Advancing Clinical NLP with RoBERTa & BART

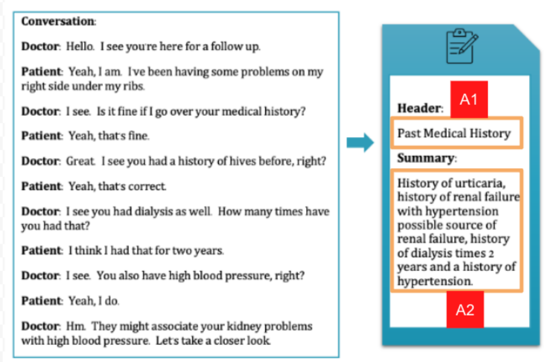
From Classification to Summarization: Unveiling New Insights

Gaurav Narasimhan
University of California, Berkeley

gaurav.narasimhan@berkeley.edu December 11, 2023

1 Abstract

In this retrospective study aligned with MediQA-Chat-2023 (from the ACL Clinical-NLP conference), I present a unique approach to medical dialogue analysis focused on classification and summarization while prioritizing data privacy and computational efficiency. This project, conducted outside the conference, utilizes an ensemble model for classification (Task A1) and a fine-tuned BART model with SAMSum dataset for summarization (Task A2), demonstrating the effectiveness of combining unprompted training with prompted inference. Despite not using public LLMs like GPT or Cohere and not having augmented data from Task C, the methods achieved competitive results, with top-tier scores for Task A1 and top 10 for Task A2, highlighting the robustness and practicality of the approach. This work, undertaken with the aim of contributing to ongoing discussions in the field underscores the feasibility of using LLMs responsibly in sensitive domains and sets a groundwork for future research covering a more comprehensive analysis.



Conversation:

Doctor: Hello. I see you're here for a follow up.

Patient: Yeah, I am. I've been having some problems on my right side under my ribs.

Doctor: I see. Is it fine if I go over your medical history?

Patient: Yeah, that's fine.

Doctor: Great. I see you had a history of hives before, right?

Patient: Yeah, that's correct.

Doctor: I see you had dialysis as well. How many times have you had that?

Patient: I think I had that for two years.

Doctor: I see. You also have high blood pressure, right?

Patient: Yeah, I do.

Doctor: Hm. They might associate your kidney problems with high blood pressure. Let's take a closer look.

Header:

Past Medical History

Summary:

History of urticaria, history of renal failure with hypertension possible source of renal failure, history of dialysis times 2 years and a history of hypertension.

Figure 1: Task A: Header and Summary

2 Introduction

Recent advancements in clinical natural language processing (NLP) have spotlighted the importance of summarizing doctor-patient conversations effectively. This complex task involves condensing doctor-patient conversations into concise, informative summaries, aiding medical practitioners in assimilating key points from past interactions ([4]; [23]). The unique challenges of this task arise from the necessity to comprehend complex dialogues often constrained by limited data availability and the sensitive nature of medical information ([19]).

This task is critical for reducing doctors' workload and improving patient interactions. The MEDIQA-Chat 2023 initiative has played a significant role in advancing research in this area, focusing on automatic summarization and generation of doctor-patient conversations for data augmentation [4].

Large Language Models (LLMs) like GPT-4 have shown promise in medical dialogue summarization, but their application in clinical settings demands careful adaptation and scrutiny. The challenge lies in the unstructured nature of medical conversations and the need to accurately identify key information across multiple symptom sets [17]. Various teams have employed different strategies for this task, including novel N-pass strategies, data augmentation techniques, and fine-tuning on existing language models like T5, BART, and BioGPT [23]; [3].

One significant concern with LLMs is their tendency to produce factual inaccuracies or 'hallucinations' in outputs, which can be problematic in clinical applications. Research teams have worked on addressing these issues by employing ensemble-based methods, rule-based systems, and exploring different text generation strategies based on the note's length [19]; [18]. Additionally, leveraging classical machine learning methods like Support Vector Machine in combination with GPT-3 prompts has shown robust performance in classifying and summarizing medical dialogues [25].

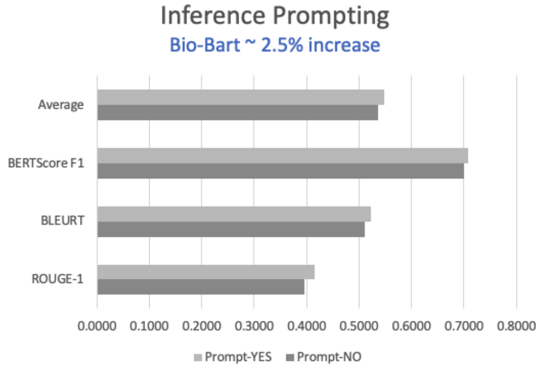


Figure 2: Key Project Insight: Inference Prompting

In this paper, I present a novel approach to the classification and summarization of medical dialogues, retrospectively aligned with the challenges of the MediQA2023-Chat workshop ([4]). This research project, conducted with a strong emphasis on data privacy, deliberately avoids using public LLMs like GPT and Cohere due to the sensitive nature of medical data. For Task A1 (classification), I employed an ensemble model, while Task A2 (summarization) involved fine-tuning a BART model with the SAMSum dataset ([9]). My approach innovatively combines unprompted training with prompted inference, a technique commonly seen in GPT models, leading to a significant improvement in Task A2's average scores ([17]).

The experiments were conducted with a focus on developing cost-effective and production-ready services using prior-generation NVIDIA A10G Tensor Core GPUs. Despite the constraints of not participating in the conference and lacking augmented data from Task C, my methods yielded competitive results, highlighting the robustness of the approach under resource limitations ([21]).

This research aims to contribute retrospectively to the ongoing discussion in the field, with the intention to refine these solutions with augmented data from Task C and present a comprehensive analysis covering Tasks B and C in future work.

3 About MEDIQA-Chat 2023

The MEDIQA-Chat 2023 initiative, as detailed by Ben Abacha et al. (2023) [4], comprises a series of shared tasks aimed at fostering advancements in automatic clinical note generation from doctor-patient conversations. This initiative included three primary tasks: Short Dialogue2Note Summarization (Task A) [5], Full Dialogue2Note Summarization (Task B) [5], and Note2Dialogue Generation (Task C) [5]. My focus primarily lies in Task A, which involves generating a section summary, inclusive of both the section header and text, based on short snippets of doctor-patient conversation.

3.1 Task A1: Header Classification

Task A1 of MEDIQA-Chat 2023 presents the challenge of classifying short doctor-patient conversations into predefined section headers. The task demands a nuanced understanding of the content to categorize each dialogue into one of twenty possible sections. These headers, which include categories like 'Allergy', 'Diagnosis', and 'Medications', form the basis of structuring clinical notes in a way that is both accessible and informative for healthcare providers. The classification model's success hinges on its ability to discern subtle details within conversations and to align them with the corresponding section header that best encapsulates the discussion's focal points. Table 1 provides a snapshot of the twenty classifica-

Category	Description
ALLERGY	Documented allergies, particularly to medications, including adverse reactions.
ASSESSMENT	Physician's interpretation and summarization of patient's health issues.
CC (Chief Complaint)	Primary reason for the patient seeking medical attention.
DIAGNOSIS	Final diagnosis determined by the physician based on evaluation.
DISPOSITION	Patient's status at the end of the visit and instructions for follow-up.
EDCOURSE	Details of the patient's experience and treatment in the emergency department.
EXAM	Findings from the physical examination, covering various systems.
FAM/SOCHX	Patient's family health history and social lifestyle factors.
GENHX	General history including the history of present illness and demographics.
GYNHX	Patient's gynecological and obstetrical history, if applicable.
IMAGING	Results and findings from diagnostic imaging studies.
IMMUNIZATIONS	Record of vaccinations and current immunization status.
LABS	Results from laboratory tests and their clinical interpretations.
MEDICATIONS	List of current medications and prescriptions being taken by the patient.
OTHER HISTORY	Additional relevant historical information not covered in other categories.
PASTMEDICALHX	Comprehensive history of the patient's past medical conditions.
PASTSURGICAL	Record of past surgical procedures the patient has undergone.
PLAN	Outline of the treatment plan including any recommended actions or follow-up.
PROCEDURES	Details of any medical procedures performed on the patient.
ROS (Review Of Systems)	Systematic review of each major body system.

Table 1: List of Categories (Task A1) in MEDIQA-Chat 2023

tion categories used for Task A1, each representing a distinct facet of the clinical note-taking process.

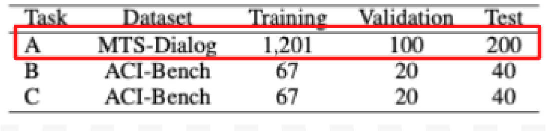
3.2 Task A2: Dialogue Summarization

Conversely, Task A2 expands upon the classification foundation laid by Task A1 and delves into the summarization of the doctor-patient dialogues. The objective is to distill the essence of these conversations into concise, coherent summaries that are aligned with the classified section headers. This task is intricate, as it involves not just the reduction of text but also the preservation of clinical relevance and the maintenance of narrative coherence. The model must navigate complex medical terminologies, patient concerns, and diagnostic details to produce a summary that accurately reflects the content and context of the interactions. Figure 1 depicts a schematic that encompasses both Task A1 and Task A2, illustrating the nature of classification and summarization in the generation of clinical notes.

4 Dataset Description

The MEDIQA-Chat 2023 dataset, provided by Ben Abacha et al. (2023) [4], comprises doctor-patient dialogue transcripts, each paired with relevant section headers and summary notes. This dataset is segmented into training, validation, and test sets. The training set includes 1,201 conversation pairs along with their corresponding section headers and summaries. The validation set contains 100 pairs, while the test set encompasses 200 conversations. These dialogues span across 20 diverse section headers, ranging from Medications and Review of Systems to Past Surgical History and more, offering a comprehensive representation of clinical scenarios.

For this project, I relied exclusively on the MTS-Dialog dataset as defined for Task A. However, it



Task	Dataset	Training	Validation	Test
A	MTS-Dialog	1,201	100	200
B	ACI-Bench	67	20	40
C	ACI-Bench	67	20	40

Figure 3: Task A: Data - MTS Dialog

is worth noting that additional datasets (Dialog-sum and Samsum) were tested during fine-tuning but discarded after no noticeable improvements were seen in preliminary research. In addition to MIMIC-IV Notes, both these datasets will be considered for future research.

5 Evaluation Metrics

The evaluation of MEDIQA-Chat 2023 tasks leverages an ensemble of metrics for a thorough and precise assessment. ROUGE (Lin, 2004) [15], a pivotal metric in summarization, evaluates the F1 scores across ROUGE-1, ROUGE-2, ROUGE-L and ROUGE-L-Sum. Complementing ROUGE, BLEURT scores (Sellam et al., 2020) [22] are employed to gauge semantic parallels between the generated and reference summaries. Additionally, BERTScore (Zhang et al., 2020) [26] provides sentence-level semantic analysis. These metrics, carefully chosen based on their strong correlation with human judgment in clinical note generation, collectively offer a holistic evaluation perspective. The average score derived from ROUGE-1, BLEURT-20, and BERTScore (microsoft/deberta-xlarge-mnli) serves as the primary criterion for ranking my results in Task A1, ie short note generation (Average-Score).

For Task A, the evaluation also includes the accuracy of section header classification, adding another dimension to the assessment process.

6 Baseline Models

In my approach to establishing a baseline for the MEDIQA-Chat 2023 tasks, I followed the guidelines set forth by the workshop's authors, focusing on utilizing OpenAI's GPT-3.5-turbo. This decision was based on GPT-3.5-turbo's proficiency in language understanding and generation tasks, making it a suitable choice for handling the complexities of medical dialogue summarization and classification.

In hindsight, I find that the actual leaderboard shows a baseline of GPT-4, which I have tested with as well and the results are in line with those of GPT-3-Turbo which is the current baseline.

6.1 Task A1: Section Header Classification

Task A1 required the classification of doctor-patient conversations into one of twenty distinct section headers. To guide GPT-3.5-turbo in this task, I formulated a specific prompt that directed the model to first classify the conversation into a category such as FAMILY HISTORY/SOCIAL HISTORY, HISTORY OF PRESENT ILLNESS, PAST MEDICAL HISTORY, and so on. Subsequently, the model was instructed to provide a summary of the conversation in the style of a clinical note.

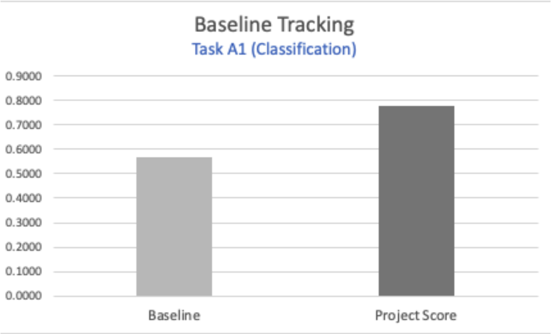


Figure 4: Task A1: Baseline

The prompt was structured verbatim with that used by the authors and is as follows:

"Classify the conversation into one of these 20 classes: FAMILY HISTORY/SOCIAL HISTORY, HISTORY of PRESENT ILLNESS, PAST MEDICAL HISTORY, CHIEF COMPLAINT, PAST SURGICAL HISTORY, Allergy, REVIEW OF SYS-

TEMS, Medications, Assessment, Exam, Diagnosis, Disposition, Plan, EMERGENCY DEPARTMENT COURSE, Immunizations, Imaging, GYNECOLOGIC HISTORY, Procedures, Other history, Labs. The response should start with the selected class, followed by the summary of the conversation in a clinical note style. The conversation is: ”

6.2 Task A2: Section Summary

For Task A2, the aim was to assess the quality of section headers and summaries generated by the model. I evaluated the performance of GPT-3.5-turbo using a set of metrics: Accuracy for Task A1, and Rouge-1, BLEURT, and BERT Score F1 for Task A2. These metrics were selected due to their proven effectiveness in measuring the relevance and quality of generated summaries.

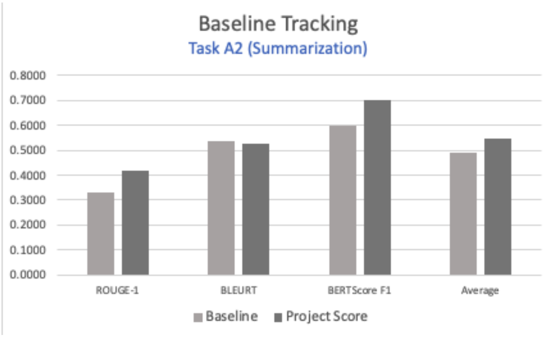


Figure 5: Task A2: Baseline

7 Solution

This paper details the development of a novel two-stage approach to tackle the challenges of medical dialogue classification and summarization. By leveraging the strengths of RoBERTa and Clinical Longformer for Task A1, the study achieved an impressive classification accuracy of 0.78, rivaling the highest scores at the conference. The methodology’s success is attributed to the design and application of an arbitration process and the careful curation of model hyperparameters.

For Task A2, the research ventured into uncharted territory, applying prompt engineering to a BART-Samsum model traditionally used without prompts. This strategy led to a noticeable improvement in summarization quality, marking a departure from conventional BART applications and offering new insights into the versatility of transformer-based models in the clinical domain.

7.1 Task A1: Section Header

For Task A1 I implemented an ensemble model merging RoBERTa and Clinical Longformer (alternatively, Clinical BigBird). This model aimed to classify doctor-patient conversations into one of twenty categories with high precision. The challenge was considerable due to the complex nature of medical dialogues.

RoBERTa-Based Classification: The initial stage involved fine-tuning RoBERTa for the 20-class problem. This model, trained to identify the most likely category for each conversation, achieved an accuracy of 0.77. The dialogues underwent tokenization and processing through RoBERTa, yielding category predictions. These were then compared to actual labels to assess accuracy.

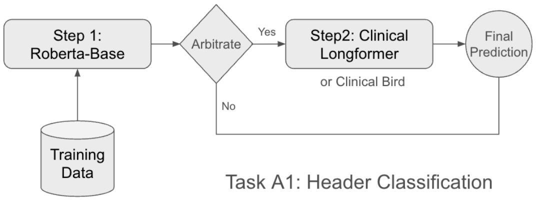


Figure 6: Task A1: Solution

Results

Arbitration and Clinical Longformer Refinement: The second stage introduced an arbitration process, focusing initially on ’GENHX’ and ’Medications’. Selected dialogues were reanalyzed using

Table 2: RESULTS: Task A1: Classification

Category	Label	Total Predictions	Correct Predictions	Precision	Total Actuals	Correct Actuals	Recall	Accuracy	F1
0	ALLERGY	11	11	1.000000	12	11	0.916667	0.78	0.956522
1	ASSESSMENT	5	4	0.800000	11	4	0.363636	0.78	0.500000
2	CC	9	5	0.555556	11	5	0.454545	0.78	0.500000
3	DIAGNOSIS	2	0	0.000000	1	0	0.000000	0.78	0.000000
4	DISPOSITION	2	1	0.500000	1	1	1.000000	0.78	0.666667
5	EDCOURSE	1	0	0.000000	4	0	0.000000	0.78	0.000000
6	EXAM	8	3	0.375000	5	3	0.600000	0.78	0.461538
7	FAM/SOCHX	49	44	0.897959	45	44	0.977778	0.78	0.936170
8	GENHX	48	42	0.875000	53	42	0.792453	0.78	0.831683
9	GYNHX	1	1	1.000000	1	1	1.000000	0.78	1.000000
10	IMAGING	2	1	0.500000	1	1	1.000000	0.78	0.666667
11	IMMUNIZATIONS	1	1	1.000000	1	1	1.000000	0.78	1.000000
12	LABS	0	0	0.000000	1	0	0.000000	0.78	0.000000
13	MEDICATIONS	12	10	0.833333	10	10	1.000000	0.78	0.909091
14	OTHER$_{H}ISTORY$	0	0	0.000000	3	0	0.000000	0.78	0.000000
15	PASTMEDICALHX	24	13	0.541667	14	13	0.928571	0.78	0.684211
16	PASTSURGICAL	6	6	1.000000	7	6	0.857143	0.78	0.923077
17	PLAN	3	1	0.333333	1	1	1.000000	0.78	0.500000
18	PROCEDURES	0	0	0.000000	1	0	0.000000	0.78	0.000000
19	ROS	16	13	0.812500	17	13	0.764706	0.78	0.787879

a fine-tuned Clinical Longformer model. This step refined predictions through detailed examination of specific dialogues, utilizing Longformer's extensive context analysis capabilities.

This dual-model approach enhanced overall accuracy, achieving 0.78, equating the conference's highest accuracy. The selection process for the second step's dialogues was adaptable, based on the first step's outcomes, highlighting potential areas for further exploration.

The ensemble model's success in achieving high accuracy in medical dialogue classification showcases the effectiveness of combining different model strengths. The flexible arbitration process between the two steps offers interesting research possibilities, emphasizing the ensemble method's potential in complex tasks like medical NLP.

7.2 Task A2: Section Summary

for Task A2, I developed a novel approach for section summary generation that focused exclusively on finetuning a BART-Samsum model with the Samsum dataset. This task involved the challenge of accurately summarizing complex medical dialogues.

BART-Samsum Finetuning and Inference:

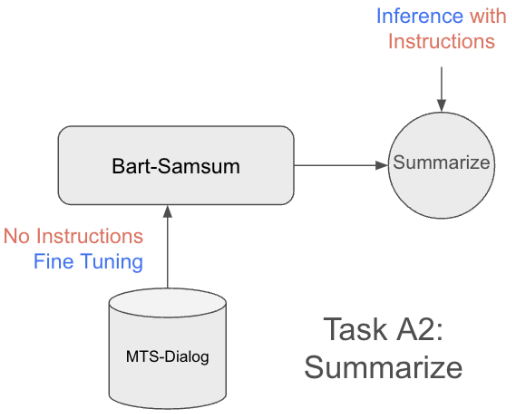


Figure 7: Task A2: Solution

Table 3: RESULTS: Task A2: Summarization

HuggingFace Model	**Model Type**	**Rouge-1**	**BertScore-F1**	**BLEURT**	**AvgScore**
20231130_BioBart-Base_5ep_Summ_Loss_0.76	Bio-Bart	0.4145	0.7086	0.5226	0.5486
20231205_Bart-lg-samsum_8ep_Summ_Loss_1.09	Bart-Samsum	0.4192	0.6986	0.5269	0.5482
20231130_BioBart-Base_10ep_Summ_Loss_0.77	Bio-Bart	0.4045	0.7196	0.5148	0.5463
20231205_Bart-lg-samsum_8ep_Summ_Loss_1.09	Bart-Samsum	0.4109	0.6910	0.5212	0.5410
20231129_Bart-Lg-samsum_3ep_Summ_Loss_0.81_R1_0.32	Bart-Samsum	0.4076	0.6939	0.5178	0.5398
20231130_BioBart-Base_5ep_Summ_Loss_0.76	Bio-Bart	0.4037	0.6986	0.5122	0.5382
20231207_Step_98_Retrain_Instrn_Bart-S_9ep_Loss_0.42	Bart-Samsum	0.3901	0.6916	0.5079	0.5299
20231207_Step_101_Retrain_Augmn_Instrn_BioBart_Xep_Loss_0.45	Bio-Bart	0.3796	0.6911	0.5097	0.5268
20231201_Clinic-T5-Lrg_9ep_Summ_Loss_0.93	Clinical-T5	0.3434	0.6535	0.4979	0.4982
20231130_Clinic-T5-Sci_20ep_Summ_Loss_0.84	Clinical-T5	0.3025	0.6300	0.4627	0.4651
20231130_Clinic-T5-Base_18ep_Summ_Loss_0.85	Clinical-T5	0.2833	0.6240	0.4694	0.4589

The core of the solution centered around finetuning the BART-Samsum model, initially trained on the SAMSum dataset, specifically on the MTS-Dialog training dataset. This finetuning was conducted without the use of explicit prompts, relying solely on the dialogue data. During the inference phase, a carefully crafted prompt was introduced to the finetuned model. This method of prompt engineering during the inference stage represents a deviation from typical BART model usage, which generally does not involve prompts for summarization.

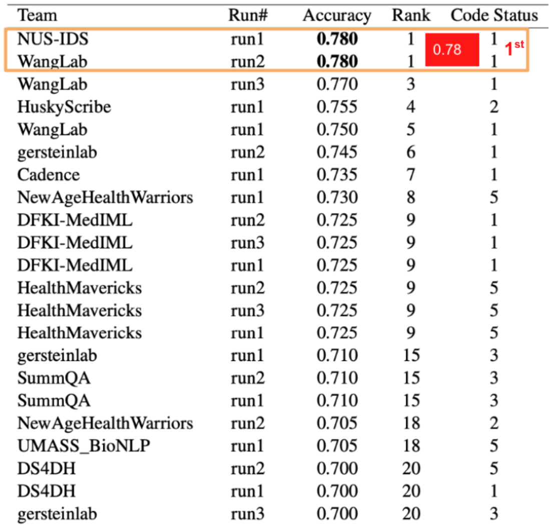
The inclusion of prompts during inference led to a notable improvement in the quality of the summarization, demonstrating a 2%-2.5% increase in performance. This finding is meaningful as it contrasts with the common practice in GPT-style models, where prompt-based inference is more typical. Applying prompts in a BART-based model for summarization, which is generally prompt-agnostic, underscores a novel and effective application of prompt engineering in transformer-based models for medical dialogue summarization.

Results

This approach, streamlining the process into a single step of fine-tuning and inference with prompt engineering, simplifies the methodology while maintaining effectiveness. It demonstrates the potential of specialized model training combined with strategic prompting in enhancing the quality of medical dialogue summaries.

8 Ranking

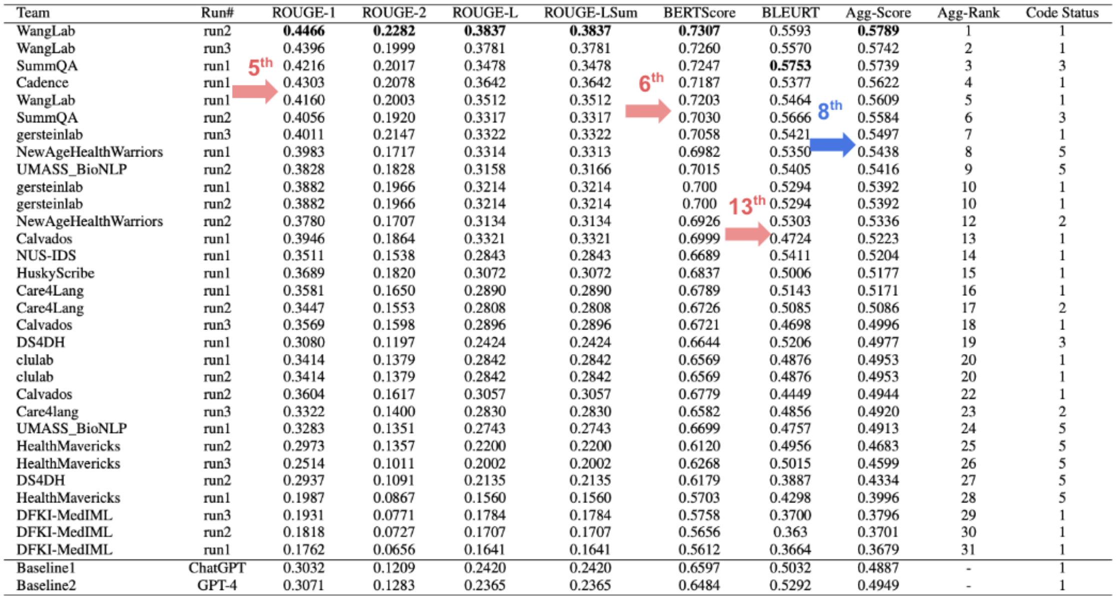
8.1 Task A1: Header Classification



Team	Run#	Accuracy	Rank	Code Status
NUS-IDS	run1	**0.780**	1	1
WangLab	run2	**0.780**	1	1
WangLab	run3	0.770	3	1
HuskyScribe	run1	0.755	4	2
WangLab	run1	0.750	5	1
gersteinlab	run2	0.745	6	1
Cadence	run1	0.735	7	1
NewAgeHealthWarriors	run1	0.730	8	5
DFKI-MedIML	run2	0.725	9	1
DFKI-MedIML	run3	0.725	9	1
DFKI-MedIML	run1	0.725	9	1
HealthMavericks	run2	0.725	9	5
HealthMavericks	run3	0.725	9	5
HealthMavericks	run1	0.725	9	5
gersteinlab	run1	0.710	15	3
SummQA	run2	0.710	15	3
SummQA	run1	0.710	15	3
NewAgeHealthWarriors	run2	0.705	18	2
UMASS_BioNLP	run1	0.705	18	5
DS4DH	run2	0.700	20	5
DS4DH	run1	0.700	20	1
gersteinlab	run3	0.700	20	3

Figure 8: Task A1: Ranking - partial

In Task A1, my project attained a top-tier accuracy rate of 0.78, matching the performance of leading group submissions. This achievement is noteworthy given the project's constraints: it was completed within a shorter timeframe, as a solo endeavor, and without leveraging augmented data from Task



Team	Run#	ROUGE-1	ROUGE-2	ROUGE-L	ROUGE-LSum	BERTScore	BLEURT	Agg-Score	Agg-Rank	Code Status
WangLab	run2	**0.4466**	**0.2282**	**0.3837**	**0.3837**	**0.7307**	0.5593	**0.5789**	1	1
WangLab	run3	0.4396	0.1999	0.3781	0.3781	0.7260	0.5570	0.5742	2	1
SummQA	run1	0.4216	0.2017	0.3478	0.3478	0.7247	**0.5753**	0.5739	3	3
Cadence	run1	0.4303	0.2078	0.3642	0.3642	0.7187	0.5377	0.5622	4	1
WangLab	run1	0.4160	0.2003	0.3512	0.3512	0.7203	0.5464	0.5609	5	1
SummQA	run2	0.4056	0.1920	0.3317	0.3317	0.7030	0.5666	0.5584	6	3
gersteinlab	run3	0.4011	0.2147	0.3322	0.3322	0.7058	0.5421	0.5497	7	1
NewAgeHealthWarriors	run1	0.3983	0.1717	0.3314	0.3313	0.6982	0.5350	0.5438	8	5
UMASS_BioNLP	run2	0.3828	0.1828	0.3158	0.3166	0.7015	0.5405	0.5416	9	5
gersteinlab	run1	0.3882	0.1966	0.3214	0.3214	0.700	0.5294	0.5392	10	1
gersteinlab	run2	0.3882	0.1966	0.3214	0.3214	0.700	0.5294	0.5392	10	1
NewAgeHealthWarriors	run2	0.3780	0.1707	0.3134	0.3134	0.6926	0.5303	0.5336	12	2
Calvados	run1	0.3946	0.1864	0.3321	0.3321	0.6999	0.4724	0.5223	13	1
NUS-IDS	run1	0.3511	0.1538	0.2843	0.2843	0.6689	0.5411	0.5204	14	1
HuskyScribe	run1	0.3689	0.1820	0.3072	0.3072	0.6837	0.5006	0.5177	15	1
Care4Lang	run1	0.3581	0.1650	0.2890	0.2890	0.6789	0.5143	0.5171	16	1
Care4Lang	run2	0.3447	0.1553	0.2808	0.2808	0.6726	0.5085	0.5086	17	2
Calvados	run3	0.3569	0.1598	0.2896	0.2896	0.6721	0.4698	0.4996	18	1
DS4DH	run1	0.3080	0.1197	0.2424	0.2424	0.6644	0.5206	0.4977	19	3
clulab	run1	0.3414	0.1379	0.2842	0.2842	0.6569	0.4876	0.4953	20	1
clulab	run2	0.3414	0.1379	0.2842	0.2842	0.6569	0.4876	0.4953	20	1
Calvados	run2	0.3604	0.1617	0.3057	0.3057	0.6779	0.4449	0.4944	22	1
Care4lang	run3	0.3322	0.1400	0.2830	0.2830	0.6582	0.4856	0.4920	23	2
UMASS_BioNLP	run1	0.3283	0.1351	0.2743	0.2743	0.6699	0.4757	0.4913	24	5
HealthMavericks	run2	0.2973	0.1357	0.2200	0.2200	0.6120	0.4956	0.4683	25	5
HealthMavericks	run3	0.2514	0.1011	0.2002	0.2002	0.6268	0.5015	0.4599	26	5
DS4DH	run2	0.2937	0.1091	0.2135	0.2135	0.6179	0.3887	0.4334	27	5
HealthMavericks	run1	0.1987	0.0867	0.1560	0.1560	0.5703	0.4298	0.3996	28	5
DFKI-MedIML	run3	0.1931	0.0771	0.1784	0.1784	0.5758	0.3700	0.3796	29	1
DFKI-MedIML	run2	0.1818	0.0727	0.1707	0.1707	0.5656	0.363	0.3701	30	1
DFKI-MedIML	run1	0.1762	0.0656	0.1641	0.1641	0.5612	0.3664	0.3679	31	1
Baseline1	ChatGPT	0.3032	0.1209	0.2420	0.2420	0.6597	0.5032	0.4887	-	1
Baseline2	GPT-4	0.3071	0.1283	0.2365	0.2365	0.6484	0.5292	0.4949	-	1

Figure 9: Task A2: Ranking - partial

C, a resource likely utilized by other teams to boost training effectiveness. Furthermore, the success was realized without relying on public Large Language Models (LLMs) such as GPT-4 or Cohere, adhering to data privacy considerations. The project's result can be attributed to the innovative application of prompt-enhanced BART models and the strategic use of ensemble models, carefully optimized through hyperparameter selection.

8.2 Task A2: Section Summary

For Task A2, which focused on section summary, my model's performance varied between the 5th and 13th positions when compared to current team benchmarks, depending on the metric being considered, such as ROUGE-1 or average score. This variation can be observed despite the mentioned project constraints and the absence of public LLMs in the development process. The application of prompt engineering to BART models, which is atypical for such models traditionally not prompted like GPT-style models, provided a novel edge. The careful balance of hyperparameters and model selection resulted in an edge for this project.

9 Limitations

The study presented in this document, while comprehensive, does not encapsulate the full array of potential methods for generating clinical notes. The utilized dataset's scope, both in size and range of medical specialties, is constrained, suggesting the necessity for additional validation across broader datasets and clinical contexts. Additionally, the absence of external medical knowledge integration, a factor known to bolster performance in similar tasks, marks a delineated path for future enhancements.

The development of the models described was constrained by the limited access to advanced computational resources, which restricted the exploration of larger, potentially more effective language models. Moreover, reliance on publicly available datasets and the deliberate avoidance of commercial language models with proprietary APIs, due to data privacy considerations, presented challenges that were addressed through novel methodological approaches such as prompt engineering for non-interactive models. The variability observed in the generated summaries, even under controlled settings, highlights the stochastic nature of the language models and underscores the imperative for more stable and deterministic output in clinical applications.

10 Conclusion

The methodologies outlined in this paper, reflecting a retrospective analysis aligned with the MediQA2023-Chat challenge, underscore the potential and adaptability of current NLP technologies in the realm of medical dialogue summarization and classification. The implementation of an ensemble model for Task A1 and the application of prompted inference in a fine-tuned BART model for Task A2 have demonstrated that competitive results can be achieved even when operating under resource constraints, such as limited computational power and data privacy concerns. The project's alignment with top-tier results, notably without the aid of augmented data or the use of public LLMs, attests to the effectiveness of the employed strategies.

This endeavor not only navigated the challenges of data scarcity and computational limitations but also introduced techniques like prompt engineering in traditionally non-prompted models, thus providing a blueprint for effective utilization of NLP in sensitive domains. The results, showcasing competitive standings despite the absence of collaborative team efforts and extended timelines, lay the groundwork for further exploration and refinement of these techniques.

Looking forward to 2024, the intention is to expand upon these foundations with augmented datasets and deeper analyses covering a broader spectrum of tasks. It is anticipated that this progressive research will contribute to the collective understanding and development of NLP applications within clinical settings, fostering advancements that are both technically sound and ethically responsible.

References

[1] Asma Ben Abacha, Wen wai Yim, George Michalopoulos, and Thomas Lin. An investigation of evaluation metrics for automated medical note generation, 2023.

[2] Griffin Adams, Jason Zucker, and Noémie Elhadad. A meta-evaluation of faithfulness metrics for long-form hospital-course summarization, 2023.

[3] Amal Alqahtani, Rana Salama, Mona Diab, and Abdou Youssef. Care4Lang at MEDIQA-chat 2023: Fine-tuning language models for classifying and summarizing clinical dialogues. In Tristan Naumann, Asma Ben Abacha, Steven Bethard, Kirk Roberts, and Anna Rumshisky, editors, *Proceedings of the 5th Clinical Natural Language Processing Workshop*, pages 524–528, Toronto, Canada, July 2023. Association for Computational Linguistics.

[4] Asma Ben Abacha, Wen-wai Yim, Griffin Adams, Neal Snider, and Meliha Yetisgen. Overview of the MEDIQA-chat 2023 shared tasks on the summarization & generation of doctor-patient conversations. In Tristan Naumann, Asma Ben Abacha, Steven Bethard, Kirk Roberts, and Anna Rumshisky, editors, *Proceedings of the 5th Clinical Natural Language Processing Workshop*, pages 503–513, Toronto, Canada, July 2023. Association for Computational Linguistics.

[5] Asma Ben Abacha, Wen-wai Yim, Yadan Fan, and Thomas Lin. An empirical study of clinical note generation from doctor-patient encounters. In Andreas Vlachos and Isabelle Augenstein, editors, *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2291–2302, Dubrovnik, Croatia, May 2023. Association for Computational Linguistics.

[6] Yulong Chen, Yang Liu, Liang Chen, and Yue Zhang. DialogSum: A real-life scenario dialogue summarization dataset. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli, editors, *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 5062–5074, Online, August 2021. Association for Computational Linguistics.

[7] Bharath Chintagunta, Namit Katariya, Xavier Amatriain, and Anitha Kannan. Medically aware GPT-3 as a data generator for medical dialogue summarization. In Chaitanya Shivade, Rashmi Gangadharaiah, Spandana Gella, Sandeep Konam, Shaoqing Yuan, Yi Zhang, Parminder Bhatia, and Byron Wallace, editors, *Proceedings of the Second Workshop on Natural Language Processing for Medical Conversations*, pages 66–76, Online, June 2021. Association for Computational Linguistics.

[8] John Giorgi, Augustin Toma, Ronald Xie, Sondra S. Chen, Kevin R. An, Grace X. Zheng, and Bo Wang. Wanglab at mediqa-chat 2023: Clinical note generation from doctor-patient conversations using large language models, 2023.

[9] Bogdan Gliwa, Iwona Mochol, Maciej Biesek, and Aleksander Wawer. SAMSum corpus: A human-annotated dialogue dataset for abstractive summarization. In Lu Wang, Jackie Chi Kit Cheung, Giuseppe Carenini, and Fei Liu, editors, *Proceedings of the 2nd Workshop on New Frontiers in Summarization*, pages 70–79, Hong Kong, China, November 2019. Association for Computational Linguistics.

[10] Colin Grambow, Longxiang Zhang, and Thomas Schaaf. In-domain pre-training improves clinical note generation from doctor-patient conversations. In Emiel Krahmer, Kathy McCoy, and Ehud Reiter, editors, *Proceedings of the First Workshop on Natural Language Generation in Healthcare*, pages 9–22, Waterville, Maine, USA and virtual meeting, July 2022. Association for Computational Linguistics.

[11] Sarthak Jain, Ramin Mohammadi, and Byron C. Wallace. An analysis of attention over clinical notes for predictive tasks. In Anna Rumshisky, Kirk Roberts, Steven Bethard, and Tristan Naumann, editors, *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 15–21, Minneapolis, Minnesota, USA, June 2019. Association for Computational Linguistics.

[12] Sanjeev Kumar Karn, Rikhiya Ghosh, Kusuma P, and Oladimeji Farri. shs-nlp at RadSum23: Domain-adaptive pre-training of instruction-tuned LLMs for radiology report impression generation. In Dina Demner-fushman, Sophia Ananiadou, and Kevin Cohen, editors, *The 22nd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks*, pages 550–556, Toronto, Canada, July 2023. Association for Computational Linguistics.

[13] Peter Lee, Sebastien Bubeck, and Joseph Petro. Benefits, limits, and risks of gpt-4 as an ai chatbot for medicine. *New England Journal of Medicine*, 388(13):1233–1239, 2023. PMID: 36988602.

[14] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online, July 2020. Association for Computational Linguistics.

[15] Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain, July 2004. Association for Computational Linguistics.

[16] Sri Macharla, Ashok Madamanchi, and Nikhilesh Kancharla. nav-nlp at RadSum23: Abstractive summarization of radiology reports using BART finetuning. In Dina Demner-fushman, Sophia Ananiadou, and Kevin Cohen, editors, *The 22nd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks*, pages 541–544, Toronto, Canada, July 2023. Association for Computational Linguistics.

[17] Yash Mathur, Sanketh Rangreji, Raghav Kapoor, Medha Palavalli, Amanda Bertsch, and Matthew Gormley. SummQA at MEDIQA-chat 2023: In-context learning with GPT-4 for medical summarization. In Tristan Naumann, Asma Ben Abacha, Steven Bethard, Kirk Roberts, and Anna Rumshisky, editors, *Proceedings of the 5th Clinical Natural Language Processing Workshop*, pages 490–502, Toronto, Canada, July 2023. Association for Computational Linguistics.

[18] Kirill Milintsevich and Navneet Agarwal. Calvados at MEDIQA-chat 2023: Improving clinical note generation with multi-task instruction finetuning. In Tristan Naumann, Asma Ben Abacha, Steven Bethard, Kirk Roberts, and Anna Rumshisky, editors, *Proceedings of the 5th Clinical Natural Language Processing Workshop*, pages 529–535, Toronto, Canada, July 2023. Association for Computational Linguistics.

[19] Prakhar Mishra and Ravi Theja Desetty. NewAgeHealthWarriors at MEDIQA-chat 2023 task a: Summarizing short medical conversation with transformers. In Tristan Naumann, Asma Ben Abacha, Steven Bethard, Kirk Roberts, and Anna Rumshisky, editors, *Proceedings of the 5th Clinical Natural Language Processing Workshop*, pages 414–421, Toronto, Canada, July 2023. Association for Computational Linguistics.

[20] Francesco Moramarco, Alex Papadopoulos Korfiatis, Mark Perera, Damir Juric, Jack Flann, Ehud Reiter, Anya Belz, and Aleksandar Savkov. Human evaluation and correlation with

automatic metrics in consultation note generation. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5739–5754, Dublin, Ireland, May 2022. Association for Computational Linguistics.

[21] Kadir Bulut Ozler and Steven Bethard. clulab at MEDIQA-chat 2023: Summarization and classification of medical dialogues. In Tristan Naumann, Asma Ben Abacha, Steven Bethard, Kirk Roberts, and Anna Rumshisky, editors, *Proceedings of the 5th Clinical Natural Language Processing Workshop*, pages 144–149, Toronto, Canada, July 2023. Association for Computational Linguistics.

[22] Thibault Sellam, Dipanjan Das, and Ankur Parikh. BLEURT: Learning robust metrics for text generation. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892, Online, July 2020. Association for Computational Linguistics.

[23] Ashwyn Sharma, David Feldman, and Aneesh Jain. Team cadence at MEDIQA-chat 2023: Generating, augmenting and summarizing clinical dialogue with large language models. In Tristan Naumann, Asma Ben Abacha, Steven Bethard, Kirk Roberts, and Anna Rumshisky, editors, *Proceedings of the 5th Clinical Natural Language Processing Workshop*, pages 228–235, Toronto, Canada, July 2023. Association for Computational Linguistics.

[24] Hongyi Yuan, Zheng Yuan, Ruyi Gan, Jiaxing Zhang, Yutao Xie, and Sheng Yu. Biobart: Pretraining and evaluation of a biomedical generative language model, 2022.

[25] Boya Zhang, Rahul Mishra, and Douglas Teodoro. DS4DH at MEDIQA-chat 2023: Leveraging SVM and GPT-3 prompt engineering for medical dialogue classification and summarization. In Tristan Naumann, Asma Ben Abacha, Steven Bethard, Kirk Roberts, and Anna Rumshisky, editors, *Proceedings of the 5th Clinical Natural Language Processing Workshop*, pages 536–545, Toronto, Canada, July 2023. Association for Computational Linguistics.

[26] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert, 2020.