Using LLMs to Generate Patient Journeys in Portuguese: an Experiment

Tahsir Ahmed Munna[1,2,*], Ana Luísa Fernandes[1,3], Purificação Silvano[1,3,4], Nuno Guimarães[1,2] and Alípio Jorge[1,2]

[1]INESC TEC, Portugal
[2]Faculdade de Ciências da Universidade do Porto (FCUP), Portugal
[3]Faculdade de Letras da Universidade do Porto (FLUP), Portugal
[4]Centro de Linguística da Universidade do Porto (CLUP), Portugal

Abstract

The relationship of a patient with a hospital from admission to discharge is often kept in a series of textual documents that describe the patient's journey. These documents are important to analyze the different steps of the clinical process and to make aggregated studies of the paths of patients in the hospital. In this paper, we explore the potential of Large Language Models (LLMs) to generate realistic and comprehensive patient journeys in European Portuguese, addressing the scarcity of medical data in this specific context. We employed Google's Gemini 1.5 Flash model and utilized a dataset of 285 European Portuguese published case reports from the SPMI website, published by the Portuguese Society of Internal Medicine, as references for generating synthetic medical reports. Our methodology involves a sequential approach to generating a synthetic patient journey. Initially, we generate an admission report, followed by a discharge report. Subsequently, we generate a comprehensive patient journey that integrates the admission, multiple daily progress reports, and the discharge into a cohesive narrative. This end-to-end process ensures a realistic and detailed representation of the patient's clinical pathway as a patient's journey. The generated reports were rigorously evaluated by medical and linguistic professionals, as well as automatic metrics to measure the inclusion of key medical entities, similarity to the case report, and correct Portuguese variant. Both qualitative and quantitative evaluations confirmed that the generated synthetic reports are predominantly written in European Portuguese without the loss of important medical information from the case reports. This work contributes to developing high-quality synthetic medical data for training LLMs and advancing AI-driven healthcare applications in under-resourced language settings.

Keywords

Large Language Model, Patient Journey, Medical Text Generation, Gemini, Prompt Engineering, European Portuguese, Contextual Coherence, Semantic Accuracy

1. Introduction

In recent years, Large Language Models (LLMs) have provided advancements across a variety of complex tasks, including question answering [1, 2], code generation [3, 4], and text generation [5, 6]. The multimodal capabilities [7] of LLMs are another notable feature. Models like GPT-4 [8] combine textual and visual inputs that expand the realm of possible application of these models.

LLMs have also proven outstanding achievements in domain-specific tasks, with the medical field being a prime example [9]. For instance, in the medical context, LLMs are used to generate clinical summaries [6], improve diagnostic processes [10], and provide medical decision support [11]. The capacity to process unstructured clinical narratives and combine multimodal data, including textual inputs and medical images, has considerably improved their value in healthcare [8]. These features establish LLMs as transformational appliances for improving healthcare service and research.

In: R. Campos, A. Jorge, A. Jatowt, S. Bhatia, M. Litvak (eds.): Proceedings of the Text2Story'25 Workshop, Lucca (Italy), 10-April-2025

*Corresponding author.
tahsir.a.munna@inesctec.pt (T. A. Munna); ana.l.fernandes@inesctec.pt (A. L. Fernandes); msilvano@letras.up.pt (P. Silvano); nuno.r.guimaraes@inesctec.pt (N. Guimarães); amjorge@fc.up.pt (A. Jorge)
https://github.com/tahsirmunna/patients_journey.git (T. A. Munna)
0000-0001-9269-502X (T. A. Munna); 0009-0009-0552-3904 (A. L. Fernandes); 0000-0001-8057-5338 (P. Silvano); 0000-0003-2854-2891 (N. Guimarães); 0000-0002-5475-1382 (A. Jorge)

Regardless of these developments, the lack of high-quality, annotated medical data prevents LLMs from being widely used in the medical field. Strict privacy laws like GDPR[1] and HIPAA[2], as well as the difficulties in getting patient consent [12], usually restrict access to real-world medical datasets. Furthermore, developing complete and useful real-world medical datasets is challenging due to the high variability of medical information, which adds complexity to the process. Although a significant amount of medical data is available in English, accessing it in other languages, such as Portuguese, especially in its European variant, is far more challenging. This lack of data impacts the training and fine-tuning of LLMs in specific languages [13]. In order to address these obstacles, different approaches for synthetic data generation [5] have been proposed. In particular, the injection of synthetic data in the training process of LLMs has shown improvements in their performance in specific medical domain tasks [14].

To support healthcare applications, synthetic data generation has typically focused on generating individual medical reports such as admission and discharge reports [15], summarization of medical text [16, 6], and supporting tasks like answering medical questions [1, 2]. On the other hand, most medical synthetic data generated by LLMs, including ClinicalBERT [17], MedPaLM [18], BioGPT [19], PMC-LLaMA [20], BioMedLM [21] are mostly in English, with limited utilization of other languages such as Portuguese. In particular, European Portuguese remains comparatively underrepresented, especially in the medical field, and is often classified as a low- to mid-resource language in this context [22]. Despite efforts by the research community and industry to develop language-unified or language-specific LLMs [23], significant gaps continue to persist.

In this paper, we contribute to the mitigation of the problem of lack of clinical corpora in European Portuguese by proposing a method to generate a specific corpus that can also be adapted to other languages. To the best of our knowledge, no prior research has explicitly focused on generating comprehensive synthetic medical reports in European Portuguese that encapsulate a patient's entire hospitalization journey—from admission and daily progress updates to discharge, all integrated into a cohesive narrative.

This work has the following main contributions:

- **Generation of Synthetic Medical Dataset:** The goal of this research is to generate synthetic datasets that convey the full journey of a patient's hospital stay, including admission reports, daily progress reports, and discharge summaries. In contrast to traditional datasets, which usually contain single reports, the method proposed in this study captures the entire range of a patient's experience. Fine-tuning LLMs on this synthetic medical dataset enables LLMs-based medical support systems to better understand patient hospitalization processes, improving diagnoses, personalized treatments, and overall care.
- **Mitigating the Data Scarcity Problem:** This work is part of a project aimed at establishing Portugal as a global hub for innovative healthcare solutions. It seeks to address the scarcity of data in European Portuguese for AI-driven medical decision support. By generating comprehensive synthetic medical datasets, we aim to meet the specific demands of this project, as well as contribute to broader advancements in this field.
- **Evaluation of LLM for Generating Patients' Journeys in European Portuguese (PT-PT):** This study makes a unique contribution by assessing the LLMs' proficiency in generating medical text in European Portuguese.

The remainder of the paper is structured as follows: Section 2 provides an overview of existing literature on natural language generation, with a focus on advancements in medical text generation. Section 3 outlines the proposed pipeline for generating synthetic medical reports. Section 4 details the qualitative and quantitative evaluation including exploratory results of our work. Finally, Section 5 concludes the paper with limitations and explores potential directions for future research.

[1] https://gdpr-info.eu/

[2] https://www.hhs.gov/programs/hipaa/index.html

2. Related Work

The exponential growth in LLMs technology, such as OpenAI's GPT-4 [8], DeepSeek [24] and similar architectures, have sparked significant interest in their potential applications within the healthcare domain. These models, trained on vast amounts of data, demonstrate remarkable capabilities in understanding and generating human-like text, which can be leveraged for tasks such as medical documentation [25], patient communication [26], clinical decision support [27]. Several studies also demonstrate the effectiveness of LLMs in medical text summarization, where they achieve superior performance compared to traditional methods in terms of both speed and accuracy [28, 29]. Beyond summarization, LLMs have also demonstrated promising results in other important and critical medical tasks. LLMs can identify patterns and relationships within large datasets that help to enable applications in clinical decision support, as explained by Benary et al. [27]. Furthermore, LLMs are also used for analyzing medical images, such as X-rays and CT scans, and extracting relevant diagnostic features [30, 31, 32]. Moreover, the potential of LLMs in drug discovery is increasingly recognized as researchers leverage these models to predict molecular properties, optimize drug design, and accelerate the identification of potential drug candidates [33].

On the other hand, Omiye et al. [34] explain in their study that LLMs in the medical field are facing a significant limitation in developing robust and reliable AI models due to a lack of high-quality annotated medical data. This issue is particularly acute when dealing with specific languages and variants, such as European Portuguese, where data scarcity directly impacts the ability to create and validate accurate models within the targeted language [14]. The limitations imposed by real-world data scarcity have fueled research into methods for generating synthetic medical data. This approach aims to generate synthetic electronic health records (EHRs) and clinical notes, offering a potential solution to the data scarcity problem [35, 36, 37]. However, the generation of realistic synthetic reports that can capture the temporal sequence of events of a patient journey in European Portuguese presents unique challenges, demanding a more sophisticated approach beyond existing techniques [14]. The generation of a patient journey requires not only accurate medical information but also a deep understanding of the linguistic and cultural context [38]. While these applications showcase the versatility of LLMs in healthcare, their adaptation to the specific challenge of patient journey generation remains largely unexplored. Existing research primarily focuses on individual tasks such as generating admission reports and discharge notes [15].

Finally, constructing a comprehensive patient journey necessitates the integration of diverse elements, including the temporal progression of medical events and settings. Notably, in our study, we showed that existing LLMs, such as Gemini, demonstrate the capability to generate resource-constrained medical language by leveraging case reports as references, effectively generating comprehensive patient journeys that provide valuable insights for training and evaluating LLMs.

3. Proposed Approach

We propose a method for generating synthetic patient journeys that fully represent a patient's hospital experience, from admission and daily progress notes to discharge, while existing research often concentrates on generating individual medical reports. This holistic approach offers several key advantages. First, it provides a more complete and nuanced picture of disease progression, enabling a more thorough analysis. Second, it is important to train AI-driven medical support systems that require comprehensive patient information to make accurate and intelligent decisions. Our approach encompasses a sequence of generations: the initial admission report, which details the patient's condition upon arrival; daily progress notes documenting examinations, medications, treatments, and procedures (including surgeries, if applicable); and finally, the discharge report, which summarizes the patient's stay and overall outcome.

To generate synthetic patient journeys, we utilized the Gemini 1.5 Flash model[3] via its API. This

[3] https://ai.google.dev/gemini-api/docs/models/gemini

model was selected due to its fast and versatile performance across a wide range of tasks, as well as our access to its paid version. However, other powerful LLMs could potentially achieve similar results.

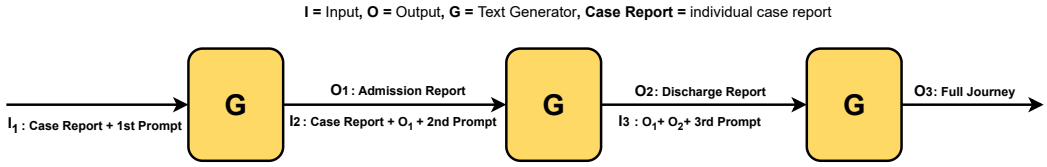

Figure 1: Our proposed pipeline for generating patient's full journey. Here, I_1, I_2, I_3 represent the sequential inputs to G, which acts as a text generator. Similarly, O_1, O_2, O_3 denote the sequential outputs of G.

The pipeline for the generation of the synthetic patient's journeys is presented in Figure 1. The process begins by utilizing a generator G (Gemini 1.5 Flash model), which is responsible for generating the admission, discharge, and the patient's full journey reports, including daily progress notes. To generate these medical reports, a reference *case report* and first *prompt* are provided as input (I_1).

The reference case reports were extracted from a publicly available dataset of Portuguese clinical articles for internal medicine, sourced from the SPMI website[4]. A total of 863 articles were initially collected. Articles that did not contain case reports or were written in languages other than Portuguese were excluded, resulting in a final dataset of 285 case reports. Each case report includes textual descriptions of a patient's clinical case, providing key information such as symptoms, signs, and relevant medical history for the admission report, as well as treatment summaries, exam results, discharge medications, and follow-up instructions for the discharge report. These case reports were selected because they can be used without raising privacy or ethical concerns. However, significant differences in the textual structure between case reports and medical reports introduce additional challenges to our task. These case reports are also important for generating synthetic reports because they ensure accuracy, consistency, and relevance to the specific context or domain. They provide a reliable foundation for the model to produce coherent and contextually appropriate outputs, reducing the risk of errors or irrelevant information.

The prompt specifies the desired output format and content, ensuring clarity and precision in the generated text. As highlighted in the study by Jin et al. [39], refining prompts often requires multiple iterations to achieve clear and well-defined instructions. For this reason, a wide-ranging experimentation was conducted to optimize all the prompt wording and ensure the accuracy and completeness of the patient journey. Interestingly, during the experiment, we found that using English prompts to instruct the model to generate Portuguese text yielded better results than prompts written directly in Portuguese. As a result, all prompts in this study were written in English. To further enhance prompt quality, ChatGPT was employed to test and refine the prompts before their use with the Gemini model. This step was taken as an additional precaution to mitigate potential biases or errors that might arise during generation by the Gemini model. However, using the Gemini model alone may not necessarily have a significant impact. Additionally, the prompts were also evaluated by a linguist, ensuring the linguistic accuracy of the prompt. To make the generated reports more realistic and aligned with a human perspective, we instructed the prompt to include occasional typos in the generated text, ensuring they appear not infrequently and only in a few samples. This rigorous approach ensures the reliability and quality of the final outputs.

We started by providing the generator with the first prompt instructing it to generate an *admission report* (O_1). Once the admission report was generated, we proceeded to the second generation step. Here, the admission report, along with the case report and a new second prompt was used as input (I_2) to generate a *discharge report* (O_2). We generate the discharge report before generating the full patient journey because the full journey includes the admission report, discharge report, and daily updates during the hospital stay. Therefore, we need both the admission and discharge reports to ensure the G can generate an accurate and well-structured full patient journey. Finally, for the third and principal

[4]https://casosclinicosonline.spmi.pt/index.php

generation step of the patient's full journey, we used the generated discharge and admission report as input (I_3), excluding the case report. This omission allowed the model some freedom, while the inclusion of the synthetic admission and discharge reports provided necessary constraints to maintain realism. At the end of all generations, a sample of journeys' was reviewed by a specialist with expertise in linguistics and pharmaceutical sciences to evaluate their coherence, accuracy, and overall suitability as authentic medical records. An example of our generated full patient journey is presented in Figure 2.

*Relatório de Admissão
Doente masculino, 36 anos, autónomo, com história de asma na infância e rinite alérgica, medicado com fluticasona/salmeterol SOS há 1 ano. Tabagista (10 maços/ano), consumo alcoólico <12g/dia. Viagem à República Dominicana (zonas rurais) há 2 anos. Apresenta dor e edema escrotal assimétrico (direito > esquerdo) e edema perineal após traumatismo perineal minor (queda de bicicleta). Queixa de início súbito. Edema de membros inferiores e superiores (predominante em membros inferiores). Observam-se placas edematosas não eritematosas, pouco dolorosas, na região extensora dos antebraços e perna esquerda. Artrite tibiotársica direita. Agravamento progressivo do edema escrotal e perineal com áreas equimóticas e descolamento epidérmico, sem necrose. Eosinofilia (670 µL, 9%), trombocitopenia (85000 µL). PCR levemente elevada (2,4 mg/dL). Negava febre, queixas digestivas, respiratórias e urinárias. Iniciou doxiciclina e AINE em consulta de urologia e posteriormente amoxicilina/ácido clavulânico. Administrou-se metilprednisolona 40mg VO. Diagnóstico inicial: suspeita de gangrena de Fournier. Internamento no serviço de cirurgia.
*Relatório do Internamento - Dias 1-10
Paciente internado com diagnóstico inicial de gangrena de Fournier. Evolução clínica com agravamento do edema e celulite perineal. Manutenção da metilprednisolona 40mg/dia. Realizados exames microbiológicos (hemoculturas, uroculturas, coproculturas) - todos negativos. Serologias para várias infecções também negativas. Eosinofilia em aumento (670 µL no dia 1 para 1980 µL no dia 10). Trombocitopenia persistente. TC pélvica demonstrou celulite perineal sem envolvimento fascial. Sem melhora significativa do quadro clínico.
*Relatório do Internamento - Dia 11
Iniciada corticoterapia com metilprednisolona 40mg/dia. Observada melhoria clínica e analítica, com redução da eosinofilia e normalização da trombocitopenia.
*Relatório do Internamento - Dias 12-24
Realizadas biópsias de pele: paniculite eosinofílica com vasculite. Inicialmente considerada Granulomatose Eosinofílica com Poliangeíte (Síndrome de Churg-Strauss), ANCA negativo. Broncofibroscopia com Lavado Bronco-Alveolar (LBA): trofozoítos e cistos compatíveis com Entamoeba histolytica. Diagnóstico revisto para infecção pulmonar por E. histolytica com vasculite associada. Iniciado tratamento com metronidazol 750mg t.i.d. durante 14 dias, com redução gradual da corticoterapia. Melhora clínica progressiva, com resolução completa das lesões cutâneas e melhoria do derrame pleural (não descrito anteriormente, mas inferido pela melhora clínica). Azatioprina suspensa devido à trombocitopenia possivelmente iatrogênica.
*Relatório de Alta
Doente com 25 dias de internamento. Alta hospitalar com resolução completa das lesões cutâneas e melhoria significativa do quadro clínico. Diagnóstico final: infecção pulmonar por Entamoeba histolytica com vasculite associada. Orientação para seguimento em consulta externa de doenças autoimunes e repetição de LBA para confirmar erradicação da E. histolytica. Recomenda-se seguimento próximo para monitorização da eosinofilia e função renal e hepática. Observa-se uma possível relação entre a viagem à República Dominicana e a infecção parasitária.

Figure 2: An example of a synthetically generated patient's full journey, including admission notes, daily progress reports, and discharge summaries. (* denotes report titles).

Additionally, all sample materials, including code, case reports, prompts, generated admission, discharge, and comprehensive patient journey reports are available via *GitHub repository* [5].

4. Evaluation

In this section, we assess the quality of the generated patient journeys. The evaluation protocol includes qualitative and quantitative assessments, done by experts and by automated benchmarking, respectively.

4.1. Evaluation Protocol (EP)

Qualitative EP: The linguistic and content analysis of the reports generated by the LLM was conducted by an expert in linguistics and pharmaceutical sciences with experience in the analysis of medical reports. Drawing on the expert's prior experience analyzing in previous projects and following a comprehensive examination of the real medical reports used in this study, six key parameters were identified for assessment based on their distinct characteristics:

[5] https://github.com/tahsirmunna/patients_journey.git

1. **Specialised Medical Language**: The case reports employ specialized medical language that is precise and appropriate to the clinical context.
2. **Narrative Nature**: The case reports exhibit a narrative character (i.e. presenting a story organised around a sequence of events, with a structure comprising a beginning, middle, and end.)
3. **Coherence**: The case reports adhere to a logical and consistent structure, without any internal contradictions.
4. **Use of Inter-sentential Connectors**: The case reports feature few or no inter-sentential connectors (e.g. "as a result", "in conclusion", etc).
5. **Occurrence of Typographical Errors**: The case reports often contain minor typographical errors; however, these do not hinder the comprehension of the affected words.
6. **Essential Medical Information**: The case reports include all the necessary medical information to understand the clinical case, such as the reason for hospitalisation (in the case of admission reports), the patient's progress during hospitalisation up to discharge (in discharge reports), and the patient's complete clinical journey (in full journey reports).

Each parameter was assessed by the expert using Likert scales [40], with scores ranging from 1 to 5. For instance, for the first parameter, "Specialized Medical Language," the question posed was: "Does the report use specialized medical language (technical terms appropriate for the medical context)?" Five response options were provided: (1) Not specialized; (2) Slightly specialised; (3) Moderately specialised; (4) Quite specialised;(5) Fully specialized. A detailed analysis of the parameters can be found in the project's GitHub repository. For the development and interpretation of the Likert scale, the recommendations of [41] were followed.

Quantitative EP: To evaluate the quality of the generated patient journeys, we used three different quantitative methods in order to measure (1) the inclusion of key information presented in the case report (such as symptoms, diagnoses, and exams) in the generated reports; (2) the semantic and textual similarity between the case report and the generated reports; and (3) the identification of European Portuguese in the generated text.

To evaluate the inclusion of key medical information in the generated reports, we applied MediAlbertina [42], a state-of-the-art model for **Named Entity Recognition (NER)** specifically designed to extract medical entities (such as diagnoses, medications, and procedures) from medical texts in European Portuguese. We extract the entities from individual case reports and the corresponding generated reports. After extracting, we verify if the generated reports included the key medical entities that existed in the case reports in the following way: Let E_s be the set of entities in the generated individual report (admission, discharge, and full journey) and E_o be the set of entities in the corresponding individual case report. We want to verify that all entities in the generated report(E_s) are a subset of the entities in the case report (E_o). We can demonstrate this in this way $E_s \subseteq E_o$. After that, we find the number of matches between individuals E_s and E_o to get a score.

To assess the semantic similarity between a candidate (generated report) text and a reference (case report) text, we utilized **BERTscore** [43], which uses BERT embeddings to measure contextual alignment. It works by first generating contextual BERT embeddings for each token in both texts. Then, it computes the cosine similarity[44] between each token in the candidate and reference texts. Precision is calculated as the average similarity of candidate tokens to their closest reference tokens, while recall is the average similarity of reference tokens to their closest candidate tokens. The final BERTScore is the harmonic mean (F1 score) of precision and recall. A higher BERTscore indicates greater semantic similarity between the original and generated text. In addition, we also applied **BLEU score** [45], a common metric for evaluating machine translation that measures the amount of word overlap between the original and generated texts. BLEU scores are lower if there are many differences in the exact words used.

Finally, we applied a **Portuguese Language Variety Identifier (LVI)** [46]. This LVI system is specifically trained to distinguish between European and Brazilian Portuguese texts. Using the LVI, we

Table 1
Arithmetic Means and Standard Deviations of the Likert Scales Results

	Admission Reports		Discharge Reports		Full Journey Reports		Ideal Results	
	$\bar{x}$	σ	$\bar{x}$	σ	$\bar{x}$	σ	$\bar{x}$	σ
Specialised Medical Language	5 – Fully Specialised	0	5 – Fully Specialised	0	5 – Fully Specialised	0.92	5 – Fully Specialised	0
Narrative Nature	5 – Fully Narrative	0	5 – Fully Narrative	0	5 – Fully Narrative	0	5 – Fully Narrative	0
Coherence	5 – Fully Coherent	0	5 – Fully Coherent	0	5 – Fully Coherent	0	5 – Fully Coherent	0
Inter-sentential Connectors	1 – Completely absent	0	1 – Completely absent	0.64	1 – Completely absent	0.3	1 – Completely absent	0
Typos	1 – None	0.4	1 – None	0.4	1 – None	0	1 – None	0
Essential Medical Information	3 – Moderately sufficient	1	5 – Fully sufficient	1	5 – Fully sufficient	0.92	5 – Fully sufficient	0

can assess if the output aligns with the linguistic characteristics of European Portuguese, ensuring that the patient's journey is created in this language variant.

4.2. Qualitative Results

The six previously mentioned quality measurement parameters were analyzed across 30 synthetic reports corresponding to the medical histories of 10 patients. Each patient had an admission report, a discharge report, and a full journey report. To ensure diversity in our sample, we selected 10 unique patients from the 285 available using the k-means clustering technique [47]. Table 1 presents the results of the arithmetic means and standard deviations of the sample of 10 patients' journeys.

As evidenced by the results presented in Table 1, all the synthetic reports analysed, similar to the case reports, exhibited specialised, consistent and concise language appropriate to the clinical context. Similarly, all reports displayed a narrative nature, characterised by coherence and logical structure. As with the case reports, the use of inter-sentential connectors was rare (only one connector in the full journey report of id_41, two connectors in the discharge report of id_52, and one connector in the discharge report of id_270 or entirely absent.

Regarding typographical errors, some occurrences were noted, similar to those found in the case reports, without compromising the understanding of the words. However, in the discharge report for the patient identified as *id_270*, the LLM introduced an unusual alteration by replacing the word "cotovelos" (elbows) with "coelhos" (rabbits) and adding the following sentence at the end of the report: "Desculpe pelo erro tipográfico em "coelho" - era "cotovelo"" ("Apologies for the typographical error in "coelho" – it should have been "cotovelo""). This change does not represent a typographical error, and the sentence at the end is not something typically found in a case report. Another relevant example was observed in *id_41*, where the term "influenzais" was used in the sentence "Recomenda-se vacinação anti-pneumocócica e influenzais" ("Pneumococcal and influenza vaccination is recommended"). This term is neither dictionary-recognised nor commonly used in medical jargon. Lastly, in the full journey report for the same patient, the expression "36 UMA" (Units of cigarettes per Year) was incorrectly translated by the LLM as "36 anos-pack", a non-existent unit of measurement. This "hallucination" was isolated, as the other reports maintained the correct terminology.

Regarding the inclusion of essential medical information, the admission reports exhibited significant gaps. In several cases, essential details, such as results from medical and laboratory tests, were missing, which are necessary to justify the therapeutic decisions made. In contrast, the discharge and full journey reports were notably comprehensive, with the LLM adding coherent and contextually relevant information (not present in the case reports), significantly enriching the synthetic reports.

In terms of the variety of Portuguese, the synthetic reports were predominantly written in European Portuguese. Only one case displayed a feature of Brazilian Portuguese, specifically in the sentence "O quadro clínico se agravou progressivamente" ("The clinical condition progressively worsened"), from *id_91*. In this sentence, the clitic pronoun "se" appeared in a proclitic position ("se agravou"), typical of Brazilian Portuguese. In European Portuguese, the typical construction would have been in an enclitic position ("agravou-se").

Overall, the synthetic reports generated by the LLM demonstrated good quality in terms of specialised language, narrative structure, and clinical context appropriateness, aligning closely with the standards observed in the case reports. However, occasional flaws were identified, such as errors in the translation

of units of measurement and omissions of relevant information in some admission reports. Despite these limitations, the results suggest that the LLM holds great potential for generating medical reports.

4.3. Quantitative Results

Table 2
Portuguese Language Variety Identification, NER inclusion and Similarity Scores

Report	PT-PT (%)	PT-BR (%)
Admission	**98.24**	1.76
Discharge	**95.42**	4.58
Full Journey	**97.18**	2.82

Metric	Score
Admission NER inclusive score	1.00
Discharge NER inclusive score	1.00
Full journey NER inclusive score	**1.00**
Admission BERTscore	0.75
Discharge BERTscore	0.76
Full journey BERTscore	0.77
Admission BLEU score	0.025
Discharge BLEU score	0.13
Full Journey BLEU score	0.14

In addition to qualitative analysis, we also measured quantitative results, as shown in Table 2, to provide a comprehensive evaluation of the performance and effectiveness of the generated outputs. The right side of Table 2 shows that NER-inclusive average scores of **1.00** were achieved across all three report types, indicating that the medical entities extracted from the generated texts matched those in the case reports. These results were achieved through multiple iterations of prompt refinement to ensure appropriate generation and the inclusion of key entities from the case reports. The NER-inclusive average score for the full journey report also remains 1.00 even though the case report was not provided during the generation of the full journey. This happens because the admission and discharge reports linked to the case report were used as input. On the other hand, during full journey generation, the model generates text with some degree of freedom and occasionally adds treatments or medications not present in the case report, as explained in Section 4.2.

Furthermore, the BERTscores, which measure semantic similarity, also demonstrate strong performance, ranging in average score from **0.75** to **0.77**. This indicates a high level of semantic alignment and coherence between the generated and case reports. In contrast, BLEU scores were much lower, ranging in average score from **0.026** to **0.148**, showing that the generated texts are not word-for-word copies of the case reports. This doesn't mean the texts are poor—instead, it highlights that they capture the intended meaning without repeating the exact words.

Finally, on the left side of Table 2, the results demonstrate the distribution of Portuguese variants in the generated texts. A strong predominance of European Portuguese (PT-PT) is evident, with an average score consistently exceeding **95%** across all three report types. The presence of Brazilian Portuguese (PT-BR) was minimal, with an average score never surpassing **4.58%** in any of the generated reports. This indicates that the model predominantly adheres to European Portuguese linguistic norms.

4.4. Exploratory Results

The exploratory results provide a detailed analysis of 285 generated medical reports, focusing on key metrics such as the average number of tokens, most frequent terms, and their occurrences. This analysis highlights the distinct characteristics of admission, discharge, and full journey reports, demonstrating the model's ability to generate coherent and contextually appropriate medical narratives in European Portuguese.

Table 3 shows the exploratory analysis of the generated medical reports reveals distinct patterns in the text and its structure. The admission reports are the most concise, averaging 133.80 tokens,

Table 3
Exploratory Results of Generated Reports

Generated Reports	Avg. # Tokens	Most 10 Frequent Terms
Admission	133.80	anos: 417, exame: 278, antecedentes: 230, dor: 162, sexo: 154, refere: 144, história: 136, físico: 134, urgência: 131, febre: 124
Discharge	323.57	alta: 586, paciente: 410, anos: 375, seguimento: 360, consulta: 301, doente: 298, tratamento: 265, sexo: 259, avaliação: 258, revelou: 257
Full Journey	565.99	relatório: 1424, dia: 1110, paciente: 968, alta: 853, anos: 664, internamento: 614, dias: 606, dor: 471, tratamento: 452, realizada: 451

with frequent terms such as "anos" (age), "exame" (exam), "dor" (pain), and "febre" (fever) highlighting their focus on initial patient assessments, symptoms, and diagnostic procedures. The discharge reports are more detailed, averaging 323.57 tokens, and they emphasize terms like "alta" (discharge), "seguimento" (follow-up), and "tratamento" (treatment), reflecting their role in summarizing hospital stays, treatment outcomes, and post-discharge care plans. On the other hand, full journey reports are the most comprehensive, averaging 565.99 tokens, with terms such as "relatório" (report), "internamento" (hospitalization), and "tratamento" (treatment) indicating thorough documentation of the patient's entire hospital experience, from admission to discharge. This information is important as it ensures diversity across the three types of medical reports, demonstrating the model's ability to generate coherent and contextual reports in European Portuguese. This also includes capturing key medical terminology and structural nuances and supporting the model's effectiveness in producing realistic synthetic patient journeys. For better understanding, we present a word cloud visualization [48] in Figure 3.

(a) Admission Report (b) Discharge Report (c) Full Journey Report

Figure 3: Word cloud visualizations for generated admission, discharge, and full journey reports.

Figure 3 visualizes the most frequently occurring terms in generated admission, discharge, and full patient journey reports. This visualization helps identify key themes and terminology used across different stages of the patient journey. In the admission report, frequent terms include "anos" (age),"exame" (examination), "antecedentes" (background), "urgência" (emergency), and "dor" (pain). This reflects a focus on initial patient evaluations, medical history, and acute symptoms that led to hospitalization. The discharge report prominently features terms such as "alta" (discharge), "seguimento" (follow-up), "tratamento" (treatment) and "avaliação" (evaluation). This indicates an emphasis on therapeutic interventions, clinical progress, and the patient's condition at the time of discharge. The full journey report

combines terms from both admission and discharge, with dominant words like "relatório" (report), "internamento" (hospitalization), " "tratamento" (treatment) and "realizada" (carry out). This highlights the comprehensive nature of the full journey, encompassing the entire patient trajectory from diagnosis and treatment to follow-up assessments. Overall, the word clouds demonstrate that each stage of the hospital journey has distinct medical focuses—admission reports concentrate on symptoms and history, discharge reports emphasize treatment outcomes, and the full journey provides a detailed and cohesive medical narrative.

5. Conclusions, Limitations and Future Work

This research highlights the potential of LLMs to address the scarcity of open clinical textual records, particularly in European Portuguese, by generating realistic and comprehensive patient journeys. This marks a significant advancement, given the limited data available for training and evaluating LLMs in this under-resourced language and domain. Our findings highlight the effectiveness of the Gemini 1.5 Flash model in producing synthetic patient journeys that closely mirror the structure and content of real-world medical records. The proposed generation approach, referencing real-world clinical case reports, proved particularly effective in ensuring both the coherence and clinical accuracy of the generated texts. Quantitative analysis confirms the accurate generation and transference of key medical entities from the case reports to the generated reports. BERTscores demonstrated strong semantic alignment with the case reports; the lower BLEU scores indicate that the generated texts are not exact replicas of the case reports, confirming the successful creation of high-quality European Portuguese patient full journeys. Additionally, qualitative evaluations by a linguistics and pharmaceutical sciences expert experienced in assessing medical reports further validated the clinical accuracy and linguistic coherence of the generated report. The generated text was also found to feature a high level of association with European Portuguese, as verified by both evaluation processes and finally, exploratory results ensuring the diversity across the three types of medical reports.

While our paper demonstrates the promising capability of LLMs in generating realistic European Portuguese patient journeys, it is important to acknowledge the limitations. The dataset we used for reference, while carefully selected, comprises only 285 anonymized case reports for internal medicine, which does not fully represent the diversity of patient experiences or clinical scenarios. In addition, the evaluation process conducted, although comprehensive, relies on automated metrics such as BertScore, which have their own biases and limitations, especially in capturing the nuanced semantic meaning of medical language. Furthermore, the study's focus on a single language limits the generalization of these findings to other under-resourced languages and healthcare settings. Finally, healthcare also demands high precision, as minor errors can have severe consequences. However, LLMs, designed for general responses, may produce inaccurate or misleading information, increasing the risk of harmful or unreliable outputs in medical contexts.

Future research directions are multifaceted. First, expanding the dataset to include a wider variety of patient cases and clinical scenarios will improve the generalizability and robustness of the proposed approach for generating patient's journey. The integration of additional data modalities, such as medical images and laboratory results, presents a promising avenue for generating even more comprehensive and realistic synthetic patient journeys. Moreover, adapting this methodology to other under-resourced languages will contribute significantly to the development of diverse and widely accessible AI-driven healthcare tools.

Acknowledgments

This work is co-financed by Component 5 - Capitalization and Business Innovation, integrated in the Resilience Dimension of the Recovery and Resilience Plan within the scope of the Recovery and Resilience Mechanism (MRR) of the European Union (EU), framed in the Next Generation EU, for the period 2021 - 2026, within project HfPT, with reference 41.

References

[1] K. Singhal, T. Tu, J. Gottweis, R. Sayres, E. Wulczyn, M. Amin, L. Hou, K. Clark, S. R. Pfohl, H. Cole-Lewis, et al., Toward expert-level medical question answering with large language models, Nature Medicine (2025) 1–8.

[2] J. Wang, Z. Yang, Z. Yao, H. Yu, Jmlr: Joint medical llm and retrieval training for enhancing reasoning and professional question answering capability, arXiv preprint arXiv:2402.17887 (2024).

[3] C. Qian, X. Cong, C. Yang, W. Chen, Y. Su, J. Xu, Z. Liu, M. Sun, Communicative agents for software development, arXiv preprint arXiv:2307.07924 6 (2023).

[4] F. Lin, D. J. Kim, et al., When llm-based code generation meets the software development process, arXiv preprint arXiv:2403.15852 (2024).

[5] G. Kumichev, P. Blinov, Y. Kuzkina, V. Goncharov, G. Zubkova, N. Zenovkin, A. Goncharov, A. Savchenko, Medsyn: Llm-based synthetic medical text generation framework, in: Joint European Conference on Machine Learning and Knowledge Discovery in Databases, Springer, 2024, pp. 215–230.

[6] H. Jung, Y. Kim, H. Choi, H. Seo, M. Kim, J. Han, G. Kee, S. Park, S. Ko, B. Kim, et al., Enhancing clinical efficiency through llm: Discharge note generation for cardiac patients, arXiv preprint arXiv:2404.05144 (2024).

[7] S. Wu, H. Fei, L. Qu, W. Ji, T.-S. Chua, Next-gpt: Any-to-any multimodal llm, arXiv preprint arXiv:2309.05519 (2023).

[8] H. Nori, N. King, S. M. McKinney, D. Carignan, E. Horvitz, Capabilities of gpt-4 on medical challenge problems, arXiv preprint arXiv:2303.13375 (2023).

[9] S. Pal, M. Bhattacharya, S.-S. Lee, C. Chakraborty, A domain-specific next-generation large language model (llm) or chatgpt is required for biomedical engineering and research, Annals of biomedical engineering 52 (2024) 451–454.

[10] E. Ullah, A. Parwani, M. M. Baig, R. Singh, Challenges and barriers of using large language models (llm) such as chatgpt for diagnostic medicine with a focus on digital pathology–a recent scoping review, Diagnostic pathology 19 (2024) 43.

[11] D. Oniani, X. Wu, S. Visweswaran, S. Kapoor, S. Kooragayalu, K. Polanska, Y. Wang, Enhancing large language models for clinical decision support by incorporating clinical practice guidelines, arXiv preprint arXiv:2401.11120 (2024).

[12] S. Wiertz, J. Boldt, Ethical, legal, and practical concerns surrounding the implemention of new forms of consent for health data research: Qualitative interview study, Journal of Medical Internet Research 26 (2024) e52180.

[13] V. Kumar, E. Ntoutsi, P. S. Rajawat, G. Medda, D. R. Recupero, Unlocking llms: Addressing scarce data and bias challenges in mental health, arXiv preprint arXiv:2412.12981 (2024).

[14] L. F. P. Henriques, Narrative Extraction from Synthetic Clinical Texts in Portuguese, Master's thesis, Universidade do Porto (Portugal), 2024.

[15] I. Hartsock, G. Rasool, Vision-language models for medical report generation and visual question answering: A review, Frontiers in Artificial Intelligence 7 (2024) 1430984.

[16] D. Van Veen, C. Van Uden, L. Blankemeier, J.-B. Delbrouck, A. Aali, C. Bluethgen, A. Pareek, M. Polacin, E. P. Reis, A. Seehofnerova, et al., Clinical text summarization: adapting large language models can outperform human experts, Research Square (2023).

[17] E. Alsentzer, J. R. Murphy, W. Boag, W.-H. Weng, D. Jin, T. Naumann, M. McDermott, Publicly available clinical bert embeddings, arXiv preprint arXiv:1904.03323 (2019).

[18] A. Chowdhery, S. Narang, J. Devlin, M. Bosma, G. Mishra, A. Roberts, P. Barham, H. W. Chung, C. Sutton, S. Gehrmann, et al., Palm: Scaling language modeling with pathways, Journal of Machine Learning Research 24 (2023) 1–113.

[19] R. Luo, L. Sun, Y. Xia, T. Qin, S. Zhang, H. Poon, T.-Y. Liu, Biogpt: generative pre-trained transformer for biomedical text generation and mining, Briefings in bioinformatics 23 (2022) bbac409.

[20] C. Wu, W. Lin, X. Zhang, Y. Zhang, W. Xie, Y. Wang, Pmc-llama: toward building open-source language models for medicine, Journal of the American Medical Informatics Association (2024)

ocae045.
[21] E. Bolton, A. Venigalla, M. Yasunaga, D. Hall, B. Xiong, T. Lee, R. Daneshjou, J. Frankle, P. Liang, M. Carbin, et al., Biomedlm: A 2.7 b parameter language model trained on biomedical text, arXiv preprint arXiv:2403.18421 (2024).
[22] A. Névéol, H. Dalianis, S. Velupillai, G. Savova, P. Zweigenbaum, Clinical natural language processing in languages other than english: opportunities and challenges, Journal of biomedical semantics 9 (2018) 1–13.
[23] J. Y. Wang, N. Sukiennik, T. Li, W. Su, Q. Hao, J. Xu, Z. Huang, F. Xu, Y. Li, A survey on human-centric llms, arXiv preprint arXiv:2411.14491 (2024).
[24] Z. Wu, X. Chen, Z. Pan, X. Liu, W. Liu, D. Dai, H. Gao, Y. Ma, C. Wu, B. Wang, et al., Deepseek-vl2: Mixture-of-experts vision-language models for advanced multimodal understanding, arXiv preprint arXiv:2412.10302 (2024).
[25] S. Goyal, E. Rastogi, S. P. Rajagopal, D. Yuan, F. Zhao, J. Chintagunta, G. Naik, J. Ward, Healai: A healthcare llm for effective medical documentation, in: Proceedings of the 17th ACM International Conference on Web Search and Data Mining, 2024, pp. 1167–1168.
[26] C. R. Subramanian, D. A. Yang, R. Khanna, Enhancing health care communication with large language models—the role, challenges, and future directions, JAMA Network Open 7 (2024) e240347–e240347.
[27] M. Benary, X. D. Wang, M. Schmidt, D. Soll, G. Hilfenhaus, M. Nassir, C. Sigler, M. Knödler, U. Keller, D. Beule, et al., Leveraging large language models for decision support in personalized oncology, JAMA Network Open 6 (2023) e2343689–e2343689.
[28] H. Jin, Y. Zhang, D. Meng, J. Wang, J. Tan, A comprehensive survey on process-oriented automatic text summarization with exploration of llm-based methods, arXiv preprint arXiv:2403.02901 (2024).
[29] D. Van Veen, C. Van Uden, L. Blankemeier, J.-B. Delbrouck, A. Aali, C. Bluethgen, A. Pareek, M. Polacin, E. P. Reis, A. Seehofnerová, et al., Adapted large language models can outperform medical experts in clinical text summarization, Nature medicine 30 (2024) 1134–1142.
[30] S. Wang, Z. Zhao, X. Ouyang, Q. Wang, D. Shen, Chatcad: Interactive computer-aided diagnosis on medical image using large language models, arXiv preprint arXiv:2302.07257 (2023).
[31] D. Tian, S. Jiang, L. Zhang, X. Lu, Y. Xu, The role of large language models in medical image processing: a narrative review, Quantitative Imaging in Medicine and Surgery 14 (2023) 1108.
[32] S. Lee, J. Youn, H. Kim, M. Kim, S. H. Yoon, Cxr-llava: a multimodal large language model for interpreting chest x-ray images, European Radiology (2025) 1–13.
[33] J.-P. Vert, How will generative ai disrupt data science in drug discovery?, Nature Biotechnology 41 (2023) 750–751.
[34] J. A. Omiye, H. Gui, S. J. Rezaei, J. Zou, R. Daneshjou, Large language models in medicine: the potentials and pitfalls: a narrative review, Annals of Internal Medicine 177 (2024) 210–220.
[35] R. J. Chen, M. Y. Lu, T. Y. Chen, D. F. Williamson, F. Mahmood, Synthetic data in machine learning for medicine and healthcare, Nature Biomedical Engineering 5 (2021) 493–497.
[36] R. Tang, X. Han, X. Jiang, X. Hu, Does synthetic data generation of llms help clinical text mining?, arXiv preprint arXiv:2303.04360 (2023).
[37] A. Bauer, S. Trapp, M. Stenger, R. Leppich, S. Kounev, M. Leznik, K. Chard, I. Foster, Comprehensive exploration of synthetic data generation: A survey, arXiv preprint arXiv:2401.02524 (2024).
[38] J. C. Chow, K. Li, Ethical considerations in human-centered ai: Advancing oncology chatbots through large language models, JMIR Bioinformatics and Biotechnology 5 (2024) e64406.
[39] H. Jin, H. Che, Y. Lin, H. Chen, Promptmrg: Diagnosis-driven prompts for medical report generation, in: Proceedings of the AAAI Conference on Artificial Intelligence, volume 38, 2024, pp. 2607–2615.
[40] R. Likert, A technique for the measurement of attitudes, Archives of Psychology, Nova Iorque, 1932.
[41] L. South, D. Saffo, O. Vitek, C. Dunne, M. A. Borkin, Effective use of likert scales in visualization evaluations: A systematic review, Computer Graphics Forum 41 (2022) 43–55. URL: https://doi.org/10.1111/cgf.14521. doi:10.1111/cgf.14521.

[42] M. J. B. Nunes, MediAlbertina: A family of European Portuguese medical language models, Master's thesis, 2024.
[43] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, Y. Artzi, Bertscore: Evaluating text generation with bert, arXiv preprint arXiv:1904.09675 (2019).
[44] F. Rahutomo, T. Kitasuka, M. Aritsugi, et al., Semantic cosine similarity, in: The 7th international student conference on advanced science and technology ICAST, volume 4, University of Seoul South Korea, 2012, p. 1.
[45] M. Post, A call for clarity in reporting bleu scores, arXiv preprint arXiv:1804.08771 (2018).
[46] H. Sousa, R. Almeida, P. Silvano, I. Cantante, R. Campos, A. Jorge, Enhancing portuguese variety identification with cross-domain approaches, arXiv preprint arXiv:2502.14394 (2025).
[47] A. Ahmad, L. Dey, A k-mean clustering algorithm for mixed numeric and categorical data, Data & Knowledge Engineering 63 (2007) 503–527.
[48] F. Heimerl, S. Lohmann, S. Lange, T. Ertl, Word cloud explorer: Text analytics based on word clouds, in: 2014 47th Hawaii international conference on system sciences, IEEE, 2014, pp. 1833–1842.