CASO Latam-GPT

IA y acceso a datos de entrenamiento en América Latina

¿Qué sabemos sobre los datos de entrenamiento de los LLM más populares?

Reported parameters	Gemini		Llama		ChatGPT			Mistral	Qwen	Deepseek
	1.0	*1.5*	*2*	*3*	*3*	*4*	*4o*	*large*	*2.5*	*R1*
Model in Open access	No	No	Yes	Yes	No	No	No	Yes	Yes	Yes
Organization	Google	Google	Meta	Meta	OpenAI	OpenAI	OpenAI	Mistral	Alibaba	Deepseek
Tokens Pre-training Dataset weight	N/A	N/A	2 T	15 T	300 B	N/A	N/A	N/A	18 T	14.8 T
Dataset availability	No	No	No	No	No	No	No	No	No	No
Reported Data sources	No	No	No	No	Yes	No	No	No	No	No

https://www.latamgpt.org/

EQUIPO DE RECOLECCIÓN DE DATOS PROYECTO Latam-GPT

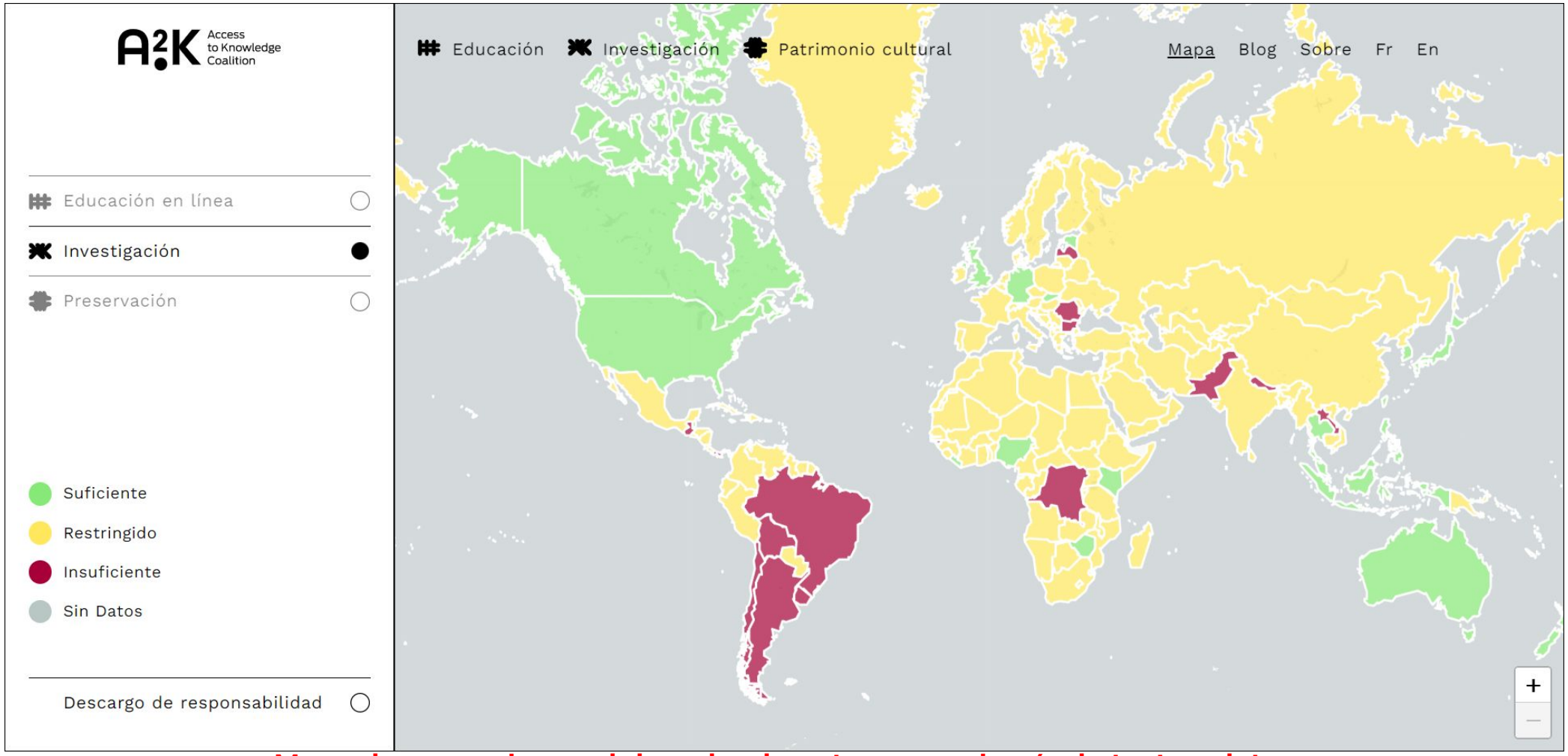
Modelo de Gobernanza de Datos:

- **Procurar la mayor transparencia y disponibilidad posible:**
  - del corpus de datos de entrenamiento
  - del modelo
- **Recolección no predatoria basada en:**
  - aportes comunitarios (convenios y socios)
  - datos con licencias libres
  - web scraping ético

“Se dice que la luz solar es el mejor desinfectante” (Louis Brandeis, 1913)

Latam-GPT

MÚLTIPLES PROBLEMAS RELACIONADOS CON DERECHOS DE AUTOR



Mapa de excepciones al derecho de autor para minería de texto y datos

https://www.a2k-coalition.org/map/

MÚLTIPLES PROBLEMAS RELACIONADOS CON DERECHOS DE AUTOR

PROBLEMAS 3 EJEMPLOS:

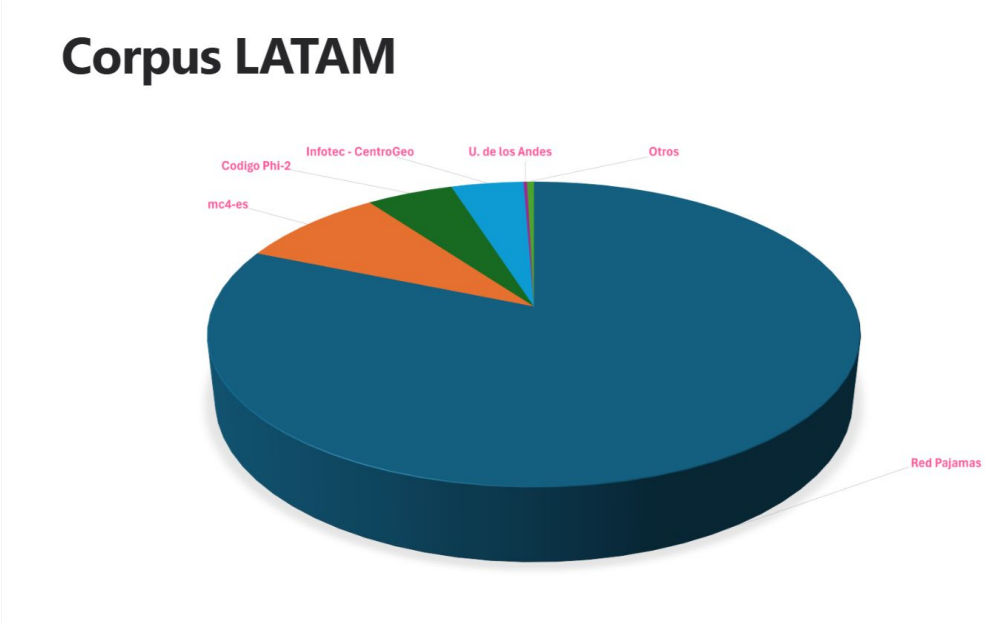
1. EL REUSO DE OTROS DATA-SETS LIBRES
2. EL USO DE INFORMACIÓN PÚBLICA (las taquigráficas de sesiones y comisiones parlamentarias)
3. EL USO DE SUBTÍTULOS DE YOUTUBE (de sesiones parlamentarias)

1 PROBLEMAS CON EL REUSO DE OTROS DATA-SETS LIBRES

En noviembre el Proyecto lanzó dataset masivo de texto de alta calidad en español 🤗

El corpus contiene 700 GB de texto, compuesto por 129 millones de documentos filtrados por calidad académica de otro Dataset: RedPajamas-v2 publicado con licencia libre (Apache).

Descubrimos que este dataset contiene fuentes minadas bajo el principio de fair use de USA



corpus latam-gpt/red_pajama_es_hq

1

PROBLEMAS CON EL REUSO DE OTROS DATA-SETS LIBRES

SE DECIDIÓ PUBLICAR EL NUEVO DATASET CON EL SIGUIENTE AVISO:

“This dataset is licensed under the Apache 2.0 license. The text documents of the source database (RedPajama-Data-v2) also provided under an Apache 2.0 license by the Together Computer team under the jurisdiction of the United States of America.

Understanding that there may be differences between the jurisdiction of the USA and Latin American countries** and in the search for the greatest possible transparency, **we provide the following contact to ask any questions, comments or complaints: eugenio.herrera@cenia.cl”

2

PROBLEMAS CON EL USO DE INFORMACIÓN PÚBLICA

¿es legal incluir las taquigráficas de las sesiones parlamentarias en un dataset de entrenamiento?

- **Los discursos parlamentarios son obras alcanzadas por las leyes de derechos de autor**

- **Las leyes de acceso a la información pública no solucionan los problemas de derechos de autor**

Ejemplo: *Artículo 25 de la Ley de Derechos de Autor de Uruguay*

Los discursos políticos, científicos o literarios y, en general, las conferencias sobre temas intelectuales, no podrán ser publicados si el autor no lo hubiera autorizado. **Los discursos parlamentarios podrán ser publicados libremente salvo cuando se haga la publicación con fines de lucro, caso en el cual será necesaria la autorización del autor.**

Exceptúase la información periodística.

PROBLEMAS CON EL USO DE INFORMACIÓN PÚBLICA

2 ¿es legal incluir las taquigráficas en un dataset de entrenamiento?

País	¿Los ToS de los sitios web parlamentarios permiten descargar y procesar los contenidos del sitio?
Argentina	Autorización expresa
Bolivia	No se publican taquigráficas
Chile	Autorización expresa
Colombia	Tos prohíben usar, "salvo que esté permitido por la ley"
Costa Rica	Autorización expresa
Cuba	No se publican taquigráficas
Ecuador	No existen ToS
El Salvador	No existen ToS
Guatemala	No existen ToS
Honduras	No se publican taquigráficas
México	ToS sin referencias al Derecho de Autor
Nicaragua	No existen ToS
Panamá	No existen ToS
Paraguay	Autorización expresa
Perú	No existen ToS
Rep. Dominicana	No, prohibición expresa
Uruguay	No, prohibición expresa
Venezuela	No se publican taquigráficas

PASO 1:

Analizar los Términos y Condiciones de los sitios web parlamentarios

2

PROBLEMAS CON EL USO DE INFORMACIÓN PÚBLICA

¿es legal incluir las taquigráficas en un dataset de entrenamiento?

País	**Excepciones para discursos o documentos públicos (en sentido amplio) que cubran el web scraping de las taquigráficas y actas parlamentarias**	**Excepciones para actividades de investigación que cubran el web scraping de forma genérica**
Argentina	Sí	No
Bolivia	No	No
Chile	Incierto	No
Colombia	Incierto	No
Costa Rica	Sí	No
Cuba	Sí	Sí
Ecuador	Sí	Incierto
El Salvador	No	No
Guatemala	Incierto	No
Honduras	Incierto	No
México	Incierto	No
Nicaragua	Sí	No
Panamá	Sí	No
Paraguay	No	No
Perú	No	No
Rep. Dominicana	No	No
Uruguay	Sí	No
Venezuela	No	No

PASO 2:

Analizar las leyes buscando excepciones al derecho de autor

3

PROBLEMAS CON EL USO DE SUBTÍTULOS DE YOUTUBE

NO PODREMOS EXTRAER LOS SUBTÍTULOS DE LAS CUENTAS DE YOUTUBE DE CADA PARLAMENTO, **AÚN CON LA AUTORIZACIÓN DEL PARLAMENTO**

- Los **ToS de youtube** prohíben la extracción de datos por fuera de la API
- La **API de Youtube** solo permite la extracción de subtítulos de cuentas propias.
- El **texto de robot.txt** de youtube prohíbe expresamente scrapear subtítulos
- **ZONA GRIS:**

  1) extraer los datos sin utilizar una cuenta y sin haber aceptado los ToS

  2) utilizar servicios de extracción de subtítulos (Crawlbase o Apify)

El régimen de derechos de autor en Chile (y del resto de los países socios del proyecto) plantea un panorama de alta inseguridad jurídica

- En Brasil, Chile, Ecuador y Uruguay **existen Proyectos Ley que proponen excepciones al Derecho de Autor** para la investigación basada en Ciencia de Datos.
- Estos proyectos están siendo **ampliamente resistidos** por editoriales, gestoras colectivas y otros titulares de derechos.
- La solución que se impulsa a nivel internacional es el **licenciamiento y la creación de nuevos derechos de remuneración** ¿serán suficientes para abarcar los datos necesarios para los grandes LLM?, ¿qué efectos tendrán sobre el desarrollo de la IA en el Sur Global?

¡GRACIAS!

@PatoDiazGNU

ACCESO JUSTO al CONOCIMIENTO

Alianza de la Sociedad Civil Latinoamericana