A Privacy-preserving Approach to Ingest Knowledge from Proprietary Web-based to Locally Run Models for Medical Progress Note Generation

Sarvesh Soni and **Dina Demner-Fushman**
Lister Hill National Center for Biomedical Communications
National Library of Medicine, National Institutes of Health, Bethesda, MD, USA
sarvesh.soni@nih.gov, ddemner@mail.nih.gov

Abstract

Clinical documentation is correlated with increasing clinician burden, leading to the rise of automated methods to generate medical notes. Due to the sensitive nature of patient electronic health records (EHRs), locally run models are preferred for a variety of reasons including privacy, bias, and cost. However, most open-source locally run models (including medical-specific) are much smaller with limited input context size compared to the more powerful closed-source large language models (LLMs) generally available through web APIs (Application Programming Interfaces). In this paper, we propose a framework to harness superior reasoning capabilities and medical knowledge from closed-source online LLMs in a privacy-preserving manner and seamlessly incorporate it into locally run models. Specifically, we leverage a web-based model to distill the vast patient information available in EHRs into a clinically relevant subset without sending sensitive patient health information online and use this distilled knowledge to generate progress notes by a locally run model. Our ablation results indicate that the proposed framework improves the performance of the Mixtral model on progress note generation by 4.6 points on ROUGE (a text-matching based metric) and 7.56 points on MEDCON F1 (a metric that measures the clinical concepts overlap).

1 Introduction

Physicians document progress or SOAP (subjective, objective, assessment, and plan) notes in electronic health records (EHRs) periodically to document patient care journey. While abundant patient chart data (e.g., regularly collected lab values) enhances physician assessment of patient progress, it leads to information overload and clinician burden, giving rise to clinician burnout (Tai-Seale et al., 2017), emphasizing the importance of automating this task.

The increasing popularity and capabilities of large language models (LLMs) led to their numerous applications in both general and medical domains (Chen et al., 2024). While the closed-source LLMs available via web APIs (Application Programming Interfaces) generally outperform the locally run alternatives, there is a growing popularity and community support for on-premise models, especially in the medical domain because of several advantages that these models offer such as transparency, adaptability, and information security (Tian et al., 2024). We propose to reap the benefits offered by locally run models while harnessing the strong reasoning capabilities of API-based proprietary LLMs. To this end, there have been numerous efforts toward distilling knowledge from proprietary LLMs (e.g., GPT-4) to train smaller or locally run models (Xu et al., 2024). In the medical domain, most work on such distillation has focused on curating instruction-tuning datasets using superior LLMs for training or tuning smaller models (Wu et al., 2023; Zhang et al., 2023, 2024). Differently, our framework exploits web-based LLMs for achieving a *bottleneck* task for locally run models formulated in a way that does not spill sensitive patient information to online API-based models.

We formulate the task of progress note generation (PNG) to automatically generate the next note given a patient's prior progress note and all interim structured chart data (e.g., vital signs). One of the main limitations of the locally run models in tackling PNG is processing and clinically analyzing the vast amount of interim structured chart data (an average of over 1400 rows of tabular data between any pair of subsequent progress notes) – the *bottleneck*. To overcome this barrier, we leverage an advanced API-based proprietary model to choose clinically relevant structured data rows without sending any real patient information to the online model server. This distilled structured chart information, along with the prior progress note, is used by a locally run model to generate the next progress note.

Proceedings of the Fifth Workshop on Privacy in Natural Language Processing, pages 178–183

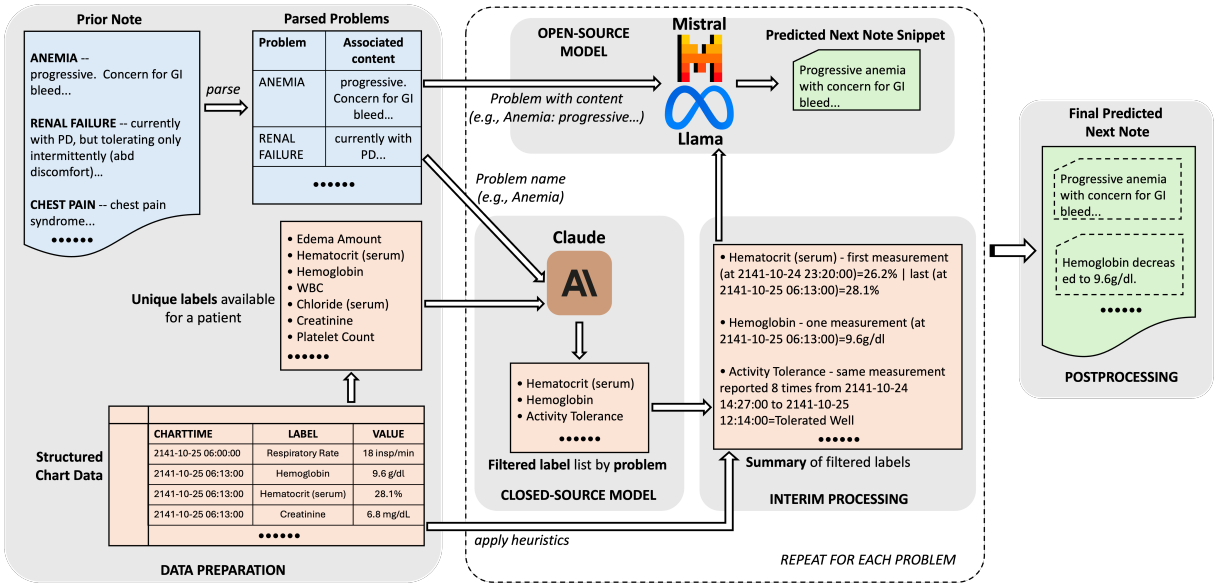

Figure 1: Proposed framework with example snippets.

2 Methods

2.1 Data

We sample the progress notes used in our evaluation from MIMIC-III, a publicly accessible database collected from an intensive care unit (ICU) setting (Johnson et al., 2016). The included pairs of subsequent progress notes were selected if they (1) belong to the same ICU admission and, between their documentation times, there is (2) no other documented progress note and (3) non-empty structured chart data. This resulted in a total of 7089 annotation instances (note pairs) associated with 1616 unique patients and a mean of 1474.9 rows of structured chart data per instance. Due to resource constraints, we randomly sample 100 instances for quantitative evaluation. We additionally perform manual analysis on a sub-sample. The instructions to access the dataset and code used for evaluations are available at GitHub[1].

The information in the subjective part of the progress notes is provided by the patient (information more likely to be found in patient-physician conversations) while the objective part is mainly comprised of factual patient data such as laboratory values (oftentimes directly fetched into the note without major modifications). Differently, writing the assessment and plan sections requires a careful examination of the past notes and structured chart data. Thus, in this work, we focus on automatically generating the assessment and plan sections of a progress note given the previous note and all interim structured chart data.

2.2 Framework

Figure 1 shows the proposed framework's architecture. The pair of notes in an annotation instance is referred to as *Prior* and *Next* notes and the interim structured chart data as *Structured Chart Data*.

2.2.1 Data Preparation

The *Prior* note is segmented into different *problem*-specific sections by (1) identifying clinical problem entities using a clinical concept extraction system, Stanza (Zhang et al., 2021), and (2) applying heuristics over the annotations (e.g., the identified problem entity must be at the beginning of a sentence). Further, we extract the unique available data labels from *Structured Chart Data*, without the associated clinical data values.

2.2.2 Proprietary Web-based Model

We call the online API-based model once for each problem segment identified from the Data Preparation step. Only the problem entity text span (e.g., *Anemia*) identified by concept extractor and the unique data labels (e.g., *Hemoglobin*) without any corresponding values (e.g., *9.6 g/dl*) are sent to the web-based model (Figure 2). Multiple structured data elements are collected routinely for subsets of the patients with similar problems. Thus, despite the problem names and data labels coming from a real patient, it is safe to assume that this step does not raise any major privacy concerns, especially in

[1] github.com/soni-sarvesh/png-privacy-preserving

The following is the list of available structured chart data elements from a patient's electronic health records.

[STRUCTURED CHART DATA LABELS]

For a specific problem of "[PROBLEM DESCRIPTION]", which of the data elements from the provided list above will be useful for a clinician to assess the progress of the patient and why?

Note: Only output the data elements from the provided list above. Do not output data elements that are not part of the provided list above.

The output should be a JSON snippet formatted in the following schema, including the leading and trailing "```json" and "```".

```json
{{
"selected element #1": "reason",
"selected element #2": "reason",
and so on
}}
```

Figure 2: The prompt used for instructing the web-based model. Text in [*] is replaced with data.

the absence of any identifiable patient information and the specific data values.

We prompt the model to filter the list of data labels using the supplied problem name such that the resultant labels are useful to document the progress of the patient. The model outputs a list of filtered labels, picking the most important attributes in context of the provided problem name. We chose Anthropic's Claude 3 Opus (Anthropic, 2024) as our web-based model owing to its superior performance among other proprietary models.

2.2.3 Interim processing

Though the count of filtered data labels for each problem was much smaller, the resultant structured data table with only these labels still contained substantial number of rows. To overcome this, we summarize the rows by aggregating the values associated with data labels based on their data types using simple rules. For numerical values, we reduce the numbers to include only the first and the last measurements with associated chart times along with the mean, minimum, and maximum values. For categorical data, we include the first and the last measurements with chart times along with the most frequent value with its frequency. General corner cases were covered such as reporting the value directly in the case of a single value.

You are given the following initial assessment and plan note for a patient for the specific problem of "[PROBLEM DESCRIPTION]" written at [PRIOR NOTE CHARTTIME]:

[PRIOR ASSESSMENT AND PLAN NOTE]

The following is the summary of relevant structured patient chart data with selected chart times:

[FILTERED STRUCTURED CHART DATA]

Current time is [NEXT NOTE CHARTTIME]. Generate a new assessment and plan note for the problem of "[PROBLEM DESCRIPTION]" by incorporating the recent events from the patient's chart. Restrict the length of the new note to a maximum of 50 words.

Figure 3: The prompt used for instructing the locally run models. Text in [*] is replaced with data.

2.2.4 Locally Run Models

The resultant summary from the interim processing step is fed to the locally run model for each problem individually along with the entire problem-specific note text. Additionally, we include the chart times of the *Prior* (for temporal context) and *Next* (acting as the note generation time for a fair comparison with ground truth) notes (Figure 3). The model predicts the *Next* note text for the input problem. We experiment using three locally run models–Biomistral 7B (Labrak et al., 2024), Mixtral 8x7B (Jiang et al., 2024), and LLaMa 2 70B (Touvron et al., 2023). Biomistral is developed by further pre-training the Mistral model (Jiang et al., 2023), an open-weight locally run model, on the PubMed Central Open Access Subset while Mixtral is a mixture-of-experts model based on Mistral. LLaMa 2 is the next generation model from the LLaMa family of LLMs and has shown to outperform the web-based models in some cases.

2.2.5 Post-processing

We combine the generated notes for individual problems to produce a coherent predicted *Next* note. We use three metrics for our quantitative evaluation–ROUGE (Lin, 2004), BERTScore (Zhang et al., 2019) using $\text{RoBERTa}_{\text{LARGE}}$ (Liu et al., 2019), and MEDCON (Yim et al., 2023). ROUGE-N calculates N-gram overlap between the predicted and original *Next* notes while ROUGE-L uses the length of the longest common subsequence and ROUGE-Lsum splits the text into sentences before calculating ROUGE-L. BERTScore measures the cosine similarity between BERT-based contextual embeddings of the tokens in predicted and orig-

Table 1: Evaluation results on 100 sampled instances. Ablations are performed on 30 instances due to hardware constraints. In the ablation section, rows starting with *"− knowledge"* indicate the model results without the use of problem segments and knowledge distillation using a web-based model. The best model results in each category are **bolded**. *Prior* – return the prior note as prediction.

Baseline	ROUGE				BERTScore			MEDCON
	1	2	L	Lsum	Precision	Recall	F1	F1
Prior	51.24	35.33	41.87	50.55	88.58	88.44	88.50	55.46
Biomistral 7B	20.97	5.09	11.32	20.19	80.46	**78.65**	79.52	23.06
Mixtral 8x7B	**23.67**	**6.61**	**13.69**	**22.76**	**81.13**	78.55	**79.80**	**26.88**
LLaMa 2 70B	19.24	4.61	10.60	18.63	79.33	77.97	78.63	23.19
Ablation analysis on a sub-sample								
Prior	51.77	35.04	42.13	50.73	89.62	89.94	89.77	55.46
Biomistral 7B	20.10	4.55	10.97	19.36	80.81	79.96	80.37	23.63
− knowledge	20.85	**7.36**	13.81	19.88	**82.08**	**80.87**	**81.42**	21.99
Mixtral 8x7B	**24.68**	6.09	**14.57**	**23.79**	81.99	79.64	80.78	**27.60**
− knowledge	20.29	3.84	10.99	19.19	80.96	78.44	79.66	20.04
LLaMa 2 70B	18.43	4.22	10.27	17.89	79.97	79.04	79.49	23.74
− knowledge	16.75	2.64	8.97	16.03	80.21	77.68	78.91	16.75

inal text. Differently, MEDCON calculates the overlap (using F1-score) between Unified Medical Language System (UMLS) concepts identified in the generated and real notes text.

3 Results

The performance measures in automatically generating progress notes are shown in Table 1. Interestingly, the baseline results from merely returning the same note text as the prior note achieves highest automated evaluation metric scores. Note that this is due to the high textual similarity between the next and previous notes as the progress notes are oftentimes copied forward for editing. The larger models, Mixtral and LLaMa, performed better than Biomistral on the MEDCON metric, while Mixtral performed the best on all three metrics. The ablation results in the sub-sample demonstrate the advantage of our proposed framework that uses problem segments (as opposed to the entire note as input) and distilled structured chart data labels (instead of providing all available data as input). All the models gained improvement in their MEDCON scores with the incorporation of the proposed framework while all the larger models (Mixtral and LLaMa) saw improvements on ROGUE, BERTScore F1 and MEDCON. Of note, Mixtral achieved the largest performance improvements across all the metrics (with as much as 4.6 points on ROGUE-Lsum and 7.56 on MEDCON).

Our qualitative analysis of the predictions by the best and worst performing models on 20 instances (Table 2) aligns well with the quantitative results. Further, in our manual evaluation, we found that in most cases the predicted notes contained the relevant interim change information. For instance, *"pain and fluid status"* in the original next note is appropriately captured in the system prediction by *"pain and possible dehydration"*. There was minimal evidence of hallucinations (the inclusion of incorrect or irrelevant information in the output) where, in one instance, Biomistral suggested *"increasing the dose of vasopressor"* while the original note mentioned *"off pressors"*. Notably, Mixtral did not include incorrect information in the manually evaluated predictions.

4 Discussion

Our results indicate the advantage of tackling the task of PNG by considering individual component problems at a time and leveraging advanced web-based models to transfer knowledge by filtering relevant clinical attributes in structured chart data. Our manual evaluation suggests the predicted notes capture the important updates on patient's progress. Importantly, Mixtral exhibited capabilities in capturing overall status changes (e.g., *sepsis improving*), whereas the Biomistral demonstrated

Table 2: Common prediction characteristics from a manual evaluation of the models predictions on 20 annotation instances. *Info* – Information; *Gold* – Original next note; *Pred* – Predicted next note.

Category	Prediction description	Example	Biomistral	Mixtral
			% (#)	
Relevant Info	Updated the note with relevant information	**Gold:** Tachycardia: ... Likely due to pain and fluid status. **Pred:** Tachycardia ... is likely related to pain and possible dehydration ... (**Good**)	65.0 (13)	80.0 (16)
		Gold: a-fib: ... No evidence for dvt. **Pred:** *could not capture* (**Bad**)		
Wrong Info	Included content that is incorrect or unrelated to patient	**Gold:** Septic shock- resolved, off pressors since yesterday ... **Pred:** #Septic shock ... recommend increasing the dose of vasopressor support ...	5.0 (1)	0.0 (0)

its ability to capture domain knowledge-related updates (e.g., *add digoxin 0.25mg daily*). Fine-tuning LLMs leads to specialized domain knowledge (as exhibited by Biomistral), however, it is also shown to reduce general in-context learning abilities (Wang et al., 2023), as seen in Table 1.

Overall, the findings from this paper provide support for the feasibility of the complex task of PNG. Further, it provides a framework for harnessing the reasoning capabilities of proprietary API-based models in a privacy-preserving manner while using a locally run model for handling sensitive patient information.

5 Limitations

The limitations of our framework include its inability to capture new problems that may have emerged in the interval, which is an interesting avenue for future research. Moreover, physicians use information beyond the structured chart data while writing progress notes, e.g., radiology reports. As described earlier, it is challenging to incorporate the interim structured data along with the previous note text in the limited context size of existing on-premise models. Thus, we leave the inclusion of other information sources to future work.

Acknowledgments

This work was supported by the intramural research program at the U.S. National Library of Medicine, National Institutes of Health, and utilized the computational resources of the NIH HPC Biowulf cluster (http://hpc.nih.gov).

References

Anthropic. 2024. The Claude 3 Model Family: Opus, Sonnet, Haiku. https://paperswithcode.com/paper/the-claude-3-model-family-opus-sonnet-haiku.

Hailin Chen, Fangkai Jiao, Xingxuan Li, Chengwei Qin, Mathieu Ravaut, Ruochen Zhao, Caiming Xiong, and Shafiq Joty. 2024. ChatGPT's One-year Anniversary: Are Open-Source Large Language Models Catching up? *Preprint*, arxiv:2311.16989.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7B. *Preprint*, arxiv:2310.06825.

Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, Gianna Lengyel, Guillaume Bour, Guillaume Lample, Lélio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, Szymon Antoniak, Teven Le Scao, Théophile Gervet, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2024. Mixtral of Experts. *Preprint*, arxiv:2401.04088.

Alistair E. W. Johnson, Tom J. Pollard, Lu Shen, Li-wei H. Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G. Mark. 2016. MIMIC-III, a freely accessible critical care database. *Scientific Data*, 3(1):160035.

Yanis Labrak, Adrien Bazoge, Emmanuel Morin, Pierre-Antoine Gourraud, Mickael Rouvier, and Richard

Dufour. 2024. BioMistral: A Collection of Open-Source Pretrained Large Language Models for Medical Domains. *Preprint*, arxiv:2402.10373.

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.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *Preprint*, arxiv:1907.11692.

Ming Tai-Seale, Cliff W. Olson, Jinnan Li, Albert S. Chan, Criss Morikawa, Meg Durbin, Wei Wang, and Harold S. Luft. 2017. Electronic Health Record Logs Indicate That Physicians Split Time Evenly Between Seeing Patients And Desktop Medicine. *Health Affairs*, 36(4):655–662.

Shubo Tian, Qiao Jin, Lana Yeganova, Po-Ting Lai, Qingqing Zhu, Xiuying Chen, Yifan Yang, Qingyu Chen, Won Kim, Donald C Comeau, Rezarta Islamaj, Aadit Kapoor, Xin Gao, and Zhiyong Lu. 2024. Opportunities and challenges for ChatGPT and large language models in biomedicine and health. *Briefings in Bioinformatics*, 25(1):bbad493.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open Foundation and Fine-Tuned Chat Models. *Preprint*, arxiv:2307.09288.

Yihan Wang, Si Si, Daliang Li, Michal Lukasik, Felix Yu, Cho-Jui Hsieh, Inderjit S. Dhillon, and Sanjiv Kumar. 2023. Two-stage LLM Fine-tuning with Less Specialization and More Generalization. In *The Twelfth International Conference on Learning Representations*.

Chaoyi Wu, Weixiong Lin, Xiaoman Zhang, Ya Zhang, Yanfeng Wang, and Weidi Xie. 2023. PMC-LLaMA: Towards Building Open-source Language Models for Medicine. *Preprint*, arxiv:2304.14454.

Xiaohan Xu, Ming Li, Chongyang Tao, Tao Shen, Reynold Cheng, Jinyang Li, Can Xu, Dacheng Tao, and Tianyi Zhou. 2024. A Survey on Knowledge Distillation of Large Language Models. *Preprint*, arxiv:2402.13116.

Wen-wai Yim, Yujuan Fu, Asma Ben Abacha, Neal Snider, Thomas Lin, and Meliha Yetisgen. 2023. Aci-bench: A Novel Ambient Clinical Intelligence Dataset for Benchmarking Automatic Visit Note Generation. *Sci Data*, 10(1):586.

Hongbo Zhang, Junying Chen, Feng Jiang, Fei Yu, Zhihong Chen, Guiming Chen, Jianquan Li, Xiangbo Wu, Zhang Zhiyi, Qingying Xiao, Xiang Wan, Benyou Wang, and Haizhou Li. 2023. HuatuoGPT, Towards Taming Language Model to Be a Doctor. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10859–10885, Singapore. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2019. BERTScore: Evaluating Text Generation with BERT. In *International Conference on Learning Representations*.

Xinlu Zhang, Chenxin Tian, Xianjun Yang, Lichang Chen, Zekun Li, and Linda Ruth Petzold. 2024. AlpaCare:Instruction-tuned Large Language Models for Medical Application. *Preprint*, arxiv:2310.14558.

Yuhao Zhang, Yuhui Zhang, Peng Qi, Christopher D Manning, and Curtis P Langlotz. 2021. Biomedical and clinical English model packages for the Stanza Python NLP library. *Journal of the American Medical Informatics Association*, 28(9):1892–1899.