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, Vivek Seshadri
Karya Inc., Bengaluru, Karnataka, India

Deepthi Sudharsan, Kalika Bali
Microsoft Research
Bengaluru, Karnataka, India

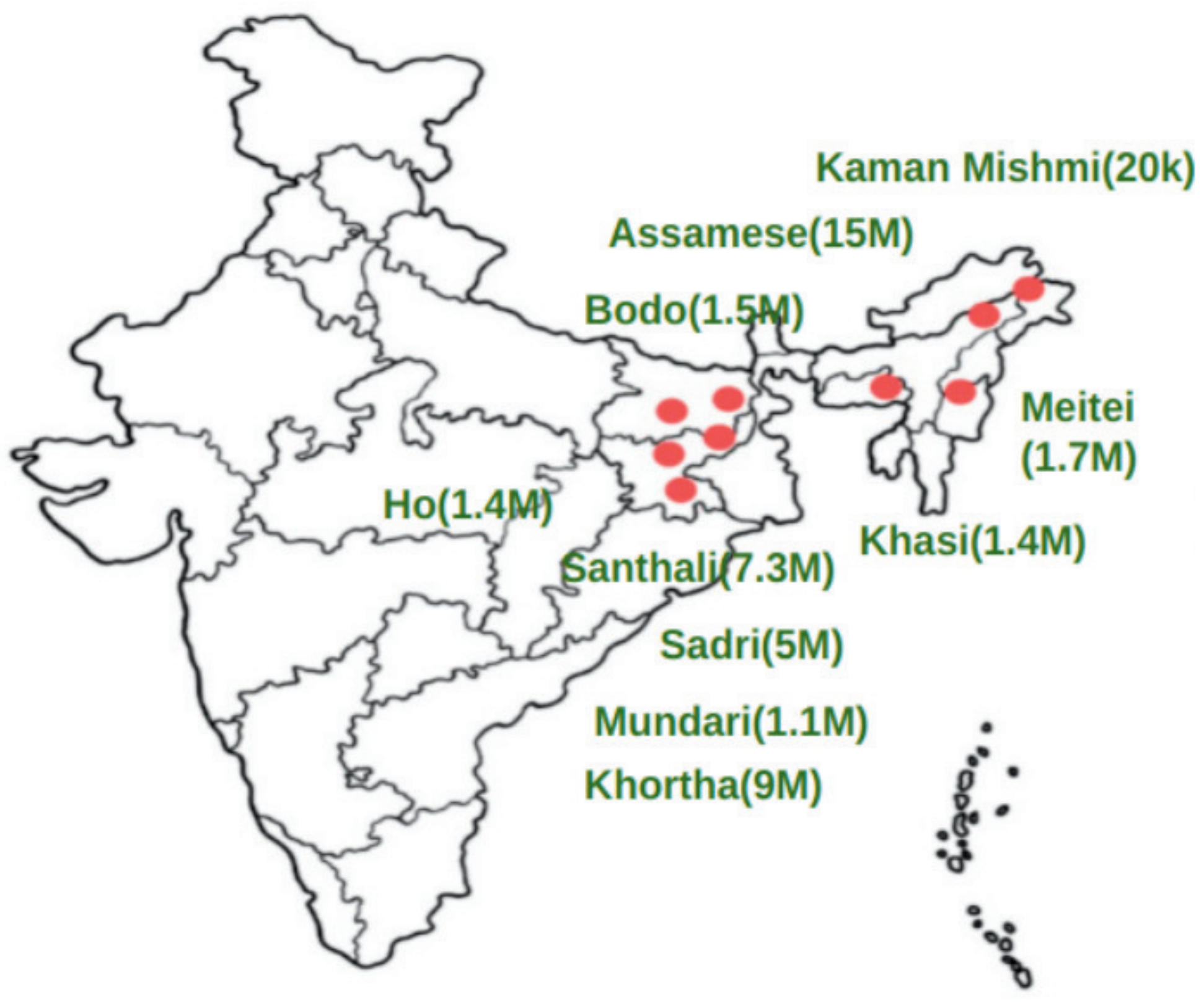
ABSTRACT

We introduce ELR-1000, a culturally grounded multimodal dataset of 1,060 traditional recipes collected from rural communities across Eastern India, representing 10 endangered languages. The dataset captures both culinary practices and the socio-cultural knowledge embedded in Indigenous food traditions, gathered via a mobile interface designed for low-digital-literacy contributors. Evaluating state-of-the-art LLMs reveals notable challenges in translating low-resource, culturally specific language—but providing contextual cues and cultural preservation guidelines leads to substantial improvements. By releasing ELR-1000, we aim to support equitable, culturally aware language technologies for underrepresented languages and domains.

RESEARCH GAP

- Existing benchmarks primarily focus on linguistic accuracy; our work explores cultural authenticity.
- Most available datasets follow translation workflows from English; community-authored approaches offer an alternative paradigm.
- Benchmarks incorporating endangered Indic languages with cultural context remain limited, presenting research opportunities.

OUR CONTRIBUTION: ELR-1000, A COMMUNITY-GENERATED DATASET



1,060 Traditional Recipes	10 Endangered Languages	6 Indian States
338 Rurual Women Contributors	46+ Hours of Audio	500+ Unique Ingredients

Data Samples

Chambai,
Kaman Mishmi dish

Amaltas flower
(Cassia fistula)

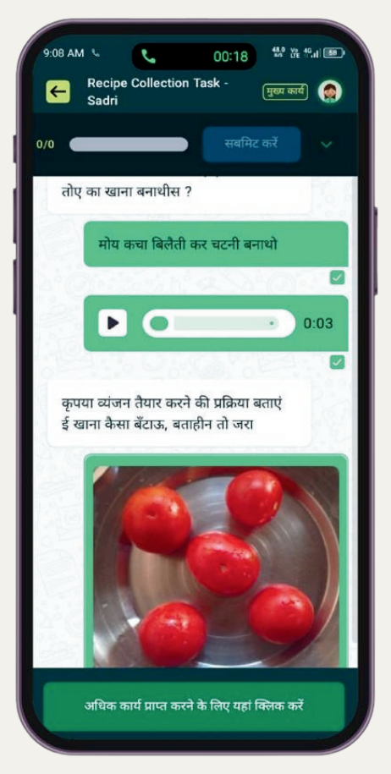
Red ant eggs

Kachnar flower *dish*
(Bauhinia variegata)

Language	State	Recipes	Unique Ingredients	Images	Audio Duration (hh:mm:ss)
Mundari	Jharkhand	82	85	703	09:33:52
Sadri	Jharkhand	107	104	1103	06:27:19
Santhali	Bihar	120	98	1004	08:52:36
Khortha	Bihar	126	73	1129	11:15:33
Ho	Jharkhand	91	80	875	04:13:26
Assamese	Assam	113	148	1415	04:21:54
Bodo	Assam	95	190	1532	25:46:36
Meitei	Manipur	100	97	580	12:13:28
Khasi	Meghalaya	98	89	1928	17:59:00
Kaman Mishmi	Arunachal Pradesh	128	92	1129	20:22:07

Data Collection Methodology

- Grassroots Partnerships Partnered with local NGOs in Jharkhand & the Northeast—regions with rich tribal cultures but facing linguistic vulnerability (UNESCO "Vulnerable" status)
- **Capacity Building Training Sessions:** Karya trained coordinators → Coordinators mobilized & onboarded workers → First-time digital participants collected recipes in their local languages
- **Continuous Support** Daily handholding + WhatsApp groups + Weekly monitoring = Successful community-driven data collection



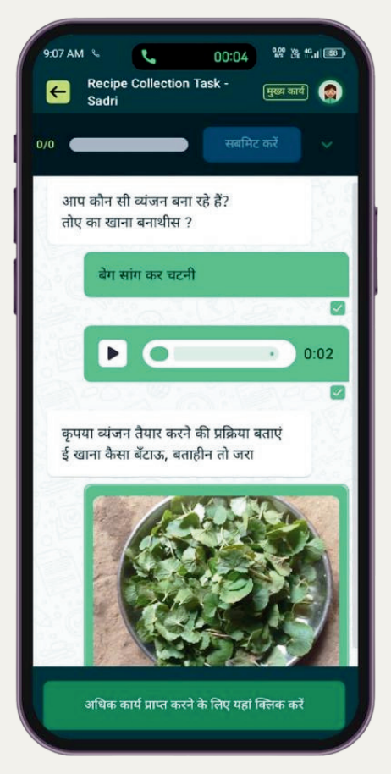
Recipe Collection Task - Sadri

तोए का खाना बनाथीस ?

मोय कचा बिलैती कर चटनी बनाथो

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

अधिक कार्य प्राप्त करने के लिए यहां क्लिक करें



Recipe Collection Task - Sadri

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

बेग सांग कर चटनी

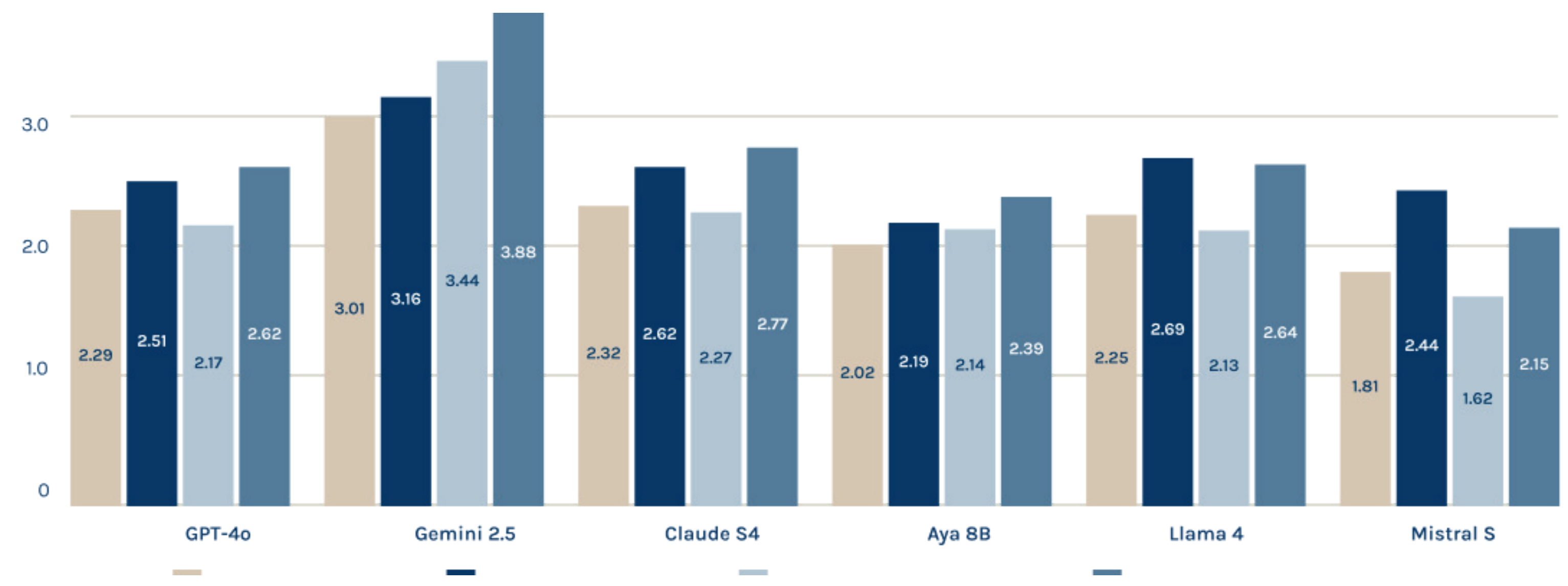
कृपया व्यंजन तैयार करने की प्रक्रिया बताएं ई खाना कैसा बँटाऊ, बताहीन तो जरा

अधिक कार्य प्राप्त करने के लिए यहां क्लिक करें

Existing LLM Knowledge Gap

- High Awareness, Low Capability: LLMs show strong factual knowledge about East Indic languages (e.g., demographics, regions, language families) but lack practical linguistic competence.
- The Functional Gap: They struggle with functional translation—the step from "knowing about" a language to accurately "generating" it remains weak.
- Cultural Complexity: Models often miss nuances in indigenous ingredients and cooking methods, leading to oversimplified or incorrect translations.

Overview of Existing Work and Remaining Gaps in Food Knowledge Documentation



Experimental Methodology

Model selection

- 6 models: 3 proprietary (Gemini 2.5, GPT-4o, and Claude sonnet 4) & 3 open source (Llama, Mistral, and Aya) to ensure fair comparison across licensing types.
- Prompting: 3 Recipes from each Language
- Conditions: Generated content under no context vs. contextual settings

Evaluation metrics

- All LLM outputs compared against human-verified gold standard translations.
- Adequacy , Fluency, Comprehensibility, and Cultural Appropriateness.

Hybrid judging

- **LLM Judges:** Gemini 2.5 Pro and GPT-5 were used to perform initial grading.
- **Human Oversight:** Native Speakers reviewed a sample to validate the LLM Judge scoring, focusing on instances of **low cultural appropriateness**

Key Findings

- **Normalization Bias:** Persistent replacement of indigenous tools/methods with globally dominant, Western equivalents (e.g., "chopping board")
- **Epistemic Implication:** Errors are not just linguistic; they show a profound lack of cultural grounding.
- **Takeaway**: For endangered languages, Cultural Context is the foundation for functional translation, not just an optional enhancement.

Source / Gold Standard Item	LLM Mistranslation (Systematic Error)
Dish ingredient: **Star Fruit**	Bamboo Shoot
Traditional Tool: **Mortar and Pestle**	Chopping Board (Western Bias)
Key Ingredient: **Silkworms**	Replaced with **Mushrooms** or **generic Chicken Curry** (Hallucination)

ACKNOWLEDGEMENTS

We gratefully acknowledge the local coordinators and contributors whose time, knowledge, and effort made this work possible. And special thanks to Divyansh Kushwaha and Abhijeet Shukla from Karya Tech team developed and maintained the mobile data collection platform