EasyChair Preprint

№ 10534

NewAgeHealthWarriors at MEDIQA-Chat 2023 Task A: Summarizing Short Medical Conversation with Transformers

Prakhar Mishra and Ravi Theja Desetty

July 10, 2023

NewAgeHealthWarriors at MEDIQA-Chat 2023 Task A: Summarizing Short Medical Conversation with Transformers

Prakhar Mishra
IIIT Bangalore, India
prakhar.mishra@iiitb.org

Ravi Theja Desetty
IIIT Bangalore, India
ravi.theja@iiitb.org

Abstract

This paper presents the MEDIQA-Chat 2023 shared task organized at the ACL-Clinical NLP workshop. The shared task is motivated by the need to develop methods to automatically generate clinical notes from doctor-patient conversations. In this paper, we present our submission for *MEDIQA-Chat 2023 Task A: Short Dialogue2Note Summarization*. Manual creation of these clinical notes requires extensive human efforts, thus making it a time-consuming and expensive process. To address this, we propose an ensemble-based method over GPT-3, BART, BERT variants, and Rule-based systems to automatically generate clinical notes from these conversations. The proposed system achieves a score of 0.730 and 0.544 for both the sub-tasks on the test set (ranking 8th on the leaderboard for both tasks) and shows better performance compared to a baseline system using BART variants.

1 Introduction

Telecare has experienced an exponential increase in utilization since the onset of the COVID-19 pandemic, leading to the emergence of a vast network of healthcare providers and patients Garfan et al. (2021). We consider this to be a significant use case within the Telecare domain, medical personnel often need to provide a concise summary of the conversation they had with their patient in order to ensure that a colleague is able to follow up on the next consultation. Both patients and medical professionals can use these summaries to refer back to their interactions in the future. Unfortunately, manually creating conversation notes after each encounter consumes a significant amount of time, and energy, and also poses challenges when done at scale.

Recent advancements in Natural Language Processing (NLP) with large language models (LLMs) like GPT-3 have shown promising results in their ability to generate convincing natural language and successfully solve tasks including classification, answering questions, and summarization even in zero-shot and few-shot environments Brown et al. (2020). This makes them a popular choice as opposed to a pre-trained model, which needs to be adjusted separately for each downstream task. In this paper, we propose an ensemble of rule-based methods, traditional sequence models, large language models, and BERT-based models to develop an automated system for generating these notes from doctor-patient conversations. We also show that few-shot large language models outperform traditional sequence-to-sequence models in the setting of limited data.

2 Related Work

Summarization is a crucial task in NLP, particularly for extracting key information from multi-speaker conversations. Various approaches have been proposed for meeting summarization, such as DialogLM Zhong et al. (2022), a pre-trained neural encoder-decoder model. In the context of medical dialogues between doctors and patients, identifying symptoms, diagnoses, and treatments is essential for deriving a medical solution. Song et al. (2020) introduced the hierarchical encoder-tagger model (HET) to specifically identify important utterances in medical conversations for summarizing medical conversations. Krishna et al. (2021) introduced pointer generator networks for deep summarization of physician-patient dialogues. Joshi et al. (2020) introduced a variant of the pointer generator network that handles negations and imposes a penalty on the generator distribution and Zhang et al. (2021) fine-tuned BART models for summarizing doctor-patient interactions.

3 Task and Dataset Details

The MEDIQA-Chat 2023 Ben Abacha et al. (2023a) shared task has been developed to foster research in the field of automatic clinical note

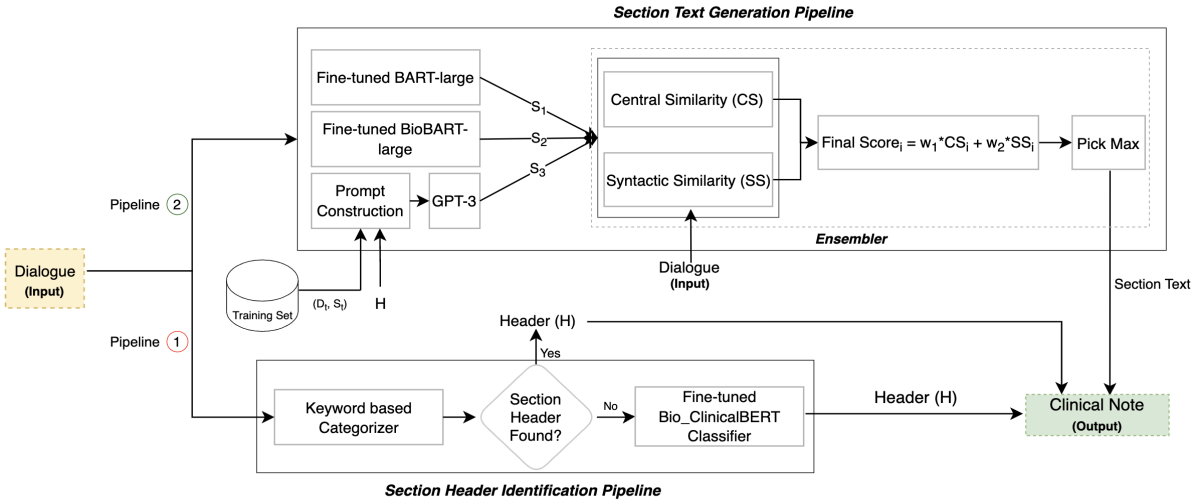

Figure 1: Pipeline for conversation to clinical note generation. Here, S_i, D_t, S_t, CS_i, and SS_i are section texts generated from independent models, dialogue from training data, section text from training data, central similarity score for S_i, syntactic similarity score for S_i respectively.

generation derived from doctor-patient conversations. It comprises three tasks[1], namely, Short Dialogue2Note Summarization (Task A) Ben Abacha et al. (2023b), Full Dialogue2Note Summarization (Task B) wai Yim et al. (2023), and Note2Dialogue Generation (Task C) wai Yim et al. (2023). Our work primarily focuses on Task A. This task requires us to create a section summary, encompassing both the section header and text, based on a short input of a doctor-patient conversation.

The dataset consists of 20 distinct section headers for each conversation, such as Medications, Review of Systems, Past Surgical History, Chief Complaint, etc. The training set contains 1,201 pairs of conversations, each accompanied by their relevant section headers and text, while the validation set is composed of 100 pairs of conversations and their respective summaries. Table 1 shows statistics around the train/val/test data splits. Table 3 shows a few snippets of actual training data containing Section Header, Section Text, and doctor-patient conversation.

4 System Description

In this section, we give a detailed explanation of our proposed system[2]. We propose two separate pipelines - one for section header identification and the other for section text generation. We also discuss our ensemble strategy and related intuition.

[1]Task Page: https://github.com/abachaa/MEDIQA-Chat-2023

[2]Code: https://github.com/prakhar21/MEDIQA-CHAT-2023-NewAgeHealthWarriors

Data	Dialogue len	Sec. len	# Samples
Train	105.6	40.5	1201
Val	89.9	36.0	100
Test	100.0	-	200

Table 1: Dataset Statistics. Dialogue len, and Sec. len denotes the average number of words at the Dialogue and section level respectively.

Category	Coverage Text
Allergy	Incase of no allergies, reply with keyword 'no known allergies'
Fam/Sochx	Incase of no family medical history found, reply with keyword 'noncontributory'
Genhx	Don't forget to mention age and gender of the patient, if present.

Table 2: Examples of Coverage Text

Finally, we output results from both pipelines to generate final summaries.

Section Header Identification: The task of Section Header Identification involves categorizing a given doctor-patient conversation to the relevant header from a list of pre-defined headers. Table 4 lists down all the available headers along with their expanded form which we received as a part of the task description.

We developed a 2-step strategy for detecting the accurate section header for a given doctor-patient conversation. Figure 1 shows the inference flow

section_header	section_text	dialogue
PASTMEDICALHX	Asthma.	**Doctor**: How's your asthma since you started using your inhaler again? **Patient**: Much better. I don't know why I didn't take it with me everywhere I went. **Doctor**: It's important to carry it with you, especially during times where you're exercising or walking more than usual. **Patient**: Yeah. I think I've learned my lesson. **Doctor**: Besides asthma, do you have any other medical problems?
CC	Burn, right arm.	**Doctor**: Hi, how are you? **Patient**: I burned my hand. **Doctor**: Oh, I am sorry. Wow! **Patient**: Yeah. **Doctor**: Is it only right arm? **Patient**: Yes.
FAM/SOCHX	His brothers had prostate cancer. Father had brain cancer. Heart disease in both sides of the family. Has diabetes in his brother and sister.	**Doctor**: Can you tell me about any diseases that run in your family? **Patient**: Sure, my brother has a prostate cancer. **Doctor**: Okay, brother. **Patient**: My father had brain cancer. **Doctor**: Okay, dad. **Patient**: Then on both sides of my family there are many heart related issues. **Doctor**: Okay. **Patient**: And my brother and sister both have diabetes. **Doctor**: Okay. **Patient**: Yes, that's it.

Table 3: Sample data from training set

of the Section Header Identification pipeline. In step-1, we categorize a given conversation to its section header using our Keyword lookup list. We refer to this as 'Keyword based Categorizer' in the diagram. We manually curated this list by going through many examples of section texts for every section header from the training data. If no section header is identified in this step, we pass the same conversation to the Bio_ClinicalBERT Alsentzer et al. (2019) model, variant of BERT Devlin et al. (2018), which we had fine-tuned on our dataset of conversation and section header pairs. Please refer to Section 5 and 6 for more details on model description, implementation, and results.

Section Text Generation: The task of Section Text Generation involves generating a summary of the given doctor-patient conversation. We propose an ensemble of 3 transformer-based models, i.e, BART-large, BioBART-large, and few-shot GPT-3 for the same. Here, we fine-tune BART-large Lewis et al. (2019) and BioBART-large Yuan et al. (2022) transformer models in a sequence-to-sequence paradigm on our training dataset. The models were fine-tuned with input as doctor-patient conversations and output as associated section text with the training objective of maximizing the likelihood of the generated summary.

For GPT-3, we adopt a few-shot prompt engineering-based Liu et al. (2023) approach for generating our section text. Few-shot prompting helps enable in-context learning for large language models like GPT-3. Figure 2 shows a detailed annotation of the GPT-3 prompt that we use for our purpose. In the figure, <*Dialogue Example*> is an example dialogue from the training dataset that we sample randomly based on the predicted section header(<*Section Header*>) on the <*Test Dialogue*> sequence. The intuition behind adding this extra knowledge to our prompt was to help our model learn the writing style of actual section text. We also experimented by giving multiple examples of dialogue, section header, and section text as a part of our prompt. Please refer to Section 6 for more details on experiments.

During our initial analysis of the training dataset, we observed that there were certain specific writing

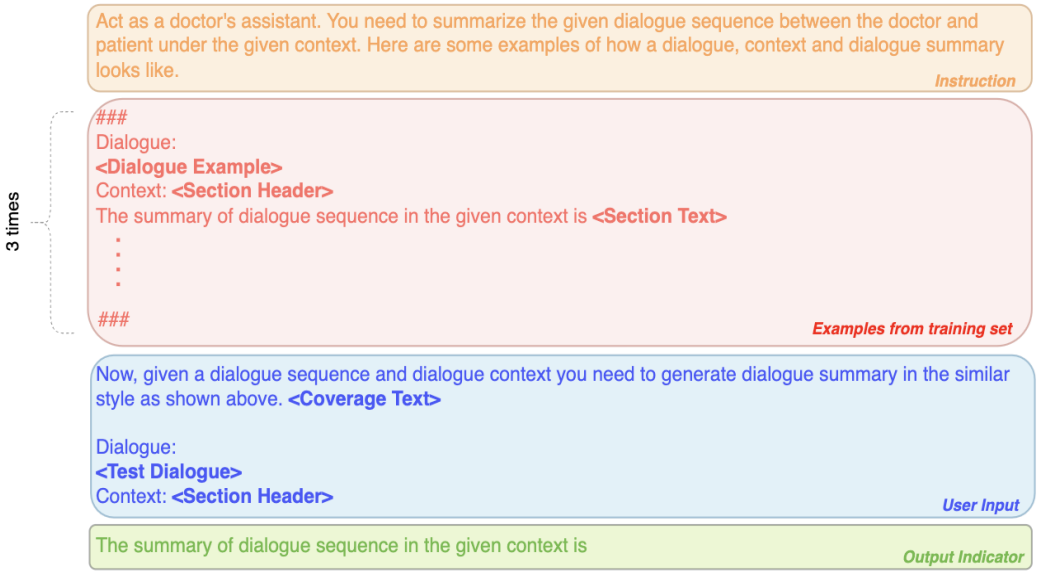

Figure 2: GPT-3 prompt structure

Header	Full Header
Fam/Sochx	Family History/Social History
Genhx	History of Present Illness
Pastmedicalhx	Past Medical History
Cc	Chief Complaint
Pastsurgical	Past Surgical History
Allergy	
Ros	Review of Systems
Medications	
Assessment	
Exam	
Diagnosis	
Disposition	
Plan	
Edcourse	Emergency Department Course
Immunizations	
Imaging	
Gynhx	Gynecologic History
Procedures	
Other_history	
Labs	

Table 4: List of all section headers and their full forms

patterns present in the section text. For example - in numerous cases, the section text for 'History of Present Illness' starts with the patient's age and gender. There were also cases where section text had words like 'Noncontributory', 'None', etc. To accommodate such a writing style in the model's output, we added an additional text called <*Coverage Text*> for some of the selected section headers. We have shown some examples of coverage text in Table 2.

Figure 1 shows the inference flow of the Section Text Generation pipeline. Once the section text from each of the models is generated, we score all of them and pick the one with the maximum score as our choice of final generated section text. For each section text, we calculate a final score based on a weighted scoring scheme that combines both central and syntactic similarity scores with weights w_1 and w_2 respectively. We found $w_1 = w_2 = 0.5$ to work best for our use case. For calculating central similarity for each of the generated section text, we implement the work done in Kobayashi (2018) and for syntactic similarity, we calculate the token-level Jaccard similarity between the section text and input dialogue. Jaccard similarity is defined as the ratio between the intersection of two sets and the union of two sets, and it is often used as a metric of similarity. Intuitively, section text that centrally captures the majority theme across all the generated section texts will have a high central similarity score, whereas, the syntactic similarity would help ensure faithfulness. Goel et al. (2021)

Dialogue	AS	GS	R1
Doctor: Do you drink alcohol or smoke cigarettes? **Patient**: No, I do not. **Doctor**: Are you sure? **Patient**: Yes.	Denies the use of alcohol or tobacco.	Denies the use of alcohol or tobacco.	1.0
Doctor: Have you ever had surgery? **Patient**: One too many times. **Doctor**: Which ones? **Patient**: I had my appendix taken out and glaucoma surgery fairly recently. I also had my gallbladder taken out ten years ago and a partial colon resection due to colon cancer in nineteen sixty one. **Doctor**: Any recurring episodes of colon cancer? **Patient**: No, thankfully.	Partial colon resection of colon carcinoma in 1961 with no recurrence, cholecystectomy 10 years ago, appendectomy, and glaucoma surgery.	Appendectomy and glaucoma surgery. Cholecystectomy 10 years ago and partial colon resection due to colon cancer in 1961.	70.3
Doctor: Any difficulty in hearing? **Patient**: No. **Doctor**: Difficulty swallowing? **Patient**: Um no. **Doctor**: Any double vision or blurred vision or difficulty seeing things properly? **Patient**: No, no problem at all. **Doctor**: Okay. **Doctor**: How about headaches or migraine? **Patient**: No headache. **Doctor**: Did you notice any change in your bowel moment? **Patient**: No, it is the same. **Doctor**: Any pain while urinating or change in frequency? **Patient**: No. **Doctor**: Okay.	No headaches. No visual, hearing, or swallowing difficulties. No changes in bowel or urinary habits.	NEUROLOGICAL: No difficulty in hearing, swallowing, double vision, blurred vision, headaches, or migraines. GASTROINTESTINAL: No change in bowel movements or difficulty urinating.	48.7
Doctor: It seems like you are not feeling very well today? **Patient**: Yeah. I have had diarrhea and pain in my stomach. **Doctor**: Have you experienced any vomiting? **Patient**: Yes. I threw up this morning."	Diarrhea, vomiting, and abdominal pain.	The patient states that he is not feeling very well today. He has had diarrhea and pain in his stomach this morning, and he has had vomiting this morning.	23.5
Doctor: I will do some examinations on you. I will check your chest and then I will talk to you as I move forward, okay? **Patient**: I'm okay with that. **Doctor**: So, let's see what we have here. Hm, Yeah, just looks good. I do not find anything abnormal.	CHEST: The chest examination is unremarkable.	Chest x-ray without any abnormality.	16.7
Guest_clinician: I did a review of her systems, and everything looks normal other than what was mentioned earlier. **Doctor**: Okay, thanks for your help. **Guest_clinician**: No problem.	The remaining ROS is unremarkable.	Review of Systems: Everything appears to be normal other than what was mentioned earlier.	0.0
Doctor: Are you allergic to anything, food or medicines? **Patient**: No allergies that I know of.	None.	No known drug allergies.	0.0

Table 5: Examples are arranged in decreasing order of R1. Here, AS, GS, and R1 refer to the actual, generated section texts, and Rouge-1 respectively.

defines a faithful summary to be one that contains minimal information outside the source text. Finally, output from both pipelines is used to report the final generated clinical note.

5 Model Background

In this section, we discuss in brief the background of various machine learning models that we have used in our implementation.

Bio_ClinicalBERT: The Bio_ClinicalBERT[3] model is initialized from BioBERT Lee et al. (2020) and trained on all notes from MIMIC III Johnson et al. (2016), a database containing electronic health records from ICU patients. The model was pre-trained on a GeForce GTX TITAN X 12 GB GPU, on a batch size of 32, a maximum sequence length of 128, and a learning rate of 5x10-5.

BART & BioBART: BART Lewis et al. (2019) [4] is a transformer encoder-decoder (seq2seq) model with a bidirectional (BERT-like) encoder and an autoregressive decoder to perform complex NLG tasks like summarization, translation, etc. BART is pre-trained to reconstruct the original text from the noisy text. BioBART Yuan et al. (2022) [5], a BART variant was pre-trained on PubMed abstracts to achieve biomedical domain adaption.

GPT-3: GPT-3 Brown et al. (2020) is an autoregressive language model with 175 billion parameters. It achieves strong performance on many NLP datasets, including translation, and question-answering, as well as several tasks that require on-the-fly reasoning or domain adaptation.

6 Experiments and Results

In this section, we discuss experiments, implementation details, and results obtained for both our tasks.

Section Header Identification: We trained Fasttext Joulin et al. (2016) along with multiple BERT variants like BERT-base Devlin et al. (2018), RoBERTa Liu et al. (2019), and Bio_ClinicalBERT Alsentzer et al. (2019) for identifying section headers from doctor-patient dialogues. We use simpletransformers Rajapakse (2019) python library for fine-tuning all our transformer models. Amongst all of them, Bio_ClinicalBERT gave us the best score on the validation dataset. Our final model also incorporates a Keyword-based categorizer in the pipeline giving us the best accuracy of 77% on the validation set. We use weighted cross-entropy as our loss function because of the skewed distribution of headers in the training data. The corresponding weights per header category were calculated using sklearn's *compute_class_weight* function on the training dataset and we train our best model for 10 epochs.

Section Text Generation: We fine-tuned BART, BioBART architectures and, also inferred GPT-3 model in a few-shot setting. Interestingly, GPT-3 in the few-shot setting outperforms all our fully supervised models by **1.2+** Rouge-Avg points (Refer Table 7). Rouge-Avg(RA) is the average score of Rouge-1 (R1), Rouge-2 (R2), Rouge-L (RL), and RougeLSum (RLS). We train both our BART-large and BioBART-large models for 5 epochs, set a beam size of 5 while decoding the sequence, and use the cross-entropy loss as our objective function. We tuned all the hyperparameters based on the performance of our model on the validation set. Before coming up with the final prompt structure for few-shot GPT-3, we experimented with a couple of things. We tested by keeping <*Section Header*> in their original form (as received in training data - acronymized) and also by replacing them to their full form, as received in the tasks description. Please refer to table 4 for a list of all headers and their expanded forms.

We experimented with the number of ground truth examples for our prompt. We tested with a maximum of 3 examples, because, for values higher than 3, the rate of getting the maximum token limit error from the GPT-3 API had increased significantly. Across 0, 1, 2, and 3, we found 3 to be giving the best results on the validation set. Finally, our ensemble of BART, BioBART, and GPT-3 outperforms all our individual models by **0.9+** Rouge-Avg points (Refer Table 7) on the validation dataset. Some sample dialogues and generated section text are shown in Table 5.

Table 6 and 7 show the result of our evaluation for the Section Header Identification and Section Text Generation pipeline on the validation and test datasets respectively. The evaluators report a few more metrics such as BERTScore Zhang

[3]Bio_ClinicalBERT Model:https://huggingface.co/emilyalsentzer/Bio_ClinicalBERT

[4]BART Model: https://huggingface.co/facebook/bart-large

[5]BioBART Model: https://huggingface.co/GanjinZero/biobart-large

Split	Method	Accuracy
Val	Fine-tuned Bio_ClinicalBERT	0.75
	Fine-tuned Bio_ClinicalBERT + KW Classifier	0.77
Test	Fine-tuned Bio_ClinicalBERT + KW Classifier	0.73

Table 6: Section Header classification from Dialogue on Val set. Here, KW stands for Keyword-based.

Split	Method	R1	R2	RL	RLS	RA	BSF	BLEURT
Val	Fine-tuned BioBART-large	38.1	14.8	31.0	31.0	28.7	-	-
	Fine-tuned BART-large	39.0	14.6	31.6	31.4	29.2	-	-
	Few-shot GPT-3	40.3	16.5	32.4	32.2	30.4	-	-
	Ensemble	41.6	17.2	33.1	33.3	31.3	-	-
Test	Ensemble	39.8	17.17	33.14	33.13	30.81	69.82	53.5

Table 7: Section text generation from Dialogue on the validation set. Here, R1, R2, RL, RLS, RA, and BSF refer to Rouge-1, Rouge-2, Rouge-L, RougeLSum, Rouge-Avg, and BertScore-F1 respectively.

et al. (2019), a metric that focuses on computing semantic similarity between tokens of reference and hypothesis, and BLEURT Sellam et al. (2020), a learned evaluation metric based on BERT for evaluating the generated summaries. The default models used for calculating BERTScore and BLEURT were 'microsoft/deberta-xlarge-mnli' [6] and 'BLEURT-20'[7] respectively. We report the score for these metrics in Table 7 for the test datasets due to computing constraints.

7 Observations

Here we discuss some observations that we made on results as shown in Table 5.

- With reference to examples 1 and 2, our model was able to correctly capture the year, duration, and other diagnostic details.
- With reference to example 3, our model was able to capture more details and also attempted to categorize diagnosis under relevant categories, which was not originally present in the ground truth summary.
- With reference to examples 4 and 5, our model generated some made-up facts such as the duration of the day, and the chest examination being an x-ray.
- With reference to examples 6 and 7, our model was accurately able to generate text with the same findings. However, it wrote it in an elaborate manner pushing R1 to 0.0.

8 Conclusion and Future Work

We have presented a novel ensemble-based approach for the task of automatic short medical dialogue to note summarization. Our method effectively combines fully supervised transformer models, few-shot GPT-3, and rule-based systems, generating accurate and coherent summaries of doctor-patient conversations. The proposed system demonstrates competitive performance on the MEDIQA-Chat 2023 Task A, highlighting its potential to enhance telecare and healthcare services.

As part of future work, we plan to explore advanced pre-trained models and techniques to further improve our system's performance in the medical context. Additionally, we aim to investigate the applicability of our approach in handling more complex dialogues. We also plan to conduct an in-depth analysis of the generated summaries to identify areas for further fine-tuning.

[6]DeBERTa Model: https://huggingface.co/microsoft/deberta-xlarge-mnli

[7]BLEURT-20 Model: https://huggingface.co/lucadiliello/BLEURT-20

References

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

Asma Ben Abacha, Wen wai Yim, Griffin Adams, Neal Snider, and Meliha Yetisgen. 2023a. 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. 2023b. An empirical study of clinical note generation from doctor-patient encounters. In *EACL 2023*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Salem Garfan, Abdullah Hussein Alamoodi, BB Zaidan, Mohammed Al-Zobbi, Rula A Hamid, Jwan K Alwan, Ibraheem YY Ahmaro, Eman Thabet Khalid, FM Jumaah, Osamah Shihab Albahri, et al. 2021. Telehealth utilization during the covid-19 pandemic: A systematic review. *Computers in biology and medicine*, 138:104878.

Karan Goel, Nazneen Rajani, Jesse Vig, Samson Tan, Jason Wu, Stephan Zheng, Caiming Xiong, Mohit Bansal, and Christopher Ré. 2021. Robustness gym: Unifying the nlp evaluation landscape. *arXiv preprint arXiv:2101.04840*.

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.

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*.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2016. Bag of tricks for efficient text classification. *arXiv preprint arXiv:1607.01759*.

Hayato Kobayashi. 2018. Frustratingly easy model ensemble for abstractive summarization. In *Proceedings of the 2018 conference on empirical methods in natural language processing*, pages 4165–4176.

Kundan Krishna, Sopan Khosla, Jeffrey P. Bigham, and Zachary C. Lipton. 2021. Generating soap notes from doctor-patient conversations using modular summarization techniques.

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

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2023. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, 55(9):1–35.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

T. C. Rajapakse. 2019. Simple transformers. `https://github.com/ThilinaRajapakse/simpletransformers`.

Thibault Sellam, Dipanjan Das, and Ankur P Parikh. 2020. Bleurt: Learning robust metrics for text generation. *arXiv preprint arXiv:2004.04696*.

Yan Song, Yuanhe Tian, Nan Wang, and Fei Xia. 2020. Summarizing medical conversations via identifying important utterances. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 717–729, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Wen wai Yim, Yujuan Fu, Asma Ben Abacha, Neal Snider, Thomas Lin, and Meliha Yetisgen. 2023. The aci demo corpus: An open dataset for benchmarking the state-of-the-art for automatic note generation from doctor-patient conversations. *Submitted to Nature Scientific Data*.

Hongyi Yuan, Zheng Yuan, Ruyi Gan, Jiaxing Zhang, Yutao Xie, and Sheng Yu. 2022. Biobart: pretraining and evaluation of a biomedical generative language model. *arXiv preprint arXiv:2204.03905*.

Longxiang Zhang, Renato Negrinho, Arindam Ghosh, Vasudevan Jagannathan, Hamid Reza Hassanzadeh, Thomas Schaaf, and Matthew R. Gormley. 2021. Leveraging pretrained models for automatic summarization of doctor-patient conversations. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3693–3712, Punta Cana, Dominican Republic. Association for Computational Linguistics.

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

Ming Zhong, Yang Liu, Yichong Xu, Chenguang Zhu, and Michael Zeng. 2022. Dialoglm: Pre-trained model for long dialogue understanding and summarization.