SimpliMED: Automatic Simplification of Cardiology Discharge Reports Using Large Language Models

Lucas Molino Piñar[1], **Manuel Carlos Díaz Galiano**[1], **María Teresa Martín Valdivia**[1], **José Ángel Urbano-Moral**[2], **Elena Sola-García**[2]
[1]SINAI Research Group, University of Jaén, Spain
[2]ReCaH Research Group, University Hospital of Jaén, Spain
[1]{lmolino, mcdiaz, maite}@ujaen.es, [2]{jaurbanomoral, esola.cardiologia}@gmail.com

Abstract

Medical discharge reports frequently contain highly technical language that creates significant communication barriers between healthcare professionals and patients, potentially compromising treatment adherence and post-discharge care quality. In this paper, we present SimpliMED, a modular system designed to automatically simplify cardiology discharge reports using Large Language Models (LLMs) and advanced Natural Language Processing techniques (NLP). Our architecture integrates section-based preprocessing with specialized prompts, explicit handling of medical abbreviations, and therapeutic explanations of medications to enhance accessibility. We evaluate our system using a corpus of 307 anonymized cardiology discharge reports from a Spanish medical center. For abbreviation detection, our fine-tuned Small Language Model (SLM) achieves an F1-score of 0.90, significantly outperforming regex-based approaches (F1: 0.67). For medication recognition, we achieve F1-scores of 0.91 for commercial names and 0.70 for active principles. We also contribute a therapeutic dictionary containing 14,611 medications with patient-friendly explanations extracted from the Spanish Agency of Medicines. Expert evaluation by two cardiologists yields an overall quality score of 75%, with highest performance for admission reason (91%) and current illness (75%) sections. While results demonstrate the potential of LLM-based medical text simplification for Spanish clinical language, we identify areas requiring further development before clinical deployment.

Keywords: Medical text simplification, clinical NLP, large language models, discharge reports, patient accessibility, Spanish clinical language, plain language

1. Introduction

Effective medical communication is fundamental to the relationship between healthcare professionals and patients. However, hospital discharge reports frequently employ specialized technical language that hinders comprehension by patients and their families (Kandula et al., 2010). This communication barrier not only generates uncertainty but can also compromise treatment adherence and subsequent medical follow-up. When patients leave the hospital without fully understanding their diagnosis, prescribed treatments, or warning signs requiring immediate attention, the consequences can range from medication errors to delayed recognition of serious complications.

This challenge is particularly relevant given that a significant portion of the population has limited reading competence (OECD, 2023). The demand for accessible texts has led to initiatives such as Easy-to-Read guidelines (AENOR, 2018) and the Plain Language movement (Adler, 2012). Among the first systems addressing text simplification for Spanish, Bott and Saggion (2012) proposed an automatic approach targeting e-Accessibility. Recently, the Spanish Organic Law 5/2024 on the Right of Defense mandates that communications be drafted in clear, universally accessible language, highlighting the growing legal recognition of text accessibility.

Recent advances in NLP, particularly the development of LLMs (Brown et al., 2020), open new possibilities for addressing this challenge. These tools are capable of transforming complex texts while maintaining their precise content and adapting the register and complexity of language according to the receiver's needs (Joseph et al., 2023). The emergence of the Transformer architecture (Vaswani et al., 2017) revolutionized the field, enabling models that capture long-range dependencies and generate coherent, contextually appropriate text.

In this paper, we present SimpliMED, an automated system designed to transform cardiology discharge reports into versions that are accessible and comprehensible to patients. Cardiology reports present particular challenges: dense technical terminology, numerous abbreviations with context-dependent meanings, complex medication regimens, and temporal narratives describing clinical evolution. Through advanced NLP techniques, we seek to develop a solution that preserves the integrity of medical information while facilitating its interpretation.

Our main contributions include a modular architecture for medical report simplification combining LLMs with specialized modules for abbreviation and medication handling (Figure 1), featuring section-

Proceedings of the Third Workshop on Patient-Oriented Language Processing (CL4Health), pages 72–81
12 May 2026.

based preprocessing that enables tailored simplification for each report component. We developed two fine-tuned Small Language Models (SLMs) achieving F1-scores of 0.90 for clinical abbreviation detection and 0.91 for medication recognition, as well as a therapeutic dictionary containing 14,611 medications with simplified explanations extracted from the AEMPS (*Agencia Española de Medicamentos y Productos Sanitarios*—Spanish Agency of Medicines and Medical Devices)[1]. Finally, we present expert clinical validation by cardiologists with structured evaluation methodology.

The remainder of this paper is organized as follows: Section 2 reviews related work in text simplification. In section 3 we describe the SimpliMED system architecture. Section 4 presents the experimental setup, showing the results in Section 5. Section 6 discusses limitations. Finally we conclude with Section 7.

2. Related Work

The field of automatic text simplification has undergone a paradigm shift in recent years, evolving from early rule-based systems to the current dominance of Large Language Models (LLMs). As detailed in the systematic review by Bednarczyk et al. (2025), this evolution reflects a transition from rigid syntactic operations towards semantic-aware generative approaches. Their analysis of 45 recent studies concludes that LLMs significantly outperform previous architectures in clinical summarization and simplification tasks, particularly when fine-tuned on domain-specific corpora.

Historically, simplification relied on manual rules and synonym replacement (Carroll et al., 1998), an approach that, while interpretability-friendly, struggled with the rich variability of natural language. The statistical machine translation era (Wubben et al., 2012) and subsequent neural sequence-to-sequence models (e.g., *T5*) (Botarleanu et al., 2020) improved fluency but often required massive parallel corpora, Simplification/Original pairs, which are notoriously scarce in specialized domains like medicine.

In the healthcare context, simplification is not merely a readability task but a patient safety necessity. Kandula et al. (2010) identified early on that medical simplification requires handling not just vocabulary but also syntactic complexity and logical explanations. Unlike general domain simplification, medical applications must prioritize content preservation above all else. Bednarczyk et al. (2025) emphasize that while general-purpose models (like *GPT-4* or *LLaMA*) exhibit strong zero-shot capabilities, superior performance in clinical settings is often achieved through specialized fine-tuning. The comprehensive survey by Ondov et al. (2022) further systematizes automated methods for biomedical text simplification, identifying key challenges including terminology handling, evaluation methodology, and the scarcity of domain-specific parallel corpora.

Recent work addresses simplification from complementary perspectives. Rust et al. (2025) demonstrated that LLMs can effectively simplify discharge summaries while augmenting them with lifestyle recommendations. Vásquez-Rodríguez et al. (2023) proposed document-level simplification with coherence evaluation, highlighting the importance of maintaining discourse structure. Vásquez-Rodríguez et al. (2021) further investigated evaluation methodologies for text simplification, revealing limitations of standard metrics. Notably, Xia et al. (2025) introduced JEBS, a fine-grained biomedical lexical simplification task that explicitly separates the identification of hard-to-understand terms from the simplification process, an approach conceptually aligned with our modular architecture. However, Joseph et al. (2023) warn that these models require rigorous domain-specific evaluation beyond standard metrics like BLEU (Papineni et al., 2002) or SARI (Xu et al., 2016), which correlate poorly with clinical accuracy.

Despite these global advances, resources for Spanish medical simplification remain limited compared to English. The CLARA-MeD project (Campillos-Llanos et al., 2022) represents a significant milestone, providing the first comparable corpus for Spanish medical text simplification, yet the ecosystem lacks the robust pipeline tools available for English. Most existing Spanish systems rely on direct translation of English models or general-purpose LLMs without clinical grounding. Our work addresses this gap by proposing a modular architecture that explicitly separates abbreviation and medication handling from the simplification generation, utilizing the generative power of modern LLMs while imposing the structural constraints necessary for clinical reliability.

3. Methodology

This section presents the SimpliMED system architecture, including model selection rationale, the specialized modules for abbreviation and medication handling, and the prompt engineering strategy.

3.1. System Architecture

SimpliMED follows a modular pipeline architecture (Figure 1) explicitly designed to address the linguistic challenges of clinical documentation through four integrated components: preprocessing and

[1] https://cima.aemps.es/

segmentation, abbreviation detection, medication recognition, and LLM-based simplification.

Our approach prioritizes the structural modularity of clinical reports, recognizing that each section serves a distinct communicative function (Denny et al., 2009). For instance, the admission reason typically introduces the clinical episode concisely, whereas diagnostic tests present technical results requiring detailed explanation. Consequently, the system processes each section independently using specialized prompts (Figure 2). This granular processing strategy allows the incorporation of appropriate levels of detail, structural formatting, and explanatory depth specifically suited to the unique requirements of each section type, ensuring higher quality simplification compared to global processing approaches.

To further enhance accessibility, the architecture incorporates specialized mechanisms for handling technical terminology. Medical reports frequently employ abbreviations like AF (Atrial Fibrillation—abbreviated as "FA" in Spanish), HT (HyperTension—abbreviated as "HTA" in Spanish), and LVEF (Left Ventricular Ejection Fraction—abbreviated as "FEVI" in Spanish) which create significant barriers for patients. The system addresses this issue by explicitly detecting and expanding these terms, maintaining the original abbreviation in parentheses to facilitate learning while ensuring comprehension. Similarly, treatment sections often present medications in condensed formats with cryptic dosing schedules. The system enriches this information by identifying medications and retrieving their specific therapeutic purposes (e.g., "anticoagulation", "lipid control"), thereby providing patients with essential context about their treatment regimen.

The implementation of this pipeline begins with preprocessing and segmentation, where raw reports, often containing digitization artifacts, are transformed into structured text. This stage handles irregularities and divides the content into up to ten standard clinical sections including admission reason, medical history, current illness, events during admission, diagnostic tests, surgical interventions, discharge diagnosis, treatment, discharge plan, and follow-up. Following segmentation, the text flows through the abbreviation and medication modules (detailed in Section 3.4 and Section 3.5 respectively) before the final LLM-based simplification phase (Section 3.6).

3.2. Model Selection

After preliminary evaluation of multiple architectures including *LLaMA 2/3* variants (1B–8B parameters), *Mistral 7B*, *Gemma 7B*, *Phi-2*, and OpenAI *GPT* models (3.5/4/4o), we selected *LLaMA 3.1 8B Instruct* as our base model. The selection was driven primarily by privacy requirements: sensitive medical data requires local processing without external API calls, and healthcare data protection regulations prohibit transmission of patient information to third-party cloud services. Consequently, non-open-weight models were discarded for production use. While OpenAI models were explored in preliminary phases for methodology development, they were not used for processing real clinical reports.

The model demonstrated exceptional contextual sensitivity, maintaining coherence across extended reports and preserving relationships between sections, crucial when findings in one section depend on information presented elsewhere, such as when the discharge diagnosis must be understood in the context of diagnostic test results. *LLaMA 3.1 8B* also showed basic clinical reasoning capacity sufficient to identify causal relationships between symptoms, findings, and recommendations, which proved adequate for the simplification task that primarily requires reformulation rather than diagnosis. Finally, the 8B parameter model offers an optimal balance between capability and computational resources for clinical deployment scenarios, with inference performed on a single GPU within acceptable time constraints.

3.3. Prompt Engineering Strategy

The design of prompts for each section follows a systematic architecture where the model assumes the role of an expert in simplifying cardiology reports for elderly patients with limited medical training. Each prompt specifies simplification rather than summarization, maintaining semantic integrity while ensuring comprehensibility. The prompts emphasize concision through short sentences, avoid unnecessary jargon, and require a formal third-person style with respectful tone.[2]

A key strategy involves preserving medical terms in parentheses when simplifying technical vocabulary, allowing patients to recognize original terminology. Semantic fidelity remains the main criterion: any reformulation must preserve clinical meaning. The framework provides homogeneous structure while allowing section-specific nuances, for admission reason or medical history, causal clarity is reinforced; for diagnostic tests or treatment, the need to explain technical terms is emphasized.

3.4. Abbreviation Detection Module

For abbreviation detection, we compared two approaches. The first is a regex-based detection approach using the pattern `(?<![a-zA-Z0-`

[2]The complete code including all section-specific prompts is available at https://github.com/lmolino03/simplimed.

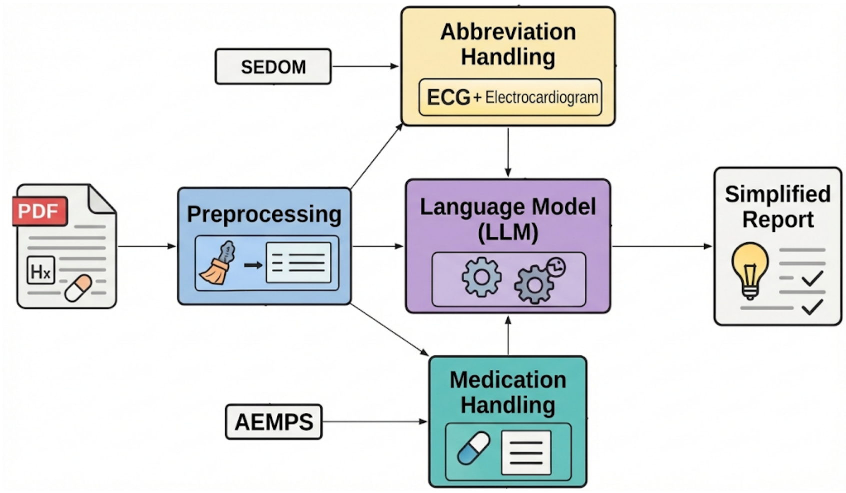

Figure 1: SimpliMED system architecture. Clinical reports are preprocessed, segmented into sections, enriched with abbreviation meanings and medication information, then simplified using specialized prompts per section type.

MOTIVO DE INGRESO
Dolor torácico.
ANTECEDENTES
HTA en tratamiento. DM tipo 2. Dislipemia. Exfumador desde hace 5 años.
ENFERMEDAD ACTUAL
Varón de 68 años que acude por dolor torácico típico con cortejo vegetativo.
EVOLUCIÓN Y CURSO CLÍNICO
Paciente que ingresa en UCI coronaria. Se realiza coronariografía urgente con implante de stent en DA proximal.
PRUEBAS COMPLEMENTARIAS
ECG: elevación ST en V1-V4. Troponinas elevadas. Ecocardiograma: FEVI 45%.
INTERVENCIÓN - PROCEDIMIENTOS QUIRÚRGICOS
Coronariografía con implante de stent farmacoactivo en DA proximal.
JUICIO CLÍNICO
SCACEST anterior. Enfermedad coronaria de un vaso (DA proximal).
TRATAMIENTO
AAS 100mg 1-0-0, Clopidogrel 75mg 1-0-0, Atorvastatina 80mg 0-0-1.
PLAN DE ACTUACIÓN
Rehabilitación cardíaca. Control de factores de riesgo cardiovascular.
REVISIONES
Consulta de Cardiología en 4 semanas. Control analítico previo.

(a) Original Report (Spanish)

REASON FOR ADMISSION
Chest pain.
MEDICAL HISTORY
HTA under treatment. Type 2 DM. Dyslipidemia. Ex-smoker for 5 years.
CURRENT ILLNESS
68-year-old male presenting with typical chest pain with autonomic symptoms.
EVENTS DURING ADMISSION
Patient admitted to coronary ICU. Urgent coronary angiography performed with stent implantation in proximal LAD.
DIAGNOSTIC TESTS
ECG: ST elevation in V1-V4. Elevated troponins. Echocardiogram: LVEF 45%.
INTERVENTION - SURGICAL PROCEDURES
Coronary angiography with drug-eluting stent implantation in proximal LAD.
DISCHARGE DIAGNOSIS
Anterior STEMI. Single-vessel coronary disease (proximal LAD).
TREATMENT
ASA 100mg 1-0-0, Clopidogrel 75mg 1-0-0, Atorvastatin 80mg 0-0-1.
DISCHARGE PLAN
Cardiac rehabilitation. Cardiovascular risk factor control.
FOLLOW-UP
Cardiology consultation in 4 weeks. Prior bloodwork.

(b) Translated Report for legibility reason(English)

Figure 2: Example of a cardiology discharge report structure showing typical sections processed by SimpliMED.

`9])([A-Z]{2,})(?![a-zA-Z0-9])` to identify sequences of two or more uppercase characters not adjacent to alphanumerics, with detected candidates validated against the SEDOM (*Sociedad Española de Documentación Médica*—Spanish Society of Medical Documentation) dictionary[3] containing 7,054 medical abbreviation entries. The second approach uses a *LLaMA 3.2 3B Instruct* model fine-tuned on a *GPT-4o*-annotated corpus, where the training corpus was generated from 70% of the clinical reports with *GPT-4o* identifying all abbreviations present. Fine-tuning configuration included batch size 2, 3 epochs, learning rate 1×10^{-4}, maximum sequence length 2,048 tokens, and gradient accumulation steps 4.

The SLM approach offers advantages over regex because it considers semantic context rather than purely orthographic patterns. For example, it can distinguish "DE" as "*Desviación Estándar*" (Standard Deviation) in a laboratory results context versus the Spanish preposition "*de*" (of/from) appearing in uppercase text. For abbreviations with multiple possible meanings, we delegate disambiguation to the simplification LLM, where the prompt includes the detected abbreviation along with its possible meanings from the SEDOM dictionary, requesting the model to select the contextually appropriate interpretation based on the full report content.

[3] `https://sedom.es/`

3.5. Medication Recognition Module

Medication recognition relies exclusively on a fine-tuned SLM, deliberately excluding regex-based approaches. This design decision stems from the high variability in how medications appear in clinical text: misspellings, informal abbreviations, and non-standard forms are common in handwritten or rapidly dictated reports.

The *LLaMA 3B* model was fine-tuned for medication and active principle recognition using the same training configuration as the abbreviation module. The annotated corpus was generated using *GPT-4o* to identify all medication mentions in the training set.

To provide meaningful explanations of medication purposes, we constructed a therapeutic dictionary from the AEMPS database. We collected 14,611 authorized medications in Spain, extracting information from both patient leaflets and technical specifications. For each medication, *GPT-4o mini* extracted the therapeutic use from Section 1 of patient leaflets (Accessible language) and Section 4.1 of technical specifications (Clinical indications), resulting in a dictionary that links each medication and active principle to a concise therapeutic explanation. This dictionary is publicly available[4].

Medication lookup uses Levenshtein distance for fuzzy matching to handle spelling variations and typos common in clinical text. For combined medications with multiple active principles, the system searches for entries containing all detected components.

3.6. LLM-Based Simplification Module

The final simplification stage integrates the outputs of the abbreviation and medication modules into a structured prompt for *LLaMA 3.1 8B Instruct*. For each section, the prompt includes: (1) the original clinical text, (2) a list of detected abbreviations with their expanded meanings as provided by the abbreviation module, and (3) for treatment sections, the identified medications with their therapeutic purposes retrieved from the AEMPS dictionary.

The model processes each section independently using section-specific prompts. In the "with module" configuration, the LLM receives preprocessed abbreviation expansions and medication information, enabling it to produce accurate simplifications. In contrast, the "without module" baseline relies solely on the LLM's internal knowledge to interpret abbreviations and medications during simplification, without any prior detection or dictionary lookup step.

[4] https://github.com/sinai-uja/DICCIONARIO_MEDICAMENTOS

4. Experimental Setup

4.1. Corpus Description

Our corpus comprises 307 anonymized cardiology discharge reports from University Hospital Jaen (Spain), provided under an institutional collaboration framework and anonymized following data protection protocols. This cardiology-focused dataset presents thematic and terminological coherence that reduces linguistic variability, enabling development of focused systems that are easier to control and evaluate.

Reports follow a standard structure with up to ten sections (Table 1), though not all sections appear in every report. Sections are explicitly labeled in the reports using recognizable headers, which our preprocessing module detects through pattern matching. The most frequent sections are current illness, medical history, and diagnostic tests, appearing in approximately 95% of reports, while surgical intervention sections appear only in 42 of 307 reports (14%).

Section	Freq.	Avg. Tokens
Admission reason	98%	∼5
Medical history	95%	∼85
Current illness	92%	∼70
Discharge diagnosis	90%	∼45
Treatment	88%	∼120
Diagnostic tests	75%	∼180
Events during admission	70%	∼150
Discharge plan	65%	∼50
Follow-up	45%	∼25
Surgical intervention	14%	∼80

Table 1: Report section frequency and average token count.

Section length varies considerably by type. Admission reason sections are typically brief (median ∼5 tokens), providing concise statements of the reason for hospitalization. In contrast, diagnostic test sections show the highest average length (∼180 tokens), reflecting the need to document multiple diagnostic study results including numerical values, reference ranges, and detailed technical descriptions.

4.2. Abbreviation Analysis

Analysis of the corpus reveals systematic patterns in abbreviation usage. The most frequent abbreviations include cardiovascular-specific terms such as AF, HT, and LVEF (abbreviated as "FA, HTA, and FEVI" in Spanish).

The abbreviation density varies considerably by section type. Treatment sections show the highest abbreviation density, with an average of 8.3 abbreviations per 100 tokens, reflecting the condensed nature of medication prescriptions. In contrast, admission reason sections show lower density (2.1 per 100 tokens), as they typically use more complete clinical descriptions. This pattern suggests that sections requiring quick clinical reference tend to use more abbreviated forms, while patient-facing sections, those most likely to be read and interpreted by patients, such as admission reason, discharge plan, and follow-up, maintain more explicit terminology.

4.3. Evaluation Methodology

Given the absence of established automatic metrics for medical text simplification, we employ a combination of automatic evaluation for subtasks and expert medical evaluation for overall simplification quality.

For automatic evaluation of subtasks, we computed Precision, Recall, and F1-score for abbreviation detection and medication recognition. The corpus was split into 70% for training (215 reports) and 30% for testing (92 reports), with the test set held out entirely from fine-tuning. The test set was annotated using *GPT-4o*, which was prompted to identify all abbreviations and medication mentions in each report. These annotations were then manually reviewed by the first author to correct errors, particularly false positives from non-medical uppercase acronyms and missed active principles. While we acknowledge that using LLM-generated annotation, even after manual revision, introduces potential bias when evaluating other language models, this approach was necessitated by the lack of expert-annotated clinical NLP resources for Spanish, and the cost of full manual annotation by clinicians.

For expert clinical evaluation, two cardiologists evaluated simplified versions of 8 randomly selected reports. Evaluation was performed by section using a 3-level scale where 0 indicates incorrect simplifications containing factual errors or misinterpretations that could mislead patients, 0.5 indicates partial success where the simplification is generally acceptable but contains minor issues or room for improvement, and 1 indicates correct simplifications that accurately convey the original information in accessible language suitable for patients. Additionally, experts provided qualitative feedback including observations, improvement suggestions, and general comments on simplification quality. This dual approach captures both structured ratings and nuanced clinical perspectives that numerical scores cannot fully represent.

5. Results

This section presents quantitative results for the specialized subtasks (abbreviation detection, medication recognition) and qualitative results from expert evaluation of the complete simplification system.

5.1. Abbreviation Detection

The fine-tuned SLM significantly outperforms the regex-based approach (Table 2).

System	Precision	Recall	F1
Regex-based	0.63	0.77	0.67
SLM (fine-tuned)	**0.89**	**0.93**	**0.90**

Table 2: Abbreviation detection performance comparison.

The SLM achieves 23 percentage points higher F1-score than the regex approach, which suffers from lower precision (0.63) due to false positives from uppercase text segments that are not abbreviations.

5.2. Medication Recognition

Table 3 presents results for medication and active principle recognition.

Category	Precision	Recall	F1
Active substances	0.70	0.71	0.70
Brand names	**0.91**	**0.91**	**0.91**
Combined (union)	0.89	0.93	0.90

Table 3: Medication recognition performance by category.

Medication brand names achieve the highest F1 (0.91), while active substances show lower performance (0.70) due to category confusion. The combined evaluation achieves F1 of 0.90, demonstrating strong overall medication recognition.

5.3. Expert Evaluation

Two cardiologists independently evaluated simplified versions of 8 randomly selected reports, scoring each section on a 0–1 scale where 0 indicates incorrect simplifications with factual errors, 0.5 indicates partial success with minor issues, and 1 indicates correct simplifications suitable for patients. Table 4 summarizes the evaluation scores across all sections.

Performance varies considerably across section types. Admission reason achieves the highest score (0.91), reflecting brief, factual content requiring minimal interpretation, while diagnostic tests

Section	Score
Admission reason	0.91
Current illness	0.75
Discharge plan	0.71
Medical history	0.69
Intervention	0.67
Discharge diagnosis	0.56
Follow-up	0.56
Treatment	0.53
Events during admission	0.50
Diagnostic tests	0.44
Global Score	**0.75**

Table 4: Expert evaluation scores by section (n=8 reports, scale 0-1).

receives the lowest (0.44), as technical laboratory values and electrocardiogram readings require clinical interpretation beyond current model capabilities. The global score of 0.75 indicates reasonable quality, though significant improvement remains necessary before clinical deployment.

Analysis of inter-rater agreement yields a weighted Cohen's Kappa of 0.72 (Landis and Koch, 1977), indicating substantial agreement, with 73% of ratings achieving exact agreement. The results suggest the model excels at reformulating declarative statements but struggles with content requiring domain expertise for interpretation.

5.4. Qualitative Examples

The following examples illustrate how each specialized module contributes to improved simplification. Below we present the examples with bilingual support.

Abbreviation expansion. Figure 3 contrasts correct vs. incorrect abbreviation handling.

Medication explanations. Figure 4 shows how therapeutic purpose explanations improve treatment comprehension.

Abbreviation Expansion Comparison

Note: Abbreviation handling is optimized for Spanish.

Spanish:

Original: *"Antecedentes de HTA y FA permanente. Ecocardiograma con FEVI preservada."*

Without module: "Antecedentes de **Hiper-Tensión Área** y **Factor de Angustia** permanente. Ecocardiograma con **Factor Evidente** preservada."

With module: "Antecedentes de **Hipertensión Arterial (HTA)** y **Fibrilación Auricular (FA)** permanente. Ecocardiograma con **Fracción de Eyección del Ventrículo Izquierdo (FEVI)** preservada."

English:

Original: *"History of HT and permanent AF. Echocardiogram showing preserved LVEF."*

Without module: "History of **Hyper-Tension Area** and permanent **Anguish Factor**. Echocardiogram with **Evident Factor** preserved."

With module: "History of **Arterial Hypertension (HT)** and permanent **Atrial Fibrillation (AF)**. Echocardiogram showing preserved **Left Ventricular Ejection Fraction (LVEF)**."

Figure 3: Without module: incorrect expansions. With module: correct clinical meanings.

6. Discussion

This section evaluates the system's performance, analyzes error patterns, and identifies future development directions.

6.1. System Performance

SimpliMED demonstrates the effectiveness of a modular pipeline and domain-adapted Small Language Models (SLMs). Our fine-tuned models for abbreviation detection and medication recognition achieved F1 scores of 0.90 and 0.91, respectively, proving that specialized models can be both efficient and accurate. Integrating validated resources like SEDOM and AEMPS reduces hallucinations, while expert evaluation by cardiologists provides deeper insights than standard readability metrics. However, a key limitation is the lack of direct patient testing, which remains necessary to fully validate comprehension improvements.

The use of *LLaMA 3.1 8B* involves a trade-off between quality and computational cost. While requiring significant GPU resources, this model size offers a balance that smaller models cannot yet match. Furthermore, our section-based segmentation approach ensures consistency and prevents information loss, highlighting the importance of structural awareness in clinical text processing.

6.2. Error Analysis

Critical error patterns included incorrect abbreviation disambiguation (e.g., confusing portal and pulmonary hypertension) and incomplete explanations for combination medications. Occasional factual hallucinations in complex clinical interpretations (like Tako-Tsubo syndrome) underscore the limitations of current model reasoning in technical sections.

Future iterations will prioritize Retrieval-Augmented Generation (RAG) to ground outputs in

Treatment Section with Therapeutic Explanations

Spanish:

Original: *"Apixaban 5mg 1-0-1; Aspirina 100mg 1-0-0; Atorvastatina 80mg 0-0-1"*

With treatment module:

- **Anticoagulación:** Apixaban 5mg. Tomar por la mañana y por la noche. Sirve para prevenir coágulos.
- **Antiagregante:** Aspirina 100mg. Tomar por la mañana. Sirve para proteger las arterias.
- **Lípidos:** Atorvastatina 80mg. Tomar por la noche. Sirve para bajar el colesterol y proteger el corazón.

English:

Original: *"Apixaban 5mg 1-0-1; Aspirin 100mg 1-0-0; Atorvastatin 80mg 0-0-1"*

With treatment module:

- **Anticoagulation:** Apixaban 5mg morning and evening—prevents clots.
- **Antiplatelet:** Aspirin 100mg morning—protects arteries.
- **Lipids:** Atorvastatin 80mg evening—lowers cholesterol and protects the heart.

Figure 4: Example of medication simplification. The treatment module categorizes medications and adds therapeutic purpose explanations.

verified knowledge bases and clinical knowledge graphs. Additionally, implementing a human-in-the-loop validation step will ensure patient safety during the transition toward fully automated systems.

6.3. Evaluation Considerations

Our pragmatic evaluation combines automatic metrics with expert review, addressing the limitations of readability indices which often ignore semantic fidelity. Despite the small sample size, expert feedback correctly prioritizes clinical accuracy and patient safety.

We acknowledge the potential for bias when using LLM-generated annotations as reference data. While manual review mitigates major errors, this "circular" evaluation highlights the need for expert-annotated benchmarks specifically for the Spanish clinical domain.

6.4. Future Work

Future research will focus on extending the system to other specialties, developing automated metrics tailored for medical simplification, and conducting direct patient comprehension studies through interactive interfaces.

7. Conclusions

We presented SimpliMED, a modular system for simplifying cardiology discharge reports using Large Language Models and specialized NLP components. Our approach addresses access barriers that prevent patients from understanding their medical documentation through an architecture that integrates three complementary components.

The section-based processing approach preserves clinical completeness by applying tailored prompts to each report section, avoiding the information loss observed with global processing approaches. The abbreviation handling component, implemented through fine-tuned detection achieving F1 of 0.90 combined with contextual disambiguation, significantly improves upon LLM-only approaches that frequently misexpand medical abbreviations with potential clinical consequences. The medication explanation module transforms cryptic prescription lists into categorized, purposeful treatment summaries through integration of the AEMPS therapeutic dictionary containing 14,611 medications.

Expert evaluation by two cardiologists yields 75% overall quality, with strong performance on admission reason sections at 91% and current illness at 75%, while identifying challenges with technical sections like diagnostic tests at 44% that require clinical interpretation beyond current model capabilities.

While demonstrating significant potential, SimpliMED requires further development before clinical deployment. Critical needs include enhanced disambiguation mechanisms to prevent dangerous misinterpretations, clinical validation pipelines to catch conceptual errors, and robust handling of complex medical concepts. The expert feedback particularly emphasizes restructuring outputs toward direct, patient-centered communication rather than technical reformulation.

This work represents a meaningful step toward improving patient access to medical information and reducing health literacy barriers in Spanish healthcare contexts. As legal requirements for accessible institutional communications extend to the medical domain, systems like SimpliMED will become increasingly essential for ensuring patients' right to understand their own health information.

8. Limitations

Several limitations constrain the interpretation and generalization of our results. First, expert evaluation covered only 8 reports due to resource constraints; while informative, this sample size limits generalization of quality assessments and necessitates larger-scale clinical validation before deployment. Second, no established automatic metrics exist for medical text simplification quality, traditional readability indices do not account for semantic fidelity or clinical accuracy, complicating iterative development and benchmark comparisons.

The system was developed and evaluated exclusively on cardiology reports. Generalization to other medical specialties requires additional adaptation and validation, as terminology, abbreviations, and section structures vary significantly across clinical domains. Furthermore, while expert evaluation assesses clinical accuracy, we did not conduct comprehension testing with actual patients, leaving the ultimate measure of simplification success, improved patient understanding, to be validated in future work.

9. Ethical Considerations

This research was conducted in accordance with the ethical principles of the Declaration of Helsinki and was approved by the corresponding institutional Ethics Committee (code: SICEIA-2024-001925). Informed consent was waived by this committee due to the retrospective design and the use of anonymized data from hospital discharge reports.

The study complied with applicable data protection regulations, including the European General Data Protection Regulation and Spanish healthcare data protection laws. The corpus of 307 cardiology discharge reports was fully anonymized prior to use, with all patient identifiers, dates, and potentially identifying information removed by the providing medical institution.

The system is designed as a clinical decision support tool and is not intended to replace professional medical judgment. Simplified reports are meant to complement, not substitute, direct patient-physician communication. We acknowledge the risk of oversimplification potentially omitting clinically relevant nuances, and recommend that any deployment include physician review of simplified outputs before patient delivery.

10. Acknowledgements

This work is funded by the Ministerio para la Transformación Digital y de la Función Pública and Plan de Recuperación, Transformación y Resiliencia - Funded by EU – NextGenerationEU within the framework of the project Desarrollo Modelos ALIA. This work has also been partially supported by Project ROMANET (CERV-2024-CHAR-LITI-101215052), funded by the European Union under the Citizens, Equality, Rights and Values programme, Project CONSENSO (PID2021-122263OB-C21) Project HEART-NLP-UJA (PID2024-156263OB-C21) and project VERITAS-H (AIA2025-163322-C64) funded by MICIU/AEI/10.13039/501100011033 and by ERDF/EU, Project GALENO-IA (DGP_PIDI_2024_00852:) funded by Junta de Andalucía

11. Bibliographical References

Mark Adler. 2012. The plain language movement. In Lawrence M. Solan and Peter M. Tiersma, editors, *The Oxford Handbook of Language and Law*. Oxford University Press.

AENOR. 2018. *Lectura Fácil. Pautas y recomendaciones para la elaboración de documentos*. AENOR INTERNACIONAL SAU, Madrid.

Lydie Bednarczyk, Daniel Reichenpfader, Christophe Gaudet-Blavignac, Amon Kenna Ette, Jamil Zaghir, Yuanyuan Zheng, Adel Bensahla, Mina Bjelogrlic, and Christian Lovis. 2025. Scientific evidence for clinical text summarization using large language models: Scoping review. *J Med Internet Res*, 27:e68998.

Robert-Mihai Botarleanu, Mihai Dascalu, Scott Andrew Crossley, and Danielle S. McNamara. 2020. Sequence-to-sequence models for automated text simplification. In *Artificial Intelligence in Education: 21st International Conference, AIED 2020, Ifrane, Morocco, July 6–10, 2020, Proceedings, Part II*, pages 31–36, Berlin, Heidelberg. Springer-Verlag.

Stefan Bott and Horacio Saggion. 2012. Automatic simplification of spanish text for e-accessibility. In *Computers Helping People with Special Needs*, pages 527–534, Berlin, Heidelberg. Springer Berlin Heidelberg.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901.

Leonardo Campillos-Llanos, Ana Rosa Terroba Reinares, Sofía Zunder Puig, Ana Valverde-Mateos, and Ana Capllonch-Carrión. 2022. Building a comparable corpus and a benchmark for Spanish medical text simplification. *Procesamiento del Lenguaje Natural*, 69:189–196.

John Carroll, Guido Minnen, Yvonne Canning, Siobhan Devlin, and John Tait. 1998. Practical simplification of English newspaper text to assist aphasic readers. *Proc. of AAAI-98 Workshop on Integrating Artificial Intelligence and Assistive Technology*.

Joshua C. Denny, Anderson Spickard, Kevin B. Johnson, Neeraja B. Peterson, Josh F. Peterson, and Randolph A. Miller. 2009. Evaluation of a method to identify and categorize section headers in clinical documents. *Journal of the American Medical Informatics Association*, 16(6):806–815.

Sebastian Joseph, Kathryn Kazanas, Keziah Reina, Vishnesh Ramanathan, Wei Xu, Byron Wallace, and Junyi Jessy Li. 2023. Multilingual simplification of medical texts. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 16662–16692, Singapore. Association for Computational Linguistics.

Sasikiran Kandula, Dorothy Curtis, and Qing Zeng-Treitler. 2010. A semantic and syntactic text simplification tool for health content. *AMIA ... Annual Symposium proceedings / AMIA Symposium. AMIA Symposium*, 2010:366–70.

J. Richard Landis and Gary G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174.

OECD. 2023. Survey of adult skills (PIAAC): 2023 country note – Spain.

Brian Ondov, Kush Attal, and Dina Demner-Fushman. 2022. A survey of automated methods for biomedical text simplification. *Journal of the American Medical Informatics Association*, 29(11):1976–1988.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing 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, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Paul Rust, Julian Frings, Sven Meister, and Leonard Fehring. 2025. Evaluation of a large language model to simplify discharge summaries and provide cardiological lifestyle recommendations. *Communications Medicine*, 5.

Laura Vásquez-Rodríguez, Matthew Shardlow, Piotr Przybyła, and Sophia Ananiadou. 2021. Investigating text simplification evaluation. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 876–882, Online. Association for Computational Linguistics.

Laura Vásquez-Rodríguez, Matthew Shardlow, Piotr Przybyła, and Sophia Ananiadou. 2023. Document-level text simplification with coherence evaluation. In *Proceedings of the Second Workshop on Text Simplification, Accessibility and Readability*, pages 85–101, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Sander Wubben, Antal van den Bosch, and Emiel Krahmer. 2012. Sentence simplification by monolingual machine translation. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1015–1024, Jeju Island, Korea. Association for Computational Linguistics.

William Xia, Ishita Unde, Brian David Ondov, and Dina Demner-Fushman. 2025. JEBS: A fine-grained biomedical lexical simplification task. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 17654–17666, Vienna, Austria. Association for Computational Linguistics.

Wei Xu, Courtney Napoles, Ellie Pavlick, Quanze Chen, and Chris Callison-Burch. 2016. Optimizing statistical machine translation for text simplification. *Transactions of the Association for Computational Linguistics*, 4:401–415.

12. Language Resource References

Thomas Scialom, Paul-Alexis Dray, Sylvain Lamprier, Benjamin Piwowarski, and Jacopo Staiano. 2020. Mlsum: The multilingual summarization corpus.