

Behind-the-Scenes Challenges in Sino-Tibetan Low-Resource Language Technology Development

Shuheng Liu and Michael L. Best

Georgia Institute of Technology
sliu775@gatech.edu, mikeb@gatech.edu

Abstract

Despite persistent efforts to develop low-resource language technologies (LRLTs), little attention has been paid to LRLT practitioners’ experiences, challenges, and opportunities. We employ a qualitative approach and conduct 15 semi-structured interviews with LRLT practitioners working on three Sino-Tibetan low-resource languages. Through our analysis, we identify not only technical challenges such as a lack of textual data and reliable language processing tools, but also linguistic, social, and financial challenges that cause and exacerbate them. Fortunately, we also find that participants are generally optimistic about LRLT development and mitigate these challenges by fostering and seeking support from diverse sources and leveraging large language models. We discuss the complex nature of the challenges in relation to prior literature and provide recommendations for practitioners, governments, and the broader NLP community.

1 Introduction

Low-resource language technologies (LRLTs) development has become one of the most prominent and fastest-growing natural language processing (NLP) research topics over the years (Schopf et al., 2023). From cross-lingual transfer learning (Conneau et al., 2018; Ruder et al., 2019) to data augmentation (Fadaee et al., 2017; Liu, 2022), researchers have proposed various technical approaches to extend the development of language technologies (LTs, including digital artifacts for languages such as input methods, fonts, and language corpora, as well as language models) in low-resource settings.

Nevertheless, the challenges associated with LRLT development extend beyond algorithmic and technical considerations. While analyses of LRLT artifacts produced by different communities (Yu et al., 2022) can infer certain challenges, the NLP

literature rarely documents those associated with the prerequisites for successful LRLT development. Making these often-overlooked challenges explicit can help direct the efforts of those who support, or seek to support, LRLT development toward the areas of greatest need, while reducing the burden on LRLT practitioners by better aligning support with their actual needs.

Accordingly, we adopt a qualitative research approach, specifically interviews with LRLT practitioners, to surface the “behind-the-scenes” challenges that are typically absent from published accounts. We recruited 15 practitioners who had worked on Cantonese, Southern Min, or Tibetan LTs and conducted semi-structured interviews covering broad topics (Knott et al., 2022). These languages have relatively high levels of resource availability (Joshi et al., 2020) and the largest bodies of published NLP research (Liu and Best, 2025) among Sino-Tibetan low-resource languages (LRLs). As we aim to find general patterns across different LRLT communities, these languages serve as a pragmatic and informative set that allows us to study these challenges under comparatively favorable conditions. We asked participants about their experiences developing LRLTs for the language or languages on which they had worked. Specifically, we are interested in the following research questions:

RQ1. What are the different challenges faced by the Sino-Tibetan LRLT practitioners?

RQ2. What do practitioners identify as ways for them to mitigate these challenges?

Our analysis finds that LRLT practitioners must also overcome the lack of standardization in both written and spoken language when developing technologies (§ 4.1.1), as well as the lack of collaboration and resource support inherent in LRLT development (§ 4.1.3), beyond the technical challenge

of the need for more reliable language processing tools (§ 4.1.2). These challenges are not standalone but often entangled with each other such that they collectively aggravate the already restrictive low-resource settings. It is also surprising how challenges differ by language, even when practitioners work in similar low-resource settings. However, LRLT practitioners have attempted to mitigate these challenges with some success through support from the existing LRLT development communities and related areas (§ 4.2.1), further highlighting the need for increasing support and collaboration. More recently, they have leveraged advances in large language models (LLMs) to directly address some critical issues in LRLT development (§ 4.2.2). We discuss what the broader NLP community can learn from these challenges and opportunities, and we recommend how LRLT practitioners can better overcome them. The contribution of this paper can be summarized as:

1. We identify the challenges faced by Sino-Tibetan LRLT practitioners. The challenges may also be applicable to NLP practitioners working on other low-resource settings.
2. We identify the ways through which these practitioners are able to mitigate the challenges, and we suggest recommendations on how these challenges can be overcome for the advancement of LRLT development.

Through interviews and qualitative data analysis, we reveal deeper insights about LRLT communities. We hope this work inspires more qualitative and participatory research in NLP, advancing LTs through a human-centered approach.

2 Related Work

2.1 LRLT Development

LRLT development has become an important sub-field of NLP, with its dedicated track in major conferences (e.g. "Low-Resource Methods for NLP" in ACL and EMNLP), workshops (e.g. LoResLM¹), conferences (e.g. LaTeLL²) and journals (e.g. TAL-LIP³). Specifically for Cantonese, Tibetan, and Southern Min, most LT research centers around topics such as corpus creation (Meelen et al., 2021), machine translation (Hong et al., 2024; Zhou et al.,

2024), and speech recognition (Yang et al., 2023; Li et al., 2025). There has also been more research on how LLMs can support these languages, creating benchmarks for language models (Cheng et al., 2025), or training LLMs for LRLs. (Lai et al., 2024; Pan et al., 2025; Sha et al., 2025).

2.2 Qualitative and Participatory Research

Participatory research (Cornwall and Jewkes, 1995; Vaughn and Jacquez, 2020), i.e., collaborative research between researchers and community members to understand a research topic from a community's perspective. Community involvement can broaden how the success of LT development is evaluated by shifting attention beyond technical performance to whether technologies advance community-defined goals, support culturally meaningful practices, and contribute to language vitality (Bird, 2020). NLP researchers have been increasingly employing participatory research. Dolinska et al. (2024) present the study most similar in design to ours. Through structured interviews with 16 speakers representing 6 endangered-language communities in Thailand, they examined speakers' perceptions of their languages, their communities' interest in and needs for digital tools, and the challenges surrounding their development. Harvey et al. (2025) conducted semi-structured interviews to investigate challenges faced by practitioners when using publicly available NLP tools to measure representational harms. Siminyu et al. (2023) also conducted interviews with content creators and language technologists to investigate their motivations and challenges in African LT development, and provide recommendations. Beyond interview-based studies, participatory research can involve native speakers in other stages of the LT development pipeline. Native speakers may participate in collecting and validating speech data covering domains relevant to their communities (Huaman et al., 2026). Researchers may also consult speaker communities to understand their priorities and incorporate community feedback into the design and continued development of tools such as bilingual dictionaries (Hunt et al., 2023).

Together, these participatory and qualitative studies demonstrate how community involvement can reveal local goals and technology needs. Participatory approaches can also support more sustainable LT research communities (Nekoto et al., 2020), and community ownership and governance of language data (Hunt et al., 2023).

¹<https://loreslm.github.io>

²<https://latell.org/2026/>

³<https://dl.acm.org/journal/TALLIP>

3 Method

We conducted 15 semi-structured interviews, i.e., interviews with a topic guide instead of a closed set of questions, with researchers and software developers who have worked on LT development for Cantonese, Southern Min, and Tibetan after approval from the Georgia Institute of Technology Institutional Review Board. We detail the procedure for recruitment, interview, and analysis below.

3.1 Recruitment

We followed the purposive sampling (Etikan et al., 2016) by sending email invitations to recruit participants who have worked in the technologies for at least one of the three targeted languages and have at least three years of experience working in LTs for developing applications for LRLs and/or publishing academic papers about LRLTs. The participants’ email addresses were found on their publications and through Internet searches. In addition, we employed snowball sampling (Parker et al., 2019) by asking interview participants to connect us to potential participants. We stopped recruiting participants when we began to notice repetitive names of potential participants and themes (Czarniawska, 2014). In the end, we recruited 15 participants to conduct the interviews. The demographics of the participants are listed in Table 1.

3.2 Interview Procedure

The first author conducted semi-structured interviews with 8 participants in-person and 7 online. Interviews were conducted in either Mandarin or English, based on the participant’s preference. Most interviews lasted from 60-80 minutes and were audio-recorded after obtaining the participants’ consent. Each interview was divided into three major sections. First, we asked about the participant’s background in developing LTs and LRLTs. Next, we proceeded with questions about the current development of technologies for the target Sino-Tibetan language, including the main research foci and challenges. Lastly, we ended the interview by asking about their opinions on the future of LRLT development.

3.3 Data Analysis

To ensure confidentiality, we used a modified version of the OpenAI Whisper model (Radford et al., 2023) on GitHub⁴ that could be downloaded and

PID	Language	Base	Years
P1	Cantonese	East Asia	10+
P2	Cantonese	East Asia	10+
P3	Cantonese	East Asia	5-10
P4	Cantonese	North America	5-10
P5	Cantonese	East Asia	3-5
P6	Southern Min	East Asia	10+
P7	Southern Min	East Asia	10+
P8	Southern Min	East Asia	10+
P9	Southern Min	East Asia	10+
P10	Southern Min	East Asia	10+
P11	Southern Min	East Asia	5-10
P12	Tibetan	Europe	10+
P13	Tibetan	Europe	10+
P14	Tibetan	East Asia	5-10
P15	Tibetan	North America	3-5

Table 1: Participants’ information showing their assigned participant IDs (**PID**), the Sino-Tibetan LRL that they are working on (**Language**), where they are currently working (**Base**), and the number of years that they have worked in LT development (**Years**).

run locally to automatically generate transcriptions for the recorded interviews. We used the medium model size with all other hyperparameters set to default. The generated transcripts were examined and coded by the first author using a computer-assisted qualitative data analysis software (CAQ-DAS), QualCoder⁵ for data analysis. With our research questions borne in mind, we conducted an inductive thematic analysis, drawing on grounded theory (Glaser and Strauss, 1967) coding techniques including descriptive coding, i.e., annotating conversation segments by summarizing with a short phrase (Saldaña, 2009), and in vivo coding, i.e., annotating conversation segments using the actual language of the conversation (Glaser and Strauss, 1967), to code everything related to the research questions. Finally, these codes were re-organized and grouped into different categories to answer our research questions.

4 Findings

Our analysis yielded three main types of challenges. The lack of standardization in both written and spoken language (§ 4.1.1) creates an additional barrier to data collection and limits the usefulness of developed LRLTs. Basic processing tools

⁴<https://github.com/SYSTRAN/faster-whisper>

⁵<https://github.com/ccbogel/QualCoder>

for LRLs often perform unreliably (§ 4.1.2), even though further LT development and real-life deployment may require near-perfect accuracy. Practitioners receive limited support within their communities, constraining collaboration and artifact creation and sharing, and inadequate government support, including limited funding and policy backing. (§ 4.1.3). Practitioners identified two ways of mitigating these challenges: seeking collaborative and financial support from a diverse range of communities, institutions, and individuals, where such support was available (§ 4.2.1), and leveraging emerging LLM capabilities to assist with various stages of the LRLT development pipeline (§ 4.2.2).

4.1 Three Types of Challenges

The core challenge for LRLT development lies in the term "low-resource" itself, i.e., there is a lack of data (more specifically, textual data, in the current paradigm of NLP) for these languages. Yu et al. (2022) found that most input text (unlabeled text) is collected from the web. However, such a data collection method is limited in low-resource contexts. P6 gave specific figures to illustrate the struggle in Southern Min data collection through web scraping: *"The text that we can get from the Internet is only about 10 million characters, and that is because some writers are donating their work. Otherwise, what we could have scraped was only about 5 million characters."* 10 million Chinese characters take up a storage size of around 20-30 MB, which is even smaller than the classic SQuAD dataset (Rajpurkar et al., 2016), the famous English Question-Answering dataset that was annotated and released roughly a decade ago. The large pre-training corpora these days for LLMs, such as C4 (Raffel et al., 2020), which was used to train many LLMs including LLaMA (Touvron et al., 2023), are on the order of terabytes.

While data scarcity is the main challenge for the development of LRLTs, our analysis found challenges with the linguistic difficulties of the languages themselves, the limited resources and financial support, and the need for developing basic tools – all of which collectively contribute to the low-resource-ness of these languages.

4.1.1 Lack of Standardization

Textual data can only be collected when people's thoughts are written down. This is not trivial as reflected by P5 when they were trying to extend their work to Shanghaiese (a dialect of Wu Chi-

nese spoken mainly in Shanghai): *"If you have 10 million speaker, you ought to have a standardized orthography, or standardized romanization, and the romanization need to come with a character dictionary, and a compatible input method, and of course, textbooks. I just assume these are going to be taken for granted, but that is not the case."* While all three languages in our study have some prescribed orthographies, we found that each of them is associated with a reason why collecting LRL textual data might be proven difficult.

1. Written Cantonese *"lacks standardization"* (P1). Even when there is a standard orthography published by the government, *"We might not [write] according to that standard"* (P2). People write Cantonese that often deviates from existing standards, with homophonic characters and the Latin alphabet.
2. Southern Min can be written in *"Chinese characters and two forms of romanizations"* (P10), but most researchers (and speakers) *"cannot read or write"* (P7) the language even when they speak and understand it.
3. While there is a single standard orthography used to write Tibetan, *"We have very, very little text of people writing in their mother tongue"* (P13) and *"even newspapers now in Tibetan sound a lot like classical Buddhist texts, much more than they sound like how people speak on the street"* (P12).

Resource-rich languages use standardized writings adopted by the government and used by their native speakers, often to reflect how they use the language in everyday life. While there might be dialectal and individual differences, spellings remain relatively homogeneous for each dialect. However, for LRLs with non-standardized writing, no writing at all, or if the standards are not used by the native speakers, creating clean, homogeneous textual corpora becomes challenging.

What about spoken data – if we only collect spoken data instead of fixating on writing them down in some prescribed orthography, would this be an easier alternative? Unfortunately, this form of data collection could be equally challenging for LRLs, if not more. Since a lot of LRLs are also marginalized languages where they are spoken, instead of having one standardized form of speech, dialects of these languages are spoken. P6, P7 and P11 all

pointed out how Southern Min spoken in northern versus southern Taiwan carries different accents. Many ASR models would only recognize one dialect of the language, struggling with other dialects (P11) or foreign accents (P9). In some cases, what people often understand as dialects can be mutually unintelligible, such as Amdo Tibetan and Lhasa Tibetan (P12; Tournadre, 2014), so that developing LTs for only one "dialect" of the language will not be usable by speakers of another "dialect".

Even just in the collection of written or spoken data, our findings show that this lack of standardization is slowing down LRLT development under the current development paradigm. The standardized orthography and speech adopted by speakers are often the hidden prerequisites for large-scale data collection, and the absence for these LRLs becomes the underlying reason holding back LRLT development, yet changing the status quo is out of reach for these LT practitioners.

4.1.2 Unreliable LRL Processing Tools

Participants identified the unreliable language processing tools for LT development as a critical challenge for LRLT development. This insufficiency ranges from the need of evaluation tools such as benchmarks (P3, P15) and auxiliary tools such as OCR (P14) and grapheme-to-phoneme (G2P, P5), to the absence of a good tokenizer (P12, P13), font (P14), or even dictionary (P5). Sometimes practitioners are forced to work with low-performing tools and noisy corpora (P3, P8) without alternative choices. Participants raised concerns about the quality of available tools and corpora, emphasizing the need for more reliable tools and higher-quality data to support NLP development:

"The core problem is that data are essential for training an AI model, whether an LLM or another type of model, yet such data are extremely scarce, especially structured datasets. Much of the available data is like what I mentioned earlier - they are Cantonese data, but they are also Mandarin data. They are woven together. To put it bluntly, the quality is extremely poor." (P3)

"The bar for AI is that [...] if it can do 80% of [a task] right, [...] that is pretty good. [...] I just find a lot of it is noise that doesn't have the level of precision needed to build other stuff on top. If it is

directly consumer facing, I think that's okay in certain contexts, but if you are looking to build NLP stuff that other people lean on, it is a really shaky foundation." (P5)

However, deployable LRLTs in real life are often expected to be perfect. Tools such as subtitle generators (P10) or G2P (P5) for labeling character pronunciation are expected to have an error rate within 5% or even smaller to be usable in real life. P5 elaborated that *"It was accepted that [G2P] is not going to get any better than 1 in 19 [...] 1 in 19 means that if you want to do a relatively thin book of 200 pages, or 150 pages, about 50,000 characters, you're looking at 5,000 errors that you need to find. Your error management needs an error management system of its own."* Moving beyond concerns about technical performance, P14 warned that uncontrolled outputs from generative AI systems could pose broader risks for marginalized LRL communities: *"This could potentially involve sensitive ethnic issues. [...] once it produces content that cannot be controlled, no one may be able to contain it, and it could be very difficult for anyone to bear the consequences of the resulting public controversy or political problems."*

Developing reliable foundational tools is thus crucial both to advancing LRLT development and to ensuring that LRL communities can benefit from these technologies while minimizing potential harms. Continued reliance on existing tools risks perpetuating a vicious cycle in which weak foundational components impede the development of more reliable and useful LRLTs.

4.1.3 Limited Support and Funding

Data scarcity can also be exacerbated by the lack of support from LRLT development communities and local governments. Success in data collection for resource-rich languages, such as English and Mandarin Chinese, depends not only on data availability, but also on the collective effort of the NLP community to open-source the collective resources. This includes open-source data collected from large-scale web scraping projects such as the CommonCrawl project⁶, as well as uploaded to and shared through platforms such as GitHub and HuggingFace. For LRLs, however, we found an interesting pattern that even when corpora are collected,

⁶<http://commoncrawl.org>

practitioners are often unwilling to open-source them to their communities:

"No, I have never encountered cases of sharing corpora, and it is unlikely [...] I bought a corpus from [a university]. They wouldn't share it with me. I had to buy it from them." (P11)

"So some papers of [nationality] researchers saying 'Oh, we did this and that, and we have this big corpus, and we have these numbers', but there's never any data. So it's like, well, good for you, but can't use any of it [...] It's probably very well funded, but nothing comes out of it [...] One paper, and, you know, there's no corpus [...] It's probably good, but we have no idea what it is." (P13)

Some participants justified such practice. P14 commented that *"If you spent all the funding and finally made it [...] you might not be willing to share."* Similarly, P6 attributed this practice to some members of the LRLT community and the *"immaturity of the field"*, saying that *"More opportunities of sharing might come from more people. And [English and Mandarin NLP] is relatively more mature. When something is immature – still in its early stage – such sharing will be difficult."*

The lack of support within the LRLT ecosystem also manifests in limited collaboration and difficulties recruiting and retaining new talent. There is a lack of collaboration between academia and industry (P3), as well as within academia (P11). P1 believed that there should be more conversations across different communities, but there had not been *"much interaction between the AI people with linguists"*, and people working on Sinitic languages kept *"reinventing the wheel"* and *"doing silly things again and again"*. Moreover, fewer students are entering the field, and *"it's hard to keep them"* when the alternative job options are often more lucrative (P8). Together with limited resource sharing, these challenges weaken the collective capacity for LRLT development.

Another major part of the support comes from the government, which drafts policies and provides funding for practitioners of technological development. For LRLTs, however, such support can often be insufficient, or sometimes entirely missing. P6 commented that government support can be limited and *"nominal but not real"*, as both P6 and P7 noted

that there had not been specialized governmental agencies and policies for the equality of Southern Min despite recognizing the endangeredness of the language. For example, while Southern Min is now taught in school, the education, with only 1-2 hours of lessons every week, is insufficient for students' mastery of the language (P6).

When discussing practitioners gathering in various conferences, P1 pointed out that while there are conferences for data collection and even LT development for specific languages, *"conference participation is an expensive thing, and low-resource language people don't have that [money]"*. Almost every phase of LT development, from data collection and model training (P5, P12, P14) to student (P8, P9) / participant recruitment (P8, P12) and conference traveling, depends on funding, yet for LRLT practitioners, securing funding is a major hurdle as what they develop can be deemed as unprofitable (P4, P7, P11), not easily commercialized (P8), or having less impact in academia and industry (P5, P12). They often have to *"compete with [...] not just NLP projects"* (P1), and attempt to improve their chances for funding, whether to expand their work to more profitable and commercializable product with the addition of resource-rich languages to attract more users (P6, P7), or tweak their proposal by adding *"buzzwords"* (P5) / *"sprinkles on a cupcake"* (P12), demonstrating their projects' values appealing to those providing funding. When funding of 10 thousand USD cannot be guaranteed (P5), LRLT development becomes hardly possible. New and existing practitioners might switch their focus to areas with more available funding and compensation, exacerbating the aforementioned challenge of recruiting and retaining talent.

Overall, the lack of standardization (and thus textual and spoken data), collaboration, funding, and tools all influence and affect each other. They collectively contribute to the overall problem of low-resource-ness and amplify the hardship in LRLT development. This leads LRLT development into the downward spiral described by the Matthew Effect (Merton, 1968), i.e., *"the poor get poorer"*, causing these challenges to become even more serious in the future.

4.2 Means of Mitigation

Despite the challenges, our interview participants deeply care about LRLT development and are highly motivated to preserve the identities and cultures of LRL speakers (P5, P11, P14, etc.) espe-

cially when they themselves are native speakers, improve the lives of vulnerable speaker population (P8, P9), and by intellectual curiosity (P12, P15). These motivations encourage practitioners to find ways to mitigate the challenges in LRLT development, including seeking existing community support and utilizing LLMs in their work.

4.2.1 Seeking Support from Diverse Sources

In Section 4.1.3, we showed that many practitioners struggled to obtain support, particularly through resource sharing and funding. However, such support was uneven rather than absent. Participants' accounts suggest that progress in LRLT development often depends on assistance from research communities, government bodies, and interested individuals. For practitioners who received it, this assistance helped mitigate several challenges associated with LRLT development.

Participants pointed to international research conferences and workshops for potential venues to seek collaboration for LRLT development. Oriental COCOSDA⁷, EURALI⁸, and workshops for regional LRLs in LREC⁹ and CoLang¹⁰ were mentioned. Participants commented that these are ways that "allow me to learn what neighboring countries are working on" (P6) and for people to "gather" and "talk to each other" (P1).

Research labs will also collaborate with IT companies (P1, P6) and other research labs, sometimes in another country (P3, P15). To seek collaboration with limited funding, practitioner-initiated communities for LT development are formed. Researchers and developers communicate through social media platforms to share their interests and discuss progress in LRLT development (P1, P4). The resulting networks could provide crucial support to practitioners entering the field, as reflected by P5: *"The one thing that is very, very important in the whole project, making any way forward, is support from people like [a fellow practitioner], who, when I have brought nothing to the table yet, still helped me try to get something on the table. And the Cantonese Linguistics community tends to be quite welcoming, and they're willing to provide advice, provide help, even though they may already be legends in what they do."*

Support for LRLT projects can come from gov-

ernments, nongovernmental organizations, private individuals, and speaker communities, taking both financial and nonfinancial forms. Although participants generally described public funding as limited, P6 and P8 reported receiving it, while P1 and P14 were aware of other government-funded projects. P13 noted that private individuals and nongovernmental organizations could also provide alternative sources of funding. Support from speakers extended beyond direct financial assistance: P10 described how they could promote new LT products on social media and purchase them, while P1 noted that they could also participate in beta testing.

Although substantially more support is still needed from governments, nongovernmental organizations, research communities, and individual benefactors, participants' accounts suggest that current progress in LRLT development relies heavily on the financial, emotional, and practical assistance provided by these actors.

4.2.2 Emergence of LLMs

Prior work has revealed LLMs' limited performance on LRLs (Court and Elsner, 2024; Jadhav et al., 2025) and raised social and ethical considerations of using LLMs for LRLs (Smart et al., 2024). Nevertheless, the majority of our participants express a positive attitude towards the effect of LLMs on LRLT development.

Because all three languages in our study are represented to varying degrees in existing LLMs, participants reported that these models could help address critical challenges in LRLT development, including shortages of data and basic processing tools. P1, P3, and P11 discussed leveraging LLMs to generate LRL data through data augmentation for model training. P12 described a pipeline for human-LLM collaboration. LLMs can help with data annotation, which is often expensive if done manually, and practitioners develop filters using different heuristics to improve the quality of the automatic annotations. This aligns with recent literature describing the usage of LLMs to help with various aspects of LRLT development, including annotation (Kholodna et al., 2024) and data augmentation (Wang et al., 2025). With the challenge in language education (§ 4.1.3), P6 proposed that LLMs can help language instructors prepare teaching materials or be customized to check students' learning progress. Most participants believed that "LLM is a trend" (P15), and generally felt optimistic about the positive impact, i.e., helping with

⁷<https://www.oriental-cocosda.org>

⁸<https://sites.google.com/view/eurali/>

⁹<https://aclanthology.org/venues/lrec/>

¹⁰<https://www.colanginstitute.org/attend>

basic language tasks, assisting with language education, and potential future capabilities that LLMs can bring to LRLT development.

However, despite all the benefits and practical assistance listed above, some participants also expressed concerns about how LLMs will shape the future for LRLs and LRLTs, even including participants who find LLMs useful. *"If we envision our products to be able to translate everything and we don't have to learn but just speak our own languages, then it becomes a challenge [...] If everyone is too lazy to learn, then wouldn't it accelerate the extinction of low-resource languages?"*, a concern of LLMs discouraging people from learning other languages expressed by P9 and also shared by P7. Similarly, P6 was worried about the LLMs' outstanding performance on resource-rich languages potentially forcing users to give up on using LRLs. Some voiced their distrust in using LLMs for LRLT development due to LLMs' lack of interpretability, unexplainable reasoning ability (P8), and hallucination in their generations (P5).

LRLT practitioners are actively enlisting the assistance of existing communities, organizations, and interested individuals that they know of, as well as the capabilities of LLMs to help them combat the innate challenges of LRLT development. However, the help that they are receiving is far from enough, and we call for increasing attention and assistance from relevant stakeholders.

5 Discussion

In this section, we situate our findings within the existing literature and consider their implications. We first examine how our findings corroborate and extend prior research. Subsequently, we translate these insights into recommendations for relevant stakeholders, with the aim of supporting more inclusive and sustainable LRLT development.

5.1 Connections with Prior Literature

The identified challenges correspond to the aspects of low-resourcedness categorized in Nigatu et al. (2024). The lack of standardization (§ 4.1.1) and the unreliability of existing tools (§ 4.1.2) primarily concern the **Artifacts** aspect, whereas insufficient support (§ 4.1.3) cuts across the **Socio-Political**, **Resource**, and **Agency** aspects. In addition, our findings highlight the complexity of some of the aspects. Prior work identifies limited standardization as both a feature of low-resourcedness (Nigatu

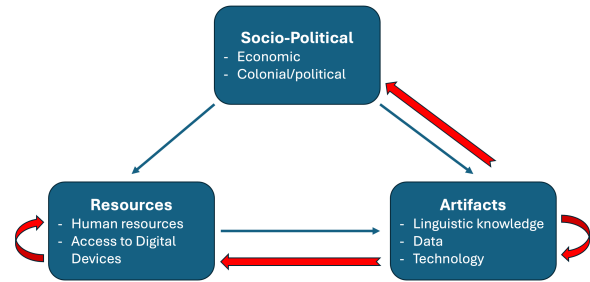


Figure 1: Illustration of how three aspects defined in Nigatu et al. (2024) affect each other. An arrow pointing from aspect A to aspect B suggests that A affects B. Thin blue arrows are identified in Nigatu et al. (2024), and broad red arrows are discovered in our study.

et al., 2024; Coto-Solano, 2022) and a barrier to LT development across communities (Dolinska et al., 2024). Our findings (§ 4.1.1) extend this observation by showing that the existence of a standardized orthography is insufficient: digital resources may remain scarce when competing standards co-exist, speakers do not adopt the standard, or its use is restricted to particular contexts. Similarly, although all practitioners faced the same overarching challenge of securing funding, the particular funding sources that were difficult to access varied among them (§ 4.1.3), leading them to seek support through different channels (§ 4.2.1). The situated nature of these challenges calls for case-specific investigations and recommendations.

Nigatu et al. (2024) identify three causal relationships among the **Socio-Political**, **Resource**, and **Artifacts** dimensions. We set aside **Agency** in this discussion because it functions as a cross-cutting dimension that interacts with each of the other three. Specifically, (1) **Socio-Political** conditions constrain **Resource** availability; (2) **Resource** availability influences the number of **Artifacts** produced; and (3) **Socio-Political** conditions also directly affect the production of **Artifacts**, including documented linguistic knowledge. Our findings suggest additional causal links among these dimensions, forming bidirectional and self-reinforcing feedback loops (Figure 1). First, shortages of funding and human resources can reduce the quality of datasets and technologies. Tools developed under these constraints may subsequently have limited commercial viability, giving funders less incentive to support further development (§ 4.1.3); this represents an influence of **Artifacts** on **Socio-Political** decisions. Second, dataset development often de-

depends on access to shared data, yet data scarcity limits the material available for sharing and reuse (§ 4.1.1), creating a loop in which a lack of **Artifacts** constrains the production of further **Artifacts**. Finally, a larger practitioner community and greater availability of data and tools can help newcomers enter the field, whereas scarcity in both areas raises barriers to entry (§ 4.1.3; § 4.2.1). This illustrates how existing **Artifacts** and **Resources** can influence the future availability of **Resources**, particularly human resources.

Participants’ accounts indicate that a central motivation for LRLT development is its connection to practitioners’ linguistic and cultural identities (§ 4.2). This connection is especially salient because many languages described as low-resource in NLP are also socially or politically marginalized in the societies where they are spoken (Smart et al., 2024). Siminyu et al. (2023) similarly report that native-speaker practitioners are motivated to develop LTs for their own languages. Sociolinguistic research suggests that identity is not merely reflected in language use but is also constructed through linguistic practice and interaction (Bucholtz and Hall, 2005; Eckert, 2012). By expanding opportunities for speakers to use their languages, LRLTs developed with and for speaker communities may therefore support linguistic identity, language maintenance, and cultural continuity.

5.2 Recommendations

Building interconnected communities grounded in speaker participation. We recommend expanding the existing LRLT communities and increasing connections among them. Such communities can be modeled after Masakhane (Orife et al., 2020). Acknowledging the lack of focus on African LTs, the opaqueness of research effort, and the socio-economic factors hindering connections, Masakhane has expanded from 144 to over 3,000 participants in five years (Issaka et al., 2025), publishing and releasing high-quality research artifacts. This example demonstrates how a well-connected community could overcome challenges faced by individual practitioners. LRLT practitioners could also involve LRL speakers in various stages of LT development to understand their needs and disagreements, and evaluate existing LTs to better identify future directions (Matias and Price, 2025).

Enabling collaboration and language use through supportive government policies. Com-

munications and collaboration from different regions can be restricted by local governments. For the collective effort in LRLT development, we encourage governments to ease the restrictions on research collaborations. Local governments should understand the connection between languages and identities, and children’s ability to achieve bilingualism (Byers-Heinlein and Lew-Williams, 2013). Policies could promote the use of marginalized languages by making their study compulsory in primary and secondary education, granting them official status, and encouraging their use in government communications.

Coordinating multi-stakeholder support for sustainable LRLT development. The broader NLP community should continue the effort to welcome LRLT research and practitioners by lowering the cost for conference registration, having dedicated paper submission tracks, workshops, and conferences. NLP practitioners should continue to develop LRL-specific or language-independent tools. The prevalent use of LLMs in LRLT development suggests the need to build more inclusive LLMs. Funding opportunities dedicated to LRLT development can be increased and made available by governments, NLP communities, companies, and individuals. Different stakeholders can also convene to discuss and understand each other’s perspectives, shaping a more sustainable path for LRLT development.

6 Conclusion

In this paper, we used semi-structured interviews with Sino-Tibetan LRLT practitioners to investigate the challenges they face and the ways they seek to mitigate them. Beyond data scarcity, participants identified challenges related to the standardization of written and spoken language, the reliability of basic LRL processing tools, and limited funding, collaboration, and research support. These interrelated challenges not only hinder LRLT development but also perpetuate the low-resource status of these languages. Nevertheless, participants generally expressed optimism about the field’s progress. They described drawing on support from diverse sources and leveraging LLMs for basic language tasks. By making these challenges and potential means of mitigation explicit, this study calls for greater attention and support from relevant stakeholders and advocates for sustainable LRLT development with, for, and by native-speaker communities.

Limitations

Our recruitment process involves emailing practitioners whom we can find online and snowball sampling by connecting with more participants through practitioners we have already connected with. We acknowledge that this recruitment process likely resulted in our participants being those who are already connected with other practitioners and who are open to voicing their experience. We also faced challenges in our recruitment process when participants did not respond to emails, or even refused to participate in interviews. The three languages selected, i.e., Cantonese, Southern Min and Tibetan, are languages with the richest resources among all Sino-Tibetan LRLs. Practitioners working on other Sino-Tibetan LRLs might be facing more severe challenges and fewer ways to mitigate them. We therefore consider our findings to be a subset of all challenges and mitigation faced by Sino-Tibetan LRLT practitioners (and therefore LRLT practitioners in general), and limited by the participants and communities that we engaged with. Future research should try to expand the investigation to practitioners who are less connected with the broader NLP community and other practitioners, those working on other LRLs (especially some extremely low-resource ones), and comparing challenges across different communities.

Acknowledgments

We are indebted to all our interview participants for taking the time to discuss these important issues with us, helping us and the broader NLP community better understand what is needed for LRLT development. We greatly appreciate the help of Dr. Chin-Hui Lee and Dr. Daai-Syut Yeung in establishing connections with some of our early participants. The final version of this paper would not have been possible without the assistance and feedback of members of the Georgia Tech T+ID Lab and the anonymous ARR reviewers.

References

- Steven Bird. 2020. [Decolonising speech and language technology](#). In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3504–3519, Barcelona, Spain (Online). International Committee on Computational Linguistics.
- Mary Bucholtz and Kira Hall. 2005. [Identity and interaction: a sociocultural linguistic approach](#). *Discourse Studies*, 7(4-5):585–614.
- Krista Byers-Heinlein and Casey Lew-Williams. 2013. Bilingualism in the early years: What the science says. *Learn Landsc*, 7(1):95–112.
- Tsz Chung Cheng, Chung Shing Cheng, Chaak-ming Lau, Eugene Lam, Wong Chun Yat, Hoi On Yu, and Cheuk Hei Chong. 2025. [HKCanto-eval: A benchmark for evaluating Cantonese language understanding and cultural comprehension in LLMs](#). In *Proceedings of the 29th Conference on Computational Natural Language Learning*, pages 1–11, Vienna, Austria. Association for Computational Linguistics.
- Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. [XNLI: Evaluating cross-lingual sentence representations](#). In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2475–2485, Brussels, Belgium. Association for Computational Linguistics.
- Andrea Cornwall and Rachel Jewkes. 1995. [What is participatory research?](#) *Social Science & Medicine*, 41(12):1667–1676.
- Rolando Coto-Solano. 2022. [Evaluating word embeddings in extremely under-resourced languages: A case study in Bribri](#). In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 4455–4467, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.
- Sara Court and Micha Elsner. 2024. [Shortcomings of LLMs for low-resource translation: Retrieval and understanding are both the problem](#). In *Proceedings of the Ninth Conference on Machine Translation*, pages 1332–1354, Miami, Florida, USA. Association for Computational Linguistics.
- Barbara Czarniawska. 2014. *Social Science Research: From Field to Desk*, first edition edition. SAGE Publications Ltd, 55 City Road, London.
- Joanna Dolinska, Shekhar Nayak, and Sumittra Suraratcha. 2024. [Akha, dara-ang, karen, khamu, Mlabri and urak lawoi’ language minorities’ subjective perception of their languages and the outlook for development of digital tools](#). In *Proceedings of the Seventh Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pages 94–99, St. Julians, Malta. Association for Computational Linguistics.
- Penelope Eckert. 2012. [Three waves of variation study: The emergence of meaning in the study of sociolinguistic variation](#). *Annual Review Anthropology*, 41:87–100.
- Ilker Etikan, Sulaiman Abubakar Musa, Rukayya Sunusi Alkassim, and 1 others. 2016. Comparison of convenience sampling and purposive sampling. *American journal of theoretical and applied statistics*, 5(1):1–4.

- Marzieh Fadaee, Arianna Bisazza, and Christof Monz. 2017. [Data augmentation for low-resource neural machine translation](#). In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 567–573, Vancouver, Canada. Association for Computational Linguistics.
- Barney G. Glaser and Anselm L. Strauss. 1967. *The discovery of grounded theory: Strategies for qualitative research*. Aldine de Gruyter.
- Emma Harvey, Emily Sheng, Su Lin Blodgett, Alexandra Chouldechova, Jean Garcia-Gathright, Alexandra Olteanu, and Hanna Wallach. 2025. [Understanding and meeting practitioner needs when measuring representational harms caused by LLM-based systems](#). In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 18423–18440, Vienna, Austria. Association for Computational Linguistics.
- Kung Yin Hong, Lifeng Han, Riza Batista-Navarro, and Goran Nenadic. 2024. [CantonMT: Cantonese-English neural machine translation looking into evaluations](#). In *Proceedings of the 16th Conference of the Association for Machine Translation in the Americas (Volume 2: Presentations)*, pages 133–144, Chicago, USA. Association for Machine Translation in the Americas.
- Elwin Huaman, Adrian Gamarra Lafuente, Johanna Cordova, and Anna Korhonen. 2026. [Building community-centred NLP resources for Puno Quechua](#). In *Proceedings of the Sixth Workshop on NLP for Indigenous Languages of the Americas (AmericasNLP)*, pages 147–152, San Diego, California, USA. Association for Computational Linguistics.
- Benjamin Hunt, Lane Schwartz, Sylvia Schreiner, and Emily Chen. 2023. [Community consultation and the development of an online akuzipik-English dictionary](#). In *Proceedings of the Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP)*, pages 134–143, Toronto, Canada. Association for Computational Linguistics.
- Sheriff Issaka, Keyi Wang, Yinka Ajibola, Oluwatumininu Samuel-Ipaye, Zhaoyi Zhang, Nicté Aguilón Jimenez, Evans Kofi Agyei, Abraham Lin, Rohan Ramachandran, Sadick Abdul Mumin, Faith Nchifor, Mohammed Shuraim, Lieqi Liu, Erick Rosas Gonzalez, Sylvester Kpei, Jemimah Osei, Carlene Ajeneza, Persis Boateng, Prisca Adwoa Dufie Yeboah, and Saadia Gabriel. 2025. [The african languages lab: A collaborative approach to advancing low-resource african nlp](#). *Preprint*, arXiv:2510.05644.
- Suramya Jadhav, Abhay Shanbhag, Amogh Thakurdesai, Ridhima Sinare, and Raviraj Joshi. 2025. [On limitations of LLM as annotator for low resource languages](#). In *Proceedings of the 8th International Conference on Natural Language and Speech Processing (ICNLSP-2025)*, pages 277–282, Southern Denmark University, Odense, Denmark. Association for Computational Linguistics.
- Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. [The state and fate of linguistic diversity and inclusion in the NLP world](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online. Association for Computational Linguistics.
- Nataliia Kholodna, Sahib Julka, Mohammad Khodadadi, Muhammed Nurullah Gumus, and Michael Granitzer. 2024. [Llms in the loop: Leveraging large language model annotations for active learning in low-resource languages](#). In *Machine Learning and Knowledge Discovery in Databases. Applied Data Science Track: European Conference, ECML PKDD 2024, Vilnius, Lithuania, September 9–13, 2024, Proceedings, Part X*, page 397–412, Berlin, Heidelberg. Springer-Verlag.
- Eleanor Knott, Aliya Hamid Rao, Kate Summers, and Chana Teeger. 2022. Interviews in the social sciences. *Nat. Rev. Methods Primers*, 2(1).
- Yen-Chun Lai, Yi-Jun Zheng, Wen-Han Hsu, Yan-Ming Lin, Cheng-Hsiu Cho, Chih-Chung Kuo, Chao-Shih Huang, and Yuan-Fu Liao. 2024. [Construction of large language models for taigi and hakka using transfer learning](#). In *2024 27th Conference of the Oriental COCOSA International Committee for the Coordination and Standardisation of Speech Databases and Assessment Techniques (O-COCOSA)*, pages 1–6.
- Longhao Li, Zhao Guo, Hongjie Chen, Yuhang Dai, Ziyu Zhang, Hongfei Xue, Tianlun Zuo, Chengyou Wang, Shuiyuan Wang, Jie Li, Jian Kang, Xin Xu, Hui Bu, Binbin Zhang, Ruibin Yuan, Ziya Zhou, Wei Xue, and Lei Xie. 2025. [Wenetspeech-yue: A large-scale cantonese speech corpus with multi-dimensional annotation](#). *Preprint*, arXiv:2509.03959.
- Evelyn Kai-Yan Liu. 2022. [Low-resource neural machine translation: A case study of Cantonese](#). In *Proceedings of the Ninth Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 28–40, Gyeongju, Republic of Korea. Association for Computational Linguistics.
- Shuheng Liu and Michael Best. 2025. [A survey of NLP progress in Sino-Tibetan low-resource languages](#). In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 7804–7825, Albuquerque, New Mexico. Association for Computational Linguistics.
- J. Nathan Matias and Megan Price. 2025. [How public involvement can improve the science of ai](#). *Proceedings of the National Academy of Sciences*, 122(48):e242111122.
- Marieke Meelen, Élie Roux, and Nathan Hill. 2021. [Optimisation of the largest annotated tibetan corpus combining rule-based, memory-based, and deep-learning](#)

- methods. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, 20(1).
- Robert K. Merton. 1968. [The matthew effect in science](#). *Science*, 159(3810):56–63.
- Wilhelmina Nekoto, Vukosi Marivate, Tshinondiwa Matsila, Timi Fasubaa, Taiwo Fagbohunge, Solomon Oluwole Akinola, Shamsuddeen Muhammad, Salomon Kabongo Kabenamualu, Salomey Osei, Freshia Sackey, Rubungo Andre Niyongabo, Ricky Macharm, Perez Ogayo, Orevaoghene Ahia, Musie Meressa Berhe, Mofetoluwa Adeyemi, Masabata Mokgesi-Seling, Lawrence Okegbemi, Laura Martinus, and 28 others. 2020. [Participatory research for low-resourced machine translation: A case study in African languages](#). In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2144–2160, Online. Association for Computational Linguistics.
- Hellina Hailu Nigatu, Atnafu Lambebo Tonja, Benjamin Rosman, Thamar Solorio, and Monojit Choudhury. 2024. [The Zeno’s paradox of ‘low-resource’ languages](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 17753–17774, Miami, Florida, USA. Association for Computational Linguistics.
- Iroro Orife, Julia Kreutzer, Blessing Sibanda, Daniel Whitenack, Kathleen Siminyu, Laura Martinus, Jamiil Toure Ali, Jade Abbott, Vukosi Marivate, Salomon Kabongo, Musie Meressa, Espoir Murhabazi, Orevaoghene Ahia, Elan van Biljon, Arshath Ramkilowan, Adewale Akinfaderin, Alp Öktem, Wole Akin, Ghollah Kioko, and 6 others. 2020. [Masakhane – machine translation for africa](#). *Preprint*, arXiv:2003.11529.
- Leiyu Pan, Bojian Xiong, Lei Yang, Renren Jin, Shaowei Zhang, Yue Chen, Ling Shi, Jiang Zhou, Junru Wu, Zhen Wang, Jianxiang Peng, Juesi Xiao, Tianyu Dong, Zhuowen Han, Zhuo Chen, Yuqi Ren, and Deyi Xiong. 2025. [Advancing large language models for tibetan with curated data and continual pre-training](#). *Preprint*, arXiv:2507.09205.
- Charlie Parker, Sam Scott, and Alistair Geddes. 2019. Snowball sampling. *SAGE research methods foundations*.
- Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *Proceedings of the 40th International Conference on Machine Learning, ICML’23*. JMLR.org.
- Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21(1).
- Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. [SQuAD: 100,000+ questions for machine comprehension of text](#). In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas. Association for Computational Linguistics.
- Sebastian Ruder, Ivan Vulić, and Anders Søgaard. 2019. [A survey of cross-lingual word embedding models](#). *J. Artif. Int. Res.*, 65(1):569–630.
- Johnny Saldaña. 2009. *The coding manual for qualitative researchers*. Sage Publications Ltd.
- Tim Schopf, Karim Arabi, and Florian Matthes. 2023. [Exploring the landscape of natural language processing research](#). In *Proceedings of the 14th International Conference on Recent Advances in Natural Language Processing*, pages 1034–1045, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.
- Jiu Sha, Mengxiao Zhu, Chong Feng, and Jizhuoma Ci. 2025. [Tibetan-llama 2: Large language model for tibetan](#). *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, 24(12).
- Kathleen Siminyu, Jade Abbott, Kólá Túbòsún, Aremu Anuoluwapo, Blessing K. Sibanda, Kofi Yeboah, David Adelani, Masabata Mokgesi-Seling, Frederick R. Apina, Angela Thandizwe Mthembu, Arshath Ramkilowan, and Babatunde Oladimeji. 2023. [Consultative engagement of stakeholders toward a roadmap for african language technologies](#). *Patterns*, 4(8):100820.
- Andrew Smart, Ben Hutchinson, Lameck Mbangula Amugongo, Suzanne Dikker, Alex Zito, Amber Ebinama, Zara Wudiri, Ding Wang, Erin van Liemt, João Sedoc, Seyi Olojo, Stanley Uwakwe, Edem Wornyo, Sonja Schmer-Galunder, and Jamila Smith-Loud. 2024. [Socially responsible data for large multilingual language models](#). *Preprint*, arXiv:2409.05247.
- Nicolas Tournadre. 2014. [The tibetic languages and their classification](#). In Thomas Owen-Smith and Nathan Hill, editors, *Trans-Himalayan Linguistics*, pages 105–130. De Gruyter Mouton, Berlin, Boston.
- Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurélien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. [Llama: Open and efficient foundation language models](#). *ArXiv*, abs/2302.13971.
- Lisa M. Vaughn and Farrah Jacquez. 2020. [Participatory Research Methods – Choice Points in the Research Process](#). *Journal of Participatory Research Methods*, 1(1).
- Yurong Wang, Min Lin, Qitu Hu, Longjie Bao, Shuangcheng Bai, and Yanling Li. 2025. [Large and small models for collaborative cross-lingual data augmentation in entity relationship extraction for low-resource languages](#). *Journal of King Saud University Computer and Information Sciences*, 37(4):56.

Jiahao Yang, Jianguo Wei, Kuntharrrgyal Khysru, Junhai Xu, Wenhuan Lu, Wenjun Ke, and Xiaokang Yang. 2023. [Effective fine-tuning method for tibetan low-resource dialect speech recognition](#). In *2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pages 960–965.

Xinyan Yu, Trina Chatterjee, Akari Asai, Junjie Hu, and Eunsol Choi. 2022. [Beyond counting datasets: A survey of multilingual dataset construction and necessary resources](#). In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 3725–3743, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Mingjun Zhou, Quzong Gesang, Nuo Qun, Tashi Nyima, and Dongrub Rinchen. 2024. [Tibetan-chinese machine translation enhanced on cross-lingual pre-trained model](#). In *2024 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pages 1618–1623.

A Appendix

A.1 Participant Consent Form

Here is the full text for the consent form with contact information redacted.

You are being asked to be a volunteer in a research study. The purpose of this study is to investigate the assets and challenges of the research communities working on low-resource Sino-Tibetan language technology by interviewing experts in this field. The interview will take approximately 1 hour to complete.

To protect your privacy, your records will be kept under a code number rather than by name. Your name will only be collected when signing the consent forms. The signed consent forms will be kept in a secured locker, and unless you give specific consent otherwise, only study staff will be allowed to look at them. Your name and any other fact that might point to you will not appear when results of this study are presented or published.

The risks involved are no greater than those involved in daily activities. You will not likely directly benefit for joining this study, but your participation will help the NLP research community to understand the direction for future development of low-resource language technology.

We will comply with any applicable laws and regulations regarding confidentiality. To make sure that this research is being carried out in the proper way, the Georgia Institute of Technology IRB may review study records. The Office of Human Research Protections may also look at study records.

If you have any questions about the study, you may contact Dr. Michael Best at telephone [REDACTED], or via email at [REDACTED]. If you have any questions about your rights as a research subject, you may contact Georgia Institute of Technology Office of Research Integrity Assurance at [REDACTED]. Thank you for participating in this study.