Yale at "Discharge Me!": Evaluating Constrained Generation of Discharge Summaries with Unstructured and Structured Information

Vimig Socrates MS[1,3]* Thomas Huang BS[1,2]*
Xuguang Ai MS[1], Soraya Fereydooni BS[1] Qingyu Chen PhD[1]
R. Andrew Taylor MD MHS[1,2] David Chartash PhD[1,4]
[1]Department of Biomedical Informatics and Data Science, Yale University School of Medicine
[2]Department of Emergency Medicine, Yale School of Medicine
[3]Program of Computational Biology and Bioinformatics, Yale University
[4]School of Medicine, University College Dublin - National University of Ireland, Dublin
{vimig.socrates,t.huang}@yale.edu | * = Equal Contribution

Abstract

In this work, we propose our top-ranking (2nd place) pipeline for the generation of discharge summary subsections as a part of the BioNLP 2024 Shared Task 2: "Discharge Me!". We evaluate both encoder-decoder and state-of-the-art decoder-only language models on the generation of two key sections of the discharge summary. To evaluate the ability of NLP methods to further alleviate the documentation burden on physicians, we also design a novel pipeline to generate the brief hospital course directly from structured information found in the EHR. Finally, we evaluate a constrained beam search approach to inject external knowledge about relevant patient problems into the text generation process. We find that a BioBART model fine-tuned on a larger fraction of the data without constrained beam search outperforms all other models.

1 Introduction

Discharge summaries are vital sources of information that provide a bridge between inpatient treatment and continuation of care in rehabilitation, outpatient, or other intermediate settings. These summaries are often the only form of communication that follows a patient to their next setting of care (Kind and Smith, 2011). This documentation serves many roles, including next action items necessary for the patient, clear identification of incidental findings necessitating follow-up, new treatment regiments, and many other important components of patient treatment plan (Chatterton et al., 2023). The discharge summary is a complex document that addresses not only a wide array of members of the care team including the patient's primary care physician, specialists, ancillary departments, but also the patient themselves. Within the discharge summary, two sections are particularly instrumental in the continuity of care and complex in their content: the Brief Hospital Course (BHC) and the Discharge Instructions.

The BHC summarizes the course of events that occurred from the moment a patient presents to the emergency department (ED) through their hospital course, ending in discharge. This summary is often structured by problem list or procedure and depends heavily on the discharging service (medical vs. surgical etc.) Discharge instructions serve to inform the patient through lay language about key details of their hospital stay, as well as to structure the complex follow-up care that patients will navigate after discharge, enabling them to manage their health effectively in collaboration with their outpatient medical team (Becker et al., 2021; Dubb et al., 2022).

Large Language Models (LLMs), such as ChatGPT, offer a potential solution to the long-standing issue of inaccessible medical communication and the time-demanding nature of synthesis of discharge summaries. Creating high-quality discharge summaries is a challenging and time-consuming task. Significant prior work has demonstrated the broad capabilities of LLMs in clinical natural language processing (Gilson et al., 2023; Nayak et al., 2023; Eppler et al., 2023; Zaretsky et al., 2024). This work suggests that LLMs could be leveraged for the automated generation of discharge summaries. The automatic generation of discharge summaries from inpatient documentation could support alleviating the burden of clinical documentation, particularly under the significant pressures of the inpatient setting (Searle et al., 2023; O'Donnell et al., 2009).

2 Background

2.1 Task Description

The BioNLP 2024 Task 2 Challenge, *Discharge Me!* (Xu et al., 2024), consists of two subtasks: (1) generation of Brief Hospital Courses (BHC) and (2) generation of Discharge Instructions.

Proceedings of the 23rd Workshop on Biomedical Language Processing, pages 724–730

Task	Model	Overall	BLEU	ROUGE-1	ROUGE-2	ROUGE-L	BERTScore	Meteor	AlignScore	MEDCON
Brief Hospital Course	AIMI-Baseline	0.1141	0.0171	0.1184	0.0698	0.1348	0.1726	0.0889	0.1714	0.1398
	GPT 3.5 (0-shot)	0.2035	0.0210	0.3472	0.0983	0.2289	0.2815	0.2232	0.1865	0.2410
	Clinical-T5	0.2068	0.0357	0.3145	0.1378	0.2273	0.3180	0.1678	**0.2251**	0.2285
	BioBART	0.2198	0.0576	0.3161	0.1100	0.2021	**0.3383**	**0.2823**	0.2007	**0.2515**
	BioBART v2	0.2227	**0.0600**	0.3310	0.1239	0.2231	0.3354	0.2802	0.1941	0.2340
	BioBART-Shuffled	**0.2464**	0.0488	**0.3807**	**0.2052**	**0.3003**	0.3278	0.2661	0.1959	0.2463
	GPT-3.5 + Pseudo-SOAP notes*	0.1498	0.0032	0.2603	0.0345	0.1233	0.2374	0.2037	0.2000	0.1360
	BioBART + Constrained Generation	0.1255	0.0045	0.2015	0.0381	0.0900	0.1607	0.2070	0.1739	0.1282
Discharge Instructions	AIMI-Baseline	0.0909	0.0119	0.1343	0.0335	0.0910	0.1026	0.0889	0.1622	0.1025
	GPT-3.5 (0-shot)	0.2289	0.0299	0.3761	0.1312	0.2271	0.3047	0.3109	0.1821	0.2690
	BioBART	**0.3308**	**0.1458**	**0.4465**	**0.1796**	**0.2679**	**0.4382**	**0.3976**	**0.3527**	**0.4183**

Table 1: Results for Black-box GPT models, BART, and T5 pipelines for brief hospital course (subtask 1) and discharge instruction (subtask 2) generation. * indicates that only 121/250 summaries were used, as not all patients had transfer events in structured data.

Formally, we define both problems as sequence-to-sequence text generation tasks. Subtask 1 can be seen as abstractive summarization of the text preceding the BHC. As much of this text is auto-generated by the EHR (e.g. demographics, past medical history, pertinent labs), we can leverage this information without increasing the burden of documentation on physicians. Subtask 2 can also be considered summarization, but requires that the generated text be patient-readable. In this setting, we use the BHC that would have already been generated and attempt to simplify the hospital course, while also providing recommendations for follow-up care. In this work, we evaluate both encoder-decoder models (e.g. BART, T5) and decoder-only models (black-box Azure GPT-3.5). We also propose 2 additional pipelines: (1) a structured data-only BHC generation pipeline that completely removes the need for physician documentation and (2) a constrained beam search approach to improve recall of clinical concepts in BHCs.

2.2 Dataset

The challenge dataset included discharge summaries from 109,168 visits to the Emergency Department (ED) from the **note** and **ED** modules of MIMIC-IV. MIMIC-IV is a publicly available database sourced from the Beth Israel Deaconess Medical Center electronic health record (EHR) that provides a wide array of de-identified patient information containing both structured and unstructured data (Johnson et al., 2023). The text data consisted of a discharge summary, chief complaints, and at least one radiology report. The dataset also included demographics and ED diagnoses as structured data. For our models developed using only structured information, we used data from the MIMIC-IV **hosp** module that included additional demographics (e.g. admission times, treatment wards), hospital diagnoses, procedures, laboratory values, inpatient medications, and lab culture results. These structured data elements were used in a GPT-3.5 pipeline described in Section 3.2.3. The data set was divided into training, validation, and testing (phase I and II) testing sets, of which 250 discharge summaries were selected for standardized final evaluation (Xu et al., 2024).

2.3 SOAP Notes

The subjective, objective, assessment, and plan (SOAP) note is a widely used standard method of documentation used by healthcare providers. The SOAP note is a method of standardizing medical documentation to help physicians streamline clinical decision making (Weed, 1968). The subjective commonly contains the chief complaint, history of present illness (HPI), past medical history, and review of systems. Objective information contains vital signs, physical exam findings, and diagnostic data such as labs, imaging, and other testing. The assessment represents a synthesis of the information collected in prior sections and a presentation of a differential diagnosis. The plan reflects the next steps, frequently including important action items such as consults, additional testing, medications, and other interventions (Tait, 1979).

3 Method

3.1 Data Preprocessing

For the generation of BHCs, we extract all preceding text prior to the brief hospital course. For the generation of discharge instructions, we use the provided ground-truth BHC. For all models that we fine-tuned, we tokenized text based on the encoding scheme used during model training. BART-based models (Lewis et al., 2020) use Byte-pair encoding (Radford et al., 2019) and truncate input and output tokens to the max sequence length of 1024.

Unlike BART-based models, T5 models (Raffel et al., 2020) use Sentencepiece tokenization (Kudo and Richardson, 2018) and relative position embeddings, so while input tokens are truncated to 512 (max context length), output tokens are set to the maximum for our dataset (3903 tokens).

3.2 Subtask 1: Brief Hospital Course

3.2.1 Generation from Unstructured Data

We train three model classes to generate brief hospital courses: BART-based, T5-based, and black box GPT models. We opt for continuously-pretrained biomedical encoder-decoder models as previous work has demonstrated that these models outperform those trained from scratch (Gu et al., 2021). *BioBART-Large* (Yuan et al., 2022) is a 12-layer encoder-decoder model with 406M parameters initialized from a general domain model and continuously pretrained on biomedical paper abstracts from PubMed. The other encoder-decoder model, *Clinical-T5-Large* (Hernandez et al., 2023), with 770M parameters, was instead trained from scratch on MIMIC-III (Johnson et al., 2016) and MIMIC-IV notes. Due to the relative position embeddings in Clinical-T5, it can generate longer summaries, unlike the BioBART model, which is limited to 1024 tokens of output, due to its fixed position embeddings. Given that 10.3% of the challenge dataset has greater than 1024 tokens, we hypothesize that the Clinical-T5 model will achieve better performance.

We also compare fine-tuning with 0-shot performance of an Azure OpenAI GPT-3.5 Turbo model* with human-based abuse monitoring switched off, in keeping with MIMIC's data use agreement. During preliminary evaluations, no significant differences were observed between GPT-3.5 and GPT-4, leading us to choose the more economical option.

3.2.2 Constrained Generation

Upon manual review during phase II testing, we noticed that our encoder-decoder models often failed to provide key formatting or content sections in the BHC. For example, summaries generated by CMED (Cardiac Medical) services tend to contain summaries structured by problem list (e.g. *# UTI:...# Cough...*). Due to the variability in discharge summaries based on individual physician preferences, discharge ward, and patient context, encoder-decoder models seemed to struggle to learn summary structure. Therefore, we attempted to enforce the inclusion of important problems through constrained beam search generation (Anderson et al., 2016; Post and Vilar, 2018; Hu et al., 2019). Constrained beam search injects external knowledge into the generative beam search process by including additional beams for tokens of interest. To identify relevant concepts of interest, we used MedCat (Kraljevic et al., 2021) to tag the history of present illness section preceding the BHC with UMLS (Bodenreider, 2004) concepts. We then constrained our best-performing BioBART model to include these concepts during its beam search. We called this model **BioBART + Constrained Generation**.

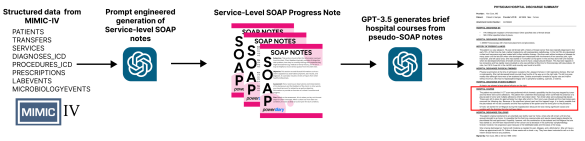

Figure 1: GPT-3.5 Pipeline for generation of brief hospital courses using only structured data

3.2.3 Generation from Structured Data

To evaluate the ability of GPT-based models to further alleviate documentation burden, we develop a pipeline to generate BHCs directly from structured data. As shown in Figure 1, we first extract all relevant structured information for each patient: demographics, ED diagnoses, procedures, inpatient medications, lab values, and lab culture results. As SOAP notes are generally generated either daily or for each service, we generate individual SOAP notes for each transfer during the patient's hospital admission. These SOAP notes are then provided to the GPT-3.5 model in a 0-shot setting to generate brief hospital courses.

3.3 Subtask 2: Discharge Instructions

Similar to subtask 1, we evaluate both fine-tuning and in-context learning (ICL) in the generation of discharge instructions. Namely, we fine-tune BioBART-Large on the brief hospital course text, under the assumption that discharge summaries are generated sequentially and this information would be available to the model in practice. A limited subset of BHCs was provided to GPT-3.5 LLM in a 2-shot approach. In this setting, we noticed that in-context learning did not necessarily improve generation structure so we opted to not evaluate the full test set in the 2-shot setting. Therefore, we

*version: 2023-07-01-preview

provided GPT-3.5 with BHCs and evaluated it in a 0-shot setting. GPT-4 LLM also did not demonstrate performance improvement as measured by ROUGE in a limited subset of 375 notes (GPT-3.5 R-1: 0.306, R-2: 0.083, R-L: 0.159, GPT-4 R-1: 0.309, R-2: 0.079, R-L: 0.157, respectively). As a result, GPT-3.5 was provided the entire test set of BHCs in a 0-shot setting for discharge instruction generation.

4 Experiments and Results

4.1 Quantitative Evaluation

To evaluate the performance of the models, a suite of automated summarization metrics including BLEU, ROUGE-1, ROUGE-2, ROUGE-L, BERTScore, Meteor, AlignScore, and MEDCON were calculated (Papineni et al., 2002; Lin, 2004; Zhang et al., 2019; Banerjee and Lavie, 2005; Zha et al., 2023; Yim et al., 2023). We report summarization metrics for all model variations in Table 1. On BHC generation, we trained two encoder-decoder models: the *Clinical-T5* and *BioBART* models finding that BioBART performed better. Manual review showed that while Clinical-T5 was fine-tuned on larger generations (up to 3903 tokens), its original pre-training truncated generation to 512 tokens, and therefore the model remained biased towards shorter generations. To evaluate the impact of in-domain vocabulary on BHC generation, we also tested *BioBART v2*, continuously pre-trained with a larger, cross-domain vocabulary, as opposed to the standard general domain vocabulary (Yuan et al., 2022). We found that BioBART outperformed BioBART v2, potentially due to the expanded vocabulary coming from biomedical literature rather than the clinical notes found in this challenge. Finally, we also tested the impact of increased training data size (*BioBART-Shuffled*) by shuffling the phase I training, validation and testing data set, before recombining for an additional 14,690 discharge summaries in the training set. Across the encoder-decoder models, *BioBART-Shuffled* performed best, yielding us 2nd place on the challenge leaderboard.

We also compared these results to a *GPT-3.5 0-shot* model, finding that black-box GPT-3.5 performed worse than the best performing fine-tuned model. When repeating this experiment with our structured data-only method (*GPT-3.5 + Pseudo-SOAP notes*) as well as constrained generation, we found that neither of these methods offered improvement over our best-performing model, BioBART. In the generation of discharge instructions, the BioBART outperformed GPT-3.5, and so was included as our challenge submission.

Evaluation Criteria	Brief Hospital Course	Discharge Instructions
Completeness	3.52	4.27
Correctness	2.57	3.95
Readability	2.11	-
Overall	1.53	2.36

Table 2: Average ratings across 3 criteria for 3 clinicians (Discharge instruction readability was not assessed as the target audience are patients)

4.2 Qualitative Evaluation

To evaluate the clinical validity of generated brief hospital courses and discharge instructions, a team of 3 clinicians reviewed a random sample of 25 generations from the hidden set of 250 discharge summaries. Each clinician rated both the brief hospital courses and discharge instructions according to 3 criteria: Completeness (captures important information), Correctness (contains less false information), and Readability. They also provide a holistic assessment as an overall score (Xu et al., 2024). All metrics are averaged, and results are presented in Table 2, showing that both the brief hospital course and discharge instructions received their highest grades in completeness and lowest in the overall evaluation criteria. Selected discharge instructions received higher grades across completeness, correctness, and overall criteria compared to brief hospital course. This is likely due to the nature of the increased complexity and wider range of information often necessary for inclusion in a brief hospital course.

5 Limitations

While our methods were able to produce reasonable BHCs and discharge instructions, there are several important limitations to our study. Computationally, we were unable to perform a rigorous hyperparameter search across all our experimental conditions due to computational constraints. There is potential for improvement given further resources. Namely, we were surprised that *Constrained Generation* performed significantly worse than vanilla BioBART models. This is potentially due to the additional hyperparameters that need to be tuned, including the tokens of interest that MedCat identified and beam sizes.

Furthermore, we believe that there is limited clin-

ical validity in the current task as it has been framed. The automated generation of BHC and discharge instructions utilizing physician generated preceding text does not truly automate the task, nor does it obviate the need for the core summarization task of note-writing on the part of the physician (rather than documentation of findings). We attempted to model a more representative use case by including the generation of Pseudo-SOAP notes but found significantly worse performance, demonstrating this difficulty of the real-world clinical task. Furthermore, the format and physician- and institution-specific stylistic choices had a significant impact on automated performance, as demonstrated by the significant variation in documentation length and lack of standard templates even within services that discharged patients. The Challenge organizers did attempt to alleviate concerns around generalizability with a qualitative analysis by clinicians, but further efforts in automated metrics involving semantic comparison are necessary.

6 Conclusion

In this work, we present experiments for the automated generation of brief hospital courses and discharge instructions from structured and unstructured data captured during an ED encounter. We find that a BioBART model with increased training data performed better than both other encoder-decoder models and a black-box decoder-only model. We also find that constraining generation to emphasize generation of UMLS concepts worsens performance. Finally, we show that GPT-3.5 can generate brief hospital courses purely from structured information, further reducing the annotation burden for physicians.

Acknowledgments

We would like to thank Justin Xu and the other members of the "Discharge Me!" team for organizing this competition, as well as all their feedback and prompt responses to questions along the way.

References

Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. 2016. Guided open vocabulary image captioning with constrained beam search. *arXiv preprint arXiv:1612.00576*.

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.

Christoph Becker, Samuel Zumbrunn, Katharina Beck, Alessia Vincent, Nina Loretz, Jonas Müller, Simon A Amacher, Rainer Schaefert, and Sabina Hunziker. 2021. Interventions to improve communication at hospital discharge and rates of readmission: a systematic review and meta-analysis. *JAMA Network Open*, 4(8):e2119346–e2119346.

Olivier Bodenreider. 2004. The unified medical language system (umls): integrating biomedical terminology. *Nucleic acids research*, 32(suppl_1):D267–D270.

Brittany Chatterton, Jennifer Chen, Eleanor Bimla Schwarz, and Jennifer Karlin. 2023. Primary care physicians' perspectives on high-quality discharge summaries. *Journal of General Internal Medicine*, pages 1–6.

Simran Dubb, Gurmeet Kaur, Sweta Kumari, Krishna Murti, and Biplab Pal. 2022. Comprehension and compliance with discharge instructions among pediatric caregivers. *Clinical Epidemiology and Global Health*, 17:101137.

Michael B Eppler, Conner Ganjavi, J Everett Knudsen, Ryan J Davis, Oluwatobiloba Ayo-Ajibola, Aditya Desai, Lorenzo Storino Ramacciotti, Andrew Chen, Andre De Castro Abreu, Mihir M Desai, et al. 2023. Bridging the gap between urological research and patient understanding: the role of large language models in automated generation of layperson's summaries. *Urology practice*, 10(5):436–443.

Aidan Gilson, Conrad W Safranek, Thomas Huang, Vimig Socrates, Ling Chi, Richard Andrew Taylor, David Chartash, et al. 2023. How does chatgpt perform on the united states medical licensing examination (usmle)? the implications of large language models for medical education and knowledge assessment. *JMIR Medical Education*, 9(1):e45312.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2021. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare (HEALTH)*, 3(1):1–23.

Evan Hernandez, Diwakar Mahajan, Jonas Wulff, Micah J Smith, Zachary Ziegler, Daniel Nadler, Peter Szolovits, Alistair Johnson, Emily Alsentzer, et al. 2023. Do we still need clinical language models? In *Conference on Health, Inference, and Learning*, pages 578–597. PMLR.

J. Edward Hu, Huda Khayrallah, Ryan Culkin, Patrick Xia, Tongfei Chen, Matt Post, and Benjamin Van Durme. 2019. Improved lexically constrained decoding for translation and monolingual rewriting. In *Proceedings of the 2019 Conference of the North*

American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pages 839–850, Minneapolis, Minnesota. Association for Computational Linguistics.

Alistair EW Johnson, Lucas Bulgarelli, Lu Shen, Alvin Gayles, Ayad Shammout, Steven Horng, Tom J Pollard, Sicheng Hao, Benjamin Moody, Brian Gow, et al. 2023. Mimic-iv, a freely accessible electronic health record dataset. *Scientific data*, 10(1):1.

Alistair EW 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(1):1–9.

Amy JH Kind and Maureen A Smith. 2011. Documentation of mandated discharge summary components in transitions from acute to subacute care.

Zeljko Kraljevic, Thomas Searle, Anthony Shek, Lukasz Roguski, Kawsar Noor, Daniel Bean, Aurelie Mascio, Leilei Zhu, Amos A Folarin, Angus Roberts, Rebecca Bendayan, Mark P Richardson, Robert Stewart, Anoop D Shah, Wai Keong Wong, Zina Ibrahim, James T Teo, and Richard J B Dobson. 2021. Multi-domain clinical natural language processing with MedCAT: The medical concept annotation toolkit. *Artif. Intell. Med.*, 117:102083.

Taku Kudo and John Richardson. 2018. Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. *arXiv preprint arXiv:1808.06226*.

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

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Ashwin Nayak, Matthew S Alkaitis, Kristen Nayak, Margaret Nikolov, Kevin P Weinfurt, and Kevin Schulman. 2023. Comparison of history of present illness summaries generated by a chatbot and senior internal medicine residents. *JAMA Internal Medicine*, 183(9):1026–1027.

Heather C O'Donnell, Rainu Kaushal, Yolanda Barrón, Mark A Callahan, Ronald D Adelman, and Eugenia L Siegler. 2009. Physicians' attitudes towards copy and pasting in electronic note writing. *Journal of general internal medicine*, 24:63–68.

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.

Matt Post and David Vilar. 2018. Fast lexically constrained decoding with dynamic beam allocation for neural machine translation. *arXiv preprint arXiv:1804.06609*.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67.

Thomas Searle, Zina Ibrahim, James Teo, and Richard JB Dobson. 2023. Discharge summary hospital course summarisation of in patient electronic health record text with clinical concept guided deep pre-trained transformer models. *Journal of Biomedical Informatics*, 141:104358.

Ian Tait. 1979. *The History and Function of Clinical Records*. Md thesis, University of Cambridge.

Lawrence L Weed. 1968. Medical records that guide and teach. *New England Journal of Medicine*, 278(12):593–600.

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.

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.

Hongyi Yuan, Zheng Yuan, Ruyi Gan, Jiaxing Zhang, Yutao Xie, and Sheng Yu. 2022. BioBART: Pretraining and evaluation of a biomedical generative language model. In *Proceedings of the 21st Workshop on Biomedical Language Processing*, pages 97–109, Dublin, Ireland. Association for Computational Linguistics.

Jonah Zaretsky, Jeong Min Kim, Samuel Baskharoun, Yunan Zhao, Jonathan Austrian, Yindalon Aphinyanaphongs, Ravi Gupta, Saul B Blecker, and Jonah Feldman. 2024. Generative artificial intelligence to transform inpatient discharge summaries to

patient-friendly language and format. *JAMA network open*, 7(3):e240357–e240357.

Yuheng Zha, Yichi Yang, Ruichen Li, and Zhiting Hu. 2023. Alignscore: Evaluating factual consistency with a unified alignment function. *arXiv preprint arXiv:2305.16739*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.