



Catálogo Referencial: **LATAMGPT-CORPUS V1.0** LatamGPT



Agradecimientos

La liberación de LatamGPT v1.0 como el primer modelo de lenguaje abierto y soberano de Latinoamérica y el Caribe marca un punto relevante en la historia de soberanía digital de América Latina y el Caribe. Este logro no es solo el resultado de una excelencia técnica, sino el fruto de una visión compartida que se atrevió a dar un paso audaz e innovador en un entorno de grandes desafíos. Es, ante todo, un testimonio de lo que podemos alcanzar mediante la colaboración estratégica y un propósito común.

Este proyecto nace con la misión de democratizar el acceso a la tecnología, garantizando que las herramientas del mañana hablen nuestras lenguas y reflejen nuestra identidad, y, al mismo tiempo, desarrollar capacidades locales que contribuyan a una mayor soberanía tecnológica. Al promover el acceso abierto al conocimiento, LatamGPT v1.0 se constituye en un bien público digital orientado a reducir brechas y fortalecer a nuestras sociedades desde sus propias raíces culturales e históricas mediante la IA.

Reconocimientos

Expresamos nuestro más sincero agradecimiento a todas las organizaciones y aliados estratégicos que hicieron posible la custodia y el procesamiento de este patrimonio tecnológico:

1. Liderazgo, Coordinación e Infraestructura de Cómputo

Reconocemos el respaldo estratégico de las organizaciones que hicieron posible el desarrollo de LatamGPT mediante la provisión de infraestructura de cómputo, soporte técnico y financiamiento:

- **Amazon Web Services (AWS):** Por proveer la infraestructura de servidores de alto rendimiento indispensable para el pre-entrenamiento a gran escala, además del acompañamiento de expertos y expertas que han desarrollado modelos similares en otras latitudes.
- **Data Observatory (DO):** Por articular el apoyo de AWS a esta iniciativa, brindar soporte técnico especializado en tecnologías Cloud y gestionar los créditos AWS de infraestructura.
- **Banco de Desarrollo de América Latina y el Caribe (CAF):** Por el financiamiento y respaldo estratégico fundamentales para hacer viable el proyecto y reducir las brechas digitales de la región.
- **Ministerio de Ciencia, Tecnología, Conocimiento e Innovación de Chile:** Por su acompañamiento institucional y su rol en la gestión de los fondos y la articulación de alianzas en toda Latinoamérica y el Caribe.

2. Red de Cooperación Institucional, Académica y Técnica

Nuestra gratitud a los organismos del Estado, instituciones académicas y organizaciones de la sociedad civil que, bajo principios de transparencia y rigor científico, contribuyeron con datos públicos, bases de datos especializadas, labor técnica y desarrollo de publicaciones fundamentales para el modelo:

- **Ecosistema Regional** (en orden alfabético): Academia Chilena de la Lengua; Agencia de Gobierno Electrónico y Sociedad de la Información y del Conocimiento –AGESIC– (Uruguay); Aimara Lab (Perú); Amity (México); Artificyan; Banco Interamericano de Desarrollo –BID–; Bibliotecas UC; Cámara de Diputados y Diputadas (Chile); Centro de Investigación e Innovación en Tecnologías de la Información y Comunicación –INFOTEC– (México); Centro de Ingeniería para la Investigación, Desarrollo e Innovación Tecnológica –CIDIT– (Paraguay); CentroGeo (México); Comisión Económica para América Latina y el Caribe –CEPAL–; Consejo de Rectoras y Rectores de las Universidades Chilenas (CRUCH); Consejo para la Transparencia (CPLT) (Chile); Corporación Universitaria Minuto de Dios –UNIMINUTO– (Colombia); Datysoc - Data Uruguay (Uruguay); Departamento de Ciencias de la Computación (DCC) de la Universidad de Chile; Ejército de Chile; Facultad de Filosofía y Humanidades de la Universidad de Chile; Facultad de Matemática, Astronomía, Física y Computación (FAMAF) (Argentina); Foundation for Argentinean Development (FUNDAR) (Argentina); Fundación Huésped (Argentina); Fundación Instituto Profesional Duoc UC; Fundación Vía Libre (Argentina); Instituto Nacional de Astrofísica, Óptica y Electrónica –INAOE– (México); JhedAI; LAIA Argentina (Argentina); Laboratorio de Entornos Virtuales de Aprendizaje –LabEva– (Uruguay); La Referencia - Red Clara; Ministerio de Agricultura (Chile); Ministerio de Bienes Nacionales (Chile);

Ministerio de Ciencia, Tecnología, Conocimiento e Innovación (Chile); Ministerio de Ciencia, Tecnología e Innovación (Brasil); Ministerio de Economía (Chile); Ministerio de Educación (Chile); Ministerio de Hacienda (Chile); Ministerio de Salud (Chile); Ministerio de Tecnologías de la Información y las Comunicaciones –MinTIC– (Colombia); Oficina Gubernamental de Tecnologías de la Información y Comunicaciones –OGTIC– (República Dominicana); Organización de los Estados Americanos –OEA–; Pontificia Universidad Javeriana (Colombia); Presidencia del Consejo de Ministros (Perú); Red Divulga Ciencia (Ecuador); Secretaría de Ciencia, Humanidades, Tecnología e Innovación –SECHTI– (México); Secretaría Nacional de Ciencia, Tecnología e Innovación –SENACYT– (Panamá); Servicio Nacional del Patrimonio Cultural (SERPAT) (Chile); Subsecretaría de Desarrollo Regional y Administrativo (SUBDERE) (Chile); Tabuga (República Dominicana); Theodora AI; Universidad Adolfo Ibáñez (UAI) (Chile); Universidad Alberto Hurtado (UAH) (Chile); Universidad Central de Venezuela (Venezuela); Universidad Continental (Perú); Universidad de La Salle (Colombia); Universidad de los Andes (Colombia); Universidad Gabriela Mistral (UGM) (Chile); Universidad Gran Asunción –UNIGRAN– (Paraguay); Universidad Nacional de Avellaneda (Argentina); Universidad Nacional de General San Martín (UNSAM) (Argentina); Universidad Nacional de Hurlingham (UNAHUR) (Argentina); Universidad Particular de Especialidades Espíritu Santo –UEES– (Ecuador); Universidad Ricardo Palma (Perú); Universidad Técnica Particular de Loja –UTPL– (Ecuador); Universidad Tecnológica de Panamá (Panamá); y Wikimedia Chile.

3. Especialistas y Expertos en Datos

Un agradecimiento especial a los profesionales que colaboraron en el monitoreo ético, la recolección estratégica y la validación del corpus:

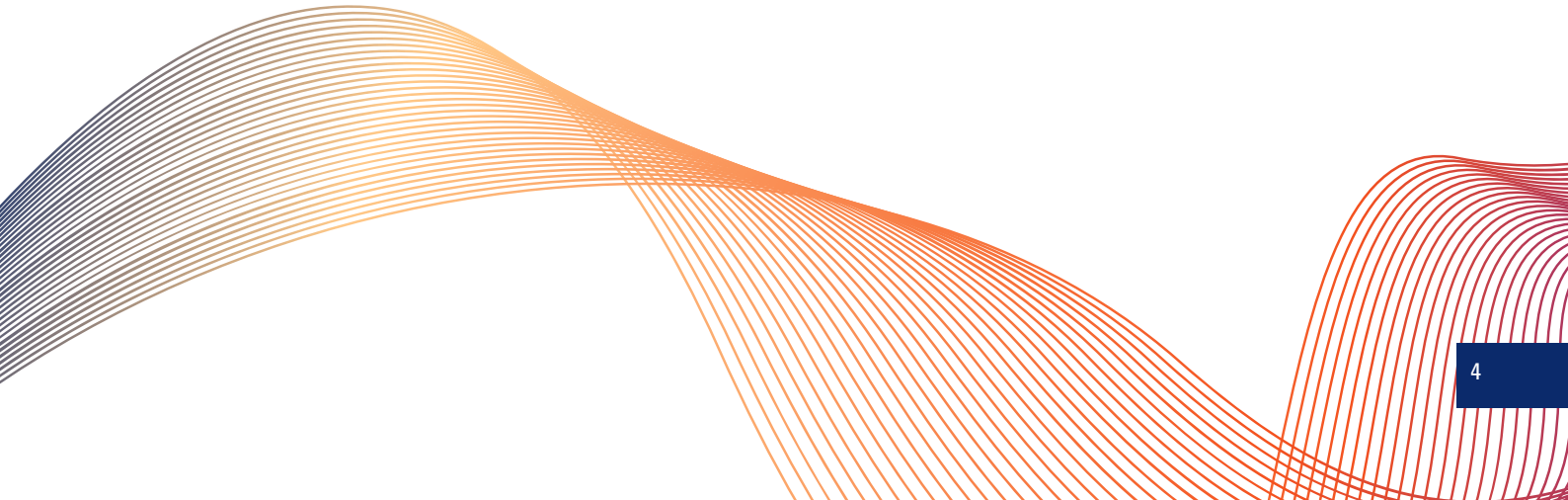
- **Especialistas:** Extendemos nuestra gratitud personal a Milton Alfredo Reyes Herrera (Académico Pontificia Universidad Católica), Juan Raúl Loaiza Arias (Académico Universidad Alberto Hurtado), Maryan Gil (Banco de Desarrollo de América Latina y el Caribe), María Grandury (SomosNLP), Lisette Fuentes Lorca (Centro de Estudios e Investigaciones Laborales –CEIL–), Catalina Gascon Barcaza (Universidad de Playa Ancha), Cristina Benalcazar (Universidad Alberto Hurtado), Freddy Vilches (Universidad Lewis and Clark) y Sebastián Breguel (Corporación Educacional Villa Alemana - Mi Aula), cuyo conocimiento y labor técnica fueron vitales para la integridad y el desarrollo de las publicaciones asociadas a este proyecto.

Finalmente, nuestro agradecimiento al equipo técnico detrás de esta iniciativa en Cenia, el Centro Nacional de IA de Chile. En orden alfabético: Andrés Carvallo; Alexandra Arriagada; Alexandra Davidoff; Alexandra García; Álvaro Soto; Bianca Solar; Bárbara Muñoz; Carlos Aspillaga; Christ Devia; Clemente Henríquez; Denis Parra; Eugenio Herrera-Berg; Felipe Bravo; Gabriela Arriagada; Gonzalo Fuentes; Manuel Cifuentes; Marcelo Mendoza; Mónica Soto; Omar U. Florez; Rodrigo Contreras; Rodrigo Durán; Romina Torres; Rubén Quintupil; Sebastián Cifuentes; Sebastián Donoso; Sebastián Martínez; Tomás Couso; y Tomás Gómez.



Índice

- Prólogo 6
- Resumen ejecutivo10
- 1. Introducción 12
- 2. Dimensiones 14
 - 2.1. Dimensión I: Estrategias de recopilación de datos 15
 - 2.2. Dimensión II: Red de alianzas estratégicas y alcance regional 16
 - 2.2.1. Distribución geográfica de la red de colaboración regional 16
 - 2.2.2. Tipología e instrumentos de colaboración y gobernanza de datos 18
 - 2.3. Dimensión III: Arquitectura de curaduría y procesamiento de datos (2024–2025) 20
 - 2.3.1. Estrategias de adquisición y criterios de incorporación 20
 - 2.3.2. Asesoría jurídica y gestión de licencias 23
 - 2.3.3. Arquitectura del pipeline de procesamiento y refinamiento 23
 - 2.3.3.1. Depuración mediante filtros heurísticos 23
 - 2.3.3.2. Consolidación de unicidad (deduplicación) 23
 - 2.3.3.3. Etiquetado automatizado y clasificación multidimensional 23
 - 2.3.3.4. Validación de calidad sintáctica y estructural 24
 - 2.3.3.5. Anonimización y Resguardo de la Privacidad 24
 - 2.3.4. Validación de calidad sintáctica y estructural 24
 - 2.3.5. Anonimización y Resguardo de la Privacidad 24
 - 2.4. Dimensión IV: Gestión de licencias y marco legal 26
 - 2.4.1. Niveles de apertura y gobernanza del LatamGPT-Corpus v1.0 26
 - 2.4.1.1. Niveles de acceso (disponibilidad de los datos) 26
 - 2.4.1.2. Restricción de Uso para Grandes Empresas 27
 - 2.4.1.3. Metodología de evaluación y transparencia 28
 - 2.4.1.4. Criterios de web scraping para información pública y fundamento constitucional 29
 - 2.4.1.5. Mecanismos de Transparencia y Validación Científica 31
- 3. Análisis estadístico del corpus 32
- 4. Principales hallazgos 35
 - 4.1. La brecha de conocimiento latinoamericano en los grandes modelos de IA 41
 - 4.2. Por qué la gobernanza de datos importa 42
 - 4.3. El laberinto legal: riesgos y cómo navegarlos 42
- 5. Conclusiones y transparencia de datos 43
- 6. Anexo: Glosario de términos y referencias 46
- 7. Testimonios de alianzas 51



Figuras

Figura 1: Alcance Territorial y Densidad de Alianzas Estratégicas (2023 – 2026)	16
Figura 2: Diversidad de Instrumentos de Colaboración y Adaptación Normativa por País	19
Figura 3: Dimensiones Temáticas del LatamGPT-Corpus v1.0.	21
Figura 4: Arquitectura Operativa y Despliegue de Talento Especializado para el Corpus v1.0.	22
Figura 5: Arquitectura del Pipeline de Procesamiento del LatamGPT-Corpus v1.0.	25
Figura 6: Esquema de clasificación y gobernanza de licencias del corpus	30
Figura 7: Composición porcentual de licencias en el LatamGPT-Corpus v1.0.	31
Figura 8: Pipeline de procesamiento del corpus	33
Figura 9: Distribución de tokens por país y tópicos en el conjunto Stable v2	34

Tablas

Tabla 1: Número de instrumentos de colaboración formalizados por país (2023–2026)	17
Tabla 2: Modelos de lenguaje globales de propósito general	37
Tabla 3: Modelos soberanos y regionales. (a) Europa y América Latina, (b) África e India	38
Tabla 4: Datasets de voz y corpus de referencia	40
Tabla 5: Licencias identificadas y condiciones generales de uso	49

Prólogo

Más que un modelo de lenguaje: una nueva forma de colaborar

El desarrollo de la primera versión de LatamGPT culmina con la publicación de este catálogo de datos. A lo largo del proceso hemos hecho carne una vocación institucional por la colaboración radical, que creemos es la mejor, y quizás la única forma, de relevar la importancia de la región en la orquesta global de Inteligencia Artificial.

LatamGPT puede tomarse como un piloto a gran escala de esta visión: colaboración entre academia, sociedad civil, gobiernos, organismos multilaterales y empresas privadas del sur global. Lo que puede parecer marginal ante la magnitud del avance del estado del arte en materia de modelos de frontera, resulta trascendental para la situación local de América Latina y el Caribe. Constituye el primer esfuerzo colectivo de carácter técnico que se proyecta hacia el futuro más allá del objetivo original, justamente porque a lo largo del camino descubrimos que la riqueza no estaba en la publicación de un código, sino que en la articulación de capacidades que se retroalimentan mutuamente en la interacción.

LatamGPT pasó de ser un modelo de 70B de parámetros a un ecosistema de bienes públicos digitales, con mecanismos funcionales de identificación, ingesta y curatoría de datos; publicación de los mismos; entrenamiento de modelo y, quizás lo más importante, un modelo de gobernanza y coordinación que habilitó un proyecto ambicioso en ingeniería y austero en recursos financieros y computacionales.

Para nosotros en CENIA es un orgullo haber podido coordinar este proyecto, y queremos dar las gracias a todas las organizaciones que lo hicieron posible, pero particularmente al Banco de Desarrollo para América Latina y el Caribe CAF y al Ministerio de Ciencia, Tecnología, Conocimiento e Innovación de Chile, quienes financiaron a través de una cooperación técnica parte relevante del costo operativo del último año. También a la Fundación Data Observatory y a Amazon Web Services, quienes dispusieron y habilitaron el cómputo necesario para esta versión.

Esperamos, realmente, que LatamGPT sea más que un modelo de lenguaje: sea un modelo de cómo podemos colaborar para construir más allá de nuestra frontera.

Álvaro Soto

Director CENIA y LatamGPT

Rodrigo Durán

Gerente CENIA

Hay momentos en que una región deja de mirarse desde afuera y comienza, por fin, a narrarse a sí misma.

Durante demasiado tiempo, América Latina ha sido más hablada que hablante; más interpretada que autora de su propio relato. En 1943, el artista uruguayo Joaquín Torres García tomó el mapa de América del Sur y lo dio vuelta. Con ese gesto, simple pero radical, nos recordó que el norte no siempre tiene que estar arriba, y que el sur también puede ser centro. Que la mirada puede, y debe, cambiar.

Y hoy, décadas después, ese gesto vuelve a cobrar sentido en un nuevo territorio: el digital. Porque si antes los mapas definían cómo nos ubicábamos en el espacio, hoy son los datos y los modelos que los procesan los que definen cómo somos representados, interpretados y comprendidos. La revolución tecnológica actual está redefiniendo las bases del conocimiento global. Pero esa transformación no ha sido neutra. Ha sido entrenada, en gran medida, desde otras latitudes y otras realidades.

Y en ese contexto, emerge una verdad incómoda pero necesaria: lo que no está en los datos, no existe en el modelo; y lo que no existe en el modelo queda, inevitablemente, fuera de las decisiones del futuro. No es sólo un desafío técnico, es un desafío cultural y profundamente humano. Porque entendimos que no basta con usar la inteligencia artificial; también debemos participar en la construcción del conocimiento que la hace posible, para que piense con nuestras voces, nuestras lenguas y nuestras historias.

Y, en ese sentido, LatamGPT es, en esencia, un acto de América invertida: dar vuelta el mapa de la inteligencia artificial, redefinir desde dónde se produce el conocimiento y dejar de ser únicamente fuente de datos y convertirnos en autores de nuestras propias infraestructuras.

Pero este no es un proyecto individual. Es, ante todo, un esfuerzo colectivo: Más de 79 instituciones de 15 países decidieron compartir conocimiento y construir confianza en torno a un propósito común. En un mundo donde los datos tienden a concentrarse, este proyecto apuesta por la apertura, la colaboración y una gobernanza responsable que entienda el conocimiento como un bien que se cuida para poder compartirse.

LatamGPT se sostiene sobre datos, talento y cómputo. Pero su propósito es más profundo: construir una infraestructura cognitiva latinoamericana sobre la cual podamos desarrollar tecnología con sentido, contexto e identidad. Porque la pregunta de fondo no es sólo qué tecnología construimos. Es desde dónde la construimos. Y para quién. Por primera vez, tenemos la posibilidad real de crear nuestros propios modelos, con nuestras lenguas, nuestros datos y nuestra visión del mundo. De participar, en nuestros propios términos, en la definición del futuro digital. Y en ese proceso, cambia la forma en que nos vemos y, con ello, el futuro que somos capaces de imaginar.

Tal como Torres García invirtió el mapa, hoy estamos invirtiendo la dirección desde donde se produce el conocimiento. Y quizás, en ese gesto, también estamos respondiendo a ese grito que nos precede. Un grito que no era sólo advertencia, sino también posibilidad. Porque el futuro no está escrito. Se entrena. Y hoy, por primera vez, lo estamos entrenando juntas y juntos. Desde el sur, y esta vez, con voz propia.

Mónica Soto

Directora de Operaciones CENIA

LatamGPT: una voz latinoamericana

“¡América, América! ¡Todo por ella; porque nos vendrá de ella desdicha o bien! Somos aún México, Venezuela, Chile, el azteca-español, el quechua-español, el araucano-español; pero seremos mañana, cuando la desgracia nos haga crujir entre su dura quijada, un solo dolor y no más que un anhelo”.

Así comienza El Grito, de Gabriela Mistral, obra cuyos extractos fueron recitados en el evento de lanzamiento de LatamGPT. La selección de este texto estuvo lejos de ser anecdótica. Gabriela Mistral no solo es, a la fecha, la única mujer iberoamericana laureada con el Nobel de literatura, sino también una intelectual que se atrevió a pensar y escribir sobre la identidad latinoamericana. El Grito es un llamado de atención a los mismos latinoamericanos, más propensos a la admiración extranjera que a forjar con orgullo el propio destino.

LatamGPT podría entenderse como una respuesta a ese grito desesperado que fue proferido hace más de un siglo. En un momento en que la Inteligencia Artificial se ha vuelto la tecnología del momento, cuando está definiendo la geopolítica global, y en medio de los reclamos por los sesgos WEIRD (Western, Educated, Industrialized, Rich and Democratic) de los modelos del Norte Global, LatamGPT es una apuesta y un llamado a la acción. Porque, tal como denunciaba la poetisa de Vicuña, “hablamos, alegamos, mientras él [el Norte] siembra, funde, asierra, labra, multiplica, forja”.

Crear el primer modelo de lenguaje latinoamericano podría parecer un pequeño paso dentro de la oferta de modelos conversacionales, pero representa un gran salto para toda la región. Esta iniciativa pone en valor el lenguaje, la cultura y la tradición de América Latina. Más aún, crea capacidades productivas para una región relegada a un rol de consumidor. Su sello diferenciador no solo radica en el contenido, que rescata estos aspectos idiosincráticos ya descritos, sino también en la forma de construir el modelo. La curaduría de datos, la transparencia y su apertura son parte fundamental del proyecto.

Así, LatamGPT representa mucho más que un modelo de lenguaje. Es una piedra angular sobre la cual se proyecta un trabajo colaborativo, una iniciativa que le da voz a una región que estaba acostumbrada al silencio. Es, en definitiva, la respuesta a un eco distante de hace más de cien años: “¡América y sólo América! ¡Qué embriaguez semejante futuro, qué hermosura, qué reinado vasto para la libertad y las excelencias mayores!”.

Sebastián Adasme

Jefe Técnico de Estudios ILIA

Resumen ejecutivo



El LatamGPT-Corpus v1.0 nace de la voluntad de construir un conjunto de datos de entrenamiento para la primera versión del modelo, que reuniera contenido representativo de América Latina y el Caribe, bajo rigurosos estándares de calidad y una política abierta, transparente y soberana de acceso y puesta a disposición de los datos. El presente Catálogo Referencial constituye la memoria técnica de la iniciativa de datos que dio origen al corpus de LatamGPT. En él se documentan sus fuentes, metodologías de curaduría, marcos legales y mecanismos de gobernanza. Asimismo, busca profundizar en el modelo de alianzas que sustenta el proyecto y servir como complemento del reporte técnico publicado junto con el modelo.

El imperativo de una voz propia

Históricamente, el desarrollo de la inteligencia artificial a nivel global se ha sustentado en conjuntos de datos que representan de manera insuficiente la realidad de nuestra región, debido a la dispersión de las fuentes y a sus limitaciones de calidad. Diversos estudios advierten una asimetría estructural: las fuentes predominantes favorecen a las regiones con mayor infraestructura digital y relegan la riqueza del español y el portugués latinoamericanos a una posición periférica. Esta falta de representación diluye nuestras variantes lingüísticas y contextos sociales en marcos ajenos y, con ello, limita el desempeño de los modelos de IA generativa en contextos locales.

Frente a este escenario, LatamGPT v1.0 ofrece una respuesta concreta: un corpus construido desde y para la región, con criterios rigurosos de calidad y una política de acceso abierta y transparente. El conjunto consolidado reúne 296,5 mil millones de tokens, de los cuales 58,8 mil millones corresponden a datos curados provenientes de alianzas institucionales, obtenidos mediante un proceso de filtrado de alta precisión aplicado a una base bruta inicial de 157 mil millones de tokens. Esta base de conocimiento integra cuatro fuentes principales:

- **Extracciones regionales optimizadas:** Capturando la esencia del discurso local.
- **Repositorios de ciencia y código abierto:** Integrando el conocimiento académico global.
- **Recopilación dirigida:** Selección exhaustiva de sitios con alto valor estratégico.
- **Transferencias institucionales:** Datos provenientes directamente de las organizaciones aliadas, que constituyen el núcleo más riguroso y soberano del corpus.

El resultado es un mosaico de conocimiento que abarca desde las ciencias exactas y la economía hasta las humanidades y la diversidad cultural de nuestra región.

Una red de colaboración sin precedentes

Este hito es el fruto de una voluntad colectiva que movilizó a 79 instituciones de 15 naciones. Ministerios, universidades y organizaciones de la sociedad civil se unieron bajo un marco de confianza, formalizado mediante instrumentos legales que garantizan la transparencia, la atribución y la confidencialidad. Es un modelo de cooperación que demuestra que la tecnología de frontera y el respeto a la ética pueden avanzar en conjunto.

Frente a la opacidad que en ocasiones caracteriza a la industria, LatamGPT v1.0 adopta un modelo de transparencia y responsabilidad. Para ello, implementamos un sistema de gobernanza de tres niveles: abierto, controlado y cerrado, que promueve el uso legítimo de la información y protege los derechos de propiedad intelectual. En una región que aún enfrenta desafíos normativos para el desarrollo de la ciencia de datos, este proyecto se rige por un gran respeto a un uso legítimo y validado de sus fuentes de información.

Documentar un corpus con este nivel de trazabilidad constituye un esfuerzo relevante en el panorama internacional. Lo hacemos porque LatamGPT v1.0 no es solo una herramienta tecnológica, sino también un precedente regional. Representa la convicción de que es posible desarrollar tecnología de vanguardia con identidad propia, promoviendo la legitimidad y, sobre todo, respondiendo con excelencia ante la sociedad que le da origen.

Introducción

El presente Catálogo Referencial se constituye como el marco documental oficial que acompaña el conjunto de datos LatamGPT v1.0. En un ecosistema tecnológico, donde la integridad de la información define la fiabilidad y seguridad de los modelos de lenguaje, este documento funciona como la hoja de ruta técnica del corpus de datos de LatamGPT.

Objetivos y Alcance

El propósito de este catálogo es documentar los procesos de recopilación y curaduría de datos de LatamGPT. Este esfuerzo representa un ejercicio de responsabilidad proactiva que busca sentar un precedente articulándose bajo tres ejes:

1. **Hacia una Gobernanza de Datos:** Este documento busca alinearse con los principios de la OCDE (Organización para la Cooperación y el Desarrollo Económicos [OCDE], s.f.) sobre IA, promoviendo la responsabilidad (Accountability) y los valores centrados en el ser humano. Representa un primer paso hacia sistemas más auditables que permitan reducir la opacidad algorítmica.
2. **Transparencia y Representación:** Tomando como referencia las recomendaciones de la UNESCO (Organización de las Naciones Unidas para la Educación, la Ciencia y la Cultura [UNESCO], 2022), este catálogo ofrece una primera mirada profunda a las distribuciones lingüísticas del modelo. Nuestra meta es que, en sucesivas iteraciones, el proyecto LatamGPT refleje con mayor fidelidad la pluralidad y riqueza de nuestra región.
3. **Trazabilidad y Gestión de Riesgos:** Inspirados por los estándares internacionales emergentes, establecemos la trazabilidad de las fuentes como un pilar fundamental. Este registro es nuestro punto de partida para una gestión ética y responsable de los riesgos asociados a los modelos de gran escala en nuestro contexto local.

Ética por Diseño y Mejora Continua

LatamGPT v1.0 se ha desarrollado bajo el paradigma de Data Ethics by Design, integrando consideraciones éticas desde su concepción. Siguiendo los principios de Floridi y Taddeo (2016), la ética de datos se ha establecido como un marco macroscópico que guía el diseño técnico, asegurando que la mitigación de sesgos y el respeto a la propiedad intelectual sean pilares estructurales y no medidas correctivas posteriores. Si bien hemos realizado esfuerzos profundos, reconocemos que este es un proceso en constante evolución hacia lo que se denomina una ventaja competitiva basada en la integridad del dato (Hasselbalch, 2016). Presentamos entonces esta versión v1.0 como un punto de partida responsable; una base sólida sobre la cual perfeccionaremos nuestros métodos en futuras iteraciones.

Este reporte responde además al llamado de soberanía impulsado por la Declaración de Santiago (2023) y sus posteriores actualizaciones (Montevideo 2024 y San José 2026) suscrita por más de 20 naciones de la región, la cual insta a los países de la región a promover una IA ética, inclusiva y soberana que preserve nuestra diversidad cultural. Asimismo, se alinea con los hallazgos del Índice Latinoamericano de Inteligencia Artificial (ILIA, 2025), que identifica la transición de "consumidores" a "productores" de tecnología como una urgencia histórica.

Al documentar LatamGPT v1.0, contribuimos a cerrar la brecha de disponibilidad de datos locales, promoviendo un ecosistema de código abierto que proteja nuestra identidad cultural y consolide nuestra autonomía tecnológica en el escenario global.

Dimensiones

2.1. Dimensión I: Estrategias de Recopilación de Datos

La construcción del LatamGPT-Corpus v1.0 demandó el diseño de una estrategia de recopilación diversificada, orientada a maximizar la cobertura temática, geográfica y lingüística de América Latina y el Caribe. El proyecto integró cuatro vertientes complementarias, asegurando la trazabilidad y el cumplimiento de integridad de cada insumo:

- **Alianzas e Intercambio Institucional:** Como componente central y distintivo de este corpus, se integraron datos cedidos directamente por instituciones aliadas. Mediante acuerdos de gobernanza formalizados, diversas organizaciones confiaron sus activos de información al proyecto, permitiendo acceder a una especificidad académica e institucional que no se encuentra en repositorios públicos y que constituye el núcleo soberano del corpus.
- **Extracción Dirigida de Nodos Estratégicos:** Para profundizar en la identidad regional, se ejecutó un proceso de extracción técnica sobre un conjunto de sitios web priorizados por su relevancia en América Latina y el Caribe. En esta primera versión, y tras un riguroso tamiz de calidad, se integraron fuentes reconocidas, como universidades, portales gubernamentales y plataformas culturales, asegurando que el modelo se nutra de fuentes relevantes para distintos contextos sociales, académicos e institucionales de la región.
- **Integración de Repositorios de Referencia:** Se realizó una prospección sistemática en plataformas globales como Hugging Face y GitHub. La selección de estos conjuntos de datos fue tanto técnica como ética: cada dataset fue evaluado bajo criterios de compatibilidad de licencias y relevancia científica, incorporando así recursos de conocimiento reconocidos y utilizados por la comunidad técnica internacional.
- **Curaduría de la Web Abierta (Common Crawl):** Más que una ingesta masiva, se realizó un procesamiento selectivo sobre los datos de Common Crawl. El equipo aplicó filtros lingüísticos y de geolocalización para extraer exclusivamente segmentos relevantes para el contexto de nuestra región, garantizando que el conjunto de datos web cumpla con estándares mínimos de pertinencia y calidad técnica.

En conjunto, este enfoque de fuentes múltiples ha permitido consolidar un corpus representativo y reutilizable. Uno de los objetivos más ambiciosos de esta primera versión fue lograr una distribución equilibrada entre los 20 países representados, reconociendo que cada territorio presenta condiciones distintas en términos de disponibilidad de datos y desarrollo de su ecosistema digital. Iniciativas como ILIA (2024, 2025) han contribuido a visibilizar estos matices, recordándonos que una representación genuinamente regional requiere esfuerzos sostenidos y diferenciados.

En esta primera versión hemos procurado acortar esa distancia con el mayor rigor posible, privilegiando la diversidad de fuentes y el fortalecimiento de alianzas institucionales en cada país. Sabemos que es un primer paso y que el camino es largo, pero estamos convencidos de que su composición sienta las bases metodológicas necesarias para el desarrollo de una Inteligencia Artificial con identidad propia y verdaderamente latinoamericana.

2.2. Dimensión II: Red de Alianzas Estratégicas y Alcance Regional

El desarrollo de LatamGPT v1.0 representa un hito relevante de un proceso de articulación regional que entre octubre de 2023 y febrero de 2026, desplegó un trabajo sistemático para fortalecer la disponibilidad y representatividad de los datos en América Latina y el Caribe. Coordinada por el Centro Nacional de Inteligencia Artificial (CENIA), esta iniciativa ha logrado consolidar una red estratégica de colaboración que trasciende fronteras y coyunturas, posicionando la información como un activo para avanzar en la autonomía tecnológica.

Este periodo de trabajo no solo implicó la formalización de diversos instrumentos jurídicos con actores públicos, académicos y privados, sino, primordialmente, la construcción de confianzas compartidas y la definición de estándares éticos para el uso responsable de la información. Un eje rector ha sido la implementación de una gobernanza transparente y entendible, orientada a resguardar la trazabilidad y la integridad de los datos. Reconociendo que este es un camino en constante evolución, se han establecido criterios que buscan garantizar la transparencia del proceso, respetando siempre la pluralidad de contextos institucionales y normativos presentes en nuestra región.

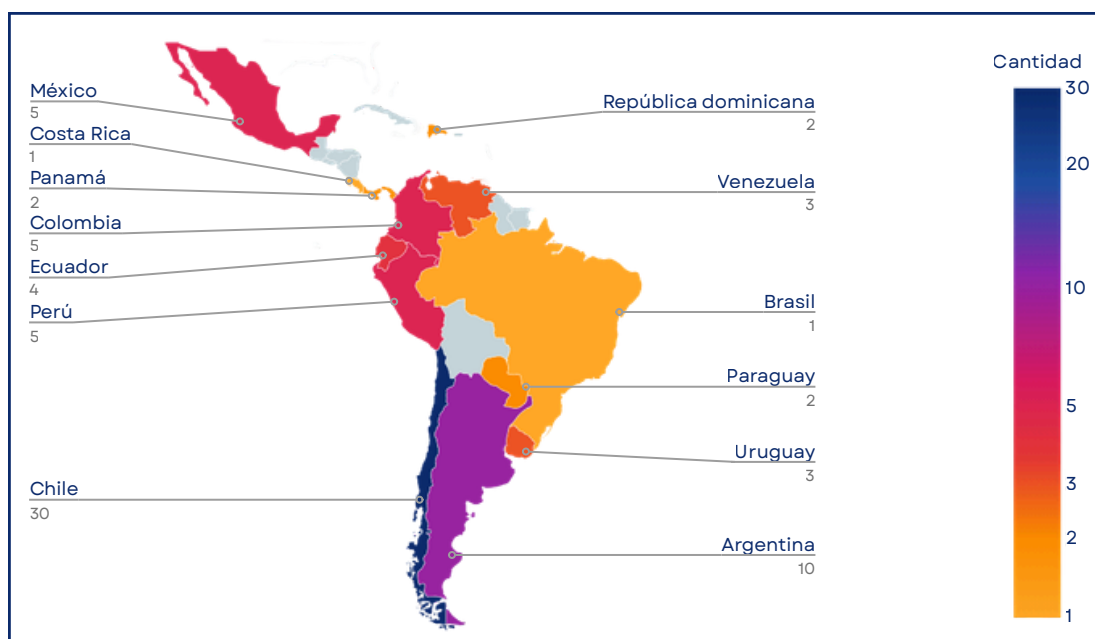
Las figuras y tablas que se presentan a continuación sintetizan este recorrido, dando cuenta tanto de la cobertura territorial alcanzada como de la diversidad de mecanismos utilizados para formalizar estas alianzas. En su conjunto, estos indicadores reflejan los avances logrados y subrayan los desafíos que persisten en la construcción de un ecosistema de datos más equitativo, colaborativo y genuinamente representativo para América Latina y el Caribe.

2.2.1. Distribución Geográfica de la Red de Colaboración Regional

A continuación, se presenta la distribución territorial de los convenios y alianzas estratégicas suscritas en el marco del proyecto en América Latina y el Caribe:

Figura 1

Alcance Territorial y Densidad de Alianzas Estratégicas (2023 – 2026)



Fuente: Elaboración propia.

La presente cartografía ilustra la distribución de las alianzas consolidadas por el proyecto LatamGPT v1.0 en América Latina y el Caribe durante el periodo comprendido entre octubre de 2023 y febrero de 2026. La intensidad cromática representa el volumen de acuerdos formalizados por nación, destacando a Chile, seguido por una sólida participación de Argentina, Colombia, México y Perú. Esta red ha permitido acceder a activos de información de alta calidad, provistos por instituciones de referencia que fortalecen la representatividad del corpus construido. Como horizonte estratégico, el mapa identifica territorios que representan oportunidades prioritarias para futuras fases de expansión, particularmente en Centroamérica y diversas naciones del Caribe. Estos ámbitos geográficos se proyectan como ejes fundamentales para profundizar la capilaridad de la red, asegurando que la diversidad cultural de toda nuestra geografía esté plenamente integrada en esta infraestructura soberana de datos. En su conjunto, este despliegue territorial no solo da cuenta de la solidez alcanzada en esta primera etapa, sino que también reafirma el compromiso del proyecto con la creación de un ecosistema de conocimiento genuinamente inclusivo para toda América Latina y el Caribe.

Tabla 1

Número de instrumentos de colaboración formalizados por país (2023–2026)

País	Suma total
Argentina	10
Brasil	1
Chile	30
Colombia	5
Costa Rica	1
Ecuador	4
EEUU	2
España	1
InternacionalArge	3
México	5
Panamá	2
Paraguay	2
Perú	5
República Dominicana	2
Uruguay	3
Venezuela	3

Fuente: Elaboración propia.

La presente tabla detalla los mecanismos de formalización jurídica suscritos en el marco del proyecto LatamGPT v1.0 durante el periodo comprendido entre octubre de 2023 y enero de 2026. Esta estructura desglosa la diversidad de figuras legales empleadas –incluyendo Memorandos de Entendimiento (MoU), Convenios Marco, Convenios Específicos, Acuerdos de Confidencialidad (NDA) y Oficios de Coordinación Institucional–, las cuales han sido fundamentales para garantizar la seguridad, la ética y la transparencia en el tratamiento de los activos de información regionales. Esta pluralidad de instrumentos refleja la capacidad de adaptación normativa del proyecto ante los diversos marcos legales de América Latina y el Caribe. Más que un registro administrativo, esta tipología evidencia la

solidez del andamiaje que viabiliza el acceso a datos de alta calidad, consolidando un corpus robusto y representativo que se sustenta en el respeto irrestricto a la propiedad intelectual y a la gobernanza soberana de cada institución aliada.

2.2.2. Tipología de Instrumentos de Colaboración y Gobernanza de Datos

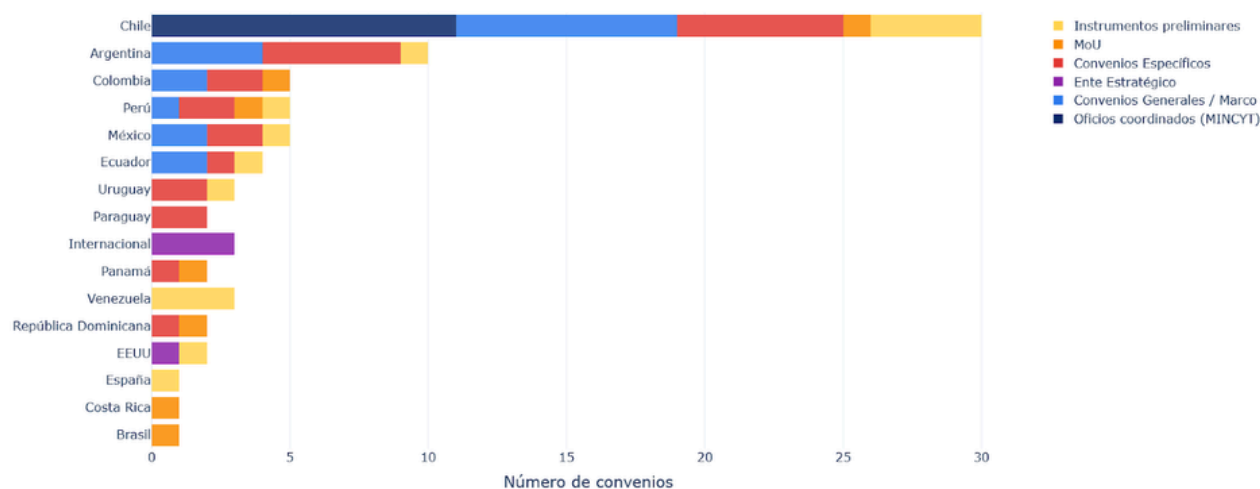
En el marco del proyecto LatamGPT v1.0, se suscribieron diversos instrumentos de colaboración orientados a formalizar la vinculación institucional y habilitar el uso de datos bajo estrictos estándares de seguridad, ética y legalidad. Esta arquitectura jurídica permitió establecer condiciones claras de cooperación, definiendo las responsabilidades de las partes y los protocolos para la transferencia y resguardo de activos de información. Los principales mecanismos empleados incluyen:

- **Oficios de Coordinación Institucional:** Gestionados a través del Ministerio de Ciencia, Tecnología, Conocimiento e Innovación de Chile, facilitando la articulación de alto nivel entre organismos del Estado.
- **Convenios Marco de Colaboración:** Acuerdos generales que sientan las bases de una relación estratégica de largo plazo entre las instituciones.
- **Acuerdos con Entes Estratégicos:** Acuerdos diseñados para colaboraciones de alto impacto con organizaciones clave del ecosistema regional.
- **Convenios Específicos de Transferencia:** Documentos técnicos que regulan de manera precisa el intercambio y uso de conjuntos de datos determinados.
- **Memorandos de Entendimiento (MoU):** Protocolos que formalizan la voluntad de cooperación y los objetivos comunes en el ámbito de la IA.
- **Instrumentos Preliminares (NDA o Cartas de Intención):** Garantías jurídicas para el tratamiento protegido de información sensible o reservada, que establecen las bases de confidencialidad antes de la formalización de acuerdos definitivos.

Complementariamente, el proyecto se sustenta en un Marco de Gobernanza de Datos que regula de forma integral la disponibilidad, el uso y la custodia de la información, con especial énfasis en las etapas posteriores al entrenamiento de los modelos. Este sistema de gobernanza es un mecanismo dinámico diseñado para asegurar un tratamiento responsable que garantice la calidad, trazabilidad y el cumplimiento normativo en todo el ciclo de vida del dato.

Figura 2

Diversidad de Instrumentos de Colaboración y Adaptación Normativa por País.



Fuente: Elaboración propia.

El presente gráfico desglosa la distribución de los mecanismos jurídicos suscritos en cada nación de América Latina y el Caribe durante el periodo 2023–2026. Las categorías representadas abarcan desde acuerdos de carácter institucional y estratégico hasta protocolos técnicos específicos para la regulación y el resguardo de activos de información. Esta visualización permite identificar la capacidad de respuesta del proyecto ante los diversos contextos normativos de cada país, evidenciando una arquitectura de colaboración flexible pero rigurosa. La predominancia de ciertos instrumentos sobre otros no es accidental; responde a una estrategia deliberada de vinculación que respeta las autonomías institucionales y garantiza que la transferencia de datos se realice bajo un marco legal sólido según cada jurisdicción. En su conjunto, esta diversidad de mecanismos consolida a LatamGPT v1.0 como un referente de gobernanza adaptable, capaz de integrar voluntades públicas y académicas bajo un estándar común de seguridad y transparencia.

2.3. Dimensión III: Arquitectura de Curaduría y Procesamiento de Datos (2024-2025)

Durante el periodo 2024-2025, se implementó un sistema para la recopilación y curaduría de los datos del LatamGPT-Corpus v1.0. Esta fase incluyó tres estrategias de recopilación de datos, apoyadas por un pipeline automatizado de procesamiento y control de calidad.

Para cumplir con los plazos de esta primera entrega y asegurar la estabilidad del modelo, se definió un corte temporal en el procesamiento de datos. Por esta razón, las alianzas y activos que no alcanzaron a integrarse en este ciclo han sido programados para futuras versiones de LatamGPT.

2.3.1. Estrategias de Adquisición y Criterios de Incorporación

La construcción del Corpus v1.0 se fundamentó en una visión interdisciplinaria que trascendía lo estrictamente tecnológico. Un pilar central de esta fase fue la definición de una estructura temática robusta, diseñada con la asesoría especializada de expertos en literatura y cultura latinoamericana contemporánea. Esta colaboración académica permitió cartografiar la identidad regional a través de diez ejes, garantizando que el modelo se nutriera de una diversidad de registros que abarcan Educación, Humanidades y Ciencias Sociales, Ciencias Exactas, Política, Economía y Finanzas, Medicina y Salud, y Comunicación y Medios. Especial relevancia se otorgó a la inclusión de contenidos sobre Arte, Deporte y Recreación, así como a la preservación del legado de Pueblos Originarios, asegurando una representatividad cultural genuina y profunda.

Bajo esta hoja de ruta, el proceso de adquisición de datos se ejecutó bajo 3 estrategias de recopilación de datos:

- **Asistencia Técnica y Extracción en Nodos Institucionales:** Reconociendo que el valor de los datos reside muchas veces en repositorios que carecen de infraestructuras de entrega automatizada, el equipo de ingeniería de LatamGPT v1.0 asumió un rol proactivo de asistencia técnica. Mediante protocolos autorizados de web scraping sobre 429 sitios, el proyecto facilitó la integración de fuentes que, de otro modo, habrían permanecido aisladas. Tras un exhaustivo tamiz de calidad y viabilidad jurídica, 125 fuentes institucionales fueron incorporadas con éxito.
- **Curaduría en Repositorios de Referencia:** De forma paralela, se realizó una prospección sistemática en plataformas globales como *GitHub* y *Hugging Face*. Se seleccionaron e integraron 15 datasets de alta relevancia, tras un riguroso proceso de validación que aseguró la apertura de sus licencias y su alineación con los estándares científicos del proyecto. Esta labor permitió dotar al corpus de recursos de conocimiento ampliamente utilizados por la comunidad técnica internacional, integrándolos orgánicamente con el contenido local.
- **Transferencia Directa y Gobernanza de Metadatos:** Finalmente, se consolidaron 11 transferencias directas de alto valor a través de infraestructuras seguras en la nube. Este mecanismo garantiza la integridad de los volúmenes de datos recibidos y se complementa con un proceso de estandarización de metadatos, asegurando la trazabilidad de la información. La supervisión de especialistas en ingeniería y gestión de proyectos permitió que este flujo de conocimiento institucional se realizara según altos estándares de rigurosidad y transparencia.

Figura 3

Dimensiones Temáticas del LatamGPT-Corpus v1.0.

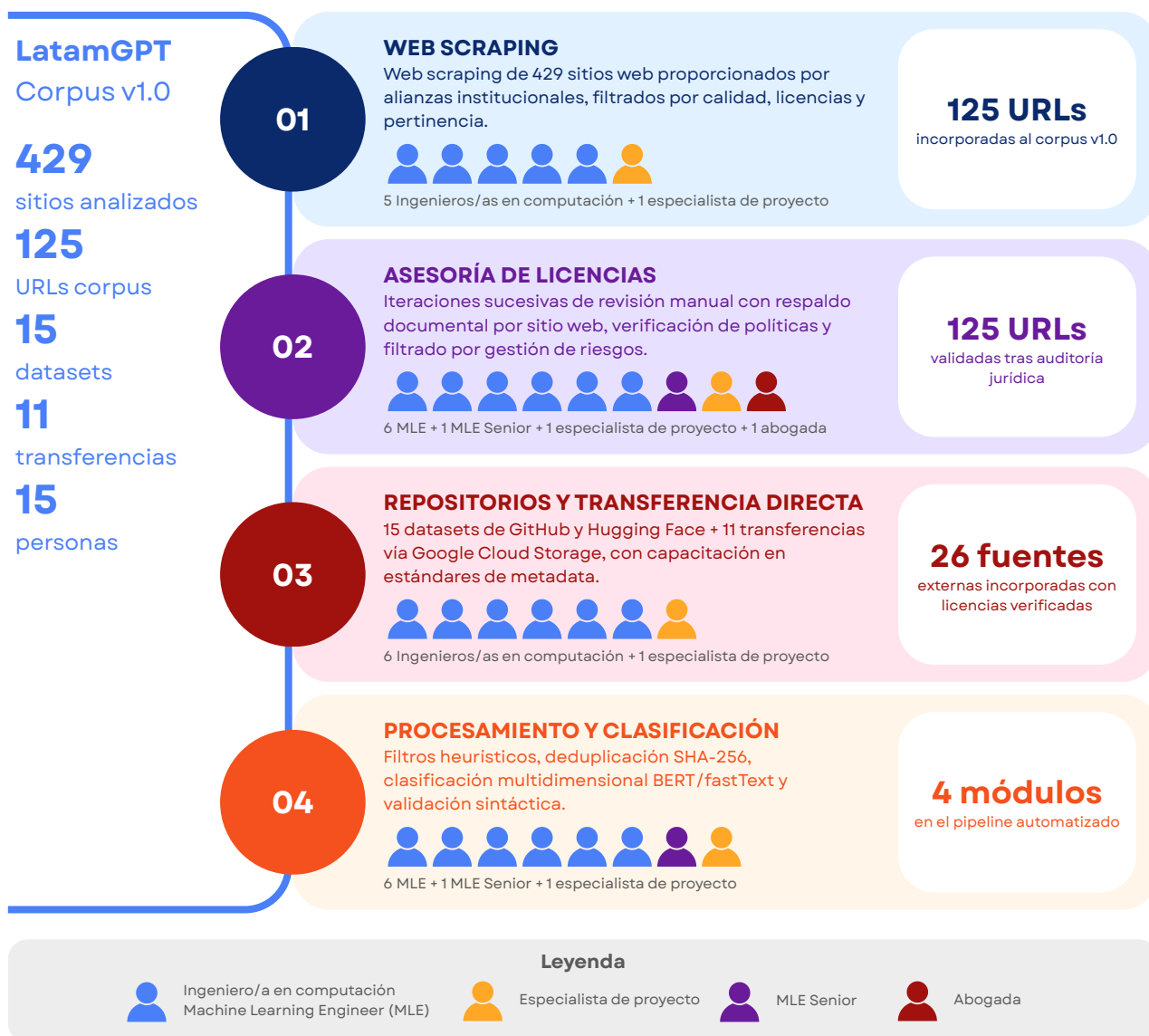


Fuente: Elaboración propia.

El gráfico muestra las 10 dimensiones temáticas que definen el contenido de esta primera versión del corpus. Estas categorías fueron seleccionadas para cubrir los pilares fundamentales de la cultura y el conocimiento de América Latina y el Caribe, incluyendo áreas que van desde las ciencias y la economía hasta las tradiciones de los pueblos originarios y la educación. Esta estructura asegura que el modelo cuente con una base de datos variada y representativa de la realidad regional.

Figura 4

Arquitectura Operativa y Despliegue de Talento Especializado para el Corpus v1.0.



Fuente: Elaboración propia.

La infografía detalla el equipo técnico que permitió la consolidación de los activos digitales del proyecto. El flujo operativo se estructuró sobre cuatro pilares estratégicos, movilizandoun equipo interdisciplinario de 15 profesionales –incluyendo Ingenieros de aprendizaje automático (MLE), especialistas en gestión de proyectos y asesoría legal– bajo un modelo de trabajo coordinado y escalable. Se destaca el rigor del proceso de curaduría y validación, en particular, de un universo de 429 sitios analizados, solo el 29% fue integrado tras superar filtros críticos de calidad y licencias. Asimismo, la figura ilustra la complejidad de la asesoría de licencias, que contó con una etapa de mitigación de riesgos guiada por expertos en derecho para asegurar el uso responsable de la información. El pipeline culmina en una fase de procesamiento de datos de alta especialización, donde se ejecutaron tareas de deduplicación, normalización y clasificación de dominios, orientadas a que el Corpus v1.0 no sea solo una acumulación de texto, sino un activo de conocimiento trazable y de alta calidad técnica.

2.3.2. Asesoría jurídica y gestión de licencias

De forma transversal al proceso de recopilación, se implementó un protocolo de auditoría jurídica diseñado para garantizar la revisión de cada fuente integrada al corpus. Este procedimiento no fue automatizado, sino que consistió en múltiples iteraciones de revisión humana hasta consolidar las 125 fuentes que cumplieran estrictamente con la política de gestión de riesgos del proyecto. Las dimensiones de este control incluyeron:

- **Trazabilidad Documental:** Generación de un respaldo individualizado por cada nodo institucional.
- **Validación de Derechos de Uso:** Verificación exhaustiva de términos, condiciones y tipologías de licenciamiento.
- **Mitigación de Riesgos:** Evaluación de contingencias legales y filtrado preventivo bajo estándares de ética de datos.

Este proceso jurídico fue ejecutado por un equipo interdisciplinario de 9 especialistas de la región, integrando la visión técnica de ingenieros, la gestión de un especialista de proyecto y el rigor de la asesoría legal, documentando atributos críticos como la procedencia geográfica, temática y licencias asociadas.

2.3.3. Arquitectura del Pipeline de Procesamiento y Refinamiento

Una vez consolidados los activos digitales, se activó un pipeline de procesamiento secuencial: un flujo computacional que orquestó algoritmos avanzados y modelos de Machine Learning para la normalización del corpus. Esta arquitectura fue diseñada por un equipo de ocho expertos en IA, estructurando la información a través de cuatro módulos críticos:

2.3.3.1. Depuración mediante Filtros Heurísticos

En la fase inicial, el sistema aplica reglas lógicas de alta granularidad para eliminar el "ruido" digital y preparar el contenido. Las acciones clave incluyen:

- **Detección Lingüística:** Selección estricta de documentos que cumplen con el idioma objetivo.
- **Saneamiento de Formato:** Eliminación de caracteres inválidos y residuos de código que comprometen la legibilidad.
- **Validación de Relevancia:** Aplicación de umbrales de longitud para asegurar que cada registro aporte valor semántico real al modelo.

2.3.3.2. Consolidación de Unicidad (Deduplicación)

Para optimizar la eficiencia del entrenamiento y mitigar sesgos, este módulo identificó y removió registros redundantes mediante algoritmos de comparación exacta. Esta etapa es vital: al garantizar que cada documento sea estrictamente único, se previene que el modelo genere distorsiones por sobreexposición a datos repetidos, optimizando también los recursos de cómputo con los que contaba el proyecto.

2.3.3.3. Etiquetado Automatizado y Clasificación Multidimensional

Una vez refinada la base de datos, el contenido fue sometido a un proceso de clasificación mediante modelos de lenguaje de arquitectura BERT, los cuales fueron específicamente entrenados por nuestro equipo para decodificar la semántica y el contexto de los textos. Esta fase operó bajo un enfoque multitiqueta que permite asignar varios atributos a un mismo documento: un documento no queda confinado a una sola categoría, sino que recibe múltiples atributos simultáneos que capturan la complejidad de la información regional. Esta estructuración se desplegó sobre tres ejes rectores:

- **Identificación y Trazabilidad Geopolítica:** El sistema fue diseñado para detectar con precisión la procedencia y el contexto de 20 países de la región que integran este esfuerzo: Argentina, Bolivia, Brasil, Chile, Colombia, Costa Rica, Cuba, Ecuador, El Salvador, Guatemala, Honduras, México, Nicaragua, Panamá, Paraguay, Perú, Puerto Rico, República Dominicana, Uruguay y Venezuela. Bajo la lógica de etiquetas concurrentes, si un texto analiza, por ejemplo, acuerdos bilaterales o flujos migratorios entre Chile y Colombia, el modelo preserva esa interconexión orgánica, asignando ambas identidades geográficas de forma simultánea.

- **Arquitectura Temática y Especialización:** El contenido fue categorizado a partir de un catálogo de 10 tópicos (Figura 3), definidos previamente por nuestra curaduría experta. Al aplicar la flexibilidad multietiqueta sobre estas áreas de interés, se asegura que un documento que aborde transversalmente disciplinas como la economía y la salud reciba todas las clasificaciones pertinentes. Esta precisión permite que la organización del conocimiento responda con exactitud a los objetivos de representatividad y profundidad del proyecto.
- **Gobernanza Ética y Detección de Toxicidad:** Para salvaguardar la integridad de LatamGPT v1.0, se implementaron modelos especializados en seguridad encargados de auditar cada texto en busca de contenido dañino. Este sistema es capaz de identificar y marcar de manera concurrente diversas dimensiones de riesgo, desde sesgos implícitos hasta discursos de odio, facilitando el aislamiento preventivo del material tóxico. Con ello, permitimos que el entrenamiento del modelo se sustente sobre un ecosistema de datos seguro, ético y alineado con valores fundamentales.

2.3.3.4. Validación de Calidad Sintáctica y Estructural

Se implementó una etapa de evaluación de calidad de los textos, en la que se asigna un puntaje basado en criterios semánticos y sintácticos. El diseño de este módulo va más allá de la limpieza convencional y se sustenta en una premisa central del entrenamiento de modelos de lenguaje: la calidad de los datos puede influir más en el rendimiento final que su volumen. Este puntaje se utiliza tanto para seleccionar datos de alta calidad en el entrenamiento de LatamGPT como para guiar la aplicación de técnicas de aprendizaje curricular (curriculum learning).

Bajo este enfoque, y tomando como referencia las prácticas avanzadas de curaduría descritas en la literatura asociada al modelo Gopher, nuestro equipo diseñó e implementó más de 30 filtros estructurales de alta precisión. Estos filtros evaluaron aspectos como la granularidad léxica, la distribución de los signos de puntuación y la coherencia sintáctica general de las oraciones.

Al aplicar esta serie de auditorías lingüísticas, fue posible detectar y descartar de forma automatizada secuencias inconexas, residuos de código o fragmentos mal formados que logran eludir los filtros tradicionales. De esta manera, el proceso fomenta que el conjunto de datos resultante posea un estándar lingüístico de alta calidad, para asegurar la precisión, fluidez y éxito del entrenamiento de LatamGPT v1.0.

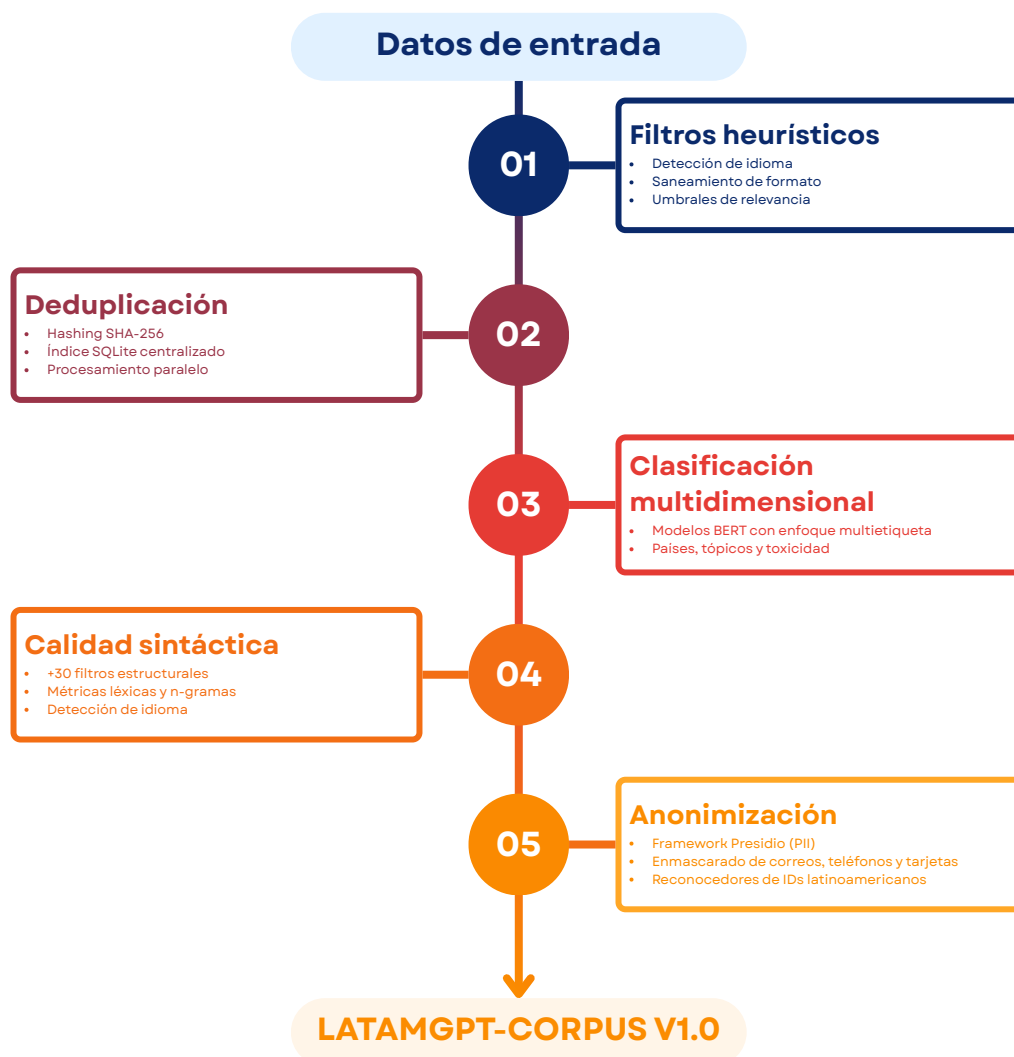
2.3.3.5. Anonimización y Resguardo de la Privacidad

Para garantizar la seguridad de la información y mitigar los riesgos de exposición de datos sensibles, el corpus fue sometido a un proceso de ofuscación utilizando el modelo estándar de anonimización Presidio. Esta fase se encargó de identificar y enmascarar sistemáticamente entidades de Información Personal Identificable (PII), neutralizando datos críticos como números de tarjetas de crédito, direcciones de correo electrónico y números telefónicos.

De manera complementaria a este estándar, el sistema fue adaptado para el contexto regional: se implementaron lógicas de detección específicas para anonimizar de forma exhaustiva los números de identificación nacional y sus respectivos códigos verifcadores (tales como el RUT en Chile) correspondientes a todos los países de Latinoamérica, asegurando así la total integridad ética de la base de datos antes de su uso.

Figura 5

Arquitectura del Pipeline de Procesamiento del LatamGPT-Corpus v1.0.



Fuente: Elaboración propia.

El flujo detalla las cinco etapas críticas de curaduría: 1) Filtros heurísticos para depuración inicial; 2) Deduplicación mediante una función hash SHA-256; 3) Clasificación multidimensional basada en modelos BERT para etiquetado geopolítico, temático y de toxicidad; 4) Validación de calidad sintáctica; y 5) Anonimización y Resguardo de la Privacidad. Este proceso da como resultado un conjunto de datos depurado, clasificado y de alta calidad.

2.4. Dimensión IV: Gestión de licencias y marco legal

2.4.1. Niveles de Apertura y Gobernanza del LatamGPT-Corpus v1.0

La clasificación de licencias del corpus es el reflejo tangible de la estrategia de gobernanza del proyecto y sus alianzas institucionales. Esta estructura se organiza en tres niveles de apertura, diseñados para equilibrar la transparencia técnica con un suficiente y adecuado respeto a los derechos de autor y dar cumplimiento al marco de gobernanza de datos del proyecto.

2.4.1.1. Niveles de Acceso (disponibilidad de los datos)

En coherencia con el principio de apertura de la iniciativa LatamGPT, se promueve que los datos aportados por las alianzas sean luego compartidos como parte del corpus abierto del proyecto. No obstante, considerando posibles restricciones legales, contractuales u otros, se definieron distintas modalidades, las cuales establecen las condiciones bajo las cuales los datos aportados podrán ser luego compartidos y reutilizados como parte del corpus abierto de LatamGPT. Las opciones ofrecidas a las alianzas fueron las siguientes:

- Completamente disponibles y abiertos (Open Data - Open Source)
- Disponibles y abiertos solo con fines de investigación
- Disponibles y abiertos para este y para futuros proyectos de CENIA, sujeto a confirmación adicional
- Disponibles y abiertos siempre y cuando sean refraseados utilizando técnicas de procesamiento natural de lenguaje
- Disponibles únicamente para este proyecto. No abiertos, luego de su uso serán eliminados de los repositorios de CENIA
- Otro (a definir de forma específica).

Según lo anterior, se implementan los siguientes niveles y vías de acceso al corpus de datos de LatamGPT:

- **Nivel de Acceso Libre (Naranja):** Integra datasets bajo licencias de libre uso (CC0 1.0, Open Data, ODC-By v1.0, Apache 2.0 y variantes CC BY). En esta categoría se han consolidado todos los activos provenientes de alianzas institucionales que optaron por compartir datos en las modalidades de:
 - "Completamente disponibles y abiertos (Open Data - Open Source)".
 - "Disponibles y abiertos siempre y cuando sean refraseados utilizando técnicas de procesamiento natural de lenguaje (NLP)".

Respecto a esta última opción, el proyecto asume el compromiso técnico de aplicar dichas transformaciones para garantizar que el output disponible en el repositorio público cumpla con los requisitos de seguridad y eventual reconocimiento de propiedad intelectual acordados. Estos activos se encuentran disponibles de forma abierta en un repositorio público de Hugging Face, permitiendo su uso, redistribución y adaptación en toda América Latina y el Caribe, condicionado solo a la "Restricción de Uso para Grandes Empresas" detallada más adelante en esta misma sección.

- **Nivel de Acceso Controlado (Azul - Investigación):** Comprende datos destinados exclusivamente a la investigación científica y académica. Su gestión se realiza mediante un repositorio restringido en Hugging Face, bajo estándares internacionales de custodia (análogos a protocolos de PhysioNet o LDC). En esta categoría se han integrado la totalidad de los activos provenientes de alianzas institucionales que, de acuerdo con el marco de gobernanza del proyecto, optaron explícitamente por la modalidad de "Disponibles y abiertos solo con fines de investigación".

Otro criterio utilizado para colocar activos en esta categoría se debe a las limitaciones para gestionar licencias de forma granular. Esto es así porque las licencias se gestionan a nivel de dataset y no es posible gestionarlas al nivel de sus componentes individuales. En datasets con componentes licenciados de forma heterogénea no es posible cumplir simultáneamente con los términos de varias licencias (aunque se trate de licencias libres) y esto condiciona la apertura de la totalidad del dataset. De esta forma, se decidió incluir en esta categoría a los datasets obtenidos de repositorios híbridos que

integran contenidos bajo diversas licencias Creative Commons (CC), ya que el uso con fines de investigación científica no es incompatible con los términos de la licencia CC más restrictiva.

Para acceder a estos activos, los interesados deberán formalizar una solicitud que acredite su afiliación institucional, declarar el propósito académico de la explotación de los datos y suscribir con CENIA un Acuerdo de Confidencialidad (NDA) que garantice el cumplimiento de los estándares éticos y legales del proyecto.

- **Nivel Cerrado (Rojo):** Define aquellos conjuntos de datos que, por determinación estratégica, legal o técnica, no serán publicados. Esta categoría integra la totalidad de los activos provenientes de alianzas institucionales que optaron por las siguientes modalidades en el documento de gobernanza:
 - "Disponibles únicamente para este proyecto. No abiertos, luego de su uso serán eliminados de los repositorios de CENIA".
 - "Disponibles y abiertos para este y para futuros proyectos de CENIA, sujeto a confirmación adicional".

Adicionalmente, este nivel incluye repositorios con licenciamientos internos altamente heterogéneos (que incluyen componentes en acceso abierto junto con otros de acceso restringido). Debido a las limitaciones de granularidad técnica de esta versión v1.0, se ha aplicado un criterio de restricción preventiva: ante la imposibilidad de segregar de forma automatizada componentes individuales con distintos permisos, se determinó que basta con que un solo elemento dentro de un conjunto sea restringido para clasificar la totalidad del dataset bajo este nivel. En esta fase inicial, hemos optado por ser deliberadamente más estrictos, priorizando la integridad del ecosistema legal y asegurando un tratamiento de la información coherente con los acuerdos de propiedad intelectual vigentes.

2.4.1.2. Restricción de Uso para Grandes Empresas

Independientemente del nivel de apertura descrito, el acceso, uso o redistribución de cualquier dato liberado por el proyecto por parte de una gran empresa está estrictamente condicionado a la obtención previa de un permiso por escrito de CENIA. El incumplimiento de esta política se considerará una violación a los términos de gobernanza. Para efectos de esta sección, se entenderá por gran empresa a toda entidad que, considerada en conjunto con sus matrices, filiales y demás entidades relacionadas, registre ingresos anuales superiores a US\$1.000 millones o que haya levantado capital acumulado superior a US\$1.000 millones, considerando aportes de capital, instrumentos convertibles en capital y financiamiento asimilable. Las entidades que no superen ninguno de estos umbrales podrán usar el corpus para los fines descritos en esta sección sin permiso previo. Se invita a estas entidades a informar su uso a CENIA con fines de trazabilidad, sin que ello constituya una condición de acceso. La condición se evaluará a la fecha de obtención del corpus; su superación posterior no afectará a los modelos ya entrenados con datos obtenidos previamente. CENIA podrá actualizar estos umbrales en futuras versiones del corpus.

Para efectos de esta restricción, se entiende por modelo fundacional aquel sistema de inteligencia artificial entrenado sobre grandes volúmenes de datos, de propósito general y adaptable a una amplia variedad de tareas posteriores mediante ajuste fino, adaptación o interfaces de programación. Se considera uso comercial el destinado a la explotación económica del modelo resultante o de servicios derivados de él, sea de forma directa o indirecta, con independencia de que el modelo se distribuya de forma abierta o cerrada.

Esta restricción no alcanza a la investigación académica, al desarrollo por parte de instituciones públicas o de la sociedad civil sin fines de lucro, ni al desarrollo de aplicaciones o modelos especializados construidos sobre modelos fundacionales preexistentes. Estas excepciones no autorizan, en ningún caso, el uso de los datos por parte de una gran empresa para el entrenamiento de modelos fundacionales destinados a uso comercial.

- **Fundamento: reciprocidad.**

Esta condición responde a un principio de reciprocidad. El LatamGPT-Corpus fue construido con el aporte voluntario de instituciones públicas, académicas y de la sociedad civil de América Latina y el

Caribe, mediante un esfuerzo colectivo orientado a fortalecer las capacidades tecnológicas de la región. Su liberación busca habilitar a esa comunidad, no subsidiar el desarrollo de modelos comerciales cuyos beneficios no retornan al ecosistema que hizo posible el corpus. CENIA reconoce que esta condición implica que el LatamGPT-Corpus, considerado como compilación, no califica como open source bajo la definición de la Open Source Initiative ni como open data bajo la Open Definition, dado que ambas excluyen restricciones por campo de uso o tipo de usuario. Se trata de una decisión deliberada, adoptada tras evaluar el trade-off entre apertura irrestricta y sostenibilidad del esfuerzo regional. CENIA evaluará esta condición en futuras versiones del corpus a la luz de la experiencia acumulada.

- **Alcance y mecanismos de aplicación.**

Esta restricción opera por dos vías complementarias:

- a. **Términos de acceso al repositorio:** El acceso a los repositorios donde CENIA pone a disposición el LatamGPT-Corpus está sujeto a la aceptación previa de los presentes términos de acceso, que el usuario acepta expresamente al momento de solicitar o efectuar la descarga. La descarga desde los repositorios, incluidos los del Nivel de Acceso Libre, estará sujeta a un mecanismo técnico de acceso condicionado (gated access) que exija la aceptación expresa de estos términos como paso previo. La obtención de los datos desde dichos repositorios constituye aceptación de estas condiciones y genera una relación contractual entre el usuario y CENIA, regida por la legislación chilena y sometida a la jurisdicción de los tribunales ordinarios de justicia con asiento en la comuna de Santiago de Chile.
- b. **Condición sobre la compilación:** Con independencia de lo anterior, esta restricción se aplica al LatamGPT-Corpus como compilación integrada, curada y distribuida por CENIA, comprendiendo su selección, organización, procesamiento, enriquecimiento, metadatos y estructura, elementos que constituyen el aporte intelectual propio del proyecto. Esta condición no modifica las licencias originales aplicables a los datos de terceros incorporados al corpus, ni limita su utilización cuando hayan sido obtenidos de manera independiente desde sus fuentes originales. Quien redistribuya el LatamGPT-Corpus como compilación, en todo o en parte sustancial, deberá incorporar estos mismos términos como condición de la redistribución y dejar constancia de ellos en la documentación o metadatos correspondientes.

- **Consecuencias del incumplimiento.**

El incumplimiento de esta restricción constituirá una infracción a los términos de acceso y a las condiciones de gobernanza y licenciamiento del LatamGPT-Corpus, y podrá dar lugar a la revocación del acceso, la suspensión de las autorizaciones concedidas y al ejercicio de las acciones legales que correspondan conforme a la legislación aplicable.

- **Solicitudes de autorización.**

Las organizaciones que requieran permiso conforme a esta sección pueden dirigir su solicitud a latam-gpt@cenia.cl. CENIA evaluará cada caso considerando, entre otros factores, el aporte de la iniciativa al ecosistema de IA de América Latina y el Caribe.

Nota sobre mecanismos de trazabilidad: los datasets liberados a los que se haya incorporado este mecanismo contienen señales técnicas de trazabilidad, como marcas de agua estadísticas, diseñadas para evaluar si fueron utilizados en el entrenamiento de un modelo de inteligencia artificial.

Nota sobre cumplimiento: Cada dataset preserva su licencia de origen en los metadatos. El usuario final asume la responsabilidad de respetar dichas licencias, mientras que el proyecto mantiene un canal permanente para identificar y retirar cualquier dato que no cuente con una licencia compatible (contacto: latam-gpt@cenia.cl).

2.4.1.3. Metodología de evaluación y transparencia

- Para cumplir con el objetivo de adoptar altos estándares de transparencia, el equipo técnico recibió capacitación especializada en gestión legal de datos. Este proceso se articuló mediante:

- **Auditoría de Compatibilidad:** Evaluación de la viabilidad técnica y legal de cada fuente, verificando existencia de licencia explícita compatible con los usos del corpus, ausencia de restricciones para entrenamiento de modelos, y tratamiento adecuado de datos personales. Las fuentes que no cumplieron estos criterios fueron excluidas del corpus.
- **Glosario de Términos:** Incorporación de una guía terminológica legal que facilita los procesos de auditoría externa y la comprensión clara por parte del usuario.

2.4.1.4. Criterios de Web Scraping para Información Pública y Fundamento Constitucional

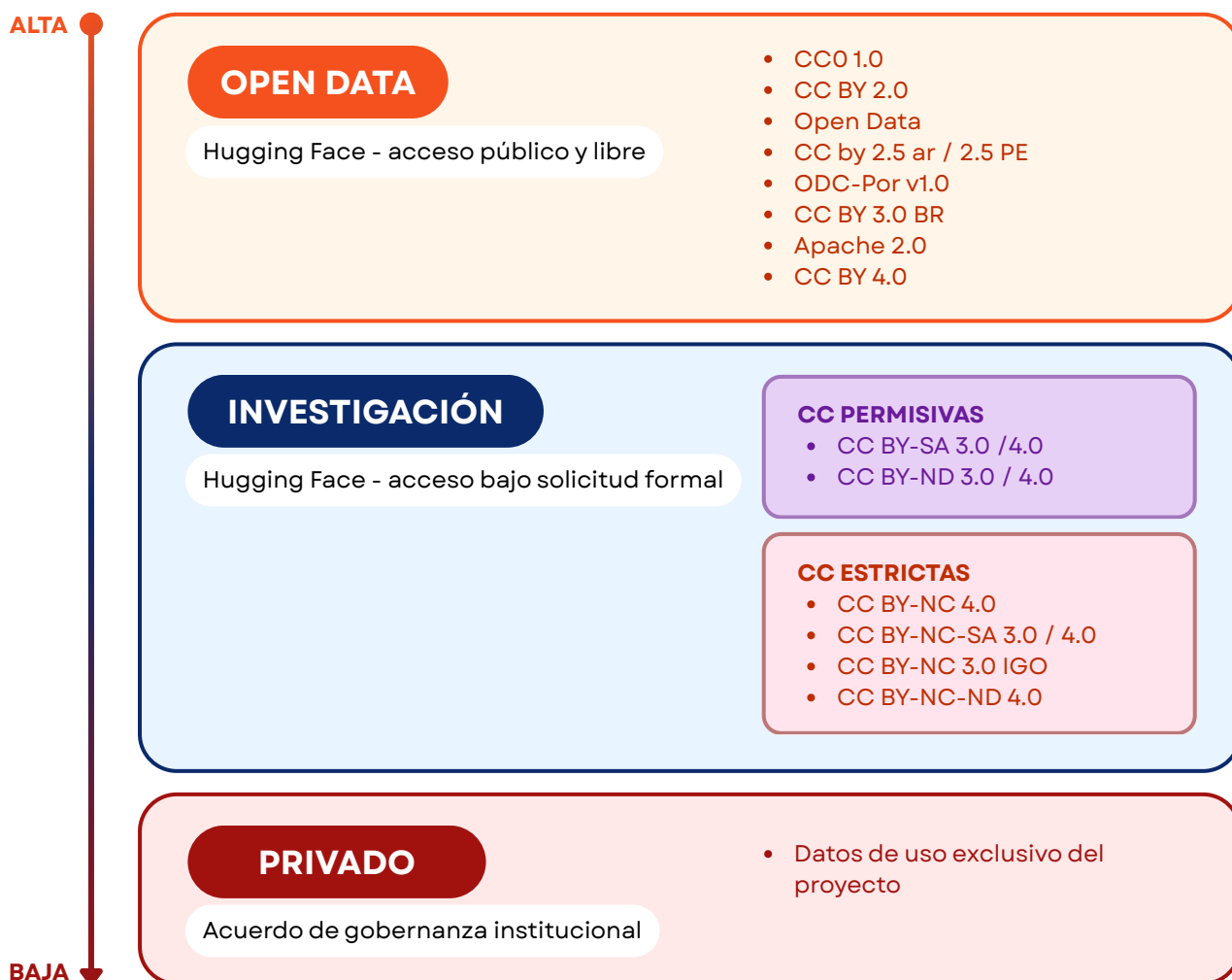
La recolección de datos provenientes de organismos del Estado se realizó partiendo de la premisa de que el derecho de acceso a la información pública es un derecho fundamental. En el marco jurídico local de Chile, este criterio se sustenta en el Artículo 8° de la Constitución Política de la República de Chile (2005), el cual establece el Principio de Publicidad, consagrandos que: "Son públicos los actos y resoluciones de los órganos del Estado, así como sus fundamentos y los procedimientos que elaboren". A su vez, la Ley Modelo Interamericana sobre Acceso a la Información Pública de la OEA (2010) prevé que toda la información publicada en cumplimiento de las obligaciones de transparencia activa por parte de los gobiernos debería "...permitir el más amplio acceso a dicha Información, de manera tal que permita su interoperabilidad en un formato de datos abiertos, así como determinar las estrategias para la identificación, generación, organización, publicación y difusión de dicha Información, permitiendo así su fácil reutilización por parte de la sociedad".

Tomando estos antecedentes como base interpretativa, se adoptaron los siguientes criterios de selección para el proceso de scraping:

- **Criterio de Inclusión:** Integración de actas de órganos deliberantes, portales gubernamentales e informes oficiales. Dado su carácter de alto interés público y su naturaleza de documento oficial del Estado, estos contenidos son esenciales para asegurar la representatividad democrática en el entrenamiento del modelo.
- **Criterio de Exclusión:** En estricto respeto a la normativa vigente, se omitieron exclusivamente aquellos dominios gubernamentales que prohíben de forma explícita la reutilización de datos o la limitan taxativamente al uso personal.

Figura 6

Esquema de clasificación y gobernanza de licencias del corpus.

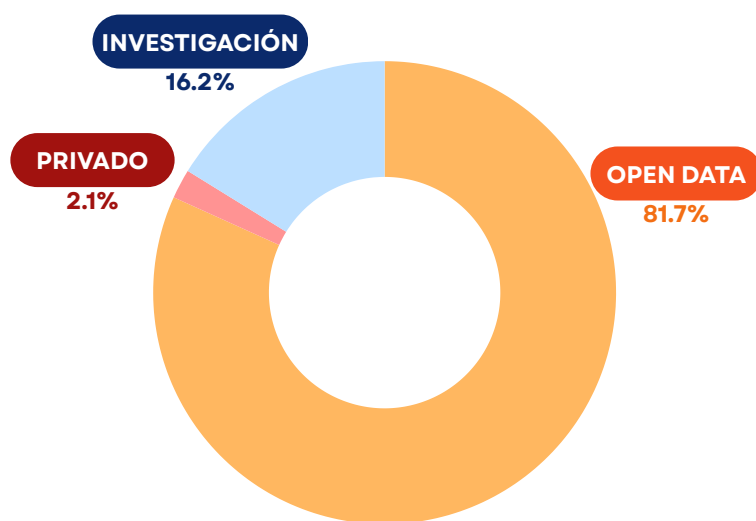


Fuente: Elaboración propia.

El diagrama organiza las fuentes de datos según su nivel de apertura, desde una disponibilidad máxima (Open Data en Hugging Face) hasta el acceso restringido (Datos Privados). Se detallan cinco categorías (Open Data, CC Permisivas, CC Estrictas, Investigación y Privado), especificando las variantes de Creative Commons (CC) aplicadas y las condiciones específicas de uso y apertura. Esta taxonomía permite una gestión de riesgos legal rigurosa y transparente, alineada con los acuerdos de gobernanza institucional.

Figura 7

Composición porcentual de licencias en el LatamGPT-Corpus v1.0.



Fuente: Elaboración propia.

El gráfico de anillo ilustra la distribución del volumen de datos según su marco legal de disponibilidad. Destaca una mayoría de Open Data (81.7%), consolidando la vocación de apertura del proyecto. El resto del corpus se distribuye entre licencias Privadas (2.1%) e Investigación (16.2%), reflejando la complejidad y diversidad de las fuentes integradas bajo protocolos de gobernanza responsable.

2.4.1.5. Mecanismos de Transparencia y Validación Científica

Establecer un modelo de gobernanza de datos con altos niveles de transparencia ha sido uno de los principales desafíos de LatamGPT v1.0. Para documentar este proceso de manera clara y verificable, se disponen los siguientes mecanismos:

- **Repositorio de Transparencia:** El inventario detallado de fuentes, metadatos y licencias se encuentra consolidado en un documento técnico disponible en <https://www.latamgpt.org/>. Este recurso está diseñado para facilitar auditorías, referencias y citas académicas por parte de la comunidad.
- **Documentación Científica:** El modelo de gobernanza y la metodología de curaduría se describen en un artículo de investigación actualmente en proceso de revisión en una revista del área. Una vez publicado, la referencia se incorporará en el repositorio del proyecto.

Análisis estadístico del corpus

El refinamiento de los datos recolectados mediante alianzas institucionales y web scraping dirigido se resume en la Figura 8 y las cifras que se presentan a continuación. El proceso de construcción de este componente del corpus partió de 157 mil millones de tokens (B) recopilados en bruto, resultado de un esfuerzo sostenido de colaboración con instituciones y organizaciones de la región. Tras la aplicación de los filtros de calidad y eliminación de duplicados descritos en la sección anterior, el volumen se redujo a 90 B, conservando únicamente el contenido que cumple con los estándares de relevancia lingüística, licenciamiento y calidad semántica establecidos para LatamGPT. El corpus final incorporado a la versión 1.0 del modelo asciende a 58,8 B tokens, lo que representa una tasa de retención del 37,4% respecto al material recopilado originalmente y constituye aproximadamente el 19,8% del corpus total de pre-entrenamiento.

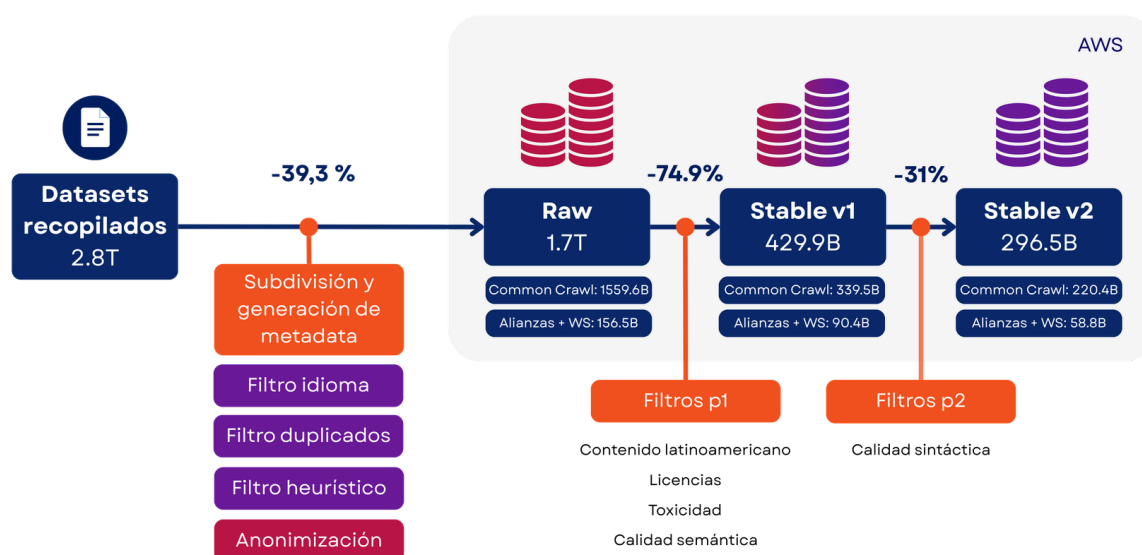
Este componente reviste particular importancia dentro del corpus, pues a diferencia de las fuentes de rastreo web generalista como Common Crawl, los datos provenientes de alianzas fueron obtenidos mediante acuerdos específicos con actores del ecosistema latinoamericano, garantizando una mayor pertinencia cultural, temática y lingüística del contenido. La distribución geográfica y temática de estos datos se detalla en las Figuras 9(a) y 9(b), respectivamente, donde se aprecia una representación amplia de los países de la región, con Argentina a la cabeza (20,0%) le sigue México (19,1%), así como una predominancia de contenidos relacionados con política (30,2%) y humanidades y ciencias sociales (24,9%).

Cabe señalar que en esta primera versión del corpus no se aplicaron criterios de normalización por país ni por categoría temática, por lo que la distribución observada refleja la disponibilidad natural de los datos recopilados y no una selección intencionalmente balanceada. Las implicancias de esta heterogeneidad sobre el comportamiento del modelo se discuten en las secciones 4 y 5.

Para consultar el desglose detallado y las estadísticas completas del corpus de pre-entrenamiento, acceda al siguiente enlace <https://www.latamgpt.org/dashboard>.

Figura 8

Pipeline de procesamiento del corpus.



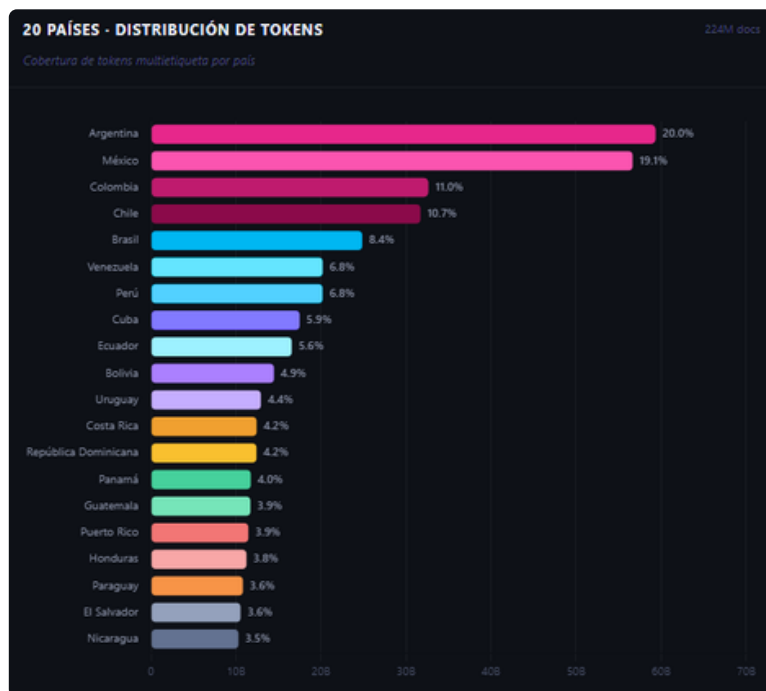
Fuente: Elaboración propia.

Partiendo de 2,8 billones de tokens (T) en datasets recopilados, el proceso aplica subdivisión y generación de metadatos, junto con filtros de idioma, duplicados, heurísticos y anonimización, reduciendo el volumen en un 39,3% hasta obtener un conjunto Raw de 1,7 T. Sobre este conjunto se aplican dos etapas de filtrado en AWS: los filtros p1 (contenido latinoamericano, licencias, toxicidad y

calidad semántica) generan el conjunto Stable v1 con 429,9 mil millones de tokens (B), y los filtros p2 (calidad sintáctica) producen el conjunto final Stable v2 con 296,5 B. En el conjunto final, las fuentes de alianzas y web scraping (WS) contribuyen con 58,8 B, equivalente a aproximadamente el 19,8% del corpus, mientras que Common Crawl aporta los 237,6 B restantes (80,2%).

Figura 9

a) Distribución de tokens por país en el conjunto Stable v2



b) Distribución de tokens por tópicos en el conjunto Stable v2



Fuente: Dashboard LatamGPT <https://www.latamgpt.org/dashboard>

(a) Argentina lidera con un 20,0% del total, seguida de México con 19,1%, seguidos de Colombia (11,0%), Chile (10,7%) y Brasil (8,4%), entre otros países de América Latina y el Caribe. **(b)** Las categorías de política (30,2%) y humanidades y ciencias sociales (24,9%) concentran casi la mitad del corpus, seguidas de economía y finanzas (11,4%), educación (11,4%) y comunicación y medios (9,4%). El desglose detallado y las estadísticas completas del corpus de pre-entrenamiento pueden consultarse en <https://www.latamgpt.org/dashboard>

Principales hallazgos



América Latina y el Caribe lleva años produciendo datos de manera acelerada –a través de sus sistemas de salud, educación, justicia y gobierno– pero una parte importante de esa información no llega a estar disponible para quienes desarrollan tecnología de inteligencia artificial. El ILIA (2024, 2025) lo documenta con claridad: aunque la región ha mejorado su capacidad para generar y organizar datos, la disponibilidad efectiva de esos datos sigue estancada en la mayoría de los países. En este contexto, el proyecto LatamGPT v1.0 se propuso enfrentar ese problema de frente, construyendo un corpus de entrenamiento que incorpora conocimiento real de la región, bajo reglas claras y con respeto a los derechos de quienes producen esa información.

Este desafío no es exclusivo de América Latina: en años recientes ha emergido a nivel global una tercera ola de modelos de lenguaje abiertos con estrategia regional explícita, donde la apertura deja de ser únicamente un principio científico y se convierte en un instrumento de posicionamiento tecnológico. Iniciativas como BLOOM (Europa, 176B parámetros), Falcon (Medio Oriente), Qwen (China), SEA-LION (Sudeste Asiático) o AfroLLaMA (África) demuestran que la optimización de los modelos fundacionales no se realiza únicamente en función de benchmarks globales, sino de utilidad en ecosistemas locales. No obstante, como evidencia la matriz de transparencia de las Tablas 2 y 3, esa apertura es profundamente desigual: modelos como Mistral o Gemma no declaran origen ni corpus, LLaMA 3 publica el modelo pero no sus datos de entrenamiento, y solo un puñado de iniciativas –Falcon, ALIA-40B, Apertus 70B– alcanzan transparencia plena en las tres dimensiones evaluadas: origen, licencia y corpus público. La Tabla 4 añade otra dimensión del problema: incluso los recursos de voz y corpus de referencia disponibles para lenguas no anglófonas son escasos, fragmentados y en su mayoría orientados a lenguas africanas o europeas minoritarias, dejando al español latinoamericano y al portugués brasileño sin infraestructura de datos comparable. La escala –medida en parámetros o tokens– sigue siendo relevante, pero la calidad, representatividad y alineación con contextos específicos emergen como variables igualmente críticas. LatamGPT v1.0 se inserta directamente en esta tendencia como una iniciativa regional de escala relevante, caracterizada por documentar su proceso de construcción y gobernanza de datos.

Tabla 2

Modelos de lenguaje globales de propósito general.

Modelo / Organización	Idiomas cubiertos / tipo lingüístico	Lenguas indígenas	Dominio / benchmarks	Matriz de transparencia		
				Origen	Licencia	Corpus público
BLOOM 176B BigScience · Jul. 2022	46 idiomas naturales + 13 lenguajes de programación Propósito general multilingüe	No (presentes lenguas africanas minoritarias: yoruba, swahili)	General SuperGLUE, WinoBias, CrowS-Pairs	Parcial	Parcial	Parcial
Notas - ROOTS: 252 fuentes. Mejor intención de transparencia; el acceso al corpus requiere solicitud. arXiv:2211.05100 arXiv:2303.03915						
LLaMA 2 (7B-70B) Meta · EE. UU. · Jul. 2023	Inglés (predominante) Monolingüe (EN)	No	General MMLU, HumanEval, GSM8K	Sí	Parcial	No
Notas - CommonCrawl, C4, GitHub, Books, ArXiv. arXiv:2307.09288						
Falcon 7B / 40B / 180B TII · EAU · 2023	Inglés (~80 %); fragmentos FR, DE, ES, PT Casi-monolingüe + europeo menor	No	General Open LLM Leaderboard, ARC, HellaSwag	Sí	Sí	Sí
Notas - RefinedWeb: 600B tokens liberados. Mayor transparencia en datos. arXiv:2306.01116						
Mistral 7B / Mixtral 8×7B Mistral AI · Francia · Sep-Dic 2023	Inglés (inferido); algunos europeos Sin declaración oficial	No declarado	General MT-Bench, MMLU (sin detalles de corpus)	No	No	No
Notas - Nula divulgación de datos. FMTI (Stanford): puntuación mínima. arXiv:2310.06825						
Gemma 1 / 2 Google DeepMind · EE. UU. · 2024	Inglés (predominante) Monolingüe + web diverso sin desglose	No	General MMLU, GSM8K, HumanEval, JGLUE (Takane)	Parcial	No	No
Notas - Categorías declaradas, sin proporciones ni dataset liberado. arXiv:2403.08295 arXiv:2408.00118						
LLaMA 3 / 3.1 / 3.2 / 3.3 Meta · EE. UU. · Abr-Dic 2024	EN (~95 %) + 8 oficiales: DE, FR, IT, PT, HI, ES, TH Multilingüe limitado	No	General MMLU Pro, IFEval, LiveCodeBench	Parcial	Parcial	No
Notas - 15 T tokens; 4× más código. Sin dataset unificado liberado. ai.meta.com/llama						
DeepSeek-V2 / V3 DeepSeek · China · May-Dic 2024	Inglés + chino (mayoría) Bilingüe EN-ZH con multilingüe parcial	No	General AIME, CNMO, LiveCodeBench, MMLU	Parcial	No	No
Notas - V3: 14,8 T tokens; 685B parámetros (MoE). arXiv:2405.04434 arXiv:2412.19437						

Panorama de los principales LLM de gran escala con predominio del inglés, ordenados cronológicamente (2022-2024). Para cada modelo se indica la organización desarrolladora, la cobertura lingüística declarada, el tipo lingüístico, la presencia de lenguas indígenas, el dominio de conocimiento y los benchmarks de evaluación reportados. La Matriz de Transparencia evalúa tres dimensiones independientes de apertura de datos –Origen, Licencia y Corpus público– según la escala de tres niveles descrita en la nota al pie. Matriz de Transparencia. Cada dimensión se codifica en tres niveles: Sí – dato declarado y públicamente verificable; Parcial – declarado pero con acceso o detalle limitado; No – sin declaración ni divulgación. Las tres dimensiones evaluadas son: Origen (procedencia de los datos de entrenamiento declarada públicamente), Licencia (licencia de los datos especificada y abierta) y Corpus público (disponibilidad del corpus para inspección o descarga). Los identificadores de artículos se expresan en notación arXiv (arXiv:AAAA.NNNNN); las URL corresponden a repositorios o sitios oficiales de cada iniciativa. La información refleja lo declarado por las organizaciones a la fecha de compilación y no constituye una auditoría independiente.

Tabla 3

Modelos soberanos y regionales.

a) Europa y América Latina

Modelo / Organización	Idiomas cubiertos / tipo lingüístico	Lenguas indígenas	Dominio / benchmarks	Matriz de transparencia		
				Origen	Licencia	Corpus público
EUROPA						
ALIA-40B BSC-CNS · España · Ene. 2025	35 lenguas europeas (ES 16 %, EN 39 %, CA 2 %, GL 0,31 %, EU 0,24 %) Europeo multilingüe soberano	No (cooficiales sí: catalán, gallego, euskera)	General MMLU (ES/CA); benchmarks del BSC	Sí	Sí	Sí
Notas - MareNostrum 5. Entrenado desde cero. Verificado por AESIA. alia.gob.es bsc.es						
Apertus 70B EPFL / ETH Zürich · Suiza · Sep. 2025	>1.000 idiomas; 40 % no-inglés; romanche, alemán suizo Masivamente multilingüe – soberano	Parcial (lenguas minorizadas europeas y otras)	General Open LLM Leaderboard; multilingüe propio	Sí	Sí	Sí
Notas - Pesos, datos, recetas y checkpoints publicados. Conforme al EU AI Act. ethz.ch hf.co/swiss-ai						
AMÉRICA LATINA						
LatamGPT CENIA · Chile · Feb. 2026	Español, portugués; inglés Regional (LATAM) – ES / PT + EN	No (incluye el tópico «pueblos originarios»; su cobertura lingüística se prevé para próximas versiones)	Regional: Latinoamérica y el Caribe CHOCLO, Trueque (conocimiento latinoamericano)	Sí	Sí	Sí
Notas - 15 países, 70+ instituciones. Base: LLaMA 3.1 70B. ~300B tokens. Presupuesto ~US \$3M. latamgpt.org						

b) África e India.

Modelo / Organización	Idiomas cubiertos / tipo lingüístico	Lenguas indígenas	Dominio / benchmarks	Matriz de transparencia		
				Origen	Licencia	Corpus público
ÁFRICA						
Masakhane (proyectos NLP) Comunidad panafricana · 2019– presente	>2.000 lenguas africanas cubiertas; modelos específicos en 50+ Comunidad distribuida – traducción, NER, voz	Sí (foco central: lenguas africanas)	General MasakhaNER, MAFAND-MT, IrokoBench, AfriSenti	Sí	Sí	Sí
Notas - ~1.000 participantes, 30 países. Ecosistema de datasets y modelos, no un único LLM. masakhane.io						
AfroLLaMA Jacaranda Health · Kenia · 2024	Swahili, Xhosa, Zulu, Yoruba, Hausa + inglés Fine-tune africano (LLaMA 3 8B)	Sí (lenguas subsaharianas de baja representación)	Salud materna; NLP general AfroBench (AfriMMLU, AfriXNLI), IrokoBench	Parcial	Sí	Parcial
Notas - Continual pretraining de LLaMA 3 8B sobre corpus africano. hf.co/Jacaranda/AfroLlama_V1						
InkubaLM 0.4B Lelapa AI · Sudáfrica · Ago. 2024	IsiZulu, Yoruba, Hausa, Swahili, IsiXhosa + EN + FR Small LM africano desde cero (SLM)	Sí (construido para lenguas africanas)	General Inkuba-Instruct, IrokoBench (AfriMMLU, AfriXNLI)	Sí	Sí	Sí
Notas - Entrenado desde cero: 2,4B tokens. Inkuba-Mono e Inkuba-Instruct liberados. arXiv:2408.17024						
INDIA						
Airavata AI4Bharat / IIT Madras · India · Ene. 2024	Hindi (principal); expansión prevista a 22 lenguas oficiales Instruction-tuned para hindi (OpenHathi / LLaMA 2)	Parcial (hindi, de baja cobertura en LLMs globales)	General IndicXTREME, Indic NLG Suite, MMLU (traducida)	Sí	Sí	Sí
Notas - Datasets curados por humanos, sin datos de GPT-4. Base: OpenHathi (LLaMA 2 extendido). arXiv:2401.15006						

Iniciativas nacionales y regionales orientadas a la soberanía lingüística y a la inclusión de lenguas subrepresentadas, agrupadas por región (Europa, América Latina, África e India). Las columnas y la Matriz de Transparencia siguen el mismo esquema que la Tabla 2.

Tabla 4

Datasets de voz y corpus de referencia.

Recurso / Organización	Idiomas cubiertos / tipo lingüístico	Lenguas indígenas	Dominio / benchmarks	Matriz de transparencia		
				Origen	Licencia	Corpus público
NaijaVoices NaijaVoices Community / Lanfrica · Nigeria · 2025	Igbo, Hausa, Yoruba (3 lenguas de Nigeria) Dataset de voz comunitario – 5.455 hablantes, 1.838 horas	Sí (lenguas con escasa representación digital)	Cultura nigeriana: 100 temas culturales (gastronomía, tradiciones, conteo, historia oral) ASR: Whisper, MMS, XLSR – hasta 75,86 % WER con fine-tuning	Sí	Sí	Sí
Notas - Dataset voz-texto. Data farming: consentimiento individual por hablante y supervisión en campo. Licencia CC BY-NC-SA 4.0. arXiv:2505.20564 naijavoices.com						
Common Corpus PleiAs · Francia · Feb. 2025	EN, FR, DE, ES, IT, PT, polaco, griego, ruso, latín y otras (>30 idiomas) Multilingüe europeo – dominio público y licencias abiertas verificadas por documento	Parcial (minoritarias: luxemburgués, gaélico irlandés, serbio, checo, griego)	Gobierno, cultura, ciencia, código, web y datos semánticos abiertos (Wikipedia, Wikidata) Pleias 1.0; Salamandra (BSC); Lucie 7B; NeKo (Nvidia); experimentos de Anthropic	Sí	Sí	Sí
Notas - ~2 billones de tokens en dominio público o licencias abiertas verificadas por documento; sin CommonCrawl ni doctrina de fair use. Seis colecciones (Open Government, Culture, Science, Code, Web, Semantic). Reconocido en el Paris AI Summit 2025. arXiv:2506.01732						

Recursos que no constituyen LLM de texto, incorporados como referencia metodológica de recolección ética de datos (data farming) y apertura legal verificada por documento. Se conservan las mismas dimensiones de la Matriz de Transparencia para permitir la comparación directa con las Tablas 2 y 3.

4.1. La brecha de conocimiento latinoamericano en los grandes modelos de IA

Cuando se entrena un gran modelo de lenguaje se utilizan enormes cantidades de texto extraído de internet. El problema es que internet no representa a todos por igual. Baack (2024) analizó en detalle Common Crawl –la principal fuente de datos para estos modelos– y mostró que su contenido está sesgado hacia los países con mayor infraestructura digital, en desmedro de regiones como América Latina y el Caribe o África. Esto tiene consecuencias concretas: los modelos más usados hoy no entienden bien el español y portugués latinoamericano en toda su diversidad, no reconocen contextos culturales locales, y prácticamente ignoran lenguas y culturas indígenas.

La brecha es más amplia de lo que parece a simple vista. América Latina representa el 6,6% del PIB mundial y el 8,8% de la población, pero solo el 1,12% de la inversión global en inteligencia artificial (ILIA, 2024, 2025). Esa desproporción se refleja también en los datos con los que se entrenan los modelos. Longpre et al. (2023) mostraron que la mayoría de los grandes corpus de entrenamiento carecen de documentación sobre su origen y sus condiciones de uso, lo que hace muy difícil saber qué se incluye y qué se deja fuera. Para América Latina y el Caribe, lo que suele quedarse fuera es precisamente lo más valioso: el conocimiento que vive en archivos, bibliotecas, universidades, parlamentos y organismos públicos que no tienen presencia en la web abierta. LatamGPT v1.0 busca recuperar exactamente ese tipo de contenido, a través de acuerdos directos con más de 79 instituciones de 15 países de la región.

Esta realidad tiene un nombre en la literatura especializada: Langlais et al. (2025) la denominan la paradoja del dato abierto –el hecho de que la mayor parte del contenido disponible bajo licencias libres o sin derechos de autor permanece prácticamente invisible en los corpus de pre-entrenamiento existentes. En su análisis de Common Corpus, el conjunto de datos abiertos más grande del mundo para pre-entrenamiento de LLMs (aproximadamente 2 billones de tokens), los autores demuestran que menos del 2% del material archivado en la web tiene licencias abiertas, y que los sistemas de recolección basados en rastreo web sistemáticamente privilegian fuentes de alta visibilidad digital –que no son necesariamente las de mayor valor cultural o institucional–. Para América Latina, donde una proporción significativa del conocimiento relevante reside en repositorios institucionales, archivos públicos y publicaciones académicas con baja presencia web, esta paradoja se agudiza especialmente.

4.2. Por qué la gobernanza de datos importa

Recolectar datos no es suficiente: lo que importa es cómo se recolectan, quién los autoriza, y bajo qué condiciones pueden usarse. Jernite et al. (2022) argumentaron que la falta de marcos claros de gobernanza de datos es uno de los principales problemas del desarrollo actual de la IA, porque permite que los modelos se entrenen con información de dudosa procedencia o sin el consentimiento de quienes la produjeron. El ILIA 2025 apunta en la misma dirección: señala que las estrategias de IA en la región tienden a quedarse en declaraciones de intención, sin planes de implementación ni mecanismos de evaluación concretos –lo que el informe llama la brecha entre disponibilidad declarada y apertura efectiva–.

LatamGPT v1.0 tomó el camino distinto. En lugar de asumir que podía usar los datos disponibles en internet, el proyecto gestó acuerdos formales con cada institución colaboradora, especificando qué datos se usarían, para qué y con qué límites. Pahune et al. (2025) sostienen que la gobernanza de datos debe funcionar como una infraestructura técnica –no como un conjunto de principios abstractos– y eso es exactamente lo que el proyecto implementó: un sistema de seis niveles de apertura que va desde datos completamente públicos hasta datos de uso exclusivo para el entrenamiento, además de un registro público donde cualquier persona puede consultar el origen y las condiciones de cada fuente. Este nivel de transparencia es inusual en el campo, y responde directamente al diagnóstico del ILIA 2025 sobre la brecha entre disponibilidad y apertura efectiva de datos en la región.

Un referente metodológico relevante en este aspecto es Common Corpus (Langlais et al., 2025), el conjunto de datos abierto más grande del mundo para pre-entrenamiento de LLMs, con



aproximadamente 2 billones de tokens desarrollado principalmente en Europa. Common Corpus documenta de forma exhaustiva la procedencia de cada fuente, incluye metadatos de licencia por documento y sigue las buenas prácticas del Dataset Convening Internacional (Baack et al., 2025), entre ellas: registrar preferencias de uso, involucrar a comunidades locales en la identificación de fuentes y publicar las herramientas de procesamiento utilizadas. LatamGPT v1.0 comparte esta filosofía de gobernanza documentada, con la especificidad adicional de haber construido acuerdos institucionales formales en un contexto legal más restrictivo y con una red de alianzas de mayor diversidad geopolítica. Ambos proyectos demuestran, desde contextos distintos, que es posible construir infraestructura de datos a gran escala sin sacrificar ni la legalidad ni la transparencia.

4.3. El laberinto legal: riesgos y cómo navegarlos

Construir un corpus de entrenamiento en América Latina y el Caribe implica navegar uno de los paisajes legales más complejos del mundo para este tipo de actividad. Flynn et al. (2022) revisaron las excepciones al derecho de autor para minería de texto y datos (TDM) a nivel global, y concluyeron que América Latina y el Caribe es la región del mundo con **menos protecciones legales para investigadores que quieran extraer datos de la web con fines científicos**. Los países del Norte Global en general cuentan con excepciones que permiten usar contenido protegido con fines de investigación sin autorización previa; en la mayoría de los países latinoamericanos, esas excepciones simplemente no existen. De hecho, Cuba es el único país de la región que cuenta con una excepción para actividades de investigación lo suficientemente amplia como para cubrir el uso de TDM (Artículo 86.1 de la Ley n.º 154 Cubana). Este contexto normativo genera un entorno de alta incertidumbre para los proyectos académicos de desarrollo de modelos. A esto se suma el hecho de que los proyectos académicos cuentan con recursos limitados y no pueden absorber el coste de posibles reclamaciones legales, como hacen las grandes empresas tecnológicas. Paradójicamente, estas mismas grandes empresas tecnológicas operan desde países cuyos marcos legales les permiten extraer datos de todo el mundo, incluidos los de nuestra región.

Con el objetivo de abordar el desafío de la gestión legal y responsable de los datos en este panorama de alta incertidumbre, LatamGPT v1.0 desarrolló una estrategia de gestión de riesgos basada en priorizar fuentes con licencias abiertas y/o con autorización explícita y documentar cada decisión de inclusión o exclusión de fuentes obtenidas mediante web scraping. Kim et al. (2025) han señalado que muchos proyectos de IA no verifican realmente las licencias de los datos que usan; LatamGPT v1.0 ha optado por hacer exactamente eso, aunque ello suponga un corpus más pequeño y un proceso más lento. En cualquier caso, es importante destacar que los elevados costes de transacción para obtener datos y la alta inseguridad jurídica generada por la ausencia de excepciones al derecho de autor, sitúan a los proyectos de investigación como LatamGPT v1.0 en una clara desventaja comparativa respecto a los que se desarrollan en países con marcos normativos adecuados para la gestión de datos.

Conclusiones y transparencia de datos

La publicación del presente Catálogo Referencial del LatamGPT-Corpus v1.0 constituye un nuevo hito en la edificación de una infraestructura digital soberana para América Latina y el Caribe. Este documento complementa el reporte técnico del modelo para ofrecer un protocolo de gobernanza, demostrando que el desarrollo de activos de Inteligencia Artificial en la región es viable bajo estándares de rigor ético y colaboración institucional.

El valor estratégico de LatamGPT v1.0 radica en su equilibrio entre especificidad regional y escala global. El corpus final de entrenamiento alcanza una dimensión de 296,5B de tokens, una cifra competitiva que garantiza la robustez y capacidad de generalización del modelo.

Cabe destacar que 58,8 mil millones de tokens provienen exclusivamente de nuestra Red de Alianzas. Este volumen es el resultado de un proceso de refinamiento técnico que depuró una ingesta inicial de 157B de tokens institucionales. Al integrar este componente de alta fidelidad con fuentes complementarias de acceso público, se ha logrado un corpus que no solo destaca por su volumen, sino por una densidad cualitativa que mitiga sesgos y asegura una representación más adecuada de los registros académicos, jurídicos y técnicos de nuestra región.

La ejecución de esta primera etapa ha requerido un esfuerzo intensivo en la articulación de alianzas y la generación de marcos de confianza desde una base de partida inexistente. La complejidad de establecer protocolos de gobernanza de datos desde cero, garantizando la seguridad jurídica de cada contribución, ha dictado un ritmo de ejecución deliberado y cauteloso.

Este proceso ha permitido cimentar las bases relacionales necesarias para la sostenibilidad del proyecto. Con una red de colaboración ya operativa y marcos de seguridad validados, se proyecta que las futuras fases de expansión del corpus se desarrollen con una mayor agilidad operativa, aprovechando la red de colaboración y la estructura de gobernanza establecidas..

En conclusión, LatamGPT v1.0 no solo entrega un corpus verificado de 296,5B de tokens; establece una hoja de ruta para la gestión de activos digitales que equilibra la vanguardia tecnológica con el respeto a la propiedad intelectual y la democratización del conocimiento soberano. Más aún, sienta las bases para que América Latina transite de un rol predominantemente consumidor de inteligencia artificial hacia uno activo en la construcción de sus propias infraestructuras de conocimiento. La continuidad de este esfuerzo dependerá de la inversión sostenida en infraestructura, la profundización de la colaboración regional y la capacidad de mantener estándares de calidad y transparencia; pero incluso en esta primera versión, LatamGPT v1.0 demuestra que es posible construir inteligencia artificial desde América Latina, con identidad propia y con vocación de impacto global.

La disponibilidad de un corpus con representación genuina de la realidad lingüística y cultural de América Latina y el Caribe tiene implicancias que van más allá del rendimiento técnico del modelo. En sectores críticos como educación, salud y servicios públicos, contar con modelos entrenados sobre datos pertinentes a la región puede mejorar significativamente la calidad de las aplicaciones y reducir los sesgos que hoy afectan a las poblaciones más vulnerables. Un sistema de salud que procesa datos en español rioplatense, en portugués brasileño o en lenguas indígenas requiere modelos que comprendan esas realidades; LatamGPT v1.0 es el primer paso estructurado hacia esa posibilidad.

Igualmente relevante es el impacto del proyecto en la formación de talento y el fortalecimiento de capacidades institucionales. La apertura del proceso –desde la metodología de curaduría hasta los mecanismos de gobernanza– permite que equipos de investigación, universidades y organismos públicos de la región aprendan haciendo: no solo consumiendo el modelo resultante, sino participando activamente en la construcción de la infraestructura que lo sustenta. Esta dimensión de transferencia de conocimiento es, quizás, uno de los legados más duraderos del proyecto.

En el contexto global, LatamGPT v1.0 se posiciona como la contraparte latinoamericana de iniciativas como Common Corpus (Langlais et al., 2025), reconocida en el AI Summit de París 2025 como la mayor base de datos abierta del mundo para el entrenamiento de LLMs. Mientras Common Corpus ha priorizado lenguas europeas y el acervo cultural del dominio público occidental, LatamGPT v1.0 aporta lo que ese ecosistema no puede proveer: representación del español y portugués latinoamericano en su diversidad real, conocimiento institucional y académico de la región, y –de manera inédita– datos



provenientes de acuerdos directos con organismos públicos latinoamericanos. Estas dos iniciativas no son competidoras sino complementarias, y abren la posibilidad de articulaciones futuras que enriquezcan mutuamente sus acervos en beneficio de toda la comunidad hispanoparlante y lusoparlante global.

Anexo: glosario de términos y referencias

Activo digital / activo de información Conjunto de datos o recursos de información que poseen valor estratégico para una organización o proyecto, y que pueden ser gestionados, protegidos y reutilizados.

Alianza estratégica Acuerdo formal entre instituciones que establece una colaboración orientada a objetivos comunes, en este caso el intercambio y custodia de datos para el desarrollo del modelo.

Auditoría jurídica Proceso de revisión legal exhaustiva de cada fuente de datos, destinado a verificar la viabilidad de su uso dentro del marco normativo aplicable.

BERT Arquitectura de modelo de lenguaje (Bidirectional Encoder Representations from Transformers) utilizada para tareas de clasificación semántica y etiquetado de texto. En el proyecto se empleó post-entrenada para clasificar contenido regional.

Common Crawl Base de datos pública que contiene rastreos masivos de páginas web de internet, utilizada como fuente de datos en el entrenamiento de modelos de lenguaje a gran escala.

Convenio específico de transferencia Instrumento jurídico que regula de forma precisa el intercambio y uso de conjuntos de datos determinados entre instituciones colaboradoras.

Convenio marco de colaboración Acuerdo general que establece las bases de una relación estratégica de largo plazo entre instituciones, sin entrar en el detalle de operaciones específicas.

Corpus Conjunto estructurado y curado de textos utilizados para entrenar un modelo de lenguaje. En este caso, representa el acervo lingüístico y temático de América Latina y el Caribe.

Curaduría de datos Proceso de selección, evaluación, clasificación y validación de fuentes y contenidos que asegura la calidad, pertinencia y representatividad de los datos incorporados al corpus.

Dataset Conjunto de datos estructurado y delimitado, generalmente disponible en repositorios como Hugging Face o GitHub, utilizado para entrenamiento o validación de modelos.

Deduplicación Proceso automatizado de identificación y eliminación de registros duplicados dentro del corpus, con el fin de evitar sobreexposición del modelo a información repetida y optimizar recursos de cómputo.

Detección de toxicidad Módulo del pipeline que identifica y marca contenido potencialmente dañino – como discurso de odio o sesgos implícitos– para su aislamiento preventivo antes del entrenamiento.

Etiquetado multietiqueta Técnica de clasificación que asigna múltiples atributos simultáneos a un documento, permitiendo capturar su complejidad temática, geográfica y ética sin reducirlo a una sola categoría.

Filtros heurísticos Reglas lógicas aplicadas en la fase inicial del pipeline para eliminar ruido digital, caracteres inválidos, textos demasiado cortos y contenido irrelevante.

Gobernanza de datos Sistema de normas, procedimientos e instrumentos jurídicos que regula la disponibilidad, el acceso, el uso y la custodia de la información a lo largo de todo su ciclo de vida.

Hugging Face Plataforma de repositorios abiertos para modelos de lenguaje y datasets de inteligencia artificial, utilizada en este proyecto tanto para la curaduría de fuentes como para la publicación del corpus.

MLE (Machine Learning Engineer) Profesional especializado en el diseño, implementación y optimización de sistemas de aprendizaje automático, incluyendo pipelines de procesamiento de datos.

Memorando de Entendimiento (MoU) Instrumento que formaliza la voluntad de cooperación entre dos o más instituciones, estableciendo objetivos comunes sin constituir un contrato vinculante.

Metadatos Información estructurada que describe las características de un conjunto de datos: su origen, licencia, temática, idioma, fecha de creación, entre otros atributos.

Minería de texto y datos (TDM) Técnica computacional que extrae información y patrones a partir de grandes volúmenes de texto o datos estructurados, relevante en el contexto legal de los derechos de autor.

NDA (Non-Disclosure Agreement / Acuerdo de Confidencialidad) Contrato legal que garantiza el tratamiento protegido de información sensible o reservada compartida entre las partes colaboradoras.

Oficio de coordinación institucional Mecanismo de comunicación formal entre organismos del Estado, gestionado a través de ministerios, para articular colaboraciones de alto nivel.

Pipeline de procesamiento Flujo computacional secuencial que organiza y automatiza las distintas etapas de transformación y refinamiento del corpus, desde la ingesta hasta la validación final.

Pre-entrenamiento Fase inicial del desarrollo de un modelo de lenguaje en la que se entrena sobre grandes volúmenes de texto para que adquiera representaciones generales del lenguaje.

Repositorio Espacio digital donde se almacenan, organizan y comparten conjuntos de datos o código, como GitHub o Hugging Face.

Soberanía digital Capacidad de una región o nación de producir, gestionar y controlar su propia infraestructura tecnológica y sus datos, sin dependencia estructural de actores externos.

Token Unidad mínima de procesamiento en un modelo de lenguaje, que puede corresponder a una palabra, parte de una palabra o un signo de puntuación. Se utiliza como medida del tamaño de un corpus.

Trazabilidad Capacidad de rastrear el origen, recorrido y condiciones de uso de cada fuente o dato integrado al corpus, garantizando transparencia y responsabilidad en todo el proceso.

Web scraping Técnica de extracción automatizada de contenido desde sitios web mediante protocolos técnicos autorizados, utilizada para incorporar fuentes institucionales al corpus.

Tabla 5

Licencias identificadas y condiciones generales de uso

Licencia	Descripción (interpretación literal)	Fuente principal
CC0 1.0	Renuncia completa a derechos de autor y derechos conexos; el dataset puede usarse sin restricciones y sin atribución.	Creative Commons (CC0 1.0 Universal Public Domain Dedication)
Open Data	Los datos pueden ser libremente usados, reutilizados y redistribuidos por cualquiera, sujeto como máximo a atribución. (incluye información pública, contenidos en dominio público o de uso autorizado sin licencia estándar).	Categoría residual basada en el análisis de los términos y condiciones de las publicaciones web y los criterios de gestión de riesgo de LatamGPT v1.0
ODC-By v1.0	Permite copiar, distribuir, transmitir y adaptar la base de datos, siempre que se atribuya al autor original.	Open Data Commons (ODC-By License v1.0)
Apache 2.0	Permite uso, reproducción, modificación y distribución, incluyendo uso comercial, con concesión expresa de derechos de patente y obligación de atribución.	Apache Software Foundation (Apache License 2.0)
CC BY 2.0	Permite distribuir, remezclar, adaptar y usar comercialmente, siempre que se otorgue el crédito correspondiente.	Creative Commons (CC BY 2.0)
CC BY 2.5 AR	Igual que CC BY: uso, distribución y adaptación permitidos con atribución, adaptado a jurisdicción argentina.	Creative Commons (CC BY 2.5 Argentina)
CC BY 2.5 PE	Igual que CC BY: uso, distribución y adaptación permitidos con atribución, adaptado a jurisdicción peruana.	Creative Commons (CC BY 2.5 Peru)
CC BY 3.0 BR	Igual que CC BY: permite uso comercial y modificaciones con atribución, adaptado a Brasil.	Creative Commons (CC BY 3.0 Brazil)
CC BY 4.0	Permite compartir, copiar, distribuir, adaptar, incluyendo uso comercial y obligación de atribución.	Creative Commons (CC BY 4.0)
CC BY-SA 3.0	Permite adaptar y usar comercialmente, siempre que los dataset derivados se licencien bajo los mismos términos.	Creative Commons (CC BY-SA 3.0)
CC BY-SA 4.0	Permite compartir y adaptar con atribución, exigiendo que los dataset derivados mantengan la misma licencia.	Creative Commons (CC BY-SA 4.0)
CC BY-ND 3.0	Permite redistribución, incluyendo uso comercial, pero no permite modificar el dataset.	Creative Commons (CC BY-ND 3.0)
CC BY-ND 4.0	Permite copiar y redistribuir sin modificaciones, incluso comercialmente, con atribución.	Creative Commons (CC BY-ND 4.0)
CC BY-NC 4.0	Permite adaptar y compartir el dataset, pero solo para fines no comerciales, con atribución.	Creative Commons (CC BY-NC 4.0)
CC BY-NC 3.0 IGO	Permite uso y adaptación solo para fines no comerciales, diseñado para organizaciones intergubernamentales.	Creative Commons (CC BY-NC 3.0 IGO)
CC BY-NC-SA 3.0	Permite adaptar para fines no comerciales, exigiendo atribución y compartir bajo la misma licencia.	Creative Commons (CC BY-NC-SA 3.0)
CC BY-NC-SA 4.0	Permite compartir y adaptar no comercialmente, con obligación de licenciar derivados igual.	Creative Commons (CC BY-NC-SA 4.0)
CC BY-NC-ND 4.0	Permite descargar y compartir el dataset con atribución, pero no modificarlo ni usarlo comercialmente.	Creative Commons (CC BY-NC-ND 4.0)
Non-Commercial / Research Only	Uso limitado exclusivamente a fines no comerciales o de investigación; condiciones definidas por el proveedor.	Términos específicos del proveedor/licenciante
Restricted	Uso sujeto a restricciones específicas; requiere autorización explícita para reutilización o distribución.	Términos contractuales del titular

Referencias

- Baack, S. (2024). A critical analysis of the largest source for generative AI training data: Common Crawl. Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency (FAccT '24), 2199–2208. <https://doi.org/10.1145/3630106.3659033>
- Baack, S., Biderman, S., Odrozek, K., Skowron, A., Bdeir, A., Bommarito, J., Ding, J., Gahntz, M., Keller, P., Langlais, P.-C., Lindahl, G., Majstorovic, S., Marda, N., Penedo, G., Van Segbroeck, M., Wang, J., von Werra, L., Baker, M., Belião, J., . . . Wolf, T. (2025). Towards best practices for open datasets for LLM training. arXiv. <https://doi.org/10.48550/arXiv.2501.08365>
- Centro Nacional de Inteligencia Artificial y Comisión Económica para América Latina y el Caribe. (2025). Índice Latinoamericano de Inteligencia Artificial (ILIA) 2025: Hallazgos principales, IA aplicada y talento humano. CEPAL / Cenia. <https://indicelatam.cl/>
- Constitución Política de la República de Chile [Const.]. Art. 8. 22 de septiembre de 2005 (Chile).
- Floridi, L., y Taddeo, M. (2016). What is data ethics? Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 374(2083), 20160360. <https://doi.org/10.1098/rsta.2016.0360>
- Flynn, S., Schirru, L., Palmedo, M., e Izquierdo, A. (2022). Research exceptions in comparative copyright. Program on Information Justice and Intellectual Property; American University Washington College of Law. <https://digitalcommons.wcl.american.edu/research/92/>
- Hasselbalch, G. (2016). Data ethics: The new competitive advantage. Publishare.
- Jernite, Y., Nguyen, H., Biderman, S., Rogers, A., Masoud, M., Danchev, V., Tan, S., Luccioni, A. S., Subramani, N., Johnson, I., Dupont, G., Dodge, J., Lo, K., Bosselut, A., Smith, E., Talat, Z., Radev, D., Beltagy, I., Bethard, S., . . . Hooker, S. (2022). Data governance in the age of large-scale data-driven language technology. Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency, 2206–2222. <https://doi.org/10.1145/3531146.3534637>
- Kim, J., Sohn, S., Jo, G. J., Choi, J., Bae, K., Lee, H., Park, Y., y Lee, H. (2025). Do not trust licenses you see: Dataset compliance and model integrity. arXiv. <https://doi.org/10.48550/arXiv.2503.02784>
- Longpre, S., Mahari, R., Chen, A., Obeng-Marnu, N., Sileo, D., Brannon, W., Muennighoff, N., Khazam, N., Kabbara, J., Perisetla, K., Wu, X., Shippole, E., Bollacker, K., Wu, T., Villa, L., Pentland, S., y Hooker, S. (2023). The Data Provenance Initiative: A large scale audit of dataset licensing & attribution in AI. arXiv. <https://doi.org/10.48550/arXiv.2310.16787>
- Mendels, O., Peled, C., Vaisman Levy, N., Hart, S., Rosenthal, T., Lahiani, L., y otros. (2018). Microsoft Presidio: Context aware, pluggable and customizable PII anonymization service for text and images [Software]. Microsoft. <https://microsoft.github.io/presidio>
- Ministerio de Ciencia, Tecnología, Conocimiento e Innovación. (2023). Declaración de Santiago: Consenso de la Cumbre Ministerial y de Altas Autoridades sobre la Ética de la Inteligencia Artificial en América Latina y el Caribe. Gobierno de Chile. <https://www.minciencia.gob.cl/noticias/declaracion-de-santiago-consenso-de-la-cumbre-ministerial-latinoamericana-y-del-caribe-sobre-ia/>
- Organización de las Naciones Unidas para la Educación, la Ciencia y la Cultura. (2022). Recomendación sobre la Ética de la Inteligencia Artificial. UNESCO. <https://www.unesco.org/es/articles/recomendacion-sobre-la-etica-de-la-inteligencia-artificial>
- Organización de los Estados Americanos. (2010). Ley Modelo Interamericana sobre Acceso a la Información Pública. https://www.oas.org/es/sla/ddi/docs/Ley_Modelo_Interamericana_Acceso_Informacion_Publica.pdf
- Organización para la Cooperación y el Desarrollo Económicos. (s.f.). AI principles. OECD. <https://www.oecd.org/en/topics/ai-principles.html>
- Parlamento Europeo y Consejo de la Unión Europea. (2024, 12 de julio). Reglamento (UE) 2024/1689 por el que se establecen normas armonizadas sobre inteligencia artificial (Ley de Inteligencia Artificial). Diario Oficial de la Unión Europea, L, 2024/1689. <https://eur-lex.europa.eu/legal-content/ES/TXT/?uri=CELEX:32024R1689>
- Schirru, L., Rocha de Souza, A., Valente, M. G., y de Perdigão Lana, A. (2024). Text and data mining exceptions in Latin America. Journal of Law, Technology & Policy, 2024(1), 125–150. <https://doi.org/10.2139/ssrn.4578523>

Testimonio Alianzas

“Si deseamos que los países de Latinoamérica logren el mayor desarrollo para sus pueblos, necesitamos tener un peso en el concierto mundial. Aunque a veces parezca un lujo, alcanzar soberanía tecnológica, y así no depender de otros en cuanto a conocimientos, técnicas, o productos se vuelve algo indispensable. Si además esa soberanía viene acompañada de transparencia, datos y código abierto, mucho mejor porque nos permite darle acceso al conocimiento a todos en nuestros países. Por eso es importante LatamGPT: no para lograr el mejor modelo de frontera, sino para desarrollar capacidades, cerrar brechas y tener productos tecnológicos adaptados a nuestra cultura, que nos den autonomía para desarrollos futuros y que podamos ofrecérselos a todos en Latinoamérica.”

Pontificia Universidad Javeriana

“A construção de Inteligências Artificiais generativas com compreensão de diversas realidades culturais é hoje um imperativo social e político. Este projeto oferece um caminho importante para esse tipo de construção na América Latina. Sem dúvida, o trabalho expandirá a investigação em técnicas modernas de Inteligência Artificial em toda a América Latina e promoverá colaborações significativas entre seus pesquisadores. Parabéns à toda equipe e em particular ao CENIA pela liderança.”

Universidade de São Paulo

“El proyecto LatamGPT demuestra que es posible la construcción de infraestructura digital pública bajo criterios de transparencia y rendición de cuentas. A su vez, la importancia del proyecto no radica únicamente en la creación de un modelo que propicie la soberanía lingüística, sino que una de sus mayores fortalezas reside en la visión estratégica para el desarrollo de capacidades regionales a largo plazo, utilizando una lógica de puertas abiertas para la participación comunitaria y compartiendo recursos y conocimientos entre cientos de socios. Agradecemos al CENIA la oportunidad de formar parte de este proceso de co-aprendizaje”

Laboratorio de Datos y Sociedad de Data Uruguay

“La soberanía tecnológica no está en pretender una neutralidad imposible, sino en la capacidad de auditar y comprender los sesgos culturales reflejados en los modelos. Frente a la opacidad de los modelos comerciales, el corpus de LatamGPT v1.0 apuesta por la transparencia: sólo mediante el acceso y la documentación de los datos de entrenamiento y el desarrollo de herramientas de evaluación situadas, como las que diseñamos junto a escuelas de Córdoba, podremos entender la visión del mundo que reproducen los modelos.”

Fundación Vía Libre y Universidad Nacional de Córdoba

“Como Universidad Continental consideramos que el enfoque de LatamGPT v1.0 es abierto y fortalece la producción científica colaborativa y crea mejores condiciones para que el conocimiento no dependa exclusivamente de plataformas extranjeras. Desde la academia, creemos que los datos abiertos, usados con criterios éticos y de calidad, son fundamentales para democratizar el acceso a estas tecnologías y ampliar sus beneficios.”

Universidad Continental

“Para las naciones en vía de desarrollo, como son las de Centroamérica y el Caribe, la IA agrupa un conjunto de tecnologías cognitivas que pueden utilizarse para analizar, identificar, optimizar y automatizar procesos operativos, haciéndolos más ágiles y eficientes, además de contribuir a la reducción de costos, aspectos críticos para su avance como naciones.

Los beneficios de esta automatización inteligente pueden erigirse en un motor de innovación y al asociarse con la Transformación Digital, su potencial se multiplica. De esta manera, la IA y sus tecnologías asociadas (Robótica y Mecatrónica) pueden utilizarse estratégicamente en áreas y procesos relacionados con la investigación y el desarrollo de productos y servicios, una necesidad imperiosa para los mercados y la competencia comercial y global.”

Sociedad Dominicana de Inteligencia Artificial (SODIA)

“La participación de LabEVA en LatamGPT v1.0 expresa una convicción: los repositorios abiertos son una infraestructura estratégica para democratizar el conocimiento y fortalecer la soberanía tecnológica regional. El aporte de materiales del repositorio institucional de la Udelar COLIBRI, muestra que la producción académica pública de América Latina puede integrarse responsablemente al desarrollo de IA, respetando licencias, marcos normativos y criterios éticos.”

José Fager-Pérez (FIC - UDELAR)

“Desde CDN entendemos que la soberanía tecnológica, los datos abiertos y la democratización del conocimiento son pilares fundamentales para el desarrollo sostenible de América Latina. Garantizar el acceso equitativo a la información y fortalecer capacidades locales no solo impulsa la innovación, sino que también promueve sociedades más informadas, participativas y resilientes. En este contexto, los medios de comunicación desempeñan un rol clave como puentes entre la información, la ciudadanía y el futuro digital de la región. Alba Nely Familia CDN”

Canal 37 y sus plataformas digitales

“Como representante de la Facultad de Humanidades y Ciencias Sociales de la Universidad Nacional de Jujuy, Argentina, considero que el desarrollo de modelos como LatamGPT v1.0 es un hito ineludible para alcanzar la soberanía digital y reducir la histórica dependencia tecnológica de nuestra región frente a las corporaciones globales.

La integración de datos abiertos y la adopción de licencias de código abierto resultan vitales para democratizar el conocimiento y asegurar una verdadera justicia epistémica, validando la pluralidad de nuestros dialectos y saberes históricamente invisibilizados. Solo a través de este tipo de colaboración regional y apertura tecnológica lograremos que la inteligencia artificial fomente el avance científico respondiendo, de manera ética y equitativa, a las verdaderas necesidades socioculturales de América Latina.”

Facultad de Humanidades y Ciencias Sociales, Universidad Nacional de Jujuy

“Desde Wikimedia Chile trabajamos activamente para que los datos abiertos sean entendidos como un bien común y una pieza de infraestructura pública digital para nuestra región. La soberanía tecnológica se fortalece con transparencia: conocer qué datos se usan, bajo qué licencias y con qué criterios, para habilitar auditoría y reutilización. Abrir también exige gobernanza y gestión de riesgos, para que el acceso y el beneficio sean realmente colectivos.”

Wikimedia Chile

“Celebramos con entusiasmo la iniciativa LatamGPT y felicitamos a quienes la impulsan. Desde la UNDAV, nos sumamos a este desafío reafirmando que la soberanía tecnológica es un imperativo estratégico para crear alternativas públicas que respondan a nuestra realidad y región.

Con una mirada geopolítica desde el Sur, este proyecto es una oportunidad para fortalecer nuestra identidad universitaria y profundizar nuestro compromiso social para asegurar que el conocimiento sea siempre un derecho compartido y soberano para las futuras generaciones.”

Universidad Nacional de Avellaneda

“Desde LAIA, entendemos la soberanía tecnológica como la capacidad de nuestras comunidades para decidir sobre sus datos, lenguajes y formas de conocimiento, no solo a nivel de los países, sino también de las personas y las culturas: en sus posibilidades creativas, cognitivas e identitarias. En un contexto de propuestas tecnológicas cerradas y concentradas, LatamGPT viene a disputar las reglas del juego de la IA, apostando a una inteligencia artificial abierta, situada y comprensible, que no nos reduzca a usuarios pasivos sino a cocreadores de las herramientas que habitamos.”

LAIA (Laboratorio Abierto de Inteligencia Artificial)

