Explainable ICD Coding via Entity Linking

Leonor Barreiros[ℑ, Ω, Ψ] **Isabel Coutinho**[Ω, Ψ] **Gonçalo M. Correia**[ℑ] **Bruno Martins**[Ω, Ψ]
[ℑ] Priberam Labs, Alameda D. Afonso Henriques, 41, 2º, 1000-123 Lisboa, Portugal
[Ω] Instituto Superior Técnico, Lisboa, Portugal
[Ψ] INESC-ID, Rua Alves Redol, 9, 1000-029, Lisboa, Portugal
{leonor.barreiros, goncalo.correia}@priberam.pt
{isabel.coutinho, bruno.g.martins}@tecnico.ulisboa.pt

Abstract

Clinical coding is a critical task in healthcare, although traditional methods for automating clinical coding may not provide sufficient explicit evidence for coders in production environments. This evidence is crucial, as medical coders have to make sure there exists at least one explicit passage in the input health record that justifies the attribution of a code. We therefore propose to reframe the task as an entity linking problem, in which each document is annotated with its set of codes and respective textual evidence, enabling better human-machine collaboration. By leveraging parameter-efficient fine-tuning of Large Language Models (LLMs), together with constrained decoding, we introduce three approaches to solve this problem that prove effective at disambiguating clinical mentions and that perform well in few-shot scenarios.

1 Introduction

Medical reports are essential documents that detail patient medical history, procedures, exams, symptoms, and diagnoses. Clinical coding involves assigning standardized codes, such as those from ICD-10, to these records. This process is crucial for hospitals, since it helps justify expenses, secure funding, or file insurance claims to cover healthcare costs. Furthermore, labeling Electronic Health Records (EHRs) through clinical coding makes their data more searchable and suitable for statistical analysis, *e.g.* potentially revealing cause-effect relationships between diseases and symptoms.

Automated solutions can help medical coders by accelerating their work and reducing errors. However, traditional automated systems that treat coding as a Multi-Label Classification (MLC) problem are often non-explainable (Teng et al., 2023; Dong et al., 2022), making it difficult for healthcare professionals to trust or verify their outputs. If systems are explainable, we can critically reason about their decisions, allowing medical practitioners to better work alongside AI tools (Arrieta et al., 2020; Goldberg et al., 2024).

To address these challenges, we propose framing clinical coding as an entity linking problem. This particular task involves annotating documents with specific entities and providing textual evidence for each one. This could enable clinical coders to understand where each code is mentioned in a record, allowing easier cooperation with AI systems. However, clinical entity linking remains largely underexplored and lacking in terms of annotated data.

Recently, we have seen several advances in Transformer-based (Vaswani et al., 2017) Large Language Models (LLMs), such as LLaMA (Touvron et al., 2023), Mistral (Jiang et al., 2023), or Gemini (Anil et al., 2023), and in the formulation of data- and compute-efficient ways to fine-tune them (Hu et al., 2021; Dettmers et al., 2023). Consequently, we focus on mitigating the above challenges by exploring clinical entity linking as a generative task through a biomedical LLM, namely BioMistral (Labrak et al., 2024). By fine-tuning an LLM, we aim to develop a system capable of solving clinical entity linking tasks effectively.

Our contributions are three-fold: (i) we propose to frame the **explainability** of ICD coding as an entity linking task; (ii) we investigate the performance gains of prompting *versus* fine-tuning a clinical LLM for this task, evaluating how different formulations for **generative entity linking** can enhance model performance; and (iii) we compare the entity linking approach to MLC, assessing the potential it has for **few-shot classification** of mentions.

2 Proposed Approaches

Traditionally, clinical coding is treated as MLC, in which a model annotates the input medical report with its set of labels. In our setting, we treat clinical coding as an entity linking problem. This

Proceedings of the Second Workshop on Patient-Oriented Language Processing (CL4Health), pages 219–227

means that given a medical report and its set of gold mentions (*i.e.*, our work assumes mentions have been pre-detected, for instance, via named entity recognition), our model must disambiguate each mention by assigning it the corresponding entity.

The following subsections detail different approaches for tackling clinical entity linking.

2.1 ICL-BioMistral

ICL-BioMistral (in-context learning) prompts a pre-trained Transformer decoder model. The prompt comprises a (pre-determined) mention, and a medical report excerpt, corresponding to the context that surrounds it. The model must output an ICD-10 code representation, corresponding to the entity which the mention refers to.

Inspired by Boyle et al. (2023), we designed a prompt with a short context and the task description. To improve the model's capability to solve the task, we use in-context learning (thoroughly analyzed by Dong et al. (2024)). As such, we add 10 random examples to the prompt. We illustrate the prompt template in Appendix A.

Similarly to GENRE (De Cao et al., 2021), we use constrained greedy decoding,[1] to ensure that the model output is always a valid ICD-10 code representation. This is implemented with a prefix tree of all possible outputs, and by forcing the generated tokens to stay within the set of possible continuations for titles of ICD codes.

2.2 SFT-BioMistral

SFT-BioMistral (supervised fine-tuning) is similiar to ICL-BioMistral, as it also outputs an in-context mention, given a report excerpt. However, instead of learning through examples, this model was fine-tuned on a causal language modeling objective, where we maximize the conditional probability for each output token, considering the input and the expected previously generated output tokens (Williams and Zipser, 1989). We consider as *input* the prompt (*i.e.*, the task description and context), and compute the cross-entropy loss over the tokens of the *output* (the title of the desired ICD-10 code). Decoding with this model again relies on a constrained decoding algorithm.

2.3 InsGenEL-BioMistral

Our last proposed model is inspired by InsGenEL (Xiao et al., 2023), which is based on GENRE (De Cao et al., 2021). Our model outputs multiple mention-entity pairs for a medical report in a single pass. This is closer to the approach clinical coders take when annotating, and it enriches predictions through the document's global context, improving coherence between predictions.

Like GENRE, our model receives a document (with gold mentions) and outputs the document with annotated mention-entity pairs. Unlike GENRE, and following InsGenEL, we use a Transformer decoder to annotate the documents. The fine-tuning process optimizes a causal language modeling objective by learning from supervised instruction-response pairs (Ren et al., 2024). A prompt template is presented in Appendix A.

During inference, we ensure a valid generation using constrained decoding. We implemented a function (based on GENRE's proposal) that receives the generated tokens and returns the possible continuations. First, it determines the state as either outside an entity—which can be the case when processing a non-mention or mention token—or inside an entity—where the model is disambiguating a mention. If outside an entity, then the possible continuation is to resume copying the input document. Otherwise, the model generates an entity representation. Similarly to our previous approaches, we use a prefix tree to ensure the model generates valid ICD-10 code representations.

3 Experimental Setup

To train and evaluate our models, we used publicly available datasets for explainable ICD coding, *i.e.* including span evidences for each code, namely CodiEsp (Miranda-Escalada et al., 2020), DisTEMIST (Miranda-Escalada et al., 2022), and MDACE (Cheng et al., 2023). Further details on these datasets are given in Appendix B. Additional experimental details are given in Appendix C.

Knowledge Base. In entity linking, entities are organized in knowledge bases. We focus on the International Classification of Diseases (ICD)[2] coding system, proposed by the World Health Organization, as a standardized way of representing diagnoses and procedures. The ICD is a hierarchical ontology, as codes are first organized into chapters, sub-chapters, and partial codes. We considered version 10, which is divided into ICD-10-CM (for diagnoses) and ICD-10-PCS (for procedures).

[1] https://huggingface.co/blog/constrained-beam-search

[2] https://www.who.int/standards/classifications/classification-of-diseases

		Micro	Macro
CodiEsp	ICL-BM	6.36	5.93
	SFT-BM	63.39	62.41
	INSGENEL-BM	**66.85**	**64.40**
MDACE	ICL-BM	10.36	7.79
	SFT-BM	**64.88**	**60.94**
	INSGENEL-BM	57.10	55.45

Table 1: Accuracy in the CodiEsp and MDACE test sets for the entity linking task. BM denotes BIOMISTRAL. We highlight in bold the best-in-class performance.

		Chap	Sub	Part
CodiEsp	ICL-BM	30.64	18.33	12.55
	SFT-BM	85.65	82.27	75.94
	INSGENEL-BM	87.79	83.60	78.81
MDACE	ICL-BM	43.88	33.90	23.35
	SFT-BM	89.17	84.84	78.91
	INSGENEL-BM	90.09	83.73	76.18

Table 2: Micro-accuracy in the CodiEsp and MDACE test sets for the entity linking task, considering only the chapter (Chap), subchapter (Sub), and partial (Part) code of each ICD-10. BM denotes BIOMISTRAL.

		1-shot	5-shot
CodiEsp	SFT-BM	**47.49**	**56.66**
	INSGENEL-BM	34.97	49.30
MDACE	SFT-BM	**36.74**	**40.89**
	INSGENEL-BM	24.39	29.66

Table 3: 1- and 5-shot micro-accuracy in the CodiEsp and MDACE test corpora. BM denotes BIOMISTRAL.

Evaluation Details. In end-to-end entity linking, we distinguish the precision, recall, and F1 metris. In our case, where we used gold mentions, these equate to a measure of accuracy, as explained by Balog (2018). We consider micro-accuracy (where we average the accuracy of all mentions) and macro-accuracy (where we compute the accuracy per document and average all values). To compare our results with existing work, we computed coding evaluation metrics. By aggregating all assignments for the entity linking task, we obtain a solution for MLC that can be evaluated with precision, recall, and F1. These were computed with the script by Miranda-Escalada et al. (2020).

4 Experimental Results

Table 1 presents our micro- and macro-accuracy on the CodiEsp and MDACE test datasets.

Practical Highlights. From Table 1, we conclude that fine-tuned models perform considerably better than ICL-BIOMISTRAL. We highlight that SFT-BIOMISTRAL has a stable performance for both evaluation corpora, whereas INSGENEL-BIOMISTRAL has limitations in MDACE, which we hypothesize might be related to the increased length of the documents. Additionally, we find that INSGENEL-BIOMISTRAL is beneficial in production scenarios: not only does it better alleviate the coder's job with its increased accuracy, but it also deals with all of a document's mentions simultaneously. Nonetheless, clinical coders receive non-annotated documents and a separate procedure must be used to recognize and annotate the textual evidence to which a code should be assigned.

Partial Results. Since the ICD-10 is organized hierarchically, a wrong prediction can be partially correct if it determines the code's ancestors up to a certain point. We assessed micro-accuracy on the chapter, subchapter, and partial code levels (a partial code contains the first three digits of an ICD), and the results are in Table 2. Both SFT-BIOMISTRAL and INSGENEL-BIOMISTRAL can provide orientation helpful in practical scenarios.

Few-shot Analysis. In Table 3, we compare the few-shot performance for all codes seen at most once or 5-times during training (1-shot and 5-shot). The number of such codes in the inference corpora is given in Appendix B. The model with the best few-shot performance was SFT-BIOMISTRAL, but INSGENEL-BIOMISTRAL is nevertheless able to predict codes trained in few-shot scenarios. We hypothesize that the reduced performance on MDACE is related to the increased document length, which may lead to hard long-range dependencies.

4.1 Comparison with Existing Results

The CodiEsp-D and CodiEsp-P tasks can be evaluated with MLC metrics, as we explain in §3. CodiEsp also proposes an end-to-end entity linking task, CodiEsp-X. It is not evaluated with entity linking metrics, since if a code is mentioned more than once in the same document, it only needs to be correctly predicted once to be considered correct. This means the evaluation micro-metrics for

	Multi-label Classification						Entity Linking		
	CodiEsp-D			CodiEsp-P			CodiEsp-X		
	P	R	F1	P	R	F1	P	R	F1
IAM CodiEsp	**81.70**	59.20	68.70	**69.10**	42.00	52.20	**75.00**	52.40	61.10
DAC-E	–	–	74.40	–	–	**56.0**	–	–	–
ICL-BM	8.91	7.19	7.96	11.34	12.45	11.87	8.42	7.19	7.76
SFT-BM	75.04	**76.20**	**75.62**	34.31	38.53	36.30	64.66	**67.10**	65.86
INSGENEL-BM	73.93	71.94	72.92	46.26	**46.78**	46.52	68.34	66.96	**67.64**

Table 4: Comparison of automated medical coding and entity linking micro performance metrics on the CodiEsp test set with existing results for the CodiEsp shared task. BM denotes BIOMISTRAL.

CodiEsp-X do not equate to our micro-accuracy.

In Table 4, we compare our results to those of the challenge's winner, *i.e.*, the IAM team (Cossin and Jouhet, 2020), and to a solution that was subsequently proposed, DAC-E (Barros et al., 2022). These systems are described in Appendix D. Although a strict comparison is not possible, since we used gold mentions contrarily to the shared tasks, our fine-tuned models had similar or better performance in most settings, indicating that our approaches remain useful in the MLC scenario.

MDACE was proposed for a different task: given the output of MLC, finding sufficient textual evidence for each code. This means that we cannot compare with the paper's benchmarking results.

5 Related Work

We briefly describe previous related work on automated ICD coding and also on entity linking.

ICD Coding & Explainability. Most solutions for automated ICD coding are based on MLC. For example, Barros et al. (2022) leverage the ICD hierarchy and propose two MLC sub-tasks on different granularities. Furthermore, many studies have addressed the importance of solving explainable ICD coding, so that clinical coders can understand the system's decisions. However, most studies focus on label-wise attention mechanisms (Glen et al., 2024; Amjad et al., 2023; Figueira et al., 2023), which are challenging to systematically evaluate, as pointed out by Teng et al. (2023) and Dong et al. (2022). More recently, researchers have developed methodologies to evaluate these interpretability solutions (Edin et al., 2024; Wu et al., 2024).

Entity Linking & Different Entity Linking Approaches. Entity linking solutions range from discriminative to generative models. Discriminative models are the most common, but many state-of-the-art models, such as those of Yamada et al. (2022), Ayoola et al. (2022), and Shavarani and Sarkar (2023), were trained on large corpora (the Wikipedia), which is not available for our domain. Generative models require less fine-tuning data to achieve similar performance. For example, Xiao et al. (2023) performed better than Ayoola et al. (2022), using 50 times less data. The model was inspired by a previous proposal from De Cao et al. (2021), which uses constrained decoding to ensure valid generation.

Clinical & Biomedical Entity Linking. The clinical and biomedical domains are specialized, and general-purpose models cannot solve clinical problems, even with a target domain fine-tuning corpus (Alekseev et al., 2022). Existing work uses methodologies similar to general-domain algorithms, but with models trained on domain corpora (Yuan et al., 2022a; Agarwal et al., 2022). For instance, Yuan et al. (2022b) propose a method similar to GENRE. In the clinical domain, most entity linking studies focus on the DisTEMIST (Miranda-Escalada et al., 2022) and CodiEsp (Miranda-Escalada et al., 2020) challenges. For example, Gallego et al. (2024) propose a Transformer encoder-based solution to DisTEMIST.

6 Conclusions

We described three approaches for the clinical entity linking problem, based on BioMistral 7B, that annotate medical reports with each mention's ICD-10 code. The models we fine-tuned, *i.e.*, SFT- and INSGENEL-BIOMISTRAL, were substantially better than the prompted ICL-BIOMISTRAL, and yielded interesting results for few-shot codes.

Limitations

Our models only deal with the disambiguation subproblem of entity linking, using pre-detected mentions. Future work should explore mention detection to obtain an end-to-end solution, which makes our models useful in production environments.

In addition, our experiments were limited to three publicly available datasets, which only represent a small subset of patients, possible medical conditions, and medical procedures. There is not a lot of clinical data publicly available to support research studies, especially annotated for entity linking. In the future, we can explore other approaches to data collection, and even leverage additional information from clinical knowledge bases, such additional information in ICD-10 itself and UMLS.

Finally, large generative models such as BioMistral 7B are generally very costly to use. For instance, the IAM system (Cossin and Jouhet, 2020), based on a dictionary, only takes 5 seconds to run on an 8 CPUs' machine. The DAC-E (Barros et al., 2022) system, while using GPU processing, is also more efficient as it uses a smaller Transformer encoder as the backbone. Future work can perhaps assess the impact of using LLMs of different sizes.

Ethical Considerations

ICD coding is a sensitive task that influences clinical and financial decisions. In our problem formulation, we facilitate keeping practitioners in charge of all clinical decisions, as they can critically assess each model decision. This allows medical coders to work alongside AI tools, fostering human-machine collaboration rather than replacing human input, with basis on the supporting evidence.

Due to restrictions in data access, we used publicly available datasets that only represent a small part of the target population. To use the MDACE corpus, we took the *Data or Specimens Only Research* training course from the CITI program.[3]

Acknowledgments

This research was supported by the Portuguese Recovery and Resilience Plan through project C645008882-00000055 (i.e., the Center For Responsible AI).

[3] https://about.citiprogram.org/

References

Dhruv Agarwal, Rico Angell, Nicholas Monath, and Andrew McCallum. 2022. Entity linking via explicit mention-mention coreference modeling. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

Anton Alekseev, Zulfat Miftahutdinov, Elena Tutubalina, Artem Shelmanov, Vladimir Ivanov, Vladimir Kokh, Alexander Nesterov, Manvel Avetisian, Andrei Chertok, and Sergey Nikolenko. 2022. Medical crossing: A cross-lingual evaluation of clinical entity linking. In *Proceedings of the Language Resources and Evaluation Conference*.

Haadia Amjad, Mohammad Shehroz Ashraf, Syed Zoraiz Ali Sherazi, Saad Khan, Muhammad Moazam Fraz, Tahir Hameed, and Syed Ahmad Chan Bukhari. 2023. Attention-based explainability approaches in healthcare natural language processing. In *Proceedings of the International Joint Conference on Biomedical Engineering Systems and Technologies*.

Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador García, Sergio Gil-López, Daniel Molina, Richard Benjamins, et al. 2020. Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, 58:82–115.

Tom Ayoola, Shubhi Tyagi, Joseph Fisher, Christos Christodoulopoulos, and Andrea Pierleoni. 2022. ReFinED: An efficient zero-shot-capable approach to end-to-end entity linking. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics*.

Krisztian Balog. 2018. *Entity-Oriented Search*, chapter 5. Springer International Publishing.

Jose Barros, Matías Rojas, Jocelyn Dunstan, and Andres Abeliuk. 2022. Divide and conquer: An extreme multi-label classification approach for coding diseases and procedures in Spanish. In *Proceedings of the International Workshop on Health Text Mining and Information Analysis (LOUHI)*.

Joseph Spartacus Boyle, Antanas Kascenas, Pat Lok, Maria Liakata, and Alison Q O'Neil. 2023. Automated clinical coding using off-the-shelf large language models. In *Conference on Neural Information Processing Systems Workshop Deep Generative Models For Health*.

Tianqi Chen, Tong He, Michael Benesty, Vadim Khotilovich, Yuan Tang, Hyunsu Cho, Kailong Chen, Rory Mitchell, Ignacio Cano, Tianyi Zhou, et al.

2015. XGBoost: Extreme gradient boosting. *R package version 0.4-2*, 1(4):1–4.

Hua Cheng, Rana Jafari, April Russell, Russell Klopfer, Edmond Lu, Benjamin Striner, and Matthew Gormley. 2023. MDACE: MIMIC documents annotated with code evidence. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*.

Sébastien Cossin and Vianney Jouhet. 2020. IAM at CLEF eHealth 2020: Concept annotation in Spanish electronic health records. In *Working Notes of the Conference and Labs of the Evaluation Forum. CEUR Workshop Proceedings*.

Nicola De Cao, Gautier Izacard, Sebastian Riedel, and Fabio Petroni. 2021. Autoregressive entity retrieval. In *Proceedings of the International Conference on Learning Representations*.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. QLoRA: Efficient finetuning of quantized LLMs. In *Proceedings of the Conference on Neural Information Processing Systems*.

Hang Dong, Matúš Falis, William Whiteley, Beatrice Alex, Joshua Matterson, Shaoxiong Ji, Jiaoyan Chen, and Honghan Wu. 2022. Automated clinical coding: what, why, and where we are? *NPJ Digital Medicine*, 5(1):159.

Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Zhiyong Wu, Baobao Chang, Xu Sun, Jingjing Xu, and Zhifang Sui. 2024. A survey on in-context learning. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.

Joakim Edin, Maria Maistro, Lars Maaløe, Lasse Borgholt, Jakob D Havtorn, and Tuukka Ruotsalo. 2024. An unsupervised approach to achieve supervised-level explainability in healthcare records. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.

João Figueira, Gonçalo M. Correia, Michalina Strzyz, and Afonso Mendes. 2023. Justifying multi-label text classifications for healthcare applications. In *Proceedings of the European Conference on Information Retrieval*.

Fernando Gallego, Guillermo López-García, Luis Gasco-Sánchez, Martin Krallinger, and Francisco J Veredas. 2024. ClinLinker: Medical entity linking of clinical concept mentions in Spanish. In *Proceedings of the International Conference of Computational Science*.

Jamie Glen, Lifeng Han, Paul Rayson, and Goran Nenadic. 2024. A comparative study on automatic coding of medical letters with explainability. *arXiv preprint arXiv:2407.13638*.

Carey Beth Goldberg, Laura Adams, David Blumenthal, Patricia Flatley Brennan, Noah Brown, Atul J Butte, Morgan Cheatham, Dave DeBronkart, Jennifer Dixon, Jeff Drazen, et al. 2024. To do no harm—and the most good—with AI in health care. *Nature Medicine*, 1(3):623–627.

Edward J Hu, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. 2021. LoRA: Low-rank adaptation of large language models. In *Proceedings of the International Conference on Learning Representations*.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Alistair E W Johnson, Tom J Pollard, Lu Shen, Li-Wei H Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. 2016. MIMIC-III, a freely accessible critical care database. *Scientific Data*, 3(1):1–9.

Yanis Labrak, Adrien Bazoge, Emmanuel Morin, Pierre-Antoine Gourraud, Mickael Rouvier, and Richard Dufour. 2024. BioMistral: A collection of open-source pretrained large language models for medical domains. In *Findings of the Association for Computational Linguistics*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *Proceedings of the International Conference on Learning Representations*.

Antonio Miranda-Escalada, Luis Gascó, Salvador Lima-López, Eulàlia Farré-Maduell, Darryl Estrada, Anastasios Nentidis, Anastasia Krithara, Georgios Katsimpras, Georgios Paliouras, and Martin Krallinger. 2022. Overview of DisTEMIST at BioASQ: Automatic detection and normalization of diseases from clinical texts: results, methods, evaluation and multilingual resources. In *Working Notes of the Conference and Labs of the Evaluation Forum. CEUR Workshop Proceedings*.

Antonio Miranda-Escalada, Aitor Gonzalez-Agirre, Jordi Armengol-Estapé, and Martin Krallinger. 2020. 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 the Conference and Labs of the Evaluation Forum. CEUR Workshop Proceedings*.

Mengjie Ren, Boxi Cao, Hongyu Lin, Cao Liu, Xianpei Han, Ke Zeng, Guanglu Wan, Xunliang Cai, and Le Sun. 2024. Learning or self-aligning? Rethinking instruction fine-tuning. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*.

Hassan Shavarani and Anoop Sarkar. 2023. Spel: Structured prediction for entity linking. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.

Fei Teng, Yiming Liu, Tianrui Li, Yi Zhang, Shuangqing Li, and Yue Zhao. 2023. A review on deep neural networks for ICD coding. *IEEE Transactions on Knowledge and Data Engineering*, 35(5):4357–4375.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. LLaMA: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of the Conference on Neural Information Processing Systems*.

Ronald J Williams and David Zipser. 1989. A learning algorithm for continually running fully recurrent neural networks. *Neural computation*, 1(2):270–280.

John Wu, David Wu, and Jimeng Sun. 2024. Beyond label attention: Transparency in language models for automated medical coding via dictionary learning. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.

Zilin Xiao, Ming Gong, Jie Wu, Xingyao Zhang, Linjun Shou, and Daxin Jiang. 2023. Instructed language models with retrievers are powerful entity linkers. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.

Ikuya Yamada, Koki Washio, Hiroyuki Shindo, and Yuji Matsumoto. 2022. Global entity disambiguation with BERT. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

Hongyi Yuan, Zheng Yuan, Ruyi Gan, Jiaxing Zhang, Yutao Xie, and Sheng Yu. 2022a. BioBART: Pretraining and evaluation of a biomedical generative language model. In *Proceedings of the Workshop on Biomedical Language Processing*.

Hongyi Yuan, Zheng Yuan, and Sheng Yu. 2022b. Generative biomedical entity linking via knowledge base-guided pre-training and synonyms-aware fine-tuning. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

A Prompt Templates

The prompt used for ICL-BIOMISTRAL is in Listing 1. For SFT-BIOMISTRAL, we used a similar prompt, without the `[Example]`s. For INSGENEL-BIOMISTRAL, we used the prompt in Listing 2. We use a prompt in English, and generate outputs in English, even with CodiEsp's Spanish reports.

```
You are a medical coder at a hospital, and you have to assign ICD-10 codes to mentions. I will give you a report excerpt and a mention that can be found in that excerpt. Your job is to associate the mention to an ICD-10 code.
Each code can be a Diagnosis in ICD-10-CM or a Procedure in ICD-10-PCS. You should give the ICD-10 code according to its type (Diagnosis or Procedure).
[Example]:
The following report excerpt, written in <language>: """<example_mention_in_context>""", contains the following mention: <example_mention>.
It corresponds to the ICD-10 entity: <example_icd>.
[Task]:
The following report excerpt, written in <language>: """<mention_in_context>""", contains the following mention: <mention>.
It corresponds to the ICD-10 entity:
```

Listing 1: Prompt for ICL-BIOMISTRAL.

```
You are a medical coder at a hospital, and you have to assign ICD-10 codes to mentions.
I will give you a medical report, whose mentions are annotated between { and }. Your job is to associate each mention to an ICD-10 code.
Each code can be a Diagnosis in ICD-10-CM or a Procedure in ICD-10-PCS. You should give the ICD-10 code according to its type (Diagnosis or Procedure) and hierarchy, that is, you should first write the chapter, then the subchapter up until the title of the ICD-10 code, separated by "-->".
ICD-10 codes should be delimited by | and |.
Annotate the following report:
<report>
```

Listing 2: Prompt for INSGENEL-BIOMISTRAL.

B Dataset Details and Statistics

We used three different corpora during training.

- CodiEsp (Miranda-Escalada et al., 2020) consists of Spanish medical reports, which were manually annotated with their ICD-10 codes and textual evidence spans. The corpus was developed for the CodiEsp shared task, which comprises three sub-tasks: automated ICD coding for ICD-10-CM (CodiEsp-D) and ICD-10-PCS (CodiEsp-P), and end-to-end clinical

		Diagnoses			Procedures		
	Reports	Samples	Codes	1-shot codes	Samples	Codes	1-shot codes
CodiEsp	500	8,199	1,720	618	2,799	435	86
DisTEMIST	750	1,912	451	176	23	4	1
MDACE	181	4,993	966	446	168	89	61
Total	1,431	15,104	2,513	912	2,990	515	138

Table 5: Datasets used for training. *Codes* refers to the number of distinct ICD-10 codes in the training data, and *1-shot codes* refers to the number of codes that only appear once.

	CodiEsp	MDACE
No. 1-shot codes	219	49
No. 5-shot codes	923	203

Table 6: Number of 1-shot and 5-shot codes in the CodiEsp and MDACE test sets, considering the number of times they were seen in the training corpus.

entity linking for ICD-10 (CodiEsp-X).

- DisTEMIST (Miranda-Escalada et al., 2022) comprises medical reports in Spanish and English (we only used the English version), manually annotated with their SNOMED CT disease codes and textual evidence spans. The authors mapped the SNOMED CT codes to ICD-10 using UMLS. This mapping was only performed for the training data, so we could not evaluate our model's performance on the DisTEMIST validation and test splits.
- MDACE (Cheng et al., 2023) consists of English medical reports, which are part of the MIMIC-III collection (Johnson et al., 2016), with manually annotated ICD-10 codes and respective textual evidence spans.

The number of test few-shot codes is in Table 6. Table 5 summarizes the training datasets.

C Experimental Details

Our models were initialized with BioMistral-7B (Labrak et al., 2024). SFT- and INSGENEL-BIOMISTRAL were fine-tuned for 5 epochs on an NVIDIA RTX A6000 GPU for 20 hours, with a batch size of 4. We used QLoRA (Dettmers et al., 2023), with rank $r = 64$ and 4-bit NF quantization, and the AdamW (Loshchilov and Hutter, 2019) optimizer with a learning rate of $2 * 10^{-4}$ and weight decay equal to 10^{-3}. For inference, models were loaded without quantization on the same GPU, and we used the same batch sizes and a greedy decoding strategy. Inference took 8 hours for all datasets.

For INSGENEL-BIOMISTRAL, to ensure all training samples did not exceed the model's context window of 8,192 tokens, we truncated all documents to 5,000 characters. During inference, the entire documents were processed.

D Comparison Systems

In Table 4, we compare our experimental results on the CodiEsp test corpus with those of the IAM and DAC-E systems, which work as follows:

- The IAM (Cossin and Jouhet, 2020) system performs explainable ICD coding. It starts by normalizing every document in the training data, and composing a dictionary whose items are the normalized mentions (denoted *terms*) and their corresponding ground-truth ICD-10 codes. Additionally, the KB entities' normalized titles are added to the dictionary. Then, each dictionary term is tokenized and stored in an n-gram tree. For inference, a matching algorithm parses each document's tokens to find matching dictionary entries. Three matching strategies are employed: perfect matching, abbreviation matching (where a hand-crafted dictionary of abbreviations is used), and Levenshtein distance-based matching.
- The DAC-E (Barros et al., 2022) approach is not as directly explainable, as it treats ICD coding as MLC. This system comprises two sub-tasks, respectively performed by *matcher* and *ranker* models. The matcher associates documents to clusters (the chapters in ICD-10), leveraging a biomedical RoBERTa model (Liu et al., 2019). The ranker computes the likelihood of each code being present in a document, considering its chapter. It was

implemented with a binary classifier for each code, trained only with documents with codes in the same cluster, for better fine-grained differentiation. The ranker was trained using the XGBoost algorithm (Chen et al., 2015).