A Continued Pretrained LLM Approach for Automatic Medical Note Generation

Anonymous ACL submission

Abstract

LLMs are revolutionizing NLP tasks. However, the most powerful LLM, like GPT-4, is too costly for most domain-specific scenarios. We present the first continuously trained 13B Llama2-based LLM that is purpose-built for medical conversations and measured on automated scribing. Our results show that our model outperforms GPT-4 in PubMedQA with 76.6% accuracy and matches its performance in summarizing medical conversations into SOAP notes. Notably, our model exceeds GPT-4 in capturing a higher number of correct medical concepts and outperforms human scribes with higher correctness and completeness.

1 Introduction

The emergence of large language model (LLM) has brought revolutionary changes to natural language processing and understanding tasks, paving the way for practical applications of AI across multiple domains such as law, finance, and healthcare. Private and open-source LLMs such as GPT-4 (OpenAI, 2023) and Llama 2 (Meta, 2023) have shown strong performance on general NLP benchmarks. However, recent studies have shown promise that with continued training on more targeted datasets, e.g. smaller LLMs like Orca (Mukherjee et al., 2023; Mitra et al., 2023) and Phi-2 (Mojan Javaheripi, 2023), can surpass much larger LLMs on general tasks.

Despite the success of LLM in general capabilities, they often fall short in niche domains like healthcare, where precision and profound understanding are crucial. This necessitates domain-specific models, particularly in healthcare where misinterpretations of facts can have significant consequences. Hence, several models such as Meditron-70B (Chen et al., 2023), PMC-LLaMA (Wu et al., 2023) have emerged.

Transcribing medical conversations is a challenging task for both humans and machines due to potential transcription errors and the innate complexity of spoken language, an issue unaddressed by existing medical LLMs. Existing LLMs which have been designed for the medical domain largely do well on problems like medical Q&A but cannot create a complete EHR-compatible medical note. Some domain-adapted LLMs (Van Veen et al., 2023) can write some components of the note, but they leave out the crucial "Subjective" section. Some fine-tuned models (Zhang et al., 2021) can generate notes from medical conversations but need human overview.

Overall, we developed a comprehensive medical LLM that understands medical dialogues. By using techniques like explanation tuning and continued pretraining on diverse set of data, including medical and general web corpora, GPT-4 task instructions, EHRs, the model was able to generate physician-approved medical SOAP notes, making it uniquely fit for industry application.

Our main contributions include:

To the best of our knowledge, we are the first to build a small-size (13B) medical LLM that can produce medical notes from doctor-patient conversations that bypass human quality.

On the task of medical note generation, our 13B pretrained model performs overall on-par against GPT-4 and higher in completeness.

We achieved 76.6% in accuracy, which beats GPT-4 75.2% on PubMedQA, with a much smaller model size.

2 Continued Pretraining

2.1 Dataset

We grouped our training data into three categories to enable the model to generate coherent English sentences, comprehend medical content, and execute complex instructions required for generating medical notes. (see Table 1)

Non-medical public datasets. To ensure that the

Dataset	Number of tokens (in billions)	Percentage of total data
Non-medical public	5.33	35.79
Medical public	5.68	38.14
Medical proprietary	3.88	26.07
Total	**14.89**	**100.00**

Table 1: Pretraining datasets

new model doesn't lose the generative capabilities of the pretrained Llama 2 model, we added general domain datasets such as C4 (Raffel et al., 2019). Continued pretraining on them was crucial for generational tasks, enhancing the model's grammar and phrase composition skills. Initially, we also included filtered subtitle data from open-subtitle and youtube. However, we decided to exclude these datasets due to their poor quality negatively impacting the model's performance.

Medical public datasets. We extracted medically relevant datasets from web corpus covering different aspects of medical concept understanding. MedDialog (Chen et al., 2020) taught medical language conversation while reading materials such as PubMed articles (Gao et al., 2020) provided the model with an overall medical context. PubMedQA (Jin et al., 2019) addressed ideas of medical Q&A.

Proprietary medical datasets. We also curated a deidentified proprietary medical dataset that consists of real-world doctor-patient conversations from the United States, Electronic Health Records (EHR), SOAP (Subjective, Objective, Assessment, and Plan) notes, and ROS templates. We also created an artificial dataset with medical instructions, that includes step-by-step thought processes. This helped the model understand multi-step reasoning, essential for the downstream medical documentation task.

While we developed a much larger high-quality custom dataset, currently only 14.89B tokens were used for this training exercise.

2.2 Training Details

We performed training using FSDP (Zhao et al., 2023) - pipeline parallelism with hybrid sharding and flash attention 2 on 32 A100 80 GB GPUs.

We continued training LLaMA2-13B using learning rate of 5e-5 and decaying it to 1e-5 following a cosine schedule. We chose a batch size of 256 with an 8K context window, maintaining a relatively small batch size to achieve about 10K effective gradient update steps. We set the weight decay at 0.1 and a warm-up step count to 50.

Robust Training. To be tolerant of machine and experiment related mishaps, we used fixed seed, checkpoints, and implemented phased training where we divided the training data into n subsets. If the loss of a particular validation subset started to stabilize, we reduced the sampling rate in the next phase for efficiency.

Data Packing & Dedup. We packed data by sentence to fit into max sequence length. We also deduplicated our data to improve data quality (Lee et al., 2021).

Loss. For the general corpus including C4, public medical materials, we calculated the gradient on every token. However, on proprietary instruction data, the loss was only calculated on response tokens.

3 Evaluations

This section shows some of our pretraining results and evaluation methodology.

3.1 Pretraining

We employed two evaluation methods to monitor pretraining. Firstly, we measured the perplexity of each data source to evaluate the model's task difficulty. Secondly, for a holistic understanding of the generation quality, we used several few-shot (3-shot) generative tasks for validation, that include:

1) Long text generation: This is a subjective section summarization task from transcript to evaluate the generation quality of different note categories.

2) Medium text generation: This is a transcription based Q&A task created by modifying the (Rohan et al., 2023) pipeline on the InstructQA dataset. We query GPT-4 on questions prompting responses ranging from a few words to a full sentence, like asking for prescribed medications.

3) Short text generation: This comprises of ROS (Review Of System) - related Q&A tasks, including questions about body system identification (multi-choice), and absence or presence of symptoms (single-choice).

We measured Rouge-cls for tasks 1, 2 and accuracy for task 3, to monitor pretraining performance.

As shown in Figure1, we found that our model's performance consistently improved for long text generation, medium text generation, and multi-choice Q&A. Surprisingly the single-choice Q&A did not show any major improvement. We believe this is because a) the accuracy numbers are

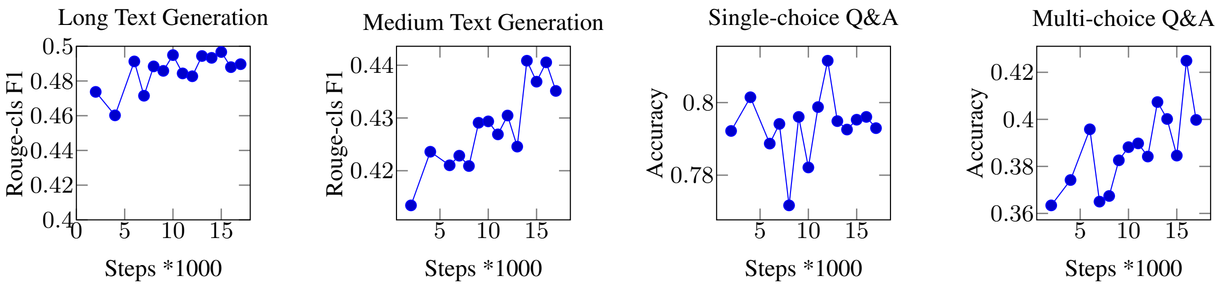

Figure 1: Pretraining validation generation capability monitoring

Training data	ROS (multi-choice) (Acc)	InstructQA Rouge-1 (f1)	InstructQA Rouge-cls (f1)
1B Total	**47.36**	**0.44**	0.41
med	37.85	0.39	0.35
pub	36.81	**0.44**	**0.42**

Table 2: **Training data ablation results**. The **med** dataset is derived from the 1B training dataset by excluding all the public datasets. Similarly, the **pub** dataset is produced by removing all medical datasets.

Model	#Incorrect	#Irrelevant	#Missed
Human	1.2	**0**	11.2
GPT-4	**0.8**	0.2	6.75
Our LLM	0.85	0.3	**4.3**

Table 3: Average entity errors per conversation

already too high b) we noticed further improvement on these numbers when we separately trained the model on a smaller related dataset, suggesting scaling up the training can lead to further improvements.

3.2 Pretraining Ablation

As shown in table 2, we examined the impact of different data proportions on a 1B token dataset, derived from a scaled-down version of our custom 15B dataset, using the 7B Llama2 model.

From ablation, we observed that removing the general datasets from the training mix inversely impacted the generative capabilities of the model leading to poorer summarization quality. We were also able to conclude that the medical datasets indeed improve the model's understanding of the medical context. This prompted us to use equal proportions of these datasets in the training to ensure that the model doesn't lose its generative capabilities while gaining medical understanding.

3.3 Medical Note Generation

Evaluation Setup. We used medical transcripts from healthcare conversations as our primary input. Both GPT-4 and the pretrained LLM used the same prompt to generate a clinical note from these transcripts. We also evaluated them against human scribes from our production system (medical students who underwent internal scribe training and received monetary compensation for their services). The evaluation included 10 doctor-patient dialogue-style conversations with an average audio duration of 12 minutes and 35 medical entities. For a fair comparison of the notes created by humans and the model, we leveraged human experts for evaluation.

As outlined in (Van Veen et al., 2023), we evaluated the generated notes on three key parameters: **Completeness**, **Correctness**, and **Conciseness**. We used human experts to create a rubric note for each transcript. This rubric note marked all the vital medical information as separate medical entities. Each entity within this note symbolizes a crucial sentence or phrase that a provider needs, to approve the note. We measured the generated note on three metrics:

Missed Information refers to the entities omitted in the test note relative to the rubric note. This metric reflects the test note's completeness.

Incorrect Information implies the entities inaccurately captured by the test note. Given the critical nature of information accuracy in healthcare, this metric is vital as misinformation can undermine trust in AI.

Irrelevant information represents the extraneous entities included in the test note when compared against the rubric note. As longer medical notes require more review time for providers, minimizing irrelevant information is key.

In our comprehensive quality assessment, the human scribe, on an average, took 1.67 times of audio time to finish the note summarization. We noticed that humans introduced some incorrectness in their notes which we attribute to ASR (Automatic Speech Recognition) errors. Both our model and GPT-4 outperformed humans in correcting these errors. Overall, our model performs on-par with GPT-

4. Thanks to our optimized training, our model was able to capture more relevant information than both human transcriptions and GPT-4.

3.4 Public Benchmark

We evaluated on two benchmarks, PubMedQA (Jin et al., 2019), MedQA (Jin et al., 2021).

In PubMedQA, our model achieved an accuracy of 76.6% after fine-tuning, which surpasses GPT-4's performance (Nori et al., 2023) of 75.2%. Our improved performance, aside from training on PubMedQA, can be attributed to our proprietary medical question-answering tasks on conversational data which focuses more on medical understanding.

In MedQA, we achieved an accuracy of 45.2%, which is lower than previous LLaMA2 based medical model (Wu et al., 2023). MedQA emphasizes medical reasoning, requiring the model to deduce diagnoses or solutions based on given problems. However, our model is designed to interpret medical conversations which does not align with this task and hence, leads to subpar performance on this dataset.

4 Discussion

We deidentified all the clinical data to remove the PHI information in accordance with our data compliance agreement. Our model is strictly used for internal medical related scribing tasks such as summarization, transcription Q&A and note review. All the prompts are audited to avoid any unintentional usage.

Our current design focuses more on in-context transcription understanding. The model's medical reasoning capability can be further improved to achieve better performance on MedQA alike datsts by introducing more medical reasoning data.

This is our first attempt to showcase the power of injecting medical reasoning on medical note generation tasks on a smaller LLM with small scale 15B dataset. We plan to further improve by scaling up our data and share future work.

5 Conclusions

This paper presents our work of developing a medical LLM capable of comprehending and summarizing medical dialogues. As a result, this is the first model, with significantly fewer parameters, to outperform humans on medical note summarization and GPT-4 on PubMedQA. Evaluations show that even small scale pretraining of smaller LLMs can show impressive gains and achieve on-par performance with GPT-4. We further believe that we can achieve improved results, if we simply scale up our training. Our work presents a promising development in healthcare documentation industry and other medical areas.

6 Related Work

Explanation tuning. Orca (Mukherjee et al., 2023; Mitra et al., 2023) models showcased that smaller Language Models (LMs) capable of sound reasoning can efficiently perform complex tasks. They were trained by explanation tuning a Llama 13B model (Touvron et al., 2023) using bigger models like GPT4 as a teacher. Despite their smaller size, they retained a majority of the quality of ChatGPT and GPT4.

Medical LLMs. Various medical LLMs such as BioGPT (Luo et al., 2022), MedGPT (Kraljevic et al., 2021), Med-PaLM (Singhal et al., 2022) and Med-PaLM 2 (Singhal et al., 2023) show how training on various medical datasets, improves model's performance on medical knowledge understanding tasks. In the LLaMA (Touvron et al., 2023) family, MEDITRON-70B (Chen et al., 2023), the state-of-the-art open-source LLM and PMC-LLaMA (Wu et al., 2023) demonstrates the importance of fundamental medical knowledge in healthcare-focused models and emphasizes the effectiveness of task-specific fine-tuning and instruction tuning.

Domain adaption LLM. Continued pretraining on unlabeled, domain-specific data, as shown by (Gururangan et al., 2020), (Beltagy et al., 2019) can successfully improve a model's performance on domain-specific tasks, offering a practical solution when resources for domain-adaptive pretraining from scratch are constrained. Furthermore, (Wu et al., 2022) highlighted the benefits of continued pretraining in improving zero-shot and few-shot compatibility.

Medical Note Generation. Prior work by (Zhang et al., 2021), (Van Veen et al., 2023) displayed potential in utilizing Language Models to generate medical summaries from dialogues. However, their focus wasn't on creating a complete note that could be directly sanctioned by the provider but rather on partial aspects or semi-automated methods, necessitating human intervention. (Chuang et al., 2023) leveraged soft prompts to control the variance of LLM outputs on medical summarization tasks.

References

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. Scibert: A pretrained language model for scientific text. *arXiv preprint arXiv:1903.10676*.

Shu Chen, Zeqian Ju, Xiangyu Dong, Hongchao Fang, Sicheng Wang, Yue Yang, Jiaqi Zeng, Ruisi Zhang, Ruoyu Zhang, Meng Zhou, Penghui Zhu, and Pengtao Xie. 2020. Meddialog: a large-scale medical dialogue dataset. *arXiv preprint arXiv:2004.03329*.

Zeming Chen, Alejandro Hernández Cano, Angelika Romanou, Antoine Bonnet, Kyle Matoba, Francesco Salvi, Matteo Pagliardini, Simin Fan, Andreas Köpf, Amirkeivan Mohtashami, et al. 2023. Meditron-70b: Scaling medical pretraining for large language models. *arXiv preprint arXiv:2311.16079*.

Yu-Neng Chuang, Ruixiang Tang, Xiaoqian Jiang, and Xia Hu. 2023. Spec: A soft prompt-based calibration on mitigating performance variability in clinical notes summarization. *arXiv preprint arXiv:2303.13035*.

Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, et al. 2020. The pile: An 800gb dataset of diverse text for language modeling. *arXiv preprint arXiv:2101.00027*.

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

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):6421.

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William Cohen, and Xinghua Lu. 2019. Pubmedqa: A dataset for biomedical research question answering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2567–2577.

Zeljko Kraljevic, Anthony Shek, Daniel Bean, Rebecca Bendayan, James Teo, and Richard Dobson. 2021. Medgpt: Medical concept prediction from clinical narratives. *arXiv preprint arXiv:2107.03134*.

Katherine Lee, Daphne Ippolito, Andrew Nystrom, Chiyuan Zhang, Douglas Eck, Chris Callison-Burch, and Nicholas Carlini. 2021. Deduplicating training data makes language models better. *arXiv preprint arXiv:2107.06499*.

Renqian Luo, Liai Sun, Yingce Xia, Tao Qin, Sheng Zhang, Hoifung Poon, and Tie-Yan Liu. 2022. Biogpt: generative pre-trained transformer for biomedical text generation and mining. *Briefings in Bioinformatics*, 23(6):bbac409.

Meta. 2023. Llama 2: Open foundation and fine-tuned chat models. *ArXiv*, abs/2307.09288.

Arindam Mitra, Luciano Del Corro, Shweti Mahajan, Andres Codas, Clarisse Simoes, Sahaj Agarwal, Xuxi Chen, Anastasia Razdaibiedina, Erik Jones, Kriti Aggarwal, Hamid Palangi, Guoqing Zheng, Corby Rosset, Hamed Khanpour, and Ahmed Awadallah. 2023. Orca 2: Teaching small language models how to reason.

Sébastien Bubeck Mojan Javaheripi. 2023. Phi-2: The surprising power of small language models.

Subhabrata Mukherjee, Arindam Mitra, Ganesh Jawahar, Sahaj Agarwal, Hamid Palangi, and Ahmed Awadallah. 2023. Orca: Progressive learning from complex explanation traces of gpt-4.

Harsha Nori, Nicholas King, Scott Mayer McKinney, Dean Carignan, and Eric Horvitz. 2023. Capabilities of gpt-4 on medical challenge problems. *arXiv preprint arXiv:2303.13375*.

OpenAI. 2023. Gpt-4 technical report. *ArXiv*, abs/2303.08774.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv e-prints*.

Taori Rohan, Gulrajani Ishaan, Zhang Tianyi, Dubois Yann, Li Xuechen, Guestrin Carlos, Liang Percy, and B. Hashimoto Tatsunori. 2023. Alpaca: A strong, replicable instruction-following model.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, et al. 2022. Large language models encode clinical knowledge. *arXiv preprint arXiv:2212.13138*.

Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Le Hou, Kevin Clark, Stephen Pfohl, Heather Cole-Lewis, Darlene Neal, et al. 2023. Towards expert-level medical question answering with large language models. *arXiv preprint arXiv:2305.09617*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Dave Van Veen, Cara Van Uden, Louis Blankemeier, Jean-Benoit Delbrouck, Asad Aali, Christian Bluethgen, Anuj Pareek, Malgorzata Polacin, William Collins, Neera Ahuja, et al. 2023. Clinical text summarization: adapting large language models can outperform human experts. *arXiv preprint arXiv:2309.07430*.

Chaoyi Wu, Weixiong Lin, Xiaoman Zhang, Ya Zhang, Yanfeng Wang, and Weidi Xie. 2023. Pmc-llama: Towards building open-source language models for medicine. *arXiv preprint arXiv:2305.10415*.

Zhaofeng Wu, Robert L Logan IV, Pete Walsh, Akshita Bhagia, Dirk Groeneveld, Sameer Singh, and Iz Beltagy. 2022. Continued pretraining for better zero-and few-shot promptability. *arXiv preprint arXiv:2210.10258*.

Longxiang Zhang, Renato Negrinho, Arindam Ghosh, Vasudevan Jagannathan, Hamid Reza Hassanzadeh, Thomas Schaaf, and Matthew R Gormley. 2021. Leveraging pretrained models for automatic summarization of doctor-patient conversations. *arXiv preprint arXiv:2109.12174*.

Yanli Zhao, Andrew Gu, Rohan Varma, Liang Luo, Chien-Chin Huang, Min Xu, Less Wright, Hamid Shojanazeri, Myle Ott, Sam Shleifer, et al. 2023. Pytorch fsdp: experiences on scaling fully sharded data parallel. *arXiv preprint arXiv:2304.11277*.