Clinical Entity Recognition and Linking in Greek Discharge Letters using Multilingual-LLM-Based Multi-Stage System

Notebook for the BioASQ Task ELCardioCC Lab at CLEF 2025

Bor-Woei Huang[1]

[1]University of Padova, Italy

Abstract

We constructed a multi-stage, hierarchical system for the ELCardioCC task of clinical entity recognition and subsequent entity linking within the domain of cardiology. We integrated the capabilities of multilingual generative large language models (LLMs) and BERT encoder architectures across different phases of our Named Entity Recognition (NER) and Entity Linking (EL) pipeline. In the initial NER phase, we designed zero-shot prompts to instruct the LLMs in the extraction of relevant clinical mentions directly from the Greek discharge letters. These prompts also guided the models in generating accurate English translations of the identified Greek terms and in producing concise biomedical entity descriptions associated with these mentions. Further, to refine the initial set of extracted entities and enhance the overall precision of our NER results, we employed a BERT bi-encoder as a sophisticated filtering mechanism, designed to identify and remove likely false positives. Then, for the EL phase, we utilized a BERT cross-encoder as the core linking component. This model took both the previously extracted clinical mentions and their generated biomedical entity descriptions as input to establish accurate mappings to standardized concepts within the ICD-10 knowledge base. Finally, the linked ICD-10 codes obtained from the EL phase were collected for the MLC-X task. Our best system achieved an F1 score of 0.5761 on the NER task, 0.5336 on the EL task, and 0.7543 on the MLC-X task.

Keywords

Clinical Named Entity Recognition, Clinical Entity Linking, Multilingual Language Model, BERT Encoder

1. Introduction

Medical records can be made more statistically analyzable by transforming unstructured clinical text into standardized, searchable codes. This can support medical research, enable extensive retrospective studies, and uncover cause-and-effect links between illnesses and symptoms, all of which contribute to a better understanding of illnesses and their treatments. Moreover, structured data from correct coding can help doctors find important information about a patient's history, symptoms, and conditions more quickly. This can help them make differential diagnoses and give more personalized care.

Early Natural Language Processing (NLP) research heavily concentrated on English text in general contexts. However, the processing of non-English clinical data has experienced a notable increase in development in recent years, supplied by the growing availability of varied datasets and advanced multilingual language models. The ELCardioCC competition [1], within the BioASQ 2025 challenge [2, 3], specifically addresses this by focusing on automatically extracting clinical entities from Greek discharge summaries. This includes identifying the chief complaint, diagnosis, prior medical history, medications, and cardiac echo mentions. A subsequent task involves linking these extracted mentions to their corresponding ICD-10 codes, which provide a universal standard for classifying health conditions. The automation of the linking process offers improved collaboration and data sharing by ensuring consistency in the exchange of medical information across various healthcare systems, regions, and countries.

While numerous pre-trained language models (PLMs) exist for various languages beyond English, research in NER has frequently focused on fine-tuning these models with limited biomedical and clinical

CLEF 2025 Working Notes, 9 – 12 September 2025, Madrid, Spain

borwoei.huang@studenti.unipd.it (B. Huang)

data to create domain-specific solutions, particularly in low-resource scenarios [4, 5, 6, 7, 8]. In contrast, EL in the medical domain for languages other than English has received considerably less attention, may due to the lack of corresponding language-specific knowledge base and readily available pre-trained langauge models. With most existing work stemming from specific challenges like CLEF eHealth, IberLEF, BioASQ/DisTEMIST, and DEFT, primarily focuses on Spanish and French. To the best of our knowledge, the application of EL to Greek clinical text remains an unexplored area.

Multilingual language models (MLMs) are useful tools, trained on enormous datasets encompassing a wide array of languages. This inherent capability to understand Greek text is particularly beneficial for the NER task on Greek clinical discharge letters. However, while MLMs excel as generalists in language comprehension and generation across diverse languages, they are not inherently medical domain experts. This lack of specialized clinical knowledge presents significant challenges, as medical terminology is replete with nuances, including an abundance of synonyms, context-dependent meanings, and specialized abbreviations. A general MLM, without explicit clinical training, often struggles to accurately disambiguate these terms. To address these complexities, we adopted an integrated approach. We leveraged several MLMs to tackle the challenges of cross-lingual NER, aiming to identify clinical entities in Greek. Following this initial clinical entity recognition phase, we then employed a two-stage hybrid retrieval approach. This approach was designed to establish a connection, linking the identified clinical entities from the Greek text to their corresponding standardized concepts within the comprehensive ICD-10 knowledge base.

2. Related Work

There are primarily two approaches to handing the NER and EL tasks: information retrieval system and multitask framework systems. Information retrieval approach system retrieves similar instance data from knowledge bases, based on lexical overlap or semantic similarity [9, 10]. Hierarchical framework NER and EL systems that employ distinct models for each task, with the first model dedicated to identifying and classifying entities, and a subsequent, separate model responsible for linking these identified mentions to a knowledge base [11]. Hierarchical NER and EL systems with separate models for NER and EL are conceptually straightforward, however, they are susceptible to error propagation. This occurs because errors made in the initial NER stage directly impact the performance of the downstream EL stage. If an entity is incorrectly identified, missed entirely, or has incorrect boundaries assigned by the NER model, the followup EL model will struggle to link it to the correct knowledge base entry, leading to a cascade of errors throughout the pipeline.

Multitask framework systems come in two main forms, joint model and end-to-end model, to reduce the error propagation issues in the hierarchical framework. Joint models combine two models, NER and EL tasks are performed in parallel by a single transformer model [12, 13, 14, 15]. End-to-End models take raw text as input and directly output linked entities, without a clearly defined intermediate NER stage. This forces the model to learn both identification and linking in a unified manner [16].

3. Methodology

3.1. Dataset

In ELCardioCC challenge, the training dataset consists of 1,000 Greek-language discharge summaries, while the test dataset includes an additional 500 discharge letters. These clinical documents contain complex information detailing patients' diseases, symptoms, diagnoses, therapeutic interventions, and clinical outcomes. The training corpus has been manually annotated to identify the exact spans of key biomedical mentions. Each annotated entity is linked to its corresponding code in the 10th revision of the International Classification of Diseases (ICD-10), a globally recognized medical taxonomy maintained by the World Health Organization (WHO)[1].

[1] https://www.cdc.gov/nchs/icd/icd-10-cm/

3.2. System

The pipeline for the cardio discharge letter NER and EL is illustrated in Figure 1. For the NER task, we prompted LLMs Gemma-3, Phi-4 and Gemini to retrieve clinical mentions and generate English translations and their descriptions. Potential false positives were then discarded by a NER filter. Moving to the EL task, the English-translated mentions from the NER phase were combined with their generated clinical entity descriptions and fed into the entity linker to find their corresponding ICD-10 codes. Finally, for the MLC-X task, we simply collected the ICD-10 codes obtained during the EL phase.

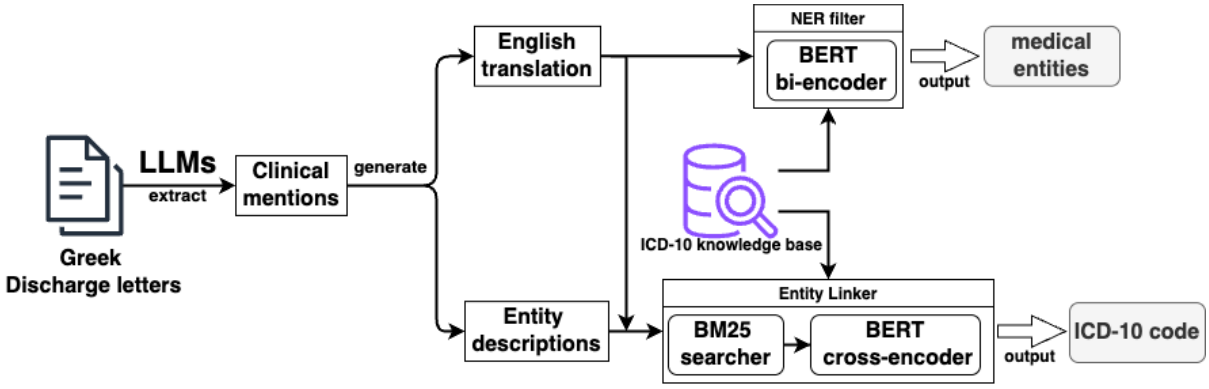

Figure 1: Overview of the cardio discharge letter NER and EL pipeline.

3.2.1. Named entity recognition phase

Multilingual large language models: Gemma-3, Phi-4 and Gemini are trained on multilingual corpora, can be used to understand Greek discharge letters and extract clinical mentions. During the NER phase, clinical terms in Greek, such as disease names, symptoms, and treatment references, were directly translated to their standardized English equivalents. Our approach is prompt engineering, where the model's outputs were guided by constructed instructions that correspond to the clinical domain. These prompts were designed to maximize recall without the need for supervised training.

Zero-shot prompt: We employed two separate prompts (detailed in Table 1) to extract medical mentions from the Greek discharge letters. Prompt 1 aimed for broad coverage, seeking to identify all medical terms and phrases present in the text. Prompt 2, however, specifically targeted mentions related to diseases, syndrome, and their treatments. The output from the prompt 1 achieves high recall but includes many general medical terms not directly relevant to the cardiovascular diseases or symptoms, necessitating a subsequent filtering step. Conversely, prompt 2 was designed to yield mentions that were more directly relevant to the entities we are seeking to extract.

Table 1
Two prompts used in zero-shot prompting.

prompt 1	Given Greek text {discharge letter}. Extract the medical terms from this text and write English translation and concise descriptions of them in JSON format (key is the Greek medical term, value is English translation and description pair tuple).
prompt 2	Given Greek text {discharge letter}. Extract the disease, syndrome, and treatment terms from this text and write English translation and concise descriptions of them in JSON format (key is the Greek medical term, value is English translation and description pair tuple).

The quantity of text provided to a language model influences the identification of medical terms within Greek discharge letters. To ensure comprehensive retrieval of all relevant mentions, we employed a dual strategy: First, we processed the entire discharge letter as a single unit, which en-

ables the language model to grasp the broader connections and interdependencies among various medical mentions throughout the document. Second, we processed the letters section by section (e.g., ΙΣΤΟΡΙΚΟ – ΑΝΤΙΚΕΙΜΕΝΙΚΗ ΕΞΕΤΑΣΗ (HISTORY – OBJECTIVE EXAMINATION), ΑΙΤΙΑ ΕΙΣΟΔΟΥ-ΑΝΤΙΚΕΙΜΕΝΙΚΗ ΕΞΕΤΕΤΑΣΗ-ΙΣΤΟΡΙΚΟ (REASON FOR ADMISSION – OBJECTIVE EXAMINATION - HISTORY), ΠΟΡΕΙΑ ΝΟΣΟΥ (COURSE OF DISEASE), ΕΡΓΑΣΤΗΡΙΑΚΕΣ ΕΞΕΤΑΣΕΙΣ (LABORATORY TESTS) etc.), afterward collected the extracted terms from each individual section. This segmented processing allows the model to concentrate on the immediate context and pick out specific mentions within each part.

Beyond identifying clinical mentions in the Greek text, our prompts also directed the LLMs to produce both English translations and brief descriptions for each extracted entity. These generated English translations and descriptions were specifically created to support the subsequent entity linking process. Table 2 provides an illustrative example of the LLM's output for a segment of a Greek discharge letter, displaying the extracted Greek mentions along with their corresponding English translations and generated descriptions.

Table 2
Example of entity extraction, translation, and description generation by multilingual large language models

Text: ΙΣΤΟΡΙΚΟ – ΑΝΤΙΚΕΙΜΕΝΙΚΗ ΕΞΕΤΑΣΗ Η ασθενής προσεκομίσθη λόγω προκάρδιου άλγους από 2 ώρου και αιμωδίες άνω άκρων και συνοδό εφίδρωση και έμετο. ΗΚΓ εισόδου : SR, LBBB, οριακή ανάσπαση ST II, III, AVF, V3- V5. Κλινική Εξέταση: ΑΠ: 180/100 mmHg, SaO2: 93 %, S1S2 : ρυθμικοί – ευκρινείς ακρόαση πνευμόνων: φυσιολογικό ΑΨ Α/α θώρακα: Διάσπαρτα επασβεστωμένα κοκκιώματα στον αριστερό πνεύμονα.Χωρίς εικόνα ενεργού πνευμονικής διήθησης. Ατομικό Αναμνηστικό: , πνευμοθώρακα, αναιμία		
Mentions	English Translations	Generated Descriptions
προκάρδιο άλγος	precordial pain	Pain in the chest, specifically around the heart.
αιμωδίες άνω άκρων	cyanosis of upper extremities	Bluish discoloration of the skin in the upper limbs due to poor oxygenation.
εφίδρωση	sweating	The excretion of fluid and electrolytes from sweat glands.
έμετο	vomiting	The forceful expulsion of contents of the stomach through the mouth.
LBBB	Left Bundle Branch Block	A conduction defect in the heart where the electrical impulse is delayed or blocked along the left bundle branch.
ανάσπαση ST	ST segment depression	A downward deflection of the ST segment on an electrocardiogram, often indicating myocardial ischemia.
επασβεστωμένα κοκκιώματα	calcified granulomas	Small nodules composed of immune cells and calcium deposits, often indicating past infection or inflammation.
αναιμία	anemia	A condition characterized by a deficiency of red blood cells or hemoglobin, resulting in reduced oxygen-carrying capacity of the blood.

Entity filter This step focuses on refining the initial set of extracted medical terms from the multilingual language models by removing likely false positives. To enhance the accuracy of these initial extractions, we filter out less relevant mentions by assessing their semantic similarity to ICD-10 English disease and symptom category definitions. This filtering employs a BERT bi-encoder to identify and eliminate irrelevant extracted mentions. Specifically, a MedEmbed BERT bi-encoder[2] [17] generates vector embeddings for each English-translated biomedical term previously identified in the Greek discharge letter. This bi-encoder independently processes each English-translated biomedical term, creating a dense, context-aware representation of its meaning. Similarly, the ICD-10 English category

[2]https://huggingface.co/abhinand/MedEmbed-large-v0.1

definitions are also embedded using the same BERT bi-encoder. For each extracted English-translated term, we calculate the cosine similarity between its embedding and the embeddings of the ICD-10 descriptions. A predefined similarity score threshold of 0.37 is then applied. This specific threshold of 0.37 was established through testing on the training dataset, primarily to enhance NER F1 scores within that dataset. Consequently, any extracted terms with a maximum cosine similarity score falling below this threshold are deemed less relevant and are subsequently discarded. Note that this similarity score threshold was determined prior to our ensemble process, meaning it might not represent the globally optimal threshold for the entire pipeline. Achieving such an optimal threshold would necessitate an end-to-end tuning approach, where the influence of the subsequent ensemble process is explicitly factored into the optimization.

Ensemble By combining the medical terms identified from the full discharge letter and its segmented sections, using our two distinct prompting strategies, we increase the likelihood of achieving a more comprehensive collection of relevant mentions as some terms that might be missed when the document is processed solely as a whole or in isolated sections.

The recognition of **long-tail entities** (highly specific or multi-word terms exceeding four words) and **nested entity mentions** (where shorter, valid entities exist within longer ones) varies across different language models. For instance, in the long-tail entity "Προκάρδιο άλγος από 3ημέρου συσφικτικό με αντανάκλαση στον ΑΡ αγκώνα" (Precordial pain from 3 days constrictive with reflection to the left elbow), some models, like Gemma-3, only extract "Προκάρδιο άλγος" (Precordial pain), completely missing the extended description. Nested entities like "Συνδρομο Ταχυκαρδιας-Βραδυκαρδιας" (Tachycardia-Bradycardia Syndrome), which contains the individual medical terms "Ταχυκαρδιας" (Tachycardia) and "Βραδυκαρδιας" (Bradycardia), often results in different identifications across models.

To mitigate the inconsistencies in identifying potential long-tail and nested entities across different language models, we aggregated the terms identified by these models and applied selection criteria based on either term length or majority voting. (1) Term length prioritization approach: for both long-tail and nested entities, the longest term length method selected the most extensive span among the extractions from different models. This ensured that more complete or encompassing phrases were preferred. (2) Majority voting approach: selected an entity span only if it was extracted by more than half of the models. This method utilizes consensus to enhance the reliability of the identified entities.

Table 3
Ensemble configurations applied for the 5 submissions

runs	Ensemble methods
1nm	Gemma-3's Prompt 2 output with section-wise processing (incorporating long-tail and nested entities from other LLMs' extractions)
2nm	union Run 1nm's output with Gemini's Prompt 2 output with entire letter processing
3nm	Gemma-3's Prompt 1 output with section-wise processing (incorporating long-tail and nested entities from other LLMs' extractions)
4nm	union Run 3nm's output with Gemini's Prompt 2 output with entire letter processing
5nm	Fused the outputs of Run 1nm, Run 3nm, and Gemini (Use majority voting method)

3.2.2. Entity linking phase

We implemented search engine style algorithm for the entity linking task, the widely used efficient two-phase hybrid retrieval system, to speed up the entity linking process. The first step is using the high-recall bag-of-words retrieval function BM25 to efficiently narrow down the vast ICD-10 knowledge base to a relatively small set of candidate codes for a given English translated mention extracted in NER phase. The correct ICD-10 code is highly likely to be within this candidate set. The second step is to take the ICD-10 candidates generated by BM25 and link to the ICD-10 code with the highest score by BERT

cross-encoder that can understand the semantic context of the mention and the entity descriptions with higher precision.

First-stage candidate generation: To generate a set of potential ICD-10 codes relevant to the extracted Greek medical mentions, we employed a BM25 searcher. Our method involved preparing the English ICD-10 knowledge base for efficient searching. We treated each ICD-10 code English category definition as a distinct document. To strengthen the content of these "documents" and improve the probable success of a match, we concatenated the textual description associated with each ICD-10 code with the descriptions of all its more specific subcategories within the ICD-10 hierarchy of categories. This effectively creates a more comprehensive textual representation for each ICD-10 concept. Following this, for each English translation of a medical mention extracted from the Greek discharge letter, we used it as a query to search across these constructed ICD-10 definition "documents" using the BM25 algorithm, which is defined as:

$$\text{Relevance Score (D, Q)} = \sum_{q_i \in Q} \text{IDF}(q_i) \cdot \frac{tf(q_i, D) \cdot (k_1 + 1)}{tf(q_i, D) + k_1 \cdot (1 - b + b \cdot \frac{|D|}{\text{avgdl}})} \tag{1}$$

where Q is an English medical term, IDF(q_i) represents the inverse document frequency weight of the medical mention q_i, and $tf(q_i, D)$ denotes the frequency of a medical mention q_i within the description document D. $|D|$ is the length of the description document, avgdl is the average document length in the collection. $k_1 = 0.5$ and $b = 0.3$ are assigned.

Second-stage linking In the second-stage of EL, to refine the candidate ICD-10 codes retrieved by BM25, we formulated new queries. These queries were constructed by combining the English translation of the extracted Greek medical mention with the concise descriptions of these mentions generated by the multilingual LLMs. Using the set of ICD-10 codes identified as potential candidates in the BM25 retrieval stage, we then employed a MedCPT cross-encoder[3] [18] to calculate a relevance score between our augmented query (English translated mention + LLM-generated English description) and constructed ICD-10 English definition documents. We established the link by selecting the ICD-10 code that received the highest relevance score from the MedCPT cross-encoder.

To clarify the roles of our BERT models, the cross-encoder in this EL stage is designed for precise identification of the definitive ICD-10 descriptions. This differs from the BERT bi-encoder, which served as a filtering mechanism in the previous NER phase, merely removing potential false positives without directly linking to ICD-10 codes. We must also note an oversight: we initially missed the ICD-10 candidate list provided in labelset.txt. This led us to implement a BM25 searcher for acceleration, though its necessity is unlikely given that the list contains only 324 candidate codes.

3.2.3. Multi-label classification phase

Our submissions for the MLC-X task were simply direct aggregations of all the ICD-10 codes mapped during the EL phases.

4. Results and Discussion

Tables 4, 5, and 6 present the results of our five runs for the NER, EL, and MLC-X tasks, respectively.

4.1. Results

Our system demonstrated generally poor precision across the NER, EL, and MLC-X tasks, which significantly hindered our overall F1 scores. The primary reason for this lies in our prompt design,

[3]https://huggingface.co/ncbi/MedCPT-Cross-Encoder

particularly Prompt 1, which was optimized for maximizing recall. This led to the extraction of an excessive number of medical terms that fell outside the cardiology domain (i.e., not in the code list in labelset.txt file).

Since we utilized a hierarchical framework system for the NER, EL and MLC-X tasks, the evaluations of the EL results are directly influenced by the outputs of the NER phase. The correlations between these two tasks are clearly visible in table 4 and 5. Among our five submissions, Run 1nm achieved the highest precision for both the NER and EL tasks. Conversely, Run 4nm demonstrated the highest recall, though its lowest precision resulted in the lowest F1 score. Run 5nm secured the highest F1 score overall, despite not having the top recall or precision in either the NER or EL tasks. This outcome suggests that our majority voting method, used to fuse the outputs from different LLMs, was particularly effective in improving both precision and F1 scores by balancing these metrics.

Table 4
The NER task evaluations for each of our five submissions

runs	Recall	Precision	F1
1nm	0.4914	0.5331	0.5114
2nm	0.6338	0.4	0.4905
3nm	0.546	0.4387	0.4865
4nm	0.6575	0.3601	0.4653
5nm*	0.6448	0.5205	0.5761

Table 5
The EL task evaluations for each of our five submissions

runs	Recall	Precision	F1
1nm	0.448	0.5	0.4726
2nm	0.5734	0.3677	0.448
3nm	0.4963	0.4211	0.4556
4nm	0.5945	0.3374	0.4305
5nm*	0.5927	0.4852	0.5336

Table 6 reveals a particularly high recall for the MLC-X task, which contributed to moderate F1 scores for these runs, despite the persistently miserable precision. As the MLC-X outputs are downstream from our NER and EL phases, a similar trend in submission results is observable. Run 4nm, for instance, achieved the highest recall among the five submissions but simultaneously suffered from the lowest precision. Run 5nm demonstrated the highest precision, which also correlated with its highest F1 score among our five submissions.

Table 6
The MLC-X task evaluations for each of our five submissions

runs	Recall	Precision	F1
1nm	0.7676	0.6205	0.6863
2nm	0.8355	0.5131	0.6358
3nm	0.8237	0.5227	0.6395
4nm	0.8576	0.4572	0.5965
5nm*	0.825	0.6947	0.7543

4.2. Discussion

Upon analyzing the performance of our system, we identified several key factors that significantly influenced the outcomes of the 3 tasks.

Translation accuracy: The accuracy of the initial translation of Greek entity names is paramount, as it directly influences the success of subsequent steps. An incorrect English translation will result

in a mismatched entity description, inevitably leading to erroneous entity linking, regardless of the linking model's inherent capabilities. Language models tend to be more accurate when processing full entity names compared to their abbreviated forms. Table 7 illustrates instances where Greek medical abbreviations can have multiple English translations and consequently generate divergent descriptions, ultimately causing the entity linking process to point to incorrect ICD-10 codes

Table 7
Ambiguity of translation examples

Mentions	English Translation	Generated Description	Entity Linking
ΣΝ	annotation : **I25** - Chronic ischemic heart disease		
	Coronary Artery Disease	A condition in which plaque builds up inside the coronary arteries.	I25
	Heart Failure	A condition in which the heart is unable to pump enough blood to meet the body's needs.	I50
ΔΛΔ	annotation : **E78** - Disorders of lipoprotein metabolism and other lipidemias		
	Dyslipidemia	An abnormal amount of lipids (e.g., cholesterol, triglycerides) in the blood.	E78
	Deep Vein Thrombosis	A blood clot that forms in a deep vein, usually in the leg.	I81
	Peripheral Artery Disease	Condition where narrowed arteries reduce blood flow to the limbs.	I79

Prompt design and section-wise processing: The higher recall but higher false positive rate observed using Prompt 1 with section-wise processing can be attributed to several factors. Language models can exhibit positional biases, leading to missed entities in lengthy documents. Furthermore, in the lengthy text, the signal for specific entities can be weakened by the presence of substantial irrelevant information, making identification more challenging. Processing the document section by section allows the model to focus on smaller chunks of text. This approach helps to overcome positional biases and reduces the amount of noise within a given processing window, thereby improving the chances of retrieving most of the desired entities. However, this segmentation can also lead to the extraction of unwanted noise mentions that might not have been identified if the model had the broader context of the entire document.

Nested entity mention and long tail entity challenges: Further complicating the NER task are the inherent challenges of nested entity mentions and long tail entities. Our NER system significantly struggled with both long-tail entities and nested entity mentions; even medical mentions extracted by the LLMs that successfully passed the NER filter were ultimately considered incorrect in the NER evaluations, despite being the terms we aimed to extract.

Different LLMs often analyze and handle nested entity mentions in varying ways, leading to discrepancies in their output. As seen in table 8, consider the nested entity "ΣΝ (PCI)". While "ΣΝ" represents a type of disease, "PCI" denotes a specific heart treatment. The mention of "PCI" within "ΣΝ" indicates a patient with the disease who has undergone PCI. Another example is "ΟΣΣ-PCI LAD," which includes both "ΟΣΣ" (acute myocardial infarction) and "PCI LAD" (PCI specifically targeting the Left Anterior Descending artery), describing a patient who received that targeted treatment. These deviations can also affect the downstream EL results. Consider the long-tail entity "Υπόχρωμη μικροκυτταρική αναιμία" (Hypochromic microcytic anemia), which is a specific type of "αναιμία" (anemia), and each of these terms could potentially link to distinct ICD-10 codes.

Despite our selection process prioritizing either the largest encompassing span or the entity span that received the most votes from different model outputs, these methods often failed to correctly capture such diverse long-tail and nested entities.

Table 8
Examples of long-tail and nested entity mentions.

Mentions	English Translation	Generated Description	Entity Linking
αναιμία	annotation : **D64** - Other anemias		
	anemia	A condition in which you lack enough healthy red blood cells to carry adequate oxygen to your body's tissues.	D64
Υπόχρωμη μικροκυτταρική αναιμία	Not annotated		
	Hypochromic microcytic anemia	A type of anemia characterized by red blood cells that are smaller than normal and have a reduced amount of hemoglobin, resulting in a pale color.	D50
ΣΝ	annotation : **I25** - Chronic ischemic heart disease		
	Coronary Artery Disease	A condition in which plaque builds up inside the coronary arteries.	I25
PCI	annotation : **Z95** - Presence of cardiac and vascular implants and grafts		
	Percutaneous Coronary Intervention	A non-surgical procedure used to treat narrowing of the coronary arteries, typically involving angioplasty and stenting.	I95
ΣΝ (PCI)	annotation : **I25** - Chronic ischemic heart disease		
	CAD (PCI)	Coronary Artery Disease with a history of percutaneous coronary intervention (angioplasty and/or stenting) of the Left Circumflex coronary artery in the specified year.	I25
ΟΣΣ	annotation : **I21** - Acute myocardial infarction		
	ACS (Acute Coronary Syndrome)	A range of conditions associated with sudden, reduced blood flow to the heart.	I21
PCI LAD	annotation : **Z95** - Presence of cardiac and vascular implants and grafts		
	Percutaneous Coronary Intervention (Left Anterior Descending artery)	A procedure to open blocked coronary arteries, specifically the left anterior descending artery.	I21
ΟΣΣ-PCI LAD	annotation : **I24** - Acute coronary thrombosis not resulting in myocardial infarction		
	NOT retrieved		

5. Conclusion

We developed a multi-stage, hierarchical system for the ELCardioCC task, focusing on medical entity recognition and entity linking. While our system demonstrated moderate recall, its precision was noticeably low. A clear trend of error propagation was observable across the multi-stage pipeline, directly impacting results from the NER phase down to the EL and MLC-X tasks. A significant contributing factor to this limitation is our system's reliance on English-translated text for processing. The crucial bridge between the Greek clinical corpus and its English translation is entirely dependent on the multilingual LLMs, and the quality control of these translations remains beyond our current capabilities.

To further improve precision and f1, we need to design more precise prompts that maintain moderate recall, and implement more robust filtering mechanisms to effectively discard false positive mentions. This is a key area for future development to refine the accuracy of our extracted and linked clinical entities. A significant oversight on our part was neglecting the ICD-10 code list provided in the labelset.txt file. This list would have dramatically reduced the search space for ICD-10 codes from over 10,000 to just 324 candidates. Using this candidate list could have substantially improved both the precision and F1 scores in our ELCardioCC challenge submissions.

Declaration on Generative AI

During the preparation of this manuscript, the author utilized Gemini as a language refinement tool. All content was reviewed and edited by the author, who takes full responsibility for the accuracy and integrity of the publication.

References

[1] D. Dimitriadis, V. Patsiou, E. Stoikopoulou, A. Toumpas, A. Kipouros, D. Papadopoulos, A. Bekiaridou, K. Barmpagiannos, A. Vasilopoulou, A. Barmpagiannos, A. Samaras, G. Giannakoulas, G. Tsoumakas, Overview of ElCardioCC Task on Clinical Coding in Cardiology at BioASQ 2025, in: G. Faggioli, N. Ferro, P. Rosso, D. Spina (Eds.), CLEF 2025 Working Notes, 2025.

[2] A. Nentidis, G. Katsimpras, A. Krithara, M. Krallinger, M. Rodríguez-Ortega, E. Rodriguez-López, N. Loukachevitch, A. Sakhovskiy, E. Tutubalina, D. Dimitriadis, G. Tsoumakas, G. Giannakoulas, A. Bekiaridou, A. Samaras, F. N. Maria Di Nunzio, Giorgio, S. Marchesin, M. Martinelli, G. Silvello, G. Paliouras, Overview of BioASQ 2025: The thirteenth BioASQ challenge on large-scale biomedical semantic indexing and question answering, in: L. P. A. G. S. d. H. J. M. F. P. P. R. D. S. G. F. N. F. Jorge Carrillo-de Albornoz, Julio Gonzalo (Ed.), Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Sixteenth International Conference of the CLEF Association (CLEF 2025), 2025.

[3] G. Tsatsaronis, G. Balikas, P. Malakasiotis, I. Partalas, M. Zschunke, M. R. Alvers, D. Weissenborn, A. Krithara, S. Petridis, D. Polychronopoulos, et al., An overview of the bioasq large-scale biomedical semantic indexing and question answering competition, BMC bioinformatics 16 (2015) 1–28.

[4] F. Liu, I. Vulić, A. Korhonen, N. Collier, Learning domain-specialised representations for cross-lingual biomedical entity linking, arXiv preprint arXiv:2105.14398 (2021).

[5] N. Boudjellal, H. Zhang, A. Khan, A. Ahmad, R. Naseem, J. Shang, L. Dai, Abioner: A bert-based model for arabic biomedical named-entity recognition, Complexity 2021 (2021) 6633213.

[6] E. T. Rubel Schneider, J. V. Andrioli de Souza, J. Knafou, L. E. Oliveira, Y. B. Gumiel, L. F. de Oliveira, D. Teodoro, E. C. Paraiso, C. Moro, et al., Biobertpt: a portuguese neural language model for clinical named entity recognition, in: Proceedings of the 3rd Clinical Natural Language Processing Workshop, 19 November 2020, 2020.

[7] Y. Kim, J.-H. Kim, J. M. Lee, M. J. Jang, Y. J. Yum, S. Kim, U. Shin, Y.-M. Kim, H. J. Joo, S. Song, A pre-trained bert for korean medical natural language processing, Scientific reports 12 (2022) 13847.

[8] M. Mitrofan, V. Păiș, Improving romanian bioner using a biologically inspired system, in: Proceedings of the 21st Workshop on Biomedical Language Processing, 2022, pp. 316–322.

[9] A. Ceccato, L. Fabbian, B.-W. Huang, I. U. Khan, H. Singh, N. Ferro, et al., Seupd@ clef: Team hiball on incremental information retrieval system with rrf and bert, in: CEUR WORKSHOP PROCEEDINGS, volume 3497, CEUR-WS, 2023, pp. 2396–2415.

[10] B.-W. Huang, Generative large language models augmented hybrid retrieval system for biomedical question answering, CLEF Working Notes (2024).

[11] L. Logeswaran, M.-W. Chang, K. Lee, K. Toutanova, J. Devlin, H. Lee, Zero-shot entity linking by reading entity descriptions, arXiv preprint arXiv:1906.07348 (2019).

[12] S. Garda, U. Leser, Belhd: improving biomedical entity linking with homonym disambiguation, Bioinformatics 40 (2024) btae474.

[13] G. López-García, J. M. Jerez, N. Ribelles, E. Alba, F. J. Veredas, Explainable clinical coding with in-domain adapted transformers, Journal of Biomedical Informatics 139 (2023) 104323.

[14] Y. Xiong, Y. Huang, Q. Chen, X. Wang, Y. Nic, B. Tang, A joint model for medical named entity recognition and normalization, Proceedings http://ceur-ws. org ISSN 1613 (2020) 17.

[15] B. Zhou, X. Cai, Y. Zhang, X. Yuan, An end-to-end progressive multi-task learning framework for medical named entity recognition and normalization, in: Proceedings of the 59th Annual Meeting

of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), 2021, pp. 6214–6224.

[16] S. Ujiie, H. Iso, S. Yada, S. Wakamiya, E. Aramaki, End-to-end biomedical entity linking with span-based dictionary matching, arXiv preprint arXiv:2104.10493 (2021).

[17] A. Balachandran, Medembed: Medical-focused embedding models, 2024. URL: https://github.com/abhinand5/MedEmbed.

[18] Q. Jin, W. Kim, Q. Chen, D. C. Comeau, L. Yeganova, W. J. Wilbur, Z. Lu, Medcpt: Contrastive pre-trained transformers with large-scale pubmed search logs for zero-shot biomedical information retrieval, Bioinformatics 39 (2023) btad651.