EasyChair Preprint

№ 10545

HealthMavericks@MEDIQA-Chat 2023: Benchmarking Different Transformer Based Models for Clinical Dialogue Summarization

Kunal Suri, Saumajit Saha and Atul Singh

July 12, 2023

HealthMavericks@MEDIQA-Chat 2023: Benchmarking different Transformer based models for Clinical Dialogue Summarization

Kunal Suri Saumajit Saha Atul Singh
{suri.kunal007, saha.saumajit, atulsingh.phd}@gmail.com

Abstract

In recent years, we have seen many Transformer based models being created to address Dialog Summarization problem. While there has been a lot of work on understanding how these models stack against each other in summarizing regular conversations such as the ones found in DialogSum dataset, there haven't been many analysis of these models on Clinical Dialog Summarization. In this article, we describe our solution to MEDIQA-Chat 2023 Shared Tasks as part of ACL-ClinicalNLP 2023 workshop which benchmarks some of the popular Transformer Architectures such as BioBart, Flan-T5, DialogLED, and OpenAI GPT3 on the problem of Clinical Dialog Summarization. We analyse their performance on two tasks - summarizing short conversations and long conversations. In addition to this, we also benchmark two popular summarization ensemble methods and report their performance.

1 Introduction

It is essential to summarise the conversation between a doctor and a patient or another doctor to maintain records for compliance, training and evaluation. However this process, at the moment, is done manually which is time consuming and expensive. This paper presents the experimental results of our explorations with state-of-the-art deep-learning techniques to summarise such conversations to accomplish both SubTask A (Ben Abacha et al., 2023b) and B (wai Yim et al., 2023) of Dialogue2Note Summarization task from MEDIQA-Chat 2023 (Ben Abacha et al., 2023a). The solution of SubTask B presented in this paper was ranked fifth among all the submissions for SubTask B. The source code for the submission can be found in GitHub[1].

The paper uses Transformer based models for both assigning conversations into a pre-defined set of clinical notes sections and summarization of conversations. Through this work, the paper also compares the performance of Transformer based models for summarization tasks. This paper benchmarks performance of several Transformer based model for summarization task on medical conversation documents. In addition to this comparison, we also evaluate performance of two ensemble techniques namely (Kobayashi, 2018) and (Chen et al., 2021). Our simulations show that finetuning of Transformer-based models works as well as in-context prompt-based finetuning of OpenAI GPT3 which has usage-based costs and the risk of compromising your internal data to an external organization.

This paper is organized as follows. Section 3 presents a brief overview of the Dialogue2Note Summarization task, including the labeled data available and the evaluation metrics. Then the paper describes current state-of-the-art for clinical note summarization in Section 2 that this paper build upon. This is followed by the description of the approach used to solve the SubTask A of Dialogue2Note Summarization task in Section 4 and SubTask B in Section 5. Then the results of our solutions for Dialogue2Note Summarization tasks are presented. Finally, the paper ends with a conclusion on the work. The paper includes an appendix containing exploratory data analysis and material that will help to better understand the solution presented in the paper.

2 Related Work

In (Zhang et al., 2021) the authors have used both a single-stage and a two-stage approach for summarization. In the single-stage approach, the authors have truncated the input sentence length to match the BART transformer model input length constraints. In the multi-stage approach, the authors summarize the input conversation and then pass these summaries through a secondary model to

[1] https://github.com/suri-kunal/acl-medi-chat-summarization

generate the final summary. The authors have only focused on summarizing the History of Present Illness (HPI) section. The current work presented in this paper extends it beyond one section. It uses a two-step approach to generate the summary but does not focus on any special processing of the data in each section to generate the summaries.

In (Krishna et al., 2021) the authors combine extractive and abstractive summarization. They have presented a wide range of algorithms. Cluster2Sent, the most elaborate among these algorithms, first identifies the noteworthy utterances in each section and then clusters them before sending them to a summarization model. The present work presented in this paper is similar to this approach in that it does a Section level summarization. The current work depends on the power of more powerful models to summarise instead of processing the text in the Sections.

In (Chintagunta et al., 2021), the authors use GPT3 for medical summarization to achieve summaries that match human annotator-generated summaries using 30x lesser data. The authors generate k-candidate summaries for an input dialogue in this work. For each candidate summary generation, the authors sample N random examples from a small labelled data set. The examples, along with the input dialogue, are sent to GPT3 for summarization. In this work, the authors select N examples for each Section. The authors have yet to identify the medical terms in the generated summary to measure its effectiveness, which could be future work.

3 Dialogue2Note Summarization Task Description

This Section provides a high-level overview of the Dialogue2Note Summarization task (including both SubTask A and B) from MEDIQA-Chat 2023[2]. The Section starts with a description of the SubTask task goals, followed by basic counts of the available labelled data. The metric used to evaluate this task is arithmetic mean of ROUGE-1 (Lin, 2004), Bertscore F1 (Zhang et al., 2019), and BLEURT (Sellam et al., 2020).

3.1 Task Definition

Given a short conversation between a Doctor and a patient or another Doctor (**Dialogue**), the goal of SubTask A is to create a system that automatically predicts the Section to which the conversation belongs to which is denoted by **Section Header**. There are twenty Sections Headers in this dataset. Some examples of Section Headers are FAM/SOCHX, GENHX, PASTMEDICALHX, CC. All of these Section Headers and their descriptions (**Section Description**) can be found in Table A2. Another part of this SubTask is to generate a summary which matches the human generated summary (**Section Text**) as closely as possible while optimizing the metric for evaluation.

The aim of SubTask B is to summarize a given Doctor-Patient conversation (**Dialogue**) in a way that the generated summary matches the clinical note written by the physician (**Note**) as closely as possible. Unlike SubTask A, this task is a lot harder to solve because average length of a conversation is significantly longer than the dialogue of SubTask A. Please refer to Figure A1 for data distribution of SubTask A and Figure A2 for Dialogue data of SubTask B to understand the difference in distribution. A clinical note consists of the following high level sections called *First Level Sections* in this paper - *Subjective*, *Objective Exam*, *Objective Result*, *Assessment and Plan*. A clinical note comprises of several Section Headers each of which can be allocated to one of the First Level sections. Given a conversation between a Doctor and a patient, we create a system that automatically generates complete clinical note with all necessary First Level Sections.

3.2 Labelled Data

In this paper we have used the labelled data provided by MEDIQA-Chat 2023 organizers for training the models. A sample data point from the labelled data set for SubTask A can be found in Table A1. An example of a Doctor-Patient Conversation and corresponding Clinical Notes generated by a human from the labelled data set for SubTask B is also shown in Figure A2. The official data consists of a training and validation split. For SubTask A, the training data contains 1201 and validation data contains 180 <dialogue, section-text, section-header> triplets. For SubTask B, the training data contains 67 and validation data contains 20 <dialogue, note> pairs.

4 SubTask A Methodology

Given a short conversation between a doctor and a patient, the goal of SubTask A is to predict its Sec-

[2] https://sites.google.com/view/mediqa2023/clinicalnlp-mediqa-chat-2023

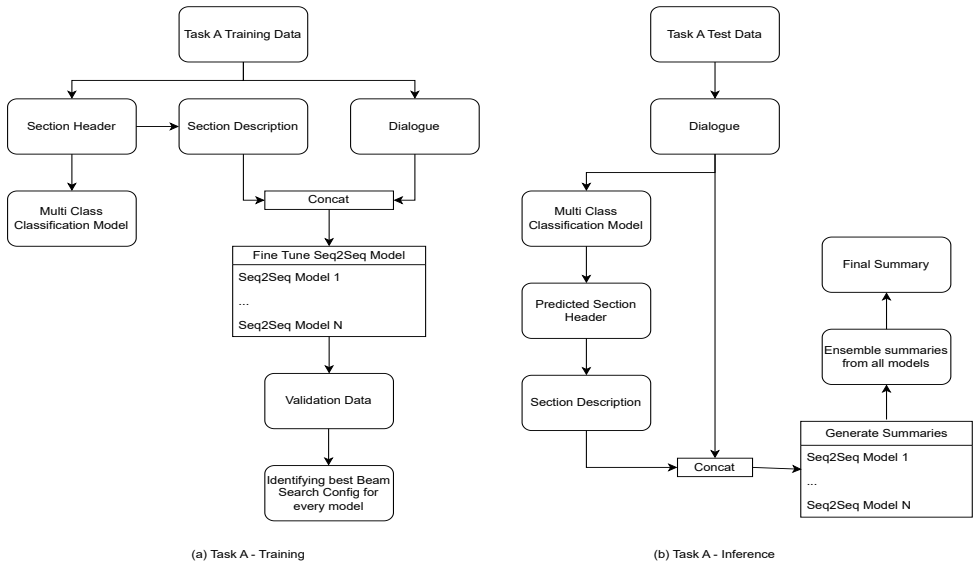


Figure 1: SubTask A - Overall Architecture

tion Header and summarize it while ensuring that the generated summary is as fluent and as close to Section Text as possible. This Section starts with a description of the approach used to predict the Section Header. This is followed by a description of the methodology used to summarize the conversation. For Dialogue Summarization, we have fine-tuned Transformer-based large language models. We have done an in-context fine-tuning of OpenAI GPT3 (Brown et al., 2020) and have fine-tuned four popular Transformer Sequence-to-Sequence (*Seq2Seq*) models. The Section describes the processed labelled data used for fine-tuning the Transformer based models, followed by the actual training steps. Then this Section looks at the steps used to generate the summary from the decoder. Finally, we discuss the two approaches used for ensembling the output of the four Transformer based models.

We achieved success using Bio-ClinicalBERT (Alsentzer et al., 2019) for classification in the healthcare domain and hence have fine-tuned this model for the classification of Dialogue to a Section Header in SubTask A. Since the target variable is highly imbalanced (see Figure A1a), we use Focal Loss (Lin et al., 2017) so that the algorithm focuses more on classes with fewer samples. We limit the number of input tokens to 300 tokens because that is the length of majority of dialogues, as shown in Figure A1. As the number of data samples available for training and validation is less, we use a 5 Fold Cross Validation approach for modelling purposes to ensure that we can capture all the information in the data. The hyper-parameters used for training and performance for all folds can be found in Table A3. During inference, we pass a given Dialogue through all five models, take an average of the logits for all the classes and output the class with the highest logit score.

We fine-tune Seq2Seq models using the labelled data (Dialogue,Section Text) for SubTask A as the (Input,Output) pair. Section Text is a part of the labelled data and is a human subject matter expert-created summary of Dialogue. As a pre-processing step, we replace all new line characters with whitespaces. The Dialogue is concatenated with the section description of its Section Header with the SEP token of the Seq2Seq architecture. While training, we use the actual section description for the actual Section Header and at inference, we use the section description corresponding to the predicted Section Header for the given Dialogue. No changes are made to Section Text.

We use a 5-fold cross validation scheme and fine-tune four Seq2Seq models - BioBart (Yuan et al., 2022), Flan-T5-Large (Chung et al., 2022), DialogLED-Base, and DialogLED-Large (Zhong et al., 2022) on each of the folds. Here we need to select the number of input tokens for encoder and decoder. For encoder we have selected token length of 512 tokens and for decoder we have selected token length of 400 tokens. All the hyper-parameters used to train each of the above architecture can be found in Table A4. To select the best model, we use early-stopping based on Validation Negative Log Loss (Yao et al., 2007). The out-of-fold results

can be found in Table A6. The distribution of tokens for Dialogue and Section Text can be found in Figure A1b and Figure A1c respectively.

To generate summaries that match the human generated summaries, we need a way to control the summary generated by the decoder component of a Seq2Seq model. This can be done by using decoding strategies such as Beam Search (Graves, 2012), Top-k Sampling (Fan et al., 2018), Top-p Sampling (Holtzman et al., 2019), Contrastive Search (Su and Collier, 2023) etc. In this module, we use Beam Search with TPESampler Algorithm from Optuna[3] to search for the optimal decoding strategy trying to maximize ROUGE-1, ROUGE-2, and BertScore rather than relying on manual tweaking of these metrics. We use TPESampler here because it supports multivariate optimization and also it handles Float, Integer, and Categorical values better than other algorithms present in Optuna[4]. We use Optuna here due to ease of implementing Hyper-parameter optimization algorithms. We did not use BLEURT during search because it is extremely time consuming. For this module, we use four hyper-parameters for Beam Search - Early Stopping, Number of Beams, No Repeat N-gram Size, Length Penalty. The search space of each of these variables can be found in the Table 1.

Variable	Data Type	Range
Early Stopping	Categorical	[True,False]
Num_Beams	Integer	5-15
No_Rep_N_Size	Integer	5-15
Len_Pen	Float	[-2,2]

Table 1: Search Space for Beam Search Decoding. Num_Beams : Number of Beams, No_Rep_N_Size : No Repeat Ngram Size, Len_Pen : Length Penalty.

The approach used for in-context finetuning using OpenAI GPT3 is as follows: For every dialog in the test set, we predict and store the Section Header. We, then, randomly pick 3 Dialog-Summary-Section Header triplet from the entire (Training + Validation) dataset with the same Section Header. We use these triplets to create three summaries. These three summaries are then merged together to get the final summary. The configuration used for this task can be found in the appendix in Table A5 and its result on the test set can be found in Table 3 against Run 3.

In this paper We have used following approaches for ensembling:

- **Generating Best Summary by semantic similarity** - We use a post-ensemble method (Kobayashi, 2018) to identify the summary which is closest to all the generated summaries. This summary is then considered to be final summary for the given Dialogue.
- **Generating Best Summary by minimizing hallucination** - The above methodology helps us to get the summary closest to all the summaries but it does not account for the faithfulness of the generated summary with the actual Dialogue. To answer this question, we use the techniques introduced in (Chen et al., 2021). They have released a model[5] which we are using out of the box.

5 SubTask B Methodology

This Section presents an end-to-end solution to convert an entire Doctor-Patient Conversation (Dialogue) to Clinical Notes as SubTask B requires. The Section starts with a description of the supervised machine learning model used to predict the Section Header to which every utterance in a conversation belongs. All of these Section Headers are mapped to the First Level Sections using the mapping in Table 2. The output clinical note will contain these First Level Sections. The description of the classification model is followed by a description of the approach used to concatenate the utterances in a Dialogue belonging to a specific First Level Section. This is followed by a description of the Transformer models used to summarise the concatenated utterances. The results from the Transformer models are passed through an ensemble technique similar to the technique proposed in Section 4 to select the final summary, which is placed in the identified First Level Section in the Clinical Note.

We train a multi-label Classifier using the Dialogue and Section Header data from SubTask A to predict the Section Header to which an utterance belongs. As the data volume is very low, we use iterative-stratification package[6] to create 5 Folds

[3] https://optuna.readthedocs.io/en/stable/reference/samplers/generated/optuna.samplers.TPESampler.html

[4] https://optuna.readthedocs.io/en/stable/reference/samplers/index.html

[5] https://github.com/CogComp/faithful_summarization

[6] https://github.com/trent-b/iterative-stratification

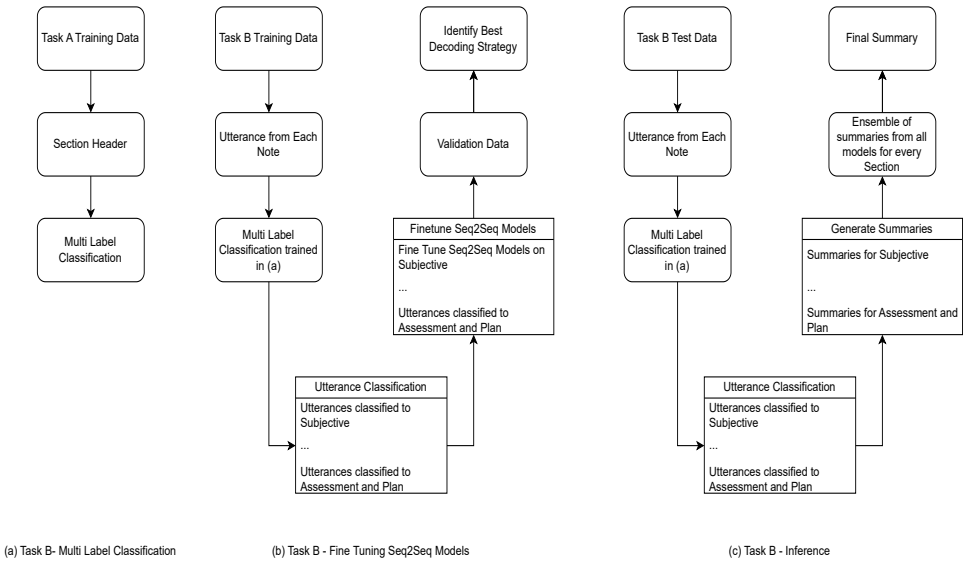


Figure 2: SubTask B - Overall Architecture

of the data and train one model for each fold - thus capturing all information in the data. Some of the labels have very low presence (see the distribution in Figure A1a), and hence we use Focal Loss instead of Binary Cross Entropy Loss so that the model can focus on labels which are harder to classify. The base model used for fine tuning is Bio-ClinicalBERT (Alsentzer et al., 2019) with number of input tokens as 512. We use early stopping to select the best model using Validation Negative Log Loss as the criteria for selecting the best model and Precision-Recall (PR) Score to evaluate performance of this model. The hyper-parameters for all folds and PR Score for all folds for all Section Headers can be found in Table A7 and Table A8 respectively.

We split every conversation on a new line character (\n) to get the list of constituent utterances. Each utterance is passed through each of the 5 Multi Label Models and we create a union of all predicted Section Headers from every model. Once every utterance has been mapped to all possible Section Headers, we transform the mapping so that we can combine all the utterances that belong to same Section Header. We ensure that utterance order should remain intact in all the sections. We then map all these Section Headers to their First Level Sections using mapping in Table 2. We have kept the mapping exhaustive to ensure that no False Negatives are left out. After this mapping, we merge all the utterances together and concat them together using whitespace character. We then split these utterances into their respective First Level Section and use the script provided by the organizers[7] to split the Note into these First Level Sections as well. The samples of dataset created after this step can be seen in the Figues A3a, A4a, A5a, and A6a. We have used the same high-level approach as in SubTask A for Dialogue Summarization. We fine-tune transformer based models, and have also used OpenAI GPT3 (Brown et al., 2020) with prompt based fine-tuning for summary generation.

We fine tune Seq2Seq models using (Utterance, Clinical Note Section) generated above as the (Input, Output) pair for every First Level Section. Before feeding the Utterance to the Encoder-Decoder models, we concatenate it with the section description of the First Level Section that the utterance belongs to, using the SEP token of the transformer architecture. We train two Seq2Seq models - DialogLED-Base and DialogLED-Large (Zhong et al., 2022) for each of the First Level Sections for each of the folds. The distribution of tokens for utterances of each First Level Sections and corresponding part of Clinical Note can be found in the Figures A3b, A3c, A4b, A4c, A5b, A5c, A6b, and A6c. All the hyper-parameters used to train each of the above architecture can also be found in Table A9.

Apart from finetuning Transformers we have also used OpenAI GPT3 (Brown et al., 2020) to generate summaries using prompt engineering. For every dialog in the test set, we pass it through Section 5 to split the Dialogue into utterances for every

[7] https://github.com/abachaa/MEDIQA-Chat-2023

First Level Section	Section Headers
Subjective	CC FAM/SOCHX GENHX PASTMEDICALHX PASTSURGICAL GYNHX OTHER_HISTORY ALLERGY ROS MEDICATIONS IMMUNIZATIONS
Objective_Exam	EXAM IMAGING LABS PROCEDURES
Objective_Results	IMAGING LABS DIAGNOSIS
Assessment_and_Plan	ASSESSMENT PLAN DISPOSITION PROCEDURES LABS MEDICATIONS EDSOURCE

Table 2: Mapping of First Level Section to Section Headers

First Level Section. We randomly pick 3 Dialog-Summary Pair for every First Level Section from the training data and truncate Dialog to 750 Tokens and Summary to number of tokens as per the First Level Section it belonged to. The number of tokens for summary section of each First Level Section can be found in Table A9. As for test dialog, we truncate it to 1000 tokens. The reason we do this is to adhere to the 4000 tokens length constraint of OpenAI GPT3 API. We concat the Train Dialog and Summary along with Test Dialog and generate a summary. This step is repeated three times. These three summaries are then merged together to get the final summary. We use the configuration in Table A5 for summarization.

To select the best model, we use early-stopping (Yao et al., 2007) using Validation Negative Log Loss and the metrics for best model for each architecture for each fold can be found in Table A3. We use the same search strategy for the optimal decoding strategy as we did for SubTask A except for one difference - we apply these techniques on the summaries generated for each First Level Sections separately. We use the same Model Ensembling Techniques that we have used in SubTask A except for one difference - we apply these techniques on the summaries generated for each First Level Sections separately.

6 SubTask A Results

This Section presents the results for SubTask A using the approach described in Section 4. We have made three submissions (mentioned as *runs* in the result tables) for predicting Section Header and three submissions for generating summaries from Dialogues. All the submissions for classification and final rankings of each of the runs can be found in Table A12. For the summarization task, we have also submitted results from three *runs*. In *run 1* and *run 2*, we have done the finetuning of the Transformer based models mentioned in Section 4 while *run 3* presents the results of summarization using OpenAI GPT3. The details for each run are as follows:

1. Run 1 - Post the summary generation, we ensemble output of all the models using *Generating Best Summary by minimizing hallucination* technique.

2. Run 2 - Post the summary generation, we ensemble output of all the models using *Generating Best Summary by semantic similarity*.

3. Run 3 - We use the an OpenAI GPT3 based approach described in Section 4.

The table containing our team's standing can be found in the Tables A12 and A13. Standings of all the teams have been calculated by calculating multi class accuracy for Section Header Classification and arithmetic mean of Rouge-1, Bertscore, BLEURT for the Dialogue summary.

The experiments show that Run1, which performs the worse on the ranking, hallucinates the most. This is counterintuitive since the goal of this approach is to minimize hallucination. We hypothesize that this could be because the model to detect hallucination was not trained on clinical data. Run2 gives the best summaries as measured by the evaluation metrics but has some hallucinations. Run3 gives the results with little to no hallucinations but has a lower score than Run2. This can happen

because ROUGE Score always favours more extended conversations over shorter ones (Schluter, 2017). This can also be seen in Table 3 where BertScore and BLEURT are better for Run3 than for Run2 whereas ROUGE Score is better for Run2 than Run3. All the summaries generated by Run1, Run2, and Run3 are available in the github repository for the interested audience.

Run	R-1	B-F1	BLEURT	MS
Run 2	0.2973	0.612	0.4956	0.4683
Run 3	0.2514	0.6268	0.5015	0.4599
Run 1	0.1987	0.5703	0.4298	0.3996

Table 3: Results of runs on Test Data. R-1: ROUGE-1, B-F1: Bertscore-F1, MS: Mean Score

6.1 Analysis of different Transformer Architectures on the data

We analyse the performance of each Transformer architectures i.e. BioBart-V2-Base, Flan-T5-Large, DialogLED-Large, DialogLED-Base, and OpenAI GPT3 on the given dataset. We find that pretrained language Models such as DialogLED-Large, DialogLED-Base have performed consistently better than large language models such as OpenAI GPT3 and Flan-T5-Large. The performance was evaluated by calculating arithmetic mean of ROUGE-1, ROUGE-2, and BertScore-F1. We do not use BLEURT here as it is extremely time consuming and based on our observations, ROUGE-2 and BLEURT have a very strong correlation. The average score across all 5 folds for each architecture can be found in the Table A6.

7 SubTask B Results

As mentioned in Section 5, we have used a two step process using first a multi-label classification model to assign a conversation to a section and then applying Transformer based models on conversations for a section. Just like in SubTask A, we have made three submissions for generating summaries from Conversations. For Run 1 and Run 3, we map the utterances into first level sections, followed by summary generation for every first level Section. The decoding strategy for each first level section for each model can be found in Table A10. In this task, we tried using beam search configuration generated from Section 4 as well but this did not work well as we were getting Out Of Memory (OOM) errors. The summaries generated by this process is used below.

1. Run 1 - Post the summary generation, we ensemble output of all the models for every First Level Section using *Generating Best Summary by minimizing hallucination* technique.

2. Run 2 - We use the approach of OpenAI GPT3 from Section 5. We don't go in-depth for this approach since we couldn't analyse it due to cost constraints.

3. Run 3 - Post the summary generation, we ensemble output of all the models for every First Level Section using *Generating Best Summary by semantic similarity*

Our team's standing in the task of summarizing full note can seen in Table A14. It has been calculated by calculating the ROUGE-1 of the full note summary. Our team's standing in the task of summarizing complete note for all First Level Section can be found in Table A18. It has been calculated by calculating the arithmetic mean of arithmetic means of Rouge-1, Bertscore, BLEURT of every First Level Section Summary.

We have analysed the pros and cons of these three runs. Run1 gives the best result on this task but has some hallucinations. Run2 gives the results with little to no hallucinations but it has a lower score than Run1 and Run3. This can be because of the information loss that has happened as we are not able to consider all the tokens in the prompt. While Run3 performs better than Run2, it still has the most hallucination. This can be seen in Table 4.

Run	ROUGE-1
Run 1	0.5311
Run 3	0.5111
Run 2	0.2759

Table 4: Results of runs on Test Data for SubTask B

7.1 Analysis of different Transformer Architectures on the data

In SubTask B we analyse the performance of two Transformer based models namely DialogLED-Large, and DialogLED-Base. We are unable to compare their performance with OpenAI GPT3 as we have done in SubTask A because of cost and funding issues. We find that DialogLED-Large

performs consistently better than DialogLED-Base. The performance is evaluated by calculating arithmetic mean of ROUGE-1, ROUGE-2, and BertScore-F1. The average score across all 5 folds for each architecture can be found in the Table 5. The performance of every model of every fold can be found in Table A11.

Arch	R-1	R-2	B-F1	MS
D-LED-B	0.5036	0.2257	0.6324	0.4539
D-LED-L	0.5235	0.2346	0.6388	0.4656

Table 5: SubTask B - Performance of different Transformer Architectures. Arch : Architecture, D-LED-B : Dialog-LED-Base, D-LED-L : Dialog-LED-Large, R-1 : ROUGE-1, R-2: ROUGE-2, B-F1 : Bertscore-F1, MS : Mean Score

8 Conclusion

The paper presents the solution and the results for SubTask A and B of Dialogue2Note Summarization task. The solution uses Transformer based models for both classification and summarization of Clinical Dialogs and the paper presents the comparison of the performance of Transformer based models on summarization. To the best of our knowledge, this is the first paper that benchmarks the performance of these models on Clinical Dialog Summarization. Our simulations and test scores show that fine-tuned transformer models work as well as in-context prompt based fine-tuning of Large Language Models such as OpenAI GPT3. This is encouraging for groups which either cannot afford huge API costs of using these Large Language Models or cannot send their data to their API due to regulatory restrictions. In addition to this, we also observe that metrics such as ROUGE might not be the suitable to gauge performance of models like OpenAI GPT3 as they focus on syntactic similarity. Metrics such as Bertscore and BLEURT seem to be more suitable for such models since they focus on semantic similarity. The paper also evaluated two different ensemble techniques and the results demonstrate that the Post Ensemble technique performs the best while also giving minimum hallucinations.

References

Emily Alsentzer, John R. Murphy, Willie Boag, Wei-Hung Weng, Di Jin, Tristan Naumann, and Matthew B. A. McDermott. 2019. Publicly available clinical BERT embeddings. *CoRR*, abs/1904.03323.

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 B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. *CoRR*, abs/2005.14165.

Sihao Chen, Fan Zhang, Kazoo Sone, and Dan Roth. 2021. Improving faithfulness in abstractive summarization with contrast candidate generation and selection. In *North American Chapter of the 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 *Proceedings of the Second Workshop on Natural Language Processing for Medical Conversations*, pages 66–76, Online. Association for Computational Linguistics.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022. Scaling instruction-finetuned language models.

Angela Fan, Mike Lewis, and Yann Dauphin. 2018. Hierarchical neural story generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 889–898, Melbourne, Australia. Association for Computational Linguistics.

Alex Graves. 2012. Sequence transduction with recurrent neural networks. *CoRR*, abs/1211.3711.

Ari Holtzman, Jan Buys, Maxwell Forbes, and Yejin Choi. 2019. The curious case of neural text degeneration. *CoRR*, abs/1904.09751.

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, Brussels, Belgium. Association for Computational Linguistics.

Kundan Krishna, Sopan Khosla, Jeffrey Bigham, and Zachary C. Lipton. 2021. Generating SOAP notes from doctor-patient conversations using modular summarization techniques. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4958–4972, Online. Association for Computational Linguistics.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Tsung-Yi Lin, Priya Goyal, Ross B. Girshick, Kaiming He, and Piotr Dollár. 2017. Focal loss for dense object detection. *CoRR*, abs/1708.02002.

Natalie Schluter. 2017. The limits of automatic summarisation according to ROUGE. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 41–45, Valencia, Spain. Association for Computational Linguistics.

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

Yixuan Su and Nigel Collier. 2023. Contrastive search is what you need for neural text generation.

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*.

Yuan Yao, Lorenzo Rosasco, and Andrea Caponnetto. 2007. On early stopping in gradient descent learning. *Constructive Approximation*, 26(2):289–315.

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

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. *CoRR*, abs/1904.09675.

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

A Appendix

A.1 Data Exploration and Explanation

This section discusses data exploration and explanation so that audience can understand why we made the decisions that we made.

A.1.1 SubTask A

A sample data point from dataset for SubTask A can be seen in Table A1.

Variable	Sample Value
Section Header	FAM/SOCHX
Section Text	The patient has been a smoker since the age of 10. So, he was smoking 2-3 packs per day. Since being started on Chantix, he says he has cut it down to half a pack per day. He does not abuse alcohol
Dialogue	Doctor: Are you a smoker? Patient: Yes. I do not drink if that is any constellation. Doctor: How much do you smoke per day? Patient: I just started taking Chantix and now I am down to a half a pack a day. Doctor: How much did you smoke per day prior to starting Chantix? Patient: I was smoking about two to three packs a day. I have been smoker since I was ten years old.

Table A1: Sample data point for SubTask A

The description of each of the Section Headers present in the data can be found in Table A2

The Data Exploration of this SubTask is give by Figure A1

The hyper-parameters and performance metrics for Predicting Section Header for every fold can be found in the Table A3. Each of the below configuration was run on Bio-ClinicalBert with Focal Loss.

The hyperparameters used to fine tune Seq2Seq Models can be found in Table A4. We run each of these models for 30 epochs with AdamW Optimizer, Learning Rate of 0.00002, and Linear Learning Scheduler.

Section Header	Section Header Description
FAM/SOCHX	FAMILY HISTORY/SOCIAL HISTORY
GENHX	HISTORY OF PRESENT ILLNESS
PASTMEDICALHX	PAST MEDICAL HISTORY
CC	CHIEF COMPLAINT
PASTSURGICAL	PAST SURGICAL HISTORY
ALLERGY	ALLERGY
ROS	REVIEW OF SYSTEMS
MEDICATIONS	MEDICATIONS
ASSESSMENT	ASSESSMENT
EXAM	EXAM
DIAGNOSIS	DIAGNOSIS
DISPOSITION	DISPOSITION
PLAN	PLAN
EDCOURSE	EMERGENCY DEPARTMENT COURSE
IMMUNIZATIONS	IMMUNIZATIONS
IMAGING	IMAGING
GYNHX	GYNECOLOGIC HISTORY
PROCEDURES	PROCEDURES
OTHER_HISTORY	OTHER_HISTORY
LABS	LABS

Table A2: Section Headers and their descriptions.

The configuration used for OpenAI GPT3 can be found in Table A5.

The average performance of these Seq2Seq Models for SubTask A can be found in Table A6. Here we didn't use BLEURT because it is a very time-consuming operation and based on our observations, ROUGE-2 is very well correlated with BLEURT.

A.1.2 SubTask B

The sample data and token distribution of this task is give by Figure A2

The sample data and token distribution of Subjective Section is give by Figure A3

The sample data and token distribution of Objective Exam Section is give by Figure A4

The sample data and token distribution of Objective Result Section is give by Figure A5

Fold	eps	WD	LR	WR	BS	Epochs	Seed	BE	BVA	BVL
4	1.00E-06	0.01	2.00E-05	0.1	16	30	42	18	0.8231	0.3796
3	1.00E-06	0.01	2.00E-05	0.1	16	30	42	18	0.7385	0.3688
2	1.00E-06	0.01	2.00E-05	0.1	16	30	42	18	0.8038	0.2131
1	1.00E-06	0.01	2.00E-05	0.1	16	30	42	18	0.7615	0.4397
0	1.00E-06	0.01	2.00E-05	0.1	16	30	42	18	0.7816	0.3912

Table A3: SubTask A - Predicting Section Header. eps : AdamW_eps, WD : AdamW Weight Decay, LR : Learning Rate, WR : Warmup Ratio, BS : Batch Size, BE : Best Epoch, BVA : Best Validation Accuracy, BVL : Best Validation Loss.

Architecture	GAS	BS	MaxSL	MaxTL	MinTL
Flan-T5-Large	3	5	512	400	8
Biobart-V2-Base	1	16	512	400	8
DialogLED-Large	3	6	512	400	8
DialogLED-Base	1	16	512	400	8

Table A4: SubTask A - Hyperparameter Tuning for Different Architectures. Optim : Optimizer, LR : Learning Rate, Sched : Scheduler, GAS : Gradient Accumulation Steps, BS : Batch Size, MaxSL : Maximum Source Length, MaxTL : Maximum Target Length, MinTL : Minimum Target Length

Hyperparameter	Value
Model	text-davinci-003
Temperature	0.5
Max Tokens	400
Top_p	1.
Frequency Penalty	0
Presence Penalty	0

Table A5: OpenAI GPT3 Hyperparameters

The sample data and token distribution of Assessment and Plan Section is give by Figure A6

The hyper-parameter setting for creating Multi Label Classification output can be seen in Table A7

The Precision Recall Score averaged over all the folds for all Section Headers can be found in the Table A8

We use the configuration in Table A9 for SubTask B summarization. Each of these models was trained for 30 Epochs with a Learning Rate of 0.00002, and a Linear Learning Scheduler.

The decoding strategy for SubTask B Summarization can be found in Table A10.

The performance of every architecture on every fold for SubTask B Summarization can be found in Table A9.

Model-Arch	R1	R2	BS-F1	MS
DL-Base	0.2471	0.0936	0.5803	0.3070
DL-Large	0.2444	0.0998	0.5741	0.3061
OpenAI GPT3	0.2233	0.0700	0.5917	0.2950
BBart-Base	0.1978	0.0767	0.5887	0.2877
FT5-Large	0.0589	0.0200	0.2458	0.1083

Table A6: SubTask A - Performance of different Transformer Architectures. Model-Arch - Model Architecture, R1 - Rouge-1, R2 - Rouge-2, BS-F1 - Bertscore-F1, MS - Mean Score, DL-Base - DialogLED-Base, DL-Large - DialogLED-Large, BBart-Base - BioBart-V2-Base, FT5-Large - Flan-T5-Large

Fold	4	3	2	1	0
AdamW_eps	0.000001	0.000001	0.000001	0.000001	0.000001
AdamW_weight_decay	0.01	0.01	0.01	0.01	0.01
batch_size	16	16	16	16	16
epochs	30	30	30	30	30
lr	0.00002	0.00002	0.00002	0.00002	0.00002
seed	42	42	42	42	42
warm_up_steps	0.1	0.1	0.1	0.1	0.1

Table A7: SubTask B - Hyperparameters used in Multi Label Classification

Model	Bio-ClinicalBERT
Section Header	**PR Score**
ALLERGY	95.77%
ASSESSMENT	34.58%
CC	55.31%
DIAGNOSIS	19.97%
DISPOSITION	71.10%
EDCOURSE	8.60%
EXAM	59.55%
FAM/SOCHX	97.16%
GENHX	90.18%
GYNHX	5.16%
IMAGING	40.61%
IMMUNIZATIONS	90.27%
LABS	25.00%
MEDICATIONS	94.87%
OTHER_HISTORY	0.43%
PASTMEDICALHX	75.84%
PASTSURGICAL	86.47%
PLAN	52.24%
PROCEDURES	2.05%
ROS	79.25%

Table A8: SubTask B - Precision Recall Scores for every Section Header

Architecture	FLS	BS	GAS	MaxSL	MinTL	MaxTL
Dial-LED-B	AP	8	2	3400	640	50
Dial-LED-L	AP	4	4	3400	640	50
Dial-LED-B	OE	8	2	3400	640	50
Dial-LED-L	OE	4	4	3400	640	50
Dial-LED-B	OR	8	2	3400	640	50
Dial-LED-L	OR	4	4	3400	640	50
Dial-LED-B	Subjective	8	2	3400	640	50
Dial-LED-L	Subjective	4	4	3400	640	50

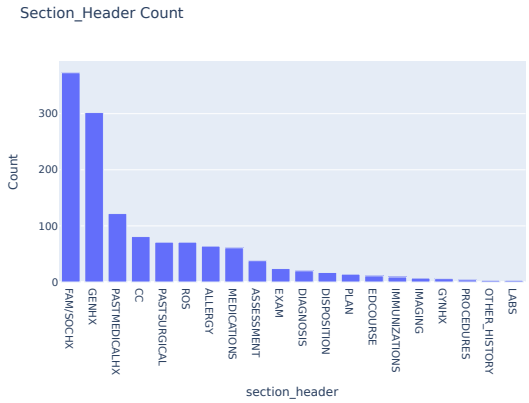
Table A9: SubTask B - Hyperparameters for every architecture for every First Level Section. Dial-LED-B : Dialog-LED-Base, Dial-LED-L : Dialog-LED-Large, FLS : First Level Section, BS : Batch Size, GAS : Gradient Accumulation Steps, LR : Learning Rate, MaxSL : Maximum Source Length, MinTL : Minimum Target Length, MaxTL : Maximum Target Length, AP : Assessment And Plan, OE : Objective Exam, OR : Objective Results

First Level Section	Architecture	# Beams	Early Stopping	Length Penalty	No Repeat Ngram Size
Objective Exam	DialogLED-Base	5	TRUE	0.2	2
Objective Exam	DialogLED-Large	5	TRUE	0.2	2
Objective Result	DialogLED-Base	5	TRUE	0.2	2
Objective Result	DialogLED-Large	5	TRUE	0.2	2
Subjective	DialogLED-Base	5	TRUE	0.2	2
Subjective	DialogLED-Large	5	TRUE	0.2	2
Assessment and Plan	DialogLED-Base	5	TRUE	0.2	2
Assessment and Plan	DialogLED-Large	5	TRUE	0.2	2

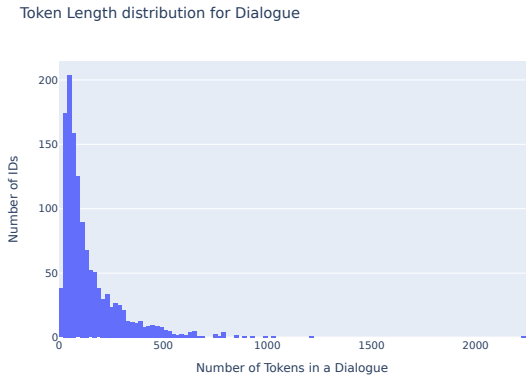
Table A10: SubTask B - Decoding Strategy

Architecture	Fold	Rouge1	Rouge2	BertScore-F1
DialogLED-Base	0	0.5189	0.2427	0.6415
DialogLED-Base	1	0.4854	0.2166	0.6231
DialogLED-Base	2	0.5345	0.2556	0.6497
DialogLED-Base	3	0.4832	0.1990	0.6198
DialogLED-Base	4	0.4958	0.2145	0.6281
DialogLED-Large	0	0.5109	0.2395	0.6467
DialogLED-Large	1	0.5459	0.2493	0.6362
DialogLED-Large	2	0.5569	0.2560	0.6530
DialogLED-Large	3	0.4917	0.2068	0.6269
DialogLED-Large	4	0.5122	0.2216	0.6310

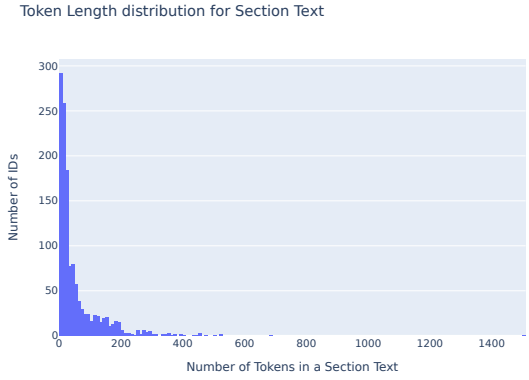
Table A11: SubTask B - Performance of every model of every fold



(a) Class distribution of Section Headers



(b) Dialogue Token Distribution

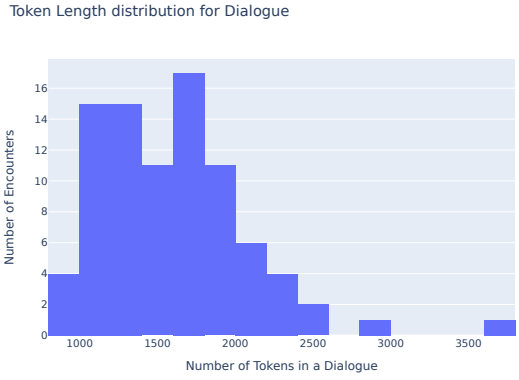


(c) Clinical Note Token Distribution

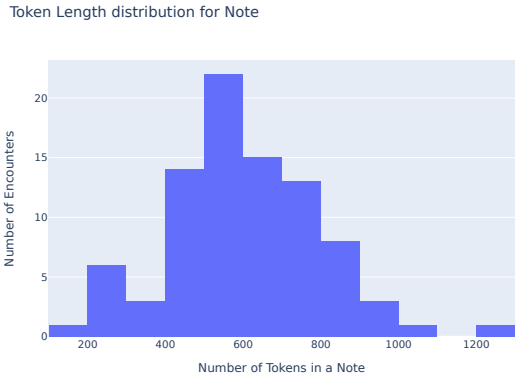
Figure A1: SubTask A - Data Exploration

Dialogue	Note
[doctor] hi , martha . how are you? [patient] i'm doing okay . how are you? [doctor] i'm doing okay . so , i know the nurse told you about dax . i'd like to tell dax a little bit about you , okay ? [patient] okay . [doctor] martha is a 50-year-old female with a past medical history significant for congestive heart failure , depression and hypertension who presents for her annual exam . so , martha , it's been a year since i've seen you . how are you doing ? ... [doctor], some time during the day to take them , okay ? [patient] that might help me remember better . [doctor] all right . that sounds good . all right , well , it's good to see you . [patient] good seeing you too . [doctor] hey , dragon , finalize the note .	CHIEF COMPLAINT Annual exam. HISTORY OF PRESENT ILLNESS REVIEW OF SYSTEMS ... ASSESSMENT AND PLAN Martha Collins is a 50-year-old female with a past medical history significant for congestive heart failure, depression, and hypertension who presents for her annual exam. Congestive heart failure. Depression. Hypertension.

(a) Sample data



(b) Dialogue Token Distribution

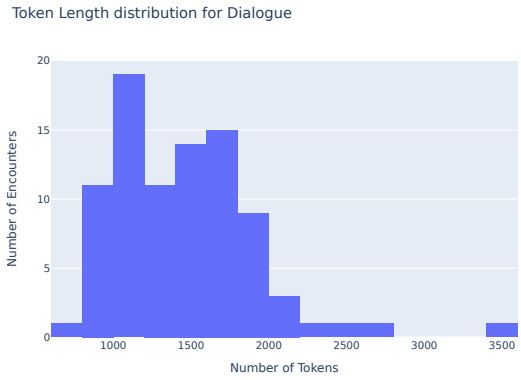


(c) Clinical Note Token Distribution

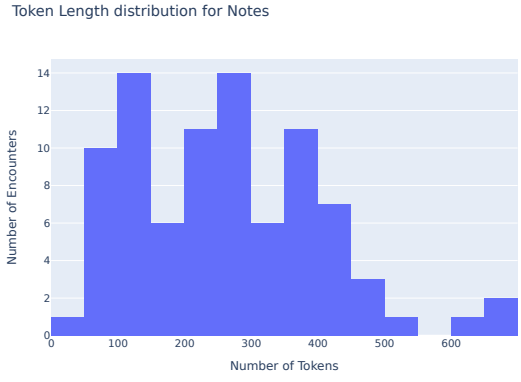
Figure A2: SubTask B - Sample Data

reference_subjective	dialogue_subjective
CHIEF COMPLAINT Annual exam. HISTORY OF PRESENT ILLNESS Martha Collins is a 50-year-old female with a past medical history significant for congestive heart failure, depression, and hypertension who presents for her annual exam. REVIEW OF SYSTEMS Ears, Nose, Mouth and Throat: Endorses nasal congestion from allergies. Cardiovascular: Denies chest pain or dyspnea on exertion. Respiratory: Denies shortness of breath. Gastrointestinal: Denies abdominal pain, nausea, or vomiting. Psychiatric: Endorses depression. Denies suicidal or homicidal ideations.	[doctor] hi , martha . how are you ? [patient] i'm doing okay . how are you ? [patient] okay . [doctor] martha is a 50-year-old female with a past medical history significant for congestive heart failure , depression and hypertension who presents for her annual exam . so , martha , it's been a year since i've seen you . how are you doing ? … [doctor] all right . that sounds good . all right , well , it's good to see you . [patient] good seeing you too . [doctor] hey , dragon , finalize the note .

(a) Sample data



(b) Dialogue Token Distribution

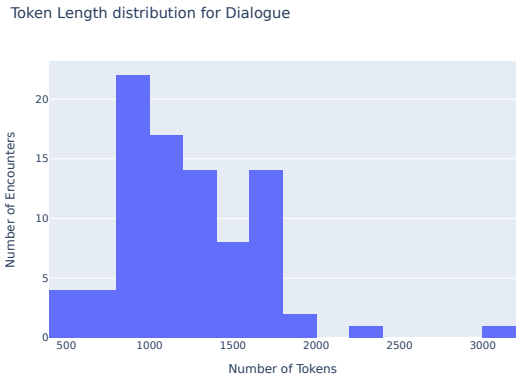


(c) Clinical Note Token Distribution

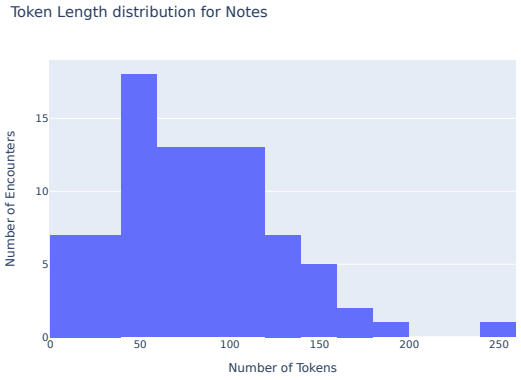
Figure A3: SubTask B - Subjective Section Sample Data

reference_objective_exam	dialogue_objective_exam
PHYSICAL EXAMINATION Cardiovascular: Grade 3/6 systolic ejection murmur. 1+ pitting edema of the bilateral lower extremities. VITALS REVIEWED Blood Pressure: Elevated.	[doctor] hi , martha . how are you ? [patient] okay . [doctor] martha is a 50-year-old female with a past medical history significant for congestive heart failure , depression and hypertension who presents for her annual exam . so , martha , it's been a year since i've seen you . how are you doing? … [doctor] all right . that sounds good . all right , well , it's good to see you . [patient] good seeing you too . [doctor] hey , dragon , finalize the note .

(a) Sample data



(b) Dialogue Token Distribution

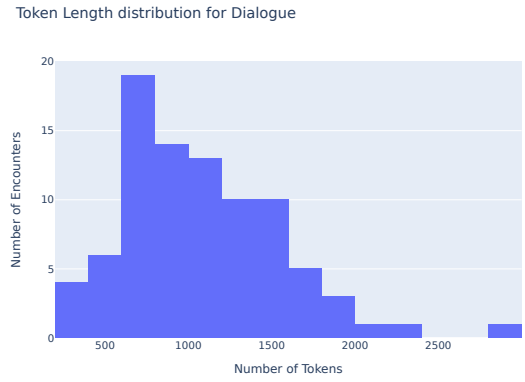


(c) Clinical Note Token Distribution

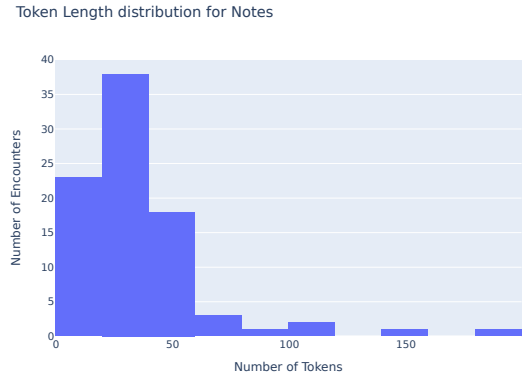
Figure A4: SubTask B - Objective Exam Section Sample Data

reference_objective_results	dialogue_objective_results
RESULTS Echocardiogram demonstrates decreased ejection fraction of 45%. Mitral regurgitation is present. Lipid panel: Elevated cholesterol.	[patient] i'm doing okay . how are you ? [doctor] i'm doing okay . so , i know the nurse told you about dax . i'd like to tell dax a little bit about you , okay ? … [patient] okay . [doctor] . some time during the day to take them , okay ? [patient] that might help me remember better . [doctor] hey , dragon , finalize the note .

(a) Sample data



(b) Dialogue Token Distribution

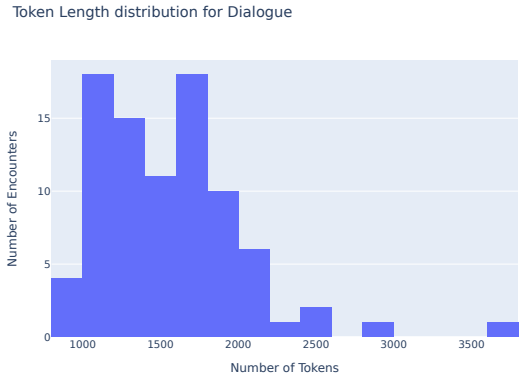


(c) Clinical Note Token Distribution

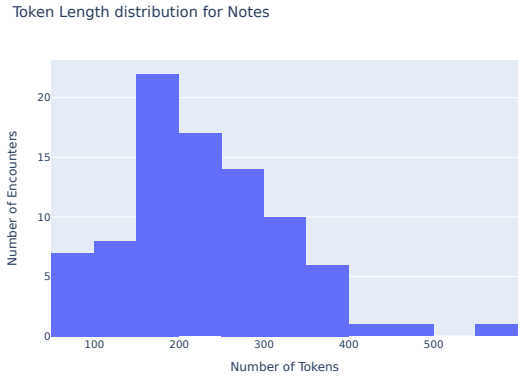
Figure A5: SubTask B - Objective Results Section sample data

reference_assessment_and_plan	dialogue_assessment_and_plan
ASSESSMENT AND PLAN Martha Collins is a 50-year-old female with a past medical history significant for congestive heart failure, depression, and hypertension who presents for her annual exam. Congestive heart failure. Medical Reasoning: She has been compliant with her medication and dietary modifications. … Medical Reasoning: The patient is due for her routine mammogram. Additional Testing: We will order a mammogram and have this scheduled for her. Patient Agreements: The patient understands and agrees with the recommended medical treatment plan.	[doctor] hi , martha . how are you ? [patient] i'm doing okay . how are you ? [doctor] i'm doing okay . so , i know the nurse told you about dax . i'd like to tell dax a little bit about you , okay ? [patient] okay . … [doctor] okay , all right . well , i'm glad to hear that . and you're taking your medication ? [patient] yes . [doctor] okay , good . and any symptoms like chest pains , shortness of breath , any swelling in your legs ? [patient] no , not that i've noticed . [patient] yeah , it's been helping a lot . i've been going every week , um , for the past year since my last annual exam . and that's been really helpful for me .

(a) Sample data



(b) Dialogue Token Distribution



(c) Clinical Note Token Distribution

Figure A6: SubTask B - Assessment and Plan Section sample data

A.2 Standing of our team

Our standings (in bold) for SubTask A - Section Header Classification is in Table A12. We omitted several teams from these standings and represent them by Ellipsis (**...**). This is done only to conserve space.

Team	Run	Accuracy	Rank
NUS-IDS	run1	0.78	1
...	...	...	...
Health-Mavericks	**run2**	**0.725**	**9**
Health-Mavericks	**run3**	**0.725**	**9**
Health-Mavericks	**run1**	**0.725**	**9**
...	...	...	...
Care4Lang	run2	0.345	31

Table A12: SubTask A - Section Header Classification Standings

Our standings (in bold) for SubTask A - Summarization is in Table A13

Team	Run	Mean Score	Rank
wanglab	run2	0.5789	1
...	...	...	...
Health-Mavericks	**run2**	**0.4683**	**25**
Health-Mavericks	**run3**	**0.4599**	**26**
ds4dh	run2	0.4334	27
Health-Mavericks	**run1**	**0.3996**	**28**
...	...	...	...
DFKI-MedIML	run1	0.3679	31

Table A13: SubTask A - Section Text Summarization Standings

Our standings (in bold) for SubTask B - Summarization is in Table A14

Team	Run	Rouge1	rank
wanglab	run3	0.6141	1
...	...	...	...
Health-Mavericks	**run1**	**0.5311**	**5**
...	...	...	...
Health-Mavericks	**run3**	**0.5111**	**11**
...	...	...	...
Health-Mavericks	**run2**	**0.2759**	**23**

Table A14: SubTask B - Notes Summarization Standings

Our standings (in bold) for Subjective Section can be found in Table A15

Team	Run	Subjective	Rank
wanglab	run1	0.6059	1
...	...	...	...
Health-Mavericks	**run1**	**0.4786**	**7**
...	...	...	...
Health-Mavericks	**run3**	**0.4657**	**12**
...	...	...	...
Health-Mavericks	**run2**	**0.3104**	**20**
...	...	...	...
Teddysum	run2	0.5353	23

Table A15: SubTask B - Subjective Section Performance

Our standings (in bold) for Objective Exam Section is in Table A16

Team	Run	Objective Exam	Rank
wanglab	run1	0.7102	1
...	...	...	...
Health-Mavericks	**run1**	**0.5374**	**7**
...	...	...	...
Health-Mavericks	**run3**	**0.4894**	**12**
...	...	...	...
Health-Mavericks	**run2**	**0.3222**	**20**
...	...	...	...
Teddysum	run2	0.1822	23

Table A16: SubTask B - Objective Exam Section Performance

Our standings (in bold) for Objective Results Section is in Table A17

Team	Run	Objective Results	Rank
wanglab	run1	0.6649	1
...	...	...	...
Health-Mavericks	**run1**	**0.5556**	**7**
...	...	...	...
Health-Mavericks	**run3**	**0.5383**	**12**
...	...	...	...
Health-Mavericks	**run2**	**0.3421**	**20**
...	...	...	...
Teddysum	run2	0.0182	23

Table A17: SubTask B - Objective Results Section Performance

Our standings (in bold) for Assessment and Plan Section is in Table A18

Team	Run	Assessment and Plan	Rank
wanglab	run1	0.6120	1
...	...	...	...
Health-Mavericks	**run1**	**0.4866**	**7**
...	...	...	...
Health-Mavericks	**run3**	**0.4854**	**12**
...	...	...	...
Health-Mavericks	**run2**	**0.3406**	**20**
...	...	...	...
Teddysum	run2	0.0968	23

Table A18: SubTask B - Assessment and Plan Section Performance