Instruction-Tuning LLaMA for Synthetic Medical Note Generation in Swedish and English

Lotta Kiefer[1], **Jesujoba O. Alabi**[1], **Thomas Vakili**[2], **Hercules Dalianis**[2], **Dietrich Klakow**[1]

[1]Department of Language Science and Technology, Saarland University, Germany
[2]Department of Computer and Systems Sciences, Stockholm University, Sweden
lkiefer@lsv.uni-saarland.de

Abstract

The increasing capabilities of large language models (LLMs) have unlocked transformative potential for medical applications, but privacy constraints limit access to high-quality training data from electronic health records (EHRs). In response, we propose a framework to generate synthetic EHRs by instruction-tuning an LLM using descriptions of diagnosis codes. We show that this framework overcomes problems of prior approaches, such as diversity reduction and medical incoherence, while maintaining strong privacy protections. Utility was measured by training models to predict diagnosis codes for EHRs. Real data still has higher utility, but synthetic data approaches real data results with increasing dataset size. The differences in utility were most likely due to noise in the synthetic data. A user study involving medical professionals confirmed no significant loss in readability or medical coherence compared to the real EHRs, even though inter-annotator agreement is low. These findings establish synthetic EHRs as a viable alternative for privacy-preserving and scalable clinical NLP applications. We release our code on GitHub.[1]

.

1 Introduction

Healthcare faces growing challenges, including staff shortages, medical errors, and unequal access – issues that artificial intelligence applications, particularly with recent advances in deep learning (DL), can help address (Goldberg et al., 2024). A critical requirement for building such systems is access to electronic health records (EHRs), which serve as a rich data source (Häyrinen et al., 2008). EHRs contain, alongside structured patient information, large amounts of unstructured text, making natural language processing (NLP) techniques particularly valuable for medical applications such as automated diagnosis coding (Huang et al., 2022), clinical text mining (Dalianis, 2018), and clinical decision support (Häyrinen et al., 2008). However, privacy concerns and strict regulations (Budu et al., 2024) often limit access, hindering the development and evaluation of medical applications.

To address this data sparsity challenge, synthetic data generation has emerged as a solution (Murtaza et al., 2023), producing artificial data that mimics the statistical properties of real EHRs while preserving patients' privacy. Hence, such data must meet two key criteria: `privacy` and `utility`. However, balancing these objectives remains challenging – even with powerful large language models (LLMs) – due to issues such as limited data diversity (Libbi et al., 2021) and insufficient medical coherence (Melamud and Shivade, 2019).

In this work, we investigate synthetic data generation to address the question: "How can LLMs balance privacy and utility for EHRs?". To answer this question, we propose a framework based on LLaMA-3.1-8B (Dubey et al., 2024) for generating synthetic medical notes in English and Swedish. By instruction-tuning the model with ICD-10 code descriptions, we improve content controllability and enhance diversity. We evaluate the generated data along four key dimensions: `fidelity`, `privacy`, `utility`, and `coherence`.

Our findings show that while real medical notes still yield better downstream performance, synthetic notes can also effectively train complex multi-label classification models for medical coding. The synthetic notes exhibit a richer vocabulary than real data, addressing common diversity issues. A user study with medical professionals found no significant loss in coherence, and privacy analyses show minimal leakage risk. These results highlight our framework's potential to generate high-quality, privacy-preserving medical notes, providing a strong foundation toward developing reliable medical NLP tools that respect patient privacy.

[1] `https://github.com/uds-lsv/synthetic-ehr-notes`

Proceedings of Recent Advances in Natural Language Processing,pages 557–566
Varna, Sep 8–10, 2025
https://doi.org/10.26615/978-954-452-098-4-065

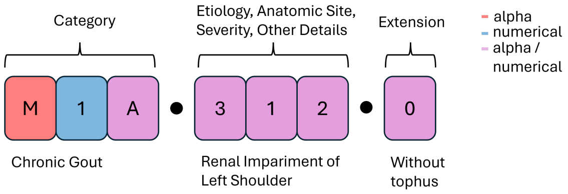

Figure 1: Example of ICD-10 structure

2 Background and Related Research

2.1 Electronic Health Records and Medical Coding

EHRs are digital repositories that securely store and share patient health data across healthcare providers, supporting clinical decision-making (Häyrinen et al., 2008). Among the structured elements, such as lab results, or diagnosis codes, EHRs also include unstructured components like discharge summaries that document a patient's hospital stay, reasons for admission, diagnoses, and treatments in free-text form (Wimsett et al., 2014).

EHRs typically have ICD-10 codes assigned by the physician. ICD-10 stands for the 10th version of the *International Classification of Diseases* (World Health Organization, 2016) and is a globally used coding system, comprising approximately 155,000 codes for diagnoses and procedures (Hirsch et al., 2016). The coding scheme follows a hierarchical structure consisting of four to seven characters as illustrated in an example in Figure 1. This structure allows for an organization into 22 distinct chapters (based on the first three characters) specifying the nature of the condition.

Assigning ICD-10 codes to medical notes is vital for administrative and analytical tasks (Edin et al., 2023) but known to be labor-intensive and error-prone due to the high amount of individual codes (Burns et al., 2012). Automating this process via multi-label classification of discharge summaries offers a scalable solution (Huang et al., 2022). However, large code sets, data scarcity, long document length, and class imbalance pose major challenges (Edin et al., 2023).

2.2 Synthetic Data Generation

Since EHRs contain sensitive patient data known as protected health information (PHI), including names, birth dates, or social security numbers, additional measures are needed to protect privacy when using it for research or development. De-identification addresses this by masking or removing PHI (Vakili et al., 2022). While it lowers privacy risks and retains utility for some tasks (Vakili et al., 2022), limitations remain, such as the potential for re-identification via quasi-identifiers and utility loss (Yogarajan et al., 2020).

As an alternative, synthetic data, generated by computational models, has gained significant attention due to its potential to address challenges related to privacy concerns, data sparsity, and bias mitigation when used for model training (Jordon et al., 2022). This work explores the emerging field of synthetic free-text medical note generation that became increasingly popular with advances in language modeling (Rankin et al., 2020). The quality of synthetic data is typically evaluated by fidelity (realism), utility (usefulness for model training), and privacy (protection from data leakage) (Budu et al., 2024).

Several approaches have been explored for generating medical notes, starting with less complex neural networks such as GANs (Guan et al., 2018), LSTMs (Melamud and Shivade, 2019), and vanilla transformers (Amin-Nejad et al., 2020), and progressing to more advanced transformer models with hundreds of millions of parameters, including decoder-only LLMs like LLaMA (Baumel et al., 2024) and GPT variants (Kumichev et al., 2024; Falis et al., 2024). There is not always a clear winner; for instance, LSTMs have shown effectiveness for downstream tasks such as named entity recognition (NER) when generating medical notes, while LLMs were better at producing more fluent text (Libbi et al., 2021). However, there is a clear trend toward leveraging pretrained state-of-the-art LLMs showing potential for a variety of different downstream tasks (Litake et al., 2024; Kumichev et al., 2024; Vakili et al., 2025).

However, findings vary; for example, Falis et al. (2024) highlight the limitations of zero-shot generations, noting lack of diversity and unnatural phrasing that impaired performance in medical coding. Such discrepancies often reflect differences in evaluation setup. For instance, the dataset used by Kumichev et al. (2024) contained substantially fewer unique codes than that in Falis et al. (2024). The highly inconsistent choice and setup of downstream evaluation tasks complicates direct comparisons. Moreover, strong task performance does not guarantee high data quality or generalizability. Notably, synthetic data has proven useful for training NER models despite being linguistically or clinically incoherent (Libbi et al., 2021), raising questions about the validity of using simple downstream tasks as proxies for data quality. Furthermore, privacy evaluation is often overlooked or deferred due to difficulties in measurement. When included, privacy is typically measured via disclosure metrics (Belkadi et al., 2025), distance metrics (Hiebel et al., 2023), or manual review (Libbi et al., 2021). Establishing consistent benchmarks and evaluation procedures is critical for progress and meaningful comparison in future work.

Three recurring challenges emerge across studies. First, there is a privacy-utility trade-off: increasing privacy often comes at the cost of reduced utility, with synthetic data typically underperforming on tasks involving real data (Melamud and Shivade, 2019; Baumel et al., 2024). Second, synthetic notes often suffer from reduced diversity, characterized by limited vocabulary that can impair performance on downstream tasks (Libbi et al., 2021; Hullmann and Hansson, 2024). Third, manual reviews frequently identify clinical inconsistencies in the generated text, indicating issues with coherence (Libbi et al., 2021; Falis et al., 2024).

Hence, in this work, we focus on understanding these challenges in the era of LLMs. To address them, we present a new generation framework that uses diagnosis code descriptions to instruction-tune an LLM for producing high-quality synthetic notes. Most similar to our work is the generation framework *MedSyn* by Kumichev et al. (2024) that relies on the incorporation of disease-specific symptoms into the prompt. Unlike MedSyn, which relies on a medical knowledge graph and is limited to a Russian dataset with few unique codes and one code per note, our simpler, scalable approach avoids external resources, enabling broader multilingual applicability and support for more codes per note. We conduct a thorough assessment of the synthetic data to evaluate this approach on two languages.

3 Data and Methods

The main goal of this work is to generate synthetic medical notes that ensure high utility and robust privacy protection, aiming at replacing real data and analyzing how well they preserve these properties.

To achieve this, we propose a novel framework (Figure 2) for conditional text generation, by instruction-tuning LLaMA-3.1-8B on ICD-10 code descriptions to produce versatile synthetic notes. We employ this framework on both English and Swedish data to test its effectiveness for both high- and comparatively lower-resourced languages. The evaluation involves four key components: comparing synthetic and real notes for similarity, assessing privacy preservation, evaluating utility in medical downstream tasks, and analyzing readability and medical coherence through a user study.

3.1 MIMIC-IV

This is an English dataset sourced from the Medical Information Mart for Intensive Care IV (MIMIC-IV) (Johnson et al., 2023),[2] comprising 524,000 admissions from over 257,000 patients at Beth Israel Deaconess Medical Center. It includes structured data and pseudonymized unstructured medical notes. Following Edin et al. (2023), discharge summaries were filtered, and ICD-10 codes occurring fewer than 10 times were excluded, yielding 122,279 documents and 7,942 unique codes. The dataset was then divided, also in line with Edin et al. (2023), ensuring broad code representation across three subsets: a large training subset (MIMIC-L, n=89,098, 72.9%), a small training subset (MIMIC-S, n=13,378, 10.9%), and an evaluation subset (MIMIC-E, n=19,802, 16.2%). To put the data in a format suitable for LLMs, highly structured sections (e.g., lab results, medication lists) and repetitive or less informative content (e.g., discharge instructions) were removed during preprocessing.

3.2 SEPR Corpus

This Swedish dataset is based on the Stockholm EPR Gastro ICD-10 Pseudo Corpus II (SEPR II),[3] derived from the Health Bank Infrastructure (Dalianis et al., 2015) by Lamproudis et al. (2024). It comprises 317,971 records from Karolinska University Hospital from 113,174 individual patients, all related to gastrointestinal conditions and annotated with 415 unique ICD-10 codes. Again, the data was split into the three subsets SEPR-L (n=237,968, 76%), SEPR-S (n=47,783, 15%), and SEPR-E (n=32,027, 10%). As the Swedish notes consist solely of concise free-text entries, no additional preprocessing was conducted.

SEPR and MIMIC-IV differ significantly in scope and granularity: MIMIC-IV covers many medical domains, whereas SEPR is restricted to

[2] A training program must be completed to obtain credentialed access to MIMIC IV.

[3] This research has been approved by the Regional Ethical Review Board in Stockholm under permission no. 2007/1625-31/5.

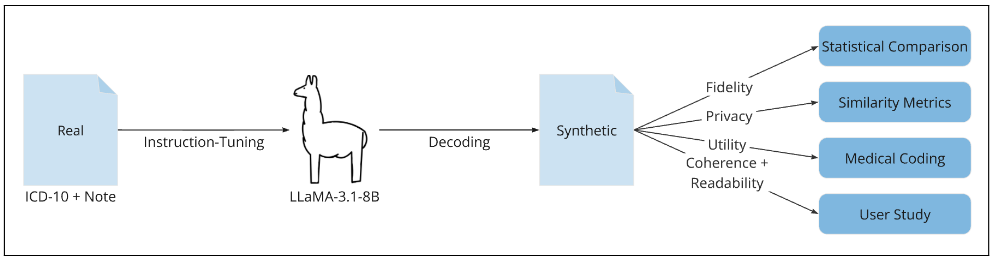

Figure 2: Overall approach employed in this work

gastrointestinal conditions. Additionally, Swedish discharge summaries are markedly more concise, averaging just 12.5% of the token length of English notes, and contain far fewer codes per document (1.09 vs. 15.65) and unique ICD-10 codes overall (415 vs. 7,942). Thus, cross-language comparison can only be made subject to these disparities.

3.3 Synthesizing Medical Notes

We used LLaMA-3.1-8B as base model for generating synthetic medical notes, chosen for its strong performance and computational efficiency. The model was instruction-tuned using ICD-10 code descriptions as prompts to enable:

(i) **Content control:** Ensuring generated notes align with specified medical domains.

(ii) **Diversity:** Producing varied notes by altering ICD-10 inputs.

(iii) **Automatic labeling:** Using ICD-10 annotations for supervised learning.

We followed the Alpaca instruction-tuning template (Taori et al., 2023), using English and Swedish prompts based on ICD-10 code descriptions as input, with outputs corresponding to discharge summaries.

The ICD-10 codes from MIMIC-L and SEPR-L were converted into descriptive texts serving as input and paired with real medical notes as output to form the training dataset. Fine-tuning was performed within the Axolotl framework (OpenAccess-AI-Collective, 2024), applying QLoRA 4-bit quantization (Dettmers et al., 2023) to reduce memory usage, and DeepSpeed ZeRO Stage 3 (Rajbhandari et al., 2020; Rasley et al., 2020) for efficient multi-GPU training. The Swedish data followed the same workflow with language-specific adjustments.

For inference, LoRA adapters were merged into the base model. Generation was accelerated using vLLM (vllm-project, 2024), which optimizes memory handling. Synthetic data was generated using ICD-10 sequences from the large and small MIMIC and SEPR datasets, leaving the evaluation datasets aside for testing.

3.4 Fidelity Evaluation

Fidelity is measured by comparing statistical features of synthetic and real MIMIC-S and SEPR-S datasets to assess alignment in data distribution and diversity. Additionally, a manual review of MIMIC-S synthetic notes was conducted to evaluate structural and linguistic similarity to real discharge summaries.

3.5 Privacy Evaluation

Following Libbi et al. (2021), we assess re-identification risk in English data using ROUGE-5 to quantify 5-gram overlaps between synthetic and real MIMIC datasets, using the overlap between real MIMIC-S and real MIMIC-L as a baseline. For Swedish, we apply the 8-gram overlap method from Hiebel et al. (2023), comparing synthetic SEPR-S to real SEPR-L. Real-real overlap provides a baseline, and results are contextualized using Hullmann and Hansson (2024), who also used the SEPR corpus for synthetic note generation.

Both similarity metrics are designed to assess memorization risks by determining whether the synthetic data is excessively similar to real data, beyond the typical overlap observed between two real datasets.

3.6 Utility Evaluation

We evaluated the synthesized notes through the medical coding task. This task was chosen due to its complexity and its importance in clinical analysis. With a high number of granular labels, accurate coding demands capturing interrelated clinical details, making it a strong proxy for note quality. Furthermore, the task can be seen as an inverse of the generation process, already providing the ICD-10 annotations needed for supervised learning.

We employed the PLM-ICD framework (Huang et al., 2022), which leverages domain-specific BERT (Devlin et al., 2019) models with segment pooling for long inputs and label-aware attention (Vu et al., 2021) for fine-grained, label-specific representations. Following Edin et al. (2023), we used RoBERTa-PM (Lewis et al., 2020) for English (with inputs truncated to 4000 tokens) and adapted the setup for Swedish using SweDeClin-BERT (Vakili et al., 2022). Each configuration was trained three times to report mean performance, with statistical significance tested using two-sample t-tests.

Models were trained on both large (MIMIC-L, SEPR-L) and small (MIMIC-S, SEPR-S) subsets in real and synthetic form, using MIMIC-E and SEPR-E for testing. To assess the suitability of model choice within our framework, we compared synthetic training sets from the base LLaMA-3.1-8B model with those produced by domain-specific variants, namely OpenBioLLM-8B (Ankit Pal, 2024) adapted to the biomedical domain and AI Sweden's LLaMA-3-8B (AI Sweden Models, 2024) adapted to Nordic languages.

3.7 User Study

To evaluate readability and clinical plausibility, medical professionals rated English and Swedish synthetic and real notes on a bipolar 1–5 scale across two dimensions: readability (from *not natural at all* to *completely natural: could be written by a doctor*) and medical coherence (from *not coherent at all* to *Perfectly coherent: Symptoms, diagnosis, procedures, etc. fit together perfectly*). Notes were presented randomized to mitigate bias, and raters were unaware that some texts were synthetic, resulting in a blind test. High coherence ratings would indicate that the synthetic notes effectively capture ICD-10 content, demonstrating successful content control via instruction-tuning.

For English, three real-synthetic pairs were sampled based on matching ICD-10 sequences, resulting in six samples with each ranging from 300–600 words for manageability. Ten German medical professionals participated as evaluators. None were native English speakers, which should be taken into account in the interpretation of the results.

For Swedish, four document pairs (100–200 words) were sampled and evaluated by four medical professionals, all native Swedish speakers.

	Total Docs	AVG Sent/Doc	AVG Token/Doc	AVG Token/Sent	Total Tokens	Unique Tokens
Statistical Comparison real vs. synthetic MIMIC-S						
Real	13,378	72	1,286	18	17,197,361	96,380
Synth	13,378	79	1,639	21	21,923,740	233,845
Statistical Comparison real vs. synthetic SEPR-S						
Real	47,783	13.96	200	14	9,553,204	191,578
Synth	47,783	13.88	252	18	12,056,615	391,152

Table 1: Statistical comparison of the real and synthetic MIMIC-S and SEPR-S datasets.

4 Results

In the following subsections, we present the evaluation results of the synthetic data.[4]

4.1 Fidelity: Do synthetic and real data match statistically?

Table 1 displays the results of the statistical comparison. Synthetic MIMIC-S notes are longer, contain more and longer sentences, and show significantly higher lexical diversity than real notes by exceeding 137,000 more unique tokens, resulting in a type-token ratio (TTR) almost twice as high. This is particularly notable since TTR typically decreases as text length increases (Tweedie and Baayen, 1998). Thus, the combination of high TTR and larger corpus size indicates a high degree of lexical diversity and contrasts prior findings that synthetic data tends to be lexically limited (Libbi et al., 2021; Hullmann and Hansson, 2024). Similarly, synthetic SEPR-S notes are more verbose and diverse, with increased token and sentence length and a 1% higher TTR. Average sentence counts per document remain consistent with the real dataset. These trends indicate that instruction-tuning LLaMA results in the generation of varied content, even exceeding the lexical range of the real data.

A manual review of MIMIC-S synthetic notes confirmed strong structural and stylistic alignment with real discharge summaries. The synthetic data replicates real-world features such as medical abbreviations, typos, and pseudonymization practices. However, some typical LLM artifacts were observed, including repetitions, pronoun mismatches, and occasional hallucinations. Although these artifacts were relatively rare, they reflect known LLM limitations and might hinder its utility. Overall, the synthetic notes closely mirror real datasets in structure and style while offering increased lexical diversity, and decoding settings could further control document length and verbosity if needed.

[4] Synthetic note examples are provided on GitHub.

	All Real/Synthetic Pairs				Highest 122 Real/Synthetic Pairs			
	AVG	Median	Min	Max	AVG	Median	Min	Max
MIMIC-S real	0.138	0.088	0.006	1.000	0.794	0.779	0.727	1.000
MIMIC-S synth	0.097	0.052	0.001	1.000	0.760	0.748	0.690	1.000

Table 2: ROUGE-5 Recall Scores: Real MIMIC-S and Synthetic MIMIC-S vs. Real MIMIC-L, reported for the full dataset and the 122 document pairs with the highest scores.

Model	**8-Gram Overlap**
Hullmann and Hansson (2024)	0.02442
Synthetic SEPR-S	0.00179
Baseline	0.00488

Table 3: 8-gram overlap between synthetic SEPR-S and real SEPR-L in comparison to baseline and Hullmann and Hansson (2024)

4.2 Privacy: Are the synthetic notes overly similar to the training set?

Table 2 compares ROUGE-5 recall scores across real and synthetic datasets. Real MIMIC-S notes were more similar to the training set (MIMIC-L) than their synthetic counterparts, with average scores nearly twice as high. Among the top 122 most similar pairs, the gap narrows but remains consistent. Similarly, the 8-gram overlap between synthetic SEPR-S and real SEPR-L was well below the baseline and prior work by Hullmann and Hansson (2024) as shown in Table 3. This supports the notion that our method produces outputs with low memorization risk, likely driven by the increased lexical variety of the synthetic data.

Overall, the synthetic datasets show low similarity to training data; nevertheless, occasional long overlaps may still persist and necessitate additional safeguards, like differential privacy (Baumel et al., 2024) or post-generation filtering, when working with non-pseudonymized data.

4.3 Utility: Are the synthetic notes useful to train a medical coding model?

Medical coding model performance across datasets are presented in Table 4. Reproducing Edin et al. (2023) with MIMIC-L training data, our results closely match previous metrics, validating our preprocessing strategy. The stable performance supports the idea that discharge summaries contain redundant information for certain tasks.

To create a smaller real-data baseline, we trained on MIMIC-S. As expected, its performance was lower than MIMIC-L, but still achieved a micro F_1 of 48.2%, providing a benchmark for synthetic data models.

Models trained on synthetic data underperformed compared to real-data models overall. Still, the synthetic MIMIC-L model outperformed the real MIMIC-S model, suggesting that synthetic data scales well with size.

For SEPR data, the trend mirrors English results: real data outperforms synthetic, but the performance gap remains moderate, with the PLM-ICD model trained on synthetic SEPR-L performing better than the one trained on real SEPR-S.

Metric differences across languages, such as higher EMR and lower Precision@k for Swedish, likely stem from SEPR's fewer codes per document. Nonetheless, the relative performance gap between real and synthetic data remained consistent, supporting the generalizability of our framework across languages.

Training on synthetic data from domain-adapted LLaMA models did not lead to performance gains for either medical or Swedish language adaptation. These results suggest that fine-tuning LLaMA-3.1-8B on discharge summaries alone is sufficient for effective adaptation to both the domain and language, reinforcing the validity of our framework and model choice. Although additional domain- or language-specific pretraining did not yield improvements in our experiments, exploring further adapted models remains important, as adaptation effectiveness might depend on the quality of data and methods used for adaptation (Lu et al., 2024).

Overall, while synthetic data is currently outperformed by real data in absolute performance, it shows strong promise, especially as dataset size increases. Given the ability to generate unlimited synthetic data, this highlights significant potential for synthetic data to replace real data in clinical NLP applications eventually.

4.4 Error Analysis: What errors are in the synthetic notes?

We analyzed errors on the MIMIC dataset to highlight limitations when training medical coding models with synthetic data. Table 5 compares correct predictions and error types between real and synthetic MIMIC-L models. We differentiate between within-family (WF) errors, where the wrong predictions still belong to the same ICD-chapter as the target, and out-of-family (OOF) errors, where the prediction falls within a different chapter. Both models show a high proportion of WF errors (79.4% for

	Classification					Ranking			
	AUC-ROC		F_1		EMR	Precision@k		R-precision	MAP
Training Data	Micro	Macro	Micro	Macro		8	15		
Edin et al. (2023)	99.2±0.0	96.6±0.2	58.5±0.7	21.1±2.3	0.4±0.0	69.9±0.6	55.0±0.6	57.9±0.8	61.9±0.9
Real MIMIC-L Short	99.2±0.0	96.6±0.1	58.7±0.4	22.8±1.8	0.4±0.0	70.0±0.3	55.1±0.4	58.0±0.4	62.0±0.5
Synth MIMIC-L	**98.9±0.0	**94.8±0.0	*54.8±0.8	*15.9±0.7	0.3±0.1	**65.5±1.0	*50.9±1.1	*53.9±0.8	*56.9±1.0
Real MIMIC-S Short	96.9±0.4	83.5±1.7	48.2±1.7	3.8±0.8	0.1±0.0	59.7±1.7	45.1±1.7	46.0±1.9	46.4±2.3
Synth MIMIC-S	95.4±0.2	*77.0±0.9	*37.6±0.6	2.1±0.3	0.1±0.0	*48.3±0.6	*35.5±0.6	*35.8±0.7	*34.8±0.7
Real SEPR-L	99.3±0.1	97.0±0.2	60.2±0.9	23.4±1.1	48.8±0.7	12.3±0.1	6.9±0.0	60.4±0.7	71.9±0.7
Synth SEPR-L	98.8±0.2	**95.3±0.1	**54.7±0.6	***14.3±0.7	*44.1±0.8	**11.7±0.1	**6.6±0.0	**54.5±0.5	*66.3±0.9
Real SEPR-S	98.5±0.0	92.2±1.2	52.4±0.5	15.0±0.9	40.5±1.2	11.5±0.0	6.6±0.0	52.1±0.8	64.4±0.6
Synth SEPR-S	*98.1±0.1	89.7±0.8	**45.9±1.1	*8.2±1.3	*30.5±2.6	*10.9±0.1	**6.3±0.0	**45.3±1.4	**58.4±1.2
Medical Pretrained	95.7±0.6	80.8±5.0	36.3±0.7	1.7±0.1	0.0±0.0	45.6±0.7	33.5±0.7	34.2±0.6	33.1±0.9
Swedish Pretrained	98.3±0.0	92.0±0.3	48.8±0.1	10.8±0.2	36.4±0.2	11.2±0.0	6.4±0.0	48.2±0.2	61.0±0.2

Table 4: Results of medical coding models trained on different MIMIC datasets. Significance: $*p < 0.05$, $**p < 0.01$, $***p < 0.001$.

Real MIMIC-L				Synth MIMIC-L			
Correct	Wrong	WF	OOF	Correct	Wrong	WF	OOF
176,270	107,514	85,356	22,158	162,939	116,232	94,733	21,499
62.1%	37.9%	79.4%	20.6%	58.4%	41.6%	81.5%	18.5%

Table 5: Counts of overall correct and wrong predictions as well as WF and OOF family errors alongside percentages for the real-data and synthetic-data MIMIC-L models.

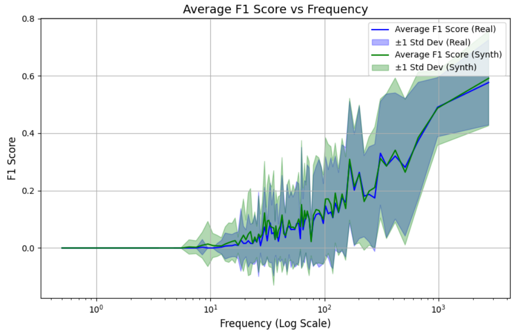

Figure 3: H_1: Macro F_1 vs. code frequencies in training data of real MIMIC-S containing 20% noise in each document (blue) and synthetic MIMIC-S (green) models.

real, 81.5% for synthetic), suggesting they capture disease categories but struggle with fine-grained distinctions. This might reflect code imbalances and a lack of detailed clinical information in discharge summaries. While the synthetic model makes more errors in general, error patterns look similar between models.

One possible reason for the synthetic model's lower performance is noise in the data. Two hypotheses were tested: (H_1) noise is spread across all synthetic notes, and (H_2) only a subset is highly noisy. To simulate noise, random word substitutions were applied to real MIMIC-S data according to each hypothesis. This was done for varying percentages of noise, and the resulting data was used to train medical coding models. To identify the best-fitting noise pattern, F_1 scores were plotted against code frequency, a factor known for its strong influence on performance (Edin et al., 2023), and compared to synthetic model curves. The closest match was found with 20% random substitution across all documents as shown in Figure 3.

This speaks in favor of H_1 and indicates that the synthetic data carries a distributed level of noise explaining utility loss. This noise is likely stemming from LLM artifacts like hallucinations, repetitions, and differences in clinical phrasing. It also implies that the high TTR may reflect noise rather than true desired lexical diversity, highlighting the need for more robust diversity metrics. Future work should aim to reduce these artifacts while preserving diversity and privacy to enhance data quality.

4.5 User Study: Are the synthetic notes medically coherent?

Figure 4 shows the ratings of all three English document pairs. Real documents scored generally slightly higher than synthetic documents in both metrics, with an average score of 3.7 vs. 3.3 for readability and 3.5 vs. 3.3 for medical coherence. However, these differences were not statistically significant ($p \geq 0.05$), neither for the single document pairs, nor on average.

As Figure 4 shows, evaluator ratings varied widely, resulting in low agreement with mean pairwise inter-rater Cohen's Kappa scores of 0.03 (readability) and 0.06 (coherence), likely reflecting differences in medical specialties, experience, and

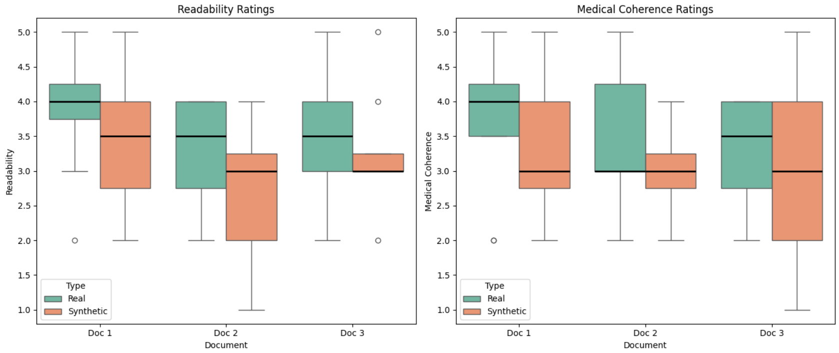

Figure 4: Comparison of readability and medical coherence scores averaged across participants for single document pairs in the MIMIC study. Boxes show the interquartile range (IQR), with the black line as the median and whiskers at 1.5× IQR. None of the differences are statistically significant with ($p \geq 0.05$).

English proficiency. Still, all documents averaged above 3 in both metrics, indicating acceptable quality. A strong correlation between readability and coherence (Pearson: 0.76) indicates shared quality factors. Swedish user study results mirrored the English study, with no significant difference in overall ratings between real and synthetic notes.

Comparable ratings and moderate scores suggest that synthetic notes are of decent quality and reflect effective content control, coherently capturing ICD-10 codes. Future work should involve a larger, more diverse evaluator pool to confirm these findings and better explain the observed disagreement.

5 Discussion and Future Direction

We present a framework for generating synthetic medical notes that aims at balancing privacy and utility by instruction-tuning LLaMA-3.1-8B on ICD-10 codes. Compared to previous methods, our approach stands out for its high privacy-preservation and improved vocabulary richness.

Synthetic data consistently falls short of real data in training medical coding models. However, non-significant differences in medical coherence evaluations and low memorization risk suggest its potential as a general substitute. Larger evaluator samples are needed to validate these findings and clarify the observed low agreement. We hypothesize that the noise contained in the synthetic data, which likely contributes to reduced utility, can be offset by increasing training set size, potentially matching real-data performance. We propose future work to investigate and mitigate this noise by minimizing distortions from real data and reducing LLM-generated artifacts.

The simplicity of the framework makes it attractive for adaptation to other models and languages. While our results demonstrate its effectiveness across languages, further multilingual evaluation is needed, particularly given the lack of transparency from recent open-source LLM creators regarding the inclusion and extent of specific languages in the model's pretraining data.

6 Conclusion

This work presents a scalable and privacy-conscious framework for generating diverse synthetic notes in both English and Swedish. By instruction tuning on ICD-10 codes, we address key challenges in synthetic clinical data generation, notably improving lexical diversity without compromising coherence. Despite some loss in utility compared to real data, our findings indicate that this gap can be narrowed through larger training sets and noise reduction strategies. Crucially, the model performs well without explicit domain or language-specific pretraining, underscoring its adaptability. These results mark a step toward replacing sensitive data with high-quality synthetic alternatives, paving the way for safer, more accessible, and multilingual clinical NLP research.

7 Limitations

Key limitations of this work include the need for deeper medical expert involvement, broader utility evaluation, and critical assessment of biases in real and synthetic data. Model comparisons were limited, and further work should explore alternative architectures, fine-tuning strategies, and privacy-preserving techniques. Developing standardized benchmarks for synthetic medical data evaluation remains an essential step.

References

AI Sweden Models. 2024. Llama-3-8B. *Hugging Face repository*. Accessed on 2024-07-02.

Ali Amin-Nejad, Julia Ive, and Sumithra Velupillai. 2020. 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.

Malaikannan Sankarasubbu Ankit Pal. 2024. OpenBioLLMs: Advancing open-source large language models for healthcare and life sciences. *Hugging Face repository*. Accessed on 2024-07-02.

Tal Baumel, Andre Manoel, Daniel Jones, et al. 2024. Controllable synthetic clinical note generation with privacy guarantees. *Computing Research Repository*, arXiv:2409.07809.

Samuel Belkadi, Libo Ren, Nicolo Micheletti, et al. 2025. Generating synthetic free-text medical records with low re-identification risk using masked language modeling. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 4: Student Research Workshop)*, pages 200–206, Albuquerque, USA. Association for Computational Linguistics.

Emmanuella Budu, Kobra Etminani, Amira Soliman, et al. 2024. Evaluation of synthetic electronic health records: A systematic review and experimental assessment. *Neurocomputing*, 603:128253.

Emily M Burns, Emma Rigby, Ravi Mamidanna, et al. 2012. Systematic review of discharge coding accuracy. *Journal of Public Health (Oxford)*, 34(1):138–148.

H. Dalianis, A. Henriksson, M. Kvist, et al. 2015. HEALTH BANK - A workbench for data science applications in healthcare. *CEUR Workshop Proceedings*.

Hercules Dalianis. 2018. *Clinical text mining: Secondary use of electronic patient records*. Springer Nature.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, et al. 2023. QLoRA: Efficient finetuning of quantized LLMs. *Advances in Neural Information Processing Systems*, 36:10088–10115.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, et al. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, et al. 2024. The Llama 3 herd of models. *Computing Research Repository*, arXiv:2407.21783.

Joakim Edin, Alexander Junge, Jakob D Havtorn, et al. 2023. Automated medical coding on MIMIC-III and MIMIC-IV: A critical review and replicability study. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2572–2582.

Matúš Falis, Aryo Pradipta Gema, Hang Dong, et al. 2024. Can GPT-3.5 generate and code discharge summaries? *Journal of the American Medical Informatics Association*, 31(10):2284–2293.

Carey Beth Goldberg, Laura Adams, David Blumenthal, et al. 2024. To do no harm - and the most good - with ai in health care. *Nat Med*, 30:623–627.

Jiaqi Guan, Runzhe Li, Sheng Yu, et al. 2018. Generation of synthetic electronic medical record text. In *2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 374–380.

Kristiina Häyrinen, Kaija Saranto, and Pirkko Nykänen. 2008. Definition, structure, content, use and impacts of electronic health records: A review of the research literature. *International journal of medical informatics*, 77(5):291–304.

Nicolas Hiebel, Olivier Ferret, Karen Fort, et al. 2023. Can synthetic text help clinical named entity recognition? A study of electronic health records in french. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2320–2338. Association for Computational Linguistics.

JA Hirsch, G Nicola, G McGinty, et al. 2016. ICD-10: history and context. *American Journal of Neuroradiology*, 37(4):596–599.

Chao-Wei Huang, Shang-Chi Tsai, and Yun-Nung Chen. 2022. PLM-ICD: Automatic ICD coding with pretrained language models. In *Proceedings of the 4th Clinical Natural Language Processing Workshop*, pages 10–20. Association for Computational Linguistics.

Tyr Hullmann and Martin Hansson. 2024. Generating synthetic training text from swedish electronic health records. Master Thesis, Stockholm University, DiVA. Accessed on 2025-08-05.

Alistair EW Johnson, Lucas Bulgarelli, Lu Shen, et al. 2023. MIMIC-IV, a freely accessible electronic health record dataset. *Scientific data*, 10(1):1.

James Jordon, Lukasz Szpruch, Florimond Houssiau, et al. 2022. Synthetic data – what, why and how? *Computing Reserach Repository*, arXiv:2205.03257.

Gleb Kumichev, Pavel Blinov, Yulia Kuzkina, et al. 2024. MedSyn: LLM-based synthetic medical text generation framework. In *Machine Learning and*

Knowledge Discovery in Databases. Applied Data Science Track: European Conference, ECML PKDD 2024, Vilnius, Lithuania, September 9-13, 2024, Proceedings, Part X, pages 215–230, Berlin, Heidelberg. Springer-Verlag.

Anastasios Lamproudis, Therese Olsen Svenning, Torbjørn Torsvik, et al. 2024. Using a large open clinical corpus for improved ICD-10 diagnosis coding. In *AMIA Annual Symposium Proceedings*, volume 2023, pages 465–473.

Patrick Lewis, Myle Ott, Jingfei Du, et al. 2020. Pre-trained language models for biomedical and clinical tasks: Understanding and extending the state-of-the-art. In *Proceedings of the 3rd Clinical Natural Language Processing Workshop*, pages 146–157. Association for Computational Linguistics.

Claudia Alessandra Libbi, Jan Trienes, Dolf Trieschnigg, et al. 2021. Generating synthetic training data for supervised de-identification of electronic health records. *Future Internet*, 13(5):136.

Onkar Litake, Brian H Park, Jeffrey L Tully, et al. 2024. Constructing synthetic datasets with generative artificial intelligence to train large language models to classify acute renal failure from clinical notes. *Journal of the American Medical Informatics Association*, 31(6):1404–1410.

Wei Lu, Rachel K. Luu, and Markus J. Buehler. 2024. Fine-tuning large language models for domain adaptation: Exploration of training strategies, scaling, model merging and synergistic capabilities. *Computing Research Repository*, arXiv:2409.03444.

Oren Melamud and Chaitanya Shivade. 2019. Towards automatic generation of shareable synthetic clinical notes using neural language models. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 35–45, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Hajra Murtaza, Musharif Ahmed, Naurin Farooq Khan, et al. 2023. Synthetic data generation: State of the art in health care domain. *Computer Science Review*, 48:100546.

OpenAccess-AI-Collective. 2024. Axolotl. Accessed on 2024-11-15.

Samyam Rajbhandari, Jeff Rasley, Olatunji Ruwase, et al. 2020. ZeRO: Memory optimizations toward training trillion parameter models. In *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis*, pages 1–16, Atlanta, Georgia. IEEE Press.

Debbie Rankin, Michaela Black, Raymond Bond, et al. 2020. Reliability of supervised machine learning using synthetic data in health care: Model to preserve privacy for data sharing. *JMIR Medical Informatics*, 8(7):e18910.

Jeff Rasley, Samyam Rajbhandari, Olatunji Ruwase, et al. 2020. Deepspeed: System optimizations enable training deep learning models with over 100 billion parameters. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 3505–3506, New York, NY, USA. Association for Computing Machinery.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, et al. 2023. Stanford alpaca: An instruction-following llama model. *GitHub repository*. Accessed on 2024-04-30.

Fiona Tweedie and Harald Baayen. 1998. How variable may a constant be? Measures of lexical richness in perspective. *Computers and the Humanities*, 32:323–352.

Thomas Vakili, Aron Henriksson, and Hercules Dalianis. 2025. Data-constrained synthesis of training data for de-identification. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 27414–27427, Vienna, Austria. Association for Computational Linguistics.

Thomas Vakili, Anastasios Lamproudis, Aron Henriksson, et al. 2022. Downstream task performance of BERT models pre-trained using automatically de-identified clinical data. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4245–4252, Marseille, France. European Language Resources Association.

vllm-project. 2024. vLLM. Accessed on 2024-11-15.

Thanh Vu, Dat Quoc Nguyen, and Anthony Nguyen. 2021. A label attention model for ICD coding from clinical text. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence*, IJCAI'20.

Jordon Wimsett, Alana Harper, and Peter Jones. 2014. Components of a good quality discharge summary: A systematic review. *Emergency Medicine Australasia*, 26(5):430–438.

World Health Organization. 2016. *International Statistical Classification of Diseases and Related Health Problems*, 10th revision, fifth edition edition, volume 1. World Health Organization, Geneva.

Vithya Yogarajan, Bernhard Pfahringer, and Michael Mayo. 2020. A review of automatic end-to-end de-identification: Is high accuracy the only metric? *Applied Artificial Intelligence*, 34(3):251–269.