataset, without additionally post-processing the outputs, we compute the F1 score and the accuracy directly between the labels and predicted values.

Question Answering We followed the same processing method as Lehman et al., 2023 utilizing SQuAD 2.0 (Rajpurkar et al., 2018).

Named Entity Recognition As instructed by Clinical-T5 (Lu et al., 2022), we referred to the evaluation of SciFive (Phan et al., 2021). BC5CDR-disease and NCBI-disease datasets label the entity by inserting "disease*" and "*disease" around disease names in the input. We convert generated sentence outputs into sequences of "B-disease", "I-disease", "O", and padding and then evaluate using the metrics seqeval (Ramshaw and Marcus, 1995; Nakayama, 2018).

Appendix F. Reproducibility Experiments

We evaluated T5, MIMIC-T5, and SciFive+MIMIC-T5 on the six datasets, following the original methodologies (i.e. a single train/dev/test split) proposed in each respective paper. Results of three clinical datasets are shown in Table 9 and results of three biomedical datasets are shown in Table 10. And these results are single-run.

Appendix G. Cross-Validation Experiments

We did 5-fold cross-validation in the following procedures: we merged the training and validation datasets and then shuffled them. Within each fold, 4 subsets are for training and one subset is for validation. For down-sampling experiments, we down-sampled the training data only while keeping the validation the same per fold. In the end, all the models were evaluated on the original test dataset. The detailed performances cross models are listed in Appendix G with paired T-test results.

Specifically, 11 shows 5-fold cross validation results on MedNLI Dataset. Table 12 shows 5-fold cross validation results on RadQA Dataset. Table 13 shows 5-fold cross validation results on CLIP dataset. Table 14 shows 5-fold cross validation results on HOC dataset. Table 15 shows 5-fold cross validation results on BC5CDR-disease dataset. Table 16 shows 5-fold cross validation results on NCBI-disease dataset. Table 17 and Table 18 are results for the stigmatizing language datasets from MIMIC-IV and Hospital System separately. Table 19, Table 20 , and Table 21 illustrate the outcomes of down-sampling experiments on the MIMIC-IV dataset; Table 22, Table 23, and Table 24 present the outcomes of down-sampling experiments on the Hospital System dataset.

Task	Dataset	Source	Evaluated By	Metrics
NLI	MedNLI	MIMIC	Both	Accuracy
Classif.	CLIP	MIMIC	MIMIC-T5	F1(Macro, Micro)
	HOC	Biomedical	SciFive+MIMIC-T5	F1, P, R
QA	RadQA	MIMIC	MIMIC-T5	EM, F1
NER	BC5CDR NCBI	Biomedical	SciFive+MIMIC-T5	F1, P, R

Table 8: Statistics of datasets used in MIMIC-T5 and SciFive+MIMIC-T5

	MedNLI	RadQA		CLIP	
	Accuracy	EM	F1	Macro-F1	Micro-F1
T5-Den	85.16	52.44	71.42	64.68	79.11
T5-Sup	84.95	54.07	71.14	64.20	78.93
MIMIC-T5					
— Ours	86.64	54.72	73.38	65.76	80.11
— Reported	87.2	55.0	74.5	66.3	80.0
SciFive+MIMIC-T5					
— Ours	85.79	52.28	70.74	63.79	78.24
— Reported	85.86	N/A	N/A	N/A	N/A

Table 9: We reproduced the results of MIMIC-T5 on three clinical datasets as reported and complemented evaluations of SciFive+MIMIC-T5 on the clinical datasets that were not used previously.

	HOC			BC5CDR			NCBI		
	F1	P	R	F1	P	R	F1	P	R
T5-Den	85.57	85.81	85.34	80.43	81.74	79.17	85.29	82.86	87.88
T5-Sup	85.91	86.01	85.80	81.70	81.46	81.94	82.53	82.49	82.58
MIMIC-T5									
— Ours	81.93	82.00	81.85	78.39	77.52	79.29	77.97	77.80	78.14
— Reported	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
SciFive+MIMIC-T5									
— Ours	85.42	85.52	85.32	82.44	81.22	83.70	85.67	83.18	88.31s
—Reported	84.78	85.37	84.79	80.35	79.24	81.49	86.73	86.37	87.09

Table 10: Evaluations on three biomedical datasets in SciFive+MIMIC-T5 were also reproduced. MIMIC-T5 and T5's performances are also shown in the table.

MedNLI

Metrics		T5-Den	T5-Sup	FLAN-T5	MIMIC-T5	SciFive+MIMIC-T5
Accuracy	Mean	85.89 ♠	84.94 ♣	86.03 ♦	86.75 ↑♠0.86, ↑♣1.81, ↑♦0.72	85.64 ↓♠0.25, ↑♣0.70, ↓♦0.39
	95% CI	[85.34, 86.24]	[84.38, 85.49]	[85.72, 86.43]	[86.27, 87.29]	[85.13, 86.37]

MedNLI-Accuracy-T-test

Models	T5-Den	T5-Sup	FLAN-T5	MIMIC-T5	SciFive+MIMIC-T5
T5-Den-Large	-	-	-	-	-
T5-Sup-Large	t(df=4)=-2.63, P=.06	-	-	-	-
FLAN-T5-Large	t(df=4)=0.31, P=.77	t(df=4)=3.17, P=.03	-	-	-
MIMIC-T5-Large	t(df=4)=1.97, P=.12	t(df=4)=4.14, P=.01	t(df=4)=1.75, P=.15	-	-
SciFive+MIMIC-T5-Large	t(df=4)=-0.45, P=.68	t(df=4)=1.18, P=.30	t(df=4)=-1.45, P=.22	t(df=4)=-2.18, P=.10	-

Table 11: 5-fold cross-validation results across models on MedNLI dataset and the t-test results among T5, MIMIC-T5, and SciFive+MIMIC-T5

RadQA

Metrics		T5-Den	T5-Sup	FLAN-T5	MIMIC-T5	SciFive+MIMIC-T5
EM	Mean	52.57 ♠	51.95 ♣	53.03 ♦	54.50 ↑♠1.92, ↑♣2.54, ↑♦1.47	53.42 ↑♠0.85, ↑♣1.47, ↑♦0.39
	95% CI	[51.9, 53.27]	[51.04, 52.95]	[51.79, 54.01]	[54.04, 55.05]	[52.69, 54.22]
F1	Mean	68.90 ♠	68.68 ♣	70.62 ♦	73.40 ↑♠4.50, ↑♣4.72, ↑♦2.78	70.37 ↑♠1.47, ↑♣1.7, ↓♦0.25
	95% CI	[68.04, 69.79]	[67.82, 69.77]	[69.76, 71.34]	[72.75, 74.00]	[69.44, 71.49]

RadQA-EM-T-test

Models	T5-Den	T5-Sup	FLAN-T5	MIMIC-T5	SciFive+MIMIC-T5
T5-Den-Large	-	-	-	-	-
T5-Sup-Large	t(df=4)=-0.91, P=.41	-	-	-	-
FLAN-T5-Large	t(df=4)=0.47, P=.67	t(df=4)=1.4, P=.23	-	-	-
MIMIC-T5-Large	t(df=4)=4.3, P=.01	t(df=4)=2.92, P=.04	t(df=4)=1.84, P=.14	-	-
SciFive+MIMIC-T5-Large	t(df=4)=1.19, P=.30	t(df=4)=1.95, P=.12	t(df=4)=0.54, P=.62	t(df=4)=-1.97, P=.12	-

RadQA-F1-T-test

Models	T5-Den	T5-Sup	FLAN-T5	MIMIC-T5	SciFive+MIMIC-T5
T5-Den-Large	-	-	-	-	-
T5-Sup-Large	t(df=4)=-0.55, P=.61	-	-	-	-
FLAN-T5-Large	t(df=4)=2.53, P=.06	t(df=4)=3.65, P=.02	-	-	-
MIMIC-T5-Large	t(df=4)=5.95, P=.004	t(df=4)=5.2, P=.007	t(df=4)=3.5, P=.02	-	-
SciFive+MIMIC-T5-Large	t(df=4)=2.19, P=.09	t(df=4)=2.15, P=.10	t(df=4)=-0.32, P=.76	t(df=4)=-5.47, P=.005	-

Table 12: 5-fold cross-validation results across models on RadQA dataset and the t-test results among T5, MIMIC-T5, and SciFive+MIMIC-T5.

CLIP

Metrics		T5-Den	T5-Sup	FLAN-T5	MIMIC-T5	SciFive+MIMIC-T5
Macro-F1	Mean	63.86 ♠	62.42 ♣	64.33 ♦	66.09 ↑♠2.23, ↑♣3.67, ↑♦1.75	63.46 ↓♠0.4, ↑♣1.04, ↓♦0.88
	95% CI	[62.18, 65.55]	[61.37, 63.4]	[62.94, 65.72]	[65.84, 66.26]	[62.2, 64.76]
Micro-F1	Mean	78.62 ♠	78.38 ♣	79.40 ♦	79.28 ↑♠0.67, ↑♣0.9, ↓♦0.12	78.23 ↓♠0.38, ↓♣0.15, ↓♦1.17
	95% CI	[77.48, 79.79]	[78.03, 78.77]	[78.85, 79.94]	[78.51, 80.01]	[77.93, 78.51]

CLIP-Macro-F1-T-test

Models	T5-Den	T5-Sup	FLAN-T5	MIMIC-T5	SciFive+MIMIC-T5
T5-Den-Large	-	-	-	-	-
T5-Sup-Large	t(df=4)=-1.39, P=.24	-	-	-	-
FLAN-T5-Large	t(df=4)=0.38, P=.72	t(df=4)=1.83, P=.14	-	-	-
MIMIC-T5-Large	t(df=4)=2.15, P=.10	t(df=4)=5.86, P=.004	t(df=4)=2.02, P=.11	-	-
SciFive+MIMIC-T5-Large	t(df=4)=-0.3, P=.78	t(df=4)=0.95, P=.39	t(df=4)=-0.73, P=.51	t(df=4)=-4.33, P=.01	-

CLIP-Micro-F1-T-test

Models	T5-Den	T5-Sup	FLAN-T5	MIMIC-T5	SciFive+MIMIC-T5
T5-Den-Large	-	-	-	-	-
T5-Sup-Large	t(df=4)=-0.34, P=.75	-	-	-	-
FLAN-T5-Large	t(df=4)=1.83, P=.14	t(df=4)=2.55, P=.06	-	-	-
MIMIC-T5-Large	t(df=4)=0.64, P=.56	t(df=4)=1.89, P=.13	t(df=4)=-0.18, P=.87	-	-
SciFive+MIMIC-T5-Large	t(df=4)=-0.67, P=.54	t(df=4)=-0.49, P=.65	t(df=4)=-4.19, P=.01	t(df=4)=-2.19, P=.09	-

Table 13: 5-fold cross-validation results across models on CLIP dataset and the t-test results among T5, MIMIC-T5, and SciFive+MIMIC-T5

HOC

Metrics		T5-Den	T5-Sup	FLAN-T5	MIMIC-T5	SciFive+MIMIC-T5
F1	Mean	84.81 ♠	84.76 ♣	84.79 ♦	82.75 ↓♠2.07, ↓♣2.02, ↓♦2.05	85.1 ↑♠0.29, ↑♣0.34, ↑♦0.31
	95% CI	[84.26, 85.49]	[84.38, 85.11]	[84.22, 85.26]	[82.53, 82.98]	[84.81, 85.4]
P	Mean	84.72 ♠	84.77 ♣	84.60 ♦	82.90 ↓♠1.82, ↓♣1.87, ↓♦1.7	85.04 ↑♠0.32, ↑♣0.27, ↑♦0.44
	95% CI	[84.04, 85.47]	[84.43, 85.11]	[83.94, 85.18]	[82.77, 83.04]	[84.67, 85.39]
R	Mean	84.91 ♠	84.76 ♣	84.99 ♦	82.60 ↓♠2.31, ↓♣2.16, ↓♦2.4	85.17 ↑♠0.26, ↑♣0.41, ↑♦0.18
	95% CI	[84.49, 85.52]	[84.28, 85.21]	[84.47, 85.4]	[82.26, 83.01]	[84.85, 85.52]

HOC-F1-T-test

Models	T5-Den	T5-Sup	FLAN-T5	MIMIC-T5	SciFive+MIMIC-T5
T5-Den-Large	-	-	-	-	-
T5-Sup-Large	t(df=4)=-0.09, P=.93	-	-	-	-
FLAN-T5-Large	t(df=4)=-0.05, P=.96	t(df=4)=0.07, P=.95	-	-	-
MIMIC-T5-Large	t(df=4)=-5.75, P=.005	t(df=4)=-6.86, P=.002	t(df=4)=-7.02, P=.002	-	-
SciFive+MIMIC-T5-Large	t(df=4)=1.05, P=.35	t(df=4)=1.31, P=.26	t(df=4)=1.08, P=.34	t(df=4)=10.06, P=¡.001	-

HOC-P-T-test

Models	T5-Den	T5-Sup	FLAN-T5	MIMIC-T5	SciFive+MIMIC-T5
T5-Den-Large	-	-	-	-	-
T5-Sup-Large	t(df=4)=0.1, P=.92	-	-	-	-
FLAN-T5-Large	t(df=4)=-0.29, P=.79	t(df=4)=-0.36, P=.73	-	-	-
MIMIC-T5-Large	t(df=4)=-4.3, P=.01	t(df=4)=-6.62, P=.003	t(df=4)=-5.46, P=.005	-	-
SciFive+MIMIC-T5-Large	t(df=4)=0.86, P=.44	t(df=4)=1.19, P=.30	t(df=4)=1.41, P=.23	t(df=4)=7.87, P=.001	-

HOC-R-T-test

Models	T5-Den	T5-Sup	FLAN-T5	MIMIC-T5	SciFive+MIMIC-T5
T5-Den-Large	-	-	-	-	-
T5-Sup-Large	t(df=4)=-0.28, P=.79	-	-	-	-
FLAN-T5-Large	t(df=4)=0.22, P=.84	t(df=4)=0.65, P=.55	-	-	-
MIMIC-T5-Large	t(df=4)=-6.84, P=.002	t(df=4)=-6.45, P=.003	t(df=4)=-8.09, P=.001	-	-
SciFive+MIMIC-T5-Large	t(df=4)=1.26, P=.27	t(df=4)=1.11, P=.33	t(df=4)=0.57, P=.60	t(df=4)=10.45, P=¡.001	-

Table 14: 5-fold cross-validation results across models on HOC dataset and the t-test results among T5, MIMIC-T5, and SciFive+MIMIC-T5

BC5CDR

Metrics		T5-Den	T5-Sup	FLAN-T5	MIMIC-T5	SciFive+MIMIC-T5
F1	Mean	82.90 ♠	82.94 ♣	83.30 ♦	81.17 ↓♠1.73, ↓♣1.77, ↓♦2.13	83.61 ↑♠0.71, ↑♣0.68, ↑♦0.31
	95% CI	[82.33, 83.38]	[82.51, 83.45]	[82.89, 83.82]	[80.92, 81.53]	[83.38, 83.82]
P	Mean	81.67 ♠	81.62 ♣	82.14 ♦	80.62 ↓♠1.05, ↓♣1.01, ↓♦1.52	83.07 ↑♠1.4, ↑♣1.44, ↑♦0.93
	95% CI	[80.77, 82.51]	[81.30, 81.96]	[81.53, 82.72]	[80.20, 80.94]	[82.51, 83.6]
R	Mean	84.19 ♠	84.29 ♣	84.50 ♦	81.73 ↓♠2.46, ↓♣2.56, ↓♦2.76	84.17 ↓♠0.02, ↓♣0.12, ↓♦0.32
	95% CI	[83.64, 84.63]	[83.55, 85.05]	[84.23, 84.93]	[81.23, 82.31]	[83.60, 84.82]

BC5CDR-F1-T-test

Models	T5-Den	T5-Sup	FLAN-T5	MIMIC-T5	SciFive+MIMIC-T5
T5-Den	-	-	-	-	-
T5-Sup	t(df=4)=0.06, P=.96	-	-	-	-
FLAN-T5	t(df=4)=0.82, P=.46	t(df=4)=2.16, P=.10	-	-	-
MIMIC-T5	t(df=4)=-9.49, P=¡.001	t(df=4)=-4.01, P=.02	t(df=4)=-5.29, P=.006	-	-
SciFive+MIMIC-T5	t(df=4)=2.38, P=.08	t(df=4)=1.7, P=.16	t(df=4)=0.76, P=.49	t(df=4)=14.45, P=¡.001	-

BC5CDR-P-T-test

Models	T5-Den	T5-Sup	FLAN-T5	MIMIC-T5	SciFive+MIMIC-T5
T5-Den	-	-	-	-	-
T5-Sup	t(df=4)=-0.08, P=.94	-	-	-	-
FLAN-T5	t(df=4)=0.71, P=.52	t(df=4)=2.37, P=.08	-	-	-
MIMIC-T5	t(df=4)=-3.07, P=.04	t(df=4)=-3.23, P=.03	t(df=4)=-3.09, P=.04	-	-
SciFive+MIMIC-T5	t(df=4)=2.08, P=.11	t(df=4)=3.41, P=.03	t(df=4)=1.73, P=.16	t(df=4)=6.04, P=.004	-

BC5CDR-R-T-test

Models	T5-Den	T5-Sup	FLAN-T5	MIMIC-T5	SciFive+MIMIC-T5
T5-Den	-	-	-	-	-
T5-Sup	t(df=4)=0.19, P=.86	-	-	-	-
FLAN-T5	t(df=4)=0.8, P=.47	t(df=4)=0.78, P=.48	-	-	-
MIMIC-T5	t(df=4)=-6.28, P=.003	t(df=4)=-3.88, P=.02	t(df=4)=-6.14, P=.004	-	-
SciFive+MIMIC-T5	t(df=4)=-0.07, P=.95	t(df=4)=-0.17, P=.88	t(df=4)=-0.61, P=.57	t(df=4)=10.35, P=¡.001	-

Table 15: 5-fold cross-validation results across models on BC5CDR-disease dataset and the t-test results among T5, MIMIC-T5, and SciFive+MIMIC-T5

NCBI

Metrics		T5-Den	T5-Sup	FLAN-T5	MIMIC-T5	SciFive+MIMIC-T5
F1	Mean	85.08 ♠	84.58 ♣	84.44 ♦	79.10 ↓♠5.98, ↓♣5.49, ↓♦5.35	85.43 ↑♠0.36, ↑♣0.85, ↑♦0.99
	95% CI	[84.61, 85.46]	[84.22, 85.02]	[83.94, 85.03]	[78.28, 79.94]	[84.89, 86.06]
P	Mean	83.49 ♠	83.58 ♣	82.20 ♦	78.35 ↓♠5.14, ↓♣5.24, ↓♦3.85	83.80 ↑♠0.31, ↑♣0.21, ↑♦1.60
	95% CI	[82.78, 84.16]	[83.13, 84.12]	[81.59, 83.04]	[77.57, 78.87]	[82.95, 84.68]
R	Mean	86.73 ♠	85.61 ♣	86.82 ♦	79.87 ↓♠6.86, ↓♣5.74, ↓♦6.95	87.14 ↑♠0.41, ↑♣1.54, ↑♦0.32
	95% CI	[86.31, 87.22]	[85.16, 86.15]	[86.31, 87.30]	[78.70, 81.30]	[86.86, 87.50]

NCBI-F1-T-test

Models	T5-Den	T5-Sup	FLAN-T5	MIMIC-T5	SciFive+MIMIC-T5
T5-Den	-	-	-	-	-
T5-Sup	t(df=4)=-0.95, P=.40	-	-	-	-
FLAN-T5	t(df=4)=-1.73, P=.16	t(df=4)=-0.33, P=.76	-	-	-
MIMIC-T5	t(df=4)=-12.52, P=¡.001	t(df=4)=-10.1, P=¡.001	t(df=4)=-11.66, P=¡.001	-	-
SciFive+MIMIC-T5	t(df=4)=0.94, P=.40	t(df=4)=1.71, P=.16	t(df=4)=1.62, P=.18	t(df=4)=8.65, P=¡.001	-

NCBI-P-T-test

Models	T5-Den	T5-Sup	FLAN-T5	MIMIC-T5	SciFive+MIMIC-T5
T5-Den	-	-	-	-	-
T5-Sup	t(df=4)=0.14, P=.90	-	-	-	-
FLAN-T5	t(df=4)=-1.88, P=.13	t(df=4)=-3.52, P=.02	-	-	-
MIMIC-T5	t(df=4)=-7.08, P=.002	t(df=4)=-14.76, P=¡.001	t(df=4)=-8.28, P=.001	-	-
SciFive+MIMIC-T5	t(df=4)=0.74, P=.50	t(df=4)=0.33, P=.76	t(df=4)=1.93, P=.13	t(df=4)=6.65, P=.003	-

NCBI-R-T-test

Models	T5-Den	T5-Sup	FLAN-T5	MIMIC-T5	SciFive+MIMIC-T5
T5-Den	-	-	-	-	-
T5-Sup	t(df=4)=-2.0, P=.12	-	-	-	-
FLAN-T5	t(df=4)=0.38, P=.72	t(df=4)=2.71, P=.05	-	-	-
MIMIC-T5	t(df=4)=-10.43, P=¡.001	t(df=4)=-7.16, P=.002	t(df=4)=-12.53, P=¡.001	-	-
SciFive+MIMIC-T5	t(df=4)=1.07, P=.34	t(df=4)=3.25, P=.03	t(df=4)=0.72, P=.51	t(df=4)=8.63, P=¡.001	-

Table 16: 5-fold cross-validation results across models on NCBI-disease dataset and the t-test results among T5, MIMIC-T5, and SciFive+MIMIC-T5

MIMIC-IV-Credibility & Obstinacy

Metrics		T5-Sup	FLAN-T5	MIMIC-T5	SciFive+MIMIC-T5
F1	Mean	76.04 ♣	77.44 ♦	74.98 ↓♣1.06, ↓♦2.46	74.65 ↓♣1.39, ↓♦2.79
	95% CI	[74.14, 77.94]	[74.96, 79.34]	[72.11, 77.75]	[70.59, 78.82]

MIMIC-IV-Compliance

Metrics		T5-Sup	FLAN-T5	MIMIC-T5	SciFive+MIMIC-T5
F1	Mean	92.94 ♣	92.93 ♦	91.57 ↓♣1.37, ↓♦1.36	90.98 ↓♣1.96, ↓♦1.95
	95% CI	[91.74, 94.18]	[91.27, 94.48]	[90.38, 92.70]	[89.90, 92.12]

MIMIC-IV-Descriptors

Metrics		T5-Sup	FLAN-T5	MIMIC-T5	SciFive+MIMIC-T5
F1	Mean	85.73 ♣	86.31 ♦	84.76 ↓♣0.97, ↓♦1.55	86.28 ↑♣0.54, ↓♦0.03
	95% CI	[83.45, 87.99]	[85.32, 87.03]	[81.90, 88.15]	[85.00, 87.79]

MIMIC-IV-Credibility & Obstinacy-F1-T-test

Models	T5-Sup	FLAN-T5	MIMIC-T5	SciFive+MIMIC-T5
T5-Sup	-	-	-	-
FLAN-T5	t(df=4)=0.64, P=.56	-	-	-
MIMIC-T5	t(df=4)=-0.52, P=.63	t(df=4)=-1.05, P=.35	-	-
SciFive+MIMIC-T5	t(df=4)=-0.81, P=.46	t(df=4)=-1.08, P=.34	t(df=4)=-0.12, P=.91	-

MIMIC-IV-Compliance-F1-T-test

Models	T5-Sup	FLAN-T5	MIMIC-T5	SciFive+MIMIC-T5
T5-Sup	-	-	-	-
FLAN-T5	t(df=4)=-0.0, P=1.00	-	-	-
MIMIC-T5	t(df=4)=-1.2, P=.30	t(df=4)=-1.2, P=.30	-	-
SciFive+MIMIC-T5	t(df=4)=-1.5, P=.21	t(df=4)=-1.93, P=.13	t(df=4)=-0.84, P=.45	-

MIMIC-IV-Descriptors-F1-T-test

Models	T5-Sup	FLAN-T5	MIMIC-T5	SciFive+MIMIC-T5
T5-Sup	-	-	-	-
FLAN-T5	t(df=4)=0.37, P=.73	-	-	-
MIMIC-T5	t(df=4)=-0.74, P=.50	t(df=4)=-0.9, P=.42	-	-
SciFive+MIMIC-T5	t(df=4)=0.33, P=.76	t(df=4)=-0.04, P=.97	t(df=4)=0.75, P=.50	-

Table 17: 5-fold cross-validation results across models on stigmatizing dataset from MIMIC-IV and the t-test results among T5-Sup, FLAN-T5, MIMIC-T5, and SciFive+MIMIC-T5

Hospital System-Credibility & Obstinacy

Metrics		T5-Sup	FLAN-T5	MIMIC-T5	SciFive+MIMIC-T5
F1	Mean	88.21 ♣	90.38 ♦	88.07 ↓ ♣0.15, ↓ ♦2.31	87.31 ↓ ♣0.9, ↓ ♦3.07
	95% CI	[86.12, 90.59]	[88.26, 92.33]	[83.11, 94.14]	[82.11, 91.76]

Hospital System-Compliance

Metrics		T5-Sup	FLAN-T5	MIMIC-T5	SciFive+MIMIC-T5
F1	Mean	86.71 ♣	88.03 ♦	84.72 ↓ ♣2.0, ↓ ♦3.31	87.21 ↑ ♣0.49, ↓ ♦0.82
	95% CI	[85.52, 88.19]	[86.84, 88.92]	[80.99, 88.67]	[86.82, 87.63]

Hospital System-Descriptors

Metrics		T5-Sup	FLAN-T5	MIMIC-T5	SciFive+MIMIC-T5
F1	Mean	90.58 ♣	86.82 ♦	88.85 ↓ ♣1.73, ↑ ♦2.03	89.98 ↓ ♣0.6, ↑ ♦3.17
	95% CI	[89.25, 91.37]	[79.94, 90.69]	[88.24, 89.46]	[89.21, 90.67]

Hospital System-Credibility & Obstinacy-F1-T-test

Models	T5-Sup	FLAN-T5	MIMIC-T5	SciFive+MIMIC-T5
T5-Sup	-	-	-	-
FLAN-T5	t(df=4)=1.02, P=.37	-	-	-
MIMIC-T5	t(df=4)=-0.07, P=.95	t(df=4)=-0.63, P=.56	-	-
SciFive+MIMIC-T5	t(df=4)=-0.34, P=.75	t(df=4)=-1.48, P=.21	t(df=4)=-0.23, P=.83	-

Hospital System-Compliance-F1-T-test

Models	T5-Sup	FLAN-T5	MIMIC-T5	SciFive+MIMIC-T5
T5-Sup	-	-	-	-
FLAN-T5	t(df=4)=1.75, P=.16	-	-	-
MIMIC-T5	t(df=4)=-1.04, P=.36	t(df=4)=-1.52, P=.20	-	-
SciFive+MIMIC-T5	t(df=4)=0.52, P=.63	t(df=4)=-0.91, P=.42	t(df=4)=1.06, P=.35	-

Hospital System-Descriptors-F1-T-test

Models	T5-Sup	FLAN-T5	MIMIC-T5	SciFive+MIMIC-T5
T5-Sup	-	-	-	-
FLAN-T5	t(df=4)=-0.99, P=.38	-	-	-
MIMIC-T5	t(df=4)=-2.5, P=.07	t(df=4)=0.6, P=.58	-	-
SciFive+MIMIC-T5	t(df=4)=-0.63, P=.57	t(df=4)=0.9, P=.42	t(df=4)=1.71, P=.16	-

Table 18: 5-fold cross-validation results across models on stigmatizing dataset from Hospital System and the t-test results among T5-Sup, FLAN-T5, MIMIC-T5, and SciFive+MIMIC-T5

MIMIC-IV-Credibility & Obstinacy - 25% downsampling data

Metrics		T5-Sup	FLAN-T5	MIMIC-T5	SciFive+MIMIC-T5
F1	Mean	63.81 ♣	71.92 ♦	64.44 ↑♣0.63, ↓♦7.48	57.67 ↓♣6.15, ↓♦14.25
	95% CI	[62.15, 65.77]	[69.6, 74.16]	[61.69, 66.63]	[54.39, 61.06]

MIMIC-IV-Compliance- 25% downsampling data

Metrics		T5-Sup	FLAN-T5	MIMIC-T5	SciFive+MIMIC-T5
F1	Mean	89.6 ♣	91.44 ♦	86.02 ↓♣3.58, ↓♦5.42	89.58 ↓♣0.02, ↓♦1.86
	95% CI	[86.8, 92.19]	[89.15, 93.09]	[84.63, 87.5]	[87.19, 92.12]

MIMIC-IV-Descriptors- 25% downsampling data

Metrics		T5-Sup	FLAN-T5	MIMIC-T5	SciFive+MIMIC-T5
F1	Mean	76.83 ♣	78.21 ♦	75.31 ↓♣1.52, ↓♦2.9	82.46 ↑♣5.63, ↑♦4.25
	95% CI	[72.49, 79.4]	[77.29, 79.19]	[72.32, 77.52]	[82.02, 82.85]

MIMIC-IV-Credibility & Obstinacy-F1-T-test

Models	T5-Sup	FLAN-T5	MIMIC-T5	SciFive+MIMIC-T5
T5-Sup	-	-	-	-
FLAN-T5	t(df=4)=3.37, P=.03	-	-	-
MIMIC-T5	t(df=4)=0.42, P=.69	t(df=4)=-2.66, P=.06	-	-
SciFive+MIMIC-T5	t(df=4)=-3.08, P=.04	t(df=4)=-4.9, P=.008	t(df=4)=-2.42, P=.07	-

MIMIC-IV-Compliance-F1-T-test

Models	T5-Sup	FLAN-T5	MIMIC-T5	SciFive+MIMIC-T5
T5-Sup	-	-	-	-
FLAN-T5	t(df=4)=1.2, P=.30	-	-	-
MIMIC-T5	t(df=4)=-4.16, P=.01	t(df=4)=-4.54, P=.01	-	-
SciFive+MIMIC-T5	t(df=4)=-0.01, P=.99	t(df=4)=-1.47, P=.21	t(df=4)=1.9, P=.13	-

MIMIC-IV-Descriptors-F1-T-test

Models	T5-Sup	FLAN-T5	MIMIC-T5	SciFive+MIMIC-T5
T5-Sup	-	-	-	-
FLAN-T5	t(df=4)=0.72, P=.51	-	-	-
MIMIC-T5	t(df=4)=-0.82, P=.46	t(df=4)=-1.67, P=.17	-	-
SciFive+MIMIC-T5	t(df=4)=2.66, P=.06	t(df=4)=5.61, P=.005	t(df=4)=4.64, P=.010	-

Table 19: 5-fold cross-validation results across models on stigmatizing dataset from MIMIC-IV with 25% downsampled training data and the t-test results among T5-Sup, FLAN-T5, MIMIC-T5, and SciFive+MIMIC-T5

MIMIC-IV-Credibility & Obstinacy - 5% downsampling data					
Metrics		T5-Sup	FLAN-T5	MIMIC-T5	SciFive+MIMIC-T5
F1	Mean	51.62 ♣	66.33 ♦	58.58 ↑♣6.96, ↓♦7.76	53.14 ↑♣1.52, ↓♦13.2
	95% CI	[42.57, 58.72]	[64.37, 67.61]	[53.25, 64.66]	[49.77, 56.59]
MIMIC-IV-Compliance- 5% downsampling data					
Metrics		T5-Sup	FLAN-T5	MIMIC-T5	SciFive+MIMIC-T5
F1	Mean	70.2 ♣	82.33 ♦	72.98 ↑♣2.78, ↓♦9.35	69.36 ↓♣0.84, ↓♦12.97
	95% CI	[67.48, 72.75]	[74.22, 87.16]	[70.25, 76.61]	[55.76, 79.52]
MIMIC-IV-Descriptors- 5% downsampling data					
Metrics		T5-Sup	FLAN-T5	MIMIC-T5	SciFive+MIMIC-T5
F1	Mean	61.5 ♣	63.16 ♦	62.5 ↑♣0.99, ↓♦0.66	67.3 ↑♣5.8, ↑♦4.14
	95% CI	[59.2, 63.54]	[49.6, 71.45]	[60.26, 64.6]	[64.92, 69.49]

MIMIC-IV-Credibility & Obstinacy-F1-T-test				
Models	T5-Sup	FLAN-T5	MIMIC-T5	SciFive+MIMIC-T5
T5-Sup	-	-	-	-
FLAN-T5	t(df=4)=2.78, P=.05	-	-	-
MIMIC-T5	t(df=4)=1.35, P=.25	t(df=4)=-2.55, P=.06	-	-
SciFive+MIMIC-T5	t(df=4)=0.27, P=.80	t(df=4)=-5.36, P=.006	t(df=4)=-1.25, P=.28	-
MIMIC-IV-Compliance-F1-T-test				
Models	T5-Sup	FLAN-T5	MIMIC-T5	SciFive+MIMIC-T5
T5-Sup	-	-	-	-
FLAN-T5	t(df=4)=2.4, P=.07	-	-	-
MIMIC-T5	t(df=4)=0.93, P=.40	t(df=4)=-1.83, P=.14	-	-
SciFive+MIMIC-T5	t(df=4)=-0.11, P=.92	t(df=4)=-1.38, P=.24	t(df=4)=-0.36, P=.73	-
MIMIC-IV-Descriptors-F1-T-test				
Models	T5-Sup	FLAN-T5	MIMIC-T5	SciFive+MIMIC-T5
T5-Sup	-	-	-	-
FLAN-T5	t(df=4)=0.25, P=.81	-	-	-
MIMIC-T5	t(df=4)=1.7, P=.16	t(df=4)=-0.11, P=.92	-	-
SciFive+MIMIC-T5	t(df=4)=9.13, P=¡.001	t(df=4)=0.58, P=.59	t(df=4)=5.05, P=.007	-

Table 20: 5-fold cross-validation results across models on stigmatizing dataset from MIMIC-IV with 5% downsampled training data and the t-test results among T5-Sup, FLAN-T5, MIMIC-T5, and SciFive+MIMIC-T5

MIMIC-IV-Credibility & Obstinacy - 1% downsampling data

Metrics		T5-Sup	FLAN-T5	MIMIC-T5	SciFive+MIMIC-T5
F1	Mean	27.7 ♣	51.02 ♦	37.08 ↑ ♣9.38, ↓ ♦13.94	33.49 ↑ ♣5.78, ↓ ♦17.54
	95% CI	[22.94, 32.17]	[45.89, 55.52]	[29.74, 42.96]	[27.39, 40.18]

MIMIC-IV-Compliance- 1% downsampling data

Metrics		T5-Sup	FLAN-T5	MIMIC-T5	SciFive+MIMIC-T5
F1	Mean	54.32 ♣	66.18 ♦	42.1 ↓ ♣12.22, ↓ ♦24.09	50.44 ↓ ♣3.88, ↓ ♦15.75
	95% CI	[49.27, 59.91]	[65.28, 66.94]	[38.57, 46.91]	[43.57, 56.06]

MIMIC-IV-Descriptors- 1% downsampling data

Metrics		T5-Sup	FLAN-T5	MIMIC-T5	SciFive+MIMIC-T5
F1	Mean	43.41 ♣	47.85 ♦	29.09 ↓ ♣14.32, ↓ ♦18.75	35.14 ↓ ♣8.27, ↓ ♦12.7
	95% CI	[40.3, 46.32]	[42.16, 55.4]	[25.76, 33.03]	[29.11, 41.98]

MIMIC-IV-Credibility & Obstinacy-F1-T-test

Models	T5-Sup	FLAN-T5	MIMIC-T5	SciFive+MIMIC-T5
T5-Sup	-	-	-	-
FLAN-T5	t(df=4)=5.92, P=.004	-	-	-
MIMIC-T5	t(df=4)=3.43, P=.03	t(df=4)=-2.43, P=.07	-	-
SciFive+MIMIC-T5	t(df=4)=1.77, P=.15	t(df=4)=-3.58, P=.02	t(df=4)=-1.76, P=.15	-

MIMIC-IV-Compliance-F1-T-test

Models	T5-Sup	FLAN-T5	MIMIC-T5	SciFive+MIMIC-T5
T5-Sup	-	-	-	-
FLAN-T5	t(df=4)=3.46, P=.03	-	-	-
MIMIC-T5	t(df=4)=-8.5, P=.001	t(df=4)=-9.03, P=¡.001	-	-
SciFive+MIMIC-T5	t(df=4)=-1.14, P=.32	t(df=4)=-4.2, P=.01	t(df=4)=2.98, P=.04	-

MIMIC-IV-Descriptors-F1-T-test

Models	T5-Sup	FLAN-T5	MIMIC-T5	SciFive+MIMIC-T5
T5-Sup	-	-	-	-
FLAN-T5	t(df=4)=1.23, P=.29	-	-	-
MIMIC-T5	t(df=4)=-3.94, P=.02	t(df=4)=-4.89, P=.008	-	-
SciFive+MIMIC-T5	t(df=4)=-1.63, P=.18	t(df=4)=-2.06, P=.11	t(df=4)=2.25, P=.09	-

Table 21: 5-fold cross-validation results across models on stigmatizing dataset from MIMIC-IV with 1% downsampled training data and the t-test results among T5-Sup, FLAN-T5, MIMIC-T5, and SciFive+MIMIC-T5

Hospital System-Credibility & Obstinacy - 25% downsampling data					
Metrics		T5-Sup	FLAN-T5	MIMIC-T5	SciFive+MIMIC-T5
F1	Mean	85.1 ♣	86.59 ♦	75.9 ↓ ♣9.19, ↓ ♦10.69	84.7 ↓ ♣0.4, ↓ ♦1.89
	95% CI	[81.26, 87.88]	[83.57, 88.98]	[71.71, 80.43]	[80.58, 88.32]
Hospital System-Compliance- 25% downsampling data					
Metrics		T5-Sup	FLAN-T5	MIMIC-T5	SciFive+MIMIC-T5
F1	Mean	82.52 ♣	85.3 ♦	79.32 ↓ ♣3.2, ↓ ♦5.98	81.53 ↓ ♣0.99, ↓ ♦3.77
	95% CI	[78.41, 84.92]	[82.98, 87.65]	[76.29, 83.15]	[75.26, 86.29]
Hospital System-Descriptors- 25% downsampling data					
Metrics		T5-Sup	FLAN-T5	MIMIC-T5	SciFive+MIMIC-T5
F1	Mean	86.45 ♣	85.98 ♦	87.19 ↑ ♣0.75, ↑ ♦1.22	86.51 ↑ ♣0.06, ↑ ♦0.53
	95% CI	[85.24, 88.16]	[85.14, 86.69]	[85.32, 89.41]	[85.09, 88.14]

Hospital System-Credibility & Obstinacy-F1-T-test				
Models	T5-Sup	FLAN-T5	MIMIC-T5	SciFive+MIMIC-T5
T5-Sup	-	-	-	-
FLAN-T5	t(df=4)=0.54, P=.62	-	-	-
MIMIC-T5	t(df=4)=-2.36, P=.08	t(df=4)=-2.94, P=.04	-	-
SciFive+MIMIC-T5	t(df=4)=-0.15, P=.89	t(df=4)=-0.54, P=.62	t(df=4)=4.8, P=.009	-
Hospital System-Compliance-F1-T-test				
Models	T5-Sup	FLAN-T5	MIMIC-T5	SciFive+MIMIC-T5
T5-Sup	-	-	-	-
FLAN-T5	t(df=4)=1.79, P=.15	-	-	-
MIMIC-T5	t(df=4)=-1.57, P=.19	t(df=4)=-2.98, P=.04	-	-
SciFive+MIMIC-T5	t(df=4)=-0.78, P=.48	t(df=4)=-1.37, P=.24	t(df=4)=0.88, P=.43	-
Hospital System-Descriptors-F1-T-test				
Models	T5-Sup	FLAN-T5	MIMIC-T5	SciFive+MIMIC-T5
T5-Sup	-	-	-	-
FLAN-T5	t(df=4)=-0.43, P=.69	-	-	-
MIMIC-T5	t(df=4)=0.53, P=.62	t(df=4)=1.05, P=.35	-	-
SciFive+MIMIC-T5	t(df=4)=0.08, P=.94	t(df=4)=0.49, P=.65	t(df=4)=-0.5, P=.65	-

Table 22: 5-fold cross-validation results across models on stigmatizing dataset from Hospital System with 25% downsampled training data and the t-test results among T5-Sup, FLAN-T5, MIMIC-T5, and SciFive+MIMIC-T5

Hospital System-Credibility & Obstinacy - 5% downsampling data

Metrics		T5-Sup	FLAN-T5	MIMIC-T5	SciFive+MIMIC-T5
F1	Mean	48.94 ♣	75.66 ♦	48.88 ↓ ♣0.06, ↓ ♦26.78	50.2 ↑ ♣1.26, ↓ ♦25.46
	95% CI	[43.03, 55.34]	[67.77, 82.11]	[46.72, 50.4]	[48.32, 52.7]

Hospital System - Compliance - 5% downsampling data

Metrics		T5-Sup	FLAN-T5	MIMIC-T5	SciFive+MIMIC-T5
F1	Mean	58.62 ♣	71.09 ♦	51.82 ↓ ♣6.8, ↓ ♦19.27	56.84 ↓ ♣1.78, ↓ ♦14.25
	95% CI	[55.48, 60.66]	[63.96, 76.14]	[49.19, 54.12]	[52.65, 60.79]

Hospital System - Descriptors - 5% downsampling data

Metrics		T5-Sup	FLAN-T5	MIMIC-T5	SciFive+MIMIC-T5
F1	Mean	69.95 ♣	79.82 ♦	69.71 ↓ ♣0.24, ↓ ♦10.11	76.41 ↑ ♣6.46, ↓ ♦3.41
	95% CI	[68.13, 71.33]	[77.17, 82.3]	[63.77, 73.81]	[75.0, 77.68]

Hospital System-Credibility & Obstinacy-F1-T-test

Models	T5-Sup	FLAN-T5	MIMIC-T5	SciFive+MIMIC-T5
T5-Sup	-	-	-	-
FLAN-T5	t(df=4)=3.89, P=.02	-	-	-
MIMIC-T5	t(df=4)=-0.02, P=.99	t(df=4)=-5.36, P=.006	-	-
SciFive+MIMIC-T5	t(df=4)=0.42, P=.70	t(df=4)=-4.42, P=.01	t(df=4)=0.93, P=.40	-

Hospital System-Compliance-F1-T-test

Models	T5-Sup	FLAN-T5	MIMIC-T5	SciFive+MIMIC-T5
T5-Sup	-	-	-	-
FLAN-T5	t(df=4)=2.52, P=.07	-	-	-
MIMIC-T5	t(df=4)=-4.8, P=.009	t(df=4)=-5.18, P=.007	-	-
SciFive+MIMIC-T5	t(df=4)=-1.17, P=.31	t(df=4)=-3.12, P=.04	t(df=4)=2.5, P=.07	-

Hospital System-Descriptors-F1-T-test

Models	T5-Sup	FLAN-T5	MIMIC-T5	SciFive+MIMIC-T5
T5-Sup	-	-	-	-
FLAN-T5	t(df=4)=6.01, P=.004	-	-	-
MIMIC-T5	t(df=4)=-0.07, P=.95	t(df=4)=-2.23, P=.09	-	-
SciFive+MIMIC-T5	t(df=4)=4.25, P=.01	t(df=4)=-3.41, P=.03	t(df=4)=1.57, P=.19	-

Table 23: 5-fold cross-validation results across models on stigmatizing dataset from Hospital System with 5% downsampled training data and the t-test results among T5-Sup, FLAN-T5, MIMIC-T5, and SciFive+MIMIC-T5

Hospital System-Credibility & Obstinacy - 1% downsampling data

Metrics		T5-Sup	FLAN-T5	MIMIC-T5	SciFive+MIMIC-T5
F1	Mean	23.29 ♣	54.45 ♦	34.95 ↑ ♣11.66, ↓ ♦19.5	31.79 ↑ ♣8.5, ↓ ♦22.66
	95% CI	[17.55, 27.46]	[45.76, 61.91]	[26.12, 46.24]	[25.45, 38.43]

Hospital System-Compliance- 1% downsampling data

Metrics		T5-Sup	FLAN-T5	MIMIC-T5	SciFive+MIMIC-T5
F1	Mean	45.44 ♣	70.6 ♦	41.92 ↓ ♣3.52, ↓ ♦28.68	38.52 ↓ ♣6.92, ↓ ♦32.08
	95% CI	[40.34, 52.88]	[68.81, 72.05]	[36.33, 46.04]	[33.42, 42.52]

Hospital System-Descriptors- 1% downsampling data

Metrics		T5-Sup	FLAN-T5	MIMIC-T5	SciFive+MIMIC-T5
F1	Mean	53.78 ♣	58.36 ♦	44.52 ↓ ♣9.26, ↓ ♦13.84	55.52 ↑ ♣1.74, ↓ ♦2.83
	95% CI	[45.07, 60.86]	[52.25, 64.75]	[32.8, 54.06]	[53.29, 58.91]

Hospital System-Credibility & Obstinacy-F1-T-test

Models	T5-Sup	FLAN-T5	MIMIC-T5	SciFive+MIMIC-T5
T5-Sup	-	-	-	-
FLAN-T5	t(df=4)=4.77, P=.009	-	-	-
MIMIC-T5	t(df=4)=1.39, P=.24	t(df=4)=-5.22, P=.006	-	-
SciFive+MIMIC-T5	t(df=4)=1.49, P=.21	t(df=4)=-6.22, P=.003	t(df=4)=-0.56, P=.61	-

Hospital System-Compliance-F1-T-test

Models	T5-Sup	FLAN-T5	MIMIC-T5	SciFive+MIMIC-T5
T5-Sup	-	-	-	-
FLAN-T5	t(df=4)=6.2, P=.003	-	-	-
MIMIC-T5	t(df=4)=-0.85, P=.44	t(df=4)=-9.06, P=¡.001	-	-
SciFive+MIMIC-T5	t(df=4)=-1.63, P=.18	t(df=4)=-14.93, P=¡.001	t(df=4)=-1.02, P=.36	-

Hospital System-Descriptors-F1-T-test

Models	T5-Sup	FLAN-T5	MIMIC-T5	SciFive+MIMIC-T5
T5-Sup	-	-	-	-
FLAN-T5	t(df=4)=0.63, P=.56	-	-	-
MIMIC-T5	t(df=4)=-2.76, P=.05	t(df=4)=-1.67, P=.17	-	-
SciFive+MIMIC-T5	t(df=4)=0.44, P=.68	t(df=4)=-0.75, P=.50	t(df=4)=2.24, P=.09	-

Table 24: 5-fold cross-validation results across models on stigmatizing dataset from Hospital System with 1% downsampled training data and the t-test results among T5-Sup, FLAN-T5, MIMIC-T5, and SciFive+MIMIC-T5