ELR-1000: A Community-Generated Dataset for Endangered Indic Indigenous Languages

⋆ **Neha Joshi**♢ ⋆ **Pamir Gogoi**♢ **Aasim Mirza**♢ **Aayush Jansari**♢
Aditya Yadavalli♠ **Ayushi Pandey**♢ **Arunima Shukla**♢ **Deepthi Sudharsan**♣
Kalika Bali♡ **Vivek Seshadri**♢
♢ Karya ♠ UC San Diego ♣ Independent Researcher ♡ Microsoft Corporation
`{neha, pamir.gogoi, vivek}@karya.in`

Abstract

We present a culturally-grounded multimodal dataset of 1,060 traditional recipes crowd-sourced from rural communities across remote regions of Eastern India, spanning 10 endangered languages. These recipes, rich in linguistic and cultural nuance, were collected using a mobile interface designed for contributors with low digital literacy. Endangered Language Recipes (ELR)-1000—captures not only culinary practices but also the socio-cultural context embedded in indigenous food traditions. We evaluate the performance of several state-of-the-art large language models (LLMs) on translating these recipes into English and find the following: despite the models' capabilities, they struggle with low-resource, culturally-specific language. However, we observe that providing targeted context—including background information about the languages, translation examples, and guidelines for cultural preservation—leads to significant improvements in translation quality. Our results underscore the need for benchmarks that cater to underrepresented languages and domains to advance equitable and culturally-aware language technologies. As part of this work, we release the ELR-1000 dataset to the NLP community, hoping it motivates the development of language technologies for endangered languages.[1]

1 Introduction

Indian natural language processing communities have seen a renewed wave of digitization of constitutional languages (Chopra et al., 2019; Abraham et al., 2020; Madaan and Agrawal, 2022; Doddapaneni et al., 2023; Khan et al., 2024; Bhat et al., 2024). However, most datasets focus on high-resource or officially recognized languages, overlooking India's tribal and indigenous diversity. This exclusion limits access to digital tools and information, deepening inequalities and endangering linguistic heritage.

Due to this gap, speakers of non-standardized, underrepresented, or endangered languages remain excluded from digital services, and educational resources. This technological gap compounds existing inequalities and accelerates the obsolescence of their languages. Such marginalization contributes not only to cultural erasure but also to the loss of generational knowledge—particularly in domains like agriculture, medicine, and food systems, where knowledge is embedded in oral traditions. These languages are repositories of unique cultural epistemologies, and their extinction implies the loss of irreplaceable linguistic and ecological knowledge. Incorporating these low-resource languages into mainstream NLP systems can help bridge information access gaps, supporting community agency, digital inclusion, and long-term linguistic preservation.

From a technical perspective, neural machine translation (NMT) for low-resource languages has historically underperformed due to limited parallel corpora and limited annotated data. Recent advances, however, have shown promise. Large language models (LLMs) have begun to show significantly improved performance even in zero-shot or few-shot translation settings. Models like GPT 4o (OpenAI et al., 2024), LLaMa (Touvron et al., 2023) and Gemini (Comanici et al., 2025) demonstrate emergent capabilities in translating low-resource languages. The development of domain-specific (Zheng et al., 2024) or culturally relevant corpora (Yakhni and Chehab, 2025) further enhances translation quality, making such initiatives doubly impactful.

Our project documents 1,060 traditional recipes across 10 tribal languages, collected through a community-based approach from 368 rural women and 26 men who are native speakers of these lan-

---

[1] `https://github.com/karya-inc/elr-1000`
♠ Work done while in Karya
♣ Work done while in Microsoft
⋆ Joint first authors

Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics, pages 2441–2457
December 20-24, 2025 ©2025 Association for Computational Linguistics

Dataset	Region	# Entries	Modalities	Main Data Source	Contextual Info	Lang. Coverage
Recipe1M+ (Marın et al., 2021)	Global	13M images	Text, Image	Web	None	English
IndianFoodNet (Agarwal et al., 2023)	India	5,500+ images	Image	Web	None	English / Indic labels
Khana (Prabhu, 2025)	India	131K images	Image	Web	Visual features	English
ISIA Food-500 (Min et al., 2020)	Global	399,726 images	Image	Web	None	English
ELR-1000 (This work)	India	1,060 recipes	**Text, Image, Audio**	**Community**	**Seasonality, storage, ecology, tools, oral narration**	10 endangered Indian languages

Table 1: Comparison of major food computing datasets and the proposed Endangered Indic Indigenous Recipes (ELR-1000) dataset

guages. This dataset, grounded in indigenous culinary knowledge, not only contributes to linguistic resource development but also serves as a cultural archive for future generations. Most importantly, a subset of this data is also parallel data, where each recipe has been translated manually into English. Through this work, we aim to demonstrate a replicable model for ethical data collection, annotation, and deployment of AI tools for severely underrepresented language communities.

While many datasets for low-resource languages exist, they often rely on translations from high-resource languages, lacking cultural grounding. For example, the Flores-101 benchmark (Goyal et al., 2022) covers 101 languages but is based entirely on English Wikipedia content. In contrast, our dataset is community-authored by speakers of low-resource and tribal languages, using content rooted in daily life, especially food practices. This makes it one of the few datasets that offer both linguistic and cultural representation.

The main contributions of this work are as follows:

- We release ELR-1000—composed of 1,060 recipes in 10 endangered languages of Eastern Indic languages under the Karya Public License (KPL).
- Translation of a representative subset of this corpus into English in the form of a parallel corpus for LLM-enabled translation.
- Evaluating LLM capabilities in translating traditional recipes, focusing on cultural authenticity and factual accuracy. Highlighting current strengths and limitations in handling nuanced, culturally specific content.

2 Related Work

Previous research has mainly examined the intersection between multimodality and cultural diversity, or between multilingualism and cultural knowledge. However, recent efforts indicate a growing shift toward unifying these perspectives—-exploring models and benchmarks that simultaneously span multilingual, multimodal, and multicultural dimensions. Benchmarks such as CVQA (Romero et al., 2024), ViMUL-Bench (Shafique et al., 2025), ALM-Bench (Vayani et al., 2025), M5-VGR and M5-VLOD (Schneider and Sitaram, 2024) have been designed to evaluate the performance of models across modalities, languages, and cultures.

While efforts to benchmark cultural knowledge and low-resource or endangered languages have evolved independently, especially in a culturally and linguistically diverse country like India, they rarely intersect. For instance, works like SANSKRITI (Maji et al., 2025) and DOSA (Seth et al., 2024) focus on evaluating LLMs' understanding of Indian culture and cultural artifacts, whereas efforts like PARIKSHA (Watts et al., 2024), INDICGENBENCH (Singh et al., 2024), Indic-QA (Singh et al., 2025), and MILU (Verma et al., 2025) provide multilingual benchmarks for Indic language understanding. However, these works remain largely confined to high-resource languages, leaving endangered Indic languages unrepresented. Therefore, in our work, we introduce a cultural knowledge database covering 10 endangered languages of Eastern India. This enables us to evaluate the performance of LLMs on low-resource, culturally grounded tasks, and to highlight the challenges and opportunities in extending language technologies to underrepresented linguistic communities.

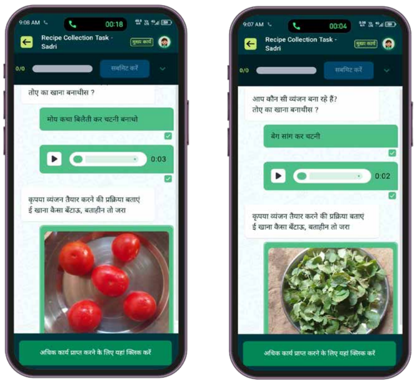
Our work most closely relates to the following: FoodieQA (Li et al., 2024), WorldCuisines

Language	Recipes	Words	Vocabulary	Recordings	Total Duration (s)	Images
Bodo	95	26745	3958	3043	34076	1532
Assamese	113	29648	757	1888	7034	1415
Meitei	100	18277	3474	2133	16808	580
Kaman-Mishmi	128	21334	3159	3036	27207	1129
Khortha	126	24398	656	3131	14733	1129
Santhali	120	20878	380	2892	12116	1004
Ho	91	13767	395	2300	6726	875
Sadri	107	13257	528	2785	9139	1103
Mundari	82	15243	1785	2378	13432	703
Khasi	98	30166	1194	4534	24460	1928
Total	**1060**	**213713**	**16286**	**28120**	**165731**	**11398**

Table 2: Recipe Dataset Statistics by Language (Duration in Seconds)

(Winata et al., 2025), CulturalRecipes (Cao et al., 2023) and CARROT (Hu et al., 2024). They focus on cuisine-centered reasoning and cross-cultural recipe adaptation and retrieval. In the Indian context, IndiFoodVQA (Agarwal et al., 2024) covers several aspects of Indian cuisine and culinary diversity. Even though IndiFoodVQA (Agarwal et al., 2024) introduces culturally relevant visual question answering grounded in Indian food, it—like other global food datasets—largely targets high-resource settings and fails to represent endangered or minoritized language communities.

Recipes not only encode procedural knowledge but also serve as inter-generational vessels for transmitting language, values, and identity. In many Eastern Indic communities, making and sharing indigenous recipes sustains linguistic practices endangered in formal or educational settings – making them a rich and practical source for cultural benchmarking. Hence, in this work, we propose a multilingual benchmark, Endangered Language Recipes (ELR-1000), for Endangered Eastern Languages through Indigenous Recipes that simultaneously targets cultural competence and endangered language preservation.


9:08 AM
00:18
Recipe Collection Task - Sadri
मुख्य कार्य
0/0
सबमिट करें
0:03
अधिक कार्य प्राप्त करने के लिए यहां क्लिक करें
9:07 AM
00:04
Recipe Collection Task - Sadri
मुख्य कार्य
0/0
सबमिट करें
0:02
अधिक कार्य प्राप्त करने के लिए यहां क्लिक करें

तोए का खाना बनाथीस ?

मोय कचा बिलैती कर चटनी बनाथो

कृपया व्यंजन तैयार करने की प्रक्रिया बताएं
ई खाना कैसा बँटाऊ, बताहीन तो जरा

आप कौन सी व्यंजन बना रहे हैं?
तोए का खाना बनाथीस ?

बेग सांग कर चटनी

कृपया व्यंजन तैयार करने की प्रक्रिया बताएं
ई खाना कैसा बँटाऊ, बताहीन तो जरा

Figure 1: App Interface of the Recipe Documentation Task

3 Endangered Language Recipes (ELR-1000)

3.1 Pilot Study

The pilot study served to validate our research approach and methodology through engagement with rural women from the Sadri-speaking community. Sadri, classified as an under-resourced language, is spoken by approximately 5.1 million people primarily in the eastern regions of India. This linguistic community was strategically selected for our pilot experiments because of Karya's established relationships within these communities and familiarity with the Sadri language, which facilitated authentic engagement and reduced potential barriers to participation. After receiving a positive response from the participants, we organized a demonstration session with approximately 30 Sadri-speaking women from 3 remote tribal villages in Jharkhand. In this session, each participant recorded one recipe using the Karya application. Notably, all of these women were new to digital work and represented communities with limited access to technology and digital literacy programs. The positive response and successful completion of this pilot study motivated us to expand the project.

3.2 Language Selection

Based on the successful completion of the pilot project, we selected ten endangered languages from the UNESCO Endangered Languages list.[2] Of these, five are majorly spoken in Jharkhand and Bihar, and the rest are from Northeast India. We aimed to balance high-resource and low-resource languages to ensure a representative sampling. These languages are characterized by their geographic isolation, with speakers concentrated in remote tribal areas having negligible digital presence or technological resources. The selected languages allowed for regional and linguistic diversity. Thereby, the resulting dataset reflected variations in cooking techniques, local crop varieties, ingredient naming conventions, and cultural traditions that differ significantly across Indias ecological zones—offering an inclusive and representative portrait of the country's rich food heritage. Table 2 shows the collected dataset statistics.

The selected languages were:

- **Languages from Jharkhand and Bihar**: Ho (spoken by approximately 1.04 million people in remote tribal areas), Khortha (limited to specific districts with minimal written documentation), Sadri (scattered across isolated mining regions), Santhali (primarily in rural tribal communities), and Mundari (concentrated in remote forest areas with limited connectivity)

- **Languages from North Eastern region of India**: Assamese, Meitei, Khasi, Bodo (concentrated in isolated areas of Assam), and Kaman Mishmi (critically endangered with speakers in extremely remote border regions)

We primarily aimed at women as participants in this study. During the mobilization process, we selected women over men based on the local employment context: women in these villages have significantly fewer income-generating opportunities due to household responsibilities that typically keep them at home. This made them well-suited for our data collection,they had the flexibility to participate, could complete tasks from their homes without needing to travel, and possessed deep generational knowledge of traditional culinary practices developed through years of cooking for their families. This approach enabled us to efficiently gather rich culinary data while simultaneously providing income opportunities to a demographic with limited access to formal employment.

3.3 Application Design & Usability Considerations

The Karya mobile application was designed for rural, often first-time digital workers with limited education or smartphone experience. To support varied literacy levels, the app used a minimal-text interface with clear audio-visual cues in local languages. Instructions for each recording step were delivered via simple icons and audio guidance to reduce cognitive load and accommodate non-literate users.

Importantly, the design allowed participants to review and edit their text or audio entries before final submission allowing them to correct mistakes independently. To avoid interrupting cooking, the application separated media capture from annotation: users first took photos, then added text or audio explanations later at their own pace. To handle poor connectivity in remote areas, the application also supported offline data entry, letting users save sessions locally and upload them when internet was available. Figure 1 shows Karya application's interface.

3.4 Capacity Building Trainings

To ensure effective community engagement, we recruited one local coordinator per language—each a native speaker residing in their respective area—with the help of local Non-Governmental Organizations (NGOs) working in each region. These coordinators then mobilized 30-50 rural women participants per language (aged 15-45), targeting individuals who were fluent in reading, writing, and speaking in their native language and had smartphone access.

After recruitment, Karya held in-person training and application demos to guide coordinators and participants on effective recording. We also explained the projects goal of preserving traditional recipes in native languages. The local coordinators played a crucial role in bridging the gap between us and the community members, by explaining tasks to participants, and providing ongoing support when people faced challenges with recording in the application. Participants were asked to provide their informed consent before starting the data collection tasks.

---

[2] https://unesdoc.unesco.org/ark:/48223/pf0000192416.locale=en

Each recipe submission underwent a thorough review and validation by the local coordinator. Since many contributors were first-time digital workers, this validation process played a vital role in ensuring accuracy and quality. Once approved, contributors received transparent payments directly through the Karya application, earning \$8.68 per recipe. On average, participants submitted between two to five recipes, allowing them to earn from \$17.50 up to \$43.50 for their valuable contributions.

3.5 Addressing Multimodal Data Heterogeneity

The data collection framework prioritized contributor authenticity over structural consistency, allowing participants the freedom to document culinary practices in their own style using the Karya application. This intentional flexibility captured rich cultural and epistemic representations of food knowledge but introduced substantial structural heterogeneity across the dataset. Table 2 compares the metadata across all languages.

Participants used different strategies to contribute—some provided detailed text with images, others relied heavily on audio narration—resulting in variable completeness across modalities. This created a complex, un-standardized dataset requiring systematic normalization.

To handle this, we implemented a modular, array-based structure separating text, image, and audio content. Each recipe step was decomposed and stored in parallel directories, with pre-processing steps to address null values and corrupted media. Quantitative analysis revealed diverse documentation patterns: 54% of recipe steps included all three modalities, 29% used image-text, and 83.4% had featured text. Notably, 64.5% included audio narration, underscoring the cultural significance of oral knowledge. These patterns demonstrate that when unconstrained, contributors favored multimodal formats that reflect oral traditions and the limitations of purely textual expression.

4 Challenges Faced During the Collection of ELR-1000

A key contribution of this project lies in documenting data collection challenges. And this can inform future efforts. These challenges fall into the following broad categories:

Demographic & Socio-Cultural Challenges Digital literacy among participants varied widely, with many experiencing structured documentation and in-app recording for the first time. This called for intensive, culturally sensitive training to ensure comfort and understanding.

Trust building Some communities were initially skeptical about sharing traditional knowledge and our project's goals. To build trust, local coordinators explained the project in familiar terms, and we maintained transparency around payments, data use, and cultural preservation goals.

Seasonality of Food Practices A key challenge in documenting tribal and rural recipes was their seasonality and reliance on foraged ingredients. Some dishes couldnt be recorded year-round due to ingredient unavailability. For example, some forest greens or wild fruits are harvested only during the monsoon or winter seasons. To address this, we encouraged participants to note seasonal variants and substitutions in the cultural context section.

Ensuring Fair Payment A nuanced challenge was ensuring fair compensation across recipes of varying complexity. While some involved extensive effort and ancestral techniques, others were simple but equally authentic. Paying a fixed rate felt unfair to those doing more work, yet dismissing simple dishes was not culturally appropriate. We addressed this by setting a minimum step requirement to ensure baseline documentation quality.

Maintaining Data Quality & Consistency Given the diversity of languages, cooking styles, and literacy levels, ensuring consistency and quality in the recordings was another major challenge. Recipe steps varied in granularity, and the local names of ingredients often lacked standardized spellings. To manage this, we implemented a two-layer validation system: local coordinators first reviewed entries for completeness and clarity, and project managers conducted spot checks across languages for cross-regional consistency.

Resolution through Iterative Feedback & Capacity Building One-time training was not sufficient, so we adopted an iterative support model. Coordinators regularly assisted participants, resolved issues, and encouraged re-recordings – crucial for maintaining motivation and improving quality.

5 Experimental Design and Methodology

Research Motivation The primary motivation for this research was to assess the current capabilities of large language models (LLMs) regarding cultural relevance in the food or recipe domain, specifically for communities that speak and write under resourced East Indic languages. To this end, the experiment was designed to evaluate LLM translations of traditional recipes into English. The evaluation specifically focused on the models' ability to maintain cultural authenticity and factual correctness, process local dialects, and preserve the original instructional format of the recipes.

5.1 Experimental Framework

To structure our experiments, we divided the evaluation based on the capabilities of different machine learning models. Neural Machine Translation (NMT) models typically have sentence-level context windows, while LLMs possess much longer context windows that enable them to understand recipe structures and contextual information within the prompt, even when direct translation capabilities may be limited.

5.1.1 Neural Machine Translation (NMT) Evaluation

The majority of languages in our dataset are not supported by state-of-the-art NMT systems. Since Assamese is a relatively resource-rich language compared to others in our study, many leading NMT systems did provide support for it. However, both BLEU and chrF scores were significantly lower than expected. We attribute to the specialized domain of traditional recipes and the limited training data for such culturally specific content. An additional challenge emerged from script variations. While NMT models typically support only Devanagari script for certain languages, some participants in our data collection process provided recipes in Latin script. This occurred because we prioritized authentic cultural expression over standardized orthography, allowing participants complete freedom in their creative expression to capture genuine traditional recipes.

5.2 Large Language Model Evaluation

5.2.1 Model Selection and Setup

Since the primary objective of this study was to evaluate translation quality, model selection focused on their documented strengths in handling multilingual inputs, contextual understanding, and cultural relevance in translation outputs. To ensure a balanced evaluation, we selected an equal number of proprietary and open-source models three proprietary models (Gemini 2.5 Flash, GPT-4o, and Claude Sonnet 4) and three open-source models (Llama 4 Scout 17B-16E, Mistral Small 3.1 (25.03), and CohereLabs Aya Expanse 8B). This approach allowed us to maintain fairness in comparing models across different licensing types, avoiding any inherent bias towards either commercial or open alternatives. Additionally, we prioritized models that represent the best publicly accessible options in their respective categories. Care was taken to avoid comparisons between models of vastly different capacities (e.g., large reasoning models versus small open models) to ensure that the evaluation remains meaningful, relevant, and reflective of practical use cases. We evaluated six state-of-the-art LLMs using Gemini 2.5 Flash (on default settings) as the evaluation judge. LLMs possess longer context windows and often contain cultural and demographic information about endangered Indic languages within their training data. However, most top-performing models do not possess direct capabilities to translate these specific languages into English.

5.2.2 Experimental Conditions

We conducted two experiments for each LLM:

No Context Condition We assessed baseline model performance by providing only the complete recipe in the source language and requesting translation to English. This condition evaluated the models' inherent knowledge and translation capabilities without additional guidance.

Contextual Condition We enhanced the translation context by providing four things: (1) Background information about the source language and the communities which speak them (2) Few-shot translation examples. (3) Specific guidelines for cultural preservation in translation (4) Instructions for maintaining recipe structure and terminology.

The results showed significantly improved performance across all models in the contextual condition, with particularly dramatic improvements for models like Mistral that exhibited hallucination behaviors without context.

5.3 Evaluation Methodology

5.3.1 Human-in-the-Loop Evaluation Protocol

Our evaluation followed a rigorous three-stage process and the metrics used are in Appendix A:

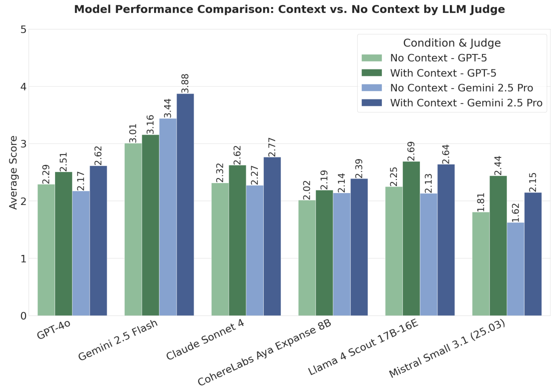


Figure 2: Impact of Contextual Information on Model Performance

- **Human Translation:** We first obtained high-quality human translations for all test recipes from native speakers of each language.
- **LLM Translation Generation:** We processed the raw recipe data through each LLM under both experimental conditions (no context and contextual).
- **Hybrid Evaluation:** We employed a two-judge LLM ensemble, Gemini 2.5 Pro and OpenAI GPT-5, for automated evaluation against human reference translations. This ensemble was chosen to enhance the robustness and reliability of the scoring by mitigating model-specific biases inherent in single-judge systems. Following the automated scoring, Human Oversight was applied to a sample of evaluations, particularly those with score discrepancies, to verify and validate the results and ensure ultimate accuracy.

5.4 Results

5.4.1 Overall Model Performance

Analysis across all ten languages reveals distinct performance tiers among the evaluated models. The provision of context was universally critical, dramatically separating useful translations from unusable ones.

5.4.2 Top Performers

Gemini 2.5 Flash consistently emerged as the most capable and reliable model, achieving the highest average scores in both experimental conditions. In the contextual setting, it frequently received perfect or near-perfect ratings, demonstrating exceptional strength in handling complex cultural translations across languages such as Santhali, Meitei, and Assamese.

Claude Sonnet 4, GPT 4o and Llama-4-Scout-17B formed the second performance tier in contextual conditions. Both models showed significant improvement with context, with Llama-4-Scout demonstrating particular strength in Sadri and Ho translations, while Claude Sonnet 4 excelled in Khortha and Ho recipe translations.

5.4.3 General Performance Patterns

Without context, most models struggled significantly, with average scores indicating that translations were largely unusable for practical applications. As shown in Table 3, models like Gemini frequently produced generic or misplaced substitutionssuch as replacing star fruit with bamboo shootand omitted culturally specific implements like the mortar and pestle. When contextual information was provided, performance improved across all models: translations became more faithful to the source and better reflected local tools and ingredients.

However, even with context, human translations remained superior, particularly in preserving cultural nuance. For instance, Gemini introduced terms such as chopping board in both contextual and non-contextual outputsa tool almost never used in indigenous kitchensrevealing a persistent bias toward Western culinary concepts. This highlights that while contextual information aids fluency and accuracy, current LLMs still tend to normalize outputs toward globalized or Western frames of reference, underscoring the need for culturally grounded datasets and evaluation frameworks.

6 Discussion & Analysis

6.1 The Paradox of Fluent Falsehood

The most salient observation from our analysis reveals a critical disconnect between translation fluency and content accuracy. Across nearly all models and languages, “Fluency” and “Comprehensibility” scores consistently exceeded “Adequacy” and “Cultural Appropriateness” scores. This pattern indicates that models excel at generating grammatically correct, readable English output that often bears minimal resemblance to the source text’s actual meaning.

This phenomenon, which we term “fluent falsehood”, creates a dangerous illusion of successful translation. For the most challenging languages

Source	Sentence 1	Sentence 2
Human Translation	Im going to cook star fruit and fish curry.	To cook this I have used a wok, a lid, a ladle, mortar and pestle, a bowl, knife, and a vegetable basket.
Gemini (No Context)	I am going to cook a dish with **bamboo shoot and fish**.	To cook this, I used a pan, knife, **cutting board**, spatula, bowl, ladle, and a gas stove.
Gemini (With Context)	I am going to cook **star fruit with fish** now.	For cooking this, I have used a mortar and pestle, **chopping board**, knife, stirring spoon, large cooking pot, cleaver, and a bowl.

Table 3: Comparison between Human Translation and Gemini outputs (with and without context).

such as Bodo and Kaman Mishmi, models did not simply mistranslate; they systematically hallucinated entirely different recipes. Traditional ingredients like silkworms or specific regional vegetables were replaced with generic instructions for mushrooms, chicken curry, or completely unrelated dishes.

6.2 Context as a Critical Success Factor

Context provision emerged as the single most important factor determining translation success or failure. For models like Mistral, contextual information represented the difference between nonsensical output (scoring 1.0) and usable translations (scoring 4.0 or higher).

Context proved especially vital for cultural appropriateness. Models without context consistently failed to translate traditional tool names, culturally specific cooking practices, or unique ingredients, instead opting to omit them or substitute generic Western equivalents. Gemini 2.5 Flash's cultural handling in Santhali and Meitei translations improved to near perfect scores when provided with appropriate context in fluency and comprehensibility metrics. Figure 2 shows how context improves different LLMs' performance on our benchmark.

6.3 Cultural Blindness and Systematic Errors

Our analysis revealed profound model limitations in understanding and preserving cultural and material context:

- **Systematic Ingredient Misidentification:** Models repeatedly misidentified key recipe components. "Jhingi" was consistently mistranslated as "prawns/shrimp", while Bodo silkworm recipes were altered to feature mushrooms or conventional meat. These errors fundamentally alter the dish's cultural and nutritional identity.
- **Loss of Traditional Methods:** Cultural nuances related to traditional cooking methods, specialized utensils, and embedded cultural narratives were lost in translation, particularly without contextual guidance.

6.4 Language-Specific Performance Variations

Model performance varied noticeably across language families. Translation quality was notably better for languages such as Khortha and Sadri, while languages including Kaman Mishmi, Bodo, Ho, Santhali and Mundari prompted near-complete translation failures across most models. This variation suggests that latent linguistic knowledge from pre-training data may provide differential support for certain languages, even within the "low-resource" category.

6.5 Implications for Endangered Language Documentation

This comprehensive analysis reveals that while LLMs possess sophisticated text generation capabilities, their application to endangered language translation presents significant challenges. The models frequently function as "fluent fabricators" rather than faithful translators, emphasizing the critical need for contextual information, evaluation beyond surface-level fluency metrics, and human oversight in endangered language NLP efforts.

7 Conclusion

This study addresses the need for culturally grounded and linguistically inclusive benchmarks in NLP, particularly for endangered and low-resource languages. By using traditional recipes as a rich, multimodal lens into language and culture, we present a dataset that bridges computational research with community-driven knowledge preservation. Our findings reveal the limitations in current translation models, especially their tendency to generate fluent yet culturally inaccurate outputs in the absence of contextual guidance. These patterns highlight the inadequacy of standard evaluation metrics and stress the importance of human-in-the-loop methodologies and culturally sensitive prompts.

Our findings further suggest that translation errors in low-resource, culturally specific data are not merely linguistic but *epistemic* —arising from a lack of cultural grounding within current large language models. As observed in ELR-1000, models often replace indigenous tools, practices, or ingredients with globally dominant equivalents, reflecting their limited awareness of cultural context. Future research could therefore move beyond surface-level evaluation and explicitly probe LLMs cultural awareness through questionanswering tasks or internal probing methods that examine how cultural concepts are represented in model embeddings. Such investigations could reveal how cultural information (or its absence) propagates within translation paradigms, and how integrating culturally grounded datasets like ELR-1000 can enable models to move from syntactic fluency to contextual fidelity, preserving the lived realities embedded in endangered languages.

As large language models continue to expand their reach, our work offers a foundation for building more equitable, ethical, and context-aware AI systems. We also document the unique challenges faced during data collection and validation as a practical guide for others seeking to build similar community-rooted multimodal datasets. We see this dataset as a vital intervention in preserving the linguistic and cultural heritage embedded in everyday food practices. It also advances more meaningful evaluation of translation systems for underrepresented languages before such knowledge disappears from both community memory and the digital world.

Limitations

There are several limitations given the fact that this is only the beginning of what we believe should be a much larger community-authored effort. We list some of them below

- Translations were not possible for the full dataset in this phase as recruiting bilingual speakers was not the focus of this study—our primary aim was to document knowledge from native speakers, many of whom were monolingual. As a result, only a three recipes per language were translated and used for evaluation. This limits the generalizability of conclusions.
- In this work, we work on building resources for 10 endangered languages. While we believe this to be a great beginning, especially considering that all of these languages are endangered, we would like to scale our efforts to cover more such languages in the future. We believe that there is a great scope in this direction given the diversity of languages present in Eastern part of India.
- We collect approximately 100 recipes per language and release it in ELR-1000. This may be sufficient for benchmarking existing LLMs but may not enough to improve the LLMs to work better in this domain and language. In the future, we would like to work with more rural communities to crowdsource more recipes so that the collected dataset can be used to improve the abilities of LLMs in this domain.
- In this project, we limited ourselves to the cuisine domain. We did this so that the collected dataset is culturally relevant to the communities we worked with. However, cuisine is not the only topic that could help us collect culturally relevant data. Some other such topics could be the agricultural and livestock farming practices that these communities follow. Covering many diverse topics like this could make the collected benchmark more valuable. In the future, we would like to expand our efforts in this direction as well.
- Finally, although our dataset is multimodal—including audio, text, and images—it is not aligned to support benchmarking LLMs for advanced tasks such as knowledge graph construction or multimodal reasoning.

Acknowledgments

We gratefully acknowledge the local coordinators and participants whose time, knowledge, and effort made this work possible. We extend our sincere thanks to Divyansh and Abhijeet for providing technical support throughout this study. We are also deeply grateful to Saroj Kumar and Pradan Katoria

for their invaluable assistance in identifying and connecting us with local coordinators.

References

Basil Abraham, Danish Goel, Divya Siddarth, Kalika Bali, Manu Chopra, Monojit Choudhury, Pratik Joshi, Preethi Jyoti, Sunayana Sitaram, and Vivek Seshadri. 2020. Crowdsourcing speech data for low-resource languages from low-income workers. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 2819–2826, Marseille, France. European Language Resources Association.

Pulkit Agarwal, Settaluri Sravanthi, and Pushpak Bhattacharyya. 2024. IndiFoodVQA: Advancing visual question answering and reasoning with a knowledge-infused synthetic data generation pipeline. In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 1158–1176, St. Julian's, Malta. Association for Computational Linguistics.

Ritu Agarwal, Tanupriya Choudhury, Neelu J Ahuja, and Tanmay Sarkar. 2023. Indianfoodnet: Detecting indian food items using deep learning. *International Journal of Computational Methods and Experimental Measurements*, 11(4):221–232.

Advait Bhat, Nidhi Kulkarni, Safiya Husain, Aditya Yadavalli, Jivat Neet Kaur, Anurag Shukla, Monali Shelar, and Vivek Seshadri. 2024. Speaking in terms of money: Financial knowledge acquisition via speech data generation. *ACM J. Comput. Sustain. Soc.*, 2(3).

Yong Cao, Yova Kementchedjhieva, Ruixiang Cui, Antonia Karamolegkou, Li Zhou, Megan Dare, Lucia Donatelli, and Daniel Hershcovich. 2023. Cultural adaptation of recipes. *Preprint*, arXiv:2310.17353.

Manu Chopra, Indrani Medhi Thies, Joyojeet Pal, Colin Scott, William Thies, and Vivek Seshadri. 2019. Exploring crowdsourced work in low-resource settings. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI '19, page 113, New York, NY, USA. Association for Computing Machinery.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, Luke Marris, Sam Petulla, Colin Gaffney, Asaf Aharoni, Nathan Lintz, Tiago Cardal Pais, Henrik Jacobsson, Idan Szpektor, Nan-Jiang Jiang, and 3290 others. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *Preprint*, arXiv:2507.06261.

Sumanth Doddapaneni, Rahul Aralikatte, Gowtham Ramesh, Shreya Goyal, Mitesh M. Khapra, Anoop Kunchukuttan, and Pratyush Kumar. 2023. Towards leaving no Indic language behind: Building monolingual corpora, benchmark and models for Indic languages. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12402–12426, Toronto, Canada. Association for Computational Linguistics.

Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc'Aurelio Ranzato, Francisco Guzmán, and Angela Fan. 2022. The Flores-101 evaluation benchmark for low-resource and multilingual machine translation. *Transactions of the Association for Computational Linguistics*, 10:522–538.

Tianyi Hu, Maria Maistro, and Daniel Hershcovich. 2024. Bridging cultures in the kitchen: A framework and benchmark for cross-cultural recipe retrieval. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 1068–1080, Miami, Florida, USA. Association for Computational Linguistics.

Mohammed Safi Ur Rahman Khan, Priyam Mehta, Ananth Sankar, Umashankar Kumaravelan, Sumanth Doddapaneni, Suriyaprasaad B, Varun G, Sparsh Jain, Anoop Kunchukuttan, Pratyush Kumar, Raj Dabre, and Mitesh M. Khapra. 2024. IndicLLMSuite: A blueprint for creating pre-training and fine-tuning datasets for Indian languages. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15831–15879, Bangkok, Thailand. Association for Computational Linguistics.

Wenyan Li, Xinyu Zhang, Jiaang Li, Qiwei Peng, Raphael Tang, Li Zhou, Weijia Zhang, Guimin Hu, Yifei Yuan, Anders Søgaard, Daniel Hershcovich, and Desmond Elliott. 2024. Foodieqa: A multimodal dataset for fine-grained understanding of chinese food culture. *Preprint*, arXiv:2406.11030.

Vishu Madaan and Prateek Agrawal. 2022. Anuvaad: A hindi-sanskrit-hindi bilingual machine translation system using rule-based approach. *International Journal of Social Ecology and Sustainable Development (IJSESD)*, 13(1):1–14.

Arijit Maji, Raghvendra Kumar, Akash Ghosh, Anushka, and Sriparna Saha. 2025. Sanskriti: A comprehensive benchmark for evaluating language models' knowledge of indian culture. *Preprint*, arXiv:2506.15355.

Javier Marın, Aritro Biswas, Ferda Ofli, Nicholas Hynes, Amaia Salvador, Yusuf Aytar, Ingmar Weber, and Antonio Torralba. 2021. Recipe1m+: A dataset for learning cross-modal embeddings for cooking recipes and food images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(1):187–203.

Weiqing Min, Linhu Liu, Zhiling Wang, Zhengdong Luo, Xiaoming Wei, Xiaolin Wei, and Shuqiang Jiang. 2020. Isia food-500: A dataset for large-scale food recognition via stacked global-local attention network. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 393–401.

OpenAI, :, Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, Aleksander Mdry, Alex Baker-Whitcomb, Alex Beutel, Alex Borzunov, Alex Carney, Alex Chow, Alex Kirillov, and 401 others. 2024. Gpt-4o system card. *Preprint*, arXiv:2410.21276.

Omkar Prabhu. 2025. Khana: A comprehensive indian cuisine dataset. *arXiv preprint arXiv:2509.06006*.

David Romero, Chenyang Lyu, Haryo Akbarianto Wibowo, Teresa Lynn, Injy Hamed, Aditya Nanda Kishore, Aishik Mandal, Alina Dragonetti, Artem Abzaliev, Atnafu Lambebo Tonja, Bontu Fufa Balcha, Chenxi Whitehouse, Christian Salamea, Dan John Velasco, David Ifeoluwa Adelani, David Le Meur, Emilio Villa-Cueva, Fajri Koto, Fauzan Farooqui, and 57 others. 2024. Cvqa: Culturally-diverse multilingual visual question answering benchmark. *Preprint*, arXiv:2406.05967.

Florian Schneider and Sunayana Sitaram. 2024. M5 – a diverse benchmark to assess the performance of large multimodal models across multilingual and multicultural vision-language tasks. *Preprint*, arXiv:2407.03791.

Agrima Seth, Sanchit Ahuja, Kalika Bali, and Sunayana Sitaram. 2024. Dosa: A dataset of social artifacts from different indian geographical subcultures. *Preprint*, arXiv:2403.14651.

Bhuiyan Sanjid Shafique, Ashmal Vayani, Muhammad Maaz, Hanoona Abdul Rasheed, Dinura Dissanayake, Mohammed Irfan Kurpath, Yahya Hmaiti, Go Inoue, Jean Lahoud, Md. Safirur Rashid, Shadid Intisar Quasem, Maheen Fatima, Franco Vidal, Mykola Maslych, Ketan Pravin More, Sanoojan Baliah, Hasindri Watawana, Yuhao Li, Fabian Farestam, and 10 others. 2025. A culturally-diverse multilingual multimodal video benchmark and model. *Preprint*, arXiv:2506.07032.

Abhishek Kumar Singh, Vishwajeet kumar, Rudra Murthy, Jaydeep Sen, Ashish Mittal, and Ganesh Ramakrishnan. 2025. Indic qa benchmark: A multilingual benchmark to evaluate question answering capability of llms for indic languages. *Preprint*, arXiv:2407.13522.

Harman Singh, Nitish Gupta, Shikhar Bharadwaj, Dinesh Tewari, and Partha Talukdar. 2024. Indicgenbench: A multilingual benchmark to evaluate generation capabilities of llms on indic languages. *Preprint*, arXiv:2404.16816.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, and 1 others. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Ashmal Vayani, Dinura Dissanayake, Hasindri Watawana, Noor Ahsan, Nevasini Sasikumar, Omkar Thawakar, Henok Biadglign Ademtew, Yahya Hmaiti, Amandeep Kumar, Kartik Kuckreja, Mykola Maslych, Wafa Al Ghallabi, Mihail Mihaylov, Chao Qin, Abdelrahman M Shaker, Mike Zhang, Mahardika Krisna Ihsani, Amiel Esplana, Monil Gokani, and 50 others. 2025. All languages matter: Evaluating lmms on culturally diverse 100 languages. *Preprint*, arXiv:2411.16508.

Sshubam Verma, Mohammed Safi Ur Rahman Khan, Vishwajeet Kumar, Rudra Murthy, and Jaydeep Sen. 2025. Milu: A multi-task indic language understanding benchmark. *Preprint*, arXiv:2411.02538.

Ishaan Watts, Varun Gumma, Aditya Yadavalli, Vivek Seshadri, Manohar Swaminathan, and Sunayana Sitaram. 2024. Pariksha: A large-scale investigation of human-llm evaluator agreement on multilingual and multi-cultural data. *Preprint*, arXiv:2406.15053.

Genta Indra Winata, Frederikus Hudi, Patrick Amadeus Irawan, David Anugraha, Rifki Afina Putri, Yutong Wang, Adam Nohejl, Ubaidillah Ariq Prathama, Nedjma Ousidhoum, Afifa Amriani, Anar Rzayev, Anirban Das, Ashmari Pramodya, Aulia Adila, Bryan Wilie, Candy Olivia Mawalim, Ching Lam Cheng, Daud Abolade, Emmanuele Chersoni, and 32 others. 2025. Worldcuisines: A massive-scale benchmark for multilingual and multicultural visual question answering on global cuisines. *Preprint*, arXiv:2410.12705.

Silvana Yakhni and Ali Chehab. 2025. Can LLMs translate cultural nuance in dialects? a case study on Lebanese Arabic. In *Proceedings of the 1st Workshop on NLP for Languages Using Arabic Script*, pages 114–135, Abu Dhabi, UAE. Association for Computational Linguistics.

Jiawei Zheng, Hanghai Hong, Xiaoli Wang, Jingsong Su, Yonggui Liang, and Shikai Wu. 2024. Fine-tuning large language models for domain-specific machine translation. *CoRR*, abs/2402.15061.

A Evaluation Metrics

We employed a comprehensive 5-point Likert scale evaluation across four critical dimensions:

1. **Adequacy (Meaning Preservation):**
    - 1: No meaning preserved
    - 2: Minimal meaning preserved
    - 3: Moderate meaning preserved
    - 4: Mostly preserved with minor omissions
    - 5: Fully preserved meaning
2. **Fluency (Grammatical Correctness):**
    - 1: Extremely poor fluency
    - 2: Poor fluency with multiple errors
    - 3: Acceptable fluency, some errors
    - 4: Good fluency, few errors
    - 5: Excellent fluency, no errors
3. **Comprehensibility (Target Language Understanding):**
    - 1: Completely incomprehensible
    - 2: Difficult to understand
    - 3: Understandable with effort
    - 4: Mostly comprehensible
    - 5: Fully comprehensible
4. **Cultural and Contextual Appropriateness:**
    - 1: Inappropriate or offensive
    - 2: Significant cultural inaccuracies
    - 3: Some inaccuracies but generally acceptable
    - 4: Mostly appropriate with minor issues
    - 5: Fully appropriate and well-adapted

Each translation pair received scores across all four metrics, with human evaluators providing justifications for scores to ensure consistency and reliability in the evaluation process.

B Prompts

This section contains the system and generation prompts used in our experiments and the evaluation prompts for the LLM Judge.

B.1 Context Free System Prompt

```
You are an expert language translator. Please translate the given text from {lang} to English.
```

B.2 Context Aware System Prompt

```
You are a specialized linguist and cultural translator with expertise in endangered languages that
have minimal digital documentation. Your mission is to provide accurate, culturally sensitive
translations from {lang} to English while preserving the linguistic and cultural integrity of
the source material.
```

Language Background: Understanding {lang}

Geographic and Cultural Context

{lang} (also known as Nagpuri or Kurukh Sadri) is an Indo-Aryan language primarily spoken in:

- **Jharkhand** (main concentration in Ranchi, Gumla, Simdega districts)
- **West Bengal** (Purulia district)
- **Odisha** (Sundargarh district)
- **Assam** (tea garden communities)

Speaker Communities and Cultural Significance

- **Primary Speakers**: Kurukh/Oraon tribal communities, Munda speakers, and other Adivasi groups
- **Total Speakers**: Approximately 2-3 million (declining)
- **Cultural Role**: Language of inter-tribal communication, traditional storytelling, folk songs, and cultural ceremonies
- **Social Context**: Often used as a lingua franca among different tribal communities in the region

Linguistic Characteristics Affecting Translation

Script and Writing System

- **Traditional**: Devanagari script (as seen in examples)
- **Status**: Limited standardized orthography; oral tradition predominant
- **Challenge**: Spelling variations common due to lack of standardization

Key Grammatical Features

1. **Word Order**: Subject-Object-Verb (SOV) structure
2. **Agglutination**: Suffixes attached to root words for grammatical meaning
3. **Case System**: Nominative, accusative, genitive, and locative markers
4. **Verb Conjugation**: Complex tense-aspect system with evidentiality markers
5. **Honorific System**: Respectful and familiar speech levels

Vocabulary Characteristics

- **Core Vocabulary**: Mix of Indo-Aryan base with significant tribal language borrowings
- **Cultural Terms**: Rich vocabulary for:
    - Traditional foods and cooking methods
    - Forest products and gathering practices
    - Agricultural terms and seasonal activities
    - Kinship and social relationships
    - Religious and ceremonial concepts
- **Code-Switching**: Frequent mixing with Hindi, local tribal languages

Cultural Translation Considerations

Traditional Knowledge Systems

- **Ecological Wisdom**: Deep knowledge of forest ecosystems, medicinal plants, seasonal cycles
- **Food Culture**: Traditional recipes using indigenous ingredients (drumsticks, forest vegetables, tribal cooking methods)
- **Social Structures**: Extended family systems, community decision-making, age-based hierarchy
- **Spiritual Practices**: Animistic beliefs, ancestor veneration, nature worship elements

Common Cultural Concepts Requiring Careful Translation

- **’Hau ants’**: Specific type of edible ant collected seasonally - cultural delicacy
- **’Hari flower’**: Specific flora with cultural/medicinal significance
- **Market vs. Forest gathering**: Distinction between purchased and traditionally collected items
- **Seasonal activities**: Many terms tied to agricultural and gathering calendars
- **Community practices**: Collective cooking, sharing, and food preparation methods

Translation Challenges Specific to {lang}

Linguistic Challenges

1. **Limited Documentation**: Few dictionaries or grammatical resources available
2. **Dialectal Variation**: Regional differences in vocabulary and pronunciation
3. **Oral Tradition**: Many concepts exist only in spoken form
4. **Compound Words**: Complex formations requiring cultural knowledge to parse
5. **Implicit Cultural Knowledge**: Meanings embedded in cultural practices

Semantic Challenges

1. **Time Concepts**: Indigenous calendar systems and seasonal markers
2. **Spatial Relationships**: Land-based orientation systems
3. **Social Deixis**: Complex system of relationship-based pronouns
4. **Cultural Metaphors**: Nature-based imagery and traditional comparisons
5. **Ceremonial Language**: Formulaic expressions for rituals and celebrations

Recognition Patterns for Translation Success

- **Food/Cooking contexts**: Look for ingredient lists, preparation methods, storage practices
- **Market/Economic contexts**: Distinguish between purchased goods and gathered resources
- **Temporal markers**: Seasonal and daily activity references
- **Social contexts**: Community activities, family relationships, traditional practices
- **Natural world**: References to specific plants, animals, ecological relationships

Your Role and Responsibilities

You understand that {lang} is an endangered language with limited digital presence, meaning:

- Standard translation resources may not exist
- Cultural context is crucial for accurate interpretation
- Each text may represent irreplaceable linguistic heritage
- Community knowledge and oral traditions inform meaning
- Dialectical variations may exist without standardized documentation

Translation Methodology

Primary Translation Approach

1. **Semantic Accuracy**: Focus on conveying the core meaning rather than word-for-word translation
2. **Cultural Preservation**: Maintain cultural concepts even when English equivalents don’t exist
3. **Contextual Interpretation**: Use linguistic patterns and cultural knowledge to interpret ambiguous passages
4. **Transparent Limitations**: Clearly indicate when meaning is uncertain or interpretative

Handling Linguistic Challenges

- **Unique Grammar**: {lang} may have grammatical structures absent in English (complex evidentiality, agglutination, tonal meaning)
- **Cultural Concepts**: Preserve terms that represent unique worldviews or practices
- **Oral Tradition Elements**: Recognize formulaic phrases, ceremonial language, and storytelling conventions
- **Temporal/Aspectual Systems**: Navigate complex verb systems that may not map to English tenses

Output Structure

Standard Translation Format:

English Translation: [Your translation]
Confidence Level: High/Medium/Low

When Additional Context Required:

English Translation: [Your translation]
Confidence Level: High/Medium/Low

For Uncertain or Complex Content:

English Translation: [Best interpretation]
Alternative Interpretations: [Other possible meanings]
Uncertainty Factors: [What makes translation ambiguous]
Confidence Level: Low

Few-Shot Examples

Example 1:
English Translation: Agasti flowers, rice, salt, garlic, chili, turmeric, tomato are needed
Confidence Level: High

Example 2:
English Translation: Grind the soaked rice and put it in a jar
Confidence Level: High

Example 3:
English Translation: Store flowers in soup/water, rest of the ingredients in containers
Confidence Level: High

Example 4:
English Translation: Bhadur saag chutney is ready to serve
Confidence Level: High

Example 5:
English Translation: Till then wash Jirhul flowers and peel the skin of the potatoes and cut 2 onions.
Confidence Level: High

Example 6:
English Translation: You can eat this recipe with chapati or rice (based on your preference).
Confidence Level: High

Ethical Guidelines and Best Practices

Cultural Sensitivity

- Treat all content as potentially sacred or culturally significant
- Avoid imposing Western concepts on indigenous worldviews
- Preserve proper nouns and culturally specific terms when appropriate
- Acknowledge when content may require community consultation for full understanding

Linguistic Integrity

- Resist over-interpretation when evidence is limited
- Clearly distinguish between certain translation and educated inference
- Maintain scholarly objectivity while respecting cultural values
- Document linguistic patterns that might inform future translation work

Transparency and Humility

- Acknowledge the limitations of working with under-documented languages
- Be explicit about confidence levels and areas of uncertainty
- Recognize that community speakers may have insights unavailable through text alone
- Frame translations as interpretations rather than definitive meanings when appropriate

Final Reminders

Every text in {lang} represents irreplaceable cultural and linguistic heritage. Approach each translation as both a linguistic challenge and a cultural responsibility. Your work may be among the few digital records of this language's richness and complexity.
When in doubt, err on the side of preservation - maintain original terms with explanation rather than forcing inadequate English substitutes. Honor both the linguistic sophistication and cultural depth of {lang} in every translation.

B.3 Generation Prompt

Translate the given recipe text in triple back ticks: “‘ “‘

B.4 LLM Judge Evaluation Prompt

You are tasked with evaluating translations produced by a machine learning model against human translations. For each translation pair (human vs. machine), please provide a score between 1 and 5 based on the following qualitative metrics:

1. **Adequacy**: Evaluate whether the translation preserves the meaning of the source text. Rate how much of the source content is accurately conveyed in the translation.
   - 1: No meaning preserved
   - 2: Minimal meaning preserved
   - 3: Moderate meaning preserved
   - 4: Mostly preserved with minor omissions
   - 5: Fully preserved meaning
2. **Fluency**: Assess the grammatical correctness and naturalness of the translation in the target language. Consider syntax, idiomatic expressions, and stylistic appropriateness.
   - 1: Extremely poor fluency
   - 2: Poor fluency with multiple errors
   - 3: Acceptable fluency, some errors
   - 4: Good fluency, few errors
   - 5: Excellent fluency, no errors
3. **Comprehensibility**: Determine if a monolingual speaker of the target language can understand the translation. This is crucial for end-user applications.
   - 1: Completely incomprehensible
   - 2: Difficult to understand
   - 3: Understandable with effort
   - 4: Mostly comprehensible
   - 5: Fully comprehensible
4. **Cultural and Contextual Appropriateness**: Evaluate whether cultural references, idioms, and context are translated appropriately or adapted for the target audience.

  - 1: Inappropriate or offensive
  - 2: Significant cultural inaccuracies
  - 3: Some inaccuracies but generally acceptable
  - 4: Mostly appropriate with minor issues
  - 5: Fully appropriate and well-adapted

Please provide your scores for each translation pair along with a brief justification for each score based on the criteria defined above.
<HUMAN TRANSLATIONS>
Drumstick recipe Pan, stirring spoon, knife None It can be cooked anytime Taught by my grandfather Drumstick, tomato, cooking oil, turmeric powder, spices, salt, onion, chilly Plucked drumstick from the tree Heated the pan Poured oil Put two chillies in the oil I cut one onion I put the drumstick in the pan After two minutes, I added turmeric powder and spices to the pan. After one minute, I sprinkled salt I added three tomatoes After five minutes, the drumstick is ready Yes, it is good It is eaten with rice It can be eaten with rice and rice beer. Yes, everyone can eat it Yes, children like to eat drumstick no Can be found at home and in the local market too After purchasing from the market, we store it in the house no </HUMAN TRANSLATIONS>

<MACHINE TRANSLATIONS>
Sujani (Vegetable Curry)
Ingredients: - Wok/pan, spatula, knife - Oil - Soap nuts (for washing) - Wait for the elderly to wake up - Boil for one hour - Sujani (vegetables), salt, turmeric, cumin, spices, onions, green chilies
Instructions: 1. Clean the sujani vegetables thoroughly 2. Heat oil in the wok 3. Add turmeric powder 4. Add finely chopped green chilies 5. Add chopped onions and fry 6. Add the sujani vegetables and cook 7. Cook with cumin and spices for a few minutes 8. Add onions and cook for a few minutes 9. Finally add salt and cook 10. Cook the sujani until tender for a few minutes 11. Now it's ready to eat 12. Serve with rice 13. Eat hot rice with this curry 14. Everyone can enjoy eating this 15. This dish is very tasty
Notes: - Prepare this when you have time - This vegetable curry goes well with rice - Enjoy!
Note: "Sujani" appears to be a local term for certain vegetables or greens commonly used in Ho cuisine.
</MACHINE TRANSLATIONS>

C Instructions for Participants for recording Recipes

- Images:
  -Take as many pictures as possible
  -Take pictures of ingredients before and after chopping, peeling or cleaning
  -Take pictures after chopping or processing
  -Take pictures of utensils, vessels if any traditional vessels are used
  -Take pictures at each step, after adding each ingredient

- Recipe Steps:
  -Try to have at least 5-7 steps for each recipe.
  -Exclude the name of the recipe and ingredients name here

- Audio or text recording:
  -After every picture, add a caption in either audio or text format.
  -Ensure you are speaking in your native language
  -Record from not more than 6-7 inches away from the mouth of the speaker.
  -Try to avoid background noise or overlapping speech.
  -One audio clip should have the voice of only one person.

- Additional note:
  -Once you have finished taking up the photos, audios and text, please review the data and then do the final submission.