Evaluating Open-Source Large Language Models for Synthetic Non-English Medical Data Generation Using Prompt-Based Techniques

Lenart Dolinar[1], Erik Calcina[2,3], Erik Novak[2,3,*]
[1]University College London, Gower Street, London, WC1E 6BT, United Kingdom
[2]Jožef Stefan Institute, Jamova cesta 39, 1000 Ljubljana, Slovenia
[3]Jožef Stefan International Postgraduate School, Jamova cesta 39, 1000 Ljubljana, Slovenia
E-mail: lenart.dolinar04@gmail.com, erik.calcina@ijs.si, erik.novak@ijs.si
[*]Corresponding author

Keywords: Synthetic data, healthcare data, multilingual data, large language models, classification

Received: November 6, 2024

Using synthetic data sets to train medicine-focused machine learning models has been shown to enhance their performance; however, most research focuses on English texts. In this paper, we explore generating non-English synthetic medical texts. We propose a methodology for generating medical synthetic data, showcasing it by generating medical texts written in a non-English mixed language. We evaluate our approach with thirteen different language models that are open-source and proprietary, and assess the quality of the data sets in two ways: performing a statistical comparison between the original data set and the generated data sets, and training a classifier to distinguish between original and synthetic examples. The `Llama-3.2-3B` *model achieves the best F1 score of* 0.821 ± 0.007 *and accuracy of* 0.816 ± 0.016*, making it most suitable for generating indistinguishable medical synthetic data. In contrast, models like* `Aya-23`*,* `Phi-3`*, and* `SmoLLM` *variants achieve high F1 scores (*0.945–0.948*), indicating their synthetic data is easily distinguishable from original data. These findings highlight the importance of model selection when generating synthetic medical data sets in non-English languages.*

Povzetek: Članek predstavi metodologijo za generiranje sintetičnih neangleških medicinskih podatkov z uporabo pozivov (promptov) in odprtokodnih LLM-jev ter oceni njihovo podobnost izvirnim podatkom z metrikami in klasifikacijo.

1 Introduction

The healthcare domain produces a lot of medical data that can be used to train machine-learning models to help medical personnel. For example, a machine-learning model designed to perform Named Entity Recognition (NER) on electronic health records (EHRs) needs extensive labeled data sets to accurately identify medical terms like diseases, treatments, and patient details. However, since the data contains sensitive personal information, hospitals are restricted from sharing it freely due to data protection regulations. In addition, there are not enough examples to train the models for some problems, such as those relating to rare diseases. Because of this, synthetic data is being used as a substitute to train the models.

Most synthetic data generation approaches focus on generating English texts. These usually utilize large language models trained on predominantly English documents retrieved from the web. However, there are only a few examples of using them to generate non-English texts. To generate data in the medical domain, particularly for data sets with similar content and distributions as the original, it is crucial to avoid third-party APIs to protect personal information. On-premise models are preferred and must be small enough to run on commercial hardware. Furthermore, language models struggle to generate texts with distributions different from their training data, especially in the case of medical texts, which are rarely available to the general public.

1.1 Contribution

The main aim of this work is to develop a methodology for generating synthetic medical examples with characteristics similar to those in the original data set.

The main research objectives are to (1) evaluate different open-source large language models in their capacity of generating synthetic medical examples with the desired characteristics and (2) examine how prompt engineering influences the quality of the generated data.

This paper proposes a methodology for generating synthetic medical data using open-source large language models. We apply the methodology to a medical data set written in a non-English mixed language, where the Latin and non-Latin script is used interchangeably. We test it with thirteen large language models and assess performance by performing a statistical comparison between the original data set and the generated ones, and training a classifier to dis-

tinguish original examples from synthetic ones. Using the same prompt, we find that the open-source `Llama-3.2-3B` model best generates synthetic data that reflects the original data set.

1.2 Paper structure

The remainder of the paper is as follows: Section 2 presents the related work on generating synthetic data using large language models. Next, the proposed methodology is described in Section 3. The experiment setting is presented in Section 4, followed by the experiment results in Section 5. We discuss the results in Section 6 and conclude the paper in Section 7.

2 Related work

This section describes the related work, focusing on large language models and methods for generating synthetic medical data.

2.1 Large language models

Large Language Models (LLMs) are models that were trained to generate human-like texts based on an extensive process of training on vast amounts of data. Models, such as Llama-3 and Llama-3.1 [1, 2], GPT-4 [3], Aya-23 [4], and Mistral [5], are often easy to work with by providing an input textual prompt, based on which the models respond. Recently, LLMs specifically designed for on-device use were also developed. Such models include the Llama-3.2 herd of models [6] and SmolLM [7]; they have fewer parameters than their bigger counterparts but are more accessible and easier to adapt.

The LLMs are helpful in specialized fields, such as medicine, since they can be fine-tuned on extensive data sets containing medical terms and concepts. This enables them to perform well in tasks such as medical synthetic data generation [8]. Despite that, they are sometimes unable to follow the instructions in the prompt accurately, leading them to hallucinate, i.e., confidently produce wrong responses [9].

In our experiments, we investigate the LLMs' performance in generating synthetic medical data given specific constraints and detailed prompts to simulate the original data set as best as possible.

2.2 Synthetic medical data generation

Recently, synthetic medical data generated using LLMs has been used to enhance the performance of models for solving different natural language processing tasks in medicine.

One work focuses on generating a synthetic data set of electronic health records of Alzheimer's Disease (AD) patients based on a label that is provided [10]. They find that the performance of their system for detecting AD-related signs and symptoms from EHRs improves vastly when trained on synthetic and original data sets as opposed to training the system only on the original one. Another work investigated using LLMs for extracting structured information from unstructured healthcare text [11]. By generating synthetic data using LLMs and fine-tuning the model, they significantly improved the models' performance for medical-named entity extraction and relation extraction tasks. Another relevant system for synthetic medical text generation uses Masked Language Modeling (MLM) to generate diverse, high-quality synthetic records, such as discharge summaries and doctor letters [12]. The system includes a de-identification component to mask Protected Health Information (PHI) and a Medical Entity Recognition model to retain key medical details. It produces synthetic data suitable for tasks like NER, with performance comparable to models trained on real data. An extensive review of machine learning approaches for generating synthetic data is available [13].

Most related works focus on English synthetic data due to the scarcity of non-English training data and the dominance of English in medical terminology [14]. However, this paper focuses on generating non-English medical texts; it extends on our previous work [15] by expanding the list of LLMs used in the experiments. The added models have fewer parameters, which will be helpful in analyzing how small a model can be to still generate meaningful examples. The related works are summarized in Table 1.

To the best of our knowledge, this is the first research to comprehensively investigate synthetic data generation for unstructured, mixed-script, non-English medical text. Our work addresses a gap in the literature by focusing on very specific characteristics that distinguish it from previous studies: extremely short doctor comments (typically 1-3 sentences), completely unstructured format without standardized medical templates, mixed-script text combining multiple writing systems, and non-English language content. While previous works have explored synthetic medical data generation in English or touched upon non-English medical texts, none have specifically targeted the unique challenges posed by this combination of structural irregularity, script diversity, and non-English linguistic characteristics that are common in real-world clinical documentation outside English-speaking healthcare systems.

3 Methodology

This section outlines our research methodology. We first present the pre-processing of the data set, followed by a description of the synthetic data generation process. Finally, we present the technical details of the proposed approach. Figure 1 shows the diagram overviewing the proposed methodology.

3.1 Data pre-processing

The data set used consisted of 1,299 examples of real medical data written in a non-English mixed language, where

Table 1: A summary of related works on synthetic medical data generation

Ref	Year	Method/Model	Data Type	Language	Key Findings
[10]	2023	Two-stage approach with LLMs	Electronic Health Records (EHRs) Alzheimer's Disease	English	Synthetic + original data significantly improved AD detection performance compared to original data only
[11]	2023	LLMs for structured information extraction	Unstructured healthcare text	English	Significantly improved medical NER and relation extraction tasks through LLM fine-tuning
[12]	2024	Masked Language Modeling (MLM)	Discharge summaries and doctor letters	English	Generated high-quality synthetic records with de-identification component, performance comparable to real data
This work	**2024**	**Multiple LLMs (including small models)**	**Short doctor comments, unstructured, mixed-script**	**Non-English**	**First comprehensive study on mixed-script non-English medical synthetic data generation**

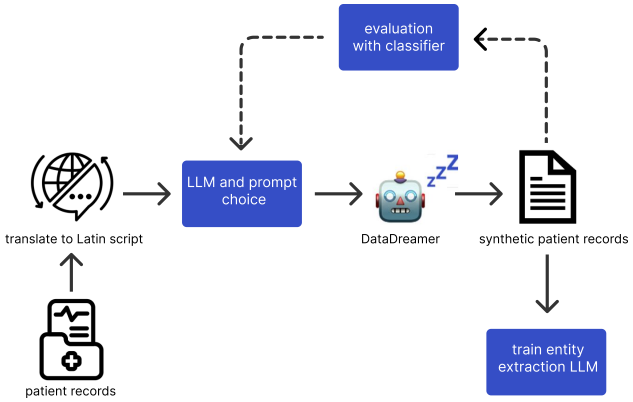

Figure 1: An overview of the methodology. The image was designed using resources from Flaticon

the Latin and non-Latin scripts were used interchangeably. It also contained 1,495 labels, most of which were in English. The labels consisted of drugs, medical events, and measurements.

To align with our non-English target language, we translated all English labels using the `NLLB-200` [16] translation model[1]. In addition, because the original data set consisted of text written in both Latin and non-Latin script, we subsequently transliterated both the labels and the examples into Latin script. This was achieved by simply transliterating each non-Latin letter to its Latin counterpart. As a result, the final data consisted entirely of Latin script text written in the target non-English language. This allowed the LLMs, predominantly trained on Latin script texts, to generate tokens with richer information.

We split the real data set into two subsets to ensure no data leakage. The first one, consisting of 930 real examples, was used for synthetic data generation. The second one, containing the remaining 369 real examples, was used for evaluation.

3.2 Synthetic data generation

We utilized the DataDreamer library [17] to generate the synthetic data set. The library enables open-source models to create synthetic data sets and was developed to work in research settings, supporting prompt templates and few-shot learning.

We developed a prompt containing the instructions and restrictions on generating the examples. To showcase the structure of the generated text better, we also provided five random examples from the original data set as few-shot examples. Next, using DataDreamer, we sent the prompt to the chosen LLM. We experimented with multiple LLMs, and about 800 examples were generated for each LLM. When using external providers such as OpenAI (`GPT-4o`, `GPT-3.5-Turbo`), we included a small number of few-shot examples containing real but short and non-identifiable patient data. All inputs were carefully reviewed to exclude any sensitive information to comply with the GDPR regulation. This process aligns with recommendations on conducting Data Protection Impact Assessments (DPIAs) in AI-based medical research projects [18].

To ensure the quality of the generated data and minimize the effect of hallucinations, we implemented a post-processing step. This included formatting the generated text into a single line, excluding overly long examples, and filtering out examples with meaningless repetition. As part of this, we carried out a simple hallucination analysis: we filtered out any generated text that exceeded 15 words or repeated the same word at least three times. These criteria proved effective in removing most hallucinated outputs while avoiding a high number of false positives. This process ensured that all generated examples followed a consistent format and were suitable for evaluation.

[1] facebook/nllb-200-distilled-600M

The generated examples showed many similarities. Therefore, rigorous methods were needed to evaluate how closely they resemble the original data set. The methods are explained in Section 4.1.

3.3 Technical details

In this section, we describe the models and the parameters used in the experiment. All models used are available via the HuggingFace's transformer library [19].

Language models. We tested eleven open-source models to generate the synthetic data sets, all of which can be run on a 32GB GPU. Our selection was guided by the following criteria: (1) we wanted to compare models that have multilingual capabilities, including our target language, and models that do not have such capabilities, and (2) we wanted to investigate the performance of small models that were considered as one of the best at the time of experimentation. For instance, we included three versions of Meta's Llama family to explore developments over time and investigate trade-offs between size. `Llama-3-8B`[2] only has support for the English language but has been fine-tuned to understand user prompts, which is a feature we expected would help a lot with the synthetic data generation [1]. Its successor, `Llama-3.1-8B`[3] , is trained on multilingual translations and can handle longer input contexts [2]. Both `Llama-3.2-1B`[4] and `Llama-3.2-3B`[5] are language models designed for on-device use, thus contain a smaller number of parameters [6]. Similarly, `SmolLM-145M`[6], `SmolLM-360M`[7], and `SmolLM-1.7B`[8] are on-device models trained on a curated collection of high-quality educational and synthetic data designed for training [7]. `Aya-23`[9] is a multilingual language model that supports 23 languages, including the target non-English language [4]. `Mistral`[10] supports a variety of languages but omits the target language [5]. The models `Gemma-2`[11] and `Phi-3`[12] were also tested and compared in the experiments [20, 21].

In addition, we experimented with two proprietary models, `GPT-4o` and `GPT-3.5-Turbo`, which are accessible via the OpenAI API [3].

Prompt details. All models received the same prompt with instructions to generate target non-English texts written in Latin script, assign a label randomly chosen from the original data set, ensure examples are at most six words long, provide concise responses, and maintain a structured format where all text is on a single line, separated by // and commas, and consistent with the provided few-shot examples. The prompt was tuned to the `Llama-3.1-8B` model and oversaw many gradual improvements. We found that repeating certain key instructions, like clearly stating that examples should be no longer than six words and that labels should be chosen randomly, helped the model follow them more reliably. Capitalizing important warnings (e.g., DO NOT generate any other text/formatting) also made a noticeable difference. Each of these small adjustments, including clarifying the format and structure, led to better overall performance. The full prompt is available in Appendix A.

4 Experiment setting

This section describes the experiment setting, which consists of the evaluation process and the metrics used to measure the approach's performance.

4.1 Evaluation approach

The quality of the generated synthetic data was measured in two parts. The first consisted of statistical measurements, such as calculating the average length of the generated examples and finding the proportion of examples that included the required labels. These statistics were then compared to the original data set.

The second part consisted of training a classifier to discern if the input text was from the original or the synthetic data set. The data set used to train and evaluate the classifier involved 369 randomly selected synthetic examples and 369 examples from the original data set, transliterated into Latin script. We chose 5-fold validation as our classification procedure and calculated the mean performance and its standard deviation across all trials.

The classifier was based on BERT [22], specifically the `bert-base-multilingual-cased` variant[13]. BERT was selected for several key reasons: (1) our data set contained multilingual text with a mixture of cased and uncased formatting, requiring a model capable of handling such linguistic diversity; (2) the relatively short length of comments in our data set aligned well with BERT's context window; and (3) BERT's proven effectiveness for text classification tasks [23]. We did not evaluate alternative classifiers in this work, however, BERT's strong track record (over 4 million downloads in April 2025 on HuggingFace) and reliable performance in prior projects made it the obvious choice for our project.

Hyperparameter selection was conducted through grid search over batch sizes 8, 16, 32, learning rates 1e-5, 2e-5, 5e-5, and training epochs 2, 3, 4, 5. Maximum sequence length was set to 128 tokens as our examples were sufficiently short to fit within this limit. The final configuration—batch size of 16, learning rate of 2e-5, max

[2] meta-llama/Meta-Llama-3-8B-Instruct
[3] meta-llama/Llama-3.1-8B-Instruct
[4] meta-llama/Llama-3.2-1B-Instruct
[5] meta-llama/Llama-3.2-3B-Instruct
[6] HuggingFaceTB/SmolLM-135M-Instruct
[7] HuggingFaceTB/SmolLM-360M-Instruct
[8] HuggingFaceTB/SmolLM-1.7B-Instruct
[9] CohereForAI/aya-23-8B
[10] mistralai/Mistral-7B-Instruct-v0.3
[11] google/gemma-2-9b-it
[12] microsoft/Phi-3-medium-4k-instruct
[13] google-bert/bert-base-multilingual-cased

length of 128, and 3 epochs—was selected based on optimal validation performance using 5-fold cross-validation. These same parameters were consistently applied across all synthetic data sets to ensure comparable results.

4.2 Metrics

To assess the quality of the generated synthetic data sets, we used the F1 score as our main metric for evaluating the classifier's performance. The target value was 0.5; if the performance is greater than 0.5, the classifier can discern the original from the synthetic examples. Hence, the synthetic data does not reflect the original data set. If the performance is less than 0.5, the classifier has difficulties separating the synthetic from the original data, which can be because the synthetic data contains copies of the original examples. In addition to the F1 score, we measured the classifier's accuracy, precision, and recall, which are also reported.

5 Results

In this section, we present the results of our experiment. We first present the statistical analysis results, followed by the classifier's evaluation.

5.1 Statistical analysis

Table 2 compares the synthetic data sets and the original one regarding label occurrence and average example length. The label occurrence in the original data set is 1.000, as all examples from the original data set are assumed to include relevant labels and information.

The most aligned synthetic data set regarding label occurrence was generated using `GPT-4o`, followed by `Llama-3-8B`. In general, high label occurrence is achieved by using all `Llama-*` models, as well as `Aya-23` and `Gemma-2`. Regarding the average example length, the data set generated using `Gemma-2` performed the best, followed by `Llama-3-8B`.

In terms of label occurrence, the worst-performing models are the `SmolLM-*` models, `Mistral`, and `Phi-3`, which did not include more than 25% of the selected label; in the case of `SmolLM-1.7B`, only 36% of the examples contained the selected label. The data set generated using the `Llama-3.1-8B`, `SmolLM-135M`, `SmolLM-360M`, and `Aya-23` had the largest difference in terms of average example length, on average generating examples with more than three extra words.

Looking at both statistics, we can conclude that `Llama-3-8B` had the best alignment to the original data set regarding label occurrence and example length, closely followed by `GPT-4o`.

5.2 The classifier evaluation

Table 3 shows the F1, precision, recall, and accuracy performances of the trained classifier on different synthetic data sets. The best performance was achieved by `Llama-3.2-3B` with an 0.821 F1 score, followed by `Llama-3.1-8B` with 0.835 F1 score and `Mistral` with 0.838 F1 score. The worst performances were on data sets generated by the `SmolLM-*`, `Aya-23`, and `GPT-3.5-Turbo` models. When comparing all metrics, `Llama-3.2-3B` performs best followed by `Mistral`.

Table 2: Statistical comparison between the original and synthetic data sets. The bold and underlined values represent the best and second-best statistics, respectively.

	LLM	Label occurrence	Avg example length
	original data set	1.000	4.682
open-source	`Llama-3-8B`	<u>0.990</u>	<u>5.330</u> (+0.648)
	`Llama-3.1-8B`	0.987	8.355 (+3.673)
	`llama-3.2-1B`	0.987	6.663 (+1.981)
	`Llama-3.2-3B`	0.973	5.533 (+0.851)
	`SmolLM-135M`	0.767	8.946 (+4.264)
	`SmolLM-360M`	0.507	9.012 (+4.330)
	`SmolLM-1.7B`	0.359	7.026 (+2.344)
	`Aya-23`	0.949	8.040 (+3.358)
	`Mistral`	0.740	6.376 (+1.694)
	`Gemma-2`	0.988	**4.207** (-0.475)
	`Phi-3`	0.782	6.071 (+1.389)
API	`GPT-4o`	**0.996**	3.691 (-0.991)
	`GPT-3.5-Turbo`	0.867	6.764 (+2.082)

6 Discussion

This section discusses the synthetic data generation performance, outlines our methodology's limitations and drawbacks, and proposes potential improvements to the approach.

6.1 LLM performance

Results in Table 2 show significant quality differences among synthetic data sets from different LLMs, with label occurrence ranging from 0.359 for `SmolLM-1.7B` to 0.996 for `GPT-4o`, and average example length from 3.691 for `GPT-4o` to 9.012 for `SmolLM-360M`.

However, Table 3 indicates no significant performance differences within a single synthetic data set, with a maximal standard deviation of the F1 score being 0.064 for the `Phi-3` data set. Other metrics also do not deviate too much from their mean value.

We can also notice that the F1 and accuracy scores are very close for all synthetic data sets, indicating that the classifier likely performed relatively similarly on both classes (synthetic and original) without significant bias. However, when comparing the `Llama-*` and `SmolLM-*` model groups, the `Llama-*` models perform better due to having

Table 3: Mean performance metrics of the classifier for synthetic data sets, with standard deviation. Performances that are closer to 0.5 are considered better. The bold and underlined values represent the best and second-best performances, respectively.

	LLM	F1	Precision	Recall	Accuracy
open-source	Llama-3-8B	0.853 ± 0.014	0.870 ± 0.038	0.853 ± 0.060	0.850 ± 0.019
	Llama-3.1-8B	$\underline{0.835 \pm 0.010}$	0.834 ± 0.002	0.853 ± 0.030	0.831 ± 0.006
	Llama-3.2-1B	0.839 ± 0.018	0.846 ± 0.031	0.860 ± 0.029	0.824 ± 0.042
	Llama-3.2-3B	$\mathbf{0.821 \pm 0.007}$	$\mathbf{0.807 \pm 0.041}$	$\underline{0.841 \pm 0.027}$	$\mathbf{0.816 \pm 0.016}$
	SmolLM-135M	0.940 ± 0.023	0.979 ± 0.016	0.914 ± 0.042	0.925 ± 0.002
	SmolLM-360M	0.948 ± 0.035	0.914 ± 0.034	0.918 ± 0.017	0.925 ± 0.000
	SmolLM-1.7B	0.948 ± 0.026	0.894 ± 0.019	0.907 ± 0.004	0.916 ± 0.015
	Aya-23	0.945 ± 0.005	0.947 ± 0.004	0.945 ± 0.005	0.945 ± 0.005
	Mistral	0.838 ± 0.050	$\underline{0.830 \pm 0.017}$	$\mathbf{0.833 \pm 0.037}$	$\underline{0.837 \pm 0.037}$
	Gemma-2	0.914 ± 0.063	0.926 ± 0.043	0.919 ± 0.035	0.917 ± 0.038
	Phi-3	0.925 ± 0.064	0.919 ± 0.047	0.920 ± 0.025	0.917 ± 0.030
API	GPT-4o	0.915 ± 0.034	0.871 ± 0.046	0.882 ± 0.028	0.890 ± 0.017
	GPT-3.5-Turbo	0.939 ± 0.032	0.947 ± 0.023	0.929 ± 0.034	0.940 ± 0.026

more parameters and being trained on diverse data sources, whereas SmolLM-* models are smaller and trained solely on educational and synthetic data. Fine-tuning the models on medical data could improve their ability to generate medical examples, but care must be taken to avoid compromising their instruction-following capabilities.

We can observe much better performance on models trained primarily on English data than on the Aya-23 data set, which is also trained on target language data. This was somewhat unexpected, as we had anticipated better results from Aya-23 due to its explicit support for the target language. Thus, it is not entirely true that when a model is trained on non-English texts, it will generate this type of synthetic medical data well.

When comparing the Llama-* models among themselves, we can see that Llama-3.2-3B performs better than Llama-3.1-8B and Llama-3-8B, even though it is a smaller model. This is most likely due to the fact that the Llama-3.2-3B model is newer and benefited from more advanced training procedures. Furthermore, as explained in Meta's report [6], the Llama-3.2-* model series also had more extensive multilingual training, which might also be the reason why it performed better for our task than the older models.

As mentioned in Section 2, we did not find any existing research focused on generating synthetic data for mixed-script, non-English medical texts. Therefore, the best F1 score among the models - 0.821 achieved by the Llama-3.2-3B model - cannot be numerically compared to prior benchmarks. Furthermore, our medical collaborators, who provided the original data, did not have the capacity to conduct the time-consuming manual evaluation of the generated outputs. However, based on internal review, and acknowledging our lack of medical expertise, we found that the generated examples appeared realistic and structurally consistent with the original data set.

6.2 Bias, representativeness, class imbalance

Alongside evaluating the quality of the generated output, it's also important to consider how well these synthetic texts reflect real-world medical language. Language models can unintentionally introduce biases based on their training data — for example, by over-representing common conditions or overlooking linguistic features specific to the original non-English texts. Ensuring that the synthetic data remains representative helps maintain both the realism and reliability of downstream evaluations.

To address this, certain distributional differences were introduced deliberately in our synthetic data. While an F1 score close to 0.5 generally indicates high-quality synthetic data, the classifier cannot establish any reliable pattern to separate the synthetic from real data - it is sometimes actually helpful to intentionally avoid exactly replicating the original data set. In our case, we wanted to address the class imbalance issues in the original data, especially for labels describing rare medical conditions. That is why we intentionally generated more synthetic examples for underrepresented labels relative to their occurrence in the original data set, to create a more balanced training set for NER tasks in future work. Hence, our classifier might have more easily distinguished synthetic from real data precisely because we had purposefully adjusted these distributional properties.

6.3 Limitations

Due to limited computing power, only one on-premise GPU with 32GB of space was available, restricting the testing of larger LLMs. To address these challenges, using cloud-based resources or distributed computing could help to run larger models and improve the variety of synthetic data generated. However, given medical data protection regulations, cloud-based resources would have to be used with

caution. Unless robust anonymization and secure processing protocols adhering to GDPR regulation were established, on-premise resources to maintain full control over data confidentiality would be prioritized.

Due to privacy concerns, when using OpenAI's `GPT-4o` and `GPT-3.5-Turbo` models, which are not locally-run models, we had to use five fixed examples when generating synthetic data instead of a larger variety. This potentially led to larger similarities of the `GPT-*` synthetic data sets to the examples instead of the original data set and, consequently, worse performance.

Furthermore, while we collaborated with medical domain specialists who provided the original data, they did not have the time or capacity to assist with a manual review of the generated texts. As a result, our evaluation relied solely on automated metrics and only a basic internal review by the authors, who lack medical expertise. This prevented the use of human-in-the-loop techniques such as reinforcement learning with human feedback, which could have iteratively improved the quality of the generated examples. The F1 scores well above 0.5 indicate that synthetic data remains somewhat distinguishable from real texts, suggesting that expert medical guidance could still meaningfully enhance generation realism.

6.4 Potential improvements

The prompt was the same for all thirteen LLMs and was primarily tested on `Llama-3-8B`. Hence, the performance might be biased towards the model. The method could be improved by tailoring the prompts to each model individually.

The evaluation of synthetic data sets could be further extended by checking for repeating examples in the synthetic data set or by checking how different the generated example is from the five provided examples. The evaluation could also be improved by checking for over-fitting to the original data set.

Furthermore, implementing adversarial fine-tuning approaches could improve the performance of our models and result in F1 scores closer to 0.5. This approach would involve training a classifier to detect synthetic examples and simultaneously improving the synthetic data generator to fool this detector by conducting error analysis of the classifier, leading to more realistic outputs.

7 Conclusion and future work

This paper presents a method for generating non-English synthetic medical data sets. To synthetically create data sets similar to the original, we carefully craft a prompt and perform pre-processing and post-processing of the data to increase performance and eliminate the effect of hallucinations.

Using a classifier and considering the inclusion of labels and generated text length, we conclude that `Llama-3.2-3B` is best for generating examples that most closely resemble the original data set. In the future, we plan to explore the underlying architectures of the models to understand their performance differences in multilingual contexts. This will allow us to refine our methods further and create more accurate data sets.

Furthermore, we intend to use the synthetic data set to train a named entity recognition (NER) system to recognize and extract medical labels from clinical history texts. To evaluate the effectiveness of this approach, we will compare the performance of NER models trained on synthetic data versus those trained on real data, using standard metrics such as precision, recall, and F1 score. This will allow us to assess how well the synthetic data preserves linguistic and semantic features necessary for downstream tasks. We also intend to create a more general pipeline enabling the code to generate synthetic medical data in various languages and formats.

Acknowledgement

This work was supported by the Slovenian Research Agency. Funded by the European Union. UK participants in Horizon Europe Project PREPARE are supported by UKRI grant number 10086219 (Trilateral Research). Views and opinions expressed are, however, those of the author(s) only and do not necessarily reflect those of the European Union or European Health and Digital Executive Agency (HADEA) or UKRI. Neither the European Union nor the granting authority nor UKRI can be held responsible for them. Grant Agreement 101080288 PREPARE HORIZON-HLTH-2022-TOOL-12-01.

Appendix

A LLM Prompt

This section provides the prompt used in the study. The prompt oversaw many revisions and was tailored towards the `Llama-3.1-8B` model. Note that due to privacy concerns, the prompt shown here is changed such that the actual language used is replaced with the label `{target_language}`. We used 5 as our `{sample_size}` and a random label from our original data as `{random_label}`.

> You are a helpful AI assistant tasked with generating synthetic medical data based on provided template data.
>
> Your task is to generate a short medical history example (one example) in `{target_language}`, but in Latin script that includes the label "`{random_label}`". The provided text contains `{sample_size}` concatenated examples, separated by newlines. Your task is to create one example that is similar to the `{sample_size}` given examples. The generated example should be about `{random_int}` words long. Follow these specific

guidelines:

Generate one example only. DO NOT generate {sample_size} examples. The example should be similar to the {sample_size} given examples, but include the label "{random_label}". The following instruction is really important. Instruction: The text should not be more than {random_int} words long. The generated text should be really short. Ensure the label "{random_label}" is included exactly as shown, somewhere within the history. DO NOT CHANGE OR TRANSLATE THE label "{random_label}" in any way! The generated text MUST NOT exceed {random_int} words. Use commas and also rarely use // as separators. The text should be concise and formatted as a list, including diseases, medications, and relevant medical events, all on the same line. They have to be relevant, do not just generate random medical conditions. Use uppercase letters. The text should be in {target_language}, but written entirely in Latin script.

The goal is to create a realistic and concise medical history, including the label "{random_label}" somewhere in the text. Only output the generated text. DO NOT include any additional commentary, explanations, or formatting notes.

References

[1] AI@Meta. Llama 3 model card, 2024. [online] https://github.com/meta-llama/llama3/blob/main/MODEL_CARD.md

[2] AI Llama Team. The llama 3 herd of models. arxiv preprint arxiv:2407.21783, 2024.

[3] OpenAI et al. GPT-4 technical report. arxiv preprint arxiv:2303.08774, 2024.

[4] V. Aryabumi, J. Dang, D. Talupuru, S. Dash, D. Cairuz, H. Lin, et al. Aya 23: open weight releases to further multilingual progress. arxiv preprint arxiv:2405.15032, 2024.

[5] A.Q. Jiang, A. Sablayrolles, A. Mensch, C. Bamford, D.S. Chaplot, D. de las Casas, et al. Mistral 7B. arxiv preprint arxiv:2310.06825, 2023.

[6] Meta AI. Llama 3.2: bringing the power of AI to edge and mobile devices. 2024. [online] https://ai.meta.com/blog/llama-3-2-connect-2024-vision-edge-mobile-devices/

[7] L.B. Allal, A. Lozhkov, and E. Bakouch. SmolLM - blazingly fast and remarkably powerful. 2024. [online] https://huggingface.co/blog/smollm

[8] K. Singhal, S. Azizi, T. Tu, S.S. Mahdavi, J. Wei, H.W. Chung, et al. Large language models encode clinical knowledge. Nature, 620, 172–180, 2023. https://doi.org/10.1038/s41586-023-06291-2

[9] X. Guo, and Y. Chen. Generative AI for synthetic data generation: methods, challenges and the future. arxiv preprint arxiv:2403.04190, 2024.

[10] R. Li, X. Wang, and H. Yu. Two directions for clinical data generation with large language models: data-to-label and label-to-data. Findings of the Association for Computational Linguistics: EMNLP 2023, 7129–7143, 2023. https://doi.org/10.18653/v1/2023.findings-emnlp.474

[11] R. Tang, X. Han, X. Jiang, and X. Hu. Does synthetic data generation of LLMs help clinical text mining? arxiv preprint arxiv:2303.04360, 2023.

[12] S. Belkadi, L. Ren, N Micheletti, L. Han, and G. Nenadic. Generating synthetic free-text medical records with low re-identification risk using masked language modeling. 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), 200–206, 2025. https://doi.org/10.18653/v1/2025.naacl-srw.20

[13] Y. Lu, L. Chen, Y. Zhang, M. Shen, H. Wang, X. Wang, et al. Machine learning for synthetic data generation: a review, arXiv preprint arXiv:2302.04062, 2024.

[14] R.E. Hamel. The dominance of English in the international scientific periodical literature and the future of language use in science. AILA Review, 20(1), 53–71, 2027. https://doi.org/10.1075/aila.20.06ham

[15] L. Dolinar, E. Calcina, and E. Novak. Generating non-English synthetic medical data sets. Proceedings of the Slovenian KDD Conference, 2024. https://doi.org/10.70314/is.2024.sikdd.4

[16] NLLB Team, et al. No language left behind: scaling human-centered machine translation. arxiv preprint arxiv:2207.04672, 2022.

[17] A. Patel, C. Raffel, and C. Callison-Burch. DataDreamer: a tool for synthetic data generation and reproducible LLM workflows. Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 3781–3799, 2024. https://doi.org/10.18653/v1/2024.acl-long.208

[18] G.G. Várkonyi, and A. Gradišek. Data protection impact assessment case study for a research project using artificial intelligence on patient data. Informatica,

44(4), 497–505, 2020. https://doi.org/10.31449/inf.v44i4.3253

[19] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, et al. Transformers: state-of-the-art natural language processing. Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, 38–45, 2020. https://doi.org/10.18653/v1/2020.emnlp-demos.6

[20] Gemma Team, et al. Gemma 2: improving open language models at a practical size. arxiv preprint arxiv:2408.00118, 2024.

[21] M. Abdin, J. Aneja, H. Awadalla, A. Awadallah, A.A. Awan, N. Bach, et al. Phi-3 technical report: a highly capable language model locally on your phone. arxiv preprint arxiv:2404.14219, 2024.

[22] T. Pires, E. Schlinger, and D. Garrette. How multilingual is multilingual BERT? Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 4996–5001, 2019. https://doi.org/10.18653/v1/p19-1493

[23] E.C. Garrido-Merchan, R. Gozalo-Brizuela, and S. Gonzalez-Carvajal. Comparing BERT against traditional machine learning models in text classification. Journal of Computational and Cognitive Engineering, 2(4), 352–356, 2023. https://doi.org/10.47852/bonviewjcce3202838