Automated Clinical Note Generation from Doctor-Patient Conversations using Large Language Models for the MEDIQA-Chat Challenge

Jennifer Yu
University of Toronto

Ronald Xie
University of Toronto

Augustin Toma
University of Toronto

Abstract

In this paper, we introduce advanced solutions to the 2023 MEDIQA-Chat challenge, focusing on the automated generation of clinical notes from physician-patient dialogues. Our models achieved first place in both Task A and Task B, which involved generating individual note excerpts and complete notes, respectively. We employed state-of-the-art large-language-model (LLM) based approaches, including fine-tuning FLAN-T5-Large, Longformer Encoder-Decoder (LED), and In-Context Learning (ICL) with GPT-4, and explored zero-shot/few-shot learning and prompt engineering techniques to enhance our solutions. In Task A, our fine-tuned FLAN-T5-Large model demonstrated competitive performance, achieving a header accuracy of 0.78, Rouge1 score of 0.4466, Rouge2 score of 0.2282, Bertscore F1 score of 0.7303, Bleurt score of 0.5593, and an aggregate score of 0.5789, outperforming the second and third place teams with aggregate scores of 0.5739 and 0.5622, respectively, achieving first place among 31 entries. In Task B, both our GPT-4+ICL approach and the fine-tuned LED approach consistently ranked higher than all other 19 entries in the MEDIQA-Chat2023 competition, with our GPT-4+ICL approach achieving the highest aggregate scores across all sections, and an average section score of 0.6483. These results demonstrate the potential of our methods in contributing to the development of automated tools that assist healthcare professionals in generating accurate clinical notes and enhancing patient engagement. The findings of this study hold implications for the future of natural language processing and machine learning applications in the healthcare domain, and indicate the promise of large-language-models in improving the documentation and communication of patient information by medical professionals.

1 Introduction

Clinical notes are essential in patient care, serving as a critical communication tool among healthcare professionals, researchers, and patients while documenting medical histories Mathioudakis et al. [2016]. However, the burden of note production can lead to physicians being more focused on their screens than engaging with patients, which may compromise the quality of care and result in omissions Gao et al. [2022]. The growing demand for automated solutions, particularly during high demand and pandemics, highlights the need for advancements in this area Sutton et al. [2020]Le Glaz et al. [2021].

The MEDIQA-Chat Tasks at ACL-ClinicalNLP 2023 addresses this need through an NLP competition focused on clinical note generation from doctor-patient conversations Abacha et al. [2023]. The Dialogue2Note Summarization task comprises generating individual note sections (Task A) and complete notes (Task B) Abacha et al. [2023].

To develop novel approaches, we created large-language-model (LLM) based solutions within the MEDIQA-Chat 2023 challenge, targeting the Dialogue2Note Summarization tasks and outperforming other participants' solutions. Our study presents advanced automatic medical note generation solutions, employing zero-shot/few-shot learning, prompt engineering, and fine-tuning. Performance was assessed using established benchmarks, including rouge Lin [2004], BERTScore Zhang et al. [2020], and BLEURT Sellam et al. [2020]. Our models secured first place in both tasks at the ACL-ClinicalNLP 2023 competition, demonstrating the effectiveness of our methods. This research carries significant implications for developing automated tools to help healthcare professionals generate accurate clinical notes while enhancing patient engagement, contributing to the advancement of clinical note generation techniques.

2 Background and Related Works

The generation of automated clinical notes from patient-physician dialogues has gained significant attention in recent years due to its potential to streamline the documentation process and enhance patient care Finley et al. [2018], Enarvi et al. [2020], Molenaar et al. [2020], Knoll et al. [2022]. Various methodologies have been proposed to address this challenge, such as employing extractive-abstractive techniques Joshi et al. [2020], Krishna et al. [2021] and fine-tuning pre-trained language models (PLMs) Zhang et al. [2021].

In addition to developing new methods, researchers have concentrated on curating high-quality datasets for training and benchmarking purposes Papadopoulos Korfiatis et al. [2022]. Some have even leveraged large language models (LLMs) to generate synthetic data for these tasks Chintagunta et al. [2021]. Furthermore, improving the evaluation of generated clinical notes has been a focus of recent studies, which have introduced both automatic metrics Moramarco et al. [2022] and human evaluation strategies Savkov et al. [2022].

Although the potential of in-context learning (ICL) for note generation has been discussed in the literature Lee et al. [2023], our work represents one of the first rigorous evaluations of this approach, thereby making a significant contribution to the field.

3 Methods

3.1 Task A Dataset and Method

Task A focuses on generating specific sections of a clinical note based on excerpts of diarized doctor-patient conversations. The training dataset for this task consists of 1200 dialogue-note-section header triplets and 100 validation examples. Participants must predict both the clinical note's subsection header (1 of 20 possible headers) and the note content derived from the patient-physician dialogue. Task A's 20 section headers are more detailed compared to the four headers used in Task B (discussed in the following subsection), with each header in Task A being a subset of those in Task B.

Figure 1 outlines two potential approaches for Task A (generating section headers and the corresponding clinical note excerpt). The first approach directly finetunes the FLAN-T5 model to predict the section header and generate the corresponding medical note in conjunction. We also briefly explored an alternative approach of predicting section headers separately by training a fully connected network (FCN) on Instructor [Su et al., 2022] embeddings of 4-utterance segments of the provided dialogue excerpt. However due to the added complexity and the limited upside as seen in Table 3, we only included the first approach in our submission. Its performance is reported and compared with other winning solutions in the challenge in Table 1.

3.2 Task B Dataset and Method

Task B aims to generate a full clinical note from complete doctor-patient dialogues. The dataset for this task contains 67 training and 20 validation examples, featuring transcribed and diarized dialogues from complete clinical encounters between patients and physicians.

Figure 2 (Left) outlines two approaches for Task B (generate the complete clinical note from the complete doctor-patient dialogue). For the first approach, we fine-tuned a Longformer-Encoder-Decoder (LED) model. Our second approach combines GPT4 with retrieval augmented in conext learning (ICL). In this approach, we retrieve the top k (k=2) most similar dialogues based on highest correlation of Instructor [Su et al., 2022] embeddings to that of the query dialogue and fetch their

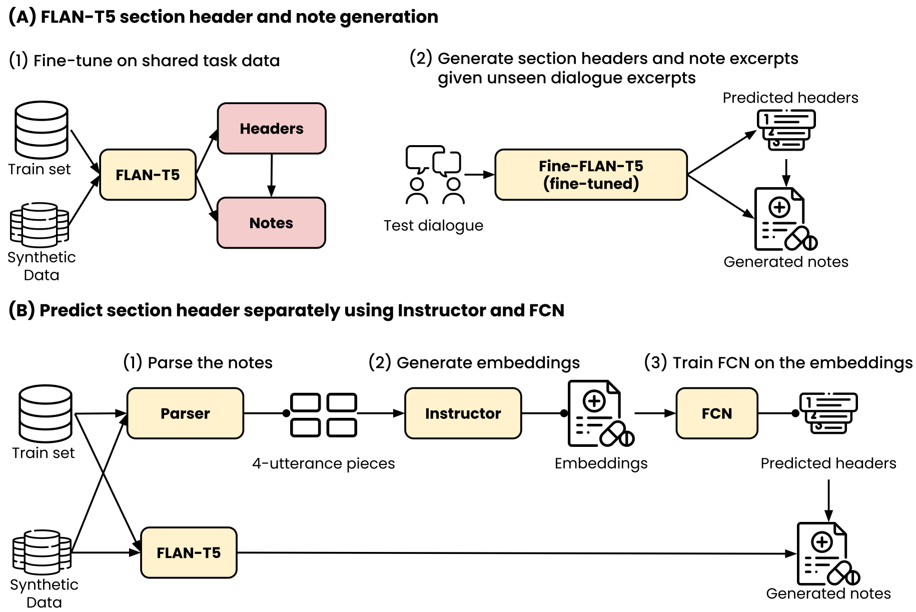

Figure 1: Method overview for Task A. (A) FLAN-T5 section header and note generation. We used FLAN-T5 to generate medical notes and headers. (B) Predict section header separately using Instructor [Su et al., 2022] and FCN. we divide the notes into 4-utterance pieces and utilized Instructor [Su et al., 2022] to generate embeddings. These embeddings are then used to train FCN solely for header prediction. Lastly, a FLAN-T5 model with the predicted section headers is used for generating notes.

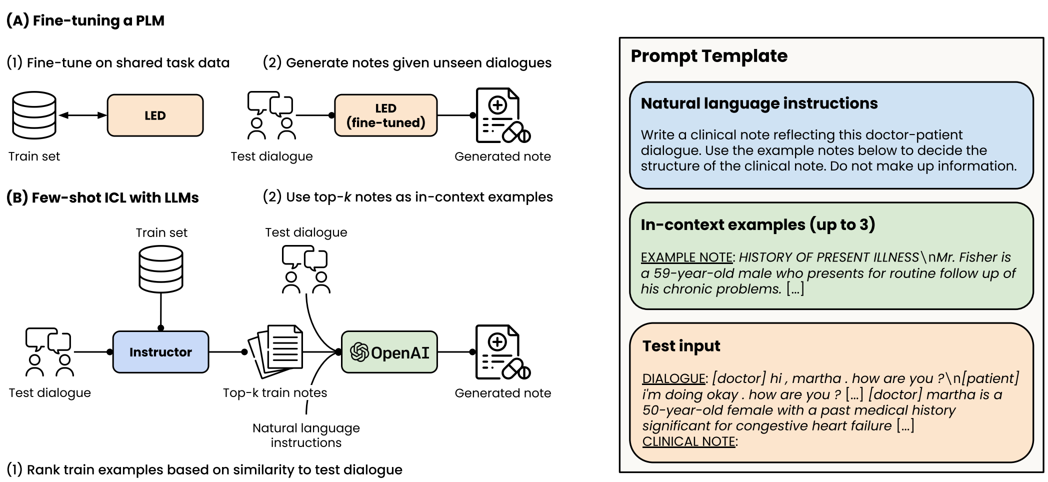

Figure 2: Method overview for Task B. **Left:** (A) Fine-tuning a pre-trained language model (PLM) on the shared task data. We used Longformer-Encoder-Decoder (LED). (B) Few-shot in-context learning (ICL) with large language models (LLMs). We rank train examples based on their similarity to the test dialogue. The notes of the top-k most similar examples are then used as the in-context examples to form a prompt alongside natural language instructions. We used GPT-4 as the LLM to generate the note given the prompt. **Right:** Prompt template for our in-context learning (ICL) with large language models (LLMs) based approach to Task B. Each prompt to the model includes natural language instructions, up to 3 in-context examples, and an unseen doctor-patient dialogue as input.

corresponding clinical notes from the 67 training examples. The retrieved notes are then used as in context examples inside the GPT4 prompt as shown in Figure2 (Right). The performance of these two approaches are reported and compared with other winning solutions in the challenge in Table 2.

3.2.1 Fine-tuning Pre-trained Language Models for Task B

For Task B, we first used a fine-tuning approach with a pre-trained language model (PLM) on the provided training set. The Longformer-Encoder-Decoder (LED) architecture was employed for this task, which has a maximum input and output size of 16,384 and 1024 tokens, respectively. Fine-tuning began from a checkpoint tuned on a PubMed summarization dataset, which we hypothesized allowed the model to leverage domain-specific knowledge.

3.2.2 In-Context Learning with LLMs for Task B

As a second approach for Task B, we employed in-context learning (ICL) with GPT-4. This method involved designing a simple prompt with natural language instructions and in-context examples, leveraging the few-shot learning capabilities of GPT-4.

The prompt size was limited to 6192 tokens, and up to 3 in-context examples were used, selected based on cosine similarity of train dialogues to the input dialogue. Dialogues were embedded using the instructor embedding model. In-context examples were restricted to the same 'dataset source' as the input dialogue, hypothesizing that this may improve performance since notes from the same dataset source likely have a similar structure and style.

4 Experiments

In this section, we present the methods and hyperparameters used for our experiments on Dialogue2Note Tasks A and B.

4.1 Task A: FLAN-T5

For Task A, we employed the large variant of the Flan-T5 model with the following hyperparameters. The maximum source length was set to 1024 tokens, and the maximum target length was set to 512 tokens. The source prefix used was: "Summarize the following patient-doctor dialogue. Include all medically relevant information, including family history, diagnosis, past medical (and surgical) history, immunizations, lab results and known allergies. You should first predict the most relevant clinical note section header and then summarize the dialogue. Dialogue:" Training and evaluation batch sizes were 8 and 12, respectively. The learning rate was 1e-4, and the optimizer used was AdamW. The model was trained for a total of 20 epochs with a warmup ratio of 0.1. Weight decay of 0.01 was applied, excluding bias and LayerNorm weights, and a label smoothing factor of 0.1 was used. BF16 was utilized during training, and the beam size during beam search decoding was set to 2.

4.2 Task B

4.2.1 LED

For Task B, we employed the Longformer-Encoder-Decoder (LED) model, with the following hyperparameters configured. We set the maximum source length to 4096 tokens and the maximum target length to 1024 tokens. The source prefix applied was identical to that in Task A. We used training and evaluation batch sizes of 8 and 6, respectively. The learning rate was established at 3e-5, and the AdamW optimizer was implemented. The model underwent training for a total of 50 epochs, with a warmup ratio of 0.1. We applied a weight decay of 0.01, excluding bias and LayerNorm weights, and utilized a label smoothing factor of 0.1. During training, we used FP16, and the beam size was set to 4 during beam search decoding. The minimum and maximum lengths of generated sequences were 1024 tokens. We incorporated a length penalty of 2.0 and restricted n-grams of size 3 to appear only once.

4.2.2 ICL

For the In-Context Learning (ICL) approach, we employed GPT-4 as the large language model. We limited the prompt size to 6192 tokens, allowing for 2000 tokens in output. We used up to 3 in-context examples, ensuring that they fit within the token limit. In-context examples were chosen based on their cosine similarity, as determined by the instructor model embeddings of dialogue. Notes associated with the most similar dialogues were provided as the in-context examples. We set the temperature parameter to 0.2 and employed the default OpenAI API hyperparameters.

5 Results

5.1 MEDIQA-Chat2023 competition results - Task A

The results of the 2023 MEDIQA-Chat competition Task A are presented in Table 1. The table shows the performance of the top three participating teams on the official held-out test set consisting of 200 extracted dialogue sections with matching header-note pairs.

Our finetuned FLAN-T5-Large achieved the highest performance across almost all metrics with a header accuracy of 0.78, Rouge1 score of 0.4466, Rouge2 score of 0.2282, Bertscore F1 score of 0.7303, Bleurt score of 0.5593, and an aggregate score of 0.5789. Our method outperformed the second and third place teams, who achieved an aggregate score of 0.5739 and 0.5622 respectively.

Method	Header Acc.	Rouge1	Rouge2	Bertscore_F1	Bleurt	Aggregate Score
Flan-T5-Large (Ours - 1st Place Team)	**0.78**	**0.4466**	**0.2282**	**0.7303**	0.5593	**0.5789**
2nd Place Team	0.71	0.4216	0.2017	0.7247	**0.5753**	0.5739
3rd Place Team	0.74	0.4303	0.2078	0.7187	0.5377	0.5622

Table 1: Task A Official 2023 MEDIQA-Chat competition results. Metrics were calculated on the offcial held-out test set consisting of 200 extracted dialogue sections with matching header-note pairs.

5.2 MEDIQA-Chat2023 competition results - Task B

The results of the 2023 MEDIQA-Chat Task B competition are presented in Table 2. Task B tests generating complete SOAP notes from clinical dialogue. Models were evaluated based on their performance on a held-out test set consisting of 40 full doctor-patient dialogues with matching SOAP notes. Aggregate scores were reported for each of the four defined SOAP note sections (Subjective, Objective Exam, Objective Results, and Assessment & Plan) and rouge1 was reported for the entire note as a whole.

The table shows the results for the top three teams, ranked by the average section score across all sections. Both our GPT4+ICL approach and the finetuned LED approach significantly outperformed all other 19 entries in the MEDIQA-Chat2023 competition. In particular, our GPT4+ICL approach achieved the highest aggregate scores across all sections, with an average section score of 0.6483. It also achieved a Rouge1 score of 0.5851 on the whole note, second only to our finetuned LED solution. We speculate that this is because the LED solution produced SOAP note lengths similar to that of the ground truths whereas the GPT4 solution often produced SOAP notes that were more clear and succinct. Content wise, the GPT4 solution outperforms LED as the latter is limited by the size of the training data. This hypothesis currently being verified and evaluated by a team of clinicians.

Overall, the competition results in Table 1 and Table 2 demonstrated that state-of-the-art models can achieve high performance on the task of generating SOAP notes from clinical dialogue.

5.3 Task A section header prediction using Instructor embeddings

We evaluate our Flan-T5-Large model, which is finetuned on both the official training set and our synthetic dataset (S2). FLAN-T5-Large achieves a test accuracy of 0.79 on test set header accuracy when trained on the combined O and S2 datasets.

We also explore the use of a fully connected network (FCN) trained on Instructor [Su et al., 2022] embeddings of 4-utterance parses of each dialogue and the entire dialogue, respectively. The resulting Instructor embeddings of 4-utterance parses are visualized in UMAP in Figure 3.

Method	S	OE	OR	AP	Average Section Score	Rouge1 (Whole Note)
GPT4 + ICL (Ours - 1st Place Team)	**0.6059**	**0.7102**	**0.6649**	**0.6120**	**0.6483**	0.5851
LED-Large (Ours - 1st Place Team)	0.5838	0.5915	0.5886	0.5607	0.5812	**0.6141**
2nd Place Team	0.4734	0.6405	0.5657	0.5368	0.5541	0.5739
3rd Place Team	0.5456	0.5307	0.5351	0.5355	0.5382	0.5622

Table 2: Task B Official 2023 MEDIQA-Chat competition results. Metrics were calculated on the offcial held-out test set consisting of 40 full doctor-patient dialogues with matching SOAP notes. Aggregate scores (From Rouge, Bert, Bleurt derived metrics) were reported for each of the four defined SOAP note sections (Subjective, Objective Exam, Objective Results, and Assessment & Plan) and rouge1 was reported for the entire note as a whole.

To mediate the class imbalance resulting from parsing the dialogues, we apply downsampling with thresholds of 500 and 1000 per class of section header to limit section headers associated with longer excerpts from dominating. The final section header is determined via majority vote. We evaluate Parsed + FCN on the official training set (O) and on combinations of O and our synthetic datasets S1 and S2. The results show that Parsed + FCN with a downsampling threshold of 1000 achieves the highest test accuracy of 0.70 on the O + S2 dataset, with a corresponding test accuracy of 0.82 when at least one section header is predicted correctly.

In addition, we evaluate All + FCN, which is trained on Instructor embeddings of the entire dialogue. All + FCN achieves a test accuracy of 0.717 on the official training set and a test accuracy of 0.77 on the combined O and S2 datasets.

From the results, it is clear that finetuning Flan-T5 for section header prediction and subsequent dialogue excerpt generation achieves superior section header prediction accuracy on the test set compared to all FCN derivatives. Therefore, although further post processing of the parsed + FCN output has potential to outperform Flan-T5-Large (81% of the dialogues has at least one vote correct), we did not pursue the option of separate section header prediction followed by note excerpt prediction owing to the simplicity of the end to end Flan-T5-large approach.

Nevertheless, we believe the parsed + FCN approach has its own interesting applications. Figure 4 is one such example where we parse entire doctor-patient dialogues in Task B and plot the resulting logits from the FCN which quite accurately reflect the content progression of the dialogue, which may then be utilized for extraction of relevant excerpts for a given section header or topic of interest.

6 Conclusion

In this paper, we present advanced solutions for automatic clinical notes generation from doctor-patient conversations using LLMs, employing a combination of techniques such as zero-shot/few-shot learning, prompt engineering, and fine-tuning. Our approach achieved first place in both Task A and B in the MEDIQA-Chat challenge in the ACL-ClinicalNLP 2023 competition, demonstrating the effectiveness of our proposed methods. Our work contributes to the development of automated tools that aid healthcare professionals in generating accurate and efficient clinical notes, reducing the workload on healthcare providers, improving the quality of care and enhancing patient satisfaction. The advancements in automated clinical note generation presented in this study have the potential to reshape the future of healthcare documentation, paving the way for more effective tools and techniques.

Some possible limitations of the current work include the reliance on GPT-4 architecture, which may not be optimal for local deployment and may run into potential data privacy concerns. Hallucinations remain a concern in both finetuned LLM and GPT based solutions. Although our solution performs

Method	Downsampling Threshold	Datasets	Test Accuracy	Test Accuracy (at least one vote correct)
Random Header	–	O	0.08	–
Majority Header	–	O	0.22	–
Flan-T5-Large	–	O + S2	**0.79**	–
Parsed + FCN	None	O	0.28	0.38
Parsed + FCN	500	O	0.66	0.78
Parsed + FCN	1000	O	0.64	0.75
Parsed + FCN	500	O + S1	0.64	0.77
Parsed + FCN	500	O + S2	0.68	0.80
Parsed + FCN	1000	O + S2	0.70	**0.82**
All + FCN	–	O	0.717	–
All + FCN	–	O + S2	**0.77**	–

Table 3: Task A section header prediction accuracy. We finetune Flan-T5-Large on the official training set and our synthetic dataset. We also explored training a fully connected network (FCN) on Instructor [Su et al., 2022] embeded 4-utterance parses of each dialogue and the entire dialogue respectively. Parsing dialogue creates large class imbalances which is mediated by downsampling. For Parsed + FCN, the final section header is determined via majority vote. The datasets used: O = official training set, S1 = synthetic dataset of 1600 dialogues and section header-note pairs generated by GPT3 and Davinci3. S2 = synthetic dataset of 24000 dialogues and section header-note pairs generated by Davinci3.

well in the tested domain, it may not generalize readily to new encounters involving drastically different medical settings. Safe guards against hallucinations is still an open topic for discussion but an important issue for transition of our method to clinical use.

Future directions for this research include exploring recently released language models, such as Med-PaLM Singhal et al. [2022] and LLaMA Touvron et al. [2023], and implementing advances in optimized exact attentionDao et al. [2022] to improve the performance of the models. Additionally, integrating a speech-to-text pipeline could pave the way for an end-to-end system that streamlines the process of medical note generation.

Lastly, the ongoing blinded randomized assessment by clinicians serves as a crucial step in validating the human preference findings from this study. Real-world validation and the potential impact on clinical practice should remain at the forefront of this research area. By continuing to rigorously assess and refine LLMs, we can work towards creating more reliable, trustworthy, and safe healthcare systems that leverage the power of artificial intelligence.

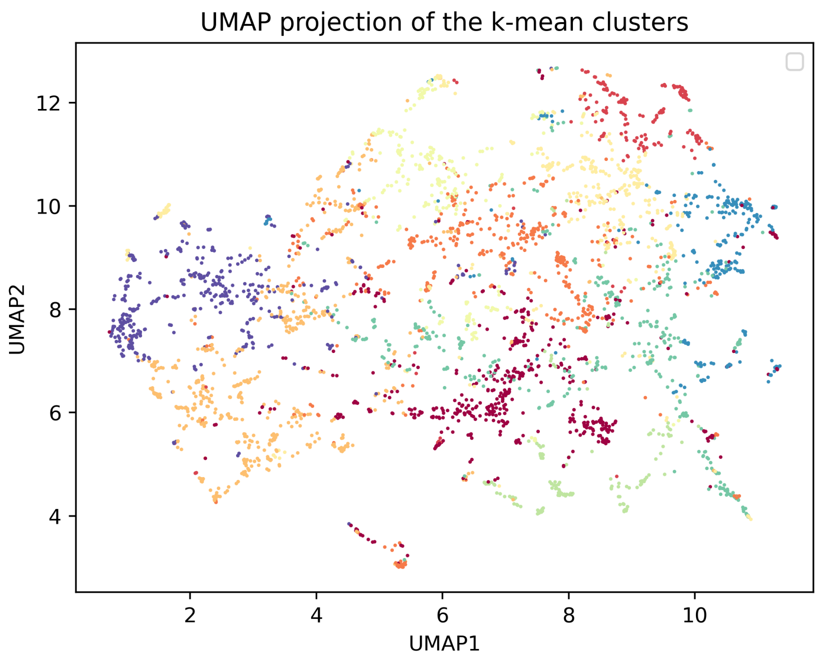

Figure 3: UMAP representation of the Instructor[Su et al., 2022] embedding for Task A dialogue excerpts, colored by the section header classes

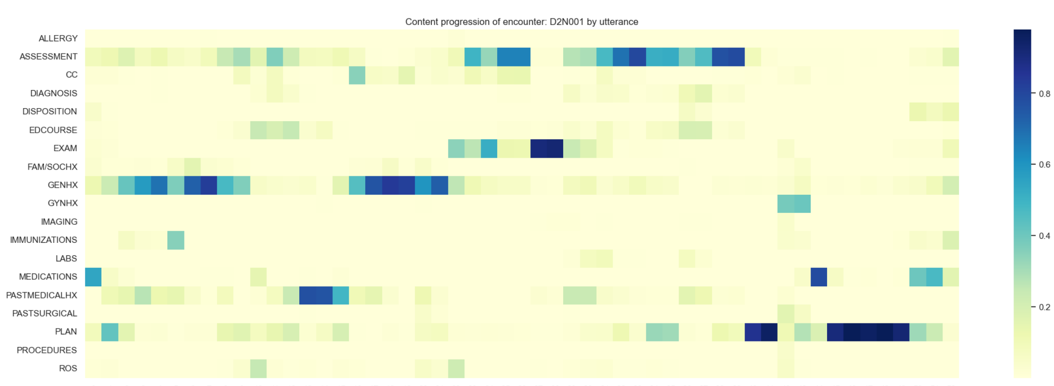

Figure 4: The full length dialogues in Task B could be parsed into 4-utterance snippets with a stride of 1, which can then be used as input to predict the content progression of the entire doctor-patient conversation using the best performing section header prediction FCN in Part A. Shown is a heatmap of the output logits for conversation D2N001 as an example.

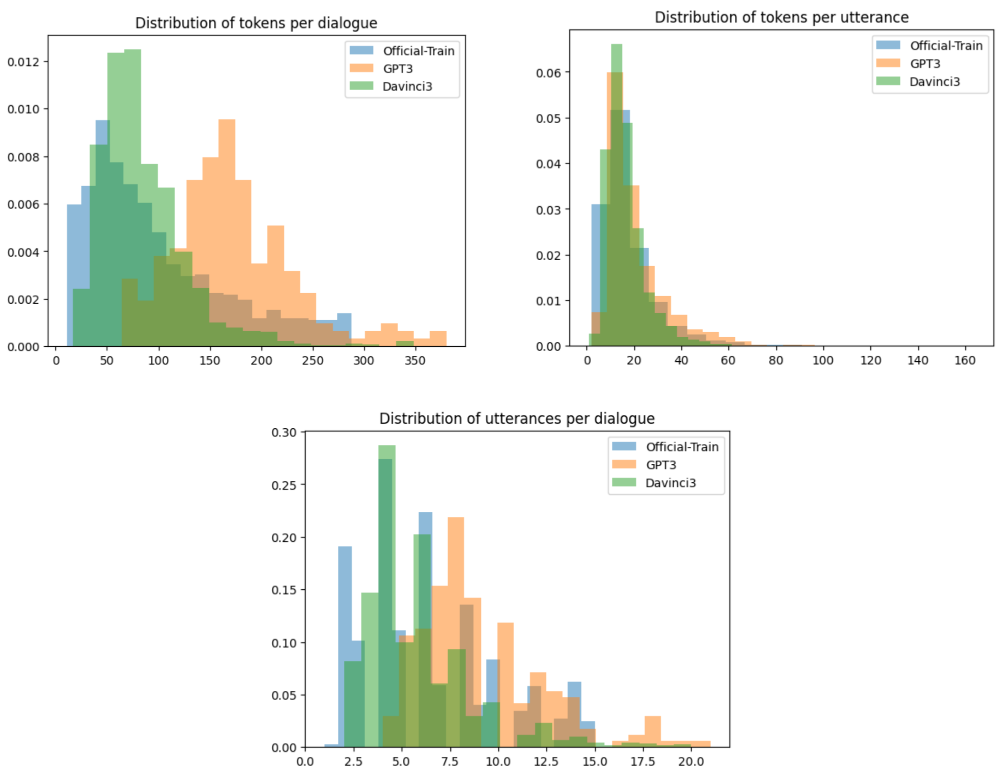

Figure 5: The distributions of tokens per dialogue (**Top left**), tokens per utterance (**Top right**), and utterances per dialogue (**Bottom**) of the official training dataset, and synthetic datasets generated by GPT3 and Davinci3 respectively for Task A.

References

Asma Ben Abacha, Wen-wai Yim, Griffin Adams, Neal Snider, and Meliha Yetisgen. Mediqa-chat tasks @ acl-clinicalnlp 2023 s. *ACL-ClinicalNLP 2023*, 2023.

Bharath Chintagunta, Namit Katariya, Xavier Amatriain, and Anitha Kannan. 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, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.nlpmc-1.9. URL `https://aclanthology.org/2021.nlpmc-1.9`.

Tri Dao, Daniel Y. Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. Flashattention: Fast and memory-efficient exact attention with io-awareness, 2022.

Seppo Enarvi, Marilisa Amoia, Miguel Del-Agua Teba, Brian Delaney, Frank Diehl, Stefan Hahn, Kristina Harris, Liam McGrath, Yue Pan, Joel Pinto, Luca Rubini, Miguel Ruiz, Gagandeep Singh, Fabian Stemmer, Weiyi Sun, Paul Vozila, Thomas Lin, and Ranjani Ramamurthy. Generating medical reports from patient-doctor conversations using sequence-to-sequence models. In *Proceedings of the First Workshop on Natural Language Processing for Medical Conversations*, pages 22–30, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.nlpmc-1.4. URL `https://aclanthology.org/2020.nlpmc-1.4`.

Gregory Finley, Erik Edwards, Amanda Robinson, Michael Brenndoerfer, Najmeh Sadoughi, James Fone, Nico Axtmann, Mark Miller, and David Suendermann-Oeft. An automated medical scribe for documenting clinical encounters. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, pages 11–15, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-5003. URL `https://aclanthology.org/N18-5003`.

Yanjun Gao, Dmitriy Dligach, Leslie Christensen, Samuel Tesch, Ryan Laffin, Dongfang Xu, Timothy Miller, Ozlem Uzuner, Matthew M Churpek, and Majid Afshar. A scoping review of publicly available language tasks in clinical natural language processing. *Journal of the American Medical Informatics Association*, 29(10):1797–1806, 08 2022. ISSN 1527-974X. doi: 10.1093/jamia/ocac127. URL `https://doi.org/10.1093/jamia/ocac127`.

Anirudh Joshi, Namit Katariya, Xavier Amatriain, and Anitha Kannan. Dr. summarize: Global summarization of medical dialogue by exploiting local structures. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3755–3763, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.findings-emnlp.335. URL `https://aclanthology.org/2020.findings-emnlp.335`.

Tom Knoll, Francesco Moramarco, Alex Papadopoulos Korfiatis, Rachel Young, Claudia Ruffini, Mark Perera, Christian Perstl, Ehud Reiter, Anya Belz, and Aleksandar Savkov. User-driven research of medical note generation software. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 385–394, Seattle, United States, July 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.naacl-main.29. URL `https://aclanthology.org/2022.naacl-main.29`.

Kundan Krishna, Sopan Khosla, Jeffrey Bigham, and Zachary C. Lipton. 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, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.384. URL `https://aclanthology.org/2021.acl-long.384`.

Aziliz Le Glaz, Yannis Haralambous, Deok-Hee Kim-Dufor, Philippe Lenca, Romain Billot, Taylor C Ryan, Jonathan Marsh, Jordan DeVylder, Michel Walter, Sofian Berrouiguet, and Christophe Lemey. Machine learning and natural language processing in mental health: Systematic review. *J Med Internet Res*, 23(5):e15708, May 2021. ISSN 1438-8871. doi: 10.2196/15708. URL `https://www.jmir.org/2021/5/e15708`.

Peter Lee, Sebastien Bubeck, and Joseph Petro. Benefits, limits, and risks of gpt-4 as an ai chatbot for medicine. *New England Journal of Medicine*, 388(13):1233–1239, 2023. doi: 10.1056/NEJMsr2214184. URL `https://doi.org/10.1056/NEJMsr2214184`. PMID: 36988602.

Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain, July 2004. Association for Computational Linguistics. URL `https://aclanthology.org/W04-1013`.

Alexander Mathioudakis, Ilona Rousalova, Ane Aamli Gagnat, Neil Saad, and Georgia Hardavella-corresponding. How to keep good clinical records. *Breathe (Sheff)*, 12(4): 369–373, 2016.

Sabine Molenaar, Lientje Maas, Verónica Burriel, Fabiano Dalpiaz, and Sjaak Brinkkemper. Medical dialogue summarization for automated reporting in healthcare. *Advanced Information Systems Engineering Workshops*, 382:76 – 88, 2020.

Francesco Moramarco, Alex Papadopoulos Korfiatis, Mark Perera, Damir Juric, Jack Flann, Ehud Reiter, Anya Belz, and Aleksandar Savkov. Human evaluation and correlation with automatic metrics in consultation note generation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5739–5754, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.394. URL `https://aclanthology.org/2022.acl-long.394`.

Alex Papadopoulos Korfiatis, Francesco Moramarco, Radmila Sarac, and Aleksandar Savkov. PriMock57: A dataset of primary care mock consultations. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 588–598, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-short.65. URL `https://aclanthology.org/2022.acl-short.65`.

Aleksandar Savkov, Francesco Moramarco, Alex Papadopoulos Korfiatis, Mark Perera, Anya Belz, and Ehud Reiter. Consultation checklists: Standardising the human evaluation of medical note generation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 111–120, Abu Dhabi, UAE, December 2022. Association for Computational Linguistics. URL `https://aclanthology.org/2022.emnlp-industry.10`.

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

Karan Singhal, Shekoofeh Azizi, Tao Tu, S. Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, Perry Payne, Martin Seneviratne, Paul Gamble, Chris Kelly, Nathaneal Scharli, Aakanksha Chowdhery, Philip Mansfield, Blaise Aguera y Arcas, Dale Webster, Greg S. Corrado, Yossi Matias, Katherine Chou, Juraj Gottweis, Nenad Tomasev, Yun Liu, Alvin Rajkomar, Joelle Barral, Christopher Semturs, Alan Karthikesalingam, and Vivek Natarajan. Large language models encode clinical knowledge, 2022.

Hongjin Su, Weijia Shi, Jungo Kasai, Yizhong Wang, Yushi Hu, Mari Ostendorf, Wen tau Yih, Noah A. Smith, Luke Zettlemoyer, and Tao Yu. One embedder, any task: Instruction-finetuned text embeddings. *ArXiv*, abs/2212.09741, 2022.

Reed T. Sutton, David Pincock, Daniel C. Baumgart, Daniel C. Sadowski, Richard N. Fedorak, and Karen I. Kroeker. An overview of clinical decision support systems: benefits, risks, and strategies for success. *npj Digital Medicine*, Med. 3, 17, 2020.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. Llama: Open and efficient foundation language models, 2023.

Longxiang Zhang, Renato Negrinho, Arindam Ghosh, Vasudevan Jagannathan, Hamid Reza Hassanzadeh, Thomas Schaaf, and Matthew R. Gormley. 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, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.findings-emnlp.313. URL `https://aclanthology.org/2021.findings-emnlp.313`.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert, 2020.