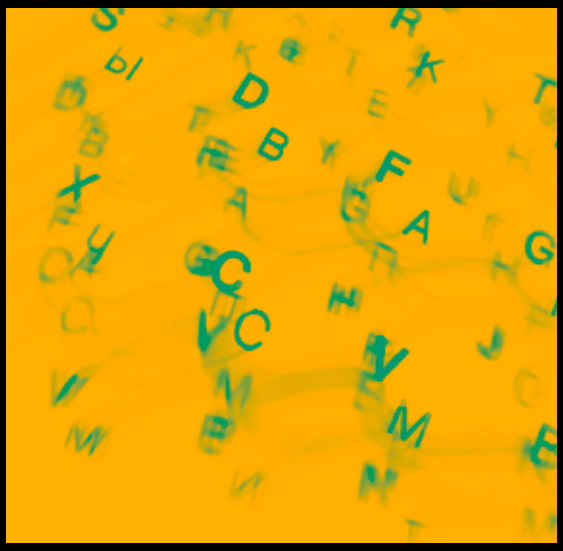
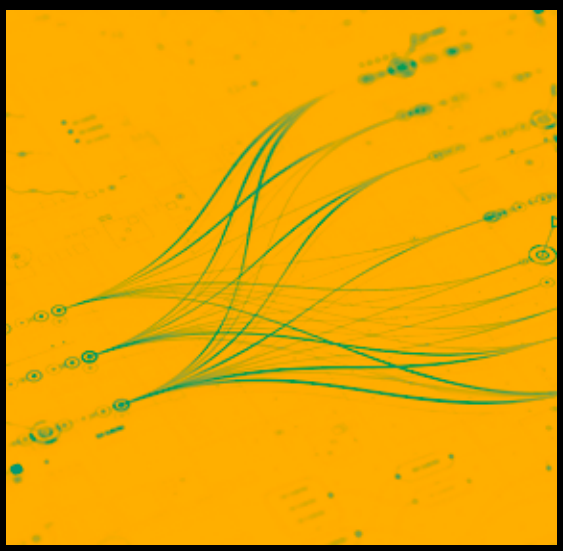
EL DESEMPEÑO DE LA INTELIGENCIA ARTIFICIAL EN EL USO DE LENGUAS INDÍGENAS AMERICANAS





Microsoft
BID | LAB

EL DESEMPEÑO DE LA INTELIGENCIA ARTIFICIAL EN EL USO DE LENGUAS INDÍGENAS AMERICANAS

EVALUACIÓN DE LA BRECHA EN IA EN LL.II.AA.

Agradecimientos: Expresamos nuestro sincero agradecimiento a los intérpretes y academias que participaron en el proyecto Lenguas Nativas y su interacción con la Inteligencia Artificial. Su valiosa colaboración fue fundamental para la evaluación del desempeño de distintas tecnologías de inteligencia artificial en lenguas originarias, permitiendo visibilizar las brechas culturales y lingüísticas presentes en estas herramientas. En particular, agradecemos a Jacob Cruz, intérprete de la lengua Náhuatl; a Armando Hueyotenco, intérprete de la lengua Náhuatl del Instituto de la Transparencia, Acceso a la Información Pública Gubernamental y Protección de Datos Personales del Estado de Hidalgo (ITAIH); a Ana Paola Quispe Quispe, intérprete de la lengua Aymara de la Academia Nacional de la Lengua Aymara (ANLA); a Elmer Machicao, intérprete de la lengua Aymara; a Tomas Rojas, intérprete de la lengua Mapuche; a Yony Mediano, intérprete de la lengua Quechua; y a Mauro Lugo, intérprete de la lengua Guaraní. Gracias a su experiencia y compromiso, se lograron evaluaciones significativas en las lenguas Náhuatl, Aymara, Mapuche, Quechua y Guaraní, fortaleciendo el diálogo inclusivo entre tecnología y diversidad lingüística en la región.

Autores: Escritura, investigación y análisis: Miguel Lucas (LLYC), Alejandro Burgueño (LLYC), Miguel Carazas (LLYC). Conceptualización y guía estratégica: César Buenadicha (BID Lab), Smeldy Ramírez (BID Lab), César Rosales (BID Lab). El equipo agradece el generoso patrocinio y apoyo de Microsoft, cuyo respaldo hizo posible la elaboración de este reporte.

Diseño y diagramación: Alejandro Scaff

https://bidlab.org

El compromiso de fAIr LAC y el Grupo BID con la IA responsable

Este producto de conocimiento fue desarrollado en el marco de fAIr LAC, la iniciativa del Grupo BID que impulsa la adopción responsable de la inteligencia artificial (IA) en América Latina y el Caribe. El documento constituye un insumo clave para el desarrollo de modelos fundacionales en lenguas indígenas americanas, complementando el trabajo de BID Lab, el laboratorio de innovación del Grupo BID, como lo demuestra su aplicación en el proyecto GuaranIA, enfocado en el desarrollo de modelos fundacionales en lengua guaraní.

La elaboración de esta publicación fue posible gracias a la valiosa contribución de Microsoft, con investigación de Llorente y Cuenca, y en estrecha colaboración con BID Lab a través de la iniciativa fAIr LAC.

fAIr LAC está plenamente alineada con el Compromiso del Grupo BID con la IA Responsable, plasmado en el Marco de IA del Grupo BID, lanzado en enero de 2025. Este marco establece una hoja de ruta estructurada en tres pilares fundamentales:

- **Instituciones y gobernanza**, para promover estándares éticos y mecanismos efectivos de supervisión;
- **Datos e infraestructura**, orientados a garantizar un acceso equitativo y abierto a activos habilitadores de IA; y
- **Personas y talento**, enfocados en el fortalecimiento de capacidades locales.

Desde su lanzamiento en 2019, fAIr LAC ha servido como plataforma para traducir principios éticos y regulatorios emergentes en herramientas prácticas. A través de BID Lab, la iniciativa ha evaluado más de 30 casos de uso utilizando metodologías como fAIr LAC 3S (Solución, Sistema, Sociedad) y fAIr Venture, diseñadas para gestionar riesgos éticos y operacionales mientras se maximiza el impacto social.

Además de su labor de asesoría técnica, fAIr LAC aborda desafíos emergentes como la sustitución laboral derivada del uso de IA, promoviendo procesos de reconversión y capacitación de trabajadores. Asimismo, fomenta la equidad lingüística mediante la investigación y desarrollo de modelos fundacionales en lenguas indígenas, con el objetivo de corregir sesgos sistémicos y mejorar el desempeño de los grandes modelos de lenguaje en idiomas subrepresentados.

BID Lab

BID Lab es el brazo de innovación y venture capital del Grupo Banco Interamericano de Desarrollo. Encontramos nuevas formas de impulsar la inclusión social, la acción ambiental y la productividad en América Latina y el Caribe. BID Lab apalanca financiamiento, conocimiento y conexiones para apoyar el emprendimiento en etapa temprana, desarrollar nuevas tecnologías, activar mercados innovadores y dinamizar sectores existentes. www.bidlab.org

AI for Good de Microsoft

El Laboratorio de AI for Good de Microsoft es una iniciativa colaborativa centrada en aprovechar la inteligencia artificial para resolver algunos de los desafíos más urgentes del mundo. Trabajando con socios globales y utilizando herramientas de IA de vanguardia, el laboratorio impulsa innovaciones en sostenibilidad ambiental, agricultura, salud y más. El Laboratorio de AI for Good se compromete a aplicar soluciones de IA para mejorar vidas, impulsar el progreso hacia los Objetivos de Desarrollo Sostenible de la ONU y generar un impacto duradero a través de la tecnología.

LLYC

LLYC es una firma global de consultoría de comunicación, marketing digital y asuntos públicos. Fue fundada en Madrid en 1995 como Llorente y Cuenca y actualmente tiene 20 oficinas en varios países, incluyendo España, Argentina, Brasil y Colombia. LLYC ayuda a sus clientes a afrontar sus retos estratégicos con soluciones basadas en la creatividad, la tecnología y la experiencia.

fAIr LAC

fAIr LAC es una alianza entre los sectores público y privado, la sociedad civil y la academia, para incidir tanto en la política pública como en el ecosistema emprendedor en la promoción del uso responsable de la IA.

1. ÍNDICE

2. RESUMEN EJECUTIVO

- **Las inteligencias artificiales generativas muestran un desempeño muy deficiente en lenguas indígenas.** Tan sólo en el 54% de los casos, ante preguntas formuladas en lenguas originarias, la respuesta es aparentemente correcta. Y cuando así lo es, en realidad esta respuesta es 4 veces más corta, y obtiene un 2.4/10 en cuanto a corrección en la expresión y un 2.3/10 en comprensión de la pregunta.
- **La baja presencia de textos escritos y otros recursos en lenguas indígenas en Internet (a diferencia de las lenguas mayoritarias) dificulta significativamente la comprensión y expresión de la IA en estas lenguas.** Este efecto es aún más pronunciado en el caso de las lenguas indígenas con menos hablantes.
- **Para incrementar la integración de lenguas indígenas en el ecosistema digital, se proponen 21 estrategias centradas tanto en incrementar los datos disponibles en estas lenguas como en el desarrollo de tecnologías habilitadoras.** Fomentar la conversación digital en lenguas indígenas, dar visibilidad a sus influencers, proteger plataformas y archivos digitales de tradiciones, y desarrollar tecnologías de traducción y voz son algunas de las más destacadas. De esa manera, tales estrategias ayudarían a entrenar los modelos de IA de una manera que mejoren su performance en estas lenguas.
- **Se propone crear un consorcio internacional para implementar las estrategias propuestas.** Compuesto de organizaciones nacionales e internacionales, instituciones dedicadas a la protección cultural y compañías de tecnología interesadas en acelerar el uso de la IA para atravesar las brechas lingüísticas.
- **Las referencias culturales indígenas son minoritarias en la IA.** El sesgo cultural apreciado en las respuestas frente a preguntas realizadas en lenguas indígenas está desviado hacia la cultura hegemónica occidental. Incluso en el caso del Quechua (la lengua que mejor comportamiento muestra en este apartado) su inclusión se sitúa por debajo de 2.3/10.
- **La IA supone una gran oportunidad para reducir el aislamiento y dar visibilidad a pueblos y culturas indígenas.** No solo es un nuevo altavoz de gran alcance para dar a conocer y perpetuar la tradición y cultura indígena, sino que puede ayudar a reducir la brecha entre sus comunidades más aisladas por cuestiones de analfabetismo y monolingüismo.
- **Los programas gubernamentales de apoyo indígena y las iniciativas de las Big Tech son dos puntales fundamentales para mejorar el rendimiento de la IA en lengua indígena.** Programas de apoyo de ONGs y las iniciativas de grandes marcas de consumo completan el panorama de aliados clave para conseguir una mejor IA.
- **Existe una altísima correlación entre el volumen de contenidos digitales disponibles en una lengua y el rendimiento que la IA muestra en esa lengua.** Rendimiento IA y representación en la Wikipedia para una lengua tienen una correlación del 91%. Cuanto mayor es el volumen de contenidos digitales en abierto en una lengua, mayor es la calidad con la que la IA habla y entiende esa lengua.

3. DEFINICIONES Y ABREVIATURAS

IA / AI: Inteligencia artificial (Artificial Intelligence). Generalmente la IA hace referencia a un campo de estudio extenso que abarca múltiples disciplinas (Machine Learning, Procesamiento de Lenguaje Natural, Razonamiento y Planificación Automática, Deep Learning, etc). En el contexto del presente informe, y simplificando, IA hace referencia a los Grandes Modelos del Lenguaje (LLMs, por Large Language Models), ya que éstas son las estructuras tecnológicas que el público general entiende en la actualidad como Inteligencias Artificiales, máxime desde la llegada de ChatGPT el 30 de noviembre de 2022.

LLM: Grandes Modelos de Lenguaje o Large Language Models. Modelos de inteligencia artificial centrados específicamente en la comprensión y generación del lenguaje verbal humano. Algunos LLMs populares del estado del arte son PHI3 de Microsoft, GPT4o de OpenAI, Sonnet 3.5 de Anthropic, Gemini Pro 1.5 de Google y Llama 3.0 de Meta.

LL.II.: Lenguas Indígenas. En el presente informe hacen referencia al Quechua, Guaraní, Aimara, Náhuatl, Quiché, Mapuche y Tupi-Guaraní moderno.

MQM: El Modelo de Calidad Multidimensional (MQM) es una herramienta analítica empleada por lingüistas para evaluar la calidad de traducciones y textos generados. Proporciona un marco estructurado que clasifica errores en categorías específicas, permitiendo un análisis detallado y cuantitativo de la precisión, fluidez, y consistencia en el lenguaje.

MMLU: La Evaluación Multitarea de Comprensión de Lenguaje (MMLU) es una métrica de benchmarking diseñada para modelos de lenguaje de última generación. Mide el desempeño en tareas de comprensión a través de múltiples dominios y niveles de dificultad, evaluando la capacidad del modelo para manejar preguntas complejas en diversas áreas del conocimiento.

Prompt: Es la instrucción o pregunta que se le introduce a la IA para ver cómo responde. En el presente estudio los prompts han sido traducidos previamente a las LL.II. correspondientes, para analizar cómo responde el modelo de IA ante una interacción directa de un usuario nativo, si identifica el idioma, y si resuelve eficientemente el resto de parámetros de medición.

Query: Una consulta (Query) es una instrucción o pregunta formulada por un usuario o sistema para recuperar información específica de una base de datos o motor de búsqueda. Es una herramienta esencial en la interacción entre usuarios y sistemas, especialmente en entornos de procesamiento de lenguaje natural.

Scraper: Un scraper es un programa automatizado diseñado para extraer información estructurada o no estructurada de páginas web. Utilizado en tareas de recopilación de datos, este proceso permite analizar grandes volúmenes de información y almacenarla para aplicaciones específicas, como análisis de tendencias o el entrenamiento de LLMs.

4. APROXIMACIÓN

En América Latina, las lenguas indígenas son un componente esencial del patrimonio cultural de la región, pero muchas enfrentan desafíos significativos debido a la globalización, la exclusión digital y la falta de recursos adaptados. Estas lenguas, habladas principalmente por comunidades rurales, reflejan una riqueza histórica y cultural que está en riesgo de desaparecer. Ante esta situación, diversas iniciativas han sido implementadas por gobiernos, organizaciones internacionales, ONGs y empresas tecnológicas para promover su preservación y revitalización. Estas iniciativas no solo buscan garantizar los derechos lingüísticos, sino también fortalecer las identidades culturales y mejorar el acceso a oportunidades educativas y económicas en las comunidades que las hablan.

Sin embargo, en el ámbito de la inteligencia artificial (IA), las lenguas indígenas presentan un muy bajo protagonismo. El rendimiento de los modelos de IA adaptados a estas lenguas perpetúa la exclusión digital y limita el acceso de estas comunidades a herramientas tecnológicas avanzadas. A diferencia de las lenguas mayoritarias, la baja presencia en Internet de textos escritos en lenguas originarias y de otros recursos merma considerablemente la comprensión y expresión de la IA en estas lenguas autóctonas. Y este efecto es tanto más acusado, cuanto menor es la población que habla una determinada lengua.

El objetivo del presente estudio es analizar la capacidad de las IAs de última generación para adaptarse y responder eficazmente en lenguas indígenas, evaluando dimensiones lingüísticas, operativas y comportamentales. Para ello, se han seleccionado siete lenguas indígenas representativas de América Latina, distribuidas a lo largo de toda su geografía, con diferentes tamaños de poblaciones de hablantes y rasgos culturales. El análisis se centra en cómo cinco modelos de lenguaje de última generación, entre ellos GPT-4o y Phi 3, abordan casos de uso prácticos como la redacción, la mensajería y la interacción con contenidos.

La metodología empleada combina estándares de evaluación lingüística y funcional, adaptando métricas como MMLU y MQM en tres categorías clave: idiomática, ejecutiva y comportamental. Analiza qué tal hablan, qué tal comprenden y en qué marco cultural se expresa la IA cuando interactuamos en lenguas indígenas. Se aborda la valoración de aspectos como corrección gramatical, coherencia cultural, precisión en la ejecución de tareas y adaptación a registros específicos. Los casos de uso se abordan a través de dos niveles de abstracción, simulando tanto usuarios habituales como expertos, lo que permite evaluar la efectividad de los modelos en contextos variados.

Este enfoque busca no solo cuantificar las brechas actuales, sino también establecer un marco para el desarrollo de estrategias inclusivas que aumenten el nivel de integración de las lenguas indígenas al ecosistema digital. De esta manera, se proponen ajustes y acciones útiles para el entrenamiento de los modelos de cara a mejorar su desempeño en estas lenguas. En definitiva, buscando el objetivo de que los resultados de este estudio sirvan como referencia para investigadores, desarrolladores y entidades gubernamentales en la formulación de políticas y herramientas que fomenten la equidad lingüística en la era de la Inteligencia artificial.

5. ALCANCE DEL ESTUDIO Y METODOLOGÍA

Este estudio analiza la capacidad y desempeño de modelos de inteligencia artificial (IA) de última generación al interactuar en lenguas indígenas americanas (LL.II.). Su objetivo es identificar y cuantificar las brechas lingüísticas, se analizaron: ***Quechua***, ***Guaraní***, ***Aimara***, ***Náhuatl***, ***Maya Quiché***, ***Mapuche*** y ***Tupí-Guaraní***. Igualmente se realizó la misma batería de pruebas sobre dos lenguas occidentales de la misma escala de hablantes, como son el *Catalán* y el *Euskera*, con el objetivo de usarlos como puntos de contraste.

- Y los LLMs de IA evaluados fueron: ***GPT-4o*** *de OpenAI,* ***Claude 3.5 Sonnet*** *de Anthropic,* ***PHI-3*** *de Microsoft,* ***Gemini 1.5 Pro*** *de Google* y ***Llama 3*** *de Meta*.

1. **Diseño de Casos de Uso:**
   Los casos de uso se centraron en redacción, mensajería e interacción con contenidos, evaluando tanto la ejecución de tareas simples como complejas. Estos escenarios se corresponden con las interacciones más frecuentemente realizadas en la actualidad con los modelos de IA.

2. **Criterios de Evaluación:**
   Se combinaron métricas establecidas como MMLU (Evaluación Multitarea de Comprensión de Lenguaje) y MQM (Modelo de Calidad Multidimensional) para evaluar tres dimensiones clave:

   - **Valoración idiomática (evaluación de la expresión):** Fluidez, corrección y coherencia.
   - **Valoración ejecutiva (evaluación de la comprensión):** Precisión, completitud y comprensión.
   - **Valoración comportamental (evaluación de la cultura reflejada):** Adaptación cultural y adecuación contextual.

3. **Análisis de Datos Digitales y Recursos Lingüísticos:**
   Se evaluó la cantidad de textos en lenguas indígenas disponibles para el entrenamiento, así como la presencia de corpus lingüísticos, herramientas digitales (traductores automáticos, detectores de idioma, speech-to-text, etc.).

4. **Comparativa Contextual:**
   Se comparó el rendimiento de la IA en lenguas indígenas frente a idiomas con mayor presencia digital (catalán, euskera) para identificar correlaciones entre la calidad de respuesta de las IA y la cantidad de recursos digitales disponibles.

5. **Dimensiones de Inclusión:**
   Se consideraron los impactos de las brechas en el rendimiento de la IA en términos económicos, sociales y culturales, evaluando cómo estas limitaciones perpetúan la exclusión digital de las comunidades indígenas.

Esta metodología busca no solo cuantificar la brecha tecnológica, sino también establecer un marco para la implementación de estrategias inclusivas que fortalezcan la presencia y funcionalidad de las lenguas indígenas en la inteligencia artificial.

6. ANÁLISIS DE DATOS Y COBERTURA LINGÜÍSTICA

Por cada 20 páginas que se navegan en la web 1 es en español, pero hacen falta 125,000 páginas para encontrar 1 en guaraní, una de las LL.II. con más presencia online.

Pero la distancia con otras lenguas minoritarias también es abrumadora: Si bien el quechua tiene 10 veces más hablantes que el euskera, éste último tiene 20 veces más artículos en la Wikipedia y 60 veces más contenido web que el quechua.

57%

de LL.II. tiene diccionarios online y sólo el 29% tiene proyectos en curso sobre la elaboración de bases de datos léxicas (quechua y guaraní).

Las herramientas digitales también son escasas. Aunque el quechua puede ser detectado por la mitad de detectores automáticos que el español, el resto de LL.II. son detectados por una tercera

Recursos fundamentales para entrenar una Inteligencia artificial en una lengua

Numerosos son los recursos digitales necesarios para entrenar a un gran modelo del lenguaje del estado del arte (también llamados **LLM -Large Language Models-**, no obstante, en este informe apelaremos a dicha tecnología como "IA" para simplificar el concepto). Estos modelos precisan de un gran volumen de datos escritos en lenguaje natural humano en cada idioma en cuestión y, para que puedan procesarlos, es preciso facilitar acceso a esta información de manera digital. Entre los ***recursos de contenidos*** más relevantes podríamos destacar:

- **Textos de conversación social**, ya que con datos de esta índole el modelo se capacita en cuestiones asociadas al registro informal y popular, así como información contemporánea y uso natural del lenguaje.
- **Textos de medios digitales y canales de información horizontal**, como la Wikipedia u otros canales, ya que con datos de esta índole se transfiere conocimiento del mundo y la realidad histórica en registros formales y de divulgación popular.
- **Textos de la web** en general, para facilitar la generalización y la abstracción a la par que se amplían contenidos. Por ejemplo, los contenidos de Common Crawl[1].
- **Otros** relativos a información gubernamental o científica, por ejemplo, que impulsen la verticalidad del modelo en otros escenarios del lenguaje como pueden ser el legal, el técnico, etc.

Entre los ***recursos de herramientas*** necesarios se destacan los siguientes:

- **Traductores automáticos**, ya que los mismos facilitan el test y la mejora de los datos de entrenamiento, además son útiles para generar datos sintéticos[2] (data augmentation) a partir de lenguas que se encuentran en un nivel más avanzado en el estado del arte de la IA.
- **Detectores automáticos de idioma**, ya que permiten filtrar los textos y contenidos que están escritos en una determinada lengua y seleccionarlos específicamente para el entrenamiento de una IA.
- **Modelos de conversión speech to text** y viceversa (también conocidos como voice to text) ya que permiten acceso a datos que provienen de fuentes en formato de audio, así como facilitar acceso a usuarios que tienen dificultades para leer o escribir una lengua específica, o bien personas no alfabetizadas.
- **Diccionarios digitales**, bases de datos léxicas, etc para facilitar el acceso universal al estudio y la comprensión de una lengua.

Datos digitales disponibles en lenguas indígenas

Aproximadamente 1 de cada 20 páginas web del mundo está escrita en español, pero para la lengua indígena americana con más información web, el Guaraní[3], harían falta 125,000 páginas web para encontrar 1 página en dicho idioma.

La brecha de las lenguas indígenas en materia de recursos digitales es tan amplio que, comparando por ejemplo con el Euskera, que posee 10 veces menos hablantes que el Quechua, el primero presenta 19 veces más entradas en la Wikipedia y 19 veces más usuarios participando en las

1 Common Crawl es una organización sin fines de lucro que recopila y proporciona de forma gratuita archivos de datos masivos extraídos de la web. Su repositorio incluye información de miles de millones de páginas web, utilizada comúnmente para entrenar modelos de inteligencia artificial y realizar investigaciones en procesamiento de lenguaje natural. https://commoncrawl.org/

2 Conjuntos de datos generados artificialmente mediante algoritmos, diseñados para replicar las características estadísticas de datos reales sin comprometer información sensible o privada. Son utilizados en pruebas, desarrollo de modelos y análisis, ofreciendo una alternativa ética y segura para manejar información.

3 Según Common Crawl https://commoncrawl.github.io/cc-crawl-statistics/plots/languages.html

publicaciones. Hay que tener presente que el Quechua posee 5 veces más contenido enciclopédico que el Guaraní, el Aimara o el Náhuatl, lo que pone de manifiesto la escasa digitalización de las LL.II. en general.

En este marco comparativo, nuestros estudios demuestran que el Euskera tiene además 60 veces más páginas webs que el Quechua, y todo esto pese a que en Quechua se publica contenido en redes sociales 2 veces y medio más que en Euskera. Entonces, ¿por qué si muchas lenguas indígenas poseen más hablantes nativos que otras lenguas (más avanzadas en materia de IA), y su presencia en la conversación social no es desdeñable, los contenidos digitales sí son escasos?

Tabla 1. Presencia de publicaciones en lenguas indígenas en Wikipedia y otras redes

	Quechua	Guaraní	Aimara	Náhuatl	Quiché	Mapuche	Tupi Guaraní	Catalán	Euskera	Español
Miles de hablantes	7,000	6,500	1,906	1,793	1,055	150	20	10,020	750	600,600
Miles de entradas en la Wikipedia	24	5	5	4	-	-	-	764	448	1,992
N° editores en la Wikipedia	40	30	15	14	-	-	-	1,503	751	14,171
Presencia en la web (Common Crawl)	0.0006%	0.0008%	0.0001%	-	-	-	-	0.20%	0.04%	4.62%
Miles de posts en X al año	7,231	985	2,387	1,650	375	20	203	92,850	2,912	5,531,599

Más allá de aspectos que abordaremos próximamente como la financiación o el soporte, una de las claves está en comprender el ratio: de promedio, cada vascoparlante al año emite 4 posts en X, mientras que cada Quechuaparlante sólo 1. Esto es un indicativo de que la población vasca está más familiarizada con las tecnologías y los canales de comunicación contemporáneos que la población Quechua.

Sin embargo, ¿cuál es el ratio estándar actualmente? El ratio varía mucho según los idiomas y la naturaleza de las poblaciones de las que dependen, pero podríamos decir que un valor estándar es 9 posts al año por hablante, algo que sucede en español y también, de manera muy similar, en catalán.

Herramientas lingüísticas digitales en lenguas indígenas

En lo que a herramientas digitales respecta, es necesario precisar que los diccionarios web están, en mayor o menor grado, disponibles solo para 4 de los 7 idiomas de estudio, siendo el Tupi Guaraní, Maya Quiché y Mapuche las lenguas indígenas que no cuentan acceso a herramientas de traducción online.

Otra cuestión muy diferente son las bases de datos léxicas[4] o lexicón, de entre las cuales uno de los ejemplos más conocidos es WordNet. El principal referente es el inglés[5], y son esenciales para entrenar y etiquetar estructuras léxicas y sintácticas. Actualmente existe un WordNet en español[6] y tanto el Euskera como el Catalán están acogidos por el framework EuroWordNet[7]. Aunque las LL.II. de estudio no presentan WordNets abiertos, sí es cierto que hay investigaciones y publicaciones que apuntan a avances en el Quechua[8] y el Guaraní[9].

4 Colección estructurada de palabras y sus características lingüísticas, como significado, pronunciación, relaciones semánticas y morfológicas. Se utiliza en procesamiento del lenguaje natural y estudios lingüísticos para analizar y modelar el comportamiento del lenguaje.

5 Wordnet Inglés, referente de Princeton: https://wordnet.princeton.edu

6 Spanish Wordnet 3.0: http://timm.ujaen.es/recursos/spanish-wordnet-3-0/

7 EuroWordnet, proyecto de agrupamiento de lenguas Europeas: https://archive.illc.uva.nl//EuroWordNet/

8 Wordnet-QU: https://aclanthology.org/2022.coling-1.390.pdf

9 Guaraní Wordnet: https://aclanthology.org/2023.gwc-1.24.pdf

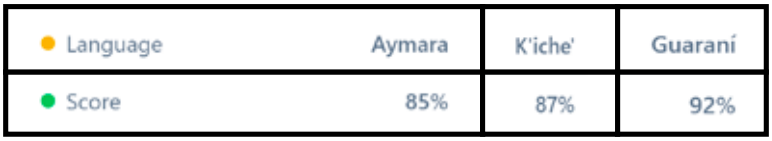
Tabla 2. Distribución de herramientas lingüísiticas y motores de búsqueda por lengua indígena

	Quechua	Guaraní	Aimara	Náhuatl	Maya Quiché	Mapuche	Tupi Guaraní	Catalán	Euskera	Español
N° de traductores automáticos	2	2	2	1	0	0	0	3	3	5
N° de detectores automáticos de idioma	5	4	4	4	3	2	0	7	8	10
Text2speech + Speech2text	2	1	1	0	0	0	0	3	3	3
Permite búsqueda en Google	Sí	No	No	No	No	No	No	Sí	Sí	No
Permite búsqueda en Bing	No	No	No	No	No	No	No	Sí	Sí	No

La detección y traducción automática es otro punto clave no sólo para medir la accesibilidad de su población, sino para calibrar la posibilidad de aumentar los volúmenes de entrenamiento y calibrar la eficiencia a la hora de inferir el idioma de los usuarios que emplean IAs. El Tupí-Guaraní moderno, la variante Amazónica con mayor presencia en Brasil del Guaraní, no dispone de detectores de idioma efectivos (siempre es confundido o anulado por el Guaraní de Paraguay) y, junto con el Quechua y el Mapuche, no dispone de traductores web al uso.

En líneas generales, idiomas como el Quechua son detectables por aproximadamente la mitad de herramientas populares (y multi-idioma) capaces de detectar el Español. Por otro lado, Guaraní, Aimara y Náhuatl se sitúan en el 40%, lo cual resulta escaso si comparamos con el 75% del Catalán o el Euskera. Algunos referentes destacados para detectar LL.II. son OpenL[10] y Originality[11], pero la poca competencia entre detectores de idioma tiende a resultar en una baja precisión de detección.

Figura 1. Muestras de confianzas de detección de algunos ejemplos en OpenL



Finalmente, es imprescindible medir la brecha existente en herramientas transformadoras de formatos, como son las Speech2text (convertidores de voz a texto). Aunque el contenido digital es mayormente escrito, recientes estudios indican que aproximadamente el 80% de la población escucha algún contenido de audio en sus dispositivos[12] (podcasts, vídeos, audiolibros, etc). A esto se le suma que muchos de los hablantes de LL.II. no saben leer ni escribir (según Unicef una quinta parte de la población indígena es analfabeta y entre aquéllos que hablan lengua indígena 1 de cada 4 no sabe leer ni escribir)[13], por lo que son formatos de mayor potencial de crecimiento en la web.

La necesidad de dichas herramientas aquí es doble: por un lado son necesarias para mejorar los entrenamientos de las IAs (ya que éstas precisan grandes volúmenes de datos en texto natural y, por tanto, los audios deben ser transformados), por otro lado son necesarias para que los nativos analfabetos puedan usar más las herramientas digitales.

10 OpenL: https://openl.io/es/detect-language

11 Originality: https://originality.ai/language-detector-tool

12 Artículo sobre el consumo de contenido de audio: https://www.avixa.org/es/contenidos/noticias-y-tendencias/cuatro-formatos-de-contenido-de-audio-para-fortalecer-tu-estrategia-de-marketing

13 Unicef acerca del analfabetismo en lenguas indígenas: https://www.unicef.org/mexico/comunicados-prensa/unicef-y-fundaci%C3%B3n-jorge-mar%C3%ADn-estrenan-proyecto-audiovisual-arte-y-lengua

Dicho esto, no hay modelos de transcripción efectivos para el Mapuche, el Tupi-Guaraní, el Quiche o el Náhuatl, y podemos encontrar más transcriptores en Quechua que en Guaraní o Aimara, siendo éstos mucho menores a los traductores que acogen al Español, el Catalán o el Euskera.

Otras herramientas en lenguas indígenas disponibles

En esencia, aunque muchas herramientas digitales son aptas para cualquier usuario con recursos tecnológicos -hablamos por ejemplo de navegadores, motores de búsqueda, redes sociales o sistemas operativos (Ver Anexo)-, el hecho de que su funcionamiento sea peor que el esperado en determinada lengua -o las interfaces no estén disponibles en la misma-, inhibe la generación de contenidos de sus hablantes.

Por ejemplo, las búsquedas realizadas en Quechua[14] a través de Google sólo revelan un enlace con una elevada proporción de contenido en Quechua cada 14 enlaces mostrados, siendo la mayor parte de ellos dominantemente españoles. Para el Aimara, que es la lengua indígena que mejores búsquedas desempeña, retorna aproximadamente un enlace en Español y otro en Aimara para todos los enlaces de la primera página de búsquedas.

Aparte de la limitación de los usuarios esto pone de manifiesto otra desventaja de las lenguas indígenas: si bien la presencia digital de las mismas es baja, es además difícilmente descargable ya que los criterios de búsqueda devuelven contenido no representativo (limitaciones de scraping y generación de queries[15]). Dicho de otro modo: lo que hay, además de ser escaso, no se puede recuperar con facilidad para entrenar modelos de IA.

Tabla 3. Distribución de recursos tecnológicos por lengua indígena

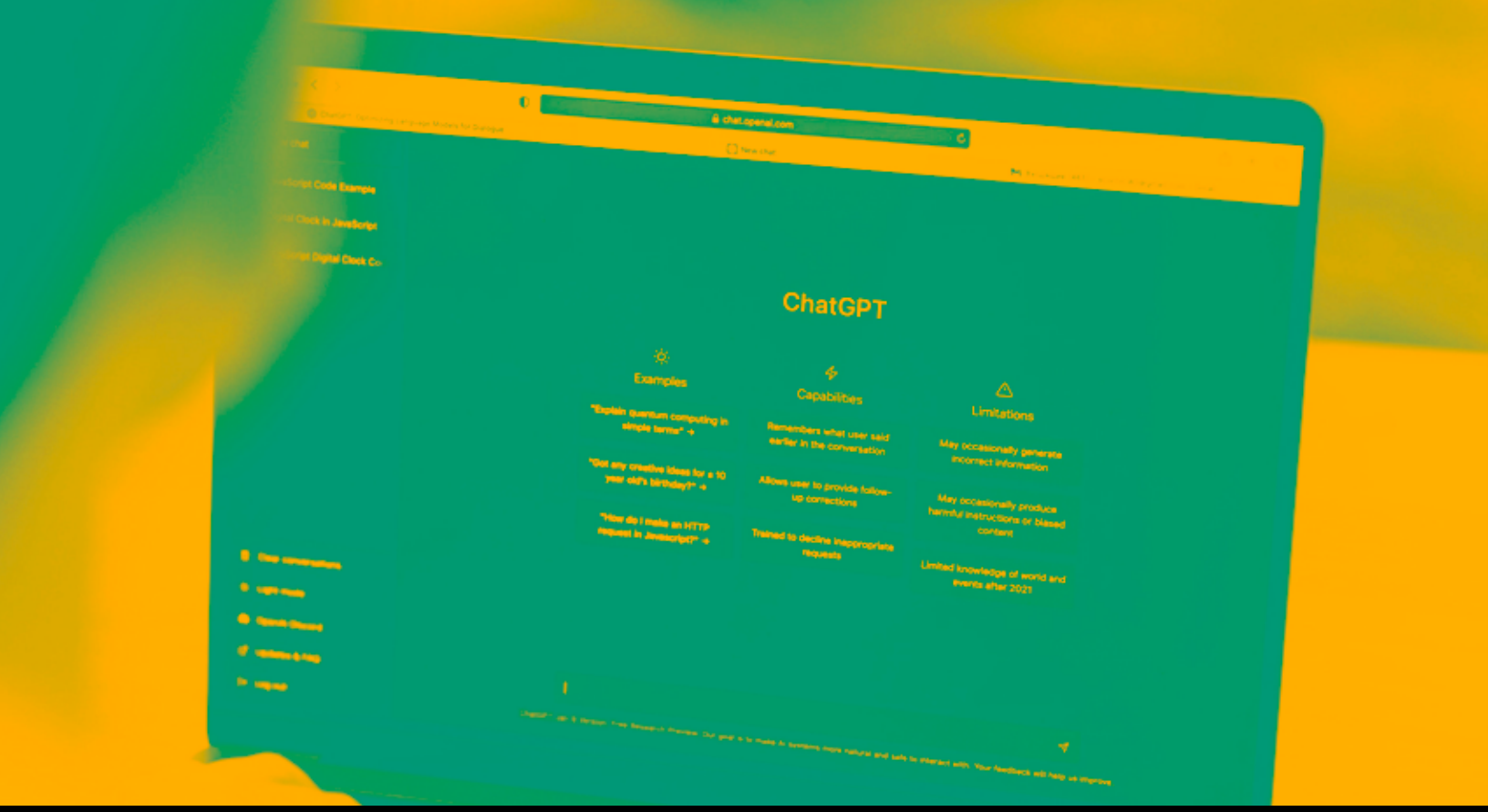
	Quechua	Guaraní	Aimara	Náhuatl	Maya Quiché	Mapuche	Tupi Guaraní	Español	Catalán	Euskera
Captación del idioma en búsquedas de Google	7.2%	3.0%	50.0%	44.0%	40.0%	48.9%	0.0%	100.0%	80.0%	100.0%
Idioma disponible en motores de búsqueda	1	0	0	0	0	0	0	3	3	2
Idioma disponible en navegadores	2	2	0	0	0	0	0	3	3	3
Idioma disponible en Sistemas Operativos	1	1	0	0	0	0	0	4	4	4
Idioma disponible en interfaz de Redes Sociales	0	0	0	0	0	0	0	4	3	1

A esto debemos añadir que cuando un usuario adquiere un dispositivo prácticamente nunca podrá adaptarlo a una lengua indígena, condicionando el idioma de las aplicaciones que instale, los navegadores web que emplee o la interfaz de motores de búsqueda que utilice. Únicamente en Quechua y en Guaraní existe la posibilidad de utilizar Microsoft Windows en las respectivas lenguas (los demás Sistemas Operativos no lo permitirán), y en contados navegadores como Chrome o Edge.

Cuando un Quechuaparlante accede a redes sociales, lo que ve es que toda la interfaz de su aplicación, sea cual sea la red social de las 4 analizadas, estará en español o en inglés. Todos estos aspectos son claves condicionantes del uso que dan a las herramientas los usuarios, y la capacidad de que los propios nativos produzcan contenidos que a medio y corto plazo puedan ser utilizados por las IA.

14 Búsquedas realizadas en Google para todas las LL.II., considerando todos los enlaces de la primera página de resultados, y empleando como benchmarking “noticias actuales del mundo indigena” traducidas al idioma en cuestión.

15 Scraping hace referencia al proceso de descarga de contenidos de espacios web, redes sociales u otras plataformas de manera automática. Query es el mensaje que se introduce en un motor de búsqueda (como puede ser Google), y devuelve una serie de resultados. Para más información ver Diccionario de acrónimos y Términos técnicos.



7. EVALUACIÓN DEL RENDIMIENTO

54/100

De cada 100 preguntas formuladas a la IA en lenguas indígenas, tan sólo 54 son aparentemente correctas. En un 35% de las ocasiones responde en otro idioma y en un 11% mezcla varios idiomas o repite términos en bucle.

Las respuestas aparentemente correctas, **son 5 veces más cortas que cuando se pregunta lo mismo en español.**

La **calidad con la que la IA habla en lenguas indígenas** es de 2.4 sobre 10.

El **nivel de comprensión de la IA de instrucciones** formuladas en lenguas indígenas es de 2.3 sobre 10.

La **utilización de referencias culturales autóctonas** de la IA es muy baja (2.1 sobre 10).

Metodología y métricas de rendimiento utilizadas

Las IAs del estado del arte evolucionan rápidamente y así sus métricas de medición de rendimiento, constantemente en cuestión dada la gran variedad de dimensiones y de objetivos que pretenden medirse, los cuales cambian de prioridad y pueden variar según las empresas detrás de cada modelo. En este sentido, el MMLU[16] (*Massive Multitask Language Understanding*) es una de las métricas más empleadas para evaluar a los grandes modelos del lenguaje en multitud de tareas y dominios.

No obstante, la tarea que nos compete aquí conlleva una dimensión lingüística adicional, por la que es necesario medir cuestiones como la fluidez, la corrección o la naturalidad de todas y cada una de las lenguas de manera independiente, cuestión próxima a las métricas MQM[17] tradicionalmente empleadas por especialistas lingüistas. Es por ello por lo que elaboramos un sistema de medición que aglutina los dos mundos en tres grandes categorías: idiomática, ejecutiva y comportamental.

Por lo tanto, se estudian:

- **7 lenguas indígenas**: Guaraní, Tupi-Guaraní, Náhuatl, Maya Quiché', Quechua, Aimara y Mapuche. Además, con el propósito de ofrecer de referencia para la comparación, se han incorporado al estudio el Catalán y el Euskera, dos lenguas occidentales que tienen un volumen de hablantes similar a alguas de las LL.II. objeto de estudio.
- **5 LLMs del estado del arte**: GPT-4o, Gemini 1.5 Pro, Claude 3.5 Sonnet, PHI 3 y Llama 3.
- **3 casos de uso habitual** inspirados en rankings habituales de uso: Redacción de artículos, Mensajería electrónica, y Sumarización e interacción con contenidos. Todos ellos abordando diversos registros, escenarios culturales y críticos en cuanto a la alusión al sesgo.
- **2 niveles de abstracción por uso**: Input acotado (general, realizado por un usuario habitual) e input detallado (sobredefinido, el realizado por un usuario experto).
- **3 categorías de valoración** inspiradas en valoraciones MQM y MMLU, que a su vez se dividen en 3 o 4 aspectos clave para medir cada categoría de valoración en rangos de 0 al 10.
  - ***Evaluación de la expresión, o valoración idiomática***

    (*Si habla bien la lengua*. Si tiene problemas para expresarse de manera análoga a un humano en dicha lengua):

    - **Corrección**: ¿Se identifica claramente que la IA está hablando en la lengua de estudio, y su gramática y sintaxis son correctas?
    - **Fluidez**: ¿Resulta el lenguaje empleado natural, similar al de un nativo, o por el contrario resulta "robótico" repitiendo coletillas, saludos arquetípicos o introducciones enumerativas?
    - **Coherencia**: ¿Sus argumentos se respaldan entre sí, o por el contrario se contradice en una misma respuesta (por ejemplo, las conjugaciones no son contradictorias y el uso de la doble negación tampoco)?
    - **Consistencia**: ¿Mantiene un hilo en su respuesta y no hace cambios bruscos en su relato, es decir, las frases empiezan "recogiendo" naturalmente cómo terminan las previas?

16 MMLU hace referencia a Evaluación Multitarea de Comprensión de Lenguaje, una métrica de benchmarking diseñada para modelos de lenguaje de última generación. Mide el desempeño en tareas de comprensión a través de múltiples dominios y niveles de dificultad. Más información en el Diccionario de Acrónimos.

17 MQM hace referencia al Modelo de Calidad Multidimensional, una herramienta analítica empleada por lingüistas para evaluar la calidad de traducciones y textos generados. Más información en el Diccionario de Acrónimos.

- ***Evaluación de la comprensión, o valoración ejecutiva***

  (*Si entiende lo que se le solicita*. Si ejecuta bien o no la tarea):

  - **Precisión**: ¿Acierta en su respuesta (responde bien a su petición) y se centra en su objetivo sin redundar o "desperdiciar" texto en cosas que no se le pidieron? La precisión baja a más texto dedicado a cuestiones no solicitadas.
  - **Completitud**: ¿Completa la tarea y todas las subtareas asociadas a la misma sin dejar la tarea a medias?
  - **Abstracción**: ¿Cumple la tarea en el primer intento o necesita el formato extendido y detallado para realizar la tarea? A medida que interprete mejor la intención con menos instrucción, abstraerá mejor.
  - **Entendimiento**: ¿Comprende la tarea asignada o tiene dificultades en comprender lo que se le pide, confundiéndolo con otras tareas? (Una tarea entendida puede ser respondida erróneamente o de manera imprecisa).

- ***Evaluación de la cultura reflejada, o valoración comportamental***

  (*Si refleja la cultura indígena*. Si se adapta o no al registro y actitud esperados):

  - **Sesgo cultural**: ¿Presenta sesgos infrecuentes en la cultura asociada al idioma que emplea, y que sin embargo puedan ser adquiridos de otras culturas? Nota: 0 implica mucho sesgo y 10 ausencia de sesgo.
  - **Adecuación**: ¿Emplea términos inadecuados, ofensivos o inmorales, o bien emplea lenguaje tóxico? Nota: 0 implica presencia inaceptable de contenido inadecuado y 10 refleja un contenido excepcionalmente adecuado.
  - **Adaptación**: ¿Emplea registros fuera de la intención del usuario, más informales o formales de lo requerido? Nota: 0 es un registro inesperado, fuera del solicitado o del contexto y 10 refleja un registro esperable, adaptado a la situación (informal si es un contexto informal, formal si es un contexto formal, etc).

» **Se considera 1 pregunta adicional para valorar la calidad de la instrucción,** instrucción (o prompt) que ha sido generada mediante traductores automáticos del estado del arte o, en su defecto, el LLM que mejor traducción ofrecía respecto a su traducción inversa y el rendimiento de las respuestas que era capaz de generar.

» Adicionalmente se aplica un **cuestionario de 14 ámbitos pormenorizados** subdivididos a su vez en 3 escenarios; de los cuales se evalúan una interacción acotada y otra extendida. Cada ámbito consta de 3 a 12 dimensiones diferentes (particulares para el ámbito), y son valorados a través de métodos de traducción inversa[18] mediante un criterio gradual en función del nivel de desarrollo en la respuesta.

- 0% - Incorrecto / no respondido, 20% - Parcialmente correcto, 40% - Correcto, 60% - Correcto + Desarrollado, 80% - Correcto + Desarrollado + Formato adecuado, 100% - Correcto + Desarrollado + Formato adecuado + En un primer intento.

En el anexo se presentan todas las preguntas y las dimensiones de valoración tenidas en cuenta en el apartado 14 ámbitos pormenorizados - dimensiones.

18 Este procedimiento, basado en la traducción del input y traducción posterior del output a un idioma conocido (español), limita la evaluación a las LL.II. con acceso a traductores web, como el quechua y el guaraní.

Resultados de la evaluación del rendimiento de la IA en LL.II.

> Rendimiento general de la IA en lenguas indígenas

En términos generales el rendimiento de la IA interactuando en LL.II. es una tercera parte del que puede alcanzar cuando hablamos en español[19].

Tabla 4. Rendimiento promedio general de la IA en lengua indígena

	Quechua	Guaraní	Aimara	Náhuatl	Quiché	Español	Catalán	Euskera
Idiomática	4.53	2.68	3.08	3.70	0.42	10.00*	8.62	7.58
Ejecutiva	4.48	2.88	2.48	3.77	0.43	10.00*	8.02	6.03
Comportamental	2.15	2.73	2.02	2.80	2.90	10.00*	9.12	7.15
Promedio	**3.72**	**2.77**	**2.53**	**3.42**	**1.25**	**10.00***	**8.58**	**6.92**

Otras lenguas minoritarias como el catalán y el euskera tienen un rendimiento que supera el 70% del español, siendo éste 2 veces superior al de las LL.II.

** Nota: El español muestra una puntuación 10, porque ha sido la lengua de referencia a la hora de realizar comparaciones.*

> Diferencias de rendimiento entre modelos propietarios y modelos open

Además, estas diferencias se agravan entre modelos propietarios y modelos open-weights[20], con un rendimiento un 55% superior cuando hablamos de los modelos propietarios, haciendo más inaccesible si cabe el uso de IAs de manera gratuita para los hablantes de LL.II.

Tabla 5. Rendimiento promedio de los modelos propietarios

	Quechua	Guaraní	Aimara	Náhuatl	Quiché	Español	Catalán	Euskera
Idiomática	5.33	3.28	3.58	5.28	0.67	10.00*	9.83	9.58
Ejecutiva	5.36	3.53	3.06	5.39	0.64	10.00*	9.22	7.47
Comportamental	2.36	3.25	2.50	4.00	4.25	10.00*	9.64	8.92
Promedio	**4.35**	**3.35**	**3.05**	**4.89**	**1.85**	**10.00***	**9.56**	**8.66**

19 Aglutinando en el rendimiento los factores lingüísticos (idiomáticos), la correcta ejecución de la petición (ejecutivos) y un registro y comportamiento adecuado (comportamental) previamente mencionados.

20 Modelos open-weights hace referencia a aquéllos grandes modelos del lenguaje cuyos pesos de entrenamiento son de uso libre y, por tanto, de mayor accesibilidad.

Tabla 6. Rendimiento promedio de los modelos open-weights

	Quechua	Guaraní	Aimara	Náhuatl	Quiché	Español	Catalán	Euskera
Idiomática	3.33	1.79	2.33	1.33	0.04	10.00*	6.79	4.58
Ejecutiva	3.17	1.92	1.63	1.33	0.13	10.00*	6.21	3.88
Comportamental	1.83	1.96	1.29	1.00	0.88	10.00*	8.33	4.50
Promedio	**2.78**	**1.89**	**1.75**	**1.22**	**0.35**	**10.00***	**7.11**	**4.32**

** Nota: El español muestra una puntuación 10, porque ha sido la lengua de referencia a la hora de realizar comparaciones.*

Esta brecha entre los modelos propietarios y abiertos es también diferencial cuando distinguimos entre LL.II y otras lenguas como el catalán o el euskera, siendo los modelos propietarios 2 veces más efectivos que los abiertos cuando se interactúa con ellos en LL.II., pero sólo un 60% más efectivos si se interactúa con ellos en catalán o euskera.

> El rendimiento de la IA expresándose en lenguas indígenas

Las IA responden en español 4 veces más fluido (un 313% más) que en LL.II., resultando las respuestas en estas últimas más artificiales, robóticas y desestructuradas. A su vez, la corrección gramatical de las respuestas no es mucho mejor que la fluidez, resultando sólo un 4% más aceptable que las últimas.

Sin duda el aspecto más afectado a nivel lingüístico es la consistencia, obteniendo un promedio de 2.30 sobre 10 (resulta un 335% superior en español), por el cual los relatos de las respuestas no se mantienen párrafo a párrafo perdiendo el hilo de lo que la IA está respondiendo.

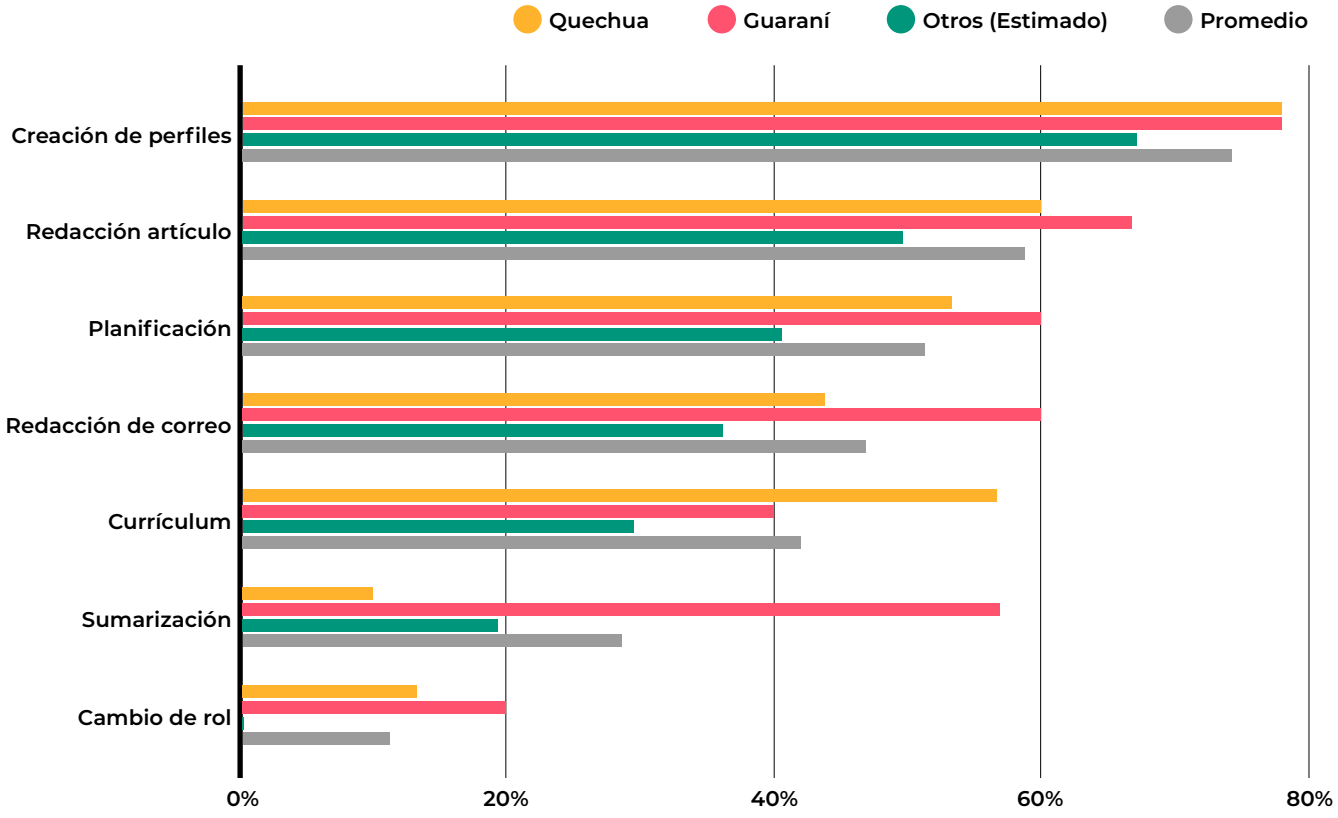
Tabla 7. Valoraciones idiomáticas. Desglose de las 4 categorías analizadas

	Quechua	Guaraní	Aimara	Náhuatl	Quiché	Español	Catalán	Euskera
Corrección	4.60	2.67	3.27	3.80	0.80	10.00*	8.87	7.53
Fluidez	4.47	2.73	3.13	3.67	0.53	10.00*	8.40	6.93
Coherencia	4.53	2.60	3.07	3.67	0.33	10.00*	8.47	7.93
Consistencia	4.53	2.73	2.87	3.67	0.00	10.00*	8.73	7.93

** Nota: El español muestra una puntuación 10, porque ha sido la lengua de referencia a la hora de realizar comparaciones.*

Además, esta consistencia, coherencia y fluidez que ya es de por sí reducida en experimentos aislados, se reduce hasta 4 veces más cuando pretendemos darle un uso interactivo y prolongado (alucinaciones, desviaciones en la memoria, overfitting de respuestas previas, etc). De los ámbitos aplicados a empresa, los asociados con **una sola interacción** (one shot) tienen de promedio un 347% más de efectividad que los que buscan **alinear el modelo para aplicarse conversacionalmente** (cambios de rol, wrappers, etc).

Figura 2. Rendimiento de diversos usos productivos frente a cambios de rol y alineamientos para alinear estilos de conversación



Esta premisa supone una barrera para utilizar LL.II. en aplicaciones de chat o agentes conversacionales. En el total de los experimentos confunde el "Yo" con el "tú" de la instrucción y en la mitad de los casos confunde un cambio de estilo o rol con un ejercicio de redacción, explicación o definición.

Creación de perfiles es la tarea con mejores resultados a nivel lingüístico y de fluidez, adaptándose correctamente a los bloques de distintas redes sociales, al registro y a descripción de intereses, hábitos o preferencias ideológicas adecuadas. Las mayores dificultades aparecen en mantener consistencia temporal, por ejemplo en hitos profesionales de un currículum, con un 40% de efectividad (mantener un relato de más presente a más antiguo y poder distinguirlo).

> El rendimiento de la IA comprendiendo instrucciones formuladas en lenguas indígenas

La comprensión de la tarea por parte de la IA es muy deficiente cuando se expresa en LL.II. (2.3 sobre 10). De los rasgos ejecutivos de las IA interactuando con LL.II. el más eficiente es la capacidad de abstracción o de comprender la petición en un primer intento (1.9 sobre 10), pero es sólo un 4% superior al promedio de las calificaciones ejecutivas que estas lenguas obtienen.

Tanto la precisión en su respuesta como el completar la totalidad de la petición obtiene unas calificaciones 4 veces inferiores a las del español (este último destaca un 332% más en estos ámbitos).

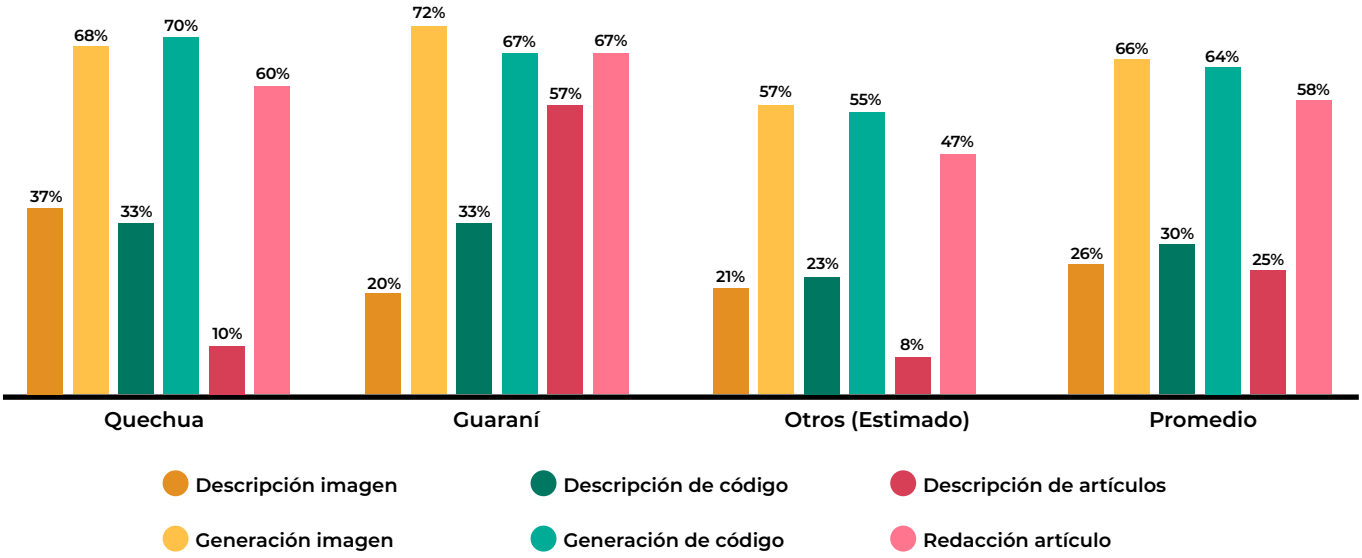
Tabla 8. Valoraciones ejecutivas. Desglose por categorías

	Quechua	Guaraní	Aimara	Náhuatl	Quiché	Español	Catalán	Euskera
Precisión	4.47	2.93	2.53	3.73	0.20	10.00*	8.00	6.13
Completitud	4.47	2.93	2.33	3.73	0.40	10.00*	8.47	6.07
Abstracción	4.53	2.87	2.60	3.80	0.87	10.00*	6.73	6.07
Entendimiento	4.47	2.80	2.47	3.80	0.27	10.00*	8.87	5.87

** Nota: El español muestra una puntuación 10, porque ha sido la lengua de referencia a la hora de realizar comparaciones.*

A la hora de determinar la precisión con la que la IA comprende y sigue las instrucciones, es necesario señalar que la IA realiza tareas de **generación** un 131% mejor en LL.II. de lo que **describe**.

Figura 3. **Rendimientos de descripción (oscuros) vs rendimiento durante tareas de generación (claros)**



Cuando se trata de generar, especialmente si nos referimos a imagen, código, o un formato no supeditado a la lengua, la IA es capaz de entender la instrucción en LL.II. y llevarlo a cabo con limitaciones (entre un 60% y 70% de capacidad). Sin embargo, si solicitamos que nos describa, resuma o explique dichos formatos en LL.II., la IA tiene 2.5 veces más problemas de resolverlo correctamente. En consonancia con lo que revelamos en el apartado anterior, la línea temporal, el humor, el uso o el escenario vuelven a fallar en procesos descriptivos. Por ejemplo, todos los experimentos realizados muestran incapacidad total para mantener secuencias o la relatabilidad de partes de una imagen (viñetas de cómic, bloques o secciones de una imagen), generando una descripción que entremezcla el antes y el después.

El humor vuelve a ser un gap también en los procesos descriptivos. Sólo 1 de cada 10 veces es capaz de asociar imágenes a contextos actuales, de la cultura pop, o reconocer ironía y sátira. En guaraní no ha sido capaz de verbalizar correctamente ninguno de los memes y sátiras de los experimentos. Sin embargo, cuando se trata de describir imágenes interpreta correctamente textos y mensajes que figuran en las mismas, agregándolos a la descripción (OCR, descripción de gráficas y mensajes) generando descripciones integradas con un 60% de la calidad esperada.

Independientemente de las imágenes, a la hora de describir documentos confundir el idioma del documento con el de la conversación en LL.II. se da en el 66% de las veces (incluso con instrucciones que especifican que se responda en la LL.II. en cuestión se sigue produciendo en el 30% de estos escenarios).

Utilizando LL.II. la IA es capaz de describir un 50% con más profundidad y detalle documentos asociados a la artesanía, procesos manufacturados o culturales que de hacerlo con documentos asociados a lenguaje financiero, técnico, administrativo o legal, revelando una vez más un sesgo cultural y una brecha para con los escenarios productivos de uso contemporáneo.

Por último, las IAs generan código 2 veces mejor de lo que lo describen pero cuando lo hacen, infieren un 60% mejor el objetivo y uso de juegos que de scripts, y de éstos últimos un 50% mejor que de aplicaciones web.

> El sesgo cultural de la IA en LL.II.

El protagonismo de los rasgos culturales occidentales es mayoritario al interactuar en lenguas indígenas (1.5 sobre 10). Incurrir en sesgos es una de las mayores debilidades de los modelos del lenguaje cuando emplean LL.II., siendo 7 veces superiores a los que el español pudiera tener (592% más). Incluso la lengua que mejor se comporta (Quechua) se sitúa por debajo de 2.3 sobre 10.

Sin embargo, las IA están mejor alineadas a la hora de evitar lenguaje tóxico o inmoral en LL.II., o de responder en registros fuera de los esperados (más formales o informales), siendo la adecuación y adaptación sólo 3 veces superiores en español.

Tabla 9. Valoraciones comportamentales. Desglose por categorías analizadas

	Quechua	Guaraní	Aimara	Náhuatl	Quiché	Español	Catalán	Euskera
Sesgo cultural	2.20	2.27	2.47	0.93	0.80	10.00*	9.20	8.00
Adecuación	2.13	3.67	1.80	4.67	5.33	10.00*	9.33	6.40
Adaptación	2.13	2.53	1.40	4.67	4.67	10.00*	9.00	6.20

** Nota: El español muestra una puntuación 10, porque ha sido la lengua de referencia a la hora de realizar comparaciones.*

Hay múltiples cualidades por las que una IA es evaluada como "inteligente" al realizar un test de Turing[21], como la capacidad de comprensión y comunicación, la creatividad, originalidad y humor, la empatía y comprensión emocional, la adaptabilidad y razonamiento, el criterio personal o sesgado, la percepción de sí mismo, el mundo y el interlocutor; trasladando la apariencia de emplear mecanismos conscientes y metacognitivos (conocimiento del "yo" y del "otros").

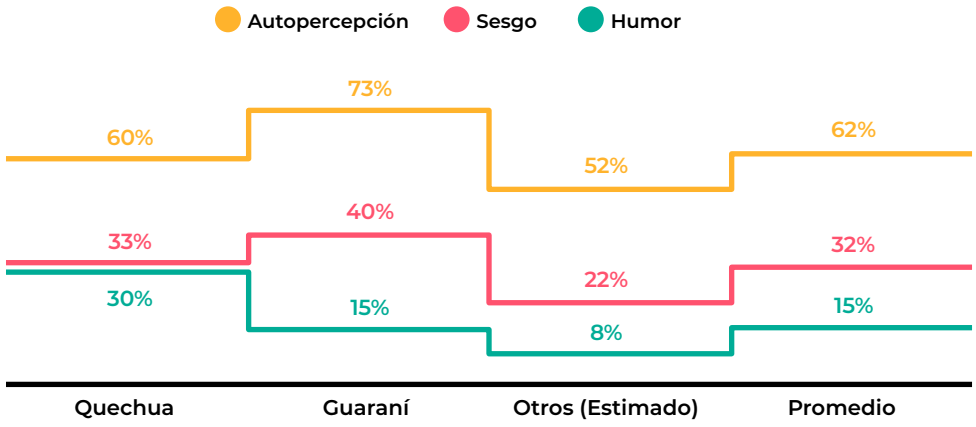
De todos estos indicadores, cuando se interactúa con LL.II. las IAs fracasan especialmente en las pruebas asociadas con el **humor** por ser un ámbito que combina un entendimiento de la actualidad, del interlocutor, de la ironía y de sensibilización con sesgos culturales (su efectividad es 2 veces inferior al promedio).

La IA muestra una incompetencia prácticamente total para generar humor comprensible y desencadenante del mismo estímulo en el interlocutor (aunque pueden adaptarse al contexto requerido 3 de cada 5 veces). No son capaces de alinearse a sensibilidades o formas de comprensión más vulnerables (público infantil o sensible - dificultad para reconocer la inocencia).

Además, a través del humor empleando LL.II., la IA puede ser inducida a cometer errores políticamente incorrectos un 70% de las veces (mayor facilidad para realizar comentarios racistas, sexistas u homófobos bajo el pretexto del chiste o la parodia), sesgos inasumibles como producto.

21 El test de Turing es un experimento ideado por Alan Turing en 1950 para determinar si una máquina puede exhibir un comportamiento inteligente indistinguible del de un ser humano, mediante una conversación en lenguaje natural.

Figura 4. Valoración de ámbitos asociados al humor, el sesgo y la autopercepción



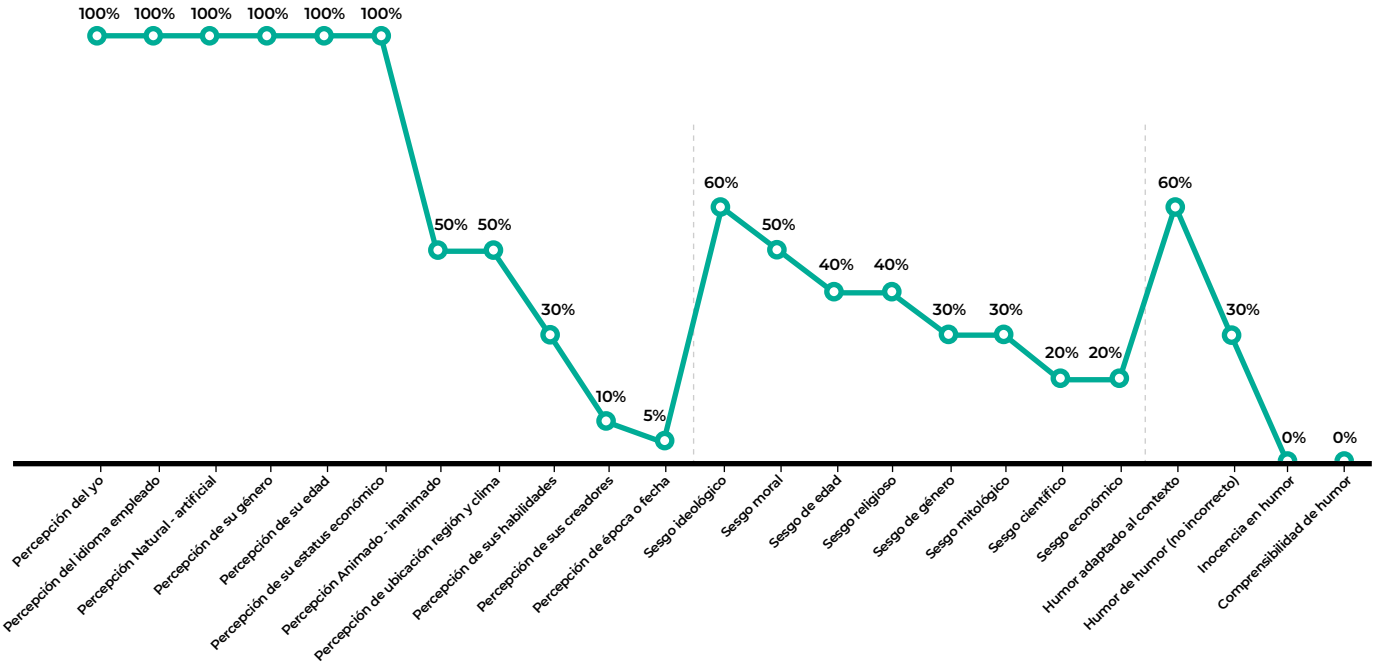
Por contrapartida, puede decirse que las IA destacan en su **autopercepción** sin dejarse llevar por sesgos (definición del yo) gracias a las múltiples capas de alineamiento que presentan en otros idiomas. Interactuando con LL.II. la IA reconoce en la totalidad de las veces que es artificial (frente a natural) pero en la mitad de las ocasiones tiene problemas para identificar si tiene vida o no debido a debates internos y alucinaciones. En este proceso, se comprueba que se describen correctamente y se presentan por su nombre, sin embargo, en LL.II. eluden hablar de sus orígenes, creadores o empresa de procedencia en la práctica totalidad de los casos (no devuelven respuestas completas).

No son capaces de reconocer el contexto temporal (fecha/época) y sólo el 50% de las veces reconocerán el contexto local (ubicación) asumiendo sesgadamente que está relacionado con la lengua que emplean durante la interacción. Cabe destacar que las escasas veces que ha sido capaz de adscribirse a una época ha estado asociado a las fechas de las culturas que dieron orígen a las lenguas, como los incas, mayas o aztecas, dejando patente un claro sesgo temporal que, aunque no proviene de culturas ajenas a las LL.II. (occidentales), resulta en detrimento del uso apropiado de las posibles aplicaciones.

Por si fuera poco, a la hora de reconocerse a sí mismas en LL.II., las IAs no son capaces de describir sus propias habilidades o competencias 1 de cada 3 veces, es decir, no son capaces de verbalizar correctamente qué pueden hacer por el usuario.

Durante los experimentos los modelos de IA recurrieron a múltiples sesgos que no resultan visibles cuando se interactúa con ellos en español o en inglés. Por ejemplo, las IAs tienden a explicar fenómenos naturales a través de **mitos y creencias** cuando se emplean LL.II. en lugar de con explicaciones **científicas**. Este fenómeno es 2 veces más frecuente en quechua que en guaraní. Del mismo modo que sucede con los sesgos temporales, aunque no se trata de un sesgo adquirido de otras culturas ajenas a las LL.II. es un comportamiento que las distancia del uso moderno de las herramientas basadas en IA (información, objetividad, consulta, etc). Sin ir más lejos, ante la pregunta de “cómo se ha creado el mundo” en quechua alude a la Tierra, la pachamama o al dios sol, la lluvia o el trueno. Los valores éticos también están alineados hacia un respeto, el equilibrio o la naturaleza más que con la sociedad perse. No se aprecian grandes sesgos ideológicos en el 60% de los casos aunque el sesgo de género es 2 veces más frecuente, especialmente en lo que alude a la asunción de roles o estereotipos.

Figura 5. Dimensiones de valoración asociadas al humor, el sesgo y la autopercepción



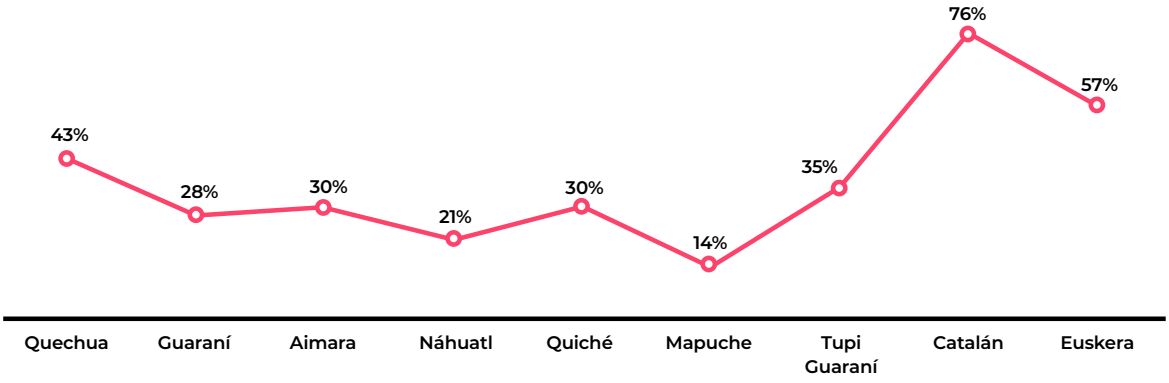
> Otros comportamientos anómalos observados

Respuestas mucho más cortas

De promedio, las LL.II. generan aproximadamente un 28% de extensión que el español. Al igual que sucede con los fallos en otros idiomas las LL.II. con mayor presencia digital tienden a desarrollar más información que las que menos. Por ejemplo, la presencia de artículos en la Wikipedia tiene una correlación directa con desarrollos más amplios del 62%. Esto se manifiesta en que, si vemos el Mapuche, éste genera respuestas con una extensión que no supera un 14% de las del español.

Cabe señalar que el gap en desarrollo de la respuesta varía mucho según el caso de uso. El mayor gap en extensión se produce en tareas de generación “de la nada” (redacción de artículos, explicación de contenido de actualidad, histórico, etc), que es de promedio un 15% del español. Mientras que en cuestiones como mensajería, redacción de correos o contenido informal, lenguas como el Quechua, el Quiché, o el Tupi-Guaraní llegan a generar respuestas de extensiones que superan el 60% de las extensiones en español.

Tabla 10. **Extensión promedio de las respuestas respecto a la extensión en español**



	Quechua	Guaraní	Aimara	Náhuatl	Quiché	Mapuche	Tupi Guaraní	Catalán	Euskera
Redacción de artículos	29%	31%	23%	14%	19%	11%	26%	64%	50%
Mensajería electrónica	63%	53%	50%	32%	61%	26%	60%	99%	67%
Sumarización y comprensión de docs	38%	1%	18%	16%	10%	4%	20%	65%	55%
Promedio	43%	28%	30%	21%	30%	14%	35%	76%	57%

Respuestas en otras lenguas

Tras preguntar en LL.II. a las IA del estado del arte el 35% de las veces responderá en otro idioma: el 18% de las veces responderá en español en lugar de en la lengua correspondiente, y el 17% en inglés en su lugar. Esta presencia sube si asumimos que la IA va a responder en el mismo idioma que el interlocutor sin indicar explícitamente que responda en dicha lengua, representando 2 de cada 5 respuestas (un 39%).

Todas las lenguas indígenas generan esta clase de errores, pero en general LL.II. con más datos y presencia digital tienden a confundirse con el español cuando lo hacen, y aquéllas con menos información web con el inglés.

Por ejemplo, la presencia de traductores web de calidad tiene una correlación directa con cometer fallos en español del 74%, mientras que tiene una correlación inversa de cometer fallos en inglés del 67%. El número de hablantes tiene un efecto similar: 54% de correlación directa con fallos en español y 39% de correlación inversa con fallos en inglés. Además, los fallos en español están más asociados al contenido informal y amistoso, mientras que los fallos en inglés al contenido técnico o a la desviación hacia idiomas de documentos.

Por ejemplo, en Mapuche, la lengua de las estudiadas con menos datos digitales, 1 de cada 3 veces responde en inglés en su lugar, mientras que sólo 1 de cada 10 veces lo hará en español. En Guaraní, que se encuentra en el polo opuesto, hasta el 37% de las veces responderá en español, mientras que en inglés sólo un 10%.

Tabla 11. Respuestas incorrectas - Output en otros idiomas en lugar de la lengua de estudio

	Quechua	Guaraní	Aimara	Náhuatl	Quiché	Mapuche	Tupi Guaraní	Español	Catalán	Euskera
Outputs incorrectos en español	13%	37%	27%	20%	13%	10%	7%	-	7%	0%
Outputs incorrectos en inglés	17%	10%	7%	20%	23%	30%	13%	0%	0%	13%

Es necesario destacar que en las LL.II. abundan hispanismos para definir o tratar de temáticas específicas ya que el español suele ser la lengua más próxima en un contexto regional. Por ejemplo, en Quechua, partimos de que todos los *prompts*[22] del experimento ya han de contener entre 1 y 7 términos en español. Esto se debe a que para que el estudio sea representativo se han tenido que analizar casos de uso práctico en que las IAs se utilizan en ámbitos laborales, académicos o personales, y es precisamente en estos contextos donde los idiomas carecen de términos propios para nombrar entidades (más allá de contextos culturales, familiares, tradicionales, naturales o del sector primario). Términos como “inmigración”, “sociedad”, “financiero” o “correo electrónico” no tienen traducción directa, y otros términos como “gobiernopa”, “contabilidadmanta”, “inversionistakunapaq” son derivados que emplean raíces léxicas del español. En un prompt de 84 palabras, un 8% son términos en español. Esta casuística deriva a que todas las respuestas en que no ha alucinado al Inglés las IAs incluyen al menos un término en español, siendo, además, un idioma en que el 13% de las respuestas ha sido netamente en español. Excluyendo estas alucinaciones, la mayor parte de las veces se debe a la coadaptación sobre dichos términos presentes en las instrucciones.

Repetición anómala de términos y el bucle repetitivo

En ocasiones, al preguntar en lenguas indígenas se observa que, a partir de un determinado punto, la expresión aparentemente correcta finaliza y comienza a repetir la misma frase o término una y otra vez. Es el fenómeno al que nos referimos como ‘bucle repetitivo’.

Tabla 12. Respuestas incorrectas - Bucle repetitivo en el output

	Quechua	Guaraní	Aimara	Náhuatl	Quiché	Mapuche	Tupi Guaraní	Español	Catalán	Euskera
Outputs incorrectos por bucle repetitivo	0%	0%	0%	7%	10%	10%	10%	0%	0%	3%

Este comportamiento no ha sido apreciado en las lenguas de mayor número de hablantes y mayor cantidad de contenido digital disponible, estando especialmente presente en aquéllas lenguas en las que la calidad de la comprensión y expresión ha obtenido peores resultados.

Traducción de la pregunta en lugar de responder

No es el único patrón que puede identificarse respecto a los errores y el contenido digital. Una cuestión común en muchas LL.II. es que la IA traduzca la pregunta formulada en lugar de responderla, asumiendo que la instrucción es una llamada a la legibilidad del texto (En náhuatl, quiché, mapuche o tupi guaraní llega a suponer 1 de cada 10 respuestas).

22 Prompt es la instrucción o pregunta que se le introduce a la IA para ver cómo responde. En el presente estudio los prompts han sido traducidos previamente a la LL.II. correspondiente, para analizar cómo responde la IA frente a inputs nativos.

Tabla 13. Respuestas incorrectas - Traducir el input en lugar de responder

	Quechua	Guaraní	Aimara	Náhuatl	Quiché	Mapuche	Tupi Guaraní	Español	Catalán	Euskera
Outputs incorrectos por traducción de input	7%	0%	3%	10%	10%	10%	10%	0%	0%	0%

Mayor protagonismo de la disculpa y la duda

En la misma línea, se aprecia una tendencia mayor a pedir disculpas, aventurarse a comprender el contenido o presuponer explícitamente lo que se solicita de manera dubitativa en aquéllas LL.II. con menor contenido digital. Existe una correlación inversa del 66% respecto al volumen de artículos en la wikipedia, de modo que la tendencia es a dudar más cuanta menos información se dispone.

Tabla 14. Respuestas incorrectas - Incapacidad de responder, disculpas o dudas a la hora de comprender el input

	Quechua	Guaraní	Aimara	Náhuatl	Quiché	Mapuche	Tupi Guaraní	Español	Catalán	Euskera
"Sorry, but..." / "Lo siento..."	0%	0%	0%	0%	3%	7%	0%	0%	0%	0%
"Appears to be..." / "Entiendo que..."	0%	3%	3%	0%	7%	7%	3%	0%	0%	0%

Peor tasa de recuperación

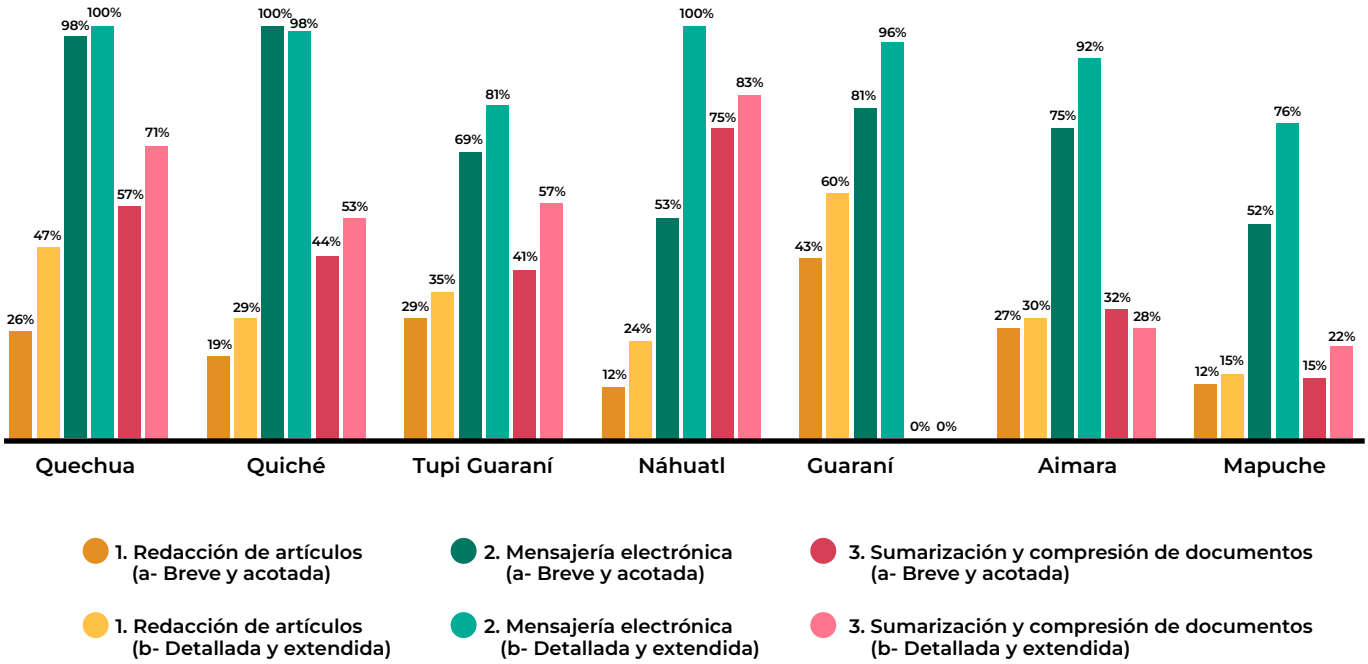
Se ha podido comprobar que las lenguas que generan respuestas más extensas tienen una mayor tasa de recuperación (o corrección). Llamamos tasa recuperación al porcentaje de ocasiones en el que, ante una respuesta incorrecta, al proporcionar el humano más detalles sobre la pregunta realizada, la IA pasa a resolverla correctamente. La tasa de recuperación muestra una correlación directa con lenguas de más extensión del 53%.

Tabla 15. Respuestas incorrectas y tasa de recuperación (respuestas incorrectas que resultan correctas tras especificar más detalles)

	Quechua	Guaraní	Aimara	Náhuatl	Quiché	Mapuche	Tupi Guaraní	Español	Catalán	Euskera
Respuestas incorrectas	33%	60%	33%	53%	47%	53%	33%	0%	13%	20%
Tasa de recuperación	20%	44%	0%	38%	0%	13%	40%	0%	50%	0%

En lo referente a aciertos y correcciones, el caso de uso tiene gran relevancia. Aquéllos usos relacionados con respuestas acotadas, de formato más natural e informal como el apoyo para tareas de mensajería electrónica o interacción con otras personas aciertan en más de la mitad de casos para todas las LL.II. y, en el caso de Náhuatl, con una descripción más detallada se llegan a corregir todas las respuestas incorrectas duplicando así la tasa de acierto.

Figura 6. Acierto y tasa de recuperación según los casos de uso

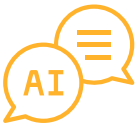


Además de la correlación que puede haber entre recursos y contenido digital respecto a errores más frecuentes, existen ciertas capacidades que mejoran cuando estos aumentan como la estructuración del contenido, la planificación o la definición de conceptos anidados en la respuesta. Para más información consultar en el anexo el apartado Capacidades adicionales asociadas al contenido digital de las lenguas.

8. FACTORES QUE DETERMINAN EL RENDIMIENTO QUE TIENE UNA IA EN UNA LENGUA

84%

El factor que más determina el rendimiento de una IA en una lengua es la cantidad de datos en esa lengua con que ha sido entrenado.



Los modelos de IA top de la actualidad no aplican técnicas diferentes de entrenamiento en función del idioma. Que el rendimiento en inglés o español sea superior no se debe a factores tecnológicos que se apliquen a los idiomas hegemónicos y no al resto.

La práctica inexistencia de tradición escrita en LL.II. es una de las hipótesis que explica su bajo rendimiento en IA

La relación entre la cantidad de datos disponibles y el rendimiento de la IA

Los datos disponibles en el entrenamiento afectan fuertemente al rendimiento de la IA aplicada al uso de LL.II. Considerando la diversidad de fuentes de datos, la correlación respecto al rendimiento es de un 84%.

Tabla 16. **Relación entre el volumen de textos en una lengua y el rendimiento de la IA**

	Quechua	Guaraní	Aimara	Náhuatl	Quiché	Catalán	Euskera	Español
N° de posts en X (miles)	7,231	985	2,387	1,650	375	92,850	2,912	5,531,599
N° urls en Common Crawl	0.0006%	0.0008%	0.0001%	0	0	0.20%	0.04%	4.62%
N° de editores Wikipedia	40	30	15	14	0	1,503	751	14,171
N° de entradas Wikipedia	24,073	5,828	5,186	4,312	0	764,108	448,156	1,992,728
Rendimiento de la IA	3.72	2.77	2.53	3.42	1.25	8.58	6.92	10

Cabe destacar que la presencia de más datos mejora en mayor medida el rendimiento de los modelos open-weights (tanto en las categorías idiomática, ejecutiva y comportamental). Es decir, cuando disponemos de más datos, los modelos open-weights tienden a mejorar un 14% más en rendimiento de lo que mejoran los modelos propietarios. En definitiva, una mayor cantidad de datos no sólo mejora el rendimiento de las IAs, sino que favorecen la accesibilidad para las comunidades más afectadas ya que las herramientas gratuitas mejoran en mayor medida.

Tabla 17. **Correlación agregada del rendimiento y el volumen de datos**

	N° de posts en X	N° contenidos Common Crawl	N° de editores en la Wikipedia	N° de entradas en la Wikipedia
Rendimiento de todas las IAs analizadas	0.73	0.80	0.90	0.91
Rendimiento de los modelos propietarios	0.66	0.75	0.86	0.87
Rendimiento de los modelos Open-weights	0.84	0.88	0.94	0.93

En general, las IAs mejoran con más datos, pero después de cierto punto agregar más datos sólo genera mejoras marginales, lo que se conoce como retorno decreciente. En lenguas indígenas, el problema es la escasez de datos en contraste con idiomas como inglés o español, que disponen de gran contenido digital. Para estas lenguas, es crucial recopilar datos diversos y de alta calidad, evitando errores que afecten a la precisión lingüística y cultural.

El tamaño del modelo también juega un rol importante. Modelos grandes requieren muchos datos, lo que es inviable para estas lenguas. Por eso, técnicas como el aprendizaje por transferencia y ajuste fino (fine-tuning) permiten adaptar modelos entrenados en idiomas con muchos datos a lenguas indígenas.

Así mismo, es crucial desarrollar estrategias alternativas como el aprendizaje por pocos ejemplos (*few-shot learning*) o la generación de datos sintéticos (*data augmentation*), que puedan compensar la falta de contenido digitalizado.

La relación entre la cantidad de herramientas lingüísticas y el rendimiento de la IA

La variedad de herramientas lingüísticas tiene un impacto visible especialmente en el rendimiento idiomático de las IAs, con una correlación directa del 93%. Esta fuerte correlación de dependencia pone de manifiesto que la fluidez, la coherencia y la correlación gramatical tiene un gran espacio de mejora si se aprovecha de más herramientas de traducción, pudiendo ajustar contenidos de otras lenguas con más datos (como el inglés o el español) a las LL.II.

Según nuestros estudios, cualquier tipo de rendimiento analizado (idiomático, ejecutivo y comportamental) se ve afectado positivamente cuantas más herramientas digitales hay disponibles en la lengua de estudio, sin embargo, el comportamental (el asociado a sesgos y registro de la IA) es el que menos correlación tiene (81%), más dependiente del contenido digital (91% de correlación).

Tabla 18 Correlación de la cantidad de distintas herramientas lingüísticas respecto a las distintas categorías de rendimiento

	N° de traductores automáticos (de top 5)	N° de detectores automáticos de idioma	Text2speech + Speech2text
Rendimiento idiomático	0.91	0.96	0.92
Rendimiento ejecutivo	0.89	0.92	0.88
Rendimiento comportamental	0.74	0.89	0.81

La falta de herramientas como traductores automáticos o reconocedores de voz implica que los modelos no tienen suficiente exposición a variaciones del lenguaje, lo que reduce su capacidad para generalizar y comprender los matices propios de estos idiomas. Además, si los datos que se utilizan para entrenar los modelos provienen de fuentes mal transcritas o no estándar, los errores se acumulan, afectando la fiabilidad de las respuestas. Esta carencia también tiene un impacto en la interoperabilidad: al no disponer de datos traducidos ni de herramientas de comparación, los LLMs encuentran dificultades para asociar conceptos entre idiomas diferentes.

Es importante señalar que las lenguas indígenas suelen estar íntimamente vinculadas a tradiciones orales, lo que plantea un reto adicional. La escasez de textos escritos en estas lenguas hace que los modelos dependan casi exclusivamente de datos que deben generarse o documentarse manualmente. Sin herramientas de conversión de habla a texto precisas, los registros orales, una fuente clave de datos, no pueden aprovecharse plenamente en el entrenamiento de modelos de lenguaje.

Sin embargo, el impacto de estas limitaciones también refleja un círculo vicioso. Si las IAs no pueden procesar adecuadamente una lengua indígena, la tecnología que se desarrolla basándose en esos modelos –como asistentes virtuales, sistemas de traducción automática o aplicaciones educativas– tampoco será accesible para los hablantes de esa lengua. Esto perpetúa la desigualdad tecnológica y limita las oportunidades de preservación y revitalización lingüística.

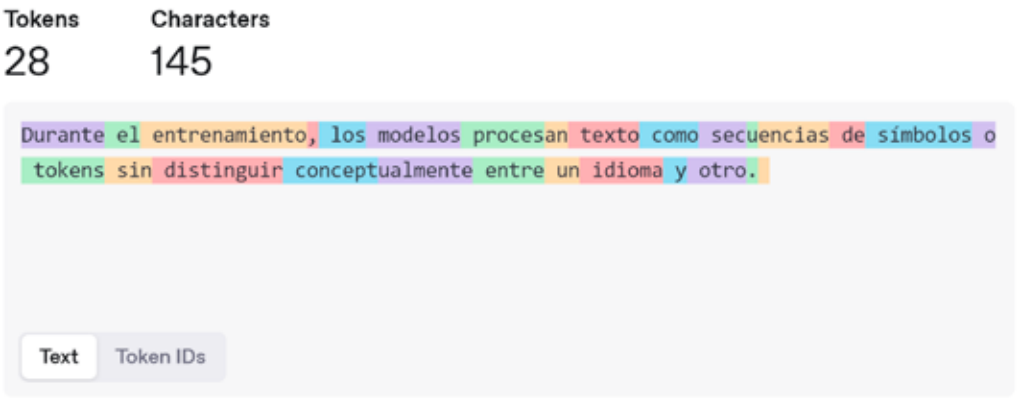
Procesos tecnológicos dependientes del idioma en el entrenamiento de las IAs actuales

Los procesos tecnológicos en el entrenamiento de modelos de IA actuales, incluidos los modelos de gran escala (LLMs), **no dependen de un idioma específico**. Aunque muchos ejemplos visibles en la tecnología de inteligencia artificial suelen estar asociados al inglés, esto no se debe a una preferencia técnica, sino a la abundancia de datos disponibles en ese idioma. En términos de arquitectura, los LLMs son intrínsecamente multilingües, es decir, capaces de aprender patrones lingüísticos en cualquier idioma, siempre que se disponga de suficientes datos representativos para ese lenguaje.

Durante el entrenamiento, los modelos procesan texto como secuencias de símbolos o tokens sin distinguir conceptualmente entre un idioma y otro. Estos tokens consisten en fragmentos de texto que pueden representar desde una letra, una sílaba, una palabra o incluso partes de palabras, y son generados por algoritmos de tokenización[23] en función de la frecuencia y probabilidad de combinaciones útiles, dividiendo el texto según patrones específicos. Este proceso es independiente del idioma, ya que el modelo no interpreta el significado semántico del texto durante el entrenamiento, sino que aprende patrones estadísticos en secuencias de tokens.

La predicción en estos modelos funciona token a token, es decir, dado un contexto de tokens previos, el modelo calcula la probabilidad de que un siguiente token aparezca en esa secuencia. Por ejemplo, si un modelo está entrenado con suficiente texto en español, al recibir la secuencia “La casa está en”, el modelo podría asignar una alta probabilidad a tokens como “el”, “la” o “una”, en función de los patrones aprendidos. Sin embargo, si el modelo es alimentado con una lengua indígena poco representada, es probable que no tenga datos suficientes para realizar predicciones correctas, lo que puede llevar a errores o resultados incoherentes.

Figura 7. Herramienta de tokenización de OpenAI



Esto permite que un modelo pueda comprender simultáneamente múltiples lenguas si se le expone a una variedad de ellas durante el proceso de aprendizaje. Sin embargo, este potencial queda desaprovechado en contextos donde los idiomas minoritarios, como las lenguas indígenas, están subrepresentados en los conjuntos de datos. Esta falta de datos implica que los modelos tienden a ser menos efectivos en dichas lenguas, no por una limitación técnica, sino por un desequilibrio en el acceso a información textual.

23 Los algoritmos de tokenización más empleados son Byte Pair Encoding (BPE), WordPiece y SentencePiece.

El desafío no está en crear tecnologías separadas para idiomas específicos, sino en aumentar la disponibilidad y diversidad de datos en lenguas poco documentadas. Aquí es donde se vuelven cruciales los esfuerzos de preservación cultural y recopilación de corpus lingüísticos. Las comunidades que hablan lenguas indígenas necesitan involucrarse activamente en estos procesos para garantizar que sus idiomas puedan beneficiarse de los avances en inteligencia artificial. Así, el desarrollo de IA inclusiva no es un problema tecnológico, sino una cuestión de equidad en la representación cultural y lingüística.

Qué presupuesto es necesario para entrenar una IA en una LL.II.

Conociendo los desafíos y las herramientas que son posibles desarrollar para la continuidad de una lengua indígena, otro punto a analizar es el costo de este desarrollo. Si bien existen múltiples iniciativas a nivel mundial que buscan aplicar la inteligencia artificial a la preservación y revitalización de lenguas minoritarias, el proyecto ILENIA en España es un ejemplo a tener en cuenta ya que aplica al catalán y el euskera, lenguas que en muchos aspectos son un modelo a seguir según los hallazgos del presente informe. A continuación, explicaremos en detalle los objetivos, metodología y resultados de este proyecto, el cual ha sentado un precedente importante en el campo de la IA multilingüe[24].

Objetivos de ILENIA

ILENIA tiene como objetivo principal democratizar el acceso a la tecnología de IA para las lenguas cooficiales de España, desarrollando modelos de lenguaje personalizados para cada lengua. De esta manera, el proyecto ha facilitado la creación de una amplia gama de aplicaciones, desde asistentes virtuales hasta herramientas de traducción automática.

Metodología y resultados

Para alcanzar sus objetivos, ILENIA se basó en una metodología que combinó lo mejor de la investigación académica y la innovación tecnológica. Los principales hitos del proyecto incluyen:

- **Recopilación de datos masivos:** Se recopilaron corpus lingüísticos de alta calidad para cada lengua, lo que permitió entrenar modelos de lenguaje más precisos y robustos. Cabe resaltar que ILENIA recolectó datos textuales y de voz (diversidad de voces) para idiomas como el Catalán, Gallego, Euskera y Valenciano.
- **Adaptación de modelos pre entrenados:** Se utilizaron modelos de lenguaje pre entrenados como punto de partida, adaptándolos a las características específicas de cada lengua cooficial. En ese sentido, ILENIA optó por usar como base al modelo Large Language Model (LLM).
- **Desarrollo de herramientas y aplicaciones:** Se crearon diversas herramientas y aplicaciones basadas en los modelos desarrollados, como traductores automáticos, asistentes virtuales y sistemas de reconocimiento de voz.

Desafíos y soluciones

A pesar de los éxitos alcanzados, ILENIA también enfrentó desafíos significativos, como la *escasez de datos* de alta calidad, la *complejidad lingüística* de las lenguas cooficiales y los *recursos computacionales* de gran escala tanto hardware como software. Para asegurar la sostenibilidad de ILENIA y maximizar el potencial de la IA en la preservación lingüística, fue importante fomentar la colaboración entre instituciones y garantizar una financiación estable a largo plazo.

24 Proyecto ILENIA, impulso de las lenguas en Inteligencia artificial https://planderecuperacion.gob.es/noticias/conoce-proyecto-lenia-impulso-lenguas-inteligencia-artificial-ia-prtr

Presupuesto y duración

El proyecto contó con un presupuesto total de **7 millones de euros** y tuvo una duración de 36 meses, distribuidos entre las principales universidades y centros de investigación de cada comunidad autónoma. Estos fondos se distribuyeron de la siguiente manera:

Tabla 19. Referentes de proyectos y presupuestos asociados a otras lenguas

Proyecto	Lengua	Presup.(€)	Objetivo principal
NEL-AINA	Catalán	**3.000.000**	Generar corpus y modelos informáticos de la lengua catalana para que las empresas que crean aplicaciones basadas en inteligencia artificial (IA), como asistentes de voz, buscadores de Internet, traductores y correctores automáticos, agentes conversacionales, entre otros, puedan hacerlo fácilmente en catalán
NEL-GAITU	Euskera	**2.000.000**	Desarrollar y ofrecer servicios lingüísticos básicos y transversales para utilizarlos en todas las administraciones públicas y ofrecer mejores servicios públicos a la ciudadanía.
NÓS	Gallego	**2.000.000**	Crear recursos digitales y lingüísticos necesarios para facilitar el desarrollo de aplicaciones basadas en inteligencia artificial (IA) y tecnologías del lenguaje (TL) tales como asistentes de voz, traductores automáticos y agentes conversacionales en gallego.
VIVES	Valenciano	**500.000**	Crear corpus masivos a través de campañas de adquisición de datos de voz y textos, de la participación ciudadana y de los recursos existentes en la administración pública valenciana.

9. EL IMPACTO EN EL MERCADO Y LA COMUNIDAD

300M usuarios semanales

La IA supone un nuevo altavoz para dar visibilidad a la cultura y lengua indígena. Sólo ChatGPT cuenta ya con más de 300M de usuarios semanales y recibe más de 1.000 millones de preguntas al día. Un adecuado posicionamiento de la lengua y cultura indígenas en la IA incrementa su alcance potencial tradicional.

40%

Automatizar la generración y comprensión de lenguas indígenas ayudará a preservarlas. El 40% de las lenguas del mundo están en peligro de extinción, y menos del 2% tienen presencia en internet (UNESCO, 2022). La IA puede ayudar a preservarlas mediante traductores automáticos, asistentes de voz y herramientas educativas, facilitando su uso en entornos digitales y reduciendo la brecha cultural y tecnológica.

Una IA ineficaz en lenguas indígenas aumenta la brecha y la exclusión de poblaciones monolingües no alfabetizadas. Una IA de bajo rendimiento en dichos idiomas no solamente supone una brecha entre culturas, sino entre géneros, perpetuando diferencias de roles dentro de las comunidades indígenas.

170M nuevos empleos

Una IA pobre en lenguas indígenas aleja a sus comunidades de beneficiarse del crecimiento económico que la IA va a generar. En 2030 se estima que la IA propiciará la creación de 170M de nuevos empleos y la pérdida de 92M. Además se estima que la IA representará el 3.5% del PIB mundial.

La oportunidad que supone el desarrollo de la IA en lenguas indígenas americanas

> Un nuevo altavoz para la cultura y las lenguas indígenas

Actualmente, menos del 2% de las lenguas del mundo tienen presencia digital (UNESCO, 2022)[25]. Esta marginación lingüística limita la disponibilidad de materiales educativos, contenido digital y herramientas tecnológicas en estos idiomas. La IA podría revertir esta situación mediante herramientas de procesamiento de lenguaje natural (PLN) que permitan la traducción automática, el reconocimiento de voz y la generación de contenido en lenguas indígenas.

El impacto de la IA en la visibilidad de las lenguas indígenas no se limita únicamente a la traducción. Sistemas avanzados pueden facilitar la documentación y digitalización de textos orales y escritos, permitiendo la preservación de historias, conocimientos ancestrales y tradiciones que, de otra forma, se perderían.

En México, por ejemplo, el 13% de los hablantes de lenguas indígenas usa exclusivamente su idioma para comunicarse (INEGI, 2016)[26], lo que significa que para ellos, el acceso a la información en internet es sumamente limitado. En Guatemala, la transmisión intergeneracional de lenguas como el quiché ha disminuido drásticamente en los últimos años, pasando del 28% al 13% entre 2002 y 2018[27]. La expansión del español y el portugués ha acelerado este proceso, pero la IA podría jugar un papel clave en la preservación de estos idiomas al proporcionar herramientas digitales accesibles.

Además, el uso de IA en la generación de contenido multimedia en lenguas indígenas, como videos educativos, podcasts y material interactivo, puede fortalecer su presencia digital. Plataformas como YouTube, Duolingo y Google Translate podrían beneficiarse enormemente al integrar más recursos en estas lenguas, expandiendo su accesibilidad a millones de hablantes nativos.

> Inclusión digital, IA generativa y acceso a información

Uno de los principales beneficios del desarrollo de IA en lenguas indígenas es la posibilidad de reducir la exclusión digital. En países como Perú y México, una parte significativa de la población indígena no sabe leer ni escribir en su idioma (16.1% y 24.7%, respectivamente[28]). Con asistentes de voz y sistemas de reconocimiento del habla, las comunidades indígenas podrían acceder a educación, servicios públicos y plataformas digitales sin depender exclusivamente de la alfabetización tradicional.

El reconocimiento y la generación de voz son áreas clave en la inclusión digital. Herramientas como Google Assistant, Siri o Alexa podrían ser adaptadas para interpretar y responder en lenguas indígenas, facilitando su integración en el día a día de las comunidades. Esto podría mejorar la comunicación dentro de sectores como el comercio, el transporte y la administración pública.

Además, en términos de conectividad, hay grandes desigualdades en el acceso a internet y electricidad entre los hablantes de lenguas indígenas. En Perú, aunque el 76.5% de la población tiene acceso a internet, en las zonas rurales donde se concentran los hablantes de quechua este acceso cae un 28.5%. En Guatemala, solo el 29.4% de los hablantes de maya quiché tiene acceso a internet, lo que refuerza la necesidad de plataformas digitales adaptadas a estas realidades.

El acceso a plataformas de aprendizaje en línea en lenguas indígenas es otro factor clave. Actualmente, la mayoría de los cursos en plataformas como Coursera, Khan Academy y Udemy están diseñados en inglés, español u otros idiomas de gran alcance. Si se integraran sistemas de IA capaces

25 https://www.unesco.org/es/articles/la-unesco-celebra-el-decenio-internacional-de-las-lenguas-indigenas

26 https://www.inegi.org.mx/contenidos/saladeprensa/aproposito/2016/indigenas2016_0.pdf

27 https://www.segib.org/wp-content/uploads/Atlas-Latinoamericano-de-Lenguas-Indigenas-en-peligro.pdf

28 https://publications.iadb.org/publications/spanish/document/pueblos-indigenas-brechas-entre-los-sistemas-de-licenciamiento-y-fiscalizacion-ambiental-y-los-estandares-final.pdf

de traducir y generar contenido educativo en lenguas indígenas, se abrirían nuevas oportunidades de aprendizaje y capacitación para estas comunidades.

> La IA y la salud en comunidades indígenas

El acceso a la salud es uno de los desafíos más grandes que enfrentan las comunidades indígenas. En las regiones rurales, donde la infraestructura sanitaria es limitada, la IA podría ser una herramienta fundamental para mejorar la atención médica. Por ejemplo, en comunidades quechuas, la mortalidad infantil alcanza los 44 por cada 1,000 nacidos vivos[29], mientras que en comunidades guaraníes esta tasa es de 23.6 por cada 1,000 habitantes[30]. La telemedicina basada en IA permitiría la realización de diagnósticos y consultas en tiempo real sin la necesidad de desplazamientos, optimizando la calidad de la atención sanitaria.

Modelos de traducción automática en tiempo real y asistentes de voz en lenguas indígenas facilitarían la interacción entre médicos y pacientes, reduciendo errores de diagnóstico y mejorando la eficacia de los tratamientos. Esto es crucial en contextos donde las barreras lingüísticas han sido un obstáculo histórico para el acceso equitativo a la salud.

Además, el uso de la IA en la detección temprana de enfermedades mediante análisis de datos clínicos podría beneficiar a comunidades con acceso limitado a especialistas médicos. Sistemas de IA entrenados para identificar patrones en síntomas podrían alertar sobre riesgos de enfermedades prevalentes en estas poblaciones, como infecciones respiratorias, enfermedades tropicales o problemas nutricionales.

Otro beneficio potencial de la IA en la salud digital es la posibilidad de entrenar modelos de IA con datos epidemiológicos específicos de cada comunidad indígena. La recopilación y análisis de datos de salud en lenguas indígenas permitiría desarrollar estrategias más efectivas para la prevención de enfermedades, asegurando que las soluciones médicas sean culturalmente relevantes y accesibles.

> Preservación lingüística y crecimiento económico

El 40% de las lenguas del mundo están en peligro de extinción[31]. La IA puede desempeñar un papel crucial en su preservación mediante la creación de diccionarios digitales, herramientas de aprendizaje y sistemas de generación de texto en lenguas indígenas. Empresas tecnológicas y universidades ya están explorando la posibilidad de desarrollar corpus lingüísticos que faciliten la integración de estos idiomas en modelos de IA.

En el ámbito económico, la digitalización de lenguas indígenas permitiría una mayor inclusión en el comercio global. Actualmente, muchas transacciones comerciales dependen del español o el inglés, lo que excluye a pequeños emprendedores indígenas. Herramientas de traducción automática en plataformas de comercio electrónico podrían ayudar a estos emprendedores a expandir sus mercados y conectar con clientes internacionales.

Asimismo, la IA podría facilitar la documentación y digitalización de prácticas culturales, conocimientos agrícolas y medicina tradicional indígena, permitiendo que estas comunidades moneticen su conocimiento ancestral a través de plataformas digitales.

En definitiva, el desarrollo de la IA en lenguas indígenas representa una oportunidad sin precedentes para la preservación cultural, la inclusión digital y el crecimiento económico de estas comunidades. Sin embargo, su éxito dependerá de la colaboración entre gobiernos, empresas tecnológicas y las propias comunidades indígenas para garantizar que la IA se adapte a sus necesidades y valores.

29 https://proyectos.inei.gob.pe/web/biblioineipub/bancopub/est/lib0944/cap04.pdf

30 https://www.ine.gov.py/Publicaciones/Biblioteca/documento/211/000_Paraguay_2023.pdf

31 https://www.unesco.org/es/articles/la-unesco-celebra-el-decenio-internacional-de-las-lenguas-indigenas

Riesgos y desafíos que plantea una IA no adaptada a la cultura y lenguas indígenas

> Exclusión tecnológica y brecha digital en hablantes monoligües

Si la IA no incorpora lenguas indígenas de manera efectiva, se amplificará la brecha digital, dejando a millones de personas sin acceso a la tecnología. La baja presencia de contenidos digitales en lenguas indígenas impide que comunidades indígenas participen plenamente en la era digital. En América Latina, la falta de acceso a servicios digitales en lenguas indígenas refuerza desigualdades estructurales.

En Perú, donde el 13.9% de la población habla una lengua indígena, el acceso a internet en comunidades rurales es un 28.5% menor que en zonas urbanas, y el acceso a electricidad es un 13.6% más bajo, lo que reduce aún más su capacidad de interactuar con tecnología avanzada. En Guatemala, solo el 29.4% de los hablantes de maya quiché tienen acceso a internet[32], lo que agrava su exclusión digital. Sin una IA adaptada a estos idiomas, estas comunidades quedarán aún más rezagadas en el acceso a información, educación y oportunidades económicas.

> Sesgos e infrarrepresentación de la cultura indígena

Los sistemas de IA entrenados con datos mayoritariamente occidentales pueden perpetuar estereotipos sobre las comunidades indígenas. Un estudio reciente reveló que los modelos generativos de IA tienden a exotizar y malinterpretar elementos culturales no occidentales, reforzando narrativas dañinas (Ghosh et al., 2024[33]). Este fenómeno podría llevar a la homogeneización de las lenguas indígenas, eliminando sus expresiones y estructuras propias.

Por ejemplo, herramientas de generación de texto o imágenes pueden reproducir estereotipos visuales o lingüísticos que refuercen una visión sesgada de estas culturas. Si los sistemas de IA no son entrenados adecuadamente con datos representativos de las lenguas y costumbres indígenas, pueden consolidar prejuicios existentes y alienar aún más a estas comunidades. La falta de diversidad en los conjuntos de datos con los que se entrenan estas tecnologías podría provocar una falta de precisión en la traducción y comprensión de las lenguas indígenas, dificultando su integración en entornos tecnológicos.

> Pérdida de la transmisión de las culturas originarias

Las lenguas indígenas se han transmitido históricamente a través de la oralidad. Sin embargo, la introducción de IA sin una integración adecuada podría desplazar la interacción directa entre generaciones. En comunidades donde el aprendizaje se da principalmente a través de la escucha y la conversación, una dependencia excesiva de la IA podría reducir el contacto con los hablantes nativos, poniendo en riesgo la continuidad lingüística y cultural.

Según la UNESCO, el 40% de las lenguas del mundo están en peligro de extinción, y muchas de ellas son indígenas. Si la IA no prioriza su digitalización y uso, muchas de estas lenguas corren el riesgo de desaparecer en una o dos generaciones. Un ejemplo crítico es el del idioma Mapuche en Chile, que, a pesar de contar con un elevado acceso a internet en su comunidad, no tiene soporte en ninguna herramienta de IA actual, lo que limita sus posibilidades de los individuos monolingües.

32 https://www.ine.gob.gt/sistema/uploads/2021/12/30/20211230192119ITIf0Taxw7mbshQNenoLw9A9K5cR4pMt.pdf

33 Ghosh, S., Venkit, P. N., Gautam, S., Wilson, S., & Caliskan, A. (2024).Do Generative AI Models Output Harm while Representing Non-Western Cultures: Evidence from A Community-Centered Approach. Recuperado de: https://arxiv.org/abs/2407.14779

> Empleo perdido y ausencia indígena en la ola de crecimiento económico asociado a la IA

Se estima que para 2030, la IA generará 170 millones de empleos y provocará la pérdida de 92 millones[34]. Sin embargo, si las comunidades indígenas no logran integrarse en esta nueva ola tecnológica, su exclusión del mercado laboral formal podría profundizarse. En Perú, la tasa de empleo informal en comunidades indígenas alcanza el 82%, lo que limita su acceso a beneficios laborales y oportunidades económicas (Ministerio de Trabajo y Promoción del Empleo, 2017[35]).

Si las herramientas digitales no se adaptan a las lenguas indígenas, las oportunidades laborales en sectores emergentes como la programación, el comercio electrónico y la educación digital seguirán siendo inaccesibles para estas comunidades. Esto perpetuaría la dependencia de trabajos informales de baja remuneración y aumentaría la vulnerabilidad económica de estos pueblos.

Si la IA no se adapta a lenguas indígenas, sectores como la educación, el comercio y la administración pública podrían excluir a estas comunidades. Algunos ejemplos concretos incluyen:

- **Educación**: Sin acceso a contenidos educativos en sus lenguas, los hablantes de lenguas indígenas tendrán menos oportunidades de capacitarse en profesiones técnicas y científicas emergentes. En América Latina, la tasa de analfabetismo entre hablantes de lenguas indígenas es hasta cinco veces mayor que la de hablantes de lenguas dominantes (INEI, 2018).
- **Comercio digital**: Las plataformas de comercio electrónico son un motor clave del crecimiento económico global, pero si las interfaces y herramientas de IA no incluyen lenguas indígenas, los emprendedores de estas comunidades no podrán acceder a mercados más amplios. En el caso de Paraguay, el 67% de la población habla guaraní[36], pero el acceso al comercio digital en esta lengua es casi inexistente.
- **Acceso a servicios públicos**: En muchas regiones, la administración pública está migrando a plataformas digitales, lo que significa que sin IA en lenguas indígenas, muchos ciudadanos no podrán realizar trámites básicos, acceder a beneficios gubernamentales o recibir información vital. En Bolivia, donde el 16.7% de la población habla aimara[37], la falta de tecnología accesible en esta lengua limita la inclusión de estas comunidades en el sistema administrativo.

Además, la ausencia de lenguas indígenas en modelos de IA impide el desarrollo de herramientas tecnológicas que puedan fortalecer la economía local, como sistemas de microfinanzas adaptados, aplicaciones de asistencia agrícola en lenguas nativas o plataformas de comercio comunitario.

> Pérdida de conocimientos ancestrales

Las lenguas indígenas no solo son un medio de comunicación, sino que también contienen vastos conocimientos sobre medicina tradicional, ecología, técnicas agrícolas sostenibles y cosmovisión. Sin una IA que pueda procesar y preservar estos conocimientos en su idioma original, el riesgo de pérdida de este saber ancestral aumenta considerablemente.

La IA tiene el potencial de documentar y sistematizar estos conocimientos en bases de datos accesibles para la investigación y el desarrollo de nuevas tecnologías. Sin embargo, si las lenguas indígenas no son una prioridad en estos desarrollos, se perderá una cantidad invaluable de información que podría contribuir a soluciones sostenibles y a la innovación en múltiples sectores.

34 El País. (2024, 18 de septiembre). La IA aportará 17,9 billones de euros a la economía mundial hasta 2030, cuando generará el 35% del PIB. Recuperado de: https://elpais.com/economia/2024-09-18/la-ia-aportara-179-billones-de-euros-a-la-economia-mundial-hasta-2030-cuando-generara-el-35-del-pib.html

35 https://www.inei.gob.pe/media/MenuRecursivo/publicaciones_digitales/Est/Lib1764/cap04.pdf

36 https://www.abc.com.py/nacionales/2023/08/25/dia-del-guarani-cuantas-personas-lo-hablan-en-paraguay/

37 https://es.wikipedia.org/wiki/Aimaras

10. DETERMINACIÓN DEL ENTORNO HABILITADOR

Programas gubernamentales de apoyo a las LL.II. existentes

> Principales programas de ámbito nacional

En los países de América Latina, los gobiernos están desarrollando esfuerzos para preservar y revitalizar las lenguas indígenas. Estas iniciativas incluyen programas educativos diseñados para fomentar su uso, promoviendo así una educación inclusiva y respetuosa hacia la diversidad lingüística de cada país.

En **Perú**, el Ministerio de Cultura, a través de la Dirección Desconcentrada de Cultura de Áncash, ha lanzado un programa en formato online destinado al aprendizaje del Quechua titulado "V edición del Curso Gratuito de Quechua Central". Este curso busca fortalecer la identidad cultural y garantizar los derechos lingüísticos de los Quechuahablantes[38].

En **Paraguay**, las autoridades del Ministerio de Educación y Ciencias (MEC) de Paraguay, junto con la Academia de la Lengua Guaraní (ALG), acordaron fortalecer la enseñanza del guaraní en el ámbito educativo y administrativo. Como parte de este compromiso, implementaron un enfoque comunicativo en guaraní en 300 escuelas dentro del "Proyecto de Jornada Escolar Extendida (JEE)". Además, impulsaron la capacitación de docentes y el desarrollo de recursos audiovisuales para apoyar la enseñanza, promoviendo así el uso y preservación del idioma en diversos contextos[39].

En **Chile**, el gobierno lanzó el programa "Las lenguas son el futuro", la cual busca revitalizar y difundir lenguas indígenas, incluyendo el Mapuche. Este esfuerzo incluye talleres de inmersión lingüística y seminarios para facilitar la preservación y el uso práctico de estas lenguas en la vida cotidiana y la administración pública[40].

38 https://www.gob.pe/institucion/cultura/campa%C3%B1as/40491-v-curso-online-gratuito-de-quechua-central

39 https://rcc.com.py/educacion-2/mec-y-academia-mejoraran-las-capacidades-de-comprension-expresion-oral-y-escrita-en-el-aprendizaje-del-idioma-guarani-2/

40 https://www.gob.cl/laslenguassonelfuturo/

En **Bolivia**, el gobierno ha incorporado las lenguas indígenas, como el Aimara, en su sistema educativo. El Ministerio de Educación anunció la inclusión de los idiomas Aimara, Quechua y Guaraní en los programas de la "franja Educa Bolivia", con el objetivo de fomentar una educación intercultural, intracultural y plurilingüe. Este programa nació como una respuesta a las necesidades educativas generadas durante la pandemia de COVID-19[41].

Además, en La Paz se implementó el curso virtual "Wiñay Aru", diseñado para promover y enseñar el idioma Aimara, con el objetivo de dignificar esta lengua indígena. Este programa, impulsado por la Delegación Municipal para el Fomento de la Interculturalidad, ofrece herramientas didácticas y audiovisuales destinadas al público en general[42].

En **Guatemala**, los programas relacionados con el Maya Quiché' incluyen la Academia de Lenguas Mayas de Guatemala (ALMG), que desarrolla materiales educativos, capacitación docente y proyectos de revitalización lingüística. Por otro lado, el Ministerio de Educación (MINEDUC) ha creado guías de autoaprendizaje en idiomas mayas, como el Quiché', dirigidas a estudiantes de preprimaria y primaria. Estos recursos están disponibles en su plataforma digital, fomentando la inclusión lingüística en la educación básica[43]. Además, el portal educativo "Aprendo en Casa y en Clase" del MINEDUC proporciona materiales en lenguas indígenas, apoyando el aprendizaje a distancia en las comunidades de hablantes de maya[44].

Por su parte, en **México**, los programas que apoyan la lengua náhuatl se canalizan a través del Instituto Nacional de los Pueblos Indígenas (INPI). Este organismo es fundamental en la implementación de políticas públicas y programas que buscan preservar y promover las lenguas indígenas, incluido el náhuatl[45]. El INPI coordina proyectos de educación intercultural bilingüe, fortalecimiento de la identidad cultural y preservación de las lenguas indígenas, mediante la capacitación docente y la creación de materiales educativos en náhuatl[46].

Finalmente, en **Brasil**, el Gobierno del Estado de Mato Grosso do Sul, a través de la Secretaría de Estado de Educación (SED), ha producido materiales didácticos en lenguas indígenas como el guaraní, con el objetivo de fomentar la alfabetización en estas lenguas entre los niños indígenas. Esta acción forma parte del programa "Alfabetiza MS Indígena", una extensión de la iniciativa MS Alfabetiza. Este programa se enfoca en promover la educación bilingüe en lenguas indígenas como el guaraní, Kaiowá, Kadiwéu y Terena, y es parte de los esfuerzos estatales para preservar y promover las lenguas indígenas en la región[47].

> Principales programas regionales e internacionales

El enfoque regional hacia la preservación de las lenguas indígenas, también se refleja en programas multipaís que fomentan la colaboración internacional para promover el uso, desarrollo y conservación de las lenguas indígenas, contribuyendo a su sostenibilidad cultural y social. Estas iniciativas tienen como objetivo no solo revitalizar lenguas marginadas o en riesgo de desaparición, sino también fortalecer los derechos culturales y lingüísticos de las comunidades indígenas.

41 https://educabolivia.com/
42 https://www.bolivia.com/tecnologia/visionarios/sdi/85847/presentan-curso-virtual-de-aymara-para-dignificar-y-aprender-la-lengua
43 https://digebi.mineduc.gob.gt/digebi/categoria-articulo/materiales-educativos/idioma/kiche/
44 https://aprendoencasayenclase.mineduc.gob.gt/index.php/guias-de-autoaprendizajeid320/
45 https://www.gob.mx/inpi
46 https://www.gob.mx/inpi/articulos/nahuatlahtolli-lengua-nahuatl-libro-ilustrado
47 https://agenciadenoticias.ms.gov.br/para-preservar-cultura-governo-do-estado-desenvolve-material-didatico-em-linguas-indigenas/

Entre estas acciones destaca la iniciativa **Rising Voices**, impulsada por la organización Global Voices. Este proyecto apoya a creadores digitales indígenas en sus esfuerzos por revitalizar y promover las lenguas indígenas y otras lenguas minoritarias[48]. Por su parte, el **Instituto Iberoamericano de Lenguas Indígenas (IIALI)**, establecido tras la XXVII Cumbre Iberoamericana en 2021, se centra en fomentar el uso, la conservación y el desarrollo de las lenguas indígenas habladas en América Latina y el Caribe. Este programa colabora con las sociedades indígenas y los Estados para garantizar el respeto y ejercicio de los derechos culturales y lingüísticos, reconociendo el valor de estas lenguas como patrimonio vivo de la región[49]. En línea con estos esfuerzos, el **Fondo para el Desarrollo de los Pueblos Indígenas de América Latina y el Caribe (FILAC)** refuerza la importancia de las lenguas indígenas en la preservación cultural y la cohesión social. Con presencia en países como Argentina, Bolivia, Chile, Brasil y Guatemala, entre otros[50]. Juntas, estas acciones constituyen un esfuerzo global por proteger y revitalizar un patrimonio lingüístico invaluable.

Iniciativas de ONG's y activismo

> Programas de apoyo de ONG's

La preservación y promoción de las lenguas indígenas ha sido un tema abordado tanto por organizaciones no gubernamentales (ONGs) como por empresas del sector privado. Estas entidades han implementado diversas iniciativas que van desde la defensa legal hasta la integración de lenguas indígenas en plataformas digitales y servicios comerciales.

Existen varias ONGs que cuentan con programas de apoyo a la preservación de la lenguas indígenas, tanto desde un enfoque de defensa legal como desde iniciativas prácticas para el uso y conservación de estas lenguas. Entre ellas, destacamos la ONG **Rising Voices**, una iniciativa que se enfoca en el apoyo a comunidades de hablantes de lenguas indígenas que están aprovechando el Internet y otras tecnologías digitales para promover sus lenguas en los espacios digitales . Esta organización proporciona herramientas para el náhuatl, maya, guaraní, Quechua[51]. Además, su colaboración con la **UNESCO**, Organización de las Naciones Unidas para la Educación, la Ciencia y la Cultura, ha tenido como resultado la publicación del instrumento "Iniciativas digitales para lenguas indígenas" (2023), a través del cual se promueven bases para la preservación, el resurgimiento y la promoción de lenguas indígenas en ocho enfoques: facilitar, multiplicar, normalizar, educar, recuperar, imaginar, defender y proteger[52].

Otra organización es **Cultural Survival**, la cual defiende los derechos de los Pueblos Indígenas de diversas regiones del mundo y apoya la autodeterminación, las culturas y la resiliencia política de éstos, desde 1972. Esta defensoría legal se enfoca en la preservación de sus lenguas y culturas[53]. También está el **Fondo para el Desarrollo de los Pueblos Indígenas de América Latina y el Caribe (FILAC)**, organismo internacional de derecho público creado en 1992, que apoya los procesos de autodesarrollo de pueblos, comunidades y organizaciones indígenas de la región, y promueve el Buen Vivir-Vivir Bien como una alternativa para garantizar la sostenibilidad ambiental, el respeto de los derechos fundamentales del ser humano, y el diálogo entre los principales actores del desarrollo indígena, que son los pueblos Indígenas, gobiernos, sociedad civil, academia, empresarios y otros.

En este mismo marco, **The Amazon Conservation Team**, es una organización que colabora con comunidades indígenas y otras comunidades locales para proteger los bosques tropicales y fortalecer la cultura tradicional. Trabaja con comunidades indígenas en la Amazonía para la conservación de la

48 https://unesdoc.unesco.org/ark:/48223/pf0000388256

49 https://www.iiali.org/objetivos-del-iiali/

50 https://www.filac.org/

51 https://rising.globalvoices.org/lenguas/

52 https://unesdoc.unesco.org/in/documentViewer.xhtml?v=2.1.196&id=p::usmarcdef_0000388256&file=/in/rest/annotationSVC/DownloadWatermarkedAttachment/attach_import_955fff98-dfc2-43d5-b353-c83691109012%3F_%3D388256spa.pdf&locale=es&multi=true&ark=/ark:/48223/pf0000388256/PDF/388256spa.pdf#INICIATIVAS.indd%3A.65862%3A857

53 https://www.culturalsurvival.org/es/node/2

biodiversidad y la cultura, incluyendo la preservación de lenguas como el Guaraní y el Tupi-Guaraní[54]. Igualmente, la **International Work Group for Indigenous Affairs (IWGIA)** es una ONG que trabaja en la defensa de los derechos de los pueblos indígenas a nivel global, incluyendo la preservación de sus lenguas y cultura. Esta organización se encuentra desarrollando una iniciativa llamada Navegador Indígena que directamente analiza la brecha de aplicación práctica mediante la recopilación y el uso de datos en acciones de incidencia y diseño de proyectos para comunidades y organizaciones indígenas[55].

> Iniciativas open source alrededor de las LL.II.

En relación con la preservación y promoción de las lenguas indígenas, es importante resaltar que, además de los programas nacionales y multipaís, han surgido varias iniciativas open source dedicadas a desarrollar recursos lingüísticos en lenguas indígenas. Existen tanto programas open source como aplicaciones de libre acceso que facilitan el desarrollo de recursos lingüísticos y la promoción de lenguas nativas. Los esfuerzos recientes se remontan al 2010, cuando se realizó la propuesta de software libre **OpenBiblio distribución náhuatl**, el cual permitiría automatizar todas las actividades realizadas en una biblioteca comunitaria para dotar de tecnología y de los manuales para su administración, sin embargo no se encontró la implementación del sistema[56].

En México, el **Proyecto Heliox (2014)**, se presentó como un sistema operativo que estaba diseñado para orientar a los usuarios a ejecutar aplicaciones, abrir archivos y navegar en sitios web mediante textos y mensajes de voz que aparecen en el lenguaje indígena seleccionado; contó con la participación del Instituto Nacional de Lenguas Indígenas de México, institución en la cual se tradujo Heliox del español al castellano que se habla en México y a lenguas indígenas como el Maya, Náhuatl y Mixe. El sistema no perduró ni tuvo actualizaciones desde su creación[57].

También podemos encontrar iniciativas como **Firefox en Guaraní, Maya Quiché', Náhuatl, Quechua**, que apuntan hacia la inclusión digital de diferentes lenguas indígenas con el fin de promover su uso y preservación a lo largo del tiempo[58]. Mozilla Nativo está a cargo de la traducción y localización del navegador Mozilla Firefox a lenguas nativas donde se busca la inclusión digital de grupos nativos, con el fin de que las lenguas nativas se sigan usando y se preserven a lo largo del tiempo[59].

Otra iniciativa es **Open Language Archives Community (OLAC)**, diseñada para facilitar la búsqueda y el acceso a bases de datos en línea de recursos lingüísticos, promoviendo la interoperabilidad entre archivos de lenguas. Este recurso promueve prácticas de open source y colaboración pero es importante mencionar que no hay información sobre si es un software open source en sí mismo[60]. Por otro lado, la herramienta libre y de código abierto **EUDICO Linguistic Annotator (ELAN)** es utilizada para la anotación de datos lingüísticos multimedia y es ampliamente requerida en trabajos de campo lingüísticos y en la documentación de lenguas como el Náhuatl[61].

La empresa y la conservación de lenguas indígenas

> La empresa de consumo y el uso de las LL.II.

En el ámbito privado, diversas empresas están adoptando iniciativas para preservar y promover las lenguas indígenas. En varios países de América Latina, estas compañías han comenzado a incorporar

54 https://www.amazonteam.org/
55 https://iwgia.org/es/
56 http://www.udgvirtual.udg.mx/apertura//index.php/apertura/article/view/132/134
57 https://www.semana.com/indigenas-mexicanos-accederan-la-tecnologia-en-su-propia-lengua/396665-3/
58 https://www.mozilla.org/gn/
59 https://mozillanativo.org/2021/webinar-localizacion-de-firefox-en-lenguas-indigenas.html
60 http://www.language-archives.org/
61 https://archive.mpi.nl/tla/elan

lenguas indígenas en sus acciones de inclusión, facilitando el acceso a servicios en estas lenguas y apoyando a las comunidades que las hablan.

El cálculo de los porcentajes por país en cuanto a iniciativas relacionadas con lenguas indígenas se realizó tomando como referencia el **Top 15 del Ranking Merco Empresas 2024** para la mayoría de los países. Sin embargo, en el caso de Paraguay, debido a su ausencia en el ranking de Merco, se utilizó la lista de las **empresas que más ganan según Forbes** y también se consideró un top 15. En ambos casos, se identificó cuántas de esas empresas han implementado alguna iniciativa relevante relacionada con las lenguas indígenas.

Aunque algunas de las principales empresas del ranking están liderando en este ámbito, también se destacó el trabajo de empresas fuera del top que han hecho esfuerzos significativos.

En Perú, al considerar las empresas del top 15 del Ranking Merco Empresas 2024, se encontró que el 20% de ellas ha implementado iniciativas relacionadas con el uso del Quechua. Por ejemplo, el **Banco de Crédito del Perú (BCP**) ha implementado el Quechua en más de 2,300 cajeros automáticos, permitiendo realizar retiros de dinero en esta lengua. Esta innovación beneficia a cerca de 4 millones de Quechuahablantes[62]. Otro banco relevante es **BBVA**, que, a través de su proyecto educativo "Aprendemos juntos 2030", dio un paso significativo al incluir el Quechua en la transmisión de los episodios de este programa, gracias a una alianza con TV Perú[63].

Además, el enfoque inclusivo también se extiende a los servicios públicos de telecomunicaciones. A pesar de no figurar en el top 15, es importante destacar que principales operadoras como **Claro, Bitel y Entel** han introducido contratos en Quechua, Aimara, ashaninka y shipibo-konibo, siendo Claro el líder en la emisión de contratos en Quechua[64].

En México, el 20% de las empresas del top 15 del Ranking Merco Empresas ha llevado a cabo campañas relacionadas con la lengua Náhuatl. Entre las destacadas se encuentran **Google**, que incorporó esta lengua en su plataforma de traducción[65], y **Grupo Modelo**, con su campaña "Tatuajes Originarios", un homenaje a las 68 lenguas indígenas de México que pone énfasis en la preservación del Náhuatl y otras lenguas en peligro[66].

Fuera del Top 15, empresas como CEMEX y AT&T México han desarrollado proyectos de preservación de lenguas indígenas. La empresa **CEMEX** participó en la producción de 14 cápsulas educativas e informativas en lenguas nativas como el náhuatl, para promover la cultura de autoprotección y resiliencia en comunidades indígenas[67]. **AT&T México** puso en marcha una iniciativa en alianza con Fundación NEMI y los locutores de Radio Huayacocotla, para desarrollar una serie de cápsulas en tres lenguas indígenas: náhuatl, otomí y tepehua, que tocan temas como: qué es la huella en Internet, e-derechos, cómo cuidarse en Internet, cómo prevenir el ciberacoso, entre otros[68].

En Guatemala, el 7% del Top 15 ha realizado alguna iniciativa en torno a la lengua maya Quiché, destacando Walmart , que ha introducido cajas de autopago con mensajes de audio en Kaqchikel, Quiché' y Q'eqchi.

62 https://forbes.pe/tecnologia/2024-07-24/el-bcp-incorporo-el-idioma-quechua-en-sus-cajeros-automaticos

63 https://www.bbva.com/es/pe/sostenibilidad/por-primera-vez-bbva-peru-presenta-aprendemos-juntos-2030-kids-en-quechua/

64 https://www.gob.pe/institucion/osiptel/noticias/961621-dia-de-las-lenguas-originarias-usuarios-de-telecomunicaciones-pueden-acceder-a-contratos-cortos-en-quechua-aimara-ashaninka-y-shipibo-kon

65 https://elpais.com/mexico/2024-06-30/maya-zapoteco-nahuatl-y-mas-de-100-idiomas-se-suman-a-google-translate-en-su-mayor-expansion-en-la-historia.html

66 https://es.rollingstone.com/tatuajes-originarios-la-campana-de-cerveza-victoria-que-preserva-la-lengua-nahuatl/

67 https://www.cemexmexico.com/-/cemex-promueve-cultura-de-prevencion-de-desastres-naturales-en-comunidades-indigenas

68 https://www.att.com.mx/noticias/att-civismo-digital-incluyente.html

Aunque fuera del top 15, el **Banco de Desarrollo Rural (Banrural)** implementó un programa para instalar cajeros automáticos multilingües con operaciones en español y otras lenguas nativas como el maya Quiché, lo cual acercó a muchos guatemaltecos al acceso a la banca[69]. Así mismo, a pesar de no contar con un programa específico para la preservación de la lengua maya, la **Fundación Patrimonio Cultural y Natural Maya (PACUNAM)**, conformada por empresas comprometidas como **Cementos Progreso, Cervecería Centro Americana S.A., Walmart de México y Centroamérica, Citibank de Guatemala, Banco Industrial, Claro**, entre otras, tiene como objetivo principal apoyar el desarrollo sostenible coordinando esfuerzos y facilitando recursos destinados a identificar, liderar y promover proyectos enfocados en la protección, preservación y rescate del patrimonio cultural y natural de Guatemala[70].

En Brasil, el 13% de las empresas del top 15 del Ranking Merco Empresas ha implementado iniciativas relacionadas con la lengua guaraní. Un ejemplo destacado es Google, que desde 2022 ha incorporado el guaraní a su plataforma de traducción. Además, **Magazine Luiza**, una empresa de comercio minorista, se ha comprometido a apoyar la diversidad cultural de Brasil, apoyando la inclusión de comunidades indígenas y la preservación de sus lenguas y costumbres[71].

Fuera del top 15, **Petrobras**, a través de su programa **"Petrobras Cultural"** ha apoyado programas de responsabilidad social dirigidos a comunidades indígenas, pero estos proyectos no se centran específicamente en la preservación del guaraní[72].

En Chile, ninguna de las empresas del top 15 del Ranking Merco ha realizado iniciativas directas relacionadas con la lengua Mapuche. Sin embargo, se ha encontrado que **Enel Generación** (que no figura en el top 15), en colaboración con las comunidades Pehuenche de Alto Biobío, inauguró la Escuela Intercultural de Quepuca Ralco. Este proyecto incorporó elementos culturales Mapuche Pehuenche en su diseño arquitectónico y educativo. A pesar de no estar directamente relacionado con la lengua Mapuche, el enfoque ha estado en respaldar a esta comunidad mediante este proyecto de infraestructura y desarrollo[73].

> Las Big Tech y sus iniciativas en LL.II.

Así como existen empresas privadas de diversos sectores que muestran su interés por la preservación de culturas indígenas, esto ha atraído la atención de empresas tecnológicas que reconocen el valor del patrimonio lingüístico y cultural como un aspecto fundamental de la diversidad global. Estas compañías han desarrollado herramientas y colaboraciones innovadoras para apoyar la inclusión digital, la enseñanza y la preservación de estas lenguas.

Entre dichas iniciativas, cabe destacar aquéllas que tienen un objetivo social y de soporte para comunidades y lenguas en riesgo. Por ejemplo, a través de la iniciativa **AI For Good Lab**[74], Microsoft plantea aplicaciones de IA desarrolladas en torno a la sostenibilidad, la acción humanitaria y la salud. Entre sus objetivos y funciones se encuentran los siguientes:

Identificar comunidades vulnerables en riesgo.

- **Asociado al clima**: Catástrofes naturales, inundaciones, sequías o subidas del nivel del mar.

69 https://www.prensalibre.com/economia/instalacion-cajeros-idioma-maya-crece_0_5111748831-html/

70 https://es.wikipedia.org/wiki/Fundaci%C3%B3n_Patrimonio_Cultural_y_Natural_Maya

71 https://jcmagazine.com/inclusion-diversidad-e-igualdad-laboral/

72 https://www.petrobras.com.br/cultural/selecoes-publicas-culturais

73 https://www.enel.cl/es/conoce-enel/prensa/press-enel-generacion/d202405-comunidades-pehuenche-de-alto-biobio-y-enel-generacion-inauguraron-la-escuela-intercultural-de-quepuca-ralco.html

74 AI For Good Lab, de Microsoft: https://www.microsoft.com/en-us/research/group/ai-for-good-research-lab/

- **Asociado al abastecimiento y accesibilidad**: comunidades en riesgo de malnutrición, en riesgo de exclusión, de carencias educativas o de elementos básicos como la ropa.
- **Asociado al entorno natural**: Riesgo extremo de deforestación, biodiversidad, agua potable, etc.

Acelerar la incorporación sanitaria de carácter público en torno a comunidades vulnerables.

- Chatbots para seguimientos clínicos, atención, control de adicciones, etc.
- Mejorando la diagnosis y detección temprana de afecciones.
- Valorando tanto individuo como población (tendencias según comunidades o áreas).

Otras iniciativas plantean la creación de ecosistemas para idiomas con poca representación en el mundo digital, como es el caso de **Ellora**[75] (*Enabling Low Resource Languages*), también de Microsoft, propuesto inicialmente para territorios de India.

A este tipo de propuestas se suman otras grandes empresas del sector, como Google con **Woolaro**[76], una iniciativa que cubre aplicaciones para describir imágenes en 17 lenguas en peligro, IBM con **Sustainability accelerator**[77] (iniciativa para proporcionar a organizaciones sin fines de lucro y agencias gubernamentales soluciones de IA para apoyar a comunidades vulnerables en todo el mundo).

Otras iniciativas regionales

En **Perú** (**Quechua**), destaca **Microsoft** donde se menciona que en noviembre del 2022, la compañía junto al Ministerio de Cultura se unieron para promover el uso de Windows y Office 365 en la lengua Quechua chanka a través del anuncio de una versión actualizada[78]. Por otra parte, la compañía **Google** anunció que el Traductor de Google sumará 24 nuevos idiomas, entre ellos, el Aimara y el Quechua[79].

En **Bolivia** (**Aimara**), se mencionó que **Traductor de Google** sumará 24 nuevos idiomas, entre ellos, el Aimara.

En **Paraguay** (**Guaraní**), la Fundación Yvy Marãe'ỹ y el Instituto Técnico Superior de Estudios Culturales y Lingüísticos Yvy Marãe'ỹ, lanzaron la Plataforma de Enseñanza de Lengua y Cultura Guaraní – Guarani Ñe'ẽ ha Arandupy Oñembo'e hagua Yvytu Pepo rehe Renda[80].

En **Chile** (**Mapuche**), el Centro de Desarrollo de Tecnologías de Inclusión, CEDETi UC presentó Mapuche Mew, una herramienta para transmitir cultura y conocimiento a través de la lengua[81]. Así mismo, **Apple** ha incluido teclados en Mapuche en sus dispositivos iOS (iPhone, iPad, Mac, etc.), lo que permite a los usuarios escribir y comunicarse en Mapuche de manera más sencilla[82].

75 Ellora, de Microsoft: https://news.microsoft.com/source/latam/features/ia/proyecto-de-microsoft-research-ayuda-a-los-idiomas-a-sobrevivir-y-prosperar/

76 Woolaroo, de Google: https://artsandculture.google.com/project/woolaroo

77 Sustainability accelerator, de IBM: https://www.ibm.com/impact/initiatives/ibm-sustainability-accelerator

78 https://www.gob.pe/institucion/cultura/noticias/665291-ministerio-de-cultura-windows-11-y-office-365-en-quechua-chanka-tras-alianza-estrategica-con-microsoft

79 https://elcomercio.pe/tecnologia/actualidad/google-incorpora-el-quechua-y-el-aimara-a-su-traductor-mexico-usa-espana-noticia/

80 https://yvymaraey.com.py/plataforma-de-ensenanza-de-lengua-guarani-fue-lanzado-hoy/

81 https://www.uc.cl/noticias/nuevo-software-para-aprender-mapudungun-y-cultura-mapuche-fue-dado-a-conocer-en-villarrica/

82 https://edition.cnn.com/tecnologias/lanzan-aplicacion-para-iphone-que-ensena-mapudungun_20120210/

En **Guatemala** (**Maya Quiché**) se conoce que **Google Translate** ha incorporado en su sistema el idioma maya quiché y ha mostrado su apoyo al idioma a través de Google Arts & Culture[83].

En **México** (**Náhuatl**), se ha señalado que a mediados de junio del 2020, se ha incorporado el Náhuatl dentro de las lenguas indígenas que viene trabajando **Google**[84].

Finalmente, en **Brasil** (**Tupi-Guaraní**), la incorporación de Google Translate comenzó en 2022[85].

A pesar de que se identificaron esfuerzos e intereses del sector privado para preservar las lenguas indígenas mediante la inclusión de traductores, a día de hoy no hay empresas de Inteligencia artificial o vinculadas a este sector que se encuentren trabajando en una IA desarrollada exclusivamente en alguna de las lenguas indígenas de este estudio.

83 https://animalpolitico.com/tendencias/ciencia-tecnologia/traductor-google-incorpora-lenguas-indigenas-mexico
84 https://www.filac.org/reportes/google-translate-incorpora-lenguas-indigenas-de-mexico-a-su-plataforma/
85 https://www.h2foz.com.br/es/fronteira/atualizacao-do-google-tradutor-tera-guarani-quechua-e-aimara/

11. ESTRATEGIAS DE INCLUSIÓN TECNOLÓGICA

A lo largo de la sección 8, vimos cuáles eran los factores clave que determinan el nivel en que la inteligencia artificial (IA) es capaz de comprender y expresarse en una lengua. Uno de los factores más determinantes es la cantidad de contenido textual disponible durante la fase de entrenamiento, que permite a los modelos de IA aprender las estructuras, el vocabulario y las reglas gramaticales de cada idioma. Esta escasez de datos es especialmente crítica en el caso de las lenguas indígenas americanas, donde el acceso a textos digitalizados y documentos escritos es considerablemente más limitado en comparación con lenguas mayoritarias como el inglés o el español. Esto sitúa a dichas lenguas en una desventaja significativa en el desarrollo de tecnologías basadas en IA.

Al profundizar en el proceso de entrenamiento de la IA orientada al lenguaje, también observamos que no existen técnicas específicas que dependan exclusivamente de un idioma. Las herramientas y los algoritmos de procesamiento del lenguaje natural son los mismos, tanto si se aplican al inglés como a otras lenguas. Entonces, ¿qué explica la diferencia de rendimiento entre unas lenguas y otras? La respuesta reside en la disponibilidad de datos. Los modelos actuales se entrenan mejor cuando cuentan con grandes volúmenes de información textual, algo que las lenguas indígenas no suelen poseer debido a factores históricos, sociales y tecnológicos relacionados con su menor presencia en entornos digitales y escritos.

En consecuencia, las estrategias que podrían tener el mayor impacto para mejorar la capacidad de la IA de comprender y expresarse en lenguas indígenas son aquellas que fomenten la creación, digitalización y documentación de contenidos en estos idiomas. Esto incluye la recopilación de materiales existentes, la conversión de textos a lenguas indígenas, el desarrollo de recursos lingüísticos y el registro de expresiones orales en formato digital. Sin embargo, aunque estas acciones son fundamentales, no son las únicas medidas posibles. Existen otras estrategias complementarias que también pueden contribuir a reducir la brecha tecnológica, incluyendo la colaboración con las comunidades hablantes, el desarrollo de herramientas adaptativas y la promoción de políticas que impulsen el desarrollo de la IA en contextos multilingües.

A continuación, se presentarán estas estrategias de manera detallada, con el objetivo de delinear un enfoque integral que permita mejorar el nivel de expresión de la IA en lenguas indígenas y avanzar hacia una mayor equidad lingüística en el ámbito tecnológico.

21 estrategias para mejorar el rendimiento de la IA en LL.II.

> Impulsar la comunicación digital en lenguas indígenas

Uno de los pilares fundamentales para reducir la brecha tecnológica en lenguas indígenas es el fomento activo de su uso en entornos digitales. Esto no solo genera contenido valioso para entrenar modelos de IA, sino que también refuerza la vitalidad y el prestigio social de estas lenguas. A continuación, se describen tres estrategias que apuntan a este objetivo.

1. Reconocimiento de influencers y creadores de contenido en lenguas indígenas

Las redes sociales se han convertido en un espacio clave para la visibilidad y la comunicación a nivel global. Reconocer e impulsar a influencers y creadores de contenido que utilizan lenguas indígenas como su principal medio de comunicación es una estrategia con gran potencial. Estos creadores, al conectar con sus comunidades y audiencias globales, no solo dan visibilidad a sus lenguas, sino que también generan contenido moderno y relevante que atrae a nuevas generaciones. Campañas de promoción, premios o colaboraciones con plataformas como YouTube, TikTok o Instagram podrían aumentar la influencia y el alcance de estos líderes digitales.

2. Formación en internet y redes sociales a través de programas de alfabetización

Muchas comunidades indígenas ya cuentan con iniciativas orientadas a la alfabetización en sus propias lenguas. Aprovechar estos programas para integrar módulos dedicados a la formación en internet y medios digitales es una estrategia eficiente para aumentar la presencia digital de las lenguas indígenas. A través de talleres prácticos, los participantes podrían aprender a crear y compartir contenidos digitales, lo que contribuiría a la producción constante de material en sus idiomas, desde blogs y foros hasta videos y publicaciones en redes sociales.

3. Desarrollo de foros y plataformas temáticas en lenguas indígenas

Crear espacios digitales donde se recojan materiales y conocimientos tradicionales en lenguas indígenas es otra estrategia crucial. Estos foros y plataformas pueden convertirse en repositorios vivos de saberes ancestrales, relatos históricos, costumbres, canciones, recetas y más. No solo preservan el patrimonio cultural, sino que también estimulan el uso continuo del idioma en contextos contemporáneos. Además, este tipo de iniciativas fomenta la colaboración intergeneracional, involucrando tanto a jóvenes creadores digitales como a portadores del conocimiento tradicional. Impulsar estas estrategias de comunicación digital no solo ayuda a nutrir los modelos de IA con más contenido en lenguas indígenas, sino que también promueve el orgullo lingüístico, la innovación tecnológica local y la inclusión digital de comunidades históricamente marginadas. Este enfoque integral sentará las bases para un mayor reconocimiento y sostenibilidad de las lenguas indígenas en el entorno digital.

> Preservar y ampliar el contenido de lenguas indígenas en línea

La preservación y ampliación del contenido en lenguas indígenas es crucial para asegurar que estas lenguas tengan una presencia duradera en el entorno digital. Aunque es vital fomentar la creación de nuevos materiales, también es necesario proteger los recursos ya existentes, evitando su desaparición o inaccesibilidad. A continuación, se detallan estrategias orientadas a la conservación digital y al desarrollo de contenidos en lenguas indígenas.

1. Programas de apoyo al mantenimiento de recursos digitales existentes

Durante el análisis realizado, se identificaron múltiples colecciones de recursos en lenguas indígenas —textos, archivos multimedia y bases de datos— que alguna vez estuvieron disponibles en línea, pero que se han perdido con el tiempo. Esto se debe a dominios expirados, servicios de alojamiento web descontinuados o falta de soporte técnico. Por ejemplo, en 2024 fue ciertamente reconocida la labor de un chileno de generar el primer traductor online español-mapuche llamado "Kulturalwaywen", sin embargo, en apenas unos años ya refleja errores de mantenimiento o conectividad[86]. Para contrarrestar esta tendencia, es fundamental desarrollar programas específicos que proporcionen servicios de hosting permanente para estos recursos, así como acceso a personal técnico que pueda garantizar su mantenimiento y actualización. De este modo, se evita la pérdida de información valiosa y se asegura su disponibilidad para futuras generaciones.

2. Digitalización y conservación de archivos físicos

En muchas comunidades indígenas, gran parte de su conocimiento cultural y lingüístico se ha transmitido de manera oral o se encuentra almacenado en archivos físicos, como manuscritos, grabaciones antiguas o libros de edición limitada. Impulsar proyectos de digitalización de estos materiales puede ampliar de manera significativa la cantidad de contenido accesible en línea. Además, el uso de formatos estándar y repositorios públicos garantiza que estos materiales puedan ser integrados en bases de datos internacionales de acceso abierto, aumentando su visibilidad.

3. Acceso a fondos para proyectos de documentación

Asegurar recursos económicos para iniciativas de documentación y conservación es una pieza clave en este proceso. A través de colaboraciones con instituciones académicas, ONG y gobiernos, se pueden crear fondos dedicados a financiar la recopilación de materiales lingüísticos, la producción de contenido en formato digital y el desarrollo de tecnologías de almacenamiento seguro. Los programas de financiación deben incluir tanto el aspecto técnico como el cultural, respetando los protocolos de consulta y participación comunitaria.

Estas estrategias permitirán no solo preservar los contenidos ya existentes, sino también garantizar que estos sigan creciendo y evolucionando con el tiempo. La continuidad digital es esencial para mantener vivas las lenguas indígenas, así como para fortalecer su representación y relevancia en el mundo tecnológico.

> Normalizar el uso de lenguas indígenas y reducir su la fragmentación

Un factor que complica la preservación y el desarrollo tecnológico de las lenguas indígenas es la gran diversidad de variedades dialectales que existen dentro de cada una. Esta fragmentación limita el alcance de los recursos disponibles, pues los materiales desarrollados en una variedad específica no siempre son comprensibles o aceptados por otras comunidades hablantes. Reducir esta fragmentación mediante acuerdos de normalización lingüística y estrategias de cooperación es clave para lograr un mayor impacto y eficiencia en los esfuerzos de preservación y digitalización.

1. Impulsar acuerdos para la unificación o convergencia dialectal

La estandarización de lenguas con múltiples variedades dialectales es una estrategia utilizada con éxito en diversos contextos. Un ejemplo notable es el del euskera, una lengua ancestral que presenta numerosas variedades dialectales. En la década de 1960, se creó el euskera batúa o euskera unificado, un estándar común que permitió que la lengua fuera utilizada en la educación, los medios de comunicación y otros contextos formales. Este proceso no significa eliminar las variedades locales, sino establecer un conjunto mínimo de normas que facilite la comunicación entre comunidades y que sirva como referencia para el desarrollo de materiales educativos, digitales y administrativos.

86 https://kutralwaywen.cl/

En el caso de las lenguas indígenas, alcanzar acuerdos similares puede aumentar considerablemente la eficacia de los recursos disponibles. Por ejemplo, un diccionario, una aplicación de aprendizaje o un corpus digital elaborado en una variedad estándar podría ser utilizado por una mayor cantidad de hablantes, maximizando el impacto de estos proyectos.

2. Respetar la diversidad en un marco de estandarización flexible

Aunque es importante llegar a acuerdos mínimos, es fundamental que estos procesos de normalización respeten la diversidad lingüística y cultural de las comunidades. Esto implica diseñar estándares flexibles que puedan adaptarse a distintos contextos. Por ejemplo, los recursos digitales pueden incorporar opciones que permitan personalizar la interfaz o el contenido según la variedad dialectal preferida. De esta forma, se mantiene un equilibrio entre la necesidad de unificar esfuerzos y la importancia de conservar las expresiones locales propias de cada comunidad.

3. Fomentar la formación de expertos locales en lingüística y tecnología

Para que estos procesos sean sostenibles, es importante formar expertos locales que puedan liderar y gestionar la estandarización de sus lenguas. Estos especialistas deben recibir formación en lingüística, tecnología y gestión de recursos digitales para que puedan contribuir a la creación de estándares y garantizar la adaptación tecnológica de sus lenguas en el tiempo.

Con estas estrategias, la normalización lingüística en lenguas indígenas puede reducir significativamente la fragmentación y permitir una mayor inclusión tecnológica. Esto, a su vez, potenciará la producción de contenido digital, mejorará el acceso a herramientas de IA y reforzará el prestigio social de las lenguas indígenas en sus respectivos territorios y más allá.

> Impulsar el desarrollo tecnológico de herramientas habilitadoras

El desarrollo de herramientas tecnológicas adaptadas a las lenguas indígenas es un factor importante para reducir la brecha digital. Estas herramientas no solo favorecen el procesado automático del contenido digital disponible, sino que también hacen más accesible la tecnología a las comunidades, especialmente a aquellas personas que no están alfabetizadas. Fomentar el desarrollo de este tipo de sistemas, puede acelerar significativamente la integración de las lenguas indígenas en el entorno digital.

1. Desarrollo de sistemas text-to-speech y speech-to-text

La creación de herramientas que conviertan texto en voz y voz en texto para lenguas indígenas tiene un doble propósito. Por un lado, estas tecnologías pueden generar nuevos contenidos digitales, enriqueciendo las bases de datos lingüísticas que los modelos de IA utilizan para aprender. Por otro lado, estos sistemas facilitan el acceso a la tecnología para personas que no saben leer o escribir en su lengua nativa, permitiéndoles interactuar con aplicaciones, servicios educativos y medios de comunicación de manera más intuitiva. Estas herramientas son especialmente útiles en contextos donde la oralidad sigue siendo la principal forma de transmisión cultural.

2. Desarrollo de herramientas para la identificación automática de lenguas en textos

Desarrollar software capaz de inferir automáticamente en qué lengua está escrito un texto es esencial para mejorar la eficiencia en la gestión de grandes volúmenes de datos. Estas herramientas permiten clasificar y filtrar contenidos digitales, seleccionando solo aquellos que pertenezcan a una lengua específica. Esto resulta especialmente útil en el caso de las lenguas indígenas, donde los textos disponibles suelen estar dispersos en plataformas generales o mezclados con materiales en otros idiomas. Con una mayor capacidad para identificar textos relevantes, se pueden crear corpus lingüísticos más sólidos y precisos, lo que mejora el entrenamiento de los modelos de IA y optimiza el desarrollo de otras tecnologías lingüísticas.

3. Mantenimiento y desarrollo de diccionarios y traductores automatizados

Los diccionarios digitales y los traductores automáticos son recursos esenciales para el acceso a las lenguas indígenas, tanto para hablantes nativos como para personas externas que desean aprenderlas o investigarlas. Estos recursos incrementan la utilidad de internet para aquellas personas que sólo hablan en su lengua indígena y ayudan a reducir las barreras lingüísticas, permitiendo que terceros, como desarrolladores, educadores o investigadores, trabajen con materiales en lenguas indígenas. Sin embargo, para que estas herramientas sean efectivas, es fundamental apoyar su mantenimiento, actualización y expansión. Esto implica agregar nuevo vocabulario, mejorar las traducciones automáticas y garantizar la precisión de los términos especializados en diversos contextos.

Con el impulso adecuado, estas herramientas habilitadoras no solo fortalecerán la representación de las lenguas indígenas en el mundo digital, sino que también contribuirán a mejorar la calidad de vida de sus hablantes, al abrirles nuevas oportunidades educativas, sociales y económicas.

> Aprovechar iniciativas de inclusión lingüística de grandes marcas del consumo

Las grandes empresas que ofrecen servicios masivos al público han comenzado a desarrollar iniciativas para incluir lenguas indígenas en las interfaces de sus productos, como aplicaciones móviles, dispositivos electrónicos y plataformas digitales. Este proceso de inclusión no solo mejora la accesibilidad para millones de hablantes, sino que también genera oportunidades para aumentar la demanda y el desarrollo de tecnologías basadas en inteligencia artificial (IA) que sean capaces de comprender y expresarse en estos idiomas.

1. Establecer alianzas con grandes empresas comprometidas con la inclusión lingüística

Grandes empresas tecnológicas (como Microsoft o Google) y de consumo masivo (como BCP, Modelo o BBVA) han dado pasos hacia la incorporación de lenguas indígenas en sus productos. Entre estas iniciativas se encuentran opciones de interfaz en lenguas indígenas en aplicaciones de mensajería, plataformas de redes sociales o asistentes virtuales. Establecer vínculos con estas empresas y colaborar en el diseño de estrategias conjuntas puede acelerar el proceso de inclusión lingüística. Por ejemplo, las empresas podrían recibir asesoría técnica y cultural para mejorar la precisión de sus traducciones automáticas y desarrollar nuevas funciones conversacionales basadas en IA.

2. Proponer la integración de tecnologías de IA conversacional

Una estrategia clave es sugerir a estas empresas que incorporen asistentes virtuales y chatbots en lenguas indígenas. Estas aplicaciones permiten a los usuarios interactuar con los servicios mediante comandos de voz o texto, mejorando significativamente la experiencia de usuario. Sin embargo, para que estos servicios funcionen de manera efectiva, es necesario desarrollar modelos de IA con capacidades avanzadas en procesamiento de lenguaje natural para dichas lenguas. Este tipo de iniciativas no solo tiene un impacto inmediato en la accesibilidad, sino que también incrementa la demanda de más contenido digital y mejores herramientas lingüísticas.

3. Generar demanda de recursos y entrenamiento lingüístico

La incorporación de lenguas indígenas en servicios populares genera una presión positiva sobre el ecosistema tecnológico, impulsando la necesidad de mejorar los recursos lingüísticos disponibles. Cuanto más visibles y utilizados sean los servicios en lenguas indígenas, mayor será el incentivo para continuar desarrollando tecnologías avanzadas como sistemas de reconocimiento de voz, traducción automática y generación de texto. Esto puede motivar a las empresas a invertir en proyectos de recopilación de datos, digitalización de textos y colaboración con comunidades locales para mejorar el rendimiento de la IA.

Al aprovechar estas iniciativas de inclusión de grandes marcas, se puede generar un efecto multiplicador en el desarrollo de tecnologías de IA para lenguas indígenas. Estas colaboraciones permiten combinar recursos técnicos, financieros y humanos, acelerando así el proceso de digitalización y fortaleciendo la presencia de estas lenguas en el ecosistema tecnológico global.

> Ampliar la conectividad de las comunidades indígenas

La conectividad es un factor esencial para la inclusión digital y tecnológica de las comunidades indígenas. Sin acceso a internet, las posibilidades de participar en la creación de contenidos digitales, el uso de herramientas basadas en IA y el aprovechamiento de recursos educativos y culturales se ven severamente limitadas. Por ello, es necesario impulsar estrategias que mejoren la infraestructura de conectividad en estas regiones y acompañarlas de iniciativas que promuevan el uso efectivo de internet. A continuación, se presentan algunas acciones clave.

1. Impulsar programas de ampliación de cobertura de internet

Muchos territorios donde residen comunidades indígenas se encuentran en zonas rurales o de difícil acceso, lo que ha dificultado históricamente la instalación de infraestructuras de conectividad. Es fundamental promover programas que amplíen la cobertura de internet en estas regiones, tanto a través de tecnologías tradicionales como mediante soluciones innovadoras, como internet satelital o redes de telecomunicaciones comunitarias. Sin embargo, la conectividad por sí sola no es suficiente. Estos programas deben ir acompañados de acciones de sensibilización y formación para que las comunidades puedan aprovechar plenamente las oportunidades digitales.

2. Acompañar la conectividad con programas de formación en el uso de internet

Una vez que las comunidades tienen acceso a internet, es importante implementar programas de formación que les enseñen a utilizar las herramientas digitales de manera efectiva y segura. Estos programas deben ser culturalmente adaptados, utilizando las lenguas indígenas siempre que sea posible, para garantizar una comprensión plena. Los contenidos de formación pueden incluir temas como la navegación en internet, el uso de redes sociales, la creación de contenido digital, la protección de la privacidad en línea y el acceso a recursos educativos y de salud.

3. Ampliar el alcance de programas de alfabetización existentes

En muchas comunidades indígenas ya existen programas de alfabetización en sus propias lenguas, que se enfocan en mejorar las habilidades de lectura y escritura. Estos programas son una base excelente para expandir el aprendizaje hacia el ámbito digital. A través de la integración de módulos específicos sobre el uso de internet y medios sociales, se puede fomentar la participación activa de los hablantes en plataformas digitales, lo que a su vez genera más contenido en lenguas indígenas y refuerza la identidad cultural en el entorno digital.

4. Establecer alianzas con actores públicos y privados

El éxito de estas iniciativas requiere la colaboración entre gobiernos, empresas tecnológicas, organizaciones no gubernamentales y las propias comunidades indígenas. Las políticas públicas deben priorizar la inclusión digital de las regiones más alejadas, mientras que las empresas pueden aportar recursos tecnológicos y experiencia técnica. Al mismo tiempo, es crucial que las comunidades sean protagonistas en el diseño e implementación de estas estrategias, asegurando que sus necesidades y expectativas sean respetadas.

Estas estrategias no solo buscan ampliar el acceso a internet, sino también garantizar que las comunidades indígenas puedan aprovechar plenamente las ventajas de la conectividad. Al integrar programas de formación, participación digital y colaboración comunitaria, se crea un entorno donde las lenguas y culturas indígenas puedan prosperar en el mundo digital.

> Incrementar la localización lingüística de los servicios de las grandes tecnológicas

La falta de localización en lenguas indígenas de los principales servicios digitales constituye una barrera significativa para millones de personas que no hablan otros idiomas. Sin interfaces en sus lenguas, muchas personas no pueden navegar fácilmente por internet, acceder a recursos en línea o utilizar herramientas digitales básicas. Esta situación limita enormemente la inclusión digital de las comunidades indígenas. Por ello, es fundamental trabajar en la localización de servicios clave, como sistemas operativos, navegadores web, buscadores y redes sociales. A continuación, se proponen estrategias para abordar este desafío.

1. Localización de sistemas operativos y navegadores web

Los sistemas operativos (Windows, macOS, Android) y navegadores web (Chrome, Firefox, Safari) son puntos clave de acceso a la tecnología. Sin localización en lenguas indígenas, los usuarios enfrentan barreras para realizar tareas básicas, lo que limita su inclusión digital. Ampliar la localización a estos sistemas mejora la accesibilidad, reconoce la importancia de estas lenguas en el entorno digital y permite una experiencia más fluida. Esto requiere colaboraciones con empresas tecnológicas para incluir dichas lenguas en procesos de traducción, actualización y personalización, asegurando una integración tecnológica más equitativa.

2. Localización de plataformas de búsqueda y redes sociales

Los servicios de búsqueda y las redes sociales son las principales puertas de acceso a la información y la comunicación en línea. Sin interfaces y menús en su lengua, muchos hablantes indígenas se ven excluidos de estas plataformas, lo que dificulta su participación en la vida digital. Establecer acuerdos con empresas tecnológicas para localizar sus servicios en lenguas indígenas puede aumentar enormemente la inclusión digital. Además, al proporcionar herramientas de búsqueda en sus idiomas, se mejora el acceso a contenidos relevantes para sus culturas y comunidades.

3. Sensibilización y colaboración con las grandes tecnológicas

Es crucial sensibilizar a las grandes empresas tecnológicas sobre la importancia de la localización lingüística. Muchas de estas empresas ya han adoptado políticas de diversidad e inclusión, lo que les permite estar abiertas a colaborar en iniciativas de localización. A través de diálogos directos, demostraciones de impacto social y estudios sobre la brecha digital, es posible persuadirlas para que amplíen su compromiso y destinen recursos a la incorporación de lenguas indígenas en sus productos.

Estas estrategias buscan reducir las barreras de acceso digital para los hablantes de lenguas indígenas. Incrementar la localización de servicios tecnológicos no solo promueve la equidad digital, sino que también refuerza el reconocimiento de las lenguas y culturas indígenas en el mundo tecnológico, permitiendo que sus hablantes participen plenamente en la sociedad digital global.

12. ANÁLISIS DE ERRORES Y MEJORA DE LA CALIDAD

El error más frecuente de las IAs interactuando en LL.II. es responder en otro idioma, superando 1 de cada 3 casos (35%).

1/10

Las respuestas resultan contaminadas en más de 1 de cada 10 respuestas que son respondidas en el idioma de instrucción (un 11%, con bucles repetitivos, sobreabundancia de hispanismos, parafraseando el prompt y traduciéndolo, pidiendo disculpas o aventurándose porque no lo entiende).

1/5

1 de cada 5 errores (23%) son corregidos con prompts más detallados, demostrando que el prompt engineering puede atajar problemas relacionados con un bajo volumen de datos, especialmente aquéllos relacionados con sesgos, formatos o la confusión de idioma.

Los errores más frecuentes de la IA en LL.II.

Como se ha podido apreciar a lo largo del informe, los errores más frecuentes se relacionan con la confusión de idiomas, la traducción de input en lugar de una respuesta directa y los sesgos culturales en las respuestas. Aproximadamente en el 35% de los casos, las respuestas generadas por las IA se producen en otro idioma, mayoritariamente en español (18%) o inglés (17%). Esto ocurre con más frecuencia en lenguas con menor presencia digital, como el Mapuche, donde uno de cada tres outputs es en inglés.

Otro error recurrente es la generación de respuestas basadas en traducciones automáticas, en lugar de interpretar y responder directamente en la lengua indígena. Este problema afecta hasta al 10% de los casos en lenguas como el Quiché, Mapuche y Tupi Guaraní. Además, la tendencia de la IA a expresar dudas o disculpas ("Sorry, but...") es más pronunciada en estos idiomas, con una correlación inversa del 66% respecto al volumen de artículos disponibles en Wikipedia.

Los errores asociados al sesgo cultural también destacan como una de las mayores debilidades de los modelos, con una tasa siete veces superior a la observada en interacciones en español. Las IA tienden a incluir referencias mitológicas o culturales en lugar de explicaciones científicas, especialmente en Quechua. Estos problemas reflejan la limitada integración de los contextos culturales propios en los modelos.

En términos estructurales, las respuestas tienden a carecer de organización en bloques o encabezados, particularmente en idiomas con escasos recursos tecnológicos. Por ejemplo, el Mapuche presenta salidas en un solo párrafo en el 80% de los casos, mientras que en el español este fenómeno es raro.

En conjunto, estos errores reflejan la importancia de fortalecer los recursos digitales y las herramientas lingüísticas específicas para cada lengua, con el fin de mejorar la precisión y adaptación cultural de los modelos de IA en lenguas indígenas.

Técnicas para mitigar los deficits de rendimiento de la IA

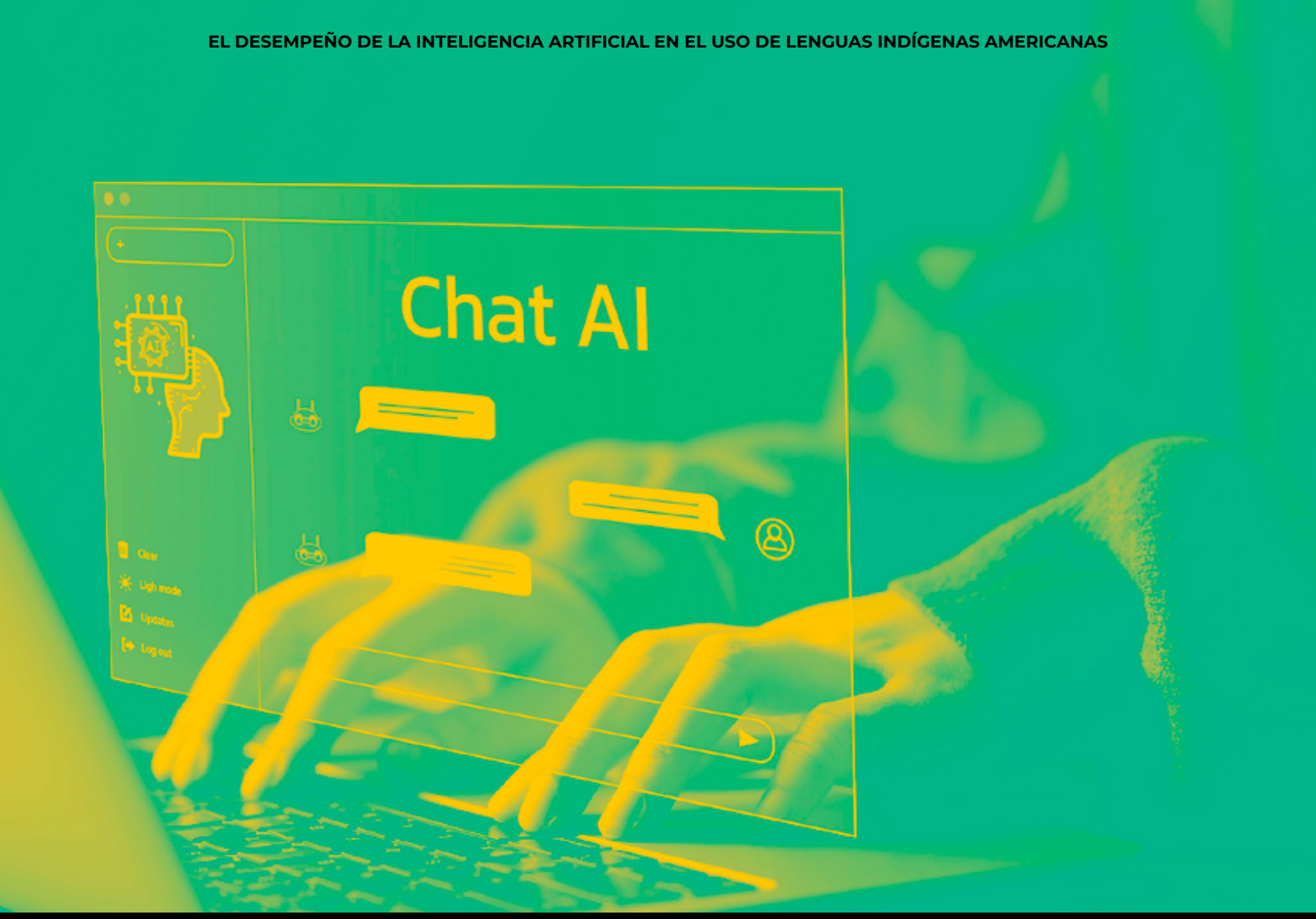
Un 23% de los experimentos que resultaron fallidos en una primera instancia (con prompts a nivel de usuario) **resultaron correctos gracias a prompts más detallados**, demostrando que los resultados pueden mejorarse a través de prompts de mayor calidad.

Este tipo de estrategias supone un apoyo cualitativo adicional a las medidas tecnológicas necesarias, y resultan efectivos no sólo para corregir errores, sino también para mitigar sesgos, corregir sobreentrenamiento de los escasos datos de las LL.II. o evitar confusiones con otros idiomas (por ejemplo, de documentos adjuntos).

Algunas de las estrategias son las siguientes:

- **Few-shot prompting**. Incluyendo ejemplos pregunta-respuesta en el prompt para que tenga una idea de cómo debe responder o con qué estructura.
- **Prompting con refuerzo del idioma**. Reforzar explícitamente el idioma en el prompt previene que la IA responda en español o inglés. También se puede reforzar al final de la instrucción, añadiendo "Si respondes en otro idioma, corrige tu respuesta inmediatamente".
- **Multi-step prompting**. Dividiendo la tarea en pasos enumerados, en lugar de incluir todas las operaciones en un sólo párrafo. Esta estrategia se puede combinar con la de few-shot prompting.

- **Context prompting**. Especialmente para mitigar sesgos. Se trata de añadir un preámbulo con información cultural para reducir respuestas que, por el escaso volumen de entrenamiento, puedan estar sesgadas o sobreajustadas.
- **Paraphrase prompting**. Exigir que reescriba, en la respuesta, la instrucción dada por el usuario, ayuda a reforzar que recuerde la instrucción y la responda correctamente.
- **Chain-of-thought (CoT)**. Especialmente para respuestas largas, alentar a la IA a razonar o explicar los razonamientos de porqué ha llegado a dichas conclusiones.
- **Refuerzo negativo**. Anticipar errores frecuentes en el prompt para que no los cometa. Por ejemplo: “no traduzcas el prompt ni respondas en español o en inglés”.



13. RECOMENDACIONES Y PLAN DE ACCIÓN

El presente plan de acción establece una serie de pasos estratégicos para disminuir la brecha tecnológica en el rendimiento de la inteligencia artificial (IA) en el contexto de lenguas indígenas americanas. Reconociendo el valor cultural, social y tecnológico de estas lenguas, el plan se enfoca en una intervención a múltiples niveles que incluye desde la formación de un consorcio internacional hasta la ejecución de proyectos comunitarios. A continuación, se describen las etapas fundamentales y sus objetivos asociados.

1. Creación de un Consorcio Internacional impulsor del proyecto

El éxito del plan depende en gran medida de la creación de un consorcio que funcione como el órgano central de dirección y coordinación. Este consorcio estará compuesto por organismos nacionales e internacionales, instituciones dedicadas a la protección cultural, empresas tecnológicas, universidades y organizaciones de la sociedad civil. Sus responsabilidades incluyen:

- **Definición de estrategia**: Establecer la visión, los objetivos a largo plazo y la planificación general del proyecto.
- **Gestión de alianzas**: Consolidar colaboraciones con entidades financiadoras, gobiernos locales y empresas tecnológicas.

- **Supervisión**: Monitorear la implementación de las acciones definidas, asegurando el cumplimiento de los plazos y objetivos.
- **Evaluación y ajustes**: Realizar análisis periódicos de los resultados, incorporando ajustes necesarios para mejorar el impacto del plan.

Este consorcio desempeñará un papel esencial como facilitador y garante de la sostenibilidad de las acciones a largo plazo.

2. Creación del Equipo de Trabajo de Implementación

Para llevar a cabo las acciones concretas definidas por el consorcio, se creará un Equipo de Trabajo de Implementación (ETI). Este equipo estará compuesto por especialistas de las entidades participantes y tendrá la misión de ejecutar, supervisar y evaluar las iniciativas locales. Entre sus principales funciones se incluyen:

- **Establecer alianzas estratégicas locales**: Identificar a actores clave (gobiernos, empresas, organizaciones no gubernamentales) para colaborar en los proyectos.
- **Supervisar proyectos**: Garantizar que los proyectos locales cumplan con los estándares de calidad y los objetivos establecidos.
- **Administrar recursos**: Asignar de manera eficiente los fondos, tecnologías y personal necesario.
- **Participación comunitaria**: Asegurar que las comunidades indígenas se involucren activamente en todas las fases del proyecto, incluyendo la toma de decisiones.

Este equipo será el puente entre las directrices del consorcio y la realidad operativa en las comunidades.

3. Organización de un evento de alta visibilidad para comunicar la iniciativa

Una vez formado el consorcio y el ETI, se realizará un evento de alcance internacional para presentar oficialmente la iniciativa. Los objetivos de este evento son los siguientes:

- **Presentación del informe**: Comunicar los resultados del estudio sobre la brecha tecnológica en lenguas indígenas.
- **Anuncio de la hoja de ruta**: Compartir con los asistentes el plan de acción estratégico, los próximos pasos y las metas a alcanzar.
- **Convocatoria del hackathon**: Promover la participación de actores interesados, incluidos desarrolladores, empresas tecnológicas, gobiernos y representantes de comunidades indígenas.

Este evento buscará atraer tanto atención mediática como nuevos aliados estratégicos que puedan sumarse al proyecto.

4. Hackaton de Innovación Tecnológica por una IA en lenguas indígenas

Como parte del plan, se organizará un hackathon enfocado en la innovación tecnológica para lenguas indígenas. Este evento reunirá a diversos actores interesados para colaborar en el desarrollo de soluciones tecnológicas. Los objetivos específicos del hackathon incluyen:

- **Desglose de estrategias**: Traducir las estrategias de inclusión en propuestas de proyectos específicos.

- **Desarrollo de soluciones**: Diseñar herramientas tecnológicas adaptadas a las necesidades de las comunidades indígenas.
- **Colaboración comunitaria**: Promover la co-creación de proyectos entre desarrolladores, empresas y miembros de las comunidades.

La estructura del hackathon contemplará la formación de equipos, la presentación de proyectos ante un jurado especializado, y la inclusión de los proyectos destacados en el plan de trabajo a largo plazo.

5. Desarrollo de alianzas estratégicas locales

El desarrollo de alianzas con actores públicos, privados, ONGs, empresas tecnológicas, universidades y medios de comunicación es fundamental, que puedan actuar como sponsors de iniciativas específicas.

- **Empresas privadas**: Involucrar a empresas tecnológicas y de otros sectores que puedan financiar o implementar proyectos de digitalización en lenguas indígenas.
- **Gobiernos locales**: Establecer colaboraciones con gobiernos interesados en promover políticas de preservación cultural y lingüística.
- **Universidades y ONGs**: Involucrar a instituciones académicas y organizaciones sin fines de lucro para aportar conocimientos, recursos y capacidades técnicas.
- **Medios de comunicación**: Difundir los avances y resultados del proyecto para generar apoyo público y atraer nuevos aliados.

Estas alianzas permitirán una mejor distribución de recursos y una mayor capacidad de ejecución en los territorios locales.

6. Ejecución de proyectos locales y monitorización del progreso de la iniciativa

La ejecución de proyectos locales constituye el núcleo del impacto del plan. A través de estos proyectos se busca mejorar directamente el rendimiento de la IA en lenguas indígenas, así como fomentar el desarrollo tecnológico dentro de las comunidades. Los pasos clave en esta fase son:

- **Asignación de recursos**: Distribuir fondos, equipamiento tecnológico y formación técnica a los equipos locales responsables.
- **Implementación**: Llevar a cabo las estrategias diseñadas en el hackathon y en los planes de trabajo, adaptándolas a las realidades locales.
- **Supervisión continua**: Establecer mecanismos de monitoreo para evaluar el progreso, documentar buenas prácticas y detectar posibles obstáculos.
- **Informes periódicos**: Los responsables de los proyectos deberán presentar informes regulares al consorcio, detallando avances, logros y desafíos.
- **Evaluación de resultados**: Analizar el impacto en términos de mejora tecnológica, creación de contenido en lenguas indígenas y participación comunitaria
- **Ajustes y mejoras**: Realizar las modificaciones necesarias para optimizar las estrategias aplicadas, basándose en los resultados obtenidos.

Este enfoque asegura que las iniciativas locales no solo sean ejecutadas de manera efectiva, sino también sostenibles y escalables en el tiempo.

A. ANEXOS Y TABLAS

A.1. KPIs de accesibilidad a distintas herramientas digitales

Lengua	Origen	País	% Población de habla el idioma	Año	Motor de búsqueda	¿Pemite el idioma?	Navegador	¿Permite el idioma?	Sistema operativo	¿Permite el idioma?	Redes sociales	¿Permite el idioma?	Cantidad de resultados	Porcentaje de resultados cuyo contenido es >70% del idioma	Concordancia con la búsqueda
Quechua	Perú	Bolivia	19%	2017	Google	Si	Chrome (Google)	Si	Windows (Microsoft)	Si	TikTok	No	13	7,19%	No
			20%	2017	Bing (Microsoft)	No	Edge (Microsoft)	Si	Linux (Libre)	No	Linkedin	No			
		Chile	0,20%	2017	Yahoo	No	Safari (Apple)	No	Mas OS (Apple)	No	Instagram	No			
									Android (Google)	No	Facebook	No			
Guaraní	Paraguay	Paraguay	33,40%	2021	Google	No	Chrome (Google)	Si	Windows (Microsoft)	Si	TikTok	No	5	0%	No
		Argentina	0,19%	2022	Bing (Microsoft)	No	Edge (Microsoft)	Si	Linux (Libre)	No	Linkedin	No			
		Bolivia	0,55%	2012	Yahoo	No	Safari (Apple)	No	Mas OS (Apple)	No	Instagram	No			
									Android (Google)	No	Facebook	No			
Mapuche	Chile	Chile	9,93%	2017	Google	No	Chrome (Google)	No	Windows (Microsoft)	No	TikTok	No	45	48,90%	No
		Argentina	0,31%	2022	Bing (Microsoft)	No	Edge (Microsoft)	No	Linux (Libre)	No	Linkedin	No			
					Yahoo	No	Safari (Apple)	No	Mas OS (Apple)	No	Instagram	No			
									Android (Google)	No	Facebook	No			
Aimara	Perú	Bolivia	11,27%	2012	Google	No	Chrome (Google)	No	Windows (Microsoft)	No	TikTok	No	2	50%	No
		Perú	1,171%	2017	Bing (Microsoft)	No	Edge (Microsoft)	No	Linux (Libre)	No	Linkedin	No			
		Argentina	0,31%	20122	Yahoo	No	Safari (Apple)	No	Mas OS (Apple)	No	Instagram	No			
									Android (Google)	No	Facebook	No			
Maya Quiché	Guatemala	Guatemala	7,00%	N/D	Google	No	Chrome (Google)	No	Windows (Microsoft)	No	TikTok	No	5	40%	Si
					Bing (Microsoft)	No	Edge (Microsoft)	No	Linux (Libre)	No	Linkedin	No			
					Yahoo	No	Safari (Apple)	No	Mas OS (Apple)	No	Instagram	No			
									Android (Google)	No	Facebook	No			
Náhuatl	México	México	1,70%	2000	Google	No	Chrome (Google)	No	Windows (Microsoft)	No	TikTok	No	9	44%	No
					Bing (Microsoft)	No	Edge (Microsoft)	No	Linux (Libre)	No	Linkedin	No			
					Yahoo	No	Safari (Apple)	No	Mas OS (Apple)	No	Instagram	No			
									Android (Google)	No	Facebook	No			
Tupi Guaraní	Brasil				Google	No	Chrome (Google)	No	Windows (Microsoft)	No	TikTok	No	1	0%	No
					Bing (Microsoft)	No	Edge (Microsoft)	No	Linux (Libre)	No	Linkedin	No			
					Yahoo	No	Safari (Apple)	No	Mas OS (Apple)	No	Instagram	No			
									Android (Google)	No	Facebook	No			

A.2. Matriz de correlación entre los escenarios digitales de los idiomas y sus errores más frecuentes

	Hablantes	Artículos Wikipedia	UsuariosWikipedia	Traductores	Extensiones respecto a español en todos los casos	Extensión respecto a español sólo en casos aparentemente correctos	Emojis en registro informal	Horarios en esquemas de planificación	Capítulos y titulares en redacciones	Define acrónimos	Respuestos español	Respuestos inglés	Repeticiones	Traduce en lugar de obedecer	"Sorry, but..."	"Appears to be..."	Casos erróneos	Casos erróneos solucionados al aportar más detalle
Hablantes	1.00	0.81	0.97	0.81	0.59	0.44	0.84	0.81	0.42	0.09	0.54	-0.39	-0.82	-0.65	-0.47	-0.54	0.08	0.27
Artículos Wikipedia	0.81	1.00	0.88	0.69	0.43	0.62	0.87	0.46	0.19	-0.09	0.07	-0.21	-0.69	-0.26	-0.41	-0.66	-0.36	0.03
UsuariosWikipedia	0.97	0.88	1.00	0.89	0.50	0.48	0.85	0.77	0.41	0.08	0.51	-0.44	-0.88	-0.62	-0.55	-0.69	-0.05	0.26
Traductores	0.81	0.69	0.89	1.00	0.28	0.29	0.67	0.68	0.60	0.07	0.74	-0.67	-0.99	-0.81	-0.62	-0.64	-0.10	0.10
Extensiones respecto a español en todos los casos	0.59	0.43	0.50	0.28	1.00	0.67	0.78	0.71	0.54	0.56	0.13	-0.48	-0.30	-0.25	-0.69	-0.46	-0.17	0.53
Extensión respecto a español sólo en casos aparentemente correctos	0.44	0.62	0.48	0.29	0.67	1.00	0.81	0.50	0.15	0.59	-0.10	-0.22	-0.25	0.12	-0.65	-0.63	-0.38	0.10
Emojis en registro informal	0.84	0.87	0.85	0.67	0.78	0.81	1.00	0.70	0.47	0.30	0.21	-0.47	-0.67	-0.36	-0.70	-0.68	-0.35	0.20
Horarios en esquemas de planificación	0.81	0.46	0.77	0.68	0.71	0.50	0.70	1.00	0.47	0.61	0.68	-0.53	-0.64	-0.60	-0.71	-0.65	0.29	0.61
Capítulos y titulares en redacciones	0.42	0.19	0.41	0.60	0.54	0.15	0.47	0.47	1.00	0.23	0.56	-0.95	-0.65	-0.77	-0.71	-0.29	-0.36	0.16
Define acrónimos	0.09	-0.09	0.08	0.07	0.56	0.59	0.30	0.61	0.23	1.00	0.26	-0.36	0.01	0.00	-0.69	-0.47	0.14	0.50
Respuestos español	0.54	0.07	0.51	0.74	0.13	-0.10	0.21	0.68	0.56	0.26	1.00	-0.60	-0.72	-0.90	-0.42	-0.24	0.43	0.20
Respuestos inglés	-0.39	-0.21	-0.44	-0.67	-0.48	-0.22	-0.47	-0.53	-0.95	-0.36	-0.60	1.00	0.69	0.73	0.84	0.49	0.37	-0.23
Repeticiones	-0.82	-0.69	-0.88	-0.99	-0.30	-0.25	-0.67	-0.64	-0.65	0.01	-0.72	0.69	1.00	0.85	0.58	0.56	0.15	-0.04
Traduce en lugar de obedecer	-0.65	-0.26	-0.62	-0.81	-0.25	0.12	-0.36	-0.60	-0.77	0.00	-0.90	0.73	0.85	1.00	0.43	0.21	-0.14	-0.12
"Sorry, but..."	-0.47	-0.41	-0.55	-0.62	-0.69	-0.65	-0.70	-0.71	-0.71	-0.69	-0.42	0.84	0.58	0.43	1.00	0.80	0.36	-0.43
"Appears to be..."	-0.54	-0.66	-0.69	-0.64	-0.46	-0.63	-0.68	-0.65	-0.29	-0.47	-0.24	0.49	0.56	0.21	0.80	1.00	0.27	-0.51
Casos erróneos	0.08	-0.36	-0.05	-0.10	-0.17	-0.38	-0.35	0.29	-0.36	0.14	0.43	0.37	0.15	-0.14	0.36	0.27	1.00	0.32
Casos erróneos solucionados al aportar más detalle	0.27	0.03	0.26	0.10	0.53	0.10	0.20	0.61	0.16	0.50	0.20	-0.23	-0.04	-0.12	-0.43	-0.51	0.32	1.00

A.3. Evaluación de rendimiento: los 14 ámbitos pormenorizados y sus dimensiones

AUTOPERCEPCIÓN	SESGO	HUMOR	DESCRIPCIÓN IMAGEN	GENERACIÓN IMAGEN
Yo - presentación	Género	Inocente (infantil)	Secuencia o cómic (mudo)	Protagonista (personaje principal)
Origen y creadores	Infancia	Comprensible	Relatabilidad	Lugar
Habilidades y capacidades	Religión	Adaptado (conecta con el escenario)	Gráfica	Público (personajes secundarios)
Idioma empleado	Mitológico	Limitado (se autocensura correctamente)	OCR y comprensión de datos	Acción
Ubicación región y clima	Científico		Meme	Verbalización (texto asociado a personaje)
Época o fecha	Económico		Humor de la sátira - comprensión actualidad	Iconografía cultural
Animado - inanimado	Moral			Etiquetado (texto en objetos)
Natural - artificial	Ideológico			Iconografía abstracta (logos)
Género				Reconocimiento de estilo
Edad				Adecuación al target (buyer persona)
Estatus económico				

DESCRIPCIÓN CÓDIGO	GENERACIÓN CÓDIGO	SUMARIZACIÓN (Descripción documentos)	REDACCIÓN DE ARTÍCULOS (generación docs 1)	REDACCIÓN DE EMAILS (generación docs 2)
Script	Script	Documentos españoles	Sostenibilidad	Imperativo (correo final para ordenar - pedir una acción)
Web	Web	Documentos ingleses	Política	Seguimiento (correo intermedio para avances)
Interacción (juego)	Interacción (juego)	Documentos portugueses	Social	Disculpas y cancelación (correo final - anular una acción)
	Estructuras (python, html)	Igualdad - sociedad - actualidad - administración		Presentación
	Sintaxis	Finanzas - economía - técnico		Registro
	Comentarios	Artesanía - tradición - espiritualidad - arte		

CREACIÓN DE PERFILES	PLANIFICACIÓN	CURRÍCULUM	CAMBIO DE ROL
Social (citas - tinder)	Calendarios (actividades) - Factor tiempo	STEM	Negocios (Cliente - Empresario)
Profesional (linkedin)	Pasos (para resolver un problema) - Factor estados	Restauración y servicios	Emocional (Mendigo - Drama)
Personal (actividades/ hobbies - instagram)	Organización (varios individuos) - Factor agentes	Comunicación	Ética (abogado defensor caso violencia g.)
Profundidad del perfil (género, edad, nombre)		Tecnicismos	
Intereses		Coherencia temporal	
Trayectoria profesional		Invención ajustada (alineada al relato)	
Adaptación al registro			
Actividades			
Ideología			

A.4. Otras funcionalidades asociadas al contenido digital en lenguas indígenas

	Quechua	Guaraní	Aimara	Náhuatl	Quiché	Mapuche	Tupi Guaraní	Español	Catalán	Euskera
Uso de capítulos, titulares o bloques	50%	70%	70%	30%	40%	20%	60%	90%	70%	70%
Uso de emojis en registro informal	60%	40%	30%	30%	30%	10%	30%	80%	70%	60%
Uso de horarios en planificación	10%	30%	10%	20%	10%	0%	10%	70%	60%	40%
Definición de acrónimos y conceptos	40%	60%	40%	80%	60%	10%	60%	100%	80%	80%

Una mejor redacción y distribución de bloques de texto (como pueden ser titulares, encabezados, capítulos, etc) que presentan una correlación directa del 60% con las lenguas que tienen más traductores online.

Aquellas lenguas con mayor volumen de comentarios en redes sociales (posts de X) tienden a hacer un uso más adecuado de jerga, emojis y registros informales con una correlación directa del 84%, y son capaces de reconocer la necesidad de emplearlo de manera abstracta sin necesidad que se les solicite explícitamente en 2 de cada 3 ocasiones.

El modo en que se estructuran las planificaciones y horarios está correlacionado con el volumen de contenido digital (Common Crawl) en un 78%.


BID
LAB