Bhaasha, Bhāṣā, Zaban: A Survey for Low-Resourced Languages in South Asia – Current Stage and Challenges

Anonymous ACL submission

Abstract

Rapid developments of pre-trained or large language models have revolutionized many NLP tasks on English datasets in recent years, unfortunately, the model developments and evaluations for low-resource languages are being overlooked, especially for languages in South Asia. While there are over 650 languages in South Asia, many of them either have very limited computational resources or are not supported in existing language models. Thus, a concrete question to be solved by this study is: *Can we assess the current stage and challenges to inform our NLP community and facilitate model developments for South Asian languages?* In this survey, we have comprehensively examined current efforts and challenges of NLP model development for low-resourced South Asian languages by retrieving studies published since 2020 with a focus on transformer-based language models, such as BERT, T5, and GPT. Our study has presented insights and issues from 3 essential aspects, data, model, and tasks, such as available data sources, fine-tuning strategies, and domain applications. Our findings highlight substantial challenges, such as missing data across critical domains (e.g., health), code-mixing, and a lack of standardized evaluation procedures. We hope that our survey efforts can raise community attentions for more targeted data curation, unified benchmarks tailored to the cultural and linguistic nuances of South Asia, and stronger collaborative efforts to ensure an equitable representation for all languages in the South Asia.

1 Introduction

South Asia is one of the most linguistically diverse regions, encompassing Indo-Aryan, Dravidian, Iranian, and Tibeto-Burman languages, along with numerous isolates (Arora et al., 2022; Borin et al., 2014). While a quarter of the world's population resides in this region (Véron et al., 2008), its languages remain severely underrepresented in NLP models (Dunn et al., 2024). Large language models (LLMs) support multiple languages (Lai et al., 2024), but South Asian ones are often missing from training corpora or present in imbalanced quantities (Khan et al., 2024). There are multiple factors behind this disparity, and it's crucial to identify and address them to ensure better representation of South Asian languages. The definition of "low-resource" varies based on data availability and digital presence (Nigatu et al., 2024; Mehta et al., 2020). We consider a language "low-resource" if it lacks computational data and standardized evaluation benchmarks for most NLP tasks. While low-resource languages has been studied for various regions (Aji et al., 2023, 2022; Adebara and Abdul-Mageed, 2022), there lacks comprehensive study on the current status of South Asian NLP, which will be fulfiled by this survey study.

Study	Inclusive Language Coverage	Data Insights	Multiple NLP Tasks	Interdisciplinary Integration	Recent LLMs
Hedderich et al.	✓[1]	✓	✓	✗	✗
Arora et al.	✓	✓	✓	✓	✗[2]
Maddu and Sanapala	✗	✓	✓	✗	✗
Ranathunga et al.	✓[1]	✓	✗	✗	✗[3]
Our Work	✓	✓	✓	✓	✓

Table 1: Comparing related surveys to ours by multiple key criteria. We denote 1-3 as [1] not specific to South-Asian languages; [2] limited discussion of LLMs; and [3] Mentions multilingual models but not for LLMs or low-resourced languages.

Study retrieval methods. We searched for relevant publications using Google Scholar, Semantic Scholar, and ACL Anthology using a combination of broad and specific search keywords such as "Indic," "Low Resource," and "Multilingual." We also include NLP tasks as additional keywords, such as "Machine Translation." To focus on recent developments, we excluded papers before 2020 and focused on recent LLMs, such as BERT, mBART, T5, and GPT. Duplicate studies were removed.

Objectives and Contributions. We aim to assess the current state of NLP research for South Asian languages, covering data, model advancements, and evaluation techniques, while identifying the key issues unique to these languages. Unlike prior surveys, as seen in Table 1, our work makes three unique contributions: 1) this study includes comprehensive language families in the South Asia, which ensures broader coverage beyond Indo-Aryan and Dravidian languages by covering other widely spoken language families in the region; 2) we examine data sources and provide data insights to accelerate low-resourced language research in South Asia; and 3) this work analyzes studies across various domains (e.g., healthcare and education) and summarizes recent LLMs and their tuning strategies (e.g., LoRA (Hu et al., 2022)). We expect this survey will inspire future directions to strengthen community efforts for underrepresented languages in South Asia.

2 Data and Resources

Data corpus is the essential to enable language models understanding complex and heterogeneous semantics and structures of South Asian languages. Indeed, there are over 650 languages are spoken in the region, yet computational resources remain relatively low, uneven, and are highly skewed toward a few languages (Zhao et al., 2025; Hasan et al., 2024; Narayanan and Aepli, 2024; Ali et al., 2024; Baruah et al., 2024). For example, the language resources are represented by small text samples, and major focus is given to languages like Hindi and Urdu (Mishra et al., 2024; Gala et al., 2023). However, the existing studies may merely answer the questions that will be answered in our study: 1) *What are the available data corpora for the low-resourced languages in South Asia?* 2) *What NLP tasks are for the corpora?* 3) *What domains are for the corpora?* To answer those questions, we summarize the data distributions by language families in Figure 1 and statistics in Table 2.

2.1 Language resources

Figure 1 illustrates the uneven distribution of South Asian languages and our collected resources. The color gradient and circle sizes indicate that a few dominant languages, such as Hindi, Bengali, and Telugu, have comparatively more resources, and the others are severely underrepresented, highlighting resource challenges and opportunities. We summarize language resources of the retrieved studies into four families, Indo-Aryan, Dravidian, Tibeto-Burman, and Iranian languages.

Indo-Aryan Languages own the largest language group in South Asia and are relatively more represented in our collected studies. For example, Hindi, Bengali, Marathi, Tamil and Urdu are among the largest bubbles in Figure 1, and Hindi corpora are available for all major NLP tasks in Table 2, which also align with existing langauge speaker populations (Gala et al., 2023). Large-scale datasets are not well-distributed across NLP tasks. For example, IndicMARCO and MultiCONER provide a vast amount of resources for information retrieval and named-entity recognition (NER) (Haq et al., 2024; Malmasi et al., 2022); Kavathekar et al. (2024) introduces a benchmark for AI-generated text detection in Hindi; and BELEBELE covers large-scale multilingual comprehension assessment (Bandarkar et al., 2024). However, Bhojpuri, Sindhi, and Assamese are represented in only a few domain-specific datasets (Baruah et al., 2024; Malmasi et al., 2022; Kumar et al., 2024): their dataset size is comparatively smaller (with less than 5000 samples) (Gala et al., 2023), such as Assamese in a back transliteration data (Baruah et al., 2024), and Angika, Magahi, and Bhojpuri in small-scale POS tagging resources (Kumar et al., 2024). Multiple code-mixed data of sentiment and emoji prediction tasks are available for Hindi, Bengali, Tamil, and Marathi (e.g., MMCQS, SENTIMOJI) (Singh et al., 2024; Ghosh et al., 2024).

Dravidian Languages include Tamil, Malayalam, Telugu, and Kannada in a number of integrated multilingual corpora (Gala et al., 2023; Haq et al., 2024; Urlana et al., 2023; Philip et al., 2021) for NLP tasks, such as diglossia classification, machine translation, and hate speech detection (Prasanna and Arora, 2024; Kumaresan et al., 2024; K et al., 2024). However, many Dravidian languages, including Kodava, Toda, and Irula, are absent from major data resources and benchmarks. A rare exception is Tulu, which is included in a recently developed parallel corpus for machine translation (Narayanan and Aepli, 2024). The language resources are relatively smaller sizes than Indo-Aryan Languages (e.g., Hindi) and cover much fewer application domains, such as healthcare.

Tibeto-Burman and Iranian Languages are critically underrepresented. Manipuri, Mizo are

Data	Language(s)	Size	NLP Task	Year	Source	Domain	Acc
INDIC-MARCO	Multiple (11)	8.8M	Neural IR	2024	Haq et al.	General	Yes
BPCC	Multiple(22)	230M	Machine Translation	2023	Gala et al.	General	Yes
TransMuCoRes	Multiple (31)	1.8M	Coreference Resolution	2024	Mishra et al.	General	Yes
Homophobia Data	Telugu, Kannada, Gujarati	38,904	Homophobia Detection	2024	Kumaresan et al.	Social Media	No
Fake News Detection	Malayalam	1,682	Fake News Detection/ Classification	2024	K et al.	News Media	No
MultiCoNER	Multiple (11)	26M	NER	2022	Malmasi et al.	Wiki&Search	Yes
POS Tagging Dataset	Angika, Magahi, Bhojpuri	2124	POS tagging	2024	Kumar et al.	News,Conversations	Yes
Assamese BackTranslit	Assamese	60K	Back transliteration	2024	Baruah et al.	Social Media	Yes
IruMozhi	Tamil	1,497	Diglossia Classification	2024	Prasanna and Arora	Wikipedia	Yes
Paraphrase Corpus	Pashto	6,727	Paraphrase detection	2024	Ali et al.	News Media	Yes
Hate Speech Data	Bengali, Hindi, Urdu	-	Sentiment Analysis, Hate Detection	2024	Hasan et al.	Social Media	No
AS-CS Dataset	Hindi, Bengali	5,062	Counter Speech Generation	2024	Das et al.	Social Media	Yes
CoPara	4 Dravidian Languages	2856	Paragraph-level alignment	2023	E et al.	News Media	Yes
PMIndiaSum	Multiple (14)	697K	Multilingual Summarization	2023	Urlana et al.	Government	Yes
CVIT-PIB v1.3	Multiple(11)	2.78M	Multilingual NMT	2021	Philip et al.	Government	Yes
CaLMQA	Multiple (23)	1.5K	LFQA	2024	Arora et al.	Culture&Society	Yes
NP Chunking Data	Persian	3,091	Noun Phrase Chunking	2022	Kavehzadeh et al.	News Media	No
Punctuation Dataset	Bengali	1.3M	Punctuation Restoration	2020	Alam et al.	News&Stories	Yes
L3Cube-MahaCorpus	Marathi	289M	Classification & NER	2022	Joshi.	News/Non-news	Yes
Flickr30K (EN-(hi-IN))	Hindi	156,915	Multimodal Machine Translation	2018	Chowdhury et al.	Image Captions	Req
SENTIMOJI	Hindi(code-mixed)	20k	Emoji Prediction	2024	Singh et al.	Social Media	Yes
Suman	Kadodi,Marathi	942	Machine Translation	2024	Dabre et al.	Conversation	Yes
WMT24 En-Hi Data	Hindi	1500	Machine Translation	2024	Bhattacharjee et al.	Mutlidomain	Yes
AGhi	Hindi	36,670	AI-generated text detection	2024	Kavathekar et al.	News	Yes
Mizo News Summariztion Dataset	Mizo	500	News Summarization	2024	(Bala et al., 2024)	News	Yes
ADIhi	Hindi	36,670	Ranking LLMs on AI Detectability	2024	Kavathekar et al.	News	Yes
En-Tcy test dataset	Tulu	1300	Machine Translation	2024	Narayanan and Aepli	Wiki,FLORES	Yes
MMCQS dataset	Hindi(code-mixed)	3,015	Multimodal Ques. Summarization	2024	Ghosh et al.	Healthcare	Yes
IN22 Benchmark	Multiple (22)	2527	Machine Translation	2023	Gala et al.	General	Yes
En-Hi Chat Translation	Hindi	16,249	Chat Translation	2022	Gain et al.	Customer Service	Yes
BELEBELE	Multiple (122 variants)	900	Multilingual Reading Comp.	2024	Bandarkar et al.	Web Articles	Yes
Multilingual DisCo	Multiple(6)	84	Geneder Bias Evaluation	2023	Vashishtha et al.	General	Yes
CounterTuringTest(CT2)	Hindi	26	Benchmarking AGTD techniques	2024	Kavathekar et al.	News	Yes
MMFCM	Hindi(code-mixed)	-	Multimodal Ques. Summarization	2024	Ghosh et al.	Healthcare	Yes
IndicNLG Benchmark	Multiple (11)	8.5M	Various Generative Tasks	2022	Kumar et al.	News, Wiki	Yes

Table 2: Available Datasets and Benchmarks for Low-Resource South Asian Languages Across Tasks and Domains. We denote 'Req' as Available on Request; 'Acc' as Public Accessibility.

Tibeto-Burman languages in our retrieved studies, such as summarization data (Urlana et al., 2023; Bala et al., 2024), however, the other languages including Bodo and Dzonkgkhe (the national language of Bhutan) are not covered. **Iranian Languages** including Pashto, Persian, and Balochi are available in our data collections, such as a paraphrase detection corpus in Pashto (Ali et al., 2024), a noun phrase chunking data in Persian (Kavehzadeh et al., 2022), and a question answering data in Balochi (Arora et al., 2024). While the IndicNLG provide one of the largest benchmark, those Tibeto-Burman and Iranian languages (e.g., Dari and Wakhi) are largely missing (Kumar et al., 2022b) besides the benchmark.

2.2 NLP Tasks

The availability of NLP tasks vary by languages in Table 2. For example, Indo-Aryan languages cover all major NLP tasks, such as machine translation, information extraction, and sentiment analysis, while the other languages only cover very few NLP tasks. This section summarizes major tasks of the South Asian languages from the data aspects in two major direction, 1) *generative* and 2) *discriminative* tasks, and we refer the models and approaches to Section 3.

Generative NLP tasks cover three major tasks, machine translation, text generation, and summarization. Machine translation is the most represented task in Table 1 and includes parallel corpora such as BPCC (Gala et al., 2023) and domain-specific datasets like CVIT-PIB v1.3 and Suman (Philip et al., 2021; Dabre et al., 2024). However, Kashmiri, Sindhi, and Tulu lack sufficient bilingual corpora—relying on back-translation (Haq et al., 2024) and cross-lingual transfer (Narayanan and Aepli, 2024). The studies indicate that the scarcity of consistent annotations and high-quality datasets may hinder translation quality. Text Summarization is mainly in general and news domains for Indo-Aryan languages, such as PMIndiaSum (Urlana et al., 2023), while missing in Dravidian and Tibeto-Burman languages. The MedSumm dataset aids in multimodal summarization for Hindi-English code-mixed clinical queries, specifically for the healthcare domain (Ghosh et al., 2024), while domain-specific summarizations are not available in other languages. Text Generation are available in two major benchmarks, IndicNLG benchmark (Kumar et al., 2022a), dialogue generation (Gain et al., 2022), and QA (question answering) (Arora et al., 2024), including biography generation, news headline generation, sentence summarization, paraphrase generation, and question generation across 11 Indic languages.

Discriminative NLP tasks mainly focus on sequential classifications, such as Named entity

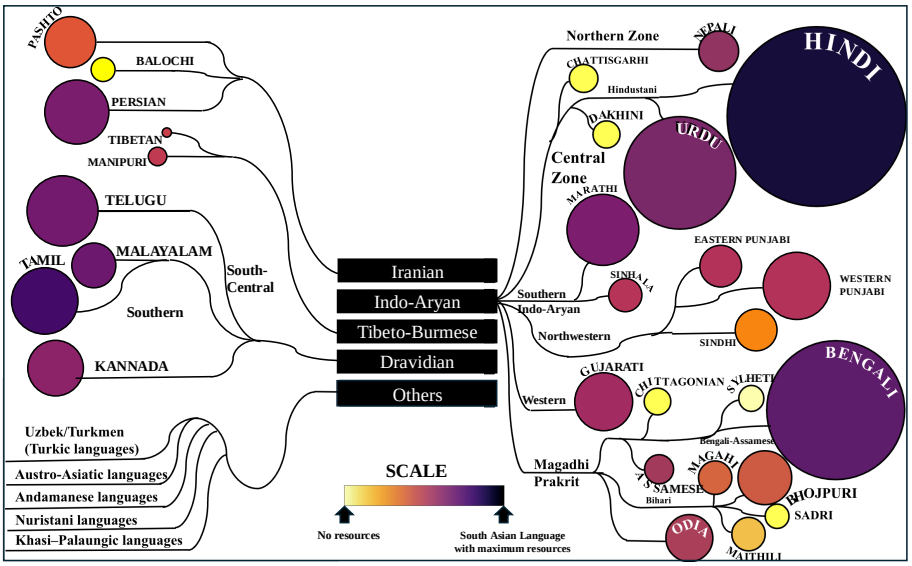

Figure 1: Language families regarding Speaker population and Resource availability. The bubble size means speaker size per language and color scale means retrieved NLP resources. Darker refers to more resources, and vice versa.

recognition (NER) and sentiment predictions. Classification data count major discriminative NLP tasks in our study, such as sentiment analysis and hate speech detection. For example, SENTIMOJI dataset is a sentiment prediction task for Hindi-English code-mixed texts (Singh et al., 2024), and hate speech detection resources are available for Hindi, Tamil, Bengali, (Hasan et al., 2024), Kannada, and Telugu (K et al., 2024). However, sentiment analysis and hate speech detection datasets remain nearly absent for Tibeto-Burman and Iranian languages. MultiCoNER (Malmasi et al., 2022) is an NER benchmark for Hindi, Bengali, Marathi, and Farsi. While POS tagging and syntactic parsing has been well explored in English, the tasks are only available with small sizes for Universal Dependencies (UD) parsing (Baruah et al., 2024) and some low-resource Indo-Aryan languages such as Angika, Magahi, and Bhojpuri (Kumar et al., 2024). Coreference resolution datasets are almost nonexistent, and limited resources are available for Hindi, where most datasets focus on sentence-level tasks rather than document-level coherence modeling (Mishra et al., 2024). Similarly, recently new data releases are primarily for Hindi, such as detectability of AI-generated text by the Counter-Turing test (Kavathekar et al., 2024).

3 Model Advances

We summarize recent model advances of South Asian languages in Table 3—covering three major topics, multilingual language models, training and fine-tuning methods, and model evaluations.

3.1 Multilingual Language Models

Code-Mixed Tokenizer is the fundamental step to encode input text containing more than multiple languages characters and usually starts by fine-tuning on existing language model tokenizers. For example, fine-tuning BERT-base to detect positive hope speech in Kannada-English (Hande et al., 2022) or predict Hindi-English emoji and sentiments (Singh et al., 2024). BUQRNN and PN-BUQRNN (Yu et al., 2024) integrated a pre-trained multilingual BERT for encoding texts, followed by quantum recurrent neural networks (QRNNs). The Overlap BPE method (Patil et al., 2022) improves tokenization consistency across Marathi, Punjabi, and Gujarati on subword-level processing for orthographically similar languages.

Transformer-based models (Vaswani et al., 2017) have dominated recent model developments for both monolingual and multilingual settings. BERT-style architecture is a common model architecture on multi-domain and several monolingual tasks, such as AxomiyaBERTa (Nath et al., 2023) for Assamese, Nepali DistilBERT and Nepali DeBERTa for Nepali (Maskey et al., 2022), and MahaBERT (Joshi, 2022) for Marathi. For multilingual models, IndicTrans2 (Gala et al., 2023) covers translation across 22 South Asian languages and IndicBART (Dabre et al., 2022) supports summarization across 2 language families, making them one of the most comprehensive multilingual mod-

Model	Architecture	Language	Training Strategy	Parameter Size	Year	Source
AxomiyaBERTa	BERT	Assamese	Continuous Pre-train + Supervised Fine-tuning	66M	2023	Nath et al.
IndicBART	BART	Multiple (11)	Continuous Pre-train on IndicCorp + Supervised Fine-tuning	244M	2022	Dabre et al.
BUQRNN	LSTM+BERT	Bengali	Supervised Training	NA	2024	Yu et al.
PN-BUQRNN	LSTM+BERT	Bengali	Supervised Training	NA	2024	Yu et al.
IndicTrans2	Transformer	Multiple (22)	Pre-train + Supervised Fine-tuning	1.1B	2023	Gala et al.
DC-LM	BERT	Kannada	Supervised Fine-tuning	110M	2022	Hande et al.
Lambani NMT	Transformer	Lambani	Pre-train + Supervised Fine-tuning	380M	2022	Chowdhury et al.
Indic-ColBERT	BERT	Multiple (11)	Supervised Fine-tuning	42M	2023	Haq et al.
MedSumm	Multiple LLMs	Hindi (Code-mixed)	Supervised Fine-tuning	7B-13B	2024	Ghosh et al.
Tri-Distil-BERT	BERT	Bengali, Hindi	Continuous Pre-train	8.3B	2024	Raihan et al.
Mixed-Distil-BERT	BERT	Bengali, Hindi	Continuous Pre-train + Supervised Finetuning	8.3B	2024	Raihan et al.
CPT-R	Llama	Multiple (5)	Continuous Pre-train	7B	2024	J et al.
IFT-R	Llama	Multiple (5)	Instruction Fine-tuning	7B	2024	J et al.
BASE	GRU	Hindi	Supervised Training	NA	2023	Lal et al.
MED	Bi-GRU	Hindi	Supervised Training	NA	2023	Lal et al.
RETRAIN^{-}	Bi-GRU	Hindi	English Gigaword Pre-train + Supervised Fine-tuning	NA	2023	Lal et al.
Nepali DistilBERT	BERT	Nepali	Nepali corpora Pre-train by Progressive Mask	66M	2022	Maskey et al.
Nepali DeBERTa	BERT	Nepali	Nepali Corpora Pre-train by Mask-LM	110M	2022	Maskey et al.
TPPoet	Transformer	Persian	Persian poetry Pretrain + Supervised Fine-tuning	33M	2023	Panahandeh et al.
MahaBERT	BERT	Marathi	L3Cube-MahaCorpus Pre-train	110M	2020	Joshi
Emoji Predictor	Transformer	Hindi (Code-mixed)	Supervised Fine-tuning	NA	2024	Singh et al.
RelateLM	BERT	Multiple (5)	Wiki/CFILT Pre-train + Supervised Fine-tuning	110M	2021	Khemchandani et al.
Multi-FAct	Mistral-7B	Bengali	Supervised Fine-tuning	7B	2024	Shafayat et al.
AI-Tutor	Transformer	Pali, Ardhamagadhi	Pre-train + Supervised Training	1.1B	2024	Dalal et al.

Table 3: Model summary by language, architecture, training strategies, and others.

els. Chowdhury et al. trained Transformer models for the machine translation from scratch on related languages (e.g., Kannada, Marathi, and Gujarati) of the target Lambani. Indic-ColBERT (Haq et al., 2024) employs retrieval-augmented supervision for multilingual search to improve document retrieval and ranking across 11 Indic languages. The common strategy mainly take supervised fine-tuning on pre-trained BERT models.

Large language models (LLMs) are being rapidly adopted for South Asian languages in the recent 3 years. MedSumm (Ghosh et al., 2024) fine-tuned 5 LLMs as base models (Llama 2 (Touvron et al., 2023), FLAN T5 (Chung et al., 2022), Mistral (Jiang et al., 2023), Vicuna (Zheng et al., 2023), and Zephyr (Tunstall et al., 2023)) on medical question summarization with visual cues for code-mixed Hindi-English patient queries. Multi-FAct (Shafayat et al., 2024) uses Mistral-7B (Jiang et al., 2023) to extract facts from LLM-generated texts. CPT-R and IFT-R (J et al., 2024a) fine-tune 7B-parameter LLaMA-2 models on romanized Indic corpora to improve transliteration-aware and mixed-script text processing. Additionally, AI-Tutor (Dalal et al., 2024) applied IndicTrans2 (Gala et al., 2023) to Pali and Ardhamagadhi. Due to LLM sizes, fine-tuning partial parameters are more practical, which are detailed in the later section of distillation and parameter-efficient fine-tuning.

3.2 Training and Fine-tuning Methods

Code-mixed and script-specific adaptations enable model understanding on text inputs with mixed languages. For example, LLMs struggled with Bengali script generation due to inefficient tokenization (Mahfuz et al., 2025). (Kumar et al., 2022a) introduced the largest Indic language paraphrasing dataset across 11 languages, IndicParaphrase, highlight that script-specific adaptations enhance paraphrase generation and improve multilingual understanding in code-mixed settings. Huzaifah et al. (2024) found the large model and diverse data may help code-switching translation by showing that GPT-4 and GPT-3.5 outperform traditional NMT models. Transliterating Indic languages into a common script could effectively improves cross-lingual transfer for NER and sentiment analysis Moosa et al. (2023). Domain-specific applications benefit from code-mixed adaptation. For example, MedSumm (Ghosh et al., 2024) integrated Hindi-English code-mixed processing for medical NLP. The Overlap BPE Algorithm (Patil et al., 2022) facilitates shared subword representations across Punjabi, Gujarati, and Marathi, which enhances consistency for orthographically similar languages. (Kirov et al., 2024) aligned transliteration patterns with phonetic structures in Tamil and Malayalam, which further improves multilingual representation. Continual pre-training strategies (Arif et al., 2024; Rajpoot et al., 2024) improve adaptation without degrading prior performance, and continual learning in NMT (Koehn, 2024) prevents catastrophic forgetting by iteratively fine-tuning with new language translation (Ramesh et al., 2023).

Supervised multilingual transfer learning. Given the linguistic similarities among South Asian languages, cross-lingual transfer learning has be-

come a key adaptation strategy. IndicBART (NMT-Adapt) (Narayanan and Aepli, 2024), IndicBART Dabre et al., and IndicTrans2 (Gala et al., 2023) show that pre-training on large multilingual corpora of related languages (that can be mapped to a single script) significantly improves translation. Llama 2-based models (CPT-R, IFT-R) (J et al., 2024a) were fine-tune on task-specific corpora, however, effectiveness varies based on linguistic proximity, with underrepresented languages facing performance declines (Hasan et al., 2024). Sabane et al. found that jointly trained NER models on Hindi-Marathi corpora outperformed monolingual ones due to shared script and grammar. Multilingual NER with Shared Neural Layers (Murthy et al., 2018) benefits from shared representations across Bengali, Tamil, and Malayalam. Back-translation and synthetic data augmentation improved Hindi, Sinhala, and Tamil NMT (Chowdhury et al., 2018).

Several studies explored finetuning approaches. Adaptive multilingual finetuning (Das et al., 2023) leverages subword embedding alignment to enhance transferability across related languages. Cultural adaptation (Zhou et al., 2023) integrated sociolinguistic factors into offensive language detection. In-context learning (ICL) and cross-lingual ICL (Cahyawijaya et al., 2024) improve generalization through query alignment, and Poudel et al. (2024) showed that domain-specific fine-tuning enhances legal English-Nepali translation.

Distillation and parameter-efficient finetuning (PEFT) reduce computational costs (Philip et al., 2021) of LLMs for the low-resource languages. For example, parameter-efficient finetuning (PEFT) methods includes the widely adopted method of low-rank adaptation (LoRA) (Hu et al., 2022). Khade et al. utilizes LoRA PEFT tuning on multilingual Gemma models for Marathi, and find that while target language generation improves, reasoning capabilities decline post finetuning. Petrov et al. introduces a multi-step PEFT approach that outperforms standard finetuning methods for languages with limited labeled data. Gurgurov et al. integrate linguistic knowledge from ConceptNet to improve sentiment analysis and NER performance.

Feature-based finetuning focus on representation learning. Bhatt et al. refine internal representations rather than updating all model weights, enabling incremental knowledge transfer from high-resource to low-resource languages. Progressive finetuning (Perera et al., 2022) incrementally adapts smaller models before transferring knowledge to larger architectures. Yadav et al. explored rank-adaptive LoRA finetuning that balanced efficiency and model performance. BenLLM-Eval (Kabir et al., 2024) finds that zero-shot LLMs like LLaMA-2-13b-chat perform significantly worse than fine-tuned models on Bengali NLP tasks. Iyer et al. (2024) demonstrate that LoRA-based finetuning improves translation tasks while maintaining efficiency, and Nag et al. show that adapter-based finetuning enhances model performance while preventing catastrophic forgetting. MedSumm (Ghosh et al., 2024) utilized QLoRA (Dettmers et al., 2023) to reduce memory overhead while enabling effective finetuning on the MMCQS dataset.

Other methods including augmentation, Zero-shot, and few-shot learning are also examined. IndiText Boost Litake et al. (2024) integrates data augmentation techniques, including Easy Data Augmentation (EDA), back-translation, and text expansion to improve text classification of 6 languages, such as Sindhi, Marathi, and Telugu. Nag et al. (2024) find that few-shot learning benefits morphologically rich languages but struggles with syntactic complexity, requiring further finetuning. Instruction tuning (Pal et al., 2024) improves multilingual translation with a few examples for LLMs.

3.3 Model Evaluations

Evaluation of multilingual models for South Asian languages varies by task, relying on metrics such as BLEU, and more recently, chrF++, COMET, and human evaluation (Gala et al., 2023; Narayanan and Aepli, 2024). Table 2 and 3 include various NLP data and tasks such as FLORES for machine translation (Goyal et al., 2022; Gala et al., 2023) with specific evaluation metrics. For NER (Venkatesh et al., 2022; Khemchandani et al., 2021) and sentiment analysis (Hande et al., 2022; Singh et al., 2024) benchmarks, accuracy, F1-score, precision, and recall are commonly evaluation metrics. For example, CPT-R and IFT-R (J et al., 2024b) with 7B parameters achieve strong zero-shot accuracy (72.5% on sentiment analysis, 78.2% F1 on NER) due to extensive pretraining. MRR (Mean Reciprocal Rank) and NDCG (Normalized Discounted Cumulative Gain) are common evaluation approaches for retrieval and ranking tasks (Haq et al., 2024). BLEU, ROUGE, METEOR, and human evaluations are popular metrics for text generation tasks, such as summarization, machine translation, and question answering (Lal et al., 2023). For example,

Challenge	Example
POS Tagging Inconsistency	"খেলা" should be tagged as NOUN in "খেলা দেখছি" (I am watching a game) and VERB in "খেলা করছি" (I am playing)
Lexical Variability	Bengali (India): "আজকে" (today); Bengali (Bangladesh): "আজগে" (today)
Diglossia	"Where are you going?" in Literary Tamil: "எங்கு செல்கிறீர்கள்"; Spoken Tamil: "எங்க போறீங்க"
Romanization	Hindi: "I am fine" can be romanized as "main theek hoon" or "mai thik hu"
Morphological Segmentation	"நடந்திருக்கிறது" (nadanthirukirathu, "has happened") can be broken into ["நட" (nada, "walk") + "ந்து" (nthu, past suffix) + "இருக்கிறது" (irukirathu, auxiliary verb)
Code mixing	Hinglish: "Mujhe ek idea aaya" (I have an idea)

Table 4: Linguistic Challenges in Low-Resource South Asian Languages for NLP

IndicTrans2 (1.1B params; Gala et al. (2023)) reported BLEU scores of 35.2 (Hindi-English) and 28.7 (Tamil-English). Recent new metrics such as COMET (Rei et al., 2020) and chrF++ (Popović, 2017) complement existing ones (e.g., BLEU) in several recent studies (Gala et al., 2023; Costa-jussà et al., 2024; Gajakos et al., 2024).

4 Trends and Challenges

Data Scarcity and Quality Issues for low-resource languages like Manipuri, Santali, and Sindhi affects model generalizability and applicability (Gala et al., 2023). Sourcing online data of the languages is constrained by limited digital content and copyright restrictions (Gain et al., 2022; Ali et al., 2024). Existing resources, especially small data, are often domain-specific (e.g., government or political) may potential introduce cultural or political biases in downstream applications (Urlana et al., 2023; Kumar et al., 2024). lacking gold-annotated resources complicates tasks, such as co-reference resolution (Mishra et al., 2024), and the rapidly evolving online discourse Bandarkar et al. (2024), hurts model long-term sustainability, such as hate speech (Kumaresan et al., 2024).

Transliteration and representation of South Asian languages are not standardized, causing biases as annotators often rely on phonetic judgment (Baruah et al., 2024). Bhattacharjee et al. (2024) noted inconsistencies in language identification and translation quality due to style and dialect differences within translations and translated text is most often not subject to human verification (Hasan et al., 2024). Also, datasets translated from English to a South Asian language can be culturally misaligned (Das et al., 2024). For culturally nuanced languages Arora et al. (2024), the requirement for proficient annotators restricts the scalability of data collection efforts. Biases due to labeler interpretation hinder sensitive tasks like hate speech detection (Kumaresan et al., 2024).

Further, certain datasets exhibit class imbalances, which lead to bias toward majority classes; solutions such as cost-sensitive learning and oversampling have been proposed (K et al., 2024), but they have not been examined in the low-resourced languages. Languages exhibiting strong diglossia (the spoken and written forms differ substantially), such as Tamil, need additional efforts as literary text cannot be used for tasks in all settings (Prasanna and Arora, 2024). Limited computing resources further restrict improvements in the curation of high-quality datasets (Philip et al., 2021).

Transliteration and Tokenization Inconsistencies reduce generalizability of multilingual models on code-mixed languages, such as Hinglish, Tanglish, and Romanized Bengali (Narayanan and Aepli, 2024; Maddu and Sanapala, 2024). These models often learn script-dependent embeddings and fail to transfer across different writing systems (Koehn, 2024). For example, transliteration ambiguity easily affects NER and speech-text alignment in ASR models (Ramesh et al., 2023).

Existing tokenization strategies such as Byte-Pair Encoding (BPE) and WordPiece frequently fragment morphologically rich words in Dravidian and Indo-Aryan languages, leading to over-segmentation and loss of meaning (Wang et al., 2024). In Devanagari-based languages like Hindi and Marathi, vowel markers (matras) are often split incorrectly (Doddapaneni et al., 2023). Similarly, agglutinative languages like Tamil and Manipuri form complex word structures that are inconsistently tokenized, affecting syntactic parsing and NMT (Narayanan and Aepli, 2024). For extremely low-resource languages, pre-trained tokenizers (Kumar et al., 2024) fail to adapt effectively as they fragment words into multiple sub-word tokens, sometimes even individual characters, introducing noise to tasks like POS tagging.

Morphological segmentation is particularly chal-

lenging for Dravidian languages since words are formed by adding multiple suffixes, as seen in Tulu (Narayanan and Aepli, 2024). Hindi, Assamese and Bengali exhibit different and complex inflectional systems complicating parsing (Chowdhury et al., 2018; Nath et al., 2023). Most Indo-Aryan languages rely heavily on dependent vowel signs (matras) and nasalization markers, where BERT tokenizers often fail to preserve and cause ambiguities (Maskey et al., 2022). For instance, the word "फूल" can be incorrectly tokenized as "फल". Assamese also possess unique sound patterns and the use of alveolar stops, which adds to the complexity of tokenization (Nath et al., 2023). Besides structural differences, administrative vocabulary in Hindi and Tamil includes Persian-origin words like "farman" (order), alongside English-origin terms (Pramodya, 2023). Additionally, Lambani, spoken by a nomadic tribe, shares linguistic features with multiple Southern Indian languages, an evident of tribal migration (Chowdhury et al., 2022).

Code mixing, Diglossia, and Ambiguity. Code-mixing, a highly domain-dependent issue, integrates English or other language words and phrases and results in hybrid forms, such as Hinglish and Tanglish (Das et al., 2024). Diglossia show substantial differences in speaking and writing. For example, Literary Tamil has retained its formal vocabulary, but spoken Tamil incorporates loanwords and phonetic simplifications (Prasanna and Arora, 2024). Additionally, polysemy and contextual ambiguities can fail many models on tasks like NER (Bhatt et al., 2022). For example, Indic languages do not capitalize proper nouns, making it difficult to distinguish named entities from common words (Philip et al., 2021). Further, entity disambiguation is context-dependent; for instance, the word "Hindustan" (हिन्दुस्थान) can refer to a location, a person, or an organization depending on its usage (Mishra et al., 2024). Many languages are grammatically gendered, and even inanimate objects are referred to with gendered nouns (Ramesh et al., 2023).

Standard evaluation benchmarks exist, but gaps have remained in evaluating multilingual models for South Asian languages. Fine-tuned multilingual models often overfit high-resource languages (e.g., Hindi) and degrade lower-resource languages due to shared embeddings (Pal et al., 2024). Catastrophic forgetting happens when adapting models to new languages or tasks, such as in LoRA and adapter-based finetuning (Nag et al., 2024). Phonetic variation across dialects within the same language family (e.g., Bengali and Assamese) results in inconsistencies in phoneme-based word embeddings (Arif et al., 2024).

Model evaluation from our collected studies in Table 3 rely on English-origin benchmarks, which can misinterpret model performance (Haq et al., 2024). Mishra et al. (2024) demonstrate how biases in back-translated datasets cause skewed results and compromise model evaluation across languages. For nuance tasks (e.g., paragraph-level translation), sentence-level evaluation methods may not be insufficient (E et al., 2023; Hasan et al., 2024). Without culturally relevant and task-specific benchmarks, evaluations fail to capture model performance accurately, especially for languages with unique structural and cultural variations (Vashishtha et al., 2023). FLORES-200 is comprehensive but lacks sufficient representation for code-mixed and dialectal variations (Gala et al., 2023). Tibeto-Burman and Austroasiatic evaluation data are almost non-existent and most studies for very low-resourced languages use manually curated datasets (Dalal et al., 2024; Chowdhury et al., 2022).

5 Conclusion

In this paper, we provide a comprehensive analysis of the current NLP research on low-resource South Asian languages. Our survey highlights recent advancements and examines persisting challenges at every stage of resource development.

One of our key takeaways is the highly uneven representation of South Asian languages in both multilingual corpora and model pre-training. While some languages have received relatively more attention, challenges remain in collecting and processing data, and adapting models to specific orthographies. Additionally, existing evaluation metrics fall short due to a lack of script- and task-specific benchmarks, as well as unaddressed demographic and sociocultural biases.

Unlike prior work, we have examined LLMs and tuning strategies, and we emphasize community-driven data collection for various tasks and domains. We call for South Asian-specific evaluation frameworks and script-specific model adaptation techniques to improve resources. We hope this survey serves as a valuable resource for researchers, and encourages further research and broader participation in advancing NLP for these languages.

Limitations

Research and development of resources for South Asian languages have been steadily advancing. Significant progress has been made in multilingual datasets and modeling, and many advancements in high-resource languages are now being adapted for low-resource South Asian languages. Since we aimed for a thorough and balanced analysis, below are some key limitations and certain measures we took to address them.

- Enumerating all studies on low-resource South Asian languages is challenging, as research is dispersed across multiple venues. Many studies are not indexed in widely used repositories like the ACL Anthology. Addressing this, we have conducted an extensive search across various sources, including Google Scholar and Semantic Scholar, and have cross-referenced key papers to ensure proper coverage.

- Identifying relevant studies is complicated due to inconsistent terminology. Papers often use non-standard or domain-specific keywords to describe work on low-resource languages. For instance, some studies refer to 'low-resource languages,' while others use 'under-resourced languages,' 'resource-scarce languages,' or 'marginalized languages.' To account for this, we have tested multiple keyword variations and have manually reviewed the related work sections of key papers to identify additional references.

- Some studies on considerably low resourced languages remain inaccessible due to their publication in regional or less widely indexed venues. We have, to our best efforts, included such publications by searching sources outside of major repositories, especially for Tibeto-Burman and Iranian languages.

References

Ife Adebara and Muhammad Abdul-Mageed. 2022. Towards afrocentric NLP for African languages: Where we are and where we can go. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3814–3841, Dublin, Ireland. Association for Computational Linguistics.

Alham Fikri Aji, Jessica Zosa Forde, Alyssa Marie Loo, Lintang Sutawika, Skyler Wang, Genta Indra Winata, Zheng-Xin Yong, Ruochen Zhang, A. Seza Doğruöz, Yin Lin Tan, and Jan Christian Blaise Cruz. 2023. Current status of NLP in south East Asia with insights from multilingualism and language diversity. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics: Tutorial Abstract*, pages 8–13, Nusa Dua, Bali. Association for Computational Linguistics.

Alham Fikri Aji, Genta Indra Winata, Fajri Koto, Samuel Cahyawijaya, Ade Romadhony, Rahmad Mahendra, Kemal Kurniawan, David Moeljadi, Radityo Eko Prasojo, Timothy Baldwin, Jey Han Lau, and Sebastian Ruder. 2022. One country, 700+ languages: NLP challenges for underrepresented languages and dialects in Indonesia. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7226–7249, Dublin, Ireland. Association for Computational Linguistics.

Tanvirul Alam, Akib Mohammed Khan, and Firoj Alam. 2020. Punctuation restoration using transformer models for high-and low-resource languages. In *W-NUT@EMNLP*.

Iqra Ali, Hidetaka Kamigaito, and Taro Watanabe. 2024. Monolingual paraphrase detection corpus for low resource pashto language at sentence level. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 11574–11581. ELRA and ICCL.

Samee Arif, Abdul Hameed Azeemi, Agha Ali Raza, and Awais Athar. 2024. Generalists vs. specialists: Evaluating large language models for urdu. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 7263–7280. Association for Computational Linguistics.

Aryaman Arora, Adam Farris, Samopriya Basu, and Suresh Kolichala. 2022. Computational historical linguistics and language diversity in South Asia. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1396–1409, Dublin, Ireland. Association for Computational Linguistics.

Shane Arora, Marzena Karpinska, Hung-Ting Chen, Ipsita Bhattacharjee, Mohit Iyyer, and Eunsol Choi. 2024. Calmqa: Exploring culturally specific long-form question answering across 23 languages. *arXiv preprint arXiv:2406.17761*.

Abhinaba Bala, Ashok Urlana, Rahul Mishra, and Parameswari Krishnamurthy. 2024. Exploring news summarization and enrichment in a highly resource-scarce indian language: A case study of mizo. In *Proceedings of the 7th Workshop on Indian Language Data: Resources and Evaluation*, pages 40–46. ELRA and ICCL.

Lucas Bandarkar, Davis Liang, Benjamin Muller, Mikel Artetxe, Satya Narayan Shukla, Donald Husa, Naman Goyal, Abhinandan Krishnan, Luke Zettlemoyer, and Madian Khabsa. 2024. The belebele benchmark: a parallel reading comprehension dataset in 122 language variants. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 749–775. Association for Computational Linguistics.

Hemanta Baruah, Sanasam Ranbir Singh, and Priyankoo Sarmah. 2024. Assamesebacktranslit: Back transliteration of romanized assamese social media text. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 1627–1637. ELRA and ICCL.

Tej K Bhatia and William C Ritchie. 2006. bilingualism in south asia. *The handbook of bilingualism*, pages 780–807.

Shaily Bhatt, Sunipa Dev, Partha Talukdar, Shachi Dave, and Vinodkumar Prabhakaran. 2022. Recontextualizing fairness in nlp: The case of india. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 727–740. Association for Computational Linguistics.

Soham Bhattacharjee, Baban Gain, and Asif Ekbal. 2024. Domain dynamics: Evaluating large language models in english-hindi translation. In *Proceedings of the Ninth Conference on Machine Translation*, pages 341–354. Association for Computational Linguistics.

Lars Borin, Anju Saxena, Taraka Rama, and Bernard Comrie. 2014. Linguistic landscaping of South Asia using digital language resources: Genetic vs. areal linguistics. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 3137–3144, Reykjavik, Iceland. European Language Resources Association (ELRA).

Samuel Cahyawijaya, Holy Lovenia, and Pascale Fung. 2024. Llms are few-shot in-context low-resource language learners. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 405–433. Association for Computational Linguistics.

Amartya Chowdhury, Deepak K. T., Samudra Vijaya K, and S R Mahadeva Prasanna. 2022. Machine translation for a very low-resource language - layer freezing approach on transfer learning. In *Proceedings of the Fifth Workshop on Technologies for Machine Translation of Low-Resource Languages (LoResMT 2022)*, pages 48–55. Association for Computational Linguistics.

Koel Dutta Chowdhury, Mohammed Hasanuzzaman, and Qun Liu. 2018. Multimodal neural machine translation for low-resource language pairs using synthetic data. In *Proceedings of the Workshop on Deep Learning Approaches for Low-Resource NLP*, pages 33–42. Association for Computational Linguistics.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022. Scaling instruction-finetuned language models. *Preprint*, arXiv:2210.11416.

Marta Ruiz Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Ken-591 neth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loïc Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, C. Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alex Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2024. Scaling neural machine translation to 200 languages. *Nature*, 630:841 – 846.

Raj Dabre, Mary Dabre, and Teresa Pereira. 2024. Machine translation of marathi dialects: A case study of kadodi. In *Proceedings of the Eleventh Workshop on Asian Translation (WAT 2024)*, pages 36–44.

Raj Dabre, Himani Shrotriya, Anoop Kunchukuttan, Ratish Puduppully, Mitesh Khapra, and Pratyush Kumar. 2022. Indicbart: A pre-trained model for indic natural language generation. In *Findings of the Association for Computational Linguistics: ACL 2022*. Association for Computational Linguistics.

Siddhartha Dalal, Rahul Aditya, Vethavikashini Chithrra Raghuram, and Prahlad Koratamaddi. 2024. Ai-tutor: Interactive learning of ancient knowledge from low-resource languages. In *Proceedings of the Eleventh Workshop on Asian Translation (WAT 2024)*, pages 56–66.

Mithun Das, Saurabh Pandey, Shivansh Sethi, Punyajoy Saha, and Animesh Mukherjee. 2024. Low-resource counterspeech generation for indic languages: The case of bengali and hindi. In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 1601–1614. Association for Computational Linguistics.

Richeek Das, Sahasra Ranjan, Shreya Pathak, and Preethi Jyothi. 2023. Improving pretraining techniques for code-switched nlp. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages

1176–1191. Association for Computational Linguistics.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: Efficient finetuning of quantized llms. *Preprint*, arXiv:2305.14314.

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.

Jonathan Dunn, Benjamin Adams, and Harish Tayyar Madabushi. 2024. Pre-trained language models represent some geographic populations better than others. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 12966–12976, Torino, Italia. ELRA and ICCL.

Nikhil E, Mukund Choudhary, and Radhika Mamidi. 2023. Copara: The first dravidian paragraph-level n-way aligned corpus. In *Proceedings of the Third Workshop on Speech and Language Technologies for Dravidian Languages*, pages 88–96. INCOMA Ltd., Shoumen, Bulgaria.

Baban Gain, Ramakrishna Appicharla, Soumya Chennabasavaraj, Nikesh Garera, Asif Ekbal, and Muthusamy Chelliah. 2022. Low resource chat translation: A benchmark for hindi–english language pair. In *Proceedings of the 15th biennial conference of the Association for Machine Translation in the Americas (Volume 1: Research Track)*, pages 83–96. Association for Machine Translation in the Americas.

Neha Gajakos, Prashanth Nayak, Rejwanul Haque, and Andy Way. 2024. The SETU-ADAPT submissions to the WMT24 low-resource Indic language translation task. In *Proceedings of the Ninth Conference on Machine Translation*, pages 762–769, Miami, Florida, USA. Association for Computational Linguistics.

Jay Gala, Pranjal A Chitale, A K Raghavan, Varun Gumma, Sumanth Doddapaneni, Aswanth Kumar M, Janki Atul Nawale, Anupama Sujatha, Ratish Puduppully, Vivek Raghavan, Pratyush Kumar, Mitesh M Khapra, Raj Dabre, and Anoop Kunchukuttan. 2023. Indictrans2: Towards high-quality and accessible machine translation models for all 22 scheduled indian languages. *Transactions on Machine Learning Research*.

Akash Ghosh, Arkadeep Acharya, Prince Jha, Sriparna Saha, Aniket Gaudgaul, Rajdeep Majumdar, Aman Chadha, Raghav Jain, Setu Sinha, and Shivani Agarwal. 2024. Medsumm: A multimodal approach to summarizing code-mixed hindi-english clinical queries. In *European Conference on Information Retrieval*, pages 106–120.

Akash Ghosh, Debayan Datta, Sriparna Saha, and Chirag Agarwal. 2025. The multilingual mind : A survey of multilingual reasoning in language models. *Preprint*, arXiv:2502.09457.

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.

Daniil Gurgurov, Mareike Hartmann, and Simon Ostermann. 2024. Adapting multilingual llms to low-resource languages with knowledge graphs via adapters. In *The First Workshop on Knowledge Graphs and Large Language Models*, page 63.

Adeep Hande, Siddhanth U Hegde, Sangeetha S, Ruba Priyadharshini, and Bharathi Raja Chakravarthi. 2022. The best of both worlds: Dual channel language modeling for hope speech detection in low-resourced kannada. In *Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion*, pages 127–135. Association for Computational Linguistics.

Saiful Haq, Ashutosh Sharma, Omar Khattab, Niyati Chhaya, and Pushpak Bhattacharyya. 2024. IndicIR-Suite: Multilingual dataset and neural information models for Indian languages. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 501–509, Bangkok, Thailand. Association for Computational Linguistics.

Md Arid Hasan, Prerona Tarannum, Krishno Dey, Imran Razzak, and Usman Naseem. 2024. Do large language models speak all languages equally? a comparative study in low-resource settings. *arXiv preprint arXiv:2408.02237*.

Michael A. Hedderich, Lukas Lange, Heike Adel, Jannik Strötgen, and Dietrich Klakow. 2021. A survey on recent approaches for natural language processing in low-resource scenarios. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2545–2568, Online. Association for Computational Linguistics.

Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Muhammad Huzaifah, Weihua Zheng, Nattapol Chanpaisit, and Kui Wu. 2024. Evaluating code-switching translation with large language models. In *International Conference on Language Resources and Evaluation*.

Vivek Iyer, Bhavitvya Malik, Pavel Stepachev, Pinzhen Chen, Barry Haddow, and Alexandra Birch. 2024.

Quality or quantity? on data scale and diversity in adapting large language models for low-resource translation. In *Proceedings of the Ninth Conference on Machine Translation*, pages 1393–1409. Association for Computational Linguistics.

Jaavid J, Raj Dabre, Aswanth M, Jay Gala, Thanmay Jayakumar, Ratish Puduppully, and Anoop Kunchukuttan. 2024a. RomanSetu: Efficiently unlocking multilingual capabilities of large language models via Romanization. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15593–15615, Bangkok, Thailand. Association for Computational Linguistics.

Jaavid J, Raj Dabre, Aswanth M, Jay Gala, Thanmay Jayakumar, Ratish Puduppully, and Anoop Kunchukuttan. 2024b. Romansetu: Efficiently unlocking multilingual capabilities of large language models via romanization. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15593–15615. Association for Computational Linguistics.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *Preprint*, arXiv:2310.06825.

Raviraj Joshi. 2022. L3cube-mahacorpus and mahabert: Marathi monolingual corpus, marathi bert language models, and resources. In *Proceedings of the WILDRE-6 Workshop within the 13th Language Resources and Evaluation Conference*, pages 97–101. European Language Resources Association.

Devika K, Hariprasath .s.b, Haripriya B, Vigneshwar E, Premjith B, and Bharathi Raja Chakravarthi. 2024. From dataset to detection: A comprehensive approach to combating malayalam fake news. In *DRAVIDIANLANGTECH*.

Mohsinul Kabir, Mohammed Saidul Islam, Md Tahmid Rahman Laskar, Mir Tafseer Nayeem, M Saiful Bari, and Enamul Hoque. 2024. Benllm-eval: A comprehensive evaluation into the potentials and pitfalls of large language models on bengali nlp. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 2238–2252. ELRA and ICCL.

Ishan Kavathekar, Anku Rani, Ashmit Chamoli, Ponnurangam Kumaraguru, Amit P Sheth, and Amitava Das. 2024. Counter turing test (ct$\hat{2}$): Investigating ai-generated text detection for hindi - ranking llms based on hindi ai detectability index. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 4902–4926. Association for Computational Linguistics.

Parsa Kavehzadeh, Mohammad Mahdi, Abdollah Pour, and Saeedeh Momtazi. 2022. A transformer-based approach for persian text chunking. *Technology Journal of Artificial Intelligence and Data Mining*, 10:373–383.

Omkar Khade, Shruti Jagdale, Abhishek Phaltankar, Gauri Takalikar, and Raviraj Joshi. 2025. Challenges in adapting multilingual LLMs to low-resource languages using LoRA PEFT tuning. In *Proceedings of the First Workshop on Challenges in Processing South Asian Languages (CHiPSAL 2025)*, pages 217–222, Abu Dhabi, UAE. International Committee on 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.

Yash Khemchandani, Sarvesh Mehtani, Vaidehi Patil, Abhijeet Awasthi, Partha Talukdar, and Sunita Sarawagi. 2021. Exploiting language relatedness for low web-resource language model adaptation: An indic languages study. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1312–1323. Association for Computational Linguistics.

Christo Kirov, Cibu Johny, Anna Katanova, Alexander Gutkin, and Brian Roark. 2024. Context-aware transliteration of romanized south asian languages. *Computational Linguistics*, 50:475–534.

Philipp Koehn. 2024. Neural methods for aligning large-scale parallel corpora from the web for south and east asian languages. In *Proceedings of the Ninth Conference on Machine Translation*, pages 1454–1466. Association for Computational Linguistics.

Aman Kumar, Himani Shrotriya, Prachi Sahu, Amogh Mishra, Raj Dabre, Ratish Puduppully, Anoop Kunchukuttan, Mitesh M. Khapra, and Pratyush Kumar. 2022a. IndicNLG benchmark: Multilingual datasets for diverse NLG tasks in Indic languages. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5363–5394, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

C S Ayush Kumar, Advaith Maharana, Srinath Murali, Premjith B, and Soman Kp. 2022b. Bert-based sequence labelling approach for dependency parsing in tamil. In *Proceedings of the Second Workshop on Speech and Language Technologies for Dravidian Languages*, pages 1–8. Association for Computational Linguistics.

Sanjeev Kumar, Preethi Jyothi, and Pushpak Bhattacharyya. 2024. Part-of-speech tagging for extremely low-resource indian languages. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 14422–14431. Association for Computational Linguistics.

Prasanna Kumar Kumaresan, Rahul Ponnusamy, Dhruv Sharma, Paul Buitelaar, and Bharathi Raja Chakravarthi. 2024. Dataset for identification of homophobia and transphobia for telugu, kannada, and gujarati. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 4404–4411. ELRA and ICCL.

Wen Lai, Mohsen Mesgar, and Alexander Fraser. 2024. LLMs beyond English: Scaling the multilingual capability of LLMs with cross-lingual feedback. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 8186–8213, Bangkok, Thailand. Association for Computational Linguistics.

Daisy Monika Lal, Paul Rayson, Krishna Pratap Singh, and Uma Shanker Tiwary. 2023. Abstractive Hindi text summarization: A challenge in a low-resource setting. In *Proceedings of the 20th International Conference on Natural Language Processing (ICON)*, pages 603–612, Goa University, Goa, India. NLP Association of India (NLPAI).

Onkar Litake, Niraj Yagnik, and Shreyas Labhsetwar. 2024. Inditext boost: Text augmentation for low resource india languages. *arXiv preprint arXiv:2401.13085*.

Sandeep Maddu and Vizianand Row Sanapala. 2024. A survey on nlp tasks, resources and techniques for low-resource telugu-english code-mixed text. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*

Tamzeed Mahfuz, Satak Kumar Dey, Ruwad Naswan, Hasnaen Adil, Khondker Salman Sayeed, and Haz Sameen Shahgir. 2025. Too late to train, too early to use? a study on necessity and viability of low-resource Bengali LLMs. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 1183–1200, Abu Dhabi, UAE. Association for Computational Linguistics.

Shervin Malmasi, Anjie Fang, Besnik Fetahu, Sudipta Kar, and Oleg Rokhlenko. 2022. Multiconer: A large-scale multilingual dataset for complex named entity recognition. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 3798–3809. International Committee on Computational Linguistics.

Utsav Maskey, Manish Bhatta, Shivangi Bhatt, Sanket Dhungel, and Bal Krishna Bal. 2022. Nepali encoder transformers: An analysis of auto encoding transformer language models for nepali text classification. In *SIGUL*.

Devansh Mehta, Sebastin Santy, Ramaravind Kommiya Mothilal, Brij Mohan Lal Srivastava, Alok Sharma, Anurag Shukla, Vishnu Prasad, Venkanna U, Amit Sharma, and Kalika Bali. 2020. Learnings from technological interventions in a low resource language: A case-study on Gondi. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 2832–2838, Marseille, France. European Language Resources Association.

Ritwik Mishra, Pooja Desur, Rajiv Ratn Shah, and Ponnurangam Kumaraguru. 2024. Multilingual coreference resolution in low-resource south asian languages. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 11813–11826. ELRA and ICCL.

Ibraheem Muhammad Moosa, Mahmud Elahi Akhter, and Ashfia Binte Habib. 2023. Does transliteration help multilingual language modeling? In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 670–685, Dubrovnik, Croatia. Association for Computational Linguistics.

Rudra Murthy, Mitesh M Khapra, and Pushpak Bhattacharyya. 2018. Improving ner tagging performance in low-resource languages via multilingual learning. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, 18.

Arijit Nag, Animesh Mukherjee, Niloy Ganguly, and Soumen Chakrabarti. 2024. Cost-performance optimization for processing low-resource language tasks using commercial llms. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 15681–15701. Association for Computational Linguistics.

Manu Narayanan and Noemi Aepli. 2024. A tulu resource for machine translation. In *International Conference on Language Resources and Evaluation*.

Abhijnan Nath, Sheikh Mannan, and Nikhil Krishnaswamy. 2023. AxomiyaBERTa: A phonologically-aware transformer model for Assamese. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 11629–11646, Toronto, Canada. 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.

Vaishali Pal, Evangelos Kanoulas, Andrew Yates, and Maarten de Rijke. 2024. Table question answering for low-resourced indic languages. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 75–92. Association for Computational Linguistics.

Amir Panahandeh, Hanie Asemi, and Esmail Nourani. 2023. Tppoet: Transformer-based persian poem generation using minimal data and advanced decoding techniques. *ArXiv*, abs/2312.02125.

Vaidehi Patil, Partha Talukdar, and Sunita Sarawagi. 2022. Overlap-based vocabulary generation improves cross-lingual transfer among related languages. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 219–233. Association for Computational Linguistics.

Ravinga Perera, Thilakshi Fonseka, Rashmini Naranpanawa, and Uthayasanker Thayasivam. 2022. Improving english to sinhala neural machine translation using part-of-speech tag. *ArXiv*, abs/2202.08882.

Aleksandar Petrov, Emanuele La Malfa, Philip Torr, and Adel Bibi. 2023. Language model tokenizers introduce unfairness between languages. In *Advances in Neural Information Processing Systems*, volume 36, pages 36963–36990. Curran Associates, Inc.

Jerin Philip, Shashank Siripragada, Vinay P Namboodiri, and C V Jawahar. 2021. Revisiting low resource status of indian languages in machine translation. In *Proceedings of the 3rd ACM India Joint International Conference on Data Science and Management of Data (8th ACM IKDD CODS 26th COMAD)*, pages 178–187. Association for Computing Machinery.

Maja Popović. 2017. chrF++: words helping character n-grams. In *Proceedings of the Second Conference on Machine Translation*, pages 612–618, Copenhagen, Denmark. Association for Computational Linguistics.

Shabdapurush Poudel, Bal Krishna Bal, and Praveen Acharya. 2024. Bidirectional english-nepali machine translation(mt) system for legal domain. In *Proceedings of the 3rd Annual Meeting of the Special Interest Group on Under-resourced Languages @ LREC-COLING 2024*, pages 53–58. ELRA and ICCL.

Ashmari Pramodya. 2023. Exploring low-resource neural machine translation for sinhala-tamil language pair. In *Recent Advances in Natural Language Processing*.

Kabilan Prasanna and Aryaman Arora. 2024. Irumozhi: Automatically classifying diglossia in tamil. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 3096–3103. Association for Computational Linguistics.

Md Nishat Raihan, Dhiman Goswami, and Antara Mahmud. 2023. Mixed-distil-bert: Code-mixed language modeling for bangla, english, and hindi. *arXiv preprint arXiv:2309.10272*.

Pawan Rajpoot, Nagaraj Bhat, and Ashish Shrivastava. 2024. Multimodal machine translation for low-resource Indic languages: A chain-of-thought approach using large language models. In *Proceedings of the Ninth Conference on Machine Translation*, pages 833–838, Miami, Florida, USA. Association for Computational Linguistics.

Krithika Ramesh, Sunayana Sitaram, and Monojit Choudhury. 2023. Fairness in language models beyond english: Gaps and challenges. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 2106–2119. Association for Computational Linguistics.

Surangika Ranathunga, En-Shiun Annie Lee, Marjana Prifti Skenduli, Ravi Shekhar, Mehreen Alam, and Rishemjit Kaur. 2023. Neural machine translation for low-resource languages: A survey. *ACM Comput. Surv.*, 55(11).

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.

Maithili Sabane, Aparna Ranade, Onkar Litake, Parth Patil, Raviraj Joshi, and Dipali Kadam. 2023. Enhancing low resource ner using assisting language and transfer learning. In *2023 2nd International Conference on Applied Artificial Intelligence and Computing (ICAAIC)*, pages 1666–1671.

Sheikh Shafayat, Eunsu Kim, Juhyun Oh, and Alice Oh. 2024. Multi-FAct: Assessing factuality of multilingual LLMs using FActscore. In *First Conference on Language Modeling*.

Gopendra Vikram Singh, Soumitra Ghosh, Mauajama Firdaus, Asif Ekbal, and Pushpak Bhattacharyya. 2024. Predicting multi-label emojis, emotions, and sentiments in code-mixed texts using an emojifying sentiments framework. *Scientific Reports*, 14:12204.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models. *Preprint*, arXiv:2307.09288.

Lewis Tunstall, Edward Beeching, Nathan Lambert, Nazneen Rajani, Kashif Rasul, Younes Belkada, Shengyi Huang, Leandro von Werra, Clémentine Fourrier, Nathan Habib, Nathan Sarrazin, Omar Sanseviero, Alexander M. Rush, and Thomas Wolf. 2023. Zephyr: Direct distillation of lm alignment. *Preprint*, arXiv:2310.16944.

Ashok Urlana, Pinzhen Chen, Zheng Zhao, Shay Cohen, Manish Shrivastava, and Barry Haddow. 2023. Pmindiasum: Multilingual and cross-lingual headline summarization for languages in india. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 11606–11628. Association for Computational Linguistics.

Aniket Vashishtha, Kabir Ahuja, and Sunayana Sitaram. 2023. On evaluating and mitigating gender biases in multilingual settings. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 307–318. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 6000–6010, Red Hook, NY, USA. Curran Associates Inc.

Gopalakrishnan Venkatesh, Abhik Jana, Steffen Remus, Özge Sevgili, Gopalakrishnan Srinivasaraghavan, and Chris Biemann. 2022. Using distributional thesaurus to enhance transformer-based contextualized representations for low resource languages. *Proceedings of the 37th ACM/SIGAPP Symposium on Applied Computing*.

Jacques Véron, Rosemary Kneipp, Godfrey Rogers, et al. 2008. The demography of south asia from the 1950s to the 2000s. *Population*, 63(1):9–89.

Lianxi Wang, Yujia Tian, and Zhuowei Chen. 2024. Enhancing hindi feature representation through fusion of dual-script word embeddings. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 5966–5976. ELRA and ICCL.

Dipendra Yadav, Sumaiya Suravee, Tobias Strauss, and Kristina Yordanova. 2024. Cross-lingual named entity recognition for low-resource languages: A hindi-nepali case study using multilingual bert models. *Proceedings of the Fourth Workshop on Multilingual Representation Learning (MRL 2024)*.

Wenbin Yu, Lei Yin, Chengjun Zhang, Yadang Chen, and Alex X Liu. 2024. Application of quantum recurrent neural network in low-resource language text classification. *IEEE Transactions on Quantum Engineering*, 5:1–13.

Xinjie Zhao, Hao Wang, Shyaman Maduranga Sriwarnasinghe, Jiacheng Tang, Shiyun Wang, Sayaka Sugiyama, and So Morikawa. 2025. Enhancing participatory development research in South Asia through LLM agents system: An empirically-grounded methodological initiative from field evidence in Sri Lankan. In *Proceedings of the First Workshop on Natural Language Processing for Indo-Aryan and Dravidian Languages*, pages 108–121, Abu Dhabi. Association for Computational Linguistics.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Preprint*, arXiv:2306.05685.

Li Zhou, Antonia Karamolegkou, Wenyu Chen, and Daniel Hershcovich. 2023. Cultural compass: Predicting transfer learning success in offensive language detection with cultural features. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 12684–12702. Association for Computational Linguistics.

A Appendix

A.1 Potential Directions

Research on code-mixing Code-mixing is very commonly observed in South Asian communication (Huzaifah et al., 2024), yet most existing resources focus on English-Hindi or English-Tamil interactions. Future work should involve code-mixing with very low-resource languages, as well as instances where speakers switch between three or more languages. Understanding which scenarios prompt code-mixing (e.g., informal communication) could improve model generalization across multilingual contexts.

Utilizing the region's Bilingualism for Data Collection Many South Asians language speakers are bilingual (Bhatia and Ritchie, 2006), which presents an opportunity for efficient data collection efforts. We recommend building parallel datasets that pair low-resource languages with better-resourced counterparts like Hindi, Urdu, Bengali or Tamil. Since English-Hindi and English-Tamil translation models are already well-developed, they could serve as pivots for translating very low-resource languages into English.

Additionally, given the script and vocabulary overlap among closely-related or geographically proximate languages (Patil et al., 2022; Chowdhury et al., 2022), future work could explore cross-lingual data augmentation techniques, particularly for regional dialects.

Addressing Bias and Improving Evaluation There are existing biases related to gender, caste, and sociocultural representation in multilingual corpora and models (Bhatt et al., 2022; Ramesh et al., 2023). However, there are no South-Asian specific large-scale bias evaluation resources. Many datasets rely on translations from English, often resulting in cultural misalignment. To address these issues, we recommend greater involvement of native speakers in dataset curation and annotation, as well as reduced reliance on English-origin resources.

Beyond bias, many widely used metrics such as BLEU and COMET do not sufficiently cover many South Asian languages. Future research should focus on developing comprehensive evaluation frameworks for these languages.

Developing Computationally Efficient NLP Models Given the computational constraints in many South Asian research facilities (Philip et al., 2021), future work should prioritize efficient fine-tuning strategies such as adapter-based tuning and LoRA. Additionally, reasoning and logical inference is being explored in multilingual contexts (Ghosh et al., 2025), but remains under-explored in South Asian NLP. Further research would improve the decision-making capabilities of models catering to South Asian languages.