

GALENO: An Intelligent LLM-Based Support System for the Generation, Analysis and Simplification of Cardiology Discharge Reports

Lucas Molino Piñar^{1,*†}, Manuel Carlos Díaz Galiano^{1,†} and María Teresa Martín Valdivia^{1,†}

¹SINAI Research Group, Computer Science Department, Universidad de Jaén, Spain

Abstract

We present GALENO, a decision-support system for cardiology physicians at the time of hospital discharge. The platform integrates three functionalities using open-source language models on local infrastructure: (1) automatic discharge report generation from clinical notes using Gemma 3 12B; (2) admission analysis via Qwen 3 8B with a three-phase extraction pipeline for clinical indicators; and (3) patient-oriented report simplification managing abbreviations and medications. This article describes the system demonstration.

Keywords

Clinical Natural Language Processing, Large Language Models, Clinical Report Generation, Medical Text Simplification, Clinical Decision Support Systems

1. Introduction

Writing hospital discharge reports constitutes one of the most time-consuming clinical tasks in health-care professionals' daily practice. In highly complex specialties such as cardiology, these reports must accurately reflect the patient's evolution during admission, prior medical history, diagnostic tests performed, interventions carried out, and the therapeutic plan to follow after discharge. The exhaustiveness required in these documents generates a considerable administrative burden that detracts from direct patient care [1].

Simultaneously, these reports present a significant communication barrier. Discharge documents are laden with specialized technical terminology, clinical abbreviations, and diagnostic jargon that remain incomprehensible to most patients and their families [2]. This lack of understanding can compromise treatment adherence, hinder subsequent medical follow-up, and ultimately increase the risk of hospital readmissions. Moreover, recent studies confirm that a significant proportion of the population possesses limited reading competence [3], which exacerbates this issue.

In this context, we present **GALENO**, an intelligent support system designed as a decision-support tool for physicians at the time of hospital discharge. Unlike previous works that address simplification [1] or clinical summary generation in isolation, GALENO integrates three complementary functionalities in a single platform:

1. **Automatic discharge report generation:** From the patient's evolution sheet and anamnesis sheet, the system generates a discharge report structured into ten standardized sections using the Gemma 3 12B model [4].
2. **Admission analysis:** A module based on Qwen 3 8B [5] extracts, normalizes, and categorizes clinical indicators and vital signs across the various records of the evolution sheet, generating interactive graphical visualizations that facilitate physician review.

SEPLN 2026: 42nd International Conference of the Spanish Society for Natural Language Processing, León, Spain, 22-25 September 2026.

*Corresponding author.

† These authors contributed equally.

✉ lmolino@ujaen.es (L. M. Piñar); mcdiaz@ujaen.es (M. C. D. Galiano); maite@ujaen.es (M. T. M. Valdivia)

ORCID 0009-0009-5990-9802 (L. M. Piñar); 0000-0001-9298-1376 (M. C. D. Galiano); 0000-0002-2874-0401 (M. T. M. Valdivia)



© 2026 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

3. **Report simplification for the patient:** The system transforms the technical discharge report into a version accessible in plain language, managing medical abbreviations and enriching pharmacological information with therapeutic explanations.

The platform runs entirely on local infrastructure (*on-premise*), ensuring compliance with health data protection regulations by not requiring the transmission of clinical information to external cloud services. This article describes the demonstration of the GALENO system and is organized as follows: Section 2 reviews related work; Section 3 details the system architecture and methodology; Section 4 describes the demonstration and its interactive graphical environment; Section 5 discusses benefits and limitations; and Section 6 presents conclusions and future work.

2. Related Work

2.1. Automated Clinical Documentation

The automation of clinical documentation has gained significant momentum with the emergence of Large Language Models (LLMs) [6], whose foundations were laid by the Transformer architecture [7]. Although LLMs offer new opportunities for producing structured reports from unstructured clinical notes, their application in real hospital environments presents inherent challenges regarding data privacy [1]. Systems based on closed-source models require cloud connections, which are incompatible with the requirements of hospital ethics committees. Hence the relevance of using open-source models with viable sizes (8–14B parameters) that can be executed on local infrastructure.

2.2. Medical Text Simplification

The field of automatic text simplification has evolved from rule-based systems [8] to current LLM-based approaches. [1] conclude in their systematic review that LLMs significantly outperform previous architectures in clinical simplification, particularly when fine-tuned on domain-specific corpora.

In the Spanish-speaking context, resources remain limited compared to English. The CLARA-Med project [9] provided a first parallel corpus of clinical sentences in Spanish. Of particular relevance to our work, modular architectures for the simplification of cardiology discharge reports have been proposed which integrate specialized modules for abbreviation handling (F1: 0.90) and medication recognition (F1: 0.91), as well as a therapeutic dictionary with 14,611 medications from the Spanish Agency of Medicines (AEMPS). GALENO integrates this methodology, complementing it with automatic generation and admission analysis.

2.3. Clinical Information Extraction

The extraction of structured information from free clinical text is an active area of biomedical NLP [10, 11]. Recently, Chain-of-Thought reasoning capabilities [12] have proven particularly effective for complex extractive tasks requiring contextual inference, even in relatively small models.

3. Architecture and Methodology

GALENO is structured as a modular system composed of three interconnected subsystems that share a common preprocessing of documentary sources. Figure 1 presents a high-level view of the processing flow. The following subsections detail the system components.

3.1. Data Input and Preprocessing

The system operates on two main documentary sources:

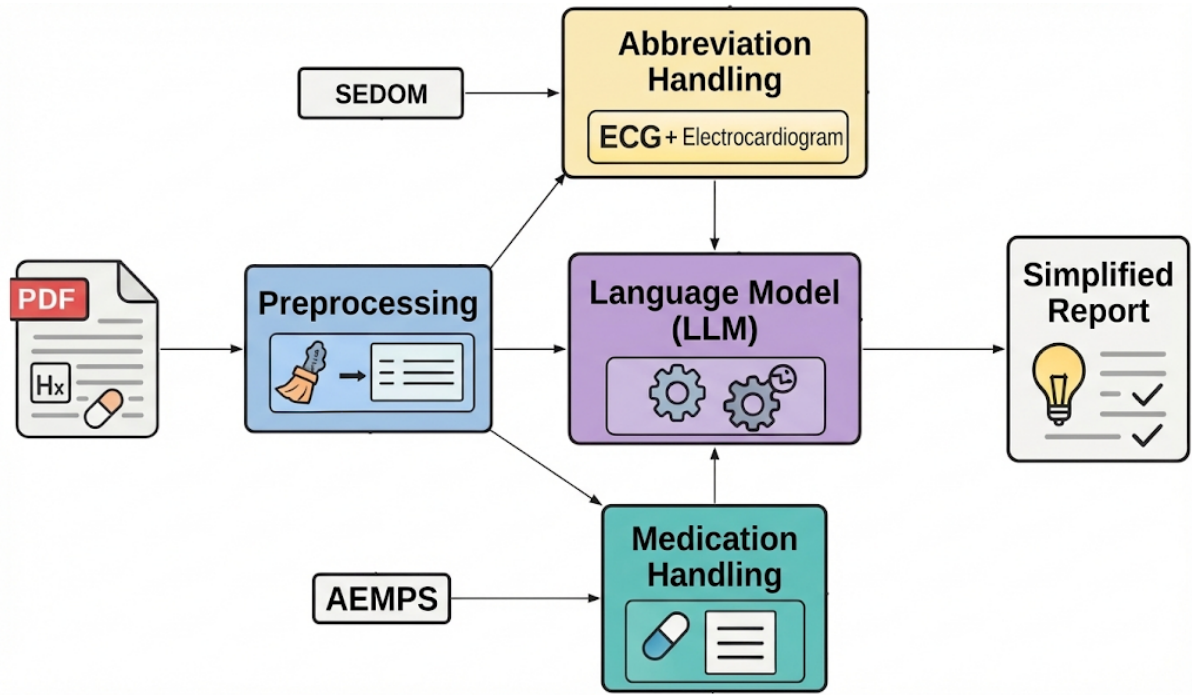


Figure 1: General architecture of the GALENO system. The flow starts from the original clinical report (PDF), passes through preprocessing that segments and structures the content, and branches into the three main modules: report generation, admission analysis, and patient simplification. The abbreviation handling (SEDOM) and medication (AEMPS) modules enrich the simplification process.

1. **Anamnesis Sheet:** Contains the patient’s prior medical history: personal background, chronic pathologies, allergies, regular medication, and relevant family history.
2. **Evolution Sheet:** A longitudinal document composed of multiple timestamped records generated during the hospital stay, accumulating clinical assessments, vital signs, laboratory results, and incidents.

The preprocessing module performs: (1) cleaning of format artifacts and invalid characters; (2) segmentation of the evolution sheet into individually time-identified records; and (3) division of the anamnesis sheet into differentiated semantic sections.

3.2. Discharge Report Generation

For the automatic generation of the discharge report, GALENO employs the **Gemma 3 12B** model [4], selected after a comparative evaluation among various open-source architectures, including LLaMA 3 (8B and 70B) [13], Mistral [14], and Mixtral. Gemma 3 12B exhibited capabilities analogous to 70B models in adherence to generation guidelines, synthesis of clinical chronologies, and grammatical quality, while maintaining inference times suitable for interactive execution on mid-to-high-range GPUs.

The process is articulated through specific prompts for each of the ten target sections of the discharge report. The model is invoked independently per section, with the preprocessed evolution and anamnesis sheets injected as context. The generated sections are: (1) Reason for Admission, (2) Personal Background, (3) Current Illness, (4) Clinical Course and Evolution, (5) Diagnostic Tests, (6) Surgical Procedures, (7) Discharge Diagnosis, (8) Treatment, (9) Action Plan, and (10) Follow-up.

This section-by-section approach, with a total of ten model calls, enables selective regeneration of individual sections. The model manages sections that lack information in the original documentation, avoiding fabrication of non-existent content.

Table 1

Illustrative example of the normalization and categorization executed in three phases by the analysis module based on Qwen 3 8B: raw value extraction (Phase 1), terminological normalization (Phase 2), and hierarchical categorization (Phase 3).

Raw Value	Normalized	Category
Blood Pressure: 120	BP	Vital Signs
Art. Press. [120/80]	BP	Vital Signs
Pressure → 125	BP	Vital Signs
Fasting Glucose: 99	Glucose	Lab Tests
Blood Sugar: 105	Glucose	Lab Tests
HbA1c = 5.6%	HbA1c	Lab Tests
HR 108 bpm	Heart Rate	Vital Signs
Rate: 70	Heart Rate	Vital Signs
Pulse 85	Heart Rate	Vital Signs

3.3. Admission Analysis

The **Admission Analysis** module provides physicians with a visual dashboard of the evolution of clinical indicators during the hospital stay. This functionality is particularly relevant in cardiovascular pathologies, where the volatility of parameters such as blood pressure, heart rate, or oxygen saturation provides critical information for decision-making.

The information is scattered in free text within the evolution sheet records, without a tabular format. GALENO employs the **Qwen 3 8B** model [5], selected for its Chain-of-Thought reasoning capability, which is particularly effective for extractive tasks with contextual inference. The process is articulated in three sequential phases, each as an independent model call:

1. **Phase 1 — Record-by-Record Extraction:** For each record in the evolution sheet, a call is made to the Qwen 3 8B model requesting the atomic extraction of the clinical numerical parameters present, isolating the value from its denotative meaning. Each record is processed independently in a separate call. For example, given a record such as “BP 124/67 mmHg. HR 108 bpm. SpO2 98%. Urine output 800 + overnight”, the model extracts the indicator-value pairs: systolic blood pressure (124), diastolic blood pressure (67), heart rate (108), oxygen saturation (98), and urine output (800).
2. **Phase 2 — Terminological Normalization:** A second model call receives the complete set of values extracted from all records and performs terminological harmonization. The goal is to identify formally distinct but semantically equivalent parameters and link them under a single normalized denomination. For example, the terms “Blood Pressure”, “BP”, “Arterial Pressure”, and “Art. Press.” are grouped under the unified identifier “Blood Pressure”. Similarly, “Fasting Glucose” and “Blood Sugar Level” are consolidated as “Glucose”.
3. **Phase 3 — Hierarchical Categorization:** A third and final model call organizes the normalized indicators into high-level clinically significant categories. Thus, heart rate, blood pressure, and oxygen saturation are grouped under “Vital Signs”; glucose, glycated hemoglobin, and creatinine under “Laboratory Tests”; and metrics such as urine output under “Fluid Balances”.

The result of this three-phase process is a typed and hierarchically organized JSON structure, ready for consumption by the frontend visualization libraries. Table 1 illustrates with a representative example how this dual grouping operates.

3.4. Report Simplification for the Patient

The third functional pillar of GALENO is the simplification of the discharge report into clear and accessible language for the patient. This module operates on the generated report or on a pre-existing report loaded by the physician.

The simplification is articulated as a sequential process integrating three specialized components:

1. **Abbreviation detection and expansion:** Through a fine-tuned Small Language Model (SLM) (F1: 0.90), the system identifies clinical abbreviations present in the text (e.g., AF, HT, LVEF). Unlike regex-based approaches (F1: 0.67), this model considers semantic context to distinguish, for instance, “SD” as “Standard Deviation” from other uses. Detected abbreviations are expanded using the SEDOM dictionary (Spanish Society of Medical Documentation), which contains 7,054 medical abbreviation entries. For abbreviations with multiple possible meanings, disambiguation is delegated to the simplification LLM, which selects the contextually appropriate interpretation.
2. **Medication recognition and enrichment:** A second fine-tuned SLM (F1: 0.91 for brand names; F1: 0.70 for active principles) identifies medications mentioned in the text. Each detected medication is matched against a therapeutic dictionary built from the AEMPS database, containing 14,611 medications with simplified therapeutic explanations. The search employs Levenshtein distance for fuzzy matching to handle the spelling variations and typos common in clinical text.
3. **Section-by-section simplification:** The report, segmented into its sections, is simplified independently via an LLM with specialized prompts for each section type. 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. Original medical terms are preserved in parentheses to facilitate patient learning.

This modular approach ensures that clinical accuracy and therapeutic plan mandates are not accidentally distorted during simplification. The hospital retains the unaltered technical version while the patient receives a pedagogical, adapted, and validated version.

4. System Demonstration

GALENO is presented as an interactive web application designed following user-centered design principles for medical professionals. The interface is structured around a top navigation bar providing access to the system’s functionalities: the home page (Figure 2), the generation module, the simplification module, and an informational section about the system operation.

The following subsections describe the interaction flow for each functionality.

4.1. Report Generation Module

The interface presents a panel divided into two tabs (Figure 3): *Evolution Sheet* and *Anamnesis Sheet*. The physician loads both documents from the hospital information system or via direct copy.

The right panel presents the generated report with all ten clinical sections, indicating the generation date and time. The physician has three actions available: **Edit** (correct sections before validation), **Simplify** (launch simplification for the patient), and **Download PDF** (generate the document for signing).

4.2. Admission Analysis Module

After executing the three extraction, normalization, and categorization phases (Section 3.3) on the evolution sheet records, the system generates a graphical visualization dashboard (Figure 4).

Each normalized indicator is visualized in its own time series chart, with color coding to facilitate quick visual reading. This allows the physician to assess trends in critical parameters such as blood pressure or oxygen saturation, contributing to the discharge decision. Information is extracted directly from narrative text, without requiring prior tabular formats.



Figure 2: GALENO-IA main screen, providing access to the two main functionalities: generation and simplification of cardiology reports.



Figure 3: Generation module. Left: input panel with evolution and anamnesis sheets. Right: automatically generated report in standardized sections, with options for editing, simplification, and PDF download.

4.3. Simplification Module

The simplification module operates on both system-generated reports and pre-existing reports loaded by the user (Figure 5).

In the simplified version, each medication is presented with: brand name and active principle in parentheses, therapeutic category in accessible language, explained dosage (e.g., “Take one tablet in the morning and one at night” instead of “1-0-1”), and a clear description of its function (e.g., “Helps prevent blood clots”). The patient can download this version as PDF.

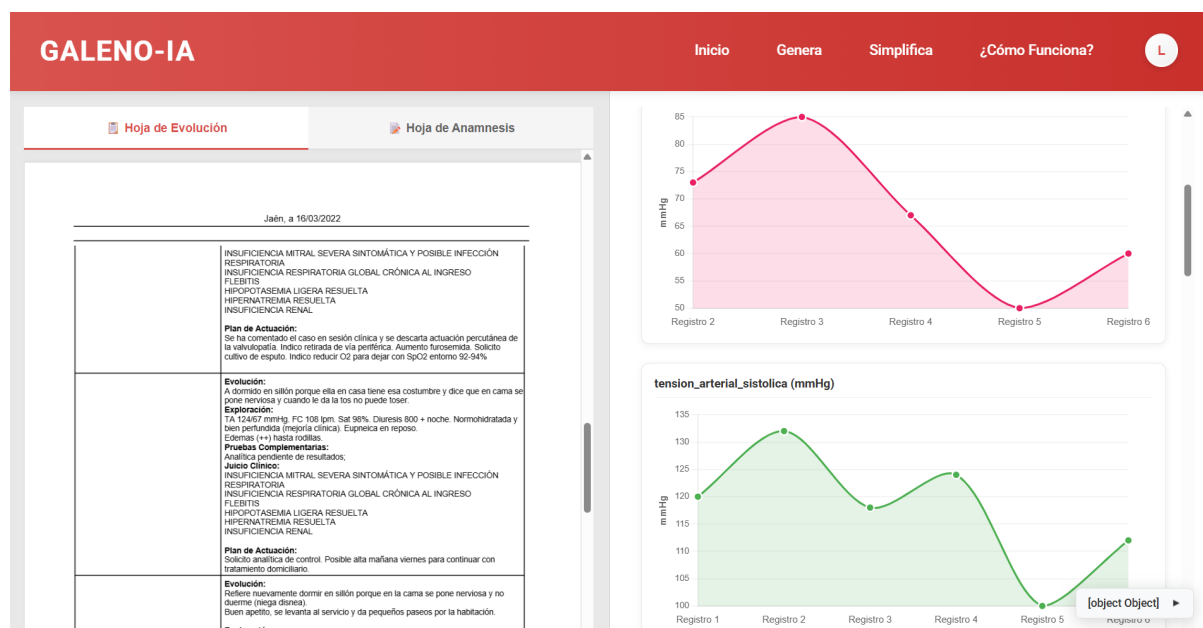


Figure 4: Admission analysis module. Left: evolution sheet with clinical records. Right: time series charts of automatically extracted indicators (diastolic and systolic blood pressure) across admission records.

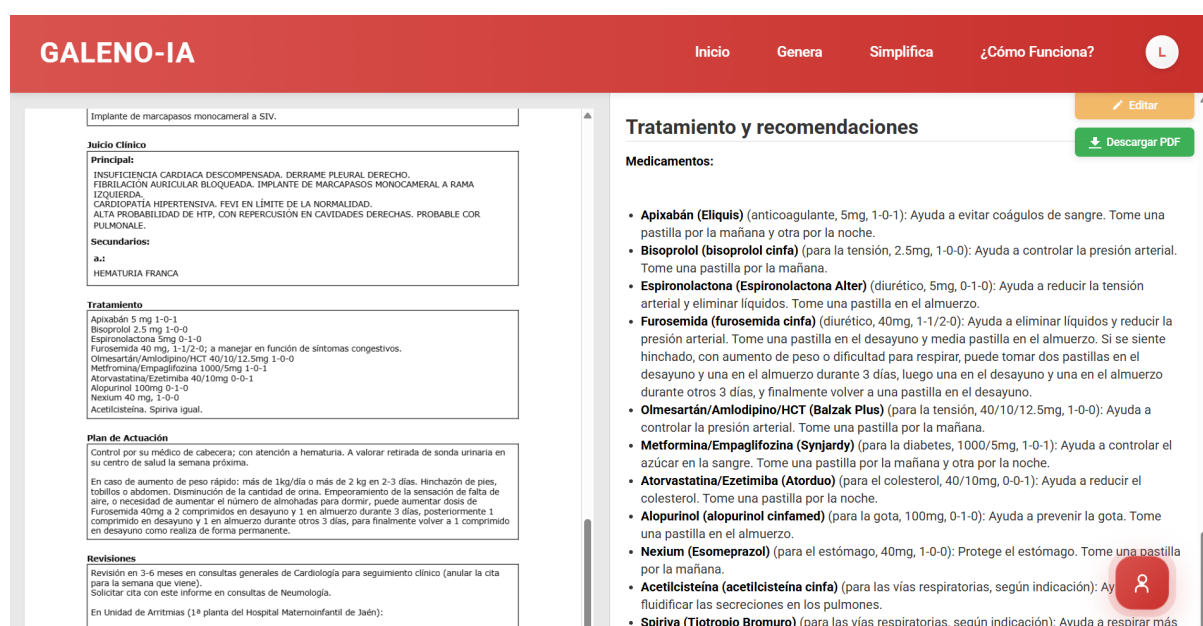


Figure 5: Simplification module. Left: original report (discharge diagnosis and treatment). Right: simplified version with brand name, active principle, explained dosage, and therapeutic indication for each medication.

5. Benefits and Limitations

GALENO's implementation provides benefits at three levels: **(1) For the physician:** significant reduction in documentation time, shifting from complete drafting to a review and fine-tuning task. The analysis dashboard provides visual support for the discharge decision. **(2) For the patient:** immediate access to a comprehensible document with medication clarification from certified sources (AEMPS), contributing to improved therapeutic adherence. **(3) At the institutional level:** standardization of discharge report format with the complete ten-section structure.

The system presents relevant limitations: as a demonstrative prototype, no formal clinical validation

has been conducted with a large panel of specialists. Generation quality depends on the exhaustiveness of the original evolution notes. Accumulated inference times may be non-negligible in a workflow demanding immediacy. Finally, the system is specialized in cardiology, and its generalization to other specialties would require adaptation of prompts and linguistic resources.

6. Conclusions and Future Work

In this article we have presented GALENO, a demonstrative intelligent support system that integrates three complementary functionalities to assist physicians at the time of hospital discharge in cardiology: automatic discharge report generation from unstructured clinical documentation using Gemma 3 12B, visual analysis of admission evolution through a three-phase extractive process based on Qwen 3 8B, and report simplification into plain language for the patient. The platform runs entirely on local infrastructure, ensuring clinical data privacy.

As future work, we plan to subject the system to formal clinical evaluation with an expanded panel of cardiology specialists, employing validated quality metrics for both generation and simplification. We also intend to refine the section-by-section generation prompts through automatic instruction optimization techniques. Additionally, we plan to integrate official coded mappings (ICD-10) in the analysis module to normalize extracted indicators at a higher semantic level. Finally, we aim to extrapolate GALENO's architecture from the cardiovascular area to other clinical specialties such as pulmonology, traumatology, or geriatrics, where clinical documentation challenges and patient communication are equally relevant.

Declaration on Generative AI

During the preparation of this work, the authors used generative AI for grammar and spelling checking. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

Acknowledgments

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.

References

- [1] M. Bednarczyk, et al., A systematic review of large language models for medical text simplification, *Journal of Medical Internet Research* 26 (2024). doi:10.2196/54850.
- [2] S. Kandula, D. Curtis, Q. Zeng-Treitler, Bootstrapping a statistical machine translation system for medical text simplification, *AMIA Annual Symposium Proceedings* 2010 (2010) 346–350.
- [3] OECD, *Education at a Glance 2023: OECD Indicators*, OECD Publishing, Paris, 2023.
- [4] Gemma Team, T. Mesnard, C. Hardin, R. Dadashi, S. Bhupatiraju, S. Pathak, L. Sifre, M. Rivière, M. S. Kale, J. Love, et al., *Gemma: Open Models Based on Gemini Research and Technology*, Technical Report, Google DeepMind, 2024. ArXiv preprint arXiv:2403.08295.

- [5] Qwen Team, J. Bai, S. Bai, Y. Chu, Z. Cui, K. Dang, X. Deng, Y. Fan, W. Ge, Y. Han, et al., Qwen technical report, arXiv preprint arXiv:2309.16609 (2024).
- [6] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al., Language models are few-shot learners, in: *Advances in Neural Information Processing Systems*, volume 33, 2020, pp. 1877–1901.
- [7] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, in: *Advances in Neural Information Processing Systems*, volume 30, 2017, pp. 5998–6008.
- [8] J. Carroll, G. Minnen, Y. Canning, S. Devlin, J. Tait, Practical simplification of English newspaper text to assist aphasic readers, in: *Proceedings of the AAAI Workshop on Integrating AI and Assistive Technology*, 1998, pp. 7–10.
- [9] L. Rodríguez, et al., CLARA-MeD: A corpus for simplification of Spanish medical literature, *Procesamiento del Lenguaje Natural* 73 (2024) 47–58.
- [10] Y. Wang, L. Wang, M. Rastegar-Mojarad, S. Moon, F. Shen, N. Afzal, S. Liu, Y. Zeng, S. Mehrabi, S. Sohn, H. Liu, Clinical information extraction applications: A literature review, *Journal of Biomedical Informatics* 77 (2018) 34–49.
- [11] H. Moen, L.-M. Peltonen, et al., Information extraction from clinical narratives: An empirical study, *Studies in Health Technology and Informatics* 192 (2013) 385–389.
- [12] J. Wei, X. Wang, D. Schuurmans, M. Bosma, B. Ichter, F. Xia, E. Chi, Q. V. Le, D. Zhou, Chain-of-thought prompting elicits reasoning in large language models, in: *Advances in Neural Information Processing Systems*, volume 35, 2022, pp. 24824–24837.
- [13] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, et al., LLaMA: Open and efficient foundation language models, arXiv preprint arXiv:2302.13971 (2023).
- [14] A. Q. Jiang, A. Sablayrolles, A. Mensch, C. Bamford, D. S. Chaplot, D. d. l. Casas, F. Bressand, G. Lengyel, G. Lample, L. Saulnier, et al., Mistral 7B, arXiv preprint arXiv:2310.06825 (2023).