Multi-Task Training with In-Domain Language Models for Diagnostic Reasoning

Brihat Sharma[1], Yanjun Gao[1], Timothy Miller[2], Matthew M. Churpek[1],
Majid Afshar[1] and Dmitriy Dligach[3]
[1]University of Wisconsin-Madison,
[2]Boston Children's Hospital and Harvard Medical School, [3]Loyola University Chicago
[1]bsharma25@wisc.edu,
[1]{ygao, mchurpek, mafshar} @medicine.wisc.edu,
[2]Timothy.Miller@childrens.harvard.edu, [3]ddligach@luc.edu

Abstract

Generative artificial intelligence (AI) is a promising direction for augmenting clinical diagnostic decision support and reducing diagnostic errors, a leading contributor to medical errors. To further the development of clinical AI systems, the Diagnostic Reasoning Benchmark (DR.BENCH) was introduced as a comprehensive generative AI framework, comprised of six tasks representing key components in clinical reasoning. We present a comparative analysis of in-domain versus out-of-domain language models as well as multi-task versus single task training with a focus on the problem summarization task in DR.BENCH (Gao et al., 2023). We demonstrate that a multitask, clinically-trained language model outperforms its general domain counterpart by a large margin, establishing a new state-of-the-art performance, with a ROUGE-L score of 28.55. This research underscores the value of domain-specific training for optimizing clinical diagnostic reasoning tasks.

1 Introduction

The electronic health record (EHR) contains daily progress notes authored by healthcare providers to represent the daily changes in care plans for their patients, including an updated list of active diagnoses. The daily progress note is one of the most important note types in the EHR and contains the daily subjective and objective details in the patient's care, which is summarized into an assessment of the overall leading diagnoses with a treatment plan section (Gao et al., 2022b). However, note bloat is a common phenomenon in medical documentation intermixed with billing requirements, non-diagnostic information, and copy and paste from prior notes (Rule et al., 2021). These additional documentation practices contribute to provider burnout and cognitive overload (Gardner et al., 2018). Problem-based charting is important
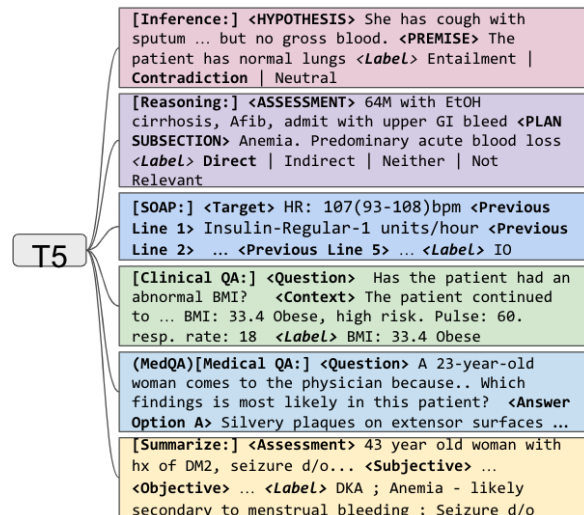

Figure 1: Training T5 with multi-task setup with six tasks from DR.BENCH (Gao et al., 2023)

to improve care throughput and help reduce diagnostic errors (Wright et al., 2012).

The medical reasoning process is complex and incorporates medical knowledge representation with analytical and experiential knowledge (Bowen, 2006). Patel and Groen developed a theory from the AI literature that experts use "forward-reasoning" from data to diagnosis 1986. The recently released benchmark DR.BENCH (Diagnostic Reasoning Benchmark) is intended to assess the ability of AI models to perform such reasoning, with multiple component tasks including diagnostic reasoning with EHR data for experiential knowledge, medical exams for knowledge representation, progress note structure prediction, and problem summarization tasks that included both extractive and abstractive medical diagnoses (Gao et al., 2023).

In this work, we focus primarily on the problem summarization task from the DR.BENCH suite, but with the hypothesis that using all tasks in DR.BENCH would improve the problem summarization task over the problem summarization task being fine-tuned alone. We make use of the T5 fam-

ily of sequence-to-sequence language models, (Raffel et al., 2020), which are first pretrained on a large unlabeled dataset and then finetuned on specific multiple downstream tasks. The text-to-text approach in our experiment makes it possible to perform multi-task training. Hence, the T5 models were ideal for experimenting with single and multi-task techniques.

Further, we experimented with a recently developed clinically-trained T5 model to quantify the value of in-domain pretraining data (Lehman and Johnson, 2023). We make our software publically available at https://git.doit.wisc.edu/smph-public/dom/uw-icu-data-science-lab-public/drbench.

2 Related Work

In the clinical domain, biomedical text summarization is a growing field. Common approaches to text summarization include feature-based methods (Patel et al., 2019), fine-tuning large language models (Lewis et al., 2020), and domain adaptation with fine-tuning methods (Xie et al., 2023). Researchers developed clinical methods for summarization from progress notes but these methods were restricted to specific diseases such as diabetes and hypertension (Liang et al., 2019). Moreover, these methods for summarization were more extractive than abstractive, using a combination of heuristics rules and deep learning techniques, and did not use large language models (Liang et al., 2019). In another work, an extractive-abstractive approach was used where meaningful sentences were extracted from the clinical notes first; these sentences were then fed into the transformer model for abstractive summarization (Pilault et al., 2020). Unfortunately, the transformer model frequently produced hallucinated outputs, and was not coherent when compared to the ground truth (Pilault et al., 2020). In a similar extractive-abstractive approach, researchers used a pointer generator network to generate a note summary cluster and a language model such as T5 to generate an abstractive summary (Krishna et al., 2021). None of these approaches used multi-task training or focused on clinically trained encoder-decoder since clinical T5

was only recently introduced. Prior work has not addressed the challenge of abstractive reasoning, or they used a two-step process to create abstractions. Recently, researchers used domain adaptive T5 model trained on the biomedical dataset but did not experiment with multi-task settings (Gao et al., 2023).

3 Methods

3.1 Dataset

In our experiments, we used DR.BENCH (Gao et al., 2023), a recently introduced benchmark designed to evaluate diagnostic reasoning capabilities of generative language models. DR.BENCH consists of three categories of tasks (two tasks per category), as shown in Figure 1. From top to the bottom, the categories and six tasks are: **Medical Knowledge Representation:** (1) Medical Natural Language Inference (MedNLI) task that considered sentence pairs with the objective to determine whether the hypothesis sentence could be inferred from the premise sentence (Shivade, 2019) (14,049 sentence pairs total); (2) Assessment and Plan Reasoning (A/P) task whose objective was to label relations between the assessment and treatment plan sections (5,897 samples). **Clinical Evidence Understanding and Integration:** (1) Electronic Medical Records Question Answering (emrQA) whose objective was to answer questions based on discharge summaries (53,199 questions total) (Pampari et al., 2018); (2) Progress Note Section Labeling task whose objective was to labels SOAP sections in progress notes (134,089 samples) (Gao et al., 2022a). **Diagnosis Generation and Summarization:** (1) Medical Board Exam Question Answering (MedQA) task that consisted of medical board exam question-answer pairs (12,725 pairs) (Jin et al., 2021); (2) Problem Summarization (ProbSumm) task whose goal was to produce the list of relevant problems and diagnoses based on the input that consisted of the SOAP sections of progress notes (2,783 samples).

In this work, we focused primarily on the problem summarization task, which was the most difficult but also believed to be the most impactful of the six DR.BENCH tasks for downstream clinical application.

3.2 Experimental Setup

In our experiments, we used six generative language models, all based on the Text-To-Text Trans-

---

[1] PubMed is a large open source biomedical and lifescience database consists of 35 million citation and abstract, and PMC (PubMed Central) consists of full articles. MIMIC-III and MIMIC-IV (Medical Information Mart for Intensive Care) are databases consisting of de-identified datasets from Beth Israel Deaconess Medical Center

Model	Training Corpus	Initialization	Citation
T5 220M T5 770M	Colossal Clean Crawled Corpus (C4)	Random Random	(Raffel et al., 2020)
SCIFIVE 220M SCIFIVE 770M	C4 + PubMed (abstracts) + PMC	T5 220M T5 770M	(Phan et al., 2021)
CLINICAL-T5 220M CLINICAL-T5 770M	MIMIC-III + MIMIC-IV	T5 220M Random	(Lehman and Johnson, 2023)

Table 1: T5 pretrained models used in the experiments. [1]

fer Transformer (T5) model (Raffel et al., 2020). The text-to-text paradigm utilized by T5 was a natural choice for our stated goal of exploring multi-task learning: transforming T5 into a multi-task learner simply involved prefixing individual task instances with a task-specific prompt after which the model could be trained using the standard cross-entropy loss.

Table 1 provides details about the models. We compared a multi-task scenario in which T5 variants were fine-tuned on all DR.BENCH tasks and a single-task scenario in which T5 was fine-tuned on the problem summarization task only. We trained T5 models as follows:

Single-task training: In single-task training for problem summarization, we used the text of the assessment, subjective and objective sections of the progress notes as input and trained T5 to generate the list of problems and diagnoses.

Multi-task training: In multi-task training, we combined all DR.BENCH tasks into a single dataset and trained T5 to generate task-specific output given task-specific input. Training examples of each task were prefixed with a task-specific prompt. The open-book setting only was used for MedQA. The rest of preprocessing follows (Gao et al., 2023).

To enable comparison with existing work (Gao et al., 2023) we used ROUGE-L score (Lin, 2004) as our evaluation metric. ROUGE-L uses the longest common subsequence statistics to compare model outputs. A resampling technique with 1000 bootstrap samples was used to estimate the 95% confidence intervals (CI) (DiCiccio and Efron, 1996).

Note that the Clinical-T5 model used in our experiment was pretrained on the same data (MIMIC-III) that was annotated by some DR.BENCH tasks (e.g. problem summarization and EmrQA). This setting is known as transductive learning. Transductive learning is a very realistic scenario for the

clinical domain where due to privacy issues, language models are likely be pretrained on the data from the same institution as the data to which they would be applied. Obviously, it would also be interesting to investigate the performance of a T5 variant that was trained on a clinical corpus that was different from which the evaluation data were sourced. Unfortunately, this was not possible due to the fact that MIMIC was the only publicly available corpus of clinical notes and it was used for training clinical language models.

The training data consisted of one progress note per unique patient. A separate cohort of unique patients was selected for the test set, ensuring no overlap between the train and test splits. All experiments used Adam optimizer with a learning rate of 1e-5, batch size of 8, beam size of 5, and 100 epochs with early stopping. The learning rate and batch size were picked based on the best hyperparameters found from the prior work (Gao et al., 2023). All experiments were completed on a single A100 GPU with 40 GB memory. The models were reviewed for error analysis by a critical care physician on the full test set of 86 progress notes and common observations were highlighted with examples in the error analysis.

4 Results and Discussion

The results of our experiments are summarized in Table 2. The full set of results including the confidence intervals is available in the Appendix (Table 4).

Clinical-T5 770M trained in the multi-task setting demonstrated the best performance (28.55) for the Summarization task, establishing a new state-of-the art for this task. The single-task setting for the same T5 variant was a close second (28.28).

T5 variants trained on in-domain data (SciFive and Clinical-T5) performed better than their general domain counterpart T5 models of the same size.

All models, except Clinical-T5 experienced a drop in performance when trained in a multi-task approach. We hypothesize that the models pretrained on non-clinical data were overwhelmed with out-of-domain (i.e. clinical) data when trained in a multi-task way and failed to generalize as a result. Predictably, larger models performed at least as well as the smaller models and outperformed the smaller models in most scenarios.

Admittedly, our work leaves open the question of whether the state-of-the-art performance obtained by Clinical-T5 770M has to do with the fact that it was pretrained on MIMIC notes, which were also annotated in the problem summarization task. At the same time, the performance of other T5 variants, such as SciFive 770M, was close, without it pretraining on MIMIC. This suggests that another T5 variant trained on a corpus of clinical notes that was different from MIMIC would perform at least as well or better depending on the size of the pre-training corpus. It should be noted that the model of this size, 770M parameters, can very likely absorb significantly larger amounts of clinical notes than what was available in MIMIC (Hoffmann et al., 2022). We leave verifying this hypothesis for future work.

Model	Training	Summarization
Gao et al., 2023	Single task	7.60 (5.31 - 9.89)
T5 220M	Single task	26.35 (22.18 - 30.52)
	Multi-task	24.84 (20.28 - 29.40)
T5 770M	Single task	26.90 (22.58 - 31.23)
	Multi-task	23.99 (19.86 - 28.13)
SCIFIVE 220M	Single task	25.31 (21.45 - 29.17)
	Multi-task	24.38 (19.99 - 28.78)
SCIFIVE 770M	Single task	27.31 (23.09 - 31.53)
	Multi-task	25.31 (21.45 - 29.17)
CLINICAL-T5 220M	Single task	25.35 (21.19 - 29.51)
	Multi-task	26.21 (21.92 - 30.49)
CLINICAL-T5 770M	Single task	28.28 (24.17 - 32.38)
	Multi-task	**28.55** (24.29 - 32.80)

Table 2: Performance of fine-tuned T5 models on the summarization task. 95% confidence intervals are included. The first row is a baseline representing the best performance on this task to date. Please see the Appendix for the full set of results.

Error Analysis: Although both clinical models produced similar ROUGE-L scores, the model trained in a single-task setting appeared to achieve better abstraction during error analysis. For the example in Table 5, the assessment described sepsis but does not mention the source of the sepsis infection in multi-task Clinical-T5 770M. The data from the subjective and objective sections of the progress note described an abdominal source and lab results were consistent with a clostridium difficile infection. The multi-task prediction was able to generate sepsis but further generated text that the source was unclear. The single task performed better abstraction and generated clostridium difficile as the source for the infection, which was more accurate during expert review. In another diagnosis, the ground truth label was "EtOH Withdrawal" (alcohol withdrawal). The multitask extracted "altered mental status, hypertensive, tachycardia," (symptoms of withdrawal) whereas the single task was able to abstract "DTs EtOH w d," (delirium tremens alcohol withdrawal - a type of severe alcohol withdrawal in critically ill patients). Again, the single task achieved greater accuracy with abstraction from symptoms of alcohol withdrawal presented in the earlier sections of the note.

Resource Utilization: The experiments were conducted on the Google Cloud Platform using one A100 40 GB NVIDIA GPU on a Linux base system. For all experiments, the total training time was approximately 250 hours for both single-task and multi-task approaches. The carbon emission footprint was 35.5 kilograms (kg) of CO_2. However, the total carbon emission was only 4.5 kg of CO_2 for the single-task experiments. (Lacoste et al., 2019)

5 Conclusion

In this work we experiment with the DR.BENCH suite of tasks and established a new state-of-the-art result on the problem list generation task, a task critical for AI-assisted diagnostic reasoning. Our other contribution indicates that multi-task learning does not work well, unless in-domain data was used for pretraining and that included (unlabeled) task data during pretraining (a scenario known as transductive learning) leads to the best performance. Finally, our work provides evidence that generative models benefit from pretraining on in-domain data. In future work, we plan to explore the utility of decoder-only LLMs for clinical diagnostic reasoning.

6 Limitations

The limitation of this work was the use of ROUGE-L as the evaluation metric. Given the many acronyms and synonyms in medical writing, ROUGE-L, based on the longest common sequence, does not capture the many nuances in its score. Researchers have shown concerns for the ROUGE score and have developed metrics for summarization that are more semantically aware of the ground truth (Akter et al., 2022), but their usability is yet to be validated.

Training large language models from scratch uses a considerable amount of carbon footprint. (Patterson et al., 2021) Fine-tuning large language models for downstream tasks is one way to reduce carbon footprint but still needs to be cost-effective. As the AI community progresses in this field, developing a cost-effective and carbon-friendly solution is needed. The NLP field is moving towards prompt-based methods with larger LLMs (Lester et al., 2021), so the next step for this research is to experiment with soft prompting approaches to address low resource settings and leverage prompt tuning in LLMs for the problem summarization task.

7 Ethics Statement

This research utilized a deidentified dataset that does not include any protected health information. This dataset operates in compliance with the PhysioNet Credential Health Data Use Agreement (v1.5.0). All experiments conducted adhered to the guidelines outlined in the PhysioNet Credentialed Health Data License Agreement. Additionally, this study has been deemed exempt from human subjects research.

References

Mousumi Akter, Naman Bansal, and Shubhra Kanti Karmaker. 2022. Revisiting automatic evaluation of extractive summarization task: Can we do better than ROUGE? In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1547–1560, Dublin, Ireland. Association for Computational Linguistics.

Judith L. Bowen. 2006. Educational strategies to promote clinical diagnostic reasoning. *New England Journal of Medicine*, 355(21):2217–2225. PMID: 17124019.

Thomas J. DiCiccio and Bradley Efron. 1996. Bootstrap confidence intervals. *Statistical Science*, 11(3):189–212.

Yanjun Gao, John Caskey, Timothy Miller, Brihat Sharma, Matthew M. Churpek, Dmitriy Dligach, and Majid Afshar. 2022a. Tasks 1 and 3 from progress note understanding suite of tasks: Soap note tagging and problem list summarization. *PhysioNet*.

Yanjun Gao, Dmitriy Dligach, Timothy Miller, John Caskey, Brihat Sharma, Matthew M. Churpek, and Majid Afshar. 2023. Dr.bench: Diagnostic reasoning benchmark for clinical natural language processing. *Journal of Biomedical Informatics*, 138:104286.

Yanjun Gao, Dmitriy Dligach, Timothy Miller, Samuel Tesch, Ryan Laffin, Matthew M. Churpek, and Majid Afshar. 2022b. Hierarchical annotation for building a suite of clinical natural language processing tasks: Progress note understanding. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 5484–5493, Marseille, France. European Language Resources Association.

Rebekah L Gardner, Emily Cooper, Jacqueline Haskell, Daniel A Harris, Sara Poplau, Philip J Kroth, and Mark Linzer. 2018. Physician stress and burnout: the impact of health information technology. *Journal of the American Medical Informatics Association*, 26(2):106–114.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katie Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, Jack W. Rae, Oriol Vinyals, and Laurent Sifre. 2022. Training compute-optimal large language models.

Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. 2021. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Applied Sciences*, 11(14).

Kundan Krishna, Sopan Khosla, Jeffrey Bigham, and Zachary C. Lipton. 2021. Generating SOAP notes from doctor-patient conversations using modular summarization techniques. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4958–4972, Online. Association for Computational Linguistics.

Alexandre Lacoste, Alexandra Luccioni, Victor Schmidt, and Thomas Dandres. 2019. Quantifying the carbon emissions of machine learning. *arXiv preprint arXiv:1910.09700*.

Eric Lehman and Alistair Johnson. 2023. Clinical-t5: Large language models built using mimic clinical text. *PhysioNet*.

Brian Lester, Rami Al-Rfou, and Noah Constant. 2021. The power of scale for parameter-efficient prompt tuning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3045–3059, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Jennifer Liang, Ching-Huei Tsou, and Ananya Poddar. 2019. A novel system for extractive clinical note summarization using EHR data. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 46–54, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Anusri Pampari, Preethi Raghavan, Jennifer Liang, and Jian Peng. 2018. emrQA: A large corpus for question answering on electronic medical records. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2357–2368, Brussels, Belgium. Association for Computational Linguistics.

Darshna Patel, Saurabh Shah, and Hitesh Chhinkaniwala. 2019. Fuzzy logic based multi document summarization with improved sentence scoring and redundancy removal technique. *Expert Systems with Applications*, 134:167–177.

Vimla L. Patel and Guy J. Groen. 1986. Knowledge based solution strategies in medical reasoning. *Cognitive Science*, 10(1):91–116.

David Patterson, Joseph Gonzalez, Quoc Le, Chen Liang, Lluis-Miquel Munguia, Daniel Rothchild, David So, Maud Texier, and Jeff Dean. 2021. Carbon emissions and large neural network training.

Long N Phan, James T Anibal, Hieu Tran, Shaurya Chanana, Erol Bahadroglu, Alec Peltekian, and Grégoire Altan-Bonnet. 2021. Scifive: a text-to-text transformer model for biomedical literature. *arXiv preprint arXiv:2106.03598*.

Jonathan Pilault, Raymond Li, Sandeep Subramanian, and Chris Pal. 2020. On extractive and abstractive neural document summarization with transformer language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9308–9319, Online. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21(1).

Adam Rule, Steven Bedrick, Michael F. Chiang, and Michelle R. Hribar. 2021. Length and Redundancy of Outpatient Progress Notes Across a Decade at an Academic Medical Center. *JAMA Network Open*, 4(7):e2115334–e2115334.

Chaitanya Shivade. 2019. Mednli - a natural language inference dataset for the clinical domain. *PhysioNet*.

Adam Wright, Justine Pang, Joshua C Feblowitz, Francine L Maloney, Allison R Wilcox, Karen Sax McLoughlin, Harley Ramelson, Louise Schneider, and David W Bates. 2012. Improving completeness of electronic problem lists through clinical decision support: a randomized, controlled trial. *Journal of the American Medical Informatics Association*, 19(4):555–561.

Qianqian Xie, Zheng Luo, Benyou Wang, and Sophia Ananiadou. 2023. A survey on biomedical text summarization with pre-trained language model.

A Appendix

This section adds two more tables to show the results of the other clinical task. The results are compared with previous results and can be seen in the baseline column. In many cases, such as MedNLI, AP, and EmrQA, we can see improvement in the multitask experiments.

Model	Training	Summarization	SOAP	A/P
Gao et al., 2023	Single task	7.60 (5.31 - 9.89)	**60.12** (59.33 - 60.90)	80.09 (79.32 - 83.23)
T5 220M	Single task	26.35 (22.18 - 30.52)	**60.12** (59.33 - 60.90)*	73.31 (71.34 - 77.65)*
	Multi-task	24.84 (20.28 - 29.40)	56.63 (55.83 - 57.42)	43.25 (41.35 - 66.59)
T5 770M	Single task	26.90 (22.58 - 31.23)	55.57 (54.78 - 56.35)*	77.96 (75.38 - 81.60)*
	Multi-task	23.99 (19.86 - 28.13)	51.10 (50.32 - 51.91)	75.15 (71.93 - 78.19)
SciFive 220M	Single task	25.31 (21.45, 29.17)	57.74 (56.95 - 58.53)*	76.76 (74.81 - 80.92)*
	Multi-task	24.38 (19.99 - 28.78)	54.86 (54.06 - 55.65)	68.87 (65.50 - 72.12)
SciFive 770M	Single task	27.31 (23.09 - 31.53)	47.65 (46.85 - 48.47)*	75.11 (73.10,79.42)*
	Multi-task	25.31 (21.45 - 29.17)	44.51 (43.72- 45.29)	77.50 (74.45 - 80.37)
Clinical-T5 220M	Single task	25.35 (21.19 - 29.51)	55.30 (54.51 - 56.11)	80.44 (77.47 - 83.35)
	Multi-task	26.21 (21.92 - 30.49)	52.41 (51.62 - 53.20)	65.49 (62.08 - 68.76)
Clinical-T5 770M	Single task	28.28 (24.17 - 32.38)	52.82 (52.03 - 53.61)	78.79 (75.76 - 81.66)
	Multi-task	**28.55** (24.29 - 32.80)	54.00 (53.21 - 54.80)	**80.58** (77.57 - 83.38)

Table 3: Finetuned T5 models on various clinical task with 95% confidence interval calculated using the bootstrapping method. A/P represents assessment and plan relational labeling task. Summarization use ROUGE L, A/P use F1-macro and SOAP use accuracy score for the evaluation metrics. The first row in the table represents best scores reported in the DR.BENCH paper and * in the other rows represent scores for the respective task in DR.BENCH paper (Gao et al., 2023)

Model	Training	EmrQA	MedNLI	MedQA
Gao et al., 2023	Single task	39.20 (34.63 - 43.78)	84.88 (82.98 - 86.64)	24.59 (22.31 - 27.02)
T5 220M	Single task	33.40 (29.27 - 37.61)*	79.75 (78.62 - 82.70)*	22.55 (20.01 - 25.69)*
	Multi-task	38.48 (37.24 - 39.79)	72.57 (70.18 - 74.82)	21.75 (19.48 - 24.12)
T5 770M	Single task	38.05 (33.56 - 42.58)*	84.04 (82.14 - 85.86)*	20.97 (18.77 - 23.25)*
	Multi-task	41.42 (40.16, 42.72)	83.19 (81.22, 85.09)	23.25 (20.97, 25.61)
SciFive 220M	Single task	37.28 (32.84 - 42.11)*	82.84 (80.87 - 84.74)*	22.78 (20.50 - 25.14)*
	Multi-task	40.08 (38.82 - 41.39)	78.83 (76.72 - 80.94)	21.52 (19.32 - 23.80)
SciFive 770M	Single task	41.21 (39.93 - 42.49)	83.89 (82.00 - 85.79)	23.09 (20.82 - 25.37)
	Multitask	41.26 (39.98 - 42.56)	84.35 (82.49 - 86.22)	23.72 (21.37 - 26.08)
Clinical-T5 220M	Single task	41.35 (40.07 - 42.65)	84.32 (82.42 - 86.15)	21.92 (19.64 - 24.19)
	Multi-task	40.30 (39.02 - 41.62)	71.23 (68.92 - 73.56)	22.46 (20.19 - 24.74)
Clinical-T5 770M	Single task	**42.69** (41.39 - 43.95)	85.86 (85.02 - 88.47)	24.27 (21.92 - 26.63)
	Multi-task	42.61 (41.34 - 43.92)	**86.14** (84.32 - 87.90)	**25.84** (23.41 - 28.28)

Table 4: Finetuned T5 models on various clinical task with 95% confidence interval calculate using the bootstrapping method. All the evaluation metrics here are the accuracy score. The first row in the table represents best scores reported in the DR.BENCH paper and * in the other rows represent scores for the respective task in DR.BENCH paper (Gao et al., 2023)

Input	Ground Truth Diagnoses/ Problems	T5 770M Single task	Clinical-T5 770M Single task	Clinical-T5 770M Multi-task
SUMMARIZE: **<*ASSESSMENT*>** 48 y/o M with HIV 47M s/p elective spinal surgery (anterior and posterior LIFs), intubated - - - - **<*SUBJECTIVE*>** Agitated, diaphoretic, altered, hypertensive and tachy this AM - - - - **<*OBJECTIVE*>** Last dose of Antibiotics: Infusions: Other ICU medications: Heparin Sodium (Prophylaxis) - - - -	EtOH withdrawal Spinal surgery	Altered MS s p elective spinal surgery	DTs EtOH w d pain h o chronic pain	Altered mental status Hypertension Tachycardia Acute renal failure s p spinal surgery
SUMMARIZE: **<*ASSESSMENT*>** SEPSIS WITHOUT ORGAN DYSFUNCTION Ms. [**Known lastname 10381**] is a 76F with multiple medical problems, who is - - - - - - - **<*SUBJECTIVE*>** FEVER - 101.7 F - [**2129-9-3**] 12:33 PM -received boluses overnight for low SBP - - - - **<*OBJECTIVE*>** Last dose of Antibiotics: Cefipime - [**2129-9-3**] 04:05 PM Metronidazole - [**2129-9-4**] 04:00 AM - - - -	Sepsis Patient has re developed fevers on 9 2 on a regimen of vancomycin ceftriaxone Possible sources include 1 Intra abdominal source	Sepsis Thrombo- cytopenia	Sepsis Likely source is clostridium difficile colitis Acute renal failure	Hypotension Likely secondary to sepsis though source unclear at this time Acute renal failure

Table 5: The table represents a snippet of the input and output sections of problem summarization. The input data contains an added prefix that denotes the task for T5, "SUMMARIZE" in this case, and <prefix> that defines the note section. Finally, "- - - -" is the continuation of the section, which was excluded here due to the space constraint.