Enigma @ ELCardioCC: Bridging NER and ICD-10 Entity Linking - A Hybrid Method for Greek Clinical Narratives

Boris Velichkov[1,*,†], Aleksis Datseris[1,2,†], Sylvia Vassileva[1,†] and Svetla Boytcheva[1,2,†]

[1]Faculty of Mathematics and Informatics, Sofia University St. Kliment Ohridski, Sofia, Bulgaria
[2]Graphwise, Sofia, Bulgaria

Abstract

This paper presents an approach for the clinical term Named Entity Recognition (NER) and Entity Linking (EL) in Greek clinical texts. The approach was developed as part of the ELCardioCC shared task for clinical coding to the International Classification of Diseases, 10th edition (ICD-10). For the NER task, we used different BERT-based models, the monolingual Greek BERT and the multilingual XLM-RoBERTa. We adapted them to the biomedical domain by additional pretraining on biomedical texts in Greek. We further fine-tuned the models for token classification on the train set to determine the ICD-10 term mentions in the text. The best F1 score we achieved was 0.7167 on the test set. For the EL, we used a hybrid approach that combined two stages. The first stage was based on a gazetteer - exact match or statistical match to unambiguous terms in a gazetteer compiled from the train set, ICD-10 specification, and other public resources. The second stage was a fine-tuned bi-encoder model (BAAI/bge-m3), applied only to mentions that did not match anything in the first stage. Our best F1 score on this task was 0.6693.

Keywords

Named Entity Recognition (NER), Biomedical NLP, Entity Linking (EL), Clinical NER, Clinical Entity Linking, Clinical Coding, Greek NER

1. Introduction

Coding medical diagnoses and procedures in patient medical records with the International Statistical Classification of Diseases and Related Health Problems - 10th Revision (ICD-10) [1] is of particular importance, both for health management and reimbursement, as well as for medical insurance, and for statistical analyses of mortality and disease prevalence. ICD-10 has established itself as an international standard in this field and is used in over 100 countries, and has been translated into over 40 languages.

The development of Natural Language Processing (NLP) models for the ICD-10 coding task would significantly help both in automating the process of coding information from patient records and in scientific research to study complex processes and relationships described in medical documentation.

In this paper, we present our results of developing models for NER and EL to ICD-10 codes of discharge summaries in Greek. The presented results are part of the ELCardioCC@CLEF 2025 shared task[2] in the BioASQ 2025 Lab competition[3] [1, 2]. The ELCardioCC task aims to develop systems for automatic recognition and coding of information from medical records with ICD-10. The organizers of the competition provided a specialized corpus of Greek-language written discharge summaries from a hospital cardiology department for this purpose. The named entities annotated with ICD-10 codes include chief complaint, diagnoses, past medical history, medications, and cardiac echo. The ELCardioCC competition is organized into three subtasks: (1) Named Entity Recognition (NER) from clinical records; (2) Entity Linking (EL) to ICD-10; and (3) Multi-Label Classification with Explainable AI (MLC-X).

CLEF 2025 Working Notes, 9 – 12 September 2025, Madrid, Spain
*Corresponding author.
boris.velichkov@fmi.uni-sofia.bg (B. Velichkov); aleksisdatseris@gmail.com (A. Datseris); svasileva@fmi.uni-sofia.bg (S. Vassileva); svetla.boytcheva@graphwise.ai (S. Boytcheva)

[1]https://icd.who.int/browse10/2019/en
[2]https://elcardiocc.web.auth.gr/
[3]https://www.bioasq.org/participate/challenges

We propose deep learning models including XLM-RoBERTa [3], Greek BERT [4], GreekDeBERTa [4], Greek-Reddit-BERT [5], BGE-M3 [6], umt5-xl [7], CohereLabs/aya-101 [8] for NER task and BGE-M3, SapBERT [9] and custom prepared dictionaries for EL to ICD-10 task.

This paper is organized as follows: Section 2 overviews NLP methods for NER and EL to ICD-10 of clinical documents; Section 3 describes the dataset provided by ELCardioCC challenge organizers and briefs the process of collection and processing of additional biomedical data related to the task; Section 4 presents in details the proposed methods and their modification and fine-tuning for language adaptation and domain adaptation; Section 5 reports evaluation results, discusses the limitation of the proposed approach and provides error analysis; Section 6 sketches further work and summarizes the proposed solution. All code used for data preprocessing, model training, and evaluation is available at: https://github.com/BorisVelichkov/enigma-elcardiocc.

2. Related Work

2.1. Greek Language Resources

NLP development for lesser-resourced languages faces a lot of challenges, mainly due to data scarcity for general-purpose tasks or in specific domains like the biomedical domain. Greek can be classified as a lesser-resourced language as it has fewer resources than the high-resource languages like English, Chinese (Mandarin), and Spanish [10]. Different tools were developed for Greek NLP, for example, the Greek NLP Toolkit[5], which addresses common NLP tasks for Greek in the general domain like NER, Part of Speech (PoS) tagging, dependency parsing, etc. In the biomedical domain, a parallel dataset (English-Greek) with abstracts and public website data was collected [6]. More resources exist consisting of term lists (Image, Sound, and Language Processing [7], and ICD-10 or GPC codes (Ketekny medical codes [8], icd10-in-Greek[9], Ketekny ICD-10 specification [10], ICD-10 guidelines in Greek[11], GPC/ETIP codes[12]).

2.2. Named Entity Recognition

NER is a critical task in processing biomedical and clinical texts [11]. Early NER systems predominantly used rule-based approaches, relying on hand-crafted rules and linguistic patterns, and strict or fuzzy dictionary lookups (gazetteers) to identify entities. While interpretable, these methods lacked generalizability and required significant manual effort [12]. Statistical machine learning models made an advancement, with Conditional Random Fields (CRFs) becoming a standard for sequence labeling due to their ability to model label sequences effectively. Support Vector Machines (SVMs) were also often applied [13]. Deep learning further transformed NER by automating feature learning. Thanks to sequential models like Recurrent Neural Networks (RNNs) and LSTMs[14], particularly Bidirectional LSTMs, have addressed sequential data challenges. The Bi-LSTM-CRF architecture became highly successful, reducing reliance on hand-engineered features [12]. The advent of transformer-based models, more specifically BERT (Bidirectional Encoder Representations from Transformers)[15], made a significant shift in the methodologies used. BERT's architecture, based on Transformer [16] encoders combined with the masked language modeling [15], allowed for the creation of a base pretrained model that, with small architectural changes like changing the top layer and a small amount of fine-tuning, outperformed most of the existing at the time state-of-the-art methods [15] [12].

[4]https://huggingface.co/AI-team-UoA/GreekDeBERTa-base
[5]https://github.com/nlpaueb/gr-nlp-toolkit
[6]https://live.european-language-grid.eu/catalogue/corpus/12599/download/
[7]http://www.iatrolexi.gr/iatrolexi/paradotea.html
[8]https://medicalcodes.instdrg.gr/search/icd/alphabetic
[9]https://github.com/drmchris21/icd10-in-Greek/blob/main/icd10
[10]https://old.instdrg.gr/
[11]https://www.oenet.gr/media/k2/attachments/iatrikes_prakseis.pdf
[12]https://medicalcodes.instdrg.gr/search/gpc/alphabetic

Multilingual models like mBERT [17] and XLM-RoBERTa (XLM-R) [3], pre-trained on extensive multilingual corpora, allowing cross-lingual knowledge transfer, offer an effective strategy for tackling low-resource languages (LRLs). XLM-R has shown strong performance, particularly for LRLs, on tasks including NER [3].

Domain adaptation is another key strategy, involving further pre-training of general models on the target language (e.g., GREEK-BERT [4]). This helps the model learn specific vocabulary and contextual patterns. For specific tasks like biomedical domain adaptations, domain adaptation is also a viable strategy to allow the model to learn the contextual patterns found in the desired domain, even if the model has already been trained extensively on the target language [18].

2.3. Entity Linking

Early methods for clinical EL, particularly for mapping text spans to standardized terminologies like ICD-10, typically relied on rule-based or gazetteer-driven systems. These approaches used exact or fuzzy string matching against curated code definitions and offered high precision but struggled with lexical variability and semantic ambiguity [19].

More recent approaches incorporate neural models into EL pipelines. Models like BioBERT [20], MedCAT [21], SapBERT [9], and BERT-XML [22] encode both mentions and ontology entries into a shared embedding space for similarity-based retrieval or multi-label classification. These models improve generalization and semantic robustness but typically require substantial annotated data and domain-specific adaptation - challenges that are particularly pronounced for low-resource languages.

Hybrid and cascading methods have shown strong performance in ICD coding by combining lexical filtering with transformer-based reranking or classification. For instance, Velichkov et al. [23] proposed a hybrid pipeline for Bulgarian that uses the ICD-10 hierarchy to enhance multi-label classification. Our approach adopts a similar strategy adapted to Greek, combining gazetteer-based filtering with neural linking using a task-adapted BGE-M3 model [6].

Greek clinical NLP remains significantly under-resourced. The IATROLEXI project [24] represents one of the earliest efforts to develop structured biomedical corpora in Greek, providing foundational resources for tasks such as information extraction and semantic annotation. More recently, Chatzimina et al. [25] demonstrated the effectiveness of transformer-based models (particularly BERT) in Greek clinical sentiment analysis, highlighting the applicability of deep language models in capturing affective dimensions of clinical discourse.

A recent survey by Papantoniou et al. [26] highlighted the limited progress in Greek biomedical NLP, with substantial gaps in areas such as EL and NER. Meanwhile, lightweight models like DistilGREEK-BERT [27] demonstrated strong performance on core tasks including NER, achieving results comparable to larger models while offering faster inference, making them promising candidates for domain-specific adaptation.

On the multilingual front, biomedical language models such as KBioXLM [28] and MMed-Llama [29] have demonstrated promising cross-lingual transfer capabilities, leveraging knowledge-aligned training and large-scale multilingual corpora. However, evaluation on Greek remains limited. These models typically rely on structured biomedical knowledge and aligned multilingual data to bridge language gaps - an issue we address through targeted domain pretraining on Greek biomedical texts.

Our work contributes to this emerging field by addressing the challenge of ICD-10 EL in Greek through a hybrid system that combines curated lexical resources with cross-lingual dense retrieval models adapted via domain-specific pretraining on Greek biomedical texts.

Recent research in the field of ICD-10 EL in clinical settings has explored the potential of using Large Language Models (LLMs) to address this task.

Simmons et al. [30, 31] evaluated the performance of several LLMs in extracting ICD-10-CM codes from discharge summaries and found that the results underperform the human coders; even with GPT-4, the highest reported agreement was only 12.4% and for Claude 3 - 12.7%. The main reason is that LLMs propose more specific codes, and ICD-10 codes for signs and symptom, that are usually not the

expected billable codes provided by human coders. Another reported issue was LLM hallucinations. For benchmark datasets like MIMIC-III ICD-10 coding the top achieved Micro-F1 is 0.589 with GPT-4 [32].

Apart from direct classification, another direction for using LLM is as an assistant that can suggest candidates or improved textual representations. Boukhers et al. [33] have investigated Llama using such an approach, and the results they obtained show increased recognition and accuracy in the shared task BioCreative VIII SympTEMIST.

Despite numerous studies in this area, many challenges still remain in automating ICD-10 code NER and EL. The studies highlight the potential of LLMs in medical informatics, while emphasizing the need for further improvements to achieve precision and recall closer to human performance in specialized tasks such as ICD-10 coding.

3. Data

3.1. ELCardioCC Dataset

The ELCardioCC dataset consists of 1,000 de-identified hospital discharge summaries written in Greek, annotated for three subtasks:

- **NER**: identifying mentions of five clinical entity types - chief complaint, diagnosis, prior medical history, drugs, and cardiac echo findings.
- **EL**: mapping each identified mention to its corresponding ICD-10 code.
- **MLC-X**: predicting all ICD-10 codes relevant to a document, along with the textual evidence supporting each prediction.

Each instance is provided in structured JSON format and includes the fields: `text` (the discharge letter), and a list of `annotations`, each containing a `mention`, its ICD-10 `code`, and character offsets (`start`, `end`) within the text.

We performed our own split of the dataset into 800 documents for training and 200 for validation (dev) (80% / 20%), as no official split was provided by the ELCardioCC task organizers. For NER, this corresponds directly to 800 and 200 annotated documents. For EL, where each mention is treated as a separate instance, this results in 8,096 training and 2,072 validation examples (79.62% / 20.38%).

All preprocessing steps were carried out by us. This focused on structural segmentation: each discharge summary was divided into sections based on visual layout (e.g., paragraph breaks), and section titles were normalized and mapped to semantic types such as `DIAGNOSIS`, `THERAPY_COURSE`, and `DISCHARGE_INSTRUCTIONS`. Mentions were aligned with their corresponding section, and offsets recalculated relative to the section text. Further, each section was split on every new line to ensure that the sample length fits BERT-based models. No additional tokenization or linguistic normalization was applied.

3.2. Additional Datasets

To complement the official ELCardioCC dataset, we collected several external resources relevant to the Greek biomedical domain. These include structured code systems, medical abbreviations, and open-domain clinical texts. All data were used solely for research purposes and to train domain-adapted models. Due to licensing restrictions and unclear redistribution terms, these resources are not publicly released.

3.2.1. Structured Medical Coding Systems

The two official portals[13,14] publish systematic catalogs of both the International Statistical Classification of Diseases and Related Health Problems (ICD-10-GrM) and the Greek Procedure Classification

[13] https://medicalcodes.instdrg.gr/home
[14] https://medicalcodesdrg.gesy.org.cy/

(GPC/ETIP). We used this information to prepare list of ICD-10 entities, which resulted 20,230 unique pairs of ICD-10 codes and labels.

3.2.2. Medical Abbreviations

We compiled 305 medical abbreviations (239 English, 66 Greek) from multiple online sources[15,16,17]. We used them to augment our dictionary with terms and their ICD-10 codes.

3.2.3. Open-Domain Clinical Texts

Using the MediaWiki API, we collected 514 Greek Wikipedia articles under the Ἰατρική¨ (Medicine) category. Articles were segmented by section, yielding 2,281 text instances in JSONL format. These texts were used for domain adaptation and representation learning.

3.3. Dictionaries

A dictionary that contains text phrases and associated with them ICD-10 codes was generated from the following sources - mentions and their ICD-10 codes from the ELCardioCC dataset keeping their number of occurences, ICD-10-GrM Alphabetic [18], ICD-10-GrM Systematic [19], list of medical abbreviations in Greek and English labeled with ICD-10 codes. The dictionary is split in two parts: unique pairs and a statistical dictionary.

The dictionary of Unique pairs comprises of all unambiguous labels from the dictionary for which a single ICD-10 code is assigned for all their occurrences in the dictionary. The result dictionary of unique pairs consists of 324 3-character ICD-10 codes and 11,552 labels in total. The distribution of the codes and labels per categories is presented in Fig. 1. The top 5 category letters are I - 22.97%, C-21.17%, R - 6.35%, E - 5.66% and Z - 5.57%. The minimum number of labels per ICD-10 code is 2, and the maximum - 294, the mean is 35.65432099. The minimum label length is 2 and the maximum label length is 384, and the mean label length is 44.01 (Fig. 2). For the experiments with the validation dataset, we excluded from the dictionary all mentions from our validation split of the ELCardioCC dataset. This dictionary is used for exact matches in our experiments.

The statistical dictionary was generated from the original dictionary in such a way that for labels that are ambiguous, i.e. more than one ICD-10 code association exists, we select the ICD-10 code with the highest frequency. The statistical dictionary contains also all pairs from the dictionary of unique pairs. The resulting dictionary contains 21,720 pairs of labels and associated ICD-10 codes. The overall percentage of labels by categories is comparable to the dictionary of unique labels. Similar to the case of the dictionary of unique pairs, the mentions from our validation split of the ELCardioCC dataset were excluded from the statistical dictionary for the experiments with the validation dataset. This dictionary is used for statistical dictionary matches in our experiments.

4. Methods

4.1. Pretraining

We experimented with models supporting different context length - BERT-based models which support a 512 token window, and BGE-M3 which supports longer context. We refer to the BERT-based models as having short context, and BGE-M3 as a long context model.

[15] https://www.bcardio.gr/el/4etos2017/53-students/syntomografies
[16] https://peptiko.gr/pos-grafetai-i-exetasi-syntomografies-exetaseon/
[17] https://www.vasiliadis-books.gr/Vasiliadis-books/wp-content/uploads/2015/10/ÎŤÎµÎŕÏĎÎµ-ÏĎÎś-ÎăÎµÏĄÎźÎµÏĞÏŒÎijÎµÎ¡Îś-30.pdf
[18] https://medicalcodes.instdrg.gr/search/icd/alphabetic
[19] https://medicalcodes.instdrg.gr/search/icd/systematic

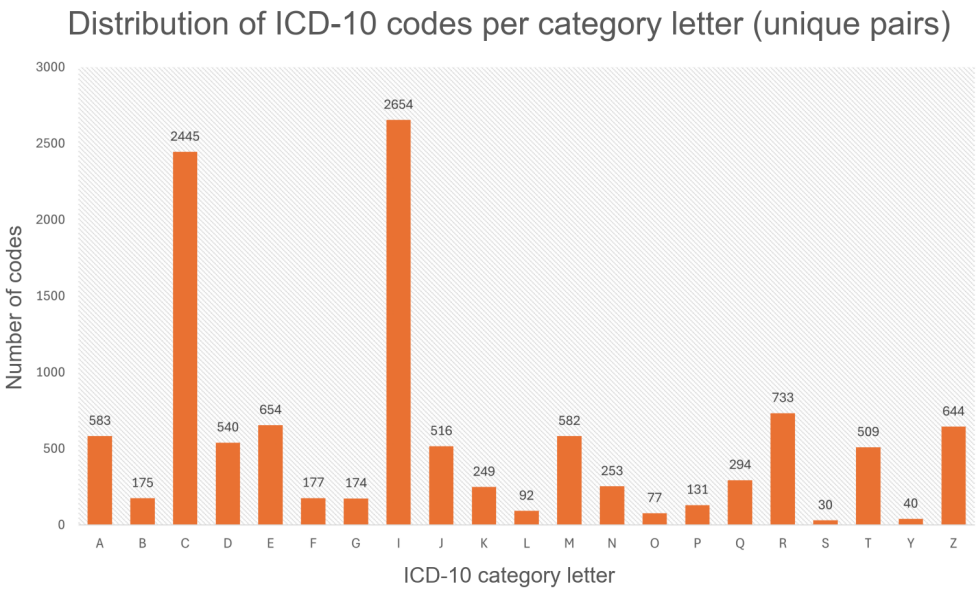

Figure 1: Distribution of ICD-10 codes per category letter for unique pairs

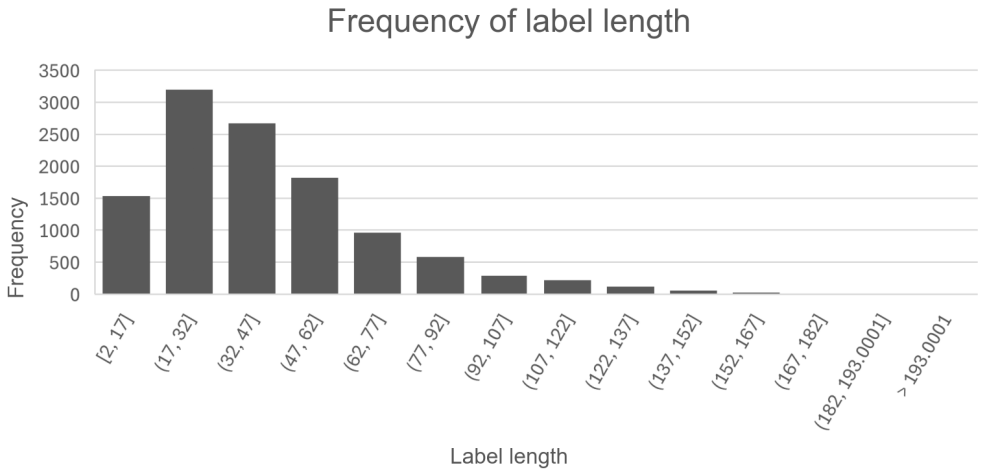

Figure 2: The frequency of the length of labels in the dictionary of unique pairs

The pretraining is split into 2 phases. **Short Context** pretraining uses the standard MLM objective and the data compiled from different sources. We use the standard hyperparameters when pretraining and train the models for 10 epochs, 2e-5 learning rate, 0.01 weight decay, batch size 64, and 15% probability of masking [15].

Long Context pretraining is pretraining using the full documents instead of splitting them into smaller chunks. The long context pretraining consists of two parts: standard MLM pretraining objective and task pretraining, consisting of pretraining on the NER objective using the full documents instead of splitting them into smaller chunks.

4.1.1. Domain Adaptation - Short Context

We compile a corpus of biomedical and clinical texts in Greek by combining the train dataset and texts crawled from public resources on the Internet (MediaWiki). The corpus consists of 3,281 documents and 987,139 tokens. Using the compiled corpus, we perform domain adaptation on two BERT-based models - Greek BERT [4] and XLM-RoBERTa [3]. We continuously pretrain these models on the masked

language modeling task for 10 epochs. We use the following hyperparameters: 2e-5 learning rate, 0.01 weight decay, batch size 64, and 15% probability of masking according to the standard configuration [15]. We used an L4 High-RAM GPU on Google Colab Pro for training the models. Due to time constraints, we have not conducted an investigation into hyperparameter optimization.

4.1.2. Domain Adaptation - Long Context

To adapt some of the models to the specific documents, we used additional pertaining. We performed Masked language modeling [15] pretraining on BGE-M3[6]. The parameters of the pretraining are listed in Table 1

Table 1
Hyperparameters for masked language modeling.

Hyperparameter	Value
Batch size	32
# Epoch	5
learning rate	5e-5
max grad norm	1
weight decay	0.01
optimizer	AdamW

We additionally performed task pretraining on the model by training the model on the NER task with the full context. We split the texts so that they can fit into the model's context length. But BGE-M3 has a context length of 8000 while the documents on average are less than 3000 tokens long. We discuss in the experiments section that classifying all of the entities at once seemed to be too difficult for the model. However, it was an effective task pretraining method; i.e., we first train the model on the full texts as a task pretraining step and then perform the final fine-tuning stage on smaller text chunks split on paragraphs and new lines. The parameters for the task pretraining are the same as the ones for MLM pretraining 1 with the exception that we found a benefit of training for a total of 20 epochs. The parameters 1 we chose are based on values that we have found that are a good starting point when fine-tuning a model.

4.2. Named Entity Recognition

We approach the task of detecting ICD-10 terms in the discharge summary as a NER task, and we finetuned different BERT-based models on token classification. The ICD-10 terms are labeled using a standard BIO tagging approach (beginning, inside, and outside of a term). We use the following models:

- Greek BERT Base [4] [20] - a Greek-specific model trained on Greek texts from Wikipedia, European Parliament Proceedings Parallel Corpus, and the Greek portion of filtered CommonCrawl. It has shown improved results on the general domain Greek NER task.
- XLM-RoBERTa Large [3] [21] - a multilingual model, trained on 2.5TB of filtered CommonCrawl data.

We use the Huggingface Transformers library to finetune the models on token classification for 5 epochs with the following hyperparameters - learning rate 2e-5, batch size 16. In order to fit in the 512-token limit of the models, we preprocess the text by splitting paragraphs and new lines. We perform initial experiments on a custom split of the train set - 80% used for training and 20% used for validation of the methods. The train split consists of 35,867 samples, the validation - 8,751, and the test set - 22,504. We train the final models used to generate predictions on the test set using the full training dataset provided by the organizers.

[20] https://huggingface.co/AI-team-UoA/GreekDeBERTa-base
[21] https://huggingface.co/FacebookAI/xlm-roberta-large

4.3. Entity Linking

We implemented dictionary based approach as a baseline using exact and fuzzy matching. For the purposes of this method, we collected and combined ICD-10 labels from different sources including: all labels from the annotated train set provided by the organizers of ELCardioCC CLEF challenge; ICD-10 Greek version, including all 3-character and 4-character codes.

Following the dictionary-based baseline, we developed a bi-encoder EL approach using a multilingual dense retrieval model. The task was framed as a mention-code semantic similarity problem, where the model learns to embed mentions and ICD-10 codes into a shared vector space and match them via cosine similarity.

We began with the publicly available multilingual dense encoder `BGE-M3` [6], and conducted exploratory pretraining using domain-specific Greek biomedical texts gathered from MediaWiki. However, simple fine-tuning on this corpus not only failed to improve performance but actually degraded it - likely due to overfitting on the limited data. To avoid repeating this process, we directly evaluated two task-adapted variants of the same model that had previously been fine-tuned for the NER subtask: `BGE-M3 + TP + FL(1) + DA + OP` and `BGE-M3 + TP + FL(1) + DA`. Both outperformed the base model on the EL task without additional pretraining, so we selected them for further fine-tuning on the mention-to-code retrieval objective.

For fine-tuning, we used `MultipleNegativesRankingLoss`, with correct mention-code pairs as positives. ICD-10 codes were represented as text strings. Models were trained for up to 50 epochs using a batch size of 32, with early stopping triggered after five epochs without macro-F1 improvement on the validation set. A default learning rate of 2e-5 was used, with linear warm-up over the first 100 steps to stabilize early training dynamics. Cosine similarity was used for inference, and top-1 and top-5 predictions were evaluated.

Finally, we experimented with a cross-encoder reranker (`bge-reranker-base`) [34], applied to rerank the top-5 candidates returned by the bi-encoder. However, it underperformed relative to the bi-encoder models and was not included in the final submission.

4.4. Multi Label Classification - eXplainable

Since BGE-M3 is capable of fitting the full documents in its context length, one of our approaches for the MLC-X subtask was a simple multi-label classification approach. The parameters for the multi-label classification fine-tuning are listed in Table 2.

Table 2
Hyperparameters for MLC-X with multi-label classification.

Hyperparameter	Value
Batch size	16
# Epoch	5
learning rate	2e-5
max grad norm	1
weight decay	0.01
optimizer	AdamW

5. Experiments and Results

5.1. Named Entity Recognition

We perform experiments with several BERT-based models on token classification, including models with short context and long context (BGE-M3). We also compare the performance of the models with and without domain adaptation pretraining on the biomedical corpus we compiled. We measure token-level

micro precision, recall, and F1 for different fine-tuned models. For our experiments on the validation set we use several different models in addition to the ones submitted in the challenge:

- GreekDeBERTaV3-base [22] - a model pretrained specifically for Greek, based on the DeBERTaV3 architecture.
- GreekDeBERTa-base [23] - a model based on DeBERTa architecture, pretrained for Greek.
- Greek-Reddit-BERT [24] [5] - a model pretrained on Greek topic classification dataset from Reddit.
- google/umt5-xl [25] [7] - a multilingual model pretrained on mC4 dataset.
- CohereLabs/aya-101 [26] [8] - a massively multilingual generative language model trained on 101 languages.

For the BGE-M3 model we experiment with different pretraining methods:

- Domain Adaptation (DA) - pretraining on Greek biomedical texts using masked language modeling objective before finetuning on the NER task.
- Task Pretraining (TP) - pretraining before the finetuning on the NER task.
- Focal Loss [35] (FL (x)) - using focal loss during NER finetuning (gamma equals to x).
- Optuna[27] hyperparameter search (OP) - using Optuna to select the best hyperparameterms for NER finetuning.

The results of our model predictions on the validation set are shown in Table 3.

Table 3
Model Performance Metrics on the validation set. The columns Dev Micro-P, Dev Micro-R, and Dev Micro-F1 correspond to token-level precision, recall, and F1-score, respectively.
DA - Domain Adaptation; TP - Task Pretraining; FL (x) - focal loss with gamma equal to x; OP - Optuna hyperparameter search; full text - the model is trained using the full text instead of chunked

Model	Dev Token Micro-P	Dev Token Micro-R	Dev Token Micro-F1
XLM-RoBERTa	84.49%	83.68%	84.08%
XLM-RoBERTa + DA	82.99%	84.21%	83.59%
bert-base-greek-uncased-v1	83.08%	82.42%	82.75%
bert-base-greek-uncased-v1 + DA	83.18%	82.99%	83.08%
GreekDeBERTaV3-base	83.19%	78.85%	80.96%
GreekDeBERTa-base	82.10%	79.33%	80.69%
Greek-Reddit-BERT	82.83%	82.66%	82.75%
BGE-M3 + TP + FL(1) + DA + OP	**87.33%**	**87.61%**	**87.47%**
BGE-M3 + TP + FL(1) + DA	86.60%	86.55%	86.57%
BGE-M3 + TP + FL(2) + DA	86.13%	85.81%	85.97%
BGE-M3 + TP + FL(2)	86.29%	85.28%	85.78%
BGE-M3 + DA	86.66%	82.68%	84.62%
BGE-M3 + full text	61.10%	58.22%	59.63%
BGE-M3	85.27%	83.11%	84.18%
umt5-xl	83.70%	83.46%	83.58%
aya-101	83.79%	83.25%	83.52%

For the encoder-decoder models umt5-xl and aya-101, we didn't have enough computational resources to perform full fine-tuning. Therefore, those models were fine-tuned using LoRA [36] adapters applied to the query, key, value, and output matrices of the attention mechanism [16] of rank 16.

[22] https://huggingface.co/AI-team-UoA/GreekDeBERTaV3-base
[23] https://huggingface.co/AI-team-UoA/GreekDeBERTa-base
[24] https://huggingface.co/IMISLab/Greek-Reddit-BERT
[25] https://huggingface.co/google/umt5-xl
[26] https://huggingface.co/CohereLabs/aya-101
[27] https://optuna.org/

Table 4
Results of evaluation of models for the NER task on the validation set using entity-level metrics.

Model	Entity Precision	Entity Recall	Entity F1
XLM-RoBERTa Large	0.8027	0.8422	0.8220
XLM-RoBERTa Large + DA	0.8126	0.8393	0.8257
Greek BERT base	0.7793	0.8012	0.7901
Greek BERT base + DA	**0.8440**	**0.8697**	**0.8567**
GreekDeBERTaV3-base	0.0900	0.0912	0.0906
GreekDeBERTa-base	0.0890	0.0917	0.0903
BGE-M3 + TP + FL(1) + DA + OP	0.0723	0.1371	0.0931
BGE-M3	0.0529	0.1170	0.0728

We performed entity-level evaluation on a subset of the models and found that even if token-level metrics are relatively high, some models show very low results on entity-level metrics. For example, the Greek DeBERTa models score about 0.80 F1 on token-level, but below 0.10 on entity level. When reviewing the predictions, we noticed that these models add extra punctuation to the predictions which renders the predicted entity completely wrong when using strict evaluation. Based on the initial experiments on the validation set using entity-level metric, we selected the models for final submission - Greek BERT and XLM-RoBERTa with domain adaptation.

The results of our model predictions on the test set for the models we submitted in the competition are shown in Table 5. Our models show slightly lower score than the Baseline provided by the organizers which is based on mBERT [28].

Table 5
Results of the submitted models for the NER task on the test set and baseline provided from the organizers.

Model	Precision	Recall	F1
Baseline (organizers)	0.6959	0.7460	0.7201
Greek BERT	0.7012	**0.7328**	**0.7167**
XLM-RoBERTa Large	**0.7079**	0.7222	0.7150

5.2. Entity Linking

We evaluated two language-adapted BGE-M3-based models for the EL subtask. These variants had previously been fine-tuned on the NER task and were selected for EL based on their strong performance during an initial exploratory phase:

- BGE-M3 + TP + FL(1) + DA
- BGE-M3 + TP + FL(1) + DA + OP

Both were subsequently fine-tuned for EL using mention–code training pairs and a contrastive learning objective. Evaluation was based on micro-averaged precision, recall, and F1 on the validation set.

While both models performed competitively, the first option (without Optuna search) was selected for submission due to its higher micro-averaged F1 (0.8871 vs. 0.8620) and superior ranking performance (MRR@5: 0.9157 vs. 0.9055). The Optuna-tuned model achieved marginally better results on several secondary metrics, including Recall@5 (0.9633 vs. 0.9527) and macro precision (0.4845 vs. 0.4804), but these gains did not outweigh the more consistent micro-level performance of the selected model.

The base BGE-M3 model without task-specific adaptation yielded substantially lower macro performance (F1 $\approx$ 0.44), emphasizing the importance of task-adaptive training for EL in this setting.

[28] https://huggingface.co/google-bert/bert-base-multilingual-cased

In addition to the BGE-M3 variants, we evaluated other approaches, including a dictionary-based method, a statistical ranking method, and hybrid models combining these with neural encoders (BGE-M3 and SapBERT). As shown in Table 6, the dictionary-based approach achieved the highest precision (0.9863), but its limited recall, due to missing exact matches for some codes, makes it unsuitable as a standalone solution. A similar limitation applies to the statistical approach. To address this, both methods were combined with neural models to improve coverage and robustness. The best overall performance was achieved by hybrid configurations, particularly those using the statistical method in combination with BGE-M3 or SapBERT, which motivated their selection for test set evaluation.

Table 6
EL performance on the validation set (micro-averaged precision, recall, and F1).

Model	Precision	Recall	F1
Dictionary	**0.9863**	0.6588	0.7899
Statistical	0.9595	**0.9595**	**0.9595**
BGE-M3 + TP + FL(1) + DA	0.8871	0.8871	0.8871
BGE-M3 + TP + FL(1) + DA + OP	0.8620	0.8620	0.8620
SapBERT	0.8446	0.8446	0.8446
Dictionary + BGE-M3 + TP + FL(1) + DA	0.9001	0.9001	0.9001
Statistical + BGE-M3 + TP + FL(1) + DA	0.9595	**0.9595**	**0.9595**
Statistical + SapBERT	0.9595	**0.9595**	**0.9595**

The results of our model predictions on the test set for the models we submitted in the competition are shown in Table 7. All of our submitted models outperformed the official baseline in terms of precision, with the best overall F1 (0.6693) achieved by combining Greek BERT with either the Dictionary + BGE-M3 or the Statistical + SapBERT approach.

Table 7
Results of the submitted models for the EL task on the test set and baseline provided from the organizers.

NER Model	EL Model	Precision	Recall	F1
Baseline (organizers)	Baseline (organizers)	0.6476	0.6942	0.6701
Greek BERT	Dictionary + BGE-M3	0.6548	**0.6844**	**0.6693**
Greek BERT	Statistical + SapBERT	0.6548	**0.6844**	**0.6693**
Greek BERT	Statistical + BGE-M3	0.6540	0.6835	0.6684
XLM-RoBERTa	Statistical + BGE-M3	**0.6594**	0.67208	0.6660
XLM-RoBERTa	Dictionary + BGE-M3	0.6585	0.6719	0.6651

5.3. Multi Label Classification - eXplainable

The multi-label classification task proved to be very difficult for BGE-M3 as both the number of labels is quite large and they are very imbalanced. The model barely achieved an F1 score of 13% on the validation set, which, combined with the time limitations, discouraged us from trying to improve the multi-label classification pipeline. Despite the fact that the multi-label classification task is more straightforward, the combination of NER followed by EL seems to be a better pipeline approach.

6. Discussion

6.1. Named Entity Recognition

Based on our experiments on the validation set, we observed that all of the models showed a high F1 score on token-level evaluation (higher than 0.80) and the best model was BGE-M3 with additional task and domain pretraining, using focal loss and hyperparameter search with Optuna. However, when

we evaluated a subset of the models on entity-level, there was a drastic difference in the performance, and the best models were BERT-based - XLM-RoBERTa and Greek BERT with domain adaptation. We spent a significant time making experiments based on the token-level metric, only to realize that the entity-level metric did not perform as well later on. This highlights the importance of using entity-level metrics from the beginning to have a more realistic evaluation of the models. The errors in prediction were mainly due to added punctuation, which did not impact the token-level significantly but reduced the score on the strict entity-level metric. We also saw that task pretraining, domain adaptation, and focal loss all bring significant improvements to the model's performance. Training the model on the full texts gave significantly lower results compared to splitting the texts, demonstrating that probably predicting all of the entities at once is a more difficult task, despite the fact that the model can use the full context. However, using the full document for model pretraining showed improved results. The results on bigger models like aya-101 and umt5-xl were not any better than on small models, either suggesting that LoRA adapters are not as effective for encoder models or that bigger encoder models need more data for pretraining to take advantage of the higher number of parameters.

6.2. Entity Linking

The error analysis of EL shows the following categories of errors:

- Misinterpretation of the mention of the Diabetes Mellitus (expected ICD-10 code E13) without specification with Diabetes Mellitus type 2 (ICD-10 code E11). Of course this is prevalent for the statistical dictionary matches.
- Low recall - not associated ICD-10 codes for entities - this is typical for the the dictionary of unique pairs used for exact matches.
- capital letters only mentions - most of the methods present poor performance on such mentions. One of the reasons is that they use functions for lower case transform of the text, that uses conversion of capital letters to lower letter by letter. In Greek this can cause some issues for example with letter Σ, that has two forms: uppercase Σ and lowercase σ (or ς in word-final position). The lower case transformation does not take into account the positions of the letters.
- Abbreviations - SapBERT can not resolve correctly most of the abbreviations. The dictionary based approaches can cope with this issue, due to enrichment with abbreviations-rich sources. This leads to many wrong ICD-10 code predictions for cardiac echo mentions. Another challenge with abbreviations is that in the discharge summaries both abbreviations in English and in Greek are used.
- Lack of capability to differentiate between specific cases with cases not classified elsewhere: for example, instead of I33 (Acute and subacute endocarditis) is predicted I39 (Endocarditis and heart valve disorders in diseases classified elsewhere).
- Predicting codes for signs and symptoms instead of assigning codes for disorders: for example, the expected ICD-10 code is J81 (Pulmonary oedema) and the predicted one is R06 (Abnormalities of breathing).
- Imprecise selection of ICD-10 for closely related conditions: for example, the expected ICD-10 code is R55 (Syncope and collapse) and the predicted code is R41 (Other symptoms and signs involving cognitive functions and awareness).

The results of most of the models are comparable. The combination of dictionary-based approaches and deep learning approaches manages to overcome some of the issues, but still some challenges remain as the ICD-10 system is very complex and the hybrid approaches cannot cover all scenarios.

7. Conclusion

In this paper, we presented approaches for NER and EL to ICD-10 of discharge summaries in Greek as part of ElcarioCC @ CLEF 2025 BioASQ challenge. We examined different solutions for NER mainly BERT-family approaches like Greek BERT and multilingual XLM-RoBERTa. Both of them were additionally

pretrained and adapted for the NER task. The best achieved result was 0.7167 F1 score on the test set. For the EL to ICD-10 codes task, we used a hybrid approach combining different dictionaries with a fine-tuned bi-encoder model (BAAI/bge-m3), achieving F1 score of 0.6693. This demonstrates that combinations between these two approaches can improve the performance of the EL. All of the presented approaches show a huge potential for solving NER and EL to ICD-10 code tasks for Greek discharge summaries.

As further work, we can experiment with LLMs to investigate their capabilities to provide solutions for domain specific tasks for language other than English. Another direction for improvement is to enrich the dictionaries and to try combinations with other transformers.

Acknowledgments

This work was partially supported by the European Union-NextGenerationEU, through the National Recovery and Resilience Plan of the Republic of Bulgaria [Grant Project No. BG-RRP-2.004-0008]. Part of this works is also supported by European Union's Horizon research and innovation programme projects RES-Q PLUS [Grant Agreement No. 101057603] and HEREDITARY [Grant Agreement No. 101137074]. Views and opinions expressed are however those of the author only and do not necessarily reflect those of the European Union. Neither the European Union nor the granting authority can be held responsible for them.

Declaration on Generative AI

During the preparation of this work, the author(s) used Grammarly in order to: Grammar and spelling check. After using these tool(s)/service(s), the author(s) reviewed and edited the content as needed and take(s) full responsibility for the publication's content.

References

[1] 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, G. M. Di Nunzio, N. Ferro, 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: J. Carrillo-de Albornoz, J. Gonzalo, L. Plaza, A. García Seco de Herrera, J. Mothe, F. Piroi, P. Rosso, D. Spina, G. Faggioli, N. Ferro (Eds.), Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Sixteenth International Conference of the CLEF Association (CLEF 2025), 2025.

[2] 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.

[3] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, V. Stoyanov, Unsupervised cross-lingual representation learning at scale, 2020. URL: https://arxiv.org/abs/1911.02116. arXiv:1911.02116.

[4] J. Koutsikakis, I. Chalkidis, P. Malakasiotis, I. Androutsopoulos, Greek-bert: The greeks visiting sesame street, in: 11th Hellenic Conference on Artificial Intelligence, SETN 2020, Association for Computing Machinery, New York, NY, USA, 2020, p. 110–117. URL: https://doi.org/10.1145/3411408.3411440.

[5] C. Mastrokostas, N. Giarelis, N. Karacapilidis, Social media topic classification on greek reddit, Information 15 (2024) 521.

[6] J. Chen, S. Xiao, P. Zhang, K. Luo, D. Lian, Z. Liu, Bge m3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation, 2024. arXiv:2402.03216.

[7] H. W. Chung, N. Constant, X. Garcia, A. Roberts, Y. Tay, S. Narang, O. Firat, Unimax: Fairer and more effective language sampling for large-scale multilingual pretraining, 2023. URL: https://arxiv.org/abs/2304.09151. arXiv:2304.09151.

[8] A. Üstün, V. Aryabumi, Z.-X. Yong, W.-Y. Ko, D. D'souza, G. Onilude, N. Bhandari, S. Singh, H.-L. Ooi, A. Kayid, F. Vargus, P. Blunsom, S. Longpre, N. Muennighoff, M. Fadaee, J. Kreutzer, S. Hooker, Aya model: An instruction finetuned open-access multilingual language model, arXiv preprint arXiv:2402.07827 (2024).

[9] F. Liu, E. Shareghi, Z. Meng, M. Basaldella, N. Collier, Self-alignment pretraining for biomedical entity representations, in: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2021, pp. 4228–4238. doi:10.18653/v1/2021.naacl-main.334.

[10] J. Pavlopoulos, J. Bakagianni, K. Pouli, M. Gavriilidou, Open or closed llm for lesser-resourced languages? lessons from greek, 2025. URL: https://arxiv.org/abs/2501.12826. arXiv:2501.12826.

[11] Warto, S. Rustad, G. Shidik, E. Noersasongko, P. Purwanto, M. Muljono, D. R. I. M. Setiadi, Systematic literature review on named entity recognition: Approach, method, and application, Statistics, Optimization & Information Computing 12 (2024) 907–942. doi:10.19139/soic-2310-5070-1631.

[12] I. Keraghel, S. Morbieu, M. Nadif, Recent advances in named entity recognition: A comprehensive survey and comparative study, 2024. URL: https://arxiv.org/abs/2401.10825. arXiv:2401.10825.

[13] N. Perera, M. Dehmer, F. Emmert-Streib, Named entity recognition and relation detection for biomedical information extraction, Frontiers in Cell and Developmental Biology 8 (2020). doi:10.3389/fcell.2020.00673.

[14] S. Hochreiter, J. Schmidhuber, Long short-term memory, Neural Computation 9 (1997) 1735–1780. doi:10.1162/neco.1997.9.8.1735.

[15] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, 2019. URL: https://arxiv.org/abs/1810.04805. arXiv:1810.04805.

[16] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, I. Polosukhin, Attention is all you need, 2023. URL: https://arxiv.org/abs/1706.03762. arXiv:1706.03762.

[17] J. Devlin, M. Chang, K. Lee, K. Toutanova, BERT: pre-training of deep bidirectional transformers for language understanding, CoRR abs/1810.04805 (2018). URL: http://arxiv.org/abs/1810.04805. arXiv:1810.04805.

[18] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, J. Kang, Biobert: a pre-trained biomedical language representation model for biomedical text mining, Bioinformatics 36 (2019) 1234–1240. URL: http://dx.doi.org/10.1093/bioinformatics/btz682. doi:10.1093/bioinformatics/btz682.

[19] C. Yan, X. Fu, X. Liu, Y. Zhang, Y. Gao, J. Wu, Q. Li, A survey of automated international classification of diseases coding: development, challenges, and applications, Intelligent Medicine 2 (2022) 161–173. URL: https://www.sciencedirect.com/science/article/pii/S2667102622000092. doi:https://doi.org/10.1016/j.imed.2022.03.003.

[20] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, J. Kang, BioBERT: a pre-trained biomedical language representation model for biomedical text mining, Bioinformatics 36 (2020) 1234–1240. doi:10.1093/bioinformatics/btz682.

[21] Z. Kraljevic, D. Bean, A. Mascio, L. Roguski, A. Folarin, A. Roberts, R. Bendayan, R. Dobson, Medcat – medical concept annotation tool, 2019. URL: https://arxiv.org/abs/1912.10166. arXiv:1912.10166.

[22] Z. Zhang, J. Liu, N. Razavian, Bert-xml: Large scale automated icd coding using bert pretraining, 2020. URL: https://arxiv.org/abs/2006.03685. arXiv:2006.03685.

[23] B. Velichkov, S. Gerginov, P. Panayotov, S. Vassileva, G. Velchev, I. Koychev, S. Boytcheva, Cascading approach for automatic icd-10 codes association to diseases in bulgarian, in: S. S. Sotirov, T. Pencheva, J. Kacprzyk, K. T. Atanassov, E. Sotirova, G. Staneva (Eds.), Contemporary Methods in

Bioinformatics and Biomedicine and Their Applications, Springer International Publishing, Cham, 2022, pp. 247–260. doi:10.1007/978-3-030-96638-6_27.

[24] C. Tsalidis, G. Orphanos, E. Mantzari, M. Pantazara, C. Diolis, A. Vagelatos, Developing a greek biomedical corpus towards text mining, Corpus Linguistics Conference 2007, University of Birmingham, 2007. Article #137. Available at https://www.birmingham.ac.uk/research/centres-institutes/centre-for-corpus-research/corpus-linguistics-conference-2007.

[25] M. E. Chatzimina, H. A. Papadaki, C. Pontikoglou, M. Tsiknakis, A comparative sentiment analysis of greek clinical conversations using bert, roberta, gpt-2, and xlnet, Bioengineering 11 (2024) 521.

[26] K. Papantoniou, Y. Tzitzikas, Nlp for the greek language: A longer survey, 2024. URL: https://arxiv.org/abs/2408.10962. arXiv:2408.10962.

[27] E. A. Karavangeli, D.-A. Pantazi, M. Iliakis, Distilgreek-bert: A distilled version of the greek-bert model, 2023.

[28] L. Geng, X. Yan, Z. Cao, J. Li, W. Li, S. Li, X. Zhou, Y. Yang, J. Zhang, Kbioxlm: A knowledge-anchored biomedical multilingual pretrained language model, arXiv preprint arXiv:2311.11564 (2023).

[29] P. Qiu, C. Wu, X. Zhang, W. Lin, H. Wang, Y. Zhang, Y. Wang, W. Xie, Towards building multilingual language model for medicine, Nature Communications 15 (2024) 8384.

[30] A. Simmons, K. Takkavatakarn, M. McDougal, B. Dilcher, J. Pincavitch, L. Meadows, J. Kauffman, E. Klang, R. Wig, G. Smith, et al., Extracting international classification of diseases codes from clinical documentation using large language models, Applied Clinical Informatics 16 (2025) 337–344.

[31] A. Simmons, K. Takkavatakarn, M. McDougal, B. Dilcher, J. Pincavitch, L. Meadows, J. Kauffman, E. Klang, R. Wig, G. Smith, et al., Benchmarking large language models for extraction of international classification of diseases codes from clinical documentation, medRxiv (2024) 2024–04.

[32] R. Li, X. Wang, H. Yu, Exploring llm multi-agents for icd coding, arXiv preprint arXiv:2406.15363 (2024).

[33] Z. Boukhers, A. Khan, Q. Ramadan, C. Yang, Large language model in medical informatics: Direct classification and enhanced text representations for automatic icd coding, in: 2024 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), IEEE, 2024, pp. 3066–3069.

[34] S. Xiao, Z. Liu, P. Zhang, N. Muennighoff, C-pack: Packaged resources to advance general chinese embedding, 2023. arXiv:2309.07597.

[35] T.-Y. Lin, P. Goyal, R. Girshick, K. He, P. Dollár, Focal loss for dense object detection, 2018. URL: https://arxiv.org/abs/1708.02002. arXiv:1708.02002.

[36] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, W. Chen, Lora: Low-rank adaptation of large language models, 2021. URL: https://arxiv.org/abs/2106.09685. arXiv:2106.09685.