



LatamGPT:

Construyendo IA desde
el sur global



LatamGPT Team

Centro Nacional de Inteligencia Artificial (CENIA), Chile.
Investigación y desarrollo de
modelos de lenguaje para América
Latina.

 latamgpt.org

Resumen / Los grandes modelos de lenguaje que hoy dominan la inteligencia artificial fueron entrenados mayoritariamente con texto en inglés. A modo de ejemplo, en los modelos con información pública, América Latina, con el 8,8% de la población mundial y seis siglos de producción escrita principalmente en español y portugués, representa menos del 2% de los datos usados para entrenar estos sistemas. LatamGPT surge para aportar a cerrar esa brecha: es un modelo de 70 mil millones de parámetros, entrenado sobre 300 mil millones de tokens provenientes de corpus latinoamericanos. Estos corpus fueron aportados por 79 instituciones en 15 países de la región, dentro de un corpus más amplio que cubre 20 países latinoamericanos. Su objetivo es contribuir a que América Latina deje de ser sólo consumidora de IA y fortalezca sus propias capacidades para desarrollarla.

El problema: los modelos de lenguaje no hablan español latinoamericano

Los grandes modelos de lenguaje aprenden del texto con el que son entrenados. Si ese texto proviene mayoritariamente de páginas web en inglés, de Wikipedia anglosajona y de libros digitalizados de bibliotecas norteamericanas y europeas, el modelo representará el mundo desde esa perspectiva. La principal fuente de datos para entrenar la mayoría de los modelos actuales es Common Crawl, una enorme colección de páginas web de Internet donde el español latinoamericano representa una fracción marginal [1].

América Latina representa el 6,6% del PIB mundial y el 8,8% de la población del planeta, pero recibe apenas el 1,12% de la inversión global en inteligencia artificial [2]. Esta brecha de inversión se traduce en una brecha de representación: los modelos que millones de latinoamericanos usan hoy fueron diseñados para otros contextos, con otros referentes culturales y otro español. El problema no es sólo lingüísti-

América Latina representa el 6,6% del PIB mundial y el 8,8% de la población del planeta, pero recibe apenas el 1,12% de la inversión global en inteligencia artificial.

co sino también contextual: ese conocimiento no se puede inyectar traduciendo textos, sino que debe surgir de haber sido entrenado sobre fuentes auténticamente regionales.

LatamGPT nace para aportar a cerrar esa brecha. Partiendo de Llama 3.1 70B [3], continuamos su entrenamiento sobre 300 mil millones de tokens provenientes de corpus que cubren 20 países latinoamericanos, con aportes institucionales de 79 organizaciones en 15 países de la región. El resultado es el mayor modelo actual de lenguaje abierto entrenado con un foco en datos sobre América Latina.

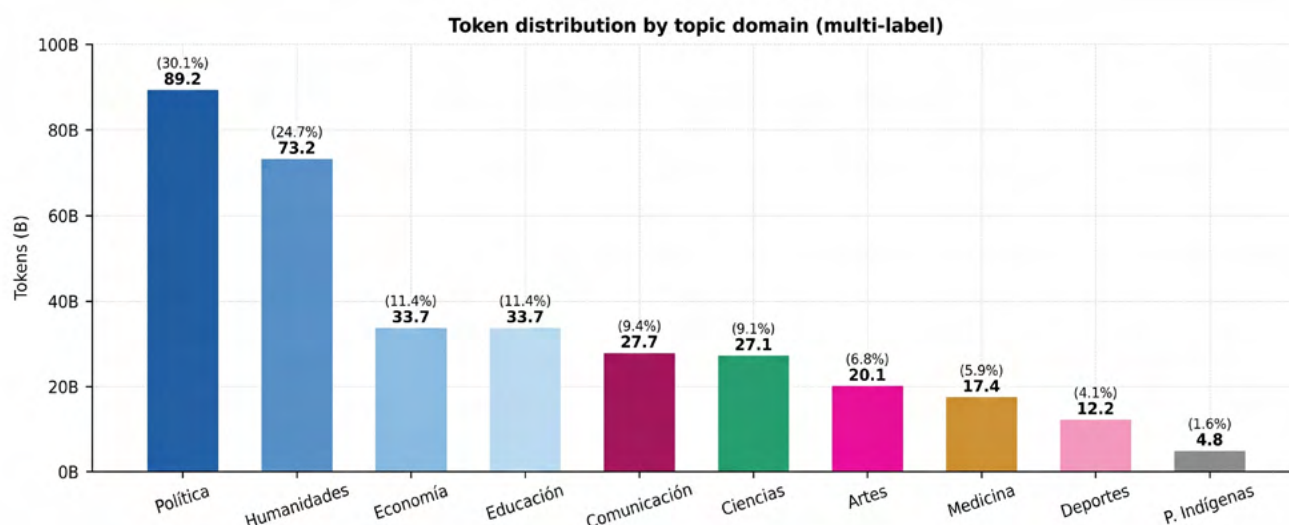


Figura 1 / Distribución por tópicos de tokens del dataset de preentrenamiento de LatamGPT.

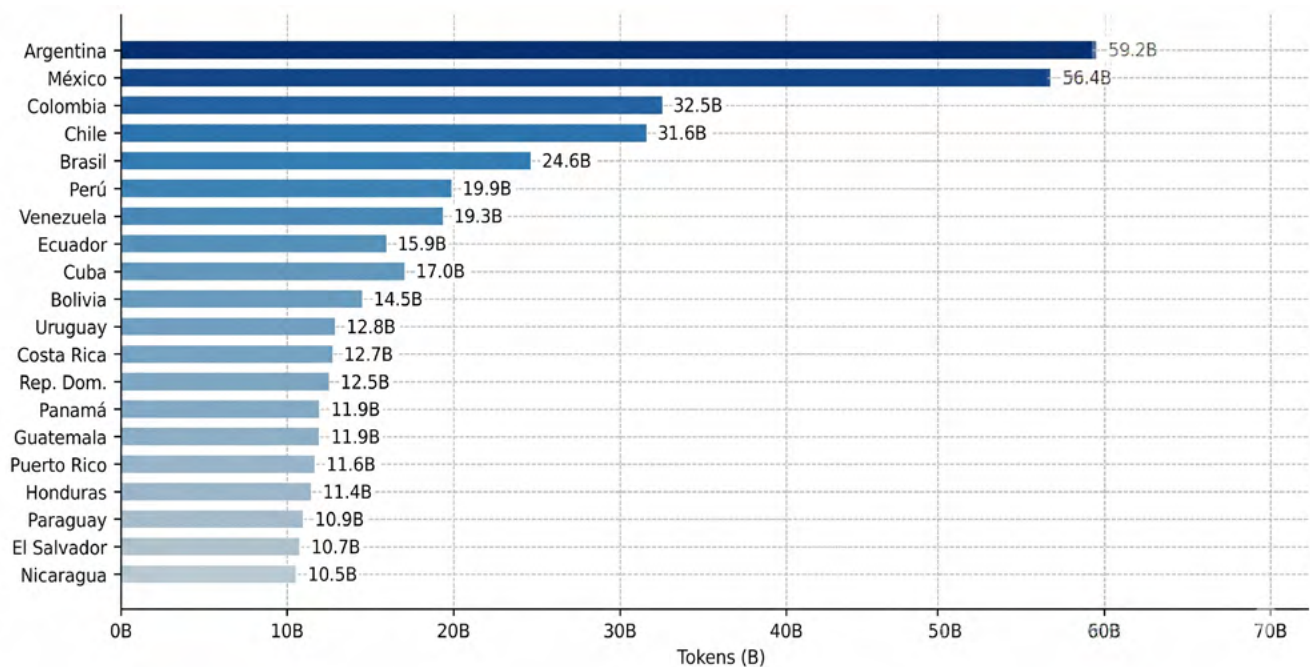


Figura 2 / Distribución de tokens asociados a los 20 países cubiertos por el corpus de LatamGPT. La clasificación geográfica es multi-etiqueta, por lo que un mismo documento puede estar asociado a más de un país y las cantidades no deben interpretarse como una partición excluyente.

Construyendo el corpus: 300 mil millones de tokens desde 79 instituciones

El corazón de cualquier modelo de lenguaje es su corpus de entrenamiento. Para LatamGPT, construir ese corpus fue una tarea que duró más de un año e involucró acuerdos formales de colaboración con 79 instituciones en 15 países de la región.

El corpus final suma 300 mil millones de tokens, unidad de medida que equivale aproximadamente a una sílaba o palabra corta. De ese total, 76,1 mil millones de tokens, casi el 20%, provienen directamente de instituciones: universidades, archivos nacionales, bibliotecas públicas, organismos gubernamentales y organizaciones de la sociedad civil que firmaron acuerdos de transferencia de datos con CENIA. El 80% restante proviene de Common Crawl, pero filtrado y seleccionado específicamente para representar el español y el portugués latinoamericanos.

La dimensión institucional es lo que distingue a LatamGPT de una simple selección del corpus web. Documentos jurídicos, producción académica regional, registros administrativos públicos, publicaciones de prensa histórica y literatura digitalizada de países con escasa presencia en Internet son fuentes que no existen en Common Crawl. Obte-

nerlas requirió establecer una amplia red de colaboración con instituciones por toda la región, firmar memorandos de entendimiento, establecer protocolos de privacidad, y en muchos casos, colaborar en la digitalización de materiales impresos. La Figura 1 muestra la distribución de temas resultante de este esfuerzo.

Esta estrategia de recolección de datos tiene consecuencias directas sobre el tipo de conocimiento que el modelo puede adquirir. A diferencia de modelos entrenados exclusivamente sobre texto web, LatamGPT ha sido expuesto a registros formales del derecho latinoamericano, a la producción científica de universidades de la región en sus idiomas originales, y a fuentes primarias de la historia de nuestros países. Eso no garantiza que el modelo sea perfecto, pero sí que conozca el mundo latinoamericano desde adentro.

La distribución geográfica del corpus refleja la desigualdad digital de la región (ver Figura 2). Bajo una clasificación geográfica multi-etiqueta, México y Argentina aparecen entre los países con mayor presencia relativa; Chile y Brasil también muestran una presencia significativa. Como un mismo documento puede estar asociado a más de un país, estas cifras no deben leerse como porcentajes excluyentes del corpus total. Los países más pequeños de Centroamérica y el Caribe tienen una presencia significativamente menor,

Con una tasa [de aprendizaje] conservadora, [...] el modelo era capaz de retener casi todo su conocimiento original e incorporaba el conocimiento latinoamericano de manera estable.

lo que indica que la brecha de representación dentro de América Latina es tan importante como la brecha entre la región y el mundo de habla inglesa.

El proceso de preparación del corpus incluyó etapas de limpieza, eliminación de duplicados, filtrado de información personal y clasificación automática de los documentos por país, idioma y dominio temático. Cada uno de estos pasos está documentado públicamente para que otras iniciativas regionales puedan replicarlo.

Preentrenamiento: aprender sin olvidar

Entrenar un modelo de lenguaje desde cero requiere cantidades de cómputo que están fuera del alcance de la mayoría de las instituciones académicas del mundo. Para LatamGPT, optamos por un enfoque diferente. Tomamos un modelo ya entrenado, el Llama 3.1 70B de Meta [3], y continuamos su entrenamiento sobre datos latinoamericanos. Esta estrategia se llama preentrenamiento continuado (*Continual Pre-Training* o CPT, por sus siglas en inglés) [4].

La motivación es exclusivamente de eficiencia. Aunque el proceso parte de un modelo existente, no corresponde a un simple *fine-tuning*: en términos de capacidades técnicas, infraestructura y desafíos de ingeniería, su complejidad es comparable a la de entrenar un modelo desde cero. Llama 3.1 ya aprendió gramática, razonamiento general, matemáticas y conocimiento del mundo durante un entrenamiento masivo. LatamGPT no necesita reaprender todo eso; sólo necesita adquirir el conocimiento específico de América Latina que Llama no tiene. Es como pedirle a un académico con amplísima formación general que realice una residencia intensiva en cultura y conocimiento latinoamericano.

El problema técnico central de un enfoque CPT es lo que en IA se llama olvido catastrófico. Cuando un modelo aprende nueva información de manera agresiva, puede degradar el conocimiento que ya tenía. Imaginemos un estudiante que, en su esfuerzo por memorizar los presidentes de todos los

países latinoamericanos, empieza a olvidar o confundir otros conceptos relevantes de la historia.

El parámetro más crítico para controlar este balance es la tasa de aprendizaje, que determina la velocidad con que el modelo ajusta sus parámetros internos frente a cada nuevo dato. Con una tasa demasiado alta, el modelo era capaz de absorber conocimiento latinoamericano muy rápidamente, pero pagaba un precio enorme. Su desempeño en pruebas generales de conocimiento en inglés caía hasta 23 puntos porcentuales respecto al punto de partida. Era un modelo más latinoamericano pero al mismo tiempo perdía capacidades previas. Con una tasa más conservadora, en cambio, el modelo era capaz de retener casi todo su conocimiento original e incorporaba el conocimiento latinoamericano de manera estable. Ese fue el punto óptimo que adoptamos para LatamGPT.

Para evaluar los resultados, usamos dos benchmarks diseñados específicamente para la realidad latinoamericana. Trueque [5] mide conocimiento factual sobre cultura, historia y geografía de América Latina en español; Choclo [6] evalúa conocimiento regional más específico sobre los 20 países cubiertos. A diferencia de MMLU o HellaSwag, que evalúan conocimiento general en inglés [7, 8], estos benchmarks fueron contruidos con preguntas sobre presidentes, escritores, platos típicos, accidentes geográficos y eventos históricos de la región, en el idioma en que esa información existe de manera natural.

Los resultados muestran que, en ambos benchmarks, LatamGPT supera consistentemente al modelo base Llama 3.1 y a modelos de tamaño similar del estado del arte como DeepSeek 67B [9], Gemma 2 27B [10] y Qwen 2.5 72B [11]. En Trueque, CPT Fase 1 alcanza 0.53 sin ejemplos de referencia, 15 puntos sobre el baseline de Llama 3.1, mientras que CPT Fase 0 alcanza 0.60 con ejemplos, 9 puntos sobre el mismo baseline (ver Figura 3). En Choclo, CPT Fase 1 obtiene 0.33 sin ejemplos y 0.35 con ellos, mejoras absolutas de 9 y 7 puntos sobre Llama 3.1, respectivamente (ver Figura 4). Crucialmente, estas ganancias se obtienen sin degradar el desempeño en benchmarks generales. En MMLU y HellaSwag, LatamGPT Fase 1 con la tasa de aprendizaje conservadora se mantiene dentro de 0,2 puntos porcentuales del modelo base, lo que confirma que aprendió conocimiento latinoamericano sin olvidar lo que ya sabía.

Postentrenamiento: aprender a seguir instrucciones

Un modelo preentrenado adquiere mucho conocimiento pero le falta seguir instrucciones. Lo que aprendió en la

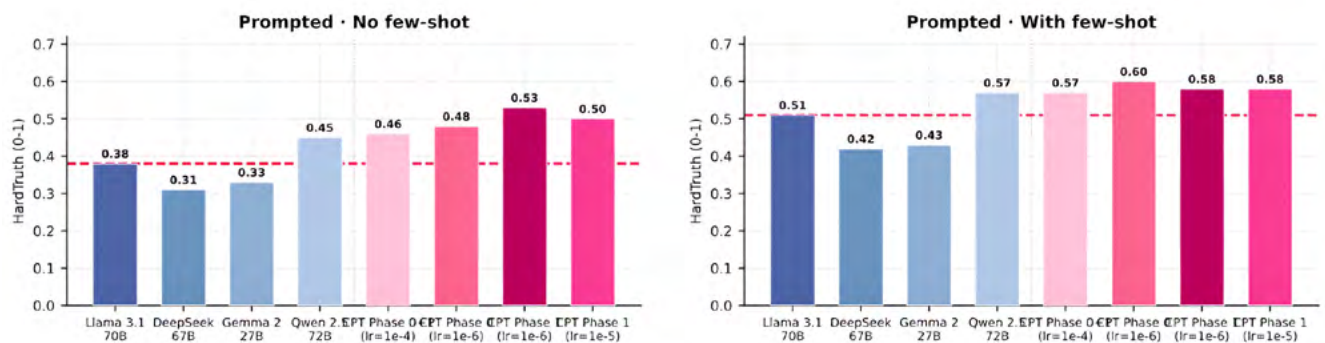


Figura 3 / Comparación de resultados en el benchmark Trueque para el modelo base de LatamGPT y modelos del estado del arte de tamaño similar.

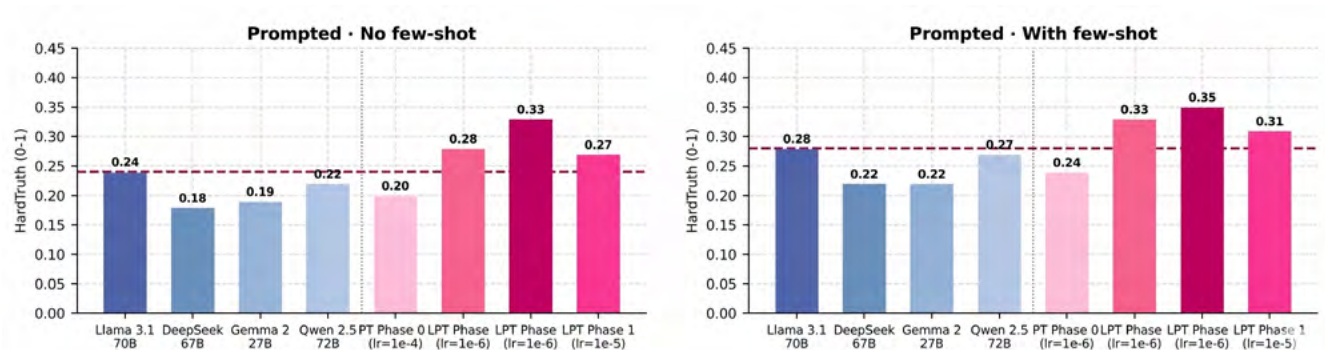


Figura 4 / Comparación de resultados en el benchmark Choclo para LatamGPT y modelos del estado del arte de tamaño similar.

etapa anterior es a predecir la siguiente palabra de un texto, una habilidad poderosa pero distinta a la de responder preguntas o sostener un diálogo útil.

Para convertir ese modelo base, producto de la fase de pre-entrenamiento, en un asistente con el que se pueda conversar, hace falta una segunda etapa de entrenamiento conocida como postentrenamiento o ajuste fino supervisado (SFT, por sus siglas en inglés).

En esta etapa el modelo aprende observando miles de ejemplos de conversaciones, donde un usuario hace una pregunta y un asistente responde con la respuesta deseada. El desafío para LatamGPT fue construir una colección de conversaciones que reflejara mejor los contextos, temas, formas de expresión y necesidades de uso propias de América Latina. Para esto, exploramos ocho fuentes distintas de datos antes de definir la mezcla final, comparando experimentalmente el aporte de cada una de ellas.

Como punto de partida usamos corpus abiertos de referencia, en particular fragmentos de Tülu 3 y Dolci, dos colecciones de instrucciones publicadas por el Allen Institute for

AI [12] que liberan datos, código y pesos de manera completa. Esa transparencia nos permitió establecer una línea base estable contra la cual medir el aporte de cada nueva fuente.

Una hipótesis razonable era que traducir conversaciones del inglés al español podía ofrecer una vía rápida y de bajo costo para obtener datos conversacionales en nuestro idioma. Sin embargo, los experimentos mostraron los límites de esa estrategia. Al entrenar con versiones traducidas de corpus como SmolTalk o UltraChat [13, 14], el modelo no mejoraba en los benchmarks regionales. La traducción preservaba parte de la estructura lingüística, pero no necesariamente la referencia cultural ni el contexto de uso. A modo de ejemplo, una conversación traducida sobre béisbol estadounidense no le enseñaba al modelo nada particularmente útil sobre fútbol latinoamericano. La excepción fue el dominio de seguridad, entendido como la capacidad del modelo de responder de forma segura, evitar toxicidad y rechazar solicitudes inapropiadas, donde esos patrones sí se transferían razonablemente bien entre idiomas.

Para enfrentar esta limitación, una contribución distintiva del proyecto fue Copuchat, una plataforma pública operada por

CENIA que entrega acceso gratuito a un modelo conversacional avanzado y, al mismo tiempo, permite construir datos conversacionales más cercanos a los usos reales de la región. De manera voluntaria y anonimizada, los usuarios pueden donar sus conversaciones, siguiendo una lógica similar a la de WildChat [15]. Esto permitió complementar los corpus existentes con interacciones más naturales, situadas en contextos latinoamericanos y alineadas con las formas concretas en que las personas de la región usan estos sistemas.

Estas conversaciones reales capturan patrones de uso regionales que los corpus importados no contienen, como preguntas en jerga local, consultas administrativas específicas de cada país o referencias culturales que un usuario latinoamericano da por sentadas. Aún así, ni los corpus abiertos ni Copuchat estaban necesariamente concentrados en contenido cultural latinoamericano. Para cubrir ese vacío generamos sintéticamente cadenas de preguntas y respuestas a partir de documentos cuidadosamente seleccionados.

Un experimento piloto con 240 conversaciones sobre cocina chilena, generadas a partir de artículos de Wikipedia, confirmó que un slice pequeño y temáticamente concentrado es capaz de mejorar la precisión del modelo en el dominio correspondiente. Escalamos esa idea a toda la Wikipedia iberoamericana y produjimos wikipedia-qa-iberoamerica, una colección de 100 mil conversaciones balanceadas por país y por tema.

Otra dificultad que tuvimos que enfrentar fue la brevedad del corpus conversacional en español. Las conversaciones disponibles tienden a ser cortas, lo que limita la capacidad del modelo para trabajar con documentos extensos y sostener interacciones de mayor profundidad. Para abordar este problema, usamos el método LongMagpie [16] y generamos cadenas largas de instrucciones sobre documentos extraídos de RedPajama-HQ-ES, una versión filtrada en español del corpus RedPajama [17].

Además, las preguntas que formula un usuario dependen de su trayectoria, profesión y circunstancias de vida. Un estudiante de medicina no interactúa con un modelo de la misma manera que un agricultor, una periodista o una funcionaria pública. Para enriquecer esa diversidad, construimos un conjunto de personas sintéticas representativas de América Latina y generamos conversaciones condicionadas a cada perfil, siguiendo el enfoque de PersonaHub [18].

Finalmente, cerramos la mezcla de datos con dos consideraciones adicionales:

La primera fue la seguridad. Además del dataset traducido ya mencionado, incorporamos Safety-SFT-50K-Sample, un

Un modelo preentrenado adquiere mucho conocimiento pero le falta seguir instrucciones.

subconjunto diseñado para calibrar la capacidad del modelo de rechazar solicitudes inapropiadas sin caer en un comportamiento excesivamente restrictivo.

La segunda fue la identidad. Como LatamGPT parte de Llama 3.1, inicialmente el modelo respondía que era Llama cuando se le preguntaba por su nombre u origen. Para reasignar esa identidad, incorporamos conversaciones en las que el usuario consulta por estos datos y el asistente responde como LatamGPT. Puede parecer un detalle menor, pero en un proyecto regional que busca afirmar una identidad propia, importa.

Tras todos los experimentos, la mezcla final de postentrenamiento incluyó Tulu-3-SFT-Mixture-No-Identity [12] como base general; Wikipedia-QA-Iberoamerica [6] para reforzar conocimiento cultural regional; Synthetic-Identity para la reasignación de identidad; los dos subconjuntos de seguridad ya mencionados; y dos conjuntos adicionales de preguntas de opción múltiple, MMLU Auxiliary Train [7] y Synthetic Latam MCQ, que mejoraron la capacidad del modelo para responder en formatos de evaluación estructurados.

El resultado es un modelo que incorpora conocimiento sobre América Latina y puede responder preguntas situadas en la región, manteniendo coherencia, utilidad y seguridad a lo largo de una conversación.

Conclusiones

Las decisiones sobre qué datos usar, qué idiomas priorizar y cómo medir el éxito de un modelo no son decisiones técnicas neutras. Determinan qué regiones del mundo estarán representadas en los sistemas de IA del futuro y cuáles quedarán dependiendo de tecnología diseñada para otros.

LatamGPT es una apuesta concreta por reducir esa asimetría. Un modelo entrenado con datos latinoamericanos, evaluado en benchmarks regionales y desarrollado por equipos de la región permite fortalecer capacidades propias en un ámbito estratégico. No se trata de rechazar la colaboración internacional, sino de participar en ella desde una posición más sólida, con la capacidad de producir, adaptar y evaluar tecnología de IA según nuestras propias realidades.



El proyecto también tiene limitaciones. Las lenguas indígenas de América Latina, más de 400 con unos 50 millones de hablantes, todavía no forman parte del corpus actual. La distribución geográfica del corpus además, refleja en parte las desigualdades digitales de la región. En CENIA ya estamos trabajando para cerrar estas brechas, a modo de ejemplo, ya hemos desarrollado traductores para Rapa Nui y Mapudungun, que esperamos integrar a LatamGPT en el futuro cercano.

La brecha de representación latinoamericana en la IA no es inevitable. Es el resultado de decisiones concretas sobre qué datos recopilar y cómo entrenar modelos. Recopilar datos latinoamericanos, entrenar modelos sobre ellos, evaluarlos con benchmarks locales y publicar los resultados para que otros construyan sobre ese trabajo es exactamente lo que estamos haciendo. La historia de quién produce inteligencia artificial en el mundo todavía se está escribiendo, y América Latina tiene algo que decir. **B**

Referencias

- [1] Baack, S. A Critical Analysis of the Largest Source for Generative AI Training Data: Common Crawl. In Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency (FAccT '24), 2024. <https://doi.org/10.1145/3630106.3659033>.
- [2] Centro Nacional de Inteligencia Artificial. Índice Latinoamericano de Inteligencia Artificial (ILIA) 2024. CENIA, 2024. <https://indicelatam.cl>.
- [3] Grattafiori, A. et al. The Llama 3 Herd of Models. arXiv preprint arXiv:2407.21783, 2024.
- [4] Ibrahim, A. et al. Simple and Scalable Strategies to Continually Pre-train Large Language Models. arXiv preprint arXiv:2403.08763, 2024.
- [5] Fuentes, G., Arriagada, A., Henríquez, C., García, M. A., and LatamGPT Team. Trueque: A Human-Reviewed Collaborative Benchmark for Latin American Knowledge and Culture. Version 0.1, 2026. <https://huggingface.co/datasets/latam-gpt/Trueque-Benchmark-beta-0.1>.
- [6] LatamGPT Team. CHOCLO: Latin American Cultural Knowledge Benchmark. 2026. <https://huggingface.co/datasets/latam-gpt/CHOCLO>.
- [7] Hendrycks, D., Burns, C., Basart, S., Zou, A., Mazeika, M., Song, D., and Steinhardt, J. Measuring Massive Multitask Language Understanding. arXiv preprint arXiv:2009.03300, 2020. Published at ICLR 2021.
- [8] Zellers, R., Holtzman, A., Bisk, Y., Farhadi, A., and Choi, Y. HellaSwag: Can a Machine Really Finish Your Sentence? In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 2019. arXiv:1905.07830.
- [9] DeepSeek-AI. DeepSeek LLM: Scaling Open-Source Language Models with Longtermism. arXiv preprint arXiv:2401.02954, 2024.
- [10] Gemma Team. Gemma 2: Improving Open Language Models at a Practical Size. arXiv preprint arXiv:2408.00118, 2024.
- [11] Bai, J. et al. Qwen Technical Report. arXiv preprint arXiv:2309.16609, 2023.
- [12] Lambert, N. et al. Tulu 3: Pushing Frontiers in Open Language Model Post-Training. arXiv preprint arXiv:2411.15124, 2024.
- [13] Ben Allal, L. et al. SmoLLM2: When Smol Goes Big—Data-Centric Training of a Small Language Model. arXiv preprint arXiv:2502.02737, 2025.
- [14] Ding, N. et al. Enhancing Chat Language Models by Scaling High-Quality Instructional Conversations. arXiv preprint arXiv:2305.14233, 2023.
- [15] Zhao, W., Ren, X., Hessel, J., Cardie, C., Choi, Y., and Deng, Y. WildChat: 1M ChatGPT Interaction Logs in the Wild. arXiv preprint arXiv:2405.01470, 2024.
- [16] Gao, C., Wu, X., Lin, Z., Zhang, D., and Hu, S. LongMagpie: A Self-Synthesis Method for Generating Large-Scale Long-Context Instructions. arXiv preprint arXiv:2505.17134, 2025.
- [17] Weber, M. et al. RedPajama: An Open Dataset for Training Large Language Models. arXiv preprint arXiv:2411.12372, 2024.
- [18] Ge, T., Chan, X., Wang, X., Yu, D., Mi, H., and Yu, D. Scaling Synthetic Data Creation with 1,000,000,000 Personas. arXiv preprint arXiv:2406.20094, 2024.