Beyond the Clutter: Investigating The Impact of "Note Bloat" in NLP

Anonymous ACL submission

Abstract

Clinical notes often suffer from "bloat" – excessive repetition and redundancy that hinder efficient information retrieval and clinical decision-making (Rule et al., 2021) [1]. This project aims to investigate the prevalence and characteristics of bloated text in clinical notes and to develop and evaluate methods for mitigating it. Using the MIMIC-III dataset, we analyzed both within-note and between-note similarity through TF-IDF and cosine similarity measures to quantify redundancy. A clinical value assessment was conducted to differentiate between clinically relevant repetitions and unnecessary bloat. To further reduce redundancy, a greedy deduplication algorithm employing clinical embeddings from the Bio_ClinicalBERT(emi, 2025)[2] model was implemented. Our analysis revealed substantial internal redundancy within notes, especially in certain clinical specialties. Although a significant portion of the redundant content contained valuable keywords, the value ratio varied—underscoring the need for a nuanced approach to deduplication. The template review process identified recurring templates that contributed to the bloat, and the deduplication algorithm significantly reduced within-note similarity while improving readability scores. Our comprehensive approach reduces redundancy and enhances readability without compromising the clinically relevant information.

1 Introduction

The healthcare industry has significantly transformed due to the shift from traditional paper-based records to electronic health records (EHRs). While EHRs streamline data collection and save medical professionals valuable time, this digital convenience has inadvertently led to an increase in the practice of copying and pasting content between clinical notes, commonly referred to as *note bloat*. Note bloat generates overly lengthy and redundant notes, diminishing documentation quality and increasing the potential for clinical errors. These inflated notes are not only cumbersome for healthcare professionals but also pose potential challenges to data-driven methodologies.

Natural Language Processing (NLP) techniques have recently emerged as promising tools for supporting clinicians by facilitating decision-making processes and effectively extracting relevant medical information from textual data. Clinical NLP models rely on concise, accurate text to generate precise predictions and meaningful insights. However, the prevalence of note bloat could adversely impact the accuracy and efficiency of these NLP-driven models(Liu et al., 2022; Zhang et al., 2020)[3] [4].

This research addresses the problem of note bloat within EHRs, specifically exploring its impact on the performance of clinical NLP models. By examining how excessive, redundant documentation influences model predictions, we aim to highlight the importance of improving note documentation practices. Enhancing documentation quality through reduced note bloat is critical not only for healthcare professionals seeking clear and accurate records but also for data scientists and analysts leveraging NLP techniques for clinical prediction and text extraction tasks(Str, 2025).

2 Methods

2.1 Datasets and Preparation

We conducted our analyses using the MIMIC-III dataset, a publicly available repository containing de-identified health-related data associated with approximately 60,000 admissions to the intensive

[1] https://jamanetwork.com/journals/jamanetworkopen/fullarticle/2782054

[2] https://huggingface.co/emilyalsentzer/Bio_ClinicalBERT

[3] http://dx.doi.org/10.1016/j.jbi.2022.104149

[4] https://dl.acm.org/doi/abs/10.1145/3374217

care unit (ICU) (Alistair E.W. Johnson, 2016; MIM, 2025), [5]. MIMIC-III includes detailed clinical information such as vital signs, laboratory measurements, procedures, medications, and free-text notes, the latter of which was the subject of this study. Before analysis, data pre-processing was performed to clean, tokenize, and structure clinical notes appropriately for NLP modeling. Specifically, preprocessing included removing duplicates, anonymizing protected health information, and tokenization of text. Stratified sampling was employed to maintain proportional representation across critical categories, ensuring robustness and reducing bias in our models. The dataset was partitioned into training, validation, and test sets using proportions of 70%, 15%, and 15%, respectively, facilitating rigorous evaluation and validation of the models.

2.2 Model Architectures

We explored the use of ClinicalBERT, specifically Bio_ClinicalBERT, to assist with identifying and mitigating note bloat in our data. ClinicalBERT is a pretrained language model specifically fine-tuned on clinical text from the MIMIC-III dataset, making it particularly suitable for our analyses. By generating clinical embeddings using ClinicalBERT, we effectively captured the semantic nuances of medical terminology and clinical expressions within notes. These embeddings were integral to our deduplication process, allowing for a sophisticated measure of semantic similarity between text segments. Employing ClinicalBERT enhanced our ability to differentiate meaningful clinical content from redundant information, thereby supporting more accurate and contextually appropriate note bloat identification.

3 Experiments

3.1 Redundancy Analysis

In clinical documentation, redundancy can lead to significant "note bloat" and may hinder the retrieval of essential patient information. This analysis leverages Term Frequency-Inverse Document Frequency (TF-IDF) vectorization and cosine similarity to quantify redundancy at two scales: within individual clinical notes (Within-Note Similarity) and across notes grouped by clinical department (Between-Note Similarity).

First, textual data is transformed using the Term Frequency-Inverse Document Frequency (TF-IDF) approach. For a term t in a document d, the TF (term frequency) is defined as

$$\text{tf}(t,d) = \frac{\text{Number of occurrences of } t \text{ in } d}{\text{Total number of terms in } d},$$

while the IDF (inverse document frequency) is given by

$$\text{idf}(t,D) = \log\left(\frac{N}{\text{df}(t)}\right),$$

where N is the total number of documents in the corpus and $\text{df}(t)$ indicates the number of documents that contain the term t. The resulting TF-IDF score is then

$$\text{TF-IDF}(t,d) = \text{tf}(t,d) \times \text{idf}(t,D).$$

Once clinical texts are vectorized, the cosine similarity between two vectors $\vec{u}$ and $\vec{v}$ is computed as:

$$\text{sim}(\vec{u},\vec{v}) = \frac{\vec{u}\cdot\vec{v}}{\|\vec{u}\|\,\|\vec{v}\|}.$$

A cosine similarity score of 1 indicates identical orientation (maximum similarity), while a score of 0 implies orthogonality (no similarity).

For **Within-Note Similarity**, each clinical note is segmented into coherent units—such as different sections (e.g., History, Physical Examination, Assessment, Plan). Each segment s_i is converted into a TF-IDF vector, and the cosine similarity between every pair of segments s_i and s_j is calculated as:

$$\text{sim}(s_i,s_j) = \frac{\vec{v}_{s_i}\cdot\vec{v}_{s_j}}{\|\vec{v}_{s_i}\|\,\|\vec{v}_{s_j}\|}.$$

High similarity scores signal internal redundancy and thus indicate areas where information might be unnecessarily repeated.

For **Between-Note Similarity**, clinical notes are first grouped by department. Let D_k be the set of notes from department k. Each note n is represented by its TF-IDF vector, and the pairwise cosine similarities between notes are computed. The average similarity $\overline{S}$ for N notes is then given by:

$$\overline{S} = \frac{2}{N(N-1)}\sum_{i<j}\text{sim}(n_i,n_j).$$

By comparing intra-department and inter-department similarity scores, one can reveal both consistent standardized language within specific specialties and the variability across different clinical areas.

[5] https://physionet.org/content/mimiciii/1.4/

It is important to note that while TF-IDF and cosine similarity provide a robust quantitative framework, these methods do not capture the full semantic context inherent in clinical documentation. Extensions involving word embeddings or context-aware models (e.g., BERT) could further enhance the analysis by addressing synonymy, polysemy, and other semantic subtleties.

Overall, this integrated approach aids in identifying redundant content and can drive improvements in clinical note quality, thereby streamlining information extraction and enhancing the efficiency of electronic health records.

3.2 Clinical Value Assessment

The clinical value assessment process aims to quantify the clinical relevance of redundant content within clinical notes by integrating expert-curated clinically valuable keywords and evaluating redundant segments identified via within-note similarity analysis.

3.2.1 Valuable Keyword Identification

A comprehensive list of clinically valuable keywords was manually curated based on domain expertise and a review of relevant literature. This list encompasses terms related to medical conditions, procedures, medications, and other clinically significant aspects. The objective is to capture keywords that are known to influence clinical decision-making and patient management.

3.2.2 Redundant Segment Evaluation

Redundant segments, identified from the within-note similarity analysis, are further examined for clinical relevance. Each segment is compared against the curated list of valuable keywords using case-insensitive matching. A segment is considered to contain valuable clinical information if it includes at least one keyword from the curated list. This evaluation helps to discern whether repetitive content within a note holds substantive clinical significance.

3.2.3 Value Ratio Calculation

The *value ratio* for a clinical note is computed as the proportion of redundant segments that were deemed valuable based on the keyword matching. Let:

$$
\begin{aligned}
n_{\text{redundant}} &= \text{Total number of redundant segments} \\
&\quad \text{in the note,} \\
n_{\text{valuable}} &= \text{Number of redundant segments} \\
&\quad \text{containing at least one} \\
&\quad \text{valuable keyword.}
\end{aligned}
$$

Then, the value ratio R is defined as:

$$
R = \frac{n_{\text{valuable}}}{n_{\text{redundant}}}.
$$

This metric provides a quantitative measure of the clinical relevance of the repeated content. A higher value ratio indicates that a greater share of the redundancy carries important clinical information, whereas a lower ratio may suggest that the note contains repetitive content with little added clinical value.

3.3 Template Review and Policy

This section outlines the process of identifying, reviewing, and optimizing templates used in clinical documentation to reduce note bloat while preserving clinically essential information.

3.3.1 Template Extraction

All clinical notes were segmented, and the frequency of each segment across the dataset was computed. Segments that appeared frequently (e.g., at least five times) were flagged as potential templates or as frequently pasted blocks of text. This quantitative extraction helps highlight recurring patterns in the notes that may contribute to redundancy (Wrenn et al., 2010).

3.3.2 Manual Review

The extracted potential templates were then manually reviewed to assess their content and evaluate their impact on note bloat. During this review, templates containing redundant or unnecessary information were flagged. This step is critical to distinguish between reusable, standardized content and excessive repetition that may impair the clarity and utility of the notes.

These measures are intended to streamline clinical documentation, minimize redundancy, and enhance the overall quality and usability of electronic health records.

3.4 Deduplication using Clinical Embeddings

3.4.1 Embedding Generation

The pre-trained Bio_ClinicalBERT model, specifically the `emilyalsentzer/Bio_ClinicalBERT` version from the SentenceTransformer library, was employed to generate clinical embeddings for text segments. This model is chosen for its capability to capture nuanced semantic relationships in clinical language. Each segment of a clinical note is fed into the model, resulting in a high-dimensional vector representation (embedding) that encapsulates its semantic meaning.

3.4.2 Greedy Deduplication

A greedy deduplication algorithm is implemented to identify and remove redundant segments within a note. The process is initiated with the first segment, and for each subsequent segment, its embedding is compared against the embeddings of all previously retained segments using cosine similarity.

If the cosine similarity between the current segment and any retained segment exceeds a predefined threshold (e.g., 0.9), the current segment is considered a duplicate and is removed. Otherwise, the segment is retained as unique content.

3.4.3 Note Reconstruction

After processing all segments, the final step involves concatenating the retained segments to reconstruct a deduplicated version of the clinical note. This reconstruction ensures that essential clinical content is preserved while redundant information is eliminated, thereby improving the clarity and efficiency of the clinical documentation.

3.5 Outcome Measurement

3.5.1 Within-Note Similarity Re-evaluation

Following deduplication, the within-note similarity metric was recomputed for the deduplicated clinical notes using the same method as previously described. This involves calculating the cosine similarity between text segments for each note. The comparison of these similarity scores with those of the original notes provides a quantitative assessment of the redundancy reduction achieved through deduplication.

3.5.2 Readability Assessment

The impact of deduplication on note readability was evaluated by calculating readability scores for both the original and deduplicated notes. The Flesch Reading Ease formula from the `textstat` package was employed, where a higher score signifies improved readability. The difference in scores between the two sets of notes serves as an indicator of the effectiveness of the deduplication process in enhancing the clarity of the clinical documentation.

4 Results and Discussion

We evaluated the effectiveness of our deduplication methodology by analyzing changes in within-note similarity and readability scores before and after applying our deduplication techniques. Tables 1 and 2 summarize these metrics for the training and test datasets.

4.1 Within-Note Similarity

Table 1 illustrates the comparison of within-note similarity before and after deduplication.

Dataset	Base	Deduplicated
Train	0.000	0.016
Test	0.000	0.017

Table 1: Within-Note Similarity Scores

After deduplication, within-note similarity increased modestly from 0.000 to approximately 0.016 and 0.017 for the training and test datasets, respectively. This indicates an improvement in content differentiation due to the removal of redundant segments.

4.2 Readability Scores

Table 2 details the dramatic improvement in readability after deduplication.

Dataset	Base	Deduplicated
Train	-347.10	36.00
Test	-296.49	33.53

Table 2: Readability Scores

Prior to deduplication, the readability scores were severely negative, indicating poor readability due to redundancy. Post-deduplication scores significantly improved to 36.00 for training and 33.53 for testing, reflecting greatly enhanced note clarity and conciseness.

4.3 Discussion

The substantial improvement in readability scores, coupled with modest yet meaningful increases in within-note similarity, underscores the utility of

our methodology. Employing ClinicalBERT embeddings in conjunction with traditional similarity measures effectively reduces redundancy and enhances textual clarity. This enhancement has implications for clinical decision-making and the reliability of downstream NLP models. Future research should aim to optimize similarity thresholds and integrate further clinical context to maximize these gains.

4.4 Future Evaluation

While the primary focus of this evaluation was on quantifying redundancy reduction and readability improvements, further assessments could extend to the influence of deduplication on downstream NLP tasks. Future studies could examine how deduplication affects tasks such as clinical concept extraction, disease prediction, or clinical decision support by comparing the performance of NLP models trained on the original notes versus the deduplicated ones

Limitations

This study faced several limitations. Firstly, due to computational constraints, processing the entire dataset, consisting of over two million clinical notes, was not feasible given the equipment available. This limitation potentially impacted the precision of the study outcomes. Additionally, employing stratified sampling further constrained the available dataset size, specifically limited to the smallest category comprising only 1,470 entries. After partitioning into training, validation, and testing sets, the dataset size was further reduced, limiting the robustness of our analysis. Finally, limitations in accessing the most current dataset (MIMIC-IV-NOTE) restricted our ability to perform a potentially more comprehensive and insightful analysis.

Future studies could benefit from utilizing high-performance computing resources or employing advanced data handling techniques, such as distributed processing, to analyze larger datasets comprehensively. Future research might also consider alternative sampling methods or aggregate smaller categories to enlarge sample sizes, thus enhancing statistical power. Lastly, acquiring access to the latest datasets in subsequent studies would allow validation and testing against more contemporary clinical documentation practices, potentially yielding richer insights and more relevant findings.

Ethics Statement

All data utilized in this study were fully de-identified and obtained from the publicly accessible MIMIC-III database, which complies with ethical standards for patient privacy and consent. Efforts were explicitly made to anonymize any protected health information (PHI) to maintain confidentiality. Furthermore, the findings of this study aim to positively impact patient care by highlighting and addressing potential pitfalls in clinical documentation practices, ultimately supporting improved clinical outcomes and NLP applications.

Acknowledgements

The author would like to thank their course instructors Amit Bhattacharyy and Mark Butler for their guidance. They are also grateful to Dr. Julien Colbert for their assistance explaining their respective work in this area and contextual experience with this problem as well as acting as the inspiration behind undertaking this project.

References

2025. emilyalsentzer/bio$_{c}linicalberthuggingface$. . Accessed: 2025-04-13.

2025. Mimic-iii clinical database v1.4. https://physionet.org/content/mimiciii/1.4/. Accessed: 2025-04-13.

2025. Streamline e/m documentation by targeting note bloat | american medical association. https://www.ama-assn.org/practice-management/cpt/streamline-em-documentation-targeting-note-bloat. Accessed: 2025-04-13.

Lu Shen Li-wei H. Lehman Mengling Feng Mohammad Ghassemi Benjamin Moody Peter Szolovits Leo Anthony Celi Roger G. Mark Alistair E.W. Johnson, Tom J. Pollard. 2016. Mimic-iii, a freely accessible critical care database. *Scientific Data*, 3(1).

Jinghui Liu, Daniel Capurro, Anthony Nguyen, and Karin Verspoor. 2022. "note bloat" impacts deep learning-based nlp models for clinical prediction tasks. *Journal of Biomedical Informatics*, 133:104149.

Adam Rule, Steven Bedrick, Michael F. Chiang, and Michelle R. Hribar. 2021. Length and redundancy of outpatient progress notes across a decade

at an academic medical center. *JAMA Network Open*, 4(7):e2115334.

J. O Wrenn, D. M Stein, S. Bakken, and P. D Stetson. 2010. Quantifying clinical narrative redundancy in an electronic health record. *Journal of the American Medical Informatics Association*, 17(1):49–53.

Wei Emma Zhang, Quan Z. Sheng, Ahoud Alhazmi, and Chenliang Li. 2020. Adversarial attacks on deep-learning models in natural language processing: A survey. *ACM Transactions on Intelligent Systems and Technology*, 11(3):1–41.

A Supplementary Figures and Tables

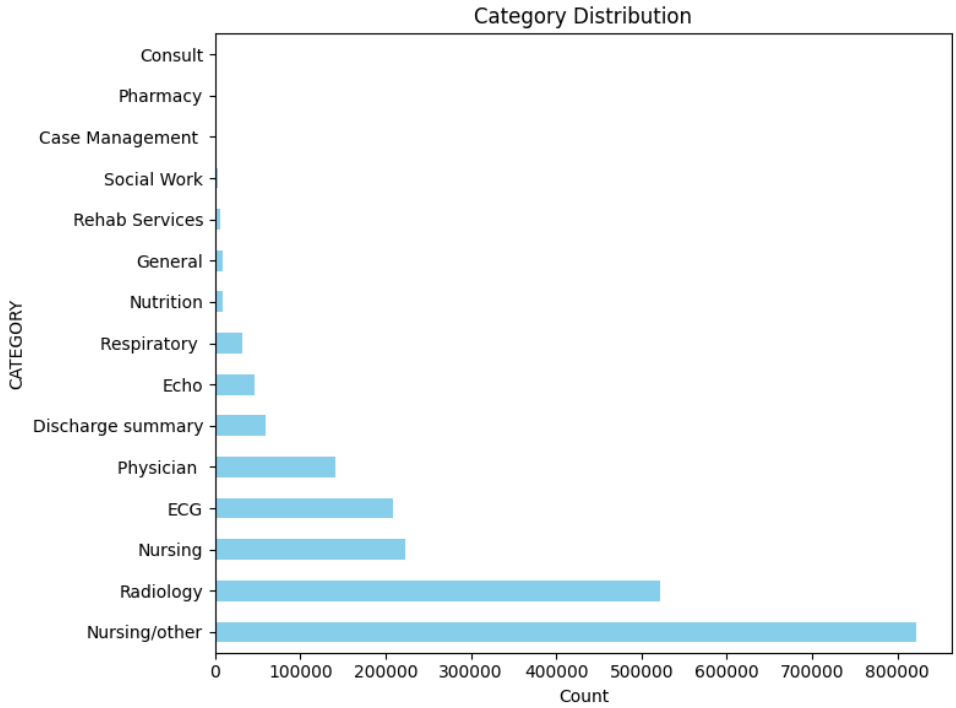

Figure 1: Original Distribution of Medical Departments MIMIC-III.

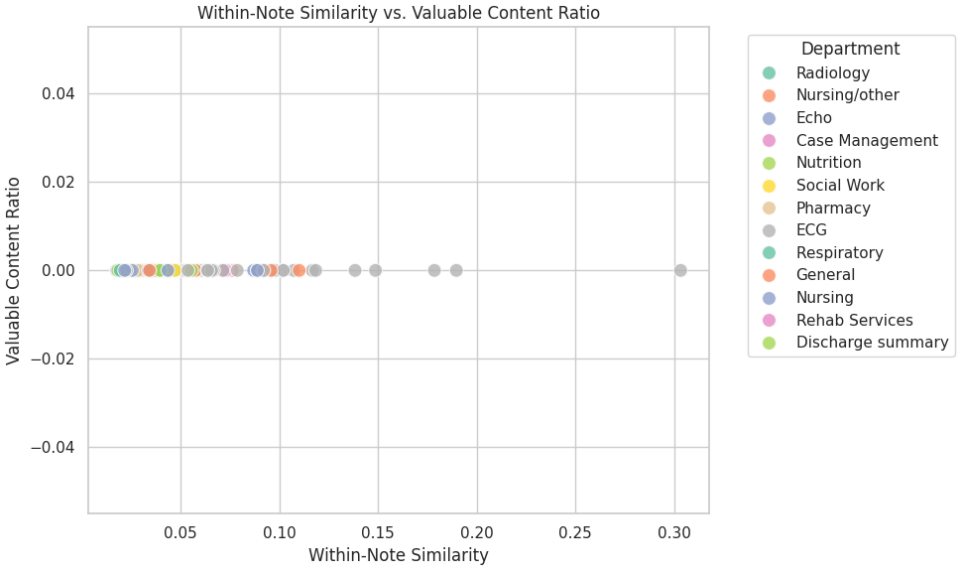

Figure 2: Valuable Content in Training Data.

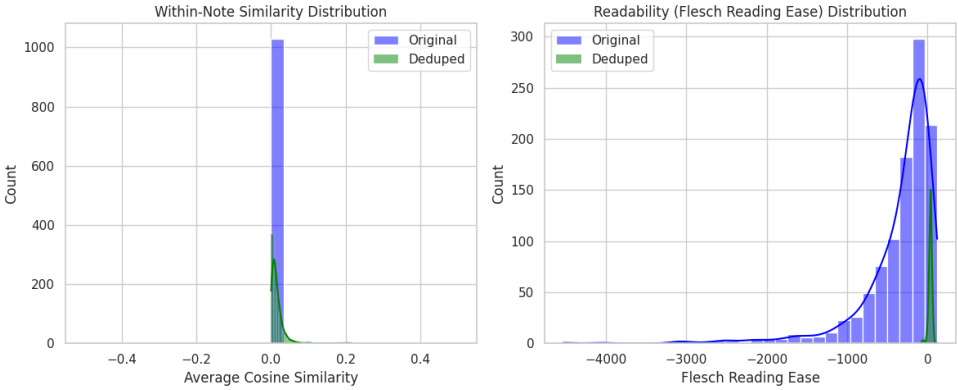

Figure 3: Training Data Deduplication Results.

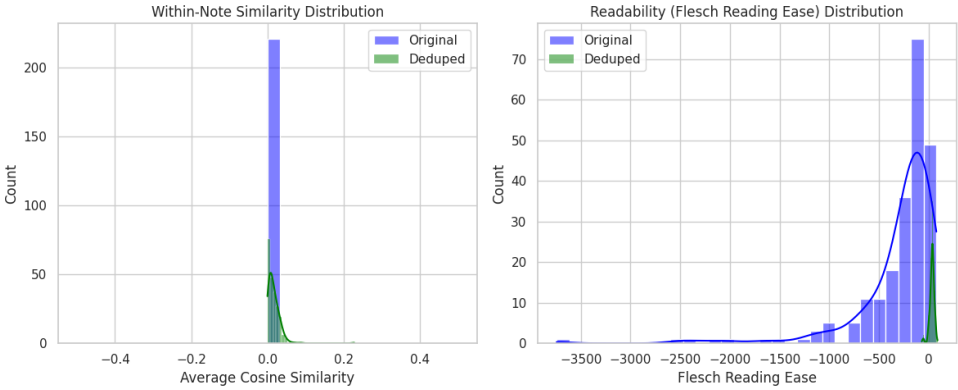

Figure 4: Test Data Deduplication Results.

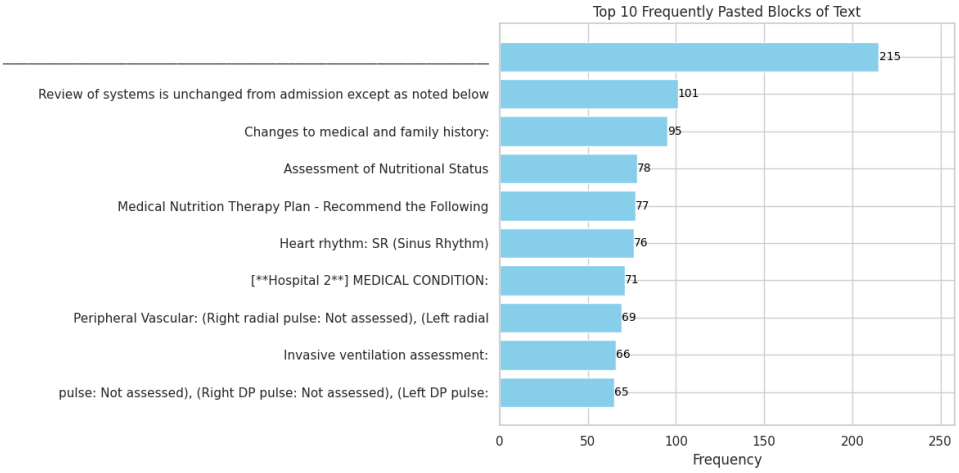

Figure 5: Training Data Commonly Used Templates.