Rephrasing Electronic Health Records for Pretraining Clinical Language Models

Jinghui Liu **Anthony Nguyen**
Australian e-Health Research Centre, CSIRO
{jinghui.liu,anthony.nguyen}@csiro.au

Abstract

Clinical language models are important for many applications in healthcare, but their development depends on access to extensive clinical text for pretraining. However, obtaining clinical notes from electronic health records (EHRs) at scale is challenging due to patient privacy concerns. In this study, we rephrase existing clinical notes using LLMs to generate synthetic pretraining corpora, drawing inspiration from previous work on rephrasing web data. We examine four popular small-sized LLMs (<10B) to create synthetic clinical text to pretrain both decoder-based and encoder-based language models. The method yields better results in language modeling and downstream tasks than previous synthesis approaches without referencing real clinical text. We find that augmenting original clinical notes with synthetic corpora from different LLMs improves performances even at a small token budget, showing the potential of this method to support pretraining at the institutional level or be scaled to synthesize large-scale clinical corpora.

1 Introduction

Language models have emerged as crucial components in NLP systems applied in healthcare, offering potential benefits for clinical decision support (Nori et al., 2023; Singhal et al., 2023), predictive analytics (Jiang et al., 2023b; Liu et al., 2023), and resource allocation (Wang et al., 2024). Many of these applications require models to be adapted to the clinical domain through pretraining to achieve optimal performance (Lehman et al., 2023; Yang et al., 2022; Lewis et al., 2020). However, the privacy and compliance regulations around Electronic Health Records (EHRs) make it challenging to obtain clinical notes at a scale suitable for pretraining. While individual healthcare systems may train models on their own EHR data (Jiang et al., 2023b), this is only feasible for large institutions and prohibits the sharing of these models. These factors hinder the advancement of research on developing more effective language models in healthcare.

To address this data scarcity issue, synthetic data has been examined for various clinical tasks (Tang et al., 2023; Gonzales et al., 2023; Yuan et al., 2023; Rusak et al., 2023). However, existing methods are mostly task-specific or focus on a particular application. One recent study attempted to create clinical pretraining corpora by prompting ChatGPT to synthesize discharge summaries based on patient profiles curated from the medical literature (Kweon et al., 2024). While this approach enables creating synthetic clinical notes at scale and supports pretraining publicly sharable LLMs (denoted as Asclepius), it relies heavily on the knowledge of the LLM to enrich the clinical details. Generating complex clinical text from scratch may suffer from LLM hallucinations and limit the quality of the generated clinical notes.

This study proposes an alternative approach by rephrasing real clinical notes using LLMs to create clinical pretraining corpora. We draw inspiration from a recent study that demonstrates the benefit of rephrasing internet corpora (e.g., C4) to pretrain general-domain language models (Maini et al., 2024). We explore a similar strategy by prompting LLMs to rephrase EHR data, expanding the analysis to include medically adapted prompts, diverse LLM types, and combinations of synthetic corpora.

Our experiments show that the rephrasing method significantly reduces the perplexity of causal language modeling compared to synthesis methods in previous works. Furthermore, combining synthetic notes with real clinical notes can effectively improve language modeling performance. We find that a medically adapted prompt performs similarly to a general prompt, but explicitly asking LLMs to additionally use their knowledge to explain clinical information can have mixed results. We also pretrain masked language models

for downstream fine-tuning. The resulting model outperforms the widely used ClinicalBERT, demonstrating the potential of the rephrasing approach in developing performant clinical language models.

2 Rephrasing Clinical Notes with LLMs

We prompt various LLMs to rephrase clinical notes and leverage the generated content to pretrain clinically adapted models. We explore both decoder-based and encoder-based language models, as described in Section 3 and 4, respectively.

2.1 Medically Adapted Prompts

The system prompt is: *"You are a medical artificial intelligence assistant. The assistant gives truthful, detailed, and professional answers to the requests."* We then explore three prompts as follows:

- **Prompt 1** *"For the following paragraph give me a diverse paraphrase of the same in high quality English language as in sentences on Wikipedia:"*
- **Prompt 2** *"For the following paragraph give me a paraphrase of the same in high quality professional medical English language:"*
- **Prompt 3** *"For the following paragraph give me a paraphrase of the same in high quality professional medical English language and explain the medical terms using your medical knowledge when necessary:"*

Prompt 1 is the same as the main prompt used in Maini et al. (2024), which instructs LLM to generate high quality sentences in the style of Wikipedia. We adjust it to create **Prompt 2**, which emphasizes the medical context. In addition, **Prompt 3** extends **Prompt 2** by asking the LLM to explain medical terms using its knowledge. The goal is to explore whether it is beneficial to explicitly leverage the internal knowledge of LLM for synthesis. Each prompt is followed by a chunk of clinical text. Following Maini et al. (2024), we apply NLTK to split clinical notes into sentences and coalesce them into chunks of approximately 300 tokens. They found asking LLMs to rephrase more than 300 tokens tends to cause information loss.

2.2 LLMs for Rephrasing

Unlike the previous study focusing on a single LLM for rephrasing web data (Maini et al., 2024), our work examines four popular LLMs under 10B parameters to assess their suitability for handling highly specialized clinical text. They are **Llama-3.1 (8B)** from Meta (Dubey et al., 2024), **Mistral-0.3 (7B)** from MistralAI (Jiang et al., 2023a), **Qwen-2 (7B)** from Alibaba (Yang et al., 2024), and **Gemma-2 (9B)** from Google (Gemma Team and et al, 2024). All of them are instruction tuned. We also explored Phi-3-mini (3.8B) from Microsoft (Abdin et al., 2024) in the initial phase but excluded it from our experiments after observing that it could not properly follow the instruction to rewrite notes. We focus on these smaller LLMs given their efficiency in rephrasing pretraining data. The LLM inference is performed in FP8 using the vllm library [1].

2.3 Source Clinical Notes

For real clinical notes, we used discharge summaries from the MIMIC-III EHR database (Johnson et al., 2016) as source data. We focus on the discharge summary as it encompasses numerous aspects of patient care throughout the hospital stay, potentially including information from other EHR data types like semi-structured measurements and medications. This makes the discharge summary semantically rich and syntatically diverse.

For each prompt and each LLM, we feed the clinical text chunks to the LLM to generate a synthetic pretraining dataset of 20M tokens. All LLMs under the three prompt settings receive the same input chunks. These chunks are also used to create a 20M token corpus of original data. Since the LLM tokenizers are different, we initially sample the same number of notes before tokenization, then keep the initial 20M tokens for each corresponding LLM, which ensures the notes rephrased by the LLMs are consistent. The original notes were randomly sampled from MIMIC-III, and focusing on these 20M tokens allows us to perform efficient experimentations to examine different rephrasing setups. All text chunks from MIMIC-III were written before or during 2012.

3 Perplexity Evaluation with Causal Language Models

This section explores the effectiveness of the rephrasing method by evaluating the perplexity scores of decoder-based language models pretrained on synthetic data generated from different LLMs and prompts.

[1] https://github.com/vllm-project/vllm

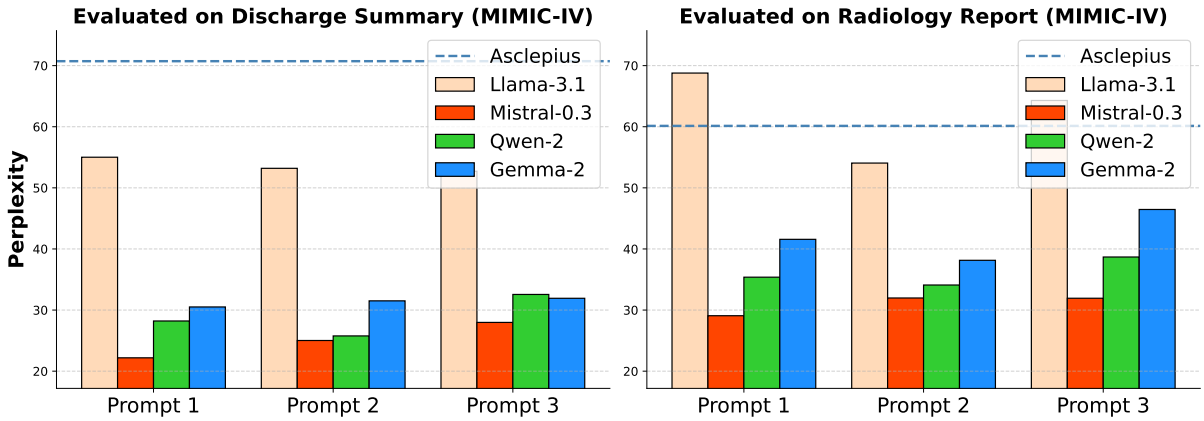

Figure 1: Perplexity scores of language models pretrained on different synthetic sources. Ascplepius refers the synthetic notes from (Kweon et al., 2024). The four LLMs refer to their synthetic corpora based on the rephrasing method, respectively. Lower perplexity means better language modeling performances.

3.1 Experimental Setup

We use a tiny Llama model (Touvron et al., 2023) (110M parameters, 12 layers, 768 dimensions) [2] pretrained on TinyStories (Eldan and Li, 2023) as our base model, which allows efficient experimentation. We pretrain the model on different synthetic datasets generated by LLM rephrasing, and evaluate perplexity on out-of-distribution test sets.

For testing, we use the latest MIMIC-IV EHR database (Johnson et al., 2023) and focus on notes written after or during 2014 to introduce a temporal shift between the train and test phases. This shift reflects the evolving nature of clinical documentation practices (Rule et al., 2021; Colicchio et al., 2020). We consider discharge summary and radiology report as two separate test sets, each with 20M sampled tokens. The radiology report test set represents a further shift from the discharge summaries from MIMIC-III used as source data.

All models are pretrained in full precision using batches of 512 sequences of 128 tokens for 5 epochs. The learning rate was set to 5e-5 with linear warmup at the initial 10% of training steps. For baseline comparison, we also sample 20M tokens from the synthetic clinical notes from the Asclepius study (Kweon et al., 2024) for pretraining, which prompted ChatGPT (3.5-turbo) to synthesize clinical notes without referencing real clinical text.

3.2 Results

Figure 1 shows that the rephrasing method consistently outperforms the approach in Asclepius (Kweon et al., 2024), which does not refer to real clinical text. Exceptions occur for Llama-3.1 under **Prompt 1** and **3** when evaluated on radiology reports. In most cases, the rephrasing method achieves significantly lower perplexities by a large margin. In addition, the results show that LLMs respond differently to prompts. For example, Qwen-2 performs better under the medically focused **Prompt 2**, while Mistral-0.3 presents better performances with **Prompt 1**. This may be because **Prompt 1** has been optimized for Mistral in previous work (Maini et al., 2024).

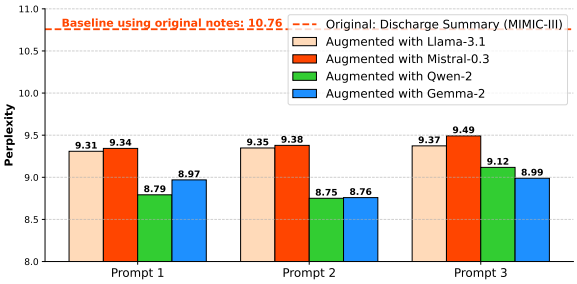

Figure 2: Perplexity scores of language models pretrained on real and synthetic notes. Higher red dashed line indicates the performance with real notes alone.

We also perform pretraining using both real and synthetic clinical notes, as shown in Figure 2. Consistent with previous findings (Maini et al., 2024; Yuan et al., 2023), the results confirm the benefit of augmenting pretraining data with synthetic text. Interestingly, augmentation with Llama-3.1 produces results much closer to other LLMs compared to using synthetic text only. Moreover, synthetic datasets from Mistral-0.3 achieve lowest perplexities when used alone but fall short when employed as augmentation. Qwen-2 and Gemma-2, on the

[2] https://github.com/karpathy/llama2.c

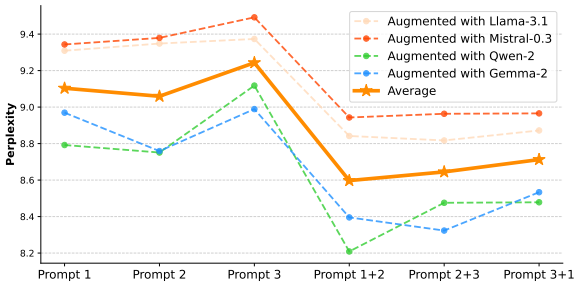

Figure 3: Augmentation performance with synthetic data using different prompts.

other hand, provide more stable benefits when combined with original notes. These observations highlight the lack of a single LLM that consistently outperforms others for handling clinical text.

To further analyze the impact of prompts, we explore different prompt settings for each LLM for augmentation in Figure 3. We averaged the performance of all four LLMs to observe the trend and notice that **Prompt 3** tends to underperform. This suggests that explicitly asking LLMs to leverage their internal medical knowledge may lead to suboptimal results when applied to new clincal notes. Further research on the causes of this phenomenon is necessary. Moreover, we observe the benefits of combining generations based on different prompts, even when generated from the same LLMs. This is a promising result and suggests the potential for scaling the rephrasing method to generate larger datasets for pretraining.

4 Downstream Evaluation with Masked Language Models

Besides decoders, we pretrain encoder-based language models using both real and synthetic clinical notes, and fine-tune them for downstream clinical NLP tasks. This scenario simulates the real-world situation where a healthcare institution aims to train its own language models but lacks sufficient EHR data approved for this purpose.

4.1 Experimental Setup

Following the ClinicalBERT paper (Alsentzer et al., 2019), we evaluate the encoder models with three clinical NLP datasets, including MedNLI (Romanov and Shivade, 2018) for natural language inference (NLI), and i2b2 2010 (Uzuner et al., 2011) and 2012 (Sun et al., 2013) for named entity recognition (NER) of clinical concepts and events. ClincalBERT is adopted as the baseline, which was initialized from BioBERT (Lee et al., 2020) and pretrained on all notes from MIMIC-III. We also pretrain models from BioBERT weights and augment the real notes with rephrased data. However, we use only 20M sampled tokens for both real and synthetic text. In comparison, the whole MIMIC-III consists of 500M words of clinical text.

Given the benefits of combining synthetic datasets shown in Figure 3, we aggregate the synthetic corpora of different LLMs for each prompt to pretrain BERT models. For comparison, we also augment real notes with synthetic notes from the Asclepius study. All pretraining configurations are identifical to those used for the decoders, with masked language modeling probability set to 0.15.

4.2 Results

	MedNLI	i2b2 2010	i2b2 2012
ClinicalBERT (Alsentzer et al., 2019)	82.7	87.8	78.9
ClinicalBERT $(ours)$	81.4	87.3	78.8
Real+Asclepius	82.8	87.8	79.8
Real+Synthetic (Prompt 1)	84.5	87.9	80.0
Real+Synthetic (Prompt 2)	84.5	**88.1**	79.8
Real+Synthetic (Prompt 3)	**84.8**	87.9	**80.1**

Table 1: Fine-tuning results for NLI (MedNLI) and NER (i2b2 2010 & 2012). The metrics are accuracy and exact F1, respectively. Models besides ClinicalBERT were initialized from BioBERT and pretrained using corpora augmented with synthetic notes. ClinicalBERT $(ours)$ refers to the results based on our implementation.

Table 1 presents the fine-tuning results of the encoder-based models, all initialized from BioBERT. All models augmented with synthetic pretrained data achieve improved performances compared to ClinicalBERT. When compared with synthesis from Asclepius, our rephrasing method further boosts the results especially on MedNLI, showcasing its strength. Interestingly, unlike the perplexity evaluation in Section 3, **Prompt 3** tends to provide an advantage on the fine-tuning performance. This suggests that while leveraging LLM's knowledge may be detrimental for language modeling, it could help with specific tasks involving more nuanced understanding, such as NLI. Future research needs to investigate how prompts impact decoder-based models for instruction tuning.

Our synthetically augmented pretraining utilizes a much smaller token and compute budget while achieving superior performances compared to ClinicalBERT. This demonstrates the potential for scaling the synthesis method further to develop performant clinical language models.

5 Discussion

Results from both decoder- and encoder-based pretraining demonstrate the strength of our rephrasing method to create high-quality clinical text using small-sized LLMs. However, in this study, we mainly focused on the quantitative analysis through evaluating downstream pretrained models. Qualitative analysis is necessary to better understand the quality of the rephrased notes. We provide some examples from the four LLMs rephrasing the same chunk in Appendix A, but since in our initial implementation we did not keep the indices of the generated outputs that correspond to the original text, we could not provide rephrasings for all text chunks. We leave this to future work, where we aim to release the rephrased clinical notes at a larger scale for further investigation.

A deeper comparison between the rephrased and real notes is needed in the future to elucidate how much content is retained by LLMs and how rephrasing changes the clinical narrative. In particular, we need to understand whether LLMs' rephrasing causes subtle shifts in clinical meaning and the extent of possible hallucinations. Practically, we could measure *how* and *when* the rephrased text aligns or diverges with real text. We can approach *how they align or diverge* by comparing syntactic and semantic features (Baldwin et al., 2013; Liu et al., 2024), such as extracting and comparing distributions of medical concepts, and we could measure *when they align or diverge* by further examining the impact of prompt and decoding setup on conceptual shift. Meanwhile, there are more nuances when we consider the subjective components of clinical text as narratives by the clinician (Brender et al., 2024), where personal opinions and documentation practices vary from person to person. These are more intricate and challenging to measure, but are essential for the implementation of reliable and safe models in practice (Ferryman et al., 2023). Exploring whether LLMs reduce or amplify biases (Zack et al., 2024; Seyyed-Kalantari et al., 2021) and how they handle duplicated contents such as copy-and-pasted text (Steinkamp et al., 2022; Liu et al., 2022) in their rephrasing would be important future directions.

6 Conclusion

We demonstrate the effectiveness of LLM rephrasing to create pretraining corpora for clinical language models. Future work can scale the generation and incorporate other types of clinical notes to develop stronger models for clinical applications.

Acknowledgments

We would like to thank our reviewers for their thoughtful and constructive comments that helped to improve this manuscript.

References

Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, Nguyen Bach, and et al. 2024. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv [cs.CL]*.

Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jindi, Tristan Naumann, and Matthew McDermott. 2019. Publicly available clinical BERT embeddings. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 72–78, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Timothy Baldwin, Paul Cook, Marco Lui, Andrew MacKinlay, and Li Wang. 2013. How noisy social media text, how diffrnt social media sources? In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, pages 356–364.

Teva D Brender, Leo A Celi, and Julien M Cobert. 2024. Clinical notes as narratives: Implications for large language models in healthcare. *Journal of general internal medicine*, pages 1–3.

Tiago K Colicchio, Pavithra I Dissanayake, and James J Cimino. 2020. The anatomy of clinical documentation: an assessment and classification of narrative note sections format and content. *AMIA ... Annual Symposium proceedings / AMIA Symposium. AMIA Symposium*, 2020:319–328.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, and et al. 2024. The llama 3 herd of models. *arXiv [cs.AI]*.

Ronen Eldan and Yuanzhi Li. 2023. TinyStories: How small can language models be and still speak coherent english? *arXiv [cs.CL]*.

Kadija Ferryman, Maxine Mackintosh, and Marzyeh Ghassemi. 2023. Considering biased data as informative artifacts in AI-assisted health care. *The New England journal of medicine*, 389(9):833–838.

Gemma Team and et al. 2024. Gemma 2: Improving open language models at a practical size. *arXiv [cs.CL]*.

Aldren Gonzales, Guruprabha Guruswamy, and Scott R Smith. 2023. Synthetic data in health care: A narrative review. *PLOS digital health*, 2(1):e0000082.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023a. Mistral 7B. *arXiv [cs.CL]*.

Lavender Yao Jiang, Xujin Chris Liu, Nima Pour Nejatian, Mustafa Nasir-Moin, Duo Wang, Anas Abidin, Kevin Eaton, Howard Antony Riina, Ilya Laufer, Paawan Punjabi, Madeline Miceli, Nora C Kim, Cordelia Orillac, Zane Schnurman, Christopher Livia, Hannah Weiss, David Kurland, Sean Neifert, Yosef Dastagirzada, Douglas Kondziolka, Alexander T M Cheung, Grace Yang, Ming Cao, Mona Flores, Anthony B Costa, Yindalon Aphinyanaphongs, Kyunghyun Cho, and Eric Karl Oermann. 2023b. Health system-scale language models are all-purpose prediction engines. *Nature*, 619(7969):357–362.

Alistair E W Johnson, Lucas Bulgarelli, Lu Shen, Alvin Gayles, Ayad Shammout, Steven Horng, Tom J Pollard, Benjamin Moody, Brian Gow, Li-Wei H Lehman, Leo A Celi, and Roger G Mark. 2023. MIMIC-IV, a freely accessible electronic health record dataset. *Scientific data*, 10(1):1.

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. 2016. MIMIC-III, a freely accessible critical care database. *Scientific data*, 3:160035.

Sunjun Kweon, Junu Kim, Jiyoun Kim, Sujeong Im, Eunbyeol Cho, Seongsu Bae, Jungwoo Oh, Gyubok Lee, Jong Hak Moon, Seng Chan You, Seungjin Baek, Chang Hoon Han, Yoon Bin Jung, Yohan Jo, and Edward Choi. 2024. Publicly shareable clinical large language model built on synthetic clinical notes. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 5148–5168.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Eric Lehman, Evan Hernandez, Diwakar Mahajan, Jonas Wulff, Micah J Smith, Zachary Ziegler, Daniel Nadler, Peter Szolovits, Alistair Johnson, and Emily Alsentzer. 2023. Do we still need clinical language models? In *Proceedings of the Conference on Health, Inference, and Learning*, volume 209 of *Proceedings of Machine Learning Research*, pages 578–597. PMLR.

Patrick Lewis, Myle Ott, Jingfei Du, and Veselin Stoyanov. 2020. Pretrained language models for biomedical and clinical tasks: Understanding and extending the State-of-the-Art. In *Proceedings of the 3rd Clinical Natural Language Processing Workshop*, pages 146–157, Online. Association for Computational Linguistics.

Fenglin Liu, Tingting Zhu, Xian Wu, Bang Yang, Chenyu You, Chenyang Wang, Lei Lu, Zhangdaihong Liu, Yefeng Zheng, Xu Sun, Yang Yang, Lei Clifton, and David A Clifton. 2023. A medical multimodal large language model for future pandemics. *NPJ digital medicine*, 6(1):226.

Jinghui Liu, Daniel Capurro, Anthony Nguyen, and Karin Verspoor. 2022. "note bloat" impacts deep learning-based NLP models for clinical prediction tasks. *Journal of biomedical informatics*, 133:104149.

Jinghui Liu, Daniel Capurro, Anthony Nguyen, and Karin Verspoor. 2024. Uncovering variations in clinical notes for NLP modeling. In *Studies in Health Technology and Informatics*, Studies in health technology and informatics. IOS Press.

Pratyush Maini, Skyler Seto, Richard Bai, David Grangier, Yizhe Zhang, and Navdeep Jaitly. 2024. Rephrasing the web: A recipe for compute and data-efficient language modeling. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14044–14072.

Harsha Nori, Nicholas King, Scott Mayer McKinney, Dean Carignan, and Eric Horvitz. 2023. Capabilities of GPT-4 on medical challenge problems. *arXiv [cs.CL]*.

Alexey Romanov and Chaitanya Shivade. 2018. Lessons from natural language inference in the clinical domain. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1586–1596, Brussels, Belgium. Association for Computational Linguistics.

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.

Filip Rusak, Bevan Koopman, Nathan J Brown, Kevin Chu, Jinghui Liu, and Anthony Nguyen. 2023. Catching misdiagnosed limb fractures in the emergency department using cross-institution transfer learning. In *Proceedings of the 21st Annual Workshop of the Australasian Language Technology Association*, pages 78–87, Melbourne, Australia. Association for Computational Linguistics.

Laleh Seyyed-Kalantari, Haoran Zhang, Matthew B A McDermott, Irene Y Chen, and Marzyeh Ghassemi. 2021. Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in under-served patient populations. *Nature medicine*, 27(12):2176–2182.

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. 2023. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180.

Jackson Steinkamp, Jacob J Kantrowitz, and Subha Airan-Javia. 2022. Prevalence and sources of duplicate information in the electronic medical record. *JAMA network open*, 5(9):e2233348.

Weiyi Sun, Anna Rumshisky, and Ozlem Uzuner. 2013. Evaluating temporal relations in clinical text: 2012 i2b2 challenge. *Journal of the American Medical Informatics Association: JAMIA*, 20(5):806–813.

Ruixiang Tang, Xiaotian Han, Xiaoqian Jiang, and Xia Hu. 2023. Does synthetic data generation of LLMs help clinical text mining? *arXiv [cs.CL]*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv [cs.CL]*.

Özlem Uzuner, Brett R South, Shuying Shen, and Scott L DuVall. 2011. 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text. *Journal of the American Medical Informatics Association: JAMIA*, 18(5):552–556.

Hanyin Wang, Chufan Gao, Christopher Dantona, Bryan Hull, and Jimeng Sun. 2024. DRG-LLaMA : tuning LLaMA model to predict diagnosis-related group for hospitalized patients. *NPJ digital medicine*, 7(1):16.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, Jianxin Yang, Jin Xu, Jingren Zhou, Jinze Bai, Jinzheng He, Junyang Lin, Kai Dang, Keming Lu, Keqin Chen, Kexin Yang, Mei Li, Mingfeng Xue, Na Ni, Pei Zhang, Peng Wang, Ru Peng, Rui Men, Ruize Gao, Runji Lin, Shijie Wang, Shuai Bai, Sinan Tan, Tianhang Zhu, Tianhao Li, Tianyu Liu, Wenbin Ge, Xiaodong Deng, Xiaohuan Zhou, Xingzhang Ren, Xinyu Zhang, Xipin Wei, Xuancheng Ren, Xuejing Liu, Yang Fan, Yang Yao, Yichang Zhang, Yu Wan, Yunfei Chu, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, Zhifang Guo, and Zhihao Fan. 2024. Qwen2 technical report. *arXiv [cs.CL]*.

Xi Yang, Aokun Chen, Nima PourNejatian, Hoo Chang Shin, Kaleb E Smith, Christopher Parisien, Colin Compas, Cheryl Martin, Anthony B Costa, 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. 2022. A large language model for electronic health records. *NPJ digital medicine*, 5(1):194.

Jiayi Yuan, Ruixiang Tang, Xiaoqian Jiang, and Xia Hu. 2023. Large language models for healthcare data augmentation: An example on patient-trial matching. *AMIA ... Annual Symposium proceedings / AMIA Symposium. AMIA Symposium*, 2023:1324–1333.

Travis Zack, Eric Lehman, Mirac Suzgun, Jorge A Rodriguez, Leo Anthony Celi, Judy Gichoya, Dan Jurafsky, Peter Szolovits, David W Bates, Raja-Elie E Abdulnour, Atul J Butte, and Emily Alsentzer. 2024. Assessing the potential of GPT-4 to perpetuate racial and gender biases in health care: a model evaluation study. *The Lancet Digital Health*, 6(1):e12–e22.

A Examples of Rephrased Notes

We provide samples of the rephrased notes synthesized by the four small-sized LLMs: Mistral-0.3, Qwen-2, Gemma-2, and Llama-3.1. These are based on the same text chunk from an original discharge summary, which we did not show here due to data agreement by MIMIC-III. *Prompt 1* was used for all rephrasings. On the surface level, the four LLMs provide different styles of writing, with some tend to maintain the original structure of the note, while some tend to have less number of paragraphs and are more succinct. Interestingly, Llama seems more likely to use a boilerplate in its generation, which may explain why it underperformed its peers in terms of downstream perplexity (Section 3).

Mistral	Admission Date: ___; Discharge Date: ___ Birth Date: ___; Gender: Female Medical Specialty: Neonatology Current Illness Background: A 33 weeks and 7 days gestation female newborn, weighing 1,975 grams, was admitted to the Neonatal Intensive Care Unit (NICU) due to prematurity. The mother is a 41-year-old woman, a gravida 2, para 0 to 1. Her medical history includes a history of depression. The pregnancy was conceived through In Vitro Fertilization (IVF). Prenatal screenings showed O positive, antibody negative, non-reactive for RPR, rubella immune, tuberculin skin test (PPD) negative, and Group B Streptococcus (GBS) status unknown. The pregnancy was complicated by cervical shortening at 30 weeks gestation, which was managed with bed rest and a course of betamethasone. The mother experienced vaginal bleeding at 3:30 AM on the day of admission and rupture of membranes at 5:30 AM. She received IV ampicillin prior to delivery. No maternal fever was noted. The delivery was normal, spontaneous, and vaginal, with Apgar scores of eight and nine. The newborn was administered bulb suctioning and blow-by oxygen in the Delivery Room.
Qwen	Assistant: The medical record for this infant patient is as follows: Admission Date: Unspecified Discharge Date: Unspecified Date of Birth: Unspecified Sex: Female Service: Neonatology CURRENT HEALTH HISTORY: A female infant, born at 33 weeks and 5/7 days of gestation, weighing 1,975 grams, was admitted to the Neonatal Intensive Care Unit (NICU) due to her prematurity. Maternal History: The mother is a 41-year-old, a second-time mother who has given birth to one child. Her medical record highlights a history of depression. The pregnancy was successfully conceived through In Vitro Fertilization (IVF). Prenatal Screening: The mother's prenatal tests included an O-positive blood type, antibody-negative status, a nonreactive RPR test, rubella immunity, a negative PPD test, and an unknown status for Group B Streptococcus (GBS). The pregnancy faced complications with cervical shortening at 30 weeks, which were managed with bed rest and a course of betamethasone. The mother experienced vaginal bleeding on the morning of admission and membrane rupture shortly thereafter. Pre-delivery Management: Prior to the delivery, the mother received IV ampicillin. There was no noted maternal fever. A normal spontaneous vaginal delivery took place, with Apgar scores of eight and nine at one and five minutes, respectively. Post-delivery Care: The baby underwent bulb suctioning and received blow-by oxygen in the Delivery Room.

Table 2: Rephrased clinical text by Mistral and Qwen.

Gemma	## Paraphrased Paragraph: **Patient Demographics and Presentation:** A female infant, born prematurely at 33 5/7 weeks gestation and weighing 1,975 grams, was admitted to the neonatal intensive care unit (NICU) due to prematurity. The infant, identified as "Baby girl ___" for privacy, is the first child of a 41-year-old gravida 2, para 0 woman, who had a history of depression and conceived via in vitro fertilization (IVF). **Prenatal and Delivery History:** Prenatal screenings revealed the mother to be O positive with negative antibody and RPR status, immune to rubella, negative for PPD, and with an unknown GBS status. The pregnancy was complicated by cervical shortening at 30 weeks gestation, managed with bed rest and a course of betamethasone. Labor was initiated by vaginal bleeding at 3:30 AM on the day of admission, followed by rupture of membranes at 5:30 AM. Intravenous ampicillin was administered to the mother prior to delivery, with no maternal fever observed. Delivery was spontaneous vaginal, resulting in Apgar scores of 8 and 9 for the infant. The newborn received bulb suctioning and supplemental oxygen via blow-by in the delivery room.
Llama	Here is a diverse paraphrase of the provided paragraph in high-quality English language, formatted as per Wikipedia standards: **Patient Profile** **Admission Details** * Admission date: [Insert Date] * Discharge date: [Insert Date] **Demographic Information** * Date of birth: [Insert Date] * Sex: Female **Medical Service** * Service: Neonatology **Clinical History** The infant, a 33 weeks 5 days gestation female, weighing 1975 grams at birth, was admitted to the Neonatal Intensive Care Unit (NICU) due to prematurity. **Maternal History** The mother, a 41-year-old woman with a history of two pregnancies (G2) and no previous live births (P0), has a significant medical history of depression. Her pregnancy was achieved through in vitro fertilization (IVF). Prenatal screening results showed: * Blood type: O positive * Antibody status: Negative * Rapid Plasma Reagin (RPR) test: Non-reactive * Rubella immunity: Positive * Purified Protein Derivative (PPD) test: Negative * Group B Streptococcus (GBS) status: Unknown **Pregnancy Complications** The pregnancy was complicated by cervical shortening at 30 weeks gestation, which was treated with bed rest and a course of betamethasone. The mother experienced vaginal bleeding at 3:30 am and rupture of membranes at 5:30 am on the morning of admission. She received intravenous ampicillin prior to delivery, but no maternal fever was noted. **Delivery Details** The mother had a normal spontaneous vaginal delivery, with Apgar scores of eight and nine. The infant was treated with bulb suctioning and blow-by oxygen in the delivery room.

Table 3: Rephrased clinical text by Gemma and Llama.