

Clinically relevant summarization of multimodal emergency medical data

Stanford University
Stanford CS224N Custom Project

Elsa Bismuth
elsabis@stanford.edu

Lucas Leanza
lleanza@stanford.edu

Jan Michael Krause
jmkrause@stanford.edu

Abstract

Emergency medical care is a high-stakes, data-intensive environment where timely and accurate information synthesis is critical. Our project aims to distill multimodal clinical data into succinct, accurate, and actionable summaries to support tasks such as patient handoffs and medical documentation. Leveraging Stanford Hospital's "MC-BEC" (Chen et al., 2023) dataset, we are optimizing data preprocessing, prompt engineering, prompt chaining, and model parametrization to generate high-quality summaries from large-scale models like GPTs. Leveraging these summaries, we fine-tune smaller models, including variations of BART, to enhance their performance without clinician-provided references. This strategy has not only validated the effectiveness of large-scale models in producing clinically useful summaries but has also led to significant advancements in the smaller models' capabilities in terms of clinical grammatical integrity, brevity, completeness, and accuracy.

1 Key Information to include

Mentor: Bessie Zhang. **External Collaborators:** Dr. David Kim, Assistant Professor of Emergency Medicine. **Contributions Elsa:** Data preprocessing + implemented the whole pipeline for generating reference summaries and evaluating them (human evaluation and quantitative metrics). **Contributions Lucas:** Coded and implemented the finetuning pipeline (pretraining, finetuning/training, model-running, and selection of best model). Engaged in human evaluation and accuracy metrics. **Contributions Michael:** Did model development, incorporated analysis tools, ran, analyzed, improved, tweaked, and evaluated all baseline + fine-tuned models.

2 Introduction

In emergency medicine, clinicians must rapidly synthesize multimodal clinical data to make time-sensitive decisions critical for patient care and medical documentation. This task involves integrating diverse data types—ranging from patient demographics and medical histories to diagnostic results—with precision and traceability. The integration of multimodal clinical data, albeit crucial, is fraught with challenges due to the complexity and depth of clinical records. Simultaneously, advancements in natural language processing (NLP), particularly with the use of large language models (LLMs), are being explored to automate the summarization of clinical records. However, this task is hindered by many challenges including redundant information, excessive irrelevant data, specialized jargon, and numerous dates and times. Cross-referencing data to provide a comprehensive view relevant to the patient's chief complaint is particularly challenging. Thus, the multimodal nature of the data and the requirement for domain expertise highlights the task as an actively acknowledged, non-trivial problem in the field.

3 Related Work

Clinical summarization, historically centered on manual methods, has evolved with the integration of NLP technologies. Early efforts focused on extracting key elements from electronic health records (EHRs) for creating patient summaries. With the advent of NLP, there was a shift towards automatic text summarization, aiming to reduce clinicians' cognitive load and improve the efficiency of information retrieval.

Shing et al.'s paper "Towards Clinical Encounter Summarization" Shing et al. (2021) introduces a method for summarizing clinical data, combining extractive and abstractive techniques. This approach extracts key information from clinical notes and reformulates it into concise summaries, focusing on accuracy and introducing 'faithfulness' and 'hallucination rate' as new evaluation metrics. Their work demonstrates the importance of NLP in healthcare and establishes a standard for handling multimodal emergency medical data.

Further emphasizing the evolution of NLP in healthcare, a recent study published in Nature Medicine, Van Veen et al. (2024) expanded on these methodologies. This study showcases how cutting-edge NLP techniques can be applied to synthesize and interpret complex medical data, and shows that "adapted large language models can outperform medical experts in clinical text summarization".

4 Approach

In this section, we will explain our two-parts approach. The first part aims at generating reference summaries using large language models while the second part focuses on leveraging these summaries for finetuning on smaller language models. Our pipeline (with our 2 fusion methods) can be visualized on figure 4.

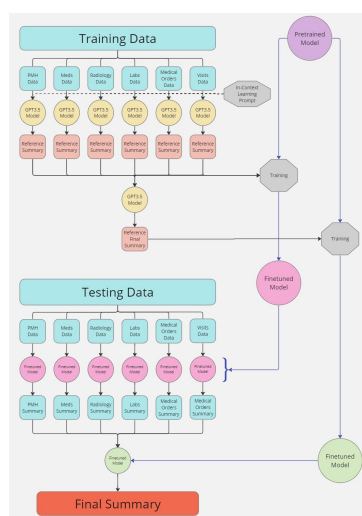


Figure 1: SOS pipeline

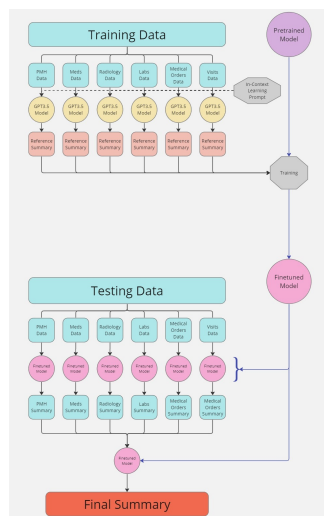


Figure 2: Non SOS pipeline

4.1 Generating reference summaries with large language models

The goal of this section is to generate a concise, accurate, and relevant summary for the 1721 visits in the ICU. We focus on model selection, hyperparameter tuning, prompt engineering, and prompt chaining. To evaluate the performance of the different combinations, we use one reference summary provided by a clinician as a reference.

We started by selecting the optimal large language model (LLM) for summarizing patient past medical histories (PMH), given their data-rich nature. After evaluating several state-of-the-art models, we adjusted key parameters—temperature, top P (nucleus sampling), presence penalty, and frequency penalty—to enhance summary clarity and data fidelity. Following model and parameter optimization, we employed an iterative approach to prompt engineering.

We explored two methodologies for generating precise and relevant summaries. The first, prompt chaining, involved creating individual summaries for each dataset component (visits, PMH, orders, labs, meds, rads) and then synthesizing these into a composite summary through late fusion. The second approach, early fusion, directly processed the raw data into a comprehensive summary. Utilizing the previously selected model and parameters, we refined our prompts with role-specific instructions, dataset descriptions, structured guidance, and in-context learning examples from clinicians. The efficacy of each method was evaluated in collaboration with a clinician, leading to the selection of the optimal summarization approach.

Overall, we conducted a thorough analysis on model selection, hyperparameter tuning, and prompt engineering to produce high-quality summaries. These summaries will then be used as reference for finetuning.

4.2 Finetuning pretrained language models

This section's goal is to use the generated summaries as reference to fine-tune smaller models, optimizing batching and training parameters. These steps involved the creation of many pipelines to handle various summary generation approaches. The pretrained models were imported from HuggingFace Library.

First, we tested pretrained models to examine which offered the most promise. Specifically, we used OpenAI's GPT-2, a transformers model pretrained on a very large corpus of English data in a self-supervised fashion (Radford et al., 2019) because of its successful use in clinical summarization as demonstrated by Van Veen et al. (2024), ElutherAI's GPT-Neo, a transformer model trained on Pile with 1.3B parameters (Black et al., 2021) because of its robust parameter architecture and finally, Facebook's Bart-Base and Bart-large-cnn models, based on a transformer encoder-decoder (seq2seq) model with a bidirectional encoder and an autoregressive decoder (Lewis et al., 2019b) chosen for their efficacy for clinical summaries (Shing et al., 2021). Visualizations for each model architecture can be found in figure 7 in the appendix. The experiments used to select the best-performing pretrained model can be found in the experiments section.

Then, we leveraged two datasets for finetuning: 1) reference summaries generated without specific prompting and 2) reference summaries from the technique described in section 4. Our team developed the training process, including batch tokenization, and did the monitoring via the "weights and biases" library. A "run_model" function was coded to manage the execution of pretrained and finetuned models, incorporating various summarization strategies through our "generate_by_batch" function. This function addresses inputs exceeding model token limits by batching and utilizing HuggingFace's generation capabilities.

We explored two summarization strategies:

- A single-summary method, generating one comprehensive summary from the complete dataset, necessitating significant input batching.
- A two-model Summary of Summaries (SOS) approach, generating summaries for each data subsection and a consolidated summary of these, requiring minimal batching.

5 Experiments

5.1 Data

We are using the "Multimodal Clinical Monitoring in the Emergency Department (MC-MED)" dataset4, which is the first dataset to combine detailed physiologic monitoring with clinical events and outcomes for a large, diverse ED population. It includes around 118,000 continuously monitored Emergency Department visits from 2020-2022, among which 1721 ICU visits which we focus on. Given its vast size, we streamlined the dataset for efficient model processing, accommodating the token limitations and minimizing memory usage. We stripped out medical codes, absolute timestamps, and non-essential information according to clinicians, while preserving the sequence of events and clinically significant details. Each visit, uniquely identified by a 'CSN', is enriched with patient-specific history and current medications, creating a clear, model-ready JSON file encapsulating all critical visit information.

5.2 Evaluation method

Assessing clinical data summarization without benchmarks presents multiple challenges. The lack of reference summaries complicates objective assessment. Additionally, ensuring the precision of clinical information is paramount to avoid adverse outcomes, requiring detailed verification to prevent inaccuracies or fabrications.

Given these considerations, we adopted a human evaluation methodology. Initially, as computer science students, we evaluated summaries for grammatical and factual accuracy, providing a foundational assessment that guided our model-selection process.

The complexity of evaluating relevance necessitated collaboration with a domain expert, clinician Dr. David Kim, which limited our evaluation to a smaller subset of summaries. This evaluation focused on two key aspects: pinpointing clinically relevant information related to the chief complaint and accurate clinical statement. In addition, we also calculated the accuracy of the generated terms by cross-checking with the factual data. This process enabled us to determine the best summary.

We also used BLEU Scores for linguistic precision and BERT Scores for semantic fidelity as mathematical measures to evaluate our summaries against a clinician's reference, aiming to determine if they can reliably identify the best summary without clinician involvement.

5.3 Experimental details

5.3.1 Generating reference summaries

We started by identifying the most suitable large language model (LLM) for summarizing patient medical histories (PMH), employing APIs from together.ai and OpenAI to evaluate various state-of-the-art models including Mistral 7B instruct (Jiang et al., 2023), Mistral 7b instruct v0.2, Alpaca (Taori et al., 2023), Vicuna (Chiang et al., 2023), Llama 2 7b (Touvron et al., 2023), gpt-3.5-turbo and gpt-4 (OpenAI), as highlighted in recent literature. The evaluation criteria included context size and model parameters, with a focus on generating concise summaries within a 512 token limit for comparative analysis presented in table 6.

Subsequently, we fine-tuned the optimal model using parameters such as temperature (1, 0.4, 0.1), top P (1, 0.4, 0.1), and penalties for presence and frequency (0, 0.5, 1). Given the need for clear, accurate, and logically consistent summaries without the necessity for creative flair, we favored settings with lower temperature and higher nucleus sampling, alongside minimal presence and frequency penalties. Our next step involved refining our prompts through iterative testing, incorporating elements like system role designation, varying guideline complexity, in-context learning (ICL) with multiple examples, thresholds for abnormal values, and detailed dataset descriptions. We explored two summarization fusion methods: one aggregating sub-dataset summaries generated with tailored prompts, and another directly summarizing raw data. Each method was assessed for effectiveness with variations in prompt structure.

5.3.2 Finetuning small language models

For fine-tuning small language models, we generated reference summaries with the best LLM with both little prompting and with prompting from section 5.3.1. In our methodology, we initially conducted environment testing and model selection to identify the optimal small model for fine-tuning. Preliminary training on 12 dataset cases with a learning rate of $5e^{-5}$ highlighted Facebook’s bart-large-cnn as the superior model, mainly due to its capability to generate coherent clinical text. Competing models tested included GPT2, bart base, gpt neo, and facebook’s bart-base.

Following selection, we adopted two strategies for model improvement: splitting the task into the summaries of the individual parts of the dataset and then performing summary of the summaries. We also implemented in-context learning.

We experimented with diverse training configurations for both the segmented model and the summary model, adjusting variables such as learning rates ($1e^{-5}$, $5e^{-5}$, $1e^{-4}$), epochs (30, 45, 70), and training set sizes (15, 45, 70 examples). Training durations ranged from 10 minutes for low-epoch-low-example-count models to 3 hours for high-epoch-high-example-count models on an Nvidia T4 GPU run through Colab.

5.4 Results

5.4.1 Reference summaries

In our study, we assessed several models for summarizing patient medical histories (PMH), finding that LLaMA 2 7B and Alpaca were suboptimal, generating either incomplete lists or irrelevant content. Mistral 7B Instruct v0.2 provided coherent summaries, but GPT-4 excelled, producing outputs that were coherent, fluent, and closely matched clinician-written summaries, as indicated in table 7. Consequently, GPT-4 was chosen for optimization. In optimizing GPT-4 for summarization, we evaluated the effects of various parameters on output quality. Lowering the temperature to 0.1 improved summary precision. Variations in the top P value showed minimal influence, thus it was set to 1. Presence and frequency penalties showed minimal impact, with no setting capturing all conditions (anemia, pneumonia, hyponatremia), highlighting consistency challenges. We standardized penalties at zero for uniformity and replicability.

Prompt design experimentation, emphasizing the system’s role as an "expert clinician" and detailed formatting instructions, significantly enhanced summary relevance and accuracy. Various strategies, including high-level guidelines, structured summaries, and in-context learning (ICL) with multiple examples, were tested for generating comprehensive summaries. These methods produced outputs that were grammatically correct, fluent, and coherent, with variations in length, content selection, and precision. Feedback from a clinician favored summaries derived directly from raw data, utilizing four examples and structured guidelines, as documented in table 10. This approach yielded summaries that met clinical standards for accuracy, relevance, and completeness, deemed sufficient for clinical integration despite potential for further refinement.

Clinician’s summary	GPT-4’s summary
An 89-year old woman with a history of atrial fibrillation on digoxin, stroke, and syncope self-presented to the ED with a breathing problem. She was tachypneic (RR 28). She was found on CT to have 5 left sided rib fractures and a clavicle fracture. Her oxygen saturation decreased periodically to 91%, and she was admitted to the ICU for further monitoring and management.	An 89-year-old female with a history of essential hypertension, hypothyroidism, and syncope presented to the ED with a breathing problem. Initial assessment revealed tachycardia (HR 147) and hypoxia (SpO2 96%), along with metabolic abnormalities including low sodium (133) and high alkaline phosphatase (230). Imaging showed bilateral pulmonary edema and pleural effusions, while labs indicated leukocytosis and abnormal electrolytes. She received oxygen, antibiotics, and fluid resuscitation, but due to persistent hypotension and abnormal labs, she was admitted to the ICU for further management.

Table 1: Comparison of clinician’s and GPT-4’s summaries. Green marks information mentioned by both GPT-4 and the clinician, orange, the correct details omitted by the clinician, and red, GPT-4 hallucinations.

When computing the BLEU score and BERT Score, we get 0.329 and 0.867, respectively. We were content to verify that the clinician’s preferred summary aligned with the highest BLEU and BERT scores across all generated summaries in table 10, bolstering our confidence in using these mathematical metrics as indicators for identifying the best reference summary.

5.4.2 Pretrained Model Baselines

We evaluated the baseline for the pretrained models. Some sample summaries can be seen in table 2 in the appendix. None of the models performed well and no model had any relevant patient data or cohesive and grammatically correct sentences.

5.4.3 Finetuning Results

Our baseline comparison revealed the pretrained Bart-Large model produced nonsensical summaries, contrasting with the coherent and accurate summaries from our optimized models (refer to table 2). Through human and expert analysis across different training parameters—sample size, epochs, and learning rate—we observed improved performance with larger data samples and increased epochs. Optimal results were achieved at 70 epochs, prompting further testing up to 140 epochs which significantly enhanced model output. The learning rate of $5e^{-5}$ was identified as most effective.

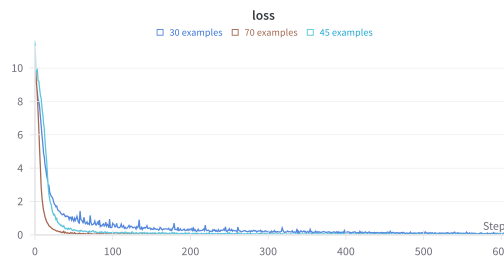


Figure 3: Training Loss Curve across various training example numbers

Parameter	Low	Medium	High
Data Sample Size	</s><s><s>Based on the data provided, the panel has a history of a series of conditions including a pattern of seiness<pad><pad><pad> The<pad> <pad>inventoryQuantity<pad>istg<pad>)...<pad></s></s><s><s> [...]	The patient presented with a history of atelectasis, including malignant melanoma of the scalp and neck. The chest X-ray showed bilateral atous lesions on the right side In the presented clinical encounter data, the patient presented with a history of at least one of the following: [...]	Thank you for providing the summary. Based on the clinical data you provided, the patient has a complex medical history including a history of various medical conditions including multiple systems injury, and a severe medical mystery. [...]
Epochs	</s><s><pad><pad><pad> medical<pad> patient<pad> history<pad> the<pad> clinical<pad>a<pad> including<pad><s>., medical medical, patient, the, history, clinical,<s> [...]	</s><s><s>The patient’s past medical history includes a variety of significant conditions and events, such as enlarged lymph nodes, nervous system disorders, irritable bowel disorders and heart conditions. [...]	Thank you for providing the summary. Based on the data you provided, the patient has a complex medical history including post-hemorrhagic anemia, post. Thank you for providing the summary. [...]
Learning Rate	</s><s><s>The study found the patient had a history of chronic heart failure, chronic pain, and chronic kidney failure. The patient also had problems with the bowels, the stomach and the back</s></s><s><s> [...]	</s><s><s>The patient has a history of chronic pain, including aortic cancer, and a host of other conditions. The patient also has issues with the eye, leg, heart and leg</s></s><s><s>The patient had a history of alcohol abuse and receiving treatment for muscle related spasms. [...]	</s><s><s>Thank you for the<pad><pad>and<pad> The<pad>The<pad> injury, heart disease, and heart injury. The chest injury revealed a heart pain,<pad> ' <pad> [END OF SUMMARY]

Table 2: Sample Summaries based on various parameters

5.4.4 Summary Approach Results: SOS vs single-summary

In table 2 are snippets (of around 80 tokens) of summaries for the same clinical encounter using various summary approaches. The learning rates were $1e^{-5}$, $5e^{-5}$, $1e^{-4}$, the epochs were 30, 45, 70, and the training set sizes were 30, 45, and 70 examples.

Model	Sample Summary
Summary of Summaries, 60 token generations - 45 epochs 70 examples 1e-05 LR	The patient presented with a history of malignant melanoma of the scalp and neck, benign neoplasms of skin, and actinic keratitis. The chest X-rayIn the presented clinical encounter data, the patient has a history of history with multiple abnormalities including low levels of protein, including elevated Troponin I levels. The clinical exam revealed ElevIn the provided clinical encounter data, the patient presented with symptoms warranting activation of a stroke code. [...]
Single Summary, Extensive Prompting - 140 epochs, 70 examples 1e - 05 LR	A 90-year-old male with a history of cerebrovascular accident (CVA), unspecified mechanism, presented via EMS with an acute cerebranal disema. Initial workup revealed abnormal vital signs including tachycardia (HR 66) and hypotension (95A 40-year-old male with a history of unspecified atrial fibrillation and hyperkalemia presented to the ED with altered mental status and hypothermia (31.6°C). Initial workup revealed critical hypoxemia (SpO2 39 mmH [...]

Table 3: Sample Summaries for various summary approaches using our finetuned Bart-Large-CNN model

5.5 Final Model Results

For the final model, which was high epochs, high examples and medium learning rate, 20% of the test set was labeled by Dr. David Kim as "good". For the five samples provided, the following complaints were noted in his expert evaluation:

Issue	Explanation and response example
Vagueness	The summaries often provided vaguer terms than desired, resulting in little overall information being conveyed. "very little interpretable information"
Lack of key term	The samples provided many terms, but omitted the most important ones. "this needs to include the chief complaint to be comprehensible (they all do)"
Confusing term categories	The model confuses diagnoses as symptoms. "the "patient's symptoms" are just more diagnoses, need the complain here"

Table 4: Dr. David Kim's Evaluation of Common Issues

The summaries presented and metrics used can be found in their entirety in the appendix. However, a sample of a summary can be found below for the purpose of illustration, where every highlighted term is a clinical term used in the generated summary, with green representing a factually correct term and red representing a hallucinated term:

Based on the clinical data provided, the patient has a complex medical history including **age-related nuclear cataract**, **bilateral disorders such as glaucoma**, and **lower eye injury**. The clinical encounter **includes bacterial degeneration**, specific eye disorders including **preglaucoma**. Thank you for providing the summary. Based on the data presented, the patient has a complex history of **pneumonitis**; **food/vomitus**, **Shock, unspecified**; and **alcohol abuse**. The radiology report indicated a **right central venous injury** with a distal tip in the **lower abdominal injury**. Thank you for providing the summary. The patient presented with a complex medical history including conditions such as **hypertension**, **diabetes**, and a history of **low blood cell count**. The patient presented with a history of significant medical issues including **obesity**, **diabetes**, and medical conditions such as **hypertension**.

In the example above the number of terms generated was 19 and the percentage of correct terms for was 53.3%. Over the five summaries seen by Dr. David Kim, **the model generated 14.43 clinical terms on average and had a 49.5% rate of correctness** when choosing said terms.

6 Analysis

6.1 Generating reference summaries

Comparing the different language models, we realized that GPT-4, the one that had the biggest context size, had the best results. Indeed, for the final best prompt that we used, we had 5655 tokens (prompt + data for CSN 99469538). Therefore, the models with a smaller context window (of 4,094 tokens for example) would not have performed with these prompt and data. Regarding the hyperparameters and role prompting, we got to the same conclusion as in Van Veen et al. (2024) for using a temperature of 0.1 and specifying a role adapted to the medical setting. Using examples for in-context learning (the more the better) also works well for direct summaries (non SOS). Finally, for the fusion methods, even if we obtained high-quality results for the summary of summaries approach, the cross information between the datasets, time-wise order, and inter-dependencies of the dataset were better understood and summarized. Globally, having a powerful model, a precise and structured prompting, and in-context learning boosts the performance of the system.

6.2 Pretrained Model Analysis

Pretrained models like GPT-2, GPT-Neo, and Bart-Large generally failed as clinical summarizers, often producing incoherent or irrelevant outputs. This is likely due to their limited input windows and the complexity of clinical data. Contrastingly, Facebook's Bart-Base, though smaller and less specialized in medical data, generated coherent, albeit incomplete, summaries. Despite this, we chose to fine-tune Bart-Large for several reasons: its architectural similarity to Bart-Base, higher parameter count suggesting better performance with extensive data, and the potential for improved accuracy through fine-tuning. Thus, Bart-Large was selected as our primary model for further development.

6.3 Finetuning Analysis

In our study of over 100 models, we found that a lower loss doesn't always mean better quality; some higher-loss models outperformed others. Surprisingly, models improved up to 140 epochs without overfitting, questioning the limits of improvement and overfitting. Our top models showed great progress from Bart-Large's initial outputs, achieving factual accuracy. However, a finer learning rate didn't improve results as expected, suggesting it might be trapped in local minima or need more epochs.

Data Sample Size: Larger data samples during finetuning lead to better results. This is because more examples help the model generalize better and identify important terms.

Epochs: More epochs enhance coherence. This is due to the model learning relationships between tokens and determining which are crucial for summaries.

Learning Rate: Its impact is complex. A low learning rate led to more nonsensical results, while a medium rate reduced such errors. A high rate caused early stopping and incoherence. We theorize that a low rate might trap the model in local minima, while a high rate could lead to overfitting, especially with fewer training samples.

6.4 Summarizing Approach Analysis: Potential Error Sources

Batching. The summary of summaries method yielded coherent results, whereas the single-summary approach encountered batching-related issues. Due to token input constraints of generative models and extensive data, the single-summary approach necessitated multiple batch processing. This led to challenges, as the model's finetuning on a fixed format caused forced replication of this format across batches, regardless of data completeness. Consequently, the model generated summaries with inaccuracies and hallucinated content when data within a batch was insufficient to match the expected structure (example: "the patient is a [age] [gender], with [past medical history], and [...]"). Overall, this highlights the relationship between flexibility and specificity: the more prompting and similar finetune examples, the less flexible the models will be and the more they will hallucinate in order to follow the formatting.

Attention. The summary of summaries model may struggle to grasp connections across data types since it generates individual summaries without integrating contextual information from other data segments, potentially overlooking relationships like that between 'asthma' in a patient's history and 'albuterol' in their medication list. Overall, this highlights the need for attention vs efficient processing. While batching might cause severe issues for strictly-finetuned models, doing a single summary still enables the model to keep processing various different parts of the data at once instead of synthesizing one at a time.

In the end, we decided to deem the summary of summaries approach better. While it might lose some of the relationships across data types, it still seems to be the best at providing coherent, full, and rich summaries

compared to the single-summary approach. The reason why is that higher batching in the single-summary model increases the chances of hallucination and creates sentences that are incomprehensible due to the need to follow stricter formatting.

6.5 Final Finetuned Model Analysis

The best finetuned model in Dr. David Kim's opinion was the summaries of summaries model with high examples, high epochs, and medium learning rate. It was the only one which had 20% of the test set being commented as a "good" summary. All other variations had mistakes or comments that made the summary imperfect. Some key aspects to note about the model, alongside potential explanations:

Coherence: The final model performs much better than its pre-trained counterpart when it comes to coherence. The model does not create nonsensical terms and makes grammatical sense in a cohesive manner, responding in full sentences. The explanation here is simple: training the model on coherent summaries increases the coherence of the pre-trained model.

Medium Accuracy: The best model only had around a 50% accuracy rate for the terms it produced. This is highly concerning as hallucination is of great concern in medical settings, where a single error can be responsible for great damages to health or the loss of life. This suggests that these models, at least for now, are not ready to be widely implemented yet within a clinical setting. We have three theories as to why the hallucination rates are this high: 1) models try to follow patterns seen in their training, but batching hides the information the model is looking for and thus hallucinates, 2) the hallucinated terms tend to be "clinically common terms", such as "diabetes" and "hypertension", which leads us to believe that the model has been trained on samples where this is common and 3) the model often "adds" too much to terms that were correct, such as stating that an accurate injury happened in an inaccurate body area, which leads us to believe the model might have a tendency to over-explain given their training, suggesting the need for more examples and further epochs for training.

Vagueness and lack of importance understanding: Dr. David Kim highlighted these issues in his expert evaluation. Vagueness seems to be a recurring theme, and we attribute this vagueness to the fact that many important clinical terms are often rare and specific, and thus the model might not have seen them when training, thus opting for vaguer, more commonly-seen terms. The lack of importance understanding is likely related to the summary of summaries approach, where much context and inter-term relationships are lost in the two-step process. As a result, the model is somewhat good at repeating terms, but not at understanding the relationships between them, meaning that it is unable to know which terms are the most important.

7 Conclusion

Our study advances emergency medical data summarization using advanced NLP and language models like GPT, producing concise, actionable summaries that enhance healthcare documentation and patient care. Key achievements include optimizing data preprocessing, prompt engineering, and model parametrization, significantly improving summary quality for critical healthcare tasks without clinician-provided references. Notably, our approach to fine-tuning smaller models with these summaries, without clinician-provided references, showcased marked improvements in grammatical integrity, brevity, completeness, and accuracy of the generated content.

Our research encountered challenges, notably the complexity of clinical data and the reliance on expert evaluation, which limited our dataset scope and added to the intricacy of our evaluation process due to the absence of standardized clinical summarization benchmarks.

Future directions include broadening evaluation techniques to cover a wider array of clinical data and incorporating mathematical metrics for summary accuracy, such as keyword relevance and conciseness. Exploring additional NLP models and hybrid approaches, as well as investigating multimodal models for enhanced data integration, represents promising avenues for advancing this field.

In conclusion, our project represents a meaningful contribution to the intersection of emergency medicine and artificial intelligence. The advances we have achieved in synthesizing multi-modal medical data can serve as a foundation for future innovations in this rapidly evolving field.

References

- Sid Black, Gao Leo, Phil Wang, Connor Leahy, and Stella Biderman. 2021. GPT-Neo: Large Scale Autoregressive Language Modeling with Mesh-Tensorflow. If you use this software, please cite it using these metadata.
- Emma Chen, Aman Kansal, Julie Chen, Boyang Tom Jin, Julia Rachel Reisler, David A Kim, and Pranav Rajpurkar. 2023. Multimodal clinical benchmark for emergency care (mc-bec): A comprehensive benchmark for evaluating foundation models in emergency medicine.
- Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality.
- Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, L elio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timoth e Lacroix, and William El Sayed. 2023. Mistral 7b.
- Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019a. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension.
- Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2019b. BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *CoRR*, abs/1910.13461.
- OpenAI. Models. Online. Accessed on: March 2, 2024.
- Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training.
- Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.
- Han-Chin Shing, Chaitanya Shivade, Nima Pourdamghani, Feng Nan, Philip Resnik, Douglas Oard, and Parminder Bhatia. 2021. Towards clinical encounter summarization: Learning to compose discharge summaries from prior notes.
- Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Stanford alpaca: An instruction-following llama model. https://github.com/tatsu-lab/stanford_alpaca.
- Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models.
- Dave Van Veen, Cara Van Uden, Louis Blankemeier, Jean-Benoit Delbrouck, Asad Aali, Christian Bluethgen, Anuj Pareek, Malgorzata Polacin, William Collins, Neera Ahuja, Curtis P. Langlotz, Jason Hom, Sergios Gatidis, John Pauly, and Akshay S. Chaudhari. 2024. Adapted large language models can outperform medical experts in clinical text summarization. *Nature Medicine*.

A Appendix

A.1 Code

Github: https://github.com/elsa-bismuth/ed_summarization/tree/master

We are using an undisclosed dataset, so please ask permission to get access to our repo and do not share nor download the data.

A.2 Figures

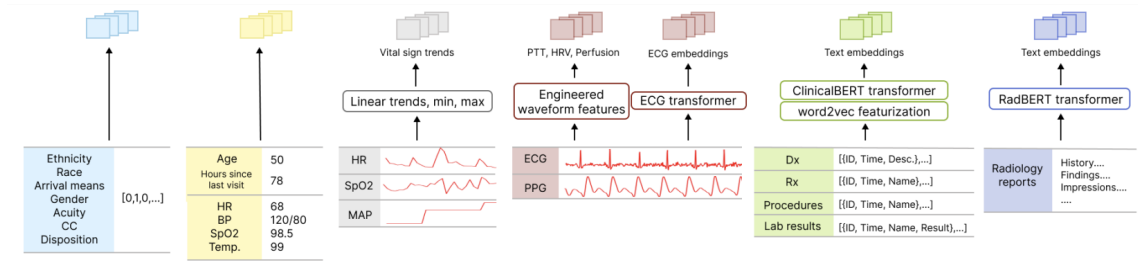


Figure 4: Summary of data modalities represented in MC-BEC (Chen et al., 2023)

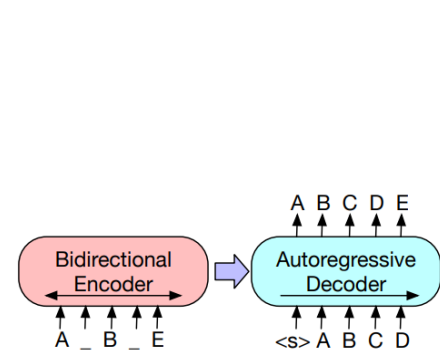


Figure 5: A simplified model architecture for Facebook's BART (Lewis et al., 2019a)

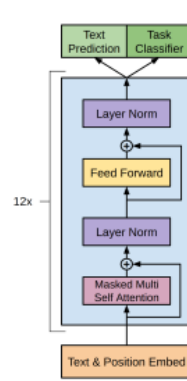


Figure 6: A classic Transformer-based architecture, similar to those used in GPT2 and GPT-Neo (Radford et al., 2018)

Figure 7: BART and Transformer architectures

Clinician's summary: "A 90 year old woman with a history of anemia, pneumonia, and hyponatremia."

Model	Summary for Past Medical History	Analysis
Mistral 7B Instruct v0.2	The patient has a complex medical history with multiple conditions. They have a history of edema, cardiac murmur, chronic venous insufficiency, unsteadiness on feet, Dupuytren's contracture, eye disorders including cataract and retinal issues, breast abnormalities, anemia, blepharitis, female genital disorders, difficulty walking, vitamin D deficiency, hypotension, pneumonia, lymphedema, sepsis, hyponatremia, urinary tract infections, and acute kidney failure. Notable abnormalities include chronic venous insufficiency, eye disorders, breast abnormalities, anemia, difficulty walking, and acute kidney failure.	Fluent, accurate, grammatically correct.
GPT-4	The patient has a history of chronic venous insufficiency, heart valve disorders, and various eye disorders including cataracts, drusen of the macula, and hypermetropia. They also present with nervous system disorders characterized by unsteadiness and difficulty in walking, as well as a history of anemia and vitamin D deficiency. The patient has recently suffered from severe sepsis, urinary tract infections, and acute kidney failure, with a record of pneumonia and hypo-osmolality and hyponatremia.	More fluent and human-like. Only selects the most important and relevant info → the best!
Alpaca	The past medical history of the patient includes edema, unspecified; cardiac murmur, unspecified; venous insufficiency (chronic) (peripheral); unsteadiness on feet; presbyopia; hypermetropia, unspecified; arcus senilis, unspecified; mammographic calcifications found on diagnostic imaging of breast; nonmalignant breast conditions; arcus senilis, unspecified; difficulty in walking, not elsewhere classified; other nervous system disorders; vitamin D deficiency, unspecified; venous insufficiency (chronic) (peripheral); weakness; encntr for general adult medical exam w/o abnormal findings; hypotension, unspecified; pneumonia, unspecified organism; lymphedema, not elsewhere classified; difficulty in walking, not elsewhere classified; severe sepsis without septic shock; urinary tract infection, site not specified; acute kidney failure, unspecified; and urinary tract infection, site not specified.	Makes a list separated with ; + not relevant to say "unspecified"
LLaMA 2 7b	Based on the past medical history provided, the patient has a history of various medical conditions, including: <ol style="list-style-type: none"> 1. Edema and residual codes, as well as unclassified conditions. 2. Cardiac murmur and heart valve disorders. 3. Venous insufficiency and other diseases of veins and lymphatics. 4. Unsteadiness on feet and other nervous system disorders. <p><i>(Summary cut for space concerns)</i></p>	Bad: makes a list instead of a 3-sentence summary and gets cut after max tokens number is reached

Table 7: Comparison of summaries on past medical history using different LLMs

Clinician's summary: "A 90 year old woman with a history of anemia, pneumonia, and hyponatremia."

Parameters	Summary for Past Medical History	Analysis
Temp = 1 Top P = 1 Presence penalty = 0 Frequency penalty = 0	The patient has a complex medical history, with notable chronic venous insufficiency, heart valve disorders, and recurrent difficulties with walking likely related to unspecified nervous system disorders. The patient has experienced severe sepsis and urinary tract infections, along with acute kidney failure and hyponatremia, suggesting a history of serious systemic infections and kidney disease. The patient also has numerous eye disorders, including unspecified cataract, drusen of macula, and hypermetropia, indicating significant vision issues.	Mentions hyponatremia.
Temp = 0.4 Top P = 1 Presence penalty = 0 Frequency penalty = 0	The patient has a history of chronic venous insufficiency and lymphedema, unspecified cardiac murmur and hypotension, and multiple nervous system disorders causing unsteadiness on feet and difficulty in walking. They also suffer from significant eye problems including presbyopia, hypermetropia, cataract, and other unspecified eye disorders. Additionally, they have a history of severe sepsis, urinary tract infections, acute kidney failure, and anemia, along with a vitamin D deficiency.	Mentions anemia.
Temp = 0.1 Top P = 1 Presence penalty = 0 Frequency penalty = 0	The patient has a history of chronic venous insufficiency, heart valve disorders, and various eye disorders including presbyopia, hypermetropia, cataract, and retinal issues. They have also experienced nervous system disorders leading to unsteadiness and difficulty in walking, as well as nutritional deficiencies like Vitamin D deficiency and anemia. Recently, the patient has been dealing with severe sepsis without septic shock, hypo-osmolality and hyponatremia, urinary tract infections, and acute kidney failure.	Mentions hyponatremia and anemia. The lowest temperature captures 2/3 disorders, outperforming higher temperatures with only 1/3.
Temp = 0.1, Top P = 0.1 Presence penalty = 0 Frequency penalty = 0	The patient has a history of chronic venous insufficiency, heart valve disorders, and other nervous system disorders causing unsteadiness and difficulty in walking. They also have a range of eye disorders including cataracts, presbyopia, hypermetropia, and retinal issues, along with a history of anemia and Vitamin D deficiency. Recently, the patient has been dealing with severe sepsis without septic shock, hypo-osmolality and hyponatremia, urinary tract infections, and acute kidney failure.	Mentions hyponatremia and anemia.
Temp = 0.1, Top P = 0.1 Presence penalty = 0.5 Frequency penalty = 0	The patient has a complex medical history with multiple chronic conditions including venous insufficiency, heart valve disorders, and various eye disorders such as cataracts, presbyopia, and retinal issues. They also have a history of nervous system disorders causing unsteadiness and difficulty walking, as well as nutritional deficiencies like Vitamin D deficiency and anemia. Recently, the patient has been dealing with severe infections including pneumonia, sepsis, and urinary tract infections, which have led to acute kidney failure and electrolyte imbalances.	Mentions anemia and pneumonia.
Temp = 0.1, Top P = 0.1 Presence penalty = 1 Frequency penalty = 0	The patient has a history of chronic venous insufficiency, unspecified cardiac murmur, and heart valve disorders. They have experienced unsteadiness on feet and difficulty in walking due to nervous system disorders, and also suffer from various eye conditions including presbyopia, hypermetropia, cataract, and other unspecified eye disorders. The patient has also had severe sepsis without septic shock, urinary tract infections, acute kidney failure, and hypo-osmolality and hyponatremia indicating fluid and electrolyte disorders.	Mentions hyponatremia.
Temp = 0.1, Top P = 0.1 Presence penalty = 1 Frequency penalty = 0.5	The patient has a history of chronic venous insufficiency, heart valve disorders, and other nervous system disorders causing unsteadiness on feet and difficulty in walking. They have multiple eye conditions including unspecified cataract, presbyopia, hypermetropia, and degenerative drusen of the macula. The patient also has a history of severe sepsis without septic shock, urinary tract infections, acute kidney failure, anemia, vitamin D deficiency and noninflammatory disorder of the uterus.	Mentions anemia.

Table 8: Comparison of hyperparameter tuning on past medical history using GPT-4.

Model	Sample Summary
Summary of Summaries, 60 token generations - 30 epochs 70 examples 1e-05 LR	"The patient presented with a history of malignant melanoma of the scalp and neck, benign neoplasms of skin, and actinic keratitis. The chest X-rayIn the presented clinical encounter data, the patient has a history of history with multiple abnormalities including low levels of protein, including elevated Troponin I levels. The clinical exam revealed ElevIn the provided clinical encounter data, the patient presented with symptoms warranting activation of a stroke code. Some of the notable abnormalities include: 2. Elevated mean corpus levels (Thank you for providing the summary. Based on the clinical data presented, the patient underwent a series of diagnostic tests and procedures including ECGs, MR brain imaging, CT scans of the "
Summary of Summaries, 60 token generations - 45 epochs 70 examples 1e-05 LR	The patient presented with a history of malignant melanoma of the scalp and neck, benign neoplasms of skin, and actinic keratitis. The chest X-rayIn the presented clinical encounter data, the patient has a history of history with multiple abnormalities including low levels of protein, including elevated Troponin I levels. The clinical exam revealed ElevIn the provided clinical encounter data, the patient presented with symptoms warranting activation of a stroke code. Some of the notable abnormalities include: 2. Elevated mean corpus levels (Thank you for providing the summary. Based on the clinical data presented, the patient underwent a series of diagnostic tests and procedures including ECGs, MR brain imaging, CT scans of the
Single Summary, Extensive Prompting - 140 epochs, 70 examples 1e - 05 LR	A 90-year-old male with a history of cerebrovascular accident (CVA), unspecified mechanism, presented via EMS with an acute cerebribral disema. Initial workup revealed abnormal vital signs including tachycardia (HR 66) and hypotension (95A 40-year-old male with a history of unspecified atrial fibrillation and hyperkalemia presented to the ED with altered mental status and hypothermia (31.6°C). Initial workup revealed critical hypoxemia (SpO2 39 mmHA 40-year-old male with a history of primary malignancy, hyponatremia, and hemangioma presented to the ED with altered mental status and hypothermia (31°C). Initial workup revealed a critical hypochloremia, elevatedA 28-year-old male with a history of Troponin I deficiency, normal CO2 (28), and normal INR (1.0) presented to the ED with findings of a meningioma. Initial workup revealed abnormal vital signs including tachycardiaA 70-year-old male with a history of cirrhosis, hyponatremia, and metabolic acidosis presented for a comprehensive metabolic panel. Initial workup revealed abnormal vital signs including hypothermia (37°C), elevated creatinine (0.94A normal low pH (7.7), normal lactic acidosis (13.6), and normal mean corpusCular Hemoglobin (31.4) and NEUTROPHIL (57.1) were abnormal. Lactic Acidosis was abnormal withA normal low pH (12.8), normal methylation (31.9), and normal chloride (83.2) on arterial blood gases revealed abnormal metabolic panel results. Labs showed abnormal low lactate (1.69), hyperkalemia (K 39.5A normal chest X-ray showed no acute cardiopulmonary findings. Labs showed normal glucose (101 mg/dL) and INR (1.1 mmol/L). Labs also showed abnormal low sodium (134 mmol) And high creatinine (1.83).A 39-year-old male with a history of hyperkalemia, elevated creatinine, and lactic acidosis presented to the ED with abnormal vital signs. Initial workup revealed abnormal low pH (9.9), high Hgb (5.7), andA normal chest X-ray showed abnormal vital signs including hypotension (32.0), hemodynamic volume (97.2), and metabolic acidosis (0.16). Labs showed metabolic acids with low sodium (134 mmol/L) and high chloride (113 mmol/LA normal low pH (7.8), normal white blood cell count (5.1), and abnormal low red blood cells (4.31). Hemoglobin was abnormal at 0.00, with hyperkalemia (K 5.2) and elevated lactate (2A 31.4-year-old male with a history of normal metabolic acidosis, HCO3 (31.7), PO2 (35), lactate (0.70), and high PH (62), presented to the ED with abnormal metabolic panel results. Initial workA 30-year-old male with a history of lactic acidosis, hypotension, and chronic kidney disease presented to the ED with altered mental status and hypothermia (31°C). Initial workup revealed critical hypochloremia, hyperkalemia (5A 40-year-old female with a history of hemophilia, hyperglycemia, and metabolic acidosis presented with altered mental status and hypothermia. Initial workup revealed critical hypochloremia (17 mg/dL), hyperhydration (393 mg).
Single summary, extensive prompting - 140 epochs, 70 examples, 1e - 05 LR	A clinician is tasked with creating concise summaries from raw patient data. The data is formatted in specific categories: past medical history, orders, labs, and medications. Each lineA 3-sentence summary of a patient's history and vital signs should highlight key information crucial for understanding the patient at the Emergency Department. The summary should be clinically relevant, highlightingA patient presented to the ED with an acute complaint of altered mental status, hypothermia (31.6°C), and elevated creatinine (1.91 mg/A 90-year-old woman with a history of hyponatremia, pneumonia, and anemia presented via EMS with altered mental status and hypothermia (31.A 63-year-old male with a history of hyperkalemia, heart failure, COPD, and hyperKALemia presented via EMS with general weakness and abnormal labA 90-year-old female with a history of lactic acidosis, heart failure on furosemide, and shortness of breath presented via EMS with tachycardA 95-year-old male with a history of unspecified atrial fibrillation and melanoma presented to the ED with altered mental status and hypothermia (36.1A 36-year-old male with a history of cirranial atherosclerotic disease, hyponatremia, and ventricular infarct presented to the EDA 40-year-old male with a history of acute intracranial hemorrhage, intracranial atherosclerosis, and no dissection presented on chest X-A normal troponin, labilirubin and acute kidney injury were noted on initial workup. Labs showed abnormal hematocrit (43.2), hemoglobinA normal chest X-ray showed low levels of red blood cell count and hyperthyroidism. Labs showed abnormal creatinine (0.94 mg/dL) and lowA normal heart rate, blood pressure and oxygen saturation were abnormal. Labs showed abnormal lactic acidosis (0.16) and hyperkalemia (K 12.8).WAM : 0.2 ; CBC with Differential (CBCD), BASOPHIL % (AUTO DIFF) ;0.6 ; NRBC, PERCENTA 27.6-year-old male with a history of cirrhoticosis, hyperkalemia, and elevated creatinine presented to the ED with altered mental statusA normal heart rate, elevated creatinine (1.91), and normal lactic acidosis (LH 60). Labs showed abnormal high listeria (43.2A normal lute, neutrophil, and hemoglobin level were abnormal, along with low low calcium (4.31), high low sodium (134), and elevated high chlorideA normal chest X-ray showed abnormal kidney function with hyperthyroidism (2.92), metabolic acidosis (AAS), and elevated lactate (0.70).A 33-year-old male with a history of venous Blood Gases And Lactate presented with abnormal results. Initial workup revealed abnormal vital signs including tachycardA 30-year-old male with a history of stromalacia, labetalol, and lactic acidosis presented to the ED with altered mental status and hypA, 'HEPATIC FUNCTION PANEL A', 'PROTHROMBIN TIME', 'MRSA SCREEN BY CULTURE', ECG 12
Single summary, less data, extensive prompting - 140 epochs, 70 examples, 1e - 05 LR	A 47-year-old male with a history of unspecified mood disorder, bradycardia, and altered mental status presented via EMS with altered mentally status and hypothermia (92°F). Initial workup revealed abnormal vital signs including bradypnea (36.8A 40-year-old male with a history of acute respiratory failure, hypoxia, and post-traumatic stress disorder presented to the ED with altered mental status and hypothermia (31°C). Initial workup revealed hypoxic hypokalemia (SpO2No acute intracranial hemorrhage, infarction, or aneurysm in the intrACrania and extracranials. No dissection in visualized neck vessels and aorta. Multiple enlarged lymph nodes in right lower neck measured up to 5 mmA 40-year-old male with a history of no flow-limiting stenosis, occlusion, or aneurysm in the intracranial and extracranials presented to the ED with altered mental status and hypothermia (31°C).A patient with abnormal POCT Arterial Blood Gases and lactate presented to the ED with altered mental status. Initial workup revealed abnormal vital signs including hypotension (10/12) and hypoxemia (SpO2 39), along with metabolic abnormalities such as low

Table 9: Sample Summaries for various summary approaches using our finetuned Bart-Large-CNN model

Here are 6 different parts of prompts:

Prompt 1:

You must write a complete, relevant, correct, and accurate 3-sentence summary. Specifically, you should focus on the abnormal data that might be relevant when a patient arrives to the Emergency Department so that the clinician who reads the summary quickly and correctly understands what is wrong with the patient.

Prompt 2:

Here is a list of the pertinent points to remember when writing the summary:

- race/ethnicity is almost never relevant
- past medical history features are often relevant to the chief complaint
- chief complaint should always be mentioned
- the abnormal vital signs relevant enough to be mentioned are when : SpO₂<90, HR>100, RR>20, MAP<65
- stating the precautions is not relevant
- you should conclude with disposition (ICU admission) after discussing what happened in ED.

In general the sequence should be:

1. Patient info + prior diagnoses and meds pertinent to the chief complaint
2. Chief complaint
3. Initial findings (abnormal vital signs, labs, imaging)
4. Interventions + response to interventions
5. Disposition

Here is the data to summarize :

Prompt 3:

Here are 4 different examples of how you should structure the merged summary and what information is relevant to mention. These examples correspond to 4 different visits:

- Example 1: A 90 year old woman with a history of anemia, pneumonia, and hyponatremia presented via EMS with altered mental status and hypothermia (31.6 C). Initial workup was notable for hyponatremia (Na 126 mmol/L), unremarkable CT head, and possible pneumonia on chest X ray. She developed hypotension (52/43) and received 2L lactated ringer's, dextrose, and broad spectrum antibiotics (cefepime and vancomycin). Despite these interventions, and gradual improvement in temperature, she remained intermittently hypotensive and was started on vasopressors (norepinephrine) then admitted to the ICU.
- Example 2: A 63 year old woman with a history of hepatitis C cirrhosis, heart failure, COPD and hyperkalemia on bumetanide and spironolactone presented via EMS with general weakness and abnormal lab results. Initial workup was notable for hyperkalemia (6.6) and elevated creatinine (1.91), and chest X ray concerning for pneumonia. She received IV fluids, broad spectrum antibiotics (vancomycin, piperacillin-tazobactam), methylprednisolone, and treatment for hyperkalemia (insulin, furosemide, sodium bicarbonate). She remained intermittently hypotensive, was started on norepinephrine, and admitted to the ICU.
- Example 3: A 63 year old man with a history of rhabdomyolysis and renal failure presented with leg pain. Workup was notable for critical hyperkalemia (8.2) and worsening kidney function (creatinine 14.5). He received treatment for hyperkalemia with furosemide, albuterol, sodium bicarbonate, calcium, and insulin, and was admitted to the ICU.
- Example 4: A 59 year old woman with a history of kidney transplant, COPD on albuterol-ipratropium, atrial fibrillation on apixaban and metoprolol, and heart failure on furosemide presented via EMS with shortness of breath. She was tachycardic (HR 148) and tachypneic (RR 20) on arrival. Labs were notable for severe lactic acidosis (6.36). She received adenosine and metoprolol for but remained tachycardic, with no significant electrolyte abnormality and no evidence of pulmonary embolus on CT scan. She was diagnosed with cardiogenic shock and admitted to the ICU.

Now, proceed with this task using the actual data provided to you, keeping in mind the need for accuracy and clinical relevance.

Prompt 4:

You are a helpful expert clinician tasked with creating concise and informative summaries from raw patient

data. Each entry of data represents different aspects of patient information collected during their visit to the Emergency Department. The data is formatted in specific categories:

- 'pmh' (Past Medical History): Each line details a medical condition with its ICD code and CCS category.
- 'orders': Lists medical orders, each line representing one order with its name.
- 'labs': Provides lab results, detailing abnormal findings, lab component names, results, and any abnormalities.
- 'visits': Describes each patient visit, including numerous details such as visit number, total visits, patient demographics, vital signs at triage, chief complaints, disposition from ED, and other relevant metrics.
- 'meds': Enumerates medications prescribed, with each line naming one medication.
- 'rads': Summarizes radiology findings, each line comprising one result with the study name and impression.

Prompt 5:

Here is an example of the structure that the summary should follow: A [age, sex] with history of [diagnoses plausibly relevant to the chief complaint] on [medications plausibly relevant to the complaint] presented to the ED with [complaint]. Initial workup showed [abnormal vital signs, lab and imaging results]. The patient received [selected interventions or types of interventions] with [any notable response to interventions, e.g. hypotension or tachycardia improving with IV fluids], and was [disposition].

Prompt 6:

Here is a list of the pertinent points to remember when writing the summary:

- race/ethnicity is almost never relevant
- past medical history features are often relevant to the chief complaint
- chief complaint should always be mentioned
- the abnormal vital signs are : SpO₂<93, HR>100, RR>20, MAP<70. Do not mention abnormalities if the vital signs are not abnormal.
- give parenthetical values for abnormal vital signs and results
- stating the precautions is not relevant
- no need to explain what tests are (e.g. "tests to assess specific pathogens") - just mention the relevant result
- no additional summary at end, the whole thing is a summary
- format electrolytes and chemical elements with symbols and formulas instead of words
- you should conclude with disposition (ICU admission) after discussing what happened in ED specificity is needed for statements about critical or dispositive findings

Clinician's summary for CSN 99469538: "An 89-year old woman with a history of atrial fibrillation on digoxin, stroke, and syncope self-presented to the ED with a breathing problem. She was tachypneic (RR 28). She was found on CT to have 5 left sided rib fractures and a clavicle fracture. Her oxygen saturation decreased periodically to 91%, and she was admitted to the ICU for further monitoring and management."

Prompt	Summary	Clinician's notes	BLEU Score	BERT Score
Prompt 1 (SOS)	An 89-year-old female patient presented to the Emergency Department with a breathing problem and was triaged as Emergency Severity Index 3-Urgent. She was admitted to the ICU for management of a closed fracture of multiple ribs of the left side, with a history of cardiac dysrhythmias, essential hypertension, and hypothyroidism. Imaging revealed traumatic chest injuries, and lab results showed abnormalities in liver and kidney function markers, as well as signs of infection or inflammation, prompting further evaluation and management in the Emergency Department. The patient's comprehensive medical profile includes a history of multiple fractures, cardiovascular and respiratory issues, and a combination of cardiac, thyroid, and other medications. She presents with interstitial and alveolar pulmonary edema, bilateral pleural effusions, and a left clavicle fracture, suggesting acute respiratory distress and subacute traumatic injuries. Isolation precautions have been implemented due to potential infectious respiratory conditions, with diagnostic tests ordered to assess specific pathogens, along with a treatment plan involving oxygen therapy, pain management, cardiovascular support, and close monitoring for respiratory and circulatory issues. In summary, the patient's presentation in the Emergency Department involves a complex medical history with multiple comorbidities, traumatic chest injuries, abnormal lab results indicating potential organ dysfunction and infection, and the need for comprehensive evaluation and management to address her respiratory distress, fractures, and underlying medical conditions.	Lots of redundancy and same order issues as previous example; no need to explain what tests are (e.g. "tests to assess specific pathogens") - just the relevant result; no additional summary at end, the whole thing is a summary	0.151	0.845
Prompt 1 + Prompt 3 (SOS, no ICL)	An 89-year-old female with a history of multiple fractures, essential hypertension, hypothyroidism, and cardiac dysrhythmias presented to the Emergency Department with a closed fracture of multiple ribs and syncope. She also has interstitial and alveolar pulmonary edema with bilateral pleural effusions, acute rib and clavicle fractures, and abnormal lab results including high alkaline phosphatase and low sodium levels. The patient is on isolation precautions for a potential infectious respiratory condition and is undergoing further evaluation with a respiratory pathogen panel, chest imaging, and cardiovascular support while being closely monitored for respiratory and circulatory issues.	Fractures are presumably diagnosis, not complaint + redundant mentions of fractures	0.221	0.855
Prompt 4 + Prompt 1 + Prompt 3 (Non SOS, ICL)	A 84-year-old female with a history of essential hypertension, hypothyroidism, and multiple fractures presented with bradycardia, hypotension, and respiratory distress. Initial evaluation revealed bilateral rib fractures, clavicle fracture, and pulmonary contusions on imaging, along with metabolic abnormalities including low sodium and high alkaline phosphatase. She was placed on airborne isolation, received oxygen therapy, and was started on medications including levothyroxine, amlodipine, and albuterol sulfate. Consults were made to orthopedics and pain management, and she was admitted for further management and monitoring.	Pretty good. Wrong age.	0.296	0.855
Prompt 1 + Prompt 2 (SOS, ICL)	An 89-year-old female with a history of multiple rib fractures presented to the Emergency Department with a chief complaint of breathing problems. On arrival, she exhibited a heart rate of 147 bpm, respiratory rate of 28/min, and SpO2 of 96%, suggestive of potential respiratory distress. The patient was promptly admitted to the ICU for further management due to acute respiratory failure with hypoxia, pleural effusion, and a nonspecific skin rash, along with abnormal vital signs and imaging findings indicating the need for close monitoring and intervention. The patient's complex medical history, including bradycardia, hypertension, hypothyroidism, cerebral infarction, atrial fibrillation, TIA, and multiple fractures, underscores the importance of comprehensive care and ICU admission for ongoing treatment and evaluation.	SpO2 96% not unusual, no need to specify need for monitoring etc., order issues as above	0.239	0.859
Prompt 1 + Prompt 2 + Prompt 3 (SOS, non ICL)	An 89-year-old female with a history of multiple fractures, atrial fibrillation, hypothyroidism, and cerebral infarction presented to the Emergency Department with shortness of breath and chest pain. Initial findings showed a heart rate of 147 bpm and SpO2 of 96%, indicating potential respiratory distress. Further evaluation revealed acute respiratory failure with hypoxia, pleural effusion, and a nonspecific skin rash, prompting ICU admission for close monitoring and management. The patient's complex medical history, including recent syncope, essential hypertension, and medication list, underscores the need for comprehensive care and ongoing assessment in the ICU setting.	This is good except history should only be mentioned once and early, and again no need for "underscores the need..."	0.204	0.865
Prompt 4 + Prompt 2 + Prompt 3 (Non SOS, ICL)	A 84-year-old female with a history of essential hypertension, hypothyroidism, and multiple fractures presented with bradycardia, hypotension, and respiratory distress. Initial evaluation revealed bilateral rib fractures, clavicle fracture, and pulmonary contusions on imaging, along with metabolic abnormalities including low sodium and high alkaline phosphatase. She was placed on airborne isolation, received oxygen therapy, and was started on medications including levothyroxine, amlodipine, and albuterol sulfate. Consults were made to orthopedics and pain management, and she was admitted for further management and monitoring.	Shouldn't include CSN - otherwise good. Wrong age.	0.296	0.855
Prompt 4 + Prompt 6 + Prompt 5 + Prompt 3 (Non SOS, ICL)	An 89-year-old female with a history of essential hypertension, hypothyroidism, and syncope presented to the ED with a breathing problem. Initial assessment revealed tachycardia (HR 147) and hypoxia (SpO2 96%), along with metabolic abnormalities including low sodium (133) and high alkaline phosphatase (230). Imaging showed bilateral pulmonary edema and pleural effusions, while labs indicated leukocytosis and abnormal electrolytes. She received oxygen, antibiotics, and fluid resuscitation, but due to persistent hypotension and abnormal labs, she was admitted to the ICU for further management.	Overlooks abnormal vital sign thresholds. Format electrolytes as "Na 133, Cl 95" instead of "low sodium 133...". "persistent hypotension" : what was the BP? Would be almost perfect if this explained the management better (what kind of oxygen therapy) → Selected as the best summary by the clinician!	0.329 → Best BLEU Score!	0.867 → Best BERT Score!

Table 10: Comparison of prompts.