Fine-Tuning Large Language Models for Structured Clinical Report Generation Using GRPO

Uday Devulapalli[1, 2], **Aarat Satsangi**[1, 2], **Apurva Narayan**[1]

[1]Western University, London, ON, Canada
[2]International Center for Applied Systems Science for Sustainable Development (ICASSSD), Cambridge, ON, Canada
udevulap@uwo.ca, asatsang@uwo.ca, apurva.narayan@uwo.ca

Abstract

The generation of structured medical reports using large language models (LLMs) presents unique challenges, particularly in maintaining clinical relevance and adhering to strict formatting requirements. In this work, we investigate the effectiveness of fine-tuning LLMs for structured report generation using DeepSeek R1 models. We conduct experiments with two model variants: DeepSeek R1 8B and DeepSeek R1 14B. For both models, we apply Group Relative Policy Optimization (GRPO) using the Medical Information Mart for Intensive Care (MIMIC-IV) dataset, leveraging Low-Rank Adaptation (LoRA) for parameter-efficient fine-tuning. Our results show that the GRPO fine-tuned DeepSeek-R1 8B and 14B models outperformed all baseline models, including the larger 32B DeepSeek-R1 model, demonstrating the effectiveness of parameter-efficient tuning. These findings underscore the potential of reinforcement learning-based fine-tuning of LLMs for generating structured reports in the medical domain.

Introduction

The automation of clinical documentation is a critical step toward improving the efficiency, consistency, and accessibility of healthcare records. Medical reports, such as discharge summaries, Subjective, Objective, Assessment, and Plan (SOAP) notes, and radiology findings, follow highly structured formats that ensure clarity, completeness, and compliance with clinical standards. However, generating such structured reports from unstructured data (e.g., physician notes or patient interactions) remains a significant challenge (Huang et al. 2024). Large Language Models (LLMs) have demonstrated remarkable capabilities in generating fluent, context-aware text (Raiaan et al. 2024). While general-purpose models like GPT-3 (Brown et al. 2020), LLaMA (Grattafiori et al. 2024), and DeepSeek (Guo et al. 2025) have been successfully applied to a variety of natural language processing (NLP) tasks, their use in structured text generation, particularly in the clinical domain, requires careful adaptation. These models often lack the innate ability to follow rigid formatting constraints and domain-specific language styles without targeted fine-tuning (Lehman et al. 2023; Kim et al. 2025). Recent advances in reinforcement learning and parameter-efficient fine-tuning methods have opened up new possibilities for aligning LLM outputs with structured target formats (Schulman et al. 2017; Ding et al. 2023). In particular, Group Relative Policy Optimization (GRPO) has emerged as a powerful strategy for optimizing model behavior with respect to custom reward functions (Shao et al. 2024), while Low-Rank Adaptation (LoRA) enables scalable fine-tuning without requiring updates to the entire model (Hu et al. 2022). In this work, we explore the effectiveness of these techniques for guiding the DeepSeek R1 8B and 14B models toward generating high-quality, structured medical reports (here 8B and 14B represent the number of parameters in billion).

Related Work

Early clinical NLP systems relied heavily on rule-based frameworks such as CHARTextract, which used handcrafted rules and regular expressions to achieve high accuracy in extracting stroke-related data from radiology reports (Gunter et al. 2022), and other deterministic engines that extracted critical information from patient records (Mykowiecka, Marciniak, and Kupć 2009). While transparent and auditable, such methods often failed to generalize beyond specific domains. Statistical models like Conditional Random Fields (CRFs) and Support Vector Machines (SVMs) addressed these limitations by leveraging labeled data for tasks such as named entity recognition (NER) and relation extraction (Meystre et al. 2007; Patrick and Li 2010). Hybrid systems combining rule-based logic with ML models, such as those by Sohn et al. and Uzuner et al., further improved performance and interpretability (Sohn et al. 2014; Uzuner et al. 2012).

The advent of pretrained language models brought significant advances in clinical NLP. BioBERT (Lee et al. 2019), pretrained on biomedical literature (National Library of Medicine 2025), outperformed BERT in biomedical NER and relation extraction, while ClinicalBERT (Huang, Altosaar, and Ranganath 2019), trained on MIMIC-III notes (Johnson et al. 2016), improved performance in clinical concept extraction. BlueBERT (Peng, Yan, and Lu 2019), combining biomedical and clinical text, demonstrated the value of blended corpora. Larger models such as GatorTron (Yang et al. 2022) and Clinical-T5 (Lu, Dou, and Nguyen 2022) showed the benefits of scaling and adapting architectures

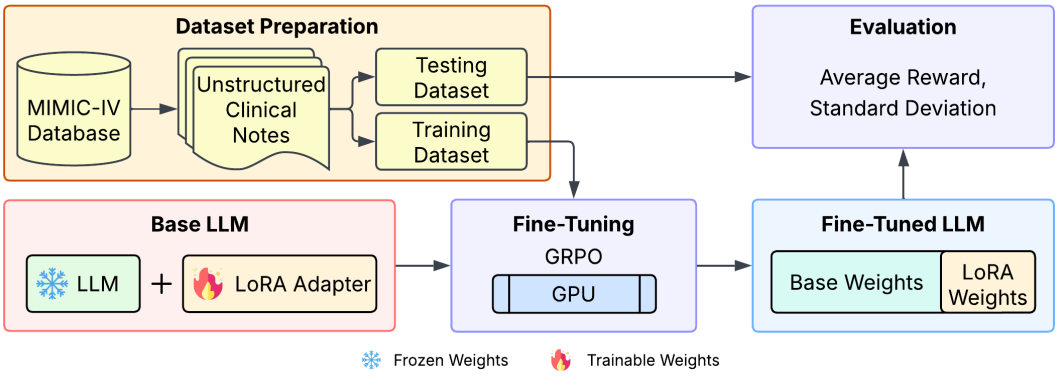


Figure 1: Overview of the training pipeline.

for summarization, classification, and structured generation. Prompt engineering has recently emerged as an alternative to fine-tuning, enabling GPT-3.5 and GPT-4 to perform clinical NLP tasks with minimal supervision (Hu et al. 2024; Taylor et al. 2024).

For structured generation, fine-tuning general-purpose LLMs for domain-specific outputs has proven effective. MediGen fine-tuned LLaMA3-8B for generating clinical reports from dialogues (Leong et al. 2024), while schema-guided generation and in-context learning improved adherence to predefined formats like JSON and XML (Du et al. 2020). Chain-of-Thought instruction tuning has enhanced faithfulness in discharge summary generation (Tang, Zhang, and Dinh 2024). Domain-specific efforts include LLM-assisted cardiology discharge summaries that increased efficiency without sacrificing accuracy (Jung et al. 2024) and attribute-aware fine-tuning to improve factuality and consistency in SOAP notes (Ramprasad, Ferracane, and Selvaraj 2023).

Methodology

Dataset

We used the MIMIC-IV-Note v2.2 dataset (Johnson et al. 2023), a large corpus of de-identified clinical notes linked to MIMIC-IV structured data made available by PhysioNet (Goldberger et al. 2000). It includes 331,794 discharge summaries from 145,915 patients and 2,321,355 radiology reports from 237,427 patients, covering admissions from 2008–2019 at Beth Israel Deaconess Medical Center. We selected a subset of 500 discharge summaries for model training and 50 discharge summaries for model testing. The preparation of data for training the models is depicted in Figure 1.

Models

In this study, we used two 4 bit quantized variants of the DeepSeek-R1 language models: DeepSeek-R1 8B distilled from the LLaMA 3.1 8B and DeepSeek-R1 14B distilled from the Qwen2.5-14B; with a LoRA module of rank 32 incorporated into each model for training. These configurations are hereafter referred to as the LoRA models. The general preparation of models for training is depicted in Figure 1. These models were selected due to their advanced reasoning capabilities, particularly their ability to engage in step-by-step logical inference and generate coherent chains of thought. Such characteristics make them well-suited for tasks involving the extraction and structuring of information from unstructured medical reports.

Fine-Tuning

The LoRA models were fine-tuned using GRPO, which is a reinforcement learning method designed to efficiently fine-tune LLMs for tasks requiring structured reasoning and output (Shao et al. 2024).

For each input prompt x, the GRPO framework samples a set of K candidate responses $\{y_1, y_2, \ldots, y_K\}$ from the current policy $\pi_\theta(y|x)$. Each response y_k is scored by a reward model $R(y_k, x)$, producing rewards $\{r_1, r_2, \ldots, r_K\}$. The group average reward is computed as

$$\bar{r} = \frac{1}{K} \sum_{k=1}^{K} r_k \tag{1}$$

The advantage for each candidate is then given by

$$A_k = r_k - \bar{r} \tag{2}$$

which centers the rewards and encourages the policy to prioritize responses better than the group mean.

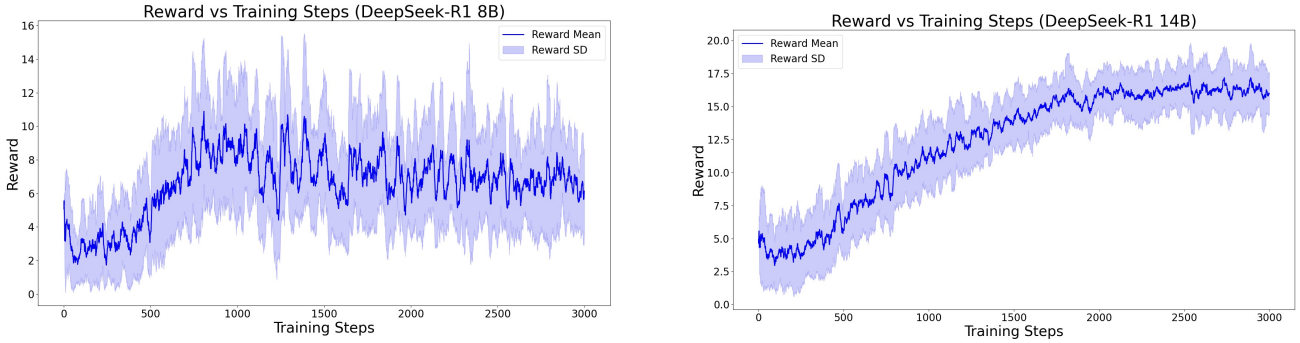
Policy updates are performed using the objective:

$$\mathcal{L}(\theta) = \frac{1}{K} \sum_{k=1}^{K} \min\big[\rho_k(\theta) A_k, \; \mathrm{clip}(\rho_k(\theta), 1-\epsilon, 1+\epsilon) A_k\big] - \beta \, \mathrm{KL}\big(\pi_\theta(\cdot|x) \,\|\, \pi_{\theta_{\text{old}}}(\cdot|x)\big) \tag{3}$$

where $\rho_k(\theta) = \frac{\pi_\theta(y_k|x)}{\pi_{\theta_{\text{old}}}(y_k|x)}$ is the importance sampling ratio, ϵ is the clipping threshold, and the KL-divergence term regularizes the updated policy toward the previous policy to maintain training stability. GRPO computes advantages

Model	Size	Avg. Reward	SD	GPU Memory
DeepSeek-R1	8B	6.86	7.32	4.9 GB
DeepSeek-R1	14B	5.54	6.87	9.0 GB
DeepSeek-R1	32B	10.02	7.23	20.0 GB
Qwen2.5	14B	11.42	6.98	9.0 GB
Llama3.1	8B	0.20	0.00	4.9 GB
DeepSeek-R1 (Fine-tuned)	8B	15.30	5.04	4.9 GB
DeepSeek-R1 (Fine-tuned)	14B	**16.38**	**3.51**	9.0 GB

Table 1: Comparison of model performance in terms of average reward, standard deviation (SD), and GPU memory usage.



(a) Deepseek-R1:8B (b) Deepseek-R1:14B

Figure 2: Rewards during fine-tuning with GRPO

across multiple candidates per prompt, enabling richer comparative learning signals for structured text generation tasks such as medical report formatting.

A rule-based reward function was developed to guide the learning process during fine-tuning. This function evaluates the generated response based on its structural completeness and length. Specifically, it awards a bonus of 0.2 to the total reward if the response exceeds 500 words. Additionally, it assigns a reward of +1 for each required section heading correctly included in the output and penalizes the model by -1 for each duplicate heading present.

Due to the large scale of LLMs, which comprise billions of parameters, fine-tuning them is computationally intensive and necessitates substantial GPU memory resources. To overcome these issues and fine-tune the base models efficiently, we adopt LoRA, which is a parameter-efficient tuning technique that inserts trainable low-rank matrices into existing linear layers, such as those used in attention mechanisms (Hu et al. 2022). This method allows for significant reduction in memory and compute requirements by freezing the pre-trained weights and updating only a small number of additional parameters. Specifically, we applied LoRA-based fine-tuning to the attention projection layers (q_proj, k_proj, v_proj, o_proj) and Multi-Layer Perceptron (MLP) layers ($gate\_proj$, up_proj, $down\_proj$).

Experimental Settings

The DeepSeek-R1 8B and DeepSeek-R1 14B models were fine-tuned with GRPO using the 8-bit paged AdamW optimizer, with a learning rate of 5×10^{-6} scheduled via a cosine decay function. A per-device batch size of 2 was employed, and gradient accumulation over 4 steps was used to simulate a larger effective batch size. For each input prompt, four candidate responses were generated to compute group-based advantages. The DeepSeek-R1 8B model was fine-tuned for 3,000 steps on an NVIDIA RTX 4090 GPU, with a total training duration of 219 hours. Similarly, the DeepSeek-R1 14B model was fine-tuned for 3,000 steps on an NVIDIA RTX A6000 GPU with a total training duration of 322 hours.

Results

Evaluation Protocol

The evaluation focused on verifying the presence or absence of specific report sections and assessing their correct sequential ordering, following the same criteria used in the reward function. The performance of the fine-tuned models was benchmarked against several base models, including DeepSeek-R1 (8B, 14B, and 32B), Qwen2.5 (14B), and LLaMA3.1 (8B), to assess improvements in structured report generation. For performance comparison, reward average and standard deviation (SD) were calculated for all the models on the test set.

Results and Analysis

Figure 2 shows the moving average (window size = 20) and SD of rewards during GRPO fine-tuning, showing a clear upward trend as the models increasingly align with the reward function. Table 1 presents the average reward and SD achieved by various large language models on the test set, along with the GPU memory required at the time of inference. The fine-tuned DeepSeek-R1 models consistently outperformed the baseline models. Among all models evaluated, the fine-tuned DeepSeek-R1 14B model achieved the highest average reward of 16.38 with an SD of 3.51, indicating both improved performance and greater consistency in generating structurally accurate clinical summaries. The fine-tuned DeepSeek-R1 8B model also achieved comparable results with the mean reward and SD of 15.30 and 5.04 respectively, outperforming all the baseline models, including the DeepSeek-R1 32B model.

Conclusion

In this study, we explored the use of LLMs for the task of structuring unstructured clinical narratives, specifically discharge summaries from the MIMIC-IV dataset. We fine-tuned the DeepSeek-R1 8B and 14B models using GRPO to enforce structural fidelity in the generated outputs. The results demonstrate that the fine-tuned DeepSeek-R1 models outperformed all baseline models, including the larger DeepSeek-R1 32B model. This work highlights the potential of reinforcement learning techniques, specifically GRPO, in tailoring LLMs to domain-specific generation tasks that require adherence to strict structural constraints.

Future work will focus on applying our methodology to a wider spectrum of clinical documents in order to rigorously assess the robustness and generalizability of the models across heterogeneous healthcare datasets. Additionally, we plan to fine-tune alternative large language models, including LLaMA 3.1 and Qwen2.5, and perform a comprehensive comparative analysis of their performance relative to the fine-tuned DeepSeek-R1 models. Furthermore, we aim to integrate these optimized models into a broader agentic framework powered by large language models, designed for the automated generation of structured clinical reports and medical knowledge graphs from unstructured textual data.

References

Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J. D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33: 1877–1901.

Ding, N.; Qin, Y.; Yang, G.; Wei, F.; Zonghan, Y.; Su, Y.; Hu, S.; Chen, Y.; Chan, C.-M.; Chen, W.; Yi, J.; Zhao, W.; Wang, X.; Liu, Z.; Zheng, H.-T.; Chen, J.; Liu, Y.; Tang, J.; Li, J.; and Sun, M. 2023. Parameter-efficient fine-tuning of large-scale pre-trained language models. *Nature Machine Intelligence*, 5: 1–16.

Du, Y.; Oraby, S.; Perera, V.; Shen, M.; Narayan-Chen, A.; Chung, T.; Venkatesh, A.; and Hakkani-Tur, D. 2020. Schema-Guided Natural Language Generation. In Davis, B.; Graham, Y.; Kelleher, J.; and Sripada, Y., eds., *Proceedings of the 13th International Conference on Natural Language Generation*, 283–295. Dublin, Ireland: Association for Computational Linguistics.

Goldberger, A. L.; Amaral, L. A. N.; Glass, L.; Hausdorff, J. M.; Ivanov, P. C.; Mark, R. G.; and Stanley, H. E. 2000. PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals. *Circulation [Online]*, 101(23): e215–e220.

Grattafiori, A.; Dubey, A.; Jauhri, A.; Pandey, A.; Kadian, A.; Al-Dahle, A.; Letman, A.; Mathur, A.; Schelten, A.; Vaughan, A.; et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Gunter, D.; Puac Polanco, P.; Miguel, O.; Thornhill, R.; Yu, A.; Liu, Z.; Mamdani, M.; Pou-Prom, C.; and Aviv, R. 2022. Rule-based natural language processing for automation of stroke data extraction: a validation study. *Neuroradiology*, 64.

Guo, D.; Yang, D.; Zhang, H.; Song, J.; Zhang, R.; Xu, R.; Zhu, Q.; Ma, S.; Wang, P.; Bi, X.; et al. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Hu, E. J.; yelong shen; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; and Chen, W. 2022. LoRA: Low-Rank Adaptation of Large Language Models. In *International Conference on Learning Representations*.

Hu, Y.; Chen, Q.; Du, J.; Peng, X.; Keloth, V.; Zuo, X.; Zhou, Y.; Li, Z.; Jiang, X.; lu, Z.; Roberts, K.; and Qi, W. 2024. Improving large language models for clinical named entity recognition via prompt engineering. *Journal of the American Medical Informatics Association*, 31.

Huang, J.; Yang, D.; Rong, R.; Nezafati, K.; Treager, C.; Chi, Z.; Wang, S.; Cheng, X.; Guo, Y.; Klesse, L.; Xiao, G.; Peterson, E.; Zhan, X.; and Xie, Y. 2024. A critical assessment of using ChatGPT for extracting structured data from clinical notes. *npj Digital Medicine*, 7.

Huang, K.; Altosaar, J.; and Ranganath, R. 2019. Clinicalbert: Modeling clinical notes and predicting hospital readmission. *arXiv preprint arXiv:1904.05342*.

Johnson, A.; Pollard, T.; Horng, S.; Celi, L. A.; and Mark, R. 2023. MIMIC-IV-Note: Deidentified free-text clinical notes (version 2.2). https://doi.org/10.13026/1n74-ne17. PhysioNet.

Johnson, A. E.; Pollard, T. J.; Shen, L.; Lehman, L.-w. H.; Feng, M.; Ghassemi, M.; Moody, B.; Szolovits, P.; Celi, L. A.; and Mark, R. G. 2016. MIMIC-III, a freely accessible critical care database. *Scientific Data*, 3: 160035.

Jung, H.; Kim, Y.; Choi, H.; Seo, H.; Kim, M.; Han, J.; Kee, G.; Park, S.; Ko, S.; Kim, B.; Kim, S.; Jun, T. J.; and Kim, Y.-H. 2024. Enhancing Clinical Efficiency through LLM: Discharge Note Generation for Cardiac Patients. arXiv:2404.05144.

Kim, J.; Podlasek, A.; Shidara, K.; Liu, F.; Alaa, A.; and Bernardo, D. 2025. Limitations of Large Language Models in Clinical Problem-Solving Arising from Inflexible Reasoning. *arXiv preprint arXiv:2502.04381*.

Lee, J.; Yoon, W.; Kim, S.; Kim, D.; Kim, S.; So, C.; and Kang, J. 2019. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics (Oxford, England)*, 36.

Lehman, E.; Hernandez, E.; Mahajan, D.; Wulff, J.; Smith, M.; Ziegler, Z.; Nadler, D.; Szolovits, P.; Johnson, A.; and Alsentzer, E. 2023. Do We Still Need Clinical Language Models? In *Conference on Health, Inference, and Learning*, 578–597. PMLR.

Leong, H. Y.; Gao, Y. F.; Shuai, J.; Zhang, Y.; and Pamuksuz, U. 2024. Efficient fine-tuning of large language models for automated medical documentation. *arXiv preprint arXiv:2409.09324*.

Lu, Q.; Dou, D.; and Nguyen, T. 2022. ClinicalT5: A Generative Language Model for Clinical Text. In Goldberg, Y.; Kozareva, Z.; and Zhang, Y., eds., *Findings of the Association for Computational Linguistics: EMNLP 2022*, 5436–5443. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics.

Meystre, S.; Savova, G.; Kipper-Schuler, K.; and Hurdle, J. 2007. Extracting Information From Textual Documents in the Electronic Health Record: A Review of Recent Research. *Yearb Med Inform*, 128–144.

Mykowiecka, A.; Marciniak, M.; and Kupć, A. 2009. Rule-based information extraction from patients' clinical data. *J. of Biomedical Informatics*, 42(5): 923–936.

National Library of Medicine. 2025. PubMed. U.S. National Institutes of Health. Accessed April 20, 2025.

Patrick, J.; and Li, M. 2010. High accuracy information extraction of medication information from clinical notes: 2009 i2b2 medication extraction challenge. *Journal of the American Medical Informatics Association : JAMIA*, 17: 524–7.

Peng, Y.; Yan, S.; and Lu, Z. 2019. Transfer Learning in Biomedical Natural Language Processing: An Evaluation of BERT and ELMo on Ten Benchmarking Datasets. arXiv:1906.05474.

Raiaan, M. A. K.; Mukta, M. S. H.; Fatema, K.; Fahad, N. M.; Sakib, S.; Mim, M. M. J.; Ahmad, J.; Ali, M. E.; and Azam, S. 2024. A Review on Large Language Models: Architectures, Applications, Taxonomies, Open Issues and Challenges. *IEEE Access*, 12: 26839–26874.

Ramprasad, S.; Ferracane, E.; and Selvaraj, S. P. 2023. Generating more faithful and consistent SOAP notes using attribute-specific parameters. In Deshpande, K.; Fiterau, M.; Joshi, S.; Lipton, Z.; Ranganath, R.; Urteaga, I.; and Yeung, S., eds., *Proceedings of the 8th Machine Learning for Healthcare Conference*, volume 219 of *Proceedings of Machine Learning Research*, 631–649. PMLR.

Schulman, J.; Wolski, F.; Dhariwal, P.; Radford, A.; and Klimov, O. 2017. Proximal Policy Optimization Algorithms. *ArXiv*, abs/1707.06347.

Shao, Z.; Wang, P.; Zhu, Q.; Xu, R.; Song, J.; Bi, X.; Zhang, H.; Zhang, M.; Li, Y. K.; Wu, Y.; and Guo, D. 2024. DeepSeekMath: Pushing the Limits of Mathematical Reasoning in Open Language Models. arXiv:2402.03300.

Sohn, S.; Clark, C.; Halgrim, S.; Murphy, S.; and Liu, H. 2014. MedXN: an Open Source Medication Extraction and Normalization Tool for Clinical Text. *Journal of the American Medical Informatics Association : JAMIA*, 21.

Tang, A. Q.; Zhang, X.; and Dinh, M. N. 2024. Ignition-Innovators at "Discharge Me!": Chain-of-Thought Instruction Finetuning Large Language Models for Discharge Summaries. In Demner-Fushman, D.; Ananiadou, S.; Miwa, M.; Roberts, K.; and Tsujii, J., eds., *Proceedings of the 23rd Workshop on Biomedical Natural Language Processing*, 731–739. Bangkok, Thailand: Association for Computational Linguistics.

Taylor, N.; Zhang, Y.; Joyce, D. W.; Gao, Z.; Kormilitzin, A.; and Nevado-Holgado, A. 2024. Clinical Prompt Learning With Frozen Language Models. *IEEE Transactions on Neural Networks and Learning Systems*, 35(11): 16453–16463.

Uzuner, O.; Bodnari, A.; Shen, S.; Forbush, T.; Pestian, J.; and South, B. 2012. Evaluating the state of the art in coreference resolution for electronic medical records. *Journal of the American Medical Informatics Association : JAMIA*, 19: 786–91.

Yang, X.; Chen, A.; PourNejatian, N.; Shin, H. C.; Smith, K. E.; Parisien, C.; Compas, C.; Martin, C.; Costa, A. B.; Flores, M. G.; Zhang, Y.; Magoc, T.; Harle, C. A.; Lipori, G.; Mitchell, D. A.; Hogan, W. R.; Shenkman, E. A.; Bian, J.; and Wu, Y. 2022. A large language model for electronic health records. *npj Digital Medicine*, 5(1): 194.