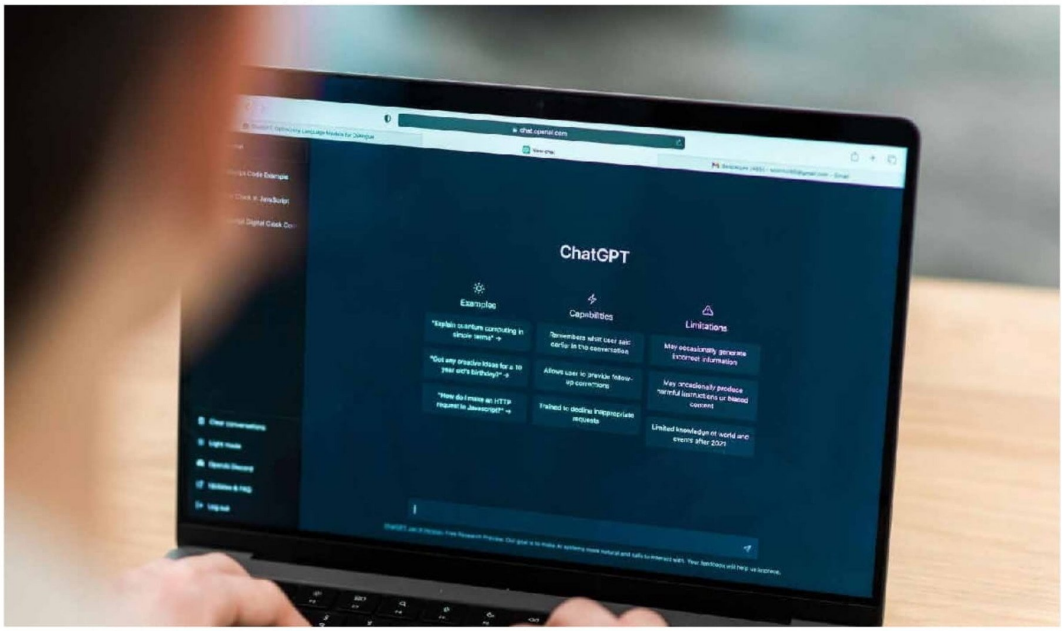
LA TERCERA
Fecha: 12/08/2025 Audiencia: 271.020 Sección: NACIONAL
Vpe: $7.238.266 Tirada: 76.017 Frecuencia: DIARIO
Vpe pág: $9.829.612 Difusión: 76.017
Vpe portada: $9.829.612 Ocupación: 73,64%

LatamGPT: la IA latinoamericana que se prepara desde Chile alimentada con más de mil millones de documentos

El desarrollo de esta inteligencia artificial, un modelo de lenguaje abierto y entrenado con datos de la región, busca representar las culturas, lenguas e historias de América Latina. Será lanzada a fines de este año para operar desde Chile hacia toda la región.

Francisco Corvalán



▶ LatamGPT busca reflejar la idiosincrasia de Latinoamérica en dicha herramienta operada con IA.

Los modelos de lenguaje actuales de inteligencia artificial (IA) son generados principalmente en el norte global del mundo, donde se construyen en base a datos de entrenamiento que no necesariamente reflejan la cultura, lenguaje e historia de los países al sur del ecuador. Es por esto que desde Chile se impulsó crear un sistema propio que mejore la precisión y representatividad de los datos arrojados al momento de ingresar un prompt.

Hasta ahora, por ejemplo, cuesta que un chatbot que utilice IA pueda realizar análisis profundo de temas relacionados con la idiosincrasia local. Cuesta también que las fuentes con que se alimenta este motor de lenguaje incluya conocimiento y cosmovisión de pueblos indígenas.

Es por esto que, con miras a desarrollar una primera versión de LatamGPT, el Centro Nacional de Inteligencia Artificial (Cenia), creado por el Estado, y el Data Observatory (DO), ONG sin fines de lucro, trabajan en la depuración de información en español, portugués e inglés que permitirán hacer de la herramienta una plataforma robusta, confiable y representativa de nuestra región.

Ambas organizaciones despliegan una hoja de ruta de ocho etapas antes de dar a luz la versión oficial de LatamGPT. Hoy, según comentan desde los organismos responsables, se ejecuta la tercera etapa de preparación de la información recogida de bibliodatos, instituciones de gobierno y universidades, entre otros. Esto permitirá, entre otras cosas, reflejar la idiosincrasia de Latinoamérica en dicha herramienta operada con IA.

El Data Observatory además participa de los desafíos técnicos de desarrollar este 'gran modelo de lenguaje' (LLM, por su sigla en inglés) de estas dimensiones. Allí colabora en proveer la enorme capacidad de cómputo requerida, y también en procesar el gran volumen de datos que existen en documentos para su entrenamiento.

Mauricio Leiva, ingeniero civil en informática y project manager del proyecto LatamGPT, señala que se han distribuido entre el Cenia y el DO el procesamiento de los datos que entrenarán la primera versión del modelo LLM generado por y para la región latinoamericana. Todo esto, financiado en primera instancia por créditos otorgados por Amazon Web Services (AWS).

El Cenia ya ha almacenado cerca de 500 gigabytes de datos, tanto en español como en portugués, y su objetivo es procesar en conjunto un total de 20,5 terabytes de datos públicos en inglés al cierre del proyecto. Estos datos son recogidos de RedPajama v2, un conjunto de datos abiertos utilizado en otros modelos como LLaMA de Meta AI, y que considera 30 mil millones de tokens o cadenas de palabras.

En la etapa actual, el Data Observatory procesa 2,5 terabytes de datos en inglés, lo que se traduce en más de mil millones de documentos correspondientes a datos web disponibles públicamente. Entre ellos hay datos obtenidos desde blogs y sitios de noticias, hasta artículos académicos y recursos educativos. Dichos documentos contemplan temáticas variadas como artes, ciencias, comunicación y medios, deportes, economía, educación, medicina, ciencias sociales o políticas.

"LatamGPT es mucho más que un proyecto tecnológico; es un hito para Latinoamérica. Estamos demostrando que la región puede liderar la construcción de IA con identidad propia, capaz de representar nuestras culturas, lenguas y realidades. Esta colaboración no sólo reúne capacidades técnicas y computacionales sin precedentes, como el procesamiento de más de mil millones de documentos, sino que marca el inicio de un ecosistema regional que genera tecnología de vanguardia sin perder de vista su raíz cultural", comenta Rodrigo Roa, director ejecutivo del DO.

Asimismo, agrega que mediante esta herramienta buscan posicionar a toda Latinoamérica "como un actor clave en la revolución de la IA, levantando soberanía tecnológica y conocimiento propio para el mundo". Si bien hay países de los cuales existe mucha información, también hay naciones del Caribe que cuentan con más información en fuentes angloparlantes, que también serán incorporados.

En los próximos pasos, DO trabajará en clasificar estos datos, con el fin seleccionar aquellos de mayor calidad y confiabilidad, para cada país y tópico para entrenar el modelo. Estos serán procesados y luego se etiquetarán y pondrán a disposición. En paralelo, LatamGPT incorporará constantemente nuevos datos para enriquecerse.

Pero ¿por qué es importante un GPT para América Latina? Según remarca Roa, no es solo un desarrollo tecnológico, sino "una forma de asegurar que la IA hable como nosotros, entienda nuestra cultura y con nuestra historia. Los modelos de empresas como OpenAI, Meta o Google están entrenados sobre todo con datos y realidades de otros países, y eso hace que muchas veces no nos representen bien, dejando entrever sesgos y errores evidentes".

El director del DO agrega que tener un LLM propio permite decidir qué conocimiento incluimos, cómo lo usamos y garantizar que la tecnología trabaje para nuestras necesidades, y no al revés. "Dejamos de ser simples usuarios pasando a convertirnos en creadores", agrega Roa.

Cabe destacar que LatamGPT será un modelo abierto, con infraestructura en nuestro país que será alojado y analizado en el Centro de Supercómputo de la U. de Tarapacá.

El proyecto también está orientado a fortalecer la soberanía tecnológica y promover la colaboración científica en la región. Se espera que esta aplicación sea lanzada en noviembre próximo. ●