LLaMA-Care

A Multimodal Medical Large Language Model for Hospital Discharge Instruction Generation

Joël Fehr
CAI - Centre for Artificial Intelligence
Zurich University of Applied Sciences
8400 Winterthur
joel.fehr@bluewin.ch

Project Page LLaMA-Care

Abstract

This study introduces LLaMA-Care, a Multimodal Large Language Model (MM-LLM) designed for the automated generation of patient hospital discharge instructions. LLaMA-Care aims to leverage various data modalities such as text, images, time series data and ICD codes, enhancing the efficiency and quality of discharge instruction generation. LLaMA-Care builds upon existing pre-trained encoders to create latent representations of different modalities which are further aligned with the LLM through Modality Bridges. The fine-tuning stage employs LoRA (Low-Rank Adaptation), ensuring efficient utilization of the existing model's strengths while adapting to the specific task in a time and memory efficient way. The model was evaluated against a baseline and a unimodal (textual input only) model, using ROUGE scores and a LLM-based evaluation method focusing on factual accuracy, completeness, and style/clarity. The findings indicate that LLaMA-Care outperforms the baseline model and demonstrates an improvement over the unimodal approach. The improvement is more pronounced in cases with limited textual data, underscoring the importance of multimodal inputs in enhancing the model's performance.

1 Introduction

In recent years, the integration of Artificial Intelligence (AI) into healthcare has marked a transformative shift in the field, offering new approaches of patient care [1]. These integrations encompass a broad spectrum of applications, from predictive analytics to applications in medical text generation. Central to the advancements in medical text generation is the role of Large Language Models (LLMs), which have started to redefine natural language processing (NLP) research and its applications in the medical domain [2]. LLMs have shown remarkable capabilities in understanding and generating complex medical texts [3].

The landscape of LLMs has expanded to include multimodal approaches, acknowledging that numerous applications do not only rely on textual input but rather on multiple modalities [4],

[5], [6], [7], [8], [9], [10]. Most of these implementations use vision and/or audio as additional input modalities to complement textual data, thereby enhancing the system's ability to process and interpret information in a manner similar to human perception. Most approaches are based on training light-weight adapters to align extracted features of modality specific pre-trained encoders to the token level representation of LLMs.

In the medical domain some approaches have been made to train LLMs on medical data in order to infuse medical knowledge into the models [11], [12], [13], [14]. But as other domains, the medical domain is heavily multimodal and there is a need for applications where multiple modalities can be used as input to these models. The need for multimodal approaches in healthcare is underscored by the diverse nature of medical data. Medical professionals often rely on a combination of textual information, visual cues from medical imaging, and electronic health records (EHR) data for accurate diagnosis and treatment planning. The multimodal LLMs (MM-LLMs), therefore represent a significant step forward in mimicking the multifaceted approach of human medical analysis, where multiple senses are engaged in patient assessment. A recent published study [15] proposes a new paradigm, referred to as generalist medical AI (GMAI), capable of processing a range of medical modalities like images, laboratory results, graphs, text and EHR data. This proposed paradigm is in sync with advancements like the XrayGPT [16] model, a conversational medical vision-language model that can analyze and answer questions about chest radiographs. Another recent study introduces HeLM [17] (Health Large Language Model for Multimodal Understanding), a framework that enables LLMs to use different medical modalities to estimate disease risks.

The ability to incorporate multiple modalities opens the door for applications like hospital discharge instructions generation from given input modalities. Discharge reports (An example can be found in Appendix: 5) summarize a patient's hospital stay, diagnoses, treatments, medications, and follow-up instructions and are important for documentation purposes and follow-up care of patients. The generation of instructions in such reports requires a lot of manual effort and is often time consuming for healthcare professionals. By leveraging the capabilities of MM-LLMs, the creation of discharge instructions can become more efficient and can help in assisting healthcare professionals.

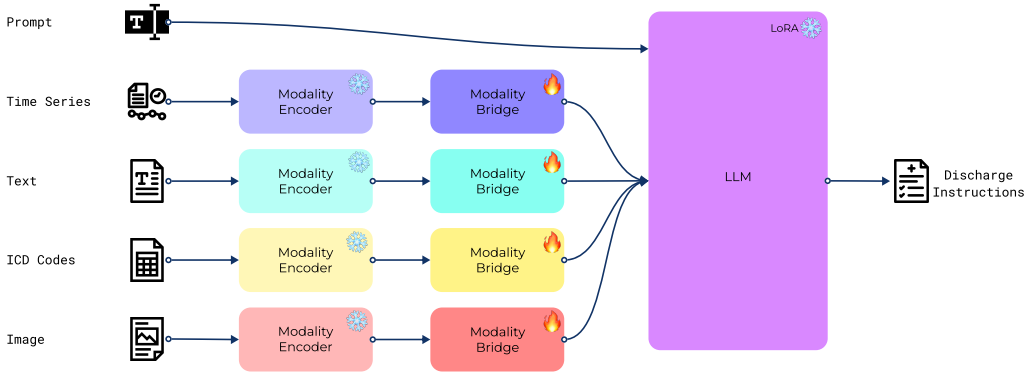

Figure 1: LLaMA-Care leverages pre-trained encoders to generate modality specific latent representations. These representations are aligned with the LLM by the Modality Bridges.

This work introduces **LLaMA-Care**, a MM-LLM specifically designed to automate the generation of hospital discharge instructions for patients. The MM-LLM leverages modalities as text, images, time series data and ICD codes to perform this task. An overview of the proposed model is depicted in Figure 1. LLaMA-Care is built on three main pillars: First, existing pre-trained encoders are leveraged to create a latent representation of the different modalities. Second LLaMA-Care uses modality specific LSTM's [18] and a linear transformation to bridge the modalities to the token embedding level of the LLM. These units are called Modality Bridges in this study. The third pillar is an open-source LLM in the form of a LLaMA model [19] which is used to generate the patient's discharge instructions based on the multimodal input.

To optimize efficiency and effectiveness, LLaMA-Care uses pre-trained encoders for different data modalities, coupled with Low-Rank Adaptation [20] (LoRA) fine-tuning. This approach strategically utilizes existing models' strengths while ensuring minimal but impactful modifications for task-specific tuning. Additionally, the Modality Bridges are trained to bridge the modalities to the token embedding level of the LLM, further aligning the representations while keeping the trainable parameters low.

2 Related Work

Multimodal Large Language Models Following the success of LLMs in various applications, researchers have turned their focus towards developing MM-LLMs to incorporate a wider range of input modalites beyond text. This shift was exemplified by the developement of models like CLIP [21] and Flamingo [22]. Most advancements in the multimodal area are based on integrating modalities such as images, videos and audio into the LLMs. A central component of these studies involves the utilization of specialized pre-trained encoders, which are tailored for individual modalities. These encoders are then coupled with an alignment procedure, which aims to bridge the modalities to the LLMs textual feature space.

MM-LLMs also emerged in the medical domain such, as HeLM [17], which incorporates multiple medical modalities to perform disease risk prediction. Other models in the domain of MM-LLMs include Med-MLLM [23] and Med-PaLM M [24]. Med-MLLM is a multimodal framework designed for rapid response in medical scenarios such as pandemics and Med-PaLM M can encode and interprete a vast range of biomedical data, such as clinical language, imaging, and genomics.

Instruction Tuning The study "Finetuned Language Models Are Zero-Shot Learners" [25] has shown an effective method to improve the zero-shot capabilities of LLMs purely based on instructions. This concept of instruction tuning has become increasingly significant in the context of LLMs, particularly for applications that require specific customization as the technique is able to serve a bridge between the general capabilities of pre-trained language models and the nature of the specific task [26].

Further, instruction tuning has found its application in the multimodal context. This expansion signifies the adaptation of the technique beyond text-based LLMs to models that can process and interpret multiple forms of data, such as images [27]. Instruction tuning represents a significant advancement in the field of AI and LLMs. It not only enhances the zero-shot learning capabilities of these models but also extends their applicability across various domains, including multimodal contexts. This makes LLMs more adaptable and effective in meeting the diverse needs of different applications.

Low-Rank Adaptation of Large Language Models LoRA [20], is a technique developed to efficiently fine-tune LLM's. It presents a viable solution to the challenges of full-parameter fine-tuning, which becomes increasingly impractical as model sizes grow. LoRA works by freezing the pre-trained model weights and introducing trainable rank decomposition matrices to each layer of the Transformer [28] architecture. These matrices operate on a lower-dimensional space and significantly reduce the number of parameters that need to be trained. This reduction in trainable parameters results in a more efficient and practical fine-tuning process. A comparative study [29] has shown significant training cost benefits compared to full-parameter fine-tuning. The study also states the importance of the base model. The effectiveness of LoRA-based tuning benefits substantially from the number of model parameters.

3 Model Architecture

The LLaMA-Care framework is built on three main modules: the Modality Encoders, the Modality Bridges and the LLM.

Modality Encoders The LLaMA-Care framework makes use of four distinct modalities. These modalities are time series data, textual data, ICD codes and images. The time series data includes lab measurements such as blood gas levels or hematology measurements and also vital signs such as heart rate or blood pressure and procedures like electroencephalograms. ICD stands for International Classification of Diseases and ICD codes provide a standardized system for diagnosing and classifying diseases and health conditions. The images are in the form of CT scans and the textual data are the

associated radiology notes. Each of these modalities was first encoded into a latent representation by leveraging pre-trained encoders.

Time Series: In order to encode the time series, for each event a set of statistical features is calculated. These statistical features include maximum, minimum, mean and variance to understand the data's range, central tendency and spread. Other features are the mean difference, mean absolute difference, maximum difference, sum of absolute differences and end-to-end difference to capture the average rate of change, variation magnitude, largest single-period change, overall change and overall directionality. To further analyze the temporal dynamics a peak detection is applied over the series of events. The peak detection identifies local maxima by comparing adjacent values in the time series. Peaks are characterized based on criteria such as height, threshold, distance, prominence, and width, allowing for a differentiation of significant peaks from minor fluctuations. Additionally, the linear trend of each time series is computed by fitting a first-degree polynomial. This trend calculation, represented by the slope of the fitted line, provides insight into the overall directional movement of the series over time. For times series with a single data point, the trend is set to zero, acknowledging the absence of directional tendencies in such cases. The resulting time series embedding per sample is a vector with 451 dimensions (9 statistical features, 1 peak variable and 1 linear trend variable for 41 distinct events). The encoding of the time series is based on the previous work of Soenksen et al. [30]. The modality encoder for the time series data can be formulated in the following way:

Consider a set of time series, denoted as $T_j = \{t_{j,1}, t_{j,2}, \ldots, t_{j,41}\}$, which is composed of 41 distinct events. Each event, represented by $t_{j,i}$, is based on a laboratory measurement, a procedures, or a vital sign. The encoding process for one sample j can be described as follows:

Set of statistical features:

$$F_j^{\text{stat}}(t_{j,i}) = \{\max(t_{j,i}), \min(t_{j,i}), \text{mean}(t_{j,i}), \text{var}(t_{j,i}), \ldots\} \in \mathbb{R}^{1\times 9}$$

Let $P_j(t_{j,i})$ denote the set of peaks in $t_{j,i}$. A peak $p \in P_j(t_{j,i})$ is identified based on criteria such as height, threshold, and distance.

If $t_{j,i}$ is represented as $\{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$, the linear trend $L_j(t_{j,i})$ is a fitted first degree polynomial.

The final encoding for a time series can be represented as a vector combining these features:

$$E_{\text{TS}}(T_j) = \text{concatenate}([F_j^{\text{stat}}(t_{j,i}), |P_j(t_{j,i})|, L_j(t_{j,i})]) \quad \forall i \in \{1, \ldots, 41\},$$

forming an embedding in $\mathbb{R}^{1\times 451}$.

ICD Codes: In this study ICD codes were encoded using the node2vec [31] algorithm within the `icdcodex`[1] framework, initially converting ICD-10 codes to their ICD-9 counterparts to maintain consistency. The node2vec algorithm embeds ICD codes into a 512-dimensional vector space. This embedding captures the hierarchical and contextual relationships among the codes. The process involves a biased random walk mechanism, which effectively balances the exploration of both local and global structural properties of the ICD code graph. Through the `icdcodex` package this study leveraged node2vec's capabilities to generate dense vector representations. These embeddings ensure that ICD codes with similar contexts and hierarchies are positioned closely in the vector space, thus reflecting their medical and categorical relationships. The encoding of ICD codes can be formulated in the following way:

Let $C_j = \{c_{j,1}, c_{j,2}, \ldots, c_{j,m}\}$ be the set of ICD codes for an individual sample j. The encoding using `node2vec` can be represented as:

$$e_{j,i}^{\text{ICD}} = \texttt{node2vec}(c_{j,i}) \quad \forall i \in \{1, \ldots, m\},$$
$$E_j^{\text{ICD}} = \text{concatenate}([e_{j,1}^{\text{ICD}}, e_{j,2}^{\text{ICD}}, \ldots, e_{j,m}^{\text{ICD}}])$$

resulting in an embedding in $\mathbb{R}^{m\times 512}$, representing the ICD codes in a hierarchical and contextual manner.

[1] https://icd-codex.readthedocs.io/en/latest/index.html

Textual data: This study utilized the BiomedVLP-BioViL-T[2] [32] model to extract radiological sentence embeddings. This model is specifically designed for processing domain specific medical texts and integrates a BERT [33] based language model. By projecting textual data into a 128-dimensional space, this model ensures that the resulting embeddings are rich in contextual and semantic details. The encoding stage for textual data can be formulated in the following manner:

Given textual notes $D_j = \{d_{j,1}, d_{j,2}, \ldots, d_{j,k}\}$, where $d_{j,i}$ are documents belonging to one sample j, the encoding using `BiomedVLP-BioViL-T` can be represented as:

$$e_{j,i}^{\text{text}} = \texttt{BiomedVLP-BioViL-T}(d_{j,i}) \quad \forall i \in \{1, \ldots, k\},$$
$$E_j^{\text{text}} = \text{concatenate}([e_{j,1}^{\text{text}}, e_{j,2}^{\text{text}}, \ldots, e_{j,k}^{\text{text}}])$$

This results in an embedding in $\mathbb{R}^{k \times 128}$, capturing contextual and semantic details of sentences.

Images: For CT scan encoding this study leveraged the `TorchXRayVision`[3] [34] library. The library is designed for working with chest X-ray scans and provides a set of classification and representation learning models. This work utilized a DenseNet [35] model from `TorchXRayVision` and made use of a feature extraction function which conducts a forward pass through the model and captures high-level features from the CT scans at a specific point in the computation graph. The resulting embeddings are 1024-dimensional vectors. The encoding of the images can be formulated in the following way:

Let $S_j = \{s_{j,1}, s_{j,2}, \ldots, s_{j,l}\}$ represent a set of CT scans of an individual sample j. Then the encoding using a DenseNet model from TorchXRayVision can be represented as:

$$e_{j,i}^{\text{img}} = \texttt{DenseNet}_{\texttt{TorchXRayVision}}(s_{j,i}) \quad \forall i \in \{1, \ldots, l\},$$
$$E_j^{\text{img}} = \text{concatenate}([e_{j,1}^{\text{img}}, e_{j,2}^{\text{img}}, \ldots, e_{j,l}^{\text{img}}])$$

The result is an embedding in $\mathbb{R}^{l \times 1024}$, capturing high-level features from the CT scans.

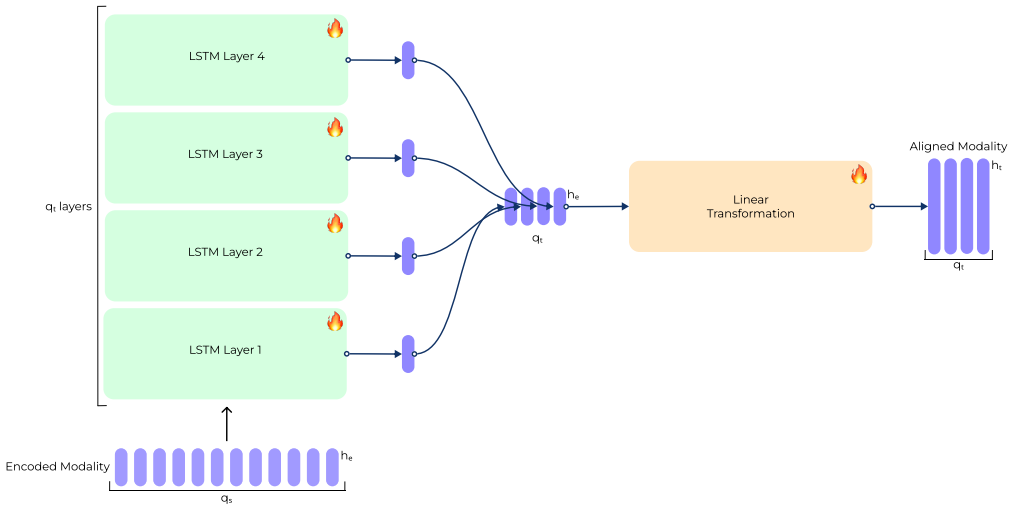

Figure 2: The Modality Bridges are based on LSTM networks and a Linear Transformation. The LSTM part has q_t layers which correspond to the number of tokens used in the LLM.

Modality Bridges: A single patient's hospital stay, which constitutes one sample, often encompasses multiple CT scans, radiology reports, and ICD codes. Due to this, the encoded embeddings representing this data are not one-dimensional but evolve into matrices that incorporate a dimension for

[2] https://huggingface.co/microsoft/BiomedVLP-BioViL-T
[3] https://github.com/mlmed/torchxrayvision

time. The LLaMA-Care framework incorporates Modality Bridges which are depicted in Figure 2. These Modality Bridges utilize LSTM networks to manage sequential data, effectively capturing the temporal relationships within the data. Additionally, they employ a linear transformation to project the encoded embeddings into the LLM token-level space. Each of the four distinct data modalities is processed through its own dedicated Modality Bridge. Within each bridge, there exist multiple layers, denoted as q_t, which correspond to the number of tokens the modality is represented in the LLM. The initial modality embedding is characterized by a shape of (q_s, h_e), signifying the sequence length and the embedding's feature dimension, respectively. This embedding is introduced into the first layer of the LSTM network. Each subsequent LSTM layer processes the information in a sequential manner and yields its final hidden state. As a result, the concatenated outputs of the last hidden states from each of the q_t LSTM layers form an embedding with the dimensions of (q_t, h_e). Following the LSTM layers, the concatenated hidden states undergo a linear transformation. This step is designed to modify the embedding dimensions to align with the feature dimensions of the LLM token embeddings, resulting in a final embedding with dimensions (q_t, h_t). Here, h_t indicates the adjusted feature dimension that corresponds to the LLMs embedded token dimension. Figure 3 shows how the modality tokens are combined with the textual prompt and gives a visual intuition about the hyperparameter q_t. A more detailed description on how the modalities and text are structured can be found in Section 4.3. The Modality Bridges for an initial modality embedding M with shape (q_s, h_e) can be formulated in the following way:

The last hidden state $h^M_{i,T}$ of each LSTM layer i is concatenated to an embedding

$$H^M = \text{concatenate}([h^M_{1,T}, h^M_{2,T}, ..., h^M_{q_t,T}])$$

with dimensions (q_t, h_e).

The last hidden state in a LSTM is calculated:

$$\begin{aligned}
i^M_t &= \sigma(W^M_{ii} x_t + b^M_{ii} + W^M_{hi} h^M_{i,t-1} + b^M_{hi}), \\
f^M_t &= \sigma(W^M_{if} x^M_t + b^M_{if} + W^M_{hf} h^M_{i,t-1} + b^M_{hf}), \\
\tilde{c}^M_{i,t} = \tanh(W^M_{ic} x^M_t + b^M_{ic} + W^M_{hc} h^M_{i,t-1} + b^M_{hc}) c^M_{i,t} &= f^M_t * c^M_{i,t-1} + i^M_t * \tilde{c}^M_{i,t}, \\
o^M_t &= \sigma(W^M_{io} x^M_t + b^M_{io} + W^M_{ho} h^M_{i,t-1} + b^M_{ho}), \\
h^M_{i,t} &= o^M_t * \tanh(c^M_{i,t}),
\end{aligned}$$

where σ denotes the sigmoid function, W the weight matrices and b the bias vectors.

The concatenated hidden states are then processed by a linear transformation:

$$E^M = H^M \cdot W^M + b^M,$$

with weight W^M and bias b^M, forming the final aligned embedding E^M for modality M with shape (q_t, h_t).

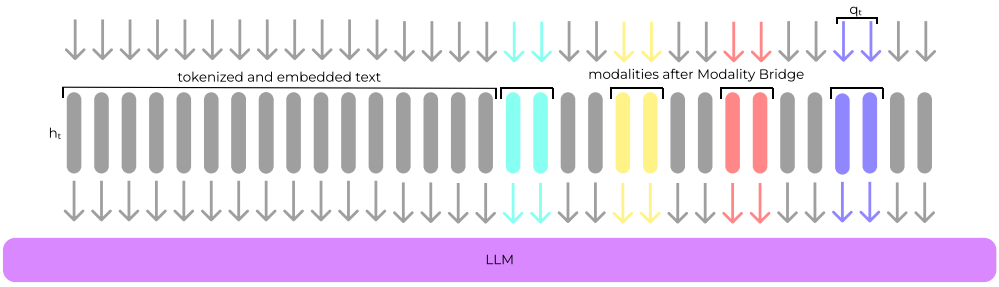

Figure 3: The aligned multimodal embeddings are concatenated with the textual prompt and fed into the LLM.

Large Language Model LLaMA-Care uses Meditron-7b [36], [37] as backbone LLM. Meditron-7b is the smallest model of the open-source Meditron family. The model is adapted to the medical

domain through continued pre-training of Llama-2-7B [19] on a curated medical corpus. Meditron-7b is a foundation model without fine-tuning or instruction-tuning and was leveraged as the building block of the LLM in the LLaMA-Care framework.

Trainable Parameters In LLaMA-Care, training is strategically divided into two stages, optimizing the learning process of the model. In the first stage only the Modality Bridges are trained in order to align the multimodal embeddings with the LLMs token level representations and all weights of the LLM are kept entirely frozen. In the second stage the Modality Bridges are still open for weight adjustments and LoRA fine-tuning is implemented for the LLM. This stage fine-tunes the LLM on the specific instruction task. LoRA only adapts a subset of the parameters of the LLM and makes sure that the initial abilities of the LLM are preserved. The exact number of parameters in each Modality Bridge differs across the different modalities because the dimensions of the encoded modality embeddings are not the same. In total the trainable parameters for all Modality Bridges sum up to approximately 95 million. The second stage extends the training to use LoRA to target all linear layers. During this stage LoRA induces 8 million trainable parameters, making **1.2 %** of the whole model's parameters trainable. An overview of the parameters and their trainability can be found in Table 1.

	Encoder		Modality Bridge		LLM	
	Model	Params	Model	Params	Model	Params
Text	BiomedVLP-BioViL-T ❄	$\approx 109M$	LSTM + Linear Transformation 🔥	$\approx 11M$		
Time Series	-	-	LSTM + Linear Transformation 🔥	$\approx 14M$	Meditron-7b ❄	$\approx 7B$
ICD Code	icdcodex ❄	$\approx 14M$	LSTM + Linear Transformation 🔥	$\approx 24M$	(LoRA 🔥)	$\approx 8M$
Image	TorchXRayVision DenseNet ❄	$\approx 7M$	LSTM + Linear Transformation 🔥	$\approx 45M$		

Table 1: Overview of trainable and frozen parameters.

4 Experiments

4.1 Dataset

This study made use of three datasets of PhysioNet [4], namely MIMIC-IV [38], MIMIC-CXR-JPG [39] and MIMIC-IV-Note[40].

MIMIC-IV is a comprehensive collection of de-identified health data from over 40,000 patients admitted to intensive care units at Beth Israel Deaconess Medical Center. The data includes detailed patient demographics, hospitalizations, lab measurements and medication prescriptions. The dataset aims to facilitate a wide range of healthcare research while maintaining patient privacy.

MIMIC-CXR-JPG includes 377,110 chest radiographs in the JPG format and is derived from the MIMIC-CXR [41] dataset. The dataset was created to facilitate research in medical image understanding and analysis, providing a more accessible JPG format of the original DICOM images. The dataset includes metadata and structured labels and is fully de-identified.

MIMIC-IV-Note includes 331,794 de-identified discharge summaries of hospitalized patients and 2,321,355 radiology reports. This dataset is particularly valuable for research in clinical natural language processing, offering a rich source of free-text clinical notes linked to MIMIC-IV's clinical data.

4.2 Data Extraction and Preprocessing

This study incorporates methodologies from Soenksen et al.'s [30] work, leveraging their approach to create a multimodal dataset from the three MIMIC datasets. The dataset for LLaMA-Care is designed to meet the specific requirements and is inspired by the original HAIM-MIMIC-MM dataset and follows similar data extraction and processing methods.

This study attempted to create single files for each unique hospital stay. This approach differs from the original approach where the data was aggregated on a file per patient level. In the newly aggregated dataset only those hospital stays that had associated electronic health records (EHR), radiological

[4] https://physionet.org

images, ICD codes and clinical notes were considered. Each admission file is a rich archive of information such as demographic details, laboratory measurements, medication administrations and prescriptions. Additionally the files contain radiological images, reports and discharge summaries. Due to the missing admission identifier in the MIMIC-CXR-JPG dataset an attempt was made to match the images to the admissions based on the subject identifier and the time stamps of the admissions.

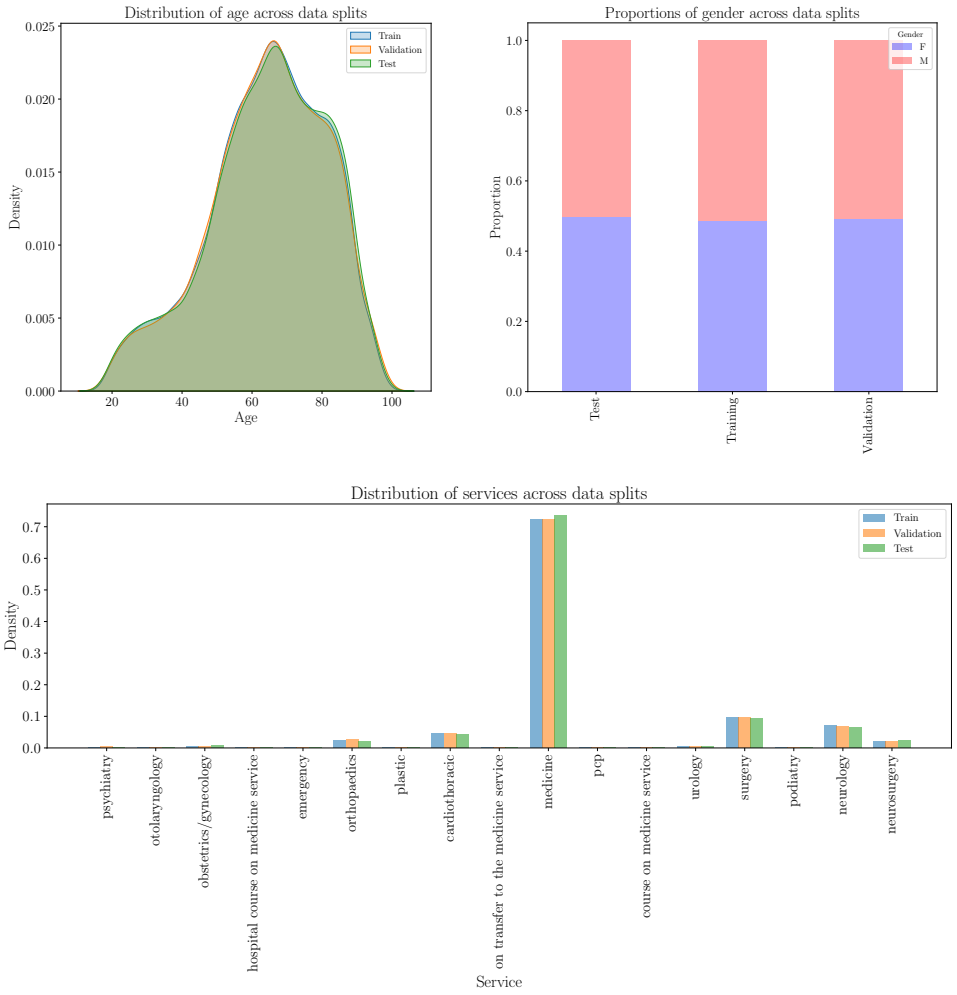

Figure 4: Visual representation of the balance across the three data splits based on age, gender, and services. The top-left plot shows the density distribution of age for the training, validation, and test sets. The top-right bar chart details the proportions of gender for each data split. The bottom plot presents the distribution of different services within the data splits.

Initially, a total of 64,571 samples (hospital admissions) matched the inclusion criteria and were extracted. To enhance the quality of the dataset 1,201 entries were removed because the relevant section for the discharge instructions was not present in the sample or could not be reliably extracted. Most of these patients without the relevant section passed away during the hospital stay and had therefore no listed follow-up instructions. Another few samples were removed based on entries without clear required services and chief complaints. The curated dataset contains 62,927 samples. To facilitate robust model training and evaluation, the dataset was divided into three random subsets: training, validation, and testing data. After splitting the sets were analyzed on age, gender, and the required services of the patients to ensure balance. An oveview of these three variables and their distributions across the datasplits can be found in Figure 4. The training set comprises 37,756 entries

(60%), forming the foundation for the model's learning process. The validation and testing sets are equally sized, each containing 12,585 and 12,586 entries (20% each), respectively.

As an additional step to improve the quality and standardization of patient discharge instructions, this study employed the capabilities of Mistral-7B-Instruct [42], a open-source LLM to refine and standardize the discharge instructions, which often varied in structure in their original form. Mistral-7B-Instruct was instructed to systematically process the existing discharge notes and to standardize them into a listed keypoint format. The prompt used for the conversion was the following:

> *"Please read the following text and summarize it by listing the most important points in a compact, bullet-point format. Each point should begin with an asterisk (*) and should strictly contain only facts derived from the text. Avoid including any opinions, interpretations, or information not explicitly mentioned in the text. TEXT:* `{text}`*".*

4.3 Experimental Setup

The LLM used as a backbone for LLaMA-Care is the medical foundation model Meditron-7B [5] [36], [37]. Meditron-7B is adapted to the medical domain through continued pre-training of Llama-2-7B [19]. The model was adapted with the help of a curated medical corpus and general domain data.

The main goal of the experimental setup was to determine the effectiveness of multimodal inputs. In order to evaluate this goal, two instruction-tuned variants of Meditron-7B were trained, a unimodal (textual data only) and a multimodal variant. The multimodal variant was first trained in an aligment stage in which the weights of the LLM were kept frozen and only the weights of the Modality Bridges were updated. The fine-tuning stage was the same for the multimodal and unimodal variant, utilizing a standard setting with LoRa rank of 8, LoRA alpha of 16 and LoRa dropout of 0.1. LoRA targeted all linear transformations in the LLM.

For the purpose of instruction tuning, the prompt is as follows:

> *### Instruction:*
> *You are a helpful medical assistant. Write hospital discharge instructions for the patient based on the given input.*
>
> *### Input:*
> *A patient, identified as* `{gender}`*, aged* `{age}` *was admitted to the hospital and required the service* `{service}` *due to* `{chief_complaint}`*. Here is some brief information about the patient: Major Procedure:* `{major_procedure}` *, Discharge Diagnosis:* `{discharge_diagnosis}` *, Discharge Medications:* `{discharge_medications}`*.* `{overview_of_stay}`*.*
> *Additionally you have information about: CT scans: <ct>*`{embedding}`*</ct>, ICD codes: <icd>*`{embedding}`*</icd>, radiology notes: <notes>*`{embedding}`*</notes> and lab/vital values </lab>*`{embedding}`*<lab>.*

In the prompt above it is visible in detail what kind of information was fed into the model. Beside the demographic variables gender and age, further details about the patient like the required hospital service, chief complaint and the major procedure applied during hospitalization were used as additional information. Further the discharge diagnosis and the discharge medications were used in combination with a short overview of the hospital stay. An example of the relevant sections in an original report can be found in the Appendix: 5. The prompt shown in the gray box also shows how the multiple modalities were fed into the multimodal model by combining it with the textual prompt. The unimodal model made use of the exact same prompt, just the part starting with *"Additionally you have"* was omitted.

Both models were then evaluated with the use of selected evaluation metrics against each other and a baseline. As a baseline meditron-7b-chat [6] was used, which is a variant of Meditron-7B which was fine-tuned using supervised fine-tuning. This model was chosen as a baseline because it has undergone a fine-tuning stage on a instruction-following dataset in constrast to the base model Meditron-7B. For the purpose of evaluation, the baseline model utilized the identical prompt as the unimodal model.

[5] https://huggingface.co/epfl-llm/meditron-7b

[6] https://huggingface.co/malhajar/meditron-7b-chat

All experiments were conducted with the same hyperparameters on a single NVIDIA A100 80GB GPU. Due to hardware limitations the batch size was set to 2 in combination with 128 gradient accumulation steps, yielding an effective batch size of 256. The learning rate was set to 9e-4 during the alignment stage of the unimodal model and 2e-4 during the fine-tuning stage for the multimodal and unimodal model. For optimization, the Adam optimizer was used, with its β_1 and β_2 parameters set to 0.9 and 0.999, respectively, and an ϵ value of 1e-8. The multimodal model was trained for one epoch during the alignment stage. The multimodal and unimodal model were both trained for 5 epochs during the fine-tuning stage. The selected hyperparameters, including the learning rates and optimizer parameters, fall within the range of standard values commonly used in similar training procedures. The choice of 128 gradient accumulation steps was specifically made to smooth the loss curve. It was observed that a single epoch for the alignment stage of the multimodal model might be sufficient, as subsequent epochs showed minimal improvement. Due to the constraints of hardware costs and limited computing time, extensive hyperparameter testing was not feasible. The hyperparameter q_t which determines the number of layers in the Modality Bridges and the number of tokens the modalities are represented in the LLM were used in the following configuration: $q_t = 6$ for ICD codes and radiology notes, $q_t = 4$ for CT scans and $q_t = 2$ for the time series data. The setting of q_t is guided by previous implementations [17], the initial embedding size and hardware contraints.

4.4 Evaluation Methods

ROUGE (Recall-Oriented Understudy for Gisting Evaluation) is a set of metrics used to evaluate summarization and machine translation tasks. The metric compares an automatically generated summary with a reference summary and calculates precision, recall, and F1-score by considering the presence and ordering of n-grams. This study specifically employed ROUGE-1 and ROUGE-L metrics.

ROUGE-1 focuses on the overlap of unigrams (single words) between the generated and reference texts and can be defined as:

Let G_{unigrams} be the set of unigrams in the generated text, and R_{unigrams} be the set of unigrams in the reference text. Then, the resulting ROUGE-1 score can be calculated as:

Precision P is the proportion of unigrams in the generated text that are also in the reference text:

$$P = \frac{|G_{\text{unigrams}} \cap R_{\text{unigrams}}|}{|G_{\text{unigrams}}|}$$

Recall R is the proportion of unigrams in the reference text that are also in the generated text:

$$R = \frac{|G_{\text{unigrams}} \cap R_{\text{unigrams}}|}{|R_{\text{unigrams}}|}$$

F1-Score for ROUGE-1, which is the harmonic mean of Precision and Recall:

$$\text{F1-Score} = 2 \times \frac{P \times R}{P + R}$$

In these formulas, $|G_{\text{unigrams}} \cap R_{\text{unigrams}}|$ represents the count of common unigrams between the generated and reference texts, $|G_{\text{unigrams}}|$ is the total count of unigrams in the generated text, and $|R_{\text{unigrams}}|$ is the total count of unigrams in the reference text.

ROUGE-L, on the other hand, assesses the longest common subsequence (LCS) between the generated and reference texts. It considers the sequence of words that appear in the same order in both texts. The ROUGE-L score can be defined as follows:

Let G_{words} be the sequence of words in the generated text, and R_{words} be the sequence of words in the reference text. LCS Length is the length of the longest common subsequence between G_{words} and R_{words}.

Precision P for ROUGE-L is the proportion of the LCS length to the length of the generated text:

$$P = \frac{\text{LCS Length}}{|G_{\text{words}}|}$$

Recall R for ROUGE-L is the proportion of the LCS length to the length of the reference text:

$$R = \frac{\text{LCS Length}}{|R_{\text{words}}|}$$

F1-Score for ROUGE-L:

$$\text{F1-Score} = 2 \times \frac{P \times R}{P + R}$$

In these formulas, $|G_{\text{words}}|$ is the total number of words in the generated text, and $|R_{\text{words}}|$ is the total number of words in the reference text. The ROUGE-L score primarily measures the longest sequence of words that the generated and reference texts have in common, thereby evaluating the fluency and structure of the generated text.

A higher ROUGE score signifies stronger alignment between the created text and the reference, suggesting improved quality and coherence. This score is a valuable quantitative tool for objectively evaluating text generation abilities.

LLM based evaluation As an additional evaluation method this study employed the new open-source Mixtral-8x7B-Instruct [7] model to evaluate the generated discharge instructions based on a given reference. The evaluation focused on three criteria: factual accuracy, completeness, and style/clarity. These criteria have been chosen because factual accuracy is important in medical notes to prevent misinformation and hallucinations, completeness is necessary to ensure that no important details are missing, and style and clarity are vital for straightforward understanding. Each of these criteria was rated on a ordinal scale from 1 to 5 by the model and model was prompted in the following way:

Factual accuracy: This criterion is used to evaluate the precision of the information provided generated notes compared to the reference notes. A score of 1 indicates significant factual errors, while a score of 5 denotes complete accuracy on a ordinal scale from 1-5.

Completeness: This criterion assesses whether the generated notes encompass all critical elements found in the reference. Scores are on a ordinal scale and range from 1, indicating many missing details, to 5, indicating a note without missing details.

Style and clarity: This criterion judges the readability and professional formatting of the notes with respect to the reference note. A score of 1 reflects poor style and clarity, while a score of 5 suggests a clear and well-structured note on a ordinal scale from 1-5.

This evaluation allows for a detailed and balanced assessment of the generated instructions. The three criterions are carefully chosen and are important aspects of medical instructions. By implementing this structured evaluation, the study aims to evaluate the generated instructions based on relevant content and not scores that are calculated in a strict mathematical manner.

4.5 Results

	ROUGE-1	ROUGE-L	Factual accuracy	Completeness	Style and clarity
baseline	0.016 ± 0.045	0.015 ± 0.041	3.124 ± 1.307	2.220 ± 1.091	3.460 ± 1.029
unimodal	0.343 ± 0.113	0.316 ± 0.112	4.056 ± 0.611	3.283 ± 0.665	$\mathbf{3.488 \pm 0.506}$
multimodal	$\mathbf{0.371 \pm 0.110}$	$\mathbf{0.349 \pm 0.111}$	$\mathbf{4.167 \pm 0.578}$	$\mathbf{3.388 \pm 0.652}$	3.480 ± 0.505

Table 2: Results of comparative experiments based on ROUGE scores and LLM evaluation, calculated on the test set.

This study made an attempt to build a MM-LLM to automatically generate discharge instructions for hospitalized patients. The experiments included a comparative analysis between a baseline model (instruction-tuned on general data) and instruction-tuned variants with multimodal or unimodal inputs. The tabulated evaluation metrics of the comparative analysis can be found in Table 2.

The comparative analysis showed that both instruction-tuned variants outperform the baseline by a large margin. It is not a fair comparison to compare the instruction-tuned variants with the baseline

[7] https://huggingface.co/mistralai/Mixtral-8x7B-Instruct-v0.1

based on ROUGE scores because the desired output instructions have a specific bullet point format which was not specifically prompted. Therefore this result can be expected because the ROUGE scores are influenced a lot by the structure of the text. The instruction-tuned variants do not only outperform the baseline based on the ROUGE score but also in the evaluation based on factual accuracy and completeness. The instruction-tuned variants are consistantly more accurate and complete than the baseline. When it comes to the style and clarity of the notes, the baseline does perform equally well as the both instruction-tuned variants. This can be explained because style and clarity is more a general than a domain specific language capability.

The comparative analysis also showed that the multimodal variant does outperform the unimodal variant by a small margin when considering the ROUGE scores. The multimodal model shows an improved performance in the LLM-based evaluation as well but does not significantly outperform the unimodal model. This outcome is likely caused by the nature of the discharge instructions which might not always depend on the multimodal input and can be largely reconstructed by the textual information in the prompt.

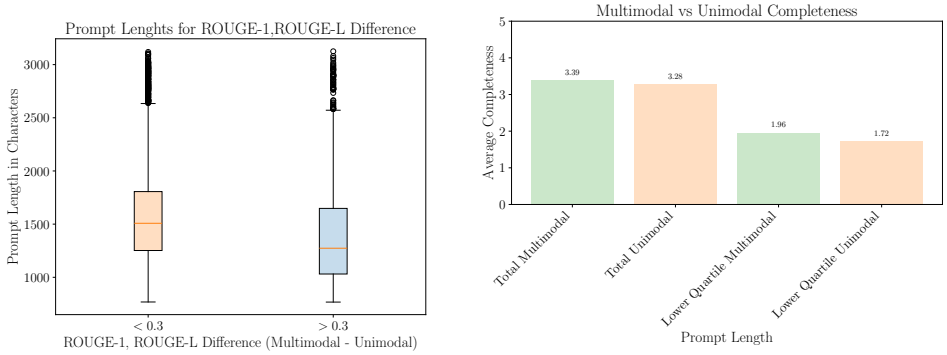

Figure 5: Comparison of ROUGE-1 and ROUGE-L scores between multimodal and unimodal completions across different prompt lengths. The box plot on the left illustrates prompt lengths categorized by ROUGE-1 and ROUGE-L score discrepancies (multimodal - unimodal). The bar chart on the right displays the average LLM rating "Completeness" for full prompt lengths and prompt lengths in the lower quartile.

A deeper analysis showed that instructions with ROUGE-1 or ROUGE-L scores at least 0.3 higher in the multimodal setting than the unimodal setting typically had about 100 characters less in their textual prompts than samples with a smaller difference. This might suggest that the multimodal setting is mostly beneficial for samples with less textual input. Moreover, it was observed that samples with prompt lengths in the lower quartile exhibited lower completeness but the performance gap between unimodal and multimodal settings grew from approximately 0.1 to 0.25. These two observations are visually shown in Figure 5. This finding suggests that the textual input is overall relevant for the performance of the model but that the multimodal variant might be able to utilize information from the multimodal sources to compensate the lack of textual information. Further analysis revealed that when the reference instruction is less than 200 characters, ROUGE scores and LLM-based scores often reflect differing performance levels. This discrepancy can be attributed to the model's maximum token limit of 250 during inference which tends to generate texts longer than 200 characters. If a reference text is short but the generated text is longer, there's a smaller chance in n-gram overlap.

5 Conclusion

This study introduced LLaMA-Care, a Multimodal Large Language Model (MM-LLM) specifically designed to automate the generation of patient discharge instructions. The model utilizes an integration of various data modalities, including text, images, time series data, and ICD codes, highlighting the potential of multimodal approaches in enhancing AI applications in healthcare.

The experiments, leveraging datasets from PhysioNet, demonstrated that LLaMA-Care could partially utilize multimodal inputs to generate coherent and accurate discharge instructions. The comparative analysis between the multimodal and unimodal variants revealed that the multimodal approach generally outperforms its unimodal counterpart by a small margin. The higher performance is especially observed in scenarios with limited textual input, where the model might effectively compensate with information from other modalities. An interesting experiment to consider would be the omission of the patient overview and other textual data to observe how both unimodal and multimodal model perform under these conditions.

It's crucial to note that this study serves primarily as a proof of principle. Further research is needed to fully exploit the potential of multimodal data integration in LLMs. The model's performance might improve with the availability of cleaner data and instructions that are more dependant on multimodal information. Additionally, exploring different Modality Bridges and modality encoders may further enhance the model's effectiveness. A limitation in this implementation is the lack of consideration for the time dimension in cross-modal contexts, as it is only handled within separate modalities. This limitation could potentially be addressed by using Transformer models with positional encoding to serve as Modality Bridges. There is a potential benefit in employing a larger LLM as the backbone for the model and should be explored further. A larger LLM might offer more sophisticated understanding and generation capabilities, further boosting the effectiveness of the multimodal approach. In conclusion, while LLaMA-Care shows the potential of automating complex tasks in healthcare using MM-LLMs, it is important to continue research in this direction.

Acknowledgements

I thank Dr. Jasmina Bogojeska for her guidance and support on this project and the Centre for Artificial Intelligence for enabling my Master's studies at the institute.

References

[1] Shuroug A. Alowais et al. "Revolutionizing healthcare: the role of artificial intelligence in clinical practice". In: *BMC Medical Education* 23.1 (Sept. 2023). DOI: 10.1186/s12909-023-04698-z. URL: https://doi.org/10.1186/s12909-023-04698-z.

[2] Harsha Nori et al. *Capabilities of GPT-4 on Medical Challenge Problems*. 2023. DOI: 10.48550/ARXIV.2303.13375. URL: https://arxiv.org/abs/2303.13375.

[3] Arun James Thirunavukarasu et al. "Large language models in medicine". In: *Nature Medicine* 29.8 (July 2023), pp. 1930–1940. DOI: 10.1038/s41591-023-02448-8. URL: https://doi.org/10.1038/s41591-023-02448-8.

[4] Junnan Li et al. *BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models*. 2023. DOI: 10.48550/ARXIV.2301.12597. URL: https://arxiv.org/abs/2301.12597.

[5] Danny Driess et al. *PaLM-E: An Embodied Multimodal Language Model*. 2023. DOI: 10.48550/ARXIV.2303.03378. URL: https://arxiv.org/abs/2303.03378.

[6] Haotian Liu et al. *Visual Instruction Tuning*. 2023. DOI: 10.48550/ARXIV.2304.08485. URL: https://arxiv.org/abs/2304.08485.

[7] Deyao Zhu et al. *MiniGPT-4: Enhancing Vision-Language Understanding with Advanced Large Language Models*. 2023. DOI: 10.48550/ARXIV.2304.10592. URL: https://arxiv.org/abs/2304.10592.

[8] Peng Gao et al. *LLaMA-Adapter V2: Parameter-Efficient Visual Instruction Model*. 2023. DOI: 10.48550/ARXIV.2304.15010. URL: https://arxiv.org/abs/2304.15010.

[9] Hang Zhang, Xin Li, and Lidong Bing. *Video-LLaMA: An Instruction-tuned Audio-Visual Language Model for Video Understanding*. 2023. DOI: 10.48550/ARXIV.2306.02858. URL: https://arxiv.org/abs/2306.02858.

[10] Shengqiong Wu et al. *NExT-GPT: Any-to-Any Multimodal LLM*. 2023. DOI: 10.48550/ARXIV.2309.05519. URL: https://arxiv.org/abs/2309.05519.

[11] Yunxiang Li et al. *ChatDoctor: A Medical Chat Model Fine-Tuned on a Large Language Model Meta-AI (LLaMA) Using Medical Domain Knowledge*. 2023. DOI: 10.48550/ARXIV.2303.14070. URL: https://arxiv.org/abs/2303.14070.

[12] Karan Singhal et al. "Large language models encode clinical knowledge". In: *Nature* 620.7972 (July 2023), pp. 172–180. DOI: 10.1038/s41586-023-06291-2. URL: https://doi.org/10.1038/s41586-023-06291-2.

[13] Chaoyi Wu et al. *PMC-LLaMA: Towards Building Open-source Language Models for Medicine*. 2023. DOI: 10.48550/ARXIV.2304.14454. URL: https://arxiv.org/abs/2304.14454.

[14] Karan Singhal et al. *Towards Expert-Level Medical Question Answering with Large Language Models*. 2023. DOI: 10.48550/ARXIV.2305.09617. URL: https://arxiv.org/abs/2305.09617.

[15] Michael Moor et al. "Foundation models for generalist medical artificial intelligence". In: *Nature* 616.7956 (Apr. 2023), pp. 259–265. DOI: 10.1038/s41586-023-05881-4. URL: https://doi.org/10.1038/s41586-023-05881-4.

[16] Omkar Thawkar et al. *XrayGPT: Chest Radiographs Summarization using Medical Vision-Language Models*. 2023. DOI: 10.48550/ARXIV.2306.07971. URL: https://arxiv.org/abs/2306.07971.

[17] Anastasiya Belyaeva et al. *Multimodal LLMs for health grounded in individual-specific data*. 2023. DOI: 10.48550/ARXIV.2307.09018. URL: https://arxiv.org/abs/2307.09018.

[18] Sepp Hochreiter and Jürgen Schmidhuber. "Long Short-Term Memory". In: *Neural Computation* 9.8 (Nov. 1997), pp. 1735–1780. ISSN: 1530-888X. DOI: 10.1162/neco.1997.9.8.1735. URL: http://dx.doi.org/10.1162/neco.1997.9.8.1735.

[19] Hugo Touvron et al. *Llama 2: Open Foundation and Fine-Tuned Chat Models*. 2023. DOI: 10.48550/ARXIV.2307.09288. URL: https://arxiv.org/abs/2307.09288.

[20] Edward J. Hu et al. *LoRA: Low-Rank Adaptation of Large Language Models*. 2021. DOI: 10.48550/ARXIV.2106.09685. URL: https://arxiv.org/abs/2106.09685.

[21] Alec Radford et al. *Learning Transferable Visual Models From Natural Language Supervision*. 2021. DOI: 10.48550/ARXIV.2103.00020. URL: https://arxiv.org/abs/2103.00020.

[22] Jean-Baptiste Alayrac et al. *Flamingo: a Visual Language Model for Few-Shot Learning*. 2022. DOI: 10.48550/ARXIV.2204.14198. URL: https://arxiv.org/abs/2204.14198.

[23] Fenglin Liu et al. "A medical multimodal large language model for future pandemics". In: *npj Digital Medicine* 6.1 (Dec. 2023). ISSN: 2398-6352. DOI: 10.1038/s41746-023-00952-2. URL: http://dx.doi.org/10.1038/s41746-023-00952-2.

[24] Tao Tu et al. *Towards Generalist Biomedical AI*. 2023. DOI: 10.48550/ARXIV.2307.14334. URL: https://arxiv.org/abs/2307.14334.

[25] Jason Wei et al. *Finetuned Language Models Are Zero-Shot Learners*. 2021. DOI: 10.48550/ARXIV.2109.01652. URL: https://arxiv.org/abs/2109.01652.

[26] Shengyu Zhang et al. *Instruction Tuning for Large Language Models: A Survey*. 2023. DOI: 10.48550/ARXIV.2308.10792. URL: https://arxiv.org/abs/2308.10792.

[27] Haotian Liu et al. *Visual Instruction Tuning*. 2023. DOI: 10.48550/ARXIV.2304.08485. URL: https://arxiv.org/abs/2304.08485.

[28] Ashish Vaswani et al. *Attention Is All You Need*. 2017. DOI: 10.48550/ARXIV.1706.03762. URL: https://arxiv.org/abs/1706.03762.

[29] Xianghui Sun et al. *A Comparative Study between Full-Parameter and LoRA-based Fine-Tuning on Chinese Instruction Data for Instruction Following Large Language Model*. 2023. DOI: 10.48550/ARXIV.2304.08109. URL: https://arxiv.org/abs/2304.08109.

[30] Luis R. Soenksen et al. "Integrated multimodal artificial intelligence framework for healthcare applications". In: *npj Digital Medicine* 5.1 (Sept. 2022). ISSN: 2398-6352. DOI: 10.1038/s41746-022-00689-4. URL: http://dx.doi.org/10.1038/s41746-022-00689-4.

[31] Aditya Grover and Jure Leskovec. *node2vec: Scalable Feature Learning for Networks*. 2016. DOI: 10.48550/ARXIV.1607.00653. URL: https://arxiv.org/abs/1607.00653.

[32] Shruthi Bannur et al. *Learning to Exploit Temporal Structure for Biomedical Vision–Language Processing*. 2023. DOI: 10.48550/ARXIV.2301.04558. URL: https://arxiv.org/abs/2301.04558.

[33] Jacob Devlin et al. *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. 2018. DOI: 10.48550/ARXIV.1810.04805. URL: https://arxiv.org/abs/1810.04805.

[34] Joseph Paul Cohen et al. *TorchXRayVision: A library of chest X-ray datasets and models*. 2021. DOI: 10.48550/ARXIV.2111.00595. URL: https://arxiv.org/abs/2111.00595.

[35] Gao Huang et al. *Densely Connected Convolutional Networks*. 2016. DOI: 10.48550/ARXIV.1608.06993. URL: https://arxiv.org/abs/1608.06993.

[36] Zeming Chen et al. *MEDITRON-70B: Scaling Medical Pretraining for Large Language Models*. 2023. DOI: 10.48550/ARXIV.2311.16079. URL: https://arxiv.org/abs/2311.16079.

[37] Zeming Chen et al. *MediTron-70B: Scaling Medical Pretraining for Large Language Models*. 2023. URL: https://github.com/epfLLM/meditron.

[38] Alistair E. W. Johnson et al. "MIMIC-IV, a freely accessible electronic health record dataset". In: *Scientific Data* 10.1 (Jan. 2023). ISSN: 2052-4463. DOI: 10.1038/s41597-022-01899-x. URL: http://dx.doi.org/10.1038/s41597-022-01899-x.

[39] Alistair E. W. Johnson et al. *MIMIC-CXR-JPG, a large publicly available database of labeled chest radiographs*. 2019. DOI: 10.48550/ARXIV.1901.07042. URL: https://arxiv.org/abs/1901.07042.

[40] Alistair Johnson et al. *MIMIC-IV-Note: Deidentified free-text clinical notes*. 2023. DOI: 10.13026/7QGP-KC16. URL: https://physionet.org/content/mimic-iv-note/.

[41] Alistair Johnson et al. *MIMIC-CXR Database*. 2020. DOI: 10.13026/S5DG-6S42. URL: https://physionet.org/content/mimic-cxr/.

[42] Albert Q. Jiang et al. *Mistral 7B*. 2023. DOI: 10.48550/ARXIV.2310.06825. URL: https://arxiv.org/abs/2310.06825.

Appendix

B: Example: Original Report

Name: ___ Unit No: ___

Admission Date: ___ Discharge Date: ___

Date of Birth: ___ Sex: M

Service: MEDICINE

Allergies: No Known Allergies / Adverse Drug Reactions

Chief Complaint: Dark urine

Major Surgical or Invasive Procedure: endotracheal intubation ___, ERCP ___

Brief Hospital Course:
Mr. ___ is a ___ homeless gentleman with a history of diastolic heart failure, CAD s/p CABG, polysubstance abuse, DM, COPD, syncope, OSA not on CPAP, and recent hospital admission for infected foot ulcer who was referred to an OSH ED for elevated bilirubin and was transferred here for emergent ERCP due to concern for cholangitis.

Discharge Medications:
1. Acetaminophen 650 mg PO Q8H:PRN pain 2. Atorvastatin 80 mg PO QPM 3. Buprenorphine-Naloxone (8mg-2mg) 1 TAB SL BID 4. Cyanocobalamin 1000 mcg PO DAILY 5. Fluticasone-Salmeterol Diskus (250/50) 1 INH IH BID 6. FoLIC Acid 1 mg PO DAILY 7. Gabapentin 1200 mg PO TID 8. Omeprazole 40 mg PO DAILY 9. Senna 17.2 mg PO QHS:PRN constipation 10. Thiamine 100 mg PO DAILY 11. GlipiZIDE 10 mg PO BID 12. Docusate Sodium 100 mg PO BID 13. Ferrous Sulfate 325 mg PO DAILY 14. Albuterol Inhaler 1 PUFF IH Q6H:PRN SOB 15. Magnesium Oxide 800 mg PO BID 16. Multivitamins 1 TAB PO DAILY 17. Glycerin Supps 1 SUPP PR PRN constipation 18. Collagenase Ointment 1 Appl TP DAILY 19. Aspirin 81 mg PO DAILY 20. Ciprofloxacin HCl 500 mg PO Q12H Duration: 4 Days 21. Lisinopril 2.5 mg PO DAILY take this in the morning 22. Metoprolol Succinate XL 25 mg PO DAILY take this in the evening 23. Potassium Chloride 40 mEq PO DAILY Hold for K > 24. Torsemide 60 mg PO DAILY

Discharge Diagnosis:
1. Cholangitis
2. CHF
3. COPD
4. Diabetes

Discharge Instructions:
You were admitted with cholangitis, or an infection in your bile duct. You were initially in the ICU and then transferred to the medical floor. We put you on oral antibiotics, and you remained stable. Please finish up four more days of antibiotics and I have sent the prescription to the ___ Pharmacy. Please followup with Dr ___ and with surgery regarding removal of your gallbladder. I have adjusted your blood pressure medication. Please restart your torsemide and potassium tomorrow, but just take 60 mg a day for now. I have sent prescriptions for lower doses of lisinopril and metoprolol to your pharmacy.

B: Loss Curves

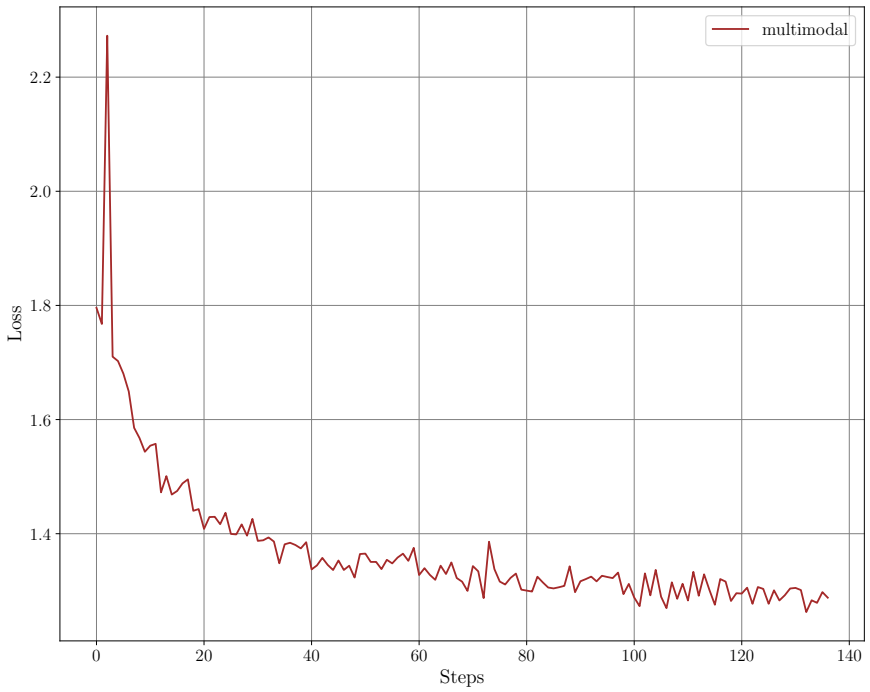

Figure 6: Training loss during the pre-training of the multimodal model.

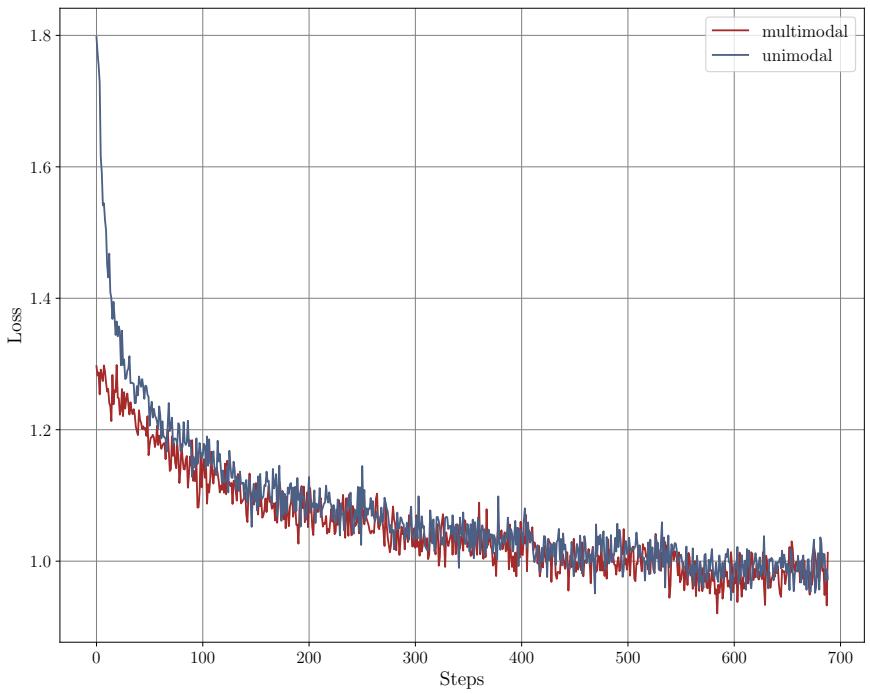

Figure 7: Training loss during the fine-tuning stage of the multi- and unimodal model.

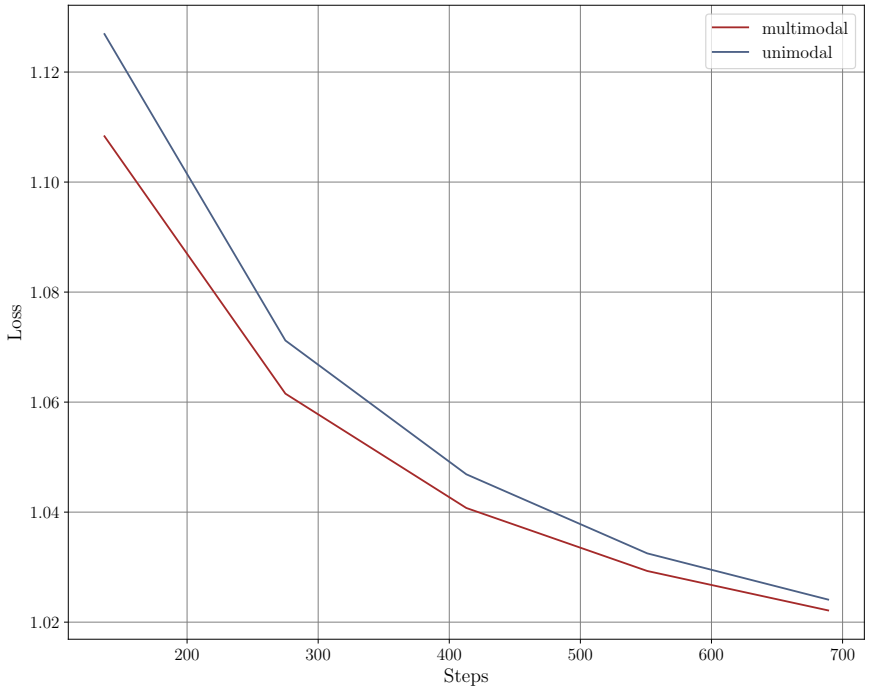

Figure 8: Validations loss during the fine-tuning stage of the multi- and unimodal model.

C: Generated examples

Original Instruction

- Shower daily with mild soap, no baths or swimming until cleared by surgeon
- Do not use lotions, cream, powder, or ointments on incisions
- Weigh yourself in the morning and take temperature in the evening, record on chart
- No driving for approximately one month and no lifting more than 10 pounds for 10 weeks
- Call with questions or concerns
- Contact cardiac surgery office with questions or concerns, answering service available during off hours

Generated Instructions

- Shower daily with mild soap, no baths or swimming until cleared by surgeon
- Do not use lotions or ointments on incisions
- Weigh yourself in the morning and take temperature in the evening
- No driving for approximately one month while taking narcotics, discussion at follow up appointment with surgeon
- Call cardiac surgery office with questions or concerns

Original Instruction

- Discharge instructions for post*surgery care
- Daily showers with mild soap, no baths or swimming until cleared by surgeon
- No lotions, cream, powder, or ointments to incisions
- Morning weight and evening temperature should be recorded on chart provided
- No driving for approximately one month and while taking narcotics
- No lifting more than 10 pounds for 10 weeks
- Females should wear a bra to reduce pulling on incision and avoid rubbing on lower edge
- Cardiac surgery office should be contacted with any questions or concerns

Generated Instructions

- Discharge instructions for patients who have undergone surgery
- Daily showering with mild soap and gentle washing of incisions recommended
- No baths or swimming until cleared by surgeon
- No lotions, cream, powder, ointments, or other products to incisions
- Morning weight and evening temperature should be recorded
- No driving for approximately one month after surgery and while taking narcotics
- Patients can call with questions or concerns during off hours

Original Instruction

- Follow up with all outpatient appointments as listed
- Take discharge paperwork to appointments
- Continue all medications as directed
- Avoid abusing alcohol and any drugs (prescription or illegal)
- Contact outpatient psychiatrist or other providers if concerns

- Call 911 or go to emergency room if feeling unsafe and unable to reach healthcare providers

Generated Instructions

- Discharge Instructions for Mr. [ANONYMIZED]
- Continue medications as directed
- Avoid abusing alcohol and drugs to worsen medical and psychiatric illnesses
- Call or go to emergency room if feeling unsafe
- Best wishes for continued health.

Original Instruction

- The patient has recovered well from lung surgery and surgery on their right leg and is ready for discharge.
- Continue to use the incentive spirometer 10 times an hour while awake.
- Check incisions daily and report any increased redness or drainage. Cover the area with a gauze pad if it is draining.
- If chest tube site starts to drain, cover it with a clean dry dressing and change it as needed to keep site clean and dry.
- Change right leg wound daily with dry sterile dressings and also change it when the gauze is wet. Visit Dr. [ANONYMIZED] next week.
- Pain medication will still be needed once the patient is home, but it can be weaned over a few weeks as discomfort resolves. Make sure to have regular bowel movements while on narcotic pain medications and use a stool softener or gentle laxative to stay regular.
- No driving while taking narcotic pain medication.
- Take 500 mg of Tylenol every 6 hours in between narcotic pain medications.
- Stay hydrated and eat well to heal incisions.

Generated Instructions

- The patient has recovered from lung surgery and is ready for discharge.
- Continue using incentive spirometer 10 times an hour while awake.
- Chest tube dressing may be removed in 48 hours, cover with clean dry dressing and change as needed to keep site clean and dry.
- Pain medication will be needed at home but can be weaned over time.
- Regular bowel movements are important when on narcotic pain medications, use stool softener or gentle laxative.
- No driving while taking narcotic pain medication.
- Take Tylenol [ANONYMIZED] mg every 6 hours in between narcotic pain medication.
- Stay hydrated