Establishing clinical NLP modelling recommendations for restricted data availability settings

Anonymous ACL submission

Abstract

When solving clinical Natural Language Processing (NLP) downstream tasks, it is well-established that incorporating clinical-specific knowledge enhances model performances. However, there are scenarios where access to data or domain-specific models is not feasible. Despite various paradigms for adapting clinical NLP-based models, such as fine-tuning already pre-trained language models, pre-training and fine-tuning models, or in-context learning, the advantages of each alternative regarding data availability still need to be explored. We determined the impact of data availability and paradigm selection in the performance of models on solving multiple clinical NLP tasks in Spanish by simulating multiple clinical data availability settings and testing various NLP modelling paradigms. Overall, the best-performing modelling strategy was pre-training a masked language model (LM) with environment-specific unannotated text starting from an off-the-shelf clinical checkpoint and then fine-tuning the LM for the downstream task. The increase in performance from the continuation of pre-training of an off-the-shelf LM is marginal, considering the high amount of resources needed for the pre-training; therefore, we recommend the fine-tuning of an off-the-shelf clinical-specific LM if the model and task-specific data are available. We recommend a few-shot learning technique using a large LM if no data is available.

1 Introduction

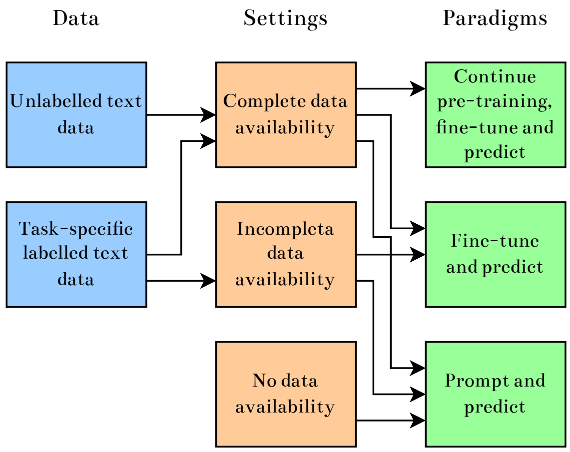
Natural language processing (NLP) has gained tremendous importance in recent years with the advent of Transformer-based pre-trained language models (PLM) (Qiu et al., 2020) and large language models (LLM) (Zhao et al., 2023). These LMs have become the new paradigm for NLP-based machine learning modelling because of their modularity and ease of transferring learning. One can fine-tune a PLM to solve any NLP task using off-the-shelf, already pre-trained LMs (Dodge et al., 2020).

It is known that using closer-to-the-domain LMs for fine-tuning downstream models improves the performance of the fine-tuned model (Gu et al., 2021; Zheng et al., 2022; Carrino et al., 2021). One of the most widespread paradigms for NLP-based modelling is the usage of a PLM for representing documents and then using a couple of layers to adapt the PLM output to solve the specific task; this framework is called the fine-tuning of PLMs. There are multiple options to optimize a model using this framework. The first option is to use a PLM and then fine-tune it to the downstream task or to pre-train an LM and then fine-tune it to the downstream task by employing unlabeled and task-labeled data. The second option is useful when no pre-trained models are available or one wants a closer-to-the-domain PLM. This second method involves initial pre-training of an LM utilizing unlabeled data specific to the target domain, followed by fine-tuning the LM's architecture with task-labeled data, mirroring the process outlined in the first option. The downside of the second option is that data is needed for both the PLM pre-training and the architecture's fine-tuning. These paradigms are described in detail in the following sections, and an overview is shown in Figure 1.

The paradigm described above requires at least some task-labeled data, but there are some settings where no data is available. A new paradigm for NLP-based modelling has arisen, where an instruction-tuned causal LLM is prompted in natural language to act as an NLP-based model with few or zero examples given (Liu et al., 2023), exploiting its in-context learning ability. This framework is also an option to consider when building NLP-based models.

Data can be restricted in clinical environments for multiple reasons, such as privacy-related issues or the lack of interoperability. These restrictions

Figure 1: Overview of the compatibilities between available data, settings and NLP paradigms that will be described in the paper.



lead to some specific settings, where in one, there is abundant data availability to apply the entire set of NLP modelling paradigms, another where the access to data is incomplete; thus, not all paradigms can be used to develop models, and in the last one no data is available; therefore a specific paradigm should be used. These settings will be carefully described in the following sections and are summarized along with their paradigm compatibility in Figure 1.

Problem A situation arises when there is an asymmetry in data availability or no data is available. In some cases, there is only task-labeled data, only domain-specific unlabeled data, or no data is available at all. Even though multiple paradigms exist for NLP modelling in clinical environments, the compatibility between data availability and the NLP modelling paradigm regarding gains in performance still needs to be explored.

Solution We performed an experimental analysis to measure the performance of solving clinical NLP tasks in Spanish with multiple data availability and NLP modelling paradigm combinations and empirically constructed recommendations for clinical NLP modelling regarding data availability.

1.1 Background

The last paradigm for deep-learning-based NLP was the usage of recurrent neural networks (RNN), which preserved the sequence nature of language in the representation of meaning (Chung et al., 2014; Hochreiter and Schmidhuber, 1997). One drawback of recurrent RNNs is their limited parallelizability, resulting in prolonged training times. Furthermore, as the sequence lengths grow, there is a tendency for information gathered at distant time steps to vanish due to inherent memory limitations. Nowadays, the Transformer completely ditches the recurrence of the architecture but also preserves word order by learning dependences without regard to their distance in the sentences (Vaswani et al., 2017). With its attention mechanism, this architecture reaches state-of-the-art in multiple NLP tasks such as text classification (Yang et al., 2019), sentiment analysis (Yang et al., 2019), dependency parsing (Mrini et al., 2020), machine translation (Edunov et al., 2018), and named entity recognition (Wang et al., 2021).

1.1.1 Pretrained language models

PLMs are LMs that were trained using self-supervised techniques over large *corpora* of unannotated text to transfer learning from the knowledge gathered in the pre-training to downstream task-specific models (Wang et al., 2022b). Early methods for PLMs consisted of static word embeddings, which were distributed word representations learned using algorithms such as Word2Vec (Mikolov et al., 2013) or GloVe (Pennington et al., 2014), and these embeddings were standard initialization parameters for deep learning architectures to solve NLP tasks. There has been a shift towards dynamic or context-aware word embeddings, which solves the problem of static word embeddings that do not consider word polysemy. These context-aware word embeddings were initially composed using RNNs (Dai and Le, 2015) such as in ELMo (Peters et al., 2018), but currently, they are based on the Transformer architecture and use web-scale unannotated text to be trained. The *de facto* standard for pre-trained Transformer-based context-aware models is BERT (Devlin et al., 2019) and BERT-alike models such as RoBERTa (He et al., 2021) and DeBERTa (He et al., 2021). This language model learns bidirectional contexts conditioning on both left and right contexts in deep stacked layers. Using BERT as a base architecture, domain-specific models have arisen, such as roberta-base-bne (Fandiño et al., 2022), a RoBERTa-based PLM for the Spanish language, PubMedBERT (Gu et al., 2021), a PLM for the biomedical domain and LEGAL-BERT (Chalkidis et al., 2020), a PLM for the legal domain, among others.

1.1.2 Large language models

LLMs are PLMs with a significantly larger model size scale (Zhao et al., 2023); for example, the PLM BERT has a model size of 0.3×10^9 parameters and the LLM GPT-3 (Brown et al., 2020), has 175×10^9 parameters. It has been found that scaling PLMs improves the performance of the models on downstream tasks (Kaplan et al., 2020); although this is true, some other surprising and more important behaviours in solving a series of complex tasks appear at LLM scales and were called *emergent abilities*. Emergent abilities are aptitudes not present in small models but arise in LLMs (Wei et al., 2022) and include *in-context learning*, where a model can generate expected outputs to natural language instructions without additional training, *instruction following*, where a model fine-tuned using natural language instructions performs well on unseen tasks that are also described in the form of instructions and *step-by-step reasoning*, where a model can solve complex problems by instructing the model involving intermediate reasoning steps for deriving the final answer. GPT-3, a closed-source privative LLM, formally introduced the concept of in-context learning, and from there, subsequent models have appeared, such as open-source models Galactica (Taylor et al., 2022), a 120×10^9 parameters model and LLaMA (Touvron et al., 2023a), a 65×10^9 parameters model. It is worth noting a significant milestone in LLMs called ChatGPT, a closed-source privative assistant-style LLM that exhibited a superior capacity to communicate with humans and has been in widespread usage by laypeople.

1.1.3 Pre-train, fine-tune and predict paradigm of PLMs

The primary adaptation method for adjusting PLM to downstream tasks is fine-tuning, where a task-specific layer is concatenated to the output of the PLM (Qiu et al., 2020). This method was proposed in the Universal Language Model Fine-Tuning (ULMFiT) framework as a transfer learning technique for domain-specific NLP, achieving state-of-the-art performances in multiple NLP tasks (Howard and Ruder, 2018). Even though the fine-tuning paradigm has been well described for adapting PLMs, LLMs have significantly higher computational complexity due to their unprecedented scale. For this reason, some special techniques have been developed, such as Parameter-Efficient Fine-Tuning (PEFT), where a small set of parameters are trained to enable a model to perform the new task (Ding et al., 2023), showing improvements over in-context learning (Liu et al., 2022).

1.1.4 Pre-train, prompt and predict paradigm of LLMs

The principal approach for interfacing with LLMs is through prompting, instructions in natural language issued to LLMs to adapt them to new scenarios with few or no labelled data (Zhao et al., 2023) by exploiting the emergent ability of in-context learning. This new NLP paradigm created a new field of prompt engineering, where prompting templates are created to achieve the most effective performance on downstream tasks (Liu et al., 2023). There is mixed evidence comparing fine-tuning vs in-context learning, whereas in some tasks such as in biomedical information extraction (Jimenez Gutierrez et al., 2022) or out-of-domain generalization (Mosbach et al., 2023) fine-tuning outperforms in-context learning, in other tasks such as code intelligence (Wang et al., 2022a), in-context learning outperforms fine-tuning.

1.1.5 Clinical NLP

For clinical NLP, domain-specific models have been explored in the literature, and their positive impact on downstream clinical NLP tasks has been proven (Kalyan and Sangeetha, 2020; Lewis et al., 2020) even in Spanish (Carrino et al., 2022). There are public pre-trained Spanish language models for the clinical domain, including masked LMs, such as the one we are going to describe in the next section and small causal character-level LMs, such as Clinical-Flair (Rojas et al., 2022), though, in the large LM category; there are very few, and only for the English language, such as BioMedLM (Bolton et al., 2023) and MEDITRON (Chen et al., 2023).

Even though most of the clinical NLP research has focused on the pre-train, fine-tune, and predict paradigm, some works have explored the prompt and predict paradigm through few-shot models (Sivarajkumar and Wang, 2022), validating that one can extract clinical information from documents through prompting (Sivarajkumar et al., 2023; Agrawal et al., 2022).

2 Data & methods

We intentionally limited data access to evaluate its impact on the performance of multiple clinical NLP modelling paradigms and foundation models.

Each restricted setting was based on a real-world simulated clinical environment.

2.1 Simulated settings

To mimic clinical settings regarding data availability, we simulated multiple settings with varying levels of data availability. We divided the data into two categories: task-specific labelled data, which can be used to fine-tune models and setting-specific unlabelled data, which can be used to continue the pre-training of the foundation models. The overall environment we are located in is a Chilean public health institution analyzing waiting list data, where the explanation of why the patient is waiting is in the form of free text, and from that dataset, multiple tasks need to be solved. Multiple reasons can restrict data availability; for example, data availability for model training can be restricted due to legal and privacy reasons or because the task trying to be solved still does not have sufficient examples due to its recent appearance.

Unannotated data The unlabelled data we used to continue the pre-training of the foundation models was the complete set of reasons for referral contained in the Chilean waiting list and is comprised of $13\,365\,476$ documents, totalling $65\,891\,568$ tokens with a vocabulary size of $513\,315$ types.

2.1.1 Complete data availability

In this data availability setting, unlabelled unstructured free-text data to continue the pre-training and task-specific labelled data are also available to fine-tune foundation models. This setting can be seen at a large healthcare provider or at a country-level public health institution such as a ministry of health, where data policies are well established, and patients must consent that their data can be used to tune machine learning models.

2.1.2 Incomplete data availability

In this data availability setting, only task-specific labelled data is available to fine-tune foundation models. The lack of unlabelled unstructured free-text data to continue the pre-training may be attributed to the fact that the provided is only acquiring data for the specific task and does not have access to close-to-the-environment unlabelled text data or according to data policies, the provider cannot merge patient data from a different source, other than the source of the task data. This setting can be seen at a medium-sized healthcare provider where the data warehousing methods are not implemented or the provider only has access to specific and segmented data sources due to the lack of interoperability.

2.1.3 No data availability

In this data availability setting, there is no unlabelled unstructured free-text data to continue the pre-training nor task-specific labelled data to fine-tune foundation models. The absence of data can be attributed to the lack of access to the electronic health record (EHR) database or policies that forbid patient data usage to tune machine learning models. This setting can be seen in a healthcare provider using an external EHR service that forbids access to the underlying database, or the provider wants to solve a new task where data is not yet available.

2.2 Clinical NLP tasks

To measure the impact of data availability on the performance of clinical NLP modelling, we used multiple clinical NLP tasks, where each is under the same environment of the analysis of unstructured waiting list data.

2.2.1 Referral prioritization

Different methods exist to prioritize patient selection to process the waiting list more fairly, and we modelled the patient prioritization through the classification of each referral regarding its state according to the Chilean Explicit Health Guarantees law (GES in Spanish), which states that specific health problems must be guaranteed to be resolved within a particular time frame. This task requires a binary classification modelling technique. The dataset (citation redacted for anonimity) contains $1\,701\,582$ examples in the training subset, $485\,649$ in the test subset and $242\,746$ in the validation subset.

2.2.2 Referral speciality classification

Each referral contained on the waiting list corresponds to a specific medical speciality. This task involves the prediction of the corresponding medical speciality given the free-text description of the reason for referral contained on the waiting list record. This task requires a multilabel modelling technique with a label space size of 48 classes. The dataset contains $3\,401\,173$ examples in the training subset, $971\,764$ in the test subset and $485\,882$ in the validation subset.

2.2.3 Clinical named entity recognition

Clinical named entity recognition is a subtype of named entity recognition in which entities of clinical interest are extracted from unstructured free-text sources. This dataset (citation redacted for anonimity) is annotated with eleven different clinical entity classes and was modelled as a token classification problem, where each of the tokens of the referrals is classified into one of the eleven clinical entity classes. The dataset contains 7987 documents in the training subset, 987 in the test subset and 887 in the validation subset.

2.3 Foundation models

We used multiple foundation models as a basis to solve the clinical NLP tasks. The attributes used to select the foundation models were the language, domain and modelling technique.

2.3.1 XLM-RoBERTa

A multilingual version of XLM-RoBERTa masked language model, pre-trained using a self-supervised technique on a *corpus* of 2.5TB of filtered CommonCrawl raw text data containing one hundred languages (Conneau et al., 2019). This model is the broadest of all of our selected foundation LMs. This model should be viewed as a baseline where no model is available for the language or the domain.

2.3.2 Spanish RoBERTa

A Spanish language version of RoBERTa masked language model, pre-trained on a *corpus* of 570GB of clean and deduplicated text, compiled from the web crawlings performed by the National Library of Spain (Biblioteca Nacional de España) from 2009 to 2019 (Fandiño et al., 2022). This model is only compatible with the language in which the clinical NLP tasks are and is a type of model (regarding language) that should be used when no domain-specific model is available.

2.3.3 Spanish biomedical and clinical RoBERTa

A Spanish language biomedical and clinical version of RoBERTa masked language model, pre-trained on a *corpus* of several biomedical *corpora* in Spanish, collected from publicly available *corpora* and crawlers, and a real-world clinical *corpus*. The entire *corpus* was comprised of more than 1B tokens (Carrino et al., 2021). This model is the closest to the domain model we used to solve the tasks, compatible with both language and domain; this should be the best-suited model to solve a domain-specific task.

2.3.4 Llama 2

Llama 2 is a causal auto-regressive language model that uses an optimized transformer architecture trained on a *corpus* of publicly available online data comprised of two trillion tokens (Touvron et al., 2023b). This model is the largest we tested but is not domain-adapted in any way, and this is the model we used for in-context learning prediction.

2.4 Modelling paradigms

We utilized various NLP modelling paradigms to tackle each clinical NLP task, experimenting with multiple paradigms for some foundational models based on their compatibility. Also, we note the compatibility of each paradigm with each data availability setting.

2.4.1 Continue pre-training, fine-tune and predict

This modelling paradigm is the most data-intensive, where we start with an already pre-trained LM checkpoint and continue the pre-training for five epochs with the closer-to-the-environment unannotated data described in 2.1. Then, with the now environment-adapted LM, we perform a fine-tuning for five epochs to solve each clinical NLP task. We continued the pre-training of all the masked LMs (XLM-RoBERTa, Spanish RoBERTa and Spanish biomedical and clinical RoBERTa) with no modification to the original vocabulary and using model-default hyperparameters. This paradigm is compatible only with the Complete data availability setting.

2.4.2 Fine-tune and predict

In this paradigm, we started with each of the off-the-shelf masked foundation models (XLM-RoBERTa, Spanish RoBERTa and Spanish biomedical and clinical RoBERTa) and performed fine-tuning for each of the clinical NLP tasks. We fine-tuned each task using the default model hyperparameters and trained for five epochs. This paradigm is compatible with both Complete data availability and Incomplete data availability settings.

2.4.3 Prompt and predict

In this paradigm, we exploited LLMs' in-context learning emergent ability through zero-shot and few-shot techniques. We prompted the LLM (Llama 2) to solve each task and parsed its answer accordingly. For the few-shot technique, we randomly sampled five examples of the training

Model	Prioritization	Specialty	CNER
xlm-roberta			
Off-the-shelf	88.85 %	51.71 %	11.09 %
Environment-pre-trained	89.03 % (+0.18)	52.36 % (+0.65)	13.85 % (+2.76)
roberta-bne			
Off-the-shelf	88.58 %	52.50 %	22.59 %
Environment-pre-trained	88.80 % (+0.22)	51.65 % (-0.85)	23.29 % (+0.70)
roberta-biomedical-clinical			
Off-the-shelf	88.80 %	53.79 %	34.46 %
Environment-pre-trained	88.85 % (+0.05)	53.85 % (+0.06)	37.25 % (+2.79)
Llama 2			
Zero-shot	6.49 %	31.41 %	5.31 %
Few-shot	56.70 % (+50.21)	31.91 % (+0.50)	15.44 % (+10.13)

Table 1: Results (macro F_1 score) for each clinical NLP task and each fine-tuned model with and without environment continuation of pre-training.

subset of each clinical NLP task. This task is compatible with Complete data availability, Incomplete data availability and No data availability settings. The prompt templates used to solve each task are available in the appendix A.

2.5 Increasing training data size and its impact on model performance

To better understand the direct impact of the number of training examples, we performed a test in which we truncated the training subset in increasing steps and measured the performance of the fine-tuned model on the complete test subset. We applied this experiment to all settings and masked LMs.

3 Results

The results for each modelling paradigm solving each clinical NLP task are presented in Table 1.

The Referral prioritization task was where the model performed the best due to its straightforward binary nature. The models could identify the prioritized health problems mentioned in the training subset and generalize the knowledge correctly in the test subset. On the other hand, the Clinical named entity recognition task was the most complex of the three tasks, and the models struggled the most to solve it. The performance of the models was directly related to the intrinsic complexity of the clinical NLP task.

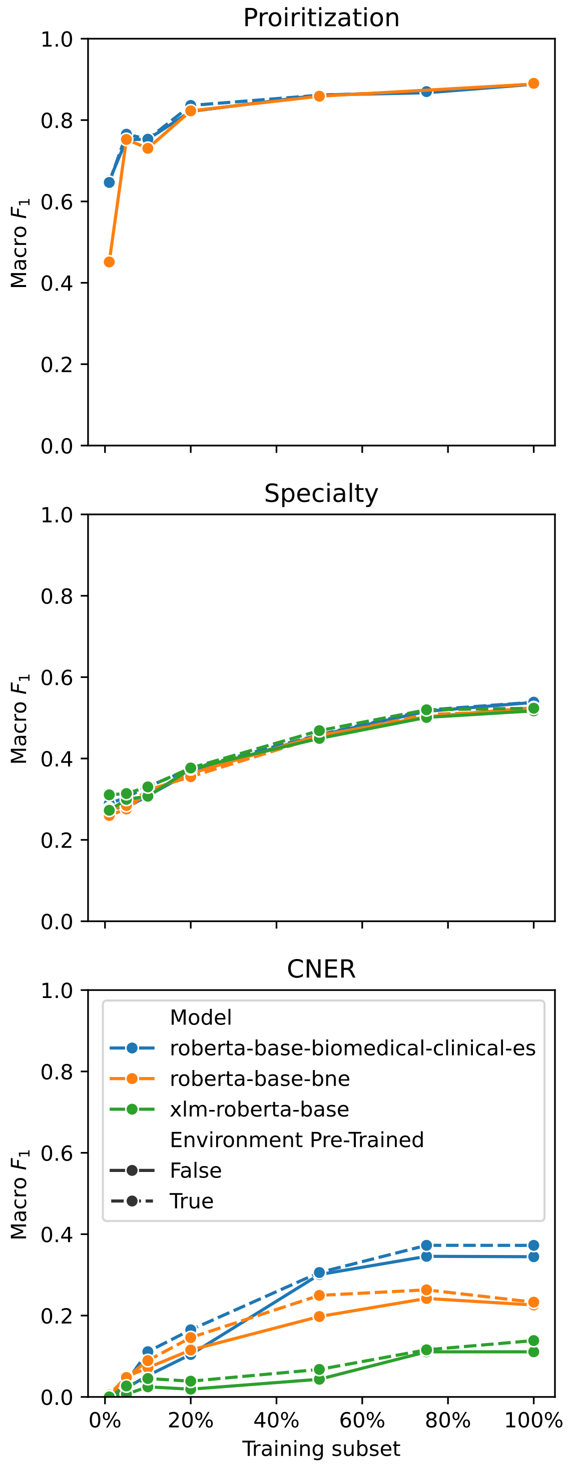
Regarding the foundation language model used to solve the clinical NLP tasks, the Spanish biomedical and clinical RoBERTa model was the best performant; this model is the closest to the domain of the clinical NLP tasks and therefore was able to transfer learning from its pre-training on a large clinical corpus. The worst-performing model was XLM-RoBERTa, which is less close to the domain foundation model; therefore, its ability to use prior knowledge to model the tasks was lacking. The closeness to the domain between the foundation model and the task is correlated with model performance on downstream tasks. Before, we only compared models that solved the tasks using the same paradigm, and overall, the worst performant model was Llama 2; however, this model was used by exploiting a different paradigm.

Environment adaptation using unlabelled unstructured free-text data improved model performance. However, the improvements are marginal, considering the computational resources used to continue the pre-training of the foundation models.

The Continue pre-training, fine-tune and predict paradigm achieved the best result in solving all the clinical NLP tasks. However, we do not recommend using it as a paradigm for clinical NLP modelling, given its high resource usage for training and its overall low gain in performance. On the other hand, the Prompt and predict paradigm was the worst performant paradigm of all three, but it is worth noting that in some specific cases, its performance was better than the other paradigms. Also, the access to few-shot examples drastically improves in-context learning performance. We recommend using this paradigm in settings with minimal access to training data.

The experiment's results on the impact of training data volume on the performance of downstream

Figure 2: Performance (macro F_1 score) by training subset for each clinical NLP task and each fine-tuned model with and without environment continuation of pre-training



tasks are presented in Figure 2.

All the models display a performance saturation even before attaining complete training data. This phenomenon is best noticed in the Referral prioritization clinical NLP task, where minimal access to training data can result in almost peak performance. This behaviour further indicates that the task has a relatively low complexity. The Referral speciality classification task exhibits a more nuanced performance saturation phenomenon than other tasks. The correlation between training data availability and performance is nearly linear, indicating that access to training data is crucial for specific complex tasks.

4 Conclusion

Our study investigated the impact of data availability on the performance of clinical NLP modelling in simulated settings with varying levels of access to task-specific labelled data and unannotated environment-specific text. We explored different paradigms, including Continue pre-training, fine-tune and predict, Fine-tune and predict, as well as Prompt and predict with few-shot learning.

Our findings indicate that choosing foundation models, especially those closer to the target domain impacts model performance. The Spanish biomedical and clinical RoBERTa model, tailored to the clinical domain, outperformed other models in our experiments. While continuing pre-training with environment-specific data improved model performance, the gains were marginal compared to the computational resources required. The fine-tuning paradigm without additional pre-training proved practical, particularly in settings with limited access to unlabelled data.

In-context learning, using the prompt and predict paradigm, demonstrated its viability, especially in scenarios where there is no labelled data available. The creation of few-shot examples significantly improved performance, highlighting the potential of this approach in data-scarce environments.

Our study also revealed a saturation point in performance concerning the amount of training data available. In some instances, minimal data access can still lead to relatively high performance, particularly for less complex tasks.

The choice of foundation models, the utilization of available data, and the selection of appropriate modelling paradigms are crucial considerations in clinical NLP tasks. While pre-training and fine-tuning with domain-specific data remain effective, in-context learning with few-shot examples offers a viable solution in settings where labelled data is unavailable.

5 Recommendations

Based on our comprehensive analysis, we provide recommendations for practitioners engaged in clinical NLP modelling:

Model selection When selecting foundation models, prioritize those that align closely with the target domain. Our results emphasize the significance of domain specificity in achieving optimal performance.

Data utilization In settings with ample access to task-specific labelled data and unannotated domain-specific text, the Continue pre-training, fine-tune and predict paradigm may be considered. However, given the resource-intensive nature of this approach, practitioners may opt for the Fine-tune and predict paradigm, especially when computational resources are constrained.

If no data is available In scenarios with no access to labelled data, the Prompt and predict paradigm, particularly with few-shot learning, emerges as a practical and effective solution. This approach allows models to leverage general knowledge and adapt to new tasks with minimal labelled examples.

Consideration of task complexity Recognize the inherent complexity of the clinical NLP task at hand. Tasks with lower complexity may achieve near-optimal performance even with minimal access to training data, highlighting the importance of task-specific considerations.

Continuous investigation Clinical NLP is dynamic, and advancements in pre-trained foundation LMs and novel paradigms are frequent. Continuously exploring emerging techniques and adapting to the evolving landscape is essential for staying at the forefront of effective healthcare information extraction.

By incorporating these recommendations, practitioners can make informed decisions based on the specific characteristics and constraints of their clinical NLP projects, ultimately enhancing the efficiency and efficacy of their models in real-world healthcare applications.

6 Limitations

We attempted to use settings that can be easily understood in real-world scenarios, but we may have unintentional biases based on our experiences in local environments. Our choice of foundation Language Models (LMs) for each category (multilingual, language-specific, and domain-specific) may require a different categorization in order to provide representative examples of LMs.

7 Ethics statement

We obtained all the data we used through a transparency law that requires public health providers to make the data available to the public. This means that anyone can access the same data that we used, provided they follow the same process that we did. The data we used is public, and we have cited the source papers where each dataset was officially released to the public.

References

Monica Agrawal, Stefan Hegselmann, Hunter Lang, Yoon Kim, and David Sontag. 2022. Large language models are few-shot clinical information extractors. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 1998–2022, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Elliot Bolton, David Hall, Michihiro Yasunaga, Tony Lee, Chris Manning, and Percy Liang. 2023. BioMedLM. `https://crfm.stanford.edu/2022/12/15/biomedlm.html`. [Accessed 14-12-2023].

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Casimiro Pio Carrino, Jordi Armengol-Estapé, Asier Gutiérrez-Fandiño, Joan Llop-Palao, Marc Pàmies, Aitor Gonzalez-Agirre, and Marta Villegas. 2021. Biomedical and clinical language models for spanish: On the benefits of domain-specific pretraining in a mid-resource scenario.

Casimiro Pio Carrino, Joan Llop, Marc Pàmies, Asier Gutiérrez-Fandiño, Jordi Armengol-Estapé, Joaquín

Silveira-Ocampo, Alfonso Valencia, Aitor Gonzalez-Agirre, and Marta Villegas. 2022. Pretrained biomedical language models for clinical nlp in spanish. In *Proceedings of the 21st Workshop on Biomedical Language Processing*. Association for Computational Linguistics.

Ilias Chalkidis, Manos Fergadiotis, Prodromos Malakasiotis, Nikolaos Aletras, and Ion Androutsopoulos. 2020. LEGAL-BERT: The muppets straight out of law school. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2898–2904, Online. Association for Computational Linguistics.

Zeming Chen, Alejandro Hernández Cano, Angelika Romanou, Antoine Bonnet, Kyle Matoba, Francesco Salvi, Matteo Pagliardini, Simin Fan, Andreas Köpf, Amirkeivan Mohtashami, Alexandre Sallinen, Alireza Sakhaeirad, Vinitra Swamy, Igor Krawczuk, Deniz Bayazit, Axel Marmet, Syrielle Montariol, Mary-Anne Hartley, Martin Jaggi, and Antoine Bosselut. 2023. Meditron-70b: Scaling medical pretraining for large language models.

Junyoung Chung, Caglar Gulcehre, Kyunghyun Cho, and Yoshua Bengio. 2014. Empirical evaluation of gated recurrent neural networks on sequence modeling. In *NIPS 2014 Workshop on Deep Learning, December 2014*.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *CoRR*, abs/1911.02116.

Andrew M Dai and Quoc V Le. 2015. Semi-supervised sequence learning. In *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Ning Ding, Yujia Qin, Guang Yang, Fuchao Wei, Zonghan Yang, Yusheng Su, Shengding Hu, Yulin Chen, Chi-Min Chan, Weize Chen, Jing Yi, Weilin Zhao, Xiaozhi Wang, Zhiyuan Liu, Hai-Tao Zheng, Jianfei Chen, Yang Liu, Jie Tang, Juanzi Li, and Maosong Sun. 2023. Parameter-efficient fine-tuning of large-scale pre-trained language models. *Nature Machine Intelligence*, 5(3):220–235.

Jesse Dodge, Gabriel Ilharco, Roy Schwartz, Ali Farhadi, Hannaneh Hajishirzi, and Noah Smith. 2020. Fine-tuning pretrained language models: Weight initializations, data orders, and early stopping.

Sergey Edunov, Myle Ott, Michael Auli, and David Grangier. 2018. Understanding back-translation at scale. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 489–500, Brussels, Belgium. Association for Computational Linguistics.

Asier Gutiérrez Fandiño, Jordi Armengol Estapé, Marc Pàmies, Joan Llop Palao, Joaquin Silveira Ocampo, Casimiro Pio Carrino, Carme Armentano Oller, Carlos Rodriguez Penagos, Aitor Gonzalez Agirre, and Marta Villegas. 2022. Maria: Spanish language models. *Procesamiento del Lenguaje Natural*, 68.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2021. Domain-specific language model pretraining for biomedical natural language processing. *ACM Trans. Comput. Healthcare*, 3(1).

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. Deberta: Decoding-enhanced bert with disentangled attention. In *International Conference on Learning Representations*.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long Short-Term Memory. *Neural Computation*, 9(8):1735–1780.

Jeremy Howard and Sebastian Ruder. 2018. Universal language model fine-tuning for text classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 328–339, Melbourne, Australia. Association for Computational Linguistics.

Bernal Jimenez Gutierrez, Nikolas McNeal, Clayton Washington, You Chen, Lang Li, Huan Sun, and Yu Su. 2022. Thinking about GPT-3 in-context learning for biomedical IE? think again. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 4497–4512, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Katikapalli Subramanyam Kalyan and S. Sangeetha. 2020. Secnlp: A survey of embeddings in clinical natural language processing. *Journal of Biomedical Informatics*, 101:103323.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling laws for neural language models. *CoRR*, abs/2001.08361.

Patrick Lewis, Myle Ott, Jingfei Du, and Veselin Stoyanov. 2020. Pretrained language models for biomedical and clinical tasks: Understanding and extending the state-of-the-art. In *Proceedings of the 3rd Clinical Natural Language Processing Workshop*. Association for Computational Linguistics.

Haokun Liu, Derek Tam, Muqeeth Mohammed, Jay Mohta, Tenghao Huang, Mohit Bansal, and Colin Raffel.

2022. Few-shot parameter-efficient fine-tuning is better and cheaper than in-context learning. In *Advances in Neural Information Processing Systems*.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2023. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, 55(9):1–35.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc.

Marius Mosbach, Tiago Pimentel, Shauli Ravfogel, Dietrich Klakow, and Yanai Elazar. 2023. Few-shot fine-tuning vs. in-context learning: A fair comparison and evaluation.

Khalil Mrini, Franck Dernoncourt, Quan Hung Tran, Trung Bui, Walter Chang, and Ndapa Nakashole. 2020. Rethinking self-attention: Towards interpretability in neural parsing. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 731–742, Online. Association for Computational Linguistics.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.

XiPeng Qiu, TianXiang Sun, YiGe Xu, YunFan Shao, Ning Dai, and XuanJing Huang. 2020. Pre-trained models for natural language processing: A survey. *Science China Technological Sciences*, 63(10):1872–1897.

Matías Rojas, Jocelyn Dunstan, and Fabián Villena. 2022. Clinical flair: A pre-trained language model for Spanish clinical natural language processing. In *Proceedings of the 4th Clinical Natural Language Processing Workshop*, pages 87–92, Seattle, WA. Association for Computational Linguistics.

Sonish Sivarajkumar, Mark Kelley, Alyssa Samolyk-Mazzanti, Shyam Visweswaran, and Yanshan Wang. 2023. An empirical evaluation of prompting strategies for large language models in zero-shot clinical natural language processing.

Sonish Sivarajkumar and Yanshan Wang. 2022. Health-Prompt: A zero-shot learning paradigm for clinical natural language processing. *AMIA Annu. Symp. Proc.*, 2022:972–981.

Ross Taylor, Marcin Kardas, Guillem Cucurull, Thomas Scialom, Anthony Hartshorn, Elvis Saravia, Andrew Poulton, Viktor Kerkez, and Robert Stojnic. 2022. Galactica: A large language model for science.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023a. Llama: Open and efficient foundation language models.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023b. Llama 2: Open foundation and fine-tuned chat models.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Chaozheng Wang, Yuanhang Yang, Cuiyun Gao, Yun Peng, Hongyu Zhang, and Michael R. Lyu. 2022a. No more fine-tuning? an experimental evaluation of prompt tuning in code intelligence. In *Proceedings of the 30th ACM Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering*, ESEC/FSE 2022, page 382–394, New York, NY, USA. Association for Computing Machinery.

Haifeng Wang, Jiwei Li, Hua Wu, Eduard Hovy, and Yu Sun. 2022b. Pre-trained language models and their applications. *Engineering*.

Xinyu Wang, Yong Jiang, Nguyen Bach, Tao Wang, Zhongqiang Huang, Fei Huang, and Kewei Tu. 2021. Automated concatenation of embeddings for structured prediction. 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), pages 2643–2660, Online. Association for Computational Linguistics.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. 2022. Emergent abilities of large language models. *Transactions on Machine Learning Research*. Survey Certification.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. 2019. *XLNet: Generalized Autoregressive Pretraining for Language Understanding*. Curran Associates Inc., Red Hook, NY, USA.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, Peiyu Liu, Jian-Yun Nie, and Ji-Rong Wen. 2023. A survey of large language models. *arXiv preprint arXiv:2303.18223*.

Zhe Zheng, Xin-Zheng Lu, Ke-Yin Chen, Yu-Cheng Zhou, and Jia-Rui Lin. 2022. Pretrained domain-specific language model for natural language processing tasks in the AEC domain. *Computers in Industry*, 142:103733.

A Prompt templates used for in-context learning

We describe the prompt templates we used to solve each clinical NLP task using in-context learning.

A.1 Referral prioritization

System prompt template `En Chile, las garantías explícitas de salud establecen prioridad para un conjunto de problemas de salud. Debes responder en español sólo la palabra "Verdadero" si la enfermedad que te entregue pertenece a uno de los 80 problemas de salud y sólo la palabra "Falso" si la enfermedad no pertenece al conjunto de problemas. Los problemas de salud son: "Accidente Cerebrovascular Isquémico en personas de 15 años y más", "Alivio del dolor y cuidados paliativos por cáncer avanzado ", "Analgesia del Parto", "Artritis Reumatoídea", "Artritis idiopática juvenil", "Asma Bronquial moderada y grave en personas menores de 15 años", "Asma bronquial en personas de 15 años y más", "Cardiopatías congénitas operables en menores de 15 años", "Colecistectomía preventiva del cáncer de vesícula en personas de 35 a 49 años", "Consumo Perjudicial o Dependencia de riesgo bajo a moderado de alcohol y drogas en personas menores de 20 años", "Cáncer Cervicouterino", "Cáncer Colorectal en personas de 15 años y más", "Cáncer Vesical en personas de 15 años y más", "Cáncer de Ovario Epitelial", "Cáncer de mama en personas de 15 años y más", "Cáncer de próstata en personas de 15 años y más", "Cáncer de testículo en personas de 15 años y más", "Cáncer en personas menores de 15 años", "Cáncer gástrico", "Depresión en personas de 15 años y más", "Desprendimiento de retina regmatógeno no traumático", "Diabetes Mellitus Tipo 1", "Diabetes Mellitus Tipo 2", "Displasia broncopulmonar del prematuro", "Displasia luxante de caderas", "Disrafias espinales", "Endoprótesis total de cadera en personas de 65 años y más con artrosis de cadera con limitación funcional severa", "Enfermedad Pulmonar Obstructiva Crónica de Tratamiento Ambulatorio", "Enfermedad Renal Crónica Etapa 4 y 5", "Enfermedad de Parkinson", "Epilepsia no refractaria en personas de 15 años y más", "Epilepsia no refractaria en personas desde 1 año y menores de 15 años", "Esclerosis múltiple remitente recurrente ", "Esquizofrenia", "Estrabismo en personas menores de 9 años", "Fibrosis Quística", "Fisura labiopalatina", "Gran Quemado", "Hemofilia", "Hemorragia Subaracnoidea secundaria a Ruptura de Aneurismas Cerebrales", "Hepatitis C", "Hepatitis crónica por Virus Hepatitis B", "Hipertensión arterial primaria o esencial en personas de 15 años y más", "Hipoacusia Bilateral en personas de 65 años y más que requieren uso de audífono", "Hipoacusia neurosensorial bilateral del prematuro", "Hipotiroidismo en personas de 15 años y más", "Infarto agudo del miocardio", "Infección respiratoria aguda (IRA) de manejo ambulatorio en personas menores de 5 años", "Leucemia en personas de 15 años y más", "Linfomas` en personas de 15 años y más", "Lupus Eritematoso Sistémico", "Neumonía adquirida en la comunidad de manejo ambulatorio en personas de 65 años y más", "Osteosarcoma en personas de 15 años y más", "Politraumatizado Grave", "Prevención de Parto Prematuro", "Prevención secundaria enfermedad renal crónica terminal", "Retinopatía del prematuro", "Retinopatía diabética", "Salud Oral Integral del adulto de 60 años", "Salud oral integral de la embarazada", "Salud oral integral para niños y niñas de 6 años", "Síndrome de Dificultad Respiratoria en el recién nacido", "Síndrome de la inmunodeficiencia adquirida VIH/SIDA", "Trastorno Bipolar en personas de 15 años y más", "Trastornos de generación del impulso y conducción en personas de 15 años y más, que requieren Marcapaso", "Tratamiento Médico en personas de 55 años y más con Artrosis de Cadera y/o Rodilla, leve o moderada", "Tratamiento Quirúrgico de Hernia del Núcleo Pulposo Lumbar", "Tratamiento Quirúrgico de lesiones crónicas de la válvula aórtica en personas de 15 años y más", "Tratamiento Quirúrgico de lesiones crónicas de las válvulas mitral y tricúspide en personas de 15 años y más", "Tratamiento de Erradicación del Helicobacter Pylori", "Tratamiento de Hipoacusia moderada en personas menores de 4 años", "Tratamiento de la hiperplasia benigna de la próstata en personas sintomáticas", "Tratamiento quirúrgico de cataratas", "Tratamiento quirúrgico de escoliosis en personas menores de 25 años", "Trauma Ocular Grave", "Traumatismo Cráneo Encefálico moderado o grave", "Tumores Primarios del Sistema Nervioso Central en personas de 15 años o más", "Urgencia Odontológica Ambulatoria", "Vicios de refracción en personas de 65 años y más" y "Órtesis (o ayudas técnicas) para personas de 65 años y más"

User prompt template ¿"<x>" pertenece a la lista de 80 problemas de salud priorizados por las garantías explícitas de salud?.

A.2 Referral speciality classification

System prompt template Eres un asistente serio que sólo da respuestas precisas y concisas que recibirá diagnósticos en Español y deberás sólo responder con el nombre de la especialidad en Español a la cual debe enviarse el diagnóstico. Las especialidades disponibles son: TRASTORNOS TEMPOROMANDIBULARES Y DOLOR OROFACIAL, REHABILITACION: PROTESIS FIJA, NUTRICION, GENETICA, ODONTOLOGIA INDIFERENCIADO, CIRUGIA TORAX, CIRUGIA INFANTIL, MEDICINA FAMILIAR, NEUROLOGIA, ONCOLOGIA, OBSTETRICIA, CIRUGIA ADULTO, DERMATOLOGIA, GERIATRIA, OTORRINOLARINGOLOGIA, BRONCOPULMONAR, MEDICINA INTERNA, PERIODONCIA, CARDIOLOGIA, OFTALMOLOGIA, REHABILITACION: PROTESIS REMOVIBLE, ENDOCRINOLOGIA, PEDIATRIA, REUMATOLOGIA, CIRUGIA PLASTICA, ORTODONCIA, CIRUGIA DE MAMAS, CIRUGIA PROCTOLOGICA, GASTROENTEROLOGIA, HEMATOLOGIA, UROLOGIA, ANESTESIOLOGIA, ENFERMEDADES DE TRANSMISION SEXUAL, OPERATORIA, NEONATOLOGIA, NEUROCIRUGIA, CIRUGIA VASCULAR PERIFERICA, GINECOLOGIA, CIRUGIA BUCAL, CIRUGIA MAXILO FACIAL, CIRUGIA ABDOMINAL, CARDIOCIRUGIA, PSIQUIATRIA, INFECTOLOGIA, TRAUMATOLOGIA, ENDODONCIA, MEDICINA FISICA Y REHABILITACION, NEFROLOGIA.

User prompt template ¿A qué especialidad debo enviar el diagnóstico "<x>"?.

A.3 Clinical named entity recognition

System prompt template Eres reconocedor de entidades nombradas que solo debe detectar las entidades en la siguiente lista: "disease": "alteracion o desviacion del estado fisiologico en una o varias partes del cuerpo, por causas en general conocidas, manifestada por sintomas y signos caracteristicos, y cuya evolucion es mas o menos previsible", - medication: "Medicamentos o drogas empleadas en el tratamiento y o prevención de enfermedades", - abbreviation: "Abreviatura", - body_part: "Órgano o

una parte anatómica de una persona", - family_member: "Miembro de la familia", - laboratory_or_test_result: "Resultado de laboratorio o test", - clinical_finding: "Observaciones, juicios o evaluaciones que se hacen sobre los pacientes", - diagnostic_procedure: "Exámenes que permiten determinar la condición del individuo ", - laboratory_procedure: "Exámenes que se realizan en diversas muestras de pacientes que permiten diagnosticar enfermedades mediante la detección de biomarcadores y otros parámetros", - therapeutic_procedure: "Actividad o tratamiento que es empleado para prevenir, reparar, eliminar o curar la enfermedad del individuo", Debes responder con el mismo texto de entrada, pero con las entidades nombradas anotadas con etiquetas en la misma línea (<nombre_entidad>lorem ipsum</nombre_entidad>), donde cada etiqueta corresponde a un nombre de entidad, por ejemplo: <entidad>Sed ut perspiciatis</entidad> unde omnis iste natus error sit voluptatem <entidad>accusantium</entidad>.
Las únicas etiquetas disponibles son: medication, abbreviation, body_part, family_member, laboratory_or_test_result, clinical_finding, diagnostic_procedure, laboratory_procedure, therapeutic_procedure, no puedes agregar más etiquetas de las incluidas en esa lista. IMPORTANTE: NO DEBES CAMBIAR EL TEXTO DE ENTRADA, SÓLO AGREGAR LAS ETIQUETAS.