Low-Resource Spanish Clinical Encoders: Architectures, Adaptation, and Evaluation under Computational Constraints

Guillem G. Subies

Abstract

The rise of large-scale language models (LLMs) has revolutionized Natural Language Processing (NLP), but their high computational demands have created significant barriers to entry for researchers operating under limited hardware budgets. This disparity is especially pronounced in specialized domains such as clinical NLP, where language, data, and resource limitations intersect. In this work, we propose a resource-efficient methodology for designing, training, and evaluating compact Spanish clinical encoder models that leverage recent architectural advances, parameter-efficient fine-tuning strategies, and domain-specific adaptation. We present a multi-stage approach that prioritizes reproducibility, open-source compatibility, and computational efficiency. Our models are developed using ClinText-SP, the largest available corpus of Spanish clinical texts, and evaluated against both general-purpose LLMs and specialized encoders across key clinical NLP tasks. The aim is to show that with careful design and judicious use of compute, low-resource encoder models can match or exceed the performance of larger systems, thereby enabling equitable access to domain-specific NLP technologies in under-resourced settings. This thesis contributes both practical tools and critical insights to the evolving field of low-resource clinical NLP in Spanish.

Keywords

Clinical NLP, Spanish Language Models, Low-Resource Learning, Encoder Architectures, Domain Adaptation, Parameter-Efficient Fine-Tuning, Open Science, Benchmarking, Transformer Models

1. Introduction

The rapid advancement of Natural Language Processing (NLP) over the past decade has been driven by two synergistic developments. First, the introduction of the Transformer architecture by Vaswani et al. [1] and its encoder–decoder variants, such as BERT [2] and GPT [3], revolutionized sequence modeling by enabling scalable attention mechanisms. Second, the exponential growth in computational power—largely enabled by NVIDIA's CUDA platform [4]—provided the hardware foundation for training ever-larger models. Together, these innovations precipitated a paradigm shift in AI, yielding breakthroughs not only in NLP but also in related domains such as speech recognition [5] and computer vision [6, 7].

Despite these achievements, the growing reliance on large, resource-intensive generative models has marginalized the development and study of more compact encoder-only architectures. Training and fine-tuning giant models require high-end GPUs and extensive budgets, putting them out of reach for many research institutions [8]. This resource gap is particularly problematic for specialized domains, where data is scarce and domain-specific performance is critical. In this work, we propose to bridge this gap by adapting recent architectural and methodological advances from large language models (LLMs) to the design of efficient, task-specific encoder models. Specifically, the thesis will focus on Spanish clinical encoder models.

Clinical NLP plays a vital role in improving patient care and facilitating medical research by extracting structured information from unstructured clinical texts. Prior studies have demonstrated that domain-adapted models, such as ClinicalBERT [9], significantly outperform general-purpose counterparts on

Doctoral Symposium on Natural Language Processing, 25 September 2025, Zaragoza, Spain.

100500844@alumnos.uc3m.es (G. G. Subies); guillem.garcia@iic.uam.es (G. G. Subies)
0000-0003-0731-4726 (G. G. Subies)
https://www.linkedin.com/in/guillemgsubies (G. G. Subies)

clinical information extraction and decision support tasks. Moreover, addressing underrepresented languages like Spanish is crucial to ensuring equitable access to NLP technologies, as highlighted by the scarcity of high-quality Spanish clinical corpora and models [10, 11]. By focusing on Spanish clinical encoder models, our research aims to reduce language and resource barriers in clinical NLP applications.

In this thesis, we want to describe a methodology for developing low-resource Spanish clinical encoders by leveraging transfer learning from multilingual and general-domain LLMs, employing parameter-efficient fine-tuning techniques, and designing targeted pretraining objectives. We will evaluate our models on a suite of clinical NLP benchmarks. Furthermore, we identify key challenges and open issues for discussion at the Doctoral Symposium, such as the trade-offs between model size, efficiency, and domain specificity.

The remainder of this paper is organized as follows: Section 2 reviews related work in domain-specific and low-resource NLP; Section 3 outlines our thesis proposal and its main hypothesis; Section 4 presents experimental methodology that will be followed and Section 5 highlights the open issues and concludes the paper.

2. Background and related work

Our preliminary Survey [11] highlighted a marked performance gap between general-domain Spanish and multilingual encoder models versus those trained on clinical Spanish texts. State-of-the-art general Spanish models (RigoBERTa [12], BETO [13], and MarIA [14]) and multilingual encoders like DeBERTa [15] and XLM-RoBERTa [16] consistently outperform Spanish Clinical Models such as Galén [17] and bsc-bio-ehr [18]. We attribute this disparity to the scarcity of in-domain data and the challenges of training from scratch under low-resource conditions. A similar—and also conunterintuitive—effect can be observed also in decoder LLMs [19].

To address these issues, we gathered ClinText-SP[1] and adopted RigoBERTa as our base to perform domain adaptation, resulting in RigoBERTa Clinical[2] [20]. ClinText-SP is the largest publicly available Spanish clinical corpus, aggregating texts from medical journals, shared-task annotations, and supplementary sources (e.g., Wikipedia, medical textbooks). It comprises 35,996 documents (average length: ~700 tokens) and totals 25.62 million tokens. The corpus balances long, structured clinical case reports with shorter schematic ones, making it suitable for a range of clinical NLP tasks.

Beyond corpus curation and domain-adaptive pretraining, recent architectural and methodological innovations offer further avenues to enhance clinical encoder models. FlashAttention [21] enables exact, memory-efficient computation of scaled dot-product attention, making it particularly suitable for long-context processing. ModernBERT [22] introduces optimizations that improve training stability and performance over traditional BERT-style encoders. GLiNER [23] presents a generalist named entity recognizer that performs well across domains without domain-specific tuning. Parameter-efficient fine-tuning approaches [24], such as adapter layers, prefix tuning, and in particular Low-Rank Adaptation (LoRA) [25], have demonstrated significant performance gains while drastically reducing the number of trainable parameters. In addition, model compression techniques like knowledge distillation and quantization [26] enable the deployment of lightweight models that retain competitive performance, which is especially important in clinical environments with constrained computational resources. Post-training quantization to 8-bit representations maintains accuracy with minimal overhead [27], whereas quantization-aware training at ternary or binary precision—as in BitNet [28]—achieves competitive results by training low-precision weights from scratch. Finally, Gemma Encoder [29] shows how to systematically convert a decoder-only model into an encoder through architectural tweaks and re-training regimes, which could yield extremely powerful encoder models without the hurdle of pre-training them.

These developments will guide future extensions of our Spanish clinical encoder suite, balancing efficiency, domain specificity, and real-world applicability.

[1] https://huggingface.co/datasets/IIC/ClinText-SP
[2] https://huggingface.co/IIC/RigoBERTa-Clinical

3. Proposal and Main Hypotheses

Building on the observations outlined in Section 2, the central hypothesis of this thesis is:

H1 *Compact, task-specific Spanish clinical encoder models—carefully engineered with recent architectural and tuning advances—can match or exceed the performance of large, general-purpose LLMs on supervised clinical NLP tasks, while greatly reducing computational cost and environmental impact.*

This overarching hypothesis decomposes into four supporting hypotheses:

H2 **Efficiency Hypothesis.** For supervised, domain-specific tasks, lightweight encoder-only models with parameter-efficient fine-tuning deliver comparable performance to full fine-tuning of large LLMs, at a fraction of the compute and energy requirements.

H3 **Domain Adaptation Hypothesis.** Integrating decoder architectural innovations into encoder-only pretraining yields significantly improved representations for clinical Spanish text, narrowing the gap with models trained on massive corpora.

H4 **Generative versus Discriminative Hypothesis.** While generative LLMs excel at open-ended text generation, specialized encoder models can outperform them on focused extraction and classification tasks in the clinical domain.

H5 **Language Equity Hypothesis.** Targeted domain adaptation and low-resource strategies can mitigate the disparity between Spanish clinical models and their English counterparts, providing high-quality tools for Spanish-speaking clinical NLP communities.

To validate these hypotheses, the thesis will pursue the following objectives:

O1 **Model Development.** Design and implement a family of Spanish clinical encoder models that incorporate (i) domain-adaptive pretraining on ClinText-SP, (ii) memory-efficient attention and encoder optimizations, and (iii) parameter-efficient fine-tuning or compression techniques.

O2 **Rigorous Evaluation.** Benchmark the proposed models against state-of-the-art Spanish and multilingual clinical encoders, as well as general-purpose LLMs, across standard clinical NLP tasks (e.g., named entity recognition, relation extraction, document classification).

O3 **Practical Validation.** Demonstrate real-world utility by integrating the best-performing model into at least two clinical use cases to quantify improvements in accuracy, latency, and resource consumption compared to LLM-based solutions.

O4 **Reproducibility and Open Science.** Release all code, trained checkpoints, and evaluation scripts under an open-source license to foster transparency, community adoption, and further research in low-resource clinical NLP.

4. Methodology and Proposed Experiments

Developing high-performance Spanish clinical encoder models under strict hardware constraints demands a rigorous, systematic methodology. In our case, “GPU poverty” is not merely a figure of speech but a lived reality: training large models on limited memory and compute forces careful budgeting of every GPU-hour and careful selection of experiments that yield maximal insight for minimal cost.

Concretely, the computational resources available for this thesis are limited to a small shared pool consisting of one NVIDIA A100 80GB GPU and two NVIDIA RTX 3090 GPUs. The total number of GPU-hours remains uncertain and will vary over time, but individual experiments must be designed to complete within a few hours each. This restricts the size and complexity of the models we can realistically train and evaluate. While 8B-parameter models are at the upper limit of what we can handle, our goal is to obtain strong performance from models with fewer than 1B parameters—an achievable target given that many high-quality encoder models tend to be smaller than their generative counterparts.

Balancing ambition with feasibility, our methodology emphasizes open, efficient, and well-validated techniques at each step, focusing on the maximum return per unit of computation.

4.1. Challenge of Limited Computing Resources

Life on a GPU-poor budget entails long queued jobs, frequent checkpoint pruning, and constant trade-offs between batch size, sequence length, and model complexity. With our computational budget, we cannot indiscriminately pretrain or fine-tune dozens of large variants. Instead, we must:

- **Prioritize efficiency:** Favor models and techniques explicitly designed for memory-efficient attention or parameter-efficient tuning.
- **Exploit transfer learning:** Leverage strong multilingual or general-domain checkpoints (e.g., RigoBERTa) as a starting point, reducing the need for full pretraining.
- **Optimize hyperparameters conservatively:** Use small-scale pilot runs to identify promising configurations before scaling up.

Acknowledging this constraint not only shapes our experimental choices but also reflects the real-world conditions of many academic and clinical research labs.

4.2. Candidate Feature Selection Pipeline

To distinguish genuine advancements from "shiny object syndrome," we will implement a multi-stage filtering process:

1. **Literature and code review:** Survey recent encoder and encoder–decoder innovations (see Section 2), cataloging techniques that claim state-of-the-art gains.
2. **Open-source viability check:** Verify that each candidate is available under a suitable open license and has an actively maintained implementation (e.g., GitHub repo with recent commits and community adoption).
3. **Hardware footprint assessment:** Estimate memory and compute requirements, discarding any feature whose resource demand exceeds our hardware budget.
4. **Empirical sanity check:** For borderline cases, inspect small-scale reproducibility reports or replicate a quick experiment on a toy dataset to confirm baseline efficacy.

This pipeline ensures we only invest scarce GPU cycles in approaches that are both credible and implementable.

4.3. Ablation Study and Incremental Validation

Once a shortlist of viable features is assembled, we perform targeted ablation experiments similar to the onces performed for the training of RigoBERTa Clinical [20]:

- **Controlled pilot runs:** Incorporate each individual feature into the RigoBERTa Clinical baseline and fine-tune on a representative clinical task.
- **Performance versus cost analysis:** Record not only improvement in primary metrics (e.g., F1 score) but also additional GPU-hours and memory usage.

Features that yield meaningful performance gains will be retained; those that fail to clear this bar are discarded.

4.4. Final Model Assembly and Benchmarking

With *n* validated improvements in hand, we assemble the final Spanish clinical encoder:

1. **Integrated training regimen:** Combine all selected architectural tweaks, attention optimizations, and tuning strategies into a unified pretraining and fine-tuning pipeline.
2. **Comprehensive evaluation suite:** Measure performance across multiple clinical tasks and datasets. The benchmarking suite from our Survey [11] will be used.
3. **Comparative analysis:** Benchmark against (i) best-in-class Spanish/multilingual encoders (e.g., BETO, XLM-RoBERTa), and (ii) a closed, general-purpose LLM via API calls (e.g., GPT-style model), tracking accuracy, latency, and cost per query.
4. **Resource and environmental reporting:** Document total GPU-hours, peak memory usage, and estimated carbon footprint savings relative to LLM baseline.

This methodological framework not only tests our central hypotheses under realistic constraints but also produces a reproducible, openly licensed research artifact that can directly inform both academic and industrial clinical NLP deployments.

Open Science Commitment. Aligned with the principles of transparency and reproducibility, we will publish all code, data splits, model checkpoints, and evaluation scripts under permissive open-source licenses. This fully documented release is intended to (i) enable fair comparisons and rapid adoption in Spanish clinical NLP, (ii) support institutions with limited resources, and (iii) foster collaborative improvements by the broader research community.

5. Discussion

This thesis raises several broader issues that merit reflection, both within the scope of our work and in the wider research landscape.

First and foremost is the growing disparity in research accessibility. The current trajectory of NLP favors massive, resource-intensive models maintained by a handful of well-funded organizations. These models often operate as black-box APIs and are prohibitively expensive to replicate or fine-tune in resource-constrained environments. This trend poses a serious challenge to the ideals of Open Science and equitable technological access. Our work is, in part, a reaction to this imbalance—an attempt to demonstrate that it is still possible to perform meaningful, high-quality NLP research with modest resources, provided that methodology and tooling are approached critically and creatively.

Another area of ongoing concern is the gap between academic NLP and real-world clinical needs. While we believe Spanish clinical encoders have the potential to contribute significantly to healthcare settings, forging meaningful collaborations with hospitals and healthcare providers remains difficult. Reaching clinicians, understanding their specific challenges with unstructured data, and aligning our tools with their workflows are non-trivial efforts. This points to the need for more interdisciplinary bridges between NLP research and the healthcare sector—particularly in Spanish-speaking countries, where resource gaps are even more pronounced.

We are also acutely aware that our focus on encoder-only architectures goes somewhat against current trends, which are heavily biased toward decoder-based generative models. These models dominate headlines and benchmarks, but they often come at immense computational and financial cost to operate. In contrast, our belief is that encoder models remain highly competitive for many structured clinical tasks, offering a far more sustainable alternative when used effectively.

Acknowledgments

This work was supported in part by the Instituto de Ingeniería del Conocimiento and Grant PID2023-148577OB-C21 (Human-Centered AI: User-Driven Adapted Language Models) by MICIU/AEI/ 10.13039/501100011033 and by FEDER/UE.

Declaration on Generative AI

During the preparation of this work, the author(s) used IIC/RigoChat-7b-v2 in order to: **Grammar and spelling check**, **Paraphrase and reword**. 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. 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.

[2] 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.

[3] A. Radford, K. Narasimhan, T. Salimans, I. Sutskever, et al., Improving language understanding by generative pre-training, 2018.

[4] J. Nickolls, I. Buck, M. Garland, K. Skadron, Scalable parallel programming with cuda: Is cuda the parallel programming model that application developers have been waiting for?, Queue 6 (2008) 40–53.

[5] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, I. Sutskever, Robust speech recognition via large-scale weak supervision, 2022. URL: https://arxiv.org/abs/2212.04356. arXiv:2212.04356.

[6] A. Ramesh, M. Pavlov, G. Goh, S. Gray, C. Voss, A. Radford, M. Chen, I. Sutskever, Zero-shot text-to-image generation, 2021. URL: https://arxiv.org/abs/2102.12092. arXiv:2102.12092.

[7] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, B. Ommer, High-resolution image synthesis with latent diffusion models, 2022. URL: https://arxiv.org/abs/2112.10752. arXiv:2112.10752.

[8] R. Agerri, E. Agirre, Lessons learned from the evaluation of spanish language models, arXiv preprint arXiv:2212.08390 (2022).

[9] K. Huang, J. Altosaar, R. Ranganath, Clinicalbert: Modeling clinical notes and predicting hospital readmission, 2020. URL: https://arxiv.org/abs/1904.05342. arXiv:1904.05342.

[10] P. Báez, F. Villena, M. Rojas, M. Durán, J. Dunstan, The Chilean waiting list corpus: a new resource for clinical named entity recognition in Spanish, in: A. Rumshisky, K. Roberts, S. Bethard, T. Naumann (Eds.), Proceedings of the 3rd Clinical Natural Language Processing Workshop, Association for Computational Linguistics, Online, 2020, pp. 291–300. URL: https://aclanthology.org/2020.clinicalnlp-1.32/. doi:10.18653/v1/2020.clinicalnlp-1.32.

[11] G. García Subies, Á. Barbero Jiménez, P. Martínez Fernández, A comparative analysis of spanish clinical encoder-based models on ner and classification tasks, Journal of the American Medical Informatics Association 31 (2024) 2137–2146. URL: https://doi.org/10.1093/jamia/ocae054. doi:10.1093/jamia/ocae054. arXiv:https://academic.oup.com/jamia/article-pdf/31/9/2137/58868058/ocae054.pdf.

[12] A. V. Serrano, G. G. Subies, H. M. Zamorano, N. A. Garcia, D. Samy, D. B. Sanchez, A. M. Sandoval, M. G. Nieto, A. B. Jimenez, Rigoberta: A state-of-the-art language model for spanish, 2022. arXiv:2205.10233.

[13] J. Cañete, G. Chaperon, R. Fuentes, J.-H. Ho, H. Kang, J. Pérez, Spanish pre-trained bert model and evaluation data, in: PML4DC at ICLR 2020, 2020.

[14] A. G. Fandiño, J. A. Estapé, M. Pàmies, J. L. Palao, J. S. Ocampo, C. P. Carrino, C. A. Oller, C. R. Penagos, A. G. Agirre, M. Villegas, Maria: Spanish language models, Procesamiento del Lenguaje Natural 68 (2022). URL: https://upcommons.upc.edu/handle/2117/367156#.YyMTB4X9A-0.mendeley. doi:10.26342/2022-68-3.

[15] P. He, J. Gao, W. Chen, Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing, 2021. arXiv:2111.09543.

[16] 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, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Associ-

ation for Computational Linguistics, Online, 2020, pp. 8440–8451. URL: https://www.aclweb.org/anthology/2020.acl-main.747. doi:10.18653/v1/2020.acl-main.747.

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

[18] C. P. Carrino, J. Llop, M. Pàmies, A. Gutiérrez-Fandiño, J. Armengol-Estapé, J. Silveira-Ocampo, A. Valencia, A. Gonzalez-Agirre, M. Villegas, Pretrained biomedical language models for clinical NLP in Spanish, in: Proceedings of the 21st Workshop on Biomedical Language Processing, Association for Computational Linguistics, Dublin, Ireland, 2022, pp. 193–199. URL: https://aclanthology.org/2022.bionlp-1.19. doi:10.18653/v1/2022.bionlp-1.19.

[19] D. P. Jeong, S. Garg, Z. C. Lipton, M. Oberst, Medical adaptation of large language and vision-language models: Are we making progress?, arXiv preprint arXiv:2411.04118 (2024).

[20] G. G. Subies, Álvaro Barbero Jiménez, P. M. Fernández, Clintext-sp and rigoberta clinical: a new set of open resources for spanish clinical nlp, 2025. URL: https://arxiv.org/abs/2503.18594. arXiv:2503.18594.

[21] T. Dao, D. Y. Fu, S. Ermon, A. Rudra, C. Ré, Flashattention: Fast and memory-efficient exact attention with io-awareness, 2022. URL: https://arxiv.org/abs/2205.14135. arXiv:2205.14135.

[22] B. Warner, A. Chaffin, B. Clavié, O. Weller, O. Hallström, S. Taghadouini, A. Gallagher, R. Biswas, F. Ladhak, T. Aarsen, N. Cooper, G. Adams, J. Howard, I. Poli, Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference, 2024. URL: https://arxiv.org/abs/2412.13663. arXiv:2412.13663.

[23] U. Zaratiana, N. Tomeh, P. Holat, T. Charnois, Gliner: Generalist model for named entity recognition using bidirectional transformer, 2023. URL: https://arxiv.org/abs/2311.08526. arXiv:2311.08526.

[24] Z. Han, C. Gao, J. Liu, J. Zhang, S. Q. Zhang, Parameter-efficient fine-tuning for large models: A comprehensive survey, 2024. URL: https://arxiv.org/abs/2403.14608. arXiv:2403.14608.

[25] 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.

[26] V. Sanh, L. Debut, J. Chaumond, T. Wolf, Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter, 2020. URL: https://arxiv.org/abs/1910.01108. arXiv:1910.01108.

[27] B. Jacob, S. Kligys, B. Chen, M. Zhu, M. Tang, A. Howard, H. Adam, D. Kalenichenko, Quantization and training of neural networks for efficient integer-arithmetic-only inference, 2017. URL: https://arxiv.org/abs/1712.05877. arXiv:1712.05877.

[28] H. Wang, S. Ma, L. Dong, S. Huang, H. Wang, L. Ma, F. Yang, R. Wang, Y. Wu, F. Wei, Bitnet: Scaling 1-bit transformers for large language models, 2023. URL: https://arxiv.org/abs/2310.11453. arXiv:2310.11453.

[29] P. Suganthan, F. Moiseev, L. Yan, J. Wu, J. Ni, J. Han, I. Zitouni, E. Alfonseca, X. Wang, Z. Dong, Adapting decoder-based language models for diverse encoder downstream tasks, arXiv preprint arXiv:2503.02656 (2025).