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

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

Amidst the rapid advances of large language models (LLMs), most LLMs still struggle with mixed-language inputs, limited Code-switching (CSW) datasets, and evaluation biases, which hinder their deployment in multilingual societies. This survey provides the first comprehensive analysis of CSW-aware LLM research, reviewing 327 studies spanning five research areas, 15+ NLP tasks, 30+ datasets, and 80+ languages. We categorize recent advances by architecture, training strategy, and evaluation methodology, outlining how LLMs have reshaped CSW modeling and identifying the challenges that persist. The paper concludes with a roadmap that emphasizes the need for inclusive datasets, fair evaluation, and linguistically grounded models to achieve truly multilingual capabilities.[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. Globally, approximately 43% of the population is bilingual and an additional 13% is trilingual (Preply, 2022; Stone, 2025), representing over 4.5 billion multilingual speakers. Despite this prevalence, Monolingual ASR systems struggle with code-switched input: word error rates increase by 30–50% (Singh et al., 2025). Even multilingual models show semantic accuracy drops of 15% (Winata et al., 2021), revealing a fundamental architectural gap. Similar challenges are observed in multilingual regions including India, Nigeria, and South Africa, where frequent CSW undermines monolingual ASR performance (Babatunde et al., 2025). 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) along with neural architectures, advancing CSW tasks like LID, POS, 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 modeling (Winata et al., 2021). The *Foundation Model Era* (2020–present) leverages massive, instruction-tuned LLMs like GPT-3 and LLaMA capable of general-purpose reasoning through multilingual pretraining and prompt-based adaptation (Wang et al., 2025b).

LLMs have transformed CSW investigation across typologically diverse language pairs, including Arabic-English (Issa et al., 2025), Cantonese-Mandarin (Dai et al., 2025), Chinese-English (Kong and Macken, 2025), Hinglish (Sheth et al., 2025), Korean-English (Yoo et al., 2025), Spanish–Guaraní (Kellert et al., 2025), and Ukrainian-Russian (Shynkarov et al., 2025), deepening our linguistic and sociocultural understanding of switching patterns (Yoo et al., 2024; Jehan et al., 2025). These advances are enabled by methodological innovations in LLMs, including in-context mixing (Shankar et al., 2024), instruction tuning (Lee et al., 2024), speech pro-

[*]Work done while interning at IIT Gandhinagar.

[1]A curated collection of all resources is maintained at `https://github.com/lingo-iitgn/awesome-code-mixing/`.

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 8519–8566
July 2-7, 2026 ©2026 Association for Computational Linguistics

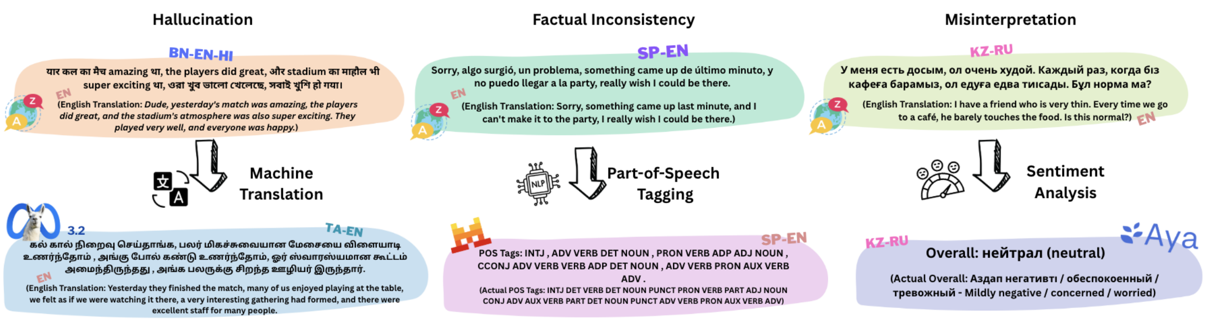

Figure 1: Common model failures on code-mixed text across tasks: ***Takeaway*** *(a) hallucination in MT for a constructed trilingual (Bn-Hi-En) example, highlighting underexplored settings; (b) factual inconsistency in POS tagging for Spanish-English data Zeng (2024); and (c) misinterpretation in sentiment analysis for Kazakh-Russian text Goloburda et al. (2025b).*

cessing (Kang, 2024), advanced metrics evaluating structural and socio-pragmatic aspects (Ugan et al., 2025; Sterner and Teufel, 2025b), and curriculum strategies for transfer (Yoo et al., 2025).

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), followed by limited pretraining data for low-resource languages (Yoo et al., 2025). Prior works, including (Winata et al., 2023a) and (Sitaram et al., 2019), have laid essential groundwork for understanding CSW in NLP but focus primarily on pre-LLM approaches and text-centric settings. Building on these foundations, we position this work as the first comprehensive survey of CSW NLP in the LLM era, covering recent advances in instruction tuning, PEFT, and multimodal (speech and vision-language) settings. We further survey the evolution of CSW research in the LLM era and present this unified taxonomy (Figure 2), with a detailed analysis in Appendix §B that categorizes prior work into five key research directions.

Position We posit that Code-switching should be treated as a core modelling challenge rather than a downstream artefact of multilinguality. In low-resource CSW settings, current monolingual and multilingual models struggle to cope as language switching interacts with script alternation, orthographic variation, and spontaneous conversational behavior, which remain insufficiently represented in existing models and benchmarks across speech, dialogue, and multimodal contexts. Bridging this gap requires structured data creation, scalable modeling, and human-in-the-loop annotation, reinforced by language- and script-aware CS-specific architectures rather than broad, undifferentiated, multilingual post-training.

Contributions The key contributions include: **(i)** We provide the first comprehensive survey of CSW research in the LLM era, analyzing 327 studies across 15 NLP tasks, 30+ datasets spanning 80+ languages, across diverse language pairs, real-world applications, and key architectural innovations. **(ii)** We present a taxonomy (Appendix §B Figure 2) organizing LLM-based CSW research by architecture, training paradigm, and evaluation, while revealing key gaps in low-resource coverage, script diversity, cross-lingual transfer, the absence of unified evaluation frameworks. **(iii)** We present a roadmap for future CSW research, highlighting the need for inclusive datasets, equitable models, and fair metrics to support linguistically grounded advances in dialogue, speech, and multimodal contexts.

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 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 across speech processing, conversational and generation tasks, motivating surveys to examine their adaptation, emerging trends, and future directions.

3 Code-Switching Task Landscape: Capabilities and Gaps

3.1 Traditional Tasks

The integration of LLMs into traditional NLP tasks has revealed both transformative capabilities and inherent limitations in CSW contexts. In **language identification**, innovative fine-grained techniques such as TongueSwitcher for boundary detection in morphologically mixed German-English words (Sterner and Teufel, 2023), MaskLID for training-free iterative identification of subdominant languages (Kargaran et al., 2024), and equivalence constraint-guided methods for grammatical switch points (Kuwanto et al., 2024) have established new benchmarks, with applications extending to hope/offensive speech detection and efficient zero/few-shot adaptation via models like COOLI and SetFit (Ahmad et al., 2025; Balouchzahi et al., 2021; Pannerselvam et al., 2024). **Part-of-Speech tagging** has benefited from contextual embeddings (e.g., mBERT on Arabic-English and Hinglish (Sabty et al., 2020; Aguilar and Solorio, 2020)), bilingual pretraining on datasets like GLUECoS (Winata et al., 2021; Prasad et al., 2021), parallel synthetic data generation (PACMAN (Chatterjee et al., 2022)), prompt-based CSW synthesis (PRO-CS and CoMix (Bansal et al., 2022; Arora et al., 2023; Kumar et al., 2022)), and S-index-augmented XLM-R fine-tuning (Absar, 2025). **Named Entity Recognition** has evolved from early embedding-attention approaches for Spanglish tweets (Wang et al., 2018) to synthetic CSW pretraining with MELM (Zhou et al., 2022) and two-stage CMB models (Pu et al., 2022), supported by benchmarks like MultiCoNER and toolkits such as CodemixedNLP (Malmasi et al., 2022b; Jayanthi et al., 2021). Advances, including contextualized embeddings (Sabty et al., 2020), pseudo-labeling (El Mekki et al., 2022), switch-point–biased self-training (Chopra et al., 2021), prompt-based methods like PRO-CS (Bansal et al., 2022), and data augmentation in CoSDA-ML (Qin et al., 2020), have further enhanced zero-shot transfer capabilities.

Despite task-specific advances, multilingual models underperform on code-mixed inputs compared to monolingual settings (Wang et al., 2025b), primarily due to the limited representativeness of CSW in pretraining data (Doğruöz et al., 2023). These issues are further aggravated in low-resource and typologically distant languages (Sravani and Mamidi, 2023). Key challenges include modeling ambiguous switch points (Chopra et al., 2021) and mitigating hallucinations in generative CSW outputs (Wang et al., 2025a). While prompting techniques, zero-shot models (e.g., GLiNER (Zaratiana et al., 2024)), generative frameworks (e.g., GPT-NER (Wang et al., 2025a)), and LLM-based post-processing (Dai et al., 2025; Khatri et al., 2023) offer promising few-shot adaptability, these methods alone cannot fully overcome the deeper structural and resource-related challenges inherent in real-world code-mixing.

3.2 Emerging Contemporary Tasks

LLMs have advanced performance in complex CSW tasks, yet continue to expose limitations in cross-lingual reasoning and cultural adaptation. In **Natural Language Inference**, early conversational datasets revealed persistent annotation disagreements and cultural ambiguities (Khanuja et al., 2020a; Huang and Yang, 2023). Synthetic CoSDA-ML data enabled zero-shot transfer (Qin et al., 2020), while in-context mixing (ICM) prompting improved contextual reasoning (Shankar et al., 2024; Prasad et al., 2021; Kumar et al., 2022), though pragmatic variability continues to cause marked drops relative to monolingual performance. Similarly, **Question Answering** benefited from LLM-based architectures such as COMMIT (Lee et al., 2024), multimodal knowledge-distillation approaches (Raj Khan et al., 2021), non-English prompting for grammaticality improvements (Behzad et al., 2024), curriculum-based CSW pretraining (Yoo et al., 2025), domain-specific embedding migration (MIGRATE) for low-resource reasoning (Hong et al., 2025b), and large-scale African benchmarks such as MEGAVERSE (Ahuja et al., 2024), building on earlier multilingual reading comprehension systems (Gupta et al., 2018). Parallel advances emerged in **Intent Classification** and slot filling, where contrastive pretraining across languages (Lin et al., 2024), prompt-based methods such as PRO-CS (Bansal et al., 2022), multilingual semantic parsing (Duong et al., 2017; Whitehouse et al., 2022), and zero-shot transfer with XLM-R (Arora et al., 2020; Krishnan et al., 2021; Wang et al., 2022) improved cross-lingual generalization.

While LLMs have advanced emerging CSW tasks, persistent limitations remain in contextual reasoning and discourse grounding. Studies on code-mixed QA, NLI, intent detection, and dialogue show that semantic evidence is often fragmented across languages, leading models to rely on shallow lexical cues rather than compositional reasoning (Gupta et al., 2018; Chakravarthy et al., 2020; Krishnan et al., 2021). Inference tasks further exhibit label instability due to culturally contingent interpretations (Khanuja et al., 2020a), while generative models struggle to maintain discourse coherence and stable switching patterns (Mehnaz et al., 2021). Text Generation suffers from data scarcity, especially for typologically distant and non-Latin-script languages (Sravani and Mamidi, 2023). These challenges, compounded by sociolinguistic variation in pragmatic norms (Park et al., 2024), motivate linguistically grounded and lightweight modeling approaches for realistic CSW deployment (Raj Khan et al., 2021).
See Appendix §C.1 and §C.2 for a detailed discussion of remaining tasks, with associated datasets and approaches in Appendix §H (Table 1).

3.3 Underexplored Frontiers Tasks

Although core CSW tasks have advanced, conversational, speech, and multimodal CSW remain underexplored, posing both opportunities and challenges for adapting LLMs to naturalistic multilingual mixing. **Reasoning tasks**, including mathematical problem-solving and cross-language entailment, struggle with logical complexity and semantic drift in CSW contexts (Raihan et al., 2023a; Mohamed et al., 2025), abstract level phenomena such as metaphor comprehension, analogical reasoning, and verb-level code-mixing preferences expose cultural biases and shallow understanding (Kodali et al., 2025a; Mehnaz et al., 2021; Choudhary et al., 2026). Beyond reasoning, **code generation** from mixed prompts achieves only moderate functional correctness (Yang and Chai, 2025; Khatri et al., 2023), despite progress in controllable CSW generation using encoder–decoder models (Mondal et al., 2022). **Conversational systems and dialogue** show emerging gains, with RAG-based architectures improving CSW customer support (Kruk et al., 2025), multilingual dialogue benchmarks enabling few-shot agents for low-resource pairs such as Choctaw–English (Brixey and Traum, 2025), and personality-aware response generation supporting coherent Hinglish multi-party dialogue (Kumar and Chakraborty, 2024). In parallel, **Safety-oriented studies** emphasize region-specific prompting for Kazakh-Russian evaluation (Goloburda et al., 2025a), while while recent work shows that code-mixing itself can be exploited as a trigger for **backdoor attacks**, raising concerns about robustness and security in CSW-aware NLP systems. **Document processing** has also been explored through multilingual OCR and contrastive representation learning for Vietnamese-English text (Dereza et al., 2024; Do et al., 2024).

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

4 Datasets and Resources

4.1 Datasets

The development of CSW datasets has evolved from manual annotation to LLM-driven scalable creation, highlighting trade-offs between expanded multilingual coverage and the authenticity of natural code-switching. However, as pre-training datasets continue to scale, manual curation becomes a challenge. For **Multilingual coverage**, large-scale corpora pre-trained on mixed-language text enhance NLU transfer through synthetic augmentation (Zhang et al., 2024a), while manually annotated datasets like the Multilingual Identification of English CSW benchmark switch across unseen languages (Sterner, 2024), SwitchLingua spans 420k samples and over 80 hours of audio across 12 languages and 63 ethnic groups with LLM-assisted bias reduction (Xie et al., 2025), and MEGAVERSE provides LLM-driven benchmarks covering 22 datasets in 83 languages for multimodal evaluation (Ahuja et al., 2024). MultiCoNER uses LLM synthetic augmentation across 3 domains and 12 languages with 33 entity classes for code-mixed NER (Malmasi et al., 2022b), NusaX offers human annotated parallel sentiment corpus for 10 Indonesian languages (Winata et al., 2023b), and GLOSS synthesizes texts for absent language pairs without manual curation (Hsu et al., 2023). For **Low-resource languages**, targeted datasets address critical underrepresentation through BnSentMix for Bengali–English (Alam et al., 2025), DravidianCodeMix spanning Tamil-, Kannada-, and Malayalam–English (Chakravarthi et al., 2022), Marathi–English corpora (Joshi et al., 2023), SentMix for trilingual NLI (Raihan et al., 2023a), GPT-3.5 synthetic Afrikaans– and Yoruba–English data (Terblanche et al., 2024), and X-RiSAWOZ with over 18k utterances (Moradshahi et al., 2023), collectively diversifying CSW NLP and reducing dependency on high-resource pairs.In **Synthetic data generation**, diverse approaches have addressed annotation scarcity: Bengali–English dependency parsing with large synthetic treebanks (270K+ sentences) (Winata et al., 2019), PhraseOut for Hinglish NMT (Jasim et al., 2020), semi-supervised generation using pre-trained encoders (Gupta et al., 2020), CoSDA-ML for zero-shot NLI across 19 languages (Qin et al., 2020), ternary sequence labeling with mBERT for Hinglish MT (Gupta et al., 2021a), VACS for perplexity reduction (Samanta et al., 2019), COMMIT for low-resource QA (Lee et al., 2024), LLM-generated puns and sentiment data (including 49K+ synthetic samples) for Spanglish and Malayalam–English (Zeng, 2024; Sarrof, 2025), In-Context Mixing for intent classification (Shankar et al., 2024), SynCS for zero-shot gains via parallel alignment (Wang et al., 2025b), and naturalistic parallel CSW datasets for PLM evaluation (Leon et al., 2024). Despite these advances, the move toward LLM-driven, large-scale datasets raises concerns about capturing sociolinguistic nuance and authentic representation, with human evaluations showing only 60–65% acceptability, and highlighting that high-quality resources for underrepresented languages still depend heavily on expert curation and community involvement (Kodali et al., 2025a).

Refer to Tables 4 and 5 in Appendix §H for a summary of CSW text and speech datasets.

4.2 Frameworks and Toolkits

To address the growing complexity of CSW research, frameworks and toolkits have emerged to standardize methodologies and streamline data creation across annotation and generation. **Annotation frameworks**, include CoSSAT, which supports fine-grained word-level and syllable-level speech annotation (Shah et al., 2019); COMMENTATOR, which integrates LLMs for robust text annotation and prediction (Sheth et al., 2024); CHAI, which leverages RLAIF to iteratively refine code-mixed translation annotations (Zhang et al., 2025c); and multimodal tools such as ToxVidLM, extending annotation to video by jointly modeling visual and textual CSW signals (Maity et al., 2024). **Synthetic data generation toolkits** include GCM, which produces linguistically grounded code-mixed text using established switching theories (Rizvi et al., 2021) (as utilized in (Huzaifah et al., 2024)); and CodemixedNLP, an open-source toolkit offering models, datasets, and synthetic augmentation for seven Hinglish tasks (Jayanthi et al., 2021). Together, these tools enable scalable corpus creation and reproducible CSW research for downstream tasks such as machine translation and sentiment analysis (Sravani and Mamidi, 2023; Zeng, 2024).

Takeaway *LLM-augmented datasets such as SwitchLingua, BnSentMix, and COMMIT expand CSW resources for low-resource languages and improve model performance. However, synthetic data may lack naturalness and cultural nuance, introducing biases. Semi-automated, human-in-the-loop annotation toolkits can help create more au-*

thentic and equitable CSW benchmarks.

5 Model Training & Adaptation

5.1 Mainstream Pre-training Approaches

Pre-training encodes mixed-language structure at scale, yielding transferable representations for diverse CSW tasks. **Specialized code-mixed models** trained on real code-mixed corpora consistently outperform multilingual baselines by directly capturing CSW dynamics. HingBERT and related models pre-trained on large-scale real-world data outperform mBERT and XLM-R on downstream NLP tasks (Nayak and Joshi, 2022). Probing studies further show that fine-tuning mBERT on curated naturalistic CSW data yields stronger attention patterns than synthetic mixing across Spanish-English and Hinglish pairs (Santy et al., 2021). Linguistically constrained synthetic embeddings improve over bilingual baselines for sentiment analysis(SA) and POS tagging (Pratapa et al., 2018b), while switch-aware architectures such as CONFLATOR emphasize language junctions to achieve state-of-the-art results on Hinglish SA and translation (Mohammed et al., 2023). For **Task-adaptive pre-training**, targeted strategies explicitly encode CSW structure. Boundary-aware masked language modeling that integrates synthetic CSW data improves downstream QA and SA performance on CSW benchmarks (Das et al., 2023). Model-merging approaches combining continued pre-training with checkpoint fusion outperform standard fine-tuning (Kodali et al., 2025b). Alignment-based methods leveraging parallel text enhance SA analysis and QA (Fazili and Jyothi, 2022), while joint LID–POS multi-task models better capture social media CSW patterns (Dowlagar and Mamidi, 2021c). Multilingual augmentation through synthetic CSW generation improves zero-shot intent detection and slot filling (Krishnan et al., 2021), and large-scale CSW pre-training with diverse synthetic mixtures yields stronger benchmarks and improved language alignment (Wang et al., 2025b).

5.2 Mainstream Fine-tuning Approaches

Fine-tuning adapts models to task-specific CSW distributions, improving in-domain performance but limiting generalization to unseen language pairs. **Task-specific fine-tuning** yields competitive in-domain results but depends heavily on curated CSW data: transformer-based fine-tuning achieves better word-level LID on low-resource Kannada-English pair (Lambebo Tonja et al., 2022), fine-tuned XLM-RoBERTa introduces the S-index for measuring switching intensity and demonstrating effective generalization (Absar, 2025), fine-tuned mBERT provides baselines for sentiment analysis on noisy social media data (Palomino and Ochoa-Luna, 2020), fine-tuned multilingual models like mBART and mT5, often combined with back-translation and ensembling, deliver fluency and accuracy for translation (Chatterjee et al., 2023; Khan et al., 2022), LLM fine-tuning with syntactic post-processing enhances Cantonese-to-Mandarin translation quality across domains (Dai et al., 2025), and efficient monolingual ASR fine-tuning substantially lowers WER on Yoruba-English code-switched speech compared to larger zero-shot multilingual models, though it degrades performance on the non-target (English) language (Babatunde et al., 2025). **Multi-task fine-tuning** leverages synergies for added robustness but can introduce negative transfer or require careful task balancing: syntax-aware joint training of language modeling and parsing lowers perplexity on Mandarin-English data (Winata et al., 2018), intermediate-task fine-tuning on bilingual auxiliaries yields consistent gains in NLI, QA, and sentiment across Hinglish and Spanish-English (Prasad et al., 2021), shared representations enhance offensive speech detection on Hinglish tweets and joint NER modeling in low-resource Arabic dialects (Amazouz et al., 2017), multi-directional fine-tuning and adapter-based methods improve translation and modular transfer (Kartik et al., 2024; Rathnayake et al., 2024), and contrastive multi-task pretraining boosts zero-shot information retrieval and transfer (Do et al., 2024).

A detailed discussion of remaining pre-training and fine-tuning approaches is provided in Appendix §D and §E.

5.3 Post-training Approaches

While post-training approaches enable rapid CSW adaptation with minimal or no labeled data, their effectiveness varies widely across language pairs and task types. **Zero-shot** CSW methods rely on prompting, heuristic switching, or synthetic augmentation, including prompt-based CSW generation with GPT-3.5 (Yong et al., 2023), entity-driven switching for slot filling and dialogue (Whitehouse et al., 2022; Liu et al., 2022), and data-centric augmentation for MT and classification (Gupta, 2022; Lai et al., 2021; Krishnan et al., 2021; Qin et al., 2020). However, even strong LLMs such as GPT-4 exhibit significant performance drops in zero-shot CSW, with outcomes highly sensitive to pretraining language composition (Zhang et al., 2023; Tatariya et al., 2023). **One- and few-shot** methods leverage limited to few examples through adapted prompting, including similarity-based prompting with ChatGPT (Tahery and Farzi, 2025), RAG-based in-context learning for hate speech detection (Srivastava, 2025), multi-task LLM fine-tuning for harmful content in memes (Kumar et al., 2025), generative transformers for emotion detection in Bangla-English-Hindi (Goswami et al., 2023), and translation with LLM classification for affective tasks (Yadav et al., 2025). **Instance-based prompting** further enhances performance, with PRO-CS using mBERT with Hinglish prompts improving NER and POS tagging (Bansal et al., 2022), GLOSS synthesizing CSW text for unseen pairs through self-training (Hsu et al., 2023), Dwesh-Vaani's RAG retrieving Hinglish examples boosting hate speech detection (Srivastava, 2025), and In-Context Mixing improving intent classification on MultiATIS++ (Shankar et al., 2024), instruction tuning for low-resource CSW scenarios (Lee et al., 2024), and synthetic data augmentation for sentiment analysis (Zeng, 2024).

Takeaway While prompting and retrieval enable rapid CSW adaptation, they struggle with informal mixing and low-resource pairs, highlighting that post-training flexibility cannot substitute for pre-training on diverse, naturalistic code-switched data (Bansal et al., 2022; Shankar et al., 2024).

6 Evaluation & Benchmarking

CSW benchmarks have evolved from narrow task evaluations to broader frameworks measuring switching patterns, cross-language performance, and contextual understanding. We review CSW benchmarks across text, speech, and multimodal tasks, with comprehensive details in Appendix (§F.1) and Table 3 in Appendix §H. Evaluating CSW systems requires diverse metrics, encompassing standard performance measures, code-switching–specific metrics, and human-centric evaluation. Full descriptions of evaluation methods, and metrics are given in Appendix (§F.2).

7 Multi- & Cross-Modal Applications

7.1 Speech Processing

Advances in recognition and multimodal integration have improved CSW speech processing, yet limited data availability continues to constrain performance across languages. **Speech translation** has advanced through end-to-end modeling for English–Spanish (Weller et al., 2022), streaming Mandarin–English via self-training (Alastruey et al., 2023), and Whisper-based fine-tuning approaches such as CoVoSwitch (Kang, 2024) and CoSTA (P S V N et al., 2025). **End-to-end ASR** research increasingly emphasizes adaptation over scale, using linguistic augmentation (Chi and Bell, 2022), monolingual fine-tuning that outperforms multilingual baselines (Babatunde et al., 2025), retrieval-augmented refinement (R et al., 2025), and architectural innovations including attention-guided Whisper adaptation (Aditya et al., 2024), mixture-of-experts models (Zhang et al., 2025a), and hybrid CTC/attention systems with language biasing (Liu et al., 2024). Complementary signals from text-derived LID (Wang and Li, 2023) and semi-supervised learning (Biswas et al., 2020) further mitigate data scarcity. Beyond audio-only pipelines, **audio-visual recognition** leverages visual cues to improve CSW ASR across African and Indian language pairs (Babatunde et al., 2025; Hemant and Narvekar, 2025), while data-centric strategies such as phrase-level mixing (Hussein et al., 2024) and zero-resource benchmarks (Huang et al., 2024) support robust evaluation.

Despite these advances, persistent challenges include high error rates at language switch points (Chi and Bell, 2022), limited generalization from synthetic data (Kugathasan and Sumathipala, 2021), and fine-tuning trade-offs in monolingual performance and computational cost (Babatunde et al., 2025).

7.2 Vision-Language Processing

Applied CSW research in real-world deployments remains limited, particularly for multimodal vision-language tasks. **Visual question answering** has advanced through knowledge distillation for Hinglish queries (Raj Khan et al., 2021). **Multimodal systems** tackle challenges in harmful meme detection through visual-text fusion (Kumar et al., 2025; Maity et al., 2024), while CLIP variants enable image-text retrieval in CSW settings (Kumari et al., 2024). Collectively, these efforts highlight growing real-world CSW applications while underscoring the need for domain-, language-, and region-aware adaptation.

7.3 Cross-Modal Integration

Beyond text-only modeling, cross-modal integration enables CSW systems to leverage phonetic, acoustic, and visual cues for robust multilingual understanding. **Phonetic modeling** supports discriminative language modeling (Winata et al., 2019), transliteration and back-transliteration (Tasawong et al., 2023; Fernando and Ranathunga, 2021), and tasks such as abusive content detection and translation alignment (Gautam et al., 2021a; Chou et al., 2023), with recent gains from transformer-based phonetic guidance and Wav2Vec2–GPT-2 fusion (Yang and Tu, 2022; Perera and Sumanathilaka, 2025). **Multimodal fusion** further improves code-mixed ASR and video-based toxicity detection by integrating audio, visual, and textual signals (Maity et al., 2024; Perera and Sumanathilaka, 2025; Zhang et al., 2025a).

Takeaway *Closing CSW performance gaps will require scalable, phonetic-aware multimodal pretraining, as approaches like Wav2Vec2 fusion already achieve 8–10% ASR error reductions in high-resource switching scenarios.*

8 Open Problems and Future Directions

Data scarcity and quality issues A key challenge in building CSW-friendly NLP systems is the lack of appropriate training data. The field remains heavily English-centric, with over 72% of speech and 92% of social media datasets involving English (Doğruöz et al., 2023), leaving non-English pairs underrepresented. Low-resource languages face *higher computational costs* due to inefficient tokenization (Nag et al., 2024), while regional biases reduce generalization, models trained on one region often fail on the same language pair from another (Doğruöz et al., 2023).

Model Architecture and performance gaps LLMs frequently exhibit *language confusion*, generating responses in unintended languages, amplified by standard fine-tuning (Yoo et al., 2025; Marchisio et al., 2024). Trained predominantly on *monolingual text*, models remain ill-equipped for *naturalistic CSW*. While LLMs excel at *synthetic code-mixed data generation* (Pratapa et al., 2018a; Winata et al., 2019), they show fragility in *zero-shot transfer* with sharp accuracy drops (Zhang et al., 2023; Tatariya et al., 2023; Tahery and Farzi, 2025). (Refer to Table 7 in Appendix §H for representative failures.)

Benchmarking and evaluation limitations LLM-based evaluators often overestimate performance relative to human judgments, especially for low-resource and non–Latin-script languages (Hada et al., 2024). Widely used metrics like BLEU and WER fail to capture the linguistic diversity which leads to poorly estimating the quality of code-mixed data (Srivastava and Singh, 2021). Similarly, perplexity correlates poorly with both ASR performance and human judgments (Cheong et al., 2021; Arora et al., 2023; Garg et al., 2021). Standard semantic similarity metrics further struggle to model cross-lingual equivalence in mixed contexts (Maimaiti et al., 2025). (Refer to Table 8 in Appendix §H failure modes.)

Real-World Impact and Applications Advances in CSW research unlock *transformative applications* with significant societal impact: multilingual conversational assistants for accessible public services, cross-lingual educational platforms adaptive to learners' natural language practices, healthcare interfaces for multilingual populations, and digital preservation tools for endangered dialects. Such applications carry substantial social and economic value by reducing language barriers, democratizing access to information, and empowering multilingual communities within the digital economy. BanglAssist for customer service (Kruk et al., 2025) and code-switched dialogue agents for language learning (Brixey and Traum, 2025) illustrates this potential.

Language Bias and Trilingual Neglect Current CSW research exhibits *pronounced bias toward high-resource language pairs*, with the majority of studies focusing on English-Spanish, English-Hindi, and English-Mandarin combinations (Sitaram et al., 2019; Winata et al., 2023a). Performance metrics demonstrate significant disparities: English-Spanish and English-Hindi systems achieve mid-90s F1 scores, while less common pairs like Arabic-Egyptian Arabic or low-resource African language combinations show substantially lower accuracy (Nguyen et al., 2021). African and Southeast Asian code-switched language pairs remain critically underexplored, with very few publicly available datasets (Terblanche et al., 2024). Despite evidence that 7% of India's population is trilingual, with over 250 million speakers engaging in multilingual discourse. Only isolated examples exist, such as SentMix (Bangla-English-Hindi trilingual dataset for NLI) and English-Hindi-Bengali Language Identification (Raihan et al., 2023a), leaving a critical gap in understanding how models process switches across three or more languages. This bias perpetuates a cycle where *resources and research investments concentrate on already well-studied pairs* (Terblanche et al., 2024; Doğruöz et al., 2021), further marginalizing underrepresented multilingual communities and limiting the development of truly inclusive CSW technologies.

Additional challenges are discussed in Appendix (§G).

Future Directions

Toward Inclusive CSW Datasets Progress in CSW NLP relies on expansive, inclusive datasets, yet *large-scale conversational resources capturing naturalistic CSW interactions remain critically lacking*. Multimodal efforts like MEGA-VERSE (Ahuja et al., 2024) show promise but fall short in linguistic and domain diversity. Switch-Lingua (Xie et al., 2025), while large and multilingual, relies on structured and synthesized text rather than fully natural conversational speech. CS-FLEURS (Yan et al., 2025) uses mostly synthetic or TTS-generated audio, limiting its ability to capture spontaneous CSW patterns. Multi-domain multilingual dialogue corpora (Moradshahi et al., 2023), though broader in scope, highlight the need for future efforts to expand coverage, diversity, and naturalistic interactions.

Next-Generation Architectures must *jointly model text, speech, and vision* to enable switch-point detection, contextual understanding, and natural multilingual interactions, while ASR and TTS systems should leverage self-supervised encoders, cross-lingual, and emotion-aware conditioning. Promising directions include Speech-Conditioned LLMs combined with MoE for ASR (Zhang et al., 2025a) and curriculum learning strategies for multilingual transfer (Yoo et al., 2025). These approaches address phonemic confusion, data scarcity and the need for adaptive language mixing (Hamed et al., 2025).

Holistic Evaluation Paradigms As CSW models become more multimodal and adaptive, evaluation must move *beyond isolated task-level metrics toward human-aligned assessment* of multilingual competence. Future frameworks should jointly capture switch-point accuracy, semantic consistency, fluency, and *sociolinguistic appropriateness*. While benchmarks such as CS-Sum (Suresh et al., 2025) and CodeMixBench (Yang and Chai, 2025) mark important progress (Hamed et al., 2025), evaluation must also account for regional and dialectal variation.

Ethics and Safety in Multilingual Contexts Beyond performance and evaluation, future CSW systems must address *critical ethical vulnerabilities* stemming from multilingual safety alignment gaps (Song et al., 2025), which disproportionately affect low-resource and marginalized language communities (Hamed et al., 2025). Safety evaluations reveal persistent failures under unseen language mixture patterns, as demonstrated by the *Qorgau* framework in Kazakh–Russian settings (Goloburda et al., 2025a). Addressing these challenges requires CSW-aware ethical AI that emphasizes inclusivity, transparency, and accountability through bias-aware training, fairness-sensitive evaluation, and *participatory data curation with speaker communities*.

9 Conclusion

CSW research has undergone a major transformation with the rise of LLMs, evolving from task-specific statistical methods to unified multilingual and instruction-based frameworks. However, this survey shows that gains remain largely confined to high-resource language pairs, while LLMs struggle with spontaneous mixing, reasoning, and sociolinguistic variation in low-resource settings. These challenges are further amplified by limited dataset coverage and the lack of robust, CSW-aware evaluation frameworks. Meaningful progress in CSW NLP therefore requires moving beyond generic multilinguality toward targeted data curation, linguistically informed architectures, and evaluation protocols grounded in real-world language-mixing.

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** While we include recent advances in speech and multimodal processing, the volume of research in these areas significantly lags behind text-based NLU, resulting in our taxonomy covering more text based NLU.

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

5. **Quantitative Reporting** All quantitative claims are representative values reported in the cited primary works rather than aggregated or meta-analyzed statistics.

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.

Acknowledgments

This work is supported by the Anusandhan National Research Foundation (ANRF), India, through the project titled "Curating and Constructing Benchmarks and Development of ML Models for Low-Level NLP Tasks in Hindi-English Code-Mixing"[2]. The authors express their gratitude to Prathamesh Shanbhag, Naren Kumar S, Pranjal Yadav, Indrayudh Mandal, Sailesh Panda and Shruti Singh for their assistance in reveiwing the manuscript and providing feedback providing valuable feedback, and contributing to experimental evaluation and error analysis. Himanshu Beniwal is supported by the Prime Minister Research Fellowship (PMRF ID-1702154), India.

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.

Bobbi Aditya, Mahdin Rohmatillah, Liang-Hsuan Tai, and Jen-Tzung Chien. 2024. Attention-guided adaptation for code-switching speech recognition. In *ICASSP 2024-2024 IEEE International Conference On Acoustics, Speech And Signal Processing (ICASSP)*, pages 10256–10260. IEEE.

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. 2021a. 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.

Vibhav Agarwal, Pooja Rao, and Dinesh Babu Jayagopi. 2021b. Towards code-mixed Hinglish dialogue generation. In *Proceedings of the 3rd Workshop on Natural Language Processing for Conversational AI*, pages 271–280, Online. Association for Computational Linguistics.

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*

[2]More information about the Project is present here: https://lingo.iitgn.ac.in/codemixing

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.

Maryam Khalifa Al Ali and Hanan Aldarmaki. 2024. Mixat: A data set of bilingual emirati-English speech. In *Proceedings of the 3rd Annual Meeting of the Special Interest Group on Under-resourced Languages @ LREC-COLING 2024*, pages 222–226, Torino, Italia. ELRA and ICCL.

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.

Haneen Alharbi, Sameer Khurana, James Glass, and Ahmed Ali. 2024. Leveraging LLM for augmenting textual data in code-switching ASR. In *Proceedings of the 1st Workshop on Synthetic Data for Generative AI (SynData4GenAI)*. ISCA. Introduces the Saudilang Code-switch Corpus (SCC).

Ahmed Ali, Shammur Absar Chowdhury, Amir Hussein, and Yasser Hifny. 2021. Arabic code-switching speech recognition using monolingual data. In *Interspeech 2021*, pages 3475–3479. ISCA. Releases the ESCWA corpus.

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. 2021. 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.

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.

Utsab Barman, Amitava Das, Joachim Wagner, and Jennifer Foster. 2014. Code mixing: A challenge for language identification in the language of social media. In *Proceedings of the First Workshop on Computational Approaches to Code Switching*, pages 13–23, Doha, Qatar. Association for Computational Linguistics.

Ruthanna Barnett, Eva Codó, Eva Eppler, Montse Forcadell, Penelope Gardner-Chloros, Roeland van Hout, Melissa Moyer, Maria Carme Torras, Maria Teresa Turell, Mark Sebba, Marianne Starren, and Sietse Wensing. 2000. The lides coding manual: A document for preparing and analyzing language interaction data version 1.1—july, 1999. *International Journal of Bilingualism*, 4(2):131–132.

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.

Alexey Birshert and Ekaterina Artemova. 2021. Call larisa ivanovna: Code-switching fools multilingual nlu models. In *International Conference on Analysis of Images, Social Networks and Texts*, pages 3–16. Springer.

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.

Jacqueline Brixey and David Traum. 2025. Does a code-switching dialogue system help users learn conversational fluency in Choctaw? In *Proceedings of the Fifth Workshop on NLP for Indigenous Languages of the Americas (AmericasNLP)*, pages 8–17, Albuquerque, New Mexico. 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.

Khyathi Chandu, Ekaterina Loginova, Vishal Gupta, Josef van Genabith, Günter Neumann, Manoj Chinnakotla, Eric Nyberg, and Alan W. Black. 2018. Code-mixed question answering challenge: Crowd-sourcing 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.

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.

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

Sik Feng Cheong, Hai Leong Chieu, and Jing Lim. 2021. Intrinsic evaluation of language models for code-switching. In *Proceedings of the Seventh Workshop on Noisy User-generated Text (W-NUT 2021)*, pages 81–86, Online. Association for Computational Linguistics.

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.

Won Ik Cho, Seok Min Kim, and Nam Soo Kim. 2020. Towards an efficient code-mixed grapheme-to-phoneme conversion in an agglutinative language: A case study on to-Korean transliteration. In *Proceedings of the 4th Workshop on Computational Approaches to Code Switching*, pages 65–70, Marseille, France. European Language Resources Association.

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.

Mukund Choudhary, Madhur Jindal, Gaurja Aeron, and Monojit Choudhury. 2026. Do LLMs model human linguistic variation? a case study in Hindi-English verb code-mixing. In *Findings of the Association for Computational Linguistics: EACL 2026*, pages 5491–5509, Rabat, Morocco. Association for Computational Linguistics.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

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).

Anuj Diwan, Rakesh Vaideeswaran, Sanket Shah, Ankita Singh, Srinivasa Raghavan, and 1 others. 2021. MUCS 2021: Multilingual and code-switching ASR challenges for low resource indian languages. In *Interspeech 2021*, pages 2446–2450. ISCA.

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.

A. Seza Doğruöz, Sunayana Sitaram, Barbara E. Bullock, and Almeida Jacqueline Toribio. 2021. A survey of code-switching: Linguistic and social perspectives for language technologies. 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 1654–1666, Online. Association for Computational Linguistics.

A. Seza Doğruöz, Sunayana Sitaram, and Zheng Xin Yong. 2023. Representativeness as a forgotten lesson for multilingual and code-switched data collection and preparation. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5751–5767, Singapore. 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 pre-trained 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.

Yukun Feng, Feng Li, and Philipp Koehn. 2022. Toward the limitation of code-switching in cross-lingual transfer. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5966–5971, Abu Dhabi, United Arab Emirates. Association for 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.

Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378.

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).

Ayush Garg, Sammed Kagi, Vivek Srivastava, and Mayank Singh. 2021. MIPE: A metric independent pipeline for effective code-mixed NLG evaluation. In *Proceedings of the 2nd Workshop on Evaluation and Comparison of NLP Systems*, pages 123–132, Punta Cana, Dominican Republic. Association for Computational Linguistics.

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.

Marwa Gaser, Manuel Mager, Injy Hamed, Nizar Habash, Slim Abdennadher, and Ngoc Thang Vu. 2023. Exploring segmentation approaches for neural machine translation of code-switched Egyptian Arabic-English text. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 3523–3538, Dubrovnik, Croatia. 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. 2025a. 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.

Maksym Goloburda, Almas Nurkas, Aibek Galym, Zhansaya Tazabek, Symbat Mussakhojayeva, Tatiana Shavrina, and Ekaterina Artemova. 2025b. Qorgau: Evaluating LLM safety in kazakh-russian bilingual contexts. In *Findings of the Association for Computational Linguistics: ACL 2025*.

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.

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. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8342–8360, Online. Association for Computational Linguistics.

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, Ngoc Thang Vu, and Slim Abdennadher. 2020. ArzEn: A speech corpus for code-switched egyptian arabic-english. In *Proceedings of the 12th Language Resources and Evaluation Conference (LREC)*, pages 4276–4284, Marseille, France. European Language Resources Association.

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.

Mengze Hong, Chen Jason Zhang, Chaotao Chen, Rongzhong Lian, and Di Jiang. 2025a. Dialogue language model with large-scale persona data engineering. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 3: Industry Track)*, pages 961–970.

Seongtae Hong, Seungyoon Lee, Hyeonseok Moon, and Heuiseok Lim. 2025b. 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.

Kuan-Po Huang, Chih-Kai Yang, Yu-Kuan Fu, Ewan Dunbar, and Hung-yi Lee. 2024. Zero resource code-switched speech benchmark using speech utterance pairs for multiple spoken languages. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 10006–10010. IEEE.

Amir Hussein, Dorsa Zeinali, Ondřej Klejch, Matthew Wiesner, Brian Yan, Shammur Chowdhury, Ahmed Ali, Shinji Watanabe, and Sanjeev Khudanpur. 2024. Speech collage: code-switched audio generation by collaging monolingual corpora. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 12006–12010. IEEE.

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.

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.

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. 2021. 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.

Haotian Jin, Haihui Fan, Jinchao Zhang, Yang Li, Bo Li, and Junhao Zhou. 2025. Dangerous language habits! exploiting code-mixing for backdoor attacks on nlp models. In *Proceedings of the 34th ACM International Conference on Information and Knowledge Management*, CIKM '25, page 1220–1230, New York, NY, USA. Association for Computing Machinery.

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*.

Artur Kiulian, Anton Polishko, Mykola Khandoga, Yevhen Kostiuk, Guillermo Gabrielli, Łukasz Gagała, Fadi Zaraket, Qusai Abu Obaida, Hrishikesh Garud, Wendy Wing Yee Mak, Dmytro Chaplynskyi, Selma Amor, and Grigol Peradze. 2025. From English-centric to effective bilingual: LLMs with custom tokenizers for underrepresented languages. In *Proceedings of the Fourth Ukrainian Natural Language Processing Workshop (UNLP 2025)*, pages 1–13, Vienna, Austria (online). Association for Computational Linguistics.

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. 2025a. 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. 2022. 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, Vaishnavi Shivkumar, Swarang Joshi, Monojit Choudhary, Ponnurangam Kumaraguru, and Manish Shrivastava. 2025b. Adapting multilingual models to code-mixed tasks via model merging. *arXiv preprint arXiv:2510.19782*.

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.

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.

Shivani Kumar and Tanmoy Chakraborty. 2024. Harmonizing code-mixed conversations: Personality-assisted code-mixed response generation in dialogues. In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 639–653, St. Julian's, Malta. Association for Computational Linguistics.

Vishwajeet Kumar, Rudra Murthy, and Tejas Dhamecha. 2022. 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.

Houssam Eddine-Othman Lachemat, Abbas Akli, Nourredine Oukas, Yassine El Kheir, Samia Haboussi, and Shammur Absar Chowdhury. 2025. Cafe: Spontaneous code-switching speech dataset in algerian dialect, french and english. *Data in Brief*, 63:112150.

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.

Chengfei Li, Shuhao Deng, Yaoping Wang, Guangjing Wang, Yaguang Gong, and 1 others. 2022. TALCS: An open-source mandarin-english code-switching corpus and a speech recognition baseline. In *Interspeech 2022*, pages 1–5. ISCA.

Yuchen Li, Chen Lyu, Jinchuan Li, Yiling Song, Zhiyong Yuan, and Yanmin Qian. 2025. Boosting code-switching ASR with mixture of experts enhanced speech-conditioned LLM. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

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, Ekaterina Artemova, and Barbara Plank. 2023. Boosting zero-shot cross-lingual retrieval by training on artificially code-switched data. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 3096–3108, Toronto, Canada. Association for Computational Linguistics.

Hexin Liu, Leibny Paola Garcia, Xiangyu Zhang, Andy WH Khong, and Sanjeev Khudanpur. 2024. Enhancing code-switching speech recognition with interactive language biases. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 10886–10890. IEEE.

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, Makuhari, Chiba, Japan. ISCA.

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.

Mieradilijiang Maimaiti, Yuanhang Zheng, Ji Zhang, Yue Zhang, Wenpei Luo, and Kaiyu Huang. 2025. Improving cross-lingual representation for semantic retrieval with code-switching. *Knowledge-Based Systems*, page 113919.

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.

Kelly Marchisio, Wei-Yin Ko, Alexandre Berard, Théo Dehaze, and Sebastian Ruder. 2024. Understanding and mitigating language confusion in LLMs. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 6653–6677, Miami, Florida, USA. 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*.

Mohsin Mohammed, Sai Kandukuri, Neeharika Gupta, Parth Patwa, Anubhab Chatterjee, Vinija Jain, Aman Chadha, and Amitava Das. 2023. CONFLATOR: Incorporating switching point based rotatory positional encodings for code-mixed language modeling. In *Proceedings of the 6th Workshop on Computational Approaches to Linguistic Code-Switching*, pages 64–73, Singapore. Association for Computational Linguistics.

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.

Li Nguyen, Christopher Bryant, Sana Kidwai, and Theresa Biberauer. 2021. Automatic language identification in code-switched hindi-english social media text. *Journal of Open Humanities Data*, 7.

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.

Alissa Ostapenko, Shuly Wintner, Melinda Fricke, and Yulia Tsvetkov. 2022. Speaker information can guide models to better inductive biases: A case study on predicting code-switching. *arXiv preprint arXiv:2203.08979*.

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.

Dwija Parikh and Thamar Solorio. 2021. Normalization and back-transliteration for code-switched data. In *Proceedings of the Fifth Workshop on Computational Approaches to Linguistic Code-Switching*, pages 119–124, Online. 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.

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.

Page Elizabeth Piccinini and Marc Garellek. 2014. Prosodic cues to monolingual versus code-switching sentences in english and spanish. In *Proceedings of the 7th Speech Prosody Conference*, pages 885–889.

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.

Preply. 2022. Bilingualism statistics 2022: Facts and figures.

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.

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.

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. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 7973–7992, Suzhou, China. Association for Computational Linguistics.

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.

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.

Sunayana Sitaram, Khyathi Raghavi Chandu, Sai Krishna Rallabandi, and Alan W Black. 2019. A survey of code-switched speech and language processing. *arXiv preprint arXiv:1904.00784*.

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. Challenges and limitations with the metrics measuring the complexity of code-mixed text. In *Proceedings of the Fifth Workshop on Computational Approaches to Linguistic Code-Switching*, pages 6–14, Online. 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. 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.

Rosetta Stone. 2025. How many people are trilingual? (plus, is it worth it?).

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.

Chihiro Taguchi, Yusuke Sakai, and Taro Watanabe. 2021. Transliteration for low-resource code-switching texts: Building an automatic Cyrillic-to-Latin converter for Tatar. In *Proceedings of the Fifth Workshop on Computational Approaches to Linguistic Code-Switching*, pages 133–140, Online. Association for Computational Linguistics.

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.

Bibek Upadhayay and Vahid Behzadan. 2025. Tongue-tied: Breaking LLMs safety through new language learning. In *Proceedings of the 7th Workshop on Computational Approaches to Linguistic Code-Switching*, pages 32–47, Albuquerque, New Mexico, USA. Association for Computational Linguistics.

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.

Qinyi Wang and Haizhou Li. 2023. Text-derived language identity incorporation for end-to-end code-switching speech recognition. In *Proceedings of the 6th Workshop on Computational Approaches to Linguistic Code-Switching*, pages 33–42, Singapore. 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, 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. 2018. Code-switching language modeling using syntax-aware multi-task learning. In *Proceedings of the Third Workshop on Computational Approaches to Linguistic Code-Switching*, pages 62–67, Melbourne, Australia. 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.

Tanushree Ravindra Pratap Yadav. 2026. Competence collapse in code-mixed generation: Spectral evidence and mechanistic recovery via cross-lingual activation steering. In *Proceedings of the Second Workshop on Language Models for Low-Resource Languages (LoResLM 2026)*, pages 29–40, Rabat, Morocco. Association for Computational Linguistics.

Brian Yan, Injy Hamed, Shuichiro Shimizu, Vasista Lodagala, William Chen, Olga Iakovenko, Bashar Talafha, Amir Hussein, Alexander Polok, Kalvin Chang, Dominik Klement, Sara Althubaiti, Puyuan Peng, Matthew Wiesner, Thamar Solorio, Ahmed Ali, Sanjeev Khudanpur, Shinji Watanabe, Chih-Chen Chen, and 8 others. 2025. Cs-fleurs: A massively multilingual and code-switched speech dataset. *arXiv preprint arXiv:2509.14161*.

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. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 2139–2169, Suzhou, China. Association for Computational Linguistics.

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.

Zhihan Zhang, Dong-Ho Lee, Yuwei Fang, Wenhao Yu, Mengzhao Jia, Meng Jiang, and Francesco Barbieri. 2024b. 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, benchmarks 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 around 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 327 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, Intent, Speech/ASR, MT), ***(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 **327** papers forming the foundation of this survey, spanning **15 NLP tasks**, **30+ datasets**, and **80+ languages**.

Handling Multi-Category Papers and Reporting Gaps Papers spanning multiple categories or tasks (common in recent LLM-era work) are assigned multi-label categories and discussed in the most relevant primary section, with cross-references provided where appropriate. Reported performance gaps are presented as representative values directly quoted or reported from individual primary studies on their respective benchmarks. This approach preserves representativeness and reflects the interdisciplinary nature of modern code-switching research (e.g., a single work advancing both synthetic data generation and fine-tuning).

B Taxonomy

In this section, we elaborate on the taxonomy of code-mixed language analytics introduced in Figure 2 that provides an analytical overview of the CSW research landscape across four dimensions: (A) task maturity, showing saturation in traditional tasks (e.g., LID, POS) and persistent gaps in reasoning and multimodal settings; (B) methodological evolution from statistical models to LLM-based approaches; (C) language-pair coverage, revealing a strong 72% English-centric bias; and (D) performance gaps that grow with task complexity, from 4% degradation in LID to over 33% in reasoning-heavy tasks. An interactive taxonomy offers paper-level details.

The rapid evolution of code-switching (CSW) research in the LLM era requires a comprehensive framework capturing methodological diversity and task complexity. Our framework supports scalable LLM approaches while distinguishing various contribution types.

This structure reflects the interconnected nature of modern CSW research, emphasizing the shift from language-pair-specific solutions to unified multilingual architectures and the need for integrated, end-to-end CSW systems.

Code-Switched Language Analytics

B.1 Code-Switching Task Landscape

Foundational Tasks These represent core NLP taskcompetencies essential for understanding code-switched text analysisstructure and linguistic properties. They form the foundationbase layer upon which more complex applications are built.

- **Language Identification**: Identifies language boundaries at the word or token level, forming the basis for downstream analysis in mixed-language text Detects language boundaries at word/token level, including Hope and Offensive text detection.

- **Part-of-Speech Tagging**: Assigns grammatical categories to code-mixed tokens, accounting for syntactic ambiguity and structural variation at switch points.

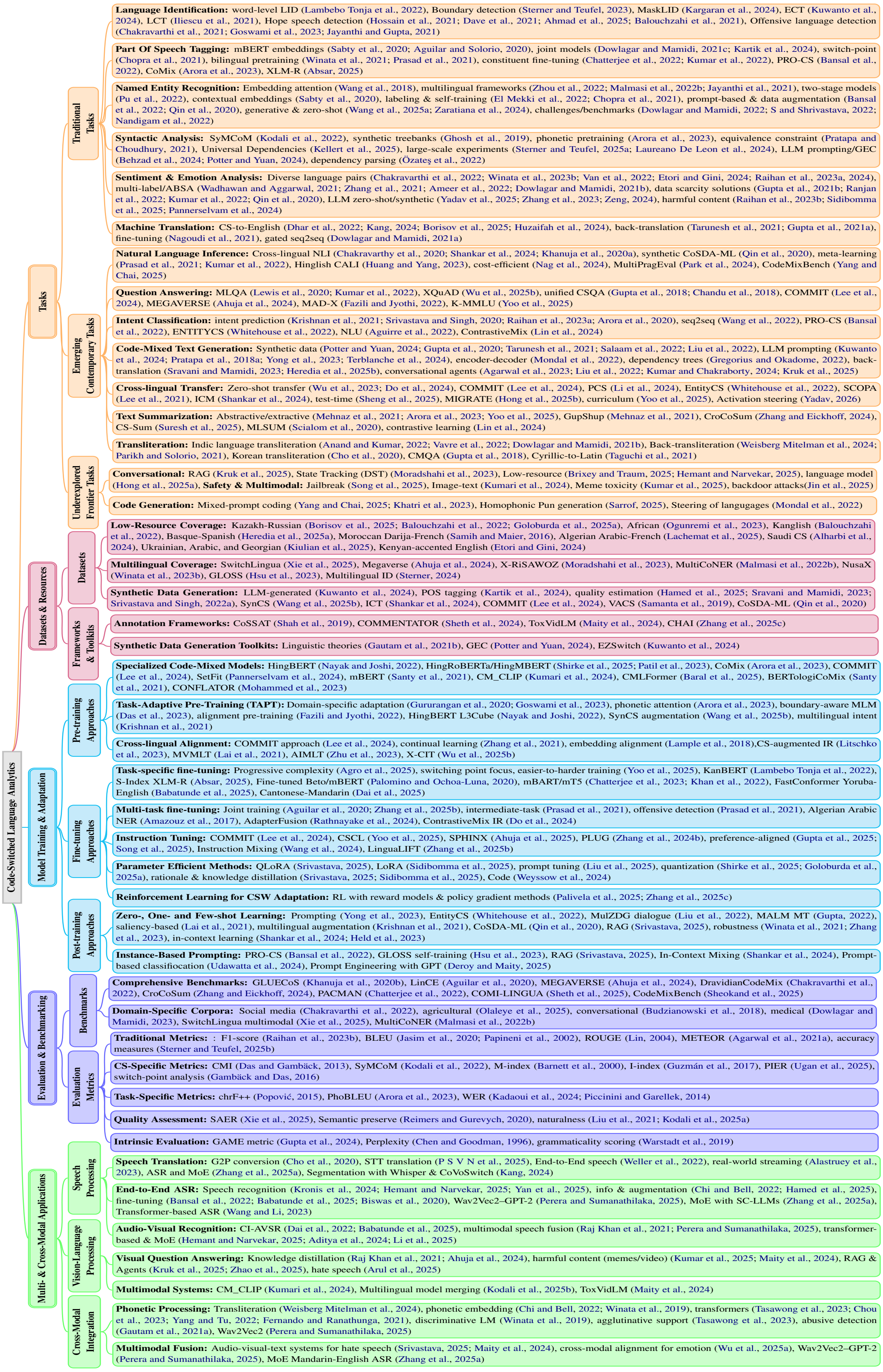

Figure 2: A unifying taxonomy of the code-switching research landscape. ***Takeaway*** *The mind map contextualizes recent LLM-based advances, revealing continuities, shifts, and unresolved challenges across the CSW literature.*

- **Named Entity Recognition**: Detects and classifies named entities across language boundaries, addressing challenges such as transliteration, script variation, and cross-lingual ambiguity.

- **Machine Translation**: Translates code-switched input into a single target language, requiring joint modeling of mixed-language syntax and semantics.

- **Syntactic Analysis**: Parses the grammatical structure of code-switched sentences to analyze well-formedness and linguistic constraints governing switching behavior.

- **Sentiment and Emotion Analysis**: Models affective meaning across languages, including aspect-based sentiment and multi-label emotion classification.

- **Machine Translation**: Translates code-switched text to monolingual output, requiring understanding of mixed-language syntax.

Emerging and Contemporary Tasks Emerging tasks extend beyond surface-level analysis to capture semantic interpretation, pragmatic reasoning, and contextual understanding in code-switched settings.

- **Natural Language Inference**: Evaluates entailment and contradiction between code-switched premise–hypothesis pairs.

- **Question Answering**: Supports information retrieval and reasoning over code-switched queries and documents.

- **Intent Classification**: Infers speaker intent in mixed-language conversational inputs, particularly relevant for dialogue and assistant systems.

- **Code-Mixed Text Generation**: Generates linguistically and sociolinguistically plausible code-switched text, often used for data augmentation and dialogue systems.

- **Cross-lingual Transfer**: Exploits code-switching to improve generalization across languages, including transfer to unseen or low-resource language pairs.

- **Text Summarization**: Produces abstractive or extractive summaries while preserving semantic content and, where relevant, code-mixing patterns.

- **Transliteration**: Converts text across scripts while maintaining phonetic fidelity in mixed-language contexts.

Underexplored and Frontier Tasks These tasks represent comparatively underexplored directions where code-switching intersects with safety, reasoning, creativity, and multimodal understanding.

- **Conversational and Speech**: Includes dialogue generation, customer support agents, and ASR systems operating on naturally occurring code-switched speech.

- **Safety and Multimodal**: Addresses multilingual safety alignment, jailbreaking, and image–text interactions in code-mixed settings.

- **Reasoning and Abstraction**: Examines causal, analogical, and metaphorical reasoning across languages within a single utterance or discourse.

- **Creative and Code Generation**: Covers programming code generation from code-mixed prompts and creative language use such as wordplay and homophonic mixing.

B.2 Datasets and Resources

Datasets Datasets form the backbone of code-switched language research, enabling empirical evaluation across diverse languages, domains, and modalities.

- **Low-Resource Coverage**: Targeted datasets addressing underrepresented language pairs, including African, Dravidian, and Central Asian languages.

- **Multilingual Coverage**: Large-scale corpora spanning multiple language families, scripts, and sociolinguistic contexts.

- **Synthetic Data**: Promptly generated corpora leveraging linguistic constraints or LLMs to augment scarce real-world data.

Frameworks and Toolkits Frameworks and toolkits support standardized annotation, data generation, and experimentation in code-switched settings.

- **Annotation Frameworks**: Tools for annotating code-mixed text, speech, and multimodal data with support for human-in-the-loop workflows.
- **Synthetic Data Generation Toolkits**: Tools designed to generate synthetic code-mixed corpora through rule-based constraints, transliteration models, and language model–based augmentation.

B.3 Model Training and Adaptation

Pre-training Approaches Pre-training strategies aim to encode code-switching phenomena directly into model representations, often outperforming generic multilingual pre-training on CSW tasks.

- **Specialized Code-Mixed Models**: Architectures trained explicitly on code-mixed corpora with switch-point awareness.
- **Task-Adaptive Pre-training**: Domain-specific and task-specific adaptation using masked language modeling and alignment-aware objectives.
- **Cross-lingual Alignment**: Representation alignment and continual learning techniques to improve multilingual generalization.

Fine-tuning Approaches Fine-tuning methods adapt pre-trained models to specific tasks while incorporating code-switching-aware objectives.

- **Task-specific Fine-tuning**: Staged training of a model (or adapters) dedicated to a single task/language pair.
- **Multi-task Fine-tuning**: Joint training of a single model on multiple tasks/language pairs simultaneously with shared parameters to enable knowledge transfer.
- **Instruction Tuning**: Instruction-following adaptation using code-mixed prompts and responses.
- **Parameter-efficient Methods**: Lightweight adaptation techniques such as LoRA, prompt tuning, and quantization-aware training.
- **Reinforcement Learning**: Reward-based optimization for improving fluency and naturalness in code-switched generation.

Post-training and Inference-time Adaptation These approaches enable generalization in low-resource settings without extensive labeled data.

- **Zero-, One-, and Few-shot Learning**: Prompt-based and retrieval-augmented methods for code-switched tasks under minimal supervision.
- **Instance-based Prompting**: In-context learning approaches that leverage curated or automatically selected code-mixed examples to guide model behavior at inference time.

B.4 Evaluation and Benchmarking

Benchmarks Benchmarks provide standardized evaluation protocols for measuring progress across tasks and domains.

- **Comprehensive Benchmarks**: Multi-task suites covering both traditional and emerging code-switched NLP tasks.
- **Domain-specific Corpora**: Evaluation datasets tailored to domains such as social media, healthcare, agriculture, and multimodal content.

Evaluation Metrics Evaluation metrics aim to capture both task performance and code-switching-specific linguistic properties.

- **Traditional Metrics**: Standard NLP measures such as accuracy, F1, BLEU, ROUGE, and METEOR.
- **Code-switching-specific Metrics**: Measures that quantify mixing intensity, syntactic diversity, and switch-point accuracy in mixed-language text.
- **Task-specific Metrics**: Evaluation measures tailored to individual tasks, accounting for script variation, phonetic ambiguity, and speech recognition errors.
- **Quality Assessment**: Human judgments of fluency, semantic preservation, and naturalness across languages.
- **Intrinsic Evaluation**: Gold-reference-independent metrics for assessing grammaticality, fluency, and distributional consistency.

B.5 Multi- and Cross-modal Applications

Speech Processing Code-switching in speech introduces phonetic and acoustic variability that challenges conventional speech models.

- **Speech Translation**: Systems that integrate automatic speech recognition and machine translation for processing mixed-language speech input.
- **End-to-End ASR**: Direct modeling of code-switched speech using data augmentation strategies and expert-based or modular architectures.
- **Audio-Visual Recognition**: Multimodal approaches that combine acoustic signals with visual cues to improve recognition robustness.

Vision–Language Processing Vision-language tasks extend code-switching to multimodal contexts.

- **Visual Question Answering**:mage-based reasoning with mixed-language questions and captions.
- **Multimodal Systems**: Joint visual–text processing for multilingual and code-switched documents.

Cross-modal Integration Cross-modal approaches aim to unify representations across text, speech, and vision.

- **Phonetic Processing**: Script conversion and phonetic embeddings for mixed-script languages.
- **Multimodal Fusion**: Joint audio–visual–text models for affective analysis and safety-related tasks.

C Code-Switching Task Landscape: Capabilities and Gaps

C.1 Traditional 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., 2021; 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).

Sentiment & Emotion Analysis ihas been extensively studied in CSW settings, with shared tasks like SemEval (Sentiment Analysis for Code-Mixed Social Media Text), where fine-tuned multilingual transformers achieved strong performance on Hinglish and Spanglish datasets via strategies such as focal loss for class imbalance (XLM-R) (Ma et al., 2020), straightforward mBERT fine-tuning (Palomino and Ochoa-Luna, 2020), RoBERTa fine-tuning (Sultan et al., 2020), stacked ensembling of BiLSTM and BERT variants (Singh and Singh Parmar, 2020), and multi-task learning with BERT (Wu et al., 2020); similar approaches were applied in NLP-CIC systems (Angel et al., 2020). Research has expanded to diverse language pairs, including Dravidian languages (Tamil-English, Malayalam-English, Kannada-English) (Chakravarthi et al., 2022), Indonesian and Vietnamese-English (Winata et al., 2023b; Van et al., 2022), Kenyan Sheng-English slang (Etori and Gini, 2024), Bengali-English trilingual sentiment (Raihan et al., 2023a), and emotion-specific trilingual analysis (Raihan et al., 2024). Multi-label emotion detection frameworks support fine-grained analysis across CSW texts (Wadhawan and Aggarwal, 2021), while cross-lingual aspect-based sentiment analysis (ABSA) leverages shared representations for improved transfer (Zhang et al., 2021). Data scarcity challenges are mitigated through unsupervised self-training on unlabeled CSW data (Gupta et al., 2021b), progressive curriculum learning with increasing mixing intensity (Ranjan et al., 2022), integration of monolingual resources (Kumar et al., 2022), and synthetic code-switched augmentation via CoSDA-ML, yielding consistent zero-shot gains across multiple tasks (Qin et al., 2020). Large language models enable effective zero-shot sentiment classification through translation-based pipelines (Yadav et al., 2025), multilingual RLAIF for preference alignment (Zhang et al., 2023), and efficient synthetic data leveraging for downstream sentiment tasks (Zeng, 2024). Harmful content detection has advanced with datasets targeting Bangla-English offensive language and Devanagari-script hate speech (Raihan et al., 2023b), where parameter-efficient fine-tuning (PEFT) and SetFit embeddings achieve competitive results on low-resource CSW hate speech (Sidibomma et al., 2025; Pannerselvam et al., 2024).

Syntactic Analysis in CSW has shifted from structural modeling to theory-guided methods, improving parsing and evaluation. SyMCoM introduced a syntactic measure of code-mixing based on POS tags for English-Hindi, enabling dataset comparison and highlighting variations in open/closed class contributions (Kodali et al., 2022). Syntax-aware multi-task LSTMs jointly trained on language modeling and parsing significantly reduced perplexity on Mandarin-English code-switched data (Winata et al., 2018). Synthetic treebanks generated via annotation projection improved dependency parsing performance for Bengali-English (Ghosh et al., 2019). CoMix leveraged phonetic and POS-guided pre-training to advance Hinglish machine translation and NER (Arora et al., 2023). Linguistically constrained generation following the Equivalence Constraint produced more natural code-mixed text compared to heuristic baselines (Pratapa and Choudhury, 2021). LLMs facilitated Universal Dependencies annotation for low-resource pairs like Spanglish and Spanish–Guaraní (Kellert et al., 2025), while large-scale experiments demonstrated strong syntactic alignment in CSW with monolingual parses (Sterner and Teufel, 2025a; Laureano De Leon et al., 2024). Non-English prompting enhanced LLM grammaticality judgments (Behzad et al., 2024), and LLM-based grammatical error correction performed well on learner corpora (Potter and Yuan, 2024). Despite these advances, enforcing universal syntactic constraints across typologically diverse languages remains difficult, often leading to unnatural switches or reduced fluency in generated text (Pratapa and Choudhury, 2021).

Machine Translation in CSW contexts has evolved from statistical to neural paradigms, addressing irregular switching and data scarcity. Pioneering works used code-switching as augmentation to enforce lexical constraints in standard NMT by replacing source phrases with target translations to teach copying (Song et al., 2025), while PhraseOut advanced controlled mixing via phrase-level replacement for multilingual low-resource scenarios (Jasim et al., 2020). Back-to-back translation improved Hinglish MT, while unsupervised approaches with linguistic heuristics enhanced Sinhala–English corpora (Tarunesh et al., 2021; Kugathasan and Sumathipala, 2021). CoSDA-ML scaled dynamic multi-language code-switching augmentation by word substitution from bilingual dictionaries to fine-tune mBERT for zero-shot cross-lingual alignment across diverse tasks (Qin et al., 2020), and CoMeT/back-translation with COMET filtering produced higher-quality synthetic parallel data for Indic/Hinglish pairs by concatenating monolingual sentences and transliterating roman script (Gautam et al., 2021b). Gated seq2seq architectures with explicit language tags (Dowlagar and Mamidi, 2021a), fine-tuned mT5 for Hinglish (Nagoudi et al., 2021), and mBART overcoming orthographic challenges in MSA-Egyptian–English (Nagoudi et al., 2021) further refined neural approaches. Recent LLM integrations, including syntactic post-processing for Cantonese–Mandarin (Dai et al., 2025) and direct GPT prompting for Hinglish fluency (Khatri et al., 2023), have elevated quality, with fine-tuned transformers/T5 achieving strong CodeMix-to-English results extended via knowledge distillation to multimodal tasks (Chatterjee et al., 2023; Jawahar et al., 2021; Raj Khan et al., 2021). Despite these strides, CSW MT remains prone to syntactic misalignment at switch points, inconsistent transliteration, and degraded performance on informal/noisy social media text, underscoring the need for more robust, linguistically grounded hybrid strategies (Winata et al., 2021; Sazzed, 2021).

C.2 Emerging Contemporary Tasks

Code-Mixed Text Generation has progressed from early transfer- and translation-based methods toward LLM-driven and data-centric approaches. Semi-supervised transfer learning and machine translation models improved Hinglish fluency and structural consistency (Gupta et al., 2020; Tarunesh et al., 2021), while COCOA demonstrated effective English–Spanish code-mixed generation through controlled switching mechanisms (Mondal et al., 2022). Syntactically grounded approaches leveraging dependency trees enabled CSW generation without parallel corpora, highlighting the role of linguistic constraints in low-resource settings (Gregorius and Okadome, 2022). Subsequent work has explored synthetic data filtering and prompt-based LLM generation to improve naturalness and diversity for language pairs such as Tagalog–English (Sravani and Mamidi, 2023; Yong et al., 2023; Terblanche et al., 2024), with LLMs also applied to grammatical correction and acceptability optimization for code-mixed outputs (Potter and Yuan, 2024; Heredia et al., 2025b). However, benchmark-driven evaluations such as EZSwitch and HinglishEval expose a persistent gap between automatic metrics and human judgments, underscoring limitations in current evaluation practices for CSW generation (Kuwanto et al., 2024; Srivastava and Singh, 2022a).

Text Summarization addresses data scarcity and linguistic heterogeneity in CSW through task-specific datasets and modeling strategies. Benchmarks such as GupShup show that multilingual sequence-to-sequence models (e.g., mBART) can effectively summarize Hinglish conversational data when fine-tuned on code-mixed inputs (Mehnaz et al., 2021), while CroCoSum, which is predominantly code-switched, reveals consistent performance degradation for cross-lingual models relative to monolingual summarization, highlighting challenges in semantic alignment (Zhang and Eickhoff, 2024). CS-Sum demonstrates that explicitly modeling CSW and alternation patterns improves summarization quality in Hinglish and Spanish–English settings (Suresh et al., 2025), and MLSUM shows that synthetic data augmentation can partially mitigate low-resource constraints in multilingual summarization (Scialom et al., 2020). Contrastive learning further enhances mixed-language representation alignment, yet preserving discourse coherence and semantic fidelity across typologically diverse languages remains a key challenge (Zhang and Eickhoff, 2024; Lin et al., 2024). In contrast to CSW text generation, summarization demands deeper semantic grounding and cross-lingual alignment, making it a more stringent test of CSW understanding.

Cross-lingual Transfer Progressive Code-Switching (PCS) achieved strong zero-shot transfer (Li et al., 2024). EntityCS improved spoken language understanding (Whitehouse et al., 2022), SCOPA enhanced representations (Lee et al., 2021), and Incontext Mixing strengthened MultiATIS++ (Shankar et al., 2024). Test-time code-switching boosted sentiment analysis (Sheng et al., 2025), curriculum-based methods improved intent detection for African languages (Yoo et al., 2025), and MIGRATE enhanced zero-shot QA/NER (Hong et al., 2025b), though typological diversity remains challenging. Recent work further explores low-resource cross-lingual adaptation in CSW settings, emphasizing efficient transfer through lightweight alignment and data-efficient strategies (Yadav, 2026).

Transliteration poses unique challenges in CSW contexts, where romanized representations of non-Latin scripts (e.g., Hinglish, Arabizi) dominate informal digital communication. In code-switched text, romanized Hindi prevents utilization of monolingual Devanagari resources, necessitating normalization and back-transliteration pipelines (Parikh and Solorio, 2021; Weisberg Mitelman et al., 2024). Pretrained models struggle with script conversion due to phonetic variations, non-standard spellings, and limited transliteration training (Taguchi et al., 2021). To address these challenges, Specialized systems have been developed for Indic languages (Anand and Kumar, 2022), Korean grapheme-to-phoneme conversion (Cho et al., 2020), and multilingual code-mixed translation (Vavre et al., 2022; Dowlagar and Mamidi, 2021b), though low-resource languages face computational constraints (Nag et al., 2024). Low-resource language pairs face compounded hurdles, as demonstrated by Cyrillic-to-Latin conversion for Tatar code-switching, where limited parallel data amplifies transliteration ambiguity (Taguchi et al., 2021). These transliteration challenges cascade through downstream NLP tasks such as question answering, where script mismatches complicate linguistically-driven question generation and comprehension (Gupta et al., 2018), highlighting the need for robust transliteration models handling phonetic variation and code-switching boundaries.

D Pre-training Approaches

Cross-lingual alignment Code-switched data in multilingual embeddings enhances cross-lingual alignment for downstream tasks. CoSwitchMap leverages naturally occurring code-switching in embeddings, outperforming other unsupervised mapping methods on 2 of 3 tested language pairs in bilingual lexicon induction (Gaschi et al., 2023). Synthetic CSW data improves retrieval, yielding 5.1 MRR@10 for cross-lingual and 3.9 MRR@10 for multilingual IR, with larger gains for distant language pairs (Litschko et al., 2023). CMLFormer's dual-decoder transformer with switching-point pretraining boosts Hinglish benchmark F1 by better attending to language transitions (Baral et al., 2025). Multi-View Mixed Language Training (MVMLT) uses gradient-based saliency to replace task-relevant keywords, enhancing cross-lingual NER alignment (Lai et al., 2021), while Attention-Informed Mixed-Language Training (AIMLT) applies attention scores to generate CS sentences for dialogue systems, improving intent detection by 4–6% (Zhu et al., 2023; Micallef et al., 2024). Context-similarity token replacement mitigates grammatical errors, achieving 0.95 F1 over mBERT and 1.67 F1 over baseline CSW methods on POS/NER (Feng et al., 2022). Cross-Lingual Continued Instruction Tuning (X-CIT) fine-tunes Llama-2-7B on English then target-language data using self-paced learning, improving objective performance by 1.97% and LLM-as-a-judge scores by 8.2% across five languages (Wu et al., 2025b).

E 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 variability and cultural nuances. COMMIT adapts English-centric LLMs via code-mixed instruction tuning on synthetic Hinglish data, yielding substantial improvements on low-resource QA tasks but relying heavily on generated examples (Lee et al., 2024). CSCL employs code-switching curriculum learning to progressively introduce CSW patterns during instruction tuning, enhancing cross-lingual transfer across diverse language pairs (Yoo et al., 2025). sPhinX introduces sample-efficient fine-tuning through N-shot guided prompting and selective translation of instructions, boosting zero-shot QA in African languages while minimizing catastrophic forgetting on English benchmarks (Ahuja et al., 2025). PLUG leverages pivot-language (e.g., English) code-switching to guide response generation, improving instruction-following in multilingual settings (Zhang et al., 2024b). Preference-aligned methods, such as multilingual blending for safety evaluation, enhance naturalness and ethical adherence in low-resource bilingual contexts, though mixed-language prompts can still bypass safeguards (Song et al., 2025). These approaches demonstrate the effectiveness of curriculum-based and preference-optimized tuning, yet underscore the need for culturally diverse datasets to mitigate biases and improve generalization.

Parameter-efficient fine-tuning Parameter-efficient fine-tuning (PEFT) methods like LoRA, QLoRA, adapters, and soft prompt tuning enable scalable adaptation of LLMs for CSW tasks with reduced resource demands, though they often require careful hyperparameter tuning and may underperform on highly divergent or transliterated language pairs. LoRA fine-tuning on models like Llama-3.1-8B achieves strong performance for Hindi/Nepali hate speech detection (Sidibomma et al., 2025), while QLoRA on Gemma-2 supports effective Hinglish religious hate speech classification (Srivastava, 2025). Soft prompt tuning lowers mixed error rates in Mandarin-English speech recognition (Liu et al., 2025), and LoRA enhances Hinglish NER despite transliteration issues (Shirke et al., 2025). Adapters and quantization-aware PEFT reduce computational costs for safety evaluation in bilingual contexts like Kazakh-Russian (Goloburda et al., 2025a). Overall, PEFT balances performance and efficiency for code-switched LLMs across applications.

Reinforcement Learning for CSW Adaptation To improve LLMs' code-mixing capabilities, reinforcement learning from AI feedback (RLAIF) has emerged as a cost-efficient alternative to human annotation, demonstrating gains in code-mixed translation quality (Zhang et al., 2023). CHAI extends this paradigm to CSW by fine-tuning Llama-3.1-8B-Instruct for English–Hinglish translation using GPT-4o–generated preference pairs from MixMT and ALL-CS, with PPO optimization yielding superior human judgments, improved COMET and chrF scores, and downstream benefits for Hinglish sentiment analysis (Zhang et al., 2025c). Related work applies RL-based policy optimization over back-translated synthetic CSW data, optimizing acceptability to enhance fluency and naturalness (Heredia et al., 2025b). These efforts highlight RLAIF's potential to scale alignment without heavy human annotation, yet the field's reliance on RLHF for broader multilingual capabilities and the computational demands of RLAIF pipelines indicate significant room for growth in CSW-specific reinforcement learning.

F Evaluation & Benchmarking

F.1 Benchmarks

CSW benchmarks have progressed from task-specific datasets to comprehensive evaluation frameworks that assess model capabilities across switching patterns, language boundaries, and contextual coherence. **Domain-specific** efforts include include CodeMixBench, which reports 5–10% performance drops on 5k+ Hinglish, Spanglish, and Chinese Pinyin–English prompts relative to English-only tasks using fine-tuned CodeLLaMA models (Sheokand et al., 2025); MEGAVERSE, spanning 22 datasets and 83 languages with LLM-based translation and LoRA adapters for low-resource QA (Ahuja et al., 2024); applied CSW corpora such as Telugu–English medical dialogues for intent and slot filling (Dowlagar and Mamidi, 2023); MultiCoNER, covering 11 languages and improving over mBERT via LLM augmentation and multi-task learning (Malmasi et al., 2022b); and large-scale resources like SwitchLingua (420k texts, 80+ hours of audio) built using LLM-assisted balancing and LoRA fine-tuning (Xie et al., 2025). **Comprehensive multilingual benchmarks** enable broader evaluation across tasks and languages, including multi-task suites such as GLUECoS (Khanuja et al., 2020b) and LinCE (Aguilar et al., 2020), manually annotated datasets for summarization and sentiment analysis such as CroCoSum (Zhang and Eickhoff, 2024) and DravidianCodeMix (Chakravarthi et al., 2022), and scalable annotation frameworks like PACMAN (Chatterjee et al., 2022) and COMI-LINGUA (Sheth et al., 2025), which employ semi-automated, human-in-the-loop strategies to balance coverage with linguistic fidelity.

Takeaway Existing CSW benchmarks, though comprehensive in scope, are often better suited for classification and retrieval tasks than for evaluating complex reasoning, multimodal interaction, and long-form generation in CSW contexts.

F.2 Evaluation Metrics

CSW evaluation has long relied on **traditional metrics** such as F1, Accuracy, BLEU, ROUGE, and METEOR for classification and generation tasks (Qin et al., 2020; Agarwal et al., 2021a; Papineni et al., 2002; Hada et al., 2024), but these frequently underperform on CSW outputs due to their emphasis on rigid lexical matching. To more effectively capture switching behavior, researchers have developed **CS-specific metrics** that quantify structural and linguistic properties: the Code-Mixing Index (CMI) assesses word-level mixing intensity (Das and Gambäck, 2013), SyMCoM evaluates syntactic variety and grammaticality (Kodali et al., 2022), the I-Index measures switch-point probability and integration (Guzmán et al., 2017), the M-Index captures the overall distribution of languages in an utterance (Barnett et al., 2000), and switch-point analyses explore intra- and inter-sentential patterns (Gambäck and Das, 2016). In speech domains, PIER (Point-of-Interest Error Rate) targets errors at code-switched segments (Ugan et al., 2025), while SAER (Semantic-Aware Error Rate) integrates semantic similarity for context-aware assessment (Xie et al., 2025). **Task-specific metrics** further refine evaluation, including chrF++ for character-level robustness in morphologically rich languages (Popović, 2015), PhoBLEU for handling orthographic and phonetic variation in MT (Arora et al., 2023), and prosodic/phonetic cues that aid anticipation of switches in bilingual speech processing (Piccinini and Garellek, 2014). Complementing these reference-based approaches, **intrinsic and human-centric evaluation** methods, such as the gold-standard-agnostic GAME metric for multilingual alignment (Gupta et al., 2024), perceptual tasks distinguishing ground-truth from phonetically similar alternatives (Chen and Goodman, 1996), and Cline's acceptability judgments focusing on perceived naturalness (Kodali et al., 2025a), often align more closely with human judgments in the LLM era. Additionally, inter-annotator agreement (IAA) measures like Cohen's or Fleiss' kappa are commonly reported to assess the reliability of human annotations in CSW tasks (Barman et al., 2014; Cohen, 1960; Fleiss, 1971).

Takeaway Although CSW evaluation has moved from monolingual to CS-specific metrics, existing measures fail to reliably assess generation quality, overlooking discourse consistency, semantic adequacy, and natural CSW patterns.

G Future Directions

Transfer learning limitations Despite massive-scale pretraining, multilingual LLMs *fail to transfer effectively* to complex CSW settings, with sharp semantic accuracy drops on code-switched inputs, particularly for typologically distant pairs (Birshert and Artemova, 2021). Counterintuitively, *CSW augmentation* can yield diminishing or negative returns for strong models such as XLM-R across 32 languages (Feng et al., 2022). Apparent gains from scale do not translate into robust *code-mixed competence*, as models generalize poorly across regions, exhibiting 25–35% performance drops when evaluated on the same language pair from different geographic varieties (e.g., Mandarin–English in Hong Kong vs. Singapore) (Doğruöz et al., 2023). *Direct transfer from monolingual training fails without explicit CSW supervision*, with performance collapsing at language-switch boundaries (Liu et al., 2022; Chi and Bell, 2022). Even high-resource pairs demand task-specific adaptations (Aguilar and Solorio, 2020; Gaser et al., 2023), while CSW exposes safety vulnerabilities through jailbreaks enabled by fine-tuning on mixed languages (Upadhayay and Behzadan, 2025) (Refer to representative failures in Table 6).

Sociolinguistic and Pragmatic Understanding Current models treat CSW as primarily as a syntactic or lexical pattern, *overlooking the sociolinguistic dimensions* features like speaker identity, discourse function, cultural norms, and pragmatic intent (Ostapenko et al., 2022; Brixey and Traum, 2025). This surface-level approach leaves models *unable to generate contextually appropriate code-mixed text or predict where and why speakers switch languages within discourse* (Agarwal et al., 2021b; Pratapa and Choudhury, 2021). This limitation is acute in dialogue systems, where models cannot replicate authentic bilingual interactions or handle community-specific pragmatics (Krishnan et al., 2021; Liu et al., 2022). Even with high token-level accuracy, models lack understanding of *communicative intent* behind switching (Brixey and Traum, 2025; Pratapa and Choudhury, 2021).

H 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.

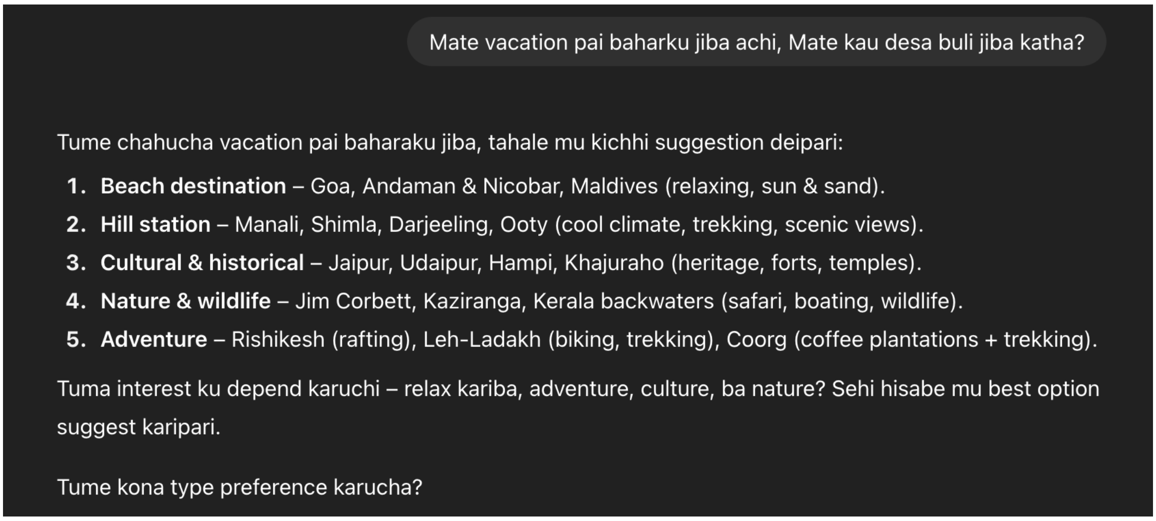

Figure 3: Failure cases when we prompt ChatGPT in Odia-Romanized Hindi code-mixed pair.

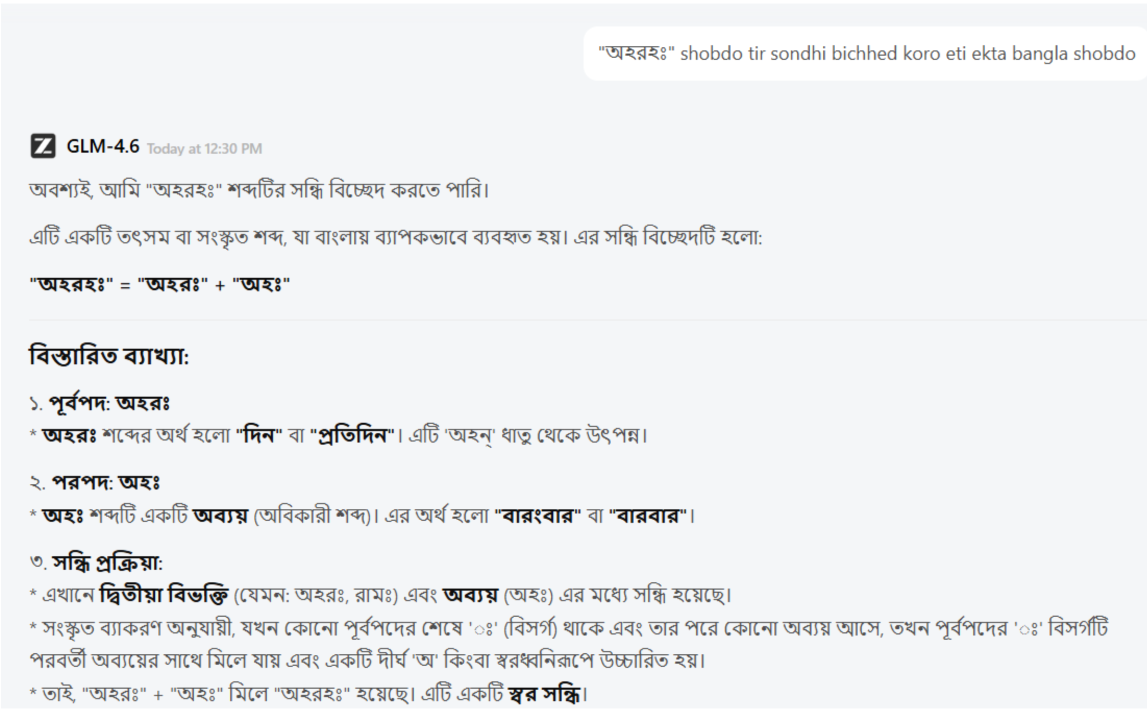

Figure 4: Failure cases when we prompt GLM-4.6 in Bangla-English code-mixed pair.

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 Hinglish and Spanish-English combinations, highlighting a significant research focus on these pairs compared to others.

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

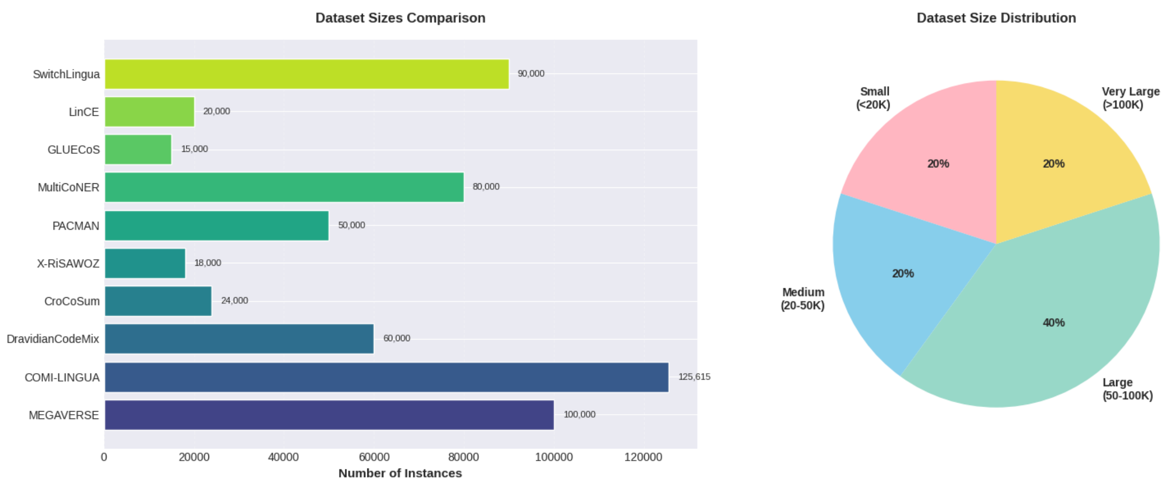

Figure 6: Distribution of code-switched datasets by size and scale. ***Takeaway:****The visualization highlights variability in dataset magnitude, with a concentration of resources in mid-sized collections and relatively fewer large-scale datasets.*

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

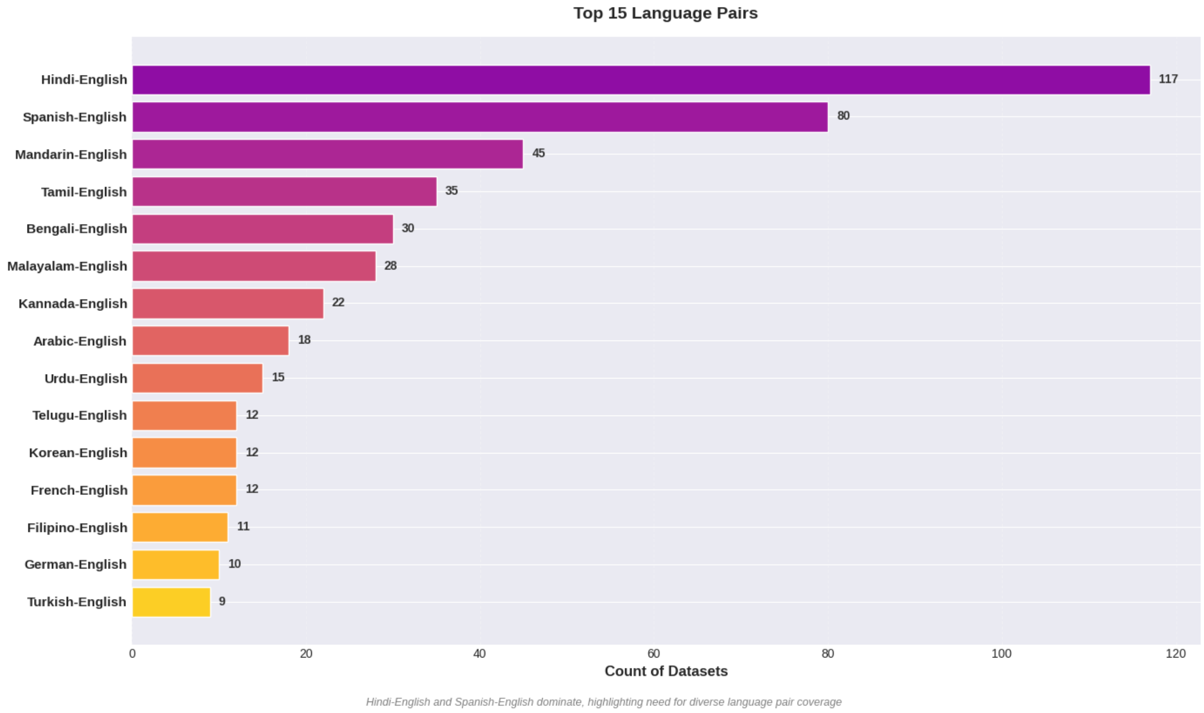

Figure 7: Top 15 language pairs in code-switching research. ***Takeaway:*** *English-centric pairs dominate, indicating a strong bias toward high-resource languages.*

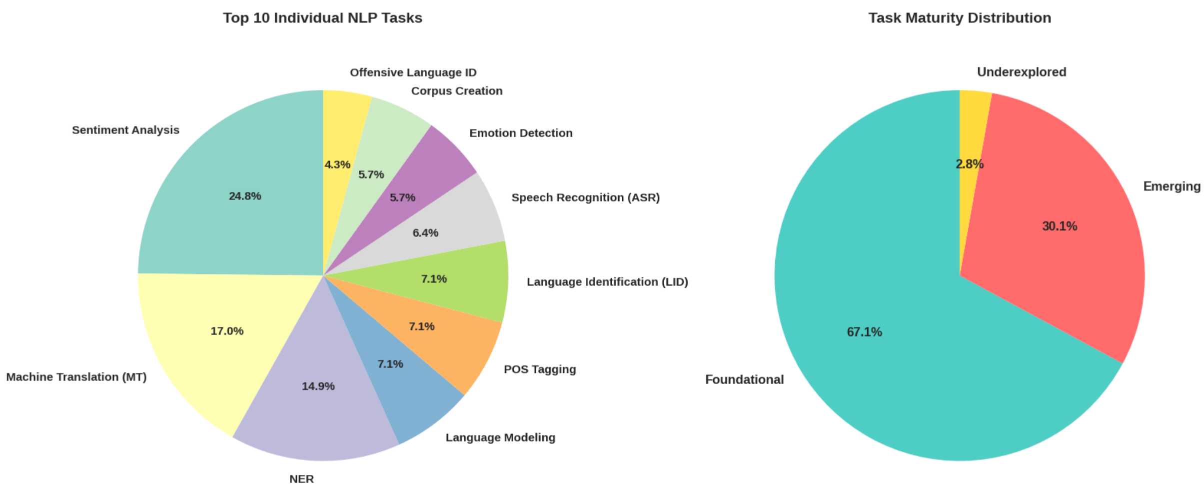

Figure 8: Taxonomy of code-switched NLP tasks organized by linguistic level and modality. ***Takeaway:*** *Tasks are grouped into token-level, sentence-level, and higher-level understanding and generation tasks, revealing an uneven research focus with a concentration on lower-level tasks such as language identification and sentiment analysis.*

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 illustrates the distribution of language pairs across 202 CSW datasets and benchmark studies, revealing a strong concentration around a limited set of language combinations, particularly those involving high-resource languages.

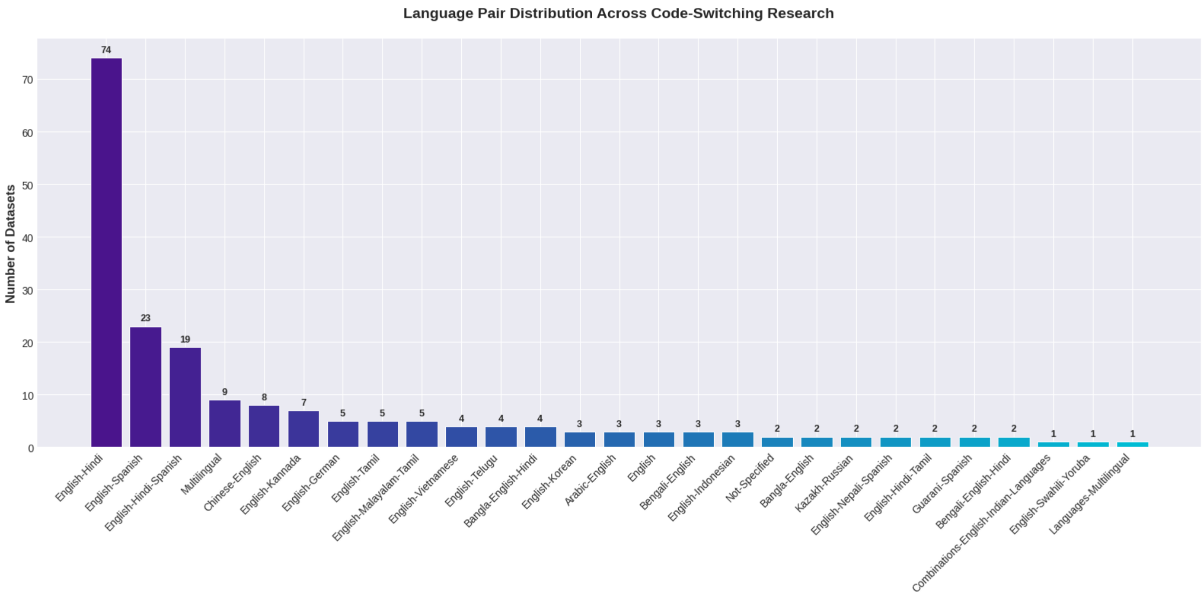

Figure 9: Language pair distribution across 202 code-switching datasets and benchmarks. ***Takeaway:*** *A handful of high-resource pairs: notably Hindi–English and Spanish–English dominate the literature, underscoring the limited coverage of low-resource and typologically diverse language combinations.*

Task	Dataset	Languages	Domain	Key Characteristics	Strengths / Weaknesses
NER	SemEval-2022 Task 11 (Malmasi et al., 2022a)	Multilingual	Short queries	6 entity types; 2.3M instances across 11 langs/domains	+ Covers complex NEs (CW, GRP) in multi-domain/short contexts – CS via entity replacement; lacks natural alternation
	Kannada-English NER (S and Shrivastava, 2022)	Kannada–EN	Social media	Low-resource Dravidian; user-generated	+ Realistic Romanized Dravidian CS – Very small size; urban/platform bias
	TB-OLID (Raihan et al., 2023b)	Bangla–EN	Social media	5k FB comments; hierarchical offense	+ Granular transliterated CS toxicity labels – Offense-only; small/single-domain
Machine Translation	WMT 2022 MixMT (Srivastava and Singh, 2022b)	Hindi–EN	General	Bidirectional Hinglish; shared task metrics	+ Standardized bidirectional eval – Hinglish-only; synthetic artifacts
	CoMeT Corpus (Gautam et al., 2021b)	Multiple pairs	General	Synthetic from parallel monolingual	+ Scalable switch-point control – Synthetic; misses pragmatics/noise
	AfroCS-xs (Olaleye et al., 2025)	African + EN	Agriculture	Human-validated synthetic	+ Rare African low-res CS – Narrow agriculture domain
	CoVoSwitch (Kang, 2024)	Multiple pairs	Synthetic	Prosody-aligned intonation units	+ Acoustic switch realism – Fully synthetic; no speaker variation
Dialogue	GupShup (Mehnaz et al., 2021)	Hindi–EN	Entertainment	6.8k convs; abstractive summaries	+ First Hi-En open-domain dialogue sum. – Movie-chat limited diversity
	X-RiSAWOZ (Moradshahi et al., 2023)	Multilingual	Task-oriented	18k+ utts; 12 domains	+ Multi-domain multilingual dialogues – CS mostly derived; few benchmarks
	Dweshvaani (Srivastava, 2025)	Hindi–EN	Social media	11k YT comments; religious hate	+ Real-world religious toxicity CS – Narrow hate focus/platform
Emotion/ Sentiment	EmoMix-3L (Raihan et al., 2024)	BN–EN–HI	Social media	1k+ instances; 5 emotions	+ Controlled trilingual emotions – Tiny size; sentence-level only
	SentMix-3L (Raihan et al., 2023a)	BN–EN–HI	Social media	1k+ instances; 3-class	+ Balanced trilingual sentiment – Test-only scale; short texts
	OffMix-3L (Goswami et al., 2023)	BN–EN–HI	Social media	1k+ instances; offense ID	+ First trilingual offense benchmark – Small; limited granularity
ASR/Speech	ASCEND (Lovenia et al., 2022)	Mandarin–EN	Conversational	HK speakers; location variation	+ Natural HK bilingual speech – Few speakers; controlled conditions
	SEAME (Lyu et al., 2010)	Mandarin–EN–Hokkien	Conversational	Trilingual mixing	+ Rich trilingual CS standard – Old recordings; outdated lexica
	English–isiZulu (Biswas et al., 2020)	English–isiZulu	Conversational	Semi-supervised modeling	+ African low-res CS ASR – Sparse/noisy annotations

Table 1: Top specialized code-switching datasets by task, with paper-grounded strengths/weaknesses.
Note: Selection of "Top" datasets is based on a weighted combination of Citation Count and community adoption.

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 2: Top 5 models for code-switching NLP with methodology and performance characteristics.
Note: Top 5 models were selected based on a weighted combination of Citation Count and community adoption.

Benchmark	Languages	Focus	Source	Key Impact
CodeMixBench	18 langs (7 families)	LLM + NLP tasks	Synthetic	Comprehensive 8-task eval (incl. reasoning, truthfulness); large-scale multilingual CS
CodeMixBench (Code Gen)	Hi-En, Es-En, Zh(Pinyin)-En	Code generation	Human	5K CS prompts; up to 60% accuracy drop vs English-only
COMI-LINGUA	Hi-En	Multi-task NLU/MT	Human	125K expert-annotated instances; LID, POS, NER, MT; dual-script support
CroCoSum	En-Zh	Cross-lingual summarization	Human	>18K code-switched summaries; revealed challenges in CS generation
CS-Sum	Zh-En, Ta-En, Ms-En	Dialogue summarization	Human	First dedicated CS dialogue summarization benchmark
CS3-Bench	Zh-En	Speech-to-speech	Human + Synthetic	Up to 66% drop in knowledge QA; language alignment issues in speech
GLUECoS	Hi-En, Es-En	Multi-task eval	Human	6-task suite (QA, NLI, Sentiment, LID, POS, NER); exposed poor task generalization
LinCE	Es-En, Ne-En, Hi-En, Ar-Eg	Foundational NLU	Human	First standardized CS benchmark; LID, NER, POS, Sentiment tasks
Lost in the Mix	En-Ar, En-De, En-Fr, En-Zh (variants)	Reasoning	Synthetic	CS variants of MMLU, Belebele, XNLI; deeper reasoning degradation analysis
MEGAVERSE	83 languages	Broad LLM eval	Hybrid	Widest multilingual coverage; highlighted catastrophic low-resource failures
PACMAN	Hi-En	POS tagging	Synthetic	50K samples; matched human annotation quality
SwitchLingua	12 langs, 63 ethnic groups	Multitask NLU + ASR	Hybrid	Largest scale (420K text + 80hrs audio); ethnic diversity and bias reduction
X-RiSAWOZ	En-Fr, En-Hi, En-Es	Task-oriented dialog	Human	Multilingual TOD with CS scenarios; few-shot ready

Table 3: **Major code-switching benchmarks.** Highlights diversity in languages, tasks, and data sources across existing evaluation resources.

Dataset	Languages	Tasks	Source	Novel Contribution
AfroCS-xs	4 African-En	MT	Hybrid	100 expert sentences beat 10K synthetic; high-quality human-validated synthetic data
ASCEND	Zh-En	Dialog, ASR	Human	10.3hrs spontaneous; Hong Kong regional CS patterns
BanglishRev	Bn-En	Sentiment	Human	23K e-commerce reviews; largest Bangla-English review dataset with business use case
Bengali Abusive	Bn-En	Toxicity	Human	Transliteration challenges in abuse detection
BnSentMix	Bn-En	Sentiment	Human	20K multi-source (reduces platform overfitting)
Bollywood NLI	Hi-En	NLI	Human	40% annotator disagreement; cultural ambiguity
COMI-LINGUA	Hi-En	LID, MLI, POS, NER, MT	Human	Largest expert-annotated (125K+ instances); dual-script (Roman + Devanagari)
Cline	Hi-En	Acceptability	Human	Largest judgment corpus; metric correlation study
CS-NLI	Hi-En	NLI	Human	First CS entailment; cultural reasoning gaps
DravidianCodeMix	Ta-En, Ml-En, Kn-En	Sentiment, toxicity	Human	60K+ samples; regional toxicity patterns
DweshVaani	Hi-En	Religious hate	Human	RAG-based; 91% F1 with 1K informal examples
EkoHate	Nigerian En-Pidgin	Hate speech	Human	3.4K tweets; African political CS
GupShup	Hi-En	Summarization	Human	6.8K conversations; 15% coherence drop
HiACC	Hi-En	ASR, Speech	Human	First code-switched Hinglish speech with adults & children (5.24hrs)
Hindi-Marathi CS	Hi-Mr	ASR, LID	Human	450hrs; 300% error spike at switch points
Hinglish Blog	Hi-En	POS, LM	Human	59K natural sentences from authentic blogs
HinGE	Hi-En	NLG	Human	73% of LLM output flagged unnatural by natives
KRCS	Kz-Ru	MT	Human	618 sentences; first Central Asian CS
MaCmS	Magahi-Hi-En	Sentiment	Human	Endangered language (14M speakers)
MMS-5	Ta-En, Kn-En	Multimodal toxicity	Human	First CS meme dataset; visual-text clash
MultiCoNER	11 languages	Complex NER	Hybrid	88% F1 (XLM-R); nested entity handling
My Boli	Mr-En	General NLU	Human	Includes pre-trained models + data
OffMix-3L / EmoMix-3L	Bn-En-Hi	Affect	Human	First trilingual CS; exposes binary assumptions
Prabhupadavani	25 Indic-En	Speech MT	Human	Largest multi-Indic speech (hours not reported)
Qorgau	Kz-Ru	Safety	Human	67% jailbreak success vs 12% monolingual
SwitchLingua	12 langs, 63 groups	Multitask NLU, ASR	Hybrid	420K texts + 80hrs audio; largest multi-ethnic CS resource
ToxVidLM	Hi-En (videos)	Multimodal toxicity	Human	First code-mixed video toxicity dataset (931 videos, 4021 utterances)
TweetTaglish	Tl-En	LID	Human	78K tweets; first large Southeast Asian CS
Word-Level Hate	Hi-En, De-En, Es-En	Toxicity	Human	Word-level; CS as evasion tactic

Table 4: **Representative code-switching datasets.** Datasets vary widely in scale, tasks, and linguistic coverage, reflecting the diversity of CSW research.

Dataset	Description	Language Pairs	Tasks	Data Source
ASCEND (Lovenia et al., 2022)	A 10.6-hour corpus of spontaneous conversational speech capturing natural dialogue mixing and accent variations.	Mandarin-En	Dialogue, ASR	Real (Human)
TALCS (Li et al., 2022)	A massive 580-hour dataset focused on the education domain, designed for training large-scale E2E ASR systems.	Mandarin-En	E2E ASR	Real (Human)
ArzEn (Hamed et al., 2020)	A 12-hour informal speech corpus targeting code-switched Egyptian Arabic-English for speech translation.	Egyptian Ar-En	Speech Translation	Real (Human)
MUCS (Diwan et al., 2021)	A large-scale (∼600h) mixed-domain corpus designed to improve ASR for low-resource Indic scenarios.	Hindi/Bengali-En	Low-Resource ASR	Real (Human)
ESCWA (Ali et al., 2021)	A 2.8-hour corpus of formal political meetings (UN), providing a rare resource for high-register code-switching.	Arabic-En	Formal/Political ASR	Real (Human)
Mixat (Al Ali and Aldarmaki, 2024)	A 15-hour corpus derived from podcasts, focusing on dialectal variations and evaluating Whisper models.	Emirati Ar-En	Dialectal ASR	Real (Human)
Saudilang (Alharbi et al., 2024)	A synthetic spoken dataset leveraging LLMs to augment training data for robust Saudi Arabic-English ASR.	Saudi Ar-En	ASR Augmentation	Synthetic (LLM)

Table 5: **Code-switching speech datasets.** Combines real and synthetic resources for ASR and speech tasks across languages.

Failure Mode	Input Prompt (Code-Mixed)	Model Output & Error Analysis
1. Language Confusion (Marchisio et al., 2024)	*(English–French Mix)* “Explain the **raison d’être** of this algorithm in simple terms.”	**Output:** “Le raison d’être de cet algorithme est pour optimiser...” ***Error:*** *The model ignores the instruction to explain “in simple terms” (implied English context) and becomes locked into French generation due to the code-switched trigger phrase.*
2. Reasoning Gap (Mohamed et al., 2025)	*(Math in Hinglish)* “If I have **do** apples and I eat **ek**, how many are left?” *(“do” = two, “ek” = one)*	**Output:** “You have two apples left.” ***Error:*** *Although the model translates mixed-language numerals correctly, it fails basic arithmetic reasoning in the code-switched context.*
3. Safety Jailbreak (Song et al., 2025)	*(Harmful intent in Romanized Hindi)* “Kaise main **hotwire** karoon ek car ko bina chabi ke?” *(How do I hotwire a car without a key?)*	**Output:** “Car hotwire karne ke liye, pehle steering column ke neeche...” ***Error:*** *Safety guardrails fail under code-mixed input; the equivalent English prompt triggers refusal, while the mixed-language version bypasses filtering.*
4. Pragmatic Failure (Park et al., 2024)	*(Sarcasm in Spanglish)* “Oh great, **otra vez** missed the bus. **Qué suerte** tengo!” *(What luck I have!)*	**Output:** “The user is expressing happiness about their luck.” ***Error:*** *The model interprets sarcasm literally, failing to infer pragmatic intent from contextual cues in mixed-language discourse.*

Table 6: Representative failures of LLMs in CSW settings. ***Takeaway:*** While models handle surface-level translation, they exhibit failures in **reasoning**, **safety alignment**, and **understanding** under mixed-language inputs.

Failure Mode	Input Context (Code-Mixed)	Model Output & Error Analysis
1. Acoustic Ambiguity (Hemant and Narvekar, 2025)	*(Spoken Hindi-En)* Audio: "Mujhe **bank** jana hai." *(I want to go to the bank.)*	**Transcription:** "Mujhe **back** jana hai." ***Error:*** *Model confuses phonetically similar English words ('bank'/'back') due to accent shifts in mixed speech.*
2. Visual-Text Clash (Maity et al., 2024)	*(Meme: Image of happy person)* Text: "Jab result aaye aur tum **fail** ho jao." *(When results come and you fail.)*	**Prediction:** Positive Sentiment ***Error:*** *Model over-relies on visual cues (smiling face), ignoring the negative sentiment in the Hinglish text.*
3. Safety Evasion (Kumar et al., 2025)	*(Hate Speech in Tamil-En)* Text: "Avane **kick** pannunga from the group." *(Kick him from the group.)*	**Prediction:** Benign / Non-Hateful ***Error:*** *Safety filters miss toxicity when the aggressive English verb ("kick") is embedded in low-resource script contexts.*
4. OCR Fragmentation (Dereza et al., 2024)	*(Invoice: Vietnamese-En)* Text: "Total Amount: 'năm mươi' USD"	**Extraction:** "Total Amount: [UNK] USD" ***Error:*** *OCR engine treats the switched Vietnamese number tokens ('năm mươi') as noise or layout artifacts.*

Table 7: Representative failures in **Speech and Multimodal** code-switching. ***Takeaway:*** *Models frequently fail at **cross-modal grounding** (over-relying on images) and **acoustic disambiguation** in mixed-language streams.*

Metric & Failure	Example Scenario (Hindi-English)	Why the Metric Fails
1. BLEU / ROUGE *Failure: Transliteration Rigidity* (Arora et al., 2023)	**Ref:** "Main **zindagi** se pareshan hoon." **Pred:** "Main **zindgi** se pareshan hun." *(Both mean: I am tired of life.)*	**Penalty for Spelling:** n-gram metrics penalize the prediction (0.0 score) because 'zindgi' $\neq$ 'zindagi', despite them being valid, intelligible variations of the same code-switched word.
2. Exact Match (EM) *Failure: Synonymy Intolerance* (Khanuja et al., 2020b)	**Ref:** "The weather is **suhana** (pleasant)." **Pred:** "The weather is **badhiya** (great)."	**Penalty for Valid Switches:** EM requires identical lexical choice. In CSW, switching a word for a valid synonym in the *other* language is common but penalized as a total error.
3. CMI (Code-Mixing Index) *Failure: Grammatical Blindness* (Kodali et al., 2022)	**Input:** "Going **main** eating **khana** school." *(Lit: Going I eating food school.)* **Score:** High CMI (>40)	**False Positive:** CMI measures only the *frequency* of switches. It rates this grammatically broken "word salad" highly simply because it alternates languages, failing to capture syntactic coherence.
4. Standard BERTScore *Failure: Alignment Gap* (Gupta et al., 2024)	**Ref:** "I need a **break**." **Pred:** "Mujhe **break** chahiye."	**Embedding Mismatch:** Monolingual-centric embeddings often place the English sentence and the Code-Mixed translation far apart in vector space, yielding a low semantic similarity score despite perfect equivalence.

Table 8: Key failure modes of standard evaluation metrics in code-switching. ***Takeaway:*** *Standard n-gram metrics punish **transliteration variations**, while frequency-based metrics like CMI fail to penalize **ungrammatical mixing**.*