e-Health CSIRO at “Discharge Me!” 2024: Generating Discharge Summary Sections with Fine-tuned Language Models

Jinghui Liu, Aaron Nicolson, Jason Dowling, Bevan Koopman, & Anthony Nguyen
Australian e-Health Research Centre, CSIRO, Brisbane, Australia
jinghui.liu@csiro.au

Abstract

Clinical documentation is an important aspect of clinicians’ daily work and often demands a significant amount of time. The BioNLP 2024 Shared Task on Streamlining Discharge Documentation (Discharge Me!) aims to alleviate this documentation burden by automatically generating discharge summary sections, including brief hospital course and discharge instruction, which are often time-consuming to synthesize and write manually. We approach the generation task by fine-tuning multiple open-sourced language models (LMs), including both decoder-only and encoder-decoder LMs, with various configurations on input context. We also examine different setups for decoding algorithms, model ensembling or merging, and model specialization. Our results show that conditioning on the content of discharge summary prior to the target sections is effective for the generation task. Furthermore, we find that smaller encoder-decoder LMs can work as well or even slightly better than larger decoder-based LMs fine-tuned through LoRA. The model checkpoints from our team (**aehrc**) are openly available.[1]

1 Introduction

Clinical documentation in the age of Electronic Health Records (EHRs) can be a significant burden to clinicians in recording clinical information effectively (Colicchio et al., 2020; Rule et al., 2021). This reduces the time clinicians spend interacting with their patients and could lead to stress and burnout (Colicchio et al., 2019), degrading both the quality of patient care and the experience of care providers (Shanafelt et al., 2016).

Language Models (LMs) have demonstrated impressive NLP capabilities and are considered to have the potential to reduce the clinical documentation burden by automatically generating clinical text (Patel and Lam, 2023; Roberts, 2024; Omiye et al., 2024). For example, a recent study (Van Veen et al., 2024) demonstrated that LMs can generate succinct clinical summaries from text including progress notes and patient-doctor dialogues, sometimes even preferred over those written by medical experts. The BioNLP 2024 Shared Task “Discharge Me!” (Xu et al., 2024) focuses on generating the discharge summary (or discharge note) to assess the potential of LMs for this specific type of clinical note, which is often more time-consuming for clinicians to document and also more challenging to model given its length and complexity.

This paper presents the submissions from e-Health CSIRO in the shared task. We approach the task by fine-tuning multiple open-sourced LMs, including both decoder-only and encoder-decoder models. We fine-tune these models to generate two specific sections from discharge notes: *brief hospital course* and *discharge instruction*, by conditioning on the prior content in the notes as context. We explore various configurations with input context, decoding, ensembling, and target specialization. We find that much smaller encoder-decoder LMs could have a slight edge over fine-tuning decoder-only LMs (all with the size of 7/8B parameters) with LoRA (Hu et al., 2022). Our best submission ranked 3^{rd} on the final leaderboard under both automatic and manual evaluation.

2 Methods

2.1 Task and Dataset

The Shared Task focuses on generating two important sections of discharge notes: brief hospital course (BHC) and discharge instruction (DI). The first section provides a snapshot of the important information about the patient care during the hospital, and the second a summary to communicate that information and instructions after leaving the hospital to patients. The audiences for the two

[1] https://github.com/JHLiu7/bionlp24-shared-task-discharge-me

Proceedings of the 23rd Workshop on Biomedical Language Processing, pages 675–684
August 16, 2024. ©2024 Association for Computational Linguistics

sections are different as the former is read by clinicians while the latter by patients. The Shared Task uses the MIMIC-IV database (Johnson et al., 2023) to curate the dataset consisting of 109,168 patients, which are split into Train (68,785), Validation (14,719), Phase I testing (14,702), and Phase II testing (10,962). Each patient has a discharge summary that includes both sections, and participants are allowed to utilize data elements in the EHR database beyond the note alone as input.

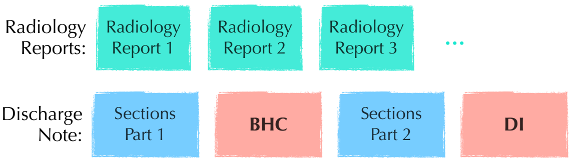

Figure 1: Illustration of the contents in clinical notes.

Our experiments focus only on the free-text clinical notes as input and do not consider other data modalities. We primarily use the content in the discharge note prior to the corresponding target section as input context. Radiology reports are considered optionally. We depict the note structures in Figure 1. Specifically, we consider the base context for BHC as C^{bhc}_{base} = *"Sections Part 1"*, and for DI as C^{di}_{base} = *"BHC"* + *"Sections Part 2"*. We consider two types of prolonged contexts: 1) $C_{base+rad} = C_{base}$ + *"Rad Reports"*, where radiology reports are concatenated with with the related sections; and 2) C^{di}_{long} = *"Sections Part 1"* + C^{di}_{base}, which extends the input context for DI. We then train models to generate the target sections T^{bhc} and T^{di} based on the corresponding contexts.

2.2 Language Models

We consider both decoder-only and encoder-decoder LMs for our experiments. For decoder-only LMs, we examine three popular open-sourced models at 7/8 billion paramater levels, including Llama3-8B [2], Mistral-7B (Jiang et al., 2023), and Gemma-7B (Gemma Team, 2024), all based on the instruction-tuned versions, denoted as Llama3-it, Mistral-it, and Gemma-it. Additionally, we examine the base version of Llama3-8B, denoted simply as Llama3. For encoder-decoder LMs, we focus on PRIMERA (447M) (Xiao et al., 2022) and Long-T5 (770M, global attention) (Guo et al., 2022), both capable of handling long input and output lengths.

	# Max Tokens (Llama3)	# Max Tokens (PRIMERA)
C^{bhc}_{base}	2816	3328
C^{di}_{base}	2048	2048
C^{di}_{long}	4608	5120
$C^{bhc}_{base+rad}$	4608	5120
$C^{di}_{base+rad}$	3840	4096
T^{bhc}	1280	1280
T^{di}	512	512

Table 1: Number of maximum tokens for modeling.

To determine the maximum lengths for modeling, we calculate the 85th percentile of the number of tokens and round it up to a multiplier of 256 for each LM. We present the statistics for Llama-3 and PRIMERA in Table 1 as examples. With each LM, we train two independent models for BHC and DI. For decoder-only LMs, we construct the prompt template similar to Alpaca (Taori et al., 2023), shown in Appendix Figure 2.

We then fine-tune these LMs for the text generation task. The decoder-only LMs, on the other hand, are loaded in half-precision (BF16) and fine-tuned through LoRA. We follow the setup from Dettmers et al. (2023) and use $lr = $ 2e-4, $r = 64$, $alpha = 16$, with LoRA attached to all linear layers. The encoder-decoder LMs are fully fine-tuned with $lr =$ 5e-5. All LMs are trained with batch size of 16 for 5 epochs using Adam, with 3% ratio for linear warmup. We use the default generation configuration, including the decoding algorithms, for the pretrained LMs. All experiments are performed on NVIDIA H100 GPU.

2.3 Evaluation

The automatic evaluation is based on 8 popular pairwise metrics, including BLEU-4 (Papineni et al., 2002), ROUGE-1/2/L (Lin, 2004), BERTScore (Zhang* et al., 2020), Meteor (Banerjee and Lavie, 2005), AlignScore (Zha et al., 2023), and MEDCON (Yim et al., 2023). They present a diverse set of measurements for string overlaps, semantic similarity, and medical concept mapping. The results for BHC and DI are averaged for each metric. The final ranking of the Shared Task is based on the average of the all scores on 250 hidden cases from Phase II testing, although participants are required to submit generation for all cases.

2.4 Experimental Setup

We investigate several factors that could impact the generation performance and compare them with

[2] https://ai.meta.com/blog/meta-llama-3/

Model	Overall	BLEU-4	ROUGE-1	ROUGE-2	ROUGE-L	BERTScore	Meteor	AlignScore	MEDCON
Fine-tuning baselines based on C_{base}									
Llama3	28.05	10.05	35.65	13.56	25.65	38.66	39.98	25.93	34.90
Llama3-it	23.53	7.88	25.56	9.66	15.70	35.13	38.90	22.73	32.69
Mistral-it	23.71	5.46	32.43	12.23	21.04	30.58	34.49	23.11	30.34
Gemma-it	25.14	6.31	35.04	11.18	24.53	32.91	36.07	23.46	31.60
PRIMERA	29.17	10.55	40.33	15.94	25.69	41.17	37.92	26.49	35.28
Long-T5	22.47	6.31	30.16	8.88	19.12	32.31	31.44	22.50	29.07
Extended Input Context									
Llama3 w/ $C_{base+rad}$	25.15	8.69	27.24	10.81	19.20	37.26	39.17	25.11	33.71
PRIMERA w/ $C_{base+rad}$	29.10	10.64	39.76	15.75	27.10	40.31	37.55	27.10	34.61
Llama3 w/ C^{di}_{long}	28.33	9.56	37.27	12.93	25.87	38.67	40.64	26.67	35.04
PRIMERA w/ C^{di}_{long}	28.26	10.14	38.93	13.48	23.73	40.68	37.95	26.80	34.37
Unified LM for both T_{bhc} and T_{di}									
Llama3 (single)	25.38	7.89	31.79	11.34	21.89	35.38	38.91	23.39	32.42
Alternative Decoding for Llama3									
Llama3 w/ beam	25.20	10.06	29.14	8.05	17.83	37.05	40.57	26.17	32.71
Llama3 w/ constrastive	24.09	8.36	27.81	10.13	18.24	36.10	35.52	26.37	30.21
Ensemble Decoding									
Llama3 + Llama3	26.17	9.67	28.79	11.36	21.02	38.31	39.71	26.29	34.24
Llama3 + Llama3-it	27.04	9.66	32.68	12.99	22.30	37.96	39.79	26.10	34.84
Merging LoRA Adapters									
Llama3 x2 LoRA	25.78	8.20	34.48	11.60	22.69	35.81	37.44	23.25	32.79
Llama3 x4 LoRA	21.80	4.50	33.05	11.97	20.45	30.79	28.31	17.35	28.00

Table 2: Results from automatic evaluation, based on 250 hidden samples from Phase II testing.

the base generation setup, in which two LMs of the same architecture are trained on C_{base} for BHC and DI, respectively. We examine the impact of extended input context by replacing C_{base} with $C_{base+rad}$ or C^{di}_{base} with C^{di}_{long}. Taking Llama3 as the example, we explore a variety of modifications, including training a unified LM that models the two targets jointly to explore the benefit of target specialization. We also apply various decoding algorithms other than greedy search, including beam search ($n = 4$), and contrastive search ($\alpha = 0.6$, $k = 6$) (Su et al., 2022). Furthermore, we explore ensemble decoding (Manakul et al., 2023) and the popular adapter merging with Llama3 as the example. The former averages the logits from two LMs for generating each token with greedy search, and the latter applied TIES (Yadav et al., 2023) to merge the paramters of several LoRA adapters (equal weights, density of 0.5) before attaching it to the main LM. Finally, we prompt instruction-tuned LMs in the zero-shot manner, including the 70B checkpoints, on a subset of validation to observe the benefit of fine-tuning for this task.

3 Results & Analysis

3.1 Both Decoder and Encoder-encoder LMs Work Well When Fine-tuned

We firstly find all LMs obtain decent results when fine-tuned with C_{base}. Meanwhile, the instruction-tuned decoder-only LMs perform worse than the base version of Llama3. This aligns with existing findings that instruction tuning could harm performance on NLP benchmarks (Ouyang et al., 2022; Ivison et al., 2023). PRIMERA performs slightly better than Llama3, despite being the smallest model we examined. On the other hand, Long-T5 seems to struggle with the task.

3.2 Prior Context of Dicharge Note is Sufficient as Input

We observe poorer results when including radiology reports as supplementary input for both Llama3 and PRIMERA. Although the input context lengths increase more than 50% with the radiology reports, it appears that no new, valuable information is added. Instead, it misleads the LMs to produce worse outputs, especially for Llama3. This shows the content in the discharge notes have well captured free-text information from the existing EHR data. Using radiology reports alone offers an overall score from 19.1 to 20.3 (Appendix Table 4).

3.3 Prolonged Context in Discharge Note Offers Little Value

In a similar fashion, we extend the input context for DI by including contents prior to BHC, namely C^{di}_{long}. Again, more context does not necessarily lead to better results. We consider this is likely due to the fact that BHC and the content between BHC and DI have provided sufficient information for generating DI. Future work may explore how to further trim down the input to reduce the noise, such as through de-duplication (Kandpal et al., 2022; Liu et al., 2022), to enhance performance.

3.4 Two Specialized LMs are Better Than One Unified LM

Instead of trainig two copies of LM for each section, we combine samples for both targets together to train a single model that is capable to produce either of the sections. We explore this with Llama3, fine-tuned with LoRA in the same setup as previous. We see the unified Llama3 performs worse than the two independant copies of Llama3, demonstrating the importance of specialization in modeling BHC and DI independently. Furthermore, as the two copies share the same base model and differs only in adapters, keeping them separately does not lead to significantly more storage cost than the unified model.

3.5 Better Decoding Methods Lead to Mixed Results

The Phase II test results in Table 2 indicate that better decoding algorithms, such as beam search and contrastive search, could lead to worse results than the baseline greedy search. Interestingly, our initial experiments on the 1000 validation samples in Appendix Table 5 show that they are at least on par and sometimes better. The mixed results show the diversity of the dataset and the need to further investigate the distribution and biases of the data.

3.6 Ensemble Decoding is Not Helpful

An ensemble of two Llama3 models trained using different data or with different base LMs at the token level is not helpful. With Llama3 + Llama3, we ensemble Llama3 fine-tuned using C_{base} and $C_{base+rad}$, and with Llama3 + Llama3-it, we ensemble the base and instruction-tuned Llama3 fine-tuned both using C_{base}. Neither of these two pairs produced improved results. Although ensembling is found helpful previously for generation (Manakul et al., 2023), for our task naively averaging the logits at token-level during decoding is both inefficient and ineffective.

3.7 Merging Adapters is Not Helpful Either

Similarly, we perform another form of ensemble by merging the LoRA adapter weights for the same base LM. *Merging with x2 LoRA* is based on adapters trained using C_{base} and $C_{base+rad}$, while *merging with x4* further merges the adapters for BHC and DI. Both substantially decrease the performance, and merging adapters trained for different targets leads to the worst result in our fine-tuning experiments. This again shows that model specialization is important for the current task. In addition, it is possible that model merging tends to prevail in generating creative contents instead of improving the specific aspects of generation quality.

3.8 Fine-tuned LMs Substantially Outperform Out-of-box LMs

Finally, we prompt the instruction-tuned LMs in the zero-shot manner to compare with fine-tuned performance. Besides Llama3-8B-it and Mistral-7B-it, we additionally prompt the 70B scale Llama3-70B-it and Mixtral-8x7b-it (Jiang et al., 2024). They achieve an overall score ranging from 15.1 to 17.4 (details in Table 5), significantly fell short compared to the fine-tuned results. Although more advanced prompting strategies are expected to enhance performance, we suspect that fine-tuning would still be the more effective solution given the amount of training data.

4 Discussions

We demonstrate that fine-tuning LMs based solely on the prior content from the discharge note is sufficient to generate BHC and DI sections. Given the heterogeneity of EHR data (Yadav et al., 2018) and variations in clinical notes (Liu et al., 2024), selecting the appropriate inputs would be crucial for both the quality and applicability of the generation. In this work, we assume that the non-BHC/DI contents of the discharge note have been populated from other available sources or clinical notes, making them readily available as model input.

The context for BHC (*"Sections Part 1"* in Figure 1) typically includes chief complaint, history of present illness, past medical history, social history, physical exam, and various pertinent results. The *"Sections Part 2"* of DI context may include admission and discharge medications, discharge disposition, dischage diagnoses.

Using these sections as input yields competitive generation results, and including additional text sources like radiology reports does not lead to improvement. One explanation is that the sections within the discharge summary, such as "pertinent results", often already include imaging findings. Future work may futher investigate how selecting

Model	Overall	BLEU-4	ROUGE-1	ROUGE-2	ROUGE-L	BERTScore	Meteor	AlignScore	MEDCON
WisPerMed	33.2	12.4	45.3	20.1	30.8	43.8	40.3	31.5	41.1
HarmonAI Lab at Yale	30.0	10.6	42.3	18.0	28.4	41.2	38.1	26.5	35.3
aehrc (ours)	29.7	9.7	41.4	19.2	28.4	38.3	39.8	27.4	33.2
EPFL-MAKE	28.9	9.8	44.4	15.5	26.2	39.9	33.6	25.5	36.0
UF-HOBI	28.6	10.2	40.1	17.4	27.5	39.5	28.9	29.6	35.5

(a) Automatic evaluation results on 250 cases from Phase II test set.

Team	Average	BHC Completeness	BHC Correctness	BHC Readability	BHC Overall	DI Completeness	DI Correctness	DI Overall
WisPerMed	3.4	3.7	3.7	3.4	2.4	3.9	4.0	2.5
HarmonAI Lab at Yale	2.9	3.5	2.6	2.1	1.5	4.3	3.9	2.4
aehrc (ours)	2.8	2.3	3.1	2.0	1.1	3.9	4.5	2.6
EPFL-MAKE	2.7	3.3	2.8	2.5	1.7	3.5	3.4	1.9
UF-HOBI	2.6	2.5	3.4	2.7	1.4	3.0	3.3	1.8

(b) Manual evaluation results by clinicians on 25 selected cases.

Table 3: Results from the top-5 teams on the final Phase II leaderboard.

relevant content (Zheng et al., 2023) or removing redundant information (Liu et al., 2022) impacts the performance. It is also unclear whether other sources of EHR information should be considered, especially those not captured by the discharge summary. These include structured EHR data and other types of clinical text, such as nursing or physician notes. Regarding structured data elements, this study does not consider diagnosis codes like ICD or DRG (Dong et al., 2022; Liu et al., 2021b), as they are typically assigned after the patient discharge. However, future work could model other measurement data or codes from prior patient encounters. Examining the end-to-end generation of discharge notes solely from structured EHR data and other clinical notes is also important to ensure that the generation model integrates into different clincial documentation workflows.

From the modeling perspective, we find that fine-tuning smaller LMs, such as PRIMERA, achieves surprisingly good results. Examination of any potential biases or overfitting is left for future work. During development, we observed that the generation qualities of Llama3 and PRIMERA were similar (examples shown in Appendix Table 6 & 7) and had better quality compared to other LMs like Mistral (see Appendix Table 7), consistent with the quantitative analysis. We noticed that Llama3 tended to generate repetitive content more often and tried to alleviate this with better decoding techniques, but were unable to improve the overall performance on quantitative metrics (see Table 2). It is possible that more hyperparameter search on either fine-tuning or decoding could lead to improvement, which we leave to future work.

Given the slight edge over Llama3 and other LMs, PRIMERA was our final submission. Table 3 shows the final leaderboard, in which we rank 3^{rd} overall and are close to 2^{nd} under both automatic [3] and manual evaluation, with the latter conducted by a team of clinicians on 25 selected samples.

Similar to previous findings (Van Veen et al., 2024), we see that the manual evaluation aligns with the automatic evaluation in ranking different systems. The manual evaluation further reports fine-grained scores on *Completeness*, *Correctness*, and *Readbility* for BHC and DI separately. Interestingly, we observe that PRIMERA obtains the best overall score for DI but worst for BHC among the top-5 teams. This may indicate the model capacity correlates with the length or complexity of the target generation, with smaller LMs potentially struggling with prolonged outputs. It is plausible that Llama3 would offer improved results on BHC, especially in terms of readability. Future work may investigate this further through separate automatic evaluations specifically for BHC and DI.

5 Conclusion

This paper describes our efforts in the "Discharge Me!" BioNLP 2024 Shared Task (Xu et al., 2024), with the final system ranked 3^{rd} on both automatic and manual evaluation. We show that fine-tuning LMs with appropriate input context has the potential to automatically synthesize high-quality discharge summary sections, which holds promise to reduce the time clinicians spend on documentation.

[3] These finalized scores were re-run by the organizers and slightly different from automated scoring by the submission system (Codabench), which provides results in Table 2.

Limitations

Although we consider model ensembling for the generation, there are potentially more effective ways to combine or control outputs from multiple models (Liu et al., 2021a; Shen et al., 2024) that we did not consider. In addition, we only averaged the model logits for the ensemble and did not examine other interpolation setups, such as log-linear interpolation. Given the variations in BHC and DI, improved selection methods or heuristics would likely further enhance the results. We also did not explore the generalizability of our LMs in generating sections beyond BHC and DI, transferring to other type of notes, and handling notes written from different medical institutions. Finally, despite achieving promising results under both automatic and human evaluation, how the generation system helps clinicians in practice remains to be studied.

Acknowledgments

We would like to thank the shared task organizers for their dedication and help along the shared task process. We also thank the reviewers for their thoughtful comments on the initial submission to improve the paper.

References

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

Tiago K Colicchio, James J Cimino, and Guilherme Del Fiol. 2019. Unintended consequences of nationwide electronic health record adoption: Challenges and opportunities in the Post-Meaningful use era. *Journal of medical Internet research*, 21(6):e13313.

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.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. QLoRA: Efficient finetuning of quantized LLMs. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Hang Dong, Matúš Falis, William Whiteley, Beatrice Alex, Joshua Matterson, Shaoxiong Ji, Jiaoyan Chen, and Honghan Wu. 2022. Automated clinical coding: what, why, and where we are? *NPJ digital medicine*, 5(1):159.

Gemma Team. 2024. Gemma: Open models based on gemini research and technology.

Mandy Guo, Joshua Ainslie, David Uthus, Santiago Ontanon, Jianmo Ni, Yun-Hsuan Sung, and Yinfei Yang. 2022. LongT5: Efficient Text-To-Text transformer for long sequences. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 724–736, Seattle, United States. Association for Computational Linguistics.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-Rank adaptation of large language models. In *International Conference on Learning Representations*.

Hamish Ivison, Yizhong Wang, Valentina Pyatkin, Nathan Lambert, Matthew Peters, Pradeep Dasigi, Joel Jang, David Wadden, Noah A Smith, Iz Beltagy, and Hannaneh Hajishirzi. 2023. Camels in a changing climate: Enhancing LM adaptation with tulu 2.

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. 2023. Mistral 7B.

Albert Q Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, Gianna Lengyel, Guillaume Bour, Guillaume Lample, Lélio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, Szymon Antoniak, Teven Le Scao, Théophile Gervet, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2024. Mixtral of experts.

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.

Nikhil Kandpal, Eric Wallace, and Colin Raffel. 2022. Deduplicating training data mitigates privacy risks in language models. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 10697–10707. PMLR.

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.

Alisa Liu, Maarten Sap, Ximing Lu, Swabha Swayamdipta, Chandra Bhagavatula, Noah A Smith, and Yejin Choi. 2021a. DExperts: Decoding-Time controlled text generation with experts and Anti-Experts. 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 6691–6706, Online. Association for Computational Linguistics.

Jinghui Liu, Daniel Capurro, Anthony Nguyen, and Karin Verspoor. 2021b. Early prediction of diagnostic-related groups and estimation of hospital cost by processing clinical notes. *NPJ digital medicine*, 4(1):103.

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.

Potsawee Manakul, Yassir Fathullah, Adian Liusie, Vyas Raina, Vatsal Raina, and Mark Gales. 2023. CUED at ProbSum 2023: Hierarchical ensemble of summarization models. In *The 22nd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks*, pages 516–523, Toronto, Canada. Association for Computational Linguistics.

Jesutofunmi A Omiye, Haiwen Gui, Shawheen J Rezaei, James Zou, and Roxana Daneshjou. 2024. Large language models in medicine: The potentials and pitfalls : A narrative review. *Annals of internal medicine*.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Sajan B Patel and Kyle Lam. 2023. ChatGPT: the future of discharge summaries? *The Lancet. Digital health*, 5(3):e107–e108.

Kirk Roberts. 2024. Large language models for reducing clinicians' documentation burden. *Nature medicine*.

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.

Tait D Shanafelt, Lotte N Dyrbye, Christine Sinsky, Omar Hasan, Daniel Satele, Jeff Sloan, and Colin P West. 2016. Relationship between clerical burden and characteristics of the electronic environment with physician burnout and professional satisfaction. *Mayo Clinic proceedings. Mayo Clinic*, 91(7):836–848.

Shannon Zejiang Shen, Hunter Lang, Bailin Wang, Yoon Kim, and David Sontag. 2024. Learning to decode collaboratively with multiple language models.

Yixuan Su, Tian Lan, Yan Wang, Dani Yogatama, Lingpeng Kong, and Nigel Collier. 2022. A contrastive framework for neural text generation. In *Advances in Neural Information Processing Systems*.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B Hashimoto. 2023. Stanford alpaca: An instruction-following LLaMA model.

Dave Van Veen, Cara Van Uden, Louis Blankemeier, Jean-Benoit Delbrouck, Asad Aali, Christian Bluethgen, Anuj Pareek, Malgorzata Polacin, Eduardo Pontes Reis, Anna Seehofnerová, Nidhi Rohatgi, Poonam Hosamani, William Collins, Neera Ahuja, Curtis P Langlotz, Jason Hom, Sergios Gatidis, John Pauly, and Akshay S Chaudhari. 2024. Adapted large language models can outperform medical experts in clinical text summarization. *Nature medicine*, 30(4):1134–1142.

Wen Xiao, Iz Beltagy, Giuseppe Carenini, and Arman Cohan. 2022. PRIMERA: Pyramid-based masked sentence pre-training for multi-document summarization. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5245–5263, Dublin, Ireland. Association for Computational Linguistics.

Justin Xu, Zhihong Chen, Andrew Johnston, Louis Blankemeier, Maya Varma, Jason Hom, William J. Collins, Ankit Modi, Robert Lloyd, Benjamin Hopkins, Curtis Langlotz, and Jean-Benoit Delbrouck. 2024. Overview of the first shared task on clinical text generation: Rrg24 and "discharge me!". In *The 23rd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks*, Bangkok, Thailand. Association for Computational Linguistics.

Pranjul Yadav, Michael Steinbach, Vipin Kumar, and Gyorgy Simon. 2018. Mining electronic health records (EHRs): A survey. *ACM Comput. Surv.*, 50(6):1–40.

Prateek Yadav, Derek Tam, Leshem Choshen, Colin Raffel, and Mohit Bansal. 2023. TIES-Merging: Resolving interference when merging models. In *Thirty-*

seventh Conference on Neural Information Processing Systems.

Wen-Wai Yim, Yujuan Fu, Asma Ben Abacha, Neal Snider, Thomas Lin, and Meliha Yetisgen. 2023. Aci-bench: a novel ambient clinical intelligence dataset for benchmarking automatic visit note generation. *Scientific data*, 10(1):586.

Yuheng Zha, Yichi Yang, Ruichen Li, and Zhiting Hu. 2023. AlignScore: Evaluating factual consistency with a unified alignment function. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11328–11348, Toronto, Canada. Association for Computational Linguistics.

Tianyi Zhang*, Varsha Kishore*, Felix Wu*, Kilian Q Weinberger, and Yoav Artzi. 2020. BERTScore: Evaluating text generation with BERT. In *International Conference on Learning Representations*.

Hongyi Zheng, Yixin Zhu, Lavender Jiang, Kyunghyun Cho, and Eric Oermann. 2023. Making the most out of the limited context length: Predictive power varies with clinical note type and note section. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop)*, pages 104–108, Toronto, Canada. Association for Computational Linguistics.

A Appendix

Model	Llama3	PRIMERA
Overall	20.34	19.10
BLEU-4	5.27	3.52
ROUGE-1	27.11	30.56
ROUGE-2	7.30	8.39
ROUGE-L	17.16	18.82
BERTScore	30.03	30.42
Meteor	32.76	27.13
AlignScore	17.38	13.46
MEDCON	25.67	20.47

Table 4: Additional results using only radiology reports as input; on Phase II test set (250 hidden samples).

Prompt template for BHC

Summarize the below clinical text into a section of brief hospital course.

### Input:
`{{input_text}}`

### Summary:
`{{target_text}}`

Prompt template for DI

Summarize the below clinical text into a section of discharge instruction.

### Input:
`{{input_text}}`

### Summary:
`{{target_text}}`

Figure 2: Template used for decoder-only LMs. $target\_text$ is removed at inference time.

Model	Overall	BLEU-4	ROUGE-1	ROUGE-2	ROUGE-L	BERTScore	Meteor	AlignScore	MEDCON
Baseline									
Llama3	30.16	11.48	38.28	18.69	25.08	41.69	31.76	31.79	42.53
Alternative decoding									
Llama3 w/ beam	28.82	11.34	33.40	16.06	22.46	40.37	33.43	31.66	41.86
Llama3 w/ nucleus	28.13	9.66	37.74	16.41	22.47	39.79	33.42	27.74	37.81
Llama3 w/ contrastive	30.98	11.98	42.28	21.49	27.33	41.38	33.34	31.35	38.70
Zero-shot prompting									
Llama3-8B-it	15.05	0.97	19.92	3.88	10.65	18.47	19.35	25.47	21.70
Llama3-70B-it	15.62	0.95	21.73	4.59	11.31	19.16	20.03	23.84	23.33
Mistral-7B-it	17.31	1.61	30.32	6.97	15.83	23.56	20.33	16.22	23.62
Mixtral-8x7B-it	17.40	1.49	30.29	7.13	15.02	22.00	19.34	20.15	23.77

Table 5: Additional results on 1000 validation samples.

BHC by Llama3	Mr. ___ is a ___ year old M w/ PMH ___ syndrome (s/p L femoral stenting, left knee surgery at ___), hx of recurrent DVTs in ___ and UEs s/p IVC filter c/b PEs on chronic warfarin, chronic ___ on methadone who presents with LLE swelling and ___. # LLE swelling: Patient presented with LLE swelling and erythema. He was found to have a chronic DVT on LLE US. He was started on IV ceftriaxone for cellulitis. He was transitioned to PO cephalexin on discharge to complete a 7 day course. # ___ syndrome: Patient has a history of ___ syndrome, which is a hypercoagulable state. He has had multiple DVTs and PEs in the past. He was found to have a chronic DVT on LLE US. He was continued on his home warfarin. His INR was therapeutic at 2.3 on discharge. # Chronic ___: Patient has a history of chronic ___ on methadone. He was continued on his home methadone. # Gout: Patient has a history of gout. He was continued on his home allopurinol. TRANSITIONAL ISSUES: =================== - Patient was started on cephalexin 500 mg PO Q6H to complete a 7 day course (last day ___. - Patient's INR was 2.3 on discharge. Please check INR on ___ and adjust warfarin dose as needed. - Please follow up with patient's ___ clinic. - Please follow up with patient's PCP. - Please follow up with patient's vascular surgeon. - Please follow up with patient's pain clinic. - Please follow up with patient's hematologist. - Please follow up with patient's PCP regarding his chronic ___. Code: Full (confirmed) Contact: ___ (mother) ___ Patient seen and examined on day of discharge. >30 minutes on discharge activities. ___ ___ Attending ___ Date: ___ Time: ___ I have read and agree with the discharge summary. ___, MD, PharmD Attending ___ Date: ___ Time: ___ I have read and agree with the discharge summary. ___, MD Attending ___ Date: ___ Time: ___ I have read and agree with the discharge summary. ___, MD Attending
BHC by PRIMERA	Mr. ___ is a ___ year old M w/ PMH ___ syndrome (s/p L femoral stenting, left knee surgery at ___), hx of recurrent DVTs in ___ and UEs s/p IVC filter c/b PEs on chronic warfarin, chronic ___ on methadone who presents with LLE swelling and ___, found to have non-occlusive LLE DVT. #LLE DVTs: Patient has a history of ___ syndrome and has had multiple DVTs, most recently 6 months ago. He has been on warfaradin for this, but has had a subtherapeutic INR. He was admitted with a sub therapeutic INR of 4. He had a LLE US which showed non-occlusion of his LLE veins. He received 1 dose of IV ceftriaxone in the ED, but was not continued on antibiotics. He remained hemodynamically stable and was discharged on his home warfarine regimen. He will follow up with his PCP and ___ clinic. #Chronic ___: Patient was continued on his methadione. #Hx of PEs: Patient with history of PE on warfain, on warafarin. INR 4 on admission. He is followed by Hematology at ___ and has been subtheraputic on warferin. He reports compliance with warfarin. - Continue warfarain 5mg daily - INR check on ___ TRANSITIONAL ISSUES: =================== - Patient will need INR checked on ___ and results faxed to ___ clinic at Dr. ___ - Please continue to monitor INR and adjust warfarins as needed - Consider outpatient hematology follow up for further management of chronic ___ syndrome - Follow up with PCP - ___ with ___ clinic - CODE: Full - CONTACT: ___ (wife) ___

Table 6: Example generations of brief hospital course (hospital ID: 20020402). More examples can be obtained using the released model checkpoints.

DI by Llama3	Dear Ms. ___, You were admitted to ___ for ongoing fevers. You were evaluated by the Infectious Disease team. You had multiple tests performed to determine the cause of your fevers. You had a CT scan of your sinuses which showed sinusitis. You were started on antibiotics for this. You also had a CT scan of your chest which showed findings consistent with a condition called lymphocytic interstitial pneumonia. You were monitored for 24 hours and did not have any recurrent fevers. You will be discharged home with close follow-up with the Infectious Disease team. They will help to arrange follow-up with Rheumatology and Pulmonology. It was a pleasure taking care of you! Sincerely, Your ___ Team
DI by PRIMERA	Dear Ms. ___, It was a pleasure caring for you at ___. You were admitted for fevers. Your flu test was negative. Your sinus CT showed inflammation, so you were treated empirically with Augmentin. Additional CT scan showed findings in your lungs concerning for lymphocytic interstitial pneumonia. Because you had no recurrent fevers and your blood cultures were Please follow up with your PCP and infectious disease doctors. ___ wish you the best! Sincerely, Your ___ Team
DI by Mistral	Discharge Instructions: Dear Ms. ___, It was a pleasure caring for you at ___ ___. You were admitted to the hospital for fevers. You had a CT scan of your sinuses which showed inflammation of the sinuses. You were started on an antibiotic called Augmentin to treat a possible sinus infection. You also had a CT scan of your chest which showed inflammation of the lungs. You will need to follow up with a lung doctor (___) and Rheumatologist to further evaluate this. The Infectious Disease doctors ___ help to arrange these appointments. Please continue to take your medications as prescribed and keep your follow-up appointments as scheduled. We wish you the best! Sincerely, Your ___ Care Team ## More Radiology Reports ___ 11:50 AM CT SINUS W/O CONTRAST IMPRESSION: 1. Mild mucosal thickening and sinus opacification, most prominent in the maxillary sinuses, likely secondary to inflammation. No focal fluid collection or osseous erosion identified. 2. Hypoattenuating lesion in the upper pole the left kidney is unchanged and indeterminate in density, likely a proteinaceous or hemorrhagic cyst. Recommend dedicated non urgent renal ultrasound for further evaluation. ___ 11:50 AM CT SINUS W/O CONTRAST IMPRESSION: 1. ## More for ___ ___ 11:50 AM CT SINUS W/O CONTRAST IMPRESSION: 1. Hypoattenu ## More for ___ ___ 11:50 AM CT SINUS W/O CONTRAST IMPRESSION: 1. Mild mucosal thickening and sinus op ## More for ___

Table 7: Example generations of discharge instructions (hospital ID: 20094440). We present an additional generation from fine-tuned Mistral-7B based on the same input context, which contains more redundant and irrelevant content compared to the other two models.