Leveraging Open Source LLMs for Clinical Coding in Spanish: They Are Here to Stay

Rosa M. Montañés-Salas[1,*,†], Natalia Sisamón Dolader[1,†], Paula Peña-Larena[1] and Rafael del-Hoyo-Alonso[1]

[1]*Aragon Institute of Technology (ITA), María de Luna, 7–8, 50018 Zaragoza, Spain*

Abstract

This work explores the application of large language models (LLMs) for automatic clinical coding in Spanish, aligned with the 2024 edition of the ICD-10-ES standard. Two main strategies are evaluated: a fully LLM-based approach and a hybrid method integrating existing transformer-based models. Experiments conducted on a translated subset of MIMIC-IV suggest improvements in classification performance, particularly for longer texts and higher levels of the coding hierarchy. Limitations remain at deeper hierarchical levels, but the approach shows potential for broader coding coverage and adaptability.

Keywords

Spanish clinical coding, Natural Language Processing, Large Language Models

1. Introduction

The growing digitalization of hospital information systems has significantly enhanced the efficiency of clinical practice, with Electronic Health Records (EHRs) playing a central role in this transformation. These systems enable streamlined management and monitoring of patient data, contributing to more effective decision-making, improved care coordination, and better allocation of resources. However, the effectiveness of EHRs in supporting healthcare delivery is dependent upon the ability to systematically extract relevant information and the accurate coding of medical records. This coding is essential for clinical decision making and administrative tasks such as billing, statistical analysis, and health research. Despite the critical role of medical coding, implementing standardized systems presents significant challenges. In Spain, the ICD-10-ES (or CIE-10) classification system serves as the official standard for coding diseases and conditions, representing a localized adaptation of the global ICD-10 system. While these standards evolve to meet new medical knowledge and societal needs, they remain complex and subject to frequent revisions. The latest version of the Spanish ICD-10 system corresponds to the 2024 edition, aligning with both international standards and national healthcare requirements. While its complexity and frequent updates reflect the depth and adaptability of the classification, they also present challenges for manual application, requiring significant effort and attention to ensure accurate coding.

The inherent nature of medical data adds an additional layer of complexity to this task. Medical texts are often unstructured, containing a significant amount of free-text content that can vary in terminology and structure. In addition, the number of potential codes is extensive, making manual assignment cumbersome and error prone. Generating and maintaining accurate reference corpora to assist in this process is not only challenging but also requires substantial resources. These challenges are further exacerbated in non-English languages, especially in Spanish, where medical language has historically

SEPLN 2025: 41st International Conference of the Spanish Society for Natural Language Processing, Zaragoza, Spain, 23-26 September 2025.

*Corresponding author.

†These authors contributed equally.

rmontanes@ita.es (R. M. Montañés-Salas); nsisamon@ita.es (N. S. Dolader); ppena@ita.es (P. Peña-Larena); rdelhoyo@ita.es (R. del-Hoyo-Alonso)

0000-0003-4636-5868 (R. M. Montañés-Salas); 0000-0001-5750-6238 (P. Peña-Larena); 0000-0003-2755-5500 (R. del-Hoyo-Alonso)

lacked the same degree of standardization and digital resources as seen in English-speaking regions.

Given these difficulties, there is an increasing need to apply Natural Language Processing (NLP) techniques to process and analyze EHRs [1]. Various models have been developed to address these challenges, including discriminative models that focus on classification tasks based on domain-specific datasets, and generative models, such as large language models (LLMs), trained on vast quantities of data for general-domain language understanding and generation. These LLMs, due to their ability to handle large-scale unstructured data and generate contextually relevant insights, offer considerable promise for automating and improving medical record analysis. Nonetheless, their deployment in healthcare settings must be approached with caution, particularly regarding privacy, security, and ethical concerns. Ensuring that these technologies adhere to stringent ethical guidelines and privacy regulations is crucial to maintaining patient trust and safety while leveraging their potential to enhance clinical workflows.

This paper explores the application of large language models for Spanish clinical coding, complying with the most recent version of ICD-10-ES codification. Several methodologies for automatic text classification are explored, covering LLM-based architectures and a hybrid transformer-based approach. The rest of the article is organized as follows: Section 2 reviews previous work on automatic classification of medical texts with ICD-10. Section 3 details the materials and methods developed in this research, while in Section 4, the experiments and the results obtained are presented. Finally, Section 5 discusses the conclusions and possible lines of future research.

2. Related Work

2.1. ICD-10-ES / CIE-10 codification

The International Classification of Diseases, 10th revision (ICD-10), developed by the World Health Organization[1], is an international standard for reporting diseases and health conditions. The Spanish adaptation, known as ICD-10-ES or CIE-10, serves as the official normative coding system within Spain's healthcare system[2].

The CIE-10 classification system is hierarchically structured and highly detailed, facilitating accurate epidemiological monitoring, statistical analysis, and administrative management. It comprises over 100,000 distinct alphanumeric codes organized into 22 chapters, each grouping diseases and conditions by organ systems or shared clinical characteristics. Chapters are denoted by letters (A–Z) and are subdivided into sections, categories, and subcategories, following the structure illustrated in Figure 1: (1) Chapter: groups code ranges into sections, generally associated with a letter; sections identify blocks of categories (e.g., A00–A09). (2) Category: a letter followed by two digits, defining a main class (e.g., A06 Amebiasis). (3) Subcategories: a letter followed by three or more digits, providing greater specificity (e.g., A06.81 Amebic cystitis). Its most recent update corresponds to the 5th edition, released in January 2024.

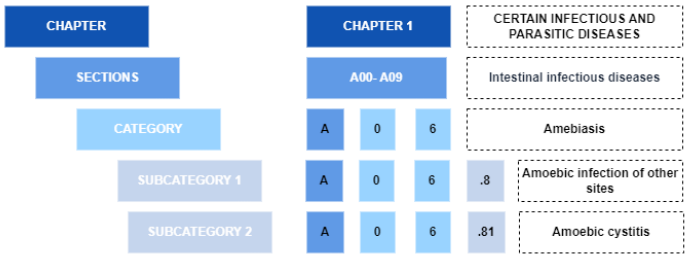

Figure 1: CIE-10 hierarchy.

[1] https://www.who.int/standards/classifications/classification-of-diseases
[2] https://www.eciemaps.sanidad.gob.es/browser/metabuscador

A primary resource for clinical text classification in Spanish is the CodiEsp [2] corpus, developed in the context of the Clinical Case Coding in Spanish Shared Task within CLEF in 2020. This dataset includes 1,000 clinical cases manually annotated with ICD-10-ES codes, divided into train, development and test splits. It covers both diagnosis and procedures, thus facilitating various shared tasks and NLP research initiatives specifically oriented towards Spanish medical text analysis.

Apart from the results obtained by the participants of the CodiEsp workshop itself, a key study in the automatic classification of medical texts in Spanish is that of López-García et al.[3], who systematically analyzed several transformer-based architectures, namely mBERT, BETO, and XLM-R, trained using Spanish annotated corpora such as the CodiEsp. Their study demonstrated that domain-specific pre-training significantly enhances the performance of transformers, achieving state-of-the-art outcomes for ICD-10-ES code assignment tasks. Specifically, the BETO-Galén model, an adapted version of BETO specifically for the medical domain in Spanish, obtained the best results outperforming previous systems by 10-11%

For English clinical text classification, the MIMIC (Medical Information Mart for Intensive Care) dataset has become a benchmark resource. It contains clinical data of patients admitted to intensive care and emergency units, including electronic medical records, laboratory results, clinical notes and ICD-10 coded medical diagnoses. Specifically, the recent version, MIMIC-IV v3.0 [4][5], released in July 2024, extends its predecessors by including comprehensive patient data for multilabel classification tasks, providing robust evaluation scenarios for various NLP methodologies [6].

Recent advancements in English clinical coding systems have extensively employed Pretrained Language Models (PLMs), as in the case of Spanish. PLM-ICD[7] a transformer-based model fine-tuned for ICD coding, enhances the potential of transfer learning strategies by incorporating mechanisms such as segment pooling and label-centric attention to manage long sequences and assign multiple codes in complex clinical texts. It has shown notable improvements on datasets like MIMIC-II and MIMIC-III, particularly in classifying infrequent codes. This model has also been evaluated in recent studies using MIMIC-IV, where Nguyen et al. [6] established it as a benchmark for extreme multi-label classification, and Edin et al. [8] conducted a comprehensive performance comparison across various MIMIC-IV-based approaches, including PLM-ICD.

2.2. Large Language Models

Numerous studies have explored the potential of Large Language Models (LLMs) in the biomedical domain, highlighting their ability to capture complex semantics and adapt to new tasks without additional fine-tuning. In ICD coding, LLMs have shown promise in processing complex clinical texts and handling extreme multi-label classification. Boukhers et al. [9] evaluated LLAMA-2 both as a direct classifier and as a generator of enriched representations for convolutional models, reporting improvements over traditional approaches. Similarly, Boyle et al. [10] showed that general-purpose LLMs such as GPT-3, GPT-4, and LLaMA-2 achieve competitive zero-shot and few-shot performance by leveraging the ICD hierarchy and code descriptions.

Hybrid approaches combining LLMs architectures with other transformer-based models have also shown encouraging results in tasks like named entity recognition and clinical coding, particularly through novel fine-tuning and prompt engineering techniques. Yang et al. [11] proposed a two-stage method integrating an LLM with an LSTM-based verifier to improve precision while maintaining recall. Mustafa et al. [12] combined SNOMED mapping, machine learning, and LLMs, demonstrating that such models can reach performance levels comparable to human experts in clinical document classification.

Moreover, recent advancements have highlighted the potential of Large Language Models (LLMs) not only as generators but also as autonomous verifiers. Weng et al.[13] showed that self-verification strategies enable LLMs to iteratively assess and improve their outputs in complex reasoning tasks. Similarly, Gu et al.[14] surveyed the use of LLMs as evaluators, addressing reliability, bias mitigation, and adaptability. These findings highlight the growing role of LLMs in tasks requiring both generation and validation.

3. Materials and Methods

3.1. Public datasets

To conduct the evaluation of the coding strategies described below, medical datasets in Spanish containing clinical texts labeled with the most recent version of ICD-10-ES were sought, focusing on the diagnosis subset. The resources considered are described below.

3.1.1. CodiEsp

The CodiEsp corpus consists of 1,000 clinical case studies in Spanish, selected manually by a practicing physician and annotated by coding professionals under the ICD-10-ES standard, in its 2nd version (from 2018). It has annotations for both the diagnosis and procedures code subsets, containing a total of 18435 tagged codes, 3427 being unique (2557 and 870 for diagnosis and procedures, respectively) within the 1,000 records. The final dataset had a total of 16504 sentences, with an average of 16.5 sentences per clinical case. It contains a total of 396,988 words, with an average of 396.2 words per clinical case.

The use of this dataset was deemed unsuitable during the initial stages of the study due to limitations such as restricted code coverage, imbalanced code distribution —hindering generalization to infrequent classes— and its reliance on an outdated version of ICD-10-ES. Since 2018, the classification system has undergone several updates, including the addition of 3,562 new codes until the 2024 edition. Notably, these updates include four new categories ('I5A', 'U07', 'U09', 'Z58') and 109 new four-digit subcategories.

3.1.2. MIMIC-IV v3.0

The MIMIC-IV v3.0 dataset is an English publicly available, de-identified, electronic health records database developed by the MIT Laboratory for Computational Physiology. It encompasses comprehensive clinical data from over 364,627 patients admitted to the Beth Israel Deaconess Medical Center in Boston, Massachusetts, USA, until 2022. The dataset includes information on patient demographics, vital signs, laboratory test results, medications, diagnoses, procedures, and administrative details. MIMIC-IV is structured modularly, linking core hospital data to multiple data sources, which enhances data usability and facilitates diverse research applications in clinical informatics and machine learning. Access to MIMIC-IV v3.0 is restricted to research purposes, requiring credentialed approval and adherence to data use agreements to ensure patient privacy and data security.

The MIMIC-IV v3.0 dataset consists of anonymized clinical records with ICD-coded admissions. Only 254,377 of these admissions are associated with ICD-10 codes, as some entries rely on the older ICD-9 standard. Furthermore, discharge summaries in text format are available for 122,288 of these ICD-10-coded admissions, which were used for this study. After filtering, we focused exclusively on the subset containing both ICD-10 codes and their corresponding text. This filtered dataset includes a total of 16,155 unique ICD codes. Their distribution is highly skewed, with a small set of codes appearing frequently across the records, while the majority are represented in only a limited number of cases. The median number of appearances per code is 4, while the mean is 109.26, highlighting the skewed nature of the distribution. Furthermore, on average, each summary discharge record contains 159.43 sentences and 1,859.98 words.

Upon analyzing the 122,288 records with available text summaries in MIMIC-IV, it is observed that 233 of the new codes added since the 2018 edition have been assigned in the labeled discharge summaries of this dataset, however, none of the new categories introduced in the 2024 version are represented, which limits the evaluation of model performance at the category level. Due to the large volume of relevant records, a subset was selected for the purposes of this study. It consists of 1,786 randomly chosen patient admissions, ensuring that the sample is representative of the overall dataset.

Since this study focuses on analyzing coding performance specifically in Spanish, the discharge summaries subset has been automatically translated using local open-source LLMs: texts were fully translated, and the annotated ICD-10 codes were mapped and validated against the ICD-10-ES standard.

3.2. Proposed approaches

This study explores two principal approaches to automatic clinical coding in Spanish in accordance with the current ICD-10-ES standard. The first approach involves the development of a classification system based solely on open-source large language models (LLMs). The second investigates the adaptation of existing Spanish finetuned models through the integration of LLM-based techniques to mitigate limitations arising from outdated training data.

3.2.1. Classification system based on LLMs

RAG-based classification: this method is based on applying a RAG (Retrieval-Augmented Generation) strategy, which combines text generation with information retrieval from external resources. This method seeks to mitigate the limitations of generative models in classification tasks by providing a grounded, restricted set of relevant codes prior to LLM inference. First, the retrieval mechanism filters and selects the most relevant CIE-10 codes related to the input text. Subsequently, the LLM model is prompted to perform the classification providing this subset as context, thus optimizing code assignment.

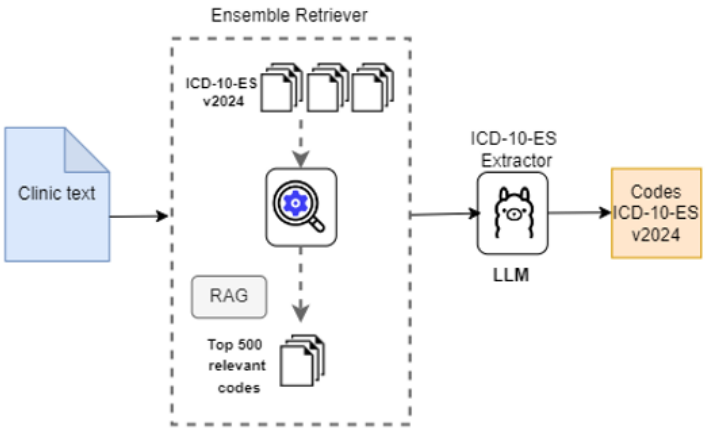

Figure 2: RAG-based classification workflow.

The workflow corresponding to this approach is depicted in figure 2. For its implementation, a vector database was built using FAISS (Facebook AI Similarity Search)[15] upon the complete CIE-10 2024 standard. Each code, along with its official Spanish descriptions, was conveniently stored and its semantic representation computed employing embedding models, which allows efficient search of similar codes to an input text. The retrieval process was carried out using an ensemble retriever, which combines two complementary techniques: a *TF-IDF Retriever*, which selects relevant codes based on lexical matches between clinical text and code descriptions, favouring specific terms present in the database; and a *Similarity Search Retriever*, which uses similarity search over the embedding space to retrieve codes semantically close to the content of the clinical report. This combination enables the capture of both exact keyword matches and contextual semantic similarities, resulting in a more accurate and representative set of codes for the clinical text under analysis. Consequently, the output of the retrieval process is a restricted set of relevant CIE-10 codes, which is then passed to the LLM module, referred to as the *ICD-10-ES Extractor*. The extractor receives the clinical text along with the retrieved codes as contextual input and, through an optimized prompt, selects those codes that are explicitly present in the document. This approach ensures that the assigned codes are strictly drawn from the updated CIE-10 taxonomy, thereby maintaining alignment with the official classification system.

Hierarchical RAG-based classification: since full CIE-10 2024 comprises over 100,000 codes, which may hinder efficient retrieval in a standard RAG scheme, a hierarchical classification approach was explored. This strategy leverages the inherent structure of the CIE-10 taxonomy to optimize code

retrieval while reducing computational complexity. By progressively narrowing the search space at each level of the hierarchy, the method contributes to improvements in both efficiency and classification accuracy. An initial RAG extracts the category-level candidates (approximately 1,000 codes). Once the most likely code groups have been identified, the following RAG modules determine the specific subcategories at level 1 (letter + three digits) and level 2 (letter + four digits).

The developed workflow is displayed in figure 3, depicting the three-phase process followed by integrating the RAG approach at each level of the hierarchy. To implement this approach, two vector

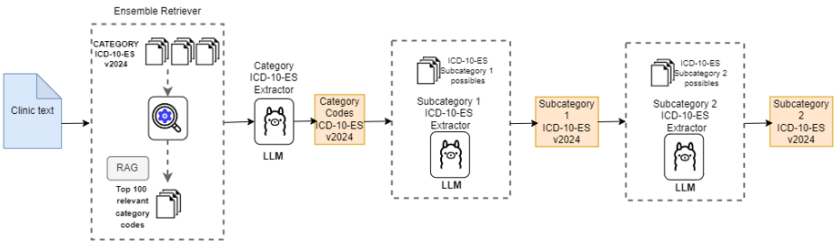

Figure 3: Hierarchical RAG-based classification workflow.

databases were built in FAISS. The first contains only the codes and descriptions up to the category level, and the second database stores lower levels of codes complying with their hierarchy.

First, an ensemble retriever (TF-IDF + Similarity Search) is applied on the vector database containing category codes. The retrieved ones are passed to the *Category ICD-10-ES Extractor*, which analyses the clinical text together with the retrieved codes and selects those categories that appear in the document. For each category identified in the previous phase, a search is performed on the second database. Unlike the previous phase, a similarity search is not applied here, but all existing subcategories associated with that category are extracted. Subsequently, an LLM called *Subcategory 1 ICD-10-ES Extractor* is used, which receives the clinical text and the possible level 1 subcategories and determines which of them appear in the document. Finally, for each level 1 subcategory identified, the process is repeated with the level 2 subcategory database, restricting the search to only those options associated with the parent subcategory. An additional LLM, *Subcategory 2 ICD-10-ES Extractor*, is used to select the most relevant level 2 subcategory regarding the clinical text content.

LLM as Classifier and Validator: in this approach, the large language model is prompted to directly generate CIE-10 codes using only the clinical text as context. This procedure may lead to assing inacurate or non-existen codes (in case of hallucination), so that a two-phase validation is performed afterwards: a filtering and a LLM-based validation steps.

Figure 4 shows the workflow developed. First, the codes extracted by the *ICD-10-ES Extractor* undergo a filtering phase in which any codes not present in the official CIE-10 nomenclature are discarded. The remaining codes are then subjected to an additional semantic validation step using a second LLM, referred to as the *ICD-10-ES Validator*. This model receives the candidate code along with its official description and the original text, and is tasked with determining whether the code is indeed supported by the content of the medical report. The result of this process is a set of codes that meet two fundamental criteria: (i) they belong to the official version of CIE-10 2024 and (ii) they have been semantically validated by the CIE Validator as being present in the clinical text analysed. This approach has the potential to enhance system accuracy and reduce the likelihood of generating irrelevant or non-existent codes.

3.2.2. Finetuned models adaptation

The alternative approach investigated is based on adapting pre-existing Spanish coding models, originally fine-tuned on an outdated version of the ICD-10-ES nomenclature, by leveraging the capabilities of large language models. The state-of-the-art model considered is BETO-Galén, a transformer-based

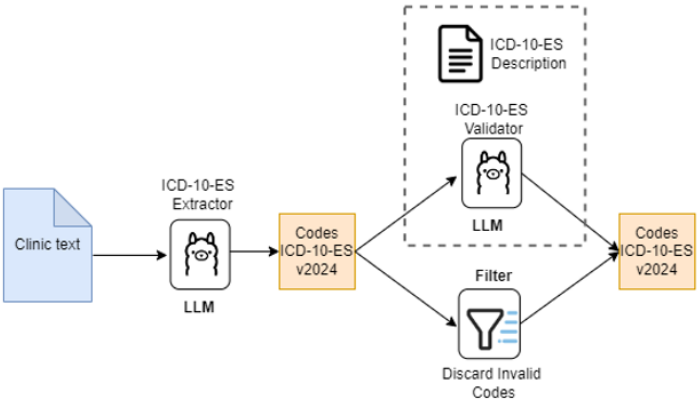

Figure 4: LLM classifier & validator workflow.

architecture fine-tuned on the CodiEsp corpus (2018 edition)[3]. To address the limitations imposed by its reliance on an older code set and to extend its classification coverage to the 2024 version of ICD-10-ES, a hybrid system is proposed. In this system, the codes predicted by BETO-Galén are combined with those generated by the LLM, using one of the retrieval strategies described previously (review section 3.2.1). Furthermore, as BETO-Galén is restricted to the label space defined by the CodiEsp dataset, the integration of LLM output enables broader coverage of codes, including those absent from the original training set.

This scheme aims to combine the prior knowledge encoded in a model trained on a curated corpus of clinical data in Spanish with the adaptability and up-to-date capacity of LLMs, potentially enhancing the overall accuracy and coverage of the classification system.

4. Evaluation

This section reports the results obtained from the evaluation of the proposed approaches over the MIMIC-IV subset assembled. The dataset is used exclusively for testing purposes, as no model training or fine-tuning is performed under the proposed LLM-based methodologies.

In addition to the performance metrics, qualitative analyses are provided to illustrate specific strengths and limitations of each method. Together, these results offer a comprehensive view of system behavior under the following experimental conditions:

- All LLM-based approaches invoke an open-source local model deployed via the Ollama[3] inference server, installed on a private server equipped with four NVIDIA H100 GPUs.
- The selected model for the coding task is 'Llama 3.1 Instruct 70B'[16], chosen for its capabilities in terms of scale, instruction alignment, and privacy.
- The selected model for the automatic machine translation of the dataset is 'Llama 3.3 70B', chosen for its performance and expanded multilingual support.
- The two approaches relying on RAG employ the native embedding representation of the Llama 3.1 model, and set the maximum number of codes retrieved during the retrieval phase to 100.
- The BETO-Galén model was downloaded from its official repository[4] and integrated unmodified.

Tables 1, 2 and 3 present the evaluation results for each of the proposed approaches, using standard metrics such as precision, recall, and F1 score. The performance of the original transformer-based model BETO-Galén is included as a baseline without LLM intervention, followed by its adaptation using the most effective LLM-based strategy (Classifier + Validator). The first and second tables assess

[3]https://ollama.com/
[4]https://github.com/guilopgar/ClinicalCodingTransformerES/tree/main/BETO/BETO-Galen

performance at higher levels of the code hierarchy (categories and three-digit subcategories respectively) while Table 3 addresses the coding of any level within the full hierarchy.

Table 1
Performance metrics on the category level on the translated MIMIC-IV subset.

Approach	Precision	Recall	F1
BETO-Galén	0.2120	0.1508	0.1763
RAG	0.0876	0.0359	0.0509
Hierarchical RAG	0.1680	0.1518	0.1595
Classifier-Validator	**0.6672**	0.1984	**0.3059**
BETO-Galén adaptation	0.3110	**0.2842**	0.2970

Table 2
Performance metrics on the subcategory 1 level on the translated MIMIC-IV subset.

Approach	Precision	Recall	F1
BETO-Galén	0.1596	0.1027	0.1250
RAG	0.0739	0.0148	0.0240
Hierarchical RAG	0.1412	0.0980	0.1577
Classifier-Validator	**0.4978**	0.1208	0.1944
BETO-Galén adaptation	0.2264	**0.1849**	**0.2036**

Table 3
Performance metrics on the full hierarchy on the translated MIMIC-IV subset.

Approach	Precision	Recall	F1
BETO-Galén	0.1544	0.0956	0.1180
RAG	0.1587	0.0108	0.0200
Hierarchical RAG	0.1424	0.0784	0.1011
Classifier-Validator	**0.4915**	0.0967	0.1617
BETO-Galén adaptation	0.2096	**0.1583**	**0.1804**

Additionally, a qualitative assessment is conducted to observe the cases where LLM-based methods produce unexpected codes. For example, in the admission 24453084 of patient 10384442, the text is tagged with the code I63231 (*Cerebral infarction due to unspecified occlusion or stenosis of the right carotid artery*), whereas the 'Classifier-Validator' model returns I639 (*Cerebral infarction, unspecified*) but also predicts I652 (*Occlusion and stenosis of carotid artery*). This shows that the models are capable of understanding the patient's condition and returning plausible codes, although a decrease in quantitative scores is observed. Another example is for patient 11524961, admission 21273404, where the text is labeled with F17210 (*Nicotine dependence, cigarettes*) and the model predicts F17200 (*Nicotine dependence*) and Z87891 (*Personal history of nicotine dependence*).

Given that one of the main goals of the research is to adapt easily to new codes, we conducted a specific assessment on these. In the evaluation sample, 248 texts were identified containing at least one new code introduced in the 2024 version. A total of 94 distinct codes were assigned, with the most frequent including I27.20, I27.21, and F10.11. When analyzing the results of the classifier-valuer model, we observed that it is capable of predicting codes not present in the 2018 version, such as I21.9 and I50.811, validating its ability to adapt to new terms. For instance, in admission 25366754 of patient 12541703, the text was labeled with I50.813 (*Acute right heart failure on chronic*), while the model predicted I50.811 (*Acute right heart failure*), both of which are new codes from the 2024 edition. Although not an exact match, the prediction remains accurate at the subcategory level (I50.81).

5. Conclusions and future work

This work has examined the challenges and opportunities associated with automatic clinical coding in Spanish, with a particular focus on aligning outputs with the 2024 version of the ICD-10-ES standard. The approaches explored, along with their empirical evaluation, provide insights into the feasibility and performance of both LLM-based and hybrid strategies within this domain.

In contrast to non-generative transformer-based models, such as BETO-Galén, which seems limited to shorter clinical narratives and a fixed code set, LLM-based methods operate over the full ICD-10-ES hierarchy, offering broader coding coverage and greater flexibility in handling extended and complex medical texts. Additionally, their architecture allows for seamless adaptation to updates in medical terminology, eliminating the need for retraining when classification standards evolve.

Quantitative results confirm the advantages of LLMs at higher hierarchical levels, where the Classifier-Validator strategy notably improves precision (0.6672) compared to the rest. However, performance diminishes at finer levels of granularity, where precision and recall decline. Despite this, qualitative analysis reveals that many predicted codes remain semantically appropriate, even when they are not exact matches, highlighting the ability of the models to approximate clinical reasoning and support coding tasks.

Overall, LLM-based approaches represent a scalable and updatable alternative for automatic clinical coding in Spanish, with demonstrated potential to enhance coverage, maintain relevance, and improve accuracy across diverse clinical scenarios.

Future work will focus on extending the study to a broader set of open-source LLMs, specifically on smaller versions, and exploring alternative configurations, including additional hybrid strategies that combine pretrained Spanish models with LLM components. An alternative worth exploring is the use of non-LLM models like PLM-ICD, trained on English clinical texts and aligned with ICD-10. This would involve translating Spanish texts into English, using PLM-ICD for coding, and then mapping results to CIE-10. Despite challenges in translation and code alignment, it presents a promising option given the lack of Spanish datasets updated to the 2024 standard.

Further efforts will also be devoted to refining the evaluation framework, incorporating metrics better suited to hierarchical and multi-label classification tasks, deeper error analysis to identify classification challenges, and inspection of the retrieval step, given the observed low recall, to detect potential gaps. Additionally, a qualitative analysis involving domain experts will be pursued to assess the clinical plausibility and semantic proximity of predicted codes. Finally, future research will consider the use of emerging LLM-based architectures, such as agentic systems, to enhance reasoning capabilities and support more complex coding workflows.

Acknowledgments

This research was funded by the Department of Big Data and Cognitive Systems at the Aragon Institute of Technology, under Retech Tourism-Spain Living Lab Agreement and by the Government of Aragon.

Declaration on Generative AI

During the preparation of this work, the authors used GPT to draft content, as well as Grammarly in order to check grammar and spelling. After using these tools, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] M. R. Hossain, S. Mahabub, A. A. Masum, I. Jahan, Natural Language Processing (NLP) in Analyzing Electronic Health Records for Better Decision Making, Journal of Computer Science and Technology Studies 6 (2024) 216–228. doi:10.32996/JCSTS.2024.6.5.18.

[2] A. Miranda-Escalada, A. Gonzalez-Agirre, J. Armengol-Estapé, M. Krallinger, Overview of Automatic Clinical Coding: Annotations, Guidelines, and Solutions for non-English Clinical Cases at CodiEsp Track of CLEF eHealth 2020, in: Working Notes of Conference and Labs of the Evaluation (CLEF) Forum, 2020.

[3] G. Lopez-Garcia, J. M. Jerez, N. Ribelles, E. Alba, F. J. Veredas, Transformers for clinical coding in spanish, IEEE Access 9 (2021) 72387–72397. doi:10.1109/ACCESS.2021.3080085.

[4] A. Johnson, L. Bulgarelli, T. Pollard, S. Horng, L. A. Celi, R. Mark, MIMIC-IV (version 3.0), PhysioNet. (2024). URL: https://physionet.org/content/mimiciv/3.0/. doi:10.13026/hxp0-hg59.

[5] A. E. Johnson, L. Bulgarelli, L. Shen, A. Gayles, A. Shammout, S. Horng, T. J. Pollard, B. Moody, B. G. et al., MIMIC-IV, a freely accessible electronic health record dataset, Scientific Data 2023 10:1 10 (2023) 1–9. doi:10.1038/s41597-022-01899-x.

[6] T.-T. Nguyen, V. Schlegel, A. Kashyap, S. Winkler, S.-S. Huang, J.-J. Liu, C.-J. Lin, Mimic-IV-ICD: A new benchmark for eXtreme MultiLabel Classification, 2023. URL: https://arxiv.org/abs/2304.13998v1. arXiv:2304.13998.

[7] C. W. Huang, S. C. Tsai, Y. N. Chen, PLM-ICD: Automatic ICD Coding with Pretrained Language Models, in: ClinicalNLP 2022 - 4th Workshop on Clinical Natural Language Processing, Association for Computational Linguistics (ACL), 2022, pp. 10–20. URL: https://aclanthology.org/2022.clinicalnlp-1.2/. doi:10.18653/V1/2022.CLINICALNLP-1.2.

[8] J. Edin, A. Junge, J. D. Havtorn, L. Borgholt, M. Maistro, T. Ruotsalo, L. Maaløe, Automated Medical Coding on MIMIC-III and MIMIC-IV: A Critical Review and Replicability Study, in: Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval, Association for Computing Machinery, Inc, 2023, pp. 2572–2582. doi:10.1145/3539618.3591918.

[9] Z. Boukhers, A. Khan, Q. Ramadan, C. Yang, Large Language Model in Medical Informatics: Direct Classification and Enhanced Text Representations for Automatic ICD Coding (2024). URL: https://arxiv.org/abs/2411.06823v1. arXiv:2411.06823.

[10] J. S. Boyle, A. Kascenas, P. Lok, M. Liakata, A. Q. O'Neil, Automated clinical coding using off-the-shelf large language models, in: Deep Generative Models for Health Workshop NeurIPS 2023, 2023. URL: https://openreview.net/forum?id=mqnR8rGWkn.

[11] Z. Yang, S. S. Batra, J. Stremmel, E. Halperin, Surpassing GPT-4 Medical Coding with a Two-Stage Approach (2023). URL: https://arxiv.org/abs/2311.13735v1. arXiv:2311.13735.

[12] A. Mustafa, U. Naseem, M. R. Azghadi, Large language models vs human for classifying clinical documents, International Journal of Medical Informatics 195 (2025) 105800. doi:10.1016/J.IJMEDINF.2025.105800.

[13] Y. Weng, M. Zhu, F. Xia, B. Li, S. He, S. Liu, B. Sun, K. Liu, J. Zhao, Large language models are better reasoners with self-verification, in: Findings of the Association for Computational Linguistics: EMNLP 2023, Association for Computational Linguistics (ACL), 2023, pp. 2550–2575. URL: https://aclanthology.org/2023.findings-emnlp.167/. doi:10.18653/V1/2023.FINDINGS-EMNLP.167.

[14] J. Gu, X. Jiang, Z. Shi, H. Tan, X. Zhai, C. Xu, W. Li, Y. Shen, S. Ma, H. L. et al., A Survey on LLM-as-a-Judge (2024). URL: https://arxiv.org/abs/2411.15594v5. arXiv:2411.15594.

[15] M. Douze, A. Guzhva, C. Deng, J. Johnson, G. Szilvasy, P.-E. Mazaré, M. Lomeli, L. Hosseini, H. Jégou, The Faiss library (2024). URL: https://arxiv.org/abs/2401.08281. arXiv:2401.08281.

[16] Meta, The llama 3 herd of models, 2024. URL: https://ai.meta.com/research/publications/the-llama-3-herd-of-models/.