Gira Presidencial, Abril 2025

Modelo de Lenguaje Latam-GPT en Brasil

Resumen

Tras la participación del Ministerio de Ciencia en el IA Summit en París en febrero del 2025, se anunció el **lanzamiento del Latam GPT**, modelo de lenguaje para inteligencia artificial diseñado por el Centro Nacional de Inteligencia Artificial (CENIA) específicamente para representar y responder a las necesidades de América Latina y el Caribe, incorporando el idioma, la cultura y el conocimiento de la región. A diferencia de otros modelos globales, LatamGPT entrena con datos representativos de los países de la región, asegurando una mejor comprensión del contexto local, con menor sesgo y mayor precisión en sus respuestas.

Este año, LatamGPT **será accesible de manera abierta, fortaleciendo la soberanía tecnológica de la región y permitiendo su aplicación en diversos ámbitos**. Entre los hitos clave para 2025, destacan:

- La culminación de su entrenamiento y alineamiento por parte de CENIA en dos versiones del modelo, asegurando respuestas éticas y alineadas con los valores de la región.
- En conjunto con Municipios y SUBDERE, el desarrollo de propuestas para nuevas aplicaciones, incluyendo herramientas para la gestión pública, la educación y la traducción automática, entre otras.

De esta manera, LatamGPT no solo representa un avance tecnológico, sino también una herramienta que democratiza el acceso a la inteligencia artificial, ofreciendo una tecnología más cercana, útil y adaptada a la realidad de las personas en América Latina y el Caribe.

Respecto a Brasil, **se planea la firma de un Memorándum de Entendimiento de cooperación en materia de IA de modelos de lenguaje latinoamericano (LATAM-GPT) entre los Ministerios de Ciencia, Tecnología, Conocimiento e Innovación y el Ministerio de Ciencia, Tecnología e Innovación de Brasil en el marco de la visita presidencial del 21 al 24 de abril 2025.** Este será un hito ya que es el primer acuerdo político que firmará Chile con otro país de Latinoamérica y el Caribe y Brasil es líder en la región en esta materia.

Por su lado CENIA no tiene acuerdos aún con ninguna institución o académico/a de dicho país y **se prevé la firma de un Protocolo de Intenciones para la Cooperación Internacional entre CENIA y la Universidad de Sao Paulo.** Se deja un listado de instituciones brasileñas que podrían ser de utilidad.

Detalles de LATAM GPT:

El proyecto "Gran Modelo de Lenguaje Latinoamericano - LATAM GPT" es una iniciativa colaborativa y de carácter abierto (*Open Sourc*e), iniciada en el 2023 y liderada por CENIA, con la participación de instituciones y gobiernos de toda América Latina y el Caribe. Su objetivo es desarrollar un modelo de lenguaje que capture el vasto conocimiento cultural, social y lingüístico de la región, sirviendo como base para aplicaciones en educación, políticas públicas, preservación de lenguas no hegemónicas y otros sectores estratégicos.

Para lograr este propósito, se utilizarán grandes volúmenes de datos de alta calidad provenientes de diversas áreas del conocimiento y la vida social de la región. Estos datos serán curados y extraídos de la web, con un importante aporte de colaboradores de distintos países latinoamericanos. Este proyecto busca fortalecer la soberanía tecnológica de la región, promover la colaboración científica y fomentar el desarrollo de capacidades en inteligencia artificial con un enfoque en la inclusión, la equidad y la preservación del patrimonio cultural latinoamericano.

Todos los modelos existentes han sido desarrollados en el norte global (EE.UU., China y Europa) debido a tres factores principales:

1. **Datos**: Entrenar un modelo de lenguaje (LLM) requiere una cantidad masiva de datos. Por ejemplo, GPT-3 fue entrenado con 300 mil millones de tokens (225 mil millones de palabras) y 175 mil millones de parámetros.

2. **Cómputo**: El costo de entrenamiento es extremadamente alto. Stanford estima que entrenar Gemini Ultra costó alrededor de 191 millones de dólares, reflejando el dominio de empresas con gran capacidad computacional.

3. **Talento especializado**: Se requiere un equipo interdisciplinario de expertos en datos, privacidad, lingüística, historia, antropología y computación para desarrollar y gestionar un LLM.

Desarrollar un modelo de lenguaje latinoamericano es un desafío comparable a una misión a Marte: exige conocimientos técnicos avanzados, una inversión significativa y un equipo altamente capacitado. Sin embargo, el impacto potencial de contar con un LLM propio supera con creces estos costos, permitiendo a la región:

- Generar conocimiento adaptado a su contexto cultural y lingüístico.
- Fomentar la soberanía tecnológica y la independencia en IA.
- Potenciar la educación, la ciencia y la innovación con herramientas propias.

LATAM GPT es un paso crucial hacia un futuro donde América Latina y el Caribe jueguen un rol central en la revolución de la inteligencia artificial, impulsando la equidad, la diversidad y el acceso al conocimiento en toda la región.

¿Por qué es importante el LLM?

Existen muchas alternativas Open Source desarrolladas en el norte global que pueden ser utilizadas para construir soluciones en la región. Entonces, ¿por qué desarrollar nuestro propio modelo en lugar de usar los ya existentes? Hay tres motivos principales:

1. **Desarrollo de conocimiento y capacidades técnicas**: Si bien el acceso a modelos abiertos ha permitido estudiar su funcionamiento, construir un LLM propio permite desarrollar conocimientos prácticos y generar capacidades técnicas locales. Este proceso fortalece la comunidad científica y técnica de la región, proporcionando una base sólida para futuras innovaciones.

2. **Soberanía tecnológica y regulación**: La regulación de la IA a nivel global se está definiendo en organismos como la ONU y la OCDE. América Latina y el Caribe deben tener una voz legítima en este debate, lo que requiere un entendimiento profundo del desarrollo y las implicancias de estos modelos. Contar con un LLM propio permitirá influir en la creación de políticas públicas adaptadas a las necesidades de la región.

3. **Preservación de la cultura y representación regional**: Los modelos actuales han sido entrenados con datos predominantemente del norte global, dejando de lado la riqueza cultural, lingüística y social de América Latina y el Caribe. Esto tiene un impacto directo en la representación de la región en el conocimiento digital y en aplicaciones prácticas como la educación, la justicia y el desarrollo sostenible. Un LLM regional garantizará que la diversidad de la región esté representada en la IA del futuro.

La importancia de esta representación se refleja en aplicaciones concretas. Por ejemplo:

- Un asistente educativo basado en un modelo entrenado en la región podrá comprender mejor los contextos culturales y lingüísticos de los estudiantes latinoamericanos.
- Un sistema de apoyo para la justicia podrá interpretar correctamente las jerarquías legales de los países de la región, evitando sesgos provenientes de sistemas jurídicos ajenos.
- Modelos entrenados con información detallada sobre la Amazonía o la Patagonia podrán contribuir de manera efectiva a la lucha contra la crisis climática.

El desarrollo de LATAM GPT es una apuesta estratégica para asegurar que América Latina y el Caribe no sólo sean consumidores de tecnología, sino también creadores activos en la revolución de la inteligencia artificial.

Desarrollo de aplicativos de software

LatamGPT será un modelo abierto disponible en plataformas opensource para el uso y goce en aplicaciones diferentes que serán definidas por comunidades de desarrollo. En ese sentido, preliminarmente no existirán aplicaciones de acceso público generadas en este proyecto, pero podrían construirse a partir de la arquitectura abierta propuesta. Para esta construcción se requiere el desarrollo de aplicativos de software y acceso a nube pública o capacidad de cómputo para los procesos de inferencia y costos recurrentes que derivan de la existencia de estas soluciones. Preliminarmente identificamos potenciales aplicaciones para las que LatamGPT podría tener un desempeño más robusto que otros modelos del estado del arte provenientes del norte global.

- **Govtech**: El conocimiento de LatamGPT sobre la cultura, idiosincrasia y procesos políticos y sociales sobre América Latina y el Caribe será más profundo que el de otros modelos debido al volumen de datos sobre Latinoamérica y el Caribe, y la importancia relativa de los mismos en los parámetros de entrenamiento.

  En ese sentido, aplicaciones orientadas al fortalecimiento de procesos de formulación, seguimiento y evaluación de políticas públicas en la región podrían mostrar un mejor desempeño y mayor precisión con esta herramienta como motor de IA que otras alternativas disponibles.

  Además, el hecho de que sea abierto permitirá a organismos públicos de la región la descarga y refinamiento del modelo en entornos digitales locales controlados, lo cual es óptimo para procesos que utilizan datos sensibles o son estratégicos.

- **Investigación universitaria de IA en los procesos educativos:** LatamGPT, al estar entrenado con datos y parámetros que reflejan de mejor manera los procesos propios de América Latina y el Caribe, tiene el potencial de ofrecer respuestas más precisas, con menos alucinaciones y resultados eventualmente tóxicos, además de mayor representación del conocimiento académico y científico generado en la región.

  Cenia está desarrollando un piloto con la universidad más grande en términos de matrícula de Chile para una integración estratégica de IA en investigación académica, identificando oportunidades de desarrollo de herramientas desarrolladas específicamente para procesos de generación de conocimiento con foco en América Latina y el Caribe. Este proceso podría escalarse para transformar estas herramientas como bienes públicos disponibles para investigadores/as e instituciones en la región.

- **Integración de lenguas no hegemónicas**: Cenia ha desarrollado el primer traductor **Rapa Nui - Español con IA,** y está desarrollando otro Mapudungun - Español. Ambos proyectos se han ejecutado en estrecha colaboración con las autoridades ancestrales locales, levantando participación activa de las comunidades desde los territorios para empoderar a quienes usualmente

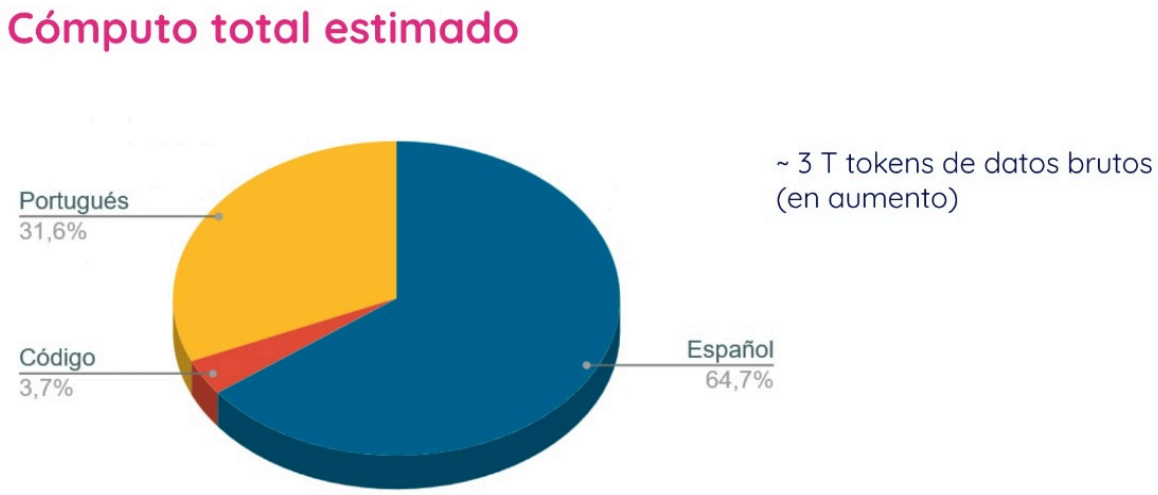
están marginados de los procesos de desarrollo tecnológico y promover la preservación de lenguas no hegemónicas.

Este proyecto está escalando para disponibilizar una solución que paulatinamente integrar otras lenguas indígenas de la región, como Guaraní y Aimara desde Perú y Paraguay, donde CENIA ya está trabajando.

Estos desarrollos podrían integrarse a LatamGPT, tanto para promover su usabilidad en comunidades marginadas como también para integrar conocimiento ancestral en el dataset.

Antecedentes para reunión con Ministra de Ciencia, Brasil

Actualmente el LatamGPT se encuentra alimentado en un **31% de datos en portugues** (esto mediante técnicas de búsqueda de datos abiertos y disponibles. Sin embargo, CENIA no cuenta con ningún convenio vigente con dicho país, con ninguna institución ni académico/a.



Considerando la relevancia estratégica de Brasil en el ecosistema científico y digital de América Latina, se deja una **lista sugerida de instituciones brasileñas** que podrían ser contrapartes ideales para la provisión de datos culturales, lingüísticos, científicos y sociales para el entrenamiento del modelo LatamGPT.

Universidades y centros de investigación

1. **Bibliotecas del Congreso Nacional de Brasil.** La Biblioteca Académico Luiz Viana Filho, del Senado Federal, fue una de las primeras que surgieron en la época del Imperio brasileño y tiene un acervo especializado en Ciencias Sociales. La Biblioteca Pedro Aleixo integra el Centro de Documentación e Información de la Cámara de Diputados y tiene un acervo de cerca de 200.000 volúmenes: una de las mayores bibliotecas de Brasilia y de Brasil.

2. **Fundação Getulio Vargas (FGV):** Especialista en políticas públicas, derecho, historia y ciencias sociales. Cuenta con bases de datos y observatorios de transparencia.
3. **Universidade Estadual de Campinas (UNICAMP):** Referente en investigación computacional, lingüística y tecnología educativa. Facultad de Ciencias Sociales y de Lingüística podrían aportar corpus especializados.
4. **Fundação Oswaldo Cruz (Fiocruz):** Institución de excelencia en salud pública con vastos repositorios de datos abiertos en salud, educación y ciencias sociales. Posee colecciones en portugués de alta calidad.
5. **Instituto Nacional de Pesquisas da Amazônia (INPA):** Relevante para la recolección de datos ambientales y culturales vinculados a pueblos amazónicos e indígenas.

Instituciones gubernamentales y archivos nacionales

6. **Instituto Brasileiro de Geografia e Estatística (IBGE):** Ofrece datos censales, sociodemográficos, territoriales y económicos. Muchos de sus conjuntos están disponibles en formatos abiertos.

7. **Arquivo Nacional do Brasil:** Repositorio central de documentos históricos del Estado brasileño, con colecciones digitalizadas disponibles para consulta.

8. **Biblioteca Nacional do Brasil:** Una de las mayores bibliotecas del hemisferio sur. Contiene textos literarios, científicos y culturales representativos del país.

9. **Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP):** Datos educativos, pruebas nacionales (como ENEM) y estudios técnicos sobre el sistema educativo brasileño.

Museos y centros culturales

10. **Museu da Língua Portuguesa (São Paulo):** Recurso clave para estudios sobre el portugués brasileño y sus variantes regionales.
11. **Museu Nacional (UFRJ):** A pesar del incendio de 2018, conserva colecciones digitalizadas de antropología, arqueología y biodiversidad.
    Instituto Moreira Salles (IMS): Posee archivos fotográficos, musicales, literarios y de prensa con foco en la identidad cultural brasileña.
12. **Fundação Casa de Rui Barbosa:** Importante acervo digital de literatura, historia y ciencias sociales, con enfoque en la cultura brasileña.

ONGs y organizaciones temáticas

14. **Instituto Socioambiental (ISA):** Especializado en pueblos indígenas y comunidades tradicionales. Tiene atlas, registros lingüísticos y mapas interactivos.
    InternetLab: Centro de investigación independiente sobre derecho,

tecnología y sociedad. Puede aportar datos sobre regulación digital y desinformación.

15. **Transparência Brasil:** Datos sobre gobierno abierto, corrupción, acceso a la información y control ciudadano.
16. **DataLabe:** Laboratorio de datos con enfoque en justicia social y análisis territorial. Trabaja con datos abiertos y accesibles desde las periferias.