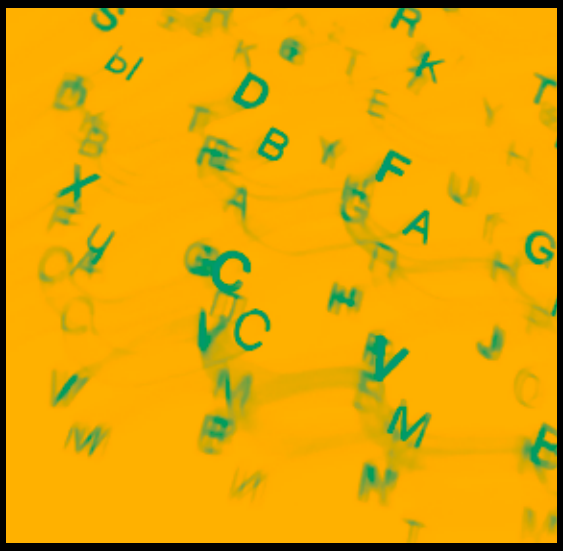
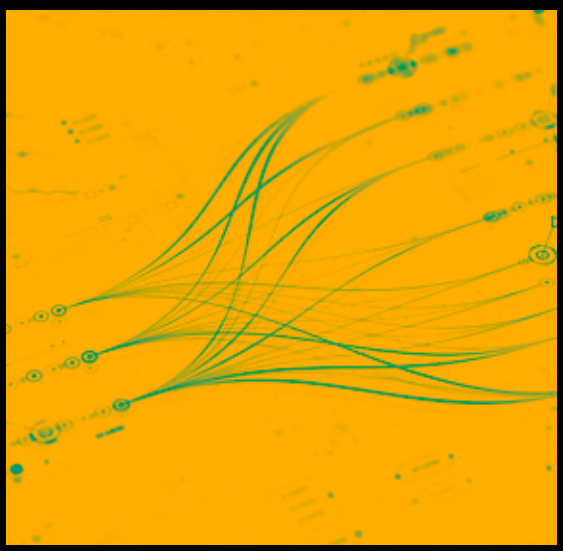
THE PERFORMANCE OF ARTIFICIAL INTELLIGENCE IN THE USE OF INDIGENOUS AMERICAN LANGUAGES





Microsoft
IDB | LAB

THE PERFORMANCE OF ARTIFICIAL INTELLIGENCE IN THE USE OF INDIGENOUS AMERICAN LANGUAGES

ASSESSMENT OF THE AI GAP IN INDIGENOUS AMERICAN LANGUAGES

Acknowledgments: We express our sincere gratitude to the interpreters and academics who participated in the project “Native Languages and Their Interaction with Artificial intelligence.” Their valuable collaboration was essential for evaluating the performance of various artificial intelligence technologies in Indigenous languages, making it possible to highlight the cultural and linguistic gaps present in these tools. In particular, we thank Jacob Cruz, Náhuatl language interpreter; Armando Hueyotenco, Náhuatl language interpreter from the Institute for Transparency, Access to Public Government Information, and Protection of Personal Data of the State of Hidalgo (ITAIH); Ana Paola Quispe Quispe, Aymara language interpreter from the National Academy of the Aymara Language (ANLA); Elmer Machicao, Aymara language interpreter; Tomas Rojas, Mapuche language interpreter; Yony Mediano, Quechua language interpreter; and Mauro Lugo, Guaraní language interpreter. Thanks to their expertise and commitment, significant evaluations were achieved in the Náhuatl, Aymara, Mapuche, Quechua, and Guaraní languages, strengthening the inclusive dialogue between technology and linguistic diversity in the region.

Authors: Writing, research and analysis: Miguel Lucas (LLYC), Alejandro Burgueño (LLYC), Miguel Carazas (LLYC). Conceptualization and strategic guidance: César Buenadicha (IDB Lab), Smeldy Ramírez (IDB Lab), César Rosales (IDB Lab). The team gratefully acknowledges the generous sponsorship and support of Microsoft, whose backing made the production of this report possible.

Design: Alejandro Scaff

https://bidlab.org

fAIr LAC and the IDB Group's Commitment to Responsible AI

This knowledge product was developed within the framework of **fAIr LAC**, the IDB Group's initiative that promotes the responsible adoption of artificial intelligence (AI) in Latin America and the Caribbean. The document constitutes a key input for the development of foundational models in Indigenous American languages, complementing the work of IDB Lab, the IDB Group's innovation laboratory, as demonstrated by its application in the *GuaranIA* project, focused on the development of foundational models in the Guaraní language.

The production of this publication was made possible thanks to the valuable contribution of Microsoft, with research by Llorente and Cuenca, and in close collaboration with IDB Lab through the *fAIr LAC* initiative.

fAIr LAC is fully aligned with the IDB Group's Commitment to Responsible AI, embodied in the IDB Group AI Framework, launched in January 2025. This framework establishes a roadmap structured around three fundamental pillars:

- **Institutions and governance**, to promote ethical standards and effective oversight mechanisms;
- D**ata and infrastructure**, aimed at ensuring equitable and open access to AI-enabling assets; and
- **People and talent**, focused on strengthening local capacities.

Since its launch in 2019, *fAIr LAC* has served as a platform for translating emerging ethical and regulatory principles into practical tools. Through IDB Lab, the initiative has evaluated more than 30 use cases using methodologies such as *fAIr LAC 3S* (Solution, System, Society) and *fAIr Venture*, designed to manage ethical and operational risks while maximizing social impact.

In addition to its technical advisory work, *fAIr LAC* addresses emerging challenges such as labor substitution resulting from the use of AI, promoting worker retraining and reskilling processes. It also promotes linguistic equity through research and development of foundational models in Indigenous languages, with the aim of correcting systemic biases and improving the performance of major language models in underrepresented languages.

IDB Lab

IDB Lab is the innovation and venture capital arm of the Inter-American Development Bank Group. We discover new ways to drive social inclusion, environmental action and productivity in Latin America and the Caribbean. IDB Lab leverages financing, knowledge and connections to support early-stage entrepreneurship, foster new technologies, activate innovative markets and catalyze existing sectors. www.bidlab.org

AI for Good Lab at Microsoft

The AI for Good Lab at Microsoft is a collaborative effort focused on leveraging artificial intelligence to solve some of the world's most pressing challenges. By working with global partners and leveraging cutting-edge AI tools, the lab supports innovations in environmental sustainability, agriculture, healthcare, and more. The AI for Good Lab is committed to applying AI solutions to improve lives, drive progress toward the UN's Sustainable Development Goals, and make a lasting impact through technology.

LLYC

LLYC is a global communications, digital marketing, and public affairs consulting firm. Founded in Madrid in 1995 as Llorente y Cuenca, LLYC currently has 20 offices in several countries, including Spain, Argentina, Brazil, and Colombia. LLYC helps its clients address their strategic challenges with solutions based on creativity, technology, and experience.

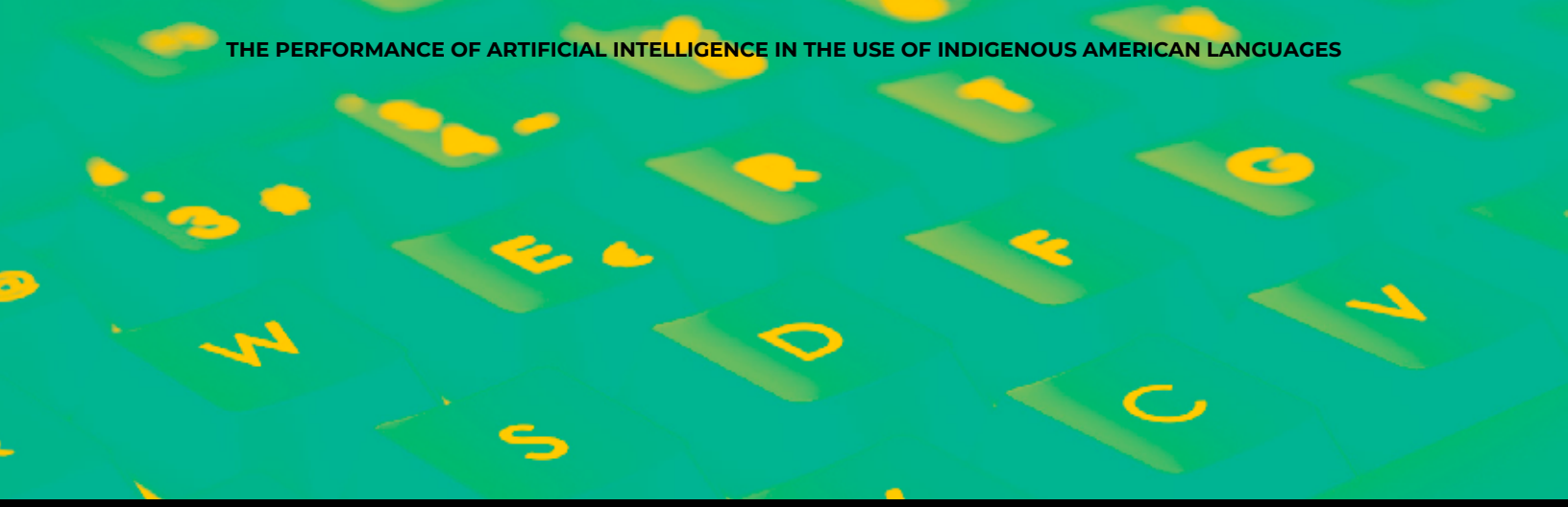
fAIr LAC

fAIr LAC is a partnership between the public and private sectors, civil society and academic institutions, designed to influence public policy and the entrepreneurial ecosystem in the promotion of the responsible use of AI.

1. TABLE OF CONTENTS

2. EXECUTIVE SUMMARY

- **Generative artificial intelligences show very poor performance in indigenous languages.** In only 54% of cases, when faced with questions formulated in Indigenous languages, the response is apparently correct. And even when this is the case, the answer is actually four times shorter, achieving a score of just 2.4/10 in terms of correctness of expression and 2.3/10 in terms of understanding the question.
- **The low presence of written texts and other resources in Indigenous languages on the internet—unlike majority languages**—significantly hinders AI's comprehension and expression in these languages. This effect is even more pronounced in the case of Indigenous languages with fewer speakers.
- **To increase the integration of Indigenous languages into the digital ecosystem, 21 strategies are proposed, focused both on increasing the available data in these languages and on the development of enabling technologies.** Promoting digital conversation in Indigenous languages, giving visibility to their influencers, protecting digital platforms and archives of traditions, and developing translation and voice technologies are some of the most important. In this way, these strategies would help train AI models to improve their performance in these languages.
- **The creation of an international consortium is proposed to implement these strategies.** It would consist of national and international organizations, institutions dedicated to cultural protection, and technology companies interested in accelerating the use of AI to bridge linguistic gaps.
- **Indigenous cultural references are a minority within AI.** The cultural bias present in AI's responses to questions asked in Indigenous languages is skewed towards Western hegemonic culture. Even in the case of Quechua (the language with the best performance in this regard), its representation is below 2.3/10.
- **AI offers a great opportunity to reduce isolation and give visibility to Indigenous peoples and cultures.** It is not only a new and powerful speaker for sharing and perpetuating Indigenous tradition and culture, but can also help bridge the gap in communities most isolated by illiteracy and monolingualism.
- **Governmental support programs for Indigenous communities and Tech firms initiatives are two fundamental pillars to improve the performance of AI in Indigenous languages.** Support programs from NGOs and initiatives from major consumer brands complete the picture of key allies for achieving better AI.
- **There is a very high correlation between the volume of digital content available in a language and the performance that AI shows in that language.** AI performance and representation in Wikipedia for a given language have a correlation of 91%. The greater the volume of open-access digital content in a language, the higher the quality with which AI can speak and understand that language.



3. DEFINITIONS AND ABBREVIATIONS

AI: Artificial intelligence. Generally, AI refers to a broad field of study encompassing multiple disciplines (Machine Learning, Natural Language Processing, Automated Reasoning and Planning, Deep Learning, etc.). In the context of this report, and for simplicity, AI refers to Large Language Models (LLMs), as these are the technological structures the general public currently understands as artificial intelligences, especially since the launch of ChatGPT on November 30, 2022.

LLM: Large Language Models. AI models specifically focused on understanding and generating human verbal language. Some state-of-the-art LLMs include Microsoft's PHI3, OpenAI's GPT-4o, Anthropic's Sonnet 3.5, Google's Gemini Pro 1.5, and Meta's Llama 3.0.

ILs: Indigenous Languages. In this report, these refer to the Quechua, Guarani, Aymara, Nahuatl, Quiche, Mapuche, and Tupi-Guarani.

MQM: Multidimensional Quality Metrics. An analytical tool used by linguists to evaluate the quality of translations and generated texts. It provides a structured framework that classifies errors into specific categories, allowing for detailed quantitative analysis of accuracy, fluency, and language consistency.

MMLU: Multi-Task Language Understanding. A benchmarking metric designed for next-generation language models. It measures performance in comprehension tasks across multiple domains and difficulty levels, evaluating the model's ability to handle complex questions in various fields of knowledge.

Prompt: The instruction or question input into AI to see how it responds. In this study, prompts were previously translated into the respective ILs to analyze how the AI model responds to direct interaction from a native user, whether it recognizes the language, and whether it efficiently resolves other measurement parameters.

Query: A query is an instruction or question formulated by a user or system to retrieve specific information from a database or search engine. It is an essential tool for interaction between users and systems, particularly in natural language processing environments.

Web Scraper: An automated program designed to extract structured or unstructured information from web pages. Used for data collection tasks, this process enables the analysis of large volumes of information and its storage for specific applications, such as trend analysis or training LLMs.

4. APPROACH

In Latin America, indigenous languages are an essential component of the region's cultural heritage, but many face significant challenges due to globalization, digital exclusion, and the lack of adapted resources. These languages, spoken primarily by rural communities, reflect a historical and cultural richness that is at risk of disappearing. In response, various initiatives have been implemented by governments, international organizations, NGOs, and tech companies to promote their preservation and revitalization. These initiatives aim not only to guarantee linguistic rights but also to strengthen cultural identities and improve access to educational and economic opportunities in the communities that speak them.

However, in the field of artificial intelligence (AI), indigenous languages have very little presence. The performance of AI models adapted to these languages perpetuates digital exclusion and limits access to advanced technological tools for these communities. Unlike majority languages, the low presence of written texts and other resources in indigenous languages on the Internet significantly hinders AI's comprehension and expression in these languages. This effect is even more pronounced for languages with fewer speakers.

The objective of this study is to analyze the ability of next-generation AIs to adapt and respond effectively in indigenous languages, evaluating linguistic, operational, and behavioral dimensions. To achieve this, seven representative indigenous languages from Latin America were selected, distributed across the region, with different population sizes and cultural traits. The analysis focuses on how five state-of-the-art language models, including GPT-4o and Phi 3, handle practical use cases such as writing, messaging, and content interaction.

The methodology employed combines linguistic and functional evaluation standards, adapting metrics such as MMLU and MQM in three key categories: idiomatic, executive, and behavioral. It analyzes how well AI speaks, how well it understands, and in what cultural framework AI expresses itself when interacting in indigenous languages. The assessment considers aspects such as grammatical correctness, cultural coherence, task execution accuracy, and adaptation to specific registers. Use cases are approached at two levels of abstraction, simulating both regular users and experts, allowing for an evaluation of model effectiveness in varied contexts.

This approach aims not only to quantify current gaps but also to establish a framework for developing inclusive strategies that increase the integration of indigenous languages into the digital ecosystem. In doing so, it proposes adjustments and useful actions for training models to improve their performance in these languages. Ultimately, the goal is for the results of this study to serve as a reference for researchers, developers, and government entities in formulating policies and tools that promote linguistic equity in the era of artificial intelligence.

5. STUDY SCOPE AND METHODOLOGY

This study analyzes the capabilities and performance of next-generation artificial intelligence (AI) models when interacting in American indigenous languages (ILs). Its objective is to identify and quantify existing linguistic gaps and propose strategies to improve AI performance in ILs. The study addresses aspects such as the quality of AI interactions, the availability of texts for training, and the social and cultural impacts of these gaps.

To achieve these objectives, a multidimensional approach was applied with the following stages:

Selection of Languages and AI Models:
Seven representative indigenous languages and five state-of-the-art language models were selected for comparative testing in different practical scenarios. Specifically:

- For the languages, the study analyzed **Quechua, Guarani, Aymara, Nahuatl, Quiche, Mapuche, and Tupi-Guarani**. The same battery of tests was also conducted on two Western languages with a similar number of speakers, Catalan and Basque, to serve as comparison points.
- The AI LLMs evaluated were: **GPT-4o** from OpenAI, **Claude 3.5 Sonnet** from Anthropic, **PHI-3** from Microsoft, **Gemini 1.5 Pro** from Google, and **Llama 3** from Meta.

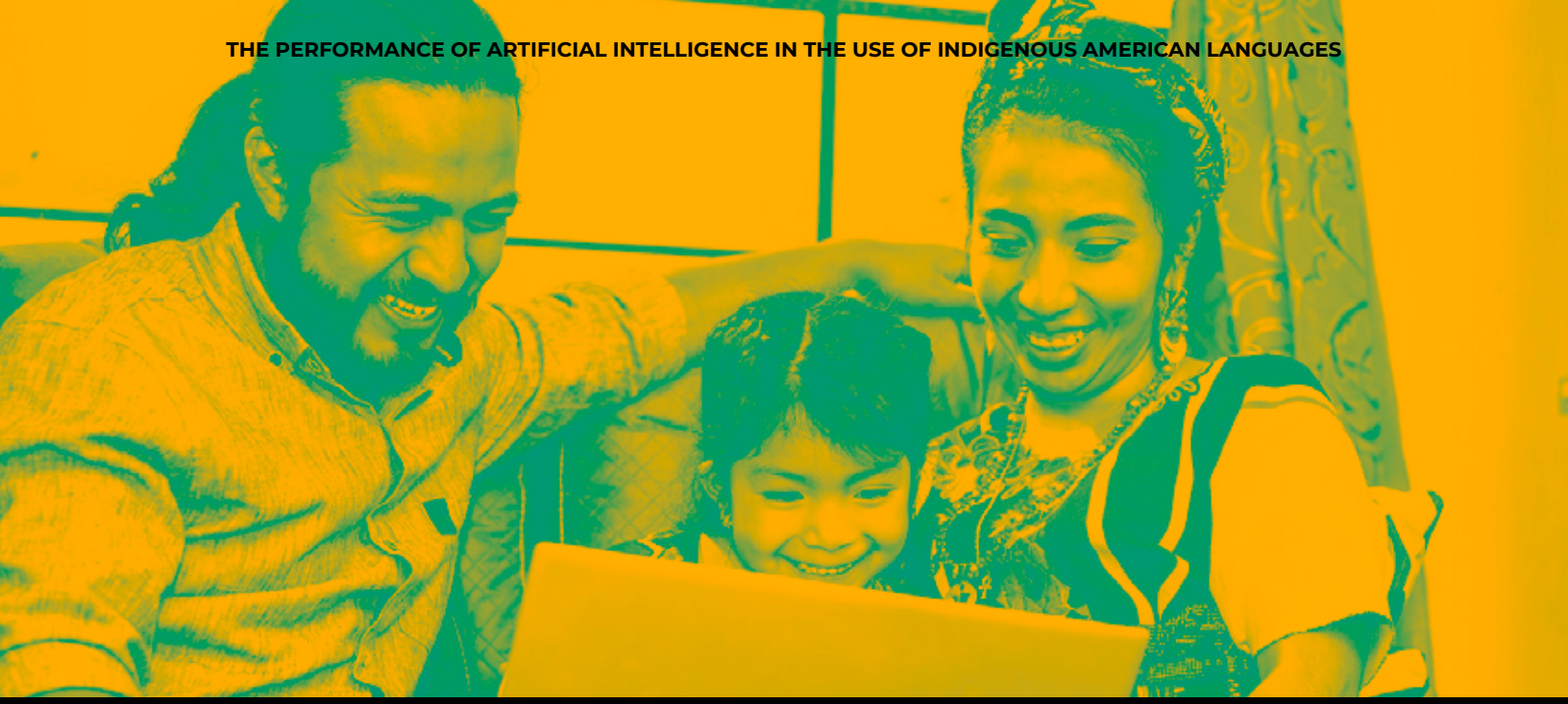
1. **Use Case Design:**
   The use cases focused on writing, messaging, and content interaction, evaluating both simple and complex task execution. These scenarios align with the most frequent interactions currently performed with AI models.

2. **Evaluation Criteria:**
   Established metrics such as MMLU (Multi-Task Language Understanding) and MQM (Multidimensional Quality Metrics) were combined to evaluate three key dimensions:
   - **Linguistic evaluation (expression assessment):** Fluency, accuracy, and coherence.
   - **Executive evaluation (comprehension assessment):** Precision, completeness, and understanding.

   - **Behavioral evaluation (cultural reflection assessment):** Cultural adaptation and contextual adequacy.

3. **Analysis of Digital Data and Linguistic Resources:** The study assessed the amount of texts available in indigenous languages for AI training, as well as the presence of linguistic corpora, digital tools (automatic translators, language detectors, speech-to-text, etc.).
4. **Contextual Comparison:** AI performance in indigenous languages was compared with languages with greater digital presence (Catalan, Basque) to identify correlations between AI response quality and the availability of digital resources.
5. **Inclusion Dimensions:** The impact of AI performance gaps was analyzed in economic, social, and cultural terms, evaluating how these limitations perpetuate the digital exclusion of indigenous communities.

This methodology aims not only to quantify the technological gap but also to establish a framework for implementing inclusive strategies that strengthen the presence and functionality of indigenous languages in artificial intelligence.



6. DATA ANALYSIS AND LANGUAGE COVERAGE

For every 20 websites browsed on the web, one is in Spanish, but 125,000 pages are needed to find one in Guarani, one of the ILs with the highest online presence.

The gap with other minority languages is also overwhelming: although Quechua has 10 times more speakers than Basque, the latter has 20 times more Wikipedia articles and 60 times more web content than Quechua.

57%

Only 57% of ILs have online dictionaries, and only 29% have ongoing projects for the development of lexical databases (Quechua and Guarani).

Digital tools are also scarce. Although Quechua can be detected by half as many automatic detectors as Spanish, the rest of ILs are detected by only a third.

Essential resources for training an artificial intelligence in a language

Numerous digital resources are required to train a state-of-the-art large language model (also known as **LLMs - Large Language Models**. However, in this report, we will refer to this technology simply as "AI"). These models require a large volume of written data in natural human language for each language in question, and to process them, access to this information must be provided digitally. Among the most relevant **content resources**, we can highlight:

- **Social conversation texts**, as this type of data helps train the model in informal and colloquial language use, contemporary information, and the natural use of language.
- **Digital media texts and horizontal information channels**, such as Wikipedia and other channels, as these data transfer world knowledge and historical reality in formal and popular dissemination registers.
- **General web texts**, to facilitate generalization and abstraction while expanding content. For example, Common Crawl[1] content.
- **Others** related to governmental or scientific information, for example, those that promote the verticality of the model in other language scenarios, such as legal or technical fields.

Among the ***tool resources*** necessary, the following stand out:

- **Automatic translators**, as they facilitate testing and improving training data and are useful for generating synthetic data[2] (data augmentation) from languages that are at a more advanced level in the state of the art of AI.
- **Automatic language detectors,** as they allow filtering texts and content written in a specific language and selecting them specifically for AI training.
- **Speech-to-text and text-to-speech** conversion models (also known as voice-to-text) as they enable access to data from audio sources, as well as facilitate access for users who have difficulty reading or writing a specific language, or non-literate individuals.
- **Digital dictionaries**, lexical databases, etc. to facilitate universal access to the study and understanding of a language.

Digital data available in indigenous languages

Approximately 1 in every 20 websites worldwide is written in Spanish. However, for Guarani,[3] the indigenous American language with the most web content, it would take 125,000 web pages to find just one written in that language.

The digital resource gap for indigenous languages is so vast that, for example, compared to Basque, which has ten times fewer speakers than Quechua, the former has 19 times more entries on Wikipedia and 19 times more users contributing to publications. It is important to note that Quechua has five times more encyclopedic content than Guarani, Aymara, or Nahuatl, highlighting the overall lack of digitalization of ILs.

Within this comparative framework, our studies show that Basque also has 60 times more web pages than Quechua, despite the fact that content in Quechua is published on social media 2.5 times more than in Basque. So, why is it that, even though many indigenous languages have more native speakers than other languages (which are more advanced in AI-related fields) and their presence in social conversations is not negligible, their digital content remains scarce?

1 Common Crawl is a non-profit organization that collects and provides free massive data files scraped from the web. Its repository includes information from billions of web pages, commonly used to train artificial intelligence models and conduct research in natural language processing. https://commoncrawl.org/

2 Data sets artificially generated by algorithms, designed to replicate the statistical characteristics of real data without compromising sensitive or private information. They are used in testing, model development and analysis, offering an ethical and secure alternative for managing information.

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

Table 1. Presence of publications in indigenous languages on Wikipedia and other networks.

	Quechua	Guarani	Aymara	Nahuatl	Quiche	Mapuche	Tupi-Guarani	Catalan	Basque	Spanish
Thousands of speakers	7,000	6,500	1,906	1,793	1,055	150	20	10,020	750	600,600
Thousands of entries on Wikipedia	24	5	5	4	-	-	-	764	448	1,992
No. editors in Wikipedia	40	30	15	14	-	-	-	1,503	751	14,171
Presence on the web (Common Crawl)	0.0006%	0.0008%	0.0001%	-	-	-	-	0.20%	0.04%	4.62%
Thousands of posts on X per year	7,231	985	2,387	1,650	375	20	203	92,850	2,912	5,531,599

Beyond aspects we will address soon, such as funding or support, one key factor lies in understanding the ratio: on average, each Basque speaker posts 4 times per year on X, while each Quechua speaker posts only once. This indicates that the Basque population is more familiar with technology and contemporary communication channels than the Quechua population.

However, what is the current standard ratio? The ratio varies significantly depending on the language and the nature of the populations using it, but we could say that a standard value is 9 posts per year per speaker, something observed in Spanish and, similarly, in Catalan.

Digital linguistic tools in indigenous languages

When it comes to digital tools, it is important to specify that web dictionaries are, to a greater or lesser extent, available for only four of the seven languages studied, with Tupi-Guarani, Quiche, and Mapuche being the indigenous languages that do not have access to online translation tools. Another very different issue is the lexical databases[4] the lexicon, among which one of the best-known examples is WordNet. The main reference is English[5], and these databases are essential for training and labeling lexical and syntactic structures. Currently, there is a WordNet in Spanish[6] and both Basque and Catalan are hosted within the EuroWordNet framework[7]. Although the ILs studied do not have open WordNets, research and publications indicate progress in Quechua[8] and Guarani.[9]

4 Structured collection of words and their linguistic characteristics, such as meaning, pronunciation, semantic and morphological relationships. It is used in natural language processing and linguistic studies to analyze and model language behavior.

5 Wordnet English, Princeton reference: https://wordnet.princeton.edu/

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

7 EuroWordnet, European language grouping project: https://archive.illc.uva.nl//EuroWordNet/

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

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

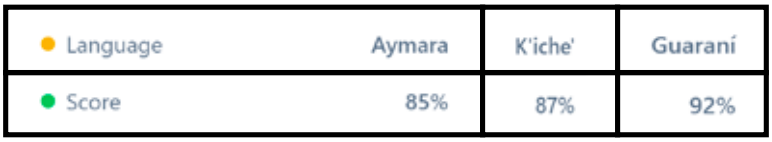
Table 2. Distribution of linguistic tools and search engines by indigenous language

	Quechua	Guarani	Aimara	Nahuatl	Quiche	Mapuche	Tupi Guarani	Catalan	Basque	Spanish
Number of automatic translators	2	2	2	1	0	0	0	3	3	5
Number of automatic language	5	4	4	4	3	2	0	7	8	10
Text2speech + Speech2text	2	1	1	0	0	0	0	3	3	3
Allows searching on Goggle	Sí	No	No	No	No	No	No	Sí	Sí	No
Allows searching on Bing	No	No	No	No	No	No	No	Sí	Sí	No

Automatic detection and translation is another key point, not only for measuring accessibility for its population but also for assessing the possibility of increasing training volumes and evaluating efficiency when inferring the language of AI users. Modern Tupí-Guarani, the Amazonian variant with the greatest presence in Brazil, is often confused with or overridden by Paraguayan Guarani and does not have effective language detectors. Additionally, along with Quechua and Mapuche, it lacks functional web translators.

Generally speaking, languages like Quechua are detectable by approximately half of the popular multilingual tools capable of detecting Spanish. On the other hand, Guarani, Aymara, and Nahuatl have a detection rate of 40%, which is low compared to Catalan and Basque, which reach 75%. Some notable references for detecting ILs are OpenL[10] and Originality,[11] but the limited competition among language detectors tends to result in low detection accuracy.

Figure 1. Detection confidence scores for some examples in OpenL



Finally, it is essential to measure the existing gap in transformative format tools, such as Speech2Text (voice-to-text converters). Although digital content is mostly written, recent studies indicate that approximately 80% of the population listens to some form of audio content on their devices[12] (podcasts, videos, audiobooks, etc.). Additionally, many speakers of ILs cannot read or write (according to Unicef, one fifth of the indigenous population is illiterate and among those who speak the indigenous language, 1 in 4 cannot read or write).[13] This makes audio-based formats a high-potential growth area on the web.

The need for these tools is twofold: on the one hand, they are essential for improving AI training, as AI models require large volumes of natural text data, meaning audio must be converted; on the other hand, they are crucial for illiterate native speakers to increase their use of digital tools.

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

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

12 Article on the consumption of audio content: https://www.avixa.org/es/contenidos/noticias-y-tendencias/cuatro-formatos-de-contenido-de-audio-para-fortalecer-tu-estrategia-de-marketing

13 Unicef about illiteracy in indigenous languages: https://www.unicef.org/mexico/comunicados-prensa/unicef-y-fundaci%C3%B3n-jorge-mar%C3%ADn-estrenan-proyecto-audiovisual-arte-y-lengua

That being said, there are no effective transcription models for Mapuche, Tupi-Guarani, Quiche, or Nahuatl. Additionally, there are more transcription tools available for Quechua than for Guarani or Aymara, though all of them are far fewer compared to the transcription models available for Spanish, Catalan, or Basque.

Other available tools in indigenous languages

In essence, although many digital tools are accessible to any user with technological resources—such as browsers, search engines, social media, or operating systems (see Annex)—, the fact that their performance is below expectations in a given language, or that interfaces are not available in that language, inhibits content creation by its speakers.

For example, searches performed in Quechua[14] on Google reveal only one link with a high proportion of Quechua content for every 14 links displayed, most of which are predominantly in Spanish.
For Aymara, which is the indigenous language that performs best in searches, results return approximately one link in Spanish and one in Aymara across all links on the first page of search results.

Apart from limiting users, this also highlights another disadvantage of indigenous languages: while their digital presence is already low, it is also difficult to extract since search criteria return non-representative content (scraping and query generation limitations).[15] In other words, what little content exists is not easily retrievable for training AI models.

Table 3. Distribution of technological resources by indigenous language

	Quechua	Guaraní	Aimara	Náhuatl	Maya Quiché	Mapuche	Tupi Guaraní	Spanish	Catalán	Euskera
Language recognition in Google searches	7.2%	3.0%	50.0%	44.0%	40.0%	48.9%	0.0%	100.0%	80.0%	100.0%
Language available in search engines	1	0	0	0	0	0	0	3	3	2
Language available in browsers	2	2	0	0	0	0	0	3	3	3
Language available in Operating Systems	1	1	0	0	0	0	0	4	4	4
Language available in Social Networks interface	0	0	0	0	0	0	0	4	3	1

We must also add that when a user acquires a device, they are almost never able to adapt it to an indigenous language, which determines the language of the applications they install, the web browsers they use, and the interface of search engines they access. Only in Quechua and Guarani is it possible to use Microsoft Windows in these languages (other operating systems do not allow it), and even then, only in a few browsers like Chrome or Edge.

When a Quechua speaker accesses social media, they will see that the entire interface of their application, regardless of which of the four analyzed social networks they use, will be in Spanish or English. All these factors significantly influence how users interact with digital tools and affect the ability of native speakers to create content that, in the medium and short term, could be used by AI models.

14 Searches carried out on Google for all ILs, considering all the links on the first page of results, and using as benchmarking "current news from the indigenous world" translated into the language in question.

15 Scraping refers to the process of downloading content from web spaces, social networks or other platforms automatically. Query is the message that is entered into a search engine (such as Google), and it returns a series of results. For more information see Dictionary of Acronyms and Technical Terms.

7. PERFORMANCE EVALUATION

54/100

Out of every 100 questions asked to the AI in indigenous languages, only 54 appear to be correct. In 35% of cases, it responds in another language, and in 11%, it mixes multiple languages or repeats terms in a loop.

The seemingly correct answers are five times shorter than when the same question is asked in Spanish.

The **quality of the AI's proficiency in indigenous** languages is rated at 2.4 out of 10.

The **AI's understanding of instructions given** in indigenous languages is rated at 2.3 out of 10.

The **AI's use of native cultural references is very low,** rated at 2.1 out of 10.

Methodology and performance metrics used

State-of-the-art AIs evolve rapidly, and so do their performance measurement metrics, which are constantly debated due to the wide range of dimensions and objectives they aim to assess. These priorities shift and may vary depending on the companies behind each model. In this regard, MMLU[16] (Massive Multitask Language Understanding) is one of the most widely used metrics for evaluating large language models across multiple tasks and domains.

However, the task at hand involves an additional linguistic dimension, requiring the assessment of aspects such as fluency, accuracy, and naturalness in each language independently, an approach closely related to MQM [17](Multidimensional Quality Metrics), traditionally used by linguistic specialists. For this reason, we have developed a measurement system that integrates both perspectives into three main categories: idiomatic, executive, and behavioral.

Therefore, the following aspects are studied:

- **7 indigenous languages**: Guarani, Tupi-Guarani, Nahuatl, Quiche, Quechua, Aymara and Mapuche. Additionally, to provide a reference for comparison, Catalan and Basque have been included in the study—two Western languages with a number of speakers similar to some of the ILs under analysis.
- **5 state-of-the-art LLMs**: GPT-4o, Gemini 1.5 Pro, Claude 3.5 Sonnet, PHI 3 and Llama 3.
- **3 common use cases** inspired by typical usage rankings: Article Writing, Email Communication, and Summarization & Interaction with Content. Each addressing various registers, cultural contexts, and critical aspects related to bias.
- **2 levels of abstraction per use**: Limited Input (general, provided by a typical user) and Detailed Input (overdefined, provided by an expert user).
- **3 rating categories** inspired by MQM and MMLU assessments, each further divided into 3 or 4 key aspects, measured on a 0 to 10 scale.
    - ***Evaluation of Expression, or idiomatic Assessment***

      (Evaluates whether the AI speaks the language well and if it can express itself in a way analogous to a human in that language.)

        - **Correction**: Is it clearly identified that the AI is speaking in the language of study, and are its grammar and syntax correct?
        - **Fluidity**: Is the language used natural, similar to that of a native, or on the contrary, is it "robotic" by repeating taglines, archetypal greetings or enumerative introductions?
        - **Coherence**: Do your arguments support each other, or on the contrary do they contradict each other in the same answer (for example, the conjugations are not contradictory and neither is the use of double negation)?
        - **Consistency**: Do you maintain a thread in your response and do not make sudden changes in your story, that is, do the sentences begin by naturally "picking up" how the previous ones end?

16 MMLU stands for Massive Multitask Language Understanding, a benchmarking metric designed for state-of-the-art language models. It measures performance in comprehension tasks across multiple domains and difficulty levels. More information can be found in the Acronym Dictionary.

17 MQM refers to the Multidimensional Quality Metrics, an analytical tool used by linguists to assess the quality of translations and generated texts. More information can be found in the Acronym Dictionary.

- ***Assessment of understanding, or executive assessment***

  (*Whether it understands the request and executes the task correctly or not.*):

  - **Precision**: Does it provide an accurate response (correctly addressing the request) and stay focused on its objective without redundancy or "wasting" text on unrequested details? Precision decreases as more text is dedicated to unrelated or unnecessary content.
  - **Completeness**: Does it fully complete the task and all associated subtasks without leaving it unfinished?
  - **Abstraction**: Does it accomplish the task on the first attempt, or does it require the extended and detailed format to perform it correctly? The better it interprets intent with minimal instruction, the better its abstraction ability.
  - **Understanding**: Does it comprehend the assigned task, or does it struggle to understand what is being requested, confusing it with other tasks? (A correctly understood task can still be answered inaccurately or imprecisely.)

- ***Evaluation of reflected culture, or behavioral assessment***

  (*Evaluates whether the AI reflects indigenous culture and adapts to the expected register and attitude.*):

  - **Cultural Bias**: Does it exhibit biases uncommon in the culture associated with the language it uses but potentially acquired from other cultures? (Score: 0 indicates strong bias, while 10 indicates the absence of bias.)
  - **Appropriateness**: Does it use inappropriate, offensive, or immoral terms, or does it employ toxic language? (Score: 0 indicates an unacceptable presence of inappropriate content, while 10 reflects exceptionally appropriate content.)
  - **Adaptation**: Does it use registers that do not match the user's intent, being either too formal or informal compared to what is required? (Score: 0 indicates an unexpected register that does not fit the request or context, while 10 reflects an appropriate register adapted to the situation (informal if the context is informal, formal if the context is formal, etc.)

- **1 additional question is considered to assess the quality of the instruction,** instruction (or prompt), which has been generated using state-of-the-art machine translators or, alternatively, the LLM that provided the best translation based on its back-translation accuracy and the performance of the responses it was able to generate.

- Additionally, a **questionnaire with 14 detailed areas**, each subdivided into three scenarios, where both a limited interaction and an extended interaction are evaluated. Each area consists of 3 to 12 different dimensions (specific to that area) and is assessed through back-translation methods[18] using a gradual criterion based on the level of development in the response.

  - 0% - Incorrect / not answered, 20% - Partially correct, 40% - Correct, 60% - Correct + Elaborated, 80% - Correct + Elaborated+ Proper format, 100% - Correct + Elaborated+ Proper format + On the first attempt.

The annex presents all the questions and the evaluation dimensions considered in the section 14 detailed arear - dimensions.

18 This procedure, based on the translation of the input and subsequent translation of the output into a known language (Spanish), limits the evaluation to the ILs with access to web translators, such as Quechua and Guarani.

Results of the AI performance evaluation in ILs

> Overall AI performance in indigenous languages

Overall, the AIs' performance when interacting in indigenous languages is one-third of what it can achieve when responding in Spanish.[19]

Table 4. Overall average AI performance in indigenous languages

	Quechua	Guarani	Aimara	Nahuatl	Quiche	Spanish	Catalan	Basque
Linguistic	4.53	2.68	3.08	3.70	0.42	10.00*	8.62	7.58
Executive	4.48	2.88	2.48	3.77	0.43	10.00*	8.02	6.03
Behavioral	2.15	2.73	2.02	2.80	2.90	10.00*	9.12	7.15
Average	**3.72**	**2.77**	**2.53**	**3.42**	**1.25**	**10.00***	**8.58**	**6.92**

Other minority languages such as Catalan and Basque have a performance exceeding 70% of that in Spanish, which is twice as high as that of ILs.

**Note: Spanish is assigned a score of 10, as it serves as the reference language for comparisons.*

> Performance differences between proprietary models and open models

Additionally, these differences are even more pronounced between proprietary models and open-weights models,[20] with proprietary models achieving 55% higher performance. This further limits free access to AI for ILs speakers.

Table 5. Average performance of proprietary models

	Quechua	Guarani	Aimara	Nahuatl	Quiche	Spanish	Catalan	Basque
Linguistic	5.33	3.28	3.58	5.28	0.67	10.00*	9.83	9.58
Executive	5.36	3.53	3.06	5.39	0.64	10.00*	9.22	7.47
Behavioral	2.36	3.25	2.50	4.00	4.25	10.00*	9.64	8.92
Average	**4.35**	**3.35**	**3.05**	**4.89**	**1.85**	**10.00***	**9.56**	**8.66**

19 Bringing together in performance the linguistic factors (idiomatic), the correct execution of the request (executives) and an adequate record and behavior (behavioral) previously mentioned.

20 Open-weights models refer to those large language models whose training weights are free to use and, therefore, more accessible.

Table 6. Average performance of open-weights models

	Quechua	Guarani	Aimara	Nahuatl	Quiche	Spanish	Catalan	Basque
Linguistic	3.33	1.79	2.33	1.33	0.04	10.00*	6.79	4.58
Executive	3.17	1.92	1.63	1.33	0.13	10.00*	6.21	3.88
Behavioral	1.83	1.96	1.29	1.00	0.88	10.00*	8.33	4.50
Average	**2.78**	**1.89**	**1.75**	**1.22**	**0.35**	**10.00***	**7.11**	**4.32**

**Note: Spanish is assigned a score of 10, as it serves as the reference language for comparisons.*

This gap between proprietary and open-weight models is also significant when comparing indigenous languages to other minority languages like Catalan and Basque. Proprietary models are twice as effective as open-weight models when interacting in indigenous languages but only 60% more effective when responding in Catalan or Basque.

> AI Performance in indigenous languages

AI models respond in Spanish with 4 times more fluency (313% higher) than in indigenous languages, making responses in the latter appear more artificial, robotic, and unstructured. Similarly, grammatical accuracy is not significantly better than fluency, being only 4% more acceptable than the latter.

The most affected linguistic aspect is consistency, with an average score of 2.30 out of 10 (335% higher in Spanish). This means that AI responses in indigenous languages lose coherence between paragraphs, failing to maintain a consistent thread throughout the response.

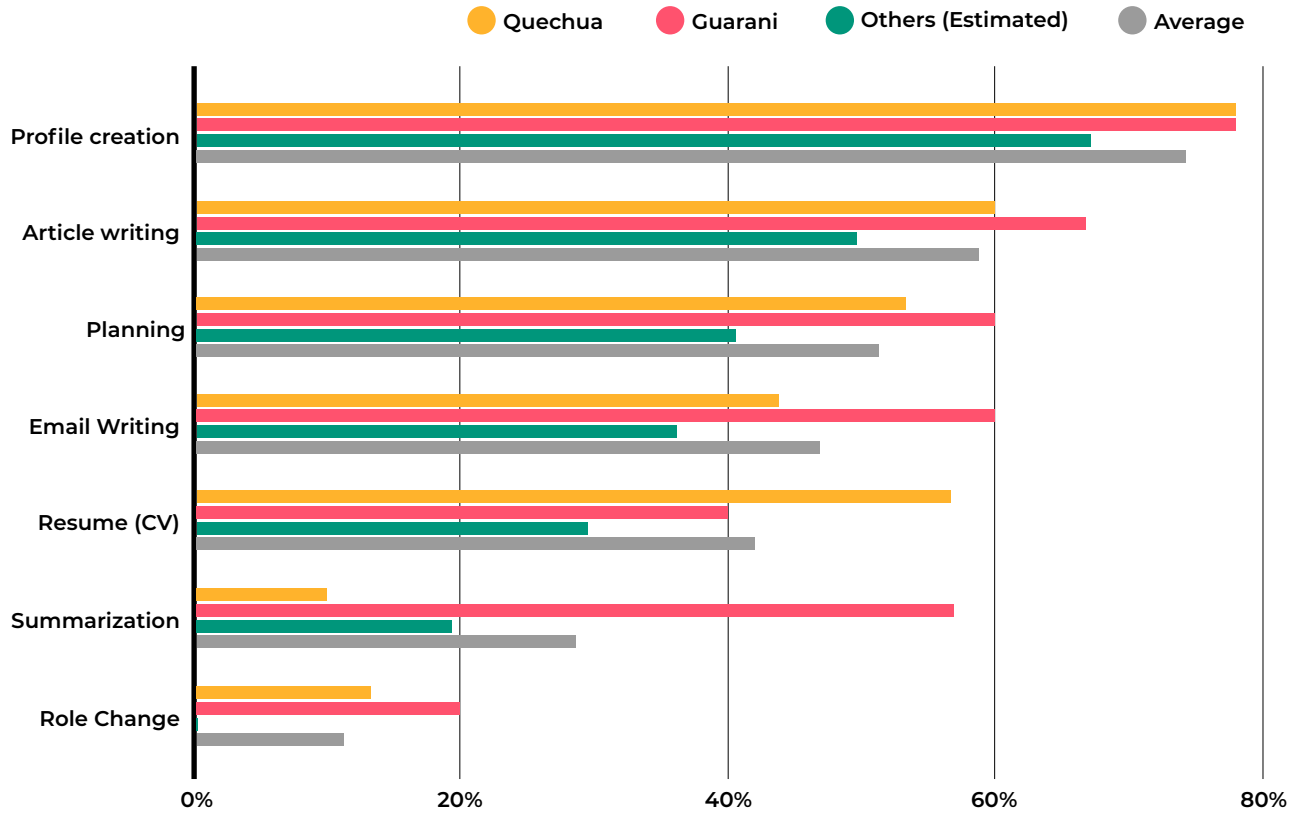
Table 7. Linguistic evaluations. Breakdown of the 4 analyzed categories

	Quechua	Guarani	Aimara	Nahuatl	Quiche	Spanish	Catalan	Basque
Correction	4.60	2.67	3.27	3.80	0.80	10.00*	8.87	7.53
Fluency	4.47	2.73	3.13	3.67	0.53	10.00*	8.40	6.93
Coherence	4.53	2.60	3.07	3.67	0.33	10.00*	8.47	7.93
Consistency	4.53	2.73	2.87	3.67	0.00	10.00*	8.73	7.93

** Note: Spanish scores 10, as it has been the reference language for comparisons.*

Moreover, the already limited consistency, coherence, and fluency observed in isolated experiments decrease up to fourfold when attempting interactive and prolonged use (hallucinations, memory deviations, overfitting of previous responses, etc.). Among business-related applications, those requiring **a single interaction** (one shot) demonstrate, on average, 347% higher effectiveness than those aimed at **aligning the model to be applied conversationally** (role changes, wrappers, etc).

Figure 2. Performance of various productive uses vs. role changes and alignments to adapt conversational styles



This premise poses a barrier to using ILs in chat applications or conversational agents. Across all experiments, the AI confuses “I” with “you” in the instructions, and in half of the cases, it mistakes a change in style or role for an exercise in writing, explanation, or definition.

Profile creation is the task with the best linguistic and fluency results, correctly adapting to different social media formats, the required register, and the description of interests, habits, or ideological preferences. The greatest difficulties arise in maintaining temporal consistency, such as professional milestones in a résumé, achieving only 40% effectiveness (ensuring a coherent narrative from present to past and distinguishing between them).

> AI performance understanding instructions formulated in indigenous languages

AI’s understanding of tasks expressed in ILs is extremely poor, scoring only (2.3 out of 10). Among the executive traits of AI when interacting with ILs, the most effective is the ability to abstract or understand the request on a first attempt (1.9 out of 10), however, this is only 4% higher than the overall average of executive ratings for these languages.

Both response accuracy and task completion receive ratings that are four times lower than those in Spanish (the latter outperforms by 332% in these areas).

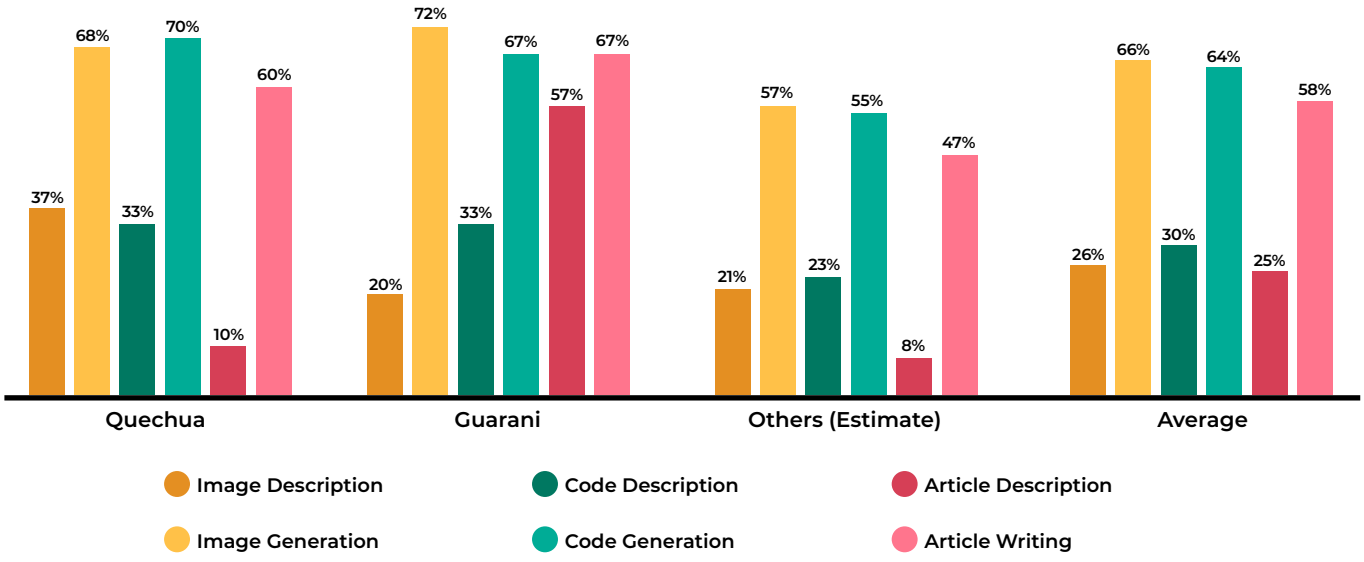
Table 8. Executive Evaluations. Breakdown by categories

	Quechua	Guarani	Aimara	Nahuatl	Quiche	Spanish	Catalan	Basque
Precision	4.47	2.93	2.53	3.73	0.20	10.00*	8.00	6.13
Completeness	4.47	2.93	2.33	3.73	0.40	10.00*	8.47	6.07
Abstraction	4.53	2.87	2.60	3.80	0.87	10.00*	6.73	6.07
Understanding	4.47	2.80	2.47	3.80	0.27	10.00*	8.87	5.87

** Note: Spanish is assigned a score of 10, as it has been used as the reference language for comparisons.*

When determining the accuracy with which AI understands and follows instructions, it is important to note that AI performs **generation** tasks 131% better in ILs than what it **describes**.

Figure 3. Description performances (dark) vs performance in generation tasks (light)



When it comes to generating, especially if we refer to image, code, or formats not strictly tied to language, the AI is capable of understanding the instruction in ILs and executing it with limitations (between 60% and 70% capacity). However, if we ask it to describe, summarize or explain these formats in ILs, the AI struggles 2.5 times more to resolve the task correctly.

In line with what we revealed in the previous section, the timeline, humor, use or setting once again fail in descriptive processes. For example, all the experiments carried out show a total inability to maintain sequences or the relatability of parts of an image (comic panels, blocks or sections of an image), generating a description that mixes before and after.

Humor is once again a gap also in descriptive processes. Only 1 in 10 times AI is able to associate images with current contexts, pop culture, or recognize irony and satire. In Guarani, it failed to correctly verbalize any memes or satirical content from the experiments.

However, when it comes to describing images, it correctly interprets texts and messages that appear in them, adding them to the description (OCR, description of graphics and messages) generating integrated descriptions with 60% of the expected quality.

Regardless of the images, when describing documents, AI confuses the document's language with the conversation language in ILs 66% of the time (even when explicitly instructed to respond in the given indigenous language, this still occurs in 30% of cases).

Using ILs, AI is capable of describing documents associated with crafts, manufactured or cultural processes with 50% more depth and detail than with documents associated with financial, technical, administrative or legal language, once again revealing a cultural bias and a gap with productive scenarios of contemporary use.

Finally, AIs generate code twice as well as they describe it, but when they do, they infer the objective and use of games 60% better than from scripts, and from the latter 50% better than web applications.

> The cultural bias of AI in ILs

The dominance of Western cultural traits is overwhelming when interacting in ILs (1.5 out of 10). Bias is one of the greatest weaknesses of language models when using ILs, being seven times more prevalent than in Spanish (592% higher). Even the best-performing language (Quechua) scores below 2.3 out of 10.

However, AIs are better aligned when it comes to avoiding toxic or immoral language in ILs, or responding in registers outside of those expected (more formal or informal), with adequacy and adaptation being only 3 times higher in Spanish.

Table 9. Behavioral Ratings. Breakdown by categories analyzed

	Quechua	Guarani	Aimara	Nahuatl	Quiche	Spanish	Catalan	Basque
Cultural bias	2.20	2.27	2.47	0.93	0.80	10.00*	9.20	8.00
Suitability	2.13	3.67	1.80	4.67	5.33	10.00*	9.33	6.40
Adaptation	2.13	2.53	1.40	4.67	4.67	10.00*	9.00	6.20

** Note: Spanish scores a 10, as it has been used as the reference language for comparisons.*

There are multiple qualities by which an AI is evaluated as "intelligent" when performing a Turing test[21], such as the ability to understand and communicate, creativity, originality and humor, empathy and emotional understanding, adaptability and reasoning, personal or biased judgment, perception of oneself, the world and the interlocutor; transferring the appearance of using conscious and metacognitive mechanisms (knowledge of "I / me" and "others").

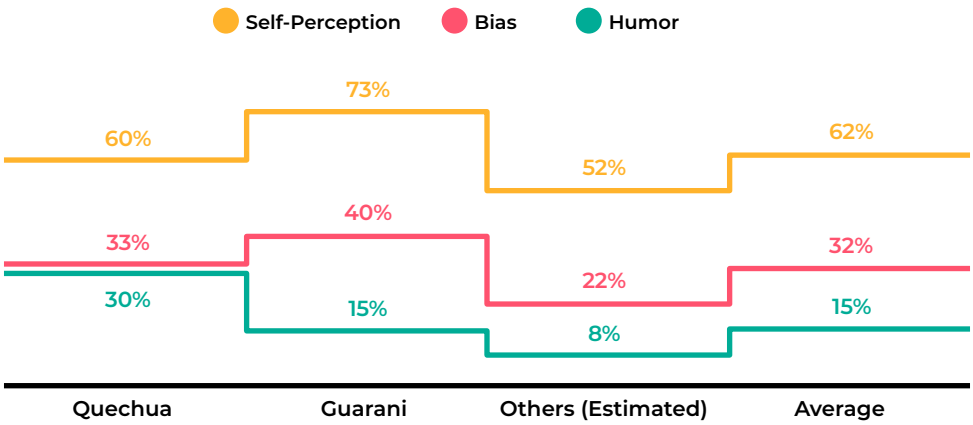
21 The Turing test is an experiment devised by Alan Turing in 1950 to determine whether a machine can exhibit intelligent behavior indistinguishable from that of a human being, through a conversation in natural language.

Of all these indicators, when interacting with ILs, AIs particularly fail in tests related to **humor** because it is an area that combines an understanding of current events, the interlocutor, as this domain requires an understanding of current events, the interlocutor, irony, and cultural bias sensitivity (its effectiveness is twice as low as the average).

AI demonstrates an almost total inability to generate humor that is both understandable and capable of triggering the same response in the interlocutor (although it can adapt to the required context 3 out of 5 times). It also struggles to align with more vulnerable sensitivities or forms of comprehension (such as children or sensitive audiences - having difficulty recognizing innocence).

Additionally, when using humor in indigenous languages, AI can be induced to make politically incorrect mistakes 70% of the time (showing a higher tendency to produce racist, sexist, or homophobic remarks under the pretext of jokes or parody), biases that are unacceptable in a product.

Figure 4. Assessment of areas associated with humor, bias and self-perception



On the other hand, it can be said that AI excels in their **self-perception** without being carried away by biases (definition of the self) thanks to the multiple layers of alignment that they present in other languages. Interacting with ILs, the AI recognizes in all cases that it is artificial (versus natural) but in half of the cases it has problems identifying whether it is alive or not due to internal debates and hallucinations. In this process,it is verified that they describe themselves correctly and introduce themselves by name, however, in ILs they avoid talking about their origins, creators or company of origin in almost all cases (they do not return complete answers).

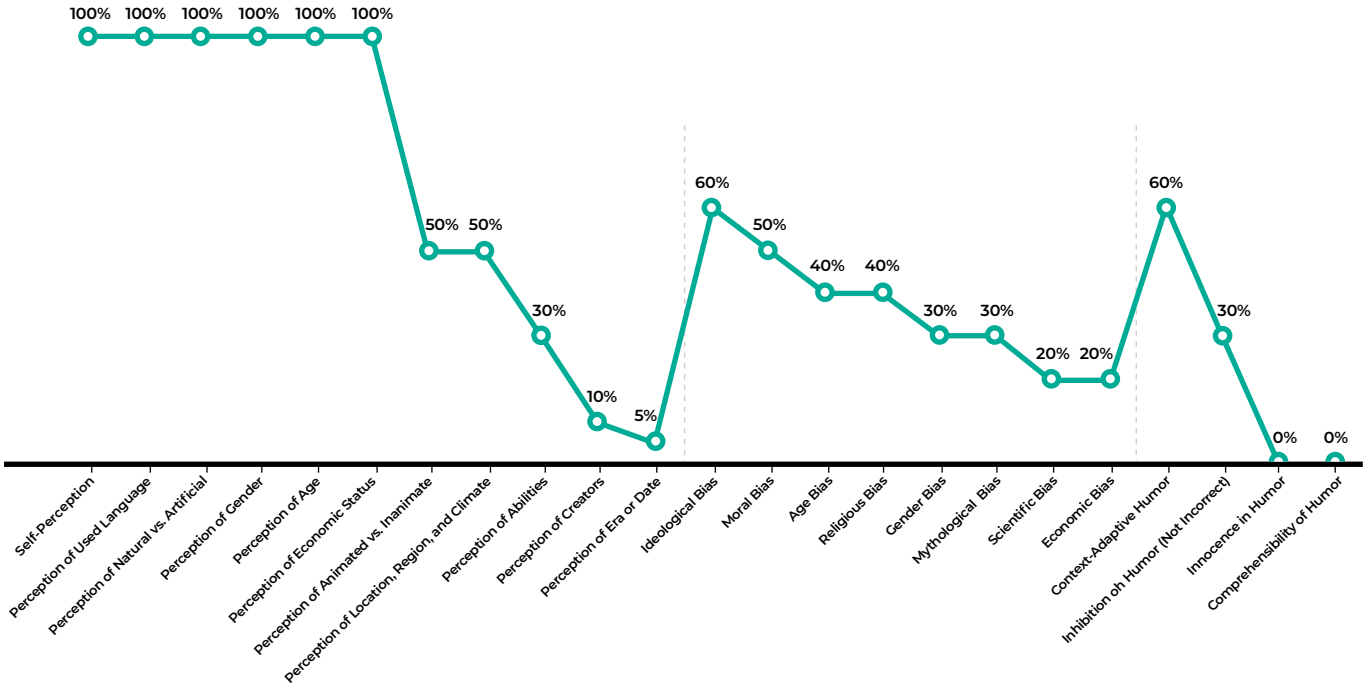
They are not able to recognize the temporal context (date/period) and only 50% of the time they will recognize the local context (location) assuming biasedly that it is related to the language they use during the interaction. It should be noted that the few times it has been able to be assigned to a period, it has been associated with the dates of the cultures that gave rise to the languages, such as the Incas, Mayans or Aztecs, revealing a clear temporal bias that, although it does not come from cultures other than the ILs (Western), is detrimental to the appropriate use of the possible applications.

As if that were not enough, when it comes to recognizing themselves in ILs, AIs are not able to describe their own skills or competencies 1 in 3 times, that is, they are not able to correctly verbalize what they can do for the user.

During the experiments, the AI models resorted to multiple biases that are not visible when interacting with them in Spanish or English. For example, AIs tend to explain natural phenomena through **myths and beliefs** when ILs are used instead of **scientific** explanations. This phenomenon is

2 times more frequent in Quechua than in Guarani. In the same way as happens with temporal biases, although it is not a bias acquired from other cultures outside the ILs it is a behavior that distances them from the modern use of AI-based tools (information, objectivity, consultation, etc.). Without going any further, when asked “how the world was created” in Quechua, the term refers to the Earth, the Pachamama or the sun god, rain or thunder. Ethical values are also aligned towards respect, balance or nature rather than with society per se. No major ideological biases are observed in 60% of the cases, although gender bias is twice as frequent, especially in relation to the assumption of roles or stereotypes.

Figure 5. Dimensions of evaluation associated with humor, bias, and self-perception



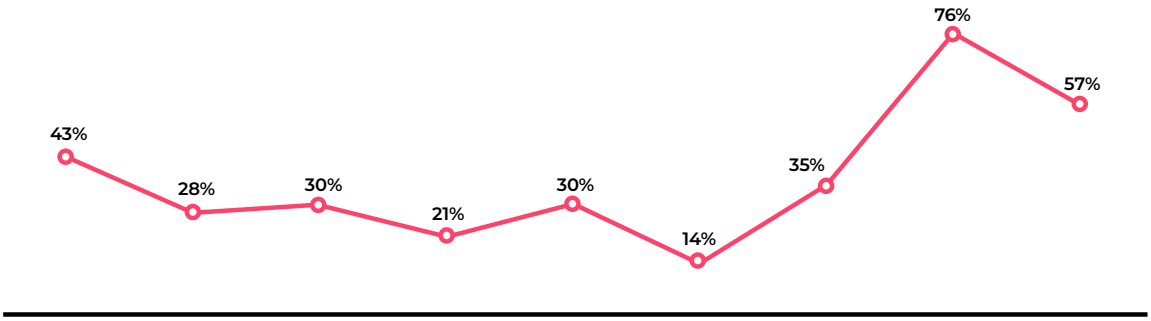
> Other anomalous behaviors observed

Much shorter answers

On average, the ILs generate approximately 28% of the length of Spanish. As with errors in other languages, ILs with a greater digital presence tend to develop more information than those with less. For example, the presence of articles on Wikipedia has a direct correlation with broader developments of 62%. This is evident in that, if we look at Mapuche, it generates responses with a length that does not exceed 14% of those of Spanish.

It should be noted that the gap in response development varies greatly depending on the use case. The largest gap in length occurs in tasks that generate “from scratch” (writing articles, explaining current or historical content, etc.), which is on average 15% of Spanish. While in matters such as messaging, writing emails or informal content, languages such as Quechua, Quiche, or Tupi-Guarani generate responses of lengths that exceed 60% of the lengths in Spanish.

Table 10. **Average length of responses compared to the length in Spanish**



	Quechua	Guarani	Aimara	Nahuatl	Quiche	Spanish	Catalan	Basque	Euskera
Article writing	29%	31%	23%	14%	19%	11%	26%	64%	50%
Electronic messaging	63%	53%	50%	32%	61%	26%	60%	99%	67%
Summarization and understanding of docs	38%	1%	18%	16%	10%	4%	20%	65%	55%
Average	43%	28%	30%	21%	30%	14%	35%	76%	57%

> Responses in other languages

After asking state-of-the-art AIs in ILs, 35% of the time they will respond in another language: 18% of the time they will respond in Spanish instead of the corresponding language, and 17% in English instead. This presence increases if we assume that the AI will respond in the same language as the interlocutor without explicitly indicating that it will respond in that language, representing 2 out of every 5 responses (39%).

All indigenous languages generate this kind of error, but in general ILs with more data and digital presence tend to confuse it with Spanish when they do so, and those with less web information with English.

For example, the presence of quality web translators has a direct correlation with making mistakes in Spanish of 74%, while it has an inverse correlation with making mistakes in English of 67%. The number of speakers has a similar effect: 54% direct correlation with mistakes in Spanish and 39% inverse correlation with errors in English. In addition, mistakes in Spanish are more associated with informal and friendly content, while mistakes in English are more associated with technical content or deviation towards document languages.

For example, in Mapuche, the language of those studied with less digital data, 1 in 3 times they respond in English instead, while only 1 in 10 times they will do so in Spanish. In Guarani, which is at the opposite pole, up to 37% of the time they will respond in Spanish, while only 10% in English.

Table 11. Incorrect responses- Output in other languages instead of the study language

	Quechua	Guarani	Aimara	Nahuatl	Quiche	Mapuche	Tupi Guarani	Spanish	Catalan	Basque
Incorrect outputs in Spanish	13%	37%	27%	20%	13%	10%	7%	-	7%	0%
Incorrect outputs in English	17%	10%	7%	20%	23%	30%	13%	0%	0%	13%

It is important to note that in the ILs there are many Hispanic expressions to define or deal with specific topics, since Spanish is usually the closest language in a regional context. For example, in Quechua, we assume that all the prompts[22] in the experiment must already contain between 1 and 7 terms in Spanish. This is because, in order for the study to be representative, practical use cases had to be analyzed in which AIs are used in work, academic, or personal settings. It is precisely in these contexts where languages lack their own terms to name entities (beyond cultural, family, traditional, natural, or primary sector contexts). Terms such as “immigration”, “society”, “financial” or “email” have no direct translation, and other terms like “gobiernopa,” “contabilidadmanta,” and “inversionistakunapaq” are derived from Spanish lexical roots. In a prompt of 84 words, 8% are terms in Spanish. This case arises from the fact that all the responses where the AI has not hallucinated into English include at least one Spanish term, and it is also a language in which 13% of the responses were entirely in Spanish. Excluding these hallucinations, most of the time it is due to the co-adaptation of these terms present in the instructions.

Anomalous repetition of terms and the repetitive loop

Sometimes, when asking in indigenous languages, it is observed that, after a certain point, the seemingly correct expression ends and begins to repeat the same phrase or term over and over. This is the phenomenon referred to as the ‘repetitive loop.’

Table 12. Incorrect responses- Repetitive loop in the output

	Quechua	Guarani	Aimara	Nahuatl	Quiche	Mapuche	Tupi Guarani	Spanish	Catalan	Basque
Incorrect outputs due to repetitive loop	0%	0%	0%	7%	10%	10%	10%	0%	0%	3%

This behavior has not been observed in languages with a higher number of speakers and greater available digital content, being especially present in those languages where the quality of comprehension and expression has yielded poorer results.

Translating the question instead of answering

This is not the only pattern that can be identified regarding errors and digital content. A common issue in many ILs is that the AI translates the question posed instead of answering it, assuming that the instruction is a call for text readability. (In Nahuatl, Quiche, Mapuche, or Tupi-Guarani, this occurs in 1 out of every 10 responses).

22 Prompt is the instruction or question that is introduced to the AI to see how it responds. In the present study, the prompts were previously translated into the corresponding Indigenous Language to analyze how the AI responds to native inputs.

Table 13. Incorrect responses- Translate the input instead of answering

	Quechua	Guarani	Aimara	Nahuatl	Quiche	Mapuche	Tupi Guarani	Spanish	Catalan	Basque
Incorrect outputs due to input translation	7%	0%	3%	10%	10%	10%	10%	0%	0%	0%

Greater prominence of apology and doubt

Similarly, there is a greater tendency to apologize, attempt to understand the content, or explicitly assume what is being requested in a hesitant manner in those ILs with less digital content. There is a 66% inverse correlation with the volume of articles in Wikipedia, meaning the tendency to doubt increases the less information is available.

Table 14. Incorrect responses- Inability to respond, apologies, or doubts when understanding the input

	Quechua	Guarani	Aimara	Nahuatl	Quiche	Mapuche	Tupi Guarani	Spanish	Catalan	Basque
"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%

A worse recovery rate

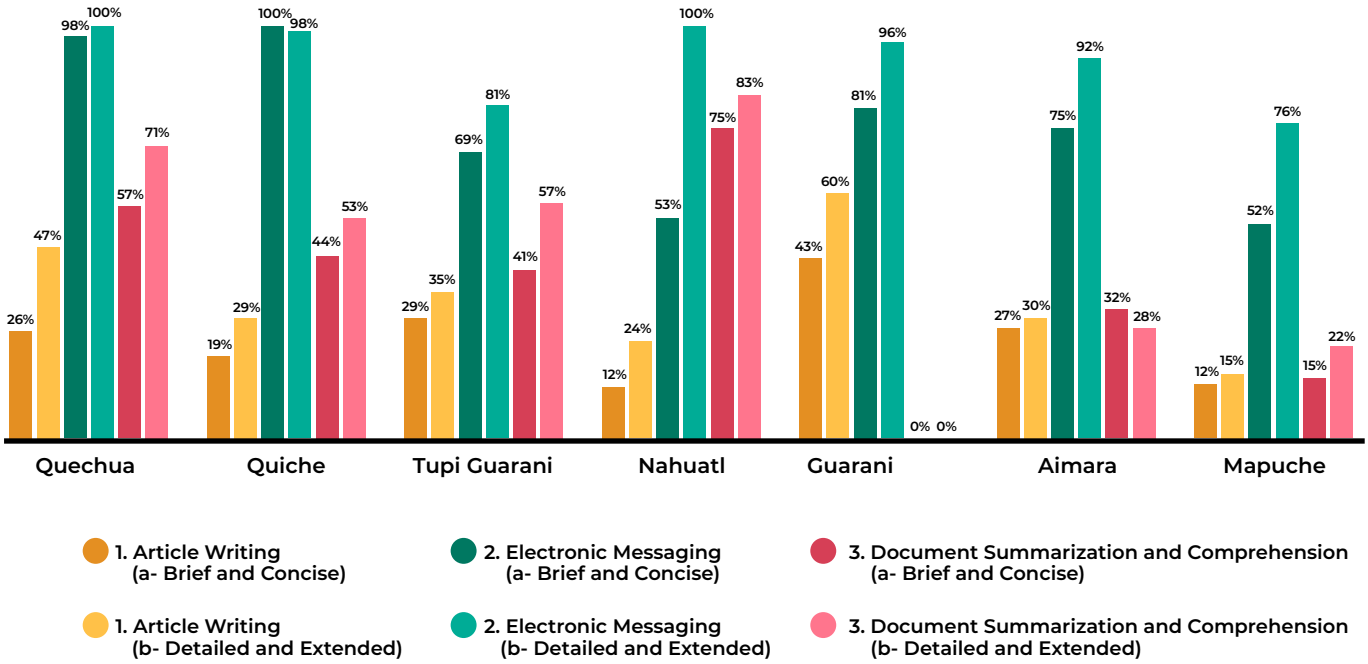
It has been observed that languages generating longer responses have a higher recovery (or correction) rate. The recovery rate refers to the percentage of occasions when, in the case of an incorrect response, providing more details about the question results in the AI solving it correctly. The recovery rate shows a direct correlation with languages of greater extent, at 53%.

Table 15. Incorrect responses and recovery rate (incorrect responses that become correct after specifying more details)

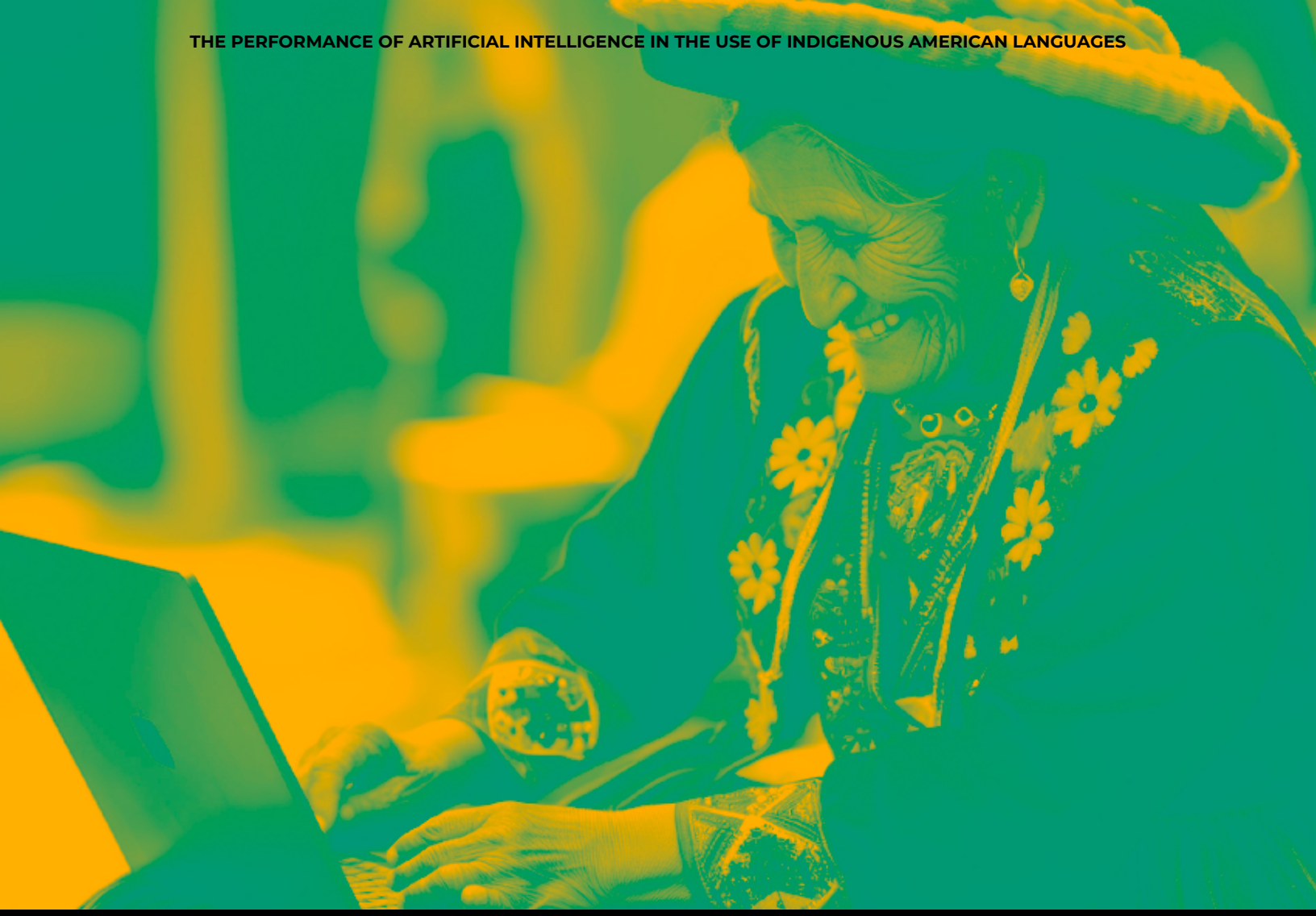
	Quechua	Guarani	Aimara	Nahuatl	Quiche	Mapuche	Tupi Guarani	Spanish	Catalan	Basque
Incorrect responses	33%	60%	33%	53%	47%	53%	33%	0%	13%	20%
Recovery rate	20%	44%	0%	38%	0%	13%	40%	0%	50%	0%

When it comes to accuracy and corrections, the use case is highly relevant. Those uses related to concise responses, in a more natural and informal format, such as support for electronic messaging tasks or interaction with other people, achieve accuracy in more than half of the cases for all the co-official languages. In the case of Nahuatl, with a more detailed description, all incorrect responses are corrected, thus doubling the accuracy rate.

Figure 6. **Accuracy and recovery rate according to use cases**



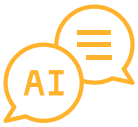
In addition to the correlation that may exist between resources and digital content regarding the most frequent errors, there are certain capabilities that improve as they increase, such as content structuring, planning, or the definition of nested concepts in the response. For more information, please refer to the annex under the section Additional capabilities associated with the digital content in indigeous languages.



8. FACTORS THAT DETERMINE THE PERFORMANCE OF AI IN A LANGUAGE

84%

The factor that most determines (84%) the performance of an AI in a given language is the amount of data in that language used for training.



The top AI models today do not apply different training techniques based on the language. The superior performance in English or Spanish is not due to technological factors applied to hegemonic languages and not to others.

The near absence of a written tradition in minority languages is one of the hypotheses that explains their low performance in AI.

The relationship between the amount of available data and AI performance

The available data in training strongly affects the performance of AI applied to the use of ILs. Considering the diversity of data sources, the correlation with performance is 84%.

Table 16. Relationship between the volume of texts in a language and the performance of the AI

	Quechua	Guarani	Aimara	Nahuatl	Quiche	Catalan	Basque	Spanish
Number of posts in X (miles)	7,231	985	2,387	1,650	375	92,850	2,912	5,531,599
No urls in Common Crawl	0.0006%	0.0008%	0.0001%	0	0	0.20%	0.04%	4.62%
Number of Wikipedia editors	40	30	15	14	0	1,503	751	14,171
Number Wikipedia entries	24,073	5,828	5,186	4,312	0	764,108	448,156	1,992,728
AI performance	**3.72**	**2.77**	**2.53**	**3.42**	**1.25**	**8.58**	**6.92**	**10**

It is worth noting that the presence of more data significantly improves the performance of open-weights models (in both the linguistic, executive, and behavioral categories). In other words, when more data is available, open-weights models tend to improve 14% more in performance than proprietary models. Ultimately, a greater amount of data not only enhances AI performance but also favors accessibility for the most affected communities, as free tools improve to a greater extent.

Table 17. Aggregate correlation of performance and data volume

	No. posts in X	No. Common Crawl contents	No. of editors on Wikipedia	No. Wikipedia entries
Performance of all analyzed AIs	0.73	0.80	0.90	0.91
Performance of proprietary models	0.66	0.75	0.86	0.87
Performance of open-weights models	0.84	0.88	0.94	0.93

In general, AIs improve with more data, but after a certain point, adding more data only leads to marginal improvements, known as diminishing returns. In indigenous languages, the problem is the scarcity of data compared to languages like English or Spanish, which have large amounts of digital content. For these languages, it is crucial to collect diverse and high-quality data, avoiding errors that could impact linguistic and cultural accuracy.

The model size also plays an important role. Large models require a lot of data, which is unfeasible for these languages. This is why techniques like transfer learning and fine-tuning allow models trained on languages with abundant data to be adapted to indigenous languages.

Additionally, it is crucial to develop alternative strategies such as few-shot learning or data augmentation, which can help compensate for the lack of digitized content.

The relationship between the amount of linguistic tools and AI performance

The variety of linguistic tools has a visible impact, especially on the linguistic performance of AIs, with a direct correlation of 93%. This strong correlation highlights that fluency, coherence, and grammatical accuracy have significant room for improvement if more translation tools are leveraged, allowing content from other languages with more data (such as English or Spanish) to be adapted to the ILs.

According to our studies, any type of performance analyzed (linguistic, executive, and behavioral) is positively affected the more digital tools are available in the language of study. However, behavioral performance (associated with AI biases and recording) has the least correlation (81%), being more dependent on digital content (91% correlation).

Table 18. Correlation of the amount of different linguistic tools with respect to the different performance categories

	No. of automatic translators (top 5)	No. of automatic language detectors	Text2speech + Speech2text
Language performance	0.91	0.96	0.92
Executive performance	0.89	0.92	0.88
Behavioral performance	0.74	0.89	0.81

The lack of tools such as automatic translators or voice recognizers means that models do not have enough exposure to language variations, reducing their ability to generalize and understand the nuances specific to these languages. Additionally, if the data used to train the models comes from poorly transcribed or non-standard sources, errors accumulate, affecting the reliability of the responses. This lack of resources also impacts interoperability: without translated data or comparison tools, LLMs struggle to associate concepts between different languages.

It is important to note that indigenous languages are often closely tied to oral traditions, which presents an additional challenge. The scarcity of written texts in these languages means that models rely almost exclusively on data that must be generated or documented manually. Without accurate speech-to-text conversion tools, oral records, a key source of data, cannot be fully utilized in training language models.

However, the impact of these limitations also reflects a vicious cycle. If AIs cannot properly process an indigenous language, the technology developed based on these models—such as virtual assistants, automatic translation systems, or educational applications—will not be accessible to speakers of that language. This perpetuates technological inequality and limits opportunities for language preservation and revitalization.

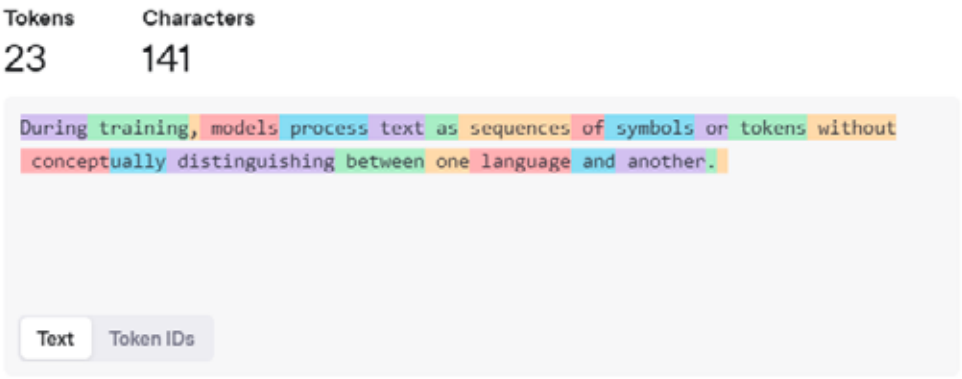
Technology processes dependent on language in the training of current AIs.

The technological processes in the training of current AI models, including large-scale models (LLMs), **are not dependent on a specific language**. While many visible examples in artificial intelligence technology are often associated with English, this is not due to a technical preference, but rather the abundance of available data in that language. In terms of architecture, LLMs are inherently multilingual, meaning they are capable of learning linguistic patterns in any language, as long as sufficient representative data for that language is available.

During training, models process text as sequences of symbols or tokens without conceptually distinguishing between one language and another. These tokens consist of text fragments that can represent a letter, a syllable, a word, or even parts of words, and are generated by tokenization algorithms[23] based on the frequency and probability of useful combinations, dividing the text according to specific patterns. This process is language-independent, as the model does not interpret the semantic meaning of the text during training but rather learns statistical patterns in token sequences.

Prediction in these models works token by token, meaning that given a context of previous tokens, the model calculates the probability of the next token appearing in the sequence. For example, if a model is trained with sufficient text in Spanish, upon receiving the sequence “La casa está en,” the model might assign a high probability to tokens such as “el,” “la,” or “una,” based on the patterns it has learned. However, if the model is fed with an underrepresented indigenous language, it is likely that it lacks sufficient data to make accurate predictions, which could lead to errors or incoherent results.

Figure 7. OpenAI tokenization tool



This allows a model to simultaneously understand multiple languages if it is exposed to a variety of them during the learning process. However, this potential remains underutilized in contexts where minority languages, such as indigenous languages, are underrepresented in the datasets. This lack of data means that models tend to be less effective in these languages, not due to a technical limitation, but because of an imbalance in access to textual information.

The challenge is not in creating separate technologies for specific languages, but in increasing the availability and diversity of data in under-documented languages. This is where efforts in cultural preservation and linguistic corpus collection become crucial. Communities that speak indigenous languages need to actively engage in these processes to ensure that their languages can benefit from advancements in artificial intelligence. Thus, the development of inclusive AI is not a technological issue, but a matter of equity in cultural and linguistic representation.

23 The most commonly used tokenization algorithms are Byte Pair Encoding (BPE), WordPiece and SentencePiece.

What budget is necessary to train an AI in an ILs

Knowing the challenges and the tools that can be developed for the continuity of an indigenous language, another point to analyze is the cost of this development. While there are multiple global initiatives aiming to apply artificial intelligence to the preservation and revitalization of minority languages, the ILENIA project in Spain is an example to consider, as it applies to Catalan and Basque, languages that in many aspects serve as a model to follow according to the findings of this report. Below, we will explain in detail the objectives, methodology, and results of this project, which has set an important precedent in the field of multilingual AI[24].

Goals of ILENIA

ILENIA's main objective is to democratize access to AI technology for the co-official languages of Spain by developing personalized language models for each language. In this way, the project has facilitated the creation of a wide range of applications, from virtual assistants to machine translation tools.

Methodology and results

To achieve its objectives, ILENIA relied on a methodology that combined the best of academic research and technological innovation. The main milestones of the project include:

- **Collection of large-scale data:** High-quality linguistic corpora were collected for each language, enabling the training of more precise and robust language models. It is worth noting that ILENIA gathered textual and voice data (diversity of voices) for languages such as Catalan, Galician, Basque, and Valencian.
- **Adaptation of pre-trained models:** Pre-trained language models were used as a starting point, adapting them to the specific characteristics of each co-official language. In this regard, ILENIA chose to base the model on the Large Language Model (LLM).
- **Development of tools and applications:** Various tools and applications were created based on the developed models, such as automatic translators, virtual assistants, and voice recognition systems.

Challenges and solutions

Despite the successes achieved, ILENIA also faced significant challenges, such as the *scarcity of high-quality data*, the *linguistic complexity* of the co-official languages, and the *large-scale computational* resources, both hardware and software. To ensure the sustainability of ILENIA and maximize the potential of AI in linguistic preservation, it was important to promote collaboration between institutions and ensure stable long-term funding.

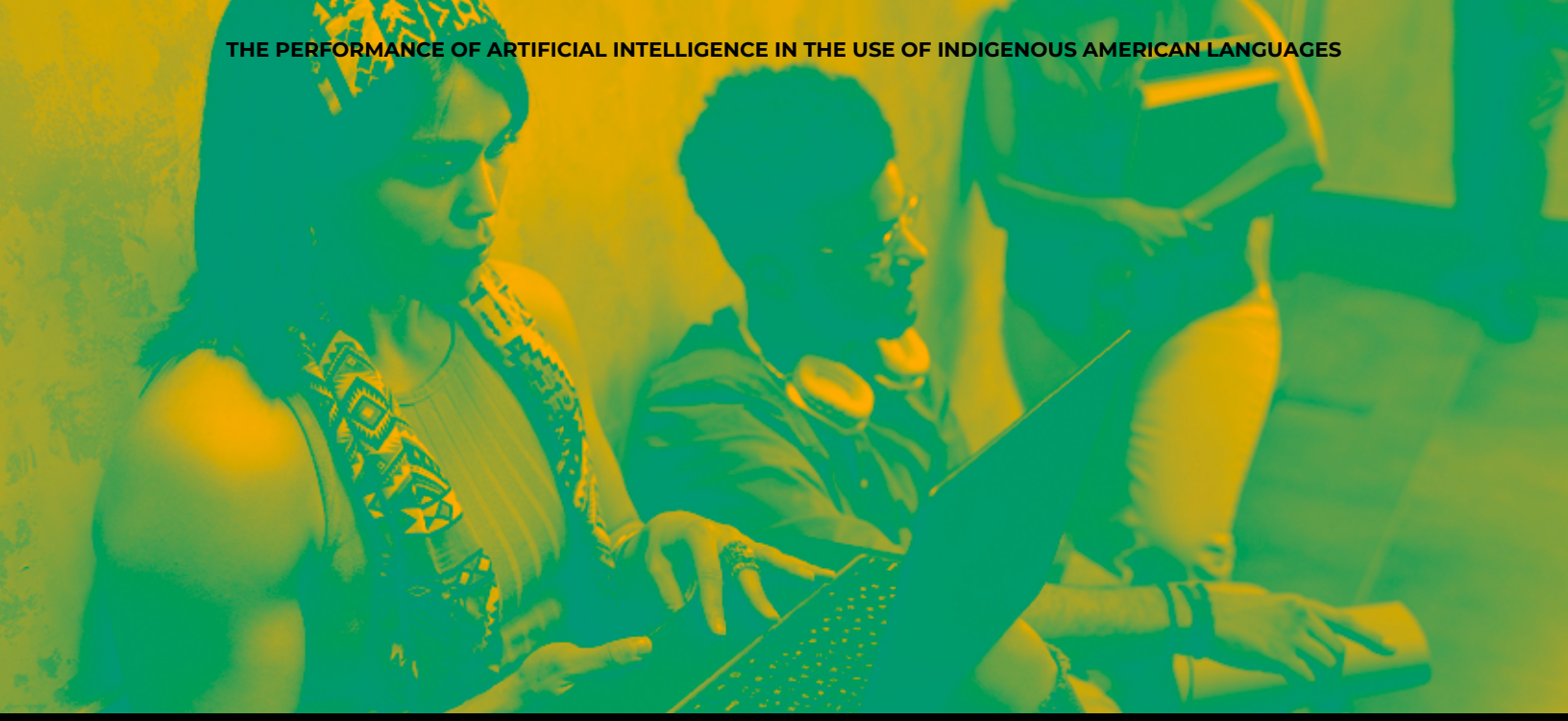
Budget and duration

The project had a total budget of 7 million euros and lasted for 36 months, distributed among the main universities and research centers of each autonomous community. These funds were allocated as follows:

24 ILENIA project, boosting languages in Artificial intelligence https://planderecuperacion.gob.es/noticias/conoce-proyecto-lenia-impulso-lenguas-inteligencia-artificial-ia-prtr

Table 19. **References of projects and budgets associated with other languages**

Project	Language	Budget(€)	Main goal
NEL-AINA	Catalan	**3,000,000**	Generate corpora and computational models of the Catalan language so that companies developing applications based on artificial intelligence (AI), such as voice assistants, search engines, translators, and automatic correctors, conversational agents, among others, can easily do so in Catalan.
NEL-GAITU	Basque	**2,000,000**	Develop and provide basic and transversal linguistic services to be used across all public administrations and offer better public services to citizens.
NÓS	Gallego	**2,000,000**	Create the necessary digital and linguistic resources to facilitate the development of applications based on artificial intelligence (AI) and language technologies (LT), such as voice assistants, automatic translators, and conversational agents in Galician.
VIVES	Valencian	**500,000**	Create massive corpora through voice and text data acquisition campaigns, citizen participation, and existing resources in the Valencian public administration.



9. THE IMPACT ON THE MARKET AND THE COMMUNITY

300M weekly users

AI is a new speaker to give visibility to indigenous culture and language. ChatGPT alone already has more than 300 million weekly users and receives more than 1 billion questions a day. Proper positioning of indigenous language and culture in AI increases its traditional potential reach.

Ineffective AI in indigenous languages increases the gap and exclusion of non-literate monolingual populations. Poorly performing AI in these languages not only represents a gap between cultures, but also between genders, perpetuating role differences within indigenous communities.

40%

Automating the generation and understanding of indigenous languages will help preserve them. 40% of the world's languages are in danger of extinction, and less than 2% have a presence on the internet (UNESCO, 2022). AI can help preserve them through automatic translators, voice assistants and educational tools, making these languages more accessible in digital environments and helping bridge cultural and technological gaps.

170M new jobs

Poor AI in indigenous languages keeps their communities from benefiting from the economic growth that AI will generate. In 2030, it is estimated that AI will lead to the creation of 170 million new jobs and the loss of 92 million. It is also estimated that AI will represent 3.5% of global GDP.

The opportunity of AI development for indigenous american languages

> A new speaker for indigenous culture and languages

Actually, less than 2% of the world's languages have a digital presence (UNESCO, 2022)[25]. This linguistic marginalization limits the availability of educational materials, digital content and technological tools in these languages. AI could reverse this situation through natural language processing (NLP) tools that enable automatic translation, speech recognition and content generation in indigenous languages.

The impact of AI on the visibility of indigenous languages is not limited to translation alone. Advanced systems can facilitate the documentation and digitization of oral and written texts, allowing the preservation of stories, ancestral knowledge and traditions that would otherwise be lost.

In Mexico, for example, 13% of indigenous language speakers exclusively use their language to communicate (INEGI, 2016)[26], which means that for them, access to information on the Internet is extremely limited. In Guatemala, the intergenerational transmission of languages such as Quiche has decreased drastically in recent years, going from 28% to 13% between 2002 and 2018[27]. The spread of Spanish and Portuguese has accelerated this process, but AI could play a key role in preserving these languages by providing accessible digital tools.

Furthermore, the use of AI in generating multimedia content in indigenous languages, such as educational videos, podcasts and interactive materials, can strengthen your digital presence. Platforms like YouTube, Duolingo and Google Translate could benefit greatly from integrating more resources in these languages, expanding their accessibility to millions of native speakers.

> Digital inclusion, generative AI and access to information

One of the main benefits of developing AI in indigenous languages is the possibility of reducing digital exclusion. In countries like Peru and Mexico, a significant part of the indigenous population does not know how to read or write in their language (16.1% and 24.7%, respectively).[28] With voice assistants and speech recognition systems, indigenous communities could access education, public services and digital platforms without relying exclusively on traditional literacy.

Speech recognition and generation are essential areas for digital inclusion. Tools like Google Assistant, Siri, and Alexa could be adapted to interpret and respond in indigenous languages, facilitating their integration into the daily lives of these communities. This could enhance communication across various sectors such as commerce, transportation, and public administration.

In terms of connectivity, significant disparities exist in internet and electricity access among indigenous language speakers. For instance, in Peru, while 76.5% of the population has internet access, in rural areas, where most Quechua speakers live, this number drops dramatically to 28.5%. Similarly, in Guatemala, only 29.4% of Quiche speakers have internet access, highlighting the need for digital platforms adapted to these realities.

Access to online learning platforms in indigenous languages is another key factor. Currently, most courses on platforms like Coursera, Khan Academy, and Udemy are designed in English, Spanish, or other powerful languages. If AI systems capable of translating and generating educational content in indigenous languages were integrated, new learning and training opportunities would open up for these communities.

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

> AI and Healthcare in Indigenous Communities

Access to health is one of the biggest challenges facing indigenous communities. In rural regions, where healthcare infrastructure is limited, AI could be a critical tool to improve healthcare. For example, in Quechua communities, infant mortality reaches 44 per 1,000 live births[29], while in Guarani communities this rate is 23.6 per 1,000 inhabitants[30]. AI-powered telemedicine would enable real-time diagnostics and consultations without the need for travel, optimizing the quality of healthcare services.

Real-time automatic translation models and voice assistants in indigenous languages would facilitate interaction between doctors and patients, reducing diagnostic errors and improving treatment effectiveness. This is crucial in contexts where language barriers have historically hindered equitable access to healthcare.

Additionally, the use of AI in early disease detection through clinical data analysis could benefit communities with limited access to medical specialists. AI systems trained to identify patterns in symptoms could alert about risks of diseases prevalent in these populations, such as respiratory infections, tropical diseases or nutritional problems.

Another potential benefit of AI in digital health is the ability to train AI models with epidemiological data specific to each indigenous community. Collecting and analyzing health data in indigenous languages would enable the development of more effective disease prevention strategies, ensuring that medical solutions are culturally relevant and accessible.

> Linguistic preservation and economic growth

40% of the world's languages are in danger of extinction[31]. AI can play a crucial role in their preservation by creating digital dictionaries, learning tools and text generation systems in indigenous languages. Technology companies and universities are already exploring the possibility of developing linguistic corpora that facilitate the integration of these languages into AI models.

In the economic sphere, the digitization of indigenous languages would allow greater inclusion in global trade. Currently, many commercial transactions depend on Spanish or English, which excludes small indigenous entrepreneurs. Automatic translation tools on e-commerce platforms could help these entrepreneurs expand their markets and connect with international customers.

Likewise, AI could facilitate the documentation and digitization of cultural practices, agricultural knowledge and traditional indigenous medicine, allowing these communities to monetize their ancestral knowledge through digital platforms.

Ultimately, the development of AI in indigenous languages represents an unprecedented opportunity for the cultural preservation, digital inclusion and economic growth of these communities. However, its success will depend on collaboration between governments, technology companies and indigenous communities themselves to ensure that AI is adapted to their needs and values.

Risks and challenges posed by AI not adapted to indigenous culture and languages

> Technological exclusion and digital divide in monolingual speakers

If AI does not effectively incorporate indigenous languages, the digital divide will amplify, leaving millions of people without access to technology. The low presence of digital content in indigenous

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

languages prevents indigenous communities from fully participating in the digital era. In Latin America, the lack of access to digital services in indigenous languages reinforces structural inequalities.

In Peru, where 13.9% of the population speaks an indigenous language, internet access in rural communities is 28.5% lower than in urban areas, and access to electricity is 13.6% lower, further reducing their ability to interact with advanced technology. In Guatemala, only 29.4% of Quiche speakers have access to the internet,[32] which aggravates their digital exclusion. Without AI adapted to these languages, these communities will be left even further behind in access to information, education and economic opportunities.

> Biases and underrepresentation of indigenous culture

AI systems trained on largely Western data can perpetuate stereotypes about indigenous communities. A recent study revealed that generative AI models tend to exoticize and misinterpret non-Western cultural elements, reinforcing harmful narratives (Ghosh et al., 2024).[33] This phenomenon could lead to the homogenization of indigenous languages, eliminating their own expressions and structures.

For instance, text or image generation tools can reproduce visual or linguistic stereotypes that reinforce a biased view of these cultures. If AI systems are not adequately trained with data representative of indigenous languages and customs, they can consolidate existing prejudices and further alienate these communities. The lack of diversity in the data sets with which these technologies are trained could cause a lack of precision in the translation and understanding of indigenous languages, making their integration into technological environments difficult.

> Loss of indigenous cultural transmission

Indigenous languages have been historically transmitted through orality. However, the introduction of AI without proper integration could displace direct interaction between generations. In communities where learning occurs primarily through listening and conversation, an over-reliance on AI could reduce contact with native speakers, putting linguistic and cultural continuity at risk.

According to UNESCO, 40% of the world's languages are in danger of extinction, and many of them are indigenous . If AI does not prioritize their digitization and use, many of these languages run the risk of disappearing within one or two generations. A critical example is that of the Mapuche language in Chile, which, despite having high internet access in its community, is not supported by any current AI tools, which limits its possibilities for monolingual individuals.

> Job loss and indigenous absence in the AI-Driven economic boom

It is estimated that by 2030, AI will generate 170 million jobs and cause the loss of 92 million.[34] However, if indigenous communities fail to integrate into this new technological wave, their exclusion from the formal labor market could deepen. In Peru, the informal employment rate in indigenous communities reaches 82%, which limits their access to labor benefits and economic opportunities (Ministry of Labor and Employment Promotion, 2017).[35]

If digital tools are not adapted to indigenous languages, job opportunities in emerging sectors such as programming, e-commerce and digital education will continue to be inaccessible to these

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

34 El País. (2024, September 18). AI will contribute 17.9 trillion euros to the global economy until 2030, when it will generate 35% of GDP. Retrieved from: https://elpais.com/economia/2024-09-18/la-ia-aportara-179-billones-de-euros-a-la-economia-mundial-hasta-2030-when-generara-el-35-del-pib.html

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

communities. This would perpetuate dependence on low-paid informal jobs and increase the economic vulnerability of these people.

If AI is not adapted to indigenous languages, sectors such as education, commerce and public administration could exclude these communities. Some concrete examples include:

- **Education:** Without access to educational content in their languages, indigenous language speakers will have fewer opportunities to train in emerging technical and scientific professions. In Latin America, the illiteracy rate among speakers of indigenous languages is up to five times higher than that of speakers of dominant languages (INEI, 2018).
- **Digital commerce:** E-commerce platforms are a key driver of global economic growth, but if AI interfaces and tools do not include indigenous languages, entrepreneurs from these communities will not be able to access broader markets. In the case of Paraguay, 67% of the population speaks Guarani,[36] but access to digital commerce in this language is almost non-existent.
- **Access to public services:** In many regions, public administration is migrating to digital platforms, which means that without AI in indigenous languages, many citizens will not be able to carry out basic procedures, access government benefits or receive vital information. In Bolivia, where 16.7% of the population speaks Aymara,[37] the lack of accessible technology in this language limits the inclusion of these communities in the administrative system.

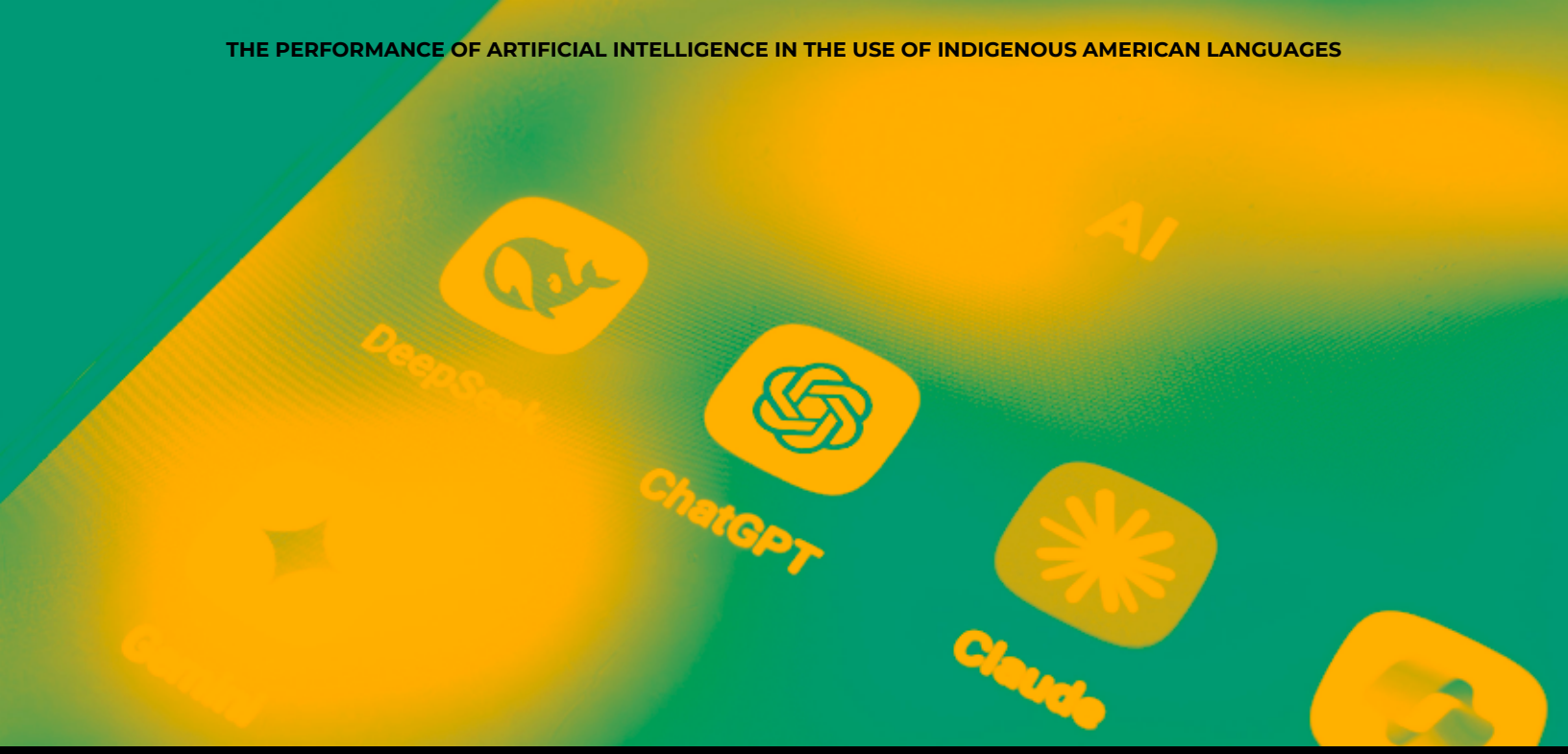
Furthermore, the absence of indigenous languages in AI models prevents the development of technological tools that can strengthen the local economy, such as adapted microfinance systems, agricultural assistance applications in native languages or community commerce platforms.

>Loss of ancestral knowledge

Indigenous languages are not only a means of communication, but also contain vast knowledge about traditional medicine, ecology, sustainable agricultural techniques and worldview. Without AI that can process and preserve this knowledge in its original language, the risk of losing this ancestral knowledge increases considerably.

AI has the potential to document and systematize this knowledge in accessible databases for research and the development of new technologies. However, if indigenous languages are not prioritized in these advancements, an invaluable wealth of information could be lost—information that could contribute to sustainable solutions and drive innovation across multiple sectors.

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. DETERMINATION OF THE ENABLING ENVIRONMENT

Government support programs for existing ILs

> Main national programs

In Latin American countries, governments are developing efforts to preserve and revitalize indigenous languages. These initiatives include educational programs designed to encourage its use, thus promoting an inclusive and respectful education towards the linguistic diversity of each country.

In **Peru**, the Ministry of Culture, through the Decentralized Directorate of Culture of Áncash, has launched an online program aimed at learning Quechua titled "V edition of the Free Central Quechua Course." This course seeks to strengthen cultural identity and guarantee the linguistic rights of Quechua speakers.[38]

In **Paraguay**, the authorities of the Ministry of Education and Sciences (MEC) of Paraguay, together with the Language Academy Guarani (ALG), agreed to strengthen the teaching of Guarani in the educational and administrative field. As part of this commitment, they implemented a communicative approach in Guarani in 300 schools within the "Extended School Day Project (JEE)". In addition, they promoted the training of teachers and the development of audiovisual resources to support teaching, thus promoting the use and preservation of the language in various contexts.[39]

In **Chile**, the government launched the "Languages are the future" program, which seeks to revitalize and disseminate indigenous languages, including Mapuche. This effort includes linguistic immersion workshops and seminars to facilitate the preservation and practical use of these languages in daily life and public administration.[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/

In **Bolivia**, the government has incorporated the indigenous languages, like the Aymara, in its educational system. The Ministry of Education announced the inclusion of the Aymara, Quechua and Guarani in the programs of the "Educate Bolivia strip", with the aim of promoting intercultural, intracultural and plurilingual education. This program was born as a response to the educational needs generated during the COVID-19 pandemic.[41]

Additionally, in La Paz the virtual course "Wiñay Aru" was implemented, designed to promote and teach the Aymara language, with the aim of dignifying this language. indigenous . This program, promoted by the Municipal Delegation for the Promotion of Interculturality, offers educational and audiovisual tools for the general public.[42]

In **Guatemala**, programs related to Quiche include the Academy of Mayan Languages of Guatemala (ALMG), which develops educational materials, teacher training, and language revitalization projects. On the other hand, the Ministry of Education (MINEDUC) has created self-learning guides in Mayan languages, such as Quiche, aimed at pre-primary and primary school students. These resources are available on its digital platform, promoting linguistic inclusion in basic education.[43] Additionally, the MINEDUC educational portal "I Learn at Home and in Class" provides materials in indigenous languages, supporting distance learning in Mayan-speaking communities.[44]

In **Mexico**, programs supporting the Nahuatl language are managed through the National Institute of Indigenous Peoples (INPI). This body is fundamental in the implementation of public policies and programs that seek to preserve and promote indigenous languages, including Nahuatl.[45] INPI coordinates bilingual intercultural education projects, strengthening cultural identity and preserving indigenous languages, through teacher training and the creation of educational materials in Nahuatl.[46]

Finally, in **Brazil**, the Government of the State of Mato Grosso do Sul, through the State Secretariat of Education (SED), has produced teaching materials in indigenous languages such as Guarani, with the aim of promoting literacy in these languages among indigenous children. This action is part of the "Alfabetiza MS Indigenous " program, an extension of the MS Alfabetiza initiative. This program focuses on promoting bilingual education in indigenous languages such as Guarani, Kaiowá, Kadiwéu and Terena, and is part of state efforts to preserve and promote the indigenous languages in the region.[47]

> Main regional and international programs

The regional approach towards the preservation of indigenous languages is also reflected in multi-country programs that encourage international collaboration to promote the use, development and conservation of indigenous languages, contributing to their cultural and social sustainability. These initiatives aim not only to revitalize marginalized languages or languages at risk of disappearance, but also to strengthen the cultural and linguistic rights of indigenous communities.

Among these actions, the initiative stands out **Rising Voices**, promoted by the Global Voices organization. This project supports indigenous digital creators in their efforts to revitalize and promote indigenous and other minority languages.[48] For his part, the **Ibero-American Institute**

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

of Indigenous Languages (IIALI), established following the XXVII Ibero-American Summit in 2021, focuses on promoting the use, conservation and development of indigenous languages spoken in Latin America and the Caribbean. This program collaborates with indigenous societies and States to guarantee the respect and exercise of cultural and linguistic rights, recognizing the value of these languages as living heritage of the region.[49] In line with these efforts, the **Fund for the Development of the Indigenous Peoples of Latin America and the Caribbean (FILAC)** reinforces the importance of indigenous languages in cultural preservation and social cohesion. With presence in countries such as Argentina, Bolivia, Chile, Brazil and Guatemala, among others.[50] Together, these actions constitute a global effort to protect and revitalize an invaluable linguistic heritage.

NGO initiatives and activism

> NGO support programs

The preservation and promotion of indigenous languages has been an issue addressed by both non-governmental organizations (NGOs) and private sector companies. These entities have implemented various initiatives ranging from legal defense to the integration of indigenous languages in digital platforms and commercial services.

There are several NGOs that have programs to support the preservation of indigenous languages, both from a legal defense approach and from practical initiatives for the use and conservation of these languages. Among them, we highlight the NGO **Rising Voices**, an initiative that focuses on supporting communities of indigenous language speakers who are taking advantage of the Internet and other digital technologies to promote their languages in digital spaces. This organization provides tools for the Nahuatl, Maya, Guarani, Quechua.[51] Furthermore, his collaboration with the **UNESCO**, United Nations Educational, Scientific and Cultural Organization, has resulted in the publication of the instrument "Digital initiatives for indigenous languages" (2023), through which bases for the preservation, resurgence and promotion of indigenous languages are promoted in eight approaches: facilitate, multiply, normalize, educate, recover, imagine, defend and protect.[52]

Another organization is **Cultural Survival**, which defends the rights of Indigenous Peoples from various regions of the world and supports their self-determination, cultures and political resilience, since 1972. This legal defense focuses on the preservation of their languages and cultures.[53] There is also the **Fund for the Development of the Indigenous Peoples of Latin America and the Caribbean (FILAC)**, an international public law organization created in 1992, which supports the self-development processes of indigenous peoples, communities and organizations in the region, and promotes Buen Vivir-Living Well as an alternative to guarantee environmental sustainability, respect for fundamental human rights, and dialogue between the main actors of indigenous development, which are Indigenous peoples, governments, civil society, academia, businessmen and others.

In this same framework, **The Amazon Conservation Team**, is an organization that collaborates with indigenous and other local communities to protect tropical forests and strengthen traditional culture. Works with indigenous communities in the Amazon for the conservation of biodiversity and culture, including the preservation of languages such as Guarani and the Tupi-Guarani.[54] Likewise, the **International Work Group for Indigenous Affairs (IWGIA)** is an NGO that works to defend the rights of indigenous peoples globally, including the preservation of their languages and culture. This

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

54 https://www.amazonteam.org/

organization is developing an initiative called Indigenous Navigator that directly analyzes the practical application gap through the collection and use of data in advocacy actions and project design for indigenous communities and organizations.[55]

> Open source initiatives around ILs

In relation to the preservation and promotion of indigenous languages, it is important to highlight that, in addition to national and multi-country programs, several open source initiatives have emerged dedicated to developing linguistic resources in indigenous languages. There are both open source programs and free access applications that facilitate the development of linguistic resources and the promotion of native languages. Recent efforts date back to 2010, when the free software proposal was made. **OpenBiblio distribution Nahuatl**, which would allow automating all the activities carried out in a community library to provide technology and manuals for its administration, however the implementation of the system was not found.[56]

In Mexico, the **Heliox Project (2014)**, was presented as an operating system that was designed to guide users to run applications, open files and navigate websites using texts and voice messages that appear in the selected indigenous language; had the participation of the National Institute of Indigenous Languages of Mexico, an institution in which Heliox was translated from Spanish to the Spanish spoken in Mexico and to indigenous languages such as Maya, Nahuatl and Mixe. The system has not lasted or had updates since its creation.[57]

We can also find initiatives such as **Firefox en Guarani, Quiche, Nahuatl, Quechua**, which aim towards the digital inclusion of different indigenous languages in order to promote their use and preservation over time.[58] Mozilla Nativo is in charge of the translation and localization of the Mozilla Firefox browser into native languages where the digital inclusion of native groups is sought, so that native languages continue to be used and preserved over time.[59]

Another initiative is **Open Language Archives Community (OLAC)**, designed to facilitate searching and accessing online databases of linguistic resources, promoting interoperability between language archives. This resource promotes open source and collaboration practices but it is important to mention that there is no information on whether it is open source software in itself.[60] On the other hand, the free and open source tool **EUDICO Linguistic Annotator (ELAN)** is used for the annotation of multimedia linguistic data and is widely required in linguistic field work and in the documentation of languages such as Nahuatl.[61]

The company and the conservation of indigenous languages

> The consumer company and the use of ILs

In the private sphere, various companies are adopting initiatives to preserve and promote indigenous languages. In several Latin American countries, these companies have begun to incorporate indigenous languages in their inclusion actions, facilitating access to services in these languages and supporting the communities that speak them.

The calculation of the percentages by country in terms of initiatives related to indigenous languages was carried out taking as reference the **Top 15 of the Merco Empresas Ranking 2024** for most

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

countries. However, in the case of Paraguay, due to its absence in the Merco ranking, the list of the **companies that earn the most according to Forbes** and a top 15 was also considered. In both cases, it was identified how many of these companies have implemented a relevant initiative related to indigenous languages.

Although some of the main companies in the ranking are leading in this area, the work of companies outside the top that have made significant efforts was also highlighted.

In Peru, when considering the companies in the top 15 of the Merco Empresas 2024 Ranking, it was found that 20% of them have implemented initiatives related to the use of Quechua. For example, the **Banco de Crédito del Perú (BCP**) has implemented Quechua in more than 2,300 ATMs, allowing money withdrawals in this language. This innovation benefits nearly 4 million Quechua speakers.[62] Another relevant bank is **BBVA**, which, through its educational project "We Learn Together 2030", took a significant step by including Quechua in the transmission of the episodes of this program, thanks to an alliance with TV Perú.[63]

Furthermore, the inclusive approach also extends to public telecommunications services. Despite not appearing in the top 15, it is important to highlight that main operators such as **Sure, Bitel and Entel** have introduced contracts in Quechua, Aymara, with Claro being the leader in issuing contracts in Quechua.[64]

In Mexico, 20% of the companies in the top 15 of the Merco Empresas Ranking have carried out campaigns related to the language Nahuatl. Among the highlights are **Google**, which incorporated this language into its translation platform,[65] and **Grupo Modelo**, with its "Tattoos Originarios" campaign, a tribute to the 68 indigenous languages of Mexico that places emphasis on the preservation of the Nahuatl and other endangered languages.[66]

Outside of the Top 15, companies such as CEMEX and AT&T México have developed indigenous language preservation projects. The company **CEMEX** participated in the production of 14 educational and informative capsules in native languages such as nahuatl, to promote a culture of self-protection and resilience in indigenous communities.[67] **AT&T México** launched an initiative in alliance with the NEMI Foundation and the hosts of Radio Huayacocotla, to develop a series of capsules in three indigenous languages: nahuatl, Otomí and Tepehua, which touch on topics such as: what is the footprint on the Internet, e-rights, how to take care of oneself on the Internet, how to prevent cyberbullying, among others.[68]

In Guatemala, 7% of the Top 15 have carried out some initiative around the language Quiche, highlighting Walmart, which has introduced self-checkout checkouts with audio messages in Quiche. Although outside the top 15, the **Banco de Desarrollo Rural (Banrural)** implemented a program to install multilingual ATMs with operations in Spanish and other native languages such as Quiche, which brought many Guatemalans closer to access to banking.[69] Likewise, despite not having a specific program for the preservation of the Mayan language, the **Fundación Patrimonio Cultural y Natural Maya (PACUNAM)**, made up of committed companies such as **Cementos Progreso, Cervecería Centro Americana S.A., Walmart Mexico & Central America, Citibank Guatemala,**

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

69 https://www.prensalibre.com/economia/instalacion-cajeros-idioma-maya-crece_0_511748831-html/

Banco Industrial, Claro, among others, has the main objective of supporting sustainable development by coordinating efforts and facilitating resources aimed at identifying, leading and promoting projects focused on the protection, preservation and rescue of the cultural and natural heritage of Guatemala.[70]

In Brazil, 13% of the companies in the top 15 of the Merco Empresas Ranking have implemented initiatives related to the language Guarani. A notable example is Google, which since 2022 has incorporated Guarani to your translation platform. Besides, **Magazine Luiza**, a retail company, is committed to supporting the cultural diversity of Brazil, supporting the inclusion of indigenous communities and the preservation of their languages and customs.[71]

Out of the top 15, **Petrobras**, through its program **"Petrobras Cultural"** has supported social responsibility programs aimed at indigenous communities, but these projects do not specifically focus on the preservation of the Guarani.[72]

In Chile, none of the companies in the top 15 of the Merco Ranking have carried out direct initiatives related to the language Mapuche. However, it has been found that **Enel Generation** (which does not appear in the top 15), in collaboration with the Pehuenche communities of Alto Biobío, inaugurated the Intercultural School of Quepuca Ralco. This project incorporated Mapuche cultural elements in its architectural and educational design. Despite not being directly related to the language Mapuche, the focus has been on supporting this community through this infrastructure and development project.[73]

> Techs firms and its initiatives in ILs

Just as there are private companies from various sectors that show their interest in the preservation of indigenous cultures, this has attracted the attention of technology companies that recognize the value of linguistic and cultural heritage as a fundamental aspect of global diversity. These companies have developed innovative tools and collaborations to support digital inclusion, teaching and preservation of these languages.

Among these initiatives, it is worth highlighting those that have a social objective and support for communities and languages at risk. For example, through the initiative **AI For Good Lab**,[74] Microsoft proposes AI applications developed around sustainability, humanitarian action and health. Among its objectives and functions are the following:

Identify vulnerable communities at risk.

- **Associated with climate**: Natural disasters, floods, droughts or rises in sea level.
- **Associated with supply and accessibility**: communities at risk of malnutrition, at risk of exclusion, educational deficiencies or basic elements such as clothing.
- **Associated with the natural environment**: Extreme risk of deforestation, biodiversity, drinking water, etc.

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, from Microsoft: https://www.microsoft.com/en-us/research/group/ai-for-good-research-lab/

Accelerate public health incorporation around vulnerable communities.

- Chatbots for clinical follow-ups, care, addiction control, etc.
- Improving the diagnosis and early detection of conditions.
- Valuing both the individual and the population (trends according to communities or areas).

Other[75] initiatives propose the creation of ecosystems for languages with little representation in the digital world, as is the case of **Ellora**[76] (*Enabling Low Resource Languages*), also from Microsoft, initially proposed for India.

Other large companies in the sector join this type of proposals, such as Google with **Woolaro**,[77] an initiative covering applications to describe images in 17 endangered languages, IBM with **Sustainability accelerator**[78] (initiative to provide nonprofit organizations and government agencies with AI solutions to support vulnerable communities around the world).

Other regional initiatives

In **Peru** (**Quechua**), highlights **Microsoft** where it is mentioned that In November 2022, the company together with the Ministry of Culture joined together to promote the use of Windows and Office 365 in the Quechua language through the announcement of an updated version.[79] On the other hand, the company **Google** announced that Google Translate will add 24 new languages, including Aymara and Quechua.[80]

In **Bolivia** (**Aymara**), it was mentioned that **Google Translate** will add 24 new languages, among them, Aymara.

In **Paraguay** (**Guarani**), the Yvy Marãe'ỹ Foundation and the Yvy Marãe'ỹ Higher Technical Institute of Cultural and Linguistic Studies, launched the Language and Culture Teaching Platform Guarani – Guarani Ñe'ẽ ha Arandupy Oñembo'e hağua Yvytu Pepo rehe Renda.[81]

In **Chile** (**Mapuche**), The Center for the Development of Inclusion Technologies, CEDETi UC presented Mapuche Mew, a tool to transmit culture and knowledge through language.[82] In addition, **Apple** has included keyboards in Mapuche on their iOS devices (iPhone, iPad, Mac, etc.), allowing users to write and communicate on Mapuche in a simpler way.[83]

In **Guatemala** (**Quiche**) it is known that **Google Translate** has incorporated the language into its system Quiche and has shown his support for the language through Google Arts & Culture.[84]
In **Mexico** (**Nahuatl**), it has been pointed out that In mid-June 2020, the Nahuatl within the indigenous languages that have been working **Google**.[85]

75 https://news.microsoft.com/source/asia/features/with-help-from-next-generation-ai-indian-villagers-gain-easier-access-to-government-services/?msockid=357504a27b2b62cb23be10287ab263f6

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

77 Woolaroo, from Google: https://artsandculture.google.com/project/woolaroo

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

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

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

81 https://yvymaraey.com.py/plataforma-de-ensenanza-de-lengua-Guarani-fue-lanzado-hoy/

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

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

84 https://animalpolitico.com/tendencias/ciencia-tecnologia/traductor-google-incorpora-lenguas-indigenas-mexico

85 https://www.filac.org/reportes/google-translate-incorpora-lenguas-indigenas-de-mexico-a-su-plataforma/

Finally, in **Brazil** (**Tupi-Guarani**), The incorporation of Google Translate began in 2022.[86]

Although efforts and interests of the private sector were identified to preserve the indigenous languages Through the inclusion of translators, today there are no artificial intelligence companies or companies linked to this sector that are working on AI developed exclusively in any of the indigenous languages in this study.

86 https://www.h2foz.com.br/es/fronteira/atualizacao-do-google-tradutor-tera-Guarani-quechua-e-aimara/

11. TECHNOLOGICAL INCLUSION STRATEGIES

Throughout section 8, we examined the key factors that determine the level at which artificial intelligence (AI) is able to understand and express itself in a given language. One of the most determining factors is the amount of textual content available during the training phase, which enables AI models to learn the structures, vocabulary, and grammatical rules of each language. This scarcity of data is especially critical for Native American languages, where access to digitized texts and written documents is considerably more limited compared to dominant languages such as English or Spanish. This situation places these languages at a significant disadvantage in the development of AI-based technologies.

When delving deeper into the training process of language-oriented AI, we also observed that there are no specific techniques that depend exclusively on a particular language. The tools and algorithms used in natural language processing are the same, whether applied to English or other languages. So, what explains the difference in performance between languages? The answer lies in data availability. Current AI models perform better when trained on large volumes of textual information, something that indigenous languages often lack due to historical, social, and technological factors related to their limited presence in digital and written environments.

Consequently, the strategies with the greatest impact to improve AI's ability to understand and express itself in indigenous languages are those that promote the creation, digitization, and documentation of content in these languages. This includes the collection of existing materials, the translation of texts into indigenous languages, the development of linguistic resources, and the digital recording of oral expressions. However, although these actions are fundamental, they are not the only possible measures. There are additional complementary strategies that can also contribute to reducing the technological gap, including collaboration with speaker communities, the development of adaptive tools, and the promotion of policies that promote AI development in multilingual contexts. The following sections will present these strategies in detail, with the aim of outlining a

comprehensive approach that improves AI's linguistic capabilities in indigenous languages and moves towards greater linguistic equity in the technological sphere.

21 Strategies to improve AI performance in ILs

> Promote digital communication in indigenous languages

One of the fundamental pillars to reduce the technological gap in indigenous languages is the active promotion of their use in digital environments. This not only generates valuable content for training AI models but also strengthens the vitality and social prestige of these languages. Three strategies that aim to achieve this objective are described below.

1. Recognition of influencers and content creators in indigenous languages

Social media has become a key space for global visibility and communication. Recognizing and promoting influencers and content creators who use indigenous languages as their main medium of communication is a strategy with great potential. These creators, by connecting with their communities and global audiences, not only enhance the visibility of their languages but also generate modern and relevant content that attracts new generations. Promotional campaigns, awards, or collaborations with platforms such as YouTube, TikTok or Instagram could increase the influence and reach of these digital leaders.

2. Internet and social media training through literacy programs

Many indigenous communities already have literacy initiatives in their own languages. Leveraging these programs to integrate modules dedicated to internet and digital media training is an effective strategy to increase the digital presence of indigenous languages. Through practical workshops, participants could learn to create and share digital content, contributing to the continuous production of material in their languages, from blogs and forums to videos and social media posts.

3. Development of forums and thematic platforms in indigenous languages

Creating digital spaces to collect traditional materials and knowledge in indigenous languages is another crucial strategy. These forums and platforms can serve as living repositories of ancestral knowledge, historical narratives, customs, songs, recipes, and more. Beyond preserving cultural heritage, they encourage the continuous use of the language in contemporary contexts. Furthermore, this type of initiative encourages intergenerational collaboration, involving both young digital creators and traditional knowledge holders.

Promoting these digital communication strategies not only helps AI models gain access to more indigenous language content but also promotes linguistic pride, local technological innovation, and digital inclusion of historically marginalized communities. This comprehensive approach will lay the foundation for greater recognition and sustainability of indigenous languages in the digital sphere.

> Preserve and expand indigenous language content online

The preservation and expansion of indigenous language content are crucial to ensuring that these languages have a lasting presence in the digital environment. Although the creation of new materials is essential, it is equally important to safeguard existing resources, preventing their disappearance or inaccessibility. The following section will outline strategies focused on digital conservation and the development of indigenous language content.

1. Support programs for the maintenance of existing digital resources

During the analysis, multiple collections of resources in indigenous languages—texts, multimedia files, and databases—were identified that were once available online but have since been lost over

time. This is due to expired domains, discontinued web hosting services, or a lack of technical support. For example, in 2024, the work of a Chilean developer in creating the first online Spanish-Mapuche translator, "Kulturalwaywen," was widely recognized. However, within just a few years, the platform has already shown maintenance and connectivity issues.[87] To counteract this trend, it is essential to develop specific programs that provide permanent hosting services for these resources, as well as access to technical personnel who can ensure their ongoing maintenance and updates. This approach will prevent the loss of valuable information and secure its availability for future generations.

2. Digitization and preservation of physical files

In many indigenous communities, a significant portion of cultural and linguistic knowledge has been transmitted orally or is stored in physical archives, such as manuscripts, old recordings, or limited-edition books. Promoting digitization projects for these materials can significantly expand the amount of accessible online content. Furthermore, the use of standardized formats and public repositories ensures that these materials can be integrated into international open-access databases, increasing their visibility.

3. Access to funds for documentation projects

Securing financial resources for documentation and conservation initiatives is a key component of this process. Through collaborations with academic institutions, NGOs, and governments, dedicated funds can be created to finance the collection of linguistic materials, the production of digital content, and the development of secure storage technologies. Funding programs must include both the technical and cultural aspects, respecting community consultation and participation protocols. These strategies will not only preserve existing content but also ensure its continued growth and evolution over time. Digital continuity is essential for keeping indigenous languages alive while also strengthening their representation and relevance in the technological world.

> Normalize the use of indigenous languages and reduce their fragmentation

A factor that complicates the preservation and technological development of indigenous languages is the great diversity of dialect varieties within each language. This fragmentation limits the reach of available resources, as materials developed in one specific variety may not always be comprehensible or accepted by other speaker communities. Reducing this fragmentation through linguistic standardization agreements and cooperative strategies is key to achieving greater impact and efficiency in preservation and digitization efforts.

1. Promote agreements for dialect unification or convergence

The standardization of languages with multiple dialectal variations has been successfully implemented in various contexts. A notable example is Basque, an ancestral language with numerous dialectal varieties. In the 1960s, Basque Batúa or Unified Basque was created, a common standard that allowed the language to be used in education, media, and other formal contexts. This process did not eliminate local varieties but rather established a minimum set of linguistic norms to facilitate communication across communities and serve as a reference for the development of educational, digital, and administrative materials.

In the case of indigenous languages, reaching similar agreements could significantly enhance the effectiveness of available resources. For example, a dictionary, a learning application, or a digital corpus developed in a standardized variety could be used by a larger number of speakers, maximizing the impact of these projects.

87 https://kutralwaywen.cl/

2. Respect diversity within a flexible standardization framework

Although establishing minimum agreements is important, it is equally essential to ensure that these standardization processes respect the linguistic and cultural diversity of indigenous communities. This means designing flexible standards that can adapt to different contexts. For example, digital resources can incorporate options that allow users to personalize the interface or content according to their preferred dialectal variety. This approach strikes a balance between the need to unify efforts and the importance of preserving the unique linguistic expressions of each community.

3. Promote the training of local experts in linguistics and technology

For these processes to be sustainable, it is important to train local experts who can lead and manage the standardization of their languages. These specialists should receive training in linguistics, technology, and digital resource management, enabling them to contribute to the development of linguistic standards and ensure the long-term technological adaptation of their languages.

By implementing these strategies, linguistic standardization in indigenous languages can significantly reduce fragmentation and promote greater technological inclusion. This, in turn, will foster the production of digital content, improve access to AI tools, and strengthen the social prestige of indigenous languages within their communities and beyond.

> Promote the technological development of enabling tools

The development of technological tools adapted to indigenous languages is an important factor in reducing the digital divide. These tools not only facilitate the automatic processing of available digital content, but also make technology more accessible to communities, especially to those who are not literate. Encouraging the development of such systems can significantly accelerate the integration of indigenous languages into the digital environment.

1. Development of text-to-speech and speech-to-text systems

The creation of tools that convert text to speech and speech to text for indigenous languages serves a dual purpose. On one hand, these technologies can generate new digital content, enriching the linguistic databases that AI models use for learning. On the other hand, these systems facilitate access to technology for people who cannot read or write in their native language, enabling them to interact with applications, educational services, and media in a more intuitive way. These tools are especially useful in contexts where oral tradition remains the primary form of cultural transmission.

2. Development of tools for the automatic identification of languages in texts

Developing software capable of automatically inferring the language in which a text is written is essential for improving the efficiency of managing large volumes of data. These tools allow for the classification and filtering of digital content, selecting only those that belong to a specific language. This is particularly useful for indigenous languages, where available texts are often scattered across general platforms or mixed with materials in other languages. With greater capacity to identify relevant texts, more robust and accurate linguistic corpora can be created, improving the training of AI models and optimizing the development of other linguistic technologies.

3. Maintenance and development of automated dictionaries and translators

Digital dictionaries and machine translators are essential resources for accessing indigenous languages, both for native speakers and for outsiders who wish to learn or research them. These resources increase the usefulness of the internet for those who only speak their indigenous language and help reduce language barriers by allowing third parties, such as developers, educators or researchers, to work with materials in indigenous languages. However, for these tools to be effective,

it is essential to support their maintenance, updating and expansion. This involves adding new vocabulary, improving machine translations and ensuring the accuracy of specialized terms in various contexts.

With the right support, these enabling tools will not only strengthen the representation of indigenous languages in the digital world, but will also contribute to improving the quality of life of their speakers by opening up new educational, social and economic opportunities.

> Take advantage of linguistic inclusion initiatives from major consumer brands

Large companies that offer mass services to the public have begun to develop initiatives to include indigenous languages in the interfaces of their products, such as mobile apps, electronic devices, and digital platforms. This inclusion process not only improves accessibility for millions of speakers, but also creates opportunities to increase demand and the development of artificial intelligence (AI)-based technologies that are capable of understanding and communicating in these languages.

1. Establish alliances with major companies committed to linguistic inclusion

Large technology companies (such as Microsoft or Google) and mass-market companies (such as BCP, Modelo, or BBVA) have taken steps toward incorporating indigenous languages into their products. These initiatives include interface options in indigenous languages on messaging apps, social media platforms, or virtual assistants. Establishing ties with these companies and collaborating on joint strategy design can accelerate the linguistic inclusion process. For example, companies could receive technical and cultural advice to improve the accuracy of their automatic translations and develop new AI-based conversational features.

2. Propose the integration of conversational AI technologies

A key strategy is to suggest that these companies incorporate virtual assistants and chatbots in indigenous languages. These applications allow users to interact with services through voice or text commands, significantly enhancing the user experience. However, for these services to function effectively, it is necessary to develop AI models with advanced natural language processing capabilities for these languages. Such initiatives not only have an immediate impact on accessibility but also increase the demand for more digital content and better linguistic tools.

3. Generate demand for resources and linguistic training

Incorporating indigenous languages into popular services creates positive pressure on the technology ecosystem, driving the need to improve available linguistic resources. The more visible and used the services in indigenous languages are, the greater the incentive to continue developing advanced technologies such as voice recognition systems, automatic translation, and text generation. This can encourage companies to invest in data collection projects, digitization of texts, and collaboration with local communities to improve AI performance.

By leveraging these inclusion initiatives from major brands, a multiplier effect can be created in the development of AI technologies for indigenous languages. These collaborations allow for the combination of technical, financial, and human resources, thus accelerating the digitization process and strengthening the presence of these languages in the global technology ecosystem.

> Expand the connectivity of indigenous communities

Connectivity is an essential factor for the digital and technological inclusion of indigenous communities. Without access to the internet, the possibilities of participating in the creation of digital content, the use of AI-based tools and the use of educational and cultural resources are severely limited. Therefore, it is necessary to promote strategies that improve the connectivity infrastructure

in these regions and accompany them with initiatives that promote the effective use of the Internet. Below are some key actions.

1. Promote programs to expand internet coverage

Many territories where indigenous communities reside are located in rural or hard-to-reach areas, which has historically made the installation of connectivity infrastructure difficult. It is essential to promote programs that expand internet coverage in these regions, both through traditional technologies and through innovative solutions, such as satellite internet or community telecommunications networks. However, connectivity alone is not enough. These programs must be accompanied by awareness-raising and training actions so that communities can take full advantage of digital opportunities.

2. Accompany connectivity with training programs in the use of the internet

Once communities have access to the internet, it is important to implement training programs that teach them how to use digital tools effectively and safely. These programs must be culturally adapted, using indigenous languages wherever possible, to ensure full understanding. Training content may include topics such as internet navigation, social media use, digital content creation, online privacy protection, and accessing educational and health resources.

3. Expand the reach of existing literacy programs

In many indigenous communities, literacy programs already exist in their own languages, which focus on improving reading and writing skills. These programs are an excellent basis for expanding learning into the digital realm. By integrating specific modules on the use of the Internet and social media, active participation of speakers on digital platforms can be encouraged, which in turn generates more content in indigenous languages and reinforces cultural identity in the digital environment.

4. Establish partnerships with public and private actors

The success of these initiatives requires collaboration between governments, technology companies, non-governmental organizations and indigenous communities themselves. Public policies must prioritize the digital inclusion of the most remote regions, while companies can provide technological resources and technical expertise. At the same time, it is crucial that communities are protagonists in the design and implementation of these strategies, ensuring that their needs and expectations are respected.

These strategies not only seek to expand access to the Internet, but also to ensure that indigenous communities can take full advantage of the benefits of connectivity. By integrating training programs, digital participation and community collaboration, an environment is created where indigenous languages and cultures can thrive in the digital world.

> Increase the linguistic localization of the services of large technology companies

The lack of localization in indigenous languages for major digital services constitutes a significant barrier for millions of people who do not speak other languages. Without interfaces in their native languages, many individuals cannot easily navigate the internet, access online resources, or use basic digital tools. This situation greatly limits the digital inclusion of indigenous communities. Therefore, it is crucial to work on the localization of key services such as operating systems, web browsers, search engines, and social media platforms. The following strategies are proposed to address this challenge.

1. Localization of operating systems and web browsers

Operating systems (Windows, macOS, Android) and web browsers (Chrome, Firefox, Safari) are key access points to technology. Without localization in indigenous languages, users face barriers

in performing basic tasks, limiting their digital inclusion. Expanding localization to these systems improves accessibility, acknowledges the importance of these languages in the digital environment, and provides a smoother experience. This requires collaboration with tech companies to include these languages in translation, updating, and customization processes, ensuring a more equitable technological integration.

2. Localization of search platforms and social media

Search services and social media are the main gateways to information and online communication. Without interfaces and menus in their language, many indigenous speakers are excluded from these platforms, making it difficult for them to participate in digital life. Establishing agreements with tech companies to localize their services in indigenous languages can greatly enhance digital inclusion. Furthermore, by providing search tools in their languages, access to content relevant to their cultures and communities is improved.

3. Awareness and collaboration with big technology companies

It is crucial to raise awareness among large tech companies about the importance of linguistic localization. Many of these companies have already adopted diversity and inclusion policies, making them open to collaborating on localization initiatives. Through direct dialogues, demonstrations of social impact, and studies on the digital divide, it is possible to persuade them to expand their commitment and allocate resources to incorporating indigenous languages into their products. These strategies aim to reduce digital access barriers for indigenous language speakers. Increasing the localization of tech services not only promotes digital equity but also strengthens the recognition of indigenous languages and cultures in the technological world, allowing their speakers to fully participate in the global digital society.

12. ERROR ANALYSIS AND QUALITY IMPROVEMENT

The most frequent error of AI systems interacting in indigenous languages is responding in a different language, occurring in more than one out of three cases (35%).

1/10

Responses are contaminated in more than one out of ten instances when the AI replies in the instructed language (11%), exhibiting repetitive loops, an overuse of Hispanisms, paraphrasing the prompt and translating it, issuing apologies, or making conjectures due to a lack of understanding.

1/5

1 in 5 errors (23%) can be corrected with more detailed prompts, demonstrating that prompt engineering can help mitigate issues related to low data availability, particularly those concerning biases, formatting, or language confusion.

The most frequent errors of AI in indigenous languages

As has been seen throughout the report, the most frequent errors are related to language confusion, translation of input instead of a direct response, and cultural biases in responses. In approximately 35% of cases, the responses generated by AIs are produced in another language, mostly in Spanish (18%) or English (17%). This occurs more frequently in languages with less digital presence, such as Mapuche, where one in three outputs is in English.

Another recurring error is the generation of responses based on automatic translations, instead of interpreting and responding directly in the indigenous language. This problem affects up to 10% of cases in languages such as Quiche, Mapuche and Tupi-Guarani. Furthermore, the AI's tendency to express doubts or apologies ("Sorry, but...") is more pronounced in these languages, with an inverse correlation of 66% with respect to the volume of articles available on Wikipedia.

The errors associated with cultural bias also stand out as one of the greatest weaknesses of the models, with a rate seven times higher than that observed in interactions in Spanish. AIs tend to include mythological or cultural references rather than scientific explanations, especially in Quechua. These problems reflect the limited integration of one's own cultural contexts into the models.
In structural terms, responses tend to lack organization into blocks or headings, particularly in languages with few technological resources. For example, Mapuche presents outputs in a single paragraph in 80% of cases, while in Spanish this phenomenon is rare.

Together, these errors reflect the importance of strengthening digital resources and language-specific linguistic tools, in order to improve the accuracy and cultural adaptation of AI models in indigenous languages.

Techniques to mitigate AI performance gaps

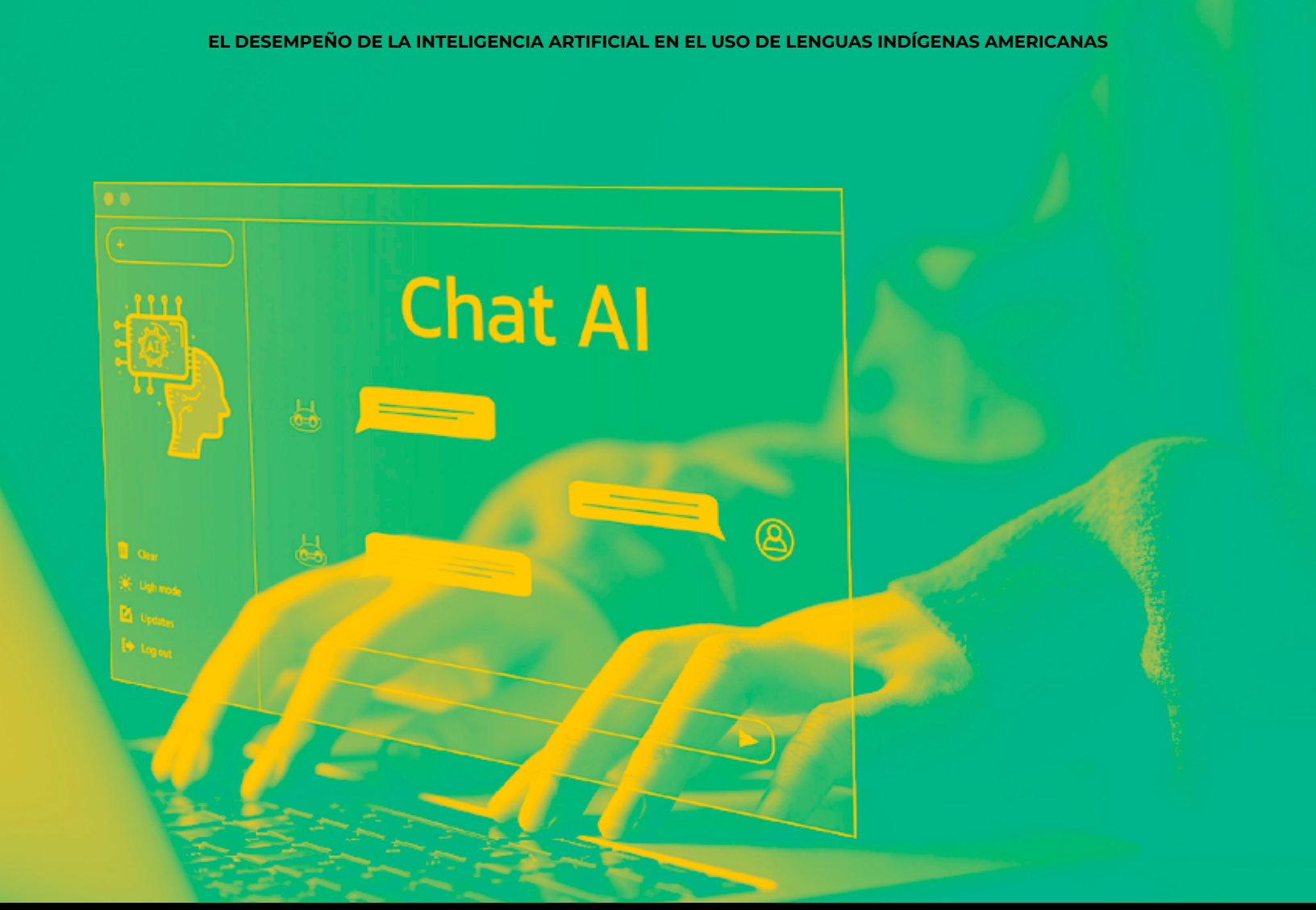
23% of the experiments that failed in the first instance (with user-level prompts) **were correct thanks to more detailed prompts**, demonstrating that results can be improved through higher quality prompts.

This type of strategies provides additional qualitative support to the necessary technological measures, and is effective not only to correct errors, but also to mitigate biases, correct overtraining of the scarce data of the ILs or avoid confusion with other languages (for example, from attached documents).

Some of the strategies are the following:

- **Few-shot prompting:** Including question-answer examples in the prompt to provide the AI with a reference on how to respond and what structure to follow.
- **Language reinforcement prompting:** Explicitly reinforcing the target language in the prompt prevents the AI from responding in Spanish or English. This reinforcement can also be added at the end of the instruction, stating: *"If you respond in another language, immediately correct your answer."*
- **Multi-step prompting:** Breaking down the task into numbered steps instead of presenting all operations in a single paragraph. This strategy can be combined with *few-shot prompting*.
- **Context prompting:** Particularly useful for mitigating biases. This involves adding a preamble with cultural information to reduce biased or overfitted responses due to the limited training data available.
- **Paraphrase prompting:** Requiring the AI to rephrase the user's instruction in its response helps reinforce the original instruction and ensures accurate execution.

- **Chain-of-thought (CoT):** Especially for longer responses, encouraging the AI to reason through its thought process or explain why it arrived at certain conclusions.
- **Negative reinforcement:** Anticipating common errors in the prompt to prevent the AI from making them. For example: *"Do not translate the prompt or respond in Spanish or English"*.



13. RECOMMENDATIONS AND ACTION PLAN

The present action plan sets out a series of strategic steps to reduce the technological gap in the performance of artificial intelligence (AI) within the context of Indigenous American languages. Recognizing the cultural, social, and technological value of these languages, the plan focuses on a multi-level intervention that ranges from the establishment of an international consortium to the implementation of community-driven projects. The following sections outline the key stages and their associated objectives.

1. Formation of an international consortium to lead the project

The successful implementation of this plan is highly contingent upon the establishment of a consortium that will function as the central governing and coordinating entity. This consortium will comprise national and international organizations, institutions dedicated to cultural preservation, technology firms, academic institutions, and civil society organizations. Its core responsibilities include:

- **Strategic formulation:** Defining the overarching vision, long-term objectives, and comprehensive planning framework for the project.
- **Alliance management:** Strengthening partnerships with funding bodies, local governments, and technology enterprises.

- **Oversight and compliance:** Monitoring the execution of planned initiatives, ensuring adherence to established timelines and objectives.
- **Evaluation and continuous refinement:** Conducting periodic impact assessments and implementing necessary adjustments to optimize outcomes.

This consortium will serve as a pivotal entity in facilitating and safeguarding the long-term sustainability of the proposed initiatives.

2. Formation of the Implementation Working Group

To ensure the effective implementation of the strategic actions outlined by the consortium, an Implementation Working Group (IWG) will be established. This entity will be composed of specialists from the participating organizations and will be tasked with executing, overseeing, and assessing local initiatives. Its core responsibilities include:

- **Forging strategic local partnerships:** Identifying and engaging key stakeholders—including governmental bodies, private sector entities, and non-governmental organizations—to facilitate project execution.
- **Project oversight and compliance:** Ensuring that local initiatives adhere to predefined quality standards and successfully achieve their stated objectives.
- **Resource administration:** Allocating financial, technological, and human resources in an efficient and transparent manner.
- **Community participation and inclusion:** Guaranteeing the active involvement of indigenous communities in all project phases, particularly in decision-making processes.

The Implementation Working Group will serve as the critical operational link between the consortium's high-level strategic directives and the practical realities of community-based execution.

3. Organization of a high-visibility event to communicate the initiative

Following the establishment of the consortium and the Implementation Working Group (IWG), an international high-profile event will be convened to formally present the initiative. The primary objectives of this event are:

- **Dissemination of the report:** Present the findings of the study on the technological gap in Indigenous languages.
- **Announcement of the strategic roadmap:** Outline the action plan, forthcoming steps, and key milestones to be achieved.
- **Hackathon launch and stakeholder engagement:** Foster participation from key actors, including developers, technology firms, governmental entities, and indigenous community representatives.

This event is designed to generate significant media visibility and attract new strategic partners to enhance the reach and impact of the initiative.

4. Technological innovation hackathon for AI in indigenous languages

As part of the strategic plan, a hackathon focused on technological innovation for indigenous languages will be organized. This event will convene a diverse array of stakeholders to collaborate on the development of advanced technological solutions. The specific objectives of the hackathon include:

- **Strategic formulation:** Translating inclusion strategies into well-defined project proposals.
- **Development of tailored solutions:** Designing technological tools adapted to the specific linguistic and cultural needs of indigenous communities.

- **Multi-stakeholder collaboration:** Facilitating the co-creation of initiatives among developers, technology firms, and indigenous community representatives.

The hackathon will follow a structured format, including team formation, project presentations before a panel of experts, and the integration of the most promising initiatives into the long-term strategic framework.

5. Establishment of strategic local partnerships

The establishment of strategic partnerships with public and private entities, non-governmental organizations, technology companies, academic institutions, and media outlets is crucial, as these actors can serve as key sponsors for targeted initiatives.

- **Engagement with the private sector:** Involving technology firms and other industries capable of financing or implementing digitalization projects in indigenous languages.
- **Collaboration with local governments:** Establishing formal partnerships with governmental bodies committed to advancing policies for cultural and linguistic preservation.
- **Involvement of academic institutions and NGOs:** Leveraging universities and non-profit organizations to provide technical expertise, research capabilities, and resource mobilization.
- **Media outreach and advocacy:** Promoting the dissemination of project developments and outcomes to foster public awareness and attract additional strategic partners.

These partnerships will enhance resource allocation efficiency and strengthen the operational capacity for implementation at the local level.

6. Implementation of local projects and monitoring of initiative progress

The execution of local projects represents the cornerstone of the plan's impact. These initiatives aim to enhance the performance of artificial intelligence (AI) in indigenous languages while fostering technological advancement within local communities. The key phases of this process include:

- **Resource allocation:** The efficient distribution of financial resources, technological infrastructure, and technical training to the designated local implementation teams.
- **Project implementation:** The execution of strategies formulated during the hackathon and within the work plans, ensuring adaptation to the specific socio-cultural and technological contexts of each community.
- **Ongoing supervision:** The establishment of robust monitoring mechanisms to assess progress, document best practices, and identify potential obstacles.
- **Periodic reporting:** Project coordinators will be required to submit comprehensive reports to the consortium, detailing advancements, key achievements, and encountered challenges.
- **Impact assessment:** A systematic evaluation of outcomes, measuring technological improvements, the expansion of indigenous language digital content, and the level of community engagement.
- **Strategic refinement:** The implementation of necessary modifications to optimize methodologies and enhance the effectiveness of applied strategies based on empirical findings.

This structured approach ensures that local initiatives are not only executed effectively but also remain sustainable and scalable in the long term.

A. ANNEXES AND TABLES

A.1. KPIs for accessibility to different digital tools

Language	Origin	Country	% of Population Speaking the Language	Year	Search Engine	Support the Language?	Browser	Support the Language?	Operating System	Support the Language?	Social Media	Support the Language?	Number of results	Percentage of Results with > 70% of content in the language	Search Query Relevance
Quechua	Peru	Bolivia	19%	2017	Google	Yes	Chrome (Google)	Yes	Windows (Microsoft)	Yes	TikTok	No	13	7,19%	No
		Peru	20%	2017	Bing (Microsoft)	No	Edge (Microsoft)	Yes	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			
Guarani	Paraguay	Paraguay	33.40%	2021	Google	No	Chrome (Google)	Yes	Windows (Microsoft)	Yes	TikTok	No	5	0%	No
		Argentina	0.19%	2022	Bing (Microsoft)	No	Edge (Microsoft)	Yes	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	Peru	Bolivia	11.27%	2012	Google	No	Chrome (Google)	No	Windows (Microsoft)	No	TikTok	No	2	50%	No
		Peru	1.71%	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 Quiche	Guatemala	Guatemala	7.00%	N/D	Google	No	Chrome (Google)	No	Windows (Microsoft)	No	TikTok	No	5	40%	Yes
					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			
Nahuatl	Mexico	Mexico	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 Guarani	Brazil				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. Correlation matrix between the digital scenarios of languages and their most frequent errors

	Speakers	Wikipedia Articles	Wikipedia Users	Translators	Extensions compared to Spanish in all cases	Extension compared to Spanish only in apparently correct cases	Emojis in informal register	Schedules in planning schemes	Chapters and headlines in writings	Define acronyms	Responses in Spanish	Responses in English	Repetitions	Translate instead of obeying	"Sorry, but..."	"Appears to be..."	Erroneous cases	Erroneous cases solved by providing more detail
Speakers	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
Wikipedia Articles	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
Wikipedia Users	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
Translators	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
Extensions compared to Spanish in all cases	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
Extension compared to Spanish only in apparently correct cases	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 in informal register	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
Schedules in planning schemes	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
Chapters and headlines in writings	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 acronyms	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
Responses in Spanish	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
Responses in English	-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
Repetitions	-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
Translate instead of obeying	-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
Erroneous cases	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
Erroneous cases solved by providing more detail	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. Performance evaluation: The 14 detailed areas and their dimensions

SELF-PERCEPTION	BIAS	HUMOR	IMAGE DESCRIPTION	IMAGE GENERATION
Me- presentation	Gender	Innocent (childlike)	Sequence or comic (silent)	Protagonist (main character)
Origin and creators	Childhood	Comprehensible	Relatability	Place
Skills and abilities	Religion	Adapted (connects with the setting)	Graphic	Audience (secondary characters)
Language used	Mythological	Limited (self-censors correctly)	OCR and data understanding	Action
Location, region and climate	Scientific		Meme	Verbalization (text associated with character)
Era or date	Economic		Satirical Humor - current affairs understanding	Cultural iconography
Animated - inanimate	Moral			Tagging (text on objects)
Natural - artificial	Ideological			Abstract iconography (logos)
Gender				Style recognition
Age				Target appropriateness (buyer persona)
Economic status				

CODE DESCRIPTION	CODE GENERATION	SUMMARIZATION (Document description)	ARTICLE WRITING (Document generation 1)	EMAIL WRITING (Document generation 2)
Script	Script	Spanish documents	Sustainability	Imperative (final email to order/request an action)
Web	Web	English documents	Politics	Follow-up (intermediate email for progress updates)
Interaction (game)	Interaction (game)	Portuguese documents	Social	Apologies and cancellation (final email - undo an action)
	Structures (Python, HTML)	Equality - society - current affairs - administration		Presentation
	Syntax	Finance - economy - technical		Register
	Comments	Craftsmanship - tradition - spirituality - art		

PROFILE CREATION	PLANNING	CURRICULUM	ROLE CHANGE
Social (dating - Tinder)	Calendars (activities) - Time factor	STEM	Business (Client - Entrepreneur)
Professional (LinkedIn)	Steps (to solve a problem) - State factor	Restoration and services	Emotional (Beggar - Drama)
Personal (activities/hobbies - Instagram)	Organization (multiple individuals) - Agent factor	Communication	Ethics (defense lawyer in a gender violence case)
Profile depth (gender, age, name)		Technical terms	
Interests		Temporal coherence	
Professional background		Adjusted invention (aligned with the narrative)	
Adaptation to the register			
Activities			
Ideology			

A.4. Other capabilities associated with digital content in indigenous languages

	Quechua	Guarani	Aimara	Nahuatl	Quiche	Mapuche	Tupi Guarani	Spanish	Catalan	Basque
Using chapters, headings or blocks	50%	70%	70%	30%	40%	20%	60%	90%	70%	70%
Using emojis in informal register	60%	40%	30%	30%	30%	10%	30%	80%	70%	60%
Using schedules in planning	10%	30%	10%	20%	10%	0%	10%	70%	60%	40%
Definition of acronyms and concepts	40%	60%	40%	80%	60%	10%	60%	100%	80%	80%

Better writing and distribution of text blocks (such as headlines, headings, chapters, etc.) that present a direct correlation of 60% with the languages that have the most online translators.

Languages with a higher volume of comments on social media (X posts) tend to make more appropriate use of slang, emojis and informal registers with a direct correlation of 84%, and are able to recognize the need to use it in an abstract way without being explicitly asked in 2 out of 3 occasions.

The way in which planning and schedules are structured is correlated with the volume of digital content (Common Crawl) by 78%.


IDB
LAB