Continual Pretraining of a Small Language Model on Cuban Spanish Corpora

No Institute Given

Abstract. Large language models have transformed natural language processing, but their effectiveness is limited for regional language variants due to underrepresentation in training data. This paper introduces Cecilia 2B, a 2-billion-parameter language model continually pretrained on nearly 1 billion tokens of Cuban Spanish text, including newspapers, encyclopedias, legal documents, literature, and song lyrics, to address the gap in language technology for Cuban Spanish. Leveraging the Salamandra 2B architecture, Cecilia 2B demonstrates the feasibility and value of adapting multilingual models to regional variants through continual pretraining, achieving improved performance on Spanish and multilingual tasks relevant to its target domain while maintaining computational efficiency suitable for resource-constrained environments. We detail the construction of a culturally rich Cuban Spanish corpus, the adaptation methodology, and a comparative evaluation with the base model, highlighting both the benefits and trade-offs of regional specialization. Cecilia 2B provides a foundational resource for Cuban Spanish NLP and establishes a path for future research in instruction tuning, corpus expansion, tokenizer retraining, and the development of larger or more specialized models.

Keywords: Language Modeling · Natural Language Processing · Corpora.

1 Introduction

Large Languages Models have revolutionized natural language processing, with an unprecedented capacity to capture important semantic and pragmatic aspects of written language. However, even though large, open-weight models such as Llama and Mistral are trained on a majority of mainstream languages, they often underperform in regional varieties of underrepresented languages. The development of domain-specific and regional language models has thus become increasingly important as large, general-purpose models often fail to capture the linguistic, cultural, and contextual nuances required for authentic communication within specific communities. Specialized models in healthcare, law, and finance outperform general models on tasks requiring domain expertise, and culturally adapted models such as CultureLLM demonstrate improved handling of language-specific phenomena [1].

In the Spanish language context, the Salamandra project [2] stands out as a family of open-source, multilingual language models designed with a strong emphasis on Spanish and co-official languages, providing a robust foundation for further adaptation to regional variants. Salamandra's architecture and training methodology make it especially well-suited for continual pretraining on regionally focused corpora, enabling the creation of models that internalize the unique features of local Spanish varieties while maintaining broad language capabilities.

Building on this foundation, the **CecilIA**[1] project aims to address the lack of high-quality language models for Cuban Spanish, a variant with distinctive lexical, syntactic, and cultural characteristics that are not adequately represented in existing models.

The Cuban Spanish variant is distinguished by unique lexical items, commonly known as *cubanismos* (e.g., "asere", "guagua", "yuma") as well as distinct phonetic and morphological patterns that influence written expression. Furthermore, the variant is rich with culturally-specific idiomatic expressions and historical references that are essential for contextual understanding. While a language model cannot capture spoken phonetics, its performance is deeply tied to its ability to process these lexical and semantic nuances. By curating a corpus from sources like local encyclopedias of *cubanismos*, national literature, and song lyrics, we aim to expose the model to these features, thereby improving its ability to generate authentic and context-aware text for this specific linguistic community.

This paper introduces *Cecilia 2B*, a 2-billion-parameter language model continually pretrained on a newly constructed corpus of approximately 1 billion tokens of Cuban written text, including newspapers, encyclopedias, legal documents, literature, and song lyrics. By adapting the powerful Salamandra 2B architecture, this work delivers the first large-scale, publicly-available foundational model specifically for Cuban Spanish. Cecilia 2B serves as a critical resource for natural language processing applications in this underrepresented variant and provides a detailed case study on the benefits and trade-offs inherent in regional language model adaptation.

Cecilia 2B is the first iteration of a larger project aimed at creating pretrained and fine-tuned language models in the Cuban Spanish variant for several model sizes, architectures, and domains. By focusing first on a relatively small model size of 2 billion parameters, Cecilia 2B balances computational efficiency with linguistic specialization, enabling deployment in resource-constrained environments common in Cuba and similar settings. This approach allows us to explore the optimal strategies for creating this type of resources, as well as facilitating broader accessibility and practical usage from the beginning of the project. The experienced obtained in this iteration of Cecilia will directly inform the development of future, larger models.

[1] https://cecilia.uhgia.org

This paper presents the design, training methodology, dataset composition, and partial evaluation of Cecilia 2B. The remainder of this paper is organized as follows: Section 2 reviews related work on small language models for regional variants, continual pretraining, and the Salamandra project. Section 3 details the design and training methodology of Cecilia 2B, including corpus construction and adaptation procedures. Section 4 presents evaluation results comparing Cecilia 2B to its base model on a suite of multilingual and Spanish NLP benchmarks. Section 5 discusses the implications, current limitations, and future directions for regional language model development. Finally, Section 6 concludes by summarizing the main contributions and outlining the potential of Cecilia for advancing Cuban Spanish NLP.

2 Background and Related Works

This section presents a review of the relevant literature in the field of language modeling, with a particular focus on small language models and their applications in regional language variants, as well as techniques for domain and linguistic adaptation. The section finalizes with a short presentation of the Salamandra models, architecture, and training procedure.

2.1 Small Language Models for Regional Variants

The development of small language models (SLMs) specifically tailored to regional language variants has emerged as a significant research direction in natural language processing. Unlike their larger counterparts, SLMs typically contain millions to a few billion parameters and are designed to operate efficiently on resource-constrained environments while maintaining competitive performance for specialized domains. These models represent a strategic response to the limitations of general-purpose large language models, which often fail to capture regional linguistic nuances, cultural contexts, and domain-specific knowledge essential for authentic communication within specific communities.

Recent research has demonstrated the effectiveness of SLMs in processing regional languages with significantly fewer parameters than traditional large language models. The Regional Tiny Stories framework exemplifies this approach, showing that models with 1-10 million parameters can produce coherent outputs when trained on language-specific datasets [3]. This work expanded the TinyStories [4] methodology to Indian languages including Hindi, Marathi, and Bengali, revealing that language-specific tokenizers consistently outperform general-purpose alternatives for regional languages.

The development of regionalized Spanish language models has gained particular attention, with projects creating word embeddings and BERT-based models trained on Twitter data from 26 Spanish-speaking countries [5]. These efforts have resulted in resources that capture lexical and semantic variations across different Spanish-speaking regions, demonstrating measurable improvements in

regional task performance. Similarly, the DADA (Dialect Adaptation via Dynamic Aggregation) framework has shown promise for adapting models to various English dialects through compositional adapter architectures that handle specific linguistic features [6].

These efforts are part of a growing focus within NLP on computational sociolinguistics, which seeks to create models that are not only domain-specific but also dialect-aware and sensitive to social contexts [7]. Research in this area explores more advanced techniques, such as dialect-aware tokenization and methods for modeling code-switching, which will inform future iterations of the Cecilia project [8].

2.2 Continual Pretraining and Domain Adaptation Techniques

Continual pretraining has emerged as a fundamental technique for adapting existing language models to new domains and regional variants while preserving previously acquired knowledge [9]. This approach aims to mitigate catastrophic forgetting—the tendency of a model to lose previously learned knowledge when trained on a new task—while simultaneously achieving knowledge transfer to improve end-task performance [10]. Research has shown that continual pretraining consistently improves models smaller than 1.5 billion parameters and demonstrates superior performance compared to traditional domain adaptation methods [9].

The efficacy of continual pretraining varies significantly based on model size and domain progression [11]. Research indicates that smaller models are particularly sensitive to continual pretraining, showing the most significant rates of both learning and forgetting [12]. Domain similarity plays a crucial role in knowledge transfer effectiveness, with semantically similar domain sequences enabling better specialization, while randomized training domains lead to improved transfer and final performance. Cross-lingual and progressive transfer learning approaches have demonstrated the ability to save up to 80% of training costs compared to random initialization when transferring models between languages [13].

2.3 The Salamandra Project

The Salamandra project, developed by the Barcelona Supercomputing Center's Language Technologies Unit, represents a comprehensive effort to create open-source, multilingual language models with particular emphasis on European languages and Spanish language variants. The project encompasses three model sizes—2B, 7B, and 40B parameters—each designed to balance computational efficiency with linguistic capability while maintaining strong performance across multiple languages [2].

Salamandra models employ a standard decoder-only Transformer architecture with several key optimizations that distinguish them from the original Transformer design. The architecture eliminates all bias terms to improve training

stability, incorporates rotary positional embeddings (RoPE) with a base frequency of 10,000 as an alternative to absolute positional embeddings, and replaces ReLU activation with SwiGLU for enhanced performance. The models utilize RMSNorm [14] instead of traditional layer normalization, with an epsilon hyperparameter set to 1e-5, and employ BFloat16 numerical precision for training stability.

Salamandra 2B, the base model for Cecilia, comprises approximately 2.25 billion parameters distributed across 24 layers with a hidden size of 2,048 and 16 attention heads. The model supports a context window of 8,192 tokens and utilizes a vocabulary size of 256,000 tokens, enabling effective processing of diverse multilingual inputs. Unlike the larger variants, Salamandra 2B relies on multi-head attention rather than grouped-query attention, reflecting optimization choices for the smaller parameter count.

The Salamandra pretraining corpus is a comprehensive multilingual datasets specifically designed for European languages, comprising text in 35 European languages and 92 programming languages. The training process utilized approximately 7.8 trillion tokens for the 2B model, with the corpus carefully curated to oversample Spanish and co-official languages of Spain (Catalan, Galician, and Basque) by a factor of two, while downsampling code and English data to achieve balanced representation [2].

A clear indicator of the growing interest in language technologies for regional and underrepresented languages, especifically in the Iberoamerican research community, is the widespread adoption of models developed within the ILENIA network[2]. Notable projects derived from Salamandra include AITANA, optimized for Valencian; Latxa, the first major Basque model; and Carballo, a foundational Galician model.

3 Design and Training of Cecilia 2B

Cecilia 2B, in its current iteration, is a 2-epoch continual pretraining checkpoint of Salamandra 2B. Salamandra 2B was chosen as the base model for Cecilia 2B due to its strong multilingual capabilities, efficient architecture, and open-source availability under an Apache 2.0 license, which facilitates fine-tuning and adaptation for specific language varieties. Its design balances model capacity and computational resource requirements, making it suitable for deployment in resource-constrained environments typical of Cuban NLP applications.

For this initial iteration, we made the deliberate decision to leave the Salamandra 2B architecture, including its original tokenizer and vocabulary, unmodified. This choice was a key element of our research design, intended to isolate the effects of continual pretraining on the new corpus. By keeping the tokenizer constant, we can attribute observed changes in performance directly to the data rather than to a combination of data and vocabulary changes.

[2] https://proyectoilenia.es/

We acknowledge this creates a known limitation: words specific to Cuban Spanish (cubanisms) that are out-of-vocabulary for the base model are tokenized into sub-optimal subword units, making them more difficult for the model to learn effectively. Consequently, developing a new tokenizer trained on our Cuban Spanish corpus and retraining the model with it remains a top priority for future work, as stated in Section 5.2.

3.1 Training Data

The training corpus for Cecilia 2B comprises approximately 1 billion tokens of Cuban Spanish text, including digitized Cuban newspapers from the last decade, the Cuban Encyclopedia, a comprehensive collection of Cuban laws, hundreds of literary works by Cuban authors, local encyclopedias documenting Cubanisms, and song lyrics from prominent Cuban artists. This diverse dataset was curated to capture the linguistic and cultural richness of Cuban Spanish.

All data was collected via web scraping under a fair use assumption and is intended solely for academic and research purposes. The web scraping process utilized custom scripts to systematically gather and parse text from targeted sources. Following collection, the raw text underwent a rigorous cleaning pipeline to ensure quality. This process included (1) removing HTML boilerplate, navigation menus, and advertisements; (2) deduplicating documents at the paragraph level to reduce redundancy; and (3) normalizing text by correcting common OCR errors and standardizing character encoding. This multi-step procedure was crucial for creating a clean, high-quality dataset suitable for language model pretraining.

To respect copyright and intellectual property rights, the raw training data is not publicly available at the moment. Table 1 presents the composition of the full corpus, showing the main sources of texts and their relative percentage within the total dataset.

Table 1: Composition of the training corpus for Cecilia 2B. {tbl-colwidths='[40,50,10]'}

Source Content	Description	(%)
Cuban Encyclopedia (Ecured)	Full snapshot of the online collaborative encyclopedia.	65.0
Cuban Newspapers (2014-2024)	Digital archives of national newspapers *Granma* and *Juventud Rebelde*.	20.0
Cuban Literature	Over 400 digitized works from key authors, e.g., *José Martí*, *Alejo Carpentier*, *José Lezama Lima*.	9.5
Official Gazette of Cuba	Comprehensive collection of laws and official government documents.	1.5

Source Content	Description	(%)
Cultural & Linguistic Dictionaries	E.g., *Diccionario de Cubanismos*, encyclopedias of Afro-Cuban culture.	0.6
Miscellaneous Cuban Texts	Academic theses, historical documents, and other curated texts.	3.4

The Cecilia 2B training corpus is extensive, as shown in Table 2, comprising nearly 300,000 text files with a total of approximately 2.6 billion characters and an estimated 385 million words. This large volume of data ensures comprehensive linguistic coverage, enabling the model to learn a wide range of lexical and syntactic patterns specific to Cuban Spanish.

Table 2: Basic corpus statistics.

Metric	Value
Total Files	296,311
Total Characters	2,631,691,355
Total Words	384,963,687
Total Lines	34,505,341
Average Document Length	8,881 characters
Average Sentence Length	17.0 words
Lexical Density	6.8 characters/word

The average document length of 8,881 characters indicates the dataset includes a balanced mix of short and long texts, which is beneficial for training a model capable of understanding various discourse structures, from brief statements to extended narratives. An average sentence length of 17 words reflects moderately complex sentence constructions typical of formal written language, supporting the model's ability to handle nuanced linguistic phenomena.

The lexical density of 6.8 characters per word suggests a rich vocabulary with a diversity of word lengths, which contributes to the model's capacity to represent the Cuban Spanish lexicon effectively. Overall, these statistics demonstrate that the dataset provides a robust foundation for continual pretraining, enabling Cecilia 2B to internalize the distinctive linguistic and cultural characteristics of Cuban Spanish.

Table 3: Tokenized corpus metrics.

Metric	Value
Total Samples	1,104,532
Total Tokens (no padding)	982,024,795

Metric	Value
Total Tokens (with padding)	1,131,040,768
Average Sequence Length (no padding)	889.3 tokens
Padding Ratio	13.2%

After tokenization, as shown in Table 3, the dataset consists of over 1.1 million samples, with nearly one billion tokens excluding padding. The average sequence length is approximately 889 tokens, with sequences ranging from a single token up to the maximum context window size of 1024 tokens. The padding ratio of 13.2% indicates that a moderate portion of sequences required padding to reach the fixed length, which is typical for datasets with variable-length texts. The data was segmented into 959,008 context windows, each containing 1024 tokens, enabling the model to process long-range dependencies effectively during training.

3.2 Training Procedure

The training of Cecilia 2B was conducted over two full epochs with a batch size of 4, combined with gradient accumulation over 16 steps to effectively simulate a larger batch size of 64. This approach balances the constraints of available GPU memory with the need for stable gradient estimates during optimization. Gradient clipping with a maximum norm of 1.0 was applied to prevent exploding gradients and improve training stability.

Optimization was performed using the AdamW optimizer with a learning rate of 2e-5, incorporating weight decay of 0.01 to regularize the model and reduce overfitting. The learning rate followed a warmup linear decay schedule, with a warmup phase covering 6% of the total training steps, allowing the model to gradually adapt to the data before reaching the peak learning rate. The AdamW hyperparameters beta1 and beta2 were set to 0.9 and 0.999, respectively, consistent with best practices for transformer training.

Mixed precision training using bfloat16 (bf16) precision was employed to accelerate computation and reduce memory consumption without sacrificing numerical stability. The training leveraged Fully Sharded Data Parallel (FSDP) parallelization with full sharding and sharded state dictionaries to optimize memory usage across multiple GPUs. Gradient checkpointing was enabled to further reduce memory footprint by trading compute for storage during backpropagation.

Validation was performed both after each epoch and periodically every 640 training steps, ensuring continuous monitoring of model performance and early detection of potential overfitting or training instability. Overall, these design choices reflect a careful balance between computational efficiency, training stability, and effective convergence on the specialized Cuban Spanish corpus, enabling Cecilia 2B to internalize linguistic nuances while operating within the constraints of available hardware resources.

Training was conducted over approximately 48 hours on a high-performance compute setup consisting of 2 NVIDIA A100 GPUs (40 GB each), an AMD EPYC CPU with 128 cores and 256 threads, and 1 TB of RAM.

Table 4 summarizes the training hyperparameters used for Cecilia 2B.

Table 4: Training Hyperparameters.

Parameter	Value
Number of epochs	2
Batch size	4
Gradient accumulation steps	16
Effective batch size	64
Learning rate	2e-5
Learning rate scheduler	Warmup linear
Warmup proportion	6%
Optimizer	AdamW
Weight decay	0.01
Beta1, Beta2	0.9, 0.999
Gradient clipping norm	1.0
Precision	bfloat16

3.3 Model and Data Availability

In line with ethical research practices and copyright law, the raw training corpus cannot be publicly released. The dataset was compiled from copyrighted sources under a fair use assumption for non-commercial academic research, and its redistribution is prohibited.

To promote responsible open science, the Cecilia 2B model is made available to the research community through a gated access protocol on the Hugging Face platform[3]. This staged-release approach allows for case-by-case evaluation of research requests, mitigating potential misuse while the model undergoes further safety and bias analysis. We are committed to a full public release under a permissive, commercially-viable license once these evaluations are complete. This strategy ensures that the research community can benefit from this work while upholding ethical standards.

4 Evaluation

[3] https://huggingface.co/gia-uh/cecilia-2b-v0.1

Table 5: Evaluation results in selected NLP tasks in English and Spanish, in comparison with Salamandra 2B.

Task	Metric	Salamandra	Cecilia	Rel. Err.
arc_challenge	acc	0.37031	0.38225	3.13%
arc_easy	acc	0.72264	0.73401	1.55%
belebele_en	acc	0.21556	0.24778	13.00%
belebele_es	acc	0.22778	0.24444	6.82%
escola	acc	0.59259	0.55461	-6.41%
openbookqa	acc	0.30000	0.28200	-6.00%
openbookqa_es	acc	0.30800	0.29400	-4.55%
paws_en	acc	0.56100	0.57350	2.18%
paws_es	acc	0.56050	0.55550	-0.89%
piqa	acc	0.73721	0.73667	-0.07%
social_iqa	acc	0.45394	0.44626	-1.69%
teca	acc	0.46481	0.43174	-7.11%
wnli	acc	0.46479	0.42254	-9.09%
wnli_es	acc	0.56338	0.59155	4.76%
xnli_en	acc	0.46225	0.47671	3.03%
xstorycloze_en	acc	0.71145	0.70483	-0.93%
xstorycloze_es	acc	0.65255	0.65189	-0.10%
arc_challenge	acc_norm	0.40700	0.41809	2.65%
arc_easy	acc_norm	0.72559	0.73990	1.93%
belebele_en	acc_norm	0.21556	0.24778	13.00%
belebele_es	acc_norm	0.22778	0.24444	6.82%
openbookqa	acc_norm	0.39600	0.40000	1.00%
openbookqa_es	acc_norm	0.40800	0.40400	-0.98%
piqa	acc_norm	0.74701	0.74701	0.00%
cocoteros_es	bleu	8.46507	6.72269	-20.58%
xlsum_es	bleu	0.80082	0.59723	-25.42%
triviaqa	exact_match	0.37595	0.35432	-5.75%
xquad_es	exact_match	0.37731	0.36050	-4.45%
xquad_es	f1	0.58413	0.56911	-2.57%
cocoteros_es	rouge1	0.33887	0.31209	-7.90%
xlsum_es	rouge1	0.13464	0.08705	-35.35%
Mean Diff				**-2.43%**

Evaluation of Cecilia 2B is still ongoing. At this stage, we present partial results focused on comparing Cecilia 2B to its base model, Salamandra 2B, across a broad suite of standard NLP benchmarks. These tasks include multiple-choice question answering, reading comprehension, paraphrase identification, natural language inference, summarization, translation, and open-domain question an-

swering, in both English and Spanish. Table 5 summarizes the results of this comparison.

The results present a nuanced picture of the trade-offs involved in regional specialization. Cecilia 2B shows clear improvements in multilingual reading comprehension (BELEBELE [15] +6.82%) and natural language inference (WNLI [16] +4.73%), suggesting the continual pretraining successfully enhanced its understanding of Spanish-language nuances as intended.

Conversely, the most significant performance decreases occurred in generation-focused tasks, particularly summarization. For instance, the model's ROUGE-L score dropped by 35.35% on XLSUM-ES [17] and its BLEU score fell by 20.58% on COCOTEROS. It also shows decreasing scores in instruction following tasks like question answering (XQuAD [18], OpenBook [19], and TriviaQA [20]).

This outcome is a direct and anticipated consequence of our training strategy. The pretraining corpus, dominated by literary, legal, and encyclopedic texts, is stylistically divergent from the news articles that comprise benchmarks like XLSUM. Likewise, the instruction following tasks are not represented in the training data, since, as explained before, Cecilia 2B has not been fine-tuned for instruction following or for downstream tasks.

By specializing on long-form, narrative text, the model's proficiency in the concise, extractive style required for summarization was diminished. This highlights a critical trade-off: gaining in-domain cultural and linguistic specialization came at the cost of performance on out-of-domain, general-purpose tasks for which the model has not been fine-tuned [10].

It is important to emphasize that these benchmarks are general-purpose and not specifically tailored to the Cuban Spanish variant for which Cecilia is intended. The observed average difference is a modest decrease of about 2.4% relative to Salamandra 2B across all tasks, with the largest drops in summarization and translation. This is consistent with expectations, as the model has not yet been fine-tuned for instruction following or for downstream tasks.

Thus, at the time of writing no results are available on downstream tasks that target the unique linguistic and cultural phenomena of Cuban Spanish, which is the primary motivation for Cecilia's development. As Cecilia 2B is presently only pretrained and has not undergone instruction tuning or task-specific fine-tuning, comprehensive evaluations on downstream tasks such as question answering, dialogue generation, or other domain-specific applications remain pending. These more specialized assessments will be addressed in future work, following the development of an instruction-tuned version of Cecilia that can better support interactive and task-oriented use cases.

5 Discussion

The evaluation results presented in Section 4 provide a practical illustration of the challenges associated with catastrophic forgetting in regional model adaptation. While Cecilia 2B demonstrated gains in Spanish-language understanding tasks, the corresponding drop in performance on out-of-domain tasks like summarization shows that specialization involves a tangible trade-off. As such, Cecilia 2B remains a work in progress and is currently most suitable for research purposes.

As the model has not yet been fine-tuned for instruction following or specific downstream tasks, its direct applicability in production environments or interactive applications is limited at this stage. However, its foundational capabilities as a Cuban Spanish-pretrained language model open promising avenues for future development.

Once fine-tuned, Cecilia's relatively small size, combined with its specialized training on Cuban Spanish, positions it as a valuable resource for a range of natural language processing tasks tailored to this linguistic variant. Potential use cases include text generation that respects Cuban cultural and linguistic nuances, sentiment analysis for Cuban social media and news, named entity recognition in local contexts, machine translation with improved handling of Cubanisms, and domain-specific question answering.

5.1 Current Limitations

Currently, the model is not quantized and requires approximately 4.5 GB of GPU memory for full loading and inference, which may exceed the hardware capabilities of smaller research teams or institutions with limited computational resources. To address this, quantized versions of Cecilia 2B are planned for release in the near future, which will significantly reduce memory requirements and enable broader accessibility and deployment on more modest hardware setups. This will facilitate wider adoption and experimentation within the Cuban and broader Spanish-speaking NLP research communities.

As with all large language models, Cecilia 2B is susceptible to issues such as biases and hallucinations. The model has not yet undergone comprehensive evaluation to determine the extent to which these problems persist or whether they are exacerbated relative to the original Salamandra 2B base model. Users should be aware that outputs may reflect unintended biases present in the training data or generate factually incorrect or misleading information.

As detailed in Section 3.3, the training corpus contains copyrighted materials and the model is currently available under a gated access policy to ensure responsible use. In due course, Cecilia 2B will be publicly released under a permissive license that allows broad use, including commercial applications, once further evaluations and refinements have been completed to ensure safety and reliability.

One additional limitation is the lack of stakeholder feedback available for this first phase of the project. The contribution of the academic community beyond the computational sciences, including linguistics, history, sociology, cultural studies, etc., is a crucial next step to ensure that the model reflects the needs and expectations of the Cuban people.

5.2 Future Work

Future efforts will focus initially on further curating and expanding the Cuban Spanish corpus that underpins Cecilia 2B. Enhancing the dataset's breadth and diversity will improve the model's linguistic coverage and cultural representation, strengthening its foundation for downstream tasks.

For this particular model, the next key step is to fine-tune Cecilia 2B on general instruction-following tasks to enable more interactive and versatile applications. Subsequently, targeted fine-tuning on specific downstream Cuban Spanish NLP tasks—such as question answering, sentiment analysis, and named entity recognition—will be pursued to maximize its practical utility within the language processing domain.

In parallel, we plan to develop increasingly powerful models by leveraging larger versions of the Salamandra architecture or exploring alternative base models that demonstrate strong performance and suitability for Cuban Spanish. These efforts aim to balance model capacity, efficiency, and cultural specificity, ultimately providing the community with a range of high-quality language models tailored to Cuban Spanish and related linguistic variants.

One specific task that remains challenging is to retrain the tokenizer to better capture cubanisms and other terms that are split into distinct tokens by the Salamandra 2B tokenizer [21]. Additionally, quantized versions of all Cecilia models will be published to enable efficient inference in production environments.

At the moment of writing an ongoing effort to create an instruction fine-tuning corpus is being conducted to enable more interactive and versatile applications. In this process, several collaborators from the larger academic community including linguists, sociologists, historians, artists, lawyers, etc., are providing active feedback to incorporate as much relevant data as possible whithin the design constraints of Cuban-only text. In any case, we acknowledge the necessity of incorporating more voices and perspectives to ensure the model's development aligns with the diverse needs and values. This is a major direction for future work.

6 Conclusions

This paper introduced Cecilia 2B, a 2-billion-parameter language model continually pretrained on a diverse Cuban Spanish corpus of nearly 1 billion tokens. By adapting the Salamandra 2B architecture to focus on Cuban linguistic and

cultural features, this project delivers a vital, publicly-available foundational resource that addresses the critical gap in language technology for this underrepresented variant. The development of Cecilia 2B provides a robust case study on balancing computational efficiency with linguistic specialization and establishes a clear path for future work, including instruction tuning, corpus expansion, and the development of more advanced models tailored to Cuban Spanish.

References

[1] C. Li, M. Chen, J. Wang, S. Sitaram, and X. Xie, "Culturellm: Incorporating cultural differences into large language models," *Advances in Neural Information Processing Systems*, vol. 37, pp. 84799–84838, 2024.

[2] A. Gonzalez-Agirre *et al.*, "Salamandra technical report," *arXiv preprint arXiv:2502.08489*, 2025.

[3] N. Patil *et al.*, "Regional tiny stories: Using small models to compare language learning and tokenizer performance," *arXiv preprint arXiv:2504.07989*, 2025.

[4] Y. Li and R. Eldan, "TinyStories: How small can language models be and still speak coherent english," 2023.

[5] E. S. Tellez, D. Moctezuma, S. Miranda, M. Graff, and G. Ruiz, "Regionalized models for spanish language variations based on twitter," *Language Resources and Evaluation*, vol. 57, no. 4, pp. 1697–1727, 2023.

[6] Y. Liu, W. Held, and D. Yang, "Dada: Dialect adaptation via dynamic aggregation of linguistic rules," *arXiv preprint arXiv:2305.13406*, 2023.

[7] D. Hovy, "The social and the neural network: How to make natural language processing about people again," in *Proceedings of the second workshop on computational modeling of people's opinions, personality, and emotions in social media*, 2018, pp. 42–49.

[8] D. Nguyen, L. Rosseel, and J. Grieve, "On learning and representing social meaning in NLP: A sociolinguistic perspective," in *Proceedings of the 2021 conference of the north american chapter of the association for computational linguistics: Human language technologies*, 2021, pp. 603–612.

[9] S. Gururangan *et al.*, "Don't stop pretraining: Adapt language models to domains and tasks," *arXiv preprint arXiv:2004.10964*, 2020.

[10] I. J. Goodfellow, M. Mirza, D. Xiao, A. Courville, and Y. Bengio, "An empirical investigation of catastrophic forgetting in gradient-based neural networks," *arXiv preprint arXiv:1312.6211*, 2013.

[11] Ç. Yıldız, N. K. Ravichandran, N. Sharma, M. Bethge, and B. Ermis, "Investigating continual pretraining in large language models: Insights and implications," *arXiv preprint arXiv:2402.17400*, 2024.

[12] E. Lee, "The impact of model size on catastrophic forgetting in online continual learning," *arXiv preprint arXiv:2407.00176*, 2024.

[13] V. Šliogeris, P. Daniušis, and A. Nakvosas, "Full-parameter continual pretraining of Gemma2: Insights into fluency and domain knowledge," *arXiv preprint arXiv:2505.05946*, 2025.

[14] B. Zhang and R. Sennrich, "Root mean square layer normalization," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[15] K. Ahuja, T. Adewumi, D. I. Adelani, *et al.*, "Belebele: A novel multilingual reading comprehension dataset for low-resource languages," in *arXiv preprint arXiv:2307.09641*, 2023.

[16] H. Levesque, E. Davis, and L. Morgenstern, "The winograd schema challenge," in *Proceedings of the thirteenth international conference on principles of knowledge representation and reasoning*, 2012, pp. 552–561.

[17] T. Hasan *et al.*, "XL-sum: Large-scale multilingual abstractive summarization for 44 languages," *arXiv preprint arXiv:2106.13822*, 2021.

[18] M. Artetxe, S. Ruder, and D. Yogatama, "XQuAD: A cross-lingual question answering dataset," in *Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP)*, 2020, pp. 8658–8663.

[19] T. Mihaylov, P. Clark, T. Khot, and A. Sabharwal, "OpenBookQA: A new benchmark for open book question answering," in *Proceedings of the 2018 conference on empirical methods in natural language processing*, 2018, pp. 2381–2391.

[20] M. Joshi, E. Choi, D. Weld, and L. Zettlemoyer, "TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension," in *Proceedings of the 55th annual meeting of the association for computational linguistics (volume 1: Long papers)*, 2017, pp. 1601–1611.

[21] P. Rust, J. Pfeiffer, I. Vulić, I. Gurevych, and S. Ruder, "How good is your tokenizer? On the monolingual performance of multilingual language models," in *Proceedings of the 59th annual meeting of the association for computational linguistics (ACL)*, 2021, pp. 3118–3135.