

Cecilia: El Modelo de Lenguaje Cubano

Cecilia: The Cuban Language Model

Alejandro Piad-Morffis^{1*}, Suilan Estevez-Velarde¹, Yudivian Almeida-Cruz¹, Ernesto Luis Estevanell-Valladares¹², Roberto García Rodríguez¹, Alejandro Beltrán Varela¹, Carla Sunami Pérez Valera¹, Daniel Alejandro Valdés Pérez¹, Elena Rodríguez Horta¹, Gabriel Hernández Rodríguez¹, Deborah Famadas Rodríguez¹, Niley González Ferrales¹, Roberto Marti Cedeño¹, Juan Pablo Consuegra Ayala¹², Robiert Sepúlveda-Torres², Yoan Gutiérrez Vazquez², Andrés Montoyo², Rafael Muñoz Guillena², Manuel Javier Palomar Sanz²

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

¹Departamento de Inteligencia Artificial 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 @matcom.uh.cu elena.rodriguez@matcom.uh.cu gabriel.hernandez@matcom.uh.cu deborah.famadas@matcom.uh.cu rmarticedeno@matcom.uh.cu

²GPLSI, 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)

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 Language 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

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 (*MM-LU* +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 recopilaban y curaban 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 EcuRed¹, 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

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

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 %
cocoterros_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 %
cocoterros_es	rouge1	0,33887	0,31209	-7,90 %
xlsum_es	rouge1	0,13464	0,08705	-35,35 %
Mean Diff				-2.43 %

WNLI²), 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 mo-

nolingües 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>

do 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 estrategia 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.

Contribución 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.: *Crowdsourced, crawled, or generated? 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

- 

