EPFL-MAKE at "Discharge Me!": An LLM System for Automatically Generating Discharge Summaries of Clinical Electronic Health Record

Haotian Wu Paul Boulenger Antonin Faure
Berta Céspedes Farouk Boukil Nastasia Morel
Zeming Chen Antoine Bosselut
École Polytechnique Fédérale de Lausanne (EPFL)
{firstname.lastname}@epfl.ch

Abstract

This paper presents our contribution to the Streamlining Discharge Documentation shared task organized as part of the ACL'24 workshop. We propose MEDISCHARGE (**ME**ditron-7B Based Medical Summary Generation System for **DISCHARGE** Me), an LLM-based system to generate Brief Hospital Course and Discharge Instruction summaries based on a patient's Electronic Health Record. Our system is build on a Meditron-7B with context window extension, ensuring the system can handle cases of variable lengths with high quality. When the length of the input exceeds the system input limitation, we use a dynamic information selection framework to automatically extract important sections from the full discharge text. Then, extracted sections are removed in increasing order of importance until the input length requirement is met. We demonstrate our approach outperforms tripling the size of the context window of the model. Our system obtains a 0.289 overall score in the leaderboard, an improvement of 183% compared to the baseline, and a ROUGE-1 score of 0.444, achieving a second place performance in the shared task.

1 Introduction

In modern healthcare, the electronic health record (EHR) is a fundamental part of clinical practices as it ensures the documentation of a patient's medical journey. Essential to this record are the clinical notes seriously crafted by physicians post-consultation. These notes encapsulate crucial details ranging from the patient's reason for the visit to their medical history, symptoms, diagnosis, and recommended treatment plan (Uslu and Stausberg, 2021). Acting as vital components within the EHR, clinical notes foster effective communication among healthcare providers, offer legal protection, and ensure continuity of care (Hay et al., 2020).

However, despite their important role, clinical notes impose a substantial time burden for physicians. Recent research in the U.S. has revealed that physicians spend an average of 1.77 hours daily on documentation tasks outside of consultation hours (Gaffney et al., 2022). This extensive time investment contributes to pressing healthcare issues such as clinician burnout, excessive workloads, and understaffing (Gesner et al., 2019; Moy et al., 2021).

One area where clinicians encounter notable time constraints is in the creation of discharge summaries and hospital course summaries. Crafting these summaries to be both concise and comprehensive demands considerable effort. To address this challenge, there is a pressing need to streamline the summary generation of these sections. People try to use machine learning to automate these summaries, but all face the difficulty of models with limited abilities, domain-specific terminologies, and reasoning over specialized knowledge (Hu et al., 2020; Ive et al., 2020; Tang et al., 2023).

BioNLP ACL'24 Shared Task on Streamlining Discharge Documentation focuses on solving the summarization challenges. In this competition, participants worked with a dataset derived from MIMIC-IV, covering 109,168 Emergency Department (ED) visits. Each patient visit record encompasses several key components: the chief complaints logged by the ED, diagnosis codes (either ICD-9 or ICD-10), at least one radiology report, and a comprehensive discharge summary. The discharge summary includes vital sections "Brief Hospital Course (BHC)" and "Discharge Instruction (DI)". The main objective of this competition is to automate the generation of these two essential sections of the discharge summary (Xu et al., 2024).

To solve this shared task, we propose MEDISCHARGE, a fully automatic system based on Meditron-7B (Chen et al., 2023) with context window extension for generating BHC and DI sections according to the patient's EHR. The model with a longer context window size helps our system process the full text of most long-context cases.

Proceedings of the 23rd Workshop on Biomedical Language Processing, pages 696–711
August 16, 2024. ©2024 Association for Computational Linguistics

Next, we propose a dynamic information selection framework that can improve the robustness of the system since it can prune EHRs with very long context to fit a limited context window size. We conduct a comprehensive evaluation of our system on the full phase II test set. In the competition, our system obtained an overall score of 0.289 on a held-out subset of this test set, improving over the official baseline (0.102) by 183% relatively. We make our code available at https://github.com/HAOTIAN89/MEDISCHARGE.

2 Related Work

Automation of Clinical Text Documentation. With the development of Natural Language Processing (NLP), the automation of clinical documentation has gradually received attention due to its huge application value. At early stages, rule-based NLP approaches have been employed to extract specific information from free-text clinical notes and populate structure fields within the EHR (Meystre et al., 2008; Demner-Fushman et al., 2009). Machine learning and deep learning techniques such as long short-term memory (LSTM) (Hochreiter and Schmidhuber, 1997) and transformers (Vaswani et al., 2017) have shown promise in generating simple clinical summaries (Hu et al., 2020; Ive et al., 2020). The appearance of large language models (LLMs) has brought unprecedented changes (Achiam et al., 2023; Ouyang et al., 2022), and demonstrated potential strong capabilities in clinical text summarization (Van Veen et al., 2023).

Medical Pretrained Large Language Model. The amazing performance of LLMs mainly depends on the large amount of knowledge learned in the pretraining stage. Given the uniqueness of medical knowledge, there is substantial research focused on medically specialized pretrained LLMs. Early work, like BioBert, focused only on pretraining BERT with large-scale biomedical corpora (Lee et al., 2020; Gu et al., 2021). However, the performance of these models was limited by the small scale of the base model. An increasing number of larger medical LLMs have emerged with time, like PMC-LLaMA with 7B and 13B parameters size (Wu et al., 2023), Meditron with 7B and 70B parameters size (Chen et al., 2023), or the model with currently best performance PaLM-2, with 540B parameters size (Anil et al., 2023). In our system, Meditron-7B is selected as the pretrained LLM to do finetuning for medical summarization.

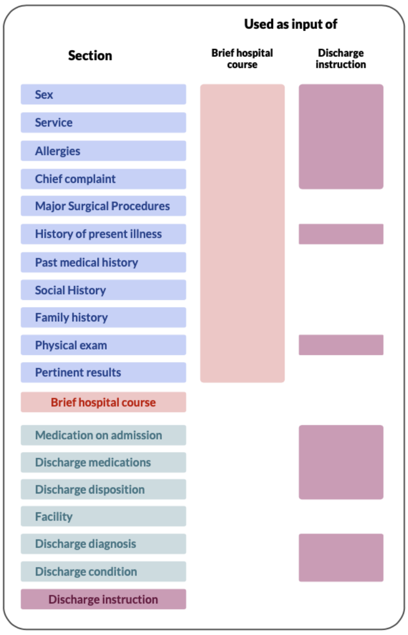

Figure 1: Full discharge text and inputs

Context Window Extension. Currently there are two main popular methods for LLM context window extension, one is Sliding Window Attention (SWA) from Mistral-7B (Jiang et al., 2023), and the other is position interpolation based on Rotary Position Embedding (RoPE) (Su et al., 2024). Although SWA can provide an extensive context window, theoretically, this method faces certain limitations as the model can utilize only a local window of restricted length at any given time. Because of this, RoPE attracted more attention, since its context window is extended actually and easy to use (Kaddour et al., 2023; bloc97, 2023).

3 The MEDISCHARGE System

Our proposed MEDISCHARGE system is an LLM-based system for the automatic generation of discharge summary sections from relevant key components of clinical EHRs (see Fig.2). Our system consists of three parts: (1) Section Extraction, (2) Instruction Fine-tuning Medical LLM, and (3) Robust Inference. We aim to utilize LLMs and refinement techniques to create summaries that ensure factual accuracy in alignment with EHRs and preserve the textual style of clinicians.

3.1 Extraction Method

To streamline the pipeline while achieving a substantial level of performance and efficiency, we design our system to operate on the full discharge summary text (excluding the target BHC and DI).

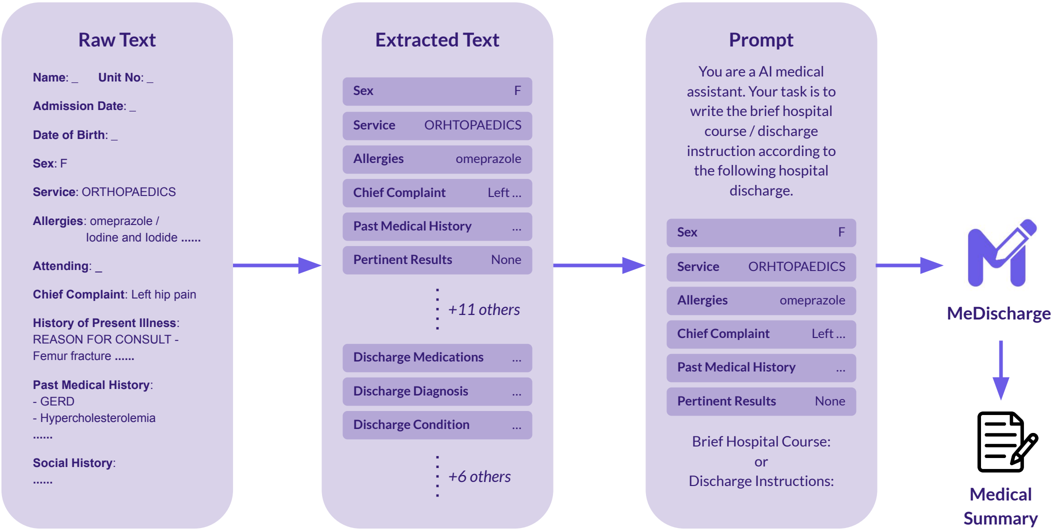

Figure 2: **Overview of MEDISCHARGE**. The raw full discharge text is the system input. First, all useful sections are extracted and combined to form new potential input. If this input is too long, our dynamic information selection framework then refines it by removing sections in increasing order of importance. Finally, the prompt will be put into an instruction fine-tuning Meditron-7B to summarize the BHC and DI, respectively.

Based on its position in the entire EHR, the discharge summary already contains the majority of the information required. Given that LLM inference is very expensive, using the summary also proves to be a more economical approach. This strategy allows us to efficiently utilize the rich features and details of the EHR while keeping computational costs manageable.

We identify 17 main sections (Fig.1) within the full discharge text by grouping consecutive sections and disregarding some sections. The extraction process encounters some challenges due to the inconsistencies in section headers, including variations in capitalization and blank headers. For instance, we find 13 different header variants for the section *Physical Exam*. Additionally, a section may appear twice if it is also a subsection of another. Our final extraction method involves a linearly ordered search of each section within the full text using *regex* matching patterns. A section is delimited by its header and the header of the next section.

We first use specific algorithm to collect all section header candidates (For more details, please see Appendix A). Upon identifying all headers for each section, the extraction process follows a specific paradigm: a section commences at one of its headers and concludes immediately before the headers of the next section, as shown in Algorithm 1. This

Algorithm 1 Algorithm for section extraction

Input: A full text discharge

```
current_discharge $\leftarrow$ full text discharge
$found$ $\leftarrow$ False
discharge_sections $\leftarrow$ { }
start_headers $\leftarrow$ []
$found$ $\leftarrow$ False
for section in all_sections do
    if start_headers is empty then
        start_headers = headers[section]
    for next_section in next_sections[section] do
        for start_header in start_headers do
            for end_header in headers[next_section] do
                s_text $\leftarrow$ find_pattern(
                  start_header ... end_header)
                  in current_discharge
                if s_text then
                    start_header $\leftarrow$ [end_header]
                    $found$ $\leftarrow$ True
                    discharge_sections[section] $\leftarrow$ s_text
                    current_discharge $\leftarrow$
                        current_discharge[(end_header):]
                    Break
                if found then
                    Break
            if found then
                Break
    if not $found$ then
        discharge_sections[section] $\leftarrow$ "None"
    $found$ $\leftarrow$ False
return discharge_sections
```

enables precise extraction of sections while ensuring no loss of text even if a header is not found or even a entire section is ignored.

3.2 Medical LLM with Context Extension

Scaling up language models has been shown to improve performance across numerous downstream tasks. As the size of the model increases, there is a greater chance of reaching a level where the phenomenon of Emergence occurs, where quantitative changes lead to qualitative shifts in behavior (Wei et al., 2022). Thus, by designing a generation system driven by LLMs, we aim to tackle complex shared tasks that are difficult for smaller models. There is also substantial evidence indicating that models pretrained in specific domains significantly outperform general-purpose models in the same domain tasks (Cui et al., 2023; Wu et al., 2023; Yang et al., 2023). Therefore, we select Meditron-7B, which is currently one of the best open-source medically pretrained LLM, with a 7B parameter scale, as our core component of the system for medical text summarization (Chen et al., 2023). We also use the Megatron-LLM, an efficiently distributed LLM trainer, to finetune our model as described in Meditron's technical report (Cano et al., 2023).

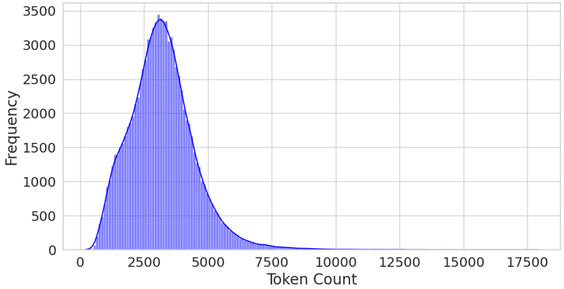

Figure 3: The input token distribution for the whole dataset (train, valid, and test set). Most samples are between 1000 to 5000. The distribution has a long tail that stretches rightwards towards higher token counts.

Additionally, we lift the limitation of the 2K fixed context window of Meditron-7B such that medical electronic files with longer text can fit in the model (Fig.3). We apply position interpolation by manipulating the RoPE positional embedding (Su et al., 2024) to effectively leverage the positional information, increasing the context window from 2K to 6K. The updated model is able to reason over more details of the EHR, effectively reducing the hallucination issue of LLMs and thus generating more factual summary sections.

3.3 Dynamic Information Selection

The dynamic information selection framework plays an important role in robust model inference under diverse cases. Once an LLM is deployed, further updating will be challenging, meaning that the context window size will remain fixed (Gao et al., 2023). When the length of a patient's EHR information exceeds the maximum length that the system can accept, the most important information will be selected to maximize utility. We explore the optimal selection method through behavior-based and result-based analyses and propose our final selection framework based on the findings.

3.3.1 Behavior-Based Analysis

We apply behavior-based analysis to determine relevance. We emulate clinicians' behavior in writing medical summaries to prioritize and select the most informative parts when facing the length limitation. We observe that clinicians often directly copy some important sentences or medical examination data from the EHRs to the summary without any modification (we show some examples in Appendix B). To measure the prevalence of direct reference in different sections, we use ROUGE-2 (Lin, 2004), since it focuses on recall, and computes scores at the word level rather than the embedding level.

$$ROUGE2 = \frac{\sum_{s\in\{Ref\}}\sum_{bi\in s} Count_{match}(bi)}{\sum_{s\in\{Ref\}}\sum_{bi\in s} Count(bi)}$$

To assess the order of importance of the sections that should be included in the BHC input, we compute the average ROUGE-2 score between each of the first 11 sections and the reference BHC. Similarly, we compute the same metric between all 17 sections and the reference DI to figure out the section importance order for the DI input. The higher score reveals which sections have a stronger direct reference to the summary target, meaning clinicians are more likely to refer to these parts when writing summaries.

The results show a clear pyramid-shaped distribution (Fig.4 and 5), where most sections have no direct reference value to the target summary. In contrast, a small number have an obvious reference value. For both BHC and DI, *History of Present Illness* has the highest direct reference score, especially in BHC, where it reaches 8.33. The sections located in the middle of the pyramid have a certain degree of differentiation. *Pertinent Results* and *Past*

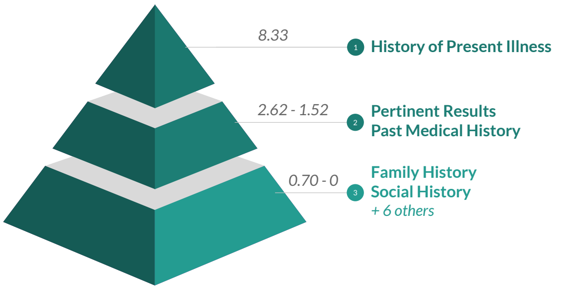

Figure 4: Pyramid of importance order for BHC

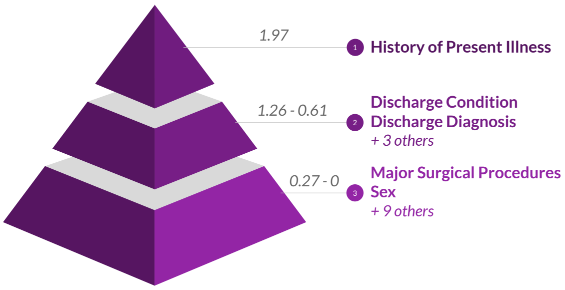

Figure 5: Pyramid of importance order for DI

Medical History are two sections that have a positive direct contribution to the target BHC. For DI, more sections are in the middle. Most of these sections appear after the BHC summary in the original unprocessed full text, such as *Discharge Condition* and *Discharge Diagnosis*. The sections at the bottom of the pyramid are not directly referenced for the writing of the summaries, so their priority will be lowered in the final decision of which sections to include. Full table results are in appendix C.

3.3.2 Result-Based Analysis

To better drive the dynamic information selection, we perform an ablation study to assess the influence of excluding specific sections on the performance metrics of discharge summary generations within the MEDISCHARGE system. Table 1 and 2 present the performance variations when compared to a baseline method that utilizes all sections. The experiment is carried out on a subset of dataset using Meditron-7B with 6K context window extension. These results are instrumental in developing a robust section selection strategy for optimizing system performance in constrained scenarios.

Marginal changes (less than 1%) in overall score may not reliably signify an impact from section removal due to the inherent variability in model performance and the small effect size. However, these minor variations still provide a qualitative understanding of section importance. Notably, sections such as *Physical Exam*, *Pertinent Results*, *History of Present Illness*, and *Discharge Diagnosis* exhibit large negative gains when omitted, ranging from -1.08% to -6.99% as shown in both tables. This suggests a substantial contribution of these sections to the overall accuracy and completeness of the generated discharge summaries.

	Sections removed	Overall	Gain (%)
1	past medical history	0.2414	0.53
2	family history	0.2413	0.46
3	social history	0.2406	0.2
4	-	0.2401	0
5	physical exam	0.2375	-1.08
6	major surgical procedures	0.2318	-3.47
7	pertinent results	0.2293	-4.53
8	history of present illness	0.2262	-5.82

Table 1: BHC overall score gains compared to the baseline depending on the sections removed

	Sections removed	Overall	Gain (%)
1	-	0.2870	0
2	medication on admission	0.2853	-0.59
3	discharge disposition	0.2832	-1.32
4	history of present illness	0.2829	1.41
5	discharge medications	0.2736	-4.67
6	discharge condition	0.2714	-5.42
7	physical exam	0.2713	-5.47
8	discharge diagnosis	0.2669	-6.99

Table 2: DI overall score gains compared to the baseline depending on the sections removed

Thus, while minor gains or losses might not constitute statistical significance, they do establish a hierarchy of importance among the sections. Sections leading to high negative gains, when omitted, are evidently crucial and should be prioritized in the dynamic information selection framework of the MEDISCHARGE system, particularly when operating under limitations such as fixed context windows or partial data availability.

3.4 Final Decision

Combining the results of the behavior-based and result-based analyses, we rank the sections by their importance in Table 4. Following this order, MEDISCHARGE extracts and integrates the important sections into the input. If the combination of the sections exceeds the model's context window size, sections with lower importance are removed based on the rank until the input can fit into the model. For details on the dynamic information

	BLEU-4	ROUGE-1	ROUGE-2	ROUGE-L	BERTScore	METEOR	AlignScore	MEDCON	Overall
BHC-Methods					**Brief Hospital Course**				
Llama2-7b-2k	0.025	0.304	0.068	0.132	0.246	0.195	0.199	0.628	0.225
Meditron-7b-2k	0.040	0.322	0.098	0.165	0.300	0.183	0.223	0.645	0.247
Meditron-7b-2k-s	0.050	0.353	0.115	0.185	0.333	0.201	0.232	0.666	0.267
Meditron-7b-d6k	0.044	0.353	0.113	0.185	0.341	0.188	0.222	0.668	0.264
Meditron-7b-i6k	0.061	0.380	0.121	0.185	0.349	**0.243**	0.245	0.696	0.285
MEDISCHARGE	**0.061**	**0.381**	**0.121**	**0.186**	**0.351**	0.242	**0.246**	**0.697**	**0.286**
DI-Methods					**Discharge Instructions**				
Llama2-7b-2k	0.026	0.270	0.062	0.130	0.189	0.222	0.222	0.536	0.207
Meditron-7b-2k	0.061	0.362	0.138	0.226	0.345	0.232	0.282	0.633	0.285
Meditron-7b-2k-s	0.088	0.418	0.177	0.268	0.402	0.281	0.341	0.674	0.331
Meditron-7b-d6k	0.074	0.400	0.170	0.265	0.399	0.239	0.337	0.658	0.318
Meditron-7b-i6k	0.099	0.416	0.186	0.275	0.402	0.285	0.363	0.670	0.337
MEDISCHARGE	**0.103**	**0.428**	**0.194**	**0.284**	**0.417**	**0.290**	**0.370**	**0.683**	**0.346**

Table 3: **The performance of our system with different methods on the full Test Phase II set**. *Llama2-7b* and *Meditron-7b* refer to the base models in our system. *2k*, *d6k* and *l6k* show the maximum sequence input of the model, where *d6k* means using "Dynamic NTK" interpolation method and *i6k* means using linear interpolation method, both to extend the context window to 6K. *s* refers to the proposed dynamic information selection framework. Otherwise, it uses a simple truncation strategy.

selection algorithm, please see the appendix D.

BHC	**DI**
sex	sex
service	service
chief complaint	chief complaint
history of present illness	discharge diagnosis
pertinent results	discharge condition
physical exam	discharge medications
major surgical procedures	physical exam
allergies	history of present illness
family history	discharge disposition
social history	medication on admission
past medical history	

Table 4: Importance section orders for BHC and DI. We just put *sex*, *service* and *chief complaint* on the top because they are always very short.

4 Experiments

4.1 Experimental Setups

We utilize the shared task dataset derived from MIMIC-IV's submodules, i.e., MIMIC-IV-Note (Johnson et al., 2023b), and MIMIC-IV-ED (Johnson et al., 2023a). In the dataset, each patient's visit information is represented by a unique number, which is associated with several medical records. The dataset comprises four subsets: training, validation, phase I testing, and phase II testing. Details on the subsets are listed in Table 5. We use the phase II testing dataset to evaluate our system.

We adopt the evaluation metrics suggested by the organizers of the competition. We use BLEU-4 (Papineni et al., 2002), ROUGE-1, -2, -L (Lin, 2004), BERTScore (Zhang et al., 2019) and METEOR (Banerjee and Lavie, 2005) as basic metrics to measure the similarity between our generated text and the ground truth. We also use AlignScore (Zha et al., 2023) and MEDCON (Yim et al., 2023) as task-specific metrics. AlignScore checks whether the generated text is factually consistent with the medical records, and MEDCON is a medical-concept-based metric to gauge the accuracy and consistency of clinical concepts.

Dataset	**Samples**	**Competition**	**Paper**
Training	68,875	Yes	Yes
Validation	14,719	Yes	Yes
Testing I	14,702	Yes	No
Testing II	10,962	Yes	Yes

Table 5: Dataset summary

4.2 Training Details

We experiment with *Llama2-7B* and *Meditron-7B* with and without linear extension. We train all of them on samples whose lengths are within the models' context windows. The main hyper-parameters are identical for the first two models. We set the max_length $= 2048$, use an AdamW optimizer with $\beta_1 = 0.9, \beta_2 = 0.95$, eps $= 10^{-5}$ and cosine learning rate schedule with 10% warmup ratio and learning rate of 2×10^{-5}, weight decay 0.1, micro_batch_size 8 and macro_batch_size 64 for 3 epochs. For the linear extension one, we increase max_length to 6144, and reduce the micro_batch_size from 8 to 2 due to the limited GPU

VRAM. All training runs are on 8 NVIDIA A100 80G GPUs.

4.3 Results

We show our system's main performance on generating BHC and DI in Table 3. For both BHC and DI generation, our proposed system MEDISCHARGE (Meditron-7B employs a linear extension to 6K with a dynamic information selection framework) outperforms the baseline with a large margin across all metrics, showing 27% and 67% relative improvements on BHC and DI respectively. Under the same configurations, the medical LLM (Meditron-7B) outperforms the general LLM (Llama2-7B) in fine-tuning tasks. Especially, the performance on DI generation increases by 38%. Dynamic and linear context window extensions both have significant increases on two tasks, and the linear one always be better (0.021 absolute gap in BHC and 0.019 in DI between two methods). Our results also suggest our proposed dynamic information selection framework is more beneficial than direct truncation when the length of the original full text is larger than the model's context window size. We show that this method improves both BHC (8% increase) and DI (16% increase) performances. Note that in DI, our selection framework even achieves a larger improvement (59.9%) than dynamic context window extension (53.6%). However, we observe that applying dynamic information selection to a model with a 6K context window shows marginal improvement. We hypothesize the benefit can be limited because a 6K context window can process most of the full text.

4.4 Section Selection Analysis

Here, we analyze the difference in performance between our dynamic information selection and the truncation method for the 2K and 6K context windows. Note that for each task, the truncation method cuts the full input text (see Fig.1) starting from the end until it fits the max input length.

4.4.1 Discharge Instruction

In Figure 6 for the dynamic information selection to DI, almost all sections are selected under a 6k context window. But for the model with only a 2K context window, the dynamic information selection works heavily, where it generates a total of 127 different kinds of section combinations on all test sets. The discharge input sections are mostly at the end of the full input text (where the truncation starts), which explains well why a heavy truncation has a greater effect on the DI generation model performance (both 2K and 6K) in Table 3.

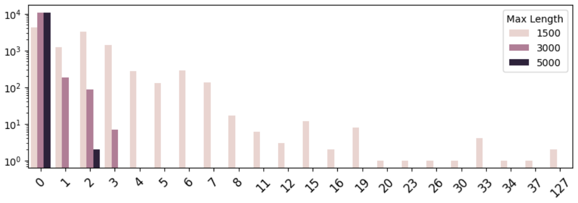

Figure 6: The number of section combinations (log scale) happened in the DI generation.

4.4.2 Brief Hospital Course

As shown in Figure 7, the dynamic information selection always generates three kinds of section combination (the first, 32nd, and 33rd ones) for the model with 6K context window, which actually is all sections, all sections without *physical exam* and *pertinent results* respectively. Since these sections are at the end of the brief hospital course input, it makes sense that these combinations have a similar effect as direct truncation. Therefore the slight performance differences between truncation and dynamic information selection are to be expected.

On the other hand, for the 2K context window on the BHC generation, the section combinations are more spread out (see Fig.7). The sections kept mostly by our framework are the *pertinent results* and *physical exam*. However, both of them are easy to remove under simple truncation as they are at the end of our full-text input. Since these are the most important sections for BHC according to Table 3, it well explains why the dynamic information selection outperforms truncation for the model with a 2K context window.

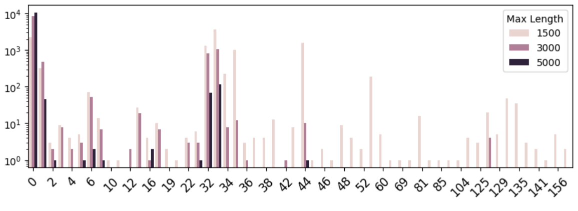

Figure 7: The number of section combinations (log scale) happened in the BHC generation.

4.5 Human Evaluation

We also do the human evaluation with the help of three clinicians on the high quality and representative 25 samples selected by the shared task organizers. Our generated sections will be evaluated for their *Completeness*, *Correctness*, and *Readability*. The detailed criteria is shown in the appendix E, and the results are in table 6.

	Completeness	Correctness	Readability
BHC	**3.5**	**3**	**2.5**
DI	**3.5**	**3.5**	**NA**

Table 6: Human evaluation result. The score is from 1 to 5, and we adjust the accuracy to 0.5 for easy reading. The Readability of Discharge Instruction aims for patients, so we didn't consider it in this evaluation.

In the human evaluation, the BHC and DI generated by our system achieve good performance in both *Completeness* and *Correctness*, which means that our generated text contains most of the important information from the EHR, and the incorrect content will not impact future care most of the time. Additionally, the score of *Readability* in BHC shows our text sometimes is slightly harder to read than the ground truth, which is most likely due to the lack of reinforcement learning of human feedback on our model. We would like to improve it by DPO (Rafailov et al., 2024) if having chances to get a suitable dataset. In summary, this result shows that the current system could replace clinicians in discharge summary writing to a certain extent and has great potential to completely free them from this time-consuming task in the future.

5 Qualitative Analysis

In addition to quantitative analysis based on different metrics, we also review specific cases manually to intuitively assess whether the generated summaries are of high quality and could potentially aid or substitute clinicians in real-life scenarios. Here is one DI generated example, and more example cases are in the appendix F.

> **Ground Truth**: Dear Mr. ___, It was a pleasure participating in your care at ___. You were admitted to the hospital with chest pain from a heart attack. You had stents placed in your coronary arteries and your symptoms improved. It is very important that you take all of your medications as prescribed and follow up with your primary care doctor and your new cardiologist as scheduled. It is extremely important that you continue to take your aspirin and plavix every day until otherwise instructed by your cardiologist in order to prevent your stents from closing.

> **Generated Text**: Dear Mr. ___, It was a pleasure taking care of you at ___. You were admitted with chest pain and found to have had a heart attack. You were taken to the cardiac catheterization lab where two blockages in your coronary arteries were found and stents were placed. It is very important that you take aspirin and plavix every day for at least one year after your procedure. Do not stop these medicines or miss any doses unless Dr. ___ that it is OK to do so. It is also very important that you quit smoking as this will help your heart heal after the procedure and prevent future heart attacks.

For DI, both texts address the key elements of the discharge information, including the most relevant details to this case, i.e. the patient's heart attack and placement of stents in the coronary arteries. They emphasize the importance of continuing medication, specifically mentioning aspirin and plavix, which are crucial for preventing clot formation on the stents. However, the generated text provides more detailed guidance on medication duration and lifestyle changes than the ground truth, which could potentially enhance patient compliance and outcomes.

6 Conclusion

The research presented in this paper highlights the significant advancements made by MEDISCHARGE system in the field of automated discharge summary generation at ACL'24 BioNLP Shared Task on Streamlining Discharge Documentation. The experiment results demonstrate that our system with efficient information usage and good costs management achieves a great performance improvement of 183% compared to the baseline, and is able to efficiently generate concise and medically accurate discharge summaries, markedly reducing the burden on healthcare professionals. The adoption of an LLM, specifically pretrained for medical data, can better complete medical summarization tasks than a general fundamental LLM. Furthermore, the dynamic information selection framework we proposed shows a robust task inference ability, significantly outperforming the simple truncation strategy and even dominating the context window extension method across several NLP metrics.

7 Limitation

While our proposed MEDISCHARGE framework has demonstrated significant achievements, we argue that there several some limitations should be noticed.

Currently, MEDISCHARGE is designed to process and generate summaries only in English. This restricts its applicability in diverse linguistic settings, which is critical in global healthcare environments where multilingual support could enhance both the utility and accessibility of automated discharge summaries.

Due to the high costs associated with human annotation, our evaluation of the model's output through clinician reviews is limited to only 25 specific samples selected by competition organizers. This sample size may not fully represent the model's performance across a wider range of discharge summaries.

Another concern is that the current implementation of MEDISCHARGE is limited to producing text-based documents only. It does not have the capability to integrate or produce image-based content, which can be an essential component of certain medical summaries, such as those including anatomical diagrams or graphical patient data.

Acknowledgments

We thank the anonymous reviewers for their thoughtful comments and Syrielle Montariol for her helpful proofreading. This work is supported by EPFL NLP lab and MAKE projects. We also thank the SWISS AI Initiative (https://www.swiss-ai.org/) for providing powerful computing infrastructure.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Rohan Anil, Andrew M Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, et al. 2023. Palm 2 technical report. *arXiv preprint arXiv:2305.10403*.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

bloc97. 2023. NTK-Aware Scaled RoPE allows LLaMA models to have extended (8k+) context size without any fine-tuning and minimal perplexity degradation.

Alejandro Hernández Cano, Matteo Pagliardini, Andreas Köpf, Kyle Matoba, Amirkeivan Mohtashami, Xingyao Wang, Olivia Simin Fan, Axel Marmet, Deniz Bayazit, Igor Krawczuk, Zeming Chen, Francesco Salvi, Antoine Bosselut, and Martin Jaggi. 2023. epfllm megatron-llm.

Zeming Chen, Alejandro Hernández Cano, Angelika Romanou, Antoine Bonnet, Kyle Matoba, Francesco Salvi, Matteo Pagliardini, Simin Fan, Andreas Köpf, Amirkeivan Mohtashami, et al. 2023. Meditron-70b: Scaling medical pretraining for large language models. *arXiv preprint arXiv:2311.16079*.

Jiaxi Cui, Zongjian Li, Yang Yan, Bohua Chen, and Li Yuan. 2023. Chatlaw: Open-source legal large language model with integrated external knowledge bases. *arXiv preprint arXiv:2306.16092*.

D Demner-Fushman, WW Chapman, and CJ McDonald. 2009. What can natural language processing do for clinical decision support? *Journal of Biomedical Informatics*, 42(5):760–772.

Adam Gaffney, Stephanie Woolhandler, Christopher Cai, David Bor, Jessica Himmelstein, Danny McCormick, and David U. Himmelstein. 2022. Medical Documentation Burden Among US Office-Based Physicians in 2019: A National Study. *JAMA Internal Medicine*, 182(5):564–566.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, and Haofen Wang. 2023. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*.

Emily Gesner, Priscilla Gazarian, and Patricia Dykes. 2019. The burden and burnout in documenting patient care: An integrative literature review. *Studies in health technology and informatics*, 264:1194—1198.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2021. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare (HEALTH)*, 3(1):1–23.

Patricia Hay, Kathy Wilton, Jennifer Barker, Julie Mortley, and Megan Cumerlato. 2020. The importance of clinical documentation improvement for Australian hospitals. *Health Information Management Journal*, 49(1):69–73.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

B Hu, A Bajracharya, and H Yu. 2020. Generating medical assessments using a neural network model: Algorithm development and validation. *JMIR Medical Informatics*, 8(1):e14971.

Julia Ive, Natalia Viani, Joyce Kam, Lucia Yin, Somain Verma, Stephen Puntis, Rudolf N Cardinal, Angus Roberts, Robert Stewart, and Sumithra Velupillai. 2020. Generation and evaluation of artificial mental health records for natural language processing. *NPJ digital medicine*, 3(1):69.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Alistair Johnson, Lucas Bulgarelli, Tom Pollard, Leo Anthony Celi, Roger Mark, and Steven Horng. 2023a. Mimic-iv-ed. PhysioNet.

Alistair Johnson, Tom Pollard, Steven Horng, Leo Anthony Celi, and Roger Mark. 2023b. Mimic-iv-note: Deidentified free-text clinical notes. PhysioNet.

Jean Kaddour, Joshua Harris, Maximilian Mozes, Herbie Bradley, Roberta Raileanu, and Robert McHardy. 2023. Challenges and applications of large language models. *arXiv preprint arXiv:2307.10169*.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

SM Meystre, GK Savova, KC Kipper-Schuler, and JF Hurdle. 2008. Extracting information from textual documents in the electronic health record: a review of recent research. *Yearbook of Medical Informatics*, pages 128–144.

Amanda J Moy, Jessica M Schwartz, RuiJun Chen, Shirin Sadri, Eugene Lucas, Kenrick D Cato, and Sarah Collins Rossetti. 2021. Measurement of clinical documentation burden among physicians and nurses using electronic health records: a scoping review. *Journal of the American Medical Informatics Association*, 28(5):998–1008.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2024. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36.

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. 2024. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063.

Xiangru Tang, Anni Zou, Zhuosheng Zhang, Yilun Zhao, Xingyao Zhang, Arman Cohan, and Mark Gerstein. 2023. Medagents: Large language models as collaborators for zero-shot medical reasoning. *arXiv preprint arXiv:2311.10537*.

Aykut Uslu and Jurgen Stausberg. 2021. Value of the Electronic Medical Record for Hospital Care: Update From the Literature. *Journal of medical Internet research*, 23(12).

Dave Van Veen, Cara Van Uden, Louis Blankemeier, Jean-Benoit Delbrouck, Asad Aali, Christian Bluethgen, Anuj Pareek, Malgorzata Polacin, Eduardo Pontes Reis, Anna Seehofnerova, et al. 2023. Clinical text summarization: Adapting large language models can outperform human experts. *Research Square*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, et al. 2022. Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682*.

Chaoyi Wu, Xiaoman Zhang, Ya Zhang, Yanfeng Wang, and Weidi Xie. 2023. Pmc-llama: Further finetuning llama on medical papers. *arXiv preprint arXiv:2304.14454*.

Justin Xu, Andrew Johnston, Zhihong Chen, Louis Blankemeier, Maya Varma, Curtis Langlotz, and Jean-Benoit Delbrouck. 2024. Overview of the first shared task on clinical text generation: Rrg24 and "discharge me!". In *The 23rd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks*, Bangkok, Thailand. Association for Computational Linguistics.

Hongyang Yang, Xiao-Yang Liu, and Christina Dan Wang. 2023. Fingpt: Open-source financial large language models. *arXiv preprint arXiv:2306.06031*.

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. *Scientific Data*, 10(1):586.

Yuheng Zha, Yichi Yang, Ruichen Li, and Zhiting Hu. 2023. Alignscore: Evaluating factual consistency with a unified alignment function. *arXiv preprint arXiv:2305.16739*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

A Extraction Method Paradigm

The extraction of the sections proves challenging. It first requires an iterative identification of the different section headers as shown in the algorithm 2. We perform this run on a subset of the discharges only and hence we may have missed some header. As an example here are the different headers we find for the section *Discharge Medications*: `['Discharge Medications:', 'Discharge medications:', '___ Medications:', '___ medications:']` ; here the *Section Basic Name* is `['Discharge Medications:'` and is the first header we consider for this section.

Algorithm 2 Algorithm for section header identification

```
Input: A section
section_headers ← [Section Basic Name]
found ← False
for discharge in all_discharges do
    for header in section_headers do
        if header in discharge then
            found ← True
            break
    if not found then
        Manually look for a new section header
        if new_header found then
            Add new_header to section_headers
    found ← False
return section_headers
```

B Direct Copy Examples in Summary

The examples provided below demonstrate how some text from the original raw sections is integrated in the BHC with minimal to none modifications. The raw text included for both examples are taken from the History of present illness section.

Example 1:

Raw text: This is a ___ yo f with h/o recently diagnosed metastatic cancer of unknown prior presenting with nausea, vomiting, and fever to 101 today.

Ground Truth BHC: ___ yo f with h/o recently diagnosed metastatic cancer of unknown primary presenting with nausea, vomiting, and fever to 101 on day of admission.

Example 2:

Raw text: ___ with HTN, HLD, & recurrent SVT on Flecainade/Toprol p/w CP/SOB and lightheadedness, found to be hypotensive with intermittent SVT without ischemic EKG changes or positive biomarkers, now admitted to the CCU for planned EP ablation.

Ground Truth BHC: ___ with HTN, HLD, & recurrent SVT on Flecainade/Toprol who presented with CP/SOB and lightheadedness, found to be hypotensive with intermittent SVT without ischemic EKG changes or positive biomarkers, admitted to the CCU for planned EP ablation.

C Full Tables of Direct Reference

D Dynamic Section Selection Algorithm

Extracting all sections enables intentional selection of sections for inclusion in our input, promoting consistency. By choosing a predefined set of sections, we ensure adherence to a standardized order (as shown in Fig.1) and consistent headers, which contrasts with the inherent variability of raw text. Furthermore, it establishes a consistent method for indicating missing sections: using

Algorithm 3 Algorithm for dynamic section selection

```
Input: All extracted sections, Section importance list, max length
$extract \leftarrow$ All extracted sections
$importance \leftarrow$ Section importance list
$max \leftarrow$ max length
Tokenize each section in $extract$
$total\_Length \leftarrow$ sum of tokenized section lengths in $extract$
if $total\_Length \leq max$ then
    return $extract$
else
    return TRY($extract$, $importance$, $max$, [])
procedure TRY($Allsections$, $importanceList$, $max\_length$, $removedSoFar$)
    for $i$ = length of $importanceList - 1$ to 0 step $-1$ do
        $currentSection \leftarrow importanceList[i]$
        $newRemovedList \leftarrow removedSoFar + [currentSection]$
        $remainingSections \leftarrow Allsections$ excluding $newRemovedList$
        $newTotalLength \leftarrow$ sum of tokenized section lengths in $remainingSections$
        if $newTotalLength \leq max\_length$ then
            return $remainingSections$
        TRY($Allsections$, $importanceList[0:i]$, $max\_length$, $newRemovedList$)
```

Section	ROUGE-2
History of Present Illness	0.01967
Discharge Condition	0.01263
Discharge Diagnosis	0.00939
Discharge Medications	0.00777
Pertinent Results	0.00673
Chief Complaint	0.00613
Past Medical History	0.00274
Physical Exam	0.00270
Major Surgical Procedures	0.00217
Medication on Admission	0.00173
Family History	0.00089
Discharge Disposition	0.00029
Allergies	0.00005
Social History	0.00004
Facility	0.00000
Service	0.00000
Sex	0.00000

Table 7: DI Direct Reference

’`Header:\nNone\n`’ instead of various representations like ’___’, empty spaces, or simply the absence of the header commonly found in raw inputs. We then create our input by concatenating the desired sections. Even if a section is not chosen for inclusion in specific samples but was generally included for the subsequent experiment (like we do in strategy selection), ’`Section Header:\nNone\n`’ is still included at the right spot to maintain consistency in input structure.

E Human Evaluation Criteria

The details of human evaluation criteria are here.

Section	ROUGE-2
History of Present Illness	0.08329
Pertinent Results	0.02621
Past Medical History	0.01515
Physical Exam	0.00702
Major Surgical Procedures	0.00575
Chief Complaint	0.00538
Family History	0.00180
Social History	0.00011
Allergies	0.00004
Sex	0.00000
Service	0.00000

Table 8: BHC Direct Reference

- **Completeness (captures important information)**
  - Captures no important information (1)
  - Captures ∼25% of the important information (2)
  - Captures ∼50% of the important information (3)
  - Captures ∼75% of the important information (4)
  - Captures all of the important information (5)
- **Correctness (contains less false information)**
  - Contains harmful content that will definitely impact future care (1)
  - Contains incorrect content that is likely to impact future care (2)

  - Contains incorrect content that may or may not impact future care (3)
  - Contains incorrect content that will not impact future care (4)
  - Contains no incorrect content (5)

- **Readability**
  - Significantly harder to read than the reference text (1)
  - Slightly harder to read than the reference text (2)
  - Neither easier nor harder to read than the reference text (3)
  - Slightly easier to read than the reference text (4)
  - Significantly easier to read than the reference text (5)

F Generated Examples

BHC Generated Example

Ground Truth: The patient presented to the emergency department and was evaluated by the orthopedic surgery team. The patient was found to have a right femoral neck fracture. Given his progressive pain, decreased functionality, and the operative nature of his injury, L discussion was had with the patient using an interpreter regarding the risk and benefit of the surgery. After discussing the operation with his family he understood not only the risks, but also the benefits that he could expect from this operation. He was then admitted to the orthopedic surgery service. The patient was taken to the operating room on ___ for closed reduction and percutaneous pinning of the right femoral neck, which the patient tolerated well. For full details of the procedure please see the separately dictated operative report. The patient was taken from the OR to the PACU in stable condition and after satisfactory recovery from anesthesia was transferred to the floor. The patient was initially given IV fluids and IV pain medications, and progressed to a regular diet and oral medications by POD#1. Despite narcotic medications being made available to him, he manage his pain control with only acetaminophen. The patient was given ___ antibiotics and anticoagulation per routine. The patient's home medications were continued throughout this hospitalization. The patient worked with ___ who determined that discharge to **** was appropriate. The ___ hospital course was otherwise unremarkable.

Generated Text: The patient presented to the emergency department and was evaluated by the orthopedic surgery team. The patient was found to have a right femoral neck fracture and was admitted to the orthopedic surgery service. The patient was taken to the operating room on ___ for closed reduction and percutaneous pinning of his right femoral neck, which the patient tolerated well. For full details of the procedure please see the separately dictated operative report. The patient was taken from the OR to the PACU in stable condition and after satisfactory recovery from anesthesia was transferred to the floor. The patient was initially given IV fluids and IV pain medications, and progressed to a regular diet and oral medications by POD#1. The patient's home medications were continued throughout this hospitalization with exception of his home dose of prednisone, which he will resume upon discharge as he is no longer in-house. The ___ hospital course was otherwise unremarkable with no acute events occurring during this admission outside of his operation.

DI Generated Example

Ground Truth: Dear Ms. ___, It was a pleasure taking care of you at ___. You came to us because of shortness of breath which you had been intermittently experiencing since ___. Your blood tests did not show anything concerning for active ischemia. Your exam was not suggestive of acute heart failure. We suspect that your shortness of breath is related to a leaky valve and that improving blood flow through the arteries of the heart with a medication called isosrbide mononitrate will improve muscular function and the leakiness of the valve. You should take this medication daily. Please be sure to keep your upcoming appointment with Dr. ___ week. We noticed that your blood pressure has been high while you were here with us. We restarted your home lisinopril, now that your kidney function has improved back to normal. Sincerely, Your ___ Care Team.

Generated Text: Dear Ms. ___, You were admitted to the hospital because you were having shortness of breath. We did a stress test which was normal and did not show any signs of worsening heart disease. We also did an ultrasound of your heart which showed that your heart is pumping normally. We think that your shortness of breath may be related to narrowing in the arteries in your legs or to spasm in the arteries in your legs or lungs. For this we started you on a new medication called Imdur (isosorbide mononitrate) which can help with these symptoms. Please follow up with Dr ___ as listed below.

Models	BLEU-4	ROUGE-1	ROUGE-2	ROUGE-L	BERTScore	METEOR	AlignScore	MEDCON	Overall
Llama2-7b-2k	0.025	0.287	0.065	0.131	0.218	0.209	0.211	0.582	0.216
Meditron-7b-2k	0.050	0.342	0.118	0.196	0.323	0.207	0.253	0.639	0.266
Meditron-7b-2k-s	0.069	0.385	0.146	0.227	0.367	0.241	0.287	0.670	0.299
Meditron-7b-d6k	0.059	0.376	0.141	0.225	0.370	0.214	0.280	0.663	0.291
Meditron-7b-l6k	0.080	0.398	0.153	0.230	0.376	0.264	0.304	0.683	0.311
Meditron-7b-l6k-s	**0.082**	**0.405**	**0.157**	**0.235**	**0.384**	**0.266**	**0.308**	**0.690**	**0.316**

Table 9: Global Models Results on the full Test Phase II set

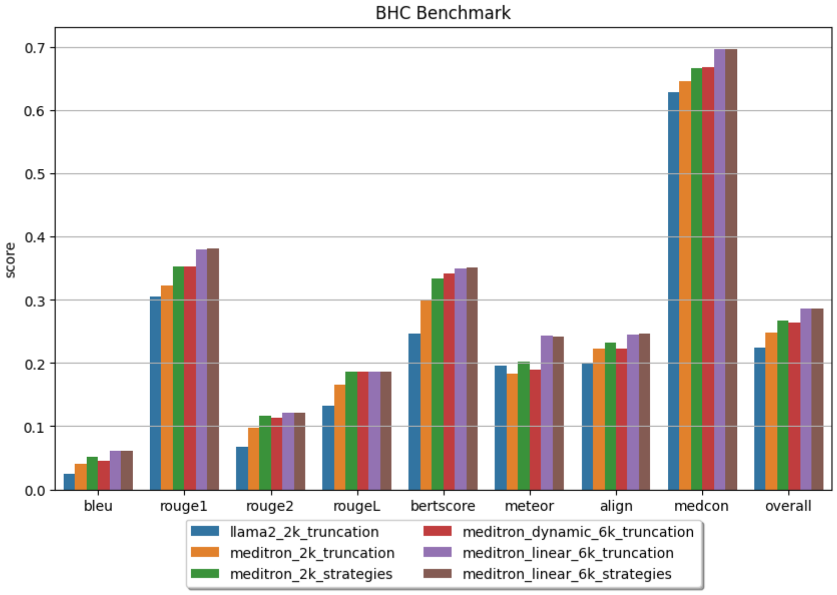

Figure 8: BHC Results on the full Test Phase II set

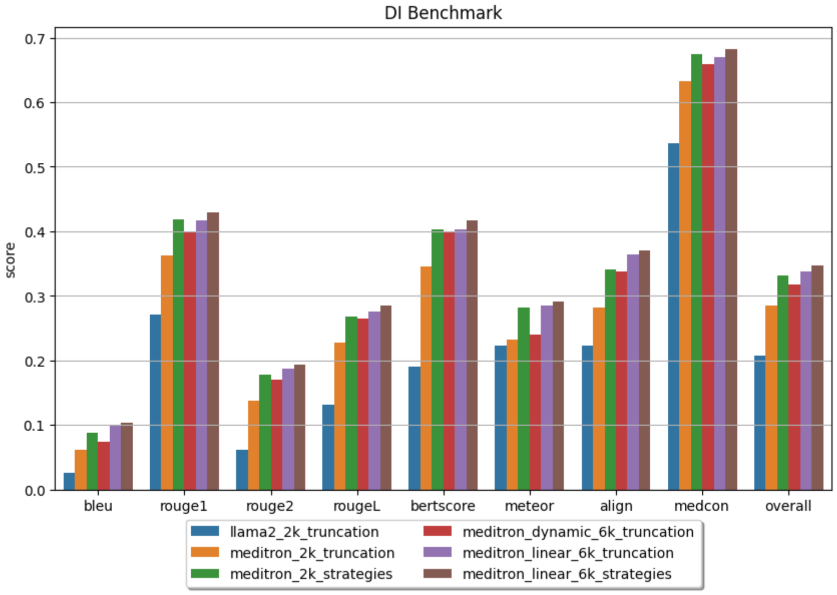

Figure 9: DI Results on the full Test Phase II set

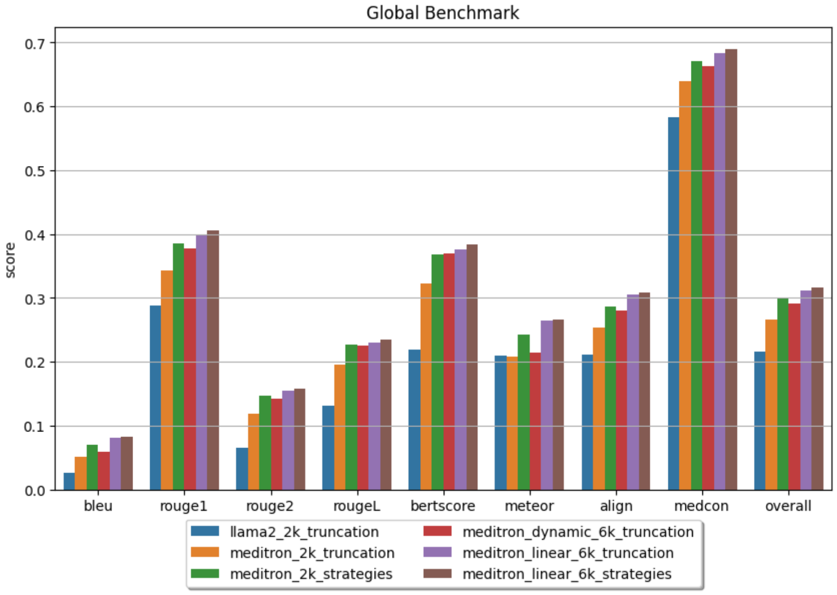

Figure 10: Global Results on the full Test Phase II set