Cecilia: The Cuban Language Model

Technical Report

Ernesto L. Estevanell Daniel A. Valdés Roberto Marti
Deborah Famadas Roberto García Gabriel Hernández
Elena Rodríguez Niley González Alejandro Beltrán
Juan Pablo Consuegra Suilan Estévez Alejandro Piad
Yudivián Almeida Robiert Sepúlveda Yoan Gutiérrez
Rafael Muñoz Andrés Montoyo Manuel Palomar

2025-05-28

Cecilia 2B is a 2-billion-parameter language model continually pretrained on a large, diverse corpus of Cuban Spanish text to capture the unique linguistic and cultural features of Cuban Spanish. Built on the Salamandra 2B architecture, Cecilia 2B adapts a robust multilingual base model through continual pretraining on approximately 1 billion tokens from Cuban newspapers, encyclopedias, legal documents, literature, and song lyrics. This approach enables efficient deployment in resource-constrained environments and provides a foundational resource for Cuban Spanish NLP tasks such as text generation, sentiment analysis, and named entity recognition. This report details the model's design, training methodology, dataset, and potential applications, highlighting its significance for Cuban Spanish language technology and future research directions.

Cecilia-2B-v0.1 (hereafter Cecilia 2B) is a compact language model continual pretrained on a diverse and extensive corpus of Cuban written text, designed to capture the unique linguistic, cultural, and social nuances of Cuban Spanish.

The motivation behind Cecilia stems from the need to develop language technologies that accurately reflect regional language variations and cultural contexts, which are often underrepresented or inadequately modeled by large, generic language models. Cuban Spanish exhibits distinct lexical, syntactic, and pragmatic features, as well as culturally specific references, that necessitate specialized modeling to improve natural language processing (NLP) performance on Cuban-specific tasks.

Cecilia 2B is the first iteration of what the authors expect to be a comprehensive project aimed at creating pretrained and fine-tunes 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.

The Cecilia 2B model is based on the Salamandra 2B architecture and was continual pretrained for two full epochs on a private corpus comprising approximately 1 billion tokens, including Cuban newspapers spanning a decade, the Cuban Online Encyclopedia, a comprehensive collection of Cuban laws, hundreds of Cuban literary works, local encyclopedias documenting Cubanisms, and song lyrics from prominent Cuban artists. This varied and culturally grounded dataset aims to guarantee Cecilia 2B internalizes both language patterns and cultural knowledge essential for Cuban Spanish NLP applications.

Cecilia 2B is aimed at a range of NLP tasks such as text generation, sentiment analysis, named entity recognition, and machine translation, all tailored to Cuban Spanish. The model's development reflects a growing trend in NLP research emphasizing the creation of smaller, domain- and dialect-specific models to democratize access to language technologies, preserve linguistic diversity, and provide more accurate and contextually relevant tools for speakers of underrepresented language varieties.

This technical report presents the design, training methodology, dataset composition, and potential applications of Cecilia 2B, highlighting its role as a foundational resource for Cuban Spanish NLP research and applications.

The remainder of this report is organized as follows: Section 2 describes the model architecture and design. Section 3 presents the training corpus and procedure. Section 4 presents some briefs notes on ongoing evaluation efforts. Section 5 discusses practical applications and potential use cases. Section 6 addresses ethical considerations, including bias and responsible deployment. Finally, Section 7 outlines future work and directions for further improving the model.

Model Architecture and Design

Cecilia 2B is built upon the Salamandra 2B model, a transformer-based decoder-only language model developed by the Barcelona Supercomputing Center's Language Technologies Unit. Salamandra 2B comprises approximately 2.25 billion parameters and employs a standard Transformer architecture with 24 layers, a hidden size of 2048, and 16 attention heads. It uses rotary positional embeddings (RoPE), SwiGLU activation functions, RMS normalization,

and flash attention to optimize training stability and computational efficiency. The model supports a large context window of 8,192 tokens and a vocabulary size of 256K tokens, enabling it to handle diverse multilingual inputs effectively.

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.

Importantly, the architecture of Salamandra 2B was left unmodified in the development of Cecilia 2B, including the tokenizer and vocabulary. The adaptation to Cuban Spanish was applied exclusively through continual pretraining on a curated Cuban text corpus, ensuring that the model's original structural and hyperparameter configurations remain intact. This approach aims to preserve the robustness and generalization properties of the base Salamandra 2B model while specializing its linguistic knowledge to the Cuban Spanish variant. However, it must be considered that words outside the original vocabulary (cubanisms and transliterated words, for example) will be harder to learn due to the nature of tokenization.

Training Data and Procedure

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. To respect copyright and intellectual property rights, the raw training data is not publicly available at the moment.

A full report of the dataset composition is available here.

Dataset Composition

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

Metric	Value
Lexical Density	6.8 characters/word

The Cecilia 2B training corpus is extensive, 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.

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.

Metric	Value
Total Samples	1,104,532
Total Tokens (no padding)	982,024,795
Total Tokens (with padding)	1,131,040,768
Average Sequence Length (no padding)	889.3 tokens
Padding Ratio	13.2%

After tokenization, 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.

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.

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

Evaluation and Benchmarking

The evaluation of Cecilia 2B is currently ongoing. The primary focus of this assessment is to measure the model's ability to capture the linguistic and cultural nuances specific to the Cuban Spanish variant, which is critical for its intended applications. In addition to these variant-specific tests, standard evaluations for biases, fairness, and general language modeling capabilities are being conducted to ensure the model's robustness and ethical soundness.

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.

Applications and Use Cases

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—approximately 2 billion parameters—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.

Currently, the model is not quantized and requires approximately 14 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.

Ethical Considerations

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.

Furthermore, the training corpus includes copyrighted materials collected under fair use assumptions strictly for academic research. Any use of Cecilia 2B must respect the intellectual property rights of the original content creators and copyright holders. Redistribution or commercial exploitation of the raw training data is prohibited.

Given these considerations and the fact that Cecilia 2B is not yet production-ready, access to the model on the Hugging Face platform is currently gated. Researchers interested in using the model must submit a request, which will be evaluated on a case-by-case basis. Approval is granted for use cases deemed ethical and aligned with responsible research practices. This controlled access aims to mitigate potential misuse and ensure that the model's deployment aligns with community standards.

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.

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. Additionally, quantized versions of all Cecilia models will be published to enable efficient inference in production environments.

Conclusions

The development of Cecilia 2B represents an initial but promising step toward creating high-quality language models tailored specifically for Cuban Spanish. This work aims not only to address the linguistic and cultural particularities of this variant but also to lay a foundation for future advancements in Cuban Spanish natural language processing.

We hope that this effort will inspire other communities across Latin America and similarly underserved language variants to build upon our experience, fostering a broader movement toward inclusive and diverse language technology development. We warmly invite researchers and practitioners interested in Cuban Spanish language modeling to collaborate, share insights, and contribute to the ongoing evolution of these resources, ultimately advancing the state of NLP for regional and minority language varieties.