Enhancing Early Detection of Critical Conditions through Natural Language Processing of Electronic Health Records

Jiacheng Zhu, Jingyi Guo, Quan Minh Nguyen, Giang Vu
{jiachzhu, jfguo, qmn103, giangvt}@upenn.edu

Abstract

In this study, we explore the efficacy of advanced language models in the healthcare sector, specifically focusing on predicting patient discharge disposition using the clinical notes in the MIMIC-IV database. Our research is centered around the adaptation of natural language processing (NLP) techniques with language models such as BERT, PubMedBERT, and GPT to enhance discharge planning and patient care. A significant finding of our study is the superior performance of PubMedBERT equipped with a linear classification head, which achieved a 78% accuracy in predicting discharge outcomes, surpassing the fine-tuned RoBERTa strong baseline models that reached a 76% accuracy. This outcome highlights the importance of domain-specific language models in the analysis of complex medical records. In contrast, our experiments with GPT-3.5, which employs in-context learning, yielded a lower accuracy of approximately 68%. This result indicates certain limitations of this approach in the specific context of medical data interpretation. The differential performance of these models underlines the necessity for careful selection and customization of models to address the unique challenges presented by healthcare data analysis. Such tailored approaches are essential for supporting efficient patient care management and improving overall healthcare outcomes.

1 Introduction

The integration of advanced data analysis and machine learning in healthcare, particularly in predicting discharge disposition, has become a vital aspect of patient care. Predictive models are increasingly used to determine whether patients will return home, require rehabilitation, or need other forms of post-hospital care. This area of research and application has seen significant growth, as evidenced by studies utilizing databases like the eICU-Collaborative Research Database. For example, machine learning techniques have been employed to identify key predictors within the first 24 hours of admission for patients in acute neurological care. This approach has shown to improve planning and resource allocation (Mickle and Deb, 2022). Similarly, the development of logistic regression-based tools demonstrates the potential of these models in predicting discharge disposition within the same timeframe. Such tools can significantly enhance early discharge planning and patient outcomes (Ballester et al., 2018; Berman et al., 2019).

In our study, we aim to develop a predictive model using advanced natural language processing (NLP) techniques to classify patients' discharge dispositions from hospital care into three groups (Home, Extended Care, and Expired), utilizing clinical notes data from MIMIC-IV dataset. That is, by analyzing the input from the clinical notes, which contains the patients' surgical history, physical exam information, allergies, etc., our model could generate the discharge disposition predictions (Fig. 1). Clinical notes, which include detailed narratives of patient care, symptoms, and medical decision-making processes, are crucial for accurate risk prediction for hospitalization and emergency department visits during home healthcare (Song et al., 2022). Likewise, their roles in enhancing models for post-acute care decision-making were also highlighted (Kennedy et al., 2022). Our study extends the application of machine learning, demonstrating the efficacy of NLP techniques in analyzing clinical notes to predict patient discharge disposition. The potential of NLP in extracting and interpreting complex clinical data offers a promising avenue for furthering precision medicine, enabling more tailored and effective patient care strategies.

2 Literature Review

The study of BERT, a deep learning model, has shown its effectiveness in understanding ambigu-

ous language in text. Originally pre-trained on the Toronto BookCorpus and English Wikipedia, BERT excels in tasks like question answering and text classification. It has been adapted for the medical field, addressing the challenge of specialized medical terminology. Gu et al., 2020 and Microsoft developed PubMedBERT, a biomedical-specific variant of BERT, using two approaches: mixed-domain pretraining and building a model from scratch with PubMed data. This model, based on the transformer architecture by Vaswani et al., 2017, outperformed its counterparts in medical context understanding, showing potential for biomedical predictions (Devlin et al., 2018; Gu et al., 2020; Vaswani et al., 2017).

Another research focused on medical dialogue summarization using GPT-4 in the MEDIQA 2023 Shared Task. The task's complexity stems from unstructured medical conversations and specialized terminology. The authors developed a system utilizing GPT-4 for summarizing medical dialogues, comparing its performance with models like T5, GPT3, and BioBERT. GPT-4 showed potential in generating abstractive, concise summaries, enhancing medical documentation efficiency. The study also assessed few-shot prompting techniques with GPT-4, recognizing their limitations and suggesting future improvements. This research is significant for projects involving classification tasks in medical documentation (Mathur et al., 2023).

Lastly, Med-BERT, tailored for Electronic Health Records (EHR), addresses the unique aspects of healthcare data. Building on the BERT model, it was further trained on a specialized EHR dataset from 20 million patients. Med-BERT excels in disease prediction and patient outcome analysis, incorporating a novel pretraining task for predicting prolonged hospital stays. Its performance in healthcare analytics, particularly in disease prediction and patient outcome analysis, demonstrates its potential for classifying patient stages within the MIMIC-IV EHR dataset (Rasmy et al., 2020).

In summary, these studies highlight the advancements in applying deep learning models like BERT and GPT-4 to the medical field, focusing on language understanding, medical dialogue summarization, and healthcare analytics. Each model demonstrates unique strengths in handling complex medical data, offering promising directions for future research and application in healthcare.

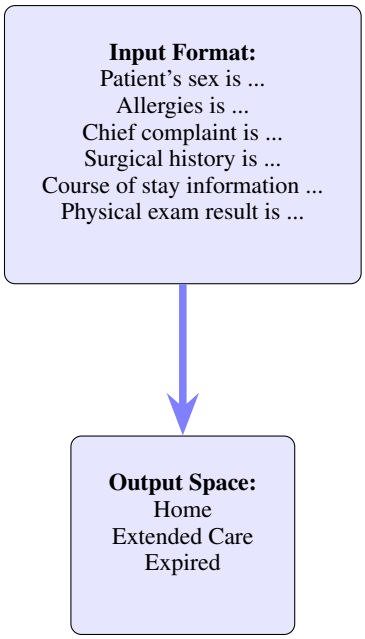

Figure 1: Illustration example of input and output

3 Experimental Design

3.1 Data

In this study, we utilized the MIMIC-IV Note dataset (version 2.2) from PhysioNet, focusing specifically on the discharge.csv file, which offers rich, detailed discharge summaries. These summaries contain de-identified patient details, medical service information, and specifics like allergies and chief complaints. They are essential for our classification task aimed at discerning critical patient stages (home, extended care, deceased) (Johnson et al., 2023). Classifying patients as "Home" indicates a diagnosis of good health without serious issues, allowing for discharge to their homes. "Extended Care" categorizes patients with varying degrees of disease symptoms who require care in services or hospitals. The label "Expired" is assigned to patients who have passed away. Due to the extensive size of the raw dataset, we selected a subset for efficiency, comprising a training set of 4,800 samples and a test set of 1,200 samples, as outlined in Table 1. A development set was omitted, as our models are language models with minimal hyperparameter tuning needs. This is

particularly relevant in the context of in-context learning, where a traditional training phase is not applicable. Instead, we used the language models to directly predict the texts in the test set, as illustrated in Fig 2.

Feature selection was a crucial step, considering BERT's 512-token limit, which is inadequate for the detailed electronic health records in our dataset (Gao, 2021). Our initial dataset encompassed six features including allergy information and chief complaints. However, to adapt to BERT's limitations (most texts exceeded 700 tokens, shown in Fig 3), we excluded less critical features. This decision was made under project constraints that prevented us from exploring models with a longer token capacity, like Longformer or BigBird (Beltagy et al., 2020; Zaheer, 2020). Our PubMedBERT analysis yielded an interesting insight: patient histories—medical, social, and family were less critical than traditionally assumed, likely due to their ambiguity and extensive use of abbreviations post de-identification. Furthermore, in the history sections, there may be mentions of medical symptoms from other individuals, such as patients' family members or acquaintances. These instances can introduce noise and potentially confuse NLP models, leading to less accurate interpretations. Excluding these histories brought the average token count within BERT's limit, facilitating effective model training (National Center for Biotechnology Information, U.S. National Library of Medicine, n.d.; Ankjær-Jensen and Lauritzen, 2013; Garg and Garg, 2019).

Moreover, initially, we had four labels including "Home with Service" in addition to the previously mentioned categories. Patients categorized as "Home with Service" exhibit mild disease symptoms but still require healthcare services after being discharged to their homes. However, we observed that the distinction between the "Home with Service" and "Extended Care" categories was inherently ambiguous. This led us to consolidate them into a single "Extended Care" category. Both groups of patients need additional support from healthcare services, regardless of the severity of their conditions. Therefore, merging these categories was a reasonable decision to eliminate ambiguity in our predictions.

Train text Example

allergies is latex / tetracycline / iv dye, iodine containing contrast media / keflin. chief complaint is doe/rv ppm lead misplacement. major surgical or invasive procedure is rv lead revision. physical exam is none. brief hospital course is primary reason for admission year old woman with past medical history of paf s/p biv ppm, copd, breast cancer s/p lumpectomy and radiation admitted from cards clinic today for periacrdial effusion without tamponade physiology for management of pericardial effusion and rv lead revision.

Train label Example

Extended Care

Figure 2: Example inputs and output labels

Dataset	Size
Train	4800
Test	1200

Table 1: Sizes of Train and Test Datasets

3.2 Evaluation Metric

In the context of our classification task, categorizing data into 'Extended Care', 'Home', and 'Expired', we have selected a suite of metrics for evaluating our model's performance. The chosen metrics include accuracy, F1 score, precision, recall, and a confusion matrix, each computed following their established mathematical formulations. The formulas for precision and recall are as follows:

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives}+\text{False Positives}}$$

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives}+\text{False Negatives}}$$

The F1 score, which combines precision and recall, is calculated using:

$$\text{F1 Score} = 2 \times \frac{\text{Precision}\times\text{Recall}}{\text{Precision}+\text{Recall}}$$

This set of metrics was chosen to provide a holistic assessment of the model's efficacy, offering not just a general success rate but also detailed insights

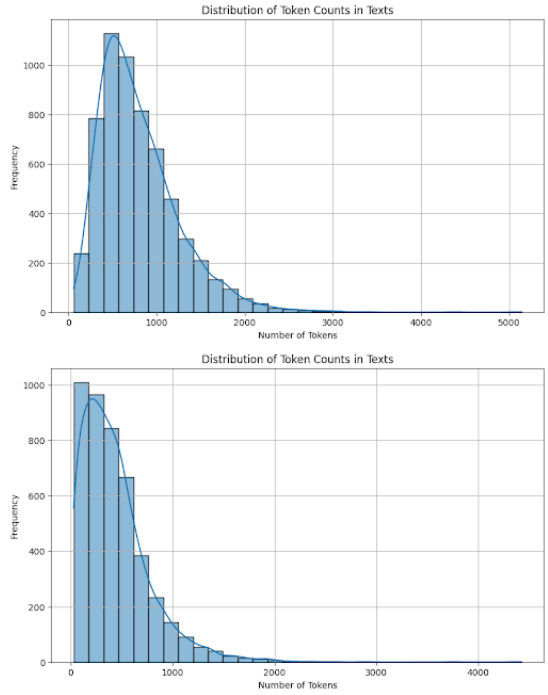

Figure 3: The Distribution of Token Counts in Medical Texts. The top plot demonstrates that the majority of texts, which include six features, are longer than 512 tokens. The bottom plot reveals that the number of tokens significantly drops after the removal of the History section, making it feasible to feed into BERT.

into the model's precision and reliability across each classification category. This comprehensive evaluation is crucial for understanding the strengths and potential areas of improvement in our classification model, aligning with standard practices in computer science research. Similar metrics have been employed in previous publications, such as the work on Med-BERT (Rasmy et al., 2020), demonstrating their relevance and effectiveness in comparable tasks (Wikipedia, n.d.b; Wikipedia, n.d.a).

3.3 Simple Baseline: RoBERTa Without Fine-Tuning

In this project, we established a baseline using the RoBERTa model, a derivative of the original BERT model, but refrained from engaging in the fine-tuning process. We retained the pre-trained BERT weights in their original state and concentrated solely on optimizing the weights of the linear classifier layer (Fig 4). This method exploits BERT's advanced language comprehension abilities, developed through training on a vast text corpus, while focusing our efforts on the task of linear classification. The basic BERT model was selected due to its ease of use, effectiveness in natural language processing (NLP) classification tasks, and its capability to understand nuanced textual contexts. In our preliminary evaluation, this baseline model achieved an accuracy of approximately 49.42%, an F1 score of 32.69%, a precision of 24.42%, and a recall of 49.42% (Table 2). These metrics serve as a benchmark, enabling us to assess the performance of our baseline model and guide our future work in enhancing and refining our linear classification techniques (Devlin et al., 2018).

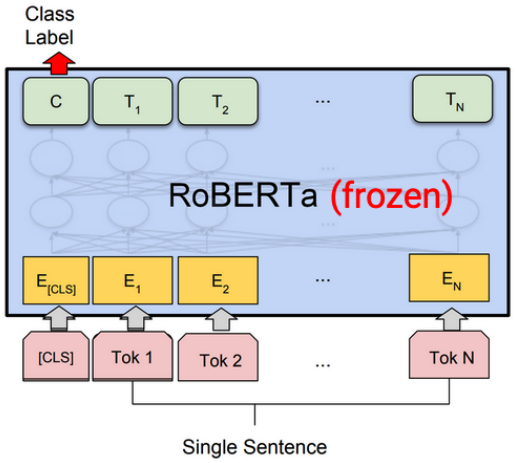

Figure 4: Basic Baseline Using RoBERTa Without Fine-Tuning

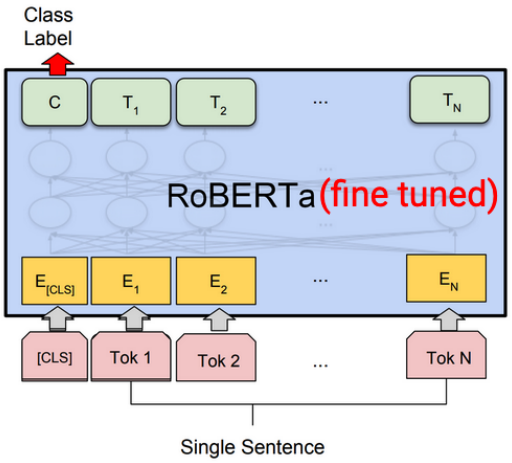

Figure 5: Strong Baseline Using RoBERTa With Fine-Tuning

4 Experimental Results

4.1 Strong Baseline: RoBERTa With Fine-Tuning

In computer linguistics, particularly for domain-specific text analysis in medical literature, adapting RoBERTa with fine-tuning has become a prevalent strategy. Models like 'Med-BERT' Rasmy et al.,

Model	Loss	Accuracy	F1	Precision	Recall	Epoch
RoBERTa w/o fine-tune	1.006448	0.494167	0.326872	0.244201	0.494167	3.0
RoBERTa w/ fine-tune	0.587303	0.759167	0.759407	0.763421	0.759167	3.0
Med-BERT w/ fine-tune	0.692040	0.781667	0.782680	0.784817	0.781667	3.0
PubMed Linear w/ fine-tune	0.673323	0.782500	0.782968	0.785014	0.782500	3.0
PubMed LSTM w/ fine-tune	0.655659	0.776667	0.777184	0.778512	0.776667	3.0
GPT-3.5 0-Shot	-	0.645833	0.602510	0.606720	0.645833	-
GPT-3.5 0-Shot w/ Explanation	-	0.669167	0.622090	0.636917	0.669167	-
GPT-3.5 2-Shot	-	0.688333	0.671471	0.674597	0.688333	-
GPT-3.5 3-Shot	-	0.655833	0.642635	0.638108	0.655833	-

Table 2: Model Performance Metrics

2020 refine RoBERTa, typically trained on various topics, to address the complexities of medical language. This process involves first continuing to pretrain on domain-specific data, and then, during actual tasks like classification, further fine-tuning the model. However, our approach differs; we use standard RoBERTa with fine-tuning, not continuing the pretraining on domain-specific data. Additionally, our task presents a threefold classification problem, in contrast to 'Med-BERT', which is focused on a binary classification task. The fine-tuning in our model adjusts both RoBERTa's pre-trained weights and its classification head using training dataset, thereby enhancing its proficiency in medical terminologies and nuances (Figure 5).

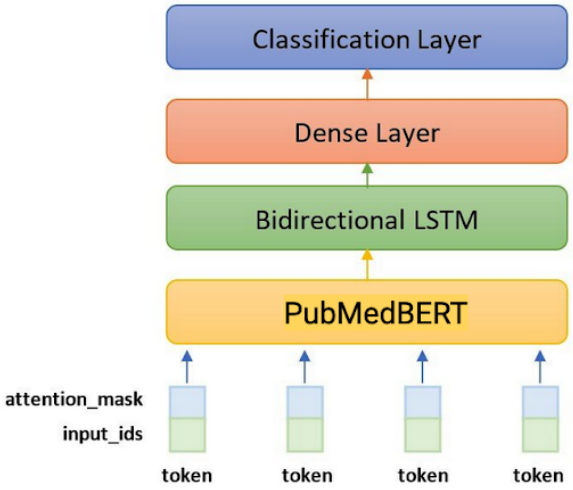

Figure 6: PubMedBERT with Bidirectional LSTM Head

Our evaluation highlights the effectiveness of our approach. The fine-tuned RoBERTa model demonstrated significant improvement, achieving an accuracy of approximately 75.17%, an F1 score of about 75.11%, and balanced precision and recall at 75.17% (Table 2). These results are very close to those of 'Med-BERT', which underwent additional pretraining on the PubMed database before fine-tuning on an Electronic Health Record (EHR) database for prediction (Table 2). The slight variance in results can be attributed to 'Med-BERT' undergoing extra pretraining before fine-tuning, while our method utilized standard RoBERTa with fine-tuning. These findings underscore the benefits and challenges of customizing RoBERTa for specialized domains, and they emphasize its adaptability and efficiency in the field of medical text analysis.

4.2 Extension 1: PubMedBERT

BioMedBERT, also known as PubMedBERT, was developed by Microsoft, has emerged as a crucial tool in natural language processing (NLP) for biomedical information extraction (Gu et al., 2020). This model, our primary extension in this project, demonstrates exceptional proficiency in understanding medical literature and terminologies, surpassing conventional models like RoBERTa in biomedical contexts. The MIMIC IV clinical notes, with their complex array of medical terminologies including allergy details, chief complaints, major procedures, physical examination insights, and brief hospital courses, present a challenging environment. These context-dependent terminologies necessitate a domain-specific model like BioMedBERT. For instance, the term "culture" has varied meanings across different features of the clinical notes, ranging from cell, tissue, or organism cultivation to aspects of human society such as arts and religions. To tackle this complexity, we have fine-tuned and optimized BioMedBERT for enhanced accuracy in predictive analysis.

A comparative evaluation shows that PubMedBERT, when compared to our refined baseline model, demonstrates incremental improvements. We based one of the PubMedBERT variants on the standard methodology in the Hugging Face framework, which incorporates a linear layer atop the pooled BERT output for sequence classification. This linear head variant showed modest improvements, with increases in accuracy and F1 scores under three percent (Table 2). Exploring more complex representations, we experimented with a bidirectional LSTM layer before the final linear classification layer. Unexpectedly, this added complexity did not improve results; the LSTM variant performed slightly worse than the linear model (Figure 6). Another notable model, Med-BERT, extends BERT's pretraining in the medical domain, and it is this approach that we aimed to mimic in our strong baseline. Med-BERT showed competitive but marginally inferior results to PubMedBERT with a linear head, with about a one percent difference in accuracy and F1 scores (Rasmy et al., 2020; Table 2). All models underwent fine-tuning on the pre-trained PubMedBERT framework, and the results, detailed in the accompanying table, underscore the importance of domain-specific pretraining in enhancing model performance.

4.3 Extension 2: GPT 3.5 Turbo

In our study, we conducted a detailed evaluation of in-context learning using GPT-3.5 Turbo, an advanced iteration of GPT-3.5 renowned for its improved capabilities in instruction following, output reproducibility, and expanded token output (OpenAI, n.d.). The experiments were designed to assess the model's performance on a complex test set comprising medical discharge summaries, with k-shot examples drawn from the training dataset. Our first experimental setup was a zero-shot variant, where GPT-3.5 Turbo classified these summaries without any prior examples, achieving an accuracy of 64.58%. To enhance its understanding, we introduced label explanations in this zero-shot scenario. This modification led to an improved performance, with accuracy increasing by 2%, reaching 66.92%.

Further experiments involved two-shot and three-shot learning setups, where the model was provided with two and three random examples from distinct

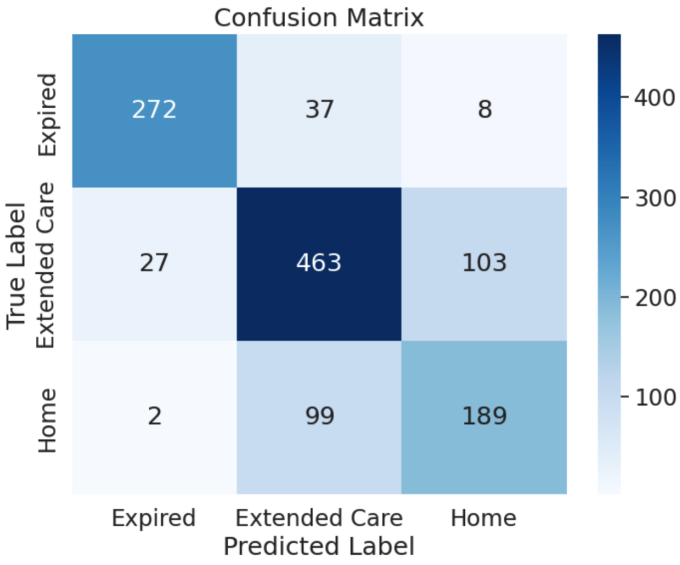

Figure 7: PubMedBert + Linear Head Confusion Matrix

classes for each prediction, respectively. These setups were intended to examine how the model's accuracy was influenced by the addition of context. The two-shot learning scenario emerged as the most effective, achieving an accuracy of 68.83%. Interestingly, the three-shot method, which provided an additional example, resulted in a slightly lower accuracy of 65.58% (Table 2). This finding suggested that an increased number of examples does not necessarily lead to higher performance in in-context learning. Throughout these experiments, a consistent challenge was the model's ability to differentiate between 'Extended Care' and 'Home' categories reliably, with non-deterministic outputs observed for similar cases within these categories. This variability in responses highlights the complexity of the task and points to the need for more refined prompting engineering. Moreover, in comparison to PubMedBERT, GPT-3.5 Turbo's in-context learning accuracy was found to be approximately 10% lower, emphasizing the challenges faced by generalized models in specialized domains and the importance of targeted, domain-specific training.

4.4 Error analysis

Our best-performing model in this project is PubMedBERT with a linear classification head, as indicated in Table 2. The confusion matrix depicted in Figure 7 shows that this model, like others evaluated in this study, tends to confuse the 'Home' and 'Extended Care' categories. This confusion is a prevalent issue and may arise from

the inherent difficulty in predicting whether a patient will return home without issues or require additional services after discharge. For instance, in one example where the model incorrectly classified a case as 'Extended Care' while the true label was 'Home', the patient's description included "left fifth toe ulcer…nondistended, nontender, no rebound or guarding, normoactive bowel sounds, no palpable masses…", suggesting a relatively minor issue. However, the model predicted 'Extended Care'. Interestingly, the model performed exceptionally well in classifying 'Expired', which indicates deceased patients. This observation suggests that the model may struggle to distinguish classes that do not have a wide gap in defining characteristics.

Comparing this to the published baseline, such as Med-BERT, which was mentioned earlier (Rasmy et al., 2020), we face limitations in direct comparison. Med-BERT was focused on binary mortality classification, a considerably different and arguably simpler task than our multi-class problem. Therefore, while our model shows some specific areas of confusion, particularly in closely related categories, the nature of the tasks and models are different enough that direct error comparison is not feasible.

5 Conclusion

In our project, we delved into the capabilities of advanced language models like PubMedBERT, BERT variants, and GPT-3.5 in the healthcare domain. Our findings highlight the substantial potential of these models, as well as the challenges they face in this specific field. Notably, we identified a critical limitation in the form of the 512-token input size cap inherent to BERT variants, including PubMedBERT, which hinders the processing of extensive patient records. Furthermore, our experiments with higher shot learning scenarios in GPT models, such as 6-shot or 12-shot learning, faced limitations due to exceeding the input token length limit and restrictions imposed by OpenAI's API rate limits.

While our implementations have shown promise, they are probably far from achieving state-of-the-art performance in this complex area, leaving considerable room for improvement. To address these issues and push the boundaries of our research, our future work will focus on integrating the Longformer attention mechanism with PubMedBERT. This integration is expected to resolve the challenges associated with longer input token sizes. Additionally, we plan to explore the potential of more advanced language models like GPT-4, particularly aiming to enhance their in-context learning capabilities.

These future initiatives are crucial in overcoming the current limitations and are expected to make significant contributions to the application of language models in healthcare. By optimizing these models, particularly in the context of patient discharge planning and decision-making processes, we aim to significantly improve efficiency and accuracy in healthcare settings. Our ultimate goal is to not only bridge the gap towards state-of-the-art performance but also to provide practical, impactful solutions in healthcare through the advancement of language model technologies.

Acknowledgements

We extend our heartfelt gratitude to our Teaching Assistant, Mona Gandhi, for her invaluable help, support, and guidance throughout the duration of this project. We would also like to express our sincere thanks to Professor Dr. Mark Yatskar for his exceptional teaching and the sharing of his extensive knowledge, which greatly enriched our learning experience and contributed significantly to the success of our project. Their combined expertise and dedication were instrumental in guiding us through the complexities of this research and in helping us achieve our objectives.

References

B. L. Ankjær-Jensen and M. B. Lauritzen. 2013. Details acquired from medical history and patients' experience of empathy – two sides of the same coin. *PubMed Central (PMC)*.

N. Ballester, P. J. Parikh, M. Donlin, E. K. May, and S. R. Simon. 2018. An early warning tool for predicting at admission the discharge disposition of a hospitalized patient. *Am J Manag Care*, 24(10):e325–31.

I. Beltagy, M. E. Peters, and A. Cohan. 2020. Longformer: The long-document transformer. *arXiv preprint*, arXiv:2004.05150.

J. E. Berman, A. Mata-Fink, H. F. Kassam, T. A. Blaine, and D. Kovacevic. 2019. Predictors of length of stay and discharge disposition after shoulder arthroplasty:

a systematic review. *JAAOS-Journal of the American Academy of Orthopaedic Surgeons*, 27(15):e696–e701.

J. Devlin, M. W. Chang, K. Lee, and K. Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint*, arXiv:1810.04805.

S. et al. Gao. 2021. Limitations of transformers on clinical text classification. *IEEE Journal of Biomedical and Health Informatics*, 25(9):3596–3607.

A. Garg and R. Garg. 2019. A descriptive analysis of patient history based on its relevance. *ResearchGate*.

Y. Gu, R. Tinn, H. Cheng, M. Lucas, N. Usuyama, X. Liu, T. Naumann, J. Gao, and H. Poon. 2020. Domain-specific language model pretraining for biomedical natural language processing. *arXiv preprint*, arXiv:2007.15779.

A. Johnson, T. Pollard, S. Horng, L. A. Celi, and R. Mark. 2023. Mimic-iv-note: Deidentified free-text clinical notes (version 2.2). *PhysioNet*.

E. E. Kennedy, K. H. Bowles, and S. Aryal. 2022. Systematic review of prediction models for post-acute care destination decision-making. *Journal of the American Medical Informatics Association*, 29(1):176–186.

Y. Mathur, S. Rangreji, R. Kapoor, M. Palavalli, A. Bertsch, and M. R. Gormley. 2023. Summqa at mediqa-chat 2023: In-context learning with gpt-4 for medical summarization. *arXiv preprint*, arXiv:2306.17384.

C. F. Mickle and D. Deb. 2022. Early prediction of patient discharge disposition in acute neurological care using machine learning. *BMC Health Services Research*, 22(1):1281.

National Center for Biotechnology Information, U.S. National Library of Medicine. n.d. *Medical History*.

OpenAI. n.d. *New Models and Developer Products Announced at DevDay*.

Laila Rasmy, Yang Xiang, Ziqian Xie, Cui Tao, and Degui Zhi. 2020. Med-bert: Pre-trained contextualized embeddings on large-scale structured electronic health records for disease prediction. *arXiv preprint*, arXiv:2005.12833.

J. Song, M. Hobensack, K. H. Bowles, M. V. McDonald, K. Cato, S. C. Rossetti, et al. 2022. Clinical notes: An untapped opportunity for improving risk prediction for hospitalization and emergency department visit during home health care. *Journal of Biomedical Informatics*, 128:104039.

A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. 2017. Attention is all you need. *arXiv preprint*, arXiv:1706.03762.

Wikipedia. n.d.a. F-score.

Wikipedia. n.d.b. Precision and recall.

M. et al. Zaheer. 2020. Big bird: Transformers for longer sequences. *arXiv preprint*, arXiv:2007.14062.

A Appendices

In the appendix, further elaboration on the confusion matrix related to in-context learning is provided, with additional details available on the following page.

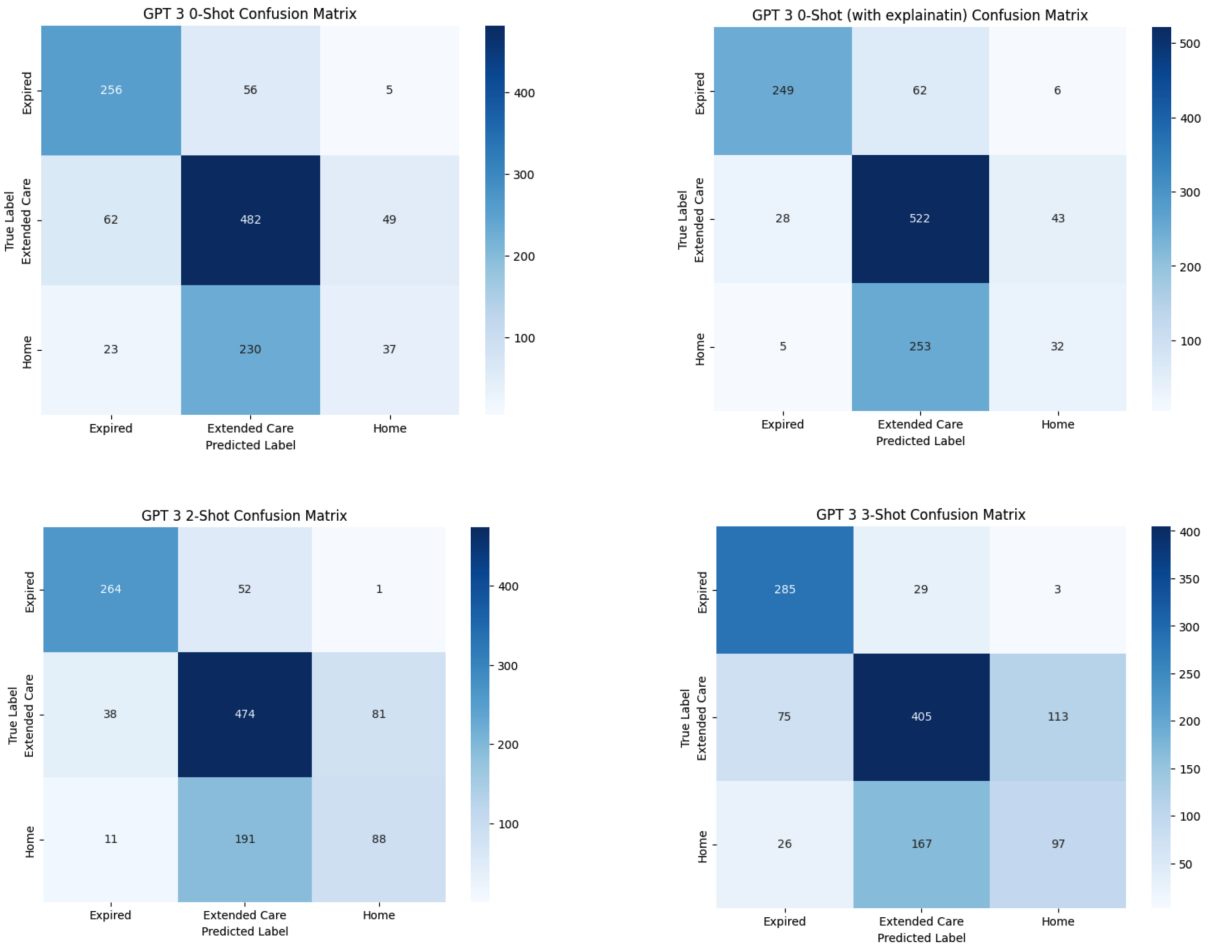

Figure 8: The confusion matrix for each variant of GPT-3.5 Turbo's in-context learning shows that all variants struggle with distinguishing between the 'Home' and 'Extended Care' categories.