Generating Readable Spanish Clinical Narratives from Structured Electronic Health Records with Multi-Agent LLMs

Generación mediante LLMs Multiagente de Narrativas Clínicas Legibles en Español a partir de Registros Electrónicos de Salud Estructurados

Mikel Idoyaga[1]*, **Ane G. Domingo-Aldama**[1]*, **Aitziber Atutxa**[1], **Josu Goikoetxea**[1] and **Ander Barrena**[1]

[1]HiTZ Basque Center for Language Technology, University of the Basque Country EHU
{mikel.idoyaga, ane.garciad, aitziber.atucha, josu.goikoetxea, ander.barrena}@ehu.eus

* Equal contribution

Abstract: Electronic Health Records (EHRs) provide longitudinal, interoperable digital histories of patient medical events across healthcare settings. However, the vast quantity of structured data and its tabular format often complicate the synthesis of a clear clinical overview. The primary objective of this work is to generate coherent Spanish clinical narratives that summarize patient trajectories, improving interpretability for clinicians and integration into decision workflows. We propose a multi-agent architecture where specialized Large Language Model (LLM) agents transform distinct clinical structured variables into a unified narrative. We compare three models with varying domain and language specialization: MedGemma-27B, Marmoka-es+it, and LLaMA 3.1. Performance is assessed through a dual evaluation framework combining expert manual annotation with an LLM-as-a-judge approach. Our results highlight the critical role of domain and language adaptation in clinical narrative generation while addressing the lack of specialized resources for Spanish clinical narratives.
Keywords: Electronic Health Records, Large Language Models, Clinical Narrative Generation, LLM-as-a-Judge.

Resumen: Los Registros Electrónicos de Salud (RES) proporcionan historiales digitales longitudinales e interoperables de los eventos médicos de los pacientes en diversos entornos sanitarios. Sin embargo, la gran cantidad de datos estructurados y su formato tabular a menudo complican la síntesis de una visión clínica clara. El objetivo principal de este trabajo es generar narrativas clínicas coherentes en español que resuman la trayectoria del paciente, haciéndola más interpretable para los clínicos y más fácil de integrar en los flujos de trabajo de decisión clínica. Proponemos un flujo de trabajo multi-agente en el que agentes especializados basados en modelos de lenguaje de gran tamaño (LLM) transforman distintas variables clínicas estructuradas en una narrativa unificada. Comparamos tres modelos con diferentes grados de especialización de dominio y lenguaje: MedGemma-27B, Marmoka-es+it y LLaMA 3.1. El rendimiento se evalúa mediante un marco dual que combina la anotación manual experta con un enfoque de "LLM-como-juez". Nuestros resultados destacan el papel crítico de la adaptación al dominio y al lenguaje en la generación de narrativas clínicas, al tiempo que abordan la falta de recursos especializados para narrativas clínicas en español.
Palabras clave: Registros Electrónicos de Salud, Modelos de Lenguaje de Gran Tamaño, Generación de Narrativas Clínica, LLM como Juez.

1 Introduction

Electronic Health Records (EHRs) have become a cornerstone of modern healthcare systems worldwide, providing a comprehensive repository of patient information collected throughout the care process. These records contain both structured and text data and play a critical role in supporting clinical decision making, medical research, and healthcare management (Holmes et al., 2021; Alzubi et al., 2021). By systematically storing patient information such as diagnoses, procedures, laboratory results, and treatments, EHRs enable healthcare professionals to access and analyze large volumes of clinical data.

A substantial portion of the information contained in EHRs is stored in structured interoperable form (ICD, ATC codes, etc). This structured data is typically generated through manual annotation into standardized coding systems and serves as a valuable source of information exchange and consultation for numerous clinical tasks, including disease evolution and diagnosis, risk assessment, and prediction of clinical outcomes. As a side result, structured EHR data has become widely used in clinical research and in the development of computational models for healthcare applications.

Despite its richness and clinical relevance, structured tabular EHR data is often overwhelming and difficult to interpret in a more comprehensive, human-centered manner. Although it contains detailed, temporally ordered information on a patient's care trajectory, its high level of redundancy, combined with its tabular and codified structure, makes it difficult to quickly interpret and grasp the full care pathway. Clinicians often rely on their own narrative summaries to describe the evolution of diseases, treatments, and key events over time; however, these narratives are highly tailored to individual practitioner documentation styles, making them static, non-interoperable, and not readily actionable.

At the same time, fully leveraging the potential of modern artificial intelligence, particularly Large Language Models (LLMs), requires textual data. While advances in natural language processing have enabled powerful methods for analyzing and generating clinical narratives, these systems are inherently designed to operate on text, limiting their effectiveness when information is primarily stored in structured formats.

Consequently, transforming structured EHR data into coherent narrative descriptions of individual patient histories offers two key advantages: it makes clinical information more human-centered and actionable, while also rendering it readily usable by LLMs for advanced AI-driven applications.

In this work, we propose a methodology for generating textual clinical narratives from structured EHR data using LLMs. Our approach follows a multi-agent architecture designed to transform structured patient information into coherent narrative summaries that reflect the clinical trajectory of each patient. To assess the quality and reliability of the generated narratives, we use both automated LLM-based judges and human expert annotation. Our research answers the following research questions:

- **RQ1**: Can we effectively transform structured EHR data into faithful and coherent clinical narratives?
- **RQ2** How does domain and language adaptation impact model performance on this transformation?
- **RQ3** Do LLM-based judges align with human experts when evaluating clinical summarized narratives?

Our main contributions are the following:

- A multi-agent methodology designed to transform structured EHR data into coherent and faithful clinical narratives.
- A comparative study of various LLMs, evaluating the impact of clinical specialization versus general-domain and language-specific training.
- An automated evaluation framework using LLMs as judges, analyzing their alignment with human expertise.

2 Related Work

Several studies have investigated the automatic generation of discharge summaries and patient documentation from structured clinical data. This field has gained significant momentum due to its diverse applications, ranging from streamlining clinical documentation workflows to enhancing data harmonization and supporting robust data augmen-

tation strategies (Woo et al., 2025). Furthermore, the widespread adoption of EHRs has made these automated approaches increasingly feasible and utilitarian, providing the rich, standardized data necessary to drive high-quality natural language generation in clinical settings.

Different methodological approaches have been proposed in the literature to address the conversion of structured data into natural language. Earlier, more simplistic methods relied on pre-configured templates to generate text (Lee et al., 2025; Contreras et al., 2025). While these template-based solutions effectively mitigate hallucination-related issues, they offer limited generalizability and require significant manual effort to develop. Consequently, the use of LLMs is gaining prominence in this research area due to their ability to produce more fluid and adaptable documentation.

Studies focusing on LLMs propose various methodologies to optimize the conversion. Some research utilizes prompt engineering and zero-shot inference with pretrained models to generate clinical summaries directly from unstructured EHR data (Ganzinger et al., 2025). In contrast, other works adopt more advanced strategies, such as fine-tuning, parameter-efficient training techniques (Li et al., 2026), or Retrieval-Augmented Generation (RAG) frameworks (Kruse et al., 2025), to enhance the accuracy and quality of the generated output. The evaluation of generated narratives is typically conducted through manual assessment by physicians or by measuring performance on downstream tasks in an extrinsic evaluation.

Furthermore, Falis et al. (2024) utilizes GPT-3.5 to generate discharge summaries based on ICD-10 code descriptions. However, in practical clinical environments, the deployment of open-source LLMs is essential to address stringent data privacy and security concerns. For instance, Ganzinger et al. (2025) employs Llama 3 as the primary generation engine, while Li et al. (2026) provides a comparative analysis of various architectures, including Llama 3, Qwen 2, and Mistral, to evaluate their performance in clinical documentation tasks.

Despite these advances, important limitations remain in the current literature. First, most existing studies focus on English language clinical data. Second, the evaluation of generated clinical narratives traditionally rely on automated string-similarity metrics against reference summaries, such as BLEU and ROUGE, which often fail to capture factual medical correctness and semantic clinical accuracy (Papineni et al. (2002); Lin (2004)). This drawback has prompted a recent shift toward automated evaluation frameworks using LLMs as judges to balance scalability with human alignment (Vasilev et al. (2025); Wei et al. (2024)).

These limitations reveal a gap in the literature regarding the automatic generation of Spanish clinical histories from structured EHR data and scalable evaluation methodologies based on LLM judges. In this work, we address these challenges by proposing a methodology for generating narrative clinical histories from structured patient data and evaluating the generated summaries using both human annotation and LLM based judges.

3 Resources

Most healthcare systems follow a standardized pipeline for documenting patient events: clinicians first record encounters as clinical notes, typically adhering to predefined formats and sections. In practice, these notes frequently reuse content from previous entries, making them prone to redundancies, potential inconsistencies, and poor interoperability. Once finalized, these narratives are subsequently transformed into standardized codes and stored within the system as structured data.

Given our goal of performing this task using real-world clinical data, the present study was made possible through resources kindly provided by a Public Healthcare System within the context of a collaboration project. The dataset consists of the final completely anonymized structured clinical information encoded by healthcare professionals using standardized coding and stored within the Health System's Business Intelligence platform.

From this platform, we extracted clinical information corresponding to a total of 65 patients. Patient selection was agnostic to any specific medical condition, and all data were retrieved entirely from codified fields. A detailed quantitative profile and descriptive statistics of the extracted patient cohort

are provided in Appendix A. The dataset includes several types of structured variables describing the patients' demographic characteristic, medical history, clinical tests, and treatments. Specifically, the following information was retrieved:

- **Past Medical History**: historical diagnoses recorded for each patient, including the corresponding diagnostic codes (ICD) and the dates on which the diagnoses were registered.
- **Laboratory Results**: results of clinical laboratory tests performed on the patient, including blood tests, urine tests, and other biochemical analyses. Each record includes the type of laboratory test, the date of the test, and the measured value.
- **Treatments**: pharmacological treatments prescribed to the patient, including the start and end dates of the treatment as well as the associated Anatomical Therapeutic Chemical (ATC) classification code.

4 Methodology

In this section, we present the proposed multi-agent architecture for transforming structured data into clinical narratives (see subsection 4.1), alongside the human evaluation framework (see subsubsection 4.2.1) and the LLM-as-a-judge evaluation methodology (see subsubsection 4.2.2). The complete prompt templates for all models are provided in Appendix B.

4.1 Multi-agent pipeline: Structured Data–to–Narrative

We propose a multi-agent architecture–based pipeline to transform structured EHR data into narrative patient histories, with *each agent* responsible for *a key section* of the final narrative. This way each agent focuses and processes specific information of the same tabular data in parallel. Once the parallel processing is complete, the individual outputs from each agent are combined into a single, cohesive clinical narrative in Spanish, ensuring an efficient and modular generation process that avoids the latencies associated with sequential execution (see Figure 1).

This modular approach enables the system to handle heterogeneous clinical variables and apply tailored processing to each

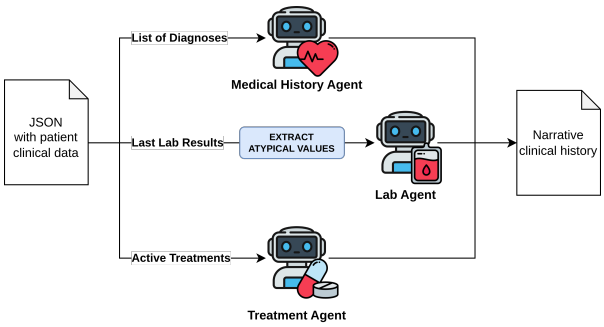

Figure 1: Pipeline for the transformation of structured data into narrative patient histories using agents.

information type, ensuring that all data are summarized according to their clinical relevance. Here is a brief description of each agent together with the implemented restrictions:

- **Past Medical History Agent:** This agent receives a list of diagnosis with timestamps, which are first temporally ordered. The ordered sequence is finally transformed into a coherent narrative that preserves clinical meaning while simplifying terminology.
- **Laboratory Results Agent:** Laboratory values are aligned with timestamps and filtered using predefined clinical reference ranges to identify abnormal results. Only tests closest to the most recent diagnosis are prioritized to reduce noise. Three prompting strategies are used:
  - *No-Filter*: All values are included, with the model implicitly instructed to focus on abnormalities.
  - *Abnormal-Explicit*: All values are included with explicit abnormality indicators.
  - *Abnormal-Only*: Only abnormal values are provided as input.
- **Treatments Agent:** This agent receives a list of treatments with start and end dates, filtering to retain only treatments active at the time of diagnosis. The resulting set is formatted into a coherent narrative.

Additionally, towards answering **RQ2**, it is necessary to assess the impact of the underlying LLM's specificity on the proposed ap-

proach. To this end, we compare three LLMs with distinct characteristics in terms of domain specialization and language adaptation, while also enabling a comparison between compact (8B) and larger-scale (27B) models (see Table 1). Llama-3.1-8B-it (AI@Meta, 2024) is a general-domain English-centric LLM, MedGemma-27B (Google, 2025) is a medical-domain English-centric LLM, and Marmoka-es+it-8B (Domingo-Aldama et al., 2026) is a medical-domain Spanish-centric LLM. Furthermore, LLaMA 3.1 8B and Marmoka-es+it-8B represent compact models that could be realistically deployed in healthcare environments with limited computational resources, whereas MedGemma-27B allows us to evaluate the impact of a larger, domain-specialized model. These models are used as the underlying generators within the multi-agent framework. The generation was performed using Hugging Face Transformers default settings except for the temperature, which was set to 0.7.

Name	Arch.	Domain	Lang.
Llama-3.1-8B-it	Llama3.1	General	multi
MedGemma-27B	Gemma3	Clinical	multi
Marmoka-es+it-8B	Llama3.1	Clinical	es

Table 1: Large language models evaluated in this study.

4.2 Evaluation

In this section, we address **RQ1**, and **RQ3** (see section 1). To this end, we first define a comprehensive set of criteria to support rigorous expert evaluation of the relevance, coherence, and clinical fidelity of the generated narratives. Based on these criteria, we then propose a two-step evaluation framework combining manual expert assessment (see subsubsection 4.2.1) with automated evaluation using LLMs as judges (see subsubsection 4.2.2).

4.2.1 Human Evaluation

The human evaluation is conducted according to the following quality criteria described under the taxonomy of quality criteria presented in Belz et al. (2025):

- **Clinical Information Coverage**: the extent to which the generated narrative accurately reflects the clinically relevant information present in the structured data. This is a Correctness criterion assessed relative to the input. Annotators explicitly identified which relevant clinical features were successfully integrated into the narrative and which were omitted.
- **Hallucination**: the presence of clinical statements or facts that are either unsupported by the original structured data or extraneous to the required narrative. This is a Correctness criterion assessed relative to the input frame of reference. Annotators highlighted the specific segments within the text classified as hallucinations. We categorized hallucinations into four distinct types:
  - *Extra*: Factually correct but non-requested content, such as normal laboratory results or general explanations of clinical features.
  - *Error*: Information that contradicts the patient's medical history or is clinically inaccurate.
  - *Prompt*: Conversational preambles, formulaic introductions, and general model chattiness.
  - *Note*: Meta-commentary or notes generated by the LLM describing the modifications it performed during the task.
- **Narrative Quality**: Rather than serving as a cumulative average of the previous criteria, this metric captures the quality of the text as a unified whole. It maps to the root node of the taxonomy (*Overall Quality*), evaluating how form and meaning interact simultaneously. Evaluators considered the linguistic readability and fluency (Goodness) alongside the aggregate impact of clinical coverage and hallucination density (Correctness). This criterion is mapped to a subjective, intrinsic, absolute evaluation mode, utilizing a global rating scale ranging from 1 (poor) to 5 (excellent).

To ensure the reliability of the evaluation framework, a subset of 45 generated narratives was selected for a preliminary annotation phase. This sample comprised five narratives from each experimental configuration, covering the three distinct LLMs and the

three types of laboratory result inputs. The primary objective of this phase was to refine the annotation guidelines and quantify the **Inter-Tagger Agreement (ITA)** between two independent annotators. Both annotators were computer scientists with previous experience in medical AI research and clinical natural language processing tasks. The evaluation was structured around the previously mentioned three evaluation criteria:

- **Clinical Information Coverage**: Inter-annotator agreement for this criterion was measured by representing each clinical feature as a binary label indicating whether it is present (1) or absent (0) in the generated narrative. Consequently, each annotator produced a binary label vector over the full set of clinical features. Agreement between annotators was then computed by comparing these two label vectors using a **Macro F1-score**, which is appropriate in this setting as it measures consistency across all features while giving equal importance to each label. The analysis yielded a near-perfect score of **0.98**, indicating that the identification of clinical facts is highly consistent and reliable across evaluators.
- **Hallucination**: To establish ITA for the hallucination criterion, we compared the highlighted text boundaries using the **Intersection over Union (IoU) metric.** The resulting score of **0.84** reflects an excellent level of spatial overlap and consistency between evaluators.
- **Narrative Quality:** Given the ordinal nature of this scale, we calculated a **Weighted Cohen's Kappa**, obtaining a value of **0.69**. According to standard benchmarks, this represents substantial agreement, validating the use of these human scores as reliable ground truth for the subsequent experimental analysis.

4.2.2 LLM-as-judge

The LLM-as-a-judge framework is implemented as a multi-agent system designed to mirror the evaluation process used by human experts focusing on the three criteria employed, namely *Clinical Information Coverage, Hallucination, Narrative Quality* (see subsubsection 4.2.1). Its goal is to produce automated quality scores that closely align with expert judgment. To this end, the system comprises three specialized LLM-based judges, each responsible for assessing different aspects of narrative quality.

- **Coverage Judge (Clinical Information Coverage criteria)**: This agent identifies omissions by systematically verifying whether each clinical feature present in the structured data is accurately reflected in the generated narrative.
- **Hallucination Judge (Hallucination criteria)**: This agent detects extraneous or fabricated information. It evaluates the narrative for the presence of "extra" data, such as unrequested normal laboratory results or clinical facts not supported by the source data.
- **Quality Judge (Narrative Quality criteria)**: Acting as the primary evaluator, this agent assigns a final score on a scale of 1–5, simulating the assessment provided by human annotators.

To analyze the interactions among the different criteria, we propose a baseline setup along with several derived combinations (see Figure 2).

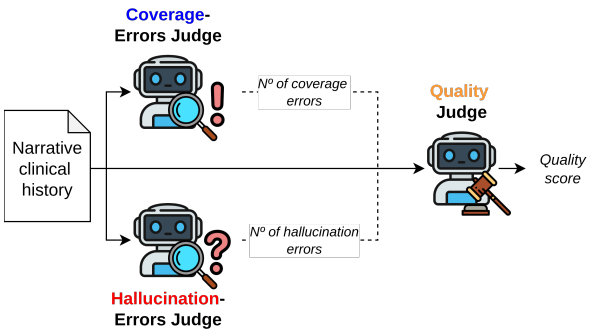

Figure 2: Multi-agent evaluation pipeline using three specialized LLM judges: Coverage (omissions), Hallucination (fabrications), and Quality (final score).

In total, four different evaluation configurations were explored:

- **Guidelines-only (G)**: The Quality Judge assesses narratives based strictly on the predefined annotator guidelines.
- **Guidelines + coverage errors (G+C)**: The Quality Judge integrates the annotator guidelines with the

specific error count identified by the Coverage Judge.

- **Guidelines + hallucination errors (G+H)**: The Quality Judge evaluates the text by combining the guidelines with the frequency of errors detected by the Hallucination Judge.
- **Combined (G+C+H)**: The Quality Judge performs a holistic assessment, incorporating the guidelines alongside error signals from both the Coverage and Hallucination Judges.

This progressive design allows us to analyze whether augmenting the judge with explicit error quantification improves its alignment with human annotators, particularly in terms of inter-annotator agreement.

We use the *gpt-oss-120b* (OpenAI, 2025) and *Qwen3-32B* (Team, 2025) models as automated evaluators. The selection of Qwen3-32B is motivated by prior work (Vasilev et al., 2025), which employs smaller variants of the Qwen family as LLM judges, supporting its suitability for evaluation tasks. In contrast, the choice of *gpt-oss-120b* is driven by the widespread use of GPT-4 (or other form the same family) as evaluators in the literature; however, in order to maintain an open-source setup, we opt for this comparable large-scale alternative. Rather than establishing a single authoritative judge, our framework deliberately incorporates multiple evaluators with varying architectural scales to study agreement patterns and alignment with human expertise across different capability levels. We set the temperature to 0 to ensure deterministic and reproducible outputs (Wei et al., 2024). Crucially, we intentionally select evaluation models that differ from those used in the generation phase to mitigate potential self-preference bias, ensuring a more objective and critical assessment of the clinical narratives.

5 Results and Discussion

This section presents the experimental results, along with a detailed analysis and discussion. We first report the human evaluation results (see subsection 5.1) to address **RQ1**, whether clinical narratives can be automatically generated from longitudinal structured data, and **RQ2**, how different LLMs perform on this task. Finally, we present the LLM-based evaluation results to address **RQ3** (see subsection 5.2).

5.1 Human Evaluation Results

In total, two annotators evaluated 180 narratives (60 per model) based on the criteria detailed in subsubsection 4.2.1. The human evaluation focused exclusively on narratives generated using the "abnormal-only" laboratory results strategy, as annotators observed in the ITA phase that LLMs cannot reliably filter laboratory data.

The results are summarized in Table 2, which details performance across the three primary criteria and provides a categorical breakdown of hallucination errors. The experimental results demonstrate a clear hierarchy in model performance, highlighting a complex trade-off between narrative fluency and clinical accuracy.

Among the evaluated systems, **Marmoka-es+it** emerged as the most reliable model, achieving a superior balance by accurately translating structured data into concise clinical reports without introducing significant distortions. This is reflected in its high Quality score of 3.57 and its comparatively low Clinical Error rate of 1.98 per narrative.

In contrast, while **Llama 3.1** produced fluid and natural-sounding prose, its performance was undermined by severe clinical misconceptions, such as the misidentification of hypercholesterolemia as hypertension. These high-level diagnostic errors resulted in a Quality score of 2.48, failing to surpass the baseline of reliability established by the domain-adapted model.

Medgemma exhibited robust clinical reasoning in some aspects, achieving the lowest Error Coverage at 0.04, but suffered from significant conversational artifacts and high verbosity. This "chattiness" is evidenced by its high scores in Hal. Prompt (2.78) and Hal. Note (2.7), which hinder readability and obscure clinical clarity. Furthermore, Medgemma struggled with temporal logic in the treatment sections, frequently making unfounded assumptions regarding medication status in the absence of an explicit reference date. Consequently, this resulted in a higher Hal. Error rate.

Beyond individual model discrepancies, several systematic error patterns were observed across the entire cohort. Human

Model	Quality (1-5)	Error Coverage	Hal. Error	Hal. Extra	Hal. Prompt	Hal. Note
Llama 3.1	2.48	0.08	2.17	5.15	2.15	0.27
Marmoka-es+it	**3.57**	0.08	**1.98**	5.17	**1.25**	**0.02**
Medgemma	2.17	**0.04**	2.25	**1.88**	2.78	2.7

Table 2: Human Annotation of Generated Narratives. The first column identifies the model evaluated; the second column presents the mean quality score across all narratives. The third column lists the mean percentage of coverage errors. The final four columns detail the specific categories of hallucination errors, indicating the average frequency of occurrence per narrative.

evaluation revealed a persistent tendency for models to either omit relevant abnormal laboratory values entirely or erroneously characterize them as normal results; this suggests that even when abnormal values were explicitly provided, the models continued to filter them based on their internal knowledge. Another recurring trend, particularly prominent in Llama 3.1 and Marmoka-es+it (both exceeding 5.1 in the Hal. Extra category), was the inclusion of extraneous content such as generalized medical explanations. While factually accurate, these elements fell outside the requested scope and contributed to unnecessary verbosity. Nevertheless, we believe that such behaviors could be further mitigated through a more rigorous and constrained prompting strategy.

5.2 LLM-as-a-Judge Results

The results presented in Table 3 and Table 4 reveal a substantial gap between human and LLM-based agreement. While the two human annotators achieved a relatively high ITA of 0.69, the agreement between LLM judges and human annotators remains consistently low across all evaluation settings. Most Cohen's Kappa values fall within the slight to fair agreement range, and in some cases even negative values are observed, suggesting systematic disagreement.

When comparing LLM judges, Qwen shows lower agreement than GPT under the Guidelines-only setting. However, when coverage errors are incorporated into the evaluation, this trend reverses: Qwen achieves higher agreement with human annotators, surpassing GPT, whose performance degrades compared to its Guidelines-only results, regardless of which model is used to compute the coverage errors. This suggests that the inclusion of coverage information affects the two judges differently.

In contrast, when hallucination errors are introduced, agreement decreases for both models, including in the Combined setting. This behavior indicates that both LLM judges struggle to reliably identify and evaluate hallucinations which is specially critical in the medical domain.

Despite the low ITA, we investigated whether a broader consensus existed regarding the relative performance of the models and strategies. We hypothesized that even if judges disagreed on absolute numerical scores, they might still share a consistent perception of which models and prompting strategies outperformed others. However, the results remained inconsistent; the lack of alignment persisted even when evaluating the ordinal ranking of models and strategies.

Given the low agreement observed between the LLM judges and the human annotators, automated evaluations were not used to derive any of the main conclusions of this study. The LLM as a Judge experiments are included only as an exploratory analysis to assess the potential of automated evaluation in this clinical summarization setting.

5.3 Analysis of Laboratory Results Identification

The Laboratory Results Agent requires special attention, as it proved to be the most challenging component and was particularly prone to coverage and hallucination errors. These issues were especially prevalent in narratives generated without filtering the abnormal laboratory results. Consequently, to determine whether these shortcomings stemmed from the narrative synthesis phase or a fundamental limitation in detecting abnormal values, we conducted a targeted analysis isolating the identification task from the generation process. Using LLaMA 3.1, Marmoka-es+it, and MedGemma, we assessed the models' intrinsic diagnostic capabilities by presenting each test description

Annotator	G		G+C				G+H				G+C+H			
			Qwen		GPT		Qwen		GPT		Qwen		GPT	
	s	w	s	w	s	w	s	w	s	w	s	w	s	w
Annotator 1	-0.071	-0.002	0.053	**0.239**	0.012	0.169	0.114	0.135	-0.058	0.080	0.051	0.194	<u>0.129</u>	0.236
Annotator 2	0.104	0.173	<u>0.172</u>	**0.334**	0.044	0.201	0.146	0.161	-0.063	0.018	0.170	0.285	-0.007	0.149

Table 3: Cohen's Kappa agreement (s = simple, w = weighted) for the QWEN model across different evaluation approaches: Guidelines only (G), Guidelines + coverage errors (G+C), Guidelines + hallucination errors (G+H), and Combined (G+C+H). Bold indicates the highest weighted Kappa (w), while underlining indicates the highest simple Kappa (s) for each annotator.

Annotator	G		G+C				G+H				G+H+C			
			Qwen		GPT		Qwen		GPT		Qwen		GPT	
	s	w	s	w	s	w	s	w	s	w	s	w	s	w
Annotator 1	<u>0.171</u>	**0.233**	0.003	0.103	0.044	0.150	0.053	0.058	0.037	0.124	0.053	0.105	0.001	0.058
Annotator 2	<u>0.186</u>	**0.293**	0.063	0.179	0.088	0.135	0.077	0.090	0.037	0.161	0.165	0.191	0.058	0.063

Table 4: Cohen's Kappa agreement (s = simple, w = weighted) for the GPT model across different evaluation approaches: Guidelines only (G), Guidelines + coverage errors (G+C), Guidelines + hallucination errors (G+H), and Combined (G+C+H). Bold indicates the highest weighted Kappa (w), while underlining indicates the highest simple Kappa (s) for each annotator.

alongside its corresponding value and requiring the LLM to classify its status. This experimental design allows for a more granular evaluation of different prompting strategies, independent of the complexities inherent in natural language generation.

The rule-based method which employs predefined clinical thresholds is used as the reference for the specific evaluation. Performance is evaluated using Precision, Recall, and F1-score, taking the rule-based filtering approach as the reference (gold standard). A prediction is considered a true positive when an abnormal value is correctly identified, a false positive when a normal value is incorrectly classified as abnormal, and a false negative when an abnormal value is not detected by the model. The obtained results are available in Table 5:

Model	Precision	Recall	F-score
Llama 3.1	0.43	<u>0.29</u>	<u>0.35</u>
Marmoka-es+it	<u>0.56</u>	0.13	0.21
Medgemma	**0.59**	**0.67**	**0.63**

Table 5: Performance of LLM-based approaches in detecting abnormal laboratory values. Results are reported in terms of Precision, Recall, and F1-score.

The results reveal clear differences across models in terms of precision–recall trade-offs. MedGemma achieves the best overall performance, with the highest F1-score (0.628), combining strong precision (0.589) and recall (0.673). This suggests that clinical domain specialization significantly improves the detection of abnormal laboratory values. In contrast, LLaMA 3.1 and Marmoka-es+it obtain notably low results, demonstrating the limitations of these models in this context. LLaMA 3.1 shows moderate and more balanced performance (F1: 0.345), but its lower precision and recall indicate limitations in handling clinically nuanced data, likely due to its general-domain nature. Marmoka-es+it achieves relatively high precision (0.558) but a very low recall (0.126), indicating a conservative behavior that fails to capture the majority of abnormal values.

Overall, these findings highlight the importance of domain-specific models for clinical tasks. Furthermore, it is important to consider that the reference values included in the rule-based approach are based on the specific hospital population; consequently, borderline values might be counted as errors when they are not clinically severe. This geographic and demographic specificity suggests that the gold standard may not fully capture the nuance of every borderline case.

6 Conclusions

RQ1: Can we effectively transform structured EHR data into faithful and coherent clinical narratives? Our findings demonstrate that a multi-agent approach is a highly effective architecture for bridging the gap between structured tabular data and coherent narratives, managing the complexity of EHR data transformation, though its success is contingent upon strict guidance and domain adaptation.

However, there is a clear trade-off between narrative quality and clinical fidelity. While the multi-agent approach enables models to produce fluid narratives, it remains vulnerable to systematic errors, such as the omission of critical laboratory values or the introduction of conversational artifacts.

RQ2: How does domain and language adaptation impact model performance on this transformation? The human evaluation demonstrates that model performance is significantly influenced by both domain specialization and linguistic alignment. General-domain models, such as LLaMA 3.1, exhibited a higher frequency of clinical errors, highlighting the risks of using non-specialized models for high-stakes medical tasks. Interestingly, while MedGemma showed strong clinical reasoning, Marmoka-es+it was often preferred due to its reduced "chattiness" and superior language adaptation for the target context. These findings suggest that for clinical narrative generation, the ideal model requires a dual-adaptation strategy that combines medical domain expertise with specific linguistic tuning.

RQ3: Do LLM-based judges align with human experts when evaluating clinical summarized narratives? While automated evaluators like gpt-oss-120b and Qwen-32B provide a scalable alternative to manual review, the results show that they do not achieve sufficient agreement with human experts to replace expert assessment in this domain. Our analysis reveals that LLM-based judges tend to be overly optimistic, often overlooking subtle clinical inaccuracies or borderline values that a human expert would flag. Consequently, human expert annotation remains the gold standard for verifying the clinical safety and nuance of generated histories. These findings are consistent with prior work, which also observed discrepancies between LLM judges and expert evaluations Vasilev et al. (2025).

7 Limitations

A key limitation of this study is the small cohort size, which limits clinical variability and the scale of evaluation, further constrained by the need for human annotation due to unreliable automatic evaluation. Despite these constraints, results were consistent across settings with Marmoka-es+it showing stable and systematic improvements over the other models.

A further limitation of this study is the use of institution specific laboratory reference ranges, which may reduce generalizability across healthcare systems. However, the framework is easily adaptable by updating these reference thresholds.

Finally, another limitation of this study is the parameter scale imbalance among generators and the lack of other Spanish-adapted models; future work will include mid-size general models and larger scales to isolate these effects. Moreover, we do not explore task-specific adaptation techniques such as fine-tuning, as this study evaluates pretrained models without additional task-specific training, and due to the limited availability of high-quality annotated clinical data for supervised adaptation. Furthermore, while our pipeline demonstrates feasibility, integrating LLMs into real-world EHR infrastructures remains an open challenge, motivating further research in this direction.

Acknowledgements

This work has been partially supported by the HiTZ Center and the Basque Government, Spain (Research group funding IT1570-22), as well as by MCIN/AEI/10.13039/5011 00011033 Spanish Ministry of Universities, Science and Innovation by means of the projects:

EDHIA PID2022-136522OB-C22 (also supported by FEDER, UE), DeepR3 TED2021-130295B-C31 (also supported by European Union NextGeneration EU/PRTR).

M. Idoyaga has been funded by the predoctoral research grant from the University of the Basque Country (EHU) (PIF24/223).

A. G. Domingo-Aldama has been funded by the Predoctoral Training Program for Non-PhD Research Personnel grant of the Basque Government (PRE_2024_1_0224).

References

AI@Meta. 2024. Llama 3.1 model card.

Alzubi, A. A., V. J. Watzlaf, and P. Sheridan. 2021. Electronic health record (ehr) abstraction. *Perspectives in health information management 18*(Spring).

Belz, A., S. Mille, and C. Thomson. 2025. The qcet taxonomy of standard quality criterion names and definitions for the evaluation of nlp systems. *arXiv preprint arXiv:2509.22064*.

Contreras, M., S. Kapoor, J. Zhang, A. Davidson, Y. Ren, Z. Guan, T. Ozrazgat-Baslanti, J. Sena, S. Nerella, A. Bihorac, et al.. 2025. A large language model for delirium prediction in the intensive care unit using structured electronic health records. *Scientific Reports 15*(1), 38890.

Domingo-Aldama, A. G., I. De La Iglesia, M. Urruela, A. Atutxa, and A. Barrena. 2026. To adapt or not to adapt: Rethinking the value of medical knowledge-aware large language models. *arXiv preprint arXiv:2604.06854*.

Falis, M., A. P. Gema, H. Dong, L. Daines, S. Basetti, M. Holder, R. S. Penfold, A. Birch, and B. Alex. 2024. Can gpt-3.5 generate and code discharge summaries? *Journal of the American Medical Informatics Association 31*(10), 2284–2293.

Ganzinger, M., N. Kunz, P. Fuchs, C. K. Lyu, M. Loos, M. Dugas, and T. M. Pausch. 2025. Automated generation of discharge summaries: leveraging large language models with clinical data. *Scientific reports 15*(1), 16466.

Google. 2025. Medgemma technical report.

Holmes, J. H., J. Beinlich, M. R. Boland, K. H. Bowles, Y. Chen, T. S. Cook, G. Demiris, M. Draugelis, L. Fluharty, P. E. Gabriel, et al.. 2021. Why is the electronic health record so challenging for research and clinical care? *Methods of information in medicine 60*(01/02), 032–048.

Kruse, M., S. Hu, N. Derby, Y. Wu, S. Stonbraker, B. Yao, D. Wang, E. Goldberg, and Y. Gao. 2025. Zero-shot large language models for long clinical text summarization with temporal reasoning. *medRxiv*, 2025–07.

Lee, S. A., S. Jain, A. Chen, K. Ono, A. Biswas, Á. Rudas, J. Fang, and J. N. Chiang. 2025. Clinical decision support using pseudo-notes from multiple streams of ehr data. *npj Digital Medicine 8*(1), 394.

Li, W., H. Feng, C. Hu, et al.. 2026. Accurate discharge summary generation using fine tuned large language models with self evaluation. *Scientific Reports 16*, 5607.

Lin, C.-Y.. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

OpenAI. 2025. gpt-oss-120b gpt-oss-20b model card.

Papineni, K., S. Roukos, T. Ward, and W.-J. Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Team, Q.. 2025. Qwen3 technical report.

Vasilev, Y., I. Raznitsyna, A. Pamova, T. Burtsev, T. Bobrovskaya, P. Kosov, A. Vladzymyrskyy, O. Omelyanskaya, and K. Arzamasov. 2025. Evaluating medical text summaries using automatic evaluation metrics and llm-as-a-judge approach: A pilot study. *Diagnostics 16*(1), 3.

Wei, H., S. He, T. Xia, F. Liu, A. Wong, J. Lin, and M. Han. 2024. Systematic evaluation of llm-as-a-judge in llm alignment tasks: Explainable metrics and diverse prompt templates. *arXiv preprint arXiv:2408.13006*.

Woo, B. F. Y., K. Cato, H. Cho, S. B. You, and J. Song. 2025. The use of large language models in clinical documentation: A scoping review. *International Journal of Nursing Studies*, 105322.

A Dataset

To provide a comprehensive overview of the cohort profile, Table 6 details both the structural clinical composition of the dataset and the linguistic properties of its underlying entities. The upper section of the table outlines the distribution of clinical features recorded per patient, highlighting the variance in clinical history complexity (ranging from a mean of 2.46 past diagnoses to highly detailed laboratory profiles averaging over 22 tests per individual).

To further contextualize the textual complexity of the data for downstream natural language processing tasks, the lower section breaks down the vocabulary of unique clinical concepts. For each distinct category (55 diagnoses, 167 treatments, and 314 laboratory components), we provide the minimum, maximum, and average lengths in both word and character counts. This dual metric exposes the lexical density of the clinical terminology, showing that while laboratory components are highly diverse, they remain concise (averaging 2.48 words), whereas diagnoses and treatment descriptions exhibit a higher semantic length, averaging approximately 5 words per concept.

Features	**Min**	**Max**	**Mean**
Past Diagnoses	1	6	2.46
Laboratory tests	1	87	22.34
Treatments	0	50	11.23
Components (per test)	1	135	25.29
Unique concepts & lengths			
Diagnoses (*Total unique: 55*)			
Words	1	15	5.27
Characters	8	101	37.64
Treatments (*Total unique: 167*)			
Words	4	10	4.95
Characters	25	72	35.77
Laboratory components (*Total unique: 314*)			
Words	1	8	2.48
Characters	6	59	22.79

Table 6: Dataset characteristics and text length statistics of the 65 patients.

B Prompts

B.1 Generation prompts

B.1.1 Laboratory Results

Regarding the three prompting strategies explored for laboratory results ("No-Filter", "Abnormal-explicit", and "Abnormal-Only"), the system message remains identical across all settings. The only variation is the content injected into the input variable (*pruebas*). In the "No-Filter" configuration, *pruebas* contains all laboratory values without further distinction. In the "Abnormal-explicit" setting, all values are provided but abnormal ones are explicitly flagged. Finally, in the "Abnormal-Only" configuration, only the out-of-range values are included in the prompt.

Past Medical History

System message
Eres un experto clínico. Resume claramente los resultados de laboratorio del paciente a partir de un JSON, incluyendo solo los valores anormales o no típicos en español. Normaliza los nombres de las pruebas y las unidades, y elimina abreviaturas o detalles técnicos innecesarios.

User message
Resultados de laboratorio del paciente:
{`pruebas`}
Summary:

B.1.2 Past Medical History

For the Past Medical History section, the input variable (*antecedentes*) contains all diagnoses recorded throughout the patient's lifetime.

Past Medical History

System message
Eres un experto en documentación clínica. Resume en español la historia médica del paciente a partir de un documento JSON de forma cronológica. Normaliza la terminología: convierte abreviaturas, elimina palabras técnicas innecesarias y formatea las enfermedades en una forma estándar y legible (por ejemplo, 'DIABETES M TIPO II O NEOM, SIN MENCION COMPL. CONTROL NEOM' → 'Diabetes M Tipo 2').

User message
Antecedentes médicos personales del paciente:
{`antecedentes`}
Resumen:

B.1.3 Treatments

For the Treatments section, the input variable (*tratamientosActivos*) contains all treatments prescribed after the date of the patient's most recent diagnosis.

Treatments

System message
Eres un farmacéutico clínico. Proporciona un resumen claro y conciso de la medicación del paciente en español a partir de un JSON. Utiliza nombres de medicamentos normalizados y legibles, eliminando abreviaturas, códigos y detalles técnicos innecesarios."

User message
Tratamientos del paciente:
{`tratamientosActivos`}
Resumen:

B.2 Evaluation prompts

We define a full evaluation prompt consisting of a system message and a structured user message. The system message instructs the model to act as a clinical expert and to evaluate AI-generated clinical summaries in Spanish, prioritizing clarity and readability, while also considering clinical accuracy and completeness. The user message includes the clinical case, the generated summary, and optional metadata in the form of coverage and hallucination error counts.

From this full configuration ((iv) Combined (G+C+H)), three additional variants are derived by selectively removing parts of the error metadata: (i) Guidelines-only (G), (ii) Guidelines + coverage errors (G+C), and (iii) Guidelines + hallucination errors (G+H). This allows us to study the impact of providing different levels of error feedback on the evaluation behavior of the model.

Evaluation Prompt

System message
You are a medical expert evaluating the quality of AI-generated clinical summaries.
The clinical case and the summaries are written in Spanish.
Your main evaluation criterion is **clarity and readability** of the summary.
You should mainly consider:

- Is the text clear and easy to understand?
- Is the information well structured?
- Is the wording concise and appropriate for a clinical summary?

You must penalize unnatural formatting or assistant-like phrasing. This includes:

- Pseudo-JSON structures
- Excessive bullet points
- Introductory or closing phrases such as:
  - "aquí tienes el resumen"
  - "sure, here is"
  - "here you have your summary"
  - Similar meta-commentary

The summary should be written as a direct clinical text.
Secondarily consider:

- Clinical accuracy (only penalize if there are clear medical errors compared to the clinical case)
- Completeness (only penalize if important information from the clinical case is missing)

You will be provided with the count of coverage and hallucination errors detected in the summary; use this as a primary factor.
Score the summary from 1 to 5.
Score meaning:

- 1 = Very poor (confusing or incorrect)
- 2 = Poor (hard to understand or with major issues)
- 3 = Acceptable but improvable
- 4 = Clear and well written
- 5 = Very clear, concise, and well organized

If the summary section is empty, assign score 1.
Return ONLY the number (1, 2, 3, 4, or 5).

User message Clinical case (Spanish):
{`caso_clinico`}
Summary section to evaluate (tipo):
{`resumen`}
Number of coverage errors detected: {`coverage`}
Number of hallucination errors detected: {`hallucination`}
Score: