WangLab at MEDIQA-Chat 2023: Clinical Note Generation from Doctor-Patient Conversations using Large Language Models

John Giorgi[1,2,3*] **Augustin Toma**[1,3,4*] **Ronald Xie**[1,2,3,4*]
Sondra S. Chen[1,5] **Kevin R. An**[1,6] **Grace X. Zheng**[1,5] **Bo Wang**[1,3,4*]

[1]University of Toronto [2]Terrence Donnelly Centre for Cellular and Biomolecular Research
[3]Vector Institute for AI [4]University Health Network [5]Sunnybrook Health Sciences Centre
[6]Department of Cardiac Surgery, University of Toronto

`{john.giorgi, augustin.toma, ronald.xie, bowang.wang}@mail.utoronto.ca`

Abstract

This paper describes our submission to the MEDIQA-Chat 2023 shared task for automatic clinical note generation from doctor-patient conversations. We report results for two approaches: the first fine-tunes a pre-trained language model (PLM) on the shared task data, and the second uses few-shot in-context learning (ICL) with a large language model (LLM). Both achieve high performance as measured by automatic metrics (e.g. ROUGE, BERTScore) and ranked second and first, respectively, of all submissions to the shared task. Expert human scrutiny indicates that notes generated via the ICL-based approach with GPT-4 are preferred about as often as human-written notes, making it a promising path toward automated note generation from doctor-patient conversations.[1]

1 Introduction

The growing burden of clinical documentation has emerged as a critical issue in healthcare, increasing job dissatisfaction and burnout rates among clinicians and negatively impacting patient experiences (Friedberg et al., 2013; Babbott et al., 2014; Arndt et al., 2017). On the other hand, timely and accurate documentation of patient encounters is critical for safe, effective care and communication between specialists. Therefore, interest in assisting clinicians by automatically generating consultation notes is mounting (Finley et al., 2018; Enarvi et al., 2020; Molenaar et al., 2020; Knoll et al., 2022).

To further encourage research on automatic clinical note generation from doctor-patient conversations, the MEDIQA-Chat Dialogue2Note shared task was proposed (Ben Abacha et al., 2023). Here, we describe our submission to subtask B: the generation of full clinical notes from doctor-patient dialogues. We explored two approaches; the first fine-tunes a pre-trained language model (PLM, §3.1), while the second uses few-shot in-context learning (ICL, §3.2). Both achieve high performance as measured by automatic natural language generation metrics (§4) and ranked second and first, respectively, of all submissions to the shared task. In a human evaluation with three expert physicians, notes generated via the ICL-based approach with GPT-4 were preferred about as often as human-written notes (§4.3).

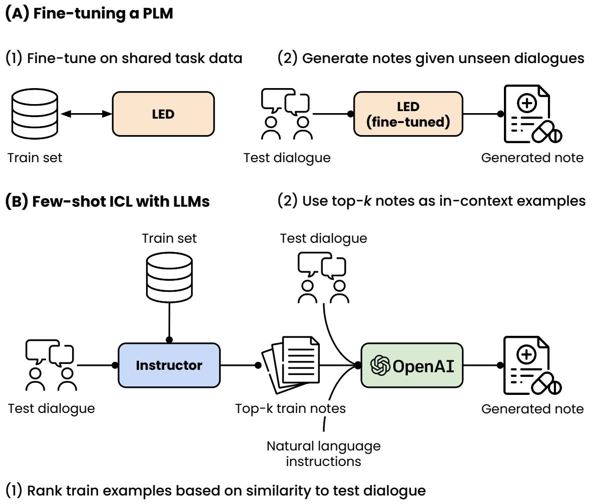

(A) Fine-tuning a PLM
(1) Fine-tune on shared task data
(2) Generate notes given unseen dialogues
LED
Train set
LED (fine-tuned)
Test dialogue
Generated note
(B) Few-shot ICL with LLMs
(2) Use top-k notes as in-context examples
Train set
Test dialogue
Instructor
OpenAI
Test dialogue
Top-k train notes
Generated note
Natural language instructions
(1) Rank train examples based on similarity to test dialogue

Figure 1: (**A**) Fine-tuning a pre-trained language model (PLM), Longformer-Encoder-Decoder (LED, Beltagy et al. 2020). (**B**) In-context learning (ICL) with large language models (LLMs). We rank train examples based on their similarity to the test dialogue using Instructor (Su et al., 2022a). Notes of the top-k most similar examples are then used as in-context examples to form a prompt alongside natural language instructions and fed to GPT-4 (OpenAI, 2023) to generate the clinical note.

2 Shared Task and Dataset

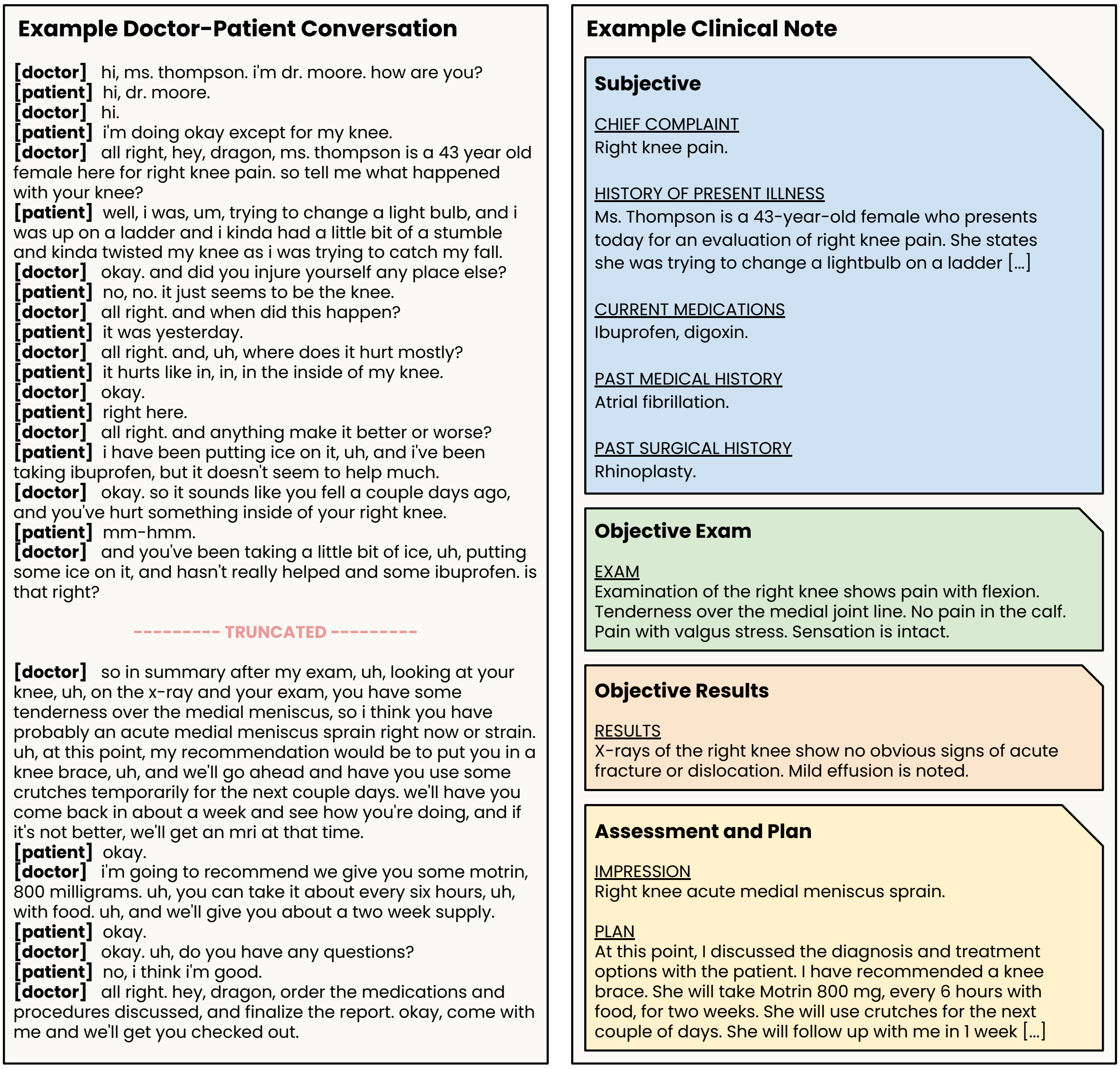
MEDIQA-Chat 2023 proposed two shared tasks:

1. **Dialogue2Note Summarization**: Given a conversation between a doctor and patient, the task is to produce a clinical note summarizing the conversation with one or more note sec-

[*] Core contributors. See author contributions

[1] `https://github.com/bowang-lab/MEDIQA-Chat-2023`

Proceedings of the 5th Clinical Natural Language Processing Workshop, pages 323–334
July 14, 2023 ©2023 Association for Computational Linguistics



Example Doctor-Patient Conversation

[doctor] hi, ms. thompson. i'm dr. moore. how are you?
[patient] hi, dr. moore.
[doctor] hi.
[patient] i'm doing okay except for my knee.
[doctor] all right, hey, dragon, ms. thompson is a 43 year old female here for right knee pain. so tell me what happened with your knee?
[patient] well, i was, um, trying to change a light bulb, and i was up on a ladder and i kinda had a little bit of a stumble and kinda twisted my knee as i was trying to catch my fall.
[doctor] okay. and did you injure yourself any place else?
[patient] no, no. it just seems to be the knee.
[doctor] all right. and when did this happen?
[patient] it was yesterday.
[doctor] all right. and, uh, where does it hurt mostly?
[patient] it hurts like in, in, in the inside of my knee.
[doctor] okay.
[patient] right here.
[doctor] all right. and anything make it better or worse?
[patient] i have been putting ice on it, uh, and i've been taking ibuprofen, but it doesn't seem to help much.
[doctor] okay. so it sounds like you fell a couple days ago, and you've hurt something inside of your right knee.
[patient] mm-hmm.
[doctor] and you've been taking a little bit of ice, uh, putting some ice on it, and hasn't really helped and some ibuprofen. is that right?

--------- TRUNCATED ---------

[doctor] so in summary after my exam, uh, looking at your knee, uh, on the x-ray and your exam, you have some tenderness over the medial meniscus, so i think you have probably an acute medial meniscus sprain right now or strain. uh, at this point, my recommendation would be to put you in a knee brace, uh, and we'll go ahead and have you use some crutches temporarily for the next couple days. we'll have you come back in about a week and see how you're doing, and if it's not better, we'll get an mri at that time.
[patient] okay.
[doctor] i'm going to recommend we give you some motrin, 800 milligrams. uh, you can take it about every six hours, uh, with food. uh, and we'll give you about a two week supply.
[patient] okay.
[doctor] okay. uh, do you have any questions?
[patient] no, i think i'm good.
[doctor] all right. hey, dragon, order the medications and procedures discussed, and finalize the report. okay, come with me and we'll get you checked out.

Example Clinical Note

Subjective

CHIEF COMPLAINT
Right knee pain.

HISTORY OF PRESENT ILLNESS
Ms. Thompson is a 43-year-old female who presents today for an evaluation of right knee pain. She states she was trying to change a lightbulb on a ladder [...]

CURRENT MEDICATIONS
Ibuprofen, digoxin.

PAST MEDICAL HISTORY
Atrial fibrillation.

PAST SURGICAL HISTORY
Rhinoplasty.

Objective Exam

EXAM
Examination of the right knee shows pain with flexion. Tenderness over the medial joint line. No pain in the calf. Pain with valgus stress. Sensation is intact.

Objective Results

RESULTS
X-rays of the right knee show no obvious signs of acute fracture or dislocation. Mild effusion is noted.

Assessment and Plan

IMPRESSION
Right knee acute medial meniscus sprain.

PLAN
At this point, I discussed the diagnosis and treatment options with the patient. I have recommended a knee brace. She will take Motrin 800 mg, every 6 hours with food, for two weeks. She will use crutches for the next couple of days. She will follow up with me in 1 week [...]

Figure 2: Example of a paired doctor-patient conversation and clinical note from the subtask B validation set. Dialogue has been lightly cleaned for legibility (e.g. remove trailing white space). Parts of the dialogue and note have been truncated. During evaluation, sections are grouped under one of four categories: "Subjective", "Objective Exam", "Objective Results", and "Assessment and Plan" (see §2.1 for details).

tions (e.g. Assessment, Past Medical History).

2. **Note2Dialogue Generation**: Given a clinical note, the task is to generate a synthetic doctor-patient conversation related to the information described in the note.

We focused on Dialogue2Note, which is divided into two subtasks. In subtask 'A' (Ben Abacha et al., 2023), the goal is to generate *specific sections* of a note given partial doctor-patient dialogues. In subtask 'B' (Yim et al., 2023), the goal is *full* note generation from complete dialogues. The remainder of the paper focuses on subtask B; see Appendix A for our approach to subtask A, which also ranks first of all submissions to the shared task.

2.1 Task definition

Each of the k examples consist of a doctor-patient dialogue, $D = d_1, \ldots, d_k$ and a corresponding clinical note, $N = n_1, \ldots, n_k$. The aim is to automatically generate a note n_i given a dialogue d_i. Each note comprises one or more sections, such as "Chief Complaint", and "Family history". During evaluation, sections are grouped under one of four categories: "Subjective", "Objective Exam", "Objective Results", and "Assessment and Plan".[2] See Figure 2 for an example doctor-patient conversation and clinical note pair.

[2] See here for the mapping

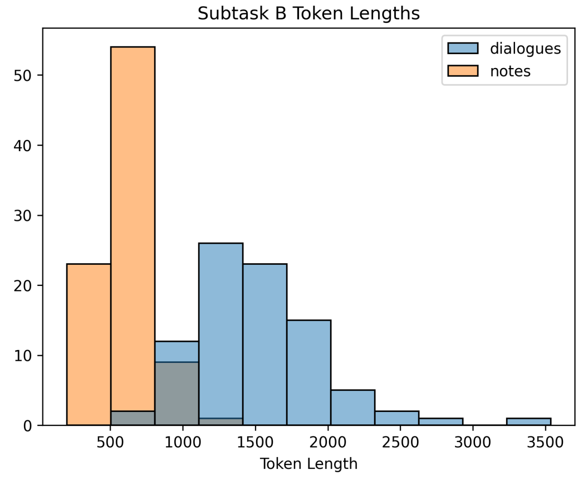


Figure 3: Histogram of token lengths for subtask B train and validation sets. Dialogues and notes were tokenized with `tiktoken` using the "`gpt-4`" encoding.

2.2 Dataset

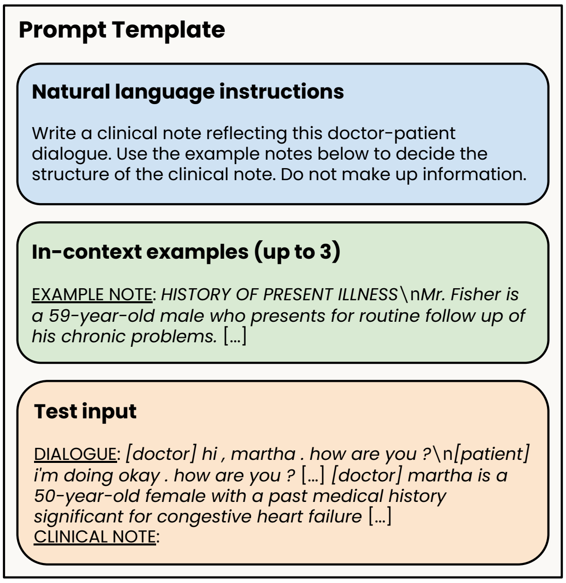
The dataset comprises 67 train and 20 validation examples, featuring transcribed dialogues from doctor-patient encounters and the resulting clinician-written notes. Each example is labelled with the 'dataset source', indicating the dialogue transcription system used to produce the note.

3 Approach

We take two high-performant approaches to the shared task. In the first, we fine-tune a pre-trained language model (PLM) on the provided training set (§3.1). In the second, we use in-context learning (ICL) with a large language model (LLM, §3.2).

3.1 Fine-tuning pre-trained language models

As a first approach, we fine-tune a PLM on the training set following a canonical, sequence-to-sequence training process (Figure 1 A; see Appendix C for details). Given the length of input dialogues (Figure 3), we elected to use Longformer-Encoder-Decoder (LED, Beltagy et al. 2020), which has a maximum input size of 16,384 tokens. We begin fine-tuning from a LED$_{\text{LARGE}}$ checkpoint tuned on the PubMed summarization dataset (Cohan et al., 2018), which performed best in preliminary experiments.[3] The model was fine-tuned using HuggingFace Transformers (Wolf et al., 2020) on a single NVIDIA A100-40GB GPU. Hyperparameters were lightly tuned on the validation set.[4]


Prompt Template

Natural language instructions

Write a clinical note reflecting this doctor-patient dialogue. Use the example notes below to decide the structure of the clinical note. Do not make up information.

In-context examples (up to 3)

EXAMPLE NOTE: *HISTORY OF PRESENT ILLNESS\nMr. Fisher is a 59-year-old male who presents for routine follow up of his chronic problems.* [...]

Test input

DIALOGUE: *[doctor] hi , martha . how are you ?\n[patient] i'm doing okay . how are you ?* [...] *[doctor] martha is a 50-year-old female with a past medical history significant for congestive heart failure* [...]
CLINICAL NOTE:

Figure 4: Prompt template for our in-context learning (ICL) based approach. Each prompt includes natural language instructions, up to 3 in-context examples, and an unseen doctor-patient dialogue as input.

3.2 In-context learning with LLMs

As a second approach, we attempt subtask B with ICL. We chose GPT-4 (OpenAI, 2023)[5] as the LLM and designed a simple prompt, which included natural language instructions and in-context examples (Figure 4). We limited the prompt size to 6192 tokens — allowing for 2000 output tokens, as the model's maximum token size is 8,192 — and used as many in-context examples as would fit within this token limit, up to a maximum of 3. We set the `temperature` parameter to 0.2 and left all other hyperparmeters of the OpenAI API at their defaults.

Natural language instructions During preliminary experiments, we found that GPT-4 was not overly sensitive to the exact phrasing of the natural language instructions in the prompt. We, therefore, elected to use short, simple instructions (Figure 4).

In-context example selection Each in-context example is a note from the train set. To select the notes, we first embed the dialogues of each training example and the input dialogue. Train dialogues are then ranked based on cosine similarity to the input dialogue; notes of the resulting top-k training examples are selected as the in-context examples (see Figure 1, B). Dialogues were embedded using Instructor (Su et al., 2022a), a text encoder that

[3] `https://huggingface.co/patrickvonplaten/led-large-16384-pubmed`

[4] See Appendix B.1 for details

[5] Specifically, the 03/14/2023 snapshot, "`gpt-4-0314`"

supports natural language instructions.[6] Lastly, we restricted in-context examples to be of the same 'dataset source' (see §2.2) as the input dialogue, hypothesizing that this may improve performance.[7]

3.3 Evaluation

Models are evaluated with the official evaluation script[8] on the validation set (as test notes are not provided). Generated notes are evaluated against the provided ground truth notes with ROUGE (Lin, 2004), BERTScore (Zhang et al., 2020) and BLEURT (Sellam et al., 2020). We report performance as the arithmetic mean of ROUGE-1 F1, BERTScore F1 and BLEURT-20 (Pu et al., 2021).

4 Results

4.1 Fine-tuning pre-trained language models

We present the results of fine-tuning LED in Table 1. Due to the non-determinism of the LED implementation,[9] we report the mean results of three training runs. Unsurprisingly, we find that scaling the model size from LED_{BASE} (12 layers, ∼162M parameters) to $\text{LED}_{\text{LARGE}}$ (24 layers, ∼460M parameters) leads to sizable gains in performance. Performance further improves by initializing the model with a checkpoint fine-tuned on the PubMed summarization dataset ($\text{LED}_{\text{LARGE-PubMed}}$). This is likely because (1) Dialouge2Note resembles a summarization task, and (2) text from PubMed is more similar to clinical text than is the general domain text used to pre-train LED.[10] Our submission to the shared task using this approach ranked second overall, outperforming the next-best submission by 2.7 average score; a difference comparable to the improvement in performance we see by doubling model size (see LED_{BASE} vs. $\text{LED}_{\text{LARGE}}$, Table 1).

Table 1: Fine-tuning LED. Mean and standard deviation (SD) of three training runs is shown. Scaling model size and pre-training on a related task improve performance. **Bold**: best scores.

Model	ROUGE-1 F1	BERTScore F1	BLEURT	Avg.
LED_{BASE}	$57.0_{0.4}$	$67.3_{0.1}$	$36.9_{0.0}$	53.8
$\text{LED}_{\text{LARGE}}$	$59.8_{0.2}$	$70.0_{0.6}$	$41.1_{0.8}$	57.0
$\text{LED}_{\text{LARGE-PubMed}}$	$\mathbf{61.7_{0.4}}$	$\mathbf{70.7_{0.2}}$	$\mathbf{41.5_{0.6}}$	**57.9**

4.2 In-context learning with LLMs

We present the results of ICL with GPT-4 in Table 2. We note several interesting trends in order of magnitude of impact. First, *selecting in-context examples based on the similarity of dialogues has a strong positive impact*, typically improving average score by 4 or more. *Using only notes as in-context examples, as opposed to dialogue-note pairs, also has a positive impact*, typically improving average score by ∼1. Surprisingly, *increasing the number of in-context examples had a marginal effect on performance*. Together these results suggest that the in-context examples' primary benefit is providing guidance with regard to the expected note structure, style and length. Finally, *filtering in-context examples to be of the same 'dataset source' as the input dialogue has a negligible impact on performance*.

The best strategy out-performs LED by almost 3 average score (60.8 vs. 57.9, see Table 1 & Table 2) and achieves first place of all submissions to the shared task, out-performing the runner up by >9 average score. We conclude that (1) few-shot ICL with GPT-4, using as little as one example, is a performant approach for note generation from doctor-patient conversations, and (2) using the notes of semantically similar dialogue-note pairs is a strong strategy for selecting the in-context examples.

4.3 Human evaluation

Automatic evaluation metrics like ROUGE, BERTScore and BLEURT are imperfect and may not correlate with aspects of human judgment.[11] Therefore, we conducted an expert human evaluation to validate our results. To make annotation feasible, we conducted it on the validation set (20 examples) using the best performing fine-tuned model: $\text{LED}_{\text{LARGE-PubMed}}$ (Table 1), and best performing ICL-based approach: 3-shot, similar, note-only examples filtered by dataset type (Table 2).

Three senior resident physicians[12] were shown a ground truth note, a note generated by the fine-tuned model, and a note generated by the ICL-based approach for each example (presented in random order as clinical note 'A', 'B' and 'C')

[6] We used the following instructions: "Represent the Medicine dialogue for clustering: `{dialogue}`"

[7] Manual review revealed that dataset source was predictive of note structure & style; likely because it indicates which clinician or electronic health record system produced the note

[8] https://github.com/abachaa/MEDIQA-Chat-2023/blob/main/scripts/evaluate_summarization.py

[9] https://github.com/huggingface/transformers/issues/12482

[10] LED is initialized from BART (Lewis et al., 2020), which was pre-trained on a combination of text from Wikipedia and BooksCorpus (Zhu et al., 2015)

[11] See §6 for an extended discussion

[12] The three annotators are a subset of the authors who did not interact with the model or model outputs before annotation

Table 2: ICL with GPT-4. Mean of ROUGE-1 F1, BERTScore F1 and BLEURT for three runs is shown. Selecting in-context examples based on similarity to input dialogue improves performance. Dialogue-note pairs as in-context examples (omitting 3-shot results due to token length limits) underperforms notes only. Filtering in-context examples to be of the same 'dataset source' as the input dialogue has little effect. **Bold**: best scores. SD < 0.1 in all cases.

	Unfiltered				Filtered by dataset source			
Example selection strategy	0-shot	1-shot	2-shot	3-shot	0-shot	1-shot	2-shot	3-shot
Dialogue-note pairs as in-context examples								
random	52.2	54.5	53.9	–	–	54.8	54.5	–
similar dialogues	–	59.4	59.4	–	–	60.1	60.3	–
Notes only as in-context examples								
random	–	56.3	56.7	56.7	–	56.3	56.5	56.7
similar dialogues	–	60.7	60.6	60.4	–	**60.8**	60.4	**60.8**

Table 3: Human evaluation. Three physicians selected their preference from human written ground-truth notes (GT), notes produced by the fine-tuned model (FT) and notes produced by in-context learning (ICL). Win rate is % of cases where note was preferred, excluding ties.

	Preferred			Ties		Win rate (%)		
Physician	GT	FT	ICL	FT/ICL	All	GT	FT	ICL
1	9	1	4	2	4	64	7	29
2	5	0	14	0	1	26	0	74
3	9	0	6	0	5	60	0	40
Total	23	1	24	2	10	48	2	50

and asked to select which note(s) they preferred, given a dialogue and some simple instructions:

Instructions: Please asses the **clinical notes A**, **B** and **C** relative to the **provided doctor-patient dialogue**. For each set of notes, you should select which note you prefer ('A', 'B', or 'C'). If you have approximately equal preference for two notes, select ('A/B', 'B/C', or 'C/A'). If you have no preference, select 'A/B/C'. A 'good' note should contain *all* **critical**, *most* **non-critical** and *very little* **irrelevant** information mentioned in a dialogue:

- **Critical**: Items medico-legally required to document the diagnosis and treatment decisions whose absence or incorrectness may lead to wrong diagnosis and treatment later on, e.g. the symptom "cough" in a suspected chest infection consultation. This is the key information a note needs to capture correctly in order to not mislead clinicians.

- **Non-critical**: Items that should be documented in a complete note but whose absence will not affect future treatment or diagnosis, e.g. "who the patient lives with" in a consultation about chest infection.

- **Irrelevant**: Medically irrelevant information covered in the consultation, e.g. the pet of a patient with a suspected chest infection just died.

The definitions of critical, non-critical and irrelevant information are taken from previous work on human evaluation of generated clinical notes (Moramarco et al., 2022; Savkov et al., 2022).

In short, notes generated by ICL are strongly preferred over notes generated by the fine-tuned model and, on average, *slightly* preferred over the human-written notes (Table 3), validating the high performance reported by the automatic metrics. We note, however, that inter-annotator agreement is low and speculate why this might be in §6.

5 Related Work

Automated note generation from doctor-patient conversations has received increasing attention in recent years (Finley et al., 2018; Enarvi et al., 2020; Molenaar et al., 2020; Knoll et al., 2022). Different methods have been proposed, such as extractive-abstractive approaches (Joshi et al., 2020; Krishna et al., 2021; Su et al., 2022b) and fine-tuning PLMs (Zhang et al. 2021, similar to our approach in §3.1). Others have focused on curating data for training and benchmarking (Papadopoulos Korfiatis et al., 2022), including the use of LLMs to produce synthetic data (Chintagunta et al., 2021). Lastly, there have been efforts to improve the evaluation of generated clinical notes, both with automatic metrics (Moramarco et al., 2022) and human evaluation (Savkov et al., 2022). While recent literature has commented on the *potential* of ICL for note generation (Lee et al., 2023), our work is among the first to evaluate this approach rigorously.

6 Conclusion

We present our submission to the MEDIQA-Chat shared task for clinical note generation from doctor-patient dialogues. We evaluated a fine-tuning-based approach with LED and an ICL-based approach with GPT-4, ranking second and first, respectively, among all submissions. Human evaluation with three physicians revealed that notes produced by GPT-4 via ICL were strongly preferred over notes produced by LED and, on average, slightly preferred over human-written notes. We conclude that ICL is a promising path toward clinical note generation from doctor-patient conversations.

Limitations

Evaluation of generated text is difficult Evaluating automatically generated text, including clinical notes, is generally hard due to the inherently subjective nature of many aspects of output quality. Automatic evaluation metrics such as ROUGE and BERTScore are imperfect (Deutsch et al., 2022) and may not correlate with aspects of expert judgment. However, they are frequently used to evaluate model-generated clinical notes and do correlate with certain aspects of quality (Moramarco et al., 2022). To further validate our findings, we also conducted a human evaluation with three expert physicians (§4.3). As noted previously (Savkov et al., 2022), even human evaluation of clinical notes is far from perfect; inter-annotator agreement is generally low, likely because physicians have differing opinions on the importance of each patient statement and whether it should be included in a consultation note. We also found low inter-annotator agreement in our human evaluation and speculate this is partially due to differences in specialties among the physicians. Physicians 1 and 3, both from family medicine, had high agreement with each other but low agreement with physician 2 (cardiac surgery, see Table 3). Investigating better automatic metrics and best practices for evaluating clinical notes (and generated text more broadly) is an active field of research. We hope to integrate novel and performant metrics in future work.

Data privacy While our GPT-4 based solution achieves the best performance, it is not compliant with data protection regulations such as HIPAA; although Azure does advertise a HIPAA-compliant option.[13] From a privacy perspective, locally deploying a model such as LED may be preferred; however, our results suggest that more work is needed for this approach to reach acceptable performance (see Table 3). In either case, when implementing automated clinical note-generation systems, healthcare providers and developers should ensure that the whole system — including text-to-speech, data transmission & storage, and model inference — adheres to privacy and security requirements to maintain trust and prevent privacy violations in the clinical setting.

Ethics Statement

Developing an automated system for clinical note generation from doctor-patient conversations raises several ethical considerations. First, informed consent is crucial: patients must be made aware of their recording, and data ownership must be prioritized. Equitable access is also important; the system must be usable for patients from diverse backgrounds, including those with disabilities, limited technical literacy, or language barriers. Addressing issues of data bias and fairness are necessary to avoid unfair treatment or misdiagnosis for certain patient groups. The system must implement robust security measures to protect patient data from unauthorized access or breaches. Establishing clear lines of accountability for errors or harms arising from using an automated system for note generation is paramount. Disclosure of known limitations or potential risks associated with using the system is essential to maintain trust in the patient-physician relationship. Finally, ongoing evaluations are necessary to ensure that system performance does not degrade and negatively impact the quality of care.

Acknowledgements

This research was enabled in part by support provided by the Digital Research Alliance of Canada (alliancecan.ca) and Compute Ontario (www.computeontario.ca). We thank all internal and external reviewers for their thoughtful feedback, which improved earlier drafts of this manuscript.

Author Contributions

John Giorgi (JG), Augustin Toma (AT) and Ronald Xie (RX) led the project in general, including

[13] https://azure.microsoft.com/en-us/products/cognitive-services/openai-service#security

data cleaning and processing, model implementation, and running experiments. JG wrote the initial manuscript and designed the human evaluation with feedback from AT and RX. AT and RX recruited Sondra S. Chen, Kevin R. An and Grace X. Zheng, who served as expert physicians in the human evaluation. Bo Wang provided high-level feedback and advice.

References

Brian G. Arndt, John W. Beasley, M. Watkinson, Jonathan L Temte, Wen-Jan Tuan, Christine A. Sinsky, and Valerie J. Gilchrist. 2017. Tethered to the ehr: Primary care physician workload assessment using ehr event log data and time-motion observations. *The Annals of Family Medicine*, 15:419–426.

Stewart F. Babbott, Linda Baier Manwell, Roger L. Brown, Enid N. H. Montague, Eric S. Williams, Mark D. Schwartz, Erik P. Hess, and Mark Linzer. 2014. Electronic medical records and physician stress in primary care: results from the memo study. *Journal of the American Medical Informatics Association : JAMIA*, 21 e1:e100–6.

Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *ArXiv*, abs/2004.05150.

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

Asma Ben Abacha, Wen-wai Yim, Yadan Fan, and Thomas Lin. 2023. An empirical study of clinical note generation from doctor-patient encounters. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2283–2294, Dubrovnik, Croatia. Association for Computational Linguistics.

Bharath Chintagunta, Namit Katariya, Xavier Amatriain, and Anitha Kannan. 2021. Medically aware GPT-3 as a data generator for medical dialogue summarization. In *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, S. Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Wei Yu, Vincent Zhao, Yanping Huang, Andrew M. Dai, Hongkun Yu, Slav Petrov, Ed Huai hsin Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022. Scaling instruction-finetuned language models. *ArXiv*, abs/2210.11416.

Arman Cohan, Franck Dernoncourt, Doo Soon Kim, Trung Bui, Seokhwan Kim, Walter Chang, and Nazli Goharian. 2018. A discourse-aware attention model for abstractive summarization of long documents. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 615–621, New Orleans, Louisiana. Association for Computational Linguistics.

Daniel Deutsch, Rotem Dror, and Dan Roth. 2022. Re-examining system-level correlations of automatic summarization evaluation metrics. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 6038–6052, Seattle, United States. Association for Computational Linguistics.

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. 2020. Generating medical reports from patient-doctor conversations using sequence-to-sequence models. In *Proceedings of the First Workshop on Natural Language Processing for Medical Conversations*, pages 22–30, Online. Association for Computational Linguistics.

Gregory Finley, Erik Edwards, Amanda Robinson, Michael Brenndoerfer, Najmeh Sadoughi, James Fone, Nico Axtmann, Mark Miller, and David Suendermann-Oeft. 2018. 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. Association for Computational Linguistics.

Mark W. Friedberg, Peggy G. Chen, Kristin R Van Busum, Frances Aunon, Chau Pham, John P. Caloyeras, Soeren Mattke, Emma Pitchforth, Denise D. Quigley, Robert Henry Brook, Francis J. Crosson, and Michael A. Tutty. 2013. Factors affecting physician professional satisfaction and their implications for patient care, health systems, and health policy. *Rand health quarterly*, 3 4:1.

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

Anirudh Joshi, Namit Katariya, Xavier Amatriain, and Anitha Kannan. 2020. 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. Association for Computational Linguistics.

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

Peter Lee, Sébastien Bubeck, and Joseph Petro. 2023. Benefits, limits, and risks of gpt-4 as an ai chatbot for medicine. *The New England journal of medicine*, 388 13:1233–1239.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, 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.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

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

Francesco Moramarco, Alex Papadopoulos Korfiatis, Mark Perera, Damir Juric, Jack Flann, Ehud Reiter, Anya Belz, and Aleksandar Savkov. 2022. 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. Association for Computational Linguistics.

OpenAI. 2023. Gpt-4 technical report. *ArXiv*, abs/2303.08774.

Alex Papadopoulos Korfiatis, Francesco Moramarco, Radmila Sarac, and Aleksandar Savkov. 2022. 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. Association for Computational Linguistics.

Amy Pu, Hyung Won Chung, Ankur Parikh, Sebastian Gehrmann, and Thibault Sellam. 2021. Learning compact metrics for MT. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 751–762, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Aleksandar Savkov, Francesco Moramarco, Alex Papadopoulos Korfiatis, Mark Perera, Anya Belz, and Ehud Reiter. 2022. 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. Association for Computational Linguistics.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. BLEURT: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892, Online. Association for Computational Linguistics.

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

Jing Su, Longxiang Zhang, Hamidreza Hassanzadeh, and Thomas Schaaf. 2022b. Extract and abstract with bart for clinical notes from doctor-patient conversations. In *Interspeech*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

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

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. 2020. Bertscore: Evaluating text generation with BERT. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Yukun Zhu, Ryan Kiros, Richard S. Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*, pages 19–27. IEEE Computer Society.

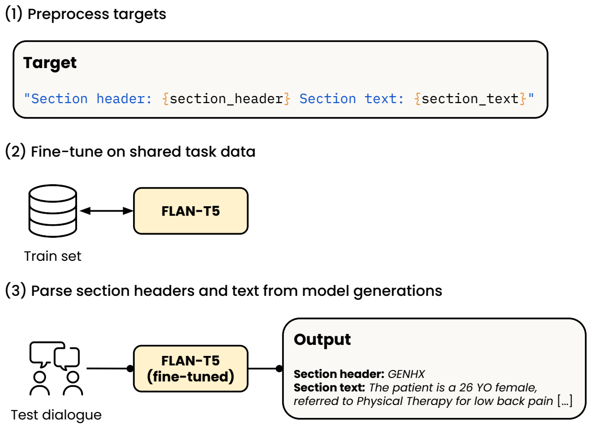

(1) Preprocess targets
Target
"Section header: {section_header} Section text: {section_text}"
(2) Fine-tune on shared task data
FLAN-T5
Train set
(3) Parse section headers and text from model generations
FLAN-T5 (fine-tuned)
Test dialogue
Output
Section header: GENHX
Section text: The patient is a 26 YO female, referred to Physical Therapy for low back pain [...]

Figure 5: Fine-tuning FLAN-T5 (Chung et al., 2022) for subtask A. Before training, targets are preprocessed as "Section header: {section_header} Section text: {section_text}". After decoding, the section header and text are parsed using regular expressions.

A Subtask A

In subtask A of the Dialogue2Note Summarization shared task, given a partial doctor-patient dialogue, the goals are to: (1) predict the appropriate section header, e.g. "PASTMEDICALHX" and (2) generate that specific section of a note. We approached this task by fine-tuning a PLM on the provided training set, following a canonical, sequence-to-sequence training process (see Appendix C for details). In preliminary experiments, we found that the instruction-tuned FLAN-T5 (Chung et al., 2022) performed particularly well at this task.

We hypothesized that jointly learning to predict the section header and generate the section text would improve overall performance. To do this, we preprocessed the training set so the targets were of the form: "Section header: {section_header} Section text: {section_text}". After decoding, the section header and text were parsed using regular expressions and evaluated separately (Figure 5). Section header prediction was evaluated as the fraction of predicted headers that match the ground truth (accuracy), and section text was evaluated similarly to subtask B (see §3.3). In cases where the model output an invalid section header,[14] we replaced it with "GENHX" (general history), which tends to summarize the contents of the other sections. The model was fine-tuned on a single NVIDIA A100-40GB GPU. Hyperparameters were lightly tuned on the validation set (Table 4).

We present the results of our approach on the validation set in Table 5. Similar to subtask B (see §4.1), we find, perhaps unsurprisingly, that scaling the model size from FLAN-T5$_{\text{BASE}}$ (24 layers, ∼250M parameters) to FLAN-T5$_{\text{LARGE}}$ (48 layers, ∼780M parameters) leads to large improvements in performance. Performance is further improved by jointly learning to predict section headers and generate note sections. Our submission to the shared task based on this approach tied for first on section header prediction (78% accuracy), and ranked first for note section generation (average ROUGE-1, BERTScore and BLEURT F1-score of 57.9).

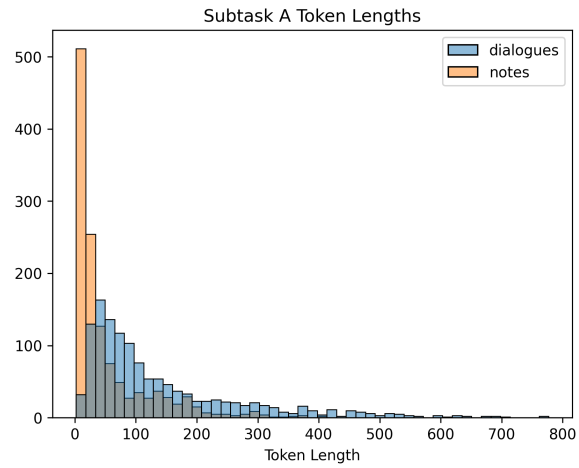


Figure 6: Histogram of token lengths for subtask A train and validation sets. Dialogues and notes were tokenized with HuggingFace Tokenizers using "google/flan-t5-large". Lengths greater than the 99th-percentile are omitted to make the plot legible.

B Subtask B

B.1 Hyperparameter tuning of LED

We lightly tuned the hyperparameters of LED$_{\text{LARGE-PubMed}}$ on the subtask B validation set against the average ROUGE-1 F1, BERTScore F1 and BLEURT-20 scores. The best hyperparameters obtained are given in Table 6. We used the same hyperparameters when fine-tuning LED$_{\text{BASE}}$ and LED$_{\text{LARGE}}$ in §4.1.

B.2 Post processing LEDs outputs

In practice, we found that the fine-tuned LED model sometimes produces invalid section headers; notably, this problem did not occur with the ICL-based approach using GPT-4. Therefore, we lightly post-processed LEDs outputs using a simple script that identifies section headers produced by the model not in the ground truth set and uses fuzzy

[14] In practice, we found that the fine-tuned model rarely, if ever, generates invalid section headers

Table 4: Hyperparameters used with FLAN-T5 on the Dialogue2Note subtask A

Hyperparameter	Value	Comment
`max_source_length`	1024	truncate input sequences to this max length
`max_target_length`	512	truncate output sequences to this max length
`source_prefix`	*"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:"*	instruction text prepended to all inputs
`train_batch_size`	8	batch size during training
`eval_batch_size`	12	batch size during inference
`learning_rate`	1e-4	learning rate during training
`optimizer`	AdamW (Loshchilov and Hutter, 2019)	optimizer used during training
`num_train_epochs`	20	total number of training epochs
`warmup_ratio`	0.1	proportion of training steps to linearly increase the learning rate to `learning_rate`
`lr_scheduler`	linear with warmup	learning rate linearly increased during first `warmup_ratio` fraction of train steps and linearly decreased to 0 afterwords
`weight_decay`	0.01	not applied to bias & `LayerNorm` weights
`label_smoothing`	0.1	label smoothing factor used during training
`bf16`	true	whether to use BF16 during training
`num_beams`	2	beam size used during beam search decoding

Table 5: Fine-tuning FLAN-T5. Accuracy of predicted section headers and score of generated note sections is shown. Jointly learning to predict section headers and generate notes improve performance. **Bold**: best scores.

		Note generation			
Model	Header prediction (%)	ROUGE-1 F1	BERTScore F1	BLEURT	Avg.
Random header	8.0	–	–	–	–
Majority header	22.0	–	–	–	–
FLAN-T5$_{\text{BASE}}$	71.0	40.1	70.5	52.7	54.5
FLAN-T5$_{\text{LARGE}}$	**79.0**	**49.8**	**74.5**	**58.0**	**60.8**
↪ *w/o header prediction*	–	48.0	74.3	57.6	59.9

string matching[15] to replace them with the closest valid header. For example, in one run, this process converted the (incorrect) predicted section header "HISTORY OF PRESENT" to the nearest valid header "HISTORY OF PRESENT ILLNESS".

C Fine-tuning Seq2Seq Models

When training the sequence-to-sequence (seq2seq) models for both subtask A (Appendix A) and B (§3.1), we followed a canonical supervised fine-tuning (SFT) process. We start with a pre-trained, encoder-decoder transformer-based language model (Vaswani et al., 2017). First, the encoder maps each token in the input to a contextual embedding. Then, the autoregressive decoder generates an output, token-by-token, attending to the outputs of the encoder at each timestep. Decoding proceeds until a special "end-of-sequence" token (e.g. `</s>`) is generated, or a maximum number of tokens have been generated. Formally, X is the *input* sequence, which in our case is a doctor-patient dialogue, and Y is the corresponding *output* sequence of length T, in our case a clinical note. We model the conditional probability:

$$p(Y|X) = \prod_{t=1}^{T} p(y_t|X, y_{<t}) \tag{1}$$

During training, we optimize over the model parameters θ the sequence cross-entropy loss:

$$\ell(\theta) = -\sum_{t=1}^{T} \log p(y_t|X, y_{<t}; \theta) \tag{2}$$

[15] We used `https://github.com/seatgeek/thefuzz`

Table 6: Hyperparameters used with Longformer-Encoder-Decoder (LED) on the Dialogue2Note subtask B

Hyperparameter	Value	Comment
max_source_length	4096	truncate input sequences to this max length
max_target_length	1024	truncate output sequences to this max length
source_prefix	*“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. Dialogue:”*	instruction text prepended to all inputs
train_batch_size	8	batch size during training
eval_batch_size	6	batch size during inference
learning_rate	3e-5	learning rate during training
optimizer	AdamW (Loshchilov and Hutter, 2019)	optimizer used during training
num_train_epochs	50	total number of training epochs
warmup_ratio	0.1	proportion of training steps to linearly increase the learning rate to learning_rate
lr_scheduler	linear with warmup	learning rate linearly increased during first warmup_ratio fraction of train steps and linearly decreased to 0 afterwords
weight_decay	0.01	not applied to bias & LayerNorm weights
label_smoothing	0.1	label smoothing factor used during training
fp16	true	whether to use FP16 during training
num_beams	4	beam size used during beam search decoding
min_length	100	min length of generated sequences
max_length	1024	max length of generated sequences
length_penalty	2.0	values $>$ 0 promote longer output sequences
no_repeat_ngram	3	ngrams of this size can only occur once

maximizing the log-likelihood of the training data. As is common, we use *teacher forcing* during training, feeding previous ground truth inputs to the decoder when predicting the next token in the sequence. During inference, we generate the output using beam search (Graves, 2012). Beams are ranked by mean token log probability after applying a length penalty. Models are fine-tuned using the HuggingFace Transformers library.[16]

[16] https://github.com/huggingface/transformers/blob/main/examples/pytorch/summarization/run_summarization.py