

Improving Large Language Models (LLMs) Systems for Underrepresented Languages

Optimización de los modelos lingüísticos grandes (LLMs) para lenguas infrarrepresentadas

Manuel M. Aguilera 

Department of Mathematical Sciences, University of Puerto Rico-Mayaguez

*Autor al que se dirige la correspondencia: alex.martinez13@upr.edu

Recibido: 13 de septiembre 2024 / Revisión: 27 de noviembre 2024 / Aceptado: 13 de agosto 2025

Abstract

According to Garcés, F. (2020) a total of 7000 languages are spoken in the world, half of which are believed to disappear by the end of the 21st century. The generative AI data engine of many AI companies has scarce data on indigenous languages, as many of them have no presence on the Internet (Kshetri, N., 2024). The present study presents a knowledge-based system that provides a stable knowledge base in indigenous languages that can be further used by Reinforcement Learning from Human Feedback (RLHF) to add indigenous languages to AI systems

Keywords: Indigenous language documentation, generative AI, Reinforcement Learning from Human Feedback (RLHF), language revitalization

Resumen

Según Garcés, F. (2020), se hablan un total de 7000 lenguas en el mundo, de las cuales se estima que la mitad desaparecerán para finales del siglo XXI. Los motores de datos generativos de muchas empresas de inteligencia artificial cuentan con datos escasos sobre lenguas indígenas, ya que muchas de estas no tienen presencia en Internet (Kshetri, N., 2024). El presente estudio presenta un sistema basado en el conocimiento que proporciona una base de datos estable en lenguas indígenas, la cual puede ser utilizada posteriormente por el Aprendizaje a partir de Retroalimentación Humana (RLHF) para incorporar lenguas indígenas a los sistemas de inteligencia artificial.

Palabras clave: Documentación de lenguas indígenas, inteligencia artificial generativa, Aprendizaje a partir de Retroalimentación Humana (RLHF), revitalización de lenguas



Introduction

In the field of generative artificial intelligence, Large Language Models (LLMs) are artificial intelligence (AI) models that retrieve data from knowledge bases, documents or texts provided by the user to answer specific questions. For indigenous languages, LLMs contain sparse data due to the methods of documenting and hosting linguistic data across history.

According to some research by (Garcia Garcia et al., 2016) the multimodal method for documenting languages, refers to an approach or methodology that involves the use of multiple modes or forms of communication (such as text, images, audio, video, etc.) to document languages or linguistic Phenomena. In addition, Semantic Domain Documentation is the process of documenting and analyzing the semantic domains present within a language or specific area of discourse (Rayson, 2008; Bouquiaux, & Thomas, 1992).

The general problem with both methods is that there is no order to the words being documented and linguists who have documented languages using these documentation methods do not host the condensed data (see Palacio et al., 2020), as an illustration, instead of hosting the words obtained from a recording, linguists host the recording, which makes AI trainers unable to condense the data, implicitly, it also creates a space where training LLMs in indigenous languages is highly expensive, much more so than training them for well-known languages.

Therefore, the following research was born with the objective of answering the following research question: *How to generate a knowledge base in specific Indigenous Languages for training in Large Language Models (LLMs)?*

General Objective

Create a knowledge base in specific indigenous languages for training in Large Language Models (LLMs).

Specific Objective

How should grammar be presented in this new knowledge base devoid of linguistic interests beyond revitalization?

Methods

This research is qualitative because three specific indigenous language knowledge bases were generated: Pech, Garifuna from Honduras and Sochiapam Chinantec, and a documentation method was created according to the needs of the AI trainer. The latter aligns with those mentioned by (Barnes et al., 2003) who suggest that qualitative studies provide a valuable tool for the generation of new hypotheses.

The implemented design is an action research design because as it aiming to propose a new method that will strengthen documentation processes for language revitalization purposes and help to AI Trainers to include indigenous languages in LLMs, as mentioned by Álvarez-Gayou (2003) and Merriam (2019): “The purpose of action research is to solve every day and immediate problems and to improve concrete practices” (p. 6-7).

This research does not involve a specific type of sample, as it was not conducted with individuals. Instead, it examines the knowledge bases created and explains the new method for language documentation in the Pech, Garifuna from Honduras and Sochiapam, Chinantec languages, respectively. Consequently, this research does not pertain to population or sample considerations, as affirmed by Hernandez-Sampieri (2018; p. 172) for some specific studies.

The data collection techniques employed are field notes and qualitative observation. It should be noted that the author of this article is the originator of these knowledge bases using this distinctive method; hence the focus of the study is on the author’s descriptions and experiences.

In addition, images produced in Midjourney 6 are used throughout the description of the Method, as forms of representation closer to the correct application of the method, as well as providing a contextualized virtual representation for indigenous peoples. It should be added that each image found within the knowledge bases are naturally taken with a camera.

Results and discussion

Based on the above considerations, it should be mentioned that the knowledge bases created as well as the documentation method are adaptations created so that the data can be used easily by the generative AI systems managed by Scale AI has worked with Microsoft, OpenAI, and Meta; today, Meta is also a major investor. Moreover, it is one of the few institutions that works with Reinforcement Learning from Human Feedback (RLHF), being this form of AI systems training the only way to include indigenous languages in LLMs. Below is the standard manner in which an RLHF behaves (Figure 1).

Reinforcement learning from human feedback (RLHF) is a method for training machine learning models by soliciting feedback from human users. RLHF enables more efficient learning. Instead of trying to write a loss function that makes the model behave more like a human, RLHF involves humans as active participants in the training process. RLHF results in models that more closely match human expectations, a typical qualitative measure of model performance. Models trained using RLHF, such as InstructGPT and ChatGPT, are generally considered to be more helpful and more goal-directed (Scale AI, Inc., 2024).

For this, valid data sources must exist at the time of training LLMs. The following is the documentation method used in this research to create the knowledge bases where the data for the following languages were stored: Pech, Garifuna from Honduras and Sochiapam, Chinatec. The methodology will be present-

ed by scale of the Common European Framework of Reference for Languages (CEFR) and the information contained in each level that can be helpful for AI Trainers through the RLHF will be exemplified.

Indigenous Language Knowledge Bases (ILKBs) is a methodology that consists of documenting indigenous languages in stages, based on the CERF, where for each level there are indications for documenting the language and training the AI model, distribution of the above is as follows, as shown in the diagram below (Figure 2).

Using complex linguistics field methods to document indigenous languages based in CEFR Category

Foundational lexicon and pronoun integration (A1)

The motor skills to be documented in any indigenous language at this level are very limited and cover mainly the knowledge of the essential vocabulary and the use of personal pronouns with this essential vocabulary. The words exposed in this essential vocabulary, are those that have a word count from Google being The latter because according to (Google, 2024) words like washing machine are in continuous use in the current context but it has a low relative frequency since the words have been used for less than 100 years. Below is a comparative analysis of the words by relative frequency (Figure 3).

Figure 1

Representation of Reinforcement Learning from Human Feedback (RLHF)

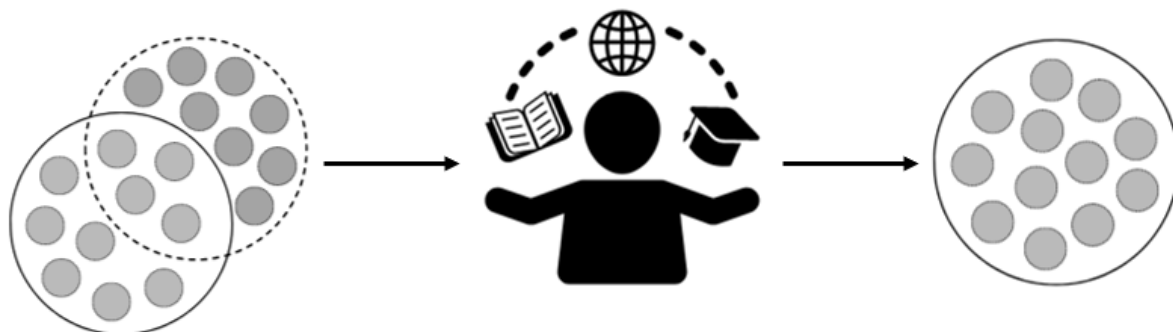


Figure 2

General approach to the construction of Indigenous Language Knowledge Bases (ILKBs) for Training Large Language Models (LLMs)

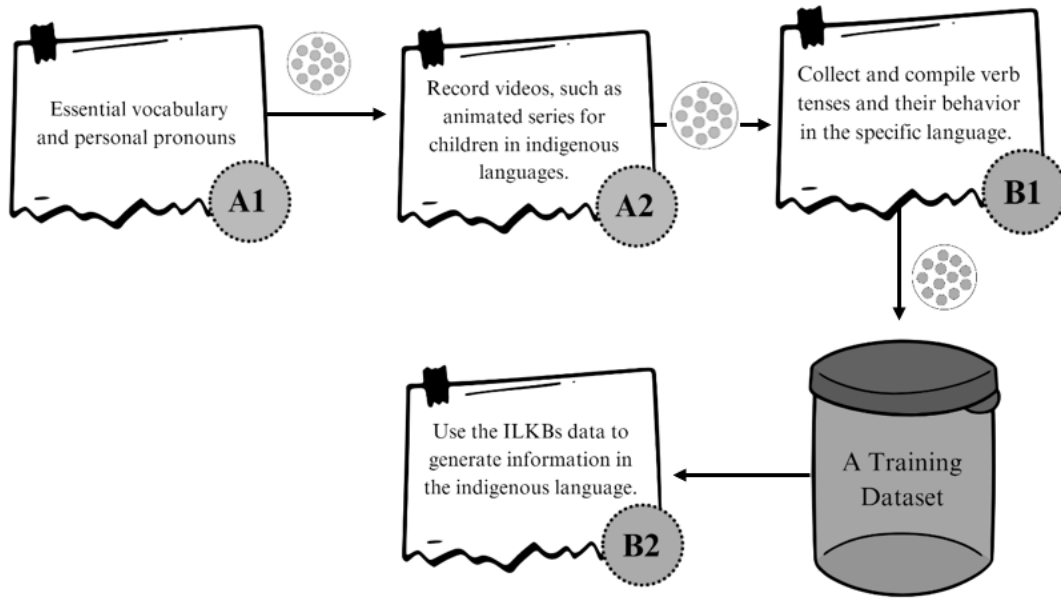
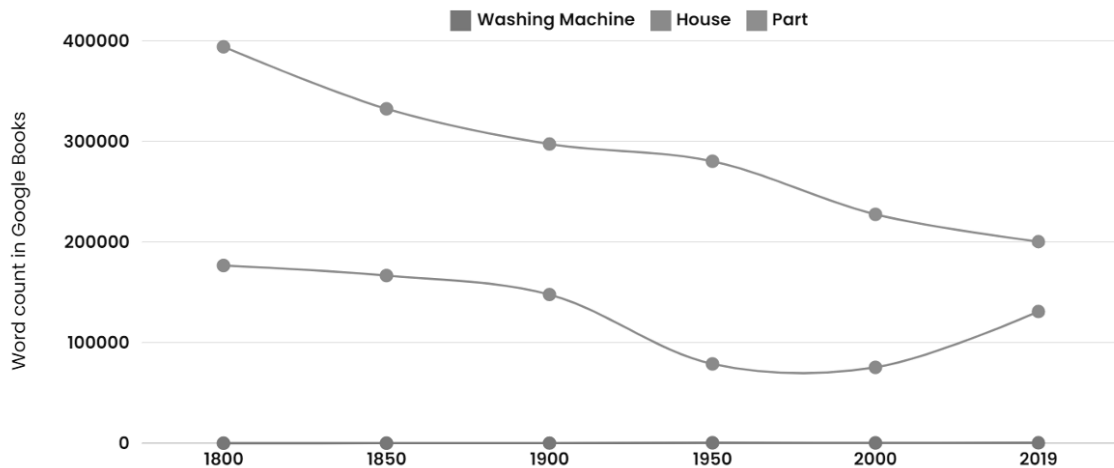


Figure 3

Analysis of the frequency of words in the hierarchy model at the moment of document an indigenous language



Note. From Google Books Ngram Viewer under the license Creative Commons (CC BY-SA 4.0) In the following graph extracted from (Google, 2024) different words can be observed with their word count of all the existing Corpus in Google Books, from these counts the words were classified into three large groups for documentary purposes in indigenous languages which are shown below.

Fundamental concepts

According to the Google Corpus, words related to essential fundamentals have a higher word count. In this case it was considered to take this group starting in the word part with a word count of 200,238 in the Google Corpus 2019 data obtaining the following classification in terms of count in Google Corpus.

$$W_G \geq 200,238$$

This group includes words of humanity, life and the world, but this category also includes prepositions, conjunctions, interrogative, personal and possessive pronouns, interjections, among others, some example words are: life, world, part, time, person, place, number, year, and, after, with, what, etc.

Daily and temporary elements

This classification was taken considering as the highest point the word house with a word count of 130,819. The lowest point is not defined. However, we consider the lowest point to be a word with no lower than 365 occurrences in Google Books, given words such as washing machine that have a low frequency but are not in disuse, leaving the frequency inequality as follows.

$$365 \leq W_G \leq 130,819$$

This group of words includes all observable human elements and actions that are still relevant to people. Some examples are: bee, fossil, hit, sleep, run, notebook, box, book, floor, house, wall, ceiling, door, etc.

Archaisms and obsolete terminology

This classification considers all those words that have fallen into disuse in contemporary language, a specific word count from Google Books is not considered because many obsolete words have a high word count due to their continuous use in previous centuries, for this reason, the linguist's task is to identify whether a word is in disuse or not by observing it and making a qualitative analysis of it, some examples of words in disuse are: abbess, quoth, petrichor, accismus, etc.

Qualitative data documentation process by the linguist

In the data collection process, the priority should be to document the personal pronouns of the group of fundamental concepts: I, you (singular), he, she, it, we, you all (Southern U.S. plural), and they. Afterwards, document the common objects of the group of daily and temporary elements through linguistic consultation. During the vocabulary review process, the linguist might have various objects representing words, such as a key, pencil, notebook, tablecloth, calculator, glue, doll, headphones, table, garbage, etc. The linguist aims to document these daily and temporary elements in the linguist native language, such as Spanish, Japanese, French, etc., to create sentences with them and incorporate subject pronouns and prepositions for translation through linguistic consultation in the particular indigenous language, as demonstrated below:

The keys *are* dirty
I *am* looking for many pencils
She has a doll

(Grammatical Phase I – English Language)

Through linguistic consultation routed in grammar, linguists have to observe how pronouns change according to conjugation in a natural way and present them in such a way that speakers and LLMs have an easy time memorizing and archiving the verb forms, of the concrete language. An interesting example is the sentence I like to see plants:

Seeing plants *is* very important to me
My plants *are* beautiful
The plants *are* mine

(Grammatical Phase I – English Language)

Therefore, at this point, LLMs can construct sentences with more linguistic elements as follows:

The tablecloth covers *the* table
My sister *has* a doll *with* a blue dress

(Grammatical Phase I – English Language)

Among the linguistic elements, relevant to LLMs are the subject and possessive pronouns and prepositions. The tasks assigned to documentation experts are limited to exploring the surrounding world to extract the most influential or important words in context. An example of fieldwork is the following (Figure 4):

Figure 4
Preparation of linguistic documentation experts before consultation



Note. From Midjourney V6 available online at the URL: https://cdn.midjourney.com/e4c88a19-2561-4e1c-b142-0d0b6b48843f/0_2.png under the license Creative Commons (CC BY-SA 4.0)

The linguist's objective in documenting an indigenous language at the CEFR Foundational lexicon and pronoun integration (A1) involves visiting locations rich in objects to extract relevant words and construct sentences. This data is subsequently transferred to a linguistic consultancy.

In Figure. 4., a young linguist is observing the natural surroundings within a forest and recording the vocabulary present in that environment for subsequent linguistic consultation.

<i>Threes</i>	<i>Stump</i>	<i>Dust</i>
<i>Trunk</i>	<i>Leaf</i>	<i>Animals</i>

<i>Soil</i>	<i>Leaves</i>	<i>Ants</i>
<i>Stick</i>	<i>Spiderweb</i>	<i>Weevils</i>
<i>Sky</i>	<i>Stone</i>	<i>Shell</i>

Now, the linguist must craft sentences or narrate a story using the words from the linguistic query process. Typically, linguists create stories for their own reference, as this helps them gauge how thoroughly they have documented the language's grammar and identify gaps that need further investigation. For example, the linguist (see Figure 4) might narrate a story about her childhood, incorporating all the words she collected from the forest, words she does not know in the indigenous language will be written in Spanish.

“I remember when I was in the house of my father and *detras* of the house it was many trees that provide *sombra* to me and to many of my friends and relatives. However, some of the threes was cutter for problems occurred with my uncle and there were just trunks. This *me enojó* because this *molestaba* the soil when fallen water of the sky. In my opinion, the best solution to this problem just was *quitar* the stumps of the garden and throw the leaf to the dump.”

As expected, the idea is for linguists to discover new words to use in forming sentences or telling stories, as there may be some words that are not currently in their vocabulary. In the text above, those words are in bold. The goal is to consult with speakers of the indigenous language to obtain the remaining vocabulary. This process must be repeated frequently to enrich the vocabulary for revitalization purposes. In the field of AI, texts written in indigenous languages can help language models analyze grammatical behavior to create new texts.

Incorporation of Videos Narrated in the Indigenous Language (A2)

At this level, linguists have a well-collected knowledge of vocabulary, word structure, and morphology. However, they face a significant challenge: students using the A1-collected data who are at this level do not know how to speak with accuracy (production) or understand naturally spoken indigenous language produced by native speakers (listening). Therefore, the linguist must address this issue at some point in the documentation process. Our training

dataset should prioritize including native language recordings in video format, specifically by incorporating translations of animated series in indigenous languages, such as *Chavo del Ocho* or *Blue's Clues*.

The use of children's videos is ideal, as they feature familiar environments and contexts that are universally relatable. An optimal portrayal of a series begins with a basic vocabulary in its initial chapters, yet as it advances, it increasingly employs more advanced and diverse vocabulary through its characters. Here is a selection of such series (Figure 5).

Figure 5

Suggested Children's series to be translated into indigenous languages of Little Fox Co., Ltd



Note. From Little Fox Co. Ltd under the license (CC BY-NC 4.0) (*Listening Phase I*)

According to (Go et al., 2023) Little Fox audiovisual videos can contribute to improve the vocabulary proficiency of some language learners. In addition, these audiovisual videos are indispensable resources for the preservation of audio in indigenous languages.

Comprehensive guide to document verb tenses (B1)

During the documentation process, language experts use images to capture clauses accurately, avoiding literal translations that could hinder revitalization efforts. This same guidance applies to AI trainers: when inputting data, they should describe

images or identify elements using a natural language. Note the following example (Figure 6).

Figure 6

Process of collecting grammar through images



Note. Generating with Midjourney 6 available online at https://cdn.midjourney.com/ee0bd8db-5b1e-4d0a-81f2-c0c62617e2f8/0_3.png under the license Creative Commons (CC BY-SA 4.0)

In Figure 6, a cow is shown in the middle of a Pech community in Honduras and it seems to be enjoying its life. It should be noted that a clause is a phrase that helps us to express actions that occur to living beings, objects and other elements of our world, it contains a subject and a verb in the sentence. A farmer from the area has a shotgun and wants to kill the cow. We can form a clause with this as follows:

The farmer wants to kill the cow with a shotgun.

(Grammatical Phase II)

The concept of documenting clauses using images is more intuitive for native speakers to construct sentences based on various real-life scenarios. Presenting these scenarios helps in understanding and sentence formation, which is essential for large language models (LLMs) during their descriptive and query stages.

The primary goal for linguists documenting in the B1 component is to capture clauses using images that depict actions. This approach allows for sentence creation by having participants identify the clause before seeing it and then refine the clause structure through repeated actions with different images. Using just one image, it is possible to get data of different

verb tenses in the particular indigenous language. Below are images of various verb tenses for documenting verb tenses in indigenous languages. The images showing a person performing an action, is effective for sentence construction and can be used to document all verb tenses (Figure 7).

Figure 7

Verb tenses illustrated with key examples and images



Note. Generating with Midjourney V6 under the license Creative Commons (CC BY-SA 4.0) (*Grammar Phase II*)

Verb tenses play a crucial role in communication. Among the most frequently used when documenting are prepositions of place, present continuous, present perfect continuous, future simple, future perfect, past simple, and past continuous.

Training datasets for indigenous languages in AI systems data entry and usage

Once the words from the daily and temporary elements section and the common grammatical elements of fundamental concepts are documented, along with their respective images and audio recordings, we have the AI Training Database. Below is an

image of an AI Training Database created for the Pech language (See Aguilera et al., n.d.) (Figure 8).

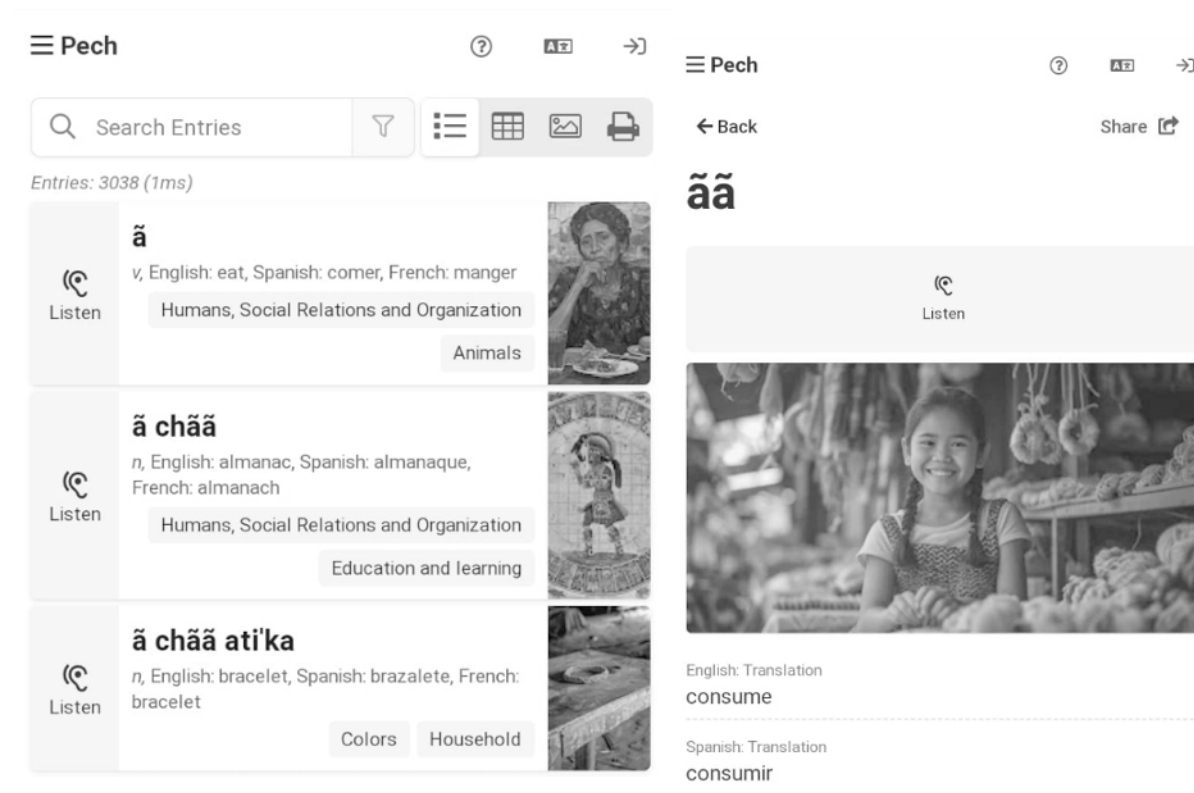
The following initial tasks must be performed by human AI Trainers to input all the data into the LLMs.

Few-shot training procedure:

In these tasks, AI Trainers, often referred to as prompt engineers in this procedure, are responsible for creating prompts and providing multiple answer pairs for each prompt. In this type of tasks, all the lexemes of an indigenous language can be added from the training database provided in the form of the following sequence.

Figure 8

A Training Database for Pech Language of Honduras



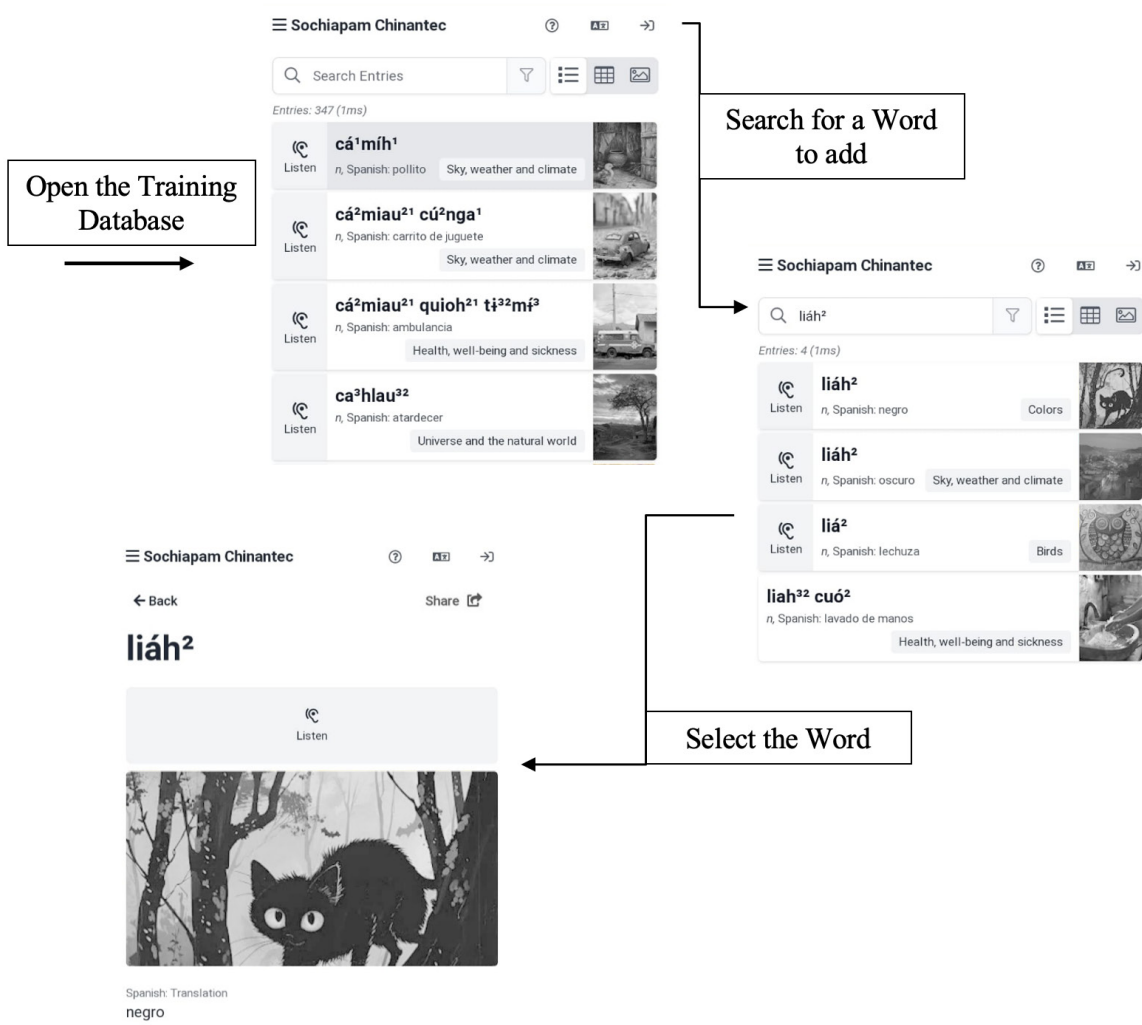
Note. From Pech Living Dictionary under the license Creative Commons (CC BY-SA 4.0)

Step 1: Use the training database to identify a word to add. In this initial step, the prompt engineer will search the provided training database to find a word for inclusion. For instance, suppose the prompt engineering explores the Sochiapam Chinantec lan-

guage training database at (Aguilera, M., & Mariscal-Sánchez, L., n.d.) and incorporate words from it using the few-shot procedure. The process would be observed as follows (Figure 9).

Figure 9

Procedure for Selecting Data to Enhance Indigenous Language Vocabulary in LLMs from the ILKBs



Note. From Sochiapam Chinantec Living Dictionary under the license Creative Commons (CC BY-SA 4.0)

Manual Integration of Indigenous Language Words into LLMs

Once a word is chosen, the prompt engineer will craft a prompt to input in the designated section. However, it is evident that all AI systems tend to exhibit hallucinations when asked to translate a word or text into an indigenous language. See the following AI response for an example.



How is translate black in Sochiapam Chinantec

The translation of “black” in Sochiapam Chinantec is “ts’ots.”



Upon further investigation, we realized that the provided translation for “black” is actually in Maya-K’iche’ and not in Sochiapam Chinantec, providing wrong information to the user. To address such discrepancies, it is essential to enhance the data on indigenous languages through Reinforcement Learning from Human Feedback (RLHF). Below, we demonstrate how a prompt engineer would write both the prompt and the response to add this word to the database correctly.

Prompt:

Translates the word black into Sochiapam Chinantec.

Response A:

The word “black” translates to “liáh” in Sochiapam Chinantec.

Response B:

According to the Sochiapam Chinantec Living Dictionary the word “black” is translate to “liáh” in Sochiapam Chinantec. For more information about this word visit: <https://livingdictionaries.app/sochiapam-chinantec/entries/list>

Performing these processes allows large language models (LLMs) to acquire annotations on vocabulary in indigenous languages, enabling more accurate translations in ancestral languages. With prompt engineering, it is feasible to comprehensively transcribe the entire vocabulary of an indigenous language within five months with a staff of 10 to 25 people, assuming a comprehensive linguistic documentation process has been completed beforehand.

Integrating Grammar in LLMs via RLHF Techniques

In the *few-shot training procedure*, the prompt engineer can also integrate the entire grammar of the language. Typically, the prompt would involve querying a specific verb tense, and the annotations performed by the linguist would provide these queried verb tenses as examples for the LLM to generate sentences from. Consider how a LLM performs in this context at this time.



Translate “The cat is under the table” in Garifuna

In Garifuna, “The cat is under the table” translates to “Nuguyaña lahañu luma luma.”

After examining the response generated by the LLM sample, it uses the Garifuna word *ahañuda* in the form *lahañu*, which means a person became angry, and *luma*, which means to arrive at a place. Therefore, the AI’s output was *He got angry until he arrived at a place*, which does not match the intended request and gives wrong information to the user because of the hallucinations. Below is an example of how a prompt engineer should add the grammar of the Language with some particular words from (Aguilera, M., Morales, J., Vargas, C., Castillo, E., & Ramirez, A., n.d.) in order for LLMs to understand the behavior of each verb tense.

Prompt:

Translates the sentence “The cat is under the table” into Garifuna.

Response A:

The sentence “The cat is under the table” translates to “Labugiñe ñen dábula mesu” in Garifuna.

Response B:

According to the Garifuna people the sentence “The cat is under the table” translates to “Labugiñe ñen dábula mesu” in Garifuna and uses the prepositions of location form to translate, to understand the grammar of Garifuna Language please visit the reference links below:

[1] Garifuna Living Dictionary Grammar Section: <https://livingdictionaries.app/honduras-garifuna/entries/list>

[2] A Grammar of Garifuna by Steffen Haurholm-Larsen: https://boris.unibe.ch/91473/1/16haurholm-larsen_s.pdf

It is recommended to perform this process repeatedly but with different elements so that the language model (LLM) can better understand the grammatical behavior of the language. By exposing it to a variety of structures and contexts, the model can learn grammatical and syntactical patterns more accurately. This is especially useful for languages with complex structures or those different from the languages the model has been predominantly trained on.



Generative AI in Indigenous Languages

Once all the vocabulary, grammar, sentence structures, and pronoun behavior in a specific language have been added, the LLM can begin to be queried for information to generate data in Indigenous languages. At present it is impossible to request this data from an LLM and expect a response in return, see the following as an illustration for the case of the Pech Language.



Translate to Pech “I remember when I was at my father’s house and behind the house there were many stumps that provided shade for me and many of my friends and family.”

Translating this sentence into Pech is *challenging* because the Pech language has limited resources and documentation. Pech (also known as Paya) does not have a direct translation available for this sentence due to limited linguistic resources. For accurate translation, consulting with a native speaker or a linguist specializing in the Pech language would be ideal.

Note that LLMs cannot generate by generative AI extensive texts in indigenous languages, because of the limited amount of data in each language, so that Generative AI (GenAI) will be possible once Few-shot methods or Plain Methods are manually added to the corpus data for each language, below is an example of how a GenAI of an LLM trained in indigenous languages would work.

Prompt:

Translate the sentence to Pech Language "I remember when I was at my father's house and behind the house there were many stumps that provided shade for me and many of my friends and family."

Response A:

The sentence “I remember when I was at my father’s house and behind the house there were many stumps that provided shade for me and many of my friends and family.” translates to “Tah takua tatus akauña chariras; kao askña. Peh ũheka chii, ũkiakua uheka kachũũwasrí tai lerí” in Pech (also known as Paya).

Response B:

According to the Pech people from Olancho and Colón the sentence “I remember when I was at my father's house and behind the house there were many stumps that provided shade for me and many of my friends and family.” translates to “Tah takua tatus akauña chariras; kao askña. Peh ũheka chii, ũkiakua uheka kachũũwasrí tai lerí” in Pech, to understand the behavior of the Pech Language please visit the reference links below:

[1] Pech Living Dictionary: <https://livingdictionaries.app/pesh/entries/list>

Although this process may be lengthy, the knowledge bases presented in this paper offer a viable alternative to the documentation of indigenous languages in order to use the documentation data to train LLMs for inclusion of other languages worldwide.

Conclusions

Throughout the research, three knowledge bases have been created in the Garifuna, Pech and Sochiapam Chinantec languages, which are useful for training artificial intelligence models by containing the word in the indigenous language, audio of the word, sentences in the native language, and their respective translations into Spanish and English, as well as a representative photographs, which makes it possible to add data from these languages to the LLMs using the Few-shot or Plain techniques (Ouyang, L., et al., 2022).

Furthermore, the structure of the Indigenous Languages Knowledge Bases (ILKBs) allows for grammar to be grouped into sentence examples, enabling repeated cases to be manually added to AI systems to improve language understanding and generation. These knowledge bases contribute to broader efforts to preserve and computationally leverage indigenous languages, supporting research in low-resource NLP, digital lexicography, and cross-linguistic AI applications (Bird, S., 2024; Adams et al., 2021).

Future work could expand these conclusions by integrating additional indigenous language resources, exploring interoperability with existing multilingual knowledge bases, and evaluating the performance of LLMs when enriched with these culturally and linguistically diverse datasets.

Author contributions

Drafting and revising the manuscript: All authors
 Conception and design of the study: All authors
 Collecting of data: All authors
 Cleaning data, performing the analysis, and/or interpretation of data: All authors
 Editing the manuscript critically for important intellectual content: All authors

Supplementary Materials

Data are within the paper

References

- Aguilera, M., & Mariscal-Sánchez, L. (n.d.). Sochiapam Chinantec Living Dictionary. Living Tongues Institute for Endangered Languages; Municipalidad de San Pedro Sochiapam. Retrieved October 31, 2025, from <https://livingdictionaries.app/sochiapam-chinantec>
- Aguilera, M., Morales, J., Vargas, C., Castillo, E., & Ramírez, A. (n.d.). Honduras Garifuna Living Dictionary. Living Tongues Institute for Endangered Languages; Instituto Hondureño de Ciencia, Tecnología y la Innovación. Retrieved October 31, 2025, from <https://livingdictionaries.app/honduras-garifuna>
- Aguilera, M., Simeón-Martínez, A., & Rafael-Palma, A. (n.d.). Pech Living Dictionary. Living Tongues Institute for Endangered Languages. Retrieved October 31, 2025, from <https://livingdictionaries.app/pesh>
- Álvarez-Gayou, J. L. (2003). *Cómo hacer investigación cualitativa. Fundamentos y metodología*.
- Barnes, R. W., Grove, J. W., & Burns, N. H. (2003). Experimental assessment of factors affecting transfer length. *Structural Journal*, 100(6), 740-748.
- Bird, S. (2024, August). Must NLP be extractive? In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL 2024)* (pp. 14915–14929). Association for Computational Linguistics.
- Bouquiaux, L., & Thomas, J. (1992). *Studying and describing unwritten languages*. SIL International.
- Garcés, F. (2020). *La revitalización de las lenguas indígenas del Ecuador: una tarea de todos*.
- García García, A., Fabiano, E., & O'Hagan, Z. (2016). *Taushiro Field Materials, 2016-09* [Collection identifier]. California Language Archive, Survey of California and Other Indian Languages, University of California, Berkeley. <http://dx.doi.org/doi:10.7297/X20V89ZG>
- Go, Y., Ying, Y., Sunjaya, M., & Saragih, J. C. F. (2023). The utilization of Little Fox Chinese video in learning Mandarin vocabulary for elementary school students. In *E3S Web of Conferences* (Vol. 426, p. 02012). EDP Sciences.

- Google. (2024). Google Ngram Viewer: 1800-2019. Retrieved May 27, 2024, from <https://books.google.com/ngrams>
- Hernández Sampieri, R. (2018). *Metodología de la investigación: Las rutas cuantitativa, cualitativa y mixta* (6.^a ed.). McGraw-Hill.
- Kshetri, N. (2024). The academic industry's response to generative artificial intelligence: An institutional analysis of large language models. *Telecommunications Policy*, 48(5), 102760.
- Little Fox. (n.d.). Series [Web page]. Retrieved October 31, 2025, from <https://chinese.littlefox.com/en/story>
- Merriam, S. B., & Grenier, R. S. (Eds.). (2019). *Qualitative research in practice: Examples for discussion and analysis*. John Wiley & Sons.
- Midjourney. (2024). AI-generated image depicting linguistic fieldwork during indigenous language documentation [Image]. Midjourney CDN. https://cdn.midjourney.com/e4c88a19-2561-4e1c-b142-0d0b6b48843f/0_2.png
- MidJourney. (2025). Untitled image generated using the prompt: "a futuristic indigenous linguist documenting endangered languages, cinematic lighting, hyper-realistic, 8k, national geographic documentary style." [AI-generated image]. Retrieved October 31, 2025, from https://cdn.midjourney.com/ee0bd8db-5b1e-4d0a-81f2-c0c62617e2f8/0_3.png
- Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P., Leike, J., & Lowe, R. (2022). Training language models to follow instructions with human feedback. arXiv. <https://doi.org/10.48550/arXiv.2203.02155>
- Philip Tim Palacio, Chundra Cathcart, I-Hsuan Chen, Emily Cibelli, Kristin Hanson, Shinae Kang, Lev Michael, Eric Prendergast, Christine Sheil, Tammy Stark, & Elise Stickles. (2020). *Berkeley Field Methods: Garifuna, 2020-06* [Collection identifier]. California Language Archive, Survey of California and Other Indian Languages, University of California, Berkeley. <http://dx.doi.org/doi:10.7297/X2RX99FN>
- Rayson, P. (2008). From key words to key semantic domains. *International Journal of Corpus Linguistics*, 13(4), 519-549.
- Scale AI., Inc. (n.d.). Reinforcement learning from human feedback (RLHF). In *Large language models*. Retrieved from [https://scale.com/guides/large-language-models#reinforcement-learning-from-human-feedback-\(rlhf\)](https://scale.com/guides/large-language-models#reinforcement-learning-from-human-feedback-(rlhf))