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

Sampoorna Poria[1], Xiaolei Huang[2]
[1] Dept of Computer Science & Engineering, West Bengal University of Technology,
[2] Department of Computer Science, University of Memphis
sampoornaporia@gmail.com, xiaolei.huang@memphis.edu

Abstract

Rapid developments of large language models have revolutionized many NLP tasks for English data. Unfortunately, the models and their evaluations for low-resource languages are being overlooked, especially for languages in South Asia. Although there are more than 650 languages in South Asia, many of them either have very limited computational resources or are missing from existing language models. Thus, a concrete question to be answered 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[1], we have comprehensively examined current efforts and challenges of NLP models for South Asian languages by retrieving studies since 2020, with a focus on transformer-based models, such as BERT, T5, & GPT. We present advances and gaps across 3 essential aspects: data, models, & tasks, such as available data sources, fine-tuning strategies, & domain applications. Our findings highlight substantial issues, including missing data in critical domains (e.g., health), code-mixing, and lack of standardized evaluation benchmarks. Our survey aims to raise awareness within the NLP community for more targeted data curation, unify benchmarks tailored to cultural and linguistic nuances of South Asia, and encourage an equitable representation of South Asian languages. The complete list of resources is available at: https://github.com/trust-nlp/LM4SouthAsia-Survey.[2]

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). However, the regional languages are often missing from training corpora or present in imbalanced quantities (Khan et al., 2024), and many of them are not supported by current large language models (LLMs) (Lai 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. Crucially, this framing moves beyond definitions based solely on speaker population, since even widely spoken languages like Hindi and Bengali remain under-resourced in terms of benchmark coverage and model support. While low-resource languages have been studied for various regions (Aji et al., 2023, 2022; Adebara and Abdul-Mageed, 2022), there is no comprehensive study on the current status of South Asian NLP, which will be fulfilled by this survey, as outlined in Table 1.

Study retrieval methods. We retrieved relevant studies from 2020 onward via ACL Anthology, Semantic Scholar, and Google Scholar by broad and specific keyword combinations. We extended the publication list by screening their citation networks in Google Scholar, such as journals or workshop venues. To assess on the latest trends, we excluded papers before 2020 and focused on neural and Transformer-based models. The detailed methodology is presented in Appendix A.1.

Objectives and Contributions. We assess the current state of NLP research for South Asian languages and summarize their key issues, evaluation limits, and research gaps unique to these languages. Unlike prior related surveys in Table 1, our work makes three unique contributions: 1) we present

[1]Bhaasha (Hindi), Bhāṣā (Bengali), and Zabān (Urdu/Persian) all mean "language" and are commonly used across South Asian language families, underscoring the paper's inclusive focus.

[2]This work was done when the first author was a remote intern at the University of Memphis.

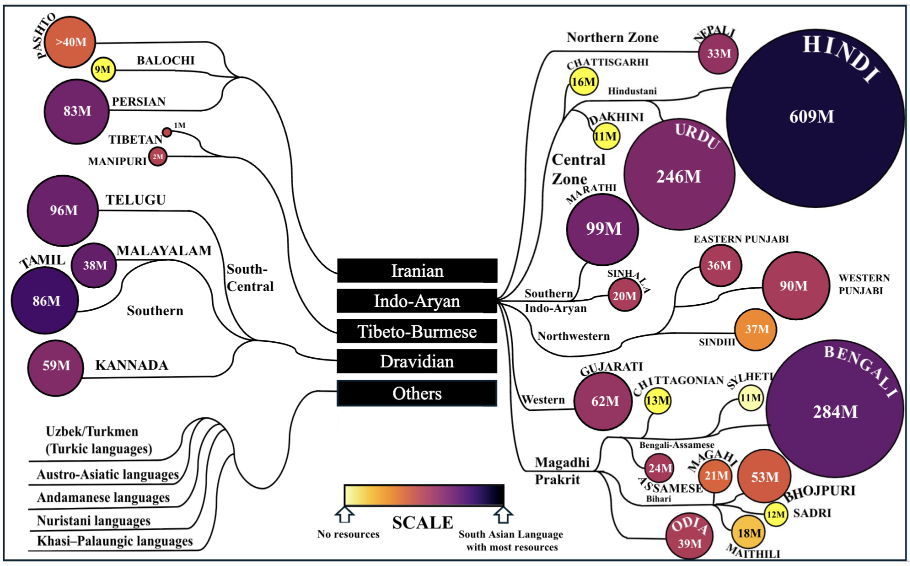

Figure 1: Language families regarding Speaker population and Resource availability. Bubble Size indicates speaker population per language and color intensity indicates the amount retrieved NLP resources. Darker color means more resources, and vice versa. "Resource size" refers to the number of papers in the ACL Anthology (until 2024) that mention the language in the title and/or abstract. Languages primarily spoken outside South Asia (e.g., Uzbek) are excluded from resource size visualization to maintain regional focus.

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 of low-resourced languages to ours by multiple key criteria. We denote superscript [1] as not specific to South-Asian languages; [2] as limited discussion of LLMs; and [3] as related to multilingual models but not for LLMs or low-resourced languages. "Interdisciplinary Integration" refers to studies connecting NLP with health, education, etc.

comprehensive language families in South Asia and broadens 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) we analyze studies across various domains (e.g., healthcare and education) and summarize recent LLMs and their tuning strategies (e.g., LoRA (Hu et al., 2022)). We hope this survey will inspire future directions to strengthen NLP community efforts for underrepresented languages in South Asia.

2 Data and Resources

A large text corpus is essential to enable language models to understand complex and heterogeneous semantics and structures of South Asian languages. Indeed, over 650 languages are spoken in the region, yet computational resources remain scarce and 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, most language resources consist of small text samples, with a major focus on languages like Hindi and Urdu (Kakwani et al., 2020; Philip et al., 2021; Gala et al., 2023). However, existing studies may merely address the questions that will be answered in our study: 1) *What are the available corpora for the low-resourced languages in South Asia?* 2) *What NLP tasks are in the corpora?* and 3) *What domains are the corpora?* To answer those questions, we summarize data distributions by language families in Figure 1 and statistics in Table 2.

2.1 Language resources

Figure 1 presents the uneven distribution of South Asian languages in our collected resources. The color gradient and circle sizes show that there are a few dominant languages with comparatively more resources, such as Hindi, Bengali, and Telugu, while the others are severely underrepresented. This highlights resource challenges and opportunities. We categorize retrieved studies by language family: Indo-Aryan, Dravidian, Tibeto-Burman, and Iranian languages.

Data	Language(s)	Size	NLP Task	Year	Source	Domain	Acc
			Datasets				
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
Samanantar	Multiple (11)	12.4M	Machine Translation	2022	Ramesh et al.	General	Yes
IndicCorp	Multiple (11)	453M	LM Pretraining	2020	Kakwani et al.	News	Yes
Sangraha	Multiple (22)	74.8M	LM Pretraining	2024	Khan et al.	General	Yes
HinDialect	Multiple (26)	-	Model Pretraining	2022	Bafna et al.	General	Yes
L3Cube-IndicNews	Multiple (11)	360K	Headline/Document Classification	2023	Mirashi et al.	News	Yes
Aksharantar	Multiple(21)	26M	Transliteration	2023	Madhani et al.	General	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
IndicSynth	Multiple (12)	4000	Audio Deepfake Detection	2025	Sharma et al.	General	Yes
CaLMQA	Multiple (23)	1.5K	LFQA	2024	Arora et al.	Culture&Society	Yes
MultiCoNER	Multiple (11)	26M	NER	2022	Malmasi et al.	Wiki&Search	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
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
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
HATS	Hindi	405	LLM Reasoning Evaluation	2025	Gupta et al.	Education	Yes
WoNBias	Bengali	31,484	Bias Classification	2025	Aupi et al.	Culture&Society	Yes
UFN2023	Urdu	4,097	Human/Machine Fake News Detection	2025	Ali et al.	News	Yes
Flickr30K (EN-(hi-IN))	Hindi	156,915	Multimodal Machine Translation	2018	Chowdhury et al.	Image Captions	Req
SENTIMOJI	Hindi	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 Summarization Dataset	Mizo	500	News Summarization	2024	Bala et al.	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	3,015	Multimodal Ques. Summarization	2024	Ghosh et al.	Healthcare	Yes
BNSENTMIX	Bengali	20K	Sentiment Analysis	2025	Alam et al.	Social Media	Yes
VACASPATI	Bengali	11M	Multiple Downstream Tasks	2023	Bhattacharyya et al.	Literature	Yes
Multi³Hate	Hindi	300	Multimodal Hate Detection	2025	Bui et al.	Social Media	Yes
MDC³	Bengali	5,007	Commercial Content Classification	2025	Shanto et al.	Social media	Yes
Hindi-BEIR	Hindi	5.89M	7 Retreival Tasks	2025	Acharya et al.	General	Yes
			Benchmarks				
IN22 Benchmark	Multiple (22)	2527	Machine Translation	2023	Gala et al.	General	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
IndicNLG Benchmark	Multiple (11)	8.5M	Various Generative Tasks	2022	Kumar et al.	News, Wiki	Yes
IndicGlue	Multiple (11)	2M	Various NLU Tasks	2020	Kakwani et al.	News, Wiki	Yes
Indic-QA	Multiple (11)	-	LLM Q&A Capabilities	2025	Singh et al.	General	Yes
MILU	Multiple (11)	79,617	Knowledge/Reasoning Evaluation	2025	Verma et al.	Multiple	Yes
En-Hi Chat Translation	Hindi	16,249	Chat Translation	2022	Gain et al.	Customer Service	Yes
CounterTuringTest(CT2)	Hindi	26	Benchmarking AGTD techniques	2024	Kavathekar et al.	News	Yes
MMFCM	Hindi	-	Multimodal Ques. Summarization	2024	Ghosh et al.	Healthcare	Yes
BenNumEval	Bengali	3.2k	LLM Numerical Reasoning Capabilities	2025	Ahmed et al.	Yes	

Table 2: Available Datasets and Benchmarks for Low-Resource South Asian Languages Across Tasks and Domains, organized by resource type (task-specific and general-purpose datasets, followed by benchmarks). We denote 'Req' as Available on Request; 'Acc' as Public Accessibility.

Indo-Aryan Languages own the largest language population in South Asia and are relatively more represented in our collected studies. For example, Hindi, Bengali, Marathi, and Urdu are among the largest bubbles in Figure 1, and Hindi corpora are available for all major NLP tasks in Table 2, aligning with existing language speaker populations (Gala et al., 2023). Large-scale data are not evenly-distributed across NLP tasks. For instance, IndicMARCO, IndicCorp, IndicGlue, MultiCONER, and BELEBELE offer large-scale datasets for IR, model pretraining, NER, and reading comprehension, particularly in high-resource Indic languages (Haq et al., 2024; Malmasi et al., 2022; Bandarkar et al., 2024; Kakwani et al., 2020). However, Bhojpuri, Sindhi, and Assamese are only in a few domain-specific datasets (Baruah et al., 2024; Malmasi et al., 2022; Kumar et al., 2024): their dataset size is comparatively smaller (less than 5,000 samples) (Gala et al., 2023).

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; Mirashi et al., 2024) 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 in size compared to Indo-Aryan Languages (e.g., Hindi) and cover much fewer application domains, such as healthcare.

Tibeto-Burman and Iranian Languages are critically underrepresented. South Asia is home to 245 Tibeto-Burman and 84 Iranian languages (Hammarström et al., 2024; Eberhard et al., 2023), yet only a handful resource appear in available datasets. Manipuri, Mizo, and Bodo are **Tibeto-Burman languages** in our retrieved studies, such as summarization data (Urlana et al., 2023; Bala et al., 2024; Madhani et al., 2023). However, the other languages including Dzonkgkhe (the national language of Bhutan) are not covered. **Iranian Languages** including Pashto, Persian, & Balochi are available in our data collections, such as a paraphrase detection corpus in Pashto (Ali et al., 2024), a noun phrase chunking corpus in Persian (Kavehzadeh et al., 2022), and a question answering corpus in Balochi (Arora et al., 2024). While IndicNLG is one of the largest benchmarks, many Tibeto-Burman and Iranian languages (e.g., Dari & Wakhi) are largely missing (Kumar et al., 2022b).

2.2 NLP Tasks

The availability of NLP tasks varies by language in Table 2. For example, Indo-Aryan languages cover all major NLP tasks, such as machine translation, information extraction, and sentiment analysis; in contrast, the other language families only cover very few NLP tasks. This section summarizes major NLP tasks from the data perspective in two major categories, 1) *generative* and 2) *discriminative* tasks. Methodologies are referred to in Section 3.

Generative NLP tasks cover three major tasks, machine translation, text generation, and summarization. <u>Machine translation</u> is the most represented task in Table 2, including BPCC (Gala et al., 2023) and domain-specific parallel corpora 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 (Baruah et al., 2024) and cross-lingual transfer (Narayanan and Aepli, 2024). The scarcity of consistent annotations and high-quality datasets can be a critical issue. <u>Text Summarization</u> is mainly in general domains (e.g., news) for Indo-Aryan languages, such as PMIndiaSum (Urlana et al., 2023), & misses coverages of Dravidian and Tibeto-Burman languages. MedSumm data aids in multimodal summarization for Hindi-English code-mixed clinical queries, specifically for the healthcare (Ghosh et al., 2024), while domain-specific summarizations are not available in other languages. <u>Text Generation</u> resources include the IndicNLG benchmark (Kumar et al., 2022a), which covers biography generation, news headline generation, sentence summarization, paraphrasing, & question generation across 11 Indic languages. Long-form question answering remains underdeveloped (Arora et al., 2024), & chat translation resources are also scarce (Gain et al., 2022)

Discriminative NLP tasks mainly focus on sequential classifications, such as Named entity recognition (NER). Classification tasks account for the majority of discriminative NLP tasks in our study, such as sentiment analysis & hate speech detection. For example, SENTIMOJI (sentiment prediction data for Hindi-English code-mixed texts) (Singh et al., 2024a), and hate detection resources are available for Hindi, Tamil, Bengali, (Hasan et al., 2024), Kannada, and Telugu (K et al., 2024). However, sentiment analysis & hate speech detection data remain nearly absent for Tibeto-Burman & Iranian languages. The table also shows that semantic or syntactic tasks are most likely available for Hindi, such as syntactic parsing & coreference resolution (Kumar et al., 2024; Mishra et al., 2024). Similarly, recently new data releases are primarily for Hindi, such as AI-generated text detectability (Kavathekar et al., 2024).

3 Model Advances

We examine 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 Tokenization is the fundamental step to encode input text containing characters from multiple languages and usually starts by fine-tuning existing language model tokenizers. For example, Kumar et al. (2023) train FastText (Bojanowski et al., 2017) on code-mixed, transliterated, and native-script social media text for multiple Indic languages, other studies fine-tune BERT (Devlin et al., 2019) or multilingual BERT tokenizers to predict positive hope speech in Kannada-English (Hande et al., 2022), Hindi-English sentiments (Singh et al., 2024a), and review ratings (Yu et al., 2024). The Overlap BPE method (Patil et al., 2022) improves tokenization consistency on subword-level processing for orthographically similar languages.

Model	Architecture	Language	Training Strategy	Parameter Size	Year	Source
AxomiyaBERTa	BERT	Assamese	Continuous Pre-train + Supervised Fine-tuning	66M	2023	Nath et al.
IndecBERT	BERT	Multiple (11)	Continuous Pre-train on IndicCorp + Supervised Fine-tuning	12M	2020	Kakwani 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.
Matina	Transformer	Persian	Domain-specific Fine-tuning	8B	2025	Hosseinbeigi et al.
IndicTrans	Transformer	Multiple (11)	Continuous Pre-train on Samanatar + Supervised Fine-tuning	1.1B	2022	Ramesh 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.
LlamaLens	Transformer	Hindi	Instruction tuning + Domain Fine-tuning; Multilingual Shuffling	8B	2025	Kmainasi et al.
NLLB-E5	Multilingual Encoder	Hindi	Knowledge Distillation + Zero-shot transfer	1.3B	2025	Acharya et al.

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

Transformer-based models (Vaswani et al., 2017) have dominated recent developments for monolingual and multilingual settings. BERT is a common architecture on multi-domain and monolingual tasks, such as AxomiyaBERTa (Nath et al., 2023), Nepali DistilBERT and DeBERTa (Maskey et al., 2022), and MahaBERT (Joshi, 2022). For multilingual models, IndicBERT (Kakwani et al., 2020) covers classification and retrieval; IndicTrans2 (Gala et al., 2023) covers translation across 22 languages; Indic-ColBERT (Haq et al., 2024) employs retrieval-augmented supervision for search to improve document retrieval across 11 languages; and IndicBART (Dabre et al., 2022) supports NMT & summarization across 2 language families. Together, these represent some of the most comprehensive models for South Asian languages. Chowdhury et al. (2022) trains Transformer models from scratch for machine translation to Lambani, using data from closely related source languages. Classification tasks mainly use supervised fine-tuning on pre-trained BERT (Devlin et al., 2019) and its variants.

Generative LLMs are being rapidly adopted for South Asian languages in the recent 3 years. MedSumm (Ghosh et al., 2024) fine-tuned 5 public LLMs (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., 2024)) 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., 2024) fine-tuned LLaMA2-7B models on romanized Indic corpora to enable transliteration-aware and mixed-script text processing. Additionally, AI-Tutor (Dalal et al., 2024) applied IndicTrans2 (Gala et al., 2023) to Pali and Ardhamagadhi. These findings suggest that multilingual models alone cannot resolve low-resource challenges in South Asia; corpus coverage and script fidelity continue to constrain their applicability, particularly for languages with limited web presence and domain coverage.

3.2 Training and Fine-tuning Methods

Code-mixed and script-specific adaptations enable model understanding of text inputs with mixed languages. For example, LLMs struggled with Bengali script generation due to inefficient tokenization (Mahfuz et al., 2025). Studies introduced related corpora to assess code-mixed capabilities, such as IndicParaphrase (Kumar et al., 2022a), the largest Indic language paraphrasing dataset across 11 languages. Transliterating Indic languages into a common script could effectively improve cross-lingual transfer, such as NER and sentiment analysis (Moosa et al., 2023). Kirov et al. (2024) aligned transliteration patterns with phonetic structures, which further improves multilingual representation. Overlap BPE (Patil et al., 2022) finds shared subword representations, which enhances consistency for orthographically similar languages. Continual pre-training strategies (Guo et al., 2025; Zheng et al., 2024) improve adaptation without degrading prior performance, for example in machine translation (Koehn, 2024), by preventing catastrophic forgetting by iteratively fine-tuning with new language pairs. Agarwal et al. (2025) introduces script-agnostic representations for Dravidian languages and show that mixing multiple writing systems during training improves robustness. While the current studies have achieved substantial progresses, script-aware tokenization remains a foundational bottleneck to enable encoding multilingual inputs of South Asian languages.

Supervised multilingual transfer learning Given the linguistic similarities in characters and morphology, cross-lingual transfer learning has become a key adaptation strategy. Narayanan and Aepli (2024), IndicBART (Dabre et al., 2022), 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 (J et al., 2024) were fine-tuned on task-specific corpora; however, effectiveness varies based on linguistic proximity, with underrepresented languages facing performance declines (Hasan et al., 2024). Studies found that jointly trained NER models on multilingual corpora outperformed monolingual ones as for shared script and grammar, such as Hindi-Marathi (Sabane et al., 2023) and Bengali-Tamil-Malayalam (Murthy 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. Zhou et al. (2023) integrates sociolinguistic factors into offensive language detection. Poudel et al. (2024) fine-tunes with domain-specific knowledge to enhance legal translation. Cross-lingual in-context learning (ICL) (Cahyawijaya et al., 2024) improve generalization by query alignment.

Distillation and parameter-efficient finetuning (PEFT) methods Adapting large models to South Asian languages often face computing and data constraints. As a result, recent work has explored PEFT strategies like LoRA, QLoRA, and multi-step PEFT (Hu et al., 2022; Petrov et al., 2023). These approaches fine-tune models like Gemma (Khade et al., 2025) with fewer parameters and lower memory cost. While LoRA improves efficiency, its effectiveness can vary across tasks: it captures dialectal variations when combined with phonological cues (Alam and Anastasopoulos, 2025) but may struggle with syntactically rich tasks. Adapter-based methods (Nag et al., 2024) offer modular, language-specific adaptation and can avoid catastrophic forgetting when tuned with domain/task-specific knowledge.

Distillation-based approaches (Ghosh et al., 2024) compress large models but typically require access to high-quality teacher models and synthetic data, which remains a bottleneck in many South Asian contexts. Feature-based finetuning (Bhatt et al., 2022) focuses on internal representation refinement to enable knowledge transfer across resource boundaries. Other strategies like rank-adaptive LoRA (Yadav et al., 2024) balance parameter savings with performance. Complementary strategies such as QLoRA (Dettmers et al., 2023) reduce memory overhead, while data-centric approaches like IndiText Boost (Litake et al., 2024) combine augmentation techniques to enhance classification for morphologically rich languages (e.g., Sindhi, Marathi). Few-shot learning offers flexibility but still struggles with syntactic generalization (Nag et al., 2024; Pal et al., 2024). While parameter-efficient & data-light methods have achieved progress, their benefits are uneven across linguistic variations, rarely extending to the least-resourced.

3.3 Model Evaluations

Model evaluation varies by task, such as BLEU for generation and human evaluation (Gala et al., 2023; Narayanan and Aepli, 2024; Duwal et al., 2025). Tables 2 and 3 summarize diverse evaluation approaches such as FLORES for machine translation (Goyal et al., 2022; Gala et al., 2023). NER (Venkatesh et al., 2022; Khemchandani et al., 2021; J et al., 2024) and sentiment analysis (Hande et al., 2022; Singh et al., 2024a) usually include accuracy, F1-score, precision, and recall. 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 standard metrics for generation tasks, such as summarization, machine translation, and question answering (Lal et al., 2023; Rajpoot et al., 2024; Gala et al., 2023). Recent new metrics such as COMET (Rei et al., 2020), phonetic-aware metrics like PhoBLEU (Arora et al.,

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

2023), SPBLEU (Alam and Anastasopoulos, 2025), and chrF++ (Popović, 2017) complement existing ones (Costa-jussà et al., 2024; Gajakos et al., 2024). Overall, current evaluation relies heavily on English-centric benchmarks and metrics (BLEU, F1, etc.), which can misrepresent true performance on South Asian languages and thus motivate the need for region-specific evaluation frameworks.

4 Trends and Challenges

Building on the contributions reviewed in the previous sections, we now synthesize emerging patterns and persisting challenges.

Data Scarcity and Quality Issues for low-resource languages affect model generalizability and applicability (Gala et al., 2023). Existing resources, especially small datasets, are often domain-specific (e.g., government or political) due to limited digital content and copyright restrictions, and may potentially introduce cultural or political biases in downstream applications (Gain et al., 2022; Ali et al., 2024; Urlana et al., 2023; Kumar et al., 2024). The lack of gold-annotated resources complicates tasks, such as co-reference resolution (Mishra et al., 2024), and the rapidly evolving online discourse hurts model long-term sustainability (Bandarkar et al., 2024; Kumaresan et al., 2024).

Non-standardized transliteration and representation of South Asian languages introduce 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, which are common as for missing human re-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 from human annotators' varying interpretation and background can harm sensitive tasks like hate speech detection (Kumaresan et al., 2024).

Further, certain data exhibit class imbalances, leading to bias toward majority classes; solutions such as cost-sensitive learning and oversampling have been proposed (K et al., 2024) but not examined. Languages exhibiting diglossia 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). Models often learn script-dependent embeddings, which limits cross-script generalization (Koehn, 2024). For example, transliteration ambiguity can easily affect speech-text alignment in ASR models (Ramesh et al., 2023).

Existing tokenization strategies such as Byte-Pair Encoding (BPE) (Gage, 1994) and WordPiece (Devlin et al., 2019) frequently fragment morphologically rich words in Dravidian and Indo-Aryan languages, leading to over-segmentation and loss of meaning (Wang et al., 2024). 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 challenging for Dravidian languages as words are formed by adding multiple suffixes (Narayanan and Aepli, 2024). Hindi, Assamese, & Bengali exhibit different, complex inflectional systems complicating parsing (Chowdhury et al., 2018; Nath et al., 2023). Most Indo-Aryan languages rely on dependent vowel signs (matras) & nasalization markers, where BERT tokenizers often split them incorrectly (Doddapaneni et al., 2023) and cause ambiguities (Maskey et al., 2022). For instance, the word "फूल" (Flower) can be incorrectly tokenized as "फल" (Fruit). Assamese possesses unique sound patterns & alveolar stops, showing the tokenization complexity (Nath et al., 2023). Besides structural differences, administrative vocabulary include Persian-origin words like "farman" (order), alongside English-origin terms (Pramodya, 2023).

Code mixing, Diglossia, and Ambiguity are highly domain-dependent issues and can integrate English letters, words, or phrases, such as Hinglish/Tanglish (Das et al., 2024). Diglossia shows substantial differences in speaking and writing. For example, Literary Tamil retains 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 typically capitalize proper nouns, making it difficult to distinguish named entities from common words (Philip et al., 2021); "Hindustan" (हिन्दुस्थान) can refer to a location, a person, or an organization (Mishra et al., 2024). Many languages are grammatically gendered, even inanimate objects being referred to with gendered pronouns (Ramesh et al., 2023).

Dialect Variations and Continua are common issues in South Asian corpus development as most studies consider a single standard variety. Recent efforts have started addressing this by creating dialect-specific resources (Kumar et al., 2024; Chowdhury et al., 2025; Khandaker et al., 2024; Alam et al., 2024). For example, Bafna et al. (2022) curated HinDialect, a folk-song corpus covering 26 Hindi-related dialects; and VACASPATI (Bhattacharyya et al., 2023) compiles 115M Bengali literature sentences sampled across West Bengal and Bangladesh to capture regional lexical differences. Several studies incorporated dialectal cues into models: AxomiyaBERTa (Nath et al., 2023) includes phonological signals via an attention network; Alam and Anastasopoulos (2025) utilized LoRA (Hu et al., 2022) to achieve dialectal normalization and translation across South Asian dialects with limited supervision.

However, existing studies show that performance is lower on underrepresented dialects compared to common varieties, which reflects biases in data coverage. Annotation and orthography for dialectal text are inconsistent—many informal dialects lack standardization and the boundary between "dialect" and "standard" is often arbitrary (Sarveswaran et al., 2025). Data frequently conflate dialectal variants with the standard language, while current benchmarks rarely consider these variants. Most multilingual benchmarks only cover a few dominant languages, so dialectal evaluations are missing. CHiPSAL and recent shared tasks (e.g., NLU of Devanagari Script Languages) have started to address this by building annotated dialectal corpora (Sarveswaran et al., 2025). Together, these findings show that dialect-specific corpora and evaluation benchmarks are essential to avoid biasing models toward standard varieties.

LLM Alignment and Reasoning Tasks Current LLM benchmarks of South Asian languages suffer with very limited coverage. For example, the MMLU-ProX covers 13 languages (e.g., Hindi, Bengali) but omits many others such as Tamil, Marathi, & Kannada (Xuan et al., 2025). Even broader tests like Global-MMLU span multiple languages (e.g., Hindi, Telugu, Nepali, etc.) (Singh et al., 2024b), yet these datasets were generated by translating English questions. This leads to cultural mismatch. Many MMLU (Hendrycks et al.) questions (e.g., US History, Law) are Western-specific and thus irrelevant in South Asia; & the translation introduces artifacts that distort evaluation (Kadiyala et al., 2025). Ghosh et al. (2025) show that Hindi, the most spoken language in the region, is only represented in 5 multilingual reasoning corpora.

Recent work on cultural and value alignment (CultureLLM) fine-tunes LLMs on global survey data; however, such efforts test broad value judgments rather than deep reasoning in vernacular settings (Li et al., 2024). For example, Chiu et al. (2025) covers Bangladesh, India, Nepal, and Pakistan, but the corpus only focuses on trivia/etiquette and not cultural knowledge in the low-resourced languages spoken in the regions. In practice, South Asian languages are severely underrepresented in reasoning & alignment tasks with cultural considerations.

Standard evaluation benchmarks exist, but gaps have remained in evaluating multilingual models of South Asian language options, distributional balances, and NLP task diversities. Fine-tuned multilingual models often overfit high-resource regional languages (e.g., Hindi), leading to degraded performance on lower-resource languages (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 & Assamese) results in inconsistencies in phoneme-based word embeddings (Arif et al., 2024). Tibeto-Burman & 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).

Model evaluation from our collected studies generally rely on English-origin benchmarks in Table 3, which can misinterpret model performance (Haq et al., 2024). Das et al. (2025) mentions biases in back-translated datasets cause skewed results, compromising model evaluation across languages. For nuanced tasks (e.g., paragraph-level translation), sentence-level evaluation methods may not be sufficient (E et al., 2023; Hasan et al., 2024). Mukherjee et al. (2025) suggests LLM-based evaluation in the text style transfer task correlates better with human judgment than existing automatic metrics on Hindi and Bengali. Indeed, without culturally relevant & task-specific benchmarks, evaluations fail to interpret performance precisely, especially for languages with rich structural/cultural variations (Vashishtha et al., 2023).

4.1 Multilingual Resources vs South Asian-Specific Efforts

Broad multilingual resources are attracting more attentions in the NLP communities, such as two recent workshops for South Asian languages (Sarveswaran et al., 2025; Weerasinghe et al., 2025). XNLI benchmark extends English NLI to 14 languages (including Urdu) (Conneau et al., 2018), and XCOPA provides commonsense reasoning examples in 11 languages (Ponti et al., 2020). Similarly, models such as XGLM-7.5B included major South Asian languages (Lin et al., 2022), and new corpora like Glot500 (Imani et al., 2023) and MaLA-500 (Lin et al., 2024) included over 500 languages. These resources bring valuable South Asian language coverage for cross-lingual evaluation. However, they rely on general-domain and synthetic data, which can overlook region-specific linguistic and cultural features. For instance, even XGLM's balanced training includes only approximately 3.4B Hindi tokens versus 803B English, while XCOPA only covers a single Indic language.

Recent efforts explicitly address resource gaps. For example, IndicLLMSuite provides 251B tokens of pretraining and 74.8M instruction-response pair data across 22 Indian languages (Khan et al., 2024), INDIC-MARCO provides MS MARCO-style retrieval queries translated into 11 Indian languages (Haq et al., 2024), BPCC parallel corpus contains 230M English-Indic sentence pairs covering 22 Indic languages (Gala et al., 2023), and TransMuCoRes is a coreference resolution data of 31 South Asian languages (Mishra et al., 2024). These initiatives incorporate regional linguistic structures (e.g., scripts, complex morphology) and cultural context beyond generic multilingual resources.

Challenges are endless. Many cross-lingual approaches depend on back translation, introducing new bias and noise and suffering on code-switch (e.g. Hindi-English) issues (Raja and Vats, 2025; Conneau et al., 2018). Standard metrics may fail on region-specific phenomena (Mishra et al., 2024) among Indic languages. These persistent gaps underscore the necessity of region-specific research to ensure equitable and diverse NLP advancements.

5 Conclusion

In this study, we provide comprehensive synthesis and analysis of recent NLP advances on low-resourced languages in South Asia. Our work examines persisting challenges at every stage of resource development—uneven representation in multilingual corpora, model availability, multilingual tuning, and evaluation benchmarks. While a few languages have received more attention, challenges remain in collecting and processing data and adapting models to specific orthographies. Moreover, existing evaluation metrics fall short due to a lack of script- and task-specific benchmarks, as well as overlooked sociocultural biases. We present model tuning guidelines that reflect current limitations of South Asian NLP, calling for South Asian-specific frameworks and script-aware model adaptation. We include our future envisions in Appendix A.2. We expect this study can encourage broader participation in advancing further research of low-resource languages in South Asia.

Acknowledgment

The authors thank anonymous reviewers for their insightful feedback. This work has been partially supported by the National Science Foundation (NSF) CNS-2318210 (Sharif et al., 2025).

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 the ACL Anthology. During the retrieval stage, we conducted an extensive search across various sources, such as 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 extremely low-resource languages remain inaccessible because they are published in regional or less widely-indexed journals. We have, to our best efforts, included such publications by searching sources outside of major repositories, especially for Tibeto-Burman and Iranian languages. Future work could benefit from engagement with regional scholars and institutions to access non-digitized resources.

References

Arkadeep Acharya, Rudra Murthy, Vishwajeet Kumar, and Jaydeep Sen. 2025. Benchmarking and building zero-shot Hindi retrieval model with Hindi-BEIR and NLLB-e5. 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 4328–4348, Albuquerque, New Mexico. Association for Computational Linguistics.

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.

Milind Agarwal, Joshua Otten, and Antonios Anastasopoulos. 2025. Script-agnosticism and its impact on language identification for Dravidian 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 7364–7384, Albuquerque, New Mexico. Association for Computational Linguistics.

Kawsar Ahmed, Md Osama, Omar Sharif, Eftekhar Hossain, and Mohammed Moshiul Hoque. 2025. BenNumEval: A benchmark to assess LLMs' numerical reasoning capabilities in Bengali. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 17782–17799, Vienna, Austria. 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.

Christopher Akiki, Giada Pistilli, Margot Mieskes, Matthias Gallé, Thomas Wolf, Suzana Ilic, and

Yacine Jernite. Bigscience: A case study in the social construction of a multilingual large language model. In *Workshop on Broadening Research Collaborations 2022*.

Md Mahfuz Ibn Alam, Sina Ahmadi, and Antonios Anastasopoulos. 2024. CODET: A benchmark for contrastive dialectal evaluation of machine translation. In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 1790–1859, St. Julian's, Malta. Association for Computational Linguistics.

Md Mahfuz Ibn Alam and Antonios Anastasopoulos. 2025. Large language models as a normalizer for transliteration and dialectal translation. In *Proceedings of the 12th Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 39–67, Abu Dhabi, UAE. Association for Computational Linguistics.

Sadia Alam, Md Farhan Ishmam, Navid Hasin Alvee, Md Shahnewaz Siddique, Md Azam Hossain, and Abu Raihan Mostofa Kamal. 2025. BnSentMix: A diverse Bengali-English code-mixed dataset for sentiment analysis. In *Proceedings of the First Workshop on Language Models for Low-Resource Languages*, pages 68–77, Abu Dhabi, United Arab Emirates. 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.

Muhammad Zain Ali, Yuxia Wang, Bernhard Pfahringer, and Tony C Smith. 2025. Detection of human and machine-authored fake news in Urdu. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3419–3428, Vienna, Austria. Association for Computational Linguistics.

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.

Gaurav Arora, Srujana Merugu, and Vivek Sembium. 2023. CoMix: Guide transformers to code-mix using POS structure and phonetics. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 7985–8002, Toronto, Canada. 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. *CoRR*.

Md. Raisul Islam Aupi, Nishat Tafannum, Md. Shahidur Rahman, Kh Mahmudul Hassan, and Naimur Rahman. 2025. WoNBias: A dataset for classifying bias & prejudice against women in Bengali text. In *Proceedings of the 6th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pages 105–110, Vienna, Austria. Association for Computational Linguistics.

Niyati Bafna, Josef van Genabith, Cristina España-Bonet, and Zdeněk Žabokrtský. 2022. Combining noisy semantic signals with orthographic cues: Cognate induction for the Indic dialect continuum. In *Proceedings of the 26th Conference on Computational Natural Language Learning (CoNLL)*, pages 110–131, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

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. Re-contextualizing 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.

Pramit Bhattacharyya, Joydeep Mondal, Subhadip Maji, and Arnab Bhattacharya. 2023. VACASPATI: A diverse corpus of Bangla literature. 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 (Volume 1: Long Papers)*, pages 1118–1130, Nusa Dua, Bali. Association for Computational Linguistics.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

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).

Minh Duc Bui, Katharina Von Der Wense, and Anne Lauscher. 2025. Multi^{3}Hate: Multimodal, multilingual, and multicultural hate speech detection with vision–language models. 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 9714–9731, Albuquerque, New Mexico. Association for Computational Linguistics.

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.

Yu Ying Chiu, Liwei Jiang, Bill Yuchen Lin, Chan Young Park, Shuyue Stella Li, Sahithya Ravi, Mehar Bhatia, Maria Antoniak, Yulia Tsvetkov, Vered Shwartz, and Yejin Choi. 2025. CulturalBench: A robust, diverse and challenging benchmark for measuring LMs' cultural knowledge through human-AI red-teaming. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 25663–25701, Vienna, Austria. 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.

Sinthia Chowdhury, Deawan Remal, Syed Pasha, and Sheak Noori. 2025. Chatgaiyyaalap: A dataset for conversion from chittagonian dialect to standard bangla. *Data in Brief*, 59:111413.

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.

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.

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.

Sudhansu Bala Das, Samujjal Choudhury, Dr Tapas Kumar Mishra, and Dr Bidyut Kr Patra. 2025. Investigating the effect of backtranslation for Indic languages. In *Proceedings of the First Workshop on Natural Language Processing for Indo-Aryan and Dravidian Languages*, pages 152–165, Abu Dhabi. Association for Computational Linguistics.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: Efficient finetuning of quantized llms. *Advances in neural information processing systems*, 36:10088–10115.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

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.

Sharad Duwal, Suraj Prasai, and Suresh Manandhar. 2025. Domain-adaptative continual learning for low-resource tasks: Evaluation on Nepali. In *Proceedings of the First Workshop on Challenges in Processing South Asian Languages (CHiPSAL 2025)*, pages 144–153, Abu Dhabi, UAE. International Committee on Computational Linguistics.

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.

David M. Eberhard, Gary F. Simons, and Charles D. Fennig. 2023. *Ethnologue: Languages of the World*, 26th edition. SIL International, Dallas, TX.

Philip Gage. 1994. A new algorithm for data compression. *C Users Journal*, 12(2):23–38.

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, Glasgow, Scotland. Springer, Cham.

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.

Yiduo Guo, Jie Fu, Huishuai Zhang, and Dongyan Zhao. 2025. Efficient domain continual pretraining by mitigating the stability gap. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 32850–32870, Vienna, Austria. Association for Computational Linguistics.

Ashray Gupta, Rohan Joseph, and Sunny Rai. 2025. HATS : Hindi analogy test set for evaluating reasoning in large language models. In *Proceedings of the 2nd Workshop on Analogical Abstraction in Cognition, Perception, and Language (Analogy-Angle II)*, pages 57–80, Vienna, Austria. Association for Computational Linguistics.

Harald Hammarström, Robert Forkel, Martin Haspelmath, and Sebastian Bank. 2024. Glottolog 5.1. Max Planck Institute for Evolutionary Anthropology, Leipzig. Accessed 2025-05-17.

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.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In *International Conference on Learning Representations*.

Sara Bourbour Hosseinbeigi, MohammadAli SeifKashani, Javad Seraj, Fatemeh Taherinezhad, Ali Nafisi, Fatemeh Nadi, Iman Barati, Hosein Hasani, Mostafa Amiri, and Mostafa Masoudi. 2025. Matina: A culturally-aligned Persian language model using multiple LoRA experts. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 20874–20889, Vienna, Austria. 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*.

Ayyoob Imani, Peiqin Lin, Amir Hossein Kargaran, Silvia Severini, Masoud Jalili Sabet, Nora Kassner, Chunlan Ma, Helmut Schmid, André Martins, François Yvon, and Hinrich Schütze. 2023. Glot500: Scaling multilingual corpora and language models to 500 languages. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1082–1117, Toronto, Canada. Association for Computational Linguistics.

Jaavid J, Raj Dabre, Aswanth M, Jay Gala, Thanmay Jayakumar, Ratish Puduppully, and Anoop Kunchukuttan. 2024. 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.

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*.

Ram Mohan Rao Kadiyala, Siddartha Pullakhandam, Siddhant Gupta, Drishti Sharma, Jebish Purbey, Kanwal Mehreen, Muhammad Arham, and Hamza Farooq. 2025. Improving multilingual capabilities with cultural and local knowledge in large language models while enhancing native performance. *arXiv preprint arXiv:2504.09753*.

Divyanshu Kakwani, Anoop Kunchukuttan, Satish Golla, Gokul N.C., Avik Bhattacharyya, Mitesh M. Khapra, and Pratyush Kumar. 2020. IndicNLPSuite: Monolingual corpora, evaluation benchmarks and pre-trained multilingual language models for Indian languages. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4948–4961, Online. Association for Computational Linguistics.

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.

Supriya Khadka and Bijayan Bhattarai. 2025. Gender bias in Nepali-English machine translation: A comparison of LLMs and existing MT systems. In *Proceedings of the 6th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pages 75–82, Vienna, Austria. Association for Computational Linguistics.

Mohammed Khan, Priyam Mehta, Ananth Sankar, Umashankar Kumaravelan, Sumanth Doddapaneni, Suriyaprasaad B, Varun G, Sparsh Jain, Anoop Kunchukuttan, Pratyush Kumar, Raj Dabre, and Mitesh 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)*, page 15831–15879. Association for Computational Linguistics.

Md Arafat Alam Khandaker, Ziyan Shirin Raha, Bidyarthi Paul, and Tashreef Muhammad. 2024. Bridging dialects: Translating standard bangla to regional variants using neural models. In *2024 27th International Conference on Computer and Information Technology (ICCIT)*, pages 885–890. IEEE.

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.

Mohamed Bayan Kmainasi, Ali Ezzat Shahroor, Maram Hasanain, Sahinur Rahman Laskar, Naeemul Hassan, and Firoj Alam. 2025. LlamaLens: Specialized multilingual LLM for analyzing news and social media content. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 5627–5649, Albuquerque, New Mexico. Association for Computational Linguistics.

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.

Adithya Kolavi, Samarth P, and Vyoman Jain. 2025. Nayana OCR: A scalable framework for document OCR in low-resource languages. In *Proceedings of the 1st Workshop on Language Models for Underserved Communities (LM4UC 2025)*, pages 86–103, Albuquerque, New Mexico. 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.

Saurabh Kumar, Ranbir Sanasam, and Sukumar Nandi. 2023. IndiSocialFT: Multilingual word representation for Indian languages in code-mixed environment.

In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 3866–3871, Singapore. 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).

Cheng Li, Mengzhuo Chen, Jindong Wang, Sunayana Sitaram, and Xing Xie. 2024. Culturellm: Incorporating cultural differences into large language models. *Advances in Neural Information Processing Systems*, 37:84799–84838.

Peiqin Lin, Shaoxiong Ji, Jörg Tiedemann, André FT Martins, and Hinrich Schütze. 2024. Mala-500: Massive language adaptation of large language models. *CoRR*.

Xi Victoria Lin, Todor Mihaylov, Mikel Artetxe, Tianlu Wang, Shuohui Chen, Daniel Simig, Myle Ott, Naman Goyal, Shruti Bhosale, Jingfei Du, Ramakanth Pasunuru, Sam Shleifer, Punit Singh Koura, Vishrav Chaudhary, Brian O'Horo, Jeff Wang, Luke Zettlemoyer, Zornitsa Kozareva, Mona Diab, Veselin Stoyanov, and Xian Li. 2022. Few-shot learning with multilingual language models. *Preprint*, arXiv:2112.10668.

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 Viziananda 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.*

Yash Madhani, Sushane Parthan, Priyanka Bedekar, Gokul Nc, Ruchi Khapra, Anoop Kunchukuttan, Pratyush Kumar, and Mitesh Khapra. 2023. Aksharantar: Open Indic-language transliteration datasets and models for the next billion users. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 40–57, Singapore. Association for Computational Linguistics.

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.

Aishwarya Mirashi, Srushti Sonavane, Purva Lingayat, Tejas Padhiyar, and Raviraj Joshi. 2024. L3cube-indicnews: News-based short text and long document classification datasets in indic languages. *Preprint*, arXiv:2401.02254.

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.

Sourabrata Mukherjee, Atul Kr. Ojha, John Philip McCrae, and Ondrej Dusek. 2025. Evaluating text style transfer evaluation: Are there any reliable metrics? In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 4: Student Research Workshop)*, pages 418–434, Albuquerque, USA. 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.

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.

Edoardo Maria Ponti, Goran Glavaš, Olga Majewska, Qianchu Liu, Ivan Vulić, and Anna Korhonen. 2020. XCOPA: A multilingual dataset for causal commonsense reasoning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2362–2376, Online. Association for Computational Linguistics.

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.

Zahra Pourbahman, Fatemeh Rajabi, Mohammadhossein Sadeghi, Omid Ghahroodi, Somayeh Bakhshaei, Arash Amini, Reza Kazemi, and Mahdieh Soleymani Baghshah. 2025. ELAB: Extensive LLM alignment benchmark in Persian language. In *Proceedings of the Fourth Workshop on Generation, Evaluation and Metrics (GEM²)*, pages 458–470, Vienna, Austria and virtual meeting. Association for Computational Linguistics.

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.

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

Rahul Raja and Arpita Vats. 2025. Parallel corpora for machine translation in low-resource indic languages: A comprehensive review. *LoResMT 2025*, page 129.

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.

Gowtham Ramesh, Sumanth Doddapaneni, Aravinth Bheemaraj, Mayank Jobanputra, Raghavan Ak, Ajitesh Sharma, Sujit Sahoo, Harshita Diddee, Divyanshu Kakwani, Navneet Kumar, et al. 2022. Samanantar: The largest publicly available parallel corpora collection for 11 indic languages. *Transactions of the Association for Computational Linguistics*, 10:145–162.

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.

Kengatharaiyer Sarveswaran, Ashwini Vaidya, Bal Krishna Bal, Sana Shams, and Surendrabikram Thapa, editors. 2025. *Proceedings of the First Workshop on Challenges in Processing South Asian Languages (CHiPSAL 2025)*. International Committee on Computational Linguistics, Abu Dhabi, UAE.

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*.

Anik Mahmud Shanto, Mst. Sanjida Jamal Priya, Fahim Shakil Tamim, and Mohammed Moshiul Hoque. 2025. MDC^3: A novel multimodal dataset for commercial content classification in Bengali. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 4: Student Research Workshop)*, pages 311–320, Albuquerque, USA. Association for Computational Linguistics.

Mayira Sharif, Guangzeng Han, Weisi Liu, and Xiaolei Huang. 2025. Cultivating multidisciplinary research and education on gpu infrastructure for mid-south institutions at the university of memphis: Practice and challenge. *Preprint*, arXiv:2504.14786.

Divya V Sharma, Vijval Ekbote, and Anubha Gupta. 2025. IndicSynth: A large-scale multilingual synthetic speech dataset for low-resource Indian languages. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 22037–22060, Vienna, Austria. Association for Computational Linguistics.

Abhishek Kumar Singh, Vishwajeet Kumar, Rudra Murthy, Jaydeep Sen, Ashish Mittal, and Ganesh Ramakrishnan. 2025. INDIC QA BENCHMARK: A multilingual benchmark to evaluate question answering capability of LLMs for Indic languages. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 2607–2626, Albuquerque, New Mexico. Association for Computational Linguistics.

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

Shivalika Singh, Angelika Romanou, Clémentine Fourrier, David Ifeoluwa Adelani, Jian Gang Ngui, Daniel Vila-Suero, Peerat Limkonchotiwat, Kelly Marchisio, Wei Qi Leong, Yosephine Susanto, et al. 2024b. Global mmlu: Understanding and addressing cultural and linguistic biases in multilingual evaluation. *CoRR*.

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 Emanuel 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. 2024. Zephyr: Direct distillation of LM alignment. In *First Conference on Language Modeling*, Philadelphia, PA. OpenReview.

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*.

Sshubam Verma, Mohammed Safi Ur Rahman Khan, Vishwajeet Kumar, Rudra Murthy, and Jaydeep Sen. 2025. MILU: A multi-task Indic language understanding benchmark. 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 10076–10132, Albuquerque, New Mexico. Association for Computational Linguistics.

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.

Ruvan Weerasinghe, Isuri Anuradha, and Deshan Sumanathilaka, editors. 2025. *Proceedings of the First Workshop on Natural Language Processing for Indo-Aryan and Dravidian Languages*. Association for Computational Linguistics, Abu Dhabi.

Weihao Xuan, Rui Yang, Heli Qi, Qingcheng Zeng, Yunze Xiao, Yun Xing, Junjue Wang, Huitao Li, Xin Li, Kunyu Yu, et al. 2025. Mmlu-prox: A multilingual benchmark for advanced large language model evaluation. *CoRR*.

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 Xing, et al. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems*, 36:46595–46623.

Wenzhen Zheng, Wenbo Pan, Xu Xu, Libo Qin, Li Yue, and Ming Zhou. 2024. Breaking language barriers: Cross-lingual continual pre-training at scale. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 7725–7738, Miami, Florida, USA. Association for Computational Linguistics.

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 Study Retrieval and Selection Methodology

To identify relevant work on natural language processing for South Asian languages, we conducted an exhaustive literature review led independently by the two authors.

We ran systematic keyword queries combining South Asian language names (e.g. Hindi, Urdu, Bengali, etc.), region-specific words (e.g., “Indic”, “South Asian”, “Low-Resource Languages”), along with task-specific keywords (e.g., “Machine Translation”, “Named Entity Recognition”, “Sentiment Analysis”, “Multilingual Pretraining”) across major databases (ACL Anthology, Semantic Scholar, and Google Scholar). This process retrieved over 1,000 initial papers. We then removed duplicates and applied inclusion criteria to focus the review: (a) study of at least one South Asian language with a speaker population $\geq$1 million, (b) use of neural or transformer-based models (e.g., BERT, mBART, T5, GPT), and (c) publication year 2020 or later. After filtering on these criteria, 188 papers remained for full analysis.

All authors independently read and annotated all 188 papers. For each paper, we recorded detailed metadata and qualitative observations using an iteratively-developed structured coding template. Disagreements in coding were resolved through discussion until consensus was reached. The annotation template included both structured metadata (for example: language(s) studied, NLP task, model architecture or family, dataset size, year, and publication venue) and emergent, inductive tags capturing noted phenomena. Examples of inductive tags include transliteration handling, dialectal variation, data scarcity, or evaluation gaps, which were added to the template as they were discovered during reading. These were added as qualitative codes and grouped into higher-order themes.

To ensure coverage of less widely reported research, we searched beyond mainstream venues using citation tracking to identify less accessible research from under-indexed sources. This included work from regional conferences like Technology Journal of Artificial Intelligence and Data Mining, etc. (Kavehzadeh et al., 2022), and workshops focused on low-resource languages. We also scanned citations of benchmark papers like IndicNLG, TransMuCoRes, and BPCC to identify follow-up work not indexed in ACL Anthology.

We prioritized the inclusion of languages with over 1 million speakers. This allowed us to include both high-resource languages like Hindi and Bengali, as well as low-resource and often overlooked ones such as Manipuri, Balochi, Santali, and Tulu. As discussed in Figure 1 and Section 2.1, the observed imbalance in dataset and model availability reflects publication patterns, not retrieval bias.

Themes for Sections 3 and 4 were identified inductively by synthesizing recurring patterns across the annotated data. As we reviewed papers, we documented recurrent patterns, gaps, and methodological approaches, which were then grouped into cohesive sections based on relevance to ongoing challenges in South Asian NLP.

A.2 Open Challenges and Future Work

Building on our survey findings, we outline several forward-looking directions to guide future NLP research for South Asian languages.

Code-Mixing Beyond Major Language Pairs Code-mixing is pervasive in South Asian communication (Huzaifah et al., 2024), yet most available corpora focus on English-Hindi or English-Tamil interactions. We encourage future work to expand toward less-resourced combinations, such as Assamese-Bodo or Hindi-Magahi, and trilingual mixing patterns. Studying the sociolinguistic contexts in which switching occurs (e.g., informal communication, shifts in topic, regional broadcasts) can inform models that generalize better to multilingual discourse. This is particularly relevant for applications like dialogue agents and education technology, where switching is frequent.

Leveraging Bilingualism and Linguistic Proximity for Parallel Data Creation Given the high rates of bilingualism in South Asia (Bhatia and Ritchie, 2006), parallel data can be efficiently constructed by pairing low-resource languages with regionally-dominant but better-resourced ones like Hindi, Tamil, or Urdu. We encourage community-driven data collection efforts that take advantage of such speaker fluency. Translation pivots using English–Hindi or English–Tamil models (Khan et al., 2024; Gala et al., 2023) can further support indirect transfer. Additionally, our findings on shared scripts and lexical similarity among related languages in Section 2.1 (e.g., Bhojpuri–Hindi, Assamese–Bengali) suggest promising avenues for cross-lingual data augmentation (Chowdhury et al., 2022; Patil et al., 2022).

Bias Mitigation and Inclusive Dataset Design As detailed in Section 4, our review identifies persistent sociocultural biases in existing resources, ranging from gender and caste under-representation to cultural misalignment in machine-translated data (Bhatt et al., 2022; Ramesh et al., 2023), with many datasets relying on translations from English. Very recent work on Nepali-English MT (Khadka and Bhattarai, 2025) also highlights that traditional systems perpetuate gender stereotypes in occupational terms (while GPT-4o demonstrates lower bias and better gender accuracy). However, there are no South-Asian specific large-scale bias evaluation resources. Future work should prioritize participatory dataset development, with native speaker involvement in both content and annotation design. Additionally, targeted efforts are needed to build corpora for languages with scheduled or official status but little NLP presence (e.g., Bodo, Sindhi, Dzongkha, Pashto).

Evaluation Frameworks Tailored to South Asia Existing benchmarks rarely capture the linguistic complexity of South Asian languages (e.g., diglossia, agglutination, script multiplicity). Metrics such as BLEU or COMET are often used by default despite them lacking sensitivity to regional variations.

We call for the creation of culturally grounded evaluation datasets across tasks like summarization, retrieval, and QA (Philip et al., 2021; Kumar et al., 2024; Pourbahman et al., 2025), alongside human-in-the-loop assessments in multilingual and code-mixed contexts.

Developing Computationally Efficient NLP Models As noted by Philip et al. (2021), South Asian research institutions often face compute constraints. Future work should prioritize efficient fine-tuning strategies such as adapter-based tuning and LoRA. For example, fine-tuning multilingual LLMs with language-specific instructions (Khan et al., 2024) or leveraging LoRA-based adapters (Huzaifah et al., 2024; Singh et al., 2024a) can yield strong performance with minimal data. 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.

Script-Robust and Transliteration-Aware Modeling South Asian languages often use multiple scripts or informal romanizations. The survey notes that transliterating text into a common script can improve cross-lingual transfer, but current models still suffer from script-specific tokenization issues (Koehn, 2024). Recent work such as Nayana (Kolavi et al., 2025) demonstrates that combining synthetic layout-aware data generation with LoRA can enable scalable OCR for 10 Indic languages wihtout requiring annotated corpora.

Future research should focus on script-agnostic modeling: for example, designing multilingual tokenizers or shared subword vocabularies that link Devanagari, Perso-Arabic, and Roman scripts. Modules that automatically transliterate or phonetically encode text (so that Hindi and Urdu versions of the same word align) could boost transfer. Such techniques (training on mixed-script data or using script-independent representations) will help models generalize across writing systems common in South Asia.

Coordinated South Asian Benchmarks and Shared Tasks We observe fragmented evaluation across studies, with little standardization. Inspired by initiatives like IndicGLUE (Kakwani et al., 2020) and BigScience (Akiki et al.), we propose community-organized shared tasks focused on regionally relevant domains (e.g., healthcare, law, government communication) and languages. These should include multilingual, multi-script benchmarks, standardized metrics, and code-mixed test sets to advance reproducibility and collaboration.