Team Cadence at MEDIQA-Chat 2023: Generating, augmenting and summarizing clinical dialogue with large language models

Ashwyn Sharma
Cadence Solutions, USA
ashwyn@cadencerpm.com

David I. Feldman, MD, MPH
Cadence Solutions, USA
Massachusetts General Hospital, Harvard University , USA
david.feldman@cadencerpm.com

Aneesh Jain
Cadence Solutions, USA
Virginia Polytechnic Institute and State University, USA
aneeshjain70@gmail.com

Abstract

This paper describes Team Cadence's winning submission to Task C of the MEDIQA-Chat 2023 shared tasks. We also present the set of methods, including a novel *N-pass* strategy to summarize a mix of clinical dialogue and an incomplete summarized note, used to complete Task A and Task B, ranking highly on the leaderboard amongst stable and reproducible code submissions. The shared tasks invited participants to summarize, classify and generate patient-doctor conversations. Considering the small volume of training data available, we took a *data-augmentation-first* approach to the three tasks by focusing on the dialogue generation task, i.e., Task C. It proved effective in improving our models' performance on Task A and Task B. We also found the BART architecture to be highly versatile, as it formed the base for all our submissions. Finally, based on the results shared by the organizers, we note that Team Cadence was the only team to submit stable and reproducible runs to all three tasks.

1 Introduction

MEDIQA-Chat 2023 Shared Tasks included three tasks on the summarization and generation of doctor-patient conversations to promote research on these topics (Ben Abacha et al., 2023). Task A (*Short Dialogue2Note Summarization*) expected a section summary (section header and content) given a short input conversation. We recognized that generating the summary content was an abstractive summarization (Chopra et al., 2016) task while predicting the section header was a multi-class (twenty normalized section labels) classification task. Task B (*Full Dialogue2Note Summarization*) was another abstractive summarization task that required submissions to generate a complete clinical note from a whole dialogue between a patient and a doctor. The complete clinical note was expected to have the following first-level section headers: *"HISTORY OF PRESENT ILLNESS", "PHYSICAL EXAM", "RESULTS", and "ASSESSMENT AND PLAN"*. Finally, Task C (*Note2Dialogue Generation*), a data augmentation (Shorten et al., 2021) task, was about generating patient-doctor conversations for complete input notes.

Aside from predicting section headers for Task A, all other tasks could be formulated as sequence-to-sequence (Sutskever et al., 2014) learning tasks. Various model architectures based on transformers (Vaswani et al., 2017) have proved to be successful at tackling such tasks. Therefore, leveraging pre-trained model checkpoints from public repositories was considered the right choice. Encouraged by the leaderboard for SAMSum (Gliwa et al., 2019) on HuggingFace (Wolf et al., 2020), a dialogue summarization dataset, we chose BART (Lewis et al., 2019) as the base model for our experiments. Specifically, we picked the *facebook/bart-*

Proceedings of the 5th Clinical Natural Language Processing Workshop, pages 228–235
July 14, 2023 ©2023 Association for Computational Linguistics

large [1] model checkpoint (referenced as *bart-large* in this text from hereon) for its effectiveness on text-generation tasks.

The SAMSum (Gliwa et al., 2019) dataset is intended to train dialogue summarization models. However, we recognized that the input and target labels could be inverted to train a dialogue generation model. We trained/validated *bart-large* on the inverse of SAMSum (Gliwa et al., 2019) dataset followed by the Task C training dataset provided by the task organizers, achieving ROUGE-1 and ROUGE-2 scores of 59.11 and 23.69, respectively, on the validation set. This model was then used to augment datasets for Task A and Task B summarization tasks. In order to generate synthetic patient-doctor conversations, we chose to sample a thousand discharge summary notes from the MIMIC-IV-Note (Johnson et al., 2023; Goldberger et al., 2000) dataset. We then added these dialogue-note pairs to the Task A and Task B training datasets provided by the organizers. The impact of this augmentation technique is noted in Section 5 below.

For Task A summarization, *bart-large* was fine-tuned on the SAMSum (Gliwa et al., 2019) dataset followed by fine-tuning on the augmented dataset for Task A, which achieved ROUGE-1 and ROUGE-2 scores of 50.7 and 21.4, respectively, on the validation set. Our methods yielded an overall improvement (over the baseline) of 13.1% and 14% in ROUGE-1 and ROUGE-2 scores, respectively. Results from fine-tuning *bart-large* on the unaugmented (original) Task A dataset were considered the baseline in this comparison.

Inspired by the significant gains exhibited by the Task A model, we decided to use it as the base model for Task B. Fine-tuning this base model on the augmented Task B dataset yielded ROUGE-1 and ROUGE-2 scores of 54.16 and 26.04, respectively - a 13.7% gain in ROUGE-2 score over the baseline. Results from fine-tuning the base model on the unaugmented (original) Task B dataset were considered the baseline in this comparison. The final submission(run1) achieved ROUGE-1 and ROUGE-2 scores of 49.5 and 23.4 on the test set. Unfortunately, the Task B dataset comprised input conversations almost twice as long as the maximum number of tokens accepted by *bart-large*, which naturally prohibits the model's ability to summarize the entire conversation. To solve this problem, we developed an *N-pass* strategy in which the model attempts to summarize the conversation in multiple steps. Each step (or pass) involves the model taking as input the summary note of the dialogue processed till that step, concatenated with the rest of the dialogue. In other words, we trained the model to summarize a partial mix of an incomplete clinical note and an incomplete patient-doctor conversation. This strategy led to a gain of 6.6% and 8.1% in ROUGE-1 (57.76) and ROUGE-2 (28.15) scores, respectively, on the validation set. We submitted the *N-pass* model as run2, which outperformed the run1 submission by 6.8%, both for ROUGE-1 (52.9) and ROUGE-2 (25) scores, on the test set. It also improved the division-based aggregate score by 16.75%. Overall, our methods improved the baseline ROUGE-2 score by 22.9% on the validation set, while the baseline ROUGE-1 score was found to be slightly better by 0.45%.

Given the promising performance of *bart-large* on the summarization tasks, we also decided to use it for Task A classification. We leveraged the *BartForSequenceClassification* wrapper offered by HuggingFace (Wolf et al., 2020), a BART model with a sequence classification head on top (a linear layer on top of the pooled output). Using this approach, we achieved an accuracy of 78% and an F1 score of 78.37%. The final submission was reported to have an accuracy of 73.5% on the test set.

2 Background and Related Work

Studies like the ones from Alkureishi MA et al. (Alkureishi et al., 2016) and Rathert et al. (Rathert et al., 2017) have presented evidence on EHRs (Electronic Health Records) impacting the quality of patient-doctor conversations. Digital scribes (van Buchem et al., 2021) and summarization tools (Shanafelt et al., 2016) can mitigate some of these problems. However, many challenges are associated with clinical dialogue summarization (Zhu and Penn, 2006). Some significant challenges include omitting key medical concepts (Knoll et al., 2022) and hallucinating unsubstantiated information.

Several attempts have been made to address said inherent challenges and automatically generate high-quality summaries of clinical encounters. Approaches like the ones used by Enarvi et al. (2020) have utilized a transformer (Vaswani et al., 2017) model to summarize doctor-patient conversations. Joshi et al. (2020) and Michalopoulos et al.

[1] https://huggingface.co/facebook/bart-large

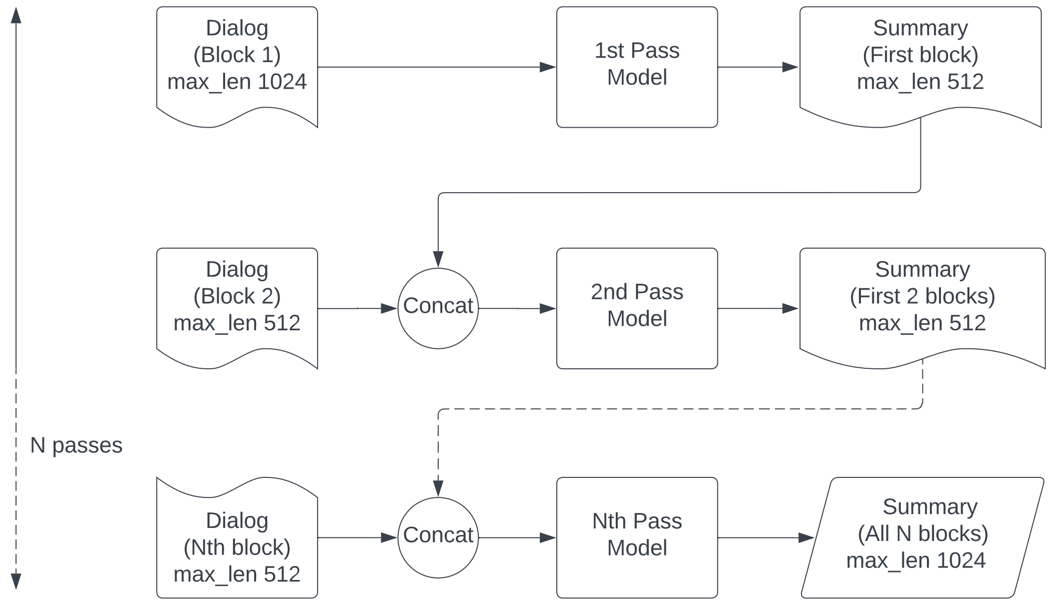

Figure 1: N-pass summarization for handling long conversations.

(2022) have also incorporated medical knowledge into these models. On the data generation front, Chintagunta et al. (2021) showed that large language models can be used for augmenting medical summarization datasets.

To the best of our knowledge, the *N-pass* strategy used to address long input sequences of Task B is novel. However, multiple multi-stage summarization approaches have been proposed so far. For example, Krishna et al. (2020) used modular summarization techniques to produce notes from patient-doctor conversations. Zhang et al. (2021) used multi-stage summarization for long inputs, whereas Gidiotis and Tsoumakas (2020) split a long document and its summary into multiple source-target pairs using sentence similarity. Recursive summarization incorporating human feedback (Wu et al., 2021) even achieved state-of-the-art results in book summarization.

3 Datasets

3.1 MEDIQA-Chat-2023

Task A training (validation) dataset (Ben Abacha et al., 2023) provided by the organizers consists of 1,201 (100) pairs of conversations and associated section headers and summaries. There were 20 unique normalized section headers overall. The Task B and Task C training (validation) set consists of 67 (20) pairs of conversations and full clinical notes (Yim et al., 2023).

3.2 SAMSum

The SAMSum dataset contains 16369 conversations and their summaries (Gliwa et al., 2019), with a train/val/test split of 14732/818/819. Several dialogue summarization models have leveraged this dataset (Ni et al., 2022) and achieved promising results on the task. We note the impact of this dataset in the ablation study (Section 5).

3.3 MIMIC-IV-Note

MIMIC-IV-Note contains 331,794 deidentified free-text clinical notes for patients included in the MIMIC-IV clinical database (Johnson et al., 2023; Goldberger et al., 2000). We sampled a thousand notes from this dataset and used the Task C (dialogue generation) model for downstream data augmentation of Task A and Task B. Ablation study (Section 5) highlights significant contributions of this dataset to improving the results.

4 Methods

4.1 Dialogue Generation

We discovered that by flipping input and target labels, the SAMSum (Gliwa et al., 2019) dataset could also train a dialogue generation model. Our

Table 1: Hyperparameters used for Task A, Task B and Task C

Parameter	Task A		Task B	Task C
	Classification	Summarization	Summarization	Generation
learning_rate	2E-05	5E-05	5E-05	5E-05
per_device_train_batch_size	8	4	4	4
per_device_eval_batch_size	8	4	2	2
weight_decay	0.01	0	0	0
num_train_epochs	30	30	30	10
fp16	TRUE	TRUE	TRUE	TRUE
gradient_accumulation_steps	4	8	8	8
gradient_checkpointing	TRUE	TRUE	TRUE	TRUE
predict_with_generate	-	TRUE	TRUE	TRUE
generation_max_length	-	512	1024	1024
max_target_length	-	512	1024	1024
max_source_length	1024	1024	1024	1024

recipe included fine-tuning *bart-large* on the inverted SAMSum (Gliwa et al., 2019) dataset for 10 epochs, followed by fine-tuning on a dataset that combined training and validation datasets from Task A and Task C for another 10 epochs. Fine-tuning was performed using the Trainer API offered by HuggingFace (Wolf et al., 2020), and the hyperparameters used are described in (Table 1). We did not perform a comprehensive sweep and recognize that a more optimal set of hyperparameters could yield better results. The model yielded by this recipe was also used for generating synthetic data for Task A and Task B summarization. Specifically, patient-doctor conversations were generated for 1000 discharge summary notes sampled from the MIMIC-IV-Note (Johnson et al., 2023; Goldberger et al., 2000) dataset. We used ROUGE-1 and ROUGE-2 scores for evaluating the model's performance on the validation set (Lin, 2004).

4.2 Dialogue Summarization

Summarization models for Task A and Task B leveraged *bart-large* fine-tuned on the SAMSum (Gliwa et al., 2019) dataset for 10 epochs as the base model. The base model was then fine-tuned on the augmented version of the Task A training dataset for 30 epochs. Like dialogue generation, fine-tuning was performed using the Trainer API offered by HuggingFace (Wolf et al., 2020), and the hyperparameters used are described in Table 1. We did not perform a comprehensive sweep and recognize that a more optimal set of hyperparameters could yield better results. With a working hypothesis that the Task A model can capture local themes in conversations with fewer turns, we used the model yielded by the above recipe as the base model for Task B.

Before augmenting the Task B dataset with the dialogue generation model, we sanitized the 1000 notes sampled from the MIMIC-IV-Note (Johnson et al., 2023; Goldberger et al., 2000) dataset. The sanitization process mainly included removing first-level section headers not accepted for evaluation by the organizers, as laid out in (Section 1). The base model was then fine-tuned on the *sanitized-and-augmented* dataset (named Augmented(Sections) in result tables) using the same process as Task A. This fine-tuned version was submitted as run1 and suffered from a significant drawback - the inability to handle input sequences longer than 1024 tokens. To address the shortcoming, we developed a novel *N-pass* approach by training a model that can generate summaries given a partial mix of incomplete summaries and incomplete dialogue. Specifically, a 2-pass version, named run2, was submitted to the shared task.

The *N-pass* approach is illustrated in Figure 1. The idea is to summarize long conversations in multiple passes, where each pass accepts as input the next *block* of the unsummarized dialogue concatenated with the summary output by the previous pass. The intuition behind this approach is to accommodate the limit on the number of input tokens accepted by the model by feeding it the dialogue in *blocks* but still propagating the context by incorporating the summary generated till that point. For run2, the model used for run1 was fine-tuned for

30 epochs on a dataset that concatenated the first *block* summary with the second *block* of the dialogue. The first *block* summaries were generated by the run1 model. A *block size* of 512 tokens was used for both the input and the output (except the final *pass* where output is 1024 tokens). We used a combination of ROUGE-1 and ROUGE-2 scores for evaluating the model's ability to summarize the conversations in the validation set (Lin, 2004).

4.3 Classification

We used a simple yet effective classification approach to producing section headers for Task A. Given the promising results from using *bart-large* on the summarization and dialogue generation tasks, we chose to stick with the same for classification. To be exact, we fine-tuned the model used for Task A submission on the classification task by leveraging the *BartForSequenceClassification* wrapper offered by HuggingFace (Wolf et al., 2020), a BART model with a sequence classification head on top (a linear layer on top of the pooled output). Again, the Trainer API was used with no hyperparameter sweep. Table 1 lists the hyperparameters used for fine-tuning the classifier.

5 Experiments and Ablation Study

Dataset	**ROUGE-1**	**ROUGE-2**
MEDIQA	47.5	19.8
Augmented (Sections)	48.12	19.9
Augmented	**50.7**	**21.4**

Table 2: Task A - results with different training datasets. Metrics evaluated on the task validation set.

Model	**ROUGE-1**	**ROUGE-2**
bart-large	44.8	18.77
bart-samsum	**47.5**	**19.8**

Table 3: Task A - impact of fine-tuning on SAMSum. Metrics evaluated on the task validation set.

5.1 Task A

In Table 2, we compare the results obtained on the Task A validation set by using three different training datasets - original Task A training data, augmented Task A training data, and *sanitized-and-augmented* (defined in Section 4.2) training data. The augmented version outperforms the original Task A training data by 6.7% (ROUGE-1) and 8% (ROUGE-2). As expected, the *sanitized-and-augmented* training data yields smaller gains because the summary notes for Task A are shorter and do not include first-level section headers in Task B training data.

An ablation study (Table 3) was also conducted on the impact of fine-tuning *bart-large* on the SAMSum(Gliwa et al., 2019) dataset. It was found that fine-tuning on the SAMSum (Gliwa et al., 2019) dataset improved performance on the validation set by 6% (ROUGE-1) and 5.4% (ROUGE-2).

Task A summarization model fine-tuned on classification achieved an accuracy of 78% and an f1 score of 78.37% on the validation set.

Version	**ROUGE-1**	**ROUGE-2**
MEDIQA	48.13	19.0
Augmented	51.86	23.42
Augmented (Sections)	54.16	26.04
2-pass	**57.76**	**28.15**

Table 4: Task B - results with different training datasets and the 2-pass strategy. Metrics evaluated on the task validation set.

Model	**ROUGE-1**	**ROUGE-2**
bart-large	**58.02**	**22.9**
bart-samsum	48.13	19.0

Table 5: Task B - impact of fine-tuning on SAMSum. Metrics evaluated on the task validation set.

5.2 Task B

Table 4 shows that the *2-pass* summarization strategy leads to a gain of 6.6% (ROUGE-1) and 8.1% (ROUGE-2). Furthermore, training on the *sanitized-and-augmented* dataset yields improvements of 12.5% (ROUGE-1) and 37% (ROUGE-2), driving home the value of data augmentation by clinical dialogue generation. Interestingly, simply fine-tuning on the SAMSum(Gliwa et al., 2019) dataset led to worse results (Table 5) on the Task B validation set, which could be explained by the discrepancy in the length of the conversations and the summaries between the two datasets.

Dataset	bart-large		bart-samsum	
	R-1	R-2	R-1	R-2
MEDIQA	53.6	17.26	56.55	20.64
Combined	58.43	22.74	**59.11**	**23.69**

Table 6: Task C - results with different training datasets and impact of fine-tuning on SAMSum. Metrics evaluated on the task validation set.

5.3 Task C

The ablation study (Table 6) for Task C highlights two significant ideas. First, adding the training data from Task A contributed a hike of 4.5% (9%) in the ROUGE-1 score and 14.7% (31.7%) in the ROUGE-2 score for the model (not) fine-tuned on the inverse of the SAMSum (Gliwa et al., 2019) dataset. Second, fine-tuning on the inverse of SAMSum (Gliwa et al., 2019) led to a gain of 5.5% (1.1%) in ROUGE-1 scores and 19.5% (4.1%) in ROUGE-2 scores when training data from Task C (Task A + Task C) was used. It shows that the additional data from Task A is more critical when fine-tuning on the inverse of SAMSum (Gliwa et al., 2019) is skipped.

6 Results

Team Cadence's submission for Task C earned ***rank-1*** amongst all participants, beating the next-best submission by 28.3% (ROUGE-1) and 99% (ROUGE-2).

The organizers shared test set results (Ben Abacha et al., 2023) along with a *code status* description where a *code status* of 1 meant that the organizers were able to run the submitted code and reproduce the results, and a *code status* of 2 meant that they were able to run the code and found minor differences with no changes in rankings. *Code statuses* 3,4, and 5 meant that the organizers found the submitted code to be unstable or not runnable under their configurations. Amongst *code statuses* 1 and 2, Team Cadence achieved the following ranks: *rank-2* on TaskB-summarization, *rank-3* on TaskA-summarization, *rank-3* on TaskB-summarization(note-divisions), and *rank-5* on TaskA-classification. The code for generating the submitted runs is being shared publicly[2].

7 System Specification

In the spirit of reproducibility, we share details of the systems used to run these experiments. The models were fine-tuned on *g4dn.12xlarge* AWS Sagemaker notebook instances [3]. HuggingFace's Python package transformers (Wolf et al., 2020) version 4.27.1 was used in a Python3.8 environment. Reported results were aggregated from 4 different runs using 4 different random seeds.

8 Limitations and Future Work

The methods described in this paper do not leverage any external medical knowledge, a technique that has been shown to be effective by other studies (Joshi et al., 2020; Michalopoulos et al., 2022). And like other methods based on large language models, in theory, our models are also prone to hallucinations and omission of key-clinical concepts. We plan to explore constrained beam search[4] as a mitigation strategy for addressing these challenges in the future.

Although the impact of the Task C model as a data augmentation tool is undoubtedly positive (Section 5), qualitative error analysis of patient-doctor conversations produced by the model showed that the output contained a small number of dialogue turns, and each individual turn was too long, packed with information. Producing conversations with a more natural flow should yield an even better boost on downstream tasks, and we leave exploring such methods to future experimentation. We also recognize that *N-pass* summarization for Task B with higher values of *N* should be able to cover the entirety of the input conversations in the Task B datasets, albeit with diminishing returns as *N* increases. We hope to evaluate them in future iterations of similar shared tasks.

9 Conclusion

The two key takeaways from the experiments and results in this paper are significant improvements in summarization results driven by data augmentation and the *N-pass* summarization technique for handling long input patient-doctor conversations. Furthermore, the fact that our submissions to all three tasks share the same base (*bart-large*) model

[2] https://github.com/ashwyn/MEDIQA-Chat-2023-Cadence

[3] https://docs.aws.amazon.com/sagemaker/latest/dg/notebooks.html

[4] https://huggingface.co/blog/constrained-beam-search

speaks volumes of its versatility. Finally, the results demonstrate the effectiveness of fine-tuning on custom datasets for specialized domains like medicine.

References

Maria Alcocer Alkureishi, Wei Wei Lee, Maureen Lyons, Valerie G Press, Sara Imam, Akua Nkansah-Amankra, Deb Werner, and Vineet M Arora. 2016. Impact of electronic medical record use on the patient–doctor relationship and communication: a systematic review. *Journal of general internal medicine*, 31:548–560.

Asma Ben Abacha, Wen-wai Yim, Griffin Adams, Neal Snider, and Meliha Yetisgen. 2023. Overview of the mediqa-chat 2023 shared tasks on the summarization and generation of doctor-patient conversations. In *ACL-ClinicalNLP 2023*.

Asma Ben Abacha, Wen-wai Yim, Yadan Fan, and Thomas Lin. 2023. An empirical study of clinical note generation from doctor-patient encounters. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2291–2302, Dubrovnik, Croatia. Association for Computational Linguistics.

Bharath Chintagunta, Namit Katariya, Xavier Amatriain, and Anitha Kannan. 2021. Medically aware gpt-3 as a data generator for medical dialogue summarization. In *Machine Learning for Healthcare Conference*, pages 354–372. PMLR.

Sumit Chopra, Michael Auli, and Alexander M Rush. 2016. Abstractive sentence summarization with attentive recurrent neural networks. In *Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: human language technologies*, pages 93–98.

Seppo Enarvi, Marilisa Amoia, Miguel Del-Agua Teba, Brian Delaney, Frank Diehl, Stefan Hahn, Kristina Harris, Liam McGrath, Yue Pan, Joel Pinto, et al. 2020. Generating medical reports from patient-doctor conversations using sequence-to-sequence models. In *Proceedings of the first workshop on natural language processing for medical conversations*, pages 22–30.

Alexios Gidiotis and Grigorios Tsoumakas. 2020. A divide-and-conquer approach to the summarization of long documents. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28:3029–3040.

Bogdan Gliwa, Iwona Mochol, Maciej Biesek, and Aleksander Wawer. 2019. SAMSum corpus: A human-annotated dialogue dataset for abstractive summarization. In *Proceedings of the 2nd Workshop on New Frontiers in Summarization*, pages 70–79, Hong Kong, China. Association for Computational Linguistics.

Ary L Goldberger, Luis AN Amaral, Leon Glass, Jeffrey M Hausdorff, Plamen Ch Ivanov, Roger G Mark, Joseph E Mietus, George B Moody, Chung-Kang Peng, and H Eugene Stanley. 2000. Physiobank, physiotoolkit, and physionet: components of a new research resource for complex physiologic signals. *circulation*, 101(23):e215–e220.

Alistair Johnson, Tom Pollard, Steven Horng, Leo Anthony Celi, and Roger Mark. 2023. Mimic-iv-note: Deidentified free-text clinical notes.

Anirudh Joshi, Namit Katariya, Xavier Amatriain, and Anitha Kannan. 2020. Dr. summarize: Global summarization of medical dialogue by exploiting local structures. *arXiv preprint arXiv:2009.08666*.

Tom Knoll, Francesco Moramarco, Alex Papadopoulos Korfiatis, Rachel Young, Claudia Ruffini, Mark Perera, Christian Perstl, Ehud Reiter, Anya Belz, and Aleksandar Savkov. 2022. User-driven research of medical note generation software. *arXiv preprint arXiv:2205.02549*.

Kundan Krishna, Sopan Khosla, Jeffrey P Bigham, and Zachary C Lipton. 2020. Generating soap notes from doctor-patient conversations using modular summarization techniques. *arXiv preprint arXiv:2005.01795*.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2019. BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *CoRR*, abs/1910.13461.

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.

George Michalopoulos, Kyle Williams, Gagandeep Singh, and Thomas Lin. 2022. Medicalsum: A guided clinical abstractive summarization model for generating medical reports from patient-doctor conversations. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 4741–4749.

Jinjie Ni, Tom Young, Vlad Pandelea, Fuzhao Xue, and Erik Cambria. 2022. Recent advances in deep learning based dialogue systems: A systematic survey. *Artificial intelligence review*, pages 1–101.

Cheryl Rathert, Jessica N Mittler, Sudeep Banerjee, and Jennifer McDaniel. 2017. Patient-centered communication in the era of electronic health records: What does the evidence say? *Patient education and counseling*, 100(1):50–64.

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. In

Mayo Clinic Proceedings, volume 91, pages 836–848. Elsevier.

Connor Shorten, Taghi M Khoshgoftaar, and Borko Furht. 2021. Text data augmentation for deep learning. *Journal of big Data*, 8:1–34.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. *Advances in neural information processing systems*, 27.

Marieke M van Buchem, Hileen Boosman, Martijn P Bauer, Ilse MJ Kant, Simone A Cammel, and Ewout W Steyerberg. 2021. The digital scribe in clinical practice: a scoping review and research agenda. *NPJ digital medicine*, 4(1):57.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations*, pages 38–45.

Jeff Wu, Long Ouyang, Daniel M Ziegler, Nisan Stiennon, Ryan Lowe, Jan Leike, and Paul Christiano. 2021. Recursively summarizing books with human feedback. *arXiv preprint arXiv:2109.10862*.

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. *Submitted to Nature Scientific Data*.

Yusen Zhang, Ansong Ni, Ziming Mao, Chen Henry Wu, Chenguang Zhu, Budhaditya Deb, Ahmed H Awadallah, Dragomir Radev, and Rui Zhang. 2021. Summˆ n: A multi-stage summarization framework for long input dialogues and documents. *arXiv preprint arXiv:2110.10150*.

Xiaodan Zhu and Gerald Penn. 2006. Summarization of spontaneous conversations. In *Ninth International Conference on Spoken Language Processing*.