Beyond Monolingual Assumptions: A Survey of Code-Switched NLP in the Era of Large Language Models

Rajvee Sheth[§◇], **Samridhi Raj Sinha**[⋆◇*], **Mahavir Patil**[†◇*], **Himanshu Beniwal**[§◇], **Mayank Singh**[§◇]

[§]IIT Gandhinagar , [⋆]NMIMS Mumbai, [†]SVNIT Surat, [◇]LINGO Research Group

Correspondence: singh.mayank@iitgn.ac.in

Abstract

Code-switching (CSW), the alternation of languages and scripts within a single utterance, remains a fundamental challenge for multiling ual NLP, even amidst the rapid advances of large language models (LLMs). Most LLMs still struggle with mixed-language inputs, limited CSW datasets, and evaluation biases, hindering deployment in multilingual societies. This survey provides the first comprehensive analysis of CSW-aware LLM research, reviewing 308 studies spanning five research areas, 12 NLP tasks, 30+ datasets, and 80+ languages. We classify recent advances by architecture, training strategy, and evaluation methodology, outlining how LLMs have reshaped CSW modeling and what challenges persist. The paper concludes with a roadmap emphasizing the need for inclusive datasets, fair evaluation, and linguistically grounded models to achieve truly multilingual intelligence. A curated collection of all resources is maintained at [1].

1 Introduction

Code-switching (CSW), the alternation between two or more languages within a single utterance or discourse, is a pervasive feature of multilingual communication worldwide (Poplack, 1988). With the rise of digital platforms, code-switched text has become ubiquitous across social media and online communication (Molina et al., 2016; Singh and Solorio, 2017), challenging NLP systems built on monolingual assumptions. In India, 26% of the population is bilingual and 7% trilingual (Chandramouli, 2011), with over 250 million speakers engaging in code-switched discourse, yet ASR systems trained on monolingual speech exhibit 30–50% higher word error rates on CSW data (Singh et al., 2025). Multilingual NLU models also suffer up to a 15% drop in semantic accuracy (Winata et al., 2021), leaving bilingual users under-served (Cihan et al., 2022). Figure 1 depicts intra- and inter-sentential code-mixing across multiple language pairs, emphasizing the linguistic variability that NLP systems must navigate.

The evolution of CSW research mirrors key milestones in NLP. The Early Statistical Era (pre-2010) relied on rule-based and probabilistic models like n-grams, HMMs, and CRFs, laying the groundwork for bilingual text processing (Solorio and Liu, 2008). The Representation Learning Era (2010–2017) introduced distributed embeddings (Word2Vec) and recurrent models, advancing CSW tasks like LID, POS tagging, and NER (Solorio et al., 2014; Sequiera et al., 2015; Molina et al., 2016). The Contextual Understanding Era (2017–2020) brought GPT, BERT, XLM, and T5, enabling fine-tuning for code-switched data, though multilingual pretraining alone proved insufficient for robust CSW performance (Winata et al., 2021). The Foundation Model Era (2020–present) leverages LLMs like GPT-3, PaLM, and LLaMA for multilingual pretraining and prompt-based adaptation (Wang et al., 2025b).

LLMs have transformed CSW investigation across typologically diverse language pairs—Spanish-English (Sheokand et al., 2025), Hindi-English (Sheth et al., 2025), Chinese-English (Kong and Macken, 2025), Korean-English (Yoo et al., 2025), Ukrainian-Russian (Shynkarov et al., 2025), Cantonese-Mandarin (Dai et al., 2025), and Arabic-English (Issa et al., 2025), deepening our linguistic and sociocultural understanding of switching patterns (Yoo et al., 2024; Jehan et al., 2025). Methodologically, current LLMs support in-context mixing (Shankar et al., 2024), instruction tuning for low-resource CSW (Lee et al., 2024), synthetic data augmentation (Zeng, 2024), and advanced metrics

[*]Work done while interning at IIT Gandhinagar.

[1]https://github.com/lingo-iitgn/awesome-code-mixing/

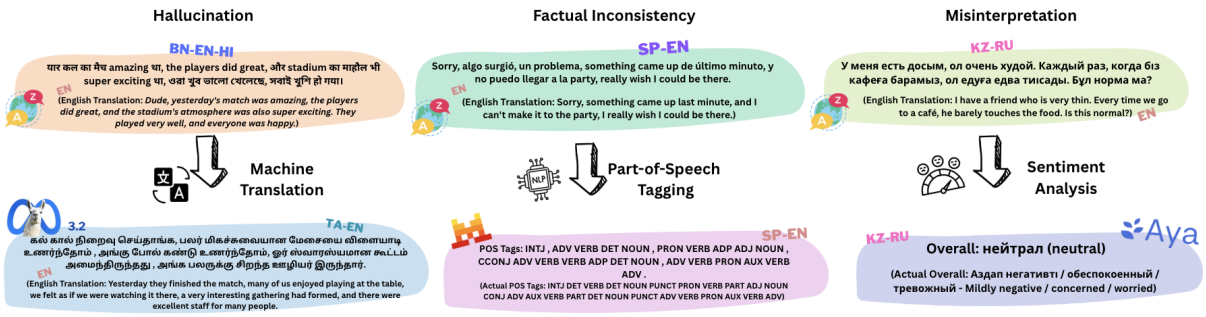

Figure 1: Common model failures on code-mixed text: ***Takeaway*** *(a) hallucination in MT translation (Bn-Hi-En), (b) factual inconsistency in POS tagging (Sp-En), and (c) misinterpretation in SA (Kz-Ru).*

evaluating structural and socio-pragmatic aspects (Ugan et al., 2025; Sterner and Teufel, 2025b).

Research Gap Despite advances, LLMs struggle with zero-shot transfer to real-world CSW scenarios (Winata et al., 2023a). Multilingual LLMs often underperform compared to fine-tuned smaller models, showing that "multilingualism" alone does not ensure CSW proficiency (Zhang et al., 2023). LLMs also exhibit asymmetric performance: non-English tokens in English contexts degrade performance, while English tokens in other languages often enhance it (Mohamed et al., 2025), exacerbated by limited pretraining data for low-resource languages (Yoo et al., 2025).

This paper surveys the evolution of CSW research in the LLM era and presents a robust taxonomy (shown in Figure 2) that categorizes it into 5 key categories. The survey traces a transition from pre-LLM statistical and rule-based approaches focused on NLU tasks to LLM-era progress in NLG, speech, and multimodal domains, which show improved fluency but still struggle with factual consistency and low-resource performance. As summarized in Appendix §Table 1, specialized models, instruction tuning, and scalable architectures show notable gains, yet major gaps remain in low-resource languages, unseen code-mix patterns, and generation consistency. Future advances require diverse, semi-annotated datasets and CSW-aware evaluation metrics to better capture linguistic blending and contextual coherence.

Contributions The key contributions include:

- We provide the first comprehensive survey of CSW research in the LLM era, covering NLU and NLG tasks, datasets and benchmarks across diverse language pairs, real-world applications, and key architectural innovations.
- This study reviews approximately 308 papers across five major CSW research areas, covering 12 core NLP tasks, analyzing over 30 datasets and benchmarks spanning more than 80 languages, and presenting a taxonomy (Figure 2) that classifies LLMs by architecture, training paradigm, and evaluations.
- We identify key gaps in LLM-based CSW research, including limited low-resource languages support, script diversity challenges, weak cross-lingual transfer, and the absence of unified evaluation frameworks. Future progress depends on inclusive datasets, equitable models, and fair metrics.

2 Pre-LLM-Era works

Early computational approaches to code-switched word processing relied on rule-based and statistical models for foundational tasks such as language identification (LID) (Molina et al., 2016; Gundapu and Mamidi, 2018; Shekhar et al., 2020; Solorio and Liu, 2008; Chittaranjan et al., 2014; King et al., 2014), part-of-speech (POS) tagging (Vyas et al., 2014; Raha et al., 2019; Pratapa et al., 2018b; Sequiera et al., 2015), named entity recognition (NER) (Ansari et al., 2018; Singh et al., 2018b,a), and sentiment analysis (SA) (Patwa et al., 2020; Joshi et al., 2016). Methods included CRFs and CNNs for LID (Solorio and Liu, 2008; Chittaranjan et al., 2014), CNNs with n-grams for POS tagging (Vyas et al., 2014), character-level RNNs and SVMs for NER (Singh et al., 2018b), and SVM-based sentiment classification (Joshi et al., 2016). BiLSTM-CRF models with embeddings later improved LID and NER, reducing perplexity (Chopra et al., 2021; Zhang et al., 2023), while switch-point sampling

enhanced Hinglish LID performance (Chatterjere et al., 2020). However, these approaches were limited by task and language-specific designs, shallow features, scarce labeled data, and poor cross-linguistic transfer (Molina et al., 2016; Shekhar et al., 2020). The fragmented nature of research led to isolated solutions, preventing the use of shared representations or unified frameworks across diverse CSW contexts (Winata et al., 2023a; Liu et al., 2022; Chi, 2025). The rise of LLMs has shifted CSW research toward unified, multilingual frameworks, motivating surveys to examine their adaptation, emerging trends, and future directions.

3 Popular CSW Tasks

3.1 Natural Language Understanding Tasks

Language Identification Word & sentence-level LID form the foundation of CSW NLP. Transformer-based models substantially improve LID accuracy, with TongueSwitcher enhancing boundary detection in German–English CSW and outperforming traditional baselines like FastText and CLD (Lambebo Tonja et al., 2022; Sterner and Teufel, 2023). MaskLID identifies subdominant language patterns, while integrating Equivalence Constraint Theory with transformers enables detection of grammatically valid switch points (Kargaran et al., 2024; Kuwanto et al., 2024). Beyond pure LID, related efforts extend to hope and offensive speech detection—addressing multilingual and adversarial switching challenges—using datasets in Roman-Urdu and Malayalam-English (Ahmad et al., 2025; Balouchzahi et al., 2021b), as well as models like COOLI and SetFit are for efficient zero and few-shot approaches (Balouchzahi et al., 2021a; Pannerselvam et al., 2024).

Part-of-Speech Tagging LLM-based approaches improve CSW POS tagging accuracy via pre-trained models, multi-task learning, and transfer strategies. mBERT embeddings capture Arabic–English CSW nuances, and character-CNNs with mBERT further boost Hinglish POS (Sabty et al., 2020; Aguilar and Solorio, 2020). Joint LID–POS transformer models and switch-point–biased self-training improve accuracy on Telugu–English data (Dowlagar and Mamidi, 2021c; Chopra et al., 2021). Bilingual pretraining on GLUECoS enhances POS (Winata et al., 2021; Prasad et al., 2021). PACMAN supports fine-tuning, while PRO-CS and CoMix leverages prompts for improved POS tagging and generation of code-switched text (Chatterjee et al., 2022; Kumar et al., 2022b; Bansal et al., 2022; Arora et al., 2023). XLM-R fine-tuned with the S-index achieves state-of-the-art on Hinglish POS improves performance (Absar, 2025).

Named Entity Recognition NER in CSW contexts has evolved significantly, from using embedding attention for Spanglish tweets (Wang et al., 2018), to multilingual frameworks like MELM that leverage synthetic CSW pretraining (Zhou et al., 2022). Two-stage CMB models improve Hinglish NER by separating span detection and classification (Pu et al., 2022). Benchmarks such as MultiCoNER and toolkits like CodemixedNLP support multilingual and Hinglish NER (Malmasi et al., 2022b; Jayanthi et al., 2021). Contextualized embeddings aid Arabic–English NER (Sabty et al., 2020), while pseudo-labeling and switch-point–biased self-training enhance boundary detection (El Mekki et al., 2022; Chopra et al., 2021). Prompt-based models like PRO-CS and data augmentation in CoSDA-ML enable zero-shot transfer (Bansal et al., 2022; Qin et al., 2020). Generative frameworks such as GPT-NER achieve strong few-shot performance but suffer from hallucinations (Wang et al., 2025a), whereas GLiNER outperforms ChatGPT in zero-shot settings (Zaratiana et al., 2024). Persistent challenges in NER include limited data, complex entities, and trilingual ambiguity (Dowlagar and Mamidi, 2022; Malmasi et al., 2022b).

Takeaway Although notable progress has been made in core CSW NLU tasks, many frontier areas remain underexplored, highlighting opportunities to extend research beyond existing linguistic and computational paradigms.

3.2 Natural Language Generation Tasks

Code-Mixed Text Generation Semi-supervised transfer learning and translation models improved Hinglish fluency (Gupta et al., 2020; Tarunesh et al., 2021), while COCOA achieved strong English–Spanish generation (Mondal et al., 2022). Dependency tree methods produced CSW text without parallel corpora (Gregorius and Okadome, 2022). Synthetic filtering and LLM prompting enhanced Tagalog–English generation (Sravani and Mamidi, 2023; Yong et al., 2023; Terblanche et al., 2024), and LLMs aided grammatical correction (Potter and Yuan, 2024; Heredia et al.,

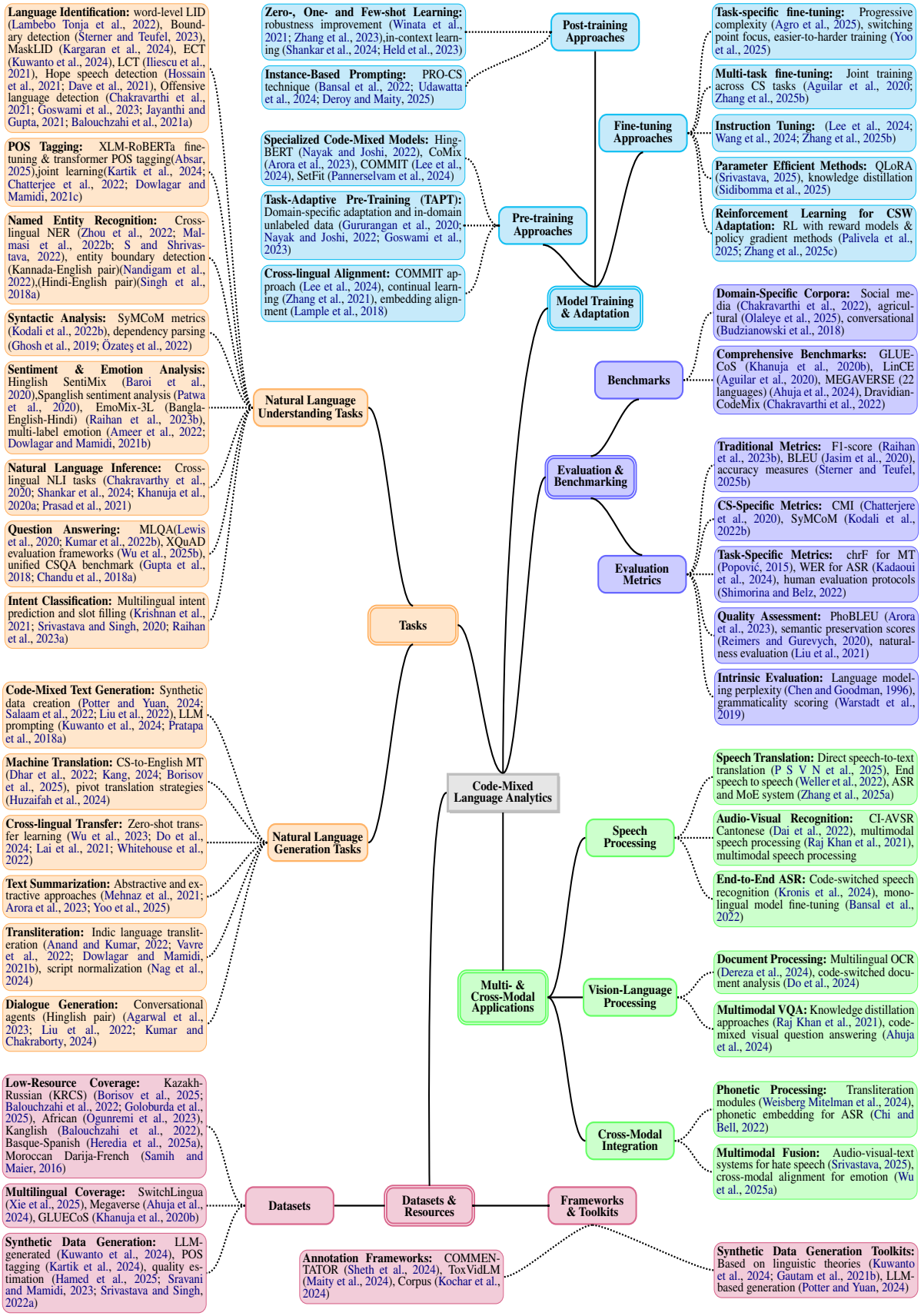

Figure 2: A taxonomy of the code-switching research landscape. ***Takeaway*** *The mind map highlights important works across the diverse categories of code-switching research.*

2025b). However, EZSwitch and HinglishEval revealed weak alignment between metrics and human judgments (Kuwanto et al., 2024; Srivastava and Singh, 2022a).

Machine Translation Back-translation improved Hinglish MT, while unsupervised approaches enhanced Sinhala–English (Tarunesh et al., 2021; Kugathasan and Sumathipala, 2021). COMET-based synthetic data boosted Indic translation (Gupta et al., 2021a; Gautam et al., 2021b). Gated seq2seq with language tags and fine-tuned mT5/mBART advanced Hinglish and MSA–Egyptian MT (Dowlagar and Mamidi, 2021a; Jawahar et al., 2021a; Nagoudi et al., 2021). LLM syntactic post-processing improved Cantonese–Mandarin translation (Dai et al., 2025), and GPT-3 prompting enhanced Hinglish fluency (Khatri et al., 2023). Fine-tuned T5 achieved strong CodeMix-to-English results (Chatterjee et al., 2023; Jawahar et al., 2021b; Raj Khan et al., 2021). Persistent challenges in MT include syntactic misalignment and LLM limitations (Winata et al., 2021; Sazzed, 2021).

Takeaway The LLM era has advanced NLG tasks by enabling generation and conversational capabilities that pre-LLM approaches could not achieve, though progress remains largely confined to high-resource pairs, with English-centric biases, limited low-resource corpora, typological diversity, and syntactic complexity still hindering generalization.

A detailed discussion of remaining NLU and NLG tasks is provided in Appendix §A.1 and §A.2, while the corresponding datasets and approaches for NLP tasks are summarized in §A.5 Table 2.

4 Datasets and Resources

4.1 Datasets

Multilingual Coverage Large-scale code-switching corpora, pre-trained on mixed-language text (e.g., millions of tokens from diverse sources), enhance NLU transfer via synthetic augmentation (Zhang et al., 2024a). The Multilingual Identification of English Code-Switching dataset, manually annotated token-level classifiers on high-resource pairs, benchmarks English switches across unseen languages (Sterner, 2024). SwitchLingua, a manually curated multi-ethnic collection, spans 420k textual samples and >80 hours audio across 12 languages/63 ethnic groups, reducing bias via LLM-assisted balancing (Xie et al., 2025). MEGAVERSE, an LLM-driven synthetic/manual benchmark, covers 22 datasets in 83 languages with translation/alignment for multimodal evaluation (Ahuja et al., 2024). MultiCoNER, manually annotated across 3 domains/11 languages (expanded to 12 in v2 with 33 entity classes), uses LLM synthetic augmentation for code-mixed NER (Malmasi et al., 2022b). NusaX, human-annotated parallel sentiment corpus, provides data for 10 Indonesian local languages to support indirect CSW analysis (Winata et al., 2023b). GLOSS, built on pre-trained MT models with a code-switching module, synthesizes CSW texts for absent language pairs, enabling generalization without manual curation (Hsu et al., 2023). This LLM-centric shift enables scalable, inclusive benchmarks, though typological gaps persist.

Low Resource Coverage Low-resource datasets includes BnSentMix, a manually annotated Bengali-English dataset (20,000 samples with 4 sentiment labels) (Alam et al., 2025) fine-tuned on XLM-R, DravidianCodeMix - manually annotated for Tamil-English (SA/LID: 64,773/64,991 samples), Kannada-English (SA/LID: 30,872 / 29,601 samples), and Malayalam-English (SA/LID: 9,751 / 8,983 samples), supports SA and offensive LID (Chakravarthi et al., 2022), COMMIT uses LLM-based instruction tuning with synthetic CSW data, blocking English and using five low-resourced languages: Greek, Hindi, Thai, Tamil, and Bengali, aligning via causal language modeling, for QA (Lee et al., 2024), Marathi–English corpora manually annotated with pre-trained models, boosts NER and SA tasks(Joshi et al., 2023), SentMix (Bangla-English-Hindi trilingual for NLI) (Raihan et al., 2023a), and GPT-3.5 synthetic Afrikaans/Yoruba–English data (Terblanche et al., 2024) improves zero-shot sentiment accuracy. These efforts diversify CSW NLP by reducing dependence on high-resource language pairs.

Synthetic Data Generation and Augmentation The scarcity of annotated code-mixed datasets has driven synthetic data generation to enhance multilingual NLP. A dependency parser for Bengali-English, augmented with a synthetic treebank (270,531 sentences), improved parsing by UAS (Winata et al., 2019). PhraseOut substitutes high-confidence phrases for multilingual NMT,

on Hindi-English (Jasim et al., 2020). A semi-supervised approach with pre-trained encoders generated Hindi-English CSW text, (Gupta et al., 2020). CoSDA-ML's synthetic multilingual data fine-tuned on mBERT, improves zero-shot NLI across 19 languages (Qin et al., 2020). mBERT-based augmentation with ternary sequence labeling enhanced Hindi-English MT (Gupta et al., 2021a). VACS, a variational autoencoder, generated synthetic CSW text achieves reduced perplexity (Samanta et al., 2019). LLM-based generation for Hindi-English puns and sentiment (49,560 synthetic corpus in Spanish-English & Malayalam-English) improved BLEU score (Zeng, 2024; Sarrof, 2025). In-Context Mixing (ICM) prompts enhanced intent classification (Shankar et al., 2024). SynCS datasets (15,500 parallel sentences) aligned representations, boosting zero-shot performance % (Wang et al., 2025b). A naturalistic CSW dataset with parallel translations improved PLM evaluation (Leon et al., 2024). However, human judgments reveal that synthetically generated CSW sentences often lack naturalness, with only 60-65% acceptability compared to human-authored text (Kodali et al., 2025).

Takeaway LLM-generated datasets expand low-resource coverage but often lack naturalness and cultural nuance. Semi-automated, human-in-the-loop annotation can yield more authentic and scalable resources. (See Table 6 in the Appendix §A.5) for the list of datasets covered).

4.2 Frameworks and Toolkits

Annotation Frameworks High-quality annotations are crucial for effective supervised models in code-mixed NLP. CoSSAT enables speech annotation (Shah et al., 2019), while COMMENTATOR integrates LLMs as an API for robust text annotation to support prediction alongside annotation. (Sheth et al., 2024). CHAI uses RL from AI feedback to enhance CSW translation annotations (Zhang et al., 2025c) and multimodal frameworks like ToxVidLM (Maity et al., 2024) manage complex video toxicity detection in CSW.

Synthetic Data Generation Toolkits To address persistent data scarcity, toolkits such as GCM (Rizvi et al., 2021), utilized by Huzaifah et al. (2024) for generating valid CSW text, and CodemixedNLP, an open-source library designed for creating synthetic CSW texts across seven Hinglish tasks (Jayanthi et al., 2021) offer robust solutions for enhancing CSW text generation. These tools facilitate large-scale corpus creation for tasks such as MT (Sravani and Mamidi, 2023) and SA (Zeng, 2024).

Takeaway Synthetic data generation for annotation workflows accelerates CSW research, but the scarcity of multilingual tools limits impact, highlighting the need for adaptable, low-resource–focused toolkits to build natural and useful benchmarks for LLM training and evaluation.

5 Model Training & Adaptation

5.1 Pre-training Approaches

Specialized Code-Mixed Models HingBERT built on L3Cube-HingCorpus outperforms mBERT on GLUECoS (Nayak and Joshi, 2022), while code-mixed embeddings improve hate speech detection over monolingual ones (Pratapa et al., 2018b; Banerjee et al., 2020). HingBERT variants enhance NLP tasks in zero-shot settings (Shirke et al., 2025; Patil et al., 2023), and natural CSW data outperforms synthetic alternatives (Santy et al., 2021). For multimodal tasks, CM-CLIP surpasses baselines by 5–10% in image-text retrieval (Kumari et al., 2024), and CMLFormer's dual-decoder improves CSW modeling (Baral et al., 2025; Suresh et al., 2024), with hierarchical transformers enhancing long-sequence handling (Suresh et al., 2024).

Task-Adaptive Pre-Training CoMix boosts translation by 12.98 BLEU (Arora et al., 2023), boundary-aware MLM improves QA/SA across Hindi-English, Spanish-English, Tamil-English, and Malayalam-English (Das et al., 2023), and alignment-based pre-training yields +7.32% SA, +0.76% NER, +1.9% QA for Hinglish (Fazili and Jyothi, 2022; Prasad et al., 2021). Multilingual augmentation enhances intent classification (Krishnan et al., 2021), and SynCS achieves +10.14 points in Chinese language while outperforming 20x equivalent monolingual data (Wang et al., 2025b).

Cross-lingual Alignment CoSwitchMap improves bilingual lexicon induction (Gaschi et al., 2023), synthetic CSW achieves 5.1 MRR@10 for cross-lingual retrieval (Litschko et al., 2023), and CMLFormer captures language transitions better (Baral et al., 2025). MVMLT enhances NER alignment (Lai et al., 2021), AIMLT improves intent detection by 4–6% (Zhu et al., 2023; Micallef

et al., 2024), and X-CIT fine-tunes Llama-2-7B across five languages (Wu et al., 2025b).

Takeaway Pretraining approaches have improved CSW performance, with specialized architectures like CM-CLIP and HingBERT enabling better handling of multilingual and multimodal inputs, though synthetic data may still miss nuances.

5.2 Mainstream Fine-tuning Approaches

Task-specific fine-tuning Fine-tuning transformers achieve state-of-the-art performance on Kannada-English LID (Lambebo Tonja et al., 2022), fine-tuned XLM-RoBERTa for POS tagging introduces the S-index for measuring switching intensity and demonstrating effective generalization (Absar, 2025). In SA and dialogue, transformer and multi-task models improve robustness on noisy social media data (Palomino and Ochoa-Luna, 2020; Wu et al., 2020). MT also benefits from fine-tuned mBART, mT5, or custom models, enhancing fluency for Hinglish and related languages (Chatterjee et al., 2023; Khan et al., 2022). For Cantonese–Mandarin speech translation, a three-component fine-tuning strategy improves BLEU scores (Dai et al., 2025), while Yoruba-English ASR fine-tuning with FastConformer Transducer reduces WER from 62.94% to 14.81% using 13× fewer parameters than Whisper Large v3 (Babatunde et al., 2025).

Multi-task fine-tuning Intermediate-task mBERT fine-tuning improves NLI, QA, and SA on GLUECoS by 2–3% for Hinglish and Spanish-English (Prasad et al., 2021). Joint training boosts offensive LID on Hinglish tweets and enhances NER and code-switch detection in Algerian-Arabic (Prasad et al., 2021; Amazouz et al., 2017), multi-directional fine-tuning raises Hinglish-English translation BLEU by 40% (Kartik et al., 2024), and modular approaches like AdapterFusion, prompt tuning, and ContrastiveMix further improve zero-shot IR and transfer (Rathnayake et al., 2024; Do et al., 2024).

Takeaway Fine-tuning LLMs on CSW corpora boosts NLP task performance via shared representations, curriculum learning, and synthetic data, but models still struggle with unseen switches, low-resource languages, syntactic variation, and spontaneous CSW.

A detailed discussion of remaining Fine-tuning approaches is provided in Appendix §A.5.

5.3 Post-training Approaches

Zero-, One- and Few-shot Learning Zero-shot methods include prompting LLMs like GPT-3.5 for CSW generation in Southeast Asian languages (Yong et al., 2023), EntityCS's entity-level CSW with Wikidata for cross-lingual slot filling (10% gain) (Chen et al., 2022), MulZDG's framework for dialogue generation (Liu et al., 2022), MALM's augmented pre-training for MT (Gupta, 2022), saliency-based multi-view mixed language training for classification (Lai et al., 2021), multilingual CSW augmentation for intent/slot filling (4.2% accuracy gain) (Krishnan et al., 2021), and CoSDA-ML's data augmentation for cross-lingual NLP (0.70 score) (Qin et al., 2020); however, LLMs like GPT-4 show 14-point accuracy drops in zero-shot CSW tasks, with pretraining language choice impacting performance significantly in such settings (Zhang et al., 2023; Tatariya et al., 2023). One and Few-shot methods include adapted prompting with ChatGPT for low-resource understanding via semantic similarity-based demonstration selection (Tahery and Farzi, 2025), RAG-based in-context learning (1-8 examples) for hate speech detection (63.03 μ-F1 with GPT-4o) (Srivastava, 2025), multi-task LLM fine-tuning for harmful content in memes (Indira Kumar et al., 2025), generative transformers for emotion detection in Bangla-English-Hindi (Goswami et al., 2023), and translation/LLM classification for affective tasks (Yadav et al., 2025); one-shot is typically grouped with few-shot, as in RAG-retrieved examples boosting F1 by 20 points (Srivastava, 2025). These methods rely on prompting, data augmentation, and entity-centric switching, but unseen language pairs and low-resource scenarios remain challenging.

Instance-based Prompting Instance-based prompting in CSW tasks enhances LLM performance, with PRO-CS using mBERT with Hinglish prompts and improving NER and POS tagging by 10–15% (Bansal et al., 2022). GLOSS synthesizes CSW text for unseen pairs and achieving 55% BLEU/METEOR gains through self-training (Hsu et al., 2023). DweshVaani's RAG retrieves 1–8 Hinglish examples, boosting hate speech detection to 63.03 score on Gemma-2 (Srivastava, 2025). In-Context Mixing improves intent classification by 5–8% on MultiATIS++ (Shankar et al., 2024). Despite these gains, informal CSW and low-resource language pairs

remain challenging (Bansal et al., 2022; Shankar et al., 2024).

Takeaway Post-training methods enable LLMs to refine their knowledge, improve reasoning, enhance factual accuracy, and align more effectively with user intents and ethical considerations in CSW context.

6 Evaluation & Benchmarking

6.1 Benchmarks

Domain-Specific & Comprehensive Benchmarks CodeMixBench, with 5k+ Hinglish, Spanglish, and Chinese Pinyin–English prompts, shows 5–10% drops over English-only tasks using fine-tuned CodeLLaMA (Sheokand et al., 2025), while MEGAVERSE spans 22 datasets and 83 languages with LLM translation and LoRA adapters for low-resource QA (Ahuja et al., 2024). Domain-targeted corpora like the Telugu–English medical dialogue dataset (3k dialogs, 29k utterances) leverage XLM-R for intent and slot filling (Dowlagar and Mamidi, 2023), and MultiCoNER covers 11 languages with LLM augmentation and multi-task transformers, improving over mBERT (Malmasi et al., 2022b). SwitchLingua contributes 420k texts and 80+ hours of multilingual audio using LLM-assisted balancing and LoRA fine-tuning (Xie et al., 2025).

Comprehensive benchmarks such as GLUECoS (Khanuja et al., 2020b) and LinCE (Aguilar et al., 2020) enable multi-task LLM evaluation, while CroCoSum (Zhang and Eickhoff, 2024) and DravidianCodeMix (Chakravarthi et al., 2022) provide manually annotated data for summarization, SA, and offensive LID tasks. PACMAN (Chatterjee et al., 2022) and COMI-LINGUA (Sheth et al., 2025) reduce annotation effort through semi-automated methods. These benchmarks combine synthetic generation, semi-automated annotation, and efficient architectures like adapters for robust, scalable CSW evaluation.

Takeaway Evaluation resources are diverse, and while synthetic augmentation can accelerate CSW research, future efforts should focus on multilinguality, multi-modality and generative tasks (See Table 5 in the Appendix A.5 for the list of benchmarks covered).

6.2 Evaluation Metrics

Traditional Metrics Early CSW NLP evaluations used monolingual metrics like F1, Precision, Recall, Accuracy for classification (Qin et al., 2020; El Mekki et al., 2022), and BLEU, ROUGE, METEOR for generative tasks (Agarwal et al., 2021; Gupta et al., 2022), but often miss CSW complexities due to rigid lexical matching (Papineni et al., 2002; Lin, 2004; Hada et al., 2024).

CS-Specific Metrics CSW-specific metrics address these limitations. Code-Mixing Index (CMI) measures word-level mixing (Das and Gambäck, 2013), SyMCoM evaluates grammatical well-formedness (Kodali et al., 2022a), I-Index assesses switch-point integration (Guzmán et al., 2017), and Switch Point-based Metrics analyze intra- and inter-sentential switches (Gambäck and Das, 2016). HinglishEval combines linguistic metrics with embeddings for better quality estimation (Kodali et al., 2022b; Srivastava and Singh, 2022a).

Task-Specific Metrics Semantic-Aware Error Rate (SAER) enhances ASR evaluation (Xie et al., 2025), PhoBLEU handles orthographic variations (Arora et al., 2023), and chrF++ evaluates character n-grams for MT in morphologically rich languages (Popović, 2015). For Quality Assessment, Metrics like SyMCoM and PIER (Kodali et al., 2022a; Ugan et al., 2025) assess syntactic and semantic quality, while Cline emphasizes human-judged naturalness (Kodali et al., 2025).

Intrinsic Evaluation Methods like distinguishing ground-truth from phonetically similar sentences (Chen and Goodman, 1996) and the gold-standard-agnostic GAME metric (Gupta et al., 2024) offer flexible CSW evaluation.

Takeaway CSW evaluation has evolved from monolingual metrics to CS-specific metrics but more linguistically informed, code-mixing aware metrics are needed to assess fairness, safety, and cross-lingual robustness in multilingual models.

7 Multi- & Cross-Modal Applications

7.1 Speech Processing

Speech Translation & End-to-End ASR CSW speech translation has progressed with end-to-end models for English–Spanish (Weller et al., 2022) and real-world streaming Mandarin–English via self-training (Alastruey et al., 2023). Speech segmentation models, such as Whisper fine-tuned on

CoVoSwitch Korean–English CSW data (Kang, 2024), and linguistic-informed data augmentation have improved ASR performance (Chi and Bell, 2022; Hamed et al., 2025). Monolingual fine-tuning outperformed multilingual models for Yoruba–English (Babatunde et al., 2025), while Hindi-Marathi datasets enabled transformer ASR (Hemant and Narvekar, 2025), Wav2Vec2 and GPT-2 enhance transcription accuracy in multilingual indian contexts with CSW pairs and dialectal variations (R et al., 2025), and MoE with SC-LLMs further boosted Mandarin-English ASR in CSW (Zhang et al., 2025a).

Audio-visual Recognition Enhanced Visual features in Mandarin-English and low-resource Yoruba-English ASR (Babatunde et al., 2025), while Hindi-Marathi CS transformers and Wav2Vec2 fusion aided Indian languages like Tamil (Hemant and Narvekar, 2025).

7.2 Vision-Language Processing

Multimodal VQA & Document Processing Multimodal VQA for CSW advanced via knowledge distillation, enables Hinglish systems to handle CS queries (Raj Khan et al., 2021). LLMs detected harmful content in code-mixed memes (Indira Kumar et al., 2025), BanglAssist used RAG for Bengali-English CSW (Kruk et al., 2025), and Kazakh-Russian LLMs evaluated safety with region-specific prompts (Goloburda et al., 2025). LLM agents analyzed Sinhala-English and Tamil-English CSW interviews (Zhao et al., 2025).

Document processing employed multilingual OCR for robust extraction (Dereza et al., 2024) and contrastive learning improved Vietnamese-English analysis (Do et al., 2024).

7.3 Cross-Modal Integration

Phonetic & Multimodal Processing Phonetic modeling enhances multilingual CSW performance through discriminative language modeling (Winata et al., 2019), transliteration for agglutinative languages (Tasawong et al., 2023), and applications such as abusive content detection (Gautam et al., 2021a), translation alignment (Chou et al., 2023), and back-transliteration (Fernando and Ranathunga, 2021). Transformer-based phonetic guidance (Yang and Tu, 2022) and Wav2Vec2–GPT-2 fusion (Perera and Sumanathilaka, 2025) further improve CSW phonetic tasks. **Multimodal fusion** integrates audio, visual, and textual cues, enhancing code-mixed ASR (Tamil, Kannada, Mandarin-English) and toxicity detection in videos (Maity et al., 2024; Perera and Sumanathilaka, 2025; Zhang et al., 2025a).

Takeaway Multimodal, phonetic, and end-to-end approaches are advancing CSW speech, vision, and document processing, but robust integration across low-resource languages and modalities remains a key challenge.

8 Emerging Frontiers in CSW Research

Futuristic Datasets The progress of CSW NLP relies on developing expansive and inclusive datasets but large-scale conversational resources capturing dynamic CSW interactions remain absent, and the lack of human-preference datasets limits adaptive modeling. Multimodal efforts such as MEGAVERSE (Ahuja et al., 2024) are promising but still restricted in linguistic and domain coverage. Building diverse corpora, such as Hindi–Marathi speech (Hemant and Narvekar, 2025) or a large multi-domain multilingual dialogue dataset (Moradshahi et al., 2023), faces persistent challenges of cost and scalability. Future work must integrate synthetic data generation with expert validation to achieve diversity and quality, particularly for underrepresented languages.

Futuristic Architectures Next-generation architectures should integrate CSW text, speech, and vision to improve switch-point detection, contextual understanding, and natural multilingual interactions, while ASR and TTS systems leverage self-supervised encoders, cross-lingual, and emotion-aware conditioning.

Futuristic Evaluation Futuristic CSW-aware evaluation will move beyond isolated metrics to holistic, human-aligned assessment of multilingual competence, jointly measuring switch-point prediction, semantic consistency, and fluency. Adaptive scoring models trained on human preferences will capture nuanced judgments of naturalness and cross-lingual balance.

Futuristic CSW-aware Ethical AI It will prioritize inclusivity and transparency, using bias-aware training for equitable performance across languages and dialects. Fairness metrics, participatory data curation, and transparent documentation will ensure accountability and foster socially and linguistically fair CSW systems.

Futuristic Applications CSW research can drive high-impact applications such as multilingual conversational assistants for accessible public services, cross-lingual education platforms that adapt to learners' language mixes, and digital preservation tools for endangered dialects. These applications are socially and economically significant, reducing language barriers, enhancing equitable access to knowledge and services, and empowering communities in the multilingual digital economy.

9 Conclusion

CSW research has undergone a major transformation with the rise of LLMs. From early rule-based and statistical methods to multilingual pretraining, fine-tuning, and instruction-based adaptation, the field has steadily moved toward more unified and scalable approaches. This survey shows that although significant progress has been made in LID, tagging, translation, and generation for high-resource pairs, major gaps persist for low-resource and culturally diverse contexts. The future of CSW research lies in building inclusive datasets, improving evaluation protocols, and designing socially grounded systems that reflect the realities of multilingual communication.

Limitations

Despite providing a broad survey, this paper has several limitations:

1. **Coverage Bias** The survey highlights widely studied language pairs and might have missed indigenous or minority code-mixed languages.

2. **Evolving Landscape** Given the rapid pace of LLM research, some approaches and benchmarks described may soon be outdated or replaced by newer paradigms.

3. **Evaluation Constraints** The analysis is largely focused on text-based tasks, whereas speech, multimodal, and conversational CSW applications remain underexplored.

4. **Practical deployment** The survey mainly covers academic progress, leaving ethical, computational, and accessibility concerns in real-world deployment less examined.

Ethics Statement

This study involves a review and synthesis of previously published research and publicly available datasets. No human or user data were collected or analyzed. All works included in this survey were cited appropriately to acknowledge original authorship. The review process was conducted with transparency and fairness, avoiding selective reporting or biased interpretations. Our study promotes fairness and inclusivity in multilingual NLP by focusing on underrepresented code-mixed language scenarios, encouraging equitable research attention toward linguistically diverse communities. The study adheres to established ethical standards for research in computational linguistics.

References

Shayaan Absar. 2025. Fine-tuning cross-lingual LLMs for POS tagging in code-switched contexts. In *Proceedings of the Third Workshop on Resources and Representations for Under-Resourced Languages and Domains (RESOURCEFUL-2025)*, pages 7–12, Tallinn, Estonia. University of Tartu Library, Estonia.

Anmol Agarwal, Jigar Gupta, Rahul Goel, Shyam Upadhyay, Pankaj Joshi, and Rengarajan Aravamudhan. 2023. CST5: Data augmentation for code-switched semantic parsing. In *Proceedings of the 1st Workshop on Taming Large Language Models: Controllability in the era of Interactive Assistants!*, pages 1–10, Prague, Czech Republic. Association for Computational Linguistics.

Vibhav Agarwal, Pooja Rao, and Dinesh Babu Jayagopi. 2021. Hinglish to English machine translation using multilingual transformers. In *Proceedings of the Student Research Workshop Associated with RANLP 2021*, pages 16–21, Online. INCOMA Ltd.

Maha Tufail Agro, Atharva Kulkarni, Karima Kadaoui, Zeerak Talat, and Hanan Aldarmaki. 2025. Code-switching in end-to-end automatic speech recognition: A systematic literature review. *Preprint*, arXiv:2507.07741.

Gustavo Aguilar, Sudipta Kar, and Thamar Solorio. 2020. LinCE: A centralized benchmark for linguistic code-switching evaluation. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 1803–1813, Marseille, France. European Language Resources Association.

Gustavo Aguilar and Thamar Solorio. 2020. From English to code-switching: Transfer learning with strong morphological clues. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8033–8044, Online. Association for Computational Linguistics.

Maia Aguirre, Manex Serras, Laura García-sardiña, Jacobo López-fernández, Ariane Méndez, and Arantza Del Pozo. 2022. Exploiting in-domain bilingual corpora for zero-shot transfer learning in NLU of intra-sentential code-switching chatbot interactions. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 138–144, Abu Dhabi, UAE. Association for Computational Linguistics.

Muhammad Ahmad, Muhammad Waqas, Ameer Hamza, Ildar Z. Batyrshin, and Grigori Sidorov. 2025. Hope speech detection in code-mixed roman urdu tweets: A positive turn in natural language processing. *ArXiv*, abs/2506.21583.

Sanchit Ahuja, Divyanshu Aggarwal, Varun Gumma, Ishaan Watts, Ashutosh Sathe, Millicent Ochieng, Rishav Hada, Prachi Jain, Mohamed Ahmed, Kalika Bali, and Sunayana Sitaram. 2024. MEGAVERSE: Benchmarking large language models across languages, modalities, models and tasks. 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 2598–2637, Mexico City, Mexico. Association for Computational Linguistics.

Sanchit Ahuja, Kumar Tanmay, Hardik Hansrajbhai Chauhan, Barun Patra, Kriti Aggarwal, Luciano Del Corro, Arindam Mitra, Tejas Indulal Dhamecha, Ahmed Hassan Awadallah, Monojit Choudhury, Vishrav Chaudhary, and Sunayana Sitaram. 2025. sPhinX: Sample efficient multilingual instruction fine-tuning through n-shot guided prompting. In *Proceedings of the Fourth Workshop on Generation, Evaluation and Metrics (GEM²)*, pages 927–946, Vienna, Austria and virtual meeting. 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.

Belen Alastruey, Matthias Sperber, Christian Gollan, Dominic Telaar, Tim Ng, and Aashish Agarwal. 2023. Towards real-world streaming speech translation for code-switched speech. In *Proceedings of the 6th Workshop on Computational Approaches to Linguistic Code-Switching*, pages 14–22, Singapore. Association for Computational Linguistics.

Djegdjiga Amazouz, Martine Adda-Decker, and Lori Lamel. 2017. Addressing code-switching in french/algerian arabic speech. In *Interspeech 2017*, pages 62–66.

Iqra Ameer, Grigori Sidorov, Helena Gómez-Adorno, and Rao Muhammad Adeel Nawab. 2022. Multi-label emotion classification on code-mixed text: Data and methods. *IEEE Access*, 10:8779–8789.

The AI Guy Anand and Jivitesh Kumar. 2022. IndicTrans: A Python library for Indic language transliteration. 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: System Demonstrations*, pages 68–75, Online. Association for Computational Linguistics.

Jason Angel, Segun Taofeek Aroyehun, Antonio Tamayo, and Alexander Gelbukh. 2020. NLP-CIC at SemEval-2020 task 9: Analysing sentiment in code-switching language using a simple deep-learning classifier. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 957–962, Barcelona (online). International Committee for Computational Linguistics.

Mohd Zeeshan Ansari, Tanvir Ahmad, and Md Arshad Ali. 2018. Cross script hindi english ner corpus from wikipedia. *Preprint*, arXiv:1810.03430.

Abhinav Arora, Akshat Shrivastava, and Lorena Sainz-Maza Lecanda. 2020. Cross-lingual transfer learning for intent detection of covid-19 utterances. In *Proceedings of the 28th International Conference on Computational Linguistics (COLING 2020)*.

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.

Maria Riveena Arul, Vigneshwaran Shanmugasundaram, S Rajalakshmi, Bharathi Raja Chakravarthi, and C. N. Subalalitha. 2025. MMS-5: A multimodal and multi-scenario hate speech dataset for dravidian languages. In *Proceedings of the First Workshop on Low-resource Languages in the Large Language Model Era (LoResLM 2025)*, Vasco da Gama, Goa, India. NLP Association of India (NLPAI).

Oreoluwa Boluwatife Babatunde, Victor Tolulope Olufemi, Emmanuel Bolarinwa, Kausar Yetunde Moshood, and Chris Chinenye Emezue. 2025. Beyond monolingual limits: Fine-tuning monolingual ASR for Yoruba-English code-switching. In *Proceedings of the 7th Workshop on Computational Approaches to Linguistic Code-Switching*, pages 18–25, Albuquerque, New Mexico, USA. Association for Computational Linguistics.

F. Balouchzahi, S. Butt, A. Hegde, N. Ashraf, H.l. Shashirekha, Grigori Sidorov, and Alexander Gelbukh. 2022. Overview of CoLI-kanglish: Word level language identification in code-mixed Kannada-English texts at ICON 2022. In *Proceedings of the 19th International Conference on Natural Language Processing (ICON): Shared Task on Word Level Language Identification in Code-mixed Kannada-English Texts*, pages 38–45, IIIT Delhi, New Delhi, India. Association for Computational Linguistics.

Fazlourrahman Balouchzahi, Aparna B K, and H L Shashirekha. 2021a. MUCS@DravidianLangTech-EACL2021:COOLI-code-mixing offensive language identification. In *Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages*, pages 323–329, Kyiv. Association for Computational Linguistics.

Fazlourrahman Balouchzahi, Aparna B K, and H L Shashirekha. 2021b. MUCS@LT-EDI-EACL2021:CoHope-hope speech detection for equality, diversity, and inclusion in code-mixed texts. In *Proceedings of the First Workshop on Language Technology for Equality, Diversity and Inclusion*, pages 180–187, Kyiv. Association for Computational Linguistics.

Shubhanker Banerjee, Bharathi Raja Chakravarthi, and John P. McCrae. 2020. Comparison of pretrained embeddings to identify hate speech in indian code-mixed text. *2020 2nd International Conference on Advances in Computing, Communication Control and Networking (ICACCCN)*, pages 21–25.

Srijan Bansal, Suraj Tripathi, Sumit Agarwal, Teruko Mitamura, and Eric Nyberg. 2022. PRO-CS : An instance-based prompt composition technique for code-switched tasks. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10243–10255, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Aditeya Baral, Allen George Ajith, Roshan Nayak, and Mrityunjay Abhijeet Bhanja. 2025. Cmlformer: A dual decoder transformer with switching point learning for code-mixed language modeling. *ArXiv*, abs/2505.12587.

Subhra Jyoti Baroi, Nivedita Singh, Ringki Das, and Thoudam Doren Singh. 2020. NITS-Hinglish-SentiMix at SemEval-2020 task 9: Sentiment analysis for code-mixed social media text using an ensemble model. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1298–1303, Barcelona (online). International Committee for Computational Linguistics.

Shabnam Behzad, Amir Zeldes, and Nathan Schneider. 2024. To ask LLMs about English grammaticality, prompt them in a different language. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 15622–15634, Miami, Florida, USA. Association for Computational Linguistics.

Astik Biswas, Febe de Wet, Ewald van der Westhuizen, and Thomas Niesler. 2020. Semi-supervised acoustic and language model training for English-isiZulu code-switched speech recognition. In *Proceedings of the 2020 Conference on Computational Approaches to Linguistic Code-Switching*, pages 61–71, Online. Association for Computational Linguistics.

Maksim Borisov, Zhanibek Kozhirbayev, and Valentin Malykh. 2025. Low-resource machine translation for code-switched Kazakh-Russian language pair. 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 66–76, Albuquerque, USA. Association for Computational Linguistics.

Paweł Budzianowski, Tsung-Hsien Wen, Bo-Hsiang Tseng, Iñigo Casanueva, Stefan Ultes, Osman Ramadan, and Milica Gašić. 2018. MultiWOZ - a large-scale multi-domain Wizard-of-Oz dataset for task-oriented dialogue modelling. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 5016–5026, Brussels, Belgium. Association for Computational Linguistics.

Bharathi Raja Chakravarthi, Ruba Priyadharshini, Navya Jose, Anand Kumar M, Thomas Mandl, Prasanna Kumar Kumaresan, Rahul Ponnusamy, Hariharan R L, John P. McCrae, and Elizabeth Sherly. 2021. Findings of the shared task on offensive language identification in Tamil, Malayalam, and Kannada. In *Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages*, pages 133–145, Kyiv. Association for Computational Linguistics.

Bharathi Raja Chakravarthi, Ruba Priyadharshini, Vigneshwaran Muralidaran, Navya Jose, Shardul Suryawanshi, Elizabeth Sherly, and John P McCrae. 2022. Dravidiancodemix: Sentiment analysis and offensive language identification dataset for dravidian languages in code-mixed text. *Language Resources and Evaluation*, 56(3):765–806.

Sharanya Chakravarthy, Anjana Umapathy, and Alan W Black. 2020. Detecting entailment in code-mixed Hindi-English conversations. In *Proceedings of the Sixth Workshop on Noisy User-generated Text (W-NUT 2020)*, pages 165–170, Online. Association for Computational Linguistics.

C. Chandramouli. 2011. Census of india 2011: Provisional population totals. Technical report, Office of Registrar General and Census Commissioner, Government of India, New Delhi.

Khyathi Chandu, Ekaterina Loginova, Vishal Gupta, Josef van Genabith, Günter Neumann, Manoj Chinnakotla, Eric Nyberg, and Alan W. Black. 2018a. Code-mixed question answering challenge: Crowdsourcing data and techniques. In *Proceedings of the Third Workshop on Computational Approaches to Linguistic Code-Switching*, pages 29–38, Melbourne, Australia. Association for Computational Linguistics.

Khyathi Chandu, Thomas Manzini, Sumeet Singh, and Alan W. Black. 2018b. Language informed modeling of code-switched text. In *Proceedings of the Third Workshop on Computational Approaches to Linguistic Code-Switching*, pages 92–97, Melbourne, Australia. Association for Computational Linguistics.

Arindam Chatterjee, Chhavi Sharma, Ayush Raj, and Asif Ekbal. 2022. PACMAN:PArallel CodeMixed dAta generatioN for POS tagging. In *Proceedings of the 19th International Conference on Natural Language Processing (ICON)*, pages 234–244, New Delhi, India. Association for Computational Linguistics.

Arindam Chatterjee, Chhavi Sharma, Yashwanth V.p., Niraj Kumar, Ayush Raj, and Asif Ekbal. 2023. Lost in translation no more: Fine-tuned transformer-based models for CodeMix to English machine translation. In *Proceedings of the 20th International Conference on Natural Language Processing (ICON)*, pages 326–335, Goa University, Goa, India. NLP Association of India (NLPAI).

Arindam Chatterjere, Vineeth Guptha, Parul Chopra, and Amitava Das. 2020. Minority positive sampling for switching points - an anecdote for the code-mixing language modeling. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 6228–6236, Marseille, France. European Language Resources Association.

Jinyu Chen, Yuxin Wang, Yujia Li, and Shuai Li. 2022. ENTITYCS: Improving cross-lingual semantics with entity-level code-switching. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 1812–1823, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Stanley F. Chen and Joshua T. Goodman. 1996. An empirical study of smoothing techniques for language modeling. *Preprint*, arXiv:cmp-lg/9606011.

Jie Chi. 2025. Understanding and modeling code-switching: metrics, triggers, and applications in multilingual nlp. *Edinburgh Research Archive*.

Jie Chi and Peter Bell. 2022. Improving code-switched ASR with linguistic information. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 7161–7172, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Gokul Chittaranjan, Yogarshi Vyas, Kalika Bali, and Monojit Choudhury. 2014. Word-level language identification using CRF: Code-switching shared task report of MSR India system. In *Proceedings of the First Workshop on Computational Approaches to Code Switching*, pages 73–79, Doha, Qatar. Association for Computational Linguistics.

Parul Chopra, Sai Krishna Rallabandi, Alan W Black, and Khyathi Raghavi Chandu. 2021. Switch point biased self-training: Re-purposing pretrained models for code-switching. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4389–4397, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Tzu Hsuan Chou, Chun-Yi Lin, and Hung-Yu Kao. 2023. Advancing multi-criteria Chinese word segmentation through criterion classification and denoising. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6460–6476, Toronto, Canada. Association for Computational Linguistics.

Helin Cihan, Yunhan Wu, Paola Peña, Justin Edwards, and Benjamin Cowan. 2022. Bilingual by default: Voice assistants and the role of code-switching in creating a bilingual user experience. In *Proceedings of the 4th Conference on Conversational User Interfaces*, CUI '22, New York, NY, USA. Association for Computing Machinery.

Wenliang Dai, Samuel Cahyawijaya, Tiezheng Yu, Elham J. Barezi, Peng Xu, Cheuk Tung Yiu, Rita Frieske, Holy Lovenia, Genta Winata, Qifeng Chen, Xiaojuan Ma, Bertram Shi, and Pascale Fung. 2022. CI-AVSR: A Cantonese audio-visual speech dataset-for in-car command recognition. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 6786–6793, Marseille, France. European Language Resources Association.

Yuqian Dai, Chun Fai Chan, Ying Ki Wong, and Tsz Ho Pun. 2025. Next-level Cantonese-to-Mandarin translation: Fine-tuning and post-processing with LLMs. In *Proceedings of the First Workshop on Language Models for Low-Resource Languages*, pages 427–436, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Amitava Das and Björn Gambäck. 2013. Code-mixing in social media text. *Traitement Automatique des Langues*, 54(3):41–64.

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, Toronto, Canada. Association for Computational Linguistics.

Bhargav Dave, Shripad Bhat, and Prasenjit Majumder. 2021. IRNLP_DAIICT@LT-EDI-EACL2021: Hope speech detection in code mixed text using TF-IDF char n-grams and MuRIL. In *Proceedings of the First Workshop on Language Technology for Equality, Diversity and Inclusion*, pages 114–117, Kyiv. Association for Computational Linguistics.

Oksana Dereza, Deirdre Ní Chonghaile, and Nicholas Wolf. 2024. "to have the 'million' readers yet": Building a digitally enhanced edition of the bilingual Irish-English newspaper an gaodhal (1881-1898). In *Proceedings of the Third Workshop on Language Technologies for Historical and Ancient Languages (LT4HALA) @ LREC-COLING-2024*, pages 65–78, Torino, Italia. ELRA and ICCL.

Aniket Deroy and Subhankar Maity. 2025. Prompt engineering using gpt for word-level code-mixed language identification in low-resource dravidian languages. *Preprint*, arXiv:2411.04025.

Rohan Dhar, Sparsh Kumar, and Akshat Kumar. 2022. Findings of the shared task on machine translation for code-mixed Hinglish text. In *Proceedings of the 19th International Conference on Natural Language Processing (ICON)*, pages 443–449, IIIT Delhi, New Delhi, India. NLP Association of India (NLPAI).

Junggeun Do, Jaeseong Lee, and Seung-won Hwang. 2024. ContrastiveMix: Overcoming code-mixing dilemma in cross-lingual transfer for information retrieval. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 197–204, Mexico City, Mexico. Association for Computational Linguistics.

Suman Dowlagar and Radhika Mamidi. 2021a. Gated convolutional sequence to sequence based learning for English-hingilsh code-switched machine translation. In *Proceedings of the Fifth Workshop on Computational Approaches to Linguistic Code-Switching*, pages 26–30, Online. Association for Computational Linguistics.

Suman Dowlagar and Radhika Mamidi. 2021b. Graph convolutional networks with multi-headed attention for code-mixed sentiment analysis. In *Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages*, pages 65–72, Kyiv. Association for Computational Linguistics.

Suman Dowlagar and Radhika Mamidi. 2021c. A pretrained transformer and CNN model with joint language ID and part-of-speech tagging for code-mixed social-media text. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 367–374, Held Online. INCOMA Ltd.

Suman Dowlagar and Radhika Mamidi. 2022. CMNEROne at SemEval-2022 task 11: Code-mixed named entity recognition by leveraging multilingual data. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 1556–1561, Seattle, United States. Association for Computational Linguistics.

Suman Dowlagar and Radhika Mamidi. 2023. A code-mixed task-oriented dialog dataset for medical domain. *Computer Speech & Language*, 78:101449.

Long Duong, Hadi Afshar, Dominique Estival, Glen Pink, Philip Cohen, and Mark Johnson. 2017. Multilingual semantic parsing and code-switching. In *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*, pages 379–389, Vancouver, Canada. Association for Computational Linguistics.

Abdellah El Mekki, Abdelkader El Mahdaouy, Mohammed Akallouch, Ismail Berrada, and Ahmed Khoumsi. 2022. UM6P-CS at SemEval-2022 task 11: Enhancing multilingual and code-mixed complex named entity recognition via pseudo labels using multilingual transformer. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 1511–1517, Seattle, United States. Association for Computational Linguistics.

Naome Etori and Maria Gini. 2024. RideKE: Leveraging low-resource Twitter user-generated content for sentiment and emotion detection on code-switched RHS dataset. In *Proceedings of the 14th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, pages 234–249, Bangkok, Thailand. Association for Computational Linguistics.

Barah Fazili and Preethi Jyothi. 2022. Aligning multilingual embeddings for improved code-switched natural language understanding. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 4268–4273, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Aloka Fernando and Surangika Ranathunga. 2021. Data augmentation to address out of VocabularyProblem in low resource Sinhala English neural machine translation. In *Proceedings of the 35th Pacific Asia Conference on Language, Information and Computation*, pages 61–70, Shanghai, China. Association for Computational Lingustics.

Björn Gambäck and Amitava Das. 2016. Comparing the level of code-switching in corpora. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1850–1855, Portorož, Slovenia. European Language Resources Association (ELRA).

Felix Gaschi, Ilias El-Baamrani, Barbara Gendron, Parisa Rastin, and Yannick Toussaint. 2023. Code-switching as a cross-lingual training signal: an example with unsupervised bilingual embedding. In *Proceedings of the 3rd Workshop on Multi-lingual Representation Learning (MRL)*, pages 208–217, Singapore. Association for Computational Linguistics.

Devansh Gautam, Kshitij Gupta, and Manish Shrivastava. 2021a. Translate and classify: Improving sequence level classification for English-Hindi code-mixed data. In *Proceedings of the Fifth Workshop on Computational Approaches to Linguistic Code-Switching*, pages 15–25, Online. Association for Computational Linguistics.

Devansh Gautam, Prashant Kodali, Kshitij Gupta, Anmol Goel, Manish Shrivastava, and Ponnurangam Kumaraguru. 2021b. CoMeT: Towards code-mixed translation using parallel monolingual sentences. In *Proceedings of the Fifth Workshop on Computational Approaches to Linguistic Code-Switching*,

pages 47–55, Online. Association for Computational Linguistics.

Urmi Ghosh, Dipti Sharma, and Simran Khanuja. 2019. Dependency parser for Bengali-English code-mixed data enhanced with a synthetic treebank. In *Proceedings of the 18th International Workshop on Treebanks and Linguistic Theories (TLT, SyntaxFest 2019)*, pages 91–99, Paris, France. Association for Computational Linguistics.

Maiya Goloburda, Nurkhan Laiyk, Diana Turmakhan, Yuxia Wang, Mukhammed Togmanov, Jonibek Mansurov, Askhat Sametov, Nurdaulet Mukhituly, Minghan Wang, Daniil Orel, Zain Muhammad Mujahid, Fajri Koto, Timothy Baldwin, and Preslav Nakov. 2025. Qorǵau: Evaluating safety in Kazakh-Russian bilingual contexts. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 9765–9784, Vienna, Austria. Association for Computational Linguistics.

Dhiman Goswami, Md Nishat Raihan, Antara Mahmud, Antonios Anastasopoulos, and Marcos Zampieri. 2023. OffMix-3L: A novel code-mixed test dataset in Bangla-English-Hindi for offensive language identification. In *Proceedings of the 11th International Workshop on Natural Language Processing for Social Media*, pages 21–27, Bali, Indonesia. Association for Computational Linguistics.

Bryan Gregorius and Takeshi Okadome. 2022. Generating code-switched text from monolingual text with dependency tree. In *Australasian Language Technology Association Workshop*.

Sunil Gundapu and Radhika Mamidi. 2018. Word level language identification in English Telugu code mixed data. In *Proceedings of the 32nd Pacific Asia Conference on Language, Information and Computation*, Hong Kong. Association for Computational Linguistics.

Abhirut Gupta, Aditya Vavre, and Sunita Sarawagi. 2021a. Training data augmentation for code-mixed translation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5760–5766, Online. Association for Computational Linguistics.

Abhirut Gupta, Aditya Vavre, and Sunita Sarawagi. 2022. Adapting multilingual models for code-mixed translation. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 7820–7832.

Akshat Gupta, Sargam Menghani, Sai Krishna Rallabandi, and Alan W Black. 2021b. Unsupervised self-training for sentiment analysis of code-switched data. In *Proceedings of the Fifth Workshop on Computational Approaches to Linguistic Code-Switching*, pages 103–112, Online. Association for Computational Linguistics.

Ayushman Gupta, Akhil Bhogal, and Kripabandhu Ghosh. 2024. Multilingual controlled generation and gold-standard-agnostic evaluation of code-mixed sentences. *arXiv preprint arXiv:2410.10580*.

Deepak Gupta, Asif Ekbal, and Pushpak Bhattacharyya. 2020. A semi-supervised approach to generate the code-mixed text using pre-trained encoder and transfer learning. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2267–2280, Online. Association for Computational Linguistics.

Deepak Gupta, Pabitra Lenka, Asif Ekbal, and Pushpak Bhattacharyya. 2018. Uncovering code-mixed challenges: A framework for linguistically driven question generation and neural based question answering. In *Proceedings of the 22nd Conference on Computational Natural Language Learning*, pages 119–130, Brussels, Belgium. Association for Computational Linguistics.

Kshitij Gupta. 2022. MALM: Mixing augmented language modeling for zero-shot machine translation. In *Proceedings of the 2nd International Workshop on Natural Language Processing for Digital Humanities*, pages 53–58, Taipei, Taiwan. Association for Computational Linguistics.

Pranav Gupta, Souvik Bhattacharyya, Niranjan Kumar M, and Billodal Roy. 2025. LexiLogic@CALCS 2025: Predicting preferences in generated code-switched text. In *Proceedings of the 7th Workshop on Computational Approaches to Linguistic Code-Switching*, pages 48–53, Albuquerque, New Mexico, USA. Association for Computational Linguistics.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. *Preprint*, arXiv:2004.10964.

Gualberto A Guzmán, Joseph Ricard, Jacqueline Serigos, Barbara E Bullock, and Almeida Jacqueline Toribio. 2017. Metrics for modeling code-switching across corpora. In *Interspeech*, pages 67–71.

Rishav Hada, Varun Gumma, Adrian de Wynter, Harshita Diddee, Mohamed Ahmed, Monojit Choudhury, Kalika Bali, and Sunayana Sitaram. 2024. Are large language model-based evaluators the solution to scaling up multilingual evaluation? In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 1051–1070, St. Julian's, Malta. Association for Computational Linguistics.

Injy Hamed, Thang Vu, and Nizar Habash. 2025. The impact of code-switched synthetic data quality is task dependent: Insights from MT and ASR. In *Proceedings of the 7th Workshop on Computational Approaches to Linguistic Code-Switching*, pages 6–17, Albuquerque, New Mexico, USA. Association for Computational Linguistics.

Adeep Hande, Ruba Priyadharshini, Anbukkarasi Sampath, Kingston Pal Thamburaj, Prabakaran Chandran, and Bharathi Raja Chakravarthi. 2021. Hope speech detection in under-resourced kannada language. *Preprint*, arXiv:2108.04616.

William Held, Christopher Hidey, Fei Liu, Eric Zhu, Rahul Goel, Diyi Yang, and Rushin Shah. 2023. DAMP: Doubly aligned multilingual parser for task-oriented dialogue. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3586–3604, Toronto, Canada. Association for Computational Linguistics.

P. Hemant and Meera Narvekar. 2025. Development of a code-switched hindi-marathi dataset and transformer-based architecture for enhanced speech recognition using dynamic switching algorithms. *Applied Acoustics*, 230:110408.

Maite Heredia, Jeremy Barnes, and Aitor Soroa. 2025a. EuskañolDS: A naturally sourced corpus for Basque-Spanish code-switching. In *Proceedings of the 7th Workshop on Computational Approaches to Linguistic Code-Switching*, pages 1–5, Albuquerque, New Mexico, USA. Association for Computational Linguistics.

Maite Heredia, Gorka Labaka, Jeremy Barnes, and Aitor Soroa. 2025b. Conditioning llms to generate code-switched text. *Preprint*, arXiv:2502.12924.

Seongtae Hong, Seungyoon Lee, Hyeonseok Moon, and Heuiseok Lim. 2025. MIGRATE: Cross-lingual adaptation of domain-specific LLMs through code-switching and embedding transfer. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 9184–9193, Abu Dhabi, UAE. Association for Computational Linguistics.

Eftekhar Hossain, Omar Sharif, and Mohammed Moshiul Hoque. 2021. NLP-CUET@LT-EDI-EACL2021: Multilingual Code-Mixed Hope Speech Detection using cross-lingual representation learner. In *Proceedings of the First Workshop on Language Technology for Equality, Diversity and Inclusion*, pages 168–174, Kyiv, Ukraine. Association for Computational Linguistics.

I-Hung Hsu, Avik Ray, Shubham Garg, Nanyun Peng, and Jing Huang. 2023. Code-switched text synthesis in unseen language pairs. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 5137–5151, Toronto, Canada. Association for Computational Linguistics.

Jing Huang and Diyi Yang. 2023. Culturally aware natural language inference. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 7591–7609, Singapore. Association for Computational Linguistics.

Chia-Chien Hung, Anne Lauscher, Ivan Vulić, Simone Ponzetto, and Goran Glavaš. 2022. Multi2WOZ: A robust multilingual dataset and conversational pretraining for task-oriented dialog. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3687–3703, Seattle, United States. Association for Computational Linguistics.

Muhammad Huzaifah, Weihua Zheng, Nattapol Chanpaisit, and Kui Wu. 2024. Evaluating code-switching translation with large language models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 6381–6394, Torino, Italia. ELRA and ICCL.

Comfort Ilevbare, Jesujoba Alabi, David Ifeoluwa Adelani, Firdous Bakare, Oluwatoyin Abiola, and Oluwaseyi Adeyemo. 2024. EkoHate: Abusive language and hate speech detection for code-switched political discussions on Nigerian Twitter. In *Proceedings of the 8th Workshop on Online Abuse and Harms (WOAH 2024)*, pages 28–37, Mexico City, Mexico. Association for Computational Linguistics.

Dana-Maria Iliescu, Rasmus Grand, Sara Qirko, and Rob van der Goot. 2021. Much gracias: Semi-supervised code-switch detection for Spanish-English: How far can we get? In *Proceedings of the Fifth Workshop on Computational Approaches to Linguistic Code-Switching*, pages 65–71, Online. Association for Computational Linguistics.

Joseph Marvin Imperial and Ethel Ong. 2022. Tweet-Taglish: A dataset of Tagalog-English code-switched tweets. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 5183–5192, Marseille, France. European Language Resources Association.

A.K. Indira Kumar, Gayathri Sthanusubramoniani, Deepa Gupta, Aarathi Rajagopalan Nair, Yousef Ajami Alotaibi, and Mohammed Zakariah. 2025. Multi-task detection of harmful content in code-mixed meme captions using large language models with zero-shot, few-shot, and fine-tuning approaches. *Egyptian Informatics Journal*, 30:100683.

Yash Ingle and Pruthwik Mishra. 2025. Ilid: Native script language identification for indian languages. *arXiv preprint arXiv:2507.11832*.

Saddam H.M. Issa, Fatima Amer Aldakhil, Amani Abdullah BinJwair, and Nizaam Kariem. 2025. Delving into bilingual dialogue: The realm of code switching and mixing in arabic-english societies. *Journal of Language Teaching and Research*.

Binu Jasim, Vinay Namboodiri, and C V Jawahar. 2020. PhraseOut: A code mixed data augmentation method for MultilingualNeural machine tranlsation. In *Proceedings of the 17th International Conference on Natural Language Processing (ICON)*, pages 470–474, Indian Institute of Technology Patna, Patna, India. NLP Association of India (NLPAI).

Ganesh Jawahar, El Moatez Billah Nagoudi, Muhammad Abdul-Mageed, and Laks Lakshmanan, V.S. 2021a. Exploring text-to-text transformers for English to Hinglish machine translation with synthetic code-mixing. In *Proceedings of the Fifth Workshop on Computational Approaches to Linguistic Code-Switching*, pages 36–46, Online. Association for Computational Linguistics.

Ganesh Jawahar, El Moatez Billah Nagoudi, Muhammad Abdul-Mageed, and Laks Lakshmanan, V.S. 2021b. Exploring text-to-text transformers for English to Hinglish machine translation with synthetic code-mixing. In *Proceedings of the Fifth Workshop on Computational Approaches to Linguistic Code-Switching*, pages 36–46, Online. Association for Computational Linguistics.

Sai Muralidhar Jayanthi and Akshat Gupta. 2021. SJ_AJ@DravidianLangTech-EACL2021: Task-adaptive pre-training of multilingual BERT models for offensive language identification. In *Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages*, pages 307–312, Kyiv. Association for Computational Linguistics.

Sai Muralidhar Jayanthi, Kavya Nerella, Khyathi Raghavi Chandu, and Alan W Black. 2021. CodemixedNLP: An extensible and open NLP toolkit for code-mixing. In *Proceedings of the Fifth Workshop on Computational Approaches to Linguistic Code-Switching*, pages 113–118, Online. Association for Computational Linguistics.

Noor Jehan, Tabassum Javed, and Shahida Banu. 2025. The evolution of code-switching in multilingual societies: A sociolinguistic perspective: https://doi. org/10.55966/assaj. 2025.4. 1.054. *ASSAJ*, 4(01):614–625.

Aditya Joshi, Ameya Prabhu, Manish Shrivastava, and Vasudeva Varma. 2016. Towards sub-word level compositions for sentiment analysis of Hindi-English code mixed text. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 2482–2491, Osaka, Japan. The COLING 2016 Organizing Committee.

Atharva Joshi, Salil Deshpande, Manali Bapat, Mrinal Kulkarni, Gowri B, Mitesh M. Khapra, and Anoop Kumar. 2023. My boli: A comprehensive suite of corpora and pre-trained models for marathi-english code-mixing. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2990–3004, Dubrovnik, Croatia. Association for Computational Linguistics.

Karima Kadaoui, Maryam Al Ali, Hawau Olamide Toyin, Ibrahim Mohammed, and Hanan Aldarmaki. 2024. PolyWER: A holistic evaluation framework for code-switched speech recognition. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 6144–6153, Miami, Florida, USA. Association for Computational Linguistics.

Yeeun Kang. 2024. CoVoSwitch: Machine translation of synthetic code-switched text based on intonation units. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop)*, pages 345–357, Bangkok, Thailand. Association for Computational Linguistics.

Amir Hossein Kargaran, François Yvon, and Hinrich Schuetze. 2024. MaskLID: Code-switching language identification through iterative masking. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 459–469, Bangkok, Thailand. Association for Computational Linguistics.

Kartik Kartik, Sanjana Soni, Anoop Kunchukuttan, Tanmoy Chakraborty, and Md. Shad Akhtar. 2024. Synthetic data generation and joint learning for robust code-mixed translation. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 15480–15492, Torino, Italia. ELRA and ICCL.

Olga Kellert, Nemika Tyagi, Muhammad Imran, Nelvin Licona-Guevara, and Carlos Gómez-Rodríguez. 2025. Parsing the switch: Llm-based ud annotation for complex code-switched and low-resource languages. *Preprint*, arXiv:2506.07274.

Abdul Khan, Hrishikesh Kanade, Girish Budhrani, Preet Jhanglani, and Jia Xu. 2022. SIT at MixMT 2022: Fluent translation built on giant pre-trained models. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 1136–1144, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Simran Khanuja, Sandipan Dandapat, Sunayana Sitaram, and Monojit Choudhury. 2020a. A new dataset for natural language inference from code-mixed conversations. In *Proceedings of the 4th Workshop on Computational Approaches to Code Switching*, pages 9–16, Marseille, France. European Language Resources Association.

Simran Khanuja, Sandipan Dandapat, Anirudh Srinivasan, Sunayana Sitaram, and Monojit Choudhury. 2020b. GLUECoS: An evaluation benchmark for code-switched NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3575–3585, Online. Association for Computational Linguistics.

Jyotsana Khatri, Vivek Srivastava, and Lovekesh Vig. 2023. Can you translate for me? code-switched machine translation with large language models. 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 2:*

Short Papers), pages 83–92, Nusa Dua, Bali. Association for Computational Linguistics.

Levi King, Eric Baucom, Timur Gilmanov, Sandra Kübler, Dan Whyatt, Wolfgang Maier, and Paul Rodrigues. 2014. The IUCL+ system: Word-level language identification via extended Markov models. In *Proceedings of the First Workshop on Computational Approaches to Code Switching*, pages 102–106, Doha, Qatar. Association for Computational Linguistics.

Rodney Kinney, Chloe Anastasiades, Russell Authur, Iz Beltagy, Jonathan Bragg, Alexandra Buraczynski, Isabel Cachola, Stefan Candra, Yoganand Chandrasekhar, Arman Cohan, and 1 others. 2023. The semantic scholar open data platform. *arXiv preprint arXiv:2301.10140*.

Chayan Kochar, Vandan Vasantlal Mujadia, Pruthwik Mishra, and Dipti Misra Sharma. 2024. Towards disfluency annotated corpora for Indian languages. In *Proceedings of the 7th Workshop on Indian Language Data: Resources and Evaluation*, pages 1–10, Torino, Italia. ELRA and ICCL.

Prashant Kodali, Anmol Goel, Likhith Asapu, Vamshi Krishna Bonagiri, Anirudh Govil, Monojit Choudhury, Ponnurangam Kumaraguru, and Manish Shrivastava. 2025. From human judgements to predictive models: Unravelling acceptability in code-mixed sentences. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 24(9):1–31.

Prashant Kodali, Anmol Goel, Monojit Choudhury, Manish Shrivastava, and Ponnurangam Kumaraguru. 2022a. SyMCoM - syntactic measure of code mixing a study of English-Hindi code-mixing. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 472–480, Dublin, Ireland. Association for Computational Linguistics.

Prashant Kodali, Tanmay Sachan, Akshay Goindani, Anmol Goel, Naman Ahuja, Manish Shrivastava, and Ponnurangam Kumaraguru. 2022b. PreCogIIITH at HinglishEval : Leveraging code-mixing metrics & language model embeddings to estimate code-mix quality. In *Proceedings of the 15th International Conference on Natural Language Generation: Generation Challenges*, pages 26–30, Waterville, Maine, USA and virtual meeting. Association for Computational Linguistics.

Delu Kong and Lieve Macken. 2025. Decoding machine translationese in English-Chinese news: LLMs vs. NMTs. In *Proceedings of Machine Translation Summit XX: Volume 1*, pages 99–112, Geneva, Switzerland. European Association for Machine Translation.

Jitin Krishnan, Antonios Anastasopoulos, Hemant Purohit, and Huzefa Rangwala. 2021. Multilingual code-switching for zero-shot cross-lingual intent prediction and slot filling. In *Proceedings of the 1st Workshop on Multilingual Representation Learning*, pages 211–223, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Martins Kronis, Askars Salimbajevs, and Mārcis Pinnis. 2024. Code-mixed text augmentation for Latvian ASR. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 3469–3479, Torino, Italia. ELRA and ICCL.

Francesco Kruk, Savindu Herath, and Prithwiraj Choudhury. 2025. Banglassist: A bengali-english generative ai chatbot for code-switching and dialect-handling in customer service. *Preprint*, arXiv:2503.22283.

Archchana Kugathasan and Sagara Sumathipala. 2021. Neural machine translation for Sinhala-English code-mixed text. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 718–726, Held Online. INCOMA Ltd.

Abhinav Kumar, Parminder Singh, and Om Singh. 2024. Prabhupadavani: A code-mixed speech translation corpus. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 8142–8149, Torino, Italia. ELRA and ICCL.

AK Indira Kumar, Gayathri Sthanusubramoniani, Deepa Gupta, Aarathi Rajagopalan Nair, Yousef Ajami Alotaibi, and Mohammed Zakariah. 2025. Multi-task detection of harmful content in code-mixed meme captions using large language models with zero-shot, few-shot, and fine-tuning approaches. *Egyptian Informatics Journal*, 30:100683.

Akshat Kumar, Sparsh Kumar, and Nandan Kumar. 2022a. Findings of the WILDRE-5 shared task on sentiment analysis of code-mixed texts. In *Proceedings of the 19th International Conference on Natural Language Processing (ICON)*, pages 426–434, IIIT Delhi, New Delhi, India. NLP Association of India (NLPAI).

Shivani Kumar and Tanmoy Chakraborty. 2024. Harmonizing code-mixed conversations: Personality-assisted code-mixed response generation in dialogues. *Preprint*, arXiv:2401.12995.

Vishwajeet Kumar, Rudra Murthy, and Tejas Dhamecha. 2022b. On utilizing constituent language resources to improve downstream tasks in Hinglish. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 3859–3865, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Vishwajeet Kumar, Rudra Murthy, and Tejas Dhamecha. 2022c. On utilizing constituent

language resources to improve downstream tasks in Hinglish. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 3859–3865, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Gitanjali Kumari, Arindam Chatterjee, Ashutosh Bajpai, Asif Ekbal, and Vinutha B. NarayanaMurthy. 2024. Cm_clip: Unveiling code-mixed multimodal learning with cross-lingual clip adaptations. In *ICON*.

Garry Kuwanto, Chaitanya Agarwal, Genta Indra Winata, and Derry Tanti Wijaya. 2024. Linguistics theory meets llm: Code-switched text generation via equivalence constrained large language models. *ArXiv*, abs/2410.22660.

Siyu Lai, Hui Huang, Dong Jing, Yufeng Chen, Jinan Xu, and Jian Liu. 2021. Saliency-based multi-view mixed language training for zero-shot cross-lingual classification. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 599–610, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Atnafu Lambebo Tonja, Mesay Gemeda Yigezu, Olga Kolesnikova, Moein Shahiki Tash, Grigori Sidorov, and Alexander Gelbukh. 2022. Transformer-based model for word level language identification in code-mixed Kannada-English texts. In *Proceedings of the 19th International Conference on Natural Language Processing (ICON): Shared Task on Word Level Language Identification in Code-mixed Kannada-English Texts*, pages 18–24, IIIT Delhi, New Delhi, India. Association for Computational Linguistics.

Guillaume Lample, Alexis Conneau, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2018. Word translation without parallel data. In *International Conference on Learning Representations*.

Frances Adriana Laureano De Leon, Harish Tayyar Madabushi, and Mark Lee. 2024. Code-mixed probes show how pre-trained models generalise on code-switched text. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 3457–3468, Torino, Italia. ELRA and ICCL.

Dohyeon Lee, Jaeseong Lee, Gyewon Lee, Byung-gon Chun, and Seung-won Hwang. 2021. Scopa: Soft code-switching and pairwise alignment for zero-shot cross-lingual transfer. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, pages 3176–3180.

Jaeseong Lee, YeonJoon Jung, and Seung-won Hwang. 2024. COMMIT: Code-mixing English-centric large language model for multilingual instruction tuning. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 3130–3137, Mexico City, Mexico. Association for Computational Linguistics.

Frances Adriana Laureano De Leon, Harish Tayyar Madabushi, and Mark Lee. 2024. Code-mixed probes show how pre-trained models generalise on code-switched text. In *International Conference on Language Resources and Evaluation*.

Patrick Lewis, Barlas Oguz, Ruty Rinott, Sebastian Riedel, and Holger Schwenk. 2020. MLQA: Evaluating cross-lingual extractive question answering. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7315–7330, Online. Association for Computational Linguistics.

Zhuoran Li, Chunming Hu, J. Chen, Zhijun Chen, Xiaohui Guo, and Richong Zhang. 2024. Improving zero-shot cross-lingual transfer via progressive code-switching. *ArXiv*, abs/2406.13361.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Wan-Ting Lin, Peng-Jen Chen, and Yang Liu. 2024. ContrastiveMix: Overcoming code-mixing dilemmas in multilingual spoken language understanding. 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 5984–6000, Mexico City, Mexico. Association for Computational Linguistics.

Robert Litschko, E. Artemova, and Barbara Plank. 2023. Boosting zero-shot cross-lingual retrieval by training on artificially code-switched data. *ArXiv*, abs/2305.05295.

Hexin Liu, Haoyang Zhang, Qiquan Zhang, Xiangyu Zhang, Dongyuan Shi, Eng Siong Chng, and Haizhou Li. 2025. Code-switching speech recognition under the lens: Model-and data-centric perspectives. *arXiv preprint arXiv:2509.24310*.

Ye Liu, Wolfgang Maier, Wolfgang Minker, and Stefan Ultes. 2021. Naturalness evaluation of natural language generation in task-oriented dialogues using BERT. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 839–845, Held Online. INCOMA Ltd.

Yongkang Liu, Shi Feng, Daling Wang, and Yifei Zhang. 2022. MulZDG: Multilingual code-switching framework for zero-shot dialogue generation. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 648–659, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Holy Lovenia, Samuel Cahyawijaya, Genta Indra Winata, Peng Xu, and Pascale Fung. 2022. ASCEND: A spontaneous conversational EN-Mandarin Dataset for code-switching. In *Proceedings of the 2022 Conference of the North American Chapter*

of the Association for Computational Linguistics: Human Language Technologies, pages 4645–4661, Seattle, United States. Association for Computational Linguistics.

Dau-Cheng Lyu, Tien-Ping Tan, Eng Siong Chng, and Haizhou Li. 2010. Seame: a mandarin-english code-switching speech corpus in south-east asia. In *Interspeech 2010*, pages 1986–1989.

Yili Ma, Liang Zhao, and Jie Hao. 2020. XLP at SemEval-2020 task 9: Cross-lingual models with focal loss for sentiment analysis of code-mixing language. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 975–980, Barcelona (online). International Committee for Computational Linguistics.

Krishanu Maity, A.S. Poornash, Sriparna Saha, and Pushpak Bhattacharyya. 2024. ToxVidLM: A multimodal framework for toxicity detection in code-mixed videos. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 11130–11142, Bangkok, Thailand. Association for Computational Linguistics.

Shervin Malmasi, Anjie Fang, Besnik Fetahu, Sudipta Kar, and Oleg Rokhlenko. 2022a. SemEval-2022 task 11: Multilingual complex named entity recognition (MultiCoNER). In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 1412–1437, Seattle, United States. Association for Computational Linguistics.

Shervin Malmasi, Marcos Zampieri, Preslav Nakov, James Glass, and Pascale Fung. 2022b. MultiCoNER: A large-scale multilingual and code-mixed dataset for complex NER. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3788–3804, Seattle, United States. Association for Computational Linguistics.

Laiba Mehnaz, Debanjan Mahata, Rakesh Gosangi, Uma Sushmitha Gunturi, Riya Jain, Gauri Gupta, Amardeep Kumar, Isabelle G. Lee, Anish Acharya, and Rajiv Ratn Shah. 2021. GupShup: Summarizing open-domain code-switched conversations. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6177–6192, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Kurt Micallef, Nizar Habash, Claudia Borg, Fadhl Eryani, and Houda Bouamor. 2024. Cross-lingual transfer from related languages: Treating low-resource Maltese as multilingual code-switching. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1014–1025, St. Julian's, Malta. Association for Computational Linguistics.

Amr Mohamed, Yang Zhang, Michalis Vazirgiannis, and Guokan Shang. 2025. Lost in the mix: Evaluating llm understanding of code-switched text. *arXiv preprint arXiv:2506.14012*.

Giovanni Molina, Fahad AlGhamdi, Mahmoud Ghoneim, Abdelati Hawwari, Nicolas Rey-Villamizar, Mona Diab, and Thamar Solorio. 2016. Overview for the second shared task on language identification in code-switched data. In *Proceedings of the Second Workshop on Computational Approaches to Code Switching*, pages 40–49, Austin, Texas. Association for Computational Linguistics.

Sneha Mondal, Ritika ., Shreya Pathak, Preethi Jyothi, and Aravindan Raghuveer. 2022. CoCoa: An encoder-decoder model for controllable code-switched generation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2466–2479, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Mehrad Moradshahi, Tianhao Shen, Kalika Bali, Monojit Choudhury, Gael de Chalendar, Anmol Goel, Sungkyun Kim, Prashant Kodali, Ponnurangam Kumaraguru, Nasredine Semmar, Sina Semnani, Jiwon Seo, Vivek Seshadri, Manish Shrivastava, Michael Sun, Aditya Yadavalli, Chaobin You, Deyi Xiong, and Monica Lam. 2023. X-RiSAWOZ: High-quality end-to-end multilingual dialogue datasets and few-shot agents. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 2773–2794, Toronto, Canada. Association for Computational Linguistics.

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, Miami, Florida, USA. Association for Computational Linguistics.

El Moatez Billah Nagoudi, AbdelRahim Elmadany, and Muhammad Abdul-Mageed. 2021. Investigating code-mixed Modern Standard Arabic-Egyptian to English machine translation. In *Proceedings of the Fifth Workshop on Computational Approaches to Linguistic Code-Switching*, pages 56–64, Online. Association for Computational Linguistics.

Poojitha Nandigam, Abhinav Appidi Reddy, Manish Shrivastava, Bharathi Raja Chakravarthi, Mujahid Shad, and Tanmoy Chakraborty. 2022. Named entity recognition for code-mixed kannada-english social media data. In *Proceedings of the 19th International Conference on Natural Language Processing (ICON)*, pages 43–49, New Delhi, India. Association for Computational Linguistics.

Ravindra Nayak and Raviraj Joshi. 2022. L3Cube-HingCorpus and HingBERT: A code mixed Hindi-English dataset and BERT language models. In *Proceedings of the WILDRE-6 Workshop within the*

13th Language Resources and Evaluation Conference, pages 7–12, Marseille, France. European Language Resources Association.

Karin Niederreiter and Dagmar Gromann. 2025. Word-level detection of code-mixed hate speech with multilingual domain transfer. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 21093–21104, Vienna, Austria. Association for Computational Linguistics.

Tolulope Ogunremi, Christopher Manning, and Dan Jurafsky. 2023. Multilingual self-supervised speech representations improve the speech recognition of low-resource African languages with codeswitching. In *Proceedings of the 6th Workshop on Computational Approaches to Linguistic Code-Switching*, pages 83–88, Singapore. Association for Computational Linguistics.

Kayode Olaleye, Arturo Oncevay, Mathieu Sibue, Nombuyiselo Zondi, Michelle Terblanche, Sibongile Mapikitla, Richard Lastrucci, Charese Smiley, and Vukosi Marivate. 2025. AfroCS-xs: Creating a compact, high-quality, human-validated code-switched dataset for African languages. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 33391–33410, Vienna, Austria. Association for Computational Linguistics.

Şaziye Betül Özateş, Arzucan Özgür, Tunga Gungor, and Özlem Çetinoğlu. 2022. Improving code-switching dependency parsing with semi-supervised auxiliary tasks. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 1159–1171, Seattle, United States. Association for Computational Linguistics.

Bhavani Shankar P S V N, Preethi Jyothi, and Pushpak Bhattacharyya. 2025. CoSTA: Code-switched speech translation using aligned speech-text interleaving. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 9194–9208, Abu Dhabi, UAE. Association for Computational Linguistics.

Hemant Palivela, Meera Narvekar, David Asirvatham, Shashi Bhushan, Vinay Rishiwal, and Udit Agarwal. 2025. Code-switching asr for low-resource indic languages: A hindi-marathi case study. *IEEE Access*, 13:9171–9198.

Daniel Palomino and José Ochoa-Luna. 2020. Palomino-ochoa at SemEval-2020 task 9: Robust system based on transformer for code-mixed sentiment classification. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 963–967, Barcelona (online). International Committee for Computational Linguistics.

Kathiravan Pannerselvam, Saranya Rajiakodi, Sajeetha Thavareesan, Sathiyaraj Thangasamy, and Kishore Ponnusamy. 2024. SetFit: A robust approach for offensive content detection in Tamil-English code-mixed conversations using sentence transfer fine-tuning. In *Proceedings of the Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 35–42, St. Julian's, Malta. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Dojun Park, Jiwoo Lee, Seohyun Park, Hyeyun Jeong, Youngeun Koo, Soonha Hwang, Seonwoo Park, and Sungeun Lee. 2024. MultiPragEval: Multilingual pragmatic evaluation of large language models. In *Proceedings of the 2nd GenBench Workshop on Generalisation (Benchmarking) in NLP*, pages 96–119, Miami, Florida, USA. Association for Computational Linguistics.

Aryan Patil, Varad Patwardhan, Abhishek Phaltankar, Gauri Takawane, and Raviraj Joshi. 2023. Comparative study of pre-trained bert models for code-mixed hindi-english data. *2023 IEEE 8th International Conference for Convergence in Technology (I2CT)*, pages 1–7.

Parth Patwa, Gustavo Aguilar, Sudipta Kar, Suraj Pandey, Srinivas PYKL, Björn Gambäck, Tanmoy Chakraborty, Thamar Solorio, and Amitava Das. 2020. SemEval-2020 task 9: Overview of sentiment analysis of code-mixed tweets. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 774–790, Barcelona (online). International Committee for Computational Linguistics.

Siyao Peng, Zihang Sun, Huangyan Shan, Marie Kolm, Verena Blaschke, Ekaterina Artemova, and Barbara Plank. 2024. Sebastian, Basti, Wastl?! recognizing named entities in Bavarian dialectal data. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 14478–14493, Torino, Italia. ELRA and ICCL.

Sandun Sameera Perera and Deshan Koshala Sumanathilaka. 2025. Machine translation and transliteration for Indo-Aryan languages: A systematic review. In *Proceedings of the First Workshop on Natural Language Processing for Indo-Aryan and Dravidian Languages*, pages 11–21, Abu Dhabi. Association for Computational Linguistics.

Shana Poplack. 1988. Contrasting patterns of code-switching in two communities. *Codeswitching: Anthropological and sociolinguistic perspectives*, 48:215–244.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the*

Tenth Workshop on Statistical Machine Translation, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

Tom Potter and Zheng Yuan. 2024. LLM-based code-switched text generation for grammatical error correction. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 16957–16965, Miami, Florida, USA. Association for Computational Linguistics.

Archiki Prasad, Mohammad Ali Rehan, Shreya Pathak, and Preethi Jyothi. 2021. The effectiveness of intermediate-task training for code-switched natural language understanding. In *Proceedings of the 1st Workshop on Multilingual Representation Learning*, pages 176–190, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Adithya Pratapa, Gayatri Bhat, Monojit Choudhury, Sunayana Sitaram, Sandipan Dandapat, and Kalika Bali. 2018a. Language modeling for code-mixing: The role of linguistic theory based synthetic data. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1543–1553, Melbourne, Australia. Association for Computational Linguistics.

Adithya Pratapa and Monojit Choudhury. 2021. Comparing grammatical theories of code-mixing. In *Proceedings of the Seventh Workshop on Noisy User-generated Text (W-NUT 2021)*, pages 158–167, Online. Association for Computational Linguistics.

Adithya Pratapa, Monojit Choudhury, and Sunayana Sitaram. 2018b. Word embeddings for code-mixed language processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3067–3072, Brussels, Belgium. Association for Computational Linguistics.

Keyu Pu, Hongyi Liu, Yixiao Yang, Jiangzhou Ji, Wenyi Lv, and Yaohan He. 2022. CMB AI lab at SemEval-2022 task 11: A two-stage approach for complex named entity recognition via span boundary detection and span classification. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 1603–1607, Seattle, United States. Association for Computational Linguistics.

Libo Qin, Minheng Ni, Yue Zhang, and Wanxiang Che. 2020. Cosda-ml: Multi-lingual code-switching data augmentation for zero-shot cross-lingual nlp. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*, pages 3853–3860. International Joint Conferences on Artificial Intelligence Organization. Main track.

Geetha R, Karthika D, and L Ashok Kumar. 2025. Enhancing asr accuracy and coherence across indian languages with wav2vec2 and gpt-2. *ICTACT Journal on Data Science and Machine Learning*, 6:761–764.

Tathagata Raha, Sainik Mahata, Dipankar Das, and Sivaji Bandyopadhyay. 2019. Development of POS tagger for English-Bengali code-mixed data. In *Proceedings of the 16th International Conference on Natural Language Processing*, pages 143–149, International Institute of Information Technology, Hyderabad, India. NLP Association of India.

Md Nishat Raihan, Dhiman Goswami, Antara Mahmud, Antonios Anastasopoulos, and Marcos Zampieri. 2023a. SentMix-3L: A novel code-mixed test dataset in Bangla-English-Hindi for sentiment analysis. In *Proceedings of the First Workshop in South East Asian Language Processing*, pages 79–84, Nusa Dua, Bali, Indonesia. Association for Computational Linguistics.

Md Nishat Raihan, Umma Tanmoy, Anika Binte Islam, Kai North, Tharindu Ranasinghe, Antonios Anastasopoulos, and Marcos Zampieri. 2023b. Offensive language identification in transliterated and code-mixed Bangla. In *Proceedings of the First Workshop on Bangla Language Processing (BLP-2023)*, pages 1–6, Singapore. Association for Computational Linguistics.

Nishat Raihan, Dhiman Goswami, Antara Mahmud, Antonios Anastasopoulos, and Marcos Zampieri. 2024. EmoMix-3L: A code-mixed dataset for Bangla-English-Hindi for emotion detection. In *Proceedings of the 7th Workshop on Indian Language Data: Resources and Evaluation*, pages 11–16, Torino, Italia. ELRA and ICCL.

Humair Raj Khan, Deepak Gupta, and Asif Ekbal. 2021. Towards developing a multilingual and code-mixed visual question answering system by knowledge distillation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 1753–1767, Punta Cana, Dominican Republic. Association for Computational Linguistics.

SB Rajeshwari and Jagadish S Kallimani. 2021. Regional language code-switching for natural language understanding and intelligent digital assistants. In *Innovations in Electrical and Electronic Engineering: Proceedings of ICEEE 2021*, pages 927–948. Springer.

Priya Rani, Theodorus Fransen, John P. McCrae, and Gaurav Negi. 2024a. MaCmS: Magahi code-mixed dataset for sentiment analysis. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 10880–10890, Torino, Italia. ELRA and ICCL.

Priya Rani, Gaurav Negi, Saroj Jha, Shardul Suryawanshi, Atul Kr. Ojha, Paul Buitelaar, and John P. McCrae. 2024b. Findings of the WILDRE shared task on code-mixed less-resourced sentiment analysis for Indo-Aryan languages. In *Proceedings of the 7th Workshop on Indian Language Data: Resources and Evaluation*, pages 17–23, Torino, Italia. ELRA and ICCL.

Sudhanshu Ranjan, Dheeraj Mekala, and Jingbo Shang. 2022. Progressive sentiment analysis for code-switched text data. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 1155–1167, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Himashi Rathnayake, Janani Sumanapala, Raveesha Rukshani, and Surangika Ranathunga. 2024. Adapterfusion-based multi-task learning for code-mixed and code-switched text classification. *Engineering Applications of Artificial Intelligence*, 127:107239.

Nils Reimers and Iryna Gurevych. 2020. Making monolingual sentence embeddings multilingual using knowledge distillation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4512–4525, Online. Association for Computational Linguistics.

Mohd Sanad Zaki Rizvi, Anirudh Srinivasan, Tanuja Ganu, Monojit Choudhury, and Sunayana Sitaram. 2021. GCM: A toolkit for generating synthetic code-mixed text. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 205–211, Online. Association for Computational Linguistics.

Sumukh S and Manish Shrivastava. 2022. "kanglish alli names!" named entity recognition for Kannada-English code-mixed social media data. In *Proceedings of the Eighth Workshop on Noisy User-generated Text (W-NUT 2022)*, pages 154–161, Gyeongju, Republic of Korea. Association for Computational Linguistics.

Caroline Sabty, Mohamed Islam, and Slim Abdennadher. 2020. Contextual embeddings for Arabic-English code-switched data. In *Proceedings of the Fifth Arabic Natural Language Processing Workshop*, pages 215–225, Barcelona, Spain (Online). Association for Computational Linguistics.

Cesa Salaam, Franck Dernoncourt, Trung Bui, Danda Rawat, and Seunghyun Yoon. 2022. Offensive content detection via synthetic code-switched text. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 6617–6624, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Bidisha Samanta, Sharmila Reddy, Hussain Jagirdar, Niloy Ganguly, and Soumen Chakrabarti. 2019. A deep generative model for code switched text. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*, pages 5175–5181. International Joint Conferences on Artificial Intelligence Organization.

Younes Samih and Wolfgang Maier. 2016. An Arabic-Moroccan Darija code-switched corpus. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4170–4175, Portorož, Slovenia. European Language Resources Association (ELRA).

Sebastin Santy, Anirudh Srinivasan, and Monojit Choudhury. 2021. BERTologiCoMix: How does code-mixing interact with multilingual BERT? In *Proceedings of the Second Workshop on Domain Adaptation for NLP*, pages 111–121, Kyiv, Ukraine. Association for Computational Linguistics.

Yash Raj Sarrof. 2025. Homophonic pun generation in code mixed Hindi English. In *Proceedings of the 1st Workshop on Computational Humor (CHum)*, pages 23–31, Online. Association for Computational Linguistics.

Sunil Saumya, Abhinav Kumar, and Jyoti Prakash Singh. 2021. Offensive language identification in Dravidian code mixed social media text. In *Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages*, pages 36–45, Kyiv. Association for Computational Linguistics.

Salim Sazzed. 2021. Abusive content detection in transliterated Bengali-English social media corpus. In *Proceedings of the Fifth Workshop on Computational Approaches to Linguistic Code-Switching*, pages 125–130, Online. Association for Computational Linguistics.

Thomas Scialom, Paul-Alexis Dray, Sylvain Lamprier, Benjamin Piwowarski, and Jacopo Staiano. 2020. MLSUM: The multilingual summarization corpus. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8051–8067, Online. Association for Computational Linguistics.

Royal Sequiera, Monojit Choudhury, and Kalika Bali. 2015. POS tagging of Hindi-English code mixed text from social media: Some machine learning experiments. In *Proceedings of the 12th International Conference on Natural Language Processing*, pages 237–246, Trivandrum, India. NLP Association of India.

Sanket Shah, Pratik Joshi, Sebastin Santy, and Sunayana Sitaram. 2019. CoSSAT: Code-switched speech annotation tool. In *Proceedings of the First Workshop on Aggregating and Analysing Crowdsourced Annotations for NLP*, pages 48–52, Hong Kong. Association for Computational Linguistics.

Mohammad Nazmush Shamael, Sabila Nawshin, Swakkhar Shatabda, and Salekul Islam. 2024. Banglishrev: A large-scale bangla-english and code-mixed dataset of product reviews in e-commerce. *Preprint*, arXiv:2412.13161.

Bhavani Shankar, Preethi Jyothi, and Pushpak Bhattacharyya. 2024. In-context mixing (ICM): Code-mixed prompts for multilingual LLMs. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4162–4176, Bangkok, Thailand. Association for Computational Linguistics.

Kogilavani Shanmugavadivel, VE Sathishkumar, Sandhiya Raja, T Bheema Lingaiah, S Neelakandan, and Malliga Subramanian. 2022. Deep learning based sentiment analysis and offensive language identification on multilingual code-mixed data. *Scientific Reports*, 12(1):21557.

Shashi Shekhar, Dilip Kumar Sharma, and Mirza Mohd. Sufyan Beg. 2020. Language identification framework in code-mixed social media text based on quantum lstm — the word belongs to which language? *Modern Physics Letters B*, 34:2050086.

Dongming Sheng, Kexin Han, Hao Li, Yan Zhang, Yucheng Huang, Jun Lang, and Wenqiang Liu. 2025. Test-time code-switching for cross-lingual aspect sentiment triplet extraction. 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 5041–5053, Albuquerque, New Mexico. Association for Computational Linguistics.

Mehak Sheokand, Sparsh Kumar, and Akshat Kumar. 2025. CodeMixBench: A new benchmark for generating code from code-mixed prompts. In *Proceedings of the 18th International Conference on Natural Language Processing (ICON)*, Vasco da Gama, Goa, India. NLP Association of India (NLPAI).

Rajvee Sheth, Himanshu Beniwal, and Mayank Singh. 2025. Comi-lingua: Expert annotated large-scale dataset for multitask nlp in hindi-english code-mixing. *arXiv preprint arXiv:2503.21670*.

Rajvee Sheth, Shubh Nisar, Heenaben Prajapati, Himanshu Beniwal, and Mayank Singh. 2024. Commentator: A code-mixed multilingual text annotation framework. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 101–109, Miami, Florida, USA. Association for Computational Linguistics.

Anastasia Shimorina and Anya Belz. 2022. The human evaluation datasheet: A template for recording details of human evaluation experiments in NLP. In *Proceedings of the 2nd Workshop on Human Evaluation of NLP Systems (HumEval)*, pages 54–75, Dublin, Ireland. Association for Computational Linguistics.

Mayur Shirke, Amey Shembade, Pavan Thorat, Madhushri Wagh, and Raviraj Joshi. 2025. Comparative study of pre-trained bert and large language models for code-mixed named entity recognition. *arXiv preprint arXiv:2509.02514*.

Yurii Shynkarov, Veronika Solopova, and Vera Schmitt. 2025. Improving sentiment analysis for Ukrainian social media code-switching data. In *Proceedings of the Fourth Ukrainian Natural Language Processing Workshop (UNLP 2025)*, pages 179–193, Vienna, Austria (online). Association for Computational Linguistics.

Rushendra Sidibomma, Pransh Patwa, Parth Patwa, Aman Chadha, Vinija Jain, and Amitava Das. 2025. LLMsAgainstHate@NLU of Devanagari script languages 2025: Hate speech detection and target identification in Devanagari languages via parameter efficient fine-tuning of LLMs. In *Proceedings of the First Workshop on Challenges in Processing South Asian Languages (CHiPSAL 2025)*, pages 301–307, Abu Dhabi, UAE. International Committee on Computational Linguistics.

Abhishek Singh and Surya Pratap Singh Parmar. 2020. Voice@SRIB at SemEval-2020 tasks 9 and 12: Stacked ensemblingmethod for sentiment and offensiveness detection in social media. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1331–1341, Barcelona (online). International Committee for Computational Linguistics.

Kushagra Singh, Indira Sen, and Ponnurangam Kumaraguru. 2018a. Language identification and named entity recognition in Hinglish code mixed tweets. In *Proceedings of ACL 2018, Student Research Workshop*, pages 52–58, Melbourne, Australia. Association for Computational Linguistics.

Shruti Singh, Muskaan Singh, and Virender Kadyan. 2025. Hiacc: Hinglish adult & children code-switched corpus. *Data in Brief*, page 111886.

Thoudam Doren Singh and Thamar Solorio. 2017. Towards translating mixed-code comments from social media. In *International Conference on Computational Linguistics and Intelligent Text Processing*, pages 457–468. Springer.

Vinay Singh, Deepanshu Vijay, Syed Sarfaraz Akhtar, and Manish Shrivastava. 2018b. Named entity recognition for Hindi-English code-mixed social media text. In *Proceedings of the Seventh Named Entities Workshop*, pages 27–35, Melbourne, Australia. Association for Computational Linguistics.

Thamar Solorio, Elizabeth Blair, Suraj Maharjan, Steven Bethard, Mona Diab, Mahmoud Ghoneim, Abdelati Hawwari, Fahad AlGhamdi, Julia Hirschberg, Alison Chang, and Pascale Fung. 2014. Overview for the first shared task on language identification in code-switched data. In *Proceedings of the First Workshop on Computational Approaches to Code Switching*, pages 62–72, Doha, Qatar. Association for Computational Linguistics.

Thamar Solorio and Yang Liu. 2008. Learning to predict code-switching points. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 973–981, Honolulu, Hawaii. Association for Computational Linguistics.

Jiayang Song, Yuheng Huang, Zhehua Zhou, and Lei Ma. 2025. Multilingual blending: Large language model safety alignment evaluation with language mixture. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 3433–3449, Albuquerque, New Mexico. Association for Computational Linguistics.

Dama Sravani and Radhika Mamidi. 2023. Enhancing code-mixed text generation using synthetic data filtering in neural machine translation. In *Proceedings of the 27th Conference on Computational Natural Language Learning (CoNLL)*, pages 211–220, Singapore. Association for Computational Linguistics.

Varad Srivastava. 2025. DweshVaani: An LLM for detecting religious hate speech in code-mixed Hindi-English. In *Proceedings of the First Workshop on Challenges in Processing South Asian Languages (CHiPSAL 2025)*, pages 46–60, Abu Dhabi, UAE. International Committee on Computational Linguistics.

Vivek Srivastava and Mayank Singh. 2020. IIT Gandhinagar at SemEval-2020 task 9: Code-mixed sentiment classification using candidate sentence generation and selection. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1259–1264, Barcelona (online). International Committee for Computational Linguistics.

Vivek Srivastava and Mayank Singh. 2021. HinGE: A dataset for generation and evaluation of code-mixed Hinglish text. In *Proceedings of the 2nd Workshop on Evaluation and Comparison of NLP Systems*, pages 200–208, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Vivek Srivastava and Mayank Singh. 2022a. HinglishEval generation challenge on quality estimation of synthetic code-mixed text: Overview and results. In *Proceedings of the 15th International Conference on Natural Language Generation: Generation Challenges*, pages 19–25, Waterville, Maine, USA and virtual meeting. Association for Computational Linguistics.

Vivek Srivastava and Mayank Singh. 2022b. HinglishEval generation challenge on quality estimation of synthetic code-mixed text: Overview and results. In *Proceedings of the 15th International Conference on Natural Language Generation: Generation Challenges*, pages 19–25, Waterville, Maine, USA and virtual meeting. Association for Computational Linguistics.

Vivek Srivastava and Mayank Singh. 2022c. Overview and results of MixMT shared-task at WMT 2022. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 806–811, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Igor Sterner. 2024. Multilingual identification of English code-switching. In *Proceedings of the Eleventh Workshop on NLP for Similar Languages, Varieties, and Dialects (VarDial 2024)*, pages 163–173, Mexico City, Mexico. Association for Computational Linguistics.

Igor Sterner and Simone Teufel. 2023. TongueSwitcher: Fine-grained identification of German-English code-switching. In *Proceedings of the 6th Workshop on Computational Approaches to Linguistic Code-Switching*, pages 1–13, Singapore. Association for Computational Linguistics.

Igor Sterner and Simone Teufel. 2025a. Code-switching and syntax: A large-scale experiment. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 11526–11533, Vienna, Austria. Association for Computational Linguistics.

Igor Sterner and Simone Teufel. 2025b. Minimal pair-based evaluation of code-switching. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 18575–18598, Vienna, Austria. Association for Computational Linguistics.

Ahmed Sultan, Mahmoud Salim, Amina Gaber, and Islam El Hosary. 2020. WESSA at SemEval-2020 task 9: Code-mixed sentiment analysis using transformers. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1342–1347, Barcelona (online). International Committee for Computational Linguistics.

Sathya Krishnan Suresh, Tanmay Surana, Lim Zhi Hao, and Eng Siong Chng. 2025. Cs-sum: A benchmark for code-switching dialogue summarization and the limits of large language models. *Preprint*, arXiv:2505.13559.

Tharun Suresh, Ayan Sengupta, Md Shad Akhtar, and Tanmoy Chakraborty. 2024. A comprehensive understanding of code-mixed language semantics using hierarchical transformer. *IEEE Transactions on Computational Social Systems*, 11(3):4139–4148.

Saedeh Tahery and Saeed Farzi. 2025. An adapted few-shot prompting technique using chatgpt to advance low-resource languages understanding. *IEEE Access*, 13:93614–93628.

Ishan Tarunesh, Syamantak Kumar, and Preethi Jyothi. 2021. From machine translation to code-switching: Generating high-quality code-switched text. 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 3154–3169, Online. Association for Computational Linguistics.

Panuthep Tasawong, Wuttikorn Ponwitayarat, Peerat Limkonchotiwat, Can Udomcharoenchaikit, Ekapol Chuangsuwanich, and Sarana Nutanong. 2023. Typo-robust representation learning for dense retrieval. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1106–1115, Toronto, Canada. Association for Computational Linguistics.

Kushal Tatariya, Heather Lent, and Miryam de Lhoneux. 2023. Transfer learning for code-mixed data: Do pretraining languages matter? In

Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis, pages 365–378, Toronto, Canada. Association for Computational Linguistics.

Michelle Terblanche, Kayode Olaleye, and Vukosi Marivate. 2024. Prompting towards alleviating code-switched data scarcity in under-resourced languages with GPT as a pivot. In *Proceedings of the 3rd Annual Meeting of the Special Interest Group on Under-resourced Languages @ LREC-COLING 2024*, pages 272–282, Torino, Italia. ELRA and ICCL.

Pasindu Udawatta, Imesha Udayangana, Charith Gamage, and G. C. De Silva. 2024. Use of prompt-based learning for code-mixed and code-switched text classification. *World Wide Web*, 27(4):2713–2742.

Enes Yavuz Ugan, Ngoc-Quan Pham, Leonard Bärmann, and Alex Waibel. 2025. Pier: A novel metric for evaluating what matters in code-switching. *Preprint*, arXiv:2501.09512.

Thin Dang Van, Hao Duong Ngoc, and Ngan Nguyen Luu-Thuy. 2022. Sentiment analysis in code-mixed Vietnamese-English sentence-level hotel reviews. In *Proceedings of the 36th Pacific Asia Conference on Language, Information and Computation*, pages 54–61, Manila, Philippines. Association for Computational Linguistics.

Aditya Vavre, Abhirut Gupta, and Sunita Sarawagi. 2022. Adapting multilingual models for code-mixed translation. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 7133–7141, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Yogarshi Vyas, Spandana Gella, Jatin Sharma, Kalika Bali, and Monojit Choudhury. 2014. POS tagging of English-Hindi code-mixed social media content. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 974–979, Doha, Qatar. Association for Computational Linguistics.

Anshul Wadhawan and Akshita Aggarwal. 2021. Towards emotion recognition in Hindi-English code-mixed data: A transformer based approach. In *Proceedings of the Eleventh Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 195–202, Online. Association for Computational Linguistics.

Changhan Wang, Kyunghyun Cho, and Douwe Kiela. 2018. Code-switched named entity recognition with embedding attention. In *Proceedings of the Third Workshop on Computational Approaches to Linguistic Code-Switching*, pages 154–158, Melbourne, Australia. Association for Computational Linguistics.

Fei Wang, Kuan-hao Huang, Anoop Kumar, Aram Galstyan, Greg Ver steeg, and Kai-wei Chang. 2022. Zero-shot cross-lingual sequence tagging as Seq2Seq generation for joint intent classification and slot filling. In *Proceedings of the Massively Multilingual Natural Language Understanding Workshop (MMNLU-22)*, pages 53–61, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Renxi Wang, Haonan Li, Minghao Wu, Yuxia Wang, Xudong Han, Chiyu Zhang, and Timothy Baldwin. 2024. Demystifying instruction mixing for fine-tuning large language models. *Preprint*, arXiv:2312.10793.

Shuhe Wang, Xiaofei Sun, Xiaoya Li, Rongbin Ouyang, Fei Wu, Tianwei Zhang, Jiwei Li, Guoyin Wang, and Chen Guo. 2025a. GPT-NER: Named entity recognition via large language models. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 4257–4275, Albuquerque, New Mexico. Association for Computational Linguistics.

Zhijun Wang, Jiahuan Li, Hao Zhou, Rongxiang Weng, Jingang Wang, Xin Huang, Xue Han, Junlan Feng, Chao Deng, and Shujian Huang. 2025b. Investigating and scaling up code-switching for multilingual language model pre-training. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 11032–11046, Vienna, Austria. Association for Computational Linguistics.

Alex Warstadt, Amanpreet Singh, and Samuel R. Bowman. 2019. Neural network acceptability judgments. *Transactions of the Association for Computational Linguistics*, 7:625–641.

Daniel Weisberg Mitelman, Nachum Dershowitz, and Kfir Bar. 2024. Code-switching and back-transliteration using a bilingual model. In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 1501–1511, St. Julian's, Malta. Association for Computational Linguistics.

Orion Weller, Matthias Sperber, Telmo Pires, Hendra Setiawan, Christian Gollan, Dominic Telaar, and Matthias Paulik. 2022. End-to-end speech translation for code switched speech. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1435–1448, Dublin, Ireland. Association for Computational Linguistics.

Martin Weyssow, Xin Zhou, Kisub Kim, David Lo, and Houari Sahraoui. 2024. Exploring parameter-efficient fine-tuning techniques for code generation with large language models. *Preprint*, arXiv:2308.10462.

Chenxi Whitehouse, Fenia Christopoulou, and Ignacio Iacobacci. 2022. EntityCS: Improving zero-shot cross-lingual transfer with entity-centric code switching. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 6698–6714, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Genta Winata, Alham Fikri Aji, Zheng Xin Yong, and Thamar Solorio. 2023a. The decades progress on code-switching research in NLP: A systematic survey on trends and challenges. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 2936–2978, Toronto, Canada. Association for Computational Linguistics.

Genta Indra Winata, Alham Fikri Aji, Samuel Cahyawijaya, Rahmad Mahendra, Fajri Koto, Ade Romadhony, Kemal Kurniawan, David Moeljadi, Radityo Eko Prasojo, Pascale Fung, Timothy Baldwin, Jey Han Lau, Rico Sennrich, and Sebastian Ruder. 2023b. NusaX: Multilingual parallel sentiment dataset for 10 Indonesian local languages. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 815–834, Dubrovnik, Croatia. Association for Computational Linguistics.

Genta Indra Winata, Alham Fikri Aji, Samuel Cahyawijaya, Rahmad Mahendra, Fajri Koto, Ade Romadhony, Kemal Kurniawan, David Moeljadi, Radityo Eko Prasojo, Pascale Fung, and 1 others. 2022. Nusax: Multilingual parallel sentiment dataset for 10 indonesian local languages. *arXiv preprint arXiv:2205.15960*.

Genta Indra Winata, Samuel Cahyawijaya, Zihan Liu, Zhaojiang Lin, Andrea Madotto, and Pascale Fung. 2021. Are multilingual models effective in code-switching? In *Proceedings of the Fifth Workshop on Computational Approaches to Linguistic Code-Switching*, pages 142–153, Online. Association for Computational Linguistics.

Genta Indra Winata, Andrea Madotto, Chien-Sheng Wu, and Pascale Fung. 2019. Code-switched language models using neural based synthetic data from parallel sentences. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 271–280, Hong Kong, China. Association for Computational Linguistics.

Chengyan Wu, Yiqiang Cai, Yang Liu, Pengxu Zhu, Yun Xue, Ziwei Gong, Julia Hirschberg, and Bolei Ma. 2025a. Multimodal emotion recognition in conversations: A survey of methods, trends, challenges and prospects. *Preprint*, arXiv:2505.20511.

Linjuan Wu, Hao-Ran Wei, Baosong Yang, and Weiming Lu. 2025b. From English to second language mastery: Enhancing LLMs with cross-lingual continued instruction tuning. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 23006–23023, Vienna, Austria. Association for Computational Linguistics.

Qi Wu, Peng Wang, and Chenghao Huang. 2020. MeisterMorxrc at SemEval-2020 task 9: Fine-tune bert and multitask learning for sentiment analysis of code-mixed tweets. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1294–1297, Barcelona (online). International Committee for Computational Linguistics.

Ting-Wei Wu, Changsheng Zhao, Ernie Chang, Yangyang Shi, Pierce Chuang, Vikas Chandra, and Biing Juang. 2023. Towards zero-shot multilingual transfer for code-switched responses. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7551–7563, Toronto, Canada. Association for Computational Linguistics.

Peng Xie, Xingyuan Liu, Tsz Wai Chan, Yequan Bie, Yangqiu Song, Yang Wang, Hao Chen, and Kani Chen. 2025. Switchlingua: The first large-scale multilingual and multi-ethnic code-switching dataset. *Preprint*, arXiv:2506.00087.

Anjali Yadav, Tanya Garg, Matej Klemen, Matej Ulčar, Basant Agarwal, and M. Robnik-Šikonja. 2025. From translation to generative llms: Classification of code-mixed affective tasks. *IEEE Transactions on Affective Computing*, 16(3):2090–2101.

Anjali Yadav, Tanya Garg, Divyanshu Sharma, Asif Ekbal, and Pushpak Bhattacharyya. 2024. From translation to generative llms: Classification of code-mixed affective tasks. *IEEE Access*.

Songlin Yang and Kewei Tu. 2022. Combining (second-order) graph-based and headed-span-based projective dependency parsing. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1428–1434, Dublin, Ireland. Association for Computational Linguistics.

Yilun Yang and Yekun Chai. 2025. Codemixbench: Evaluating code-mixing capabilities of llms across 18 languages. *Preprint*, arXiv:2507.18791.

M. Yasir, L. Chen, A. Khatoon, M. A. Malik, and F. Abid. 2021. Mixed script identification using automated dnn hyperparameter optimization. *Computational Intelligence and Neuroscience*, 2021:8415333.

Zheng Xin Yong, Ruochen Zhang, Jessica Forde, Skyler Wang, Arjun Subramonian, Holy Lovenia, Samuel Cahyawijaya, Genta Winata, Lintang Sutawika, Jan Christian Blaise Cruz, Yin Lin Tan, Long Phan, Long Phan, Rowena Garcia, Thamar Solorio, and Alham Fikri Aji. 2023. Prompting multilingual large language models to generate code-mixed texts: The case of south East Asian languages. In *Proceedings of the 6th Workshop on Computational Approaches to Linguistic Code-Switching*, pages 43–63, Singapore. Association for Computational Linguistics.

Haneul Yoo, Cheonbok Park, Sangdoo Yun, Alice Oh, and Hwaran Lee. 2025. Code-switching curriculum learning for multilingual transfer in LLMs. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 7816–7836, Vienna, Austria. Association for Computational Linguistics.

Haneul Yoo, Yongjin Yang, and Hwaran Lee. 2024. Code-switching red-teaming: Llm evaluation for

safety and multilingual understanding. In *Annual Meeting of the Association for Computational Linguistics*.

Urchade Zaratiana, Nadi Tomeh, Pierre Holat, and Thierry Charnois. 2024. GLiNER: Generalist model for named entity recognition using bidirectional transformer. 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 5364–5376, Mexico City, Mexico. Association for Computational Linguistics.

Linda Zeng. 2024. Leveraging large language models for code-mixed data augmentation in sentiment analysis. In *Proceedings of the Second Workshop on Social Influence in Conversations (SICon 2024)*, pages 85–101, Miami, Florida, USA. Association for Computational Linguistics.

Fengrun Zhang, Wang Geng, Hukai Huang, Yahui Shan, Cheng Yi, and He Qu. 2025a. Boosting code-switching asr with mixture of experts enhanced speech-conditioned llm. In *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5.

Hongbin Zhang, Kehai Chen, Xuefeng Bai, Yang Xiang, and Min Zhang. 2025b. Lingualift: An effective two-stage instruction tuning framework for low-resource language reasoning. *Preprint*, arXiv:2412.12499.

Ruochen Zhang, Samuel Cahyawijaya, Jan Christian Blaise Cruz, Genta Winata, and Alham Fikri Aji. 2023. Multilingual large language models are not (yet) code-switchers. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12567–12582, Singapore. Association for Computational Linguistics.

Ruochen Zhang and Carsten Eickhoff. 2024. CroCoSum: A benchmark dataset for cross-lingual code-switched summarization. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 4113–4126, Torino, Italia. ELRA and ICCL.

Wenbo Zhang, Aditya Majumdar, and Amulya Yadav. 2025c. Chai for llms: Improving code-mixed translation in large language models through reinforcement learning with ai feedback. *Preprint*, arXiv:2411.09073.

Wenxuan Zhang, Ruidan He, Haiyun Peng, Lidong Bing, and Wai Lam. 2021. Cross-lingual aspect-based sentiment analysis with aspect term code-switching. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9220–9230, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Yuanchi Zhang, Yile Wang, Zijun Liu, Shuo Wang, Xiaolong Wang, Peng Li, Maosong Sun, and Yang Liu. 2024a. Enhancing multilingual capabilities of large language models through self-distillation from resource-rich languages. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11189–11204, Bangkok, Thailand. Association for Computational Linguistics.

Yuhang Zhang, Yubo Chen, Jun Zhao, and Kang Liu. 2024b. REPE: Representation and prediction-level alignment for multilingual intent detection and slot filling. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 7423–7438, Miami, FL. Association for Computational Linguistics.

Zhihan Zhang, Dong-Ho Lee, Yuwei Fang, Wenhao Yu, Mengzhao Jia, Meng Jiang, and Francesco Barbieri. 2024c. PLUG: Leveraging pivot language in cross-lingual instruction tuning. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7025–7046, Bangkok, Thailand. Association for Computational Linguistics.

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.

Ran Zhou, Xin Li, Ruidan He, Lidong Bing, Erik Cambria, Luo Si, and Chunyan Miao. 2022. MELM: Data augmentation with masked entity language modeling for low-resource NER. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2251–2262, Dublin, Ireland. Association for Computational Linguistics.

Zhihong Zhu, Xuxin Cheng, Dongsheng Chen, Zhiqi Huang, Hongxiang Li, and Yuexian Zou. 2023. Mix before align: Towards zero-shot cross-lingual sentiment analysis via soft-mix and multi-view learning. In *Interspeech*.

A Methodology

This section outlines the methodology adopted to identify, review, and categorize literature relevant to this survey on code-switching in the era of LLMs. The goal was to capture key trends, modeling techniques, datasets, and challenges across NLP tasks rather than conduct an exhaustive systematic review. The approach follows established survey practices (Kinney et al., 2023).

Paper Selection We began by defining a set of search keywords targeting three core dimensions: code-mixing/code-switching, multilingual NLP, and large language models. To ensure broad linguistic coverage, the search encompassed major bilingual and multilingual language pairs documented in prior research and repositories (e.g., CoVoSwitch, GLUECoS, LinCE). Using these keywords, we queried the ACL Anthology, arXiv and Semantic Scholar databases via their APIs, with a search cutoff date of October 2025, consistent with ACL Rolling Review's recency guidelines. This process initially retrieved 500 papers.

Screening and Filtering Duplicate entries were removed using DOIs and titles, prioritizing peer-reviewed sources. The remaining papers were manually screened for relevance. A study was included if it addressed code-mixing or code-switching within any NLP task, or explored multilingual and LLM-based adaptation methods. This screening resulted in a refined set of 308 papers, covering both pre-LLM and LLM-era research.

Categorization Selected papers were categorized by ***(i)*** task type (e.g., Language Identification, POS Tagging, NER, Sentiment Analysis, Speech/ASR, MT, NLG), ***(ii)*** modality (text, speech, vision-language), and ***(iii)*** model architecture (transformer-based, instruction-tuned, multimodal). When overlaps occurred (e.g., between translation and generation), we retained the category most central to the contribution. Dataset coverage, annotation methods, and language pairs were systematically verified to map diversity and resource availability. High-, mid-, and low-resource classifications followed conventions in multilingual NLP research.

This multi-stage process of search, screening, and categorization produced **308** papers forming the foundation of this survey, spanning **12 NLP tasks**, **30+ datasets**, and **80+ languages**.

A.1 Natural Language Understanding Tasks

Language Identification Script detection remains crucial for accurate token-level processing, with Bi-GRU architectures achieving 90.17% accuracy on Roman Urdu, Hindi, Saraiki, Bengali, and English using GloVe embeddings (Yasir et al., 2021). The ILID corpus provides 250K sentences across 25 scripts and 23 languages, including dual-script instances for Manipuri and Sindhi (Ingle and Mishra, 2025). Character n-gram TF-IDF features (1–6 grams) have proven effective for Dravidian script-mixed social media text (Saumya et al., 2021). Shared-task initiatives such as LT-EDI-EACL extended hope speech detection to English, Malayalam–English, and Tamil–English, where TF-IDF features combined with MuRIL embeddings achieved F1 scores of 0.92, 0.75, and 0.57 respectively (Dave et al., 2021). Specialized datasets such as KanHope (English–Kannada) highlight persistent issues of class imbalance and preprocessing challenges involving emojis and multilingual tokens (Hande et al., 2021). Overall, methodological advances have transitioned from traditional machine learning to transformer-based architectures, where task-adaptive pre-training and multilingual contextual embeddings substantially improve performance, particularly in low-resource and morphologically rich languages (Jayanthi and Gupta, 2021; Shanmugavadivel et al., 2022). **Offensive Language Identification** in code-switched text presents unique challenges, as users often employ strategic language alternation to bypass keyword-based moderation. Foundational datasets such as OffMix-3L establish trilingual benchmarks for Bangla–English–Hindi, underscoring the difficulty of handling transliterated content where phonetic variation hinders detection accuracy (Goswami et al., 2023; Sazzed, 2021). Transformer-based systems such as COOLI explicitly target adversarial switching strategies, while synthetic code-switched data generation has emerged as a promising avenue for building linguistically diverse and robust training corpora (Balouchzahi et al., 2021a; Salaam et al., 2022). Recent paradigms incorporate transfer and multi-task learning, with approaches such as SetFit enabling efficient few-shot adaptation for Tamil–English detection, and multi-task frameworks demonstrating strong performance across zero-shot and fine-tuning scenarios for harmful multimodal content (Pannerselvam et al., 2024; Kumar et al., 2025).

Takeaway LLMs have become powerful tools for multilingual hate and offensive speech detection. Specialized models such as *DweshVaani* demonstrate effectiveness in identifying religious hate speech in Hindi–English code-mixed contexts (Srivastava, 2025). However, safety alignment evaluations reveal persistent vulnerabilities

when encountering unseen language mixture patterns, as evidenced by frameworks like *Qorgau* in Kazakh–Russian bilingual settings (Goloburda et al., 2025). These findings highlight the need for better-resourced datasets, robust multilingual alignment, and evaluation protocols attuned to code-switched phenomena.

Sentiment & Emotion Analysis Sentiment and emotion analysis in CSW contexts has advanced through tasks like SemEval, with fine-tuned transformers (mBERT, XLM-R, RoBERTa) performing well on Hindi-English and Spanish-English datasets using focal loss and ensembling (Angel et al., 2020; Ma et al., 2020; Palomino and Ochoa-Luna, 2020; Sultan et al., 2020; Singh and Singh Parmar, 2020; Wu et al., 2020). Research spans Dravidian, Indonesian, Vietnamese-English, Kenyan Sheng, and trilingual settings (Chakravarthi et al., 2022; Winata et al., 2022; Van et al., 2022; Etori and Gini, 2024; Raihan et al., 2023a, 2024). Multi-label emotion detection and ABSA enhance cross-lingual performance (Wadhawan and Aggarwal, 2021; Zhang et al., 2021). Data scarcity is addressed via unsupervised self-training, curriculum learning, monolingual data, and CoSDA-ML's synthetic augmentation (Gupta et al., 2021b; Ranjan et al., 2022; Kumar et al., 2022c; Qin et al., 2020). LLMs enable zero-shot sentiment analysis and efficient synthetic data generation (Yadav et al., 2024; Zhang et al., 2023; Zeng, 2024). Harmful content detection uses datasets for Bangla and Devanagari hate speech, with PEFT and SetFit improving results (Raihan et al., 2023b; Sidibomma et al., 2025; Pannerselvam et al., 2024).

Takeaway Transformer-based models and LLMs, enhanced by synthetic data and few-shot learning, effectively tackle data scarcity in CSW sentiment analysis, across diverse languages (Qin et al., 2020; Pannerselvam et al., 2024). Developing robust datasets for low-resource and trilingual CSW scenarios is essential to boost generalization (Raihan et al., 2024). Lightweight adaptation techniques, like PEFT, are critical for efficient deployment in real-world multilingual applications (Sidibomma et al., 2025).

Syntactic analysis Syntactic analysis in CSW contexts has advanced from structural modeling to linguistic theory-guided approaches, improving parsing and evaluation. SyMCoM achieved 93.4% POS tagging accuracy for English-Hindi CSW by capturing syntactic variety (Kodali et al., 2022a). Synthetic treebanks for Bengali-English dependency parsing yielded 76.24% UAS and 61.41% LAS (Ghosh et al., 2019). CoMix's phonetic and POS-guided pretraining boosted Hinglish translation BLEU by 12.98 and NER F1 by 2.42 (Arora et al., 2023). Linguistic constraint comparisons showed a 2:1 naturalness advantage for Equivalence Constraint over heuristics ($\alpha = 0.59$) (Pratapa and Choudhury, 2021). Recent work leverages LLMs for Universal Dependencies annotations, achieving 95.29 LAS after revision for Spanglish and Spanish–Guaraní CSW, bridging syntactic modeling with LLM capabilities (Kellert et al., 2025). Large-scale CSW syntax experiments reached 79.4% acceptability across 11 language pairs (Sterner and Teufel, 2025a), with pre-trained models showing 0.83 graph edit distance correlation to monolingual parses (Laureano De Leon et al., 2024). Non-English LLM prompting improved grammaticality detection by 12% to 79.6% accuracy (Behzad et al., 2024), and LLM-based GEC achieved 63.71 $F_{0.5}$ on learner corpora (Potter and Yuan, 2024). Challenges remain in universalizing syntactic constraints across diverse languages (Pratapa and Choudhury, 2021).

Takeaway Syntactic-aware models and LLMs achieve high CSW parsing performance, but universal syntactic constraints remain debated, limiting low-resource language generalization (Pratapa et al., 2018a; Behzad et al., 2024). Future work should prioritize lightweight LLM adaptations for non-Latin script and tonal languages to enhance cross-linguistic transfer (Potter and Yuan, 2024).

Intent Classification Intent classification in code-switched conversational AI tackles irregular language alternation and non-standard orthography. Early vectorizer–classifier pipelines achieved high accuracy and robust NER (Rajeshwari and Kallimani, 2021). Random translation augmentation improved intent and slot performance on MultiATIS++ across eight languages (Krishnan et al., 2021), while XLM-R outperformed word embeddings in zero-shot Hinglish intent detection on M-CID (Arora et al., 2020). Seq2Seq models boosted slot labeling on MASSIVE, and PRO-CS prompting enhanced Hinglish accuracy (Wang et al., 2022; Bansal et al., 2022). Multilingual semantic parsing on NLmaps and ENTITYCS further improved intent and slot perfor-

mance across languages (Duong et al., 2017; Chen et al., 2022), with bilingual corpora supporting joint multi-intent and entity recognition in chatbots (Aguirre et al., 2022). Recent approaches like ContrastiveMix pretraining enhanced intent accuracy across 51 languages (Lin et al., 2024). Persistent challenges include limited corpora for typologically distant languages and high computational costs of LLMs (Lin et al., 2024; Zhang et al., 2024b).

Takeaway Code-switched intent classification has advanced with XLM-R and prompting methods like PRO-CS, but struggles with typologically diverse languages (Lin et al., 2024; Bansal et al., 2022). Innovative augmentation and lightweight transfer learning are key to overcoming data scarcity and computational barriers (Krishnan et al., 2021).

Question Answering Code-mixed QA combines linguistic and cultural reasoning across languages. Early neural QA systems struggled with multilingual queries using synthetic monolingual data (Gupta et al., 2018). A crowdsourced dataset improved natural code-mixing with mBERT (Chandu et al., 2018a). Multimodal visual QA with knowledge distillation enhanced efficiency in low-resource settings (Raj Khan et al., 2021). LLM-based COMMIT with instruction tuning improved Hinglish extractive QA, while non-English prompting boosted accuracy (Lee et al., 2024; Behzad et al., 2024). Curriculum learning in K-MMLU/HAE-RAE improved multiple-choice QA via code-switching pretraining (Yoo et al., 2025). Embedding transfer in ARC/TruthfulQA enhanced reasoning in Arabic, Bengali, and Vietnamese (Hong et al., 2025). MEGAVERSE showed LLMs excelling in African language QA, though test contamination affected results (Ahuja et al., 2024). Alignment methods like MAD-X addressed morphological complexity, improving Hinglish QA (Fazili and Jyothi, 2022). Still, Challenges include limited evaluation frameworks for high-resource pairs and data scarcity in non-Latin scripts (Chandu et al., 2018b; Raj Khan et al., 2021).

Takeaway Code-mixed QA showcases robust progress with LLMs achieving up to 79.6% accuracy through informed prompting and synthetic data (Behzad et al., 2024; Gupta et al., 2018). The reliance on high-resource pairs like Hinglish reveals a gap in addressing diverse, low-resource language ecosystems, necessitating innovative data augmentation (Chandu et al., 2018a). Lightweight models are essential to democratize access for communities with natural code-switching practices (Raj Khan et al., 2021).

Natural Language Inference Code-mixed NLI extends entailment detection across languages. The Hindi-English CALI dataset showed mBERT's limitations due to annotation disagreements (Khanuja et al., 2020a; Huang and Yang, 2023). CoSDA-ML's synthetic data enabled zero-shot NLI across 19 languages (Qin et al., 2020), while intermediate-task fine-tuning and meta-learning improved Hinglish NLI (Prasad et al., 2021; Kumar et al., 2022b). Prompt-based ICM boosted accuracy but requires language-specific tuning (Shankar et al., 2024). Challenges in inference and preprocessing have seen innovations that reduced API costs by 90% while maintaining performance (Nag et al., 2024). However, scaling remains difficult—mBERT and XLM-R still underperform on code-mixed tasks compared to monolingual ones (Wang et al., 2025b). CodeMixEval further reported GPT-4 Turbo's accuracy drop in 18 low-resource languages, while MultiPragEval revealed persistent pragmatic inference issues tied to cultural context (Yang and Chai, 2025; Park et al., 2024).

Takeaway Code-mixed NLI achieves 0.68-0.75 F1 with fine-tuned transformers and synthetic data, but cultural and pragmatic variability hinders generalization (Huang and Yang, 2023; Park et al., 2024). Future efforts should focus on diverse language-pair benchmarks and lightweight models to improve scalability (Nag et al., 2024). Efficient preprocessing and prompt engineering are critical for cost-effective, real-world deployment.

A.2 Natural Language Generation Tasks

Cross-lingual Transfer Code-mixed cross-lingual transfer learning mitigates data scarcity by enabling models to generalize to unseen code-switched tasks. Progressive Code-Switching (PCS) incrementally augments low-resource NER data, achieving 0.82-0.87 F1 in zero-shot transfer across ten languages (Li et al., 2024). EntityCS leverages Wikidata for cross-lingual slot filling, improving SLU benchmarks by 10% F1 (Chen et al., 2022). SCOPA's soft embedding mixing and pairwise alignment boost code-switched

sentence representation by 3-5% accuracy (Lee et al., 2021). Incontext Mixing (ICM) prompts enhance intent classification by 5-8% accuracy on MultiATIS++ (Shankar et al., 2024). Test-time code-switching for aspect-based sentiment analysis yields 8-12% F1 gains on SemEval (Sheng et al., 2025). A code-switching curriculum improves intent detection by 12-15% on African languages in MASSIVE (Yoo et al., 2025). MIGRATE's domain-specific LLM adaptation aligns embeddings, boosting zero-shot QA and NER by 15-20% F1 for low-resource languages (Hong et al., 2025). Challenges persist in handling typological diversity across language families.

Takeaway Code-mixed transfer learning, with methods like PCS and MIGRATE, achieves 0.82-0.87 F1 by leveraging synthetic data and embedding alignment, but typological diversity remains a barrier (Li et al., 2024; Hong et al., 2025). Innovative prompting and curricula uniquely empower low-resource language tasks by simulating natural switching patterns (Shankar et al., 2024). Future efforts should focus on scalable, typologically inclusive frameworks to ensure robust generalization across diverse linguistic contexts (Yoo et al., 2025).

Text Summarization Text summarization in code-switched contexts tackles data scarcity and linguistic complexity with specialized datasets and models. GupShup's Hindi-English dataset (6,800+ conversations) achieved 38.2 ROUGE-L using fine-tuned mBART and multi-view seq2seq models (Mehnaz et al., 2021). CroCo-Sum's 24,000 English-Chinese article-summary pairs (92% code-switched) showed that ROUGE-1 drops for cross-lingual models (Zhang and Eickhoff, 2024). CS-Sum's Hindi-English and Spanish-English benchmark improved ROUGE-2 via alternation pattern modeling (Suresh et al., 2025). MLSUM's multilingual corpus (five languages) gained 7% ROUGE-L in low-resource settings with synthetic data fine-tuning (Scialom et al., 2020). Contrastive learning boosted ROUGE-1 by 15% on MultiATIS++ by aligning mixed-language representations (Zhang and Eickhoff, 2024). Challenges persist in capturing semantic alignments across typologically diverse languages (Lin et al., 2024).

Takeaway Code-switched NLG tasks everage LLMs and datasets like GupShup and CS-Sum to achieve strong results, but typological diversity and syntactic complexity at switch points remain hurdles. Techniques like PCS and contrastive learning improve low-resource tasks by aligning embeddings and replicating natural CSW patterns. Capturing cultural context and informal pragmatics is still challenging. Future efforts should prioritize adaptive models, user-centric evaluation, and typologically inclusive frameworks.

A.3 Emerging Contemporary Tasks

Code-switched dialogue generation has emerged as a critical application area, with frameworks like MulZDG (Liu et al., 2022) enabling zero-shot dialogue generation for low-resource languages through synthetic code-mixed data augmentation. **Conversational summarization** for code-mixed dialogues, exemplified by the GupShup dataset (Mehnaz et al., 2021) for Hinglish conversations, represents a challenging abstractive summarization task requiring understanding of informal, mixed-language discourse patterns. **Visual question answering systems** have been extended to handle code-mixed queries through knowledge distillation approaches that transfer capabilities from large multilingual models to efficient specialized architectures (Raj Khan et al., 2021). **Code-mixed text quality evaluation** has emerged as a specialized task, with approaches combining linguistic metrics and multilingual language model embeddings to assess synthetic data adequacy and fluency (Kodali et al., 2022b). **Homophonic pun generation** in code-mixed contexts demonstrates the creative potential of LLMs in cross-linguistic wordplay and humor generation (Sarrof, 2025).

A.4 Underexplored Frontiers Tasks

Although notable progress has been made in core code-switching tasks, several frontier areas remain underexplored. **Reasoning tasks** such as mathematical, and causal reasoning struggle with cross-language entailment and logical complexity (Raihan et al., 2023a; Mohamed et al., 2025). **Higher-order cognitive tasks** like metaphor understanding and analogical reasoning show limited performance and lack cultural grounding (Kodali et al., 2025; Mehnaz et al., 2021). **Coding-related tasks**, including code generation and explanation from mixed prompts, achieve moderate functional accuracy but are far from mature (Yang and Chai, 2025; Khatri et al., 2023). While controllable

language steering has shown early success (e.g., 85% switch-point control) (Mondal et al., 2022), adaptive and pragmatic CSW generation remains largely theoretical (Kuwanto et al., 2024).

Takeaway Several challenges remain: (i) data sparsity and imbalance, especially for underrepresented languages; (ii) orthographic variability and noisy social media text that reduce generalizability; (iii) limited scalability of transfer learning methods for distant language pairs; (iv) insufficient exploration of shared versus task-specific representations in joint modeling; and (v) lack of robust, community-wide benchmarks for fair comparison across diverse CSW scenarios.

A.5 Fine-tuning Approaches

Instruction Tuning Instruction tuning in multilingual (CSW) settings enhances LLMs ability to follow instructions across languages while aligning with human preferences, despite challenges like Script variabilty and cultural nuances. COMMIT adapts English-centric LLMs via code-mixed instruction tuning for low-resource QA, achieving up to 32× gains in exact match on Hinglish tasks (Lee et al., 2024). CSCL progressively introduces CSW during instruction tuning, improving cross-lingual transfer by 5–8% across 11 language pairs (Yoo et al., 2025). SPHINX achieves sample-efficient fine-tuning via translated instructions, boosting zero-shot QA performance by 10–15% in African languages (Ahuja et al., 2025). PLUG uses pivot-language code-switching to enhance instruction-following, yielding 7–12% gains in multilingual response generation (Zhang et al., 2024c). Preference-aligned methods, including Predicting Preferences in Generated CSW Text and Multilingual Blending, improve naturalness and ethical adherence by 15–20% across low-resource scenarios (Gupta et al., 2025; Song et al., 2025). These works highlight the impact of curriculum-based and preference-optimized tuning, while underscoring the need for culturally diverse datasets.

Takeaways: Curriculum and preference-guided fine-tuning enables LLMs to better handle multilingual CSW tasks with improved instruction adherence and cross-lingual generalization. Future research should emphasize robustness to spontaneous CSW, cross-domain generalization, and contextually appropriate outputs.

Parameter-efficient fine-tuning PEFT methods like LoRA, QLoRA, and soft prompt tuning efficiently adapts LLMs for CSW tasks. PEFT methods like LoRA on Llama-3.1-8B and Nemo-Instruct achieves up to 90.75% accuracy for Hindi and Nepali hate speech (Sidibomma et al., 2025), QLoRA on Gemma-2 reaches 76.19% for Hinglish religious hate speech (Srivastava, 2025), and soft prompt tuning reduces MER by 10%+ in Mandarin-English speech recognition (Liu et al., 2025). LoRA improves Hinglish NER by 5–10% despite transliteration challenges (Shirke et al., 2025). Adapters and quantization-aware PEFT lower resource demands for safety evaluation and constrained generation but require careful hyperparameter tuning (Goloburda et al., 2025; Kuwanto et al., 2024). Overall, PEFT provides scalable, resource-efficient adaptation for code-switched LLMs (Weyssow et al., 2024,?).

Takeaways: PEFT techniques like LoRA, QLoRA, and soft prompt tuning efficiently adapt LLMs to code-switched and low-resource multilingual settings with minimal parameters and hardware demands. They consistently improve performance in tasks like hate speech detection, NER, and speech recognition, while reducing computational cost.

Reinforcement Learning for CSW Adaptation To improve LLMs' code-mixing understanding, reinforcement learning from AI feedback (RLAIF) has been proposed as a cost-efficient alternative to human annotation, effectively enhancing code-mixed translation (Zhang et al., 2023). CHAI applies RLAIF on Llama-3.1-8B-Instruct for English-Hinglish translation, generating preference pairs via GPT-4o on 15,190 MixMT and ALL-CS instances, and uses PPO for 5 epochs to maximize reward while controlling KL divergence, achieving 40–46% higher human-evaluated win rates, 24–81% gains in chrF and COMET translation metrics, and 14.12% accuracy improvement in Hinglish sentiment analysis (Zhang et al., 2025c). Similarly, (Zhang et al., 2025c) enhances multilingual LLMs' handling of code-mixing using RLAIF and code-mixed machine translation, reducing reliance on human annotations by leveraging LLMs for preference labeling. (Heredia et al., 2025b) employs RL-based policy optimization with back-translation of naturally occurring CS sentences to create synthetic data, improving generation naturalness and fluency by op-

timizing for acceptability scores (Heredia et al., 2025b).These methods enhance CSW fluency and alignment, though typological diversity and computational cost remain challenges (Zhang et al., 2025c).

Takeaway RLAIF and RL-based methods for synthetic data generation represent promising but underexplored approaches to enhance LLMs code-mixed understanding while reducing reliance on human annotations.

Supplementary Material: This section provides additional resources to support our main findings, including extended tables, illustrative examples of model hallucinations, and dataset analyses for code-mixed NLP research.

Figure 3 shows a ChatGPT hallucination where a query for vacation countries results in a list of Indian states and union territories. The response also exhibits incorrect grammar, demonstrating the model's limitations in handling nuanced, code-mixed queries.

Figure 4 shows an illustration of hallucination in the open-source model GLM-4.6. When prompted to perform morphophonemic (sandhi) analysis of the Bengali word Ohoroho ("always") in Bengali-Romanized Hindi script, the model misinterprets the query and generates an incorrect response.

Figure 5 shows a failure case from Perplexity AI. When prompted in a code-mixed query for Goan temples specifically destroyed by the Portuguese, the model disregards this constraint and instead provides information on only existing ancient temples and not the destroyed ones.

Figure 6 presents a dual analysis of dataset sizes. The bar chart on the left compares the number of instances in prominent datasets, showing COMI-LINGUA (125,615) and MEGAVERSE (100,000) as the largest. The pie chart on the right categorizes the overall distribution, revealing that 'Large' datasets (50-100k instances) are the most common category, comprising 40% of the analyzed collections.

Figure 7 chart displays the prevalence of different language pairs in code-switching studies, measured by the number of available datasets and benchmarks. The data clearly indicates a strong dominance of Hindi-English and Spanish-English combinations, highlighting a significant research focus on these pairs compared to others.

Figure 8 illustrates the primary focus areas within code-switching NLP research. The left pie chart details the distribution of specific tasks, with SA (26.2%) and MT (18.5%) being the most studied. The right pie chart groups these into broader categories, where 'Other' (55.9%) and 'Understanding' (31.1%) tasks represent the vast majority of research efforts.

Figure 9 shows the distribution of language pairs in 202 code-mixing related datasets and benchmarks papers, highlighting the prevalence of specific language combinations in code-switching NLP research.

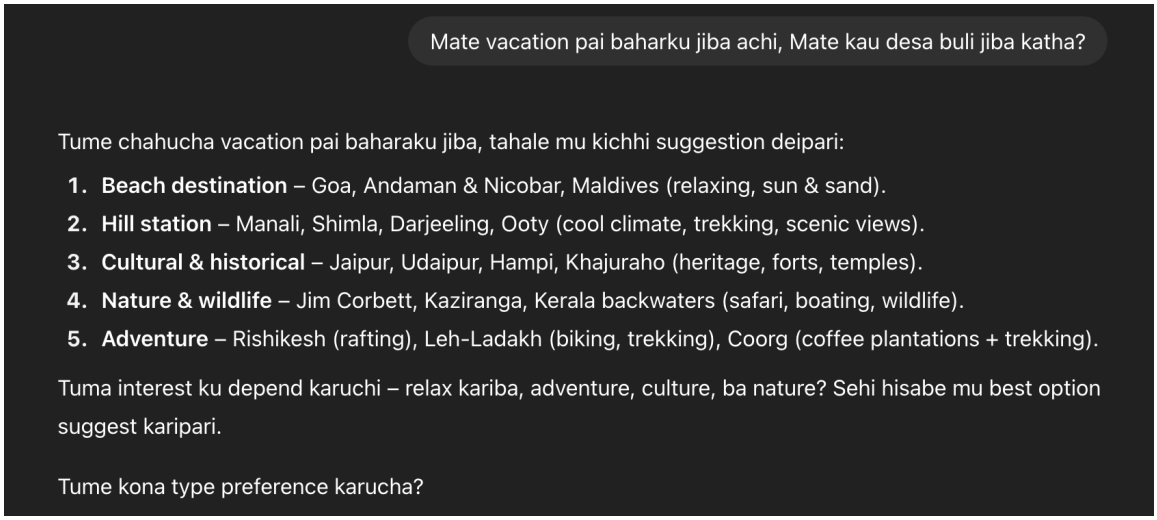

Figure 3: Failure cases when we prompt ChatGPT in Odia-Romanized Hindi code-mixed pair.

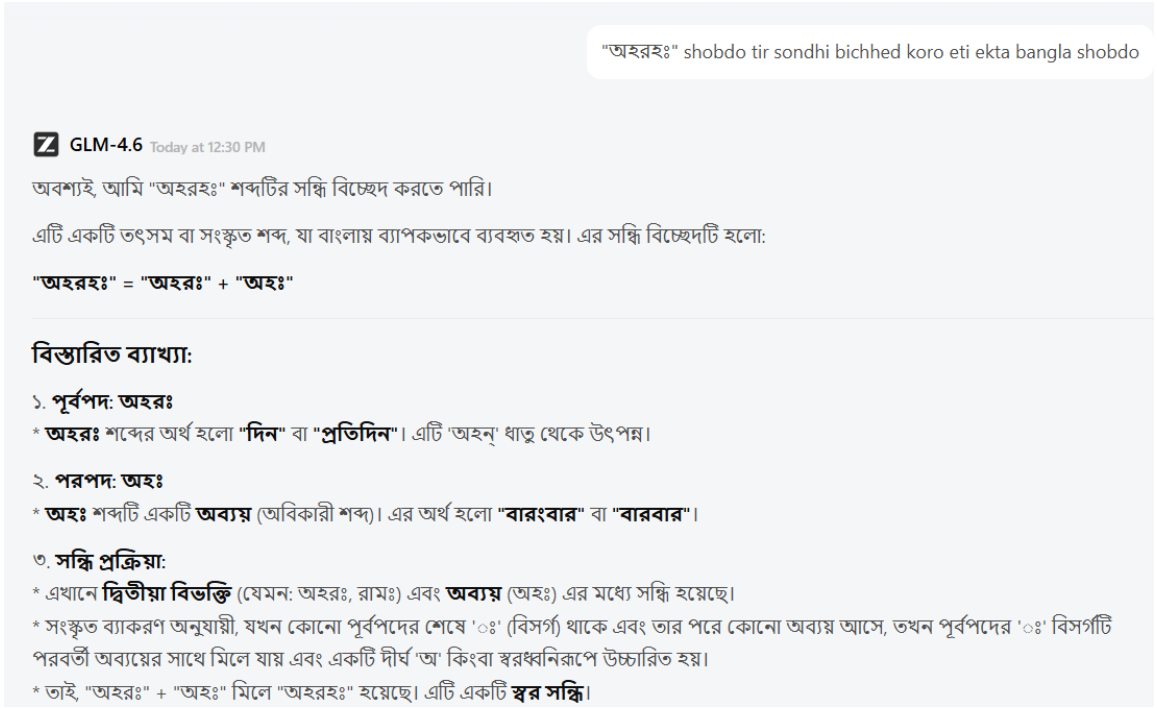

Figure 4: Failure cases when we prompt GLM-4.6 in Bangla-English code-mixed pair.

Figure 5: Failure cases when we prompt Perplexity in Konkani-English code-mixed pair.

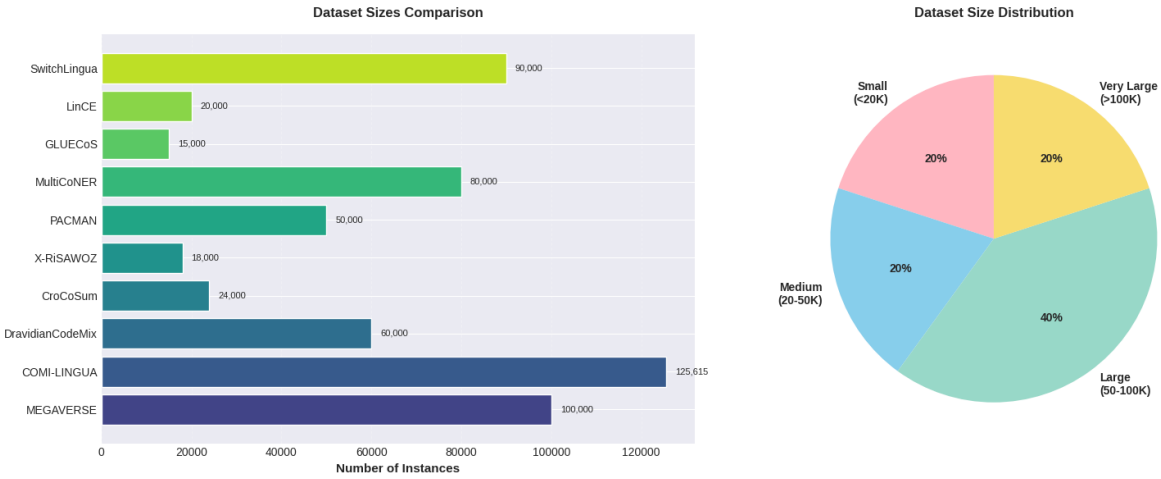

Figure 6: Analysis of Code-Switching Dataset Sizes.

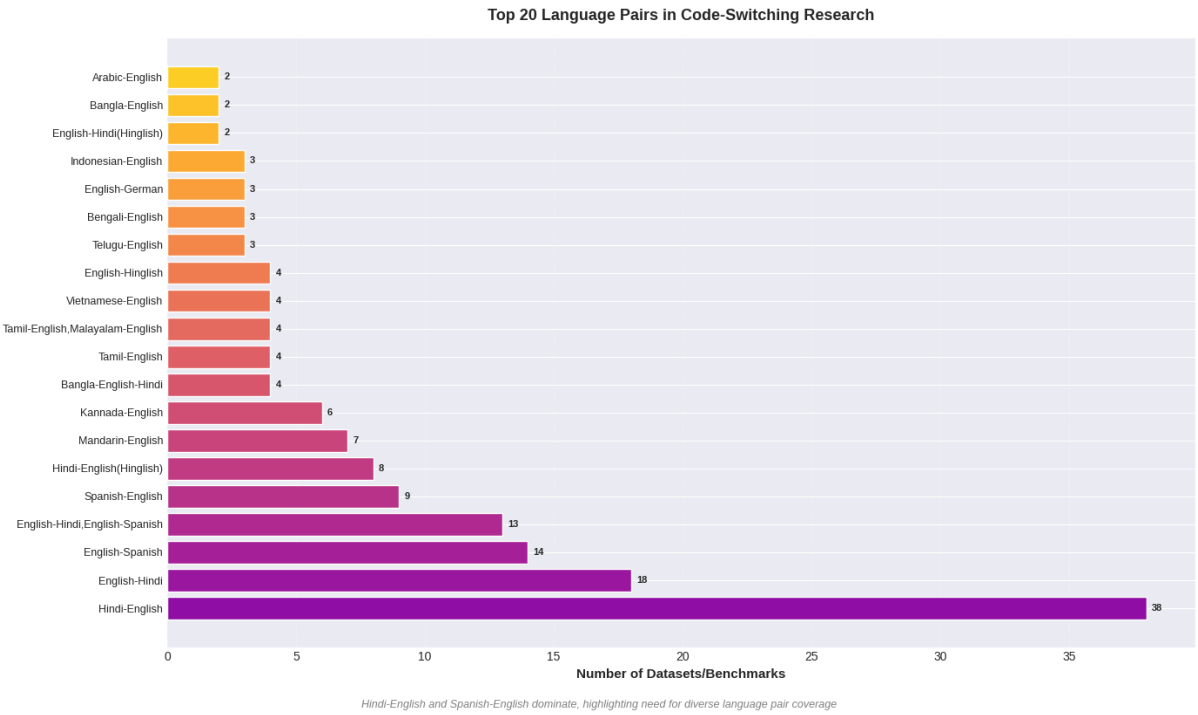

Figure 7: Top 20 Language Pairs in Code-Switching Research

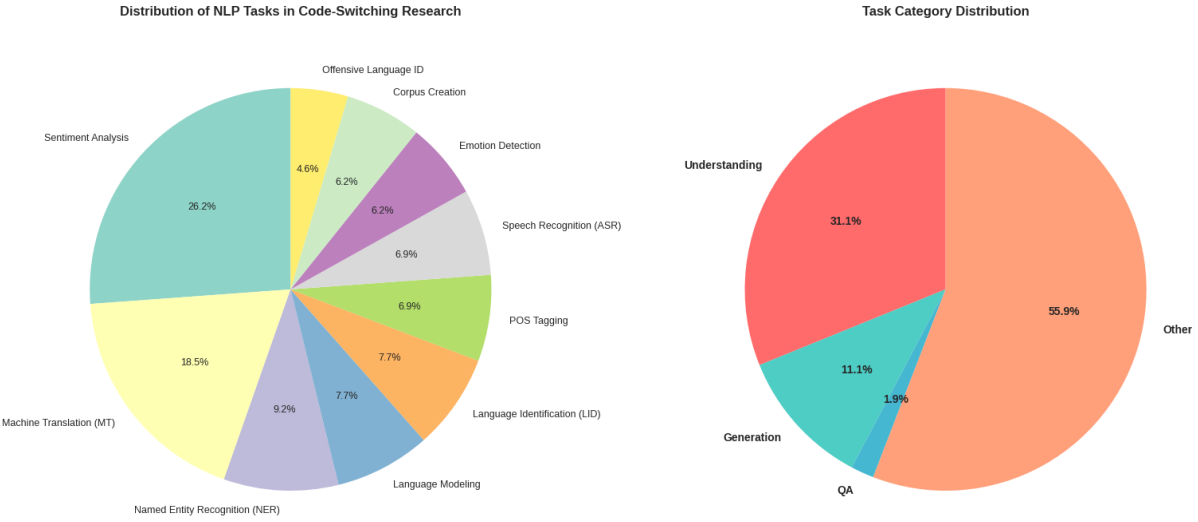

Figure 8: Distribution of NLP Tasks and Categories in Code-Switching Research.

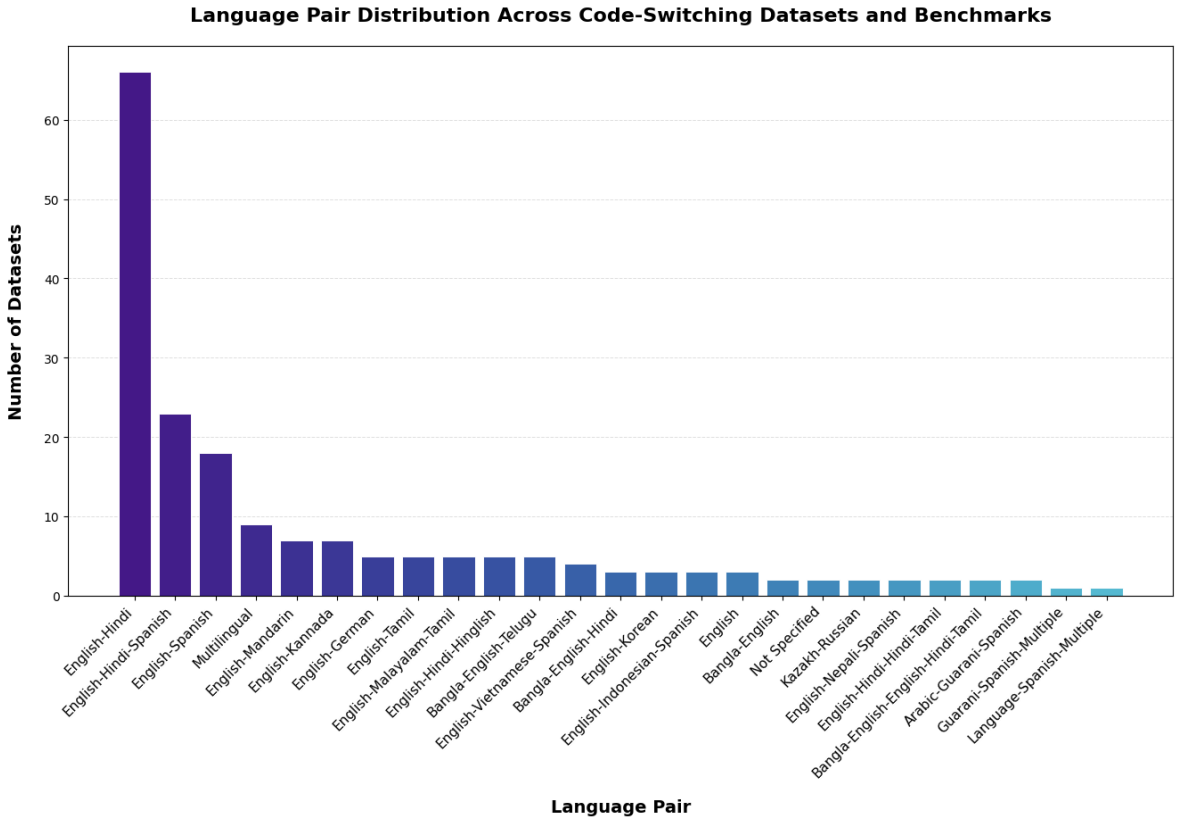

Figure 9: Language pair distribution across 202 code-mixing related datasets and benchmarks papers.

Category	Tasks/Components	Datasets & Benchmarks	Model Training & Adaptation	Evaluation & Methodology	Key Insights
NLU Tasks: Foundational	LID (MasakhaNER, ECT); POS tagging (XLM-R, IndicBERT) (Absar, 2025) ; NER (MELM, GLUECoS) (Zhou et al., 2022; Khanuja et al., 2020b); Syntactic parsing (CoStALL LLM)	LinCE (Aguilar et al., 2020), GLUECoS (Khanuja et al., 2020b), MultiCoNER (Malmasi et al., 2022b), PACMAN (50K+ samples) (Chatterjee et al., 2022), WILDRE (Kumar et al., 2022a)	XLM-RoBERTa (F1: 0.88) (Kochar et al., 2024) ; MuRIL (5-8% > mBERT) (Goswami et al., 2023); IndicBERT (F1: 0.82) (Tatariya et al., 2023); Traditional: CRF, SVM	Benchmarks: GLUECoS (6 tasks) (Khanuja et al., 2020b), LinCE (Aguilar et al., 2020); Metrics: F1, accuracy; Cross-lingual transfer	XLM-R on MultiCoNER (F1: 0.88) (Malmasi et al., 2022b); MuRIL (Goswami et al., 2023) best for Indic; handles script mixing natively
NLU Tasks: QA & Understanding	Question Answering (COMMIT) (Lee et al., 2024); Information Retrieval; Reading Comprehension; Semantic Understanding	COMMIT-based datasets (Lee et al., 2024); MLQA (multilingual QA) (Lewis et al., 2020); Low-resource QA benchmarks	COMMIT (LLaMA-7B) (Lee et al., 2024): **32x improvement** on MLQA (Lewis et al., 2020); Code-mixed instruction tuning; Template preservation with content mixing	Metrics: Exact Match, F1; Cross-lingual transfer; Few-shot evaluation	COMMIT (Lee et al., 2024) 32x improvement on MLQA (Lewis et al., 2020); demonstrates power of code-mixed instruction tuning
NLG Tasks	MT (CoMix) (Arora et al., 2023); Summarization (CroCoSum) (Zhang and Eickhoff, 2024); Dialogue Generation (X-RiSAWOZ: 18K+ utterances, 12 domains) (Moradshahi et al., 2023); Creative text (puns); CS prompting	CroCoSum (24K En articles, 18K Chinese summaries) (Zhang and Eickhoff, 2024); X-RiSAWOZ (En, Fr, Hi, Ko, En-Hi) (Moradshahi et al., 2023); IndioTrans; HinglishEval (Srivastava and Singh, 2022b)	CoMix (Arora et al., 2023): BLEU 12.98, **10x smaller** than mT5; GPT-4 strong zero-shot generation; Flan-T5-XXL (Yong et al., 2023) limited CS capability	Metrics: BLEU, ROUGE, CS-MI; Task-specific: DST, RG; Quality estimation (HinglishEval) (Srivastava and Singh, 2022b); Integration: transliteration, Indic-TTS	CoMix (Arora et al., 2023) BLEU: 12.98 with 10x efficiency; GPT-4 excels at creative generation (puns, translation)
NLU Tasks: Classification	Sentiment (DravidianCodeMix: 60K+ comments) (Chakravarthi et al., 2022); Offensive detection (OffMix-3L) (Goswami et al., 2023); Hate speech; Intent classification	DravidianCodeMix (Tamil-En, Kannada-En, Malayalam-En) (Chakravarthi et al., 2022); OffMix-3L (Goswami et al., 2023); SentMix-3L (Raihan et al., 2023a); MEGAVERSE (83 languages) (Ahuja et al., 2024)	Sentence-BERT SetFit (F1: 0.72, minimal data) (Pannerselvam et al., 2024); mBERT (F1: 0.85+ baseline); GPT-4 zero-shot strong	MEGAVERSE (Ahuja et al., 2024) holistic evaluation; Metrics: perplexity, grammaticality; Social media analysis workflows	SetFit (Pannerselvam et al., 2024) F1: 0.72 with minimal data for Dravidian (Chakravarthi et al., 2022); GPT-4 strong on SentMix/OffMix-3L (Raihan et al., 2023a; Goswami et al., 2023)
Specialized Models & Techniques	Pre-training: HingBERT (L3Cube-HingCorpus) (Nayak and Joshi, 2022); Instruction tuning; Domain adaptation; POS-guided attention	Synthetic: SynCS, LLM-generated; Low-resource: WILDRE (Magahi-En, Maithili-En) (Rani et al., 2024b); Language-specific corpora	HingBERT (Hinglish specialist, Roman script tokenizer) (Nayak and Joshi, 2022); CoMix (POS-guided + phonetic signals) (Arora et al., 2023); Flan-T5-XXL (baseline, fails at CS) (Yong et al., 2023)	Pre-training metrics; Multi-task adaptation; Phonetic feature integration; Zero-shot vs fine-tuned comparison	HingBERT outperforms mBERT on Hinglish; CoMix 10x efficiency; specialized tokenization critical
Benchmarks & Evaluation	Foundational: LID, POS, NER; Understanding: sentiment, QA; Generation: MT, summarization; Novel: code generation (CodeMixBench) (Sheokand et al., 2025)	COMI-LINGUA (125K instances, Roman+Devanagari) (Sheth et al., 2025); SwitchLingua (next-gen diversity) (Xie et al., 2025); HinglishEval (quality estimation) (Srivastava and Singh, 2022b); CodeMixBench (programming from CS) (Sheokand et al., 2025)	Evaluation across XLM-R, MuRIL, IndicBERT, GPT-4, mBERT; Zero-shot vs fine-tuned performance	Traditional: F1, BLEU; Task-specific: Prothaba; Intrinsic: perplexity; Extrinsic: GAME; Quality estimation for synthetic CS	MEGAVERSE (83 langs) (Ahuja et al., 2024); COMI-LINGUA (Hinglish) (Sheth et al., 2025); CodeMixBench (Sheokand et al., 2025) introduces novel programming task
Language Coverage & Gaps	High-resource: Hindi-En, Spanish-En; Dravidian: Tamil-En, Kannada-En, Malayalam-En; Low-resource: Magahi-En, Maithili-En; East Asian: Chinese-En	SwitchLingua (Xie et al., 2025) (multi-ethnic); X-RiSAWOZ (5 languages) (Moradshahi et al., 2023); MuRIL corpus (17M Indian) (Goswami et al., 2023); MultiCoNER (11 languages) (Malmasi et al., 2022b)	Universal: XLM-R (100 langs), mBERT (104 langs); Regional: MuRIL (17 Indic), IndicBERT (12 Indic) (Goswami et al., 2023); GPT-4 (inconsistent on low-resource)	Cross-lingual transfer; Probing tasks; Zero-shot assessment; Task-oriented dialogue evaluation (X-RiSAWOZ) (Moradshahi et al., 2023)	Low-resource pairs underserved; script mixing only by specialists; unseen CS patterns challenge all models

Table 1: Comprehensive landscape of code-switching NLP: Tasks, datasets, models, and evaluation across all categories. **Takeaway** (1) Specialized models (MuRIL, IndicBERT) outperform general-purpose on specific language families; (2) Instruction tuning yields dramatic gains (COMMIT: 32x); (3) Efficiency matters (CoMix: 10x smaller, better performance); (4) Few-shot learning viable (SetFit: F1 0.72 minimal data); (5) Major gaps: low-resource languages, unseen CS patterns, generation consistency.

Task	Dataset	Languages	Domain	Key Characteristics
NER	SemEval-2022 Task 11 (Malmasi et al., 2022a)	Multilingual	Short queries	6 entity types (Person, Location, Group, Corp, Product, Creative Work)
	Kannada-English NER (S and Shrivastava, 2022)	Kannada-English	Social media	Low-resource Dravidian language; user-generated content
	TB-OLID (Raihan et al., 2023b)	Bangla-English	Social media	5,000 Facebook comments; transliterated text; hierarchical annotation
Machine Translation	WMT 2022 MixMT (Srivastava and Singh, 2022c)	Hindi-English	General	Bidirectional translation (EN↔Hinglish); shared task with standardized metrics
	CoMeT Corpus (Gautam et al., 2021b)	Multiple pairs	General	Synthetic generation using parallel monolingual sentences
	AfroCS-xs (Olaleye et al., 2025)	4 African + EN	Agriculture	Human-validated synthetic data for low-resource African languages
	CoVoSwitch (Kang, 2024)	Multiple pairs	Synthetic/Speech	Prosody-aware synthetic data generation for MT based on intonation units
Dialogue	GupShup (Mehnaz et al., 2021)	Hindi-English	Entertainment	First code-mixed dialogue summarization; movie-based conversations
	Multilingual WOZ 2.0 (Hung et al., 2022)	Multilingual	Restaurant	Cross-lingual dialogue state tracking benchmark
	THAR (Srivastava, 2025)	Hindi-English	Social media	11,549 YouTube comments; religious hate speech detection
Emotion/ Sentiment	EmoMix-3L (Raihan et al., 2024)	Bangla-EN-Hindi	Social media	1,071 instances; 5 emotion classes (Happy, Surprise, Neutral, Sad, Angry)
	SentMix-3L (Raihan et al., 2023a)	Bangla-EN-Hindi	Social media	1,007 instances; 3-class sentiment; controlled trilingual collection
	OffMix-3L (Goswami et al., 2023)	Bangla-EN-Hindi	Social media	1,001 instances; offensive language identification; first trilingual dataset
ASR/Speech	ASCEND (Lovenia et al., 2022)	Mandarin-English	Conversational	Hong Kong speakers; location-specific variation documented
	SEAME (Lyu et al., 2010)	Mandarin-EN-Hokkien	Conversational	Singapore speakers; includes three-language mixing
	English-isiZulu CS (Biswas et al., 2020)	English-isiZulu	Conversational	Semi-supervised acoustic and language modeling; low-resource African

Table 2: Top specialized datasets by task category.

Model / Method	Type	Target Scenario	Key Innovation & Results
HingBERT (Nayak and Joshi, 2022)	Pre-trained encoder	Hindi-English (Hinglish)	Pre-trained on L3Cube-HingCorpus; outperforms mBERT on Hinglish sentiment, NER; specialized tokenizer for Roman script Hindi
COMMIT (LLaMA-7B based) (Lee et al., 2024)	Instruction-tuned LLM	Low-resource QA	Code-mixed instruction tuning; 32x improvement on MLQA; specialized code-mixing preserves templates while mixing content
CoMix (BERT/BART variants) (Arora et al., 2023)	Domain-adapted transformer	Hinglish translation, NER	POS-guided attention; phonetic signals; SOTA on LINCE (BLEU: 12.98); 10x smaller than mT5 with better performance
Flan-T5-XXL (Yong et al., 2023)	Multilingual LLMs	Zero-shot generation	Limited code-mixing capability; fails at code-mixing; useful baseline but needs fine-tuning for most tasks
Sentence-BERT (SetFit) (Pannerselvam et al., 2024)	Few-shot classifier	Low-resource offensive detection	Achieves F1: 0.72 with minimal labeled data; efficient for Tamil, Telugu offensive content; practical for data-scarce scenarios

Table 3: Specialized models and adaptation techniques for code-switching.

Model	Methodology (brief)	Strengths	Weaknesses
XLM-RoBERTa (Kochar et al., 2024)	Multilingual masked LM trained on 2.5TB CommonCrawl; 100 languages; RoBERTa architecture	+ SOTA on MultiCoNER (F1: 0.88), OffMix-3L; excellent zero-shot transfer; robust cross-lingual representations	– Struggles with unseen code-switching patterns; requires fine-tuning for best results; computationally expensive
MuRIL (Goswami et al., 2023)	BERT pre-trained on 17 Indian languages + transliterated text ; 17M Indian corpus	+ Best for Indic tasks; handles script mixing; 5-8% better than mBERT on Hindi-English	– Limited to Indian subcontinent; less effective for other language families; smaller coverage
mBERT (Goswami et al., 2023)	Multilingual BERT on 104 Wikipedia dumps; shared vocabulary	+ Strong baseline (F1: 0.85+); widely adopted; stable performance across tasks	– Curse of multilinguality; undertrained on low-resource languages; outperformed by specialized models
GPT-4 (Ahuja et al., 2024)	Decoder-only transformer; web-scale training; RLHF alignment	+ Strong zero-shot on SentMix-3L, OffMix-3L; excellent generation (puns, translation); few-shot learning	– Closed-source; expensive API; inconsistent on low-resource pairs; unpredictable behavior
IndicBERT (Tatariya et al., 2023)	BERT on 12 Indian languages; 9GB Indic corpus; language-specific tokenization	+ Best for Indian monolingual + code-mixed tasks; F1: 0.82 on Dravidian-CodeMix; efficient	– Limited to 12 languages; requires language-specific tuning; less generalizable than XLM-R

Table 4: Top 5 models for code-switching NLP with methodology and performance characteristics.

Benchmark	Description	Language Pairs	Key Focus
LinCE (Aguilar et al., 2020)	A centralized benchmark that aggregated several existing datasets (e.g., for LID, POS, NER) to standardize evaluation of foundational linguistic tasks.	Spanish-En, Nepali-En, Hindi-En, etc.	Foundational NLU Tasks
GLUECoS (Khanuja et al., 2020b)	The first multi-task benchmark suite for code-switching, modeled after the influential GLUE benchmark for English, covering 6 diverse tasks.	Hindi-En, Spanish-En	Multi-task NLU Evaluation
DravidianCodeMix (Chakravarthi et al., 2022)	Large-scale dataset with over 60,000 YouTube comments for sentiment analysis and offensive language identification in Dravidian languages.	Tamil-En, Kannada-En, Malayalam-En	Sentiment & Offensive Language
MultiCoNER (Malmasi et al., 2022b)	A large-scale multilingual benchmark for "complex" NER, featuring nested entities and code-mixing across 11 languages.	11 languages	Complex NER
PACMAN (Chatterjee et al., 2022)	A synthetically generated code-mixed POS-tagged dataset with over 50K samples, outperforming existing benchmarks in code-mixed POS tagging.	Hindi-En	POS Tagging
WILDRE Shared Tasks (Kumar et al., 2022a)	A series of shared tasks focused on promoting research for less-resourced Indo-Aryan languages on tasks like sentiment analysis.	Magahi-En, Maithili-En, etc.	Low-Resource Languages
CroCoSum (Zhang and Eickhoff, 2024)	A dataset for cross-lingual code-switched summarization of technology news, with over 24,000 English source articles and 18,000 human-written Chinese summaries.	English–Chinese	Cross-lingual Summarization
X-RiSAWOZ (Moradshahi et al., 2023)	A multi-domain, large-scale, high-quality task-oriented dialogue benchmark created by translating the Chinese RiSAWOZ dataset to English, French, Hindi, Korean, and a code-mixed English-Hindi language. Contains over 18,000 utterances per language across 12 domains.	En, Fr, Hi, Ko, En-Hi (code-mixed)	End-to-end Task-oriented Dialogue (DST, ACD, DAG, RG)
SwitchLingua (Xie et al., 2025)	A recent, large-scale multilingual and multi-ethnic dataset designed to be the next generation of code-switching benchmarks.	Multiple	Diversity & Inclusivity
MEGAVERSE (Ahuja et al., 2024)	A massive-scale, multi-modal benchmark covering 83 languages, designed for a holistic and challenging evaluation of modern LLMs.	83 languages	Broad LLM Capabilities
HinglishEval (Srivastava and Singh, 2022b)	A generation challenge focused on the difficult task of Quality Estimation for synthetically generated code-mixed text.	Hindi-En	Generative Evaluation
CodeMixBench (Sheokand et al., 2025)	The first benchmark for the novel and non-traditional task of generating programming code from code-mixed natural language prompts.	Hindi-En, Spanish-En	Non-traditional Tasks
COMI-LINGUA (Sheth et al., 2025)	Expert-annotated 1,25,615 Hinglish code-mixed instances across five core NLP tasks, covering both Roman and Devanagari scripts.	Hi-En	Multi-task: LID, MLI, POS, NER, MT

Table 5: An expanded overview of major benchmarks for standardized evaluation of code-switching NLP models.

Dataset	Description	Language Pairs	Tasks
Hinglish Blog Corpus (Chandu et al., 2018b)	A corpus from eight Hinglish blogging websites covering diverse topics, containing 59,189 unique sentences.	Hindi-En	POS Tagging, Language Modeling
CS-NLI (Chakravarthy et al., 2020)	A dataset for Hinglish conversational entailment derived from Bollywood movie scripts, containing 2,240 premise-hypothesis pairs.	Hindi-En	Natural Language Inference (NLI)
Bollywood NLI (Khanuja et al., 2020a)	A conversational NLI dataset from 18 Bollywood movie scripts, with 400 premises and 2,240 hypotheses.	Hindi-En	Natural Language Inference (NLI)
DravidianCodeMix (Chakravarthi et al., 2022)	A manually annotated dataset for sentiment and offensive language detection in Dravidian languages. Contains over 60,000 sentences.	Tamil-En, Mal.-En, Kan.-En	Sentiment, Offensive ID
HinGE (Srivastava and Singh, 2021)	A high-quality, manually annotated Hindi-English dataset for NLG tasks, with 3,872 sentences annotated by five annotators.	Hindi-En	Natural Language Generation (NLG)
Bengali Abusive Comments (Sazzed, 2021)	A balanced corpus of 3,000 YouTube comments (1,500 abusive, 1,500 non-abusive) in transliterated Bengali-English.	Bengali-En	Abusive Language Detection
MultiCoNER (Malmasi et al., 2022b)	A large-scale multilingual benchmark for "complex" NER, featuring nested entities and code-mixing across 11 languages.	11 languages	Complex Named Entity Recognition
ASCEND (Lovenia et al., 2022)	A large-scale dataset (10.3 hours) of spontaneous, multi-turn spoken conversational code-switching, capturing natural dialogue.	Mandarin-En	Dialogue, ASR
GupShup (Mehnaz et al., 2021)	The first large dataset for abstractive summarization of informal, code-switched conversations. Contains 6,800 conversations.	Hindi-En	Dialogue Summarization
TweetTaglish (Imperial and Ong, 2022)	A large dataset of 78,689 Tagalog-English code-switched tweets, expanding resources for Southeast Asian languages.	Tagalog-En	Language ID, etc.
My Boli (Joshi et al., 2023)	A comprehensive suite of corpora and pre-trained models for the under-resourced Marathi-English language pair.	Marathi-En	General NLU
OffMix-3L / EmoMix-3L (Goswami et al., 2023; Raihan et al., 2023b)	The first datasets for affective tasks involving a mix of three languages simultaneously, created from social media.	Bangla-En-Hi	Offensive ID, Emotion
EkoHate (Ilevbare et al., 2024)	A dataset of 3,398 tweets for hate speech detection in Nigerian English and Pidgin political discussions.	Nigerian En-Pidgin	Hate Speech Detection
MaCmS (Rani et al., 2024a)	The first Magahi-Hindi-English code-mixed dataset for sentiment analysis, focusing on a less-resourced minority language.	Magahi-Hindi-En	Sentiment Analysis
Prabhupadavani (Kumar et al., 2024)	A massive and unique code-mixed speech translation dataset covering 25 languages, derived from religious discourses.	25 Indic Lang.-En	Speech Translation
KRCS (Borisov et al., 2025)	The first Kazakh-Russian code-switching parallel corpus for machine translation, containing 618 parallel sentences.	Kazakh-Russian	Machine Translation
BarNER (Peng et al., 2024)	The first dialectal NER dataset for German, containing 161,000 tokens annotated from Bavarian Wikipedia and tweets.	Bavarian-German	Named Entity Recognition (NER)
BanglishRev (Shamael et al., 2024)	A large-scale (23,546 reviews) Bangla-English code-mixed dataset of e-commerce product reviews.	Bengali-En	Sentiment Analysis
BnSentMix (Alam et al., 2025)	A large (>20,000 sentences) code-mixed Bengali dataset from diverse sources (Facebook, YouTube, e-commerce).	Bengali-En	Sentiment Analysis
MMS-5 (Arul et al., 2025)	A multi-modal, multi-scenario hate speech dataset for Dravidian languages, including code-mixed text and images.	Tamil-En, Kannada-En	Multimodal Hate Speech
Qor'gau (Goloburda et al., 2025)	The first comprehensive safety evaluation dataset for bilingual Kazakh-Russian contexts, with 500 code-switched examples.	Kazakh-Russian	LLM Safety Evaluation
AfroCS-xs (Olaleye et al., 2025)	A compact, human-validated synthetic dataset for four African languages, demonstrating the power of high-quality data over quantity.	4 African Lang.-En	Machine Translation
DweshVaani (Srivastava, 2025)	An LLM fine-tuned on Gemma-2 for detecting religious hate speech in code-mixed Hindi-English.	Hindi-En	Religious Hate Speech Detection
Cline (Kodali et al., 2025)	The largest dataset of human acceptability judgments for Hinglish code-mixed text, showing low correlation with traditional metrics.	Hindi-En	Acceptability Judgments
Hindi-Marathi CS Corpus (Hemant and Narvekar, 2025)	A 450-hour annotated code-switched speech dataset covering various switching patterns for ASR and LID.	Hindi-Marathi	ASR, Language ID
Word-Level Hate Speech (Niederreiter and Gromann, 2025)	A multilingual dataset for word-level hate speech detection, exploring domain transfer across code-mixed language pairs.	Hi-En, De-En, Es-En	Word-Level Hate Speech

Table 6: An expanded overview of major datasets for code-switching NLP, highlighting language pairs, tasks, and scale.