Beyond Reconstruction: Generating Privacy-Preserving Clinical Letters

Libo Ren[1], **Samuel Belkadi**[2], **Lifeng Han**[1,3*]
Warren Del-Pinto[1], and **Goran Nenadic**[1]
[1] The University of Manchester, UK
[2] Cambridge University, UK
[3] LIACS & LUMC, Leiden University, NL
[*] *corresponding author*
l.han@lumc.nl, warren.del-pinto|g.nenadic@manchester.ac.uk
renlibo994, belkadisamuel@gmail.com

Abstract

Due to the sensitive nature of clinical letters, their use in model training, medical research, and education is limited. This work aims to generate diverse, de-identified, and high-quality synthetic clinical letters to enhance privacy protection. This study explores various pre-trained language models (PLMs) for text masking and generation, employing various masking strategies with a focus on Bio_ClinicalBERT. Both qualitative and quantitative methods are used for evaluation, supplemented by a downstream Named Entity Recognition (NER) task. Our results indicate that encoder-only models outperform encoder-decoder models. General-domain and clinical-domain PLMs exhibit comparable performance when clinical information is preserved. Preserving clinical *entities* and document *structure* yields better performance than fine-tuning alone. Masking stopwords enhances text quality, whereas masking nouns or verbs has a negative impact. BERTScore proves to be the most reliable quantitative evaluation metric in our task. Contextual information has minimal impact, indicating that synthetic letters can effectively replace original ones in downstream tasks. Unlike previous studies that focus primarily on reconstructing original letters or training a privacy-detection and substitution model, this project provides a framework for *generating diverse* clinical letters while embedding privacy detection, enabling sensitive dataset expansion and facilitating the use of real-world clinical data. Our codes and trained models will be publicly available at https://github.com/HECTA-UoM/Synthetic4Health

1 Introduction

Electronic clinical letters play a crucial role in healthcare communication. However, their sensitive nature makes them challenging to share and limits their adoption in clinical education and research (Tarur and Prasanna, 2021; Tucker et al., 2016; Spasic and Nenadic, 2020). Although public datasets such as MIMIC and i2b2 provide de-identified clinical data, they are often restricted to specific regions and institutions, limiting their representativeness of diverse clinical conditions (Humbert-Droz et al., 2022).

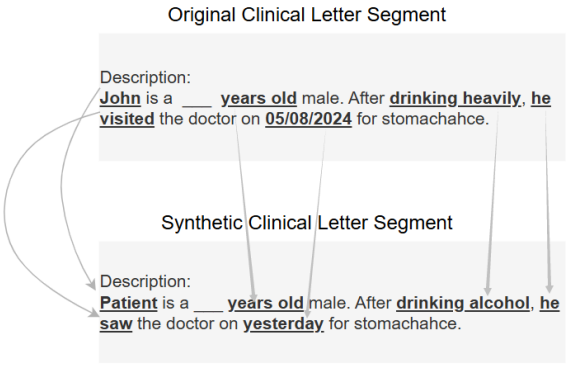

Figure 1: An Example of the Objective: generating more clinical letters from the original anonymised clinical letter segment with clinical soundness

To address these challenges, synthetic clinical letter generation has attracted growing interest. While existing methods primarily rely on structured data, Natural Language Generation (NLG) models provide a promising alternative by integrating linguistic and clinical knowledge (HÜSKE-KRAUS, 2003; Amin-Nejad et al., 2020a; Tang et al., 2023). Unlike previous studies, we go beyond training de-identification models to detect and substitute private information. This work focuses on leveraging NLG methods to generate synthetic clinical letters while indirectly minimising privacy risks. Although the dataset we used has been anonymised, we additionally apply a privacy detection and masking process as an additional verification step to further enhance the security of synthetic letters. Our findings contribute to bridging the gap in privacy-aware clinical letter generation, facilitating a more effective approach to processing real-world clinical

Proceedings of the Sixth Workshop on Privacy in Natural Language Processing, pages 60–74

letters and addressing data scarcity in the medical domain.

A brief example of our objective is shown in Figure 1. To achieve this, we investigate different *model architectures*, *segmentation strategies*, and *masking* techniques and evaluate their effectiveness both qualitatively and quantitatively. Additionally, we assess their usability in *downstream* NLP tasks such as Named Entity Recognition (NER). We ensure compliance with ethical guidelines by using only de-identified clinical data and adhering to all data use agreements.

2 Related Work

Biomedical patient data privacy protection has been an important task for clinical research, especially when it comes to big data era. Developing privacy-preserving decision support tools has been a challenge for statisticians and clinical researchers (Tucker et al., 2016; Claerhout and DeMoor, 2005; Terry, 2012; Liu et al., 2015).

Recent studies in clinical Natural Language Processing (NLP) explored various tasks, including NER, de-identification, and NLG. Several tools, such as SciSpacy (Dernoncourt et al., 2017; Kovačević et al., 2024), are designed to enhance domain-specific entity recognition, while Philter (Norgeot et al., 2020) combines both traditional and modern NLP models to identify and remove Protected Health Information (PHI). Transformer-based architectures are widely used in clinical NLG, particularly in text rewriting, discharge summary generation, and data augmentation, (Vaswani et al., 2017). For instance, LT3 (Belkadi et al., 2023) improves label-to-text generation, while DeID-GPT (Liu et al., 2023) employs GPT-4 to identify and generate substitute words for private information. Micheletti et al. (2024) demonstrate that Masked Language Models (MLMs) outperform Causal Language Models (CLMs) in text masking tasks. Existing studies either focus on training models, utilize existing LLMs *identify* to identify *private* information, or concentrate solely on *NLG* without much attention in privacy. However, few studies integrate clinical text generation with privacy-preservation and diversity considerations, which is the focus of this study.

3 Methodology

To generate clinical letters that retain the original clinical narrative without being exact duplicates, we employed various PLMs. Sensitive data is masked by and substituted with contextually predicted tokens using PLMs. Additionally, we evaluate different masking strategies to de-identify potentially sensitive information as an additional validation step. We also considered how non-sensitive elements, such as stopwords, indirectly influence the effectiveness of de-identification. A brief workflow is presented in Figure 2.

3.1 Dataset

The dataset used in this research comprises 204 clinical letters and 51,574 manually annotated clinical entities from the SNOMED CT Entity Linking Challenge (A et al., 2000; Johnson et al., 2024, 2023). Protected health information (PHI) was manually reviewed and replaced with underscores to ensure privacy. The length of the clinical letters ranges from 360 to 3,329 words, with an average length of approximately 1,450 words. Each letter contains patient information, medical history, and follow-up instructions. They are also stored in CSV format with unique identifiers and textual content. Given the input constraints of language models, clinical letters are tokenised and segmented into smaller chunks for processing before being merged. The entity annotations, sourced from SNOMED CT, cover 5,336 distinct clinical concepts and are stored in CSV format. These annotations map entity positions in the text to their corresponding SNOMED CT concepts. An excerpt from the dataset is shown in Figure 3.

3.2 Clinical Information Preserving

3.2.1 Experimental Setup

The collected dataset consists of raw clinical letters and annotations, which were first merged into a unified DataFrame. Manually annotated entities were then extracted based on their index. Since PLMs such as BERT, RoBERTa, and T5 have a token limit (typically 512 (Zeng et al., 2022)), we employed a *variable-length chunking* strategy (Subsection 3.2.2) rather than fixed-length truncation. All experiments were conducted using Google Colab Pro+ environment equipped with a T4 GPU (16GB VRAM), 52GB of system RAM, and 225GB of disk space, running Python 3.10, PyTorch 2.3.1, and Hugging Face Transformers 4.42.4.

For *feature* extraction, we used word_tokenize to preserve word integrity, which is crucial for retaining clinical entities. For masking and generation, we followed each model's native tokeniza-

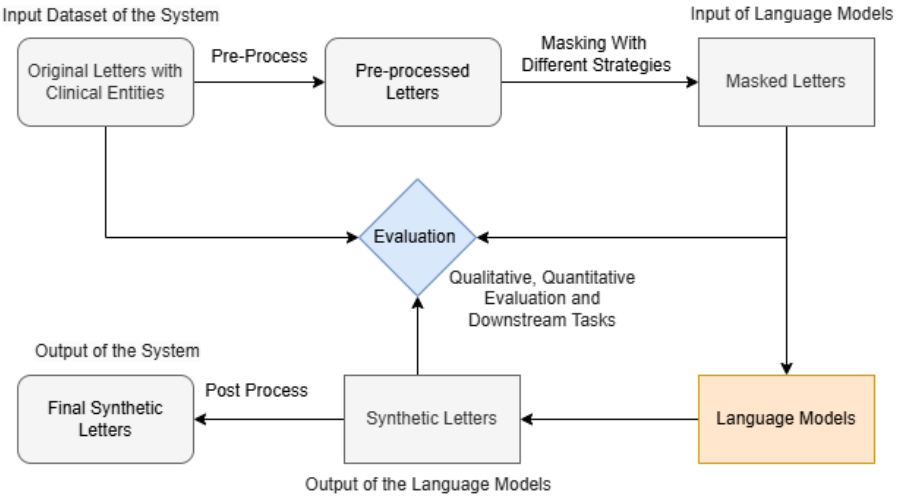

Figure 2: Overall Workflow

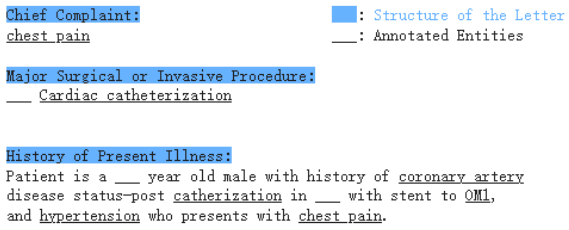

Figure 3: Text Excerpt from the Original Letter (A et al., 2000; Johnson et al., 2024, 2023) ('note_id': '17656866-DS-6')

tion method. BERT-based models utilize WordPiece tokenization, which is effective for handling out-of-vocabulary words and masked predictions. T5-based models employ Sentence-Piece tokenization, which better handles abbreviations and non-standard characters (e.g., "COVID-19")—common in clinical letters—as it does not rely on spaces for splitting. The pre-processing pipeline is shown in Figure 4.

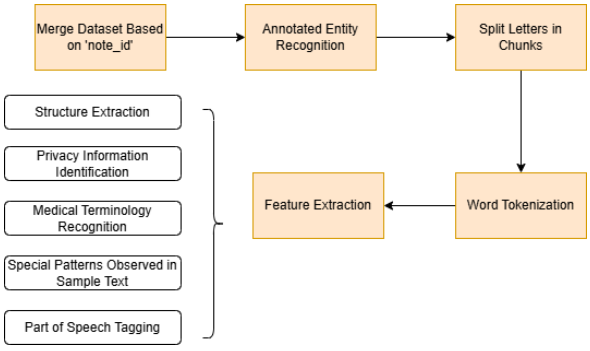

Figure 4: Pre-Processing Pipeline

3.2.2 Splitting Letters into Variable-Length Chunks

As mentioned above, pre-trained language models (PLMs) such as BERT, RoBERTa, and T5 have a token limit (typically 512 (Zeng et al., 2022)), requiring an effective strategy to process longer clinical letters. To preserve the full semantics of medical text, we adopted a Variable-Length Chunking approach based on *semantic boundaries*, instead of using tradition truncation methods like fixed-length or discarding tokens (Hou et al., 2022).

Initially, each letter was processed at the sentence level. However, this approach proved inefficient and lacked sufficient contextual information for inference. To address this, we segmented letters into *paragraph-sized chunks* while maintaining sentence integrity. Rather than strictly restricting each paragraph by '`max_tokens`' limit for each paragraph, we prioritised preserving complete sentences. To constrain fragmenting sentences, we introduced a '`max_lines`' threshold. If adding a sentence exceeds either the '`max_lines`' or `max_tokens` limit, it is moved to the next chunk. However, adhering to the `max_tokens` constraint should be our primary consideration due to model requirements. Therefore, if a single sentence does not exceed '`max_lines`' but surpasses the '`max_tokens`' limit, it is further segmented based on '`max_tokens`'. To detect sentence boundaries, we used the NLTK library. Figure 5 illustrates this process.

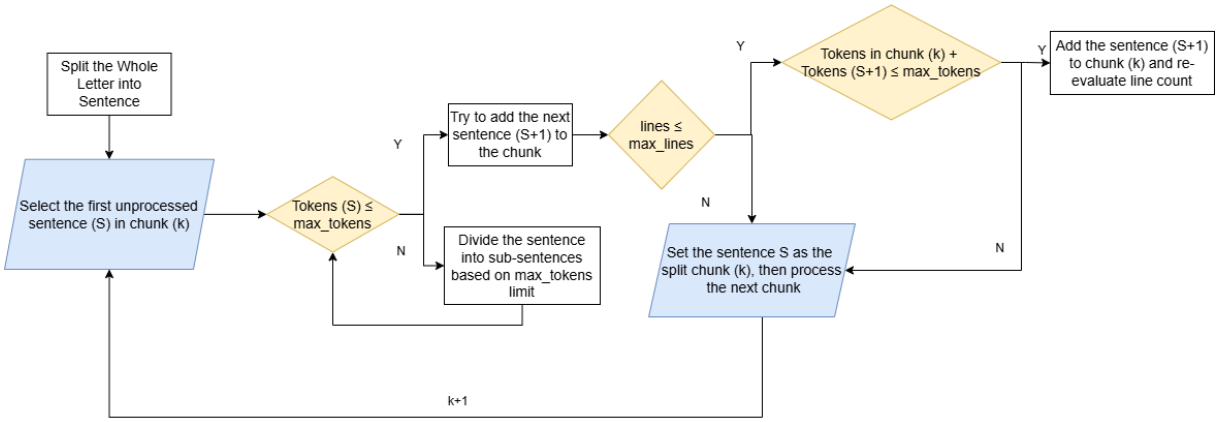

Figure 5: Text Chunking Workflow

3.2.3 Feature Extraction

To generate de-identified clinical letters while maintaining clinical narratives, we extracted key features before masking and generation. These features include:

- **Document Structure:** structural elements often correspond to capitalized headers and colons (:). They should be preserved as they define the document's format.
- **Privacy Information Identification:** An NER model (Stanza (Qi et al., 2020)) detected entities such as Name, Date, and Location, while regex masked structured data like phone numbers and emails.
- **Medical Terminology:** An NER model pretrained on i2b2 (Zhang et al., 2021) supplemented manual annotations by recognizing medical terms (e.g., Test, Treatment, Problem).
- **Special Patterns:** Medication dosages (e.g., enoxaparin 40 mg/0.4 mL) and abbreviations (e.g., b.i.d.) were retained unless classified as private.
- **POS Tagging:** To assess the impact of POS tagging on the model's understanding of clinical text, we employed a MIMIC-III-based model (Zhang et al., 2021), which outperformed NLTK and SpaCy in clinical syntactic comprehension.

3.3 Clinical Letters Generation

Our objective is to generate synthetic clinical letters that *differ* from the originals rather than producing near-identical copies, as repeated statement may indirectly reveal the patients' privacy. While fine-tuning improves precision and semantic comprehension, it risks overfitting, leading to outputs too closely aligned with the original dataset and reducing generalisability. Therefore, simply fine-tuning a model is suboptimal if PLMs can already generate readable text. Instead, the focus should be on protecting *clinical* terms and narratives while preventing *privacy* breaches. Since decoder-only models struggle with long-text processing (Amin-Nejad et al., 2020b) and require substantial computational resources, we explored both encoder-only and encoder-decoder PLMs with random masking. After evaluating their ability to generate synthetic letters, we selected Bio_ClinicalBERT for its strong domain adaptation and tested various masking strategies, as detailed in Appendix A. Additionally, given the discussion in Subsection 3.2.2, we assessed the impact of variable-length chunking on generation quality with Bio_ClinicalBERT.

3.3.1 Encoder-Only Models

Standard masked language modelling (MLM) was used in this study. First, tokens were selected for masking and then corrupted, resulting in masked text containing both masked and unmasked tokens. The model then predicted the masked tokens, replacing them with the most probable candidates. We predict all masked tokens in *parallel* within a single forward pass for each clinical letter. If processed sequentially, it might generate more coherent text, but the computational complexity would increase significantly (from *O(N)* to *O(N!)*). Given the clinical focus of this task, we explored models fine-tuned on clinical or biomedical datasets. However, since no clinically fine-tuned RoBERTa (Zhuang et al., 2021)

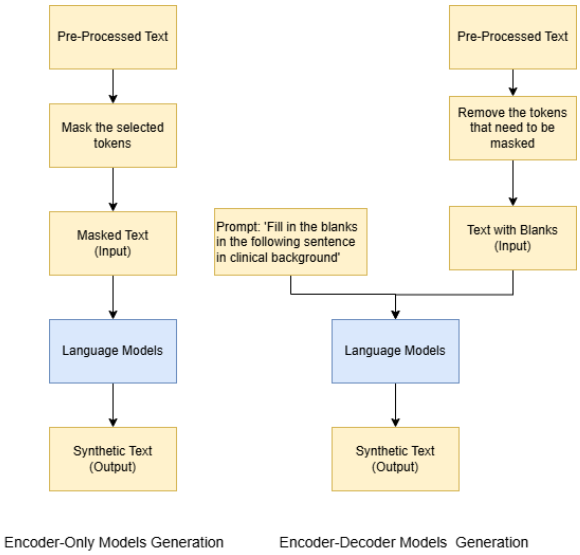

Figure 6: Comparison of Encoder-Only and Encoder-Decoder Model Architectures

variant was available, RoBERTa-base was used for comparison. The encoder-only models we evaluated include Bio_ClinicalBERT (Alsentzer et al., 2019), medicalai/ClinicalBERT (Wang et al., 2023), RoBERTa-base (Zhuang et al., 2021), and Clinical-Longformer (Li et al., 2023).

3.3.2 Encoder-Decoder Models

Although encoder-decoder models are not typically used for MLM, they excel in coherent text generation, particularly T5. Therefore, we included T5 family models in our comparisons. Unlike BERT, which replaces masked tokens with '<mask>', the T5 family models indexing masked words as 'extra_id_x'. The text, with these words removed, serves as input for generation, referred to as "text with blanks". For consistency, '<mask>' was later used when displaying masked text. Additionally, a **structured prompt** was required, formatted as "Fill in the blanks in the following sentence in clinical background" + text with blanks. Like encoder-only models, masked tokens are *predicted in parallel* across clinical letters. In this part, we experimented with T5-base (Raffel et al., 2020), Clinical-T5-Base (Eric and Johnson, 2023; Goldberger et al., 2000), Clinical-T5-Sci (Eric and Johnson, 2023; Goldberger et al., 2000), and Clinical-T5-Scratch (Eric and Johnson, 2023; Goldberger et al., 2000) for comparison. The architectures of encoder-only and encoder-decoder models are shown in Figure 6.

3.4 Evaluation

Both quantitative and qualitative methods are used to evaluate performance. Additionally, a downstream NER task assesses whether synthetic clinical letters can replace raw data. The evaluation pipeline is illustrated in Figure 8 of the Appendix.

3.4.1 Quantitative Evaluation

To assess the quality of synthetic letters, we conduct quantitative evaluation across multiple dimensions, including inference performance, readability, and similarity to raw data.

- **Standard NLG Metrics:** ROUGE, BERT Score, and METEOR assess textual similarity while ensuring generated text differs from the original. Synthetic text is compared with the original, and a baseline is established by comparing masked text to the original. The evaluation score should exceed the baseline but stay below 1.
- **Readability Metrics:** SMOG, Flesch Reading Ease, and Flesch-Kincaid Grade Level assess readability, with SMOG prioritised for clinical relevance.
- **Advanced Text Quality Metrics:** Perplexity, subjectivity, and information entropy are used to evaluate informativeness and subjectivity.
- **Invalid Prediction Rate:** Measures the ratio of invalid token predictions (e.g., subwords, punctuation) to assess the model's ability to generate meaningful text.
- **Inference Time:** Records generation time per letter, with shorter times indicating improved computational efficiency for large-scale deployment.

3.4.2 Qualitative Evaluation

While some synthetic texts performed well on most metrics, they did not always appear satisfactory upon visual inspection, whereas others with average scores appeared more natural. Although human evaluation is the most reliable method for assessing clinical letters, it is limited by time constraints and workload demands. Thus, combining qualitative and quantitative evaluations helps the identification of the most effective quantitative metrics for model evaluation. Once identified, one metric can serve as the benchmark standard, while others function

	Model Evaluation			
	RoBERTa-base	medicalai / ClinicalBERT	Clinical-Longformer	Bio _ ClinicalBERT
ROUGE-1				
Generation Performance	86.54	88.46	89.52	84.91
Baseline	84.91	84.91	84.91	84.91
ROUGE-2				
Generation Performance	74.51	78.43	79.61	73.08
Baseline	73.08	73.08	73.08	73.08
ROUGE-L				
Generation Performance	86.54	88.46	89.52	84.91
Baseline	84.91	84.91	84.91	84.91
BERTScore F1				
Generation Performance	0.81	0.83	0.84	0.85
Baseline	0.79	0.65	0.79	0.65
METEOR				
Generation Performance	0.87	0.88	0.90	0.86
Baseline	0.85	0.85	0.85	0.85
Flesch Reading Ease				
Generation Performance	10.24	18.70	9.22	16.67
Baseline (Original)	8.21	8.21	8.21	8.21
Baseline (Mask)	16.67	16.67	16.67	16.67

Table 1: Encoder-Only Models Comparison at the Sentence Level (The 'Baseline' without annotations was calculated by comparing masked text to the original text)

as complementary indicators. To address this, we selected a representative sample of clinical letters based on evaluation results, analysed the impact of different generation methods on these outcomes, and validated the findings with six additional samples to verify their consistency with quantitative metrics.

3.4.3 Downstream NER task

Beyond qualitative and quantitative evaluation, synthetic clinical letters were tested in a downstream NER task to assess their quality and potential as replacements for real clinical data. As shown in Figure 7, entities were first extracted from clinical letters using ScispaCy [1] and then used to train a base SpaCy [2] model. The trained model was applied to the test set, and the extracted entities were compared with those initially identified by ScispaCy to evaluate the consistency of entity recognition between synthetic and original clinical letters.

4 Results and Discussion

4.1 Model Comparison and Evaluation Metric Selection

4.1.1 Qualitative Results

Among encoder-only models, all four successfully generated meaningful words for masked input, correctly inferring 'r' from 'R ankle', demonstrating strong contextual understanding. Bio_ClinicalBERT further introduced relevant words absent from the input (e.g., "admitted") while maintaining clinical coherence, producing clinically sound sentences even without direct token matches, and effectively retaining clinical information while introducing diversity.

For encoder-decoder models, T5-base outperformed other variants but produced irrational outputs, including incomplete or nonsensical phrases (e.g., "open is a ___ yo male"). The other three T5 family models frequently generated de-identification (DEID) tags instead of meaningful replacements due to corpus biases. Overall, encoder-only models outperformed encoder-decoder models, aligning with previous research (Micheletti et al., 2024) showing that Masked Language Modelling (MLM) outperforms Causal Language Modelling (CLM) in medical text generation.

4.1.2 Quantitative Results

For **sentence-level** results, among encoder-only models, clinical-related models consistently outperform general domain RoBERTa-base, aligning with qualitative observations. **Bio_ClinicalBERT**, despite having no word overlap in this sample, achieves **the highest BERTScore** while maintaining a **clinically coherent output**. The encoder-decoder models generally perform poorly in most metrics compared to encoder-only models, except for METEOR. Their BERTScores are significantly

[1] https://allenai.github.io/scispacy/
[2] https://spacy.io/

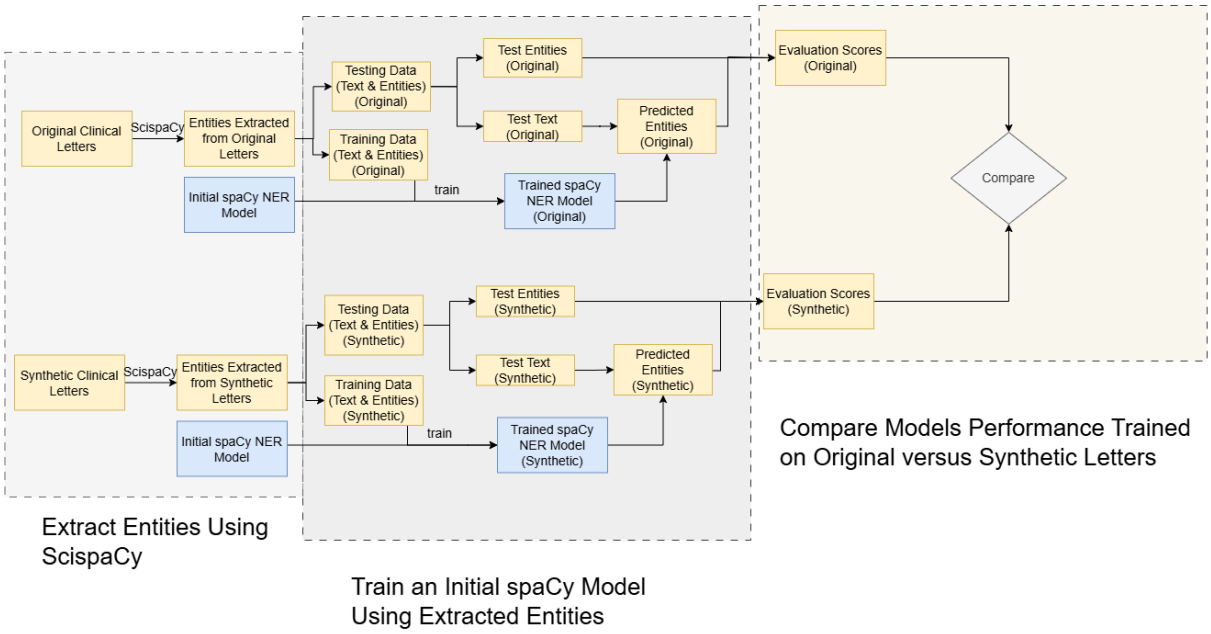

Figure 7: Workflow of Downstream NER Task

lower than the baseline, suggesting a large deviations from the original meaning. These findings further support the validity of BERTScore as the primary evaluation metric, with other metrics serving as supplementary references.

On the **full dataset**, **all encoder-only models performed similarly**, contradicting our hypothesis that clinical-related models would outperform base models. This suggests that training in clinical data does not significantly improve synthetic letter quality, likely because most clinical tokens were preserved, leaving only general tokens masked in our settings. BERTScore remains a reliable primary metric, as qualitative and quantitative evaluations align at both the sentence and dataset levels.

4.2 Variable-Length Chunk Segmentation

As mentioned in Subsection 3.2.2, we set 'max_lines' as a variable parameter and assigned a fixed value of 256 to 'max_tokens'. We tested increasing 'max_lines' values until the average tokens per chunk peaked, indicating that more clinical information could be preserved. Due to time constraints, the initial experiment on seven letters showed that 41 was the optimal 'max_lines value', where inference time decreased up to this point but rose beyond it (Table 3). This trend was consistent in 10- and 30-letter samples. However, inference time reflects only a general trend rather than precise measurements, as it is influenced by multiple factors, including chunk size and network conditions.

4.3 Masking Strategies

4.3.1 Random Masking

We evaluated the impact of masking ratios (i.e., masked tokens / total tokens) on the quality of synthetic clinical letters using Bio_ClinicalBERT. As expected, higher masking ratios led to lower similarity metrics, but all evaluation values remained above the baseline while staying below 1.0, indicating that the model preserves clinical context and generates understandable text. Notably, at a 1.0 masking ratio, BERTScore increased from 0.29 to 0.63, demonstrating Bio_ClinicalBERT's ability to *retain meaningful clinical information* despite *extensive masking*.

4.3.2 Masking Only Nouns

Masking nouns, which often correspond to Personally Identifiable Information (PII), helps verify de-identification while retaining clinical context. We found that *masking fewer nouns led to better performance across all metrics*, consistent with random masking. When the noun masking ratio reached 1.0, BERTScore increased from 0.70 to 0.89, indicating meaningful noun predictions. All evaluations are higher than the baseline but lower than 1.0. However, as the noun masking ratio increased further, BERTScore decreased significantly. To generate synthetic clinical letters that retain clinical information while being distinguishable, we

	Model Evaluation			
	T5-base	Clinical-T5-base	Clinical-T5-Scratch	Clinical-T5-Sci
ROUGE-1				
Generation Performance	86.79	85.19	87.38	80.36
Baseline	73.77	73.77	73.77	73.77
ROUGE-2				
Generation Performance	75.00	71.70	75.25	69.09
Baseline	63.33	63.33	63.33	63.33
ROUGE-L				
Generation Performance	84.91	83.33	87.38	80.36
Baseline	73.77	73.77	73.77	73.77
BERTScore F1				
Generation Performance	0.44	0.40	0.45	0.40
Baseline	0.50	0.50	0.50	0.50
METEOR				
Generation Performance	0.85	0.83	0.83	0.82
Baseline	0.85	0.85	0.85	0.85
Flesch Reading Ease				
Generation Performance	8.21	8.21	19.71	8.21
Baseline (Original)	8.21	8.21	8.21	8.21
Baseline (Mask)	8.21	8.21	8.21	8.21

Table 2: Encoder-Decoder Models Comparison at the Sentence Level (The Baseline without annotations was calculated by comparing masked text to the original text)

max_lines	10	20	30	35	40	41	42	45	50
Inference Time (min)	13:47	8:10	6:44	5:24	5:10	5:01	5:12	5:54	6:05
Average Tokens Per Chunk	51.59	90.23	131.26	136.55	144.34	146.43	146.43	146.43	146.43

Table 3: Comparison for different Chunk Size

recommend masking around 80% of nouns to maintain balanced evaluation scores. Full noun masking significantly reduces synthetic letter quality.

4.3.3 Masking Only Verbs

Masking verbs also help identify appropriate token types for masking while retaining clinical meaning. Although verbs are crucial for describing clinical events, they can often be inferred from context. Therefore, masking verbs may have a slight effect on the synthetic clinical letters quality, but can also introduce some variation. From our experimental investigations, masking verbs followed a similar trend to other masking strategies, with *both invalid prediction rates and NLG metrics decreasing as the masking ratio increased*. This is likely due to two factors: the model prioritises generating coherent sentences and may be less sensitive to verbs due to their relative scarcity in the raw data. **BERTScore remained high at 0.95** when **all** verbs were masked, compared to 0.89 when all nouns were masked.

4.3.4 Masking Only Stopwords

Masking **stopwords** aims to **reduce noise**, allowing the model to focus on clinically relevant information while enhancing **generalisation** in synthetic clinical letters to *distinguish* them from actual letters. Additionally, **varying syntax** by masking stopwords mitigates the risk of PHI reconstruction from adversarial attacks. It is often combined with other masking strategies to strengthen privacy protection. From our experiments, the results follow a similar trend to random masking, where a higher masking ratio leads to lower ROUGE Score and BERTScore. Notably, the Invalid Prediction Rate is lowest at a medium masking ratio, as higher ratios cause information loss, while lower ratios make small prediction errors more impactful. The overall **low Invalid Prediction Rate** and **high BERTScore** suggest that stopwords have minimal influence on the model's contextual understanding.

4.3.5 Comparison of Identical Actual Masking Ratios

To further observe how different masking strategies influence the generation of clinical letters, we compared the results using the same actual masking ratios but with different strategies, where the number of masked tokens remained constant. Masking only stopwords resulted in the highest BERTScore and lowest invalid prediction rate, confirming that **stopwords** have **minimal impact** on meaning. Conversely, masking **nouns and verbs**

performed **worse** than random masking, suggesting that excessive masking of these token types can **compromise** the clinical information preservation.

4.3.6 Hybrid Masking

Hybrid masking strategies are compared at the same actual masking ratio. Masking **only stopwords** yielded the **best** performance, while *adding noun masking reduced* performance, confirming that masking nouns negatively affects results. However, it still outperformed random masking, suggesting that stopwords have a greater influence than nouns. Additionally, when verbs were further masked alongside nouns and stopwords, performance deteriorated further, indicating that verbs also *negatively* impact model performance.

4.3.7 Comparison with and without Entity Preservation

To assess the impact of entity preservation, we compared results with a baseline model that did not retain entities. When 40% of nouns were masked while preserving entities, the models outperformed those without entity preservation. Additionally, with a 0.3 masking ratio, entity-preserving models had lower ROUGE-1 and ROUGE-2 scores but higher ROUGE-L and BERTScores, indicating less direct overlap with the original text but better narrative retention. These findings confirm that *preserving entities and document structure enhances model performance*, matching our goal of generating clinically coherent yet diverse synthetic letters.

4.3.8 Downstream NER Task

We evaluated whether synthetic letters can replace original (anonymised) clinical letters in NER tasks for research and model training. SpaCy models trained on synthetic letters performed similarly to those trained on original letters, achieving comparable evaluation scores with an F1 score close to ScispaCy's 0.843. This suggests that *unmasked context does not significantly impact model understanding*. Therefore, synthetic letters can be effectively used in NER tasks to replace real-world clinical letters, ensuring data privacy.

Metric	spaCy Trained on Original Letters	spaCy Trained on Synthetic Letters	Performance Delta (△)
F1	0.855	0.853	-0.002
P	0.865	0.863	-0.002
R	0.846	0.843	-0.003

Table 4: Comparisons on Downstream NER Task (Precision, Recall, F1)

5 Conclusion

This study explores de-identified synthetic clinical letters that preserve *document structure and clinical narratives* while enhancing diversity. Encoder-only models outperformed encoder-decoder models, with base models performing *comparable* to **clinical-specific** models when clinical terms were preserved. Variable-length chunking strategy effectively maintained sentence meaning, and POS-based masking influenced output quality. Masking *stopwords* improved text quality, whereas masking *nouns and verbs* had negative impacts. BERTScore was identified as the primary evaluation metric, aligning well with both quantitative and qualitative evaluations. A **downstream NER task** demonstrated the feasibility of replacing real-world letters with synthetic ones for this task. Unlike existing research that focuses on improving similarity through model fine-tuning or training a privacy detection and substitution model, this study *emphasises preserving clinically relevant information* while maintaining **diversity**. It provides a framework for better utilisation of real-world datasets while mitigating privacy risks.

Limitations

Although the strategies outlined above facilitate the generation of diverse, de-identified synthetic clinical letters, several limitations remain. One primary concern is the *quality of the data set*, which is affected by spelling errors, ambiguous polysemous words, and limited data volume, potentially impacting generalisability. Additionally, the model struggles with long-tail phenomena, frequently failing to comprehend novel words that are common in the clinical domain. Moreover, processing shorthand and abbreviations presents an additional challenge, often resulting in misinterpretations of key medical terms.

Moreover, the *limited scope of the dataset*, which includes only 204 letters, constraints generalising the findings to broader clinical scenarios. Furthermore, the *evaluation framework*, primarily based on BERTScore, focuses on textual similarity and fails to comprehensively evaluate other critical aspects such as privacy protection efficacy, text diversity, and clinical soundness.

Future work should focus on evaluating de-

identification performance using non-anonymous datasets, developing a comprehensive evaluation benchmark and enhancing clinical and general knowledge integration, e.g. (Shaji et al., 2025). The evaluation benchmark should include:

- Privacy protection evaluation using alternative PHI detection models, Membership Inference Attacks, and Model Inversion Attacks (Fang et al., 2024; Ying et al., 2020).
- Diversity evaluation through TF-IDF cosine similarity or Dependency Tree Edit Distance (Thompson et al., 2015; Tsarfaty et al., 2012).
- Clinical soundness evaluation using MEDNLI (Medical Natural Language Inference) or GPT-based assessments (Romanov and Shivade, 2018).

Additionally, techniques such as synonymous substitution, entity linking to SNOMED CT, and specialised spelling correction could be leveraged to enhance the quality and diversity of synthetic clinical letters, e.g. (Romero et al., 2025). Another potential direction is leveraging models to predict and replace privacy-sensitive content that was originally substituted with underscores.

Impact Statement

We use only de-identified clinical data from MIMIC and strictly adhere to all data use agreements. The dataset has already been anonymised, and in this project, we further applied dual anonymisation and re-generation techniques to enhance privacy protection. These strategies are described in Appendix A.

All code used in this project, which will be released, is adapted from well-known language models open-sourced in Hugging Face. However, if applied to real-world clinical letters, it must be reviewed prior to release to mitigate potential data privacy risks. Synthetic clinical letters can be reproduced using the MIMIC-IV dataset and the code provided. However, if users apply this method to process privately collected clinical letters, they should ensure compliance with data protection regulations and clarify copyright ownership.

Our findings help bridge the gap in NLG-based clinical letter generation, facilitating better utilisation of real-world clinical letters by re-generating text while masking sensitive information. This approach helps address data scarcity in medical research and education. However, challenges inherent to LLMs, such as hallucinations and data bias, still persist.

Acknowledgements

LH, WDP, and GN are grateful for the support from the grant "Assembling the Data Jigsaw: Powering Robust Research on the Causes, Determinants and Outcomes of MSK Disease", and the grant "Integrating hospital outpatient letters into the healthcare data space" (EP/V047949/1; funder: UKRI/EPSRC). LH is grateful for the 4D Picture EU project (https://4dpicture.eu/) on cancer patient journey support.

References

Goldberger A, Amaral L, Glass L, Hausdorff J, Ivanov PC, Mark R, Mietus JE, Moody GB, Peng CK, and Stanley HE. 2000. Physiobank, physiotoolkit, and physionet: Components of a new research resource for complex physiologic signals. https://physionet.org/content/.

Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jindi, Tristan Naumann, and Matthew McDermott. 2019. Publicly available clinical bert embeddings. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 72–78.

Ali Amin-Nejad, Julia Ive, and Sumithra Velupillai. 2020a. Exploring transformer text generation for medical dataset augmentation. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4699–4708.

Ali Amin-Nejad, Julia Ive, and Sumithra Velupillai. 2020b. Exploring transformer text generation for medical dataset augmentation. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4699–4708, Marseille, France. European Language Resources Association.

Samuel Belkadi, Nicolo Micheletti, Lifeng Han, Warren Del-Pinto, and Goran Nenadic. 2023. Generating medical instructions with conditional transformer. In *NeurIPS 2023 Workshop on Synthetic Data Generation with Generative AI*.

Brecht Claerhout and Georges JE DeMoor. 2005. Privacy protection for clinical and genomic data: The use of privacy-enhancing techniques in medicine. *International Journal of Medical Informatics*, 74(2-4):257–265.

Franck Dernoncourt, Ji Young Lee, Ozlem Uzuner, and Peter Szolovits. 2017. De-identification of patient notes with recurrent neural networks. *Journal of the American Medical Informatics Association*, 24(3):596–606.

Lehman Eric and Alistair Johnson. 2023. Clinical-T5: Large Language Models Built Using MIMIC Clinical Text. PhysioNet.

Hao Fang, Yixiang Qiu, Hongyao Yu, Wenbo Yu, Jiawei Kong, Baoli Chong, Bin Chen, Xuan Wang, Shu-Tao Xia, and Ke Xu. 2024. Privacy leakage on dnns: A survey of model inversion attacks and defenses. *arXiv preprint arXiv:2402.04013*.

Ary L. Goldberger, Luis A. N. Amaral, Leon Glass, Jeffrey M. Hausdorff, Plamen Ch. Ivanov, Roger G. Mark, Joseph E. Mietus, George B. Moody, Chung-Kang Peng, and H. Eugene Stanley. 2000. Physiobank, physiotoolkit, and physionet. *Circulation*, 101(23):e215–e220.

Le Hou, Richard Yuanzhe Pang, Tianyi Zhou, Yuexin Wu, Xinying Song, Xiaodan Song, and Denny Zhou. 2022. Token dropping for efficient bert pretraining. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3774–3784.

Marie Humbert-Droz, Pritam Mukherjee, and Olivier Gevaert. 2022. Strategies to address the lack of labeled data for supervised machine learning training with electronic health records: Case study for the extraction of symptoms from clinical notes. *JMIR Medical Informatics*, 10(3):e32903.

D HÜSKE-KRAUS. 2003. Text generation in clinical medicine: A review. *Methods of information in medicine*, 42(1):51–60.

Alistair Johnson, Lucas Bulgarelli, Tom Pollard, Brian Gow, Benjamin Moody, Steven Horng, Leo Anthony Celi, and Roger Mark. 2024. Mimic-iv.

Alistair EW Johnson, Lucas Bulgarelli, Lu Shen, Alvin Gayles, Ayad Shammout, Steven Horng, Tom J Pollard, Sicheng Hao, Benjamin Moody, Brian Gow, et al. 2023. Mimic-iv, a freely accessible electronic health record dataset. *Scientific data*, 10(1):1.

Aleksandar Kovačević, Bojana Bašaragin, Nikola Milošević, and Goran Nenadić. 2024. De-identification of clinical free text using natural language processing: A systematic review of current approaches. *Artificial Intelligence in Medicine*, page 102845.

Zenon Lamprou, Frank Pollick, and Yashar Moshfeghi. 2022. Role of punctuation in semantic mapping between brain and transformer models. In *International Conference on Machine Learning, Optimization, and Data Science*, pages 458–472. Springer.

Yikuan Li, Ramsey M Wehbe, Faraz S Ahmad, Hanyin Wang, and Yuan Luo. 2023. A comparative study of pretrained language models for long clinical text. *Journal of the American Medical Informatics Association*, 30(2):340–347.

Ximeng Liu, Rongxing Lu, Jianfeng Ma, Le Chen, and Baodong Qin. 2015. Privacy-preserving patient-centric clinical decision support system on naive bayesian classification. *IEEE journal of biomedical and health informatics*, 20(2):655–668.

Zhengliang Liu, Yue Huang, Xiaowei Yu, Lu Zhang, Zihao Wu, Chao Cao, Haixing Dai, Lin Zhao, Yiwei Li, Peng Shu, et al. 2023. Deid-gpt: Zero-shot medical text de-identification by gpt-4. *arXiv preprint arXiv:2303.11032*.

Nicolo Micheletti, Samuel Belkadi, Lifeng Han, and Goran Nenadic. 2024. Exploration of masked and causal language modelling for text generation. *CoRR*, abs/2405.12630.

Beau Norgeot, Kathleen Muenzen, Thomas A Peterson, Xuancheng Fan, Benjamin S Glicksberg, Gundolf Schenk, Eugenia Rutenberg, Boris Oskotsky, Marina Sirota, Jinoos Yazdany, et al. 2020. Protected health information filter (philter): accurately and securely de-identifying free-text clinical notes. *NPJ digital medicine*, 3(1):57.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D Manning. 2020. Stanza: A python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 101–108.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Alexey Romanov and Chaitanya Shivade. 2018. Lessons from natural language inference in the clinical domain. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1586–1596.

Pablo Romero, Lifeng Han, and Goran Nenadic. 2025. INSIGHTBUDDY-AI: Medication Extraction and Entity Linking using Pre-Trained Language Models and Ensemble Learning. In *NAACL-SRW, Forthcoming*, New Mexico, USA. ACL.

Dhivin Shaji, Angel Paul, Lifeng Han, Warren Del-Pinto, Goran Nenadic, and Suzan Verberne. 2025. De-identifying Clinical Texts using Biomed-Clinical BERTs and Comprehensive Risk Assessment. In *IEEE-ICHI 2025, Forthcoming*, Calabria, Italy. IEEE.

Irena Spasic and Goran Nenadic. 2020. Clinical text data in machine learning: Systematic review. *JMIR medical informatics*, 8(3):e17984.

Ruixiang Tang, Xiaotian Han, Xiaoqian Jiang, and Xia Hu. 2023. Does synthetic data generation of llms help clinical text mining? *ArXiv*, abs/2303.04360.

Sumitha Udayashankar Tarur and Sudhakar Prasanna. 2021. clinical case letter. *Indian Pediatr*, 58(188):189.

Nicolas P Terry. 2012. Protecting patient privacy in the age of big data. *UMKC L. Rev.*, 81:385.

Victor U Thompson, Christo Panchev, and Michael Oakes. 2015. Performance evaluation of similarity measures on similar and dissimilar text retrieval. In *2015 7th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K)*, volume 1, pages 577–584. IEEE.

Reut Tsarfaty, Joakim Nivre, and Evelina Andersson. 2012. Cross-framework evaluation for statistical parsing. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 44–54.

Katherine Tucker, Janice Branson, Maria Dilleen, Sally Hollis, Paul Loughlin, Mark J Nixon, and Zoë Williams. 2016. Protecting patient privacy when sharing patient-level data from clinical trials. *BMC medical research methodology*, 16:5–14.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 6000–6010, Red Hook, NY, USA. Curran Associates Inc.

Guangyu Wang, Xiaohong Liu, Zhen Ying, Guoxing Yang, Zhiwei Chen, Zhiwen Liu, Min Zhang, Hongmei Yan, Yuxing Lu, Yuanxu Gao, et al. 2023. Optimized glycemic control of type 2 diabetes with reinforcement learning: a proof-of-concept trial. *Nature Medicine*, 29(10):2633–2642.

Zuobin Ying, Yun Zhang, and Ximeng Liu. 2020. Privacy-preserving in defending against membership inference attacks. In *Proceedings of the 2020 Workshop on Privacy-Preserving Machine Learning in Practice*, pages 61–63.

Wang Zeng, Sheng Jin, Wentao Liu, Chen Qian, Ping Luo, Wanli Ouyang, and Xiaogang Wang. 2022. Not all tokens are equal: Human-centric visual analysis via token clustering transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11101–11111.

Yuhao Zhang, Yuhui Zhang, Peng Qi, Christopher D Manning, and Curtis P Langlotz. 2021. Biomedical and clinical English model packages for the Stanza Python NLP library. *Journal of the American Medical Informatics Association*.

Liu Zhuang, Lin Wayne, Shi Ya, and Zhao Jun. 2021. A robustly optimized BERT pre-training approach with post-training. In *Proceedings of the 20th Chinese National Conference on Computational Linguistics*, pages 1218–1227, Huhhot, China. Chinese Information Processing Society of China.

A Different Masking Strategies

To make the synthetic letters more readable, clinically sound, and privacy-protective, different masking strategies are experimented based on the following principles.

- **Retain Annotated Entities:** Preserve clinical knowledge and context.
- **Preserve Extracted Structures:** Keep templates for clinical letters intact.
- **Mask Detected Private Information:** Useful for de-identification, especially in real-world applications.
- **Preserve Medical Terminology:** Ensure essential clinical terms remain unmasked.
- **Preserve Non-Private Numbers:** Keep medical-related numbers (e.g., dosage, heart rate) while masking private ones (e.g., phone numbers, postal codes).
- **Preserve Punctuation:** Maintain punctuation marks such as periods ('.') and underscores ('___') to improve text clarity and coherence (Lamprou et al., 2022).
- **Retain Special Patterns in Samples:** Retain clinically relevant patterns (e.g. 'Ibuprofen > 200 mg', etc) identified from raw sample letters to preserve important clinical details.

Based on the principles above, different masking strategies were experimented with:

- **Mask Randomly:** Tokens are randomly masked in 10% increments (0%-100%) to assess how the number of masked tokens affects synthetic letter quality and provides a baseline for other masking strategies.
- **Mask Based on POS Tagging:** Tokens are masked based on their part-of-speech (POS) category (e.g., only nouns, only verbs) in 10% increments to analyse POS influence on context understanding.
- **Mask Stopwords:** Stopwords are masked to reduce noise and enhance text diversity while ensuring that crucial clinical information remains intact. This approach can also serve as an indirect strategy to prevent reconstruction by attackers leveraging the same syntactic patterns.

- **Hybrid Masking Using Different Ratio Settings:** Combines different masking strategies at varying ratios (e.g., 50% nouns + 50% stopwords) to evaluate their combined effects.

B Evaluation Pipeline

The detailed evaluation pipeline is shown in Figure 8.

C More Evaluation Details

We evaluated the performance of encoder-only and encoder-decoder models at both the sentence level (using the sample sentence in Table 1 and Table 2) and the full dataset level in Table 5. Although SMOG is commonly used for medical datasets, it is less suitable for sentence-level analysis; thus, Flesch Reading Ease was used instead.

As shown in Table 7 and Table 8, readability metrics showed minor variations, with SMOG and Flesch-Kincaid scores occasionally falling below both the masked and original baselines, likely due to punctuation or spacing errors at high masking ratios. Perplexity remained stable, suggesting that synthetic letters are effective for training clinical models, while information entropy was preserved regardless of masking ratios. Subjectivity scores remained consistent, mitigating concerns about model bias.

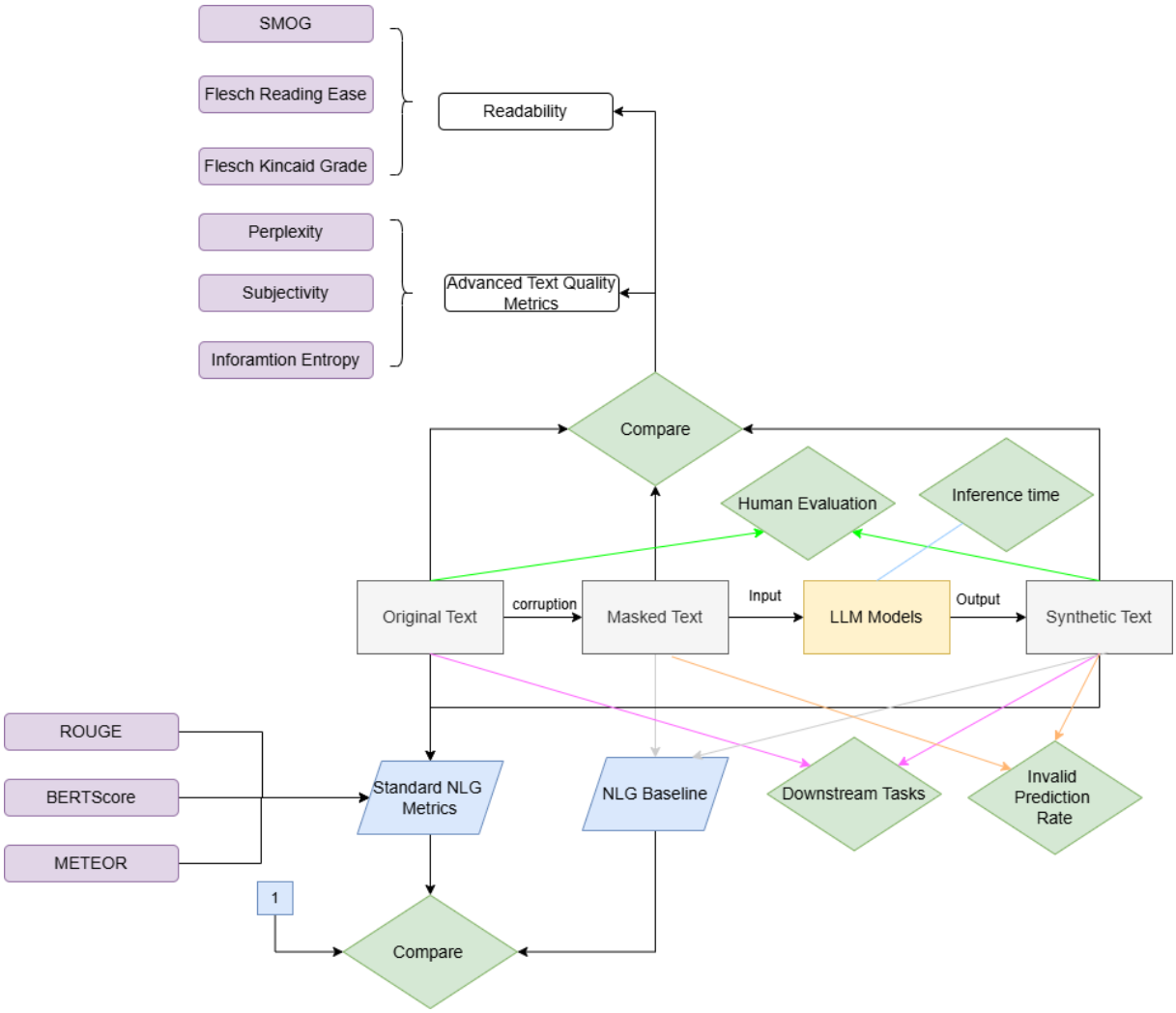

Figure 8: Evaluation Pipeline

	Model Evaluation			
	RoBERTa-base	medicalai / ClinicalBERT	Clinical-Longformer	Bio_ Clinical-BERT
ROUGE-1				
Generation Performance Baseline	92.98 85.64	93.63 85.44	94.66 85.64	93.18 85.61
ROUGE-2				
Generation Performance Baseline	86.10 74.96	87.42 74.64	89.50 74.96	86.50 74.92
ROUGE-L				
Generation Performance Baseline	92.54 85.64	93.22 85.44	94.38 85.64	92.71 85.61
BERTScore F1				
Generation Performance Baseline	0.91 0.82	0.90 0.63	0.92 0.82	0.90 0.63

Table 5: Encoder-Only Models Comparison on the Full Dataset with Masking Ratio 0.4 (The Baseline was calculated by comparing masked text to the original text)

Bio_ClinicalBERT	Masking Ratio					
	1.0	**0.8**	**0.6**	**0.4**	**0.2**	**0.0**
ROUGE-1						
Generation Performance	76.28	83.75	88.91	93.18	96.76	99.51
Baseline	64.05	71.56	78.56	85.61	92.63	99.22
ROUGE-2						
Generation Performance	62.60	70.77	78.81	86.50	93.42	99.02
Baseline	51.72	57.88	65.38	74.92	86.27	98.61
ROUGE-L						
Generation Performance	74.33	81.69	87.71	92.71	96.65	99.50
Baseline	64.05	71.56	78.56	85.61	92.63	99.22
BERTScore						
Generation Performance	0.63	0.75	0.83	0.90	0.95	0.99
Baseline	0.29	0.39	0.50	0.63	0.79	0.98
METEOR						
Generation Performance	0.70	0.80	0.87	0.93	0.97	1.00
Baseline	0.66	0.72	0.78	0.85	0.92	0.99

Table 6: Standard NLG Metrics Across Different Masking Ratios Using Bio_ClinicalBERT (The Baseline was calculated by comparing masked text to the original text)

Bio_ClinicalBERT	Masking Ratio					
	1.0	**0.8**	**0.6**	**0.4**	**0.2**	**0.0**
SMOG						
Generation Performance	8.91	9.18	9.50	9.79	10.00	10.13
Baseline (Original)	10.16	10.15	10.15	10.15	10.15	10.15
Baseline (Mask)	9.04	9.29	9.52	9.74	9.95	10.13
Flesch Reading Ease						
Generation Performance	63.77	63.44	61.41	59.54	58.06	57.02
Baseline (Original)	56.85	56.87	56.87	56.87	56.87	56.87
Baseline (Mask)	70.11	67.39	64.75	62.15	59.62	57.13
Flesch-Kincaid Grade						
Generation Performance	7.32	7.70	8.24	8.66	9.01	9.22
Baseline (Original)	9.26	9.26	9.26	9.26	9.26	9.26
Baseline (Mask)	7.41	7.79	8.16	8.52	8.87	9.22

Table 7: Readability Metrics Across Different Masking Ratios Using Bio_ClinicalBERT (The Baseline without annotations was calculated by comparing masked text to the original text)

Bio_ClinicalBERT	Masking Ratio					
	1.0	**0.8**	**0.6**	**0.4**	**0.2**	**0.0**
Perplexity						
Generation Performance	2.24	2.32	2.31	2.30	2.29	2.29
Baseline (Original)	2.22	2.28	2.28	2.28	2.28	2.28
Baseline (Mask)	250.37	65.42	24.29	8.95	4.03	2.39
Information Entropy						
Generation Performance	5.46	5.80	5.92	5.96	5.98	5.98
Baseline (Original)	5.98	5.98	5.98	5.98	5.98	5.98
Baseline (Mask)	4.51	4.93	5.29	5.60	5.85	5.97
Subjectivity						
Generation Performance	0.32	0.32	0.32	0.32	0.33	0.33
Baseline (Original)	0.33	0.33	0.33	0.33	0.33	0.33
Baseline (Mask)	0.41	0.39	0.38	0.37	0.35	0.33

Table 8: Advanced Text Quality Metrics Across Different Masking Ratios Using Bio_ClinicalBERT (The Baseline without annotations was calculated by comparing masked text to the original text)