PULSAR at MEDIQA-Sum 2023: Large Language Models Augmented by Synthetic Dialogue Convert Patient Dialogues to Medical Records

Viktor Schlegel[1,2], Hao Li[2], Yuping Wu[2], Anand Subramanian[1,3], Thanh-Tung Nguyen[1], Abhinav Ramesh Kashyap[1], Daniel Beck[4], Xiaojun Zeng[2], Riza Theresa Batista-Navarro[2], Stefan Winkler[1,3] and Goran Nenadic[2]

[1]*ASUS Intelligent Cloud Services (AICS), Singapore.*
[2]*Dept. of Computer Science, University of Manchester, United Kingdom.*
[3]*Dept. of Computer Science, National University of Singapore, Singapore.*
[4]*School of Computing and Information Systems, University of Melbourne, Australia.*

Abstract

This paper describes PULSAR, our system submission at the ImageClef 2023 MediQA-Sum task on summarising patient-doctor dialogues into clinical records. The proposed framework relies on domain-specific pre-training, to produce a specialised language model which is trained on task-specific natural data augmented by synthetic data generated by a black-box LLM. We find limited evidence towards the efficacy of domain-specific pre-training and data augmentation, while scaling up the language model yields the best performance gains. Our approach was ranked second and third among 13 submissions on task B of the challenge. Our code is available at https://github.com/yuping-wu/PULSAR.

Keywords

Abstractive Summarisation, AI for Healthcare, Dialogue Summarisation, Natural Language Processing

1. Introduction

With the recent successes of generative large language models (LLMs) on a variety of tasks [1] and domains [2], even in the face of data scarcity [3], there is vivid interest in identifying potential application scenarios that could benefit from the power of LLMs. One of the promising domains is healthcare [4] as many administrative tasks involve the transformation of textual data. LLM-based approaches that assist hospital staff in repetitive administrative tasks have the potential to improve operational efficiency and documentation quality, optimise revenue streams, reduce cognitive load on healthcare experts, and ultimately result in better and more effective patient care [5].

A range of different scenarios have been investigated for the suitability of LLM-based assistance, such as summarising patient progress notes as discharge summaries [6] or identifying problems that need treatment during a patient's hospital course [7]. One of the potential tasks is summarising doctor-patient dialogue as medical records [8]. Dialogue summarisation, an

CLEF 2023: Conference and Labs of the Evaluation Forum, September 18–21, 2023, Thessaloniki, Greece
hao.li-2@manchester.ac.uk (H. Li); yuping.wu@manchester.ac.uk (Y. Wu)

established task in the Natural Language Processing (NLP) community, aims to identify salient topics in a multi-turn dialogue [9]. State-of-the-art approaches typically formulate the problem as abstractive summarisation, making the task a prime candidate for further investigation of the potential of LLMs in clinical settings. In this scenario, conversations between patients and doctors need to be transformed into (excerpts of) clinical documentation. For example, if a 27 year old female patient mentions that they are experiencing *"Sore throat, runny nose, dry cough and fever 37.5 °C"*, the corresponding entry can be the *"Subjective"* section of a medical record excerpt, e.g., *"Patient is a 27 year old female who presents with sore throat, runny nose dry cough and a fever of 37.5 °C."* This documentation is typically performed by the consulting doctor or an attending nurse. Despite bearing potential impact for automation, with clinical staff spending at least 35 minutes of their time every other day on writing such clinical notes [10], this task was underexplored by the NLP community, compared to other hospital-related tasks, such as clinical coding [11, 12], or generating radiology reports [13]. More recently, the task has received more attention [14], however studies thus far have either focussed on narrow department selections [15, 16], did not focus on medical documentation generation [17], or have not released their data publicly [18].

To that end, the ImageClef 2023 MediSum shared task released a collection of dialogues and corresponding clinical notes in an effort to spark interest and advance the state of the art in dialogue as clinical note summarisation [8]. The task revolves around three core sub-tasks: *(A)* identifying the topic of a conversation from a selection of possible medical note sections (i.e., *"Subjective"* in the previous example), *(B)* summarising conversation snippets to appropriate sections in medical records, and, finally, *(C)* summarising full conversations to full medical records. While conversations are synthetic, the corresponding clinical notes are real, doctor-written documentation.

Our guiding objective to participate in this task was to investigate, how well a recently proposed LLM training framework can generalise to new tasks with as little adaptation as possible [19]. At its core, the framework *(i)* fine-tunes a LLM with a pre-training objective that learns to reconstruct a pseudo-summary consisting of automatically extracted medical terms and *(ii)* employs data augmentation (DA) by instructing Black-Box LLMs to obtain task-specific training data. As such, the DA framework supports any LLM, such as Bloom [20], GPT-3 [21] or GPT-3.5 [22].

Our submission for task B was ranked second best overall among all participants. Although we have not actively sought to compete in Task C, we observed that our data augmentation technique could improve the performance, particularly when the training data is scarce. These findings underline the potential of LLMs in various settings as well as the generalisability of our proposed approach.

2. Task Definition

In this section, we describe and formalise the three tasks of the ImageClef 2023 MediSum challenge.

Task A – Dialogue2Topic Classification In this task, participants need to identify the topic of a conversation. The list of possible topics corresponds to the 20 different fine-grained sections that can be part of a medical record, such as *"Subjective"*, i.e., the subjective description of symptoms by the patient.

Task B – Dialogue2Note Summarization Here, participating systems need to convert a conversation on a specific topic into a corresponding section in the medical record. This task can be regarded as conditional generation, sequence-to-sequence translation or abstractive summarisation. Approaches are evaluated on multiple natural language generation metrics, both based on n-gram overlap, i.e., ROUGE [23], as well as semantic similarity [24, 25]. 1201 training and 100 validation examples are provided. 200 examples form the test set.

Task C – Full-Encounter Dialogue2Note Summarization This task is formulated similarly to Task B, however here, the inputs are full notes and the evaluated systems need to generate medical record outputs for the four general sections *"Subjective"*, *"Objective Exam"*, *"Objective Results"* and *"Assessment and Plan"*. This task features only 67 training and 20 validation examples, with 40 examples reserved for testing. The systems are evaluated based on their output for each of the sections using the ROUGE metrics from Task B; the results are averaged across all sections. An alternative mode of evaluation combines all outputs into one single record and measures the n-gram overlap by means of the ROUGE score.

The tasks appear to be arranged as a progression, where, given a dialogue, a segmentation and classification model could segment the topics of the conversation (Task A) to be used as input for a Dialogue Snippet Summarisation Model (Task B), the output of which can be arranged as a full medical record (Task C). However, as our goal was to evaluate how well the proposed framework generalises to the tasks with as little adaptation as possible, we decide not to make any task-specific adaptations even if they could provide beneficial given the particular arrangement of the tasks. Thus, we do not rely on any additional information, treat tasks B and C in isolation, and disregard task A for not being a generative task.

3. Methodology

3.1. Language model Pre-training

Motivated by the success of predicting masked words [26] and contiguous spans [27] as self-supervised training objectives, we customised the pre-training objectives for the medical domain generation task to concatenate "gap text spans (sentences)" into a pseudo-summary. Each masked span is a medical term from the input text identified by the QuickUMLS [28] or a NER model fine-tuned on a N2C2 dataset (i2b2-2010 challenge [29]). Specifically, as shown in Figure 1, pre-training consisted of three different policies: first, when both QuickUMLS and N2C2 NER models identified entities, the QuickUMLS results were used in 70% of cases and the results of the N2C2 NER model were used in 30%. Second, when only one of them predicted any output, that output was used for masking. Third, when neither had any output, then 15% of the

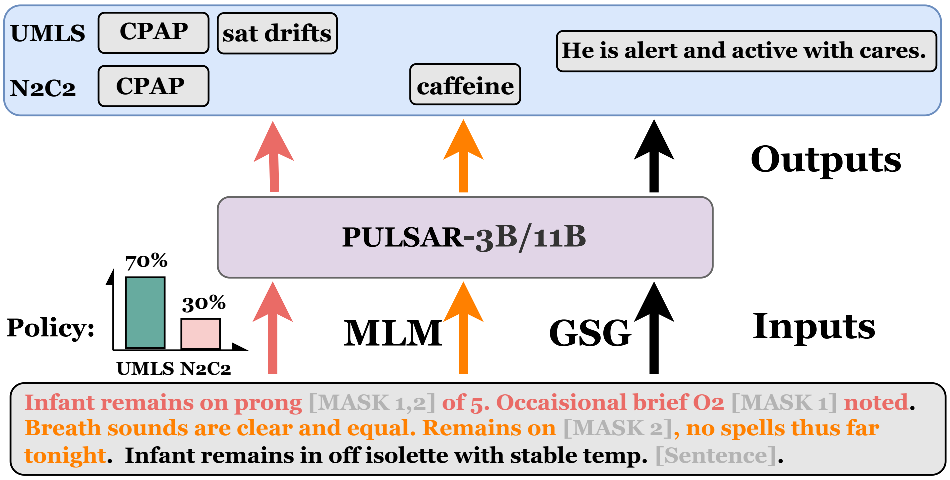

Figure 1: Example of the pre-train objective of PULSAR. Both Masked Language Model (MLM) and Gap Sentences Generation (GSG) have been employed in this scenario. Red and orange arrows exemplify the UMLS and N2C2 MLM masking strategy, respectively. Meanwhile, the black arrow shows the GSG masking strategy, where a whole sentence has been masked.

sentences were masked at random. These text spans were replaced with "sentinel" mask tokens $< extra\_id\_i >$ to inform the model that input was masked. In order to provide the model with sufficient medical knowledge, we used the MIMIC-III [11], a pre-trained corpus of 2 million data, which consists of a large number of clinical records, such as admission notes, discharge summaries or lab results.

3.2. Data Augmentation (DA)

Both tasks suffer from scarcity of training data, specifically Task C, which requires generating comprehensive clinical notes based on lengthy patient-doctor conversations based on only 67 training examples. These may be insufficient to train a model capable of performing well on the task. To address this issue, we adopt data augmentation to generate additional examples for training, as this has been shown to improve performance in data-scarce scenarios [30, 31].

Prompting Strategy We observed that Large Language Models (LLMs) such as ChatGPT are proficient in understanding clinical context and manipulating clinical data. Therefore, we utilise a pre-existing LLM to generate data for the model's training. Ideally, the data generation approach would involve providing conversations and requesting the LLM to produce the corresponding medical note. However, we are limited by the fact that we only have 67 full-length conversations in our dataset. Nonetheless, we have access to a significantly larger number of medical notes. Hence, we invert the task by prompting the LLM with a medical note (or its snippet) and ask it to generate a hypothetical conversation between the doctor and the patient. We then use the generated conversations as input to train our model to produce the corresponding clinical note.

We employ the OpenAI ChatGPT API (gpt-35-turbo) for data augmentation, utilising a two-stage prompting strategy to generate data effectively. In the first stage, we use in-context learning with one-shot prompting to prompt the LLM to generate a fictitious conversation between the doctor and patient based on the medical note, while adhering to important guidelines. We provide only one example picked from the training set as we are limited by token context windows for the API. In the second stage (only performed for task C), we prompt the model to include conversational fillers such as "ums", "uh", and "hmm" to the generated conversation from the first stage, as we noticed that the model did not include these fillers despite our instructions in the first stage.

Dataset Utilised For task B, we extract matching subsection headings from the MIMIC-III database [11], adapting the pre-processing method from Yang et al. [32] to identify section headers. We rank the generations based on their average Rouge similarity to all training instances and pick the top-scoring n conversations.

For task C, we utilise a corpus of freely available medical notes scraped from MTSamples, which is available on Kaggle[1]. Since the dataset contains medical transcriptions of notes from various medical specialities, we devise a method to pick samples from the dataset that are the closest to the medical notes in our training set. To do this, we identify and curate a list of the section headers in the training set through a heuristic approach by exploiting the fact that section headers are usually written in all capital letters. We split the document by newline and extract the lines which are fully upper-cased and add these contents to our list of section headers. We then score the medical notes in MTSamples based on the number of headers that each document has based on the curated list from the previous step and pick the top n documents from MTSamples with the highest scores to use as input for DA. We end up with a corpus of 746 data samples due to the fact that some inputs were flagged as offensive by OpenAI's content moderating policy.

4. Empirical Evaluation

4.1. Experiments set-up

We aim to empirically evaluate, how well our framework can solve the problem of converting patient dialogues to medical records. We pursue the following questions:

(i) How well can our proposed approach convert doctor-patient dialogues to Medical Records?
(ii) Does the domain-specific pre-training objective improve performance?
(iii) What is the impact of model scale on the performance?
(iv) Does synthetic data augmentation improve performance on the tasks?

To answer question *(i)* we empirically evaluate our proposed framework on the task B and C test sets of the ImageClef Challenge. For evidence towards question *(ii)*, we compare the performance of `PULSAR` to that of equally-sized `Flan-t5` models. Regarding question *(iii)*, we compare the performance of variously sized models of the same architecture and for question *(iv)*, we compare the performance of models trained on available data only to those fine-tuned on synthetically generated conversation data.

[1] https://mtsamples.com/ and https://www.kaggle.com/datasets/tboyle10/medicaltranscriptions, respectively

Table 1
Validation set performance as measured by {1,2}-gram overlap Rouge-{1,2} and longest sequence overlap Rouge-L Rouge-LSum. Models with the *-nDG were augmented with n synthetic examples. The *-HEADER suffix denotes that the section header was used as input.

ID	Setting	Rouge			
		R1	R2	RL	RLSum
	11B and 7B models				
11B2	PULSAR-11B	49.20	22.25	41.36	45.57
11B1	FLAN-T5-11B	50.75	27.92	44.93	47.58
7B1	LLaMA-7B-LoRA	39.3	15.7	33.1	33.1
	3B models				
3B4	PULSAR-3B-735DG	37.89	17.83	30.40	34.76
3B3	FLAN-T5-3B-735DG	41.44	19.05	33.93	38.41
3B2	PULSAR-3B	37.92	16.64	30.64	34.58
3B1	FLAN-T5-3B	41.91	19.41	33.76	38.04
	LARGE models (< 1B parameters)				
L4	FLAN-T5-LARGE-HEADER	38.70	16.82	31.85	36.22
L3	FLAN-T5-LARGE-1000DG	39.41	17.04	31.85	36.78
L2	FLAN-T5-LARGE	39.27	17.42	31.95	36.49
L1	CLINICAL-T5-LARGE	19.11	8.34	16.31	17.29

4.2. Implementation Details

Pre-training `PULSAR-*` is initialised with weights from the corresponding `Flan-t5-*` models [33] and pre-trained with four NVIDIA Tesla A100 80GB GPUs for 1 epoch on all MIMIC-III notes. Huggingface Accelerate is used to optimise GPU memory usage with the Fully Sharded Data Parallel (FSDP) paradigm. We set the training batch size per GPU device as 4 and the gradient accumulation step as 8 to accelerate the training process.

Fine-tuning We fine-tune all models for 3 epochs. We experiment with encoder-decoder `Flan-T5`, `PULSAR` and `Clinical-T5` [34]models, with the configurations `*-Large` (0.9B Parameters), `*-3B` and `*-11B`. Unless stated otherwise, the models are trained on two A100 80GB GPUs with cumulative batch size of 8 and the learning rate of 3^{-5}. For the largest of them, i.e., (`Flan-t5-11B` and `PULSAR-11B`), we use FSDP with CPU offloading. We also experiment with a decoder-only model, `LLAMA-13B`, freezing and quantising the base model in eight bit [35] and using the parameter-efficient LoRA [36] method. More details on hyper-parameter choice are reported in the appendix.

4.3. Results and analysis

At a glance, Table 1 shows the results of our empirical study and Table 2 shows the final ranking of all participating systems according to the official evaluation by task organisersIn the following, we discuss our findings in context of the questions outlined in the motivation of this empirical study.

Table 2
Test set performance as measured by {1,2}-gram overlap Rouge-{1,2}, longest sequence overlap Rouge-L and Rouge-LSum and semantic similarity metrics BertScore-F1 and Bleurt; ranked by their aggregation.

submission	R1	R2	RL	RLSum	BS-F1	Bleurt	Agg
SuryaKiran_run3	**43.98**	18.44	35.01	35.01	**72.31**	**55.67**	**57.32**
Flan-T5-11B	42.99	**20.04**	**35.69**	**35.69**	72.18	55.49	56.89
PULSAR-11B	41.78	19.25	34.16	34.16	72.11	55.52	56.47
Tredence_run1	42.44	17.24	35.3	35.3	72.07	53.3	55.94
SuryaKiran_run2	42.09	18.83	34.2	34.2	71.37	54.23	55.90
SuryaKiran_run1	40.56	17.59	32.72	32.72	71.09	53.24	54.96
LLaMA-7B-LoRA	38.15	17.3	31.42	31.42	71.77	51.52	53.82
HuskyScribe_run1	37.67	15.04	31.26	31.26	70.54	50.37	52.86
Tredence_run2	36.21	13.84	29.66	29.66	68.82	47.29	50.77
uetcorn_run1	29.11	10.73	22.94	22.94	65.85	49.42	48.13
uetcorn_run2	28.75	10.69	23.09	23.09	65.96	49.22	47.98
uetcorn_run3	28.72	10.7	23.04	23.04	65.92	49.13	47.92
SKKU-DSAIL_run1	26.03	11.31	18.22	18.22	59.29	53.05	46.12

Our approach generalises well to the dialogue summarisation task. Overall, our approach generalises well to Task B, with our best model (Table 1, 11B1) surpassing the 50 Rouge-1 scores mark, which means that on average, half of the prediction tokens are found in reference and vice versa. The high Rouge-L score of 44 suggests that most of these overlapping tokens indeed form a sequence. However, these scores may be "boosted" by the presence of many short target sequences in the evaluation data, such as *"Noncontributory."* or *"No known allergies."*, when a dialogue revolves around a topic that does not contribute to the patients' hospital visit.

We find that the utilising the outputs of task A—the section headers—does not contribute to improving the overall performance, compare Table 1, L2 and L4. We observed the same trend across all model sizes (not reported here for brevity).

In the absence of established baselines, we interpret the official rankings of the shared task in Table 2 as additional evidence towards the success of our approach.

There is no conclusive evidence that domain-specific pre-training is beneficial. Comparing 11B1 and 11B2, and 3B1 and 3B2 in Table 1, respectively, we observe that domain-specific pre-training by learning to predict missing medical terms in MIMIC-III notes appears not beneficial, with the gap being smaller for bigger models. One possible reason for this is the domain mismatch between pre-training and application data. MIMIC-III is dominated by inpatient progress notes which track the patients' status along the hospital stay and contain abbreviations, repetitions, incomplete sentences and medical jargon. Conversely, the medical records in the challenge are well-written and stem most likely from admission notes or outpatient encounters, where most of the *initial* documentation about a new patients' particulars, such as their chief complaint, medical history and drug allergies happens. Additionally, input dialogues have a colloquial tone, further adding to the domain mismatch between pre-training and fine-tuning.

Model scale yields the biggest performance improvements. Comparing L*, 3B* and 11B* results in Table 1, we can see a clear trend where larger models of the same family consistently perform better. The biggest hike in performance is observed between the 3B and 11B models. This observation is in line with most literature on model scale as driver of performance and the reason for emergent abilities in LLMs [37].

We also find that the model trained with adapters can learn to perform on the task successfully, despite the relatively small (around 1.1% of the full 7B model) number of trainable parameters. However, our results suggest that updating all models' parameters is more effective, as even smaller models outperform the 7B adapter model (Table 1, L2, 3B* compared to 7B1).

Data Augmentation can be helpful if training data is extremely scarce. Larger models obtain enough signal from the training data of Task B, as there is no clear improvement in scores for the 3B models (Table 1, 3B1 vs. 3B3 and 3B2 vs. 3B4). Meanwhile, data augmentation can lead to consistent, albeit minor, improvements for smaller models (Table 1, L2 vs. L3). When training data is scarce (i.e., Task C) data augmentation helps with the performance. Subjectively, models exhibit typical generation errors such as hallucination and input copying, (see Figure 2 in Appendix) and data augmentation seems to alleviate this issue. Quantitatively, data augmentation improves performance across all metrics (27.64 vs 29.41 R1, 9.79 vs 11.60 R2, 16.24 vs 19.18 RL and 23.63 vs 26.08 RLSum without and with DA, respectively). We find the results promising, as the optimised model seems to perform well without any task-specific adaptation. Ultimately, however, this simple approach does not compete with other, potentially task specific information exploiting submissions, with the best of them scoring almost 20 Rouge-1 points higher (20.32 R2, 24.30 RL and 45.06 RLSum).

5. Conclusion

In this work, we present an LLM framework and adapt it to the task of dialogue note summarisation. While we find that the approach generalises well to this new task, there is mixed evidence of the efficacy of both domain-specific pre-training and data augmentation. Our experiments seem to align with the "bitter lesson of AI"[2], in that model scale seems to trump domain-specific adaptations. This, in turn, supports the narrative of the transformative potential of LLMs in healthcare [38], as larger LLMs become more readily available.

Our findings suggest further avenues for future work: We argued that the pre-training objective may suffer from domain mismatch. As such, experimenting with other domain-specific objectives might improve the performance of the downstream tasks. Furthermore, it is unclear how the choice of hyper-parameters for both training and inference stages (i.e., decoding arguments) impacts the overall performance. Finally, we have left it for future work to investigate, whether data augmentation could provide beneficial with a more advanced filtering strategy, for example by only augmenting examples with certain length or specific section headers. As such, we will expand the work reported in this paper by experimenting with different pre-training objectives, performing a more rigorous hyper-parameter optimisation and investigating the impact of data augmentation more closely.

[2]http://www.incompleteideas.net/IncIdeas/BitterLesson.html

Limitations

The results described in this paper should be interpreted within the following context:

- The language of the conversations is English. Due to the dominance of English data during pre-training, it is expected that all LLMs that we inspected perform better on English. It is unclear how well the approach will transfer to other languages.
- The conversations are synthetic in that they have been written based on existing medical notes, rather than transcribed from real patient-doctor dialogues. While the quality has been evaluated by medical professionals, it is unclear how well the performance would translate to real-world scenarios.
- The obtained results should be regarded as preliminary, as robust empirical results such as hyper-parameter optimisation for fine-tuning, pre-training policy selection, exhaustive search for best-performing prompts for data augmentation and strategies for data selection are often impossible given the time constraints of academic challenges and shared tasks.

References

[1] V. Sanh, A. Webson, C. Raffel, S. Bach, L. Sutawika, Z. Alyafeai, A. Chaffin, A. Stiegler, A. Raja, M. Dey, et al., Multitask prompted training enables zero-shot task generalization, in: International Conference on Learning Representations 2022, 2022.

[2] M. Agrawal, S. Hegselmann, H. Lang, Y. Kim, D. Sontag, Large language models are few-shot clinical information extractors, in: Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, 2022, pp. 1998–2022.

[3] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al., Language models are few-shot learners, Advances in neural information processing systems 33 (2020) 1877–1901.

[4] K. Singhal, S. Azizi, T. Tu, S. S. Mahdavi, J. Wei, H. W. Chung, N. Scales, A. Tanwani, H. Cole-Lewis, S. Pfohl, et al., Large language models encode clinical knowledge, arXiv preprint arXiv:2212.13138 (2022).

[5] P. Rajpurkar, E. Chen, O. Banerjee, E. J. Topol, Ai in health and medicine, Nature medicine 28 (2022) 31–38.

[6] T. Searle, Z. Ibrahim, J. Teo, R. J. Dobson, 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 (2023) 104358.

[7] Y. Gao, T. Miller, M. Afshar, D. Dligach, Bionlp workshop 2023 shared task 1a: Problem list summarization, "Proceedings of the 22nd Workshop on Biomedical Language Processing" (2023).

[8] W. Yim, A. Ben Abacha, N. Snider, G. Adams, M. Yetisgen, Overview of the mediqa-sum task at imageclef 2023: Summarization and classification of doctor-patient conversations, in: CLEF 2023 Working Notes, CEUR Workshop Proceedings, CEUR-WS.org, Thessaloniki, Greece, 2023.

[9] X. Feng, X. Feng, B. Qin, A survey on dialogue summarization: Recent advances and new frontiers, in: CLEF 2023 Working Notes, CEUR Workshop Proceedings, 2022.

[10] G. Hripcsak, D. K. Vawdrey, M. R. Fred, S. B. Bostwick, Use of electronic clinical documentation: time spent and team interactions, Journal of the American Medical Informatics Association 18 (2011) 112–117.
[11] A. E. Johnson, T. J. Pollard, L. Shen, L.-w. H. Lehman, M. Feng, M. Ghassemi, B. Moody, P. Szolovits, L. Anthony Celi, R. G. Mark, Mimic-iii, a freely accessible critical care database, Scientific data 3 (2016) 1–9.
[12] T.-T. Nguyen, V. Schlegel, A. Kashyap, S. Winkler, S.-S. Huang, J.-J. Liu, C.-J. Lin, Mimic-iv-icd: A new benchmark for extreme multilabel classification, arXiv preprint arXiv:2304.13998 (2023).
[13] M. M. A. Monshi, J. Poon, V. Chung, Deep learning in generating radiology reports: A survey, Artificial Intelligence in Medicine 106 (2020) 101878.
[14] A. B. Abacha, W. Yim, Y. Fan, T. Lin, An empirical study of clinical note generation from doctor-patient encounters, in: EACL, Association for Computational Linguistics, 2023, pp. 2283–2294.
[15] N. Kazi, I. Kahanda, Automatically generating psychiatric case notes from digital transcripts of doctor-patient conversations using text mining, PeerJ Prepr. 7 (2019) e27497.
[16] S. Enarvi, M. Amoia, M. D.-A. Teba, B. Delaney, F. Diehl, S. Hahn, K. Harris, L. McGrath, Y. Pan, J. Pinto, et al., 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, 2020, pp. 22–30.
[17] W.-w. Yim, M. Yetisgen-Yildiz, Towards automating medical scribing: Clinic visit dialogue2note sentence alignment and snippet summarization, in: Proceedings of the Second Workshop on Natural Language Processing for Medical Conversations, 2021, pp. 10–20.
[18] A. Joshi, N. Katariya, X. Amatriain, A. Kannan, Dr. summarize: Global summarization of medical dialogue by exploiting local structures, arXiv preprint arXiv:2009.08666 (2020).
[19] H. Li, Y. Wu, V. Schlegel, R. Batista-Navarro, T.-T. Nguyen, A. Ramesh Kashyap, X. Zeng, D. Beck, S. Winkler, G. Nenadic, Pulsar: Pre-training with extracted healthcare terms for summarising patients' problems and data augmentation with black-box large language models, arXiv preprint (2023).
[20] T. L. Scao, A. Fan, C. Akiki, E. Pavlick, S. Ilić, D. Hesslow, R. Castagné, A. S. Luccioni, F. Yvon, M. Gallé, et al., Bloom: A 176b-parameter open-access multilingual language model, arXiv preprint arXiv:2211.05100 (2022).
[21] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, D. Amodei, Language models are few-shot learners, in: NeurIPS, 2020.
[22] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, et al., Training language models to follow instructions with human feedback, in: Advances in Neural Information Processing Systems, volume 35, 2022, pp. 27730–27744.
[23] C.-Y. Lin, Rouge: A package for automatic evaluation of summaries, in: Text summarization branches out, 2004, pp. 74–81.
[24] T. Sellam, D. Das, A. Parikh, Bleurt: Learning robust metrics for text generation, in:

Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020, pp. 7881–7892.
[25] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, Y. Artzi, Bertscore: Evaluating text generation with bert, in: International Conference on Learning Representations, 2020.
[26] M. Joshi, D. Chen, Y. Liu, D. S. Weld, L. Zettlemoyer, O. Levy, Spanbert: Improving pre-training by representing and predicting spans, Trans. Assoc. Comput. Linguistics 8 (2020) 64–77.
[27] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, P. J. Liu, Exploring the limits of transfer learning with a unified text-to-text transformer, J. Mach. Learn. Res. 21 (2020) 140:1–140:67.
[28] L. Soldaini, N. Goharian, Quickumls: a fast, unsupervised approach for medical concept extraction, in: MedIR workshop, sigir, 2016, pp. 1–4.
[29] Ö. Uzuner, B. R. South, S. Shen, S. L. DuVall, 2010 i2b2/va challenge on concepts, assertions, and relations in clinical text, J. Am. Medical Informatics Assoc. 18 (2011) 552–556.
[30] T. Schick, H. Schütze, Generating datasets with pretrained language models, in: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, 2021, pp. 6943–6951.
[31] H. Li, V. Schlegel, R. Batista-Navarro, G. Nenadic, Do you hear the people sing? key point analysis via iterative clustering and abstractive summarisation, arXiv preprint arXiv:2305.16000 (2023).
[32] Z. Yang, S. Wang, B. P. S. Rawat, A. Mitra, H. Yu, Knowledge injected prompt based fine-tuning for multi-label few-shot icd coding, arXiv preprint arXiv:2210.03304 (2022).
[33] H. W. Chung, L. Hou, S. Longpre, B. Zoph, Y. Tay, W. Fedus, E. Li, X. Wang, M. Dehghani, S. Brahma, et al., Scaling instruction-finetuned language models, arXiv preprint arXiv:2210.11416 (2022).
[34] E. Lehman, E. Hernandez, D. Mahajan, J. Wulff, M. J. Smith, Z. Ziegler, D. Nadler, P. Szolovits, A. Johnson, E. Alsentzer, Do we still need clinical language models?, arXiv preprint arXiv:2302.08091 (2023).
[35] T. Dettmers, M. Lewis, Y. Belkada, L. Zettlemoyer, Llm. int8 (): 8-bit matrix multiplication for transformers at scale, arXiv preprint arXiv:2208.07339 (2022).
[36] E. J. Hu, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, W. Chen, et al., Lora: Low-rank adaptation of large language models, in: International Conference on Learning Representations, 2022.
[37] J. Wei, Y. Tay, R. Bommasani, C. Raffel, B. Zoph, S. Borgeaud, D. Yogatama, M. Bosma, D. Zhou, D. Metzler, et al., Emergent abilities of large language models, Transactions on Machine Learning Research 8 (2011).
[38] J. Qiu, L. Li, J. Sun, J. Peng, P. Shi, R. Zhang, Y. Dong, K. Lam, F. P.-W. Lo, B. Xiao, et al., Large ai models in health informatics: Applications, challenges, and the future, arXiv preprint arXiv:2303.11568 (2023).

A. Hyper-parameters for training and inference

We initialise LoRA with $r = 16$, $\alpha = 16$ on the query, key, value and output projection weights of all layers of the base model (Q, K, V and O, respectively). The model is trained on a single A100 80GB GPU with a learning rate of 3^{-4} for the adapter weights. For both encoder-decoder and decoder only settings, during training, we optimise the parameters of the language models to minimise the cross-entropy loss between each token of the prediction and the corresponding token of the ground truth answer sequence using teacher forcing. For encoder-decoder models, we limit the length of input dialogues to at most 496 and the length of output notes to at most 214 tokens, respectively (95th percentile). For the decoder model, we limit the length of input and output combined to at most 696 tokens. During inference, we set no limits to input and output sequence lengths and decode the prediction using beam search with 6 (4 for LLaMa), temperature of 1.0, top k of 50 (40 for LLaMa) and top p of 1.0 (0.7 for LLaMa).

For task C, we use the same arguments as for task C, with the exception of limiting the input length to 2048 and output length to 990 during training, in order to fit the GPU during training.

B. Qualitative example

Figure 1 shows qualitative examples generated by our models trained on task C training data with and without data augmentation, respectively.

Flan-T5-3B	Flan-T5-3B-746DG
Alexander is a 62-year-old male, with a past medical history significant for reflux, who presents for follow-up of his chronic problems. He is so thankful you put him on that medicine for his reflux. The protonix that he had, w- made an amazing change in his life. He's really comfortable now. He eats whatever he wants, and he feels so much better. The doctor is glad to hear that. Okay. How are you doing, kind of, managing your diet? I know, you have to do some lifestyle modifications, like cutting back on caffeine and spicy foods and alcohol. How are you doing with that? *[patient] I'm doing really well. I moved over from caffeine, over to green tea. [doctor] Okay. [patient] and it, it is so, m- it doesn't cause as much problem as it did with, when he was drinking so many energy drinks a day...*	HISTORY OF PRESENT ILLNESS: Alexander is a 62-year-old male with a past medical history significant for reflux. He presents for follow-up of his chronic problems. He has a good support system at home. He has a big family. All his kids call and check on him every day. He does not have any symptoms of chest pain, shortness of breath, belly pain, nausea or vomiting. He does not have any symptoms of nausea or vomiting. On physical examination, I do not hear any carotid bruits in his neck. I do hear a slight 2/6 systolic ejection murmur, which I've heard in the past, so that's stable. His lungs are nice and clear, and he does have 1+ pitting edema bilaterally in his lower extremities. I think he is doing a good job watching his diet. He could just be retaining a little bit of fluid, maybe just from standing all day.

Figure 2: Example excerpts of outputs for Task C produced by models with and without data augmentation. An instance of input copying is highlighted in italics.