Artículo Original / Original Research

Ciencias Matemáticas, Vol. XX, No. X, XXXX, Pag. 1-10
Recibido (*Received*): XX-XX-2024, Revisado (*Revised*): XX-XX-2024
Aceptado (*Accepted*): XX-XX-2024, En línea (*Online*): XX-XX-2024

Cecilia: El Modelo de Lenguaje Cubano
Cecilia: The Cuban Language Model

Alejandro Piad-Morffis[1]*, Suilan Estevez-Velarde[1], Yudivian Almeida-Cruz[1], Ernesto Luis Estevanell-Valladares[12], Roberto García Rodríguez[1], Alejandro Beltrán Varela[1], Carla Sunami Pérez Valera[1], Daniel Alejandro Valdés Pérez[1], Elena Rodríguez Horta[1], Gabriel Hernández Rodríguez[1], Deborah Famadas Rodríguez[1], Niley González Ferrales[1], Roberto Marti Cedeño[1], Juan Pablo Consuegra Ayala[12], Robiert Sepúlveda-Torres[2], Yoan Gutiérrez Vazquez[2], Andrés Montoyo[2] Rafael Muñoz Guillena[2] Manuel Javier Palomar Sanz[2]

Resumen El artículo presenta Cecilia 2B, el primer modelo de lenguaje entrenado específicamente para el español cubano. Cecilia 2B está basado en la arquitectura Salamandra 2B, un modelo multilingüe de 2 mil millones de parámetros y adaptado mediante preentrenamiento continuo sobre un corpus cuidadosamente compilado con cerca de mil millones de tokens provenientes de fuentes diversas como prensa nacional, literatura, legislación y enciclopedias cubanas. Esta especialización permite a Cecilia 2B capturar matices lingüísticos y culturales propios del español cubano, superando las limitaciones de los modelos generalistas en la comprensión de variantes regionales. La evaluación en benchmarks multilingües estándar muestra solo una reducción promedio del $2{,}4\%$ en el desempeño respecto al modelo base, manteniendo robustez general. No obstante, aún se requieren recursos específicos para evaluar el español cubano y medir completamente sus ventajas. Este trabajo impulsa el desarrollo de tecnologías lingüísticas regionales adaptadas a culturas locales y entornos con recursos limitados, resaltando la ética en el uso de datos y la participación comunitaria.

Palabras Clave: corpus, modelado del lenguaje, procesamiento del lenguaje natural

Abstract *The article presents Cecilia 2B, the first language model specifically trained for Cuban Spanish. Cecilia 2B is based on Salamandra 2B, a robust multilingual architecture with 2 billion parameters and adapted through continual pretraining on a carefully curated corpus of nearly one billion tokens from diverse Cuban sources including national press, literature, legislation, and encyclopedias. This specialization allows Cecilia 2B to capture linguistic and cultural nuances unique to Cuban Spanish, addressing limitations of generalist large language models in regional variant understanding. The evaluation on standard multilingual benchmarks shows only a minor* 2.4% *average performance drop compared to the base model, indicating robust general language capabilities are maintained. However, dedicated Cuban Spanish evaluation resources are still needed to fully measure the model's domain-specific advantages. The work opens avenues for regional language AI technologies adapted to local cultures and resource-constrained environments, emphasizing the importance of ethical data use and community involvement.*

Keywords: *corpora, language modeling, natural language processing*

Mathematics Subject Classification: 68, 68T07, 68T50

[1]Departamento de Inteligencia Artifiacial y Sistemas Computacionales, Facultad de Matemática y Computación, Universidad de la Habana, La Habana, Cuba. Email: apiad@matcom.uh.cu sestevez@matcom.uh.cu yudi@matcom.uh.cu ernesto.estevanell@matcom.uh.cu roberto.garcia@matcom.uh.cu alejandro.beltra@matcom.uh.cu carla.sperez@matcom.uh.cu daniel.valdes@matcom.uh.cu elena.rodriguez@matcom.uh.cu gabriel.hernandez@matcom.uh.cu deborah.famadas@matcom.uh.cu niley.gonzalez@matcom.uh.cu rmarticedeno@matcom.uh.cu

[2]GPLSI, Universidad de Alicante, Alicante, España. Email: juan.consuegra@.ua.es robiert.sepulveda@.ua.es ygutierrez@.ua.es montoyo@.ua.es rafael.munoz@.ua.es mpalomar@.ua.es

*Autor para Correspondencia (***Corresponding Author***)*

Editado por: Nombre del Editor de la Sección, Institución, País. (este campo lo modifica el editor)

Citar como: Apeuno Apedos, J.J., Apetres Apecuatro, J.E., Apecinco Apesis, J.A.,& Apesite Apeocho, J.R. (202X). Plantilla Para Un Trabajo a Publicarse en La Revista Ciencias Matemáticas. *Ciencias Matemáticas*, X(X), X–XX. Recuperado a partir de https://revistas.uh.cu/rcm/ (este campo lo modifica el editor)

Pérez Díaz, A., Rodríguez Hurtado, J., Estévez Velarde, S., Piad Morffis, A., Gutiérrez Y., Estevanell-Valladares, E.L., García Rodríguez, R., Beltrán Varela, A., Pérez Valera, C.S., Valdés Pérez, D.A., Rodríguez Hurtado, E., Hernández Rodríguez, G., Famadas Rodríguez, D., González Ferrales, N., Marti Cedeño, R., Consuegra Ayala, J.P., Sepúlveda-Torres, R., Gutiérrez Vazquez, Y., Montoyo, A., Muñoz Guillena, R., & Palomar Sanz, M.J.

Introducción

El desarrollo de modelos de lenguaje específicos para variantes regionales del español es una necesidad apremiante en el campo del Procesamiento de Lenguaje Natural (PLN). Los modelos generalistas, aunque poderosos, presentan limitaciones notables cuando se aplican a lenguas con pocos recursos o a variantes regionales, ya que suelen estar entrenados principalmente con datos de idiomas dominantes y no logran capturar los matices lingüísticos, culturales y contextuales propios de comunidades específicas [18]. En el caso del español cubano, esta carencia es especialmente crítica: no existen modelos de lenguaje que reflejen de manera precisa las particularidades léxicas, sintácticas y pragmáticas de la variante cubana. La creación de un modelo adaptado a esta realidad no solo contribuiría a cerrar la brecha digital lingüística, sino que también permitiría preservar y potenciar la riqueza cultural del español cubano en el ecosistema digital.

La inexistencia de un modelo de lenguaje entrenado específicamente para el español cubano implica que las expresiones idiomáticas, referencias culturales y fenómenos lingüísticos propios de la isla no son comprendidos ni representados adecuadamente por los modelos actuales. Esta situación afecta negativamente el rendimiento de tareas como el análisis de sentimiento, la generación de texto, la traducción automática y la interacción conversacional en contextos cubanos, donde la identidad lingüística es un componente esencial de la comunicación. Afortunadamente, los avances recientes en técnicas de preentrenamiento continuo permiten aprovechar modelos multilingües robustos —como Salamandra 2B— y especializarlos eficientemente en dominios regionales mediante la exposición a corpus representativos, sin necesidad de modificar la arquitectura ni el tokenizador original. Esto abre la puerta a la creación de modelos pequeños, eficientes y culturalmente adaptados, viables incluso en entornos con recursos computacionales limitados.

En este artículo se presenta **Cecilia 2B**, el primer modelo de lenguaje entrenado específicamente para el español cubano. Cecilia 2B está basado íntegramente en la arquitectura de Salamandra 2B, un modelo multilingüe de 2 mil millones de parámetros, y ha sido adaptado a través de preentrenamiento continuo sobre un corpus cuidadosamente compilado de textos cubanos. El corpus textual incluye cerca de mil millones de *tokens* provenientes de prensa nacional, literatura, legislación, enciclopedias y letras de canciones, asegurando una cobertura amplia de registros y contextos socioculturales.

El modelo mantiene la compatibilidad total con el diseño y el vocabulario de Salamandra 2B, garantizando que las modificaciones observadas se deban exclusivamente a la especialización en datos cubanos. Cecilia 2B es, por tanto, una herramienta pionera que facilita el desarrollo de aplicaciones de PLN adaptadas a la realidad lingüística y cultural de Cuba, sentando las bases para futuros avances en la inclusión digital y la preservación del patrimonio lingüístico regional.

El artículo se organiza de la siguiente manera: en la Sección 1 se revisa el estado del arte en modelos de lenguaje para variantes regionales y lenguas con pocos recursos, destacando las limitaciones de los modelos generalistas y las ventajas de los modelos pequeños y adaptados. La Sección 2 describe en detalle la arquitectura de Cecilia 2B, el proceso de construcción del corpus cubano y los procedimientos de preentrenamiento continuo empleados. En la Sección 3 se presentan los resultados de la evaluación cuantitativa del modelo en tareas estándar multilingües, discutiendo tanto su robustez general como las limitaciones actuales en la evaluación específica del español cubano. Finalmente, en la Sección 4 se exponen las conclusiones y se plantean las líneas futuras de trabajo, enfatizando la necesidad de recursos de evaluación y aplicaciones prácticas que validen el impacto de Cecilia 2B en la comunidad cubana.

1. Estado del arte

1.1 Limitaciones de los modelos generalistas

Los modelos de lenguaje de gran escala (*Large Lenguage Models, LLMs*) han revolucionado el procesamiento del lenguaje natural, pero presentan limitaciones significativas cuando se aplican a lenguajes con pocos recursos o variantes regionales específicas. Estas limitaciones afectan tanto la representación como la funcionalidad de estos modelos en contextos lingüísticos diversos.

Los lenguajes con pocos recursos enfrentan dos limitaciones cruciales: una escasez de datos lingüísticos etiquetados y no etiquetados, y la baja calidad de los datos disponibles, que frecuentemente no son suficientemente representativos de los idiomas y sus contextos socioculturales. Esta carencia de datos de entrenamiento adecuados resulta en modelos que no capturan correctamente las particularidades lingüísticas de estos idiomas, lo que afecta su rendimiento en tareas básicas de procesamiento de lenguaje natural.

Los modelos generalistas, entrenados principalmente con datos de idiomas dominantes como el inglés, muestran un rendimiento significativamente inferior cuando se aplican a lenguajes con recursos limitados. Por ejemplo, investigaciones recientes han demostrado que incluso modelos avanzados como GPT-4o y Llama 3.1 (405B) tienen un rendimiento inferior en comparación con modelos BERT (*Bidirectional Encoder Representations from Transformers*) ajustados específicamente para idiomas como el marathi, con márgenes de precisión de 10,2 % y 14,1 % respectivamente [18].

Estos modelos también presentan limitaciones significativas en la comprensión de contextos culturales específicos. Además, carecen de la capacidad para interpretar adecuadamente expresiones idiomáticas, referencias culturales y matices lingüísticos propios de variantes regionales. Esta deficiencia se manifiesta en una menor precisión en tareas como el análisis de sentimiento, la detección de discurso de odio y la clasificación de textos en idiomas con pocos recursos.

Las evaluaciones existentes para lenguajes con pocos recursos contienen limitaciones que necesitan ser estudiadas más a fondo, ya que los marcos de evaluación actuales no capturan adecuadamente las inconsistencias culturales en los

conjuntos de datos [11]. Esto resulta en una disminución del rendimiento cuando se aplican modelos multilingües a contextos lingüísticos específicos.

Asimismo, los lenguajes con pocos recursos a menudo carecen de la infraestructura digital y las herramientas necesarias (como tokenizadores, analizadores morfológicos) que están fácilmente disponibles para lenguajes con muchos recursos. Esta carencia complica incluso los pasos iniciales del desarrollo de modelos, creando barreras tecnológicas significativas para la inclusión de estos idiomas en el panorama de la Inteligencia Artificial (IA).

La falta de representación adecuada en los modelos generalistas contribuye a ampliar la brecha digital lingüística, dejando a muchas comunidades lingüísticas sin acceso a tecnologías basadas en IA que podrían beneficiar su desarrollo educativo, cultural y económico.

1.2 Ventajas de modelos pequeños para lenguajes regionales

Los modelos de lenguaje pequeños (*Small Lenguage Models, SLMs*) ofrecen ventajas significativas para el desarrollo de soluciones lingüísticas regionales, presentando alternativas viables a los grandes modelos generalistas para contextos específicos.

Los SLMs requieren significativamente menos memoria, almacenamiento y potencia de procesamiento en comparación con los modelos grandes, lo que los hace adecuados para dispositivos con capacidades de hardware limitadas, como smartphones, tablets y dispositivos IoT(*Internet of Things*) [27]. Esta característica es particularmente relevante en regiones donde el acceso a infraestructura computacional avanzada es limitado.

La eficiencia energética de los SLMs representa otra ventaja crucial, ya que consumen menos energía que sus contrapartes más grandes. Esto no solo resulta rentable sino que también apoya el desarrollo de IA ambientalmente sostenible al reducir la huella de carbono de los modelos de aprendizaje automático. Para comunidades con recursos limitados, esta eficiencia energética puede ser determinante para la adopción de tecnologías de IA.

Los modelos pequeños son más fáciles de ajustar para aplicaciones específicas y dominios lingüísticos particulares [16]. Al no requerir grandes cantidades de datos o potencia computacional para su entrenamiento, los desarrolladores pueden adaptarlos para tareas como el procesamiento de variantes dialectales, expresiones regionales y contextos culturales específicos.

Investigaciones recientes han demostrado que los SLMs pueden proporcionar una comprensión del lenguaje de alta calidad con un consumo de recursos significativamente menor, lo que los hace ideales para habilitar el trabajo digital en contextos lingüísticos específicos [9]. Su capacidad para capturar dialectos locales, expresiones idiomáticas y matices culturales con mayor precisión que los modelos grandes contribuye no solo a la preservación del lenguaje sino también a mejorar la inclusión digital en regiones desatendidas.

Por otra parte, los SLMs pueden desplegarse localmente, lo que permite el procesamiento de datos en el dispositivo y reduce la necesidad de enviar información sensible a sistemas basados en la nube. Esta característica ofrece ventajas significativas de privacidad para todo tipo de instituciones o comunidades que pueden tener preocupaciones sobre la soberanía de sus datos o el uso de estos por terceros.

Para las lenguas indígenas y regionales, los SLMs ofrecen una solución rentable y eficiente en términos de recursos al reducir los requisitos computacionales y de datos, mientras mejoran la precisión de salida a través de conjuntos de datos específicos y contextualizados. Este enfoque permite un desarrollo más participativo, donde las comunidades lingüísticas pueden mantener mayor control sobre sus recursos lingüísticos y culturales.

1.3 Estrategias para construir modelos de lenguaje pequeños con pocos recursos

El desarrollo de modelos de lenguaje para idiomas con recursos limitados requiere enfoques innovadores que maximicen la eficiencia y efectividad del entrenamiento.

El preentrenamiento continuo (*continual pretraining*) ofrece una posibilidad para la adaptación de dominio con recursos computacionales limitados [29]. Esta técnica permite que modelos preentrenados existentes sean posteriormente entrenados con datos específicos de dominio, permitiéndoles adquirir conocimiento especializado mientras aprovechan su base de conocimiento existente.

Una publicación reciente ha demostrado mejoras significativas en el rendimiento a través del entrenamiento incremental en 400 millones de tokens, seguido de entrenamiento adicional para alcanzar mil millones de tokens. Los resultados muestran ganancias notables en tareas intensivas en conocimiento (*MMLU* $+8{,}1\,\%$) y comprensión contextual (*HellaSwag* $+7{,}6\,\%$), mientras revelan compensaciones en la especialización de dominio [17].

El preentrenamiento continuo de modelos de lenguaje pequeños en corpus específicos de dominio ha demostrado ser más efectivo que entrenar modelos desde cero. Por ejemplo, en el dominio biomédico, los modelos inicializados con *MiniLM* y continuamente preentrenados en textos específicos del dominio superaron a los modelos entrenados desde cero con el mismo vocabulario [30].

Otro enfoque es *Adapt-and-Distill* que representa una estrategia efectiva para desarrollar modelos pequeños, rápidos y efectivos para dominios específicos. Este método combina la adaptación de modelos preentrenados generales y la destilación de conocimiento específico del dominio, logrando un mejor rendimiento mientras se reduce significativamente el tamaño y se aumenta la velocidad del modelo [31].

La expansión de vocabulario específico del dominio durante la fase de adaptación y el empleo de la probabilidad de ocurrencia a nivel de corpus para elegir automáticamente el tamaño del vocabulario incremental son técnicas clave

en este enfoque. Experimentos en los dominios biomédico e informático han demostrado que esta estrategia logra un mejor rendimiento en tareas específicas del dominio mientras el modelo es 3,3 veces más pequeño y 5,1 veces más rápido que los modelos originales [31].

En el caso de lenguajes con recursos extremadamente limitados, el enfoque de "datos pequeños" ha demostrado ser sorprendentemente efectivo. La suposición común de que las lenguas con pocos recursos se benefician del entrenamiento conjunto con lenguas de mayores recursos ha sido puesta en entredicho pues se ha demostrado que es posible entrenar modelos de lenguaje multilingües competitivos con menos de un *gigabyte* (GB) de texto [25].

La combinación de datos sintéticos generados tanto por traducción automática estadística, como por modelos de traducción automática neuronal multilingües ha demostrado mejorar el rendimiento para lenguas con pocos recursos debido a la mayor diversidad de los datos sintéticos generados. Esta técnica es particularmente valiosa cuando los datos paralelos bilingües son escasos [26].

Asimismo, el uso de técnicas eficientes en parámetros como LoRA PEFT (*Parameter-Efficient Fine-Tuning*) minimiza el número de parámetros durante el ajuste fino, ofreciendo eficiencia computacional y manteniendo la robustez del modelo original al ajustar solo algunos de los parámetros [5]. Estudios más amplios han enfatizado que el uso de LoRA en entornos con pocos recursos conlleva una baja sobrecarga computacional [10, 6].

1.4 Proyectos Regionales de Modelos de Lenguaje

El proyecto SEALD (*Southeast Asian Languages in One Network Data*) constituye una de las iniciativas más ambiciosas para fortalecer la presencia digital de las lenguas del Sudeste Asiático. Mediante la colaboración entre AI Singapore y Google Research, se recopilaron y curaron grandes volúmenes de datos multilingües, abarcando idiomas como indonesio, malayo, tamil, birmano, filipino, vietnamita, tailandés, lao y jemer [24]. Este esfuerzo permitió el desarrollo de SEA-LION, una familia de modelos de lenguaje preentrenados específicamente para la región, con arquitecturas de 3 a 7 mil millones de parámetros y un vocabulario adaptado a las características lingüísticas del área, mejorando sustancialmente la comprensión y generación de texto en estos idiomas.

AfriBERTa [25] representa un enfoque innovador para lenguas africanas con pocos recursos, desafiando la suposición de que el entrenamiento conjunto con idiomas de alto recurso es siempre beneficioso. Este modelo fue entrenado exclusivamente con menos de 1 GB de texto de 11 lenguas africanas, incluyendo el primer modelo de lenguaje para cuatro de ellas. AfriBERTa demostró, en tareas de reconocimiento de entidades nombradas y clasificación de texto, que un modelo multilingüe focalizado puede superar a alternativas generalistas como mBERT y XLM-R, validando la eficacia de estrategias centradas en corpus pequeños y específicos.

Salamandra [12] es un caso paradigmático de éxito en la construcción de modelos multilingües europeos, sirviendo también como base para adaptaciones regionales como Cecilia. La arquitectura de Salamandra abarca variantes de 2, 7 y 40 mil millones de parámetros, todas entrenadas desde cero sobre un corpus multilingüe cuidadosamente curado de 7.8 billones de tokens en 35 idiomas europeos y código de programación. El modelo utiliza precisión bfloat16, embeddings RoPE, activación SwiGLU, normalización RMS, atención flash y una longitud de contexto de hasta 8192 tokens, con un vocabulario de 256000 tokens. Entrenado en el supercomputador MareNostrum 5. Salamandra ha demostrado un rendimiento competitivo en benchmarks multilingües y sirve como plataforma robusta para la especialización en variantes lingüísticas regionales.

2. Arquitectura y entrenamiento de Cecilia

Cecilia 2B está basada íntegramente en la arquitectura de Salamandra 2B (BSC-LT/salamandra-2B), un modelo de lenguaje multilingüe de 2 mil millones de parámetros. Para la adaptación al español cubano, no se realizaron modificaciones ni en la arquitectura original ni en el tokenizador empleado, manteniendo la compatibilidad total con el diseño, el tamaño de vocabulario y las capacidades de representación del modelo base. Esta decisión asegura que las mejoras en desempeño se deban exclusivamente al preentrenamiento continuo sobre datos cubanos, y no a cambios estructurales o de tokenización.

2.1 Descripción del Corpus

El corpus utilizado para el preentrenamiento de Cecilia fue cuidadosamente compilado para capturar la diversidad temática y cultural del español cubano. Incluye textos provenientes de múltiples dominios en 296311 archivos, abarcando fuentes como Ecured [1], la Enciclopedia Digital del Audiovisual Cubano, prensa nacional, literatura cubana, legislación, enciclopedias de cubanismos, y letras de canciones populares. Esta composición permite la representación de registros formales e informales, así como de distintos géneros discursivos y contextos socioculturales, proporcionando una base sólida para que el modelo aprenda matices lingüísticos y culturales propios de Cuba.

En términos cuantitativos, el corpus contiene un total de 2631691355 caracteres y 384963687 palabras distribuidas en 34505341 líneas. La longitud promedio por documento es de 8881 caracteres, mientras que la longitud promedio de las oraciones es de 17 palabras. La densidad léxica, medida como caracteres por palabra, es de 6.8, reflejando la riqueza y variedad del vocabulario presente en el dataset. Estos valores posibilitan una cobertura amplia tanto en extensión como en profundidad temática.

Para el proceso de entrenamiento, el corpus fue tokenizado en secuencias de hasta 1024 tokens utilizando el tokenizador original de Salamandra 2B, sin modificaciones [28]. El dataset

[1] https://www.ecured.cu/EcuRed:Enciclopedia_cubana

Tabla 1. Descripción del Corpus [*Corpus description*].

Métrica	Valor
Archivos totales	296,311
Caracteres totales	2,631,691,355
Palabras totales	384,963,687
Líneas totales	34,505,341
Longitud promedio doc	8,881 caracteres
Longitud promedio oración	17,0 palabras
Densidad léxica	6,8 car./palabra

tokenizado resultante consta de 1104532 muestras, sumando 982024795 tokens efectivos (sin padding) y 1131040768 tokens totales (con padding). La longitud promedio de secuencia es de 889.3 tokens, y la razón de padding es eficiente, con un 13,2 %. Este procesamiento garantiza que la mayor parte de la capacidad del modelo se utilice en datos reales y representativos.

Tabla 2. Corpus tokenizado [*Tokenized corpus*].

Métrica	Valor
Muestras totales	1,104,532
Tokens totales (sin padding)	982,024,795
Tokens totales (con padding)	1,131,040,768
Longitud promedio de secuencia (tokens)	889,3
Proporción de padding	13,2 %

El corpus fue sometido a un riguroso proceso de limpieza y validación, asegurando la ausencia de problemas de encoding y una alta integridad textual. Esto garantiza que los datos de entrada sean de calidad y representativos, fundamentales para un preentrenamiento efectivo y robusto del modelo.

2.2 Proceso de entrenamiento

El entrenamiento de Cecilia 2B se realizó mediante preentrenamiento continuo sobre el corpus cubano, manteniendo la arquitectura y el tokenizador originales de Salamandra 2B. A continuación se detallan los hiperparámetros empleados:

Número de épocas: El corpus completo se procesó dos veces, permitiendo al modelo refinar sus representaciones sobre los datos cubanos.

Learning rate y scheduler: Se empleó una tasa de aprendizaje inicial baja $(2e-5)$ con un scheduler de tipo *warmup linear*, que incrementa gradualmente la tasa de aprendizaje durante el 6 % inicial de los pasos de entrenamiento antes de decaer linealmente, con el fin de evitar inestabilidades al comienzo.

Optimizador AdamW: *AdamW* es una variante de Adam que desacopla el weight decay, permitiendo un mejor control de la regularización para prevenir el sobreajuste.

Tabla 3. Descripción del entrenamiento [*Training description*].

Parámetro	Valor
Número de épocas	2
Batch size	4
Gradient accumulation steps	16
Batch efectivo	64
Learning rate	$2e-5$
Scheduler	*Warmup linear*
Proporción de warmup	6 %
Optimizador	*AdamW*
Weight decay	,01
Betas ($\beta 1$, $\beta 2$)	,9, ,999
Gradient clipping norm	1,0
Precisión	*bfloat16*

Weight decay y betas: *Weight decay* de ,01 promueve la regularización. Los parámetros $\beta 1$ y $\beta 2$ controlan los promedios móviles de los momentos de primer y segundo orden en *AdamW*.

Gradient clipping norm: Limita la norma de los gradientes a 1.0 para evitar explosiones de gradiente.

Precisión bfloat16: Permite mayor eficiencia en memoria y cómputo, facilitando el entrenamiento de modelos grandes en hardware moderno.

Durante el entrenamiento se emplearon técnicas como *Fully Sharded Data Parallel* (FSDP) y *gradient checkpointing* para maximizar el uso eficiente de memoria y recursos computacionales, además de validaciones periódicas y monitoreo de métricas para alcanzar la robustez y generalización del modelo [21].

3. Evaluación

Para evaluar el impacto del preentrenamiento continuo de Cecilia 2B sobre datos cubanos, se comparó su desempeño con Salamandra 2B en una batería de tareas clásicas de procesamiento de lenguaje natural en inglés y español. Es importante destacar que estas tareas no están diseñadas para medir la comprensión de la variante cubana del español, sino que sirven para verificar la robustez general del modelo tras la adaptación.

La evaluación de Cecilia 2B se realizó sobre un conjunto de tareas estándar ampliamente utilizadas en la comunidad de PLN, empleando benchmarks reconocidos como SpanishBench, CatalanBench, BasqueBench, GalicianBench y tareas en inglés del *LM Evaluation Harness*[3]. Estas colecciones agrupan tareas de comprensión lectora (por ejemplo, Belebele[2], que consiste en responder preguntas de opción múltiple sobre textos breves), inferencia lógica (XNLI[8],

Tabla 4. Resultados de la evaluación cuantitativa en comparación con Salamandra 2B. [*Results of the quantitative evaluation compared to Salamandra 2B.*].

Task	Metric	Salamandra	Cecila	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 %
xnli_va	acc	0,47505	0,48523	2,10 %
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** %

WNLI[2]), razonamiento de sentido común (*XStoryCloze*[22]), identificación de paráfrasis (PAWS[19]), y preguntas de respuesta abierta y razonamiento científico (OpenBookQA[23], ARC Challenge y ARC Easy[7]). Para las tareas en español, los datasets utilizados son versiones traducidas profesionalmente o generadas y revisadas por humanos, garantizando alta calidad y relevancia para la evaluación de modelos multilingües.

Además, se incluyen benchmarks de traducción automática (como Flores [13]), tareas de resumen (XLSum [15]), y comprensión y respuesta a preguntas de trivia (TriviaQA[20], XQuAD[1]), cubriendo así un espectro amplio de habilidades lingüísticas y cognitivas. Estos benchmarks permiten comparar el desempeño general de modelos multilingües y monolingües en tareas de comprensión, inferencia, generación y traducción en diferentes idiomas, aunque, como se señaló previamente, no están diseñados para evaluar competencias específicas en variantes regionales como el español cubano.

En promedio (Tabla 4), la reducción relativa de desempeño es de apenas 2,4 % respecto al modelo base, una diferencia no significativa considerando la magnitud del cambio en los datos de entrenamiento y la especialización lograda [14]. Esto indica que el modelo mantiene su capacidad general para tareas estándar, a pesar de haber sido adaptado a un dominio lingüístico y cultural específico.

Sin embargo, aún no se dispone de benchmarks ni corpus de instrucciones diseñados para evaluar específicamente la comprensión y generación en español cubano. La creación de estos recursos será esencial para medir el verdadero valor

[2] https://huggingface.co/datasets/SetFit/wnli

añadido de Cecilia en aplicaciones donde el dominio cultural y lingüístico local es crítico.

3.1 Discusión

La principal limitación de Cecilia 2B radica en que su evaluación actual se ha realizado exclusivamente sobre tareas generales y benchmarks estándar de procesamiento de lenguaje natural, como comprensión lectora, inferencia lógica y razonamiento de sentido común, que no están diseñados para medir la comprensión ni la generación en la variante cubana del español. Esto significa que, aunque el modelo mantiene un desempeño robusto en tareas multilingües generales con solo una reducción promedio del 2,4 % respecto al modelo base, aún no es posible cuantificar su ventaja específica en aplicaciones donde los matices culturales y lingüísticos cubanos sean críticos.

Además, el corpus utilizado, si bien diverso y representativo, podría inducir sesgos hacia los dominios más representados, como prensa y enciclopedias, en detrimento de registros menos frecuentes o más formales, lo que podría limitar la cobertura de ciertos contextos y estilos lingüísticos [4]. Esta situación resalta la necesidad de continuar ampliando y balanceando el corpus, incorporando textos de géneros subrepresentados y registros formales, así como de desarrollar recursos de evaluación específicos para el español cubano.

El desarrollo de Cecilia 2B abre múltiples líneas de trabajo para el futuro. Es fundamental ampliar y diversificar el corpus de entrenamiento, integrando fuentes adicionales que reflejen una mayor variedad de registros, géneros y contextos de uso del español cubano. Asimismo, el entrenamiento con más épocas y el ajuste fino en tareas específicas permitirán mejorar la especialización y robustez del modelo, maximizando su utilidad en aplicaciones concretas.

Un objetivo clave será la creación de benchmarks y corpus de instrucciones diseñados específicamente para evaluar la comprensión y generación en español cubano, lo que permitirá medir de forma precisa el impacto y la ventaja competitiva de Cecilia en tareas relevantes para usuarios y desarrolladores locales. Finalmente, la integración de Cecilia en aplicaciones reales —como asistentes virtuales, sistemas de soporte educativo, o herramientas de procesamiento documental— será esencial para validar su utilidad práctica y fomentar su adopción en la sociedad cubana.

La creación y despliegue de modelos de lenguaje regionales como Cecilia 2B plantea importantes consideraciones éticas. Es crucial asegurar que el uso de datos respete la privacidad, los derechos de autor y la diversidad cultural de las fuentes, evitando la reproducción o amplificación de sesgos presentes en los corpus de entrenamiento. La selección y documentación cuidadosa de los datos, así como la transparencia en los procesos de construcción y evaluación del modelo, son fundamentales para mitigar riesgos de homogeneización cultural y exclusión de voces minoritarias.

Además, el diseño participativo y la consulta con comunidades lingüísticas locales deben ser una prioridad, garantizando que las tecnologías desarrolladas respondan a las necesidades, valores y expectativas de sus usuarios, y contribuyan a una inteligencia artificial más justa, inclusiva y representativa.

4. Conclusiones

En este artículo se presenta Cecilia 2B, el primer modelo de lenguaje entrenado específicamente para el español cubano, construido mediante preentrenamiento continuo sobre la arquitectura Salamandra 2B sin modificar su estructura ni tokenizador. Cecilia utiliza un corpus textual diverso y representativo de la cultura y sociedad cubanas, compuesto por cerca de mil millones de tokens, abarcando prensa, literatura, legislación y recursos enciclopédicos. El modelo se evaluó en una batería de tareas estándar multilingües, mostrando una reducción promedio de solo 2,4 % respecto al modelo base, lo que indica que la especialización en datos cubanos no compromete su robustez general. Si bien la evaluación específica en tareas propias del español cubano es aún una asignatura pendiente, Cecilia 2B constituye un paso inicial fundamental hacia la creación de tecnologías lingüísticas adaptadas a la realidad cubana y sienta las bases para futuros desarrollos y aplicaciones en procesamiento de lenguaje natural regionalizado.

Relevancia del estudio

Cecilia 2B es un modelo de lenguaje crucial para abordar la brecha digital lingüística del español cubano, una variante desatendida por los modelos generales. Su estratégia de entrenamiento demuestra una forma eficiente de adaptar la inteligencia artificial a contextos específicos. A pesar de esta adaptación, el modelo mantiene su robustez en tareas generales PLN, con una reducción mínima del rendimiento con respecto al modelo base. El verdadero valor radica en su capacidad para capturar los matices culturales y léxicos del español cubano, lo que sienta un precedente para el desarrollo de modelos regionales en entornos con recursos limitados, facilitando la inclusión digital y contribuyendo a la preservación del patrimonio lingüístico regional.

Suplementos

Este artículo contiene un suplemento de información en:
http://cecilia.uhgia.org/
https://github.com/gia-uh/cecilia
https://huggingface.co/gia-uh/cecilia-2b-v0.1

Agradecimientos

Se agradece a todos los colaboradores y proveedores de datos que hicieron posible este trabajo. El modelo no podría haber sido creado sin el compromiso y el trabajo de los miembros de los grupos **GIA-UH** y **GPLSI**.

Este trabajo fue financiado parcialmente por el proyecto ILENIA-VIVES «2022/TL22/00215334» y con financiación privada de Syalia SRL y Epistemial.

Conflictos de interés

Se declara que no existen conflictos de interés.

Contribucián de autoría

Conceptualización A.P.M., S.E.V., Y.A.C., Y.G.V., A.M, R.M.G.

Curación de datos A.P.M., S.E.V., Y.A.C., E.L.E.V., R.G.R., A.B.V., C.S.P.V., D.A.V.P., E.R.H., G.H.R., D.F.R., N.G.F., R.M.C.

Análisis formal A.P.M., Y.A.C., E.L.E.V.

Adquisición de Financiamiento A.P.M., S.E.V., Y.A.C., E.L.E.V., Y.G.V., A.M, R.M.G.

Investigación A.P.M., S.E.V., Y.A.C., E.L.E.V., R.G.R., A.B.V., C.S.P.V., D.A.V.P., E.R.H., G.H.R., D.F.R., N.G.F., R.M.C., J.P.C.A., R.S.T., Y.G.V., A.M., R.M.G., M.J.P.S.

Metodología A.P.M., S.E.V., Y.A.C., E.L.E.V.

Administración de proyecto A.P.M., S.E.V., Y.A.C., E.L.E.V.

Recursos A.P.M., S.E.V., Y.A.C., Y.G.V., A.M, R.M.G.

Software A.P.M., S.E.V., Y.A.C., E.L.E.V., R.G.R., A.B.V., C.S.P.V., D.A.V.P., E.R.H., G.H.R., D.F.R., N.G.F., R.M.C., J.P.C.A., R.S.T., Y.G.V., A.M., R.M.G., M.J.P.S.

Supervisión A.P.M., Y.A.C.

Validación S.E.V., R.S.T., E.L.E.V., A.P.M., Y.A.C.

Visualización A.P.M., S.E.V., Y.A.C., E.L.E.V., R.G.R., A.B.V., C.S.P.V., D.A.V.P., E.R.H., G.H.R., D.F.R., N.G.F., R.M.C.

Redacción: preparación del borrador original R.G.R., D.A.V.P., A.P.M., Y.A.C., E.L.E.V.

Redacción: revisión y edición A.P.M., S.E.V., Y.A.C., E.L.E.V., R.G.R., A.B.V., C.S.P.V., D.A.V.P., E.R.H., G.H.R., D.F.R., N.G.F., R.M.C., J.P.C.A., R.S.T., Y.G.V., A.M., R.M.G., M.J.P.S.

Referencias

[1] Artetxe, Mikel, Sebastian Ruder y Dani Yogatama: *On the cross-lingual transferability of monolingual representations*. arXiv preprint arXiv:1910.11856, 2019.

[2] Bandarkar, Lucas, Davis Liang, Benjamin Muller, Mikel Artetxe, Satya Narayan Shukla, Donald Husa, Naman Goyal, Abhinandan Krishnan, Luke Zettlemoyer y Madian Khabsa: *The belebele benchmark: a parallel reading comprehension dataset in 122 language variants*. arXiv preprint arXiv:2308.16884, 2023.

[3] Biderman, Stella, Hailey Schoelkopf, Lintang Sutawika, Leo Gao, Jonathan Tow, Baber Abbasi, Alham Fikri Aji, Pawan Sasanka Ammanamanchi, Sidney Black, Jordan Clive *y cols.*: *Lessons from the trenches on reproducible evaluation of language models*. arXiv preprint arXiv:2405.14782, 2024.

[4] Cahyawijaya, Samuel, Holy Lovenia, Joel Ruben Antony Moniz, Tack Hwa Wong, Mohammad Rifqi Farhansyah, Thant Thiri Maung, Frederikus Hudi, David Anugraha, Muhammad Ravi Shulthan Habibi, Muhammad Reza Qorib *y cols.*: *Crowdsource, crawl, or generate? creating sea-vl, a multicultural vision-language dataset for southeast asia*. arXiv preprint arXiv:2503.07920, 2025.

[5] Chavan, Arnav, Zhuang Liu, Deepak Gupta, Eric Xing y Zhiqiang Shen: *One-for-all: Generalized lora for parameter-efficient fine-tuning*. arXiv preprint arXiv:2306.07967, 2023.

[6] Chen, Guanduo, Yutong He, Yipeng Hu, Kun Yuan y Binhang Yuan: *CE-LoRA: Computation-Efficient LoRA Fine-Tuning for Language Models*. arXiv preprint arXiv:2502.01378, 2025.

[7] Clark, Peter, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick y Oyvind Tafjord: *Think you have solved question answering? try arc, the ai2 reasoning challenge*. arXiv preprint arXiv:1803.05457, 2018.

[8] Conneau, Alexis, Guillaume Lample, Ruty Rinott, Adina Williams, Samuel R Bowman, Holger Schwenk y Veselin Stoyanov: *XNLI: Evaluating cross-lingual sentence representations*. arXiv preprint arXiv:1809.05053, 2018.

[9] Dong, Xin, Yonggan Fu, Shizhe Diao, Wonmin Byeon, Zijia Chen, Ameya Sunil Mahabaleshwarkar, Shih Yang Liu, Matthijs Van Keirsbilck, Min Hung Chen, Yoshi Suhara *y cols.*: *Hymba: A hybrid-head architecture for small language models, 2024*. URL https://arxiv.org/abs/2411.13676.

[10] Ellison, Thayer: *LoRA-Based Lightweight Adaptation of Pretrained Models for Low-Resource Text Summarization*. Journal of Computer Science and Software Applications, 5(6), 2025.

[11] Gamboa, Lance Calvin Lim y Mark Lee: *Filipino Benchmarks for Measuring Sexist and Homophobic Bias in Multilingual Language Models from Southeast Asia*. arXiv preprint arXiv:2412.07303, 2024.

[12] Gonzalez-Agirre, Aitor, Marc Pàmies, Joan Llop, Irene Baucells, Severino Da Dalt, Daniel Tamayo, José Javier Saiz, Ferran Espuña, Jaume Prats, Javier Aula-Blasco *y cols.*: *Salamandra technical report*. arXiv preprint arXiv:2502.08489, 2025.

[13] Goyal, Naman, Cynthia Gao, Vishrav Chaudhary, Peng Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc'Aurelio Ranzato, Francisco Guzmán y Angela Fan: *The flores-101 evaluation benchmark for low-resource and multilingual machine translation*. Transactions of

the Association for Computational Linguistics, 10:522–538, 2022.

[14] Guo, Zhen y Yining Hua: *Continuous training and fine-tuning for domain-specific language models in medical question answering*. arXiv preprint arXiv:2311.00204, 2023.

[15] Hasan, Tahmid, Abhik Bhattacharjee, Md Saiful Islam, Kazi Samin, Yuan Fang Li, Yong Bin Kang, M Sohel Rahman y Rifat Shahriyar: *XL-sum: Large-scale multilingual abstractive summarization for 44 languages*. arXiv preprint arXiv:2106.13822, 2021.

[16] Hu, Zijian, Jipeng Zhang, Rui Pan, Zhaozhuo Xu, Salman Avestimehr, Chaoyang He y Tong Zhang: *Fox-1 Technical Report*. arXiv preprint arXiv:2411.05281, 2024.

[17] Ishibashi, Yoichi, Taro Yano y Masafumi Oyamada: *Mining hidden thoughts from texts: Evaluating continual pretraining with synthetic data for llm reasoning*. arXiv preprint arXiv:2505.10182, 2025.

[18] Jadhav, Suramya, Abhay Shanbhag, Amogh Thakurdesai, Ridhima Sinare y Raviraj Joshi: *On Limitations of LLM as Annotator for Low Resource Languages*. arXiv preprint arXiv:2411.17637, 2024.

[19] Jayawardena, Lasal y Prasan Yapa: *Parafusion: A large-scale llm-driven english paraphrase dataset infused with high-quality lexical and syntactic diversity*. arXiv preprint arXiv:2404.12010, 2024.

[20] Joshi, Mandar, Eunsol Choi, Daniel S Weld y Luke Zettlemoyer: *Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension*. arXiv preprint arXiv:1705.03551, 2017.

[21] Lialin, Vladislav, Vijeta Deshpande y Anna Rumshisky: *Scaling down to scale up: A guide to parameter-efficient fine-tuning*. arXiv preprint arXiv:2303.15647, 2023.

[22] Lin, Xi Victoria, Todor Mihaylov, Mikel Artetxe, Tianlu Wang, Shuohui Chen, Daniel Simig, Myle Ott, Naman Goyal, Shruti Bhosale, Jingfei Du *y cols.*: *Few-shot learning with multilingual generative language models*. En *Proceedings of the 2022 conference on empirical methods in natural language processing*, páginas 9019–9052, 2022.

[23] Mihaylov, Todor, Peter Clark, Tushar Khot y Ashish Sabharwal: *Can a suit of armor conduct electricity? a new dataset for open book question answering*. arXiv preprint arXiv:1809.02789, 2018.

[24] Ng, Raymond, Thanh Ngan Nguyen, Yuli Huang, Ngee Chia Tai, Wai Yi Leong, Wei Qi Leong, Xianbin Yong, Jian Gang Ngui, Yosephine Susanto, Nicholas Cheng *y cols.*: *Sea-lion: Southeast asian languages in one network*. arXiv preprint arXiv:2504.05747, 2025.

[25] Ogueji, Kelechi, Yuxin Zhu y Jimmy Lin: *Small data? no problem! exploring the viability of pretrained multilingual language models for low-resourced languages*. En *Proceedings of the 1st workshop on multilingual representation learning*, páginas 116–126, 2021.

[26] Shu, Peng, Junhao Chen, Zhengliang Liu, Hui Wang, Zihao Wu, Tianyang Zhong, Yiwei Li, Huaqin Zhao, Hanqi Jiang, Yi Pan *y cols.*: *Transcending language boundaries: Harnessing llms for low-resource language translation*. arXiv preprint arXiv:2411.11295, 2024.

[27] Tonja, Atnafu Lambebo, Bonaventure FP Dossou, Jessica Ojo, Jenalea Rajab, Fadel Thior, Eric Peter Wairagala, Anuoluwapo Aremu, Pelonomi Moiloa, Jade Abbott, Vukosi Marivate *y cols.*: *Inkubalm: A small language model for low-resource african languages*. arXiv preprint arXiv:2408.17024, 2024.

[28] Vo, Anh Dung, Minseong Jung, Wonbeen Lee y Daewoo Choi: *Redwhale: An adapted korean llm through efficient continual pretraining*. arXiv preprint arXiv:2408.11294, 2024.

[29] Wu, Tongtong, Linhao Luo, Yuan Fang Li, Shirui Pan, Thuy Trang Vu y Gholamreza Haffari: *Continual learning for large language models: A survey*. arXiv preprint arXiv:2402.01364, 2024.

[30] Yan, Yongyu, Kui Xue, Xiaoming Shi, Qi Ye, Jingping Liu y Tong Ruan: *Af adapter: Continual pretraining for building chinese biomedical language model*. En *2023 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, páginas 953–957. IEEE, 2023.

[31] Yao, Yunzhi, Shaohan Huang, Wenhui Wang, Li Dong y Furu Wei: *Adapt-and-distill: Developing small, fast and effective pretrained language models for domains*. arXiv preprint arXiv:2106.13474, 2021.