WisPerMed at "Discharge Me!": Advancing Text Generation in Healthcare with Large Language Models, Dynamic Expert Selection, and Priming Techniques on MIMIC-IV

Hendrik Damm[1,2],   Tabea M. G. Pakull[3,1],   Bahadır Eryılmaz[4,5],
Helmut Becker[4],   Ahmad Idrissi-Yaghir[1,2],   Henning Schäfer[3,1],
Sergej Schultenkämper[6], and Christoph M. Friedrich[1,2]

[1] Department of Computer Science, University of Applied Sciences and Arts Dortmund, Dortmund, Germany
[2] Institute for Medical Informatics, Biometry and Epidemiology (IMIBE), University Hospital Essen, Essen, Germany
[3] Institute for Transfusion Medicine, University Hospital Essen, Essen Germany
[4] Institute for AI in Medicine (IKIM), University Hospital Essen, Essen Germany
[5] Institute of Diagnostic and Interventional Radiology and Neuroradiology, University Hospital Essen, Essen, Germany
[6] Bielefeld University of Applied Sciences and Arts, Bielefeld, Germany
{hendrik.damm, christoph.friedrich}@fh-dortmund.de

Abstract

This study aims to leverage state of the art language models to automate generating the "Brief Hospital Course" and "Discharge Instructions" sections of Discharge Summaries from the MIMIC-IV dataset, reducing clinicians' administrative workload. We investigate how automation can improve documentation accuracy, alleviate clinician burnout, and enhance operational efficacy in healthcare facilities. This research was conducted within our participation in the Shared Task Discharge Me! at BioNLP @ ACL 2024. Various strategies were employed, including Few-Shot learning, instruction tuning, and Dynamic Expert Selection (DES), to develop models capable of generating the required text sections. Utilizing an additional clinical domain-specific dataset demonstrated substantial potential to enhance clinical language processing. The DES method, which optimizes the selection of text outputs from multiple predictions, proved to be especially effective. It achieved the highest overall score of 0.332 in the competition, surpassing single-model outputs. This finding suggests that advanced deep learning methods in combination with DES can effectively automate parts of electronic health record documentation. These advancements could enhance patient care by freeing clinician time for patient interactions. The integration of text selection strategies represents a promising avenue for further research.

1 Introduction

Clinical notes in electronic health records (EHRs) are used by clinicians to document patient progress in free-text format. These notes typically include the patient's experiences, symptoms, findings, diagnoses, and details of procedures and interventions performed. They serve as the foundation for Discharge Summaries (DS), which contain a section with concise overviews of the entire hospital encounter known as Brief Hospital Course (BHC) (Searle et al., 2023). They are embedded in the DS and are written by senior physicians who are responsible for the patient's overall care. In addition to BHC, DS also includes Discharge Instructions (DI), which are detailed guidelines provided to patients regarding their post-hospital care. These instructions cover the patient's ongoing care, such as medication instructions, follow-up appointments, and any necessary lifestyle adjustments to ensure proper recovery. Discharge Instructions are designed to facilitate a smooth transition from hospital care to home care and to prevent readmissions. Writing such summaries (BHC) and instructions (DI) can be time-consuming and tedious. Consequently, physicians often spend a big portion of their clinical day dedicated to EHR documentation and desk work (Sinsky et al., 2016).

This paper presents WisPerMed's contribution to the Shared Task Discharge Me! (Xu et al., 2024b), which is part of BioNLP @ ACL 2024. This Shared Task aims to ease the administrative burden on clinicians by developing automated methods to generate critical sections in DS, specifically the "Brief Hospital Course" and "Discharge Instruction". Automating the creation of these sections has the potential to improve documentation accuracy, reduce

105

clinician burnout, and ultimately optimize the processes in healthcare facilities (Patel and Lam, 2023) by allowing clinicians to allocate more time toward direct patient care.

Our work focuses on designing and implementing various innovative approaches to overcome this challenge and contribute to the overall goals of the Shared Task.

2 Dataset

The dataset (Xu, 2024) provided for this Shared Task utilizes the MIMIC-IV (Medical Information Mart for Intensive Care) database (Johnson et al., 2023a,b). MIMIC-IV is a publicly available database sourced from the EHR of the Beth Israel Deaconess Medical Center and is accessible on PhysioNet (Goldberger et al., 2000).

The task dataset is divided into four subsets: a training set consisting of 68,785 samples, a validation set containing 14,719 samples, a phase I testing set with 14,702 samples, and a phase II testing set comprising 10,962 samples. Each subset includes DS that are organized into various sections. All records contain two mandatory sections: "Brief Hospital Course" and "Discharge Instructions". The BHC section typically provides an overview of the patient's treatment and progress during their hospital stay and precedes the DI section. These DI summarize post-hospitalization care instructions and are positioned at the conclusion of the summary.

The challenge organizers provided a regular expression (regex) query to extract these two sections from the DS. The regex query ensures that the relevant information is accurately identified and separated from the rest of the DS content.

For the remainder of this paper, any reference to the "Discharge Summary" (DS) will exclude the target sections, BHC or DI.

3 Evaluation

The submissions to the Shared Task were evaluated using eight metrics, which assess the relevance and factuality of the generated target. These metrics include Bilingual Evaluation Understudy (BLEU-4) (Papineni et al., 2002), Recall-Oriented Understudy for Gisting Evaluation (ROUGE-1, ROUGE-2, ROUGE-L) (Lin, 2004), BERTScore (Zhang et al., 2020), Metric for Evaluation of Translation with Explicit Ordering (METEOR) (Banerjee and Lavie, 2005), AlignScore (Zha et al., 2023), and

Medical Concept (MEDCON) (Yim et al., 2023). The overall score was calculated by averaging the scores across these eight metrics. In addition to these evaluation metrics, readability scoring metrics were also investigated and utilized in some of the developed approaches.

After the conclusion of the competition, submissions from the highest-performing teams, determined by the overall score, were evaluated by a panel of clinicians [1]. The generated sections were assessed based on their completeness, correctness, readability, and overall comparison to the reference text. These criteria were evaluated on a scale ranging from 1 to 5, where 1 signifies performance that is considerably worse than the reference text, and 5 indicates performance that is considerably better than the reference text. Three independent clinicians scored 25 DI and 25 BHC texts from each team, using the same DS.

3.1 Relevance

Relevance was evaluated using BLEU-4, ROUGE-1, ROUGE-2, ROUGE-L and BERTScore. BLEU-4 measures the precision of 4-gram matches between the generated target and reference text, providing a quantitative measure of how closely the generated target matches the reference in terms of specific sequences of words. The ROUGE metrics measure the overlap of n-grams between the target and reference texts, providing a quantifiable measure of content overlap. Furthermore, BERTScore leverages contextual embeddings to assess the semantic similarity between texts by utilizing pretrained language models such as BERT (Devlin et al., 2019). In this Shared Task, the distilBERT model (Sanh et al., 2019), a lightweight and efficient variant of BERT, was used for the BERTScore evaluation.

3.2 Factuality

Factuality in text generation was assessed using AlignScore and Summary Consistency (SummaC) (Laban et al., 2022). AlignScore measures how well the facts in a generated summary align with those in the source text. SummaC extends the AlignScore by considering both, the alignment and consistency of the generated target, ensuring it not only contains factual information but also maintains logical coherence with the source.

---

[1] https://stanford-aimi.github.io/discharge-me/
Accessed: 2024-05-17

Furthermore, METEOR score evaluates translation quality by aligning machine-generated target with reference translations, considering synonyms, stemming, and ordering. It balances precision and recall, and penalizes non-contiguous matches to more closely reflect human judgments than simpler metrics like BLEU-4. Lastly, the MEDCON score is a medical concept-based evaluation metric that uses the F1-score to measure the similarity between the Unified Medical Language System (UMLS) concept sets found in candidate and reference clinical notes, assessing their accuracy and consistency.

3.3 Readability

Readability was assessed using the Flesch-Kincaid Grade Level (FKGL) (Kincaid et al., 1975), Dale-Chall Readability Score (DCRS) (Chall and Dale, 1995), and Coleman-Liau Index (CLI) (Coleman and Liau, 1975). FKGL estimates the educational grade level of a text based on sentence length and syllable count per word. DCRS evaluates text complexity by identifying words not recognized by typical fourth graders. CLI calculates the grade level needed to understand the text based on character counts and sentence structure. According to CLI, higher scores indicate lower readability.

4 Methods

This section describes different approaches to the Shared Task. Licenses for the used models, frameworks, and additional datasets can be found in Appendix E.

4.1 Few-Shot learning

Few-shot learning (Wang et al., 2020) enables machine learning models to quickly adapt to new tasks using only a handful of training examples, reducing the need for extensive data collection. This method has shown improved performance on new tasks with minimal input. The Few-Shot approach utilized the WizardLM-2-8x22B (WizardLM-2) (Xu et al., 2024a) model, which was released by Microsoft and is an instruction-tuned version of the Mixtral-8x22B[2] model from Mistral AI. Refer to Appendix A for prompting examples.

4.2 Instruction Tuning

The process of instruction tuning (Peng et al., 2023) in natural language processing involves guiding a pre-trained large language model to follow specific instructions or prompts. Unlike traditional fine-tuning, which focuses on adapting the model to a specific task using a task-specific dataset, instruction tuning uses diverse instruction-based datasets to train the model to generate more accurate and relevant responses to a wide range of queries. This enables the model to better generalize across different tasks by understanding and following the instructions given.

For every experiment carried out, two models were trained: One to generate DI and one to generate BHC. Between the different experiments, hyperparameters were changed only slightly to make the experiments comparable (see Appendix C). As input format, the chat template recommended by the model publishers was used for training. Chat templates[3] are structured formats that guide the interaction between the user and the model. The input consisted of a System Message and the DS taken from the MIMIC-IV dataset. Example prompts are shown in the Appendix (see Appendix A). Most models were trained on a single NVIDIA H100 80GB using the unsloth[4] framework. Only Phi-3-Mini-128K-Instruct (Abdin et al., 2024) was trained on three NVIDIA H100 80GB. It was necessary to choose Large Language Models that are capable of handling long sequences. The average DS length is about 1,775 words or 4,243 tokens, using the Mistral-7B-Instruct-v0.2 (Jiang et al., 2023) tokenizer. All models were trained with Low-Rank Adaptation (LoRA) (Hu et al., 2022). The following models were evaluated: Llama-3-8B-Instruct (AI@Meta, 2024), Llama-3-70B-Instruct (AI@Meta, 2024), OpenBioLLM-70B (Ankit Pal, 2024), Phi-3-Mini-128K-Instruct, Mistral-7B-Instruct-v0.2. In the remainder of the paper, "I" stands for Instruct in the model naming convention. Please see Appendix C for the fine-tuning setup.

Besides the classical approach of model fine-tuning, an attempt was made to prime the models to improve their understanding of "clinical language". For this, the models were instruction-tuned with the Asclepius dataset before using the task-specific MIMIC-IV dataset. For this approach, Llama-8B-I and Mistral-7B-I-v0.2 were evaluated.

Asclepius is a dataset that was released by

[Kweon et al. 2023](). This dataset contains 158,000 rows of synthetical clinical notes and instruction-answer pairs. It was built on publicly available case reports, extracted from biomedical lectures, and then transformed into clinical notes. instruction-answer pairs were built using ChatGPT-3.5-Turbo ([OpenAI, 2023]()).

4.3 MIMIC Section Identification

MIMIC Section Identification (MIMIC-SID) ([Landes et al., 2022, 2023]()) is a framework used for automatically classifying sections within unstructured clinical texts, such as patient medical records. It recognizes and defines different sections of text based on their content and context. This is particularly useful in the medical domain. Documents such as DS contain distinct sections (e.g., diagnosis, treatment, patient history) that need accurate identification for effective information retrieval and processing.

Utilizing MIMIC-SID (see Figure 1), the most important sections for the target text were identified by calculating the average BERTScore (with distilBERT) between the extracted section and the target section. The text was then ordered based on relevance, from highest to lowest BERTScore, and truncated after 2,000 words. This method assumes that relevant parts are already found at the beginning of the text, and less relevant parts would be cut out. To compare this approach to a more standardized setting, the unaltered input text was also truncated to 2,000 words. This results in two training schemes: one with 2,000 words of reordered text and one with 2,000 words of the original text.

4.4 Hyperparameters

The quality of the generated targets is strongly influenced by the inference parameters employed. The Meta-Llama-3-8B-Instruct model was utilized to establish decoding strategies for the Shared Task, specifically adopting the proposed methods by ([Minaee et al., 2024]()). Three experimental runs were conducted to examine their influence on text generation quality, each employing these decoding strategies in different configurations. The configurations and their respective parameters are detailed in Table 1.

4.5 Dynamic Expert Selection

As final approaches, five different Dynamic Expert Selections (DES) were constructed. For each DES, a set of models was pre-selected to serve as ex-
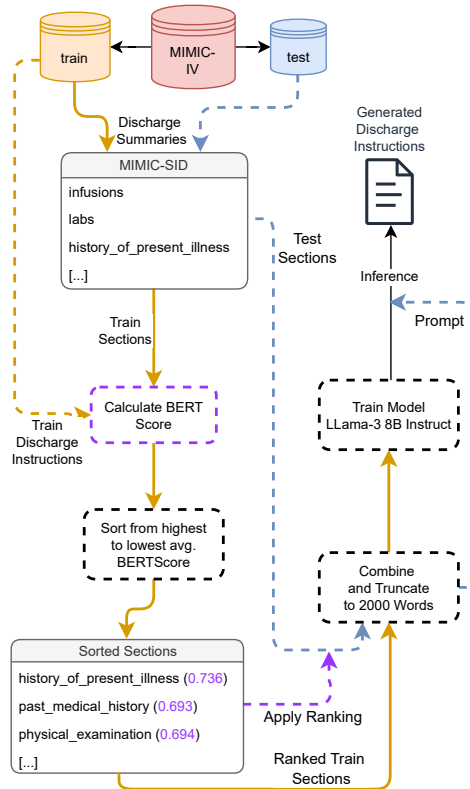

Figure 1: This workflow, exemplified by DI, is applied to BHC in the same way. With MIMIC-SID the dataset is divided into up to 50 sections. For each training section, the average BERTScore is computed using the target text as a reference. The sections are then ranked from highest to lowest BERTScore, and this ranking is applied to both the training and testing DS. The ranked training dataset is used to train the Llama-3-8B-I model. Subsequently, the ranked testing dataset is presented to the model in the form of prompts to generate DI outputs.

Parameter	Config 1	Config 2	Config 3
do_sample	False	False	True
RP	1.2	1	1
NNS	3	∞	∞
ERP	1	1.2	1
temp	0	0	0.6
top_p	0	0	0.9

Table 1: Configuration Parameters for Inference Runs. RP stands for *repetition_penalty*, NNS stands for *non_repeat_ngram_size*, and ERP stands for *encoder_repetition_penalty*.

perts. Each model generates DI and BHC for a DS, and then an expert model is selected whose text is included in the submission.

Readability and factuality scores are calculated to select the expert model. The readability scores

can be calculated without any reference text, and the factuality scores use the entire DS as a reference. Because they do not use the target texts, these scores are referred to as pre-calculated scores.

Additionally, the validation set was used to compute the pre-calculated scores for the generated targets of the Mistral-7B-I-v0.2 + Asclepius model. Furthermore, the overall scores (all challenge evaluation scores) based on the target texts were determined on the validation set. Then, the correlations between the pre-calculated scores and the overall scores were examined. Figure 2 shows these correlations as a heatmap. Taking into account these correlations, DES 1-4 were constructed. For DES 5, the lengths of the generated targets were considered instead of scores as the selection criterion. This decision was based on the observation that, particularly in longer texts, models exhibit signs of hallucination or the generation of repetitive content.

When compiling a DES, the pre-calculated scores of all included models for a DS are subjected to a min-max normalization. This means normalization over all available models. These normalized scores are then multiplied by selected weights. The model with the highest average of all normalized and weighted scores is selected as the expert for that DS.

DES 1   This DES was optimized for MEDCON and METEOR with a weight of $\frac{1}{2}$ each, as these two metrics exhibited the strongest correlation with a higher overall score. The final submission file included 6,407 texts from Mistral-7B-I-v0.2 + Asclepius, 1,210 texts from Llama-3-8B-I with greedy decoding (Minaee et al., 2024), 2,815 texts from Llama-3-8B + Asclepius, 5,600 texts from Llama-3-70B-I, 4,112 from OpenBioLLM-70B, and 1,780 from WizardLM-2.

DES 2   This DES was optimized for MEDCON and METEOR as measures for factuality, and CLI score for readability with the weights $\frac{2}{5}$ for both MEDCON and METEOR, and $\frac{1}{5}$ for CLI. The final submission file included a total of 3,471 texts from the Mistral-7B-I-v0.2 + Asclepius model, 4,374 texts from Llama-3-8B-I + Asclepius, 5,108 texts from Llama-3-70B-I, and 5,971 texts from OpenBioLLM-70B.

DES 3   This DES was optimized for all readability metrics (FKGL, DCRS, and CLI) for DI, and additionally MEDCON, METEOR, and AlignScore
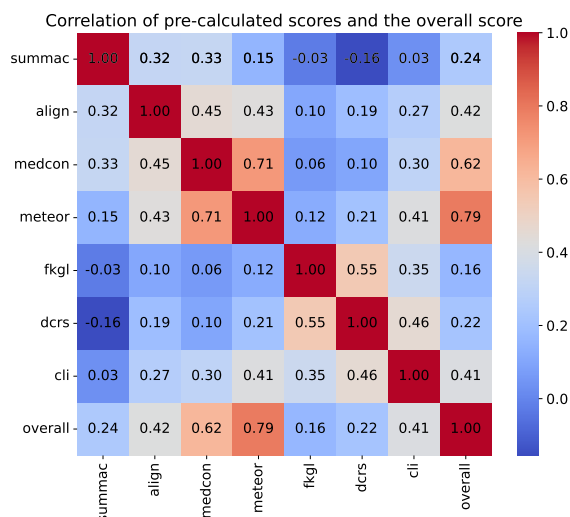

Figure 2: Heatmap of the Pearson correlations between pre-calculated scores and the overall score on the validation dataset. The pre-calculated scores include factuality scores (SummaC, AlignScore, MEDCON and METEOR), which are calculated for the generated targets of the Mistralv2 + Asclepius model with the whole DS as the reference, and readability scores (FKGL, DCRS and CLI).

as factuality metrics for both text types. For the DI, all readability metrics were assigned a weight of $-\frac{1}{9}$, and the factuality metrics were assigned a weight of $\frac{2}{9}$. For the BHC, all factuality metrics were weighted with $\frac{1}{3}$. The final submission file included 5,120 texts from Mistral-7B-I-v0.2 + Asclepius, 4,211 texts from Llama-3-8B-I + Asclepius, 5,435 texts from Llama-3-70B-I, and 7,158 texts from OpenBioLLM-70B.

DES 4   This DES used only Mistral-7B-I-v0.2 models and the values of correlation between the

---

Example given:

[...Non-repetitive text]
-Please check LFTs weekly for the next month.
-Please continue to monitor for signs of refeeding hyperglycemiAsclepius
-Please continue to monitor for signs of refeeding hypocalcemiAsclepius
[Continues Repetition...]

Figure 3: Example of repetitive and hallucinated DI output generated by Llama-3-8B-I. The words hyperglycemia and hypocalcemia are very similar but only one of them should be in the generated targets. The other one was not mentioned in the DS.

scores calculated on the generated targets against the whole DS and the overall score on the validation dataset as weights. The final submission file included 7,912 texts from Mistral-7B-I-v0.2 + Asclepius, 6,485 texts from the Mistral-7B-I-v0.2 + Asclepius model, which was further fine-tuned on the validation dataset, and 7,527 texts from Mistral-7B-I-v0.2.

DES 5 This DES considers lengths of texts instead of weighted metrics. DI in the training dataset has an average word count of approximately 196.3, whereas BHCs have an average word count of approximately 327.6. Models trained on these texts, therefore, tend to generate shorter texts for DI and longer texts for BHC. To mitigate the impact of hallucinations at the end of lengthy texts (e.g Figure 3), a strategy of preferably selecting shorter texts with the DES was adopted. The objective of the strategy was to initially rank the models based on their overall scores. Subsequently, for each DS, the text from the first model that has a word count within the range of 100 to 180 words is selected. If no model had generated a text with a word count within this range, the text with the minimum word count was selected. However, the text could not be shorter than 70 words. In the case that no text met these criteria, the text from the highest-ranked model remained.

5 Results

This section describes results of the evaluation of the developed models using the metrics described in section 3. The evaluation done by clinicians can be seen in Table 3.

5.1 Automatic Evaluation

The final scores for the Shared Task, provided by the organizers, are shown in Table 2. The table presents the approaches from the most general to the most specific, beginning with the baseline model, followed by the Few-Shot model, the instruction-tuned models, the instruction-tuned models primed with Asclepius, the MIMIC section-based approaches, and the various DES variants. The top competitors by overall score are highlighted for comparison. All inference runs for the final results were conducted with an optimized decoding strategy (temp=0.6, top_p=0.9) as described in section 4.4. Based on its high evaluation scores, Configuration 3 was chosen as the standard parameter setting, which can be seen in Table 1.

WizardLM-2, despite not being fine-tuned on the training data, surpassed the baseline with an overall score of 0.195.

Among the fine-tuned models, Llama-3-70B-I led with a score of 0.300, followed by Mistral-7B-I-v0.2 at 0.289, which has way less parameters yet outperformed several larger models, including Llama3-8B-I. Even though the OpenBioLLM-70B has been adapted for clinical use, it underperformed when compared to other models. The Phi-3-mini-128k-I (3.8 billion parameters) model matched the performance of larger models, such as the Llama-8B-I, demonstrating the efficiency of models with less parameters.

Among the configurations, incorporating the Asclepius dataset into Mistral-7B-I-v0.2 made it clearly outperform Llama3-8B-I. Excluding the DES approaches, this combination achieved the highest performance of all models.

The MIMIC-SID approaches with a shorter context length of 2,000 words displayed weak performances.

Different parameters on the Llama-8B-I, including greedy decoding and an ERP (encoder_repetition_penalty), yielded lower scores compared to setups utilizing sampling and temperature adjustments.

In the Dynamic Expert Selection category, DES 1 focused on MEDCON and METEOR scores but did not surpass the individual fine-tuned models. DES 2 achieved the highest BERTScore and AlignScore, which represent relevance and factuality, respectively. Reaching an overall score of 0.311, this DES outperformed all individual fine-tuned models. DES 3, aimed at lowering readability metrics, scored 0.296, performing better than DES 1 but lagging behind others. DES 4, using correlation values for optimization, showed negligible improvements. DES 5 achieved the highest overall score of 0.332 and topped the leaderboard by limiting text length. The approach achieved the highest scores in all metrics, except for a slightly lower BERTScore and AlignScore.

5.2 Clinical Evaluation

The ranking order of the first six teams did not change when comparing the automatic with the clinicians' evaluation results (see Table 2 and Table 3). The evaluated BHC were ranked the best over all aspects. The readability and holistic evaluation of the BHC were notably superior to that of the other teams. However, the DI scores were compa-

Model	Ovr.	BLEU	R-1	R-2	R-L	BERT	MET	Align	MED
				Baseline					
Challenge Baseline	0.102	0.015	0.126	0.051	0.113	0.138	0.098	0.167	0.121
				Few-Shot learning					
WizardLM-2 8x22B	0.195	0.017	0.257	0.074	0.158	0.331	0.310	0.193	0.218
				Instruction-tuned					
Llama-3-8B-I	0.253	0.053	0.331	0.107	0.241	0.392	0.235	0.320	0.348
Mistral-7B-I-v0.2	0.289	0.101	0.371	0.122	0.252	0.416	**0.375**	0.293	0.380
Llama-3-70B-I	**0.300**	**0.112**	0.367	**0.141**	**0.260**	**0.437**	0.347	0.334	**0.401**
OpenBioLLM-70B	0.285	0.084	0.376	0.127	0.248	0.421	0.307	**0.337**	0.383
Phi-3-mini-128k-I	0.254	0.062	**0.347**	0.128	0.217	0.359	0.310	0.275	0.330
				Instruction-tuned + Asclepius (A.)					
Llama-3-8B-I + A.	0.302	0.107	0.388	**0.150**	**0.275**	0.432	0.350	0.311	0.403
Mistral-7B-I-v0.2 + A.	**0.307**	**0.120**	**0.390**	0.140	0.258	**0.434**	**0.391**	**0.320**	**0.404**
				MIMIC Section Identification					
Llama-3-8B-I 2k	0.209	0.022	0.263	0.054	0.171	0.326	**0.199**	**0.355**	0.280
Llama-3-8B-I R 2k	**0.216**	**0.026**	**0.292**	**0.073**	**0.191**	**0.351**	0.186	0.306	**0.304**
				Hyperparameter					
Llama-3-8B-I Greedy	0.192	0.018	0.274	0.043	0.147	0.314	**0.221**	0.281	0.241
Llama-3-8B-I ERP	**0.238**	**0.032**	**0.348**	**0.093**	**0.228**	**0.372**	**0.221**	**0.307**	**0.300**
				Dynamic Expert Selection					
DES 1	0.277	0.097	0.329	0.121	0.217	0.417	0.339	0.319	0.374
DES 2	0.311	0.110	0.414	0.151	0.273	**_0.439_**	0.351	**_0.344_**	0.406
DES 3	0.296	0.108	0.366	0.128	0.242	0.435	0.352	0.335	0.400
DES 4	0.297	0.112	0.371	0.127	0.244	0.426	0.379	0.320	0.396
DES 5	**_0.332_**	**_0.124_**	**_0.453_**	**_0.201_**	**_0.308_**	0.438	**_0.403_**	0.315	**_0.411_**
				Top 5 Competitors					
HarmonAI Lab Yale	**0.300**	**0.106**	0.423	0.180	**0.284**	**0.412**	0.381	0.265	0.353
aehrc	0.297	0.097	0.414	**0.192**	**0.284**	0.383	**0.398**	0.274	0.332
EPFL-MAKE	0.289	0.098	**0.444**	0.155	0.262	0.399	0.336	0.255	**0.360**
UF-HOBI	0.286	0.102	0.401	0.174	0.275	0.395	0.289	**0.296**	0.355
de ehren	0.284	0.097	0.404	0.166	0.265	0.389	0.376	0.231	0.339

Table 2: Summary of model performance across different experimental settings. Each section represents a distinct approach: Baseline, Few-Shot Learning, instruction-tuned, instruction-tuned + Asclepius, MIMIC-SID (2k for truncation to 2k words and R for reordering the subsections in the text from most to least relevant according to BERTScore), Hyperparameter, and DES, showcasing respective strategies to address the challenge. *I* indicates that the instruction version of the model was used. Furthermore, the Top 5 runs from other challenge participants are included. Metrics include overall score (Ovr.), BLEU-4 (BLEU), ROUGE-1 (R-1), ROUGE-2 (R-2), and ROUGE-L (R-L), BERTScore (BERT), METEOR (MET), AlignScore (Align) and MEDCON (MED). Bold scores indicate the best performance in each category, with underlined bold scores highlighting the top overall scores across all experiments.

rable to those of other teams and did not achieve the highest score in any of the evaluated aspects.

In order to compare automated evaluation results with clinicians' assessments, it was necessary to normalize the scores on a scale from 0 to 1. Note that readability was not compared, as clinicians did not rate the readability of DI texts, nor did the challenge metrics include readability scores. Figure 4 illustrates that clinicians tend to assign higher scores than the automated metrics in their evaluation approach. However, the holistic evaluation aligns more closely with the overall automated scores.
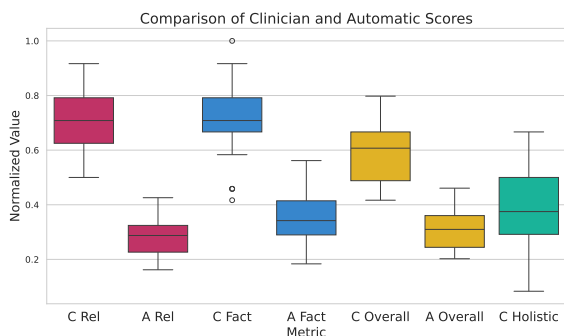

Figure 4: Boxplot of Average Clinician Scores and Average Metric Scores. C stands for Clinician and A stands for the scores caluclate with the challenge metrics. Here A Relevance includes ROUGE-1, ROUGE-2, ROUGE-L, BertScore and BLEU-4. A Factuality includes Align-Score METEOR and MEDCON.

6 Discussion

Despite being one of the less robust models evaluated, WizardLM-2 exceeded the established baseline, showing its effectiveness in a Few-Shot learning context. With minimal training examples, the model still produced high-quality texts, according to the metrics, highlighting the potential of Few-Shot learning in enhancing performance metrics.

The performance of instruction-tuned models revealed mixed outcomes. Despite being a specialized adaptation of the Llama-3-70B-I model tailored for medical contexts, OpenBioLLM-70B underperformed in relation to its base model. This behavior was unexpected, considering its design to enhance relevance and accuracy in clinical applications. Conversely, the Mistral-7B-I-v0.2 model demonstrated impressive capabilities, outperforming both the larger OpenBioLLM-70B and the Llama-8B-I models. This highlights the effectiveness of Mistral-7B-I-v0.2 in handling complex

medical text generation and summarization tasks despite its smaller size. In contrast to its reputation as one of the most promising state-of-the-art open-source LLMs, the Llama-3 models have been found to be less effective in this challenge. This is on par with the findings from LMSYS chatbot arena (Chiang et al., 2024) where LLama-3 models showed the weakest performance compared to other state-of-the-art models on the task of summarization (Dunlap et al., 2024).

Using the Asclepius dataset for priming substantially improved model performance during the fine-tuning phases. For instance, the Llama-8B-I model's score rose from 0.253 to 0.302, and the Mistral model's performance increased from 0.289 to 0.307. Notably, the Mistral-7B-I-v0.2 + Asclepius model was the top performer in the challenge, aside from DES approaches. This underscores the benefits of further training models with specialized datasets to enhance accuracy and relevance in domain-specific tasks.

The reordering of sections within the MIMIC-SID approach moderately enhanced overall model performance, demonstrating that prioritizing the most relevant sections can be beneficial. However, it is important to note that metrics sensitive to text order, such as METEOR and AlignScore, experienced a decline. This suggests that while reordering can improve general outcomes by emphasizing key information, it may simultaneously compromise the sequential integrity of the text. Therefore, this strategy confirms the utility of structurally optimizing input for task-specific relevance, albeit with some trade-offs in textual coherence.

Exploration of hyperparameter settings revealed that more complex configurations did not yield superior results. The basic approach, utilizing *do_sample=True*, *temp=0.6*, and *top_p=0.9*, consistently outperformed other tested configurations, including those with greedy decoding and encoder repetition penalties. This emphasizes the efficacy of maintaining simpler hyperparameter settings for stable and high-quality text generation. Additional complexity in parameter tuning did not always correlate with improved model performance.

The DES that relied on the pre-calculated scores had varying effects on the metrics evaluated. Using MEDCON and METEOR in combination with CLI improved the results, whereas choosing the correlation as weights resulted in no improvement. A possible reason might be that the pre-calculated scores were only calculated on the entire DS and

Team	Avg.	BHC				DI		
		Comp	Corr	Read	Hol.	Comp	Corr	Hol.
WisPerMed	**3.375**	**3.667**	**3.667**	**3.373**	**2.440**	3.947	4.000	2.533
HarmonAI Lab at Yale	2.903	3.520	2.587	2.107	1.533	**4.267**	3.947	2.360
aehrc	2.785	2.307	3.053	1.960	1.093	3.907	**4.547**	**2.627**
EPFL-MAKE	2.720	3.293	2.827	2.533	1.653	3.453	3.413	1.867
UF-HOBI	2.579	2.480	3.360	2.707	1.413	3.013	3.293	1.787
de ehren	2.335	2.280	2.987	2.680	1.120	2.813	3.053	1.413

Table 3: BHC and DI Metrics for Teams by clincans. In this Table Avg. stands for Average, Comp stands for Comperability, Corr stands for Correctness, Read stands for Readability, and Hol stands for Holistic.

not on the target text, as in the final evaluation. It may also be that the correlations are not always sufficient, and a more elaborate association analysis is needed.

Consequently, the best overall score of all DES was achieved by the approach limiting the text length, suggesting that hallucinations and repetitive sequences have a measurable impact on text quality.

The manual evaluation seen in Table 3 indicates that the holistic approach by clinicians is comparable to the automated metrics, thus reconfirming the effectiveness of the metrics used in the competition. The lower scores for the DI may be caused by information loss or distortion due to the simplification.

7 Conclusion

The research identified several opportunities for future investigation that may enhance the performance and utility of the discussed models. Initially, due to the extensive size of the training dataset and the constraints imposed by the context length of input texts, each model was trained for a maximum of only three epochs. Therefore, extending the training duration may provide improvements and merits further exploration.

Moreover, alterations to inference parameters have demonstrated notable effects on model outputs. For example, employing the ERP parameter, while maintaining other settings constant resulted in a degradation of performance metrics (from 0.253 to 0.238 overall score). This suggests a systematic evaluation of inference parameters could further enhance model output.

Additionally, priming the model has substantially improved results. Investigating additional datasets for priming purposes could further optimize model performance and expand its applica-

bility across diverse textual tasks. This could be a promising direction for future research efforts. Further opportunities lie in optimizing section reordering to balance task-specific relevance while maintaining text coherence.

The winning approach, evaluated by the automatic and clincans' evaluation, a DES, achieved the highest overall score. This suggests that generating multiple outputs and developing methodologies to select the optimal text may further improve performance. Therefore, exploring various DES techniques and selection criteria is a field for further research.

The clinicians' evaluation added valuable insights, and the automatic scores demonstrated high robustness and strong alignment with the manual assessments. This alignment indicates that the manual evaluation, even on a small subset, effectively validates the reliability of the generated texts.

Lastly, efforts to enhance the quality of medical machine learning algorithms are ongoing, along with a responsibility to report the environmental impact of the research. In this study, the total energy consumption for training and inference is estimated with 1,552.10 kWh, resulting in 591.35 kg CO_2 emission. Detailed information is provided in Appendix D.

Limitations

For the model training, only the entire discharge summaries were utilized, while the provided radiology reports and ICD 9/10 codes were not included. The decision to exclude these additional documents might have limited the comprehensiveness of our models. Future research should consider incorporating these documents to potentially improve model accuracy and contextual understanding.

Moreover, each model was trained for a max-

imum of three epochs due to the context length constraints of input texts. Extending the training duration could potentially enhance performance and merit further exploration.

Additionally, the influence of inference parameters on model outputs is notable. Systematic evaluation of these parameters is needed, as variations can measurably affect performance metrics.

Furthermore, the study did not employ advanced preprocessing or postprocessing techniques, which could substantially enhance the reliability and accuracy of the generated texts by mitigating issues such as non-factual content generation ("hallucinations"). Notably, DES 5 considered text length, which may indirectly reduce hallucinations. However, this approach does not explicitly address the issue and therefore cannot ensure their complete avoidance.

The Asclepius dataset, being synthetic and partly based on the MIMIC-III dataset, may introduce data redundancy or leakage, potentially impacting model robustness and generalizability. Future work should explore advanced data validation techniques or alternative dataset creation methodologies to mitigate these issues.

Lastly, while priming models with specialized datasets showed substantial improvements, further investigation into additional datasets for priming could optimize model performance and expand applicability across diverse textual tasks.

Acknowledgement

References

Marah Abdin, Sam Ade Jacobs, Ammar Ahmad Awan, Jyoti Aneja, Ahmed Awadallah, Hany Awadalla, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Harkirat Behl, Alon Benhaim, Misha Bilenko, Johan Bjorck, Sébastien Bubeck, Martin Cai, Caio César Teodoro Mendes, Weizhu Chen, Vishrav Chaudhary, Parul Chopra, Allie Del Giorno, Gustavo de Rosa, Matthew Dixon, Ronen Eldan, Dan Iter, Amit Garg, Abhishek Goswami, Suriya Gunasekar, Emman Haider, Junheng Hao, Russell J. Hewett, Jamie Huynh, Mojan Javaheripi, Xin Jin, Piero Kauffmann, Nikos Karampatziakis, Dongwoo Kim, Mahoud Khademi, Lev Kurilenko, James R. Lee, Yin Tat

Lee, Yuanzhi Li, Chen Liang, Weishung Liu, Eric Lin, Zeqi Lin, Piyush Madan, Arindam Mitra, Hardik Modi, Anh Nguyen, Brandon Norick, Barun Patra, Daniel Perez-Becker, Thomas Portet, Reid Pryzant, Heyang Qin, Marko Radmilac, Corby Rosset, Sambudha Roy, Olatunji Ruwase, Olli Saarikivi, Amin Saied, Adil Salim, Michael Santacroce, Shital Shah, Ning Shang, Hiteshi Sharma, Xia Song, Masahiro Tanaka, Xin Wang, Rachel Ward, Guanhua Wang, Philipp Witte, Michael Wyatt, Can Xu, Jiahang Xu, Sonali Yadav, Fan Yang, Ziyi Yang, Donghan Yu, Chengruidong Zhang, Cyril Zhang, Jianwen Zhang, Li Lyna Zhang, Yi Zhang, Yue Zhang, Yunan Zhang, and Xiren Zhou. 2024. Phi-3 technical report: A highly capable language model locally on your phone. *Preprint*, arXiv:2404.14219.

AI@Meta. 2024. Llama 3. https://github.com/meta-llama/llama3/. Accessed: 2024-05-15.

Malaikannan Sankarasubbu Ankit Pal. 2024. OpenBioLLMs: Advancing open-source large language models for healthcare and life sciences. https://huggingface.co/aaditya/OpenBioLLM-Llama3-70B. Accessed: 2024-05-15.

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.

J.S. Chall and E. Dale. 1995. *Readability Revisited: The New Dale-Chall Readability Formula*, volume 1. Brookline Books.

Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Hao Zhang, Banghua Zhu, Michael Jordan, Joseph E. Gonzalez, and Ion Stoica. 2024. Chatbot arena: An open platform for evaluating llms by human preference. *Preprint*, arXiv:2403.04132.

Meri Coleman and Ta Lin Liau. 1975. A computer readability formula designed for machine scoring. *Journal of Applied Psychology*, 60(2):283–284.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Lisa Dunlap, Evan Frick, Tianle Li, Isaac Ong, Joseph E. Gonzalez, and Wei-Lin Chiang. 2024. What's up with llama 3? arena data analysis. Accessed: 2024-05-15.

Ary L. Goldberger, Luis A. N. Amaral, Leon Glass, Jeffrey M. Hausdorff, Plamen Ch. Ivanov, Roger G.

Mark, Joseph E. Mietus, George B. Moody, Chung-Kang Peng, and H. Eugene Stanley. 2000. Physiobank, physiotoolkit, and physionet: Components of a new research resource for complex physiologic signals. *Circulation*, 101(23):e215–e220.

Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

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.

A. Johnson, L. Bulgarelli, T. Pollard, S. Horng, L. A. Celi, and R. Mark. 2023a. MIMIC-IV.

Alistair E. W. Johnson, Lucas Bulgarelli, Lu Shen, Alvin Gayles, Ayad Shammout, Steven Horng, Tom J. Pollard, Sicheng Hao, Benjamin Moody, Brian Gow, Li-wei H. Lehman, Leo A. Celi, and Roger G. Mark. 2023b. Mimic-iv, a freely accessible electronic health record dataset. *Scientific Data*, 10:1.

J Peter Kincaid, Robert P Fishburne Jr, Richard L Rogers, and Brad S Chissom. 1975. Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel.

Sunjun Kweon, Junu Kim, Jiyoun Kim, Sujeong Im, Eunbyeol Cho, Seongsu Bae, Jungwoo Oh, Gyubok Lee, Jong Hak Moon, Seng Chan You, Seungjin Baek, Chang Hoon Han, Yoon Bin Jung, Yohan Jo, and Edward Choi. 2023. Publicly shareable clinical large language model built on synthetic clinical notes. *Preprint*, arXiv:2309.00237.

Philippe Laban, Tobias Schnabel, Paul N. Bennett, and Marti A. Hearst. 2022. SummaC: Re-visiting nli-based models for inconsistency detection in summarization. *Transactions of the Association for Computational Linguistics*, 10:163–177.

Alexandre Lacoste, Alexandra Luccioni, Victor Schmidt, and Thomas Dandres. 2019. Quantifying the carbon emissions of machine learning. *CoRR*, abs/1910.09700.

Paul Landes, Barbara Di Eugenio, and Cornelia Caragea. 2023. DeepZensols: A deep learning natural language processing framework for experimentation and reproducibility. In *Proceedings of the 3rd Workshop for Natural Language Processing Open Source Software (NLP-OSS 2023)*, pages 141–146, Singapore, Singapore. Empirical Methods in Natural Language Processing.

Paul Landes, Kunal Patel, Sean S. Huang, Adam Webb, Barbara Di Eugenio, and Cornelia Caragea. 2022. A new public corpus for clinical section identification: MedSecId. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 3709–3721, Gyeongju, Republic of Korea. International Committee on 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.

Shervin Minaee, Tomas Mikolov, Narjes Nikzad, Meysam Chenaghlu, Richard Socher, Xavier Amatriain, and Jianfeng Gao. 2024. Large language models: A survey. *arXiv preprint arXiv:2402.06196*.

OpenAI. 2023. ChatGPT. https://www.openai.com. Accessed: 2024-05-15.

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, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

SB Patel and K Lam. 2023. Chatgpt: the future of discharge summaries? *Lancet Digit Health*, 5(3):e107–e108. Epub 2023 Feb 6.

Baolin Peng, Chunyuan Li, Pengcheng He, Michel Galley, and Jianfeng Gao. 2023. Instruction tuning with gpt-4. *Preprint*, arXiv:2304.03277.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of BERT: smaller, faster, cheaper and lighter. *CoRR*, abs/1910.01108.

Thomas Searle, Zina Ibrahim, James Teo, and Richard J.B. Dobson. 2023. Discharge summary hospital course summarisation of in patient electronic health record text with clinical concept guided deep pre-trained transformer models. *Journal of Biomedical Informatics*, 141:104358.

Christine Sinsky, Lacey Colligan, Ling Li, Mirela Prgomet, Sam Reynolds, Lindsey Goeders, Johanna Westbrook, Michael Tutty, and George Blike. 2016. Allocation of physician time in ambulatory practice: A time and motion study in 4 specialties. *Annals of Internal Medicine*, 165(11):753.

Dennis Ulmer, Elisa Bassignana, Max Müller-Eberstein, Daniel Varab, Mike Zhang, Rob van der Goot, Christian Hardmeier, and Barbara Plank. 2022. Experimental standards for deep learning in natural language processing research. In *Findings of the Association for Computational Linguistics: EMNLP 2022*,

pages 2673–2692, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Yaqing Wang, Quanming Yao, James T. Kwok, and Lionel M. Ni. 2020. Generalizing from a few examples: A survey on few-shot learning. *ACM Computing Surveys*, 53(3).

Can Xu, Qingfeng Sun, Kai Zheng, Xiubo Geng, Pu Zhao, Jiazhan Feng, Chongyang Tao, Qingwei Lin, and Daxin Jiang. 2024a. WizardLM: Empowering large pre-trained language models to follow complex instructions. In *The Twelfth International Conference on Learning Representations*.

J. Xu. 2024. Discharge me: Bionlp acl'24 shared task on streamlining discharge documentation.

Justin Xu, Andrew Johnston, Zhihong Chen, Louis Blankemeier, Maya Varma, Curtis Langlotz, and Jean-Benoit Delbrouck. 2024b. 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.

W. Yim, Y. Fu, A. Ben Abacha, N. Snider, T. Lin, and M. 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. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11328–11348, Toronto, Canada. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. BERTScore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

A Few-Shot Learning Prompts

This section showcases how the WizardLM-2-Model was instructed. A variety of prompts were tested, and the displayed ones (see Figure 5 and Figure 6) yielded the best results, as measured by human evaluation. Detailed instructions were provided for the DI text generation, whereas the details for the BHC generation were excluded. The BHC texts are considerably longer on average and do not follow the same pattern most of the time.

Discharge Instructions Prompt

USER: Generate a detailed discharge instruction based on the provided summary, adhering to the style of the provided examples. The instruction should comprehensively cover all aspects of the patient's care, with a total length of about 300-500 words.

Please follow the format used in previous discharge instructions:

1. Start with a polite greeting and an expression of gratitude or pleasure for having taken care of the patient.

2. Describe the reason for hospitalization succinctly.

3. Detail what occurred during the stay, including any treatments administered, patient responses, and significant changes to the patient's condition.

4. Outline clear follow-up care instructions, including medications, dietary recommendations, activity level, and scheduled follow-up visits.

5. Close with a kind farewell and additional well-wishes or reminders.

Discharge Instruction Format Example:
Dear [Patient Name],
It was a pleasure taking care of you during your hospitalization at [Hospital Name].

Why were you hospitalized?
- [Brief reason for hospitalization]

What happened while you were in the hospital?
- [Key details about treatment and patient response]
- [Any significant tests and their results]
- [Any changes to patient condition]

What should you do after you leave the hospital?
- [Medications and dosage]
- [Dietary instructions]
- [Activity recommendations]
- [Follow-up appointments]

We wish you the best in your recovery!

Sincerely,
Your [Hospital Team Name] Team

Discharge Instruction Example 1 start:
[Discharge Instruction Example from Validation set]
Discharge Instruction Example 1 end.
[· · ·]
Discharge Instruction Example 10 start:
[Discharge Instruction Example from Validation set]
Discharge Instruction Example 10 end.
Discharge Summary:
[Discharge Summary without target section]
Start with the Discharge Instructions for the Discharge Summary.
ASSISTANT:

Figure 5: Discharge Instruction Prompt for Few-Shot learning with WizradLM-2.

Brief Hospital Course Prompt

USER: Here are some Example Brief Hospital Courses.

Brief Hospital Course Example 1 start:
[Brief Hospital Course Example from Validation set]
Brief Hospital Course Example 1 end.
[· · ·]
Brief Hospital Course Example 7 start:
[Brief Hospital Course Example from Validation set]
Brief Hospital Course Example 7 end.

Now create a Brief Hospital Course in the same style as in the Examples with the information from the following Discharge Summary:
[Discharge Summary without target section]

ASSISTANT:

Figure 6: Brief Hospital Course Prompt for Few-Shot learning with WizardLM-2.

B   Instruction Tuning Prompts

Figure 7 and Figure 8 show the prompts used for instruction tuning DI and BHC. The only difference between the instruction tuning and inference prompts is that the [Target Discharge Instructions] or [Target Brief Hospital Course] was left empty for inference. For each model, the recommended chat template provided by the model inventors was followed and applied. This is especially important when using the instruction version of those models.

Discharge Instructions Prompt

<SYSTEM>You are in the world's best hospital as the best doctor. You're given a patient's details summarized by your medical staff in 'Summary'. You now need to figure out the 'Discharge Instructions' for the patient. Think carefully without error, since you might endanger a patient's life, which we do not want to happen.

<User>Summary: [Discharge Summary without target section]

Discharge Instructions:
<ASSISTANT>[Target Discharge Instructions]

Figure 7: Instruction Tuning and Inference Prompt for Discharge Instructions.

Brief Hospital Course Prompt

<SYSTEM>You are in the world's best hospital as the best doctor. You're given a patient's details summarized by your medical staff in 'Summary'. You now need to figure out a 'Brief Hospital Course' for the patient. Think carefully without error, since you might endanger a patient's life, which we do not want to happen.

<USER>Summary: [Discharge Summary without target section]

Brief Hospital Course:
<ASSISTANT>[Target Brief Hospital Course]

Figure 8: Instruction Tuning and Inference Prompt for Brief Hospital Course.

C Parameter Setup

Whenever possible, hyperparameters were only changed slightly to ensure high comparability between results. The LoRA setup is detailed in Table 4. The following modules were targeted with LoRA: "q_proj", "k_proj", "v_proj", "o_proj", "gate_proj", "up_proj", and "down_proj". While it is suggested[5] to use a LoRA Rank = LoRA Alpha * 2, this approach was not chosen due to VRAM efficiency considerations. For the detailed training setup, please see Table 5. All models were trained on 80GB H100s and 48GB

Model	LR	LA	loadIn4Bit	LD	GC	DT
Llama-3-8B-I + A. Prime	16	16	true	0	true	bfloat16
Mistral-7B-I-v0.2 + A. Prime	16	16	true	0	true	bfloat16
Llama-3-8B-I	16	16	true	0	true	bfloat16
Mistral-7B-I-v0.2	16	16	true	0	true	bfloat16
Llama-3-70B-I	16	16	true	0	true	bfloat16
OpenBioLLM-70B	16	16	true	0	true	bfloat16
Phi-3-mini-128k-I	16	16	true	0	true	bfloat16
Llama-3-8B-I 2k + A	16	16	true	0	true	bfloat16
Mistral-7B-I-v0.2 + A.	16	16	true	0	true	bfloat16
Llama-3-8B-I 2k	16	16	true	0	true	bfloat16
Llama-3-8B-I R 2k	16	16	true	0	true	bfloat16

Table 4: LoRA Setup for fine-tuning. LR means LoRA Rank, LA means LoRA Alpha, LD means LoRA Dropout, GC means Gradient Checkpointing, DT means dtype. By A. Asclepius is meant. Prime means the instruction tuning runs with Asclepius.

RTX6000s. The Unsloth open-source training framework was used because it reduced VRAM usage by at least 50% and subsequently made fine-tuning runs twice as fast. This efficiency allowed training almost all models on a single GPU. The Maximum Sequence Length for the 70B models was reduced to decrease

Model	MSL	E	GAS	WS	LR	BS	O	S	WD
Llama-3-8B-I + A. P.	15,000	1	4	5	2e-4	4	adamw_8bit	linear	0.01
Mistral-7B-I-v0.2 + A. P.	15,000	1	4	5	2e-4	4	adamw_8bit	linear	0.01
Llama-3-8B-I	10,000	3	4	5	2e-4	4	adamw_8bit	linear	0.01
Mistral-7B-I-v0.2	10,000	3	4	5	2e-4	4	adamw_8bit	linear	0.01
Llama-3-70B-I	10,000	2	4	5	2e-4	2	adamw_8bit	linear	0.01
OpenBioLLM-70B	10,000	2	4	5	2e-4	2	adamw_8bit	linear	0.01
Phi-3-mini-128k-I	12,000	2	4	10	2e-4	4	p_adamw_8bit	linear	0.01
Llama-3-8B-I + A.	13,000	2	4	5	2e-4	4	adamw_8bit	linear	0.01
Mistral-7B-I-v0.2 + A.	13,000	2	4	5	2e-4	4	adamw_8bit	linear	0.01
Llama-3-8B-I 2k	6,000	3	4	5	2e-4	4	adamw_8bit	linear	0.01
Llama-3-8B-I R 2k	6,000	3	4	5	2e-4	4	adamw_8bit	linear	0.01

Table 5: MSL means Maximum Sequence Length (Tokens), E means Epochs, GAS means Gradient Accumulation Steps, WS means Warmup Steps, LR means Learning Rate, BS means Batch Size, O means Optimizer, S means Scheduler, WD means Weight Decay. By A. Asclepius is meant. Prime (P.) means the instruction tuning runs with Asclepius.

memory consumption. For Phi-3, the optimizer was changed from adamw_8bit (Loshchilov and Hutter, 2019) to paged_adamw_8bit[6] to further optimize memory usage.

---

[5] https://magazine.sebastianraschka.com/p/practical-tips-for-finetuning-llms Accessed: 2024-05-17
[6] https://huggingface.co/docs/bitsandbytes/en/optimizers#paged-optimizers Accessed: 2024-05-15

D Environmental Impact

In scientific research, it is crucial to consider not only the direct results of experiments but also the broader implications and consequences of the research process. While the following environmental assessment is not directly tied to the primary results, reporting on the environmental footprint of the work is essential given the increasing global emphasis on sustainability and the environmental impact of computational practices. This perspective aligns with the findings of (Ulmer et al., 2022), emphasizing the importance of understanding and reporting the environmental consequences of experimental work.

The experiments were conducted using HPC resources located in Essen and Dortmund, Germany. The region's electricity generation has a carbon efficiency of 0.381 kgCO$_2$ eq/kWh[7], with approximately 41,1% [8] of the electricity being sourced from fossil fuels. To estimate the carbon footprint of our experiments, the Machine Learning Impact calculator, as presented by (Lacoste et al., 2019), is utilized. This calculator provides a comprehensive framework to quantify the carbon emissions associated with machine learning experiments, considering both the energy consumption of computational resources and the carbon efficiency of the electricity source.

Final Models	Runtime (hours)	Power (Avg. Watts)	Energy (kWh)	CO$_2$ (kg)
Mistral-7B-I-v0.2 BHC + A.	28.5	651.45	18.58	7.08
Mistral-7B-I-v0.2 + A. Prime	5	637	3.21	1.22
Mistral-7B-I-v0.2 DI + A.	27.4	681	18.71	7.13
Experiment runs	1,920	783.83	1,511.59	575.91
Overall	1,980.9	783.532	1,552.10	591.35

Table 6: Runtime, Energy Consumption and CO$_2$ Emissions for the Final models, Other Experiment Runs and Overall for All Experiments. By A. Asclepius is meant. Prime means the instruction tuning runs with Asclepius.

The carbon footprint and electricity consumption values for our optimal models, as well as for all experimental runs conducted throughout the research process presented in Table 6. The values indicate that substantial resources are expended on debugging and testing during development.

---

[7]https://ourworldindatAsclepiusorg/grapher/carbon-intensity-electricity?country=~DEU Accessed: 2024-05-14
[8]https://www.destatis.de/EN/Themes/Economic-Sectors-Enterprises/Energy/Production/Tables/gross-electricity-production.html Accessed: 2024-05-14

E   Licenses

In Table 7 the Licenses as given by the owners of the Dataset/Framework/Model are displayed.

Dataset/Framework/Model	License
Asclepius dataset[9]	Creative Commons Attribution Non Commercial Share Alike 4.0
MIMIC-IV-Note[10]	PhysioNet Credentialed Health Data License 1.5.0
MIMIC-IV-ED[11]	PhysioNet Credentialed Health Data License 1.5.0
MIMIC-SID[12]	MIT License
unsloth[13]	Apache License Version 2.0
Mistral-7B-I-v0.2[14]	Apache License Version 2.0
Llama-3-8B-I[15]	Llama 3 Community License Agreement
Llama-3-70B-I[16]	Llama 3 Community License Agreement
OpenBioLLM-70B[17]	Llama 3 Community License Agreement
WizardLM-2 8x22B[18]	MIT License
Phi-3-mini-128k-I[19]	Apache License Version 2.0

Table 7: Licenses of the dataset, Framework and Models used for this Shared Task.

---

[9]https://huggingface.co/datasets/starmpcc/Asclepius-Synthetic-Clinical-Notes Accessed: 2024-05-17
[10]https://physionet.org/content/mimic-iv-note/2.2/ Accessed: 2024-05-17
[11]https://physionet.org/content/mimic-iv-ed/2.2/ Accessed: 2024-05-17
[12]https://github.com/plandes/mimicsid Accessed: 2024-05-17
[13]https://github.com/unslothai/unsloth Accessed: 2024-05-17
[14]https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.2 Accessed: 2024-05-17
[15]https://huggingface.co/meta-llama/Meta-Llama-3-8B-Instruct Accessed: 2024-05-17
[16]https://huggingface.co/meta-llama/Meta-Llama-3-70B-Instruct Accessed: 2024-05-17
[17]https://huggingface.co/aaditya/Llama3-OpenBioLLM-70B Accessed: 2024-05-17
[18]https://huggingface.co/alpindale/WizardLM-2-8x22B Accessed: 2024-05-17
[19]https://huggingface.co/microsoft/Phi-3-mini-128k-instruct Accessed: 2024-05-17