

# Disentangling Codemixing in Chats: The NUS ABC Codemixed Corpus

Svetlana Churina, Akshat Gupta, Nur Insyirah Binte Imam Mujtahid and Kokil Jaidka  
Centre for Trusted Internet & Community,  
National University of Singapore,  
Singapore

## Abstract

Code-mixing involves the seamless integration of linguistic elements from multiple languages within a single discourse, reflecting natural multilingual communication patterns. Despite its prominence in informal interactions such as social media, chat messages and instant-messaging exchanges, there has been a lack of publicly available corpora that are author-labeled and suitable for modeling human conversations and relationships. This study introduces the first labeled and general-purpose corpus for understanding code-mixing in context while maintaining rigorous privacy and ethical standards. It includes over 355,641 messages spanning various code-mixing patterns, with a primary focus on English, Mandarin, and other languages. We expect the Codemix Corpus to serve as a foundational dataset for research in computational linguistics, sociolinguistics, and NLP applications. Code and dataset sample can be found [here](#).

## 1 Introduction

Instant messaging has become a primary communication mode across personal, professional, and educational contexts. Unlike social media platforms that promote performative, public content, messaging offers a private, real-time, and intimate space for emotional expression and relationship maintenance. Because such exchanges typically occur in dyads - reciprocal, turn-by-turn interactions - they provide a natural setting to observe how people adapt, disclose, and sustain relationships through language (Burlison, 2009; Walther, 2007). While prior studies have examined visual and affective cues such as emojis in digital communication (Boutet et al., 2021; Erle et al., 2022; Riordan, 2017), far less is known about how linguistic style - particularly code-mixing - varies with relational depth and emotional intimacy.

Across the world, roughly one-third of all countries are officially bilingual with English as a co-

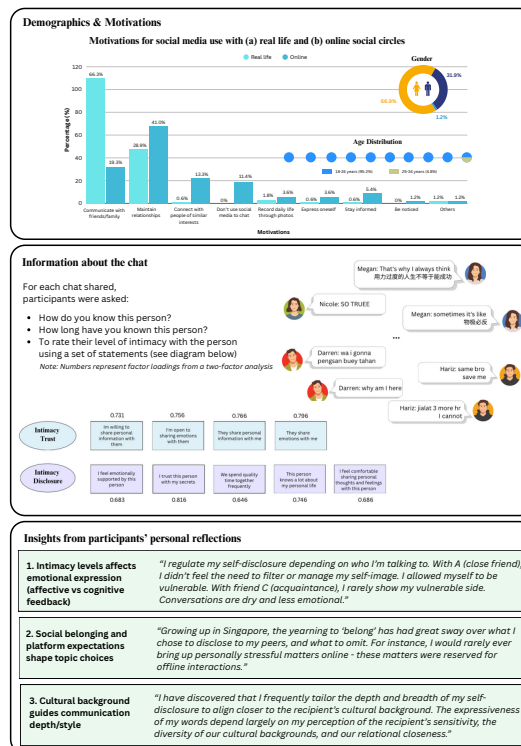


Figure 1: Top: Information collected about survey participants. Middle: Sample chat snippets illustrating varying tones and relational dynamics. Participants rated their chats on intimacy using items grouped under two dimensions—*Trust* and *Disclosure*. Bottom: Insights from participants' reflections on communication practices.

official language, and about 35% of their populations live in heterogeneous urban centers, where daily communication often spans multiple languages and scripts. These multilingual environments - especially in Southeast Asia, where English coexists with vernaculars like Malay, Mandarin, and Tamil - make code-mixed messaging a routine form of social interaction. Yet despite this linguistic reality, multilingual communication remains underexplored in computational linguistics: of the 113,590 papers in the ACL Anthology, only

455 explicitly address code-mixed language.

In this study, we introduce the **Air Batu Campur (ABC) Codemixed Corpus**—named after a layered Southeast Asian dessert that reflects the dataset’s linguistic diversity. The ABC Corpus is a publicly available collection of instant messages written in English and regional languages, gathered through voluntary data donation. Participants contributed conversations with three partners representing different levels of relational intimacy and completed validated scales measuring trust and disclosure. As an ongoing project, the corpus continues to expand through verified submissions enriched with metadata and linguistic statistics. Building on prior instant messaging corpora (Chen and Kan, 2013), it integrates linguistic and psychosocial dimensions within a privacy-aware donation framework.

The ABC Corpus links conversational dynamics with relational intimacy, providing a foundation to examine how individuals use language to negotiate closeness, express emotion, and sustain well-being in digital communication. It enables analysis of language variation and code-mixing across relationship types, as well as dyadic mechanisms such as emotional expression and linguistic adaptation.

To systematically explore these phenomena, we conducted a multi-stage analysis spanning linguistic, social, and computational dimensions. We first developed a language tagging pipeline combining rule-based, embedding-based, and LLM-assisted identification for multilingual chat data. We then applied LLM-based translation to create parallel English text for cross-lingual analysis and benchmarking. Next, we modeled linguistic accommodation and gender differences using mixed-effects regression to understand how speakers align their code-mixing and message length in response to conversational partners. Beyond structural analysis, we performed a two-level content analysis integrating psycholinguistic features and the Systemic Functional Linguistics (SFL) framework to interpret how code-mixing supports emotional and relational strategies. Finally, we demonstrate the dataset’s practical utility by fine-tuning a GPT model on the corpus to generate more contextually appropriate and authentic code-mixed conversations.

Together, these contributions offer an integrated perspective on the linguistic, social, and computational aspects of code-mixing in private digital communication. The following sections outline the theoretical motivation, data collection proce-

dures, and analyses linking linguistic behavior with intimacy and relationship context.

## 2 Background

Southeast Asia is one of the most ethnically and linguistically diverse regions in the world (Leng, 1980), with each country comprising multiple ethnic, linguistic, and religious groups (Takagi, 2009). Centuries of migration and cultural exchange have shaped its multilingual landscape, where several languages coexist. Yet globalization has accelerated the assimilation of vernaculars, as dominant languages—particularly English—gain prominence in daily communication (Singh et al., 2012; Ramli et al., 2021).

English has become central to socioeconomic mobility and national development across the region (Albury, 2016; Sercombe, 2019). In Singapore, bilingual education policies introduced in 1996 established English as the main language of instruction while promoting mother tongues as second languages (Leimgruber, 2013). This framework fostered both English dominance and the persistence of diverse vernaculars in informal contexts (Gupta, 1989).

Within this environment, code-mixing naturally emerged, giving rise to Singlish—a colloquial English variety blending Mandarin, Hokkien, Cantonese, Malay, and other elements. Singlish embodies the tension between global and local identities (Alsagoff, 2007) and serves pragmatic functions such as signaling group identity (Kipchoge, 2024), facilitating communication (Sumartono and Tan, 2018), and fostering rapport (Bolton and Botha, 2019).

## 3 Prior Work

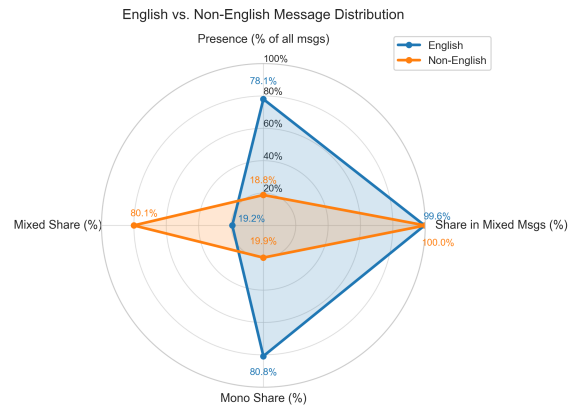
Private messaging fulfills both interpersonal and intrapersonal functions, enabling emotional regulation, social support, and sense-making after negative experiences (Zech et al., 2004; Pennebaker, 1997). These effects arise from (i) the *therapeutic impact of expressive communication*, which aids cognitive restructuring, and (ii) the *social sharing mechanism*, where disclosure in trusted relationships promotes belonging and recovery. Digital contexts further amplify such processes by lowering social constraints and encouraging authentic emotional expression. Understanding them thus requires examining *dyadic communication*, where each message responds to a partner’s prior turn.

Part A: Overall Corpus Statistics	
Metric	Value
Total Messages Collected	355,641
Unique Contributors	154
Languages Represented	English, Mandarin, Hokkien, Tamil, Malay, Japanese, Korean
Average Message Length	4.3 tokens
Code-Mixed Messages (%)	22%

Part B: Message Statistics by Intimacy Level			
Statistic	Acquaintances	Frequent Chatting	Daily Interactions
Total Messages	14,166	44,590	280,745
Mean Length (Words)	6.82	4.59	4.12
Median Length (Words)	5.0	3.0	3.0
Std. Dev. Length	9.51	33.42	5.36
Code-Mixing (%)	7.42	12.26	19.25
Transliteration (%)	2.21	4.16	0.75

a. **Corpus statistics:** Corpus composition and message statistics



b. **Radar chart:** English vs. non-English message distributions

Figure 2: **Corpus statistics and comparative message distribution.** The Table (a) summarizes corpus-level and intimacy-based message statistics. The radar chart (b) contrasts English and non-English message distributions in terms of presence, mixed-share, and mono/mixed split.

Although code-mixing has attracted growing attention, few large-scale corpora capture its real-world conversational dynamics. Existing datasets often focus on single language pairs, omit speaker metadata, or draw from scripted or social media text. Examples include L3Cube-HingCorpus (Nayak and Joshi, 2022), Bollywood Romanized Corpus (Khanuja et al., 2020), NSC (Koh et al., 2019), CoSEM (Gonzales et al., 2023), NUS-SMS Corpus (Chen and Kan, 2013), and TweetTaglish (Herrera et al., 2022). While (Foo and Ng, 2024) explore Singlish discourse particles and BOLT Phase 2 (Song et al., 2014) captures informal chats, these resources remain small or lack relational annotation. Our dataset addresses these gaps by offering authentic, multilingual private messages from Singaporeans across relationship types, with demographic and relational metadata and natural code-mixing across English, Chinese, Malay, and Tamil.

## 4 Methodology

To assess the corpus’s linguistic reliability and analytical potential, we report three complementary evaluations. First, we benchmark the natural language processing of code-mixed data through an evaluation of the language-tagging accuracy across three multilingual models—Qwen2.5, fastText, and XLM-R—using human annotations as the reference standard (Section 5.1). We also evaluate the translation quality on a stratified subset of conversations to test whether large language models can accurately render code-mixed content into coherent English (Section 5.2). Second, we model conversational

adaptation through generalized linear mixed effects analyses, testing whether speakers align their code-mixing behavior and verbosity to that of their partners, and how these tendencies vary by trust, gender, and relationship type (Section 5.3 & 5.4). Third, we examine its research potential through a study of its linguistic characteristics (Section 6) and its broader applicability to generate more fluent and culturally grounded responses in code-mixed conversations (Section 7).

### 4.1 Participant Recruitment

The ABC Codemixed Corpus comprises instant message conversations donated by Singaporean university students. Over a six-month period, 477 chat threads were collected through a campus-wide recruitment call, representing everyday exchanges across varying levels of relationship intimacy. Participants first completed a short survey capturing demographic profiles, language background, and social media use. To ensure informed and consistent participation, they were introduced to the concept of code-mixing through orientation questions within the survey. They were then invited to contribute three chat conversations by following a step-by-step procedure that guided them to open their preferred messaging application and copy a six-month excerpt of a conversation into a secure text submission box. For each donated conversation, participants also reported the demographics of their conversation partner, the nature of their relationship, and their perceived intimacy using a validated multi-item Likert scale (Laurenceau

et al., 1998), capturing subjective relational closeness rather than a dyadic measure. Due to the private nature of the data collection, only one-sided perceptions of relational closeness were obtained. Of 332 recruits, 166 contributed between one and three conversations (50% participation).

The following paragraphs describe the preprocessing and analytical steps used to prepare the dataset for linguistic analysis.

## 4.2 Corpus Statistics and Analysis

The final dataset comprises 355,641 messages from 166 participants, most aged 18-24 (Figure 1). Demographic information and platform usage patterns are detailed in Appendix D.

We analyzed code-mixing across languages using Qwen2.5-generated language tags, focusing on three complementary indicators: (i) message-level language presence, (ii) relative share of each language within mixed messages, and (iii) the balance between monolingual and mixed usage.

As shown in Figure 2, English dominates overall message volume but contributes less within mixed utterances, while regional languages—particularly Hokkien and Mandarin—play proportionally stronger roles in code-mixed exchanges. Tamil exhibited the lowest presence, and less frequent languages (Arabic, French, Spanish, etc.) were grouped under *Other*. Figure 3b further shows that most languages appear primarily in code-mixed settings, with only about 20% of their occurrences in purely monolingual form.

Overall, approximately 30% of all messages exhibit some degree of code-mixing, varying by relationship type and message length. Shorter messages (0-5 tokens) show the highest mixing rates, reaching 36.7% in intimate conversations, whereas professional exchanges retain moderate mixing even in brief turns.

## 4.3 Data Preprocessing, Language Identification, and Translation

The dataset comprises user-generated WhatsApp and Telegram messages, organized chronologically within conversation threads. Conversational structure was preserved, though each message was processed individually to support both intra- and inter-sentential code-mixing analyses.

Preprocessing removed non-linguistic content such as URLs, emails, and timestamps while retaining emojis and casing for linguistic cues (Figure 4).

Duplicated chats were excluded to maintain data integrity.

Language identification combined rule-based (WordNet, wordfreq) and model-based methods, including fastText (Joulin et al., 2017), XLM-RoBERTa (Papluca, 2020), and Qwen2.5-14B-Instruct (Hui et al., 2024), the latter prompted for Southeast Asian texting patterns. The corpus was also translated into English using Qwen2.5-14B-Instruct at the conversation level to preserve context. While the tagging and translation performance of our pipeline is reported in Section 5, the full details on the preprocessing, tagging, and translation pipelines, along with prompts and post-processing details, are provided in Appendix E and Appendix B.

## 4.4 Measured variables

We derived the following variables for analysis:

- **Code-mixing percentage:** computed sentence-wise by weighting each sentence's code-mix ratio by its length and aggregating to the message level.
- **Adaptation:** binary variable indicating whether a speaker's code-mix level shifted relative to the prior turn by the other speaker; turns from the same speaker in succession were excluded.
- **Dyad composition:** included absolute age difference and shared gender within conversational pairs (participant-partner). As most interactions occurred within age groups, age difference showed minimal variance.
- **Relationship type:** categorized as Close (family, close friends, partners), General (friends, acquaintances, community members), or Work (classmates, colleagues).
- **Intimacy factors:** factor analysis of nine self-reported items on emotional and informational sharing (Laurenceau et al., 1998) yielded two latent dimensions—*self-disclosure* and *trustworthiness* (see Appendix C).

All variables were standardized before mixed-effects models were estimated to understand the influence of code-mixing on participant behavior and perceived intimacy, with conversation as a random intercept to account for repeated measures within dyads. Full coefficient tables are reported in Appendix C.

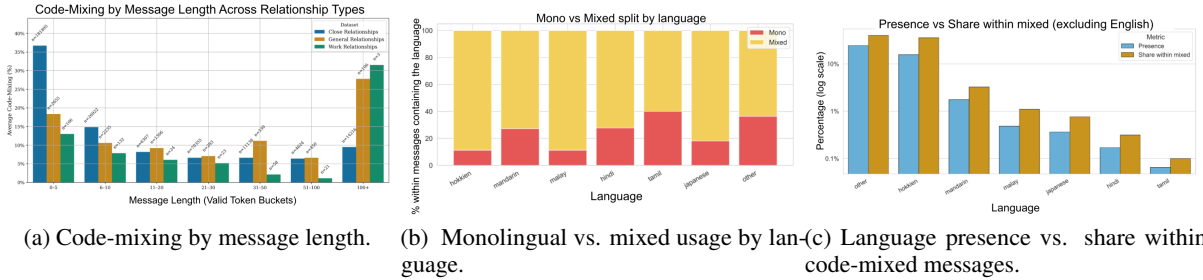


Figure 3: Overview of code-mixing patterns in the ABC Codemixed Corpus. (a) Average percentage of code-mixing across message length buckets (by token count) grouped by relationship type. (b) Distribution of monolingual (Mono) and code-mixed (Mixed) usage by language. (c) Presence of each language in the corpus versus its proportional share within code-mixed messages (excluding English).

Model	Accuracy (no English)	Accuracy (incl. English)
fastText	0.26	0.69
papluc/XLM-RoBERTa	0.18	0.71
Qwen2.5 (LLM)	<b>0.52</b>	<b>0.80</b>

Table 1: Accuracy of language tagging models compared to human annotations on 1,814 messages.

Original	Baseline_FT	Baseline_XLMR	Qwen2.5
lemme try	(lemme, sl), (try, en)	(lemme, it), (try, en)	(lemme, en), (try, en)
nihao i just had dinner	(nihao, en), (i, en), (just, en), (had, en), (dinner, en)	(nihao, sw), (i, en), (just, en), (had, en), (dinner, en)	(nihao, zh), (i, en), (just, en), (had, en), (dinner, en)
我想要脱离 nus	(我, zh), (想要, zh), (脱离, jp), (nus, vi)	(我, zh), (想要, zh), (脱离, zh), (nus, hi)	(我, zh), (想, zh), (要, zh), (脱离, zh), (nus, propernoun)

Table 2: Token-level language tags from Qwen2.5 compared to baseline methods. Codes: *zh* = Mandarin, *jp* = Japanese, *en* = English, *sw* = Swahili, *sl* = Slovenian, *vi* = Vietnamese, *hi* = Hindi, *it* = Italian.

## 5 Results

### 5.1 Tagging Performance

To evaluate the tagging performance in our preprocessing pipeline, we created a stratified random sample of 1,814 messages balanced across relationship types and code-mixing levels. Sampling combined language-tag and code-mixing based strata to ensure linguistic and contextual diversity across the most frequent languages (covering over 90% of all tokens). Language tags produced by our primary approach (Qwen2.5) were compared with two reference models—fastText (Joulin et al., 2017) and XLM-R (Papluca, 2020)—and human annotations by two Singaporean annotators (Cohen’s  $\kappa=0.74$ ).

As shown in Table 1, Qwen2.5 achieved the highest agreement with human annotations, outperforming both baselines. While overall accuracy

Baseline	BLEU	BERT_F1	ROUGE-1_F	ROUGE-L_F
Google Translate	4.55	0.87	0.49	0.46
NLLB-200	4.47	0.81	0.40	0.35

Table 3: Similarity scores of baseline translations compared to Qwen2.5 outputs.

Original	Baseline_NLLB	Baseline_GT	Qwen2.5
Omg <b>jiayou</b> he said that <b>abt</b> my prev article also I got his approval after the second article	Omg <b>jiayou</b> he said that <b>abt</b> my preview article also i got his approval after the second article	Omg <b>jiayou</b> he said that <b>abt</b> my prev article also I got his approval after the second article	Omg <b>cheer up</b> he <b>mentioned</b> that <b>about</b> my previous article too I received his approval after the second article

Table 4: Comparison of Qwen2.5 translations with baselines. Red and blue highlight code-mixed and informal words.

is higher when including English tokens (0.80), performance on non-English tokens within code-mixed segments remains substantially lower (0.52), reflecting the inherent difficulty of fine-grained language identification in informal multilingual text. Many of these errors arise from ambiguity between closely related languages and discourse particles common in Singlish. Importantly, misclassifications rarely affect the distinction between English and non-English tokens, which is the primary signal for identifying code-mixed segments. To further ensure robustness, we validate key findings on a subset of human-annotated data and observe consistent trends. Token alignment between models was conducted using the Needleman-Wunsch algorithm (Needleman and Wunsch, 1970) to resolve paraphrasing and segmentation differences. Representative examples are shown in Table 4.4 and Figure 6.

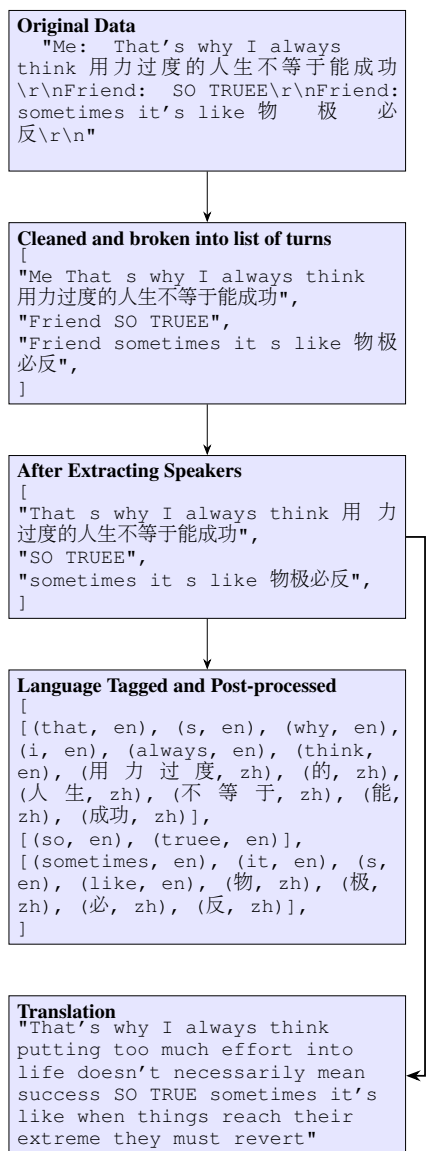


Figure 4: Language tagging and translation pipeline (en = English, zh = Mandarin).

## 5.2 Translation Performance

To evaluate translations, we used a stratified random sample of 100 conversations, each containing non-English tokens and spanning a range of code-mixing levels and language combinations, covering over 90% of all tokens. Its hard to conduct human annotation especially for such rich multilingual texts, therefore we compared our translations to two strong baselines: one generated using the Google Translate API and the other using Facebook’s NLLB-200 MT model. Both baseline translations were performed at the message level, and a single aggregated output was created for comparison. This was necessary due to the limited input handling capabilities of the two baseline models. The similarity scores for our LLM-based transla-

tions with the two baselines are shown in Table 3. We made the following observations about the translation performance:

- High BERT scores indicate strong semantic similarity with the baselines, while low BLEU scores reflect lexical divergence—likely due to paraphrasing, as seen in Table 4.
- The LLM-based approach was much better at translating transliterated words, also highlighted in red in Table 4.

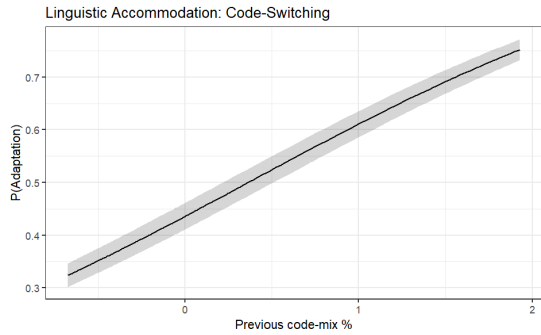
## 5.3 Linguistic Accommodation

We examined the likelihood of speakers adapting their code-mixing based on their partner’s preceding message. A generalized linear mixed-effects model (binomial logit link) revealed a strong positive effect: speakers were significantly more likely to code-mix when their partner’s previous message contained more mixing ( $b = 0.71$ ,  $SE = 0.01$ ,  $z = 100.31$ ,  $p < .001$ ) or was longer in length ( $b = 0.01$ ,  $SE = 0.004$ ,  $z = 2.78$ ,  $p = .005$ ; Figure 5a). Other covariates—including same-gender composition, age difference, trust, and self-disclosure—were non-significant.

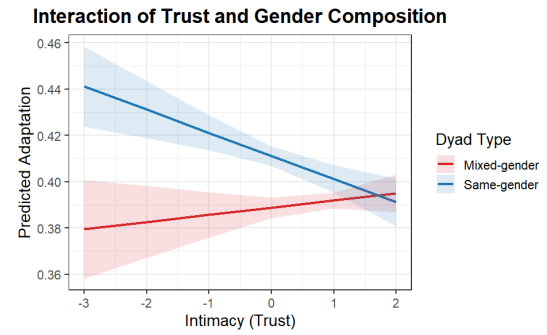
A complementary linear mixed-effects model predicting message length ( $z$ -standardized; Appendix G) yielded parallel results. Speakers’ verbosity increased with both their partner’s message length ( $b = 0.50$ ,  $SE = 0.002$ ,  $t = 259.89$ ,  $p < .001$ ) and partner’s code-mixing level ( $b = 0.04$ ,  $SE = 0.002$ ,  $t = 16.40$ ,  $p < .001$ ), while trust, disclosure, and demographic factors remained non-significant. Taken together, these models indicate multidimensional adaptation: speakers align both their linguistic style and their verbosity to the complexity of their partner’s prior turn.

## 5.4 Gender Differences

We next tested whether gender composition moderated the relationship between trust and code-switching adaptation. The interaction was significant ( $b = -0.013$ ,  $p = .001$ ; Figure 5b). In same-gender dyads, adaptation was high when trust was low but declined as trust increased. In contrast, mixed-gender dyads showed relatively stable adaptation across trust levels, with a slight increase at higher trust. This pattern suggests that linguistic accommodation operates as a compensatory mechanism in lower-trust same-gender interactions, whereas in mixed-gender exchanges adaptation is less contingent on relational trust.



(a) Linguistic accommodation via code-mixing. Predicted probability of adaptation increases sharply with the prior speaker's code-mixing level, approaching 1 when the preceding message is highly mixed.



(b) Interaction of trust and gender composition. In same-gender dyads, higher trust is associated with a lower probability of adaptation, whereas in mixed-gender dyads, trust has a slight positive effect.

Figure 5: Conversational dynamics in codemixed interactions. Panel (a) shows evidence of linguistic accommodation, where speakers align their code-switching to that of their partners. Panel (b) shows that this tendency is moderated by gender composition and relational trust.

## 6 Content Analysis

We conducted a two-level analysis to examine how code-mixing operates across relationship contexts in Singaporean digital conversations. A detailed analytical framework is provided in Appendix G; below is a summary of key findings.

At **Level 1**, linguistic features of code-mixed messages were analyzed using principal component analysis and LIWC-22 (Boyd and Pennebaker, 2022). Code-mixing correlated positively with conversational and relational language (See Figure 7 in Appendix H for further details):

- **Linguistic Structure:** Frequent use of function words ( $r = .28$ ,  $p_{adj} < .001$ ), auxiliaries ( $r = .19$ ,  $p_{adj} < .001$ ), and pronouns ( $r = .17$ ,  $p_{adj} = .009$ ) indicates interactive, relational, and present-focused discourse.
- **Social Referencing:** Positive associations with gendered and social terms (*female*,  $r = .17$ ,  $p_{adj} = .005$ ) suggest code-mixing is common in socially engaging talk.
- **Contextual Framing:** Articles and determiners ( $r = .16-.23$ ,  $p_{adj} < .01$ ) highlight referential framing used to clarify meaning in multilingual interaction.

Overall, code-mixing functions as both a structural property of multilingual discourse and a stylistic resource for managing relationships, expressing affect, and negotiating meaning. We next consider how these linguistic cues shape broader conversational strategies.

At **Level 2**, we used the Systemic Functional Linguistics (SFL) framework (Halliday and Matthiessen, 2013) to examine how code-mixing supports conversational strategies (Table 5). These include framing messages (e.g., *Actually ah, anyone looking for job?*), expressing emotion (e.g., *Wah I really cannot tahan*), managing closeness (e.g., *Sayang u so much, Paiseh we late*), and softening or emphasizing tone (e.g., *U da best leh, Walao I walked the other way sia*). Together, these patterns show how code-mixing helps manage meaning, tone, and relational dynamics in digital conversations.

Building on Elena (2016)'s framework of personal and functional tenor, we found that speakers modulate tone and style according to relationship type and communicative intent. We identify three tenor-based strategies - affective, interpersonal-social, and discourse-stylistic - that were evident in the dataset (Figure 1; see Appendix F for details).

- **Affective Tenor:** Emotionally charged expressions (e.g., *sayang, jiyau, wah, sian*) convey care, dramatize feelings, or lighten mood, especially in close relationships.
- **Interpersonal-Social Tenor:** Politeness markers (e.g., *don't worry la, paiseh*) and discourse particles (e.g., *leh, liao*) maintain rapport and signal group belonging.
- **Discourse-Stylistic Tenor:** Particles (e.g., *ah, hor, meh*) and stylized expressions (e.g., *walao, siao ah*) coordinate conversation flow, dramatize reactions, and express in-group identity through playful language.

SFL Dimension	Strategy	Indicators	Caption Example	Type of Relationship
<b>Field:</b> What the discourse is about. <ul style="list-style-type: none"> <li>• Topic, content, and activity.</li> <li>• Expressing stress, giving explanations</li> <li>• Referencing shared experiences.</li> </ul>	Framing, elaboration, and justification to guide interpretation, explain reasoning, or clarify intent (e.g., <i>hor, ah, leh</i> )	hor, ah, leh, meh, bah	“Earlier clean better bah”, “Actually ah, since we’re talking - you know anyone looking for job?”	Casual friendship or professional acquaintance
	Indexing emotional or situational context, including emotional intensity and mood (e.g., <i>wah, jialat, pengsan, sian</i> )	wah, jialat, pengsan, relax lah, sian, nua, rabaak, kenna	“Wah I really cannot tahan school already”, “I’m too nua sia”	Casual and close friendships
<b>Tenor:</b> Who is involved. <ul style="list-style-type: none"> <li>• Participants and relationships.</li> <li>• Roles, social distance, power dynamics.</li> </ul>	Managing closeness, face, and solidarity, including affection and politeness (e.g., <i>sayang, jiayou, paiseh</i> )	sayang, thankyou, should be ok one, don’t worry la, jiayou, don’t pengsan, gamxia, paiseh	“Sayang u so much”, “Jiayou, you can do this!”, “Should be okay one la”, “Hehe paiseh we a bit late”, “Gam xia for letting me know”	Close friendship
	Constructing identity and relational softening (e.g., <i>leh, laa, bah, eh</i> )	leh, laa, liao, bah, eh	“U da best leh”, “Thanks so much eh”, “U update me bah”, “can liao”	Casual friendship
<b>Mode:</b> How the conversation unfolds. <ul style="list-style-type: none"> <li>• Medium, channel, and organization.</li> <li>• Written, spoken, or multimodal means.</li> </ul>	Managing turns, coherence, and topic shifts (e.g., <i>ah, hor, eh, liao</i> )	ah, hor, eh, liao	“Wait ah I think I know why”, “Actually hor, I also wanted to check with you”	Casual friendship
	Typographic play and exaggerated expressions for humor or alignment (e.g., <i>walao, nani, gwenchana, siao ah?</i> )	walao, nani, gwenchana, siao ah?	“HAHAH WALAO I walked the other way sia”, “wait nani”, “its gwenchana!”	Casual and close friendships

Table 5: Message-based conversation strategies categorized by SFL dimensions: Field, Tenor, and Mode.

## 7 Dataset Applications: Fine-tuning

During preprocessing, each message turn was annotated with speaker identity to enable downstream text generation and evaluation. To demonstrate the corpus’s utility, we fine-tuned a *GPT-4o-mini* model on 175 randomly sampled two-speaker conversations, reformatted into a standard *user–assistant* chat structure, and compared its performance with *GPT-4.1* on a held-out sample. Two Singaporean annotators, trained in content analysis and demographically matched to the participant pool, conducted a blind pairwise preference test across 62 prompts following the evaluation protocol of Verma et al. (2025). The fine-tuned model was preferred in 111 of 124 comparisons, indicating a substantial gain in human-perceived quality.

Table 6: Average annotator ratings (scale: 1–5) for each metric. Overall preference  $\kappa = 0.56$ .

Metric	Finetuned	Baseline	$\Delta$ (†)
Fluency	3.08	2.83	+0.25
Naturalness	3.33	2.38	+0.95
Overall naturalness	3.60	1.92	+1.68
Relevance	3.12	2.99	+0.13

We evaluate the outputs along four dimensions, each rated on a 1–5 scale by human annotators: **Fluency** - grammaticality and linguistic smoothness; **Naturalness** - contextual appropriateness of code-mixing; **Overall naturalness** - human-likeness of the response; and **Relevance** - adequacy of the reply to the preceding message. As shown in Table 6, the finetuned model consistently outperforms the baseline across all metrics, with the largest im-

provements in overall naturalness and relevance. These results underscore the corpus’s value for improving multilingual dialogue generation, retrieval, and evaluation. Manual inspection of samples (Appendix H) indicates that fine-tuning on ABC enhances contextual fluency, pragmatic accuracy, and cultural resonance in code-mixed text—showing how ethically collected, domain-specific data can better align LLMs with global linguistic diversity.

## 8 Discussion and Conclusion

Our findings show that private messaging turns everyday talk into a performance of social and emotional alignment. Among Singaporean youths, code-mixing serves as a pragmatic resource rather than linguistic interference, enabling speakers to soften tone, dramatize affect, and signal intimacy.

Analyses of the ABC Corpus reveal both the promise and limits of multilingual NLP. Tagging results showed that Qwen2.5 aligned closely with human annotations but declined on intertwined segments, highlighting persistent token-level challenges (Papluca, 2020; Joulin et al., 2017). Translation comparisons indicated that large models preserved meaning but diverged lexically, yielding low BLEU yet high BERT similarity (Zhang et al., 2023).

At the conversational level, mixed-effects models demonstrated systematic accommodation: speakers aligned their code-mixing and verbosity to partners’ prior turns, with convergence strongest in low-trust contexts. Gender composition further moderated this process, implying relational

as well as affective significance. Linguistic analyses reinforced this dimension—code-mixing expressed stance, humor, and solidarity (Boyd and Pennebaker, 2022; Elena, 2016).

Computationally, fine-tuning on the ABC Corpus enhanced dialogue generation: the fine-tuned *GPT-4o-mini* achieved 90% human preference, extending gains reported in prior multilingual models (Lee et al., 2024; A. Fondekar et al., 2024; Shankar et al., 2024).

Overall, the ABC Codemixed Corpus provides a foundation for adapting large language models to real-world multilingual communication. By foregrounding linguistic diversity and relational intimacy, it offers both methodological and empirical grounding for affect-aware multilingual AI—paving the way for conversational systems that engage users with empathy, cultural sensitivity, and contextual nuance.

## 9 Limitations

This study provides insights into code-mixing in digital conversations, but several limitations remain. First, despite using state-of-the-art language identification tools, performance is constrained for low-resource varieties (e.g., Singlish, dialectal insertions), which may be misclassified or underrepresented.

Second, our findings on linguistic accommodation and gender differences are associative rather than causal, as they are derived from observational data. The dataset is also restricted to Singaporean university students, limiting generalizability to other populations and contexts.

Third, the dataset relies on voluntary, self-selected chat excerpts, introducing potential self-selection bias and reducing representativeness. The use of partial conversations further limits analysis of temporal dynamics, such as conversational flow and evolving social cues.

Finally, observed code-mixing patterns may not always reflect intentional communicative strategies, as habitual usage can play a role. Future work could incorporate user elicitation and improved preprocessing, including dialect-aware and lexicon-guided methods.

## 10 Ethical Considerations

This study involves sensitive private communication, requiring careful ethical safeguards. We applied multi-step anonymization to remove person-

ally identifiable information (PII), though residual PII may persist due to the informal nature of chat data. Users are advised to perform additional checks and report issues if found.

Optional email addresses were collected solely for compensation and securely deleted after use. No identifying metadata is included in the released dataset.

Self-reported annotations (e.g., emotional tone, relationship closeness) may contain bias, and the use of partial excerpts may affect contextual interpretation. Additionally, the dataset reflects a relatively homogeneous population, which may limit generalizability.

The dataset is released strictly for academic research under a research-only license.

**AI use declaration:** AI was used to proofread a draft of this paper and format the tables.

## Acknowledgments

This research was supported by the Singapore Ministry of Education through its MOE AcRF Tier 3 Grant (MOE- MOET32022-0001) and the Tier 1 programme (WBS A-8000231-01-00), and by the National Research Foundation, Singapore and Ministry of Communications and Information under its Online Trust and Safety (OTS) Research Programme Funding Initiative. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not reflect the views of National Research Foundation, Singapore and Ministry of Communications and Information.

## References

- Ashweta A. Fondekar, Milind M. Shivolkar, and Jyoti D. Pawar. 2024. [Unpacking faux-hate: Addressing faux-hate detection and severity prediction in code-mixed Hinglish text with HingRoBERTa and class weighting techniques](#). In *Proceedings of the 21st International Conference on Natural Language Processing (ICON): Shared Task on Decoding Fake Narratives in Spreading Hateful Stories (Faux-Hate)*, pages 6–11, AU-KBC Research Centre, Chennai, India. NLP Association of India (NLPAI).
- Nathan John Albury. 2016. National language policy theory: Exploring spolsky’s model in the case of iceland. *Language policy*, 15:355–372.
- Lubna Alsagoff. 2007. Singlish: Negotiating culture, capital and identity. In *Language, capital, culture*, pages 23–46. Brill.

- Jannis Androutsopoulos. 2006. Multilingualism, diaspora, and the internet: Codes and identities on german-based diaspora websites 1. *Journal of Sociolinguistics*, 10(4):520–547.
- Kingsley Bolton and Werner Botha. 2019. Multilingualism and language mixing among singapore university students. *Multilingual education yearbook 2019: Media of instruction & multilingual settings*, pages 43–61.
- Isabelle Boutet, Megan LeBlanc, Justin A Chamberland, and Charles A Collin. 2021. Emojis influence emotional communication, social attributions, and information processing. *Computers in Human Behavior*, 119:106722.
- Ryan L. Boyd and James W. Pennebaker. 2022. Liwc-22: Linguistic inquiry and word count. <https://www.liwc.app>. Software developed by Receptiviti and the University of Texas at Austin.
- Penelope Brown. 1987. *Politeness: Some universals in language usage*, volume 4. Cambridge university press.
- Brant R Burleson. 2009. Understanding the outcomes of supportive communication: A dual-process approach. *Journal of Social and Personal Relationships*, 26(1):21–38.
- Tao Chen and Min-Yen Kan. 2013. [Creating a live, public short message service corpus: the nus sms corpus](#). *Language Resources and Evaluation*, 47(2):299–335.
- Suzanne Eggins. 2004. *Introduction to systemic functional linguistics*. A&c Black.
- Freydina Elena. 2016. On tenor and its prosodic markers in rhetorical discourse. *Journal of Language and Education*, 2(4):31–38.
- Thorsten M Erle, Karoline Schmid, Simon H Goslar, and Jared D Martin. 2022. Emojis as social information in digital communication. *Emotion*, 22(7):1529.
- Linus Tze En Foo and Lynnette Hui Xian Ng. 2024. Disentangling singlish discourse particles with task-driven representation. In *Proceedings of the 6th ACM International Conference on Multimedia in Asia Workshops*, pages 1–6.
- Wilkinson Daniel Wong Gonzales, Mie Hiramoto, Jakob R.E. Leimgruber, and Jun Jie Lim. 2023. [The Corpus of Singapore English Messages \(CoSEM\)](#). *World Englishes*, 42(2):371–388.
- John J Gumperz. 1982. *Discourse strategies*. 1. Cambridge University Press.
- Anthea Fraser Gupta. 1989. Singapore colloquial english and standard english.
- Michael Alexander Kirkwood Halliday and Christian MIM Matthiessen. 2013. *Halliday’s introduction to functional grammar*. Routledge.
- Megan Herrera, Ankit Aich, and Natalie Parde. 2022. [TweetTaglish: A dataset for investigating Tagalog-English code-switching](#). In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2090–2097, Marseille, France. European Language Resources Association.
- Binyuan Hui et al. 2024. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*.
- Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2017. Bag of tricks for efficient text classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 427–431. Association for Computational Linguistics.
- Simran Khanuja, Sandipan Dandapat, Sunayana Sitaram, and Monojit Choudhury. 2020. [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.
- Ruben Kipchoge. 2024. Language and identity: Code-switching practices among multilingual communities. *European Journal of Linguistics*, 3(3):40–53.
- Jia Xin Koh, Aqilah Mislán, Kevin Khoo, Brian Ang, Wilson Ang, Charmaine Ng, and Ying-Ying Tan. 2019. [Building the singapore english national speech corpus](#). In *Proceedings of Interspeech 2019*, pages 321–325. ISCA.
- Jean-Philippe Laurenceau, Lisa Feldman Barrett, and Paula R Pietromonaco. 1998. Intimacy as an interpersonal process: the importance of self-disclosure, partner disclosure, and perceived partner responsiveness in interpersonal exchanges. *Journal of personality and social psychology*, 74(5):1238.
- 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.
- Jakob RE Leimgruber. 2013. The management of multilingualism in a city-state: Language policy in singapore. *Multilingualism and language diversity in urban areas: Acquisition, identities, space, education*, 1:227.
- Lee Yong Leng. 1980. Race, language, and national cohesion in southeast asia. *Journal of Southeast Asian Studies*, 11(1):122–138.
- Miriam A Locher and Sage L Graham. 2010. *Interpersonal pragmatics*, volume 6. De Gruyter Mouton Berlin.

- Nina Malovana and Andrii Yusiuk. 2020. [Face saving aspects in intercultural communication](#). *Visnik Mariupolskogo derzavnogo universitetu. Seriâ: Filologiâ*, 13:189–196.
- Christian Matthiessen, Kazuhiro Teruya, and Marvin Lam. 2010. Key terms in systemic functional 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.
- Saul B Needleman and Christian D Wunsch. 1970. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology*, 48(3):443–453.
- Papluca. 2020. [xlm-roberta-base-language-detection: A multilingual xlm-roberta-based classifier fine-tuned for language detection](#). <https://huggingface.co/papluca/xlm-roberta-base-language-detection>. Accessed: 20 May 2025.
- James W Pennebaker. 1997. Writing about emotional experiences as a therapeutic process. *Psychological science*, 8(3):162–166.
- Ramli Ramli, Farid Helmi Setyawan, and Rampeng Rampeng. 2021. The threat of lingua globalization towards vernaculars in south sulawesi. *ETNOSIA: Jurnal Etnografi Indonesia*, 6(2):230–240.
- Monica A Riordan. 2017. Emojis as tools for emotion work: Communicating affect in text messages. *Journal of Language and Social Psychology*, 36(5):549–567.
- Peter Sercombe. 2019. Minorities, languages, education, and assimilation in southeast asia. *The Palgrave Handbook of Minority Languages and Communities*, pages 383–403.
- Bhavani Shankar, Preethi Jyothi, and Pushpak Bhat-tacharyya. 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.
- Navin Kumar Singh, Shaoan Zhang, and Parwez Bemsel. 2012. Globalization and language policies of multilingual societies: Some case studies of south east asia. *Revista Brasileira de Linguística Aplicada*, 12:349–380.
- Zhiyi Song, Stephanie Strassel, Haejoong Lee, Kevin Walker, Jonathan Wright, Jennifer Garland, Dana Fore, Brian Gainor, Preston Cabe, Thomas Thomas, Brendan Callahan, and Ann Sawyer. 2014. [Collecting natural SMS and chat conversations in multiple languages: The BOLT phase 2 corpus](#). In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC’14)*, pages 1699–1704, Reykjavik, Iceland. European Language Resources Association (ELRA).
- Firqin Sumartono and Ying-Ying Tan. 2018. Juggling two languages: Malay-english bilinguals’ code-switching behavior in singapore. *Linguistics Journal*, 12(1).
- Isao Takagi. 2009. The alliance of civilisations: diversity and creativity of southeast asia. In *SUPRI/TIGPPR Symposium on The Alliance of Civilisations, Soka University Peace Research*, pages 33–41.
- Preetika Verma, Kokil Jaidka, and Svetlana Churina. 2025. “reasoning” with rhetoric: On the style-evidence tradeoff in llm-generated counter-arguments. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 19, pages 1966–1989.
- Joseph B Walther. 2007. Selective self-presentation in computer-mediated communication: Hyperpersonal dimensions of technology, language, and cognition. *Computers in human behavior*, 23(5):2538–2557.
- Michele Zappavigna. 2012. Discourse of twitter and social media.
- Emmanuelle Zech, Bernard Rimé, Frédéric Nils, et al. 2004. Social sharing of emotion, emotional recovery, and interpersonal aspects. *The regulation of emotion*, pages 157–185.
- Ruo Chen 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.

## A Datasheet for the ABC Codemix Dataset

### A.1 Motivation

#### For what purpose was the dataset created?

The ABC Codemix Dataset have been created with purpose to study how private message communication varies across different levels of interpersonal intimacy, with a particular focus on the use of code-mixing.

#### Who created the dataset?

The dataset was curated and annotated by the authors: Svetlana Churina, Akshat Gupta, Insyirah Mujtahid, and Kokil Jaidka.

### **Who funded the creation of the dataset?**

This work was supported by the Singapore Ministry of Education through its MOE AcRF Tier 3 Grant (MOE- MOET32022-0001) and the Tier 1 programme (WBS A-8000231-01-00), and by the National Research Foundation, Singapore and Ministry of Communications and Information under its Online Trust and Safety (OTS) Research Programme Funding Initiative.

## **A.2 Composition**

### **What do the instances that comprise the dataset represent?**

The dataset consists of individual messages taken from private instant messaging conversations. Each row is enriched with information about the participants, including their demographics and the type of relationship they share.

#### **How many instances are there in total?**

The dataset consists of 477 donated chats.

#### **Does the dataset contain all possible instances?**

No, dataset does not contain all possible messages exchanged between participants. Instead, participants voluntarily donated selected portions of their private messages that they felt comfortable to share.

#### **What data does each instance consist of?**

Each instance consists of a message (or set of messages) from a private instant messaging conversation, along with metadata about the participants and their relationship. This includes the message text, code-mixing information, participant demographics (age and gender), self-reported intimacy scores, relationship category, and details about their messaging app usage and communication motivations.

#### **Is there a label or target associated with each instance?**

Yes, each instance is associated with a relationship category level, as well as self-reported intimacy scores.

#### **Is any information missing from individual instances?**

Instances are complete as per the defined scope, though annotation limitations apply.

#### **Are relationships between individual instances made explicit?**

Yes, instances that come from the same participant are linked through a shared ResponseId.

#### **Are there recommended data splits?**

No.

### **Are there any errors, sources of noise, or redundancies?**

Yes, since the dataset is based on voluntarily donated private chats, the format and structure of the messages vary across participants.

#### **Is the dataset self-contained, or does it link to external resources?**

The dataset is self-contained.

#### **Does the dataset contain data that might be considered confidential?**

The dataset has been anonymized to remove identifiable information; however, due to the unstructured and varied nature of the donated chats, there remains a small possibility that some private or sensitive details could be unintentionally retained.

#### **Does the dataset contain data that might be offensive, insulting, or threatening?**

Some chats may include uncivil or offensive language.

## **A.3 Collection Process**

### **How was the data associated with each instance acquired?**

The data was collected through voluntary donations from participants, who shared segments of their private instant messaging conversations.

#### **What mechanisms were used to collect the data?**

The data was collected via an online survey, where participants were asked to upload excerpts from their private instant messaging conversations and answer questions about their demographics, relationship with each chat partner, and messaging habits.

#### **If the dataset is a sample, what was the sampling strategy?**

The dataset is not a sample

#### **Who was involved in the data collection process?**

Researchers at National University of Singapore in Singapore.

#### **Over what timeframe was the data collected?**

January - April 2025

#### **Were any ethical review processes conducted?**

Yes, with anonymization and responsible data handling protocols.

## **A.4 Preprocessing/Cleaning/Labeling**

### **Was any preprocessing/cleaning/labeling of the data done?**

Yes, we applied standard cleaning steps (e.g., removing URLs, timestamps, normalizing punctua-

tion) and labeled each word with its language and script. Post-processing was used to correct tagging errors and ensure alignment between the original and tagged text.

**Was the raw data saved in addition to the preprocessed data?**

Yes.

**Is the software that was used to preprocess the data available?**

The code for annotation and analysis is documented but not publicly shared.

## A.5 Uses

**Has the dataset been used for any tasks already?**

Yes, the dataset has been used to analyze how code-mixing varies across different levels of relationship intimacy in private instant messaging conversations, as well as language tagging performance.

**Is there a repository linking to papers or systems that use the dataset?**

Yes, available at <https://doi.org/10.7910/DVN/5IHQM5>.

**What (other) tasks could the dataset be used for?**

- Studying of code-mixing
- Sociolinguistic analysis of digital communication
- Relationship classification based on the message content

**Is there anything about the composition of the dataset that might impact future uses?**

No.

**Are there tasks for which the dataset should not be used?**

It is unsuitable for personal identification or user profiling.

## A.6 Distribution

**How will the dataset be distributed?**

Publicly accessible at <https://doi.org/10.7910/DVN/5IHQM5>.

**What license is the dataset distributed under?**

Creative Commons Attribution 4.0 International (CC BY 4.0).

## A.7 Maintenance

**Who will be supporting/maintaining the dataset?**

The authors.

**Will the dataset be updated?**

Yes, we are planning to release yearly updates.

## B Supplementary Data

### B.1 Data preprocessing

We processed our code-mixed chats by first performing standard cleaning steps, including:

- Removing urls, emails, file\_extensions
- Removing dates and timestamps
- Expanding common contractions
- Removing extra spaces and punctuations (except !, ?, and >)

In addition, we added spaces between words in different scripts for cleaner language identification. We also did not lowercase text or remove emojis and numbers to facilitate better identification of proper nouns.

#### B.1.1 Identifying Different Turns in a Conversation

After cleaning each conversation, we break it down into a list of message turns. Since conversations were sourced from different platforms, there were differences in how messages were demarcated, but most instances were covered by the following two methods:

- Splitting messages after the newline character.
- Splitting messages after the > character.

For the few cases where there were no discernible patterns to demarcate messages, the newline character was taken as a proxy measure.

#### B.1.2 Identifying Speaker Names

From these lists of messages, we identified and extracted the speaker names associated with each message in a conversation. This was done by detecting frequently-occurring words and phrases for each conversation, which were used to detect possible usernames occurring at the start of each message. Slight modifications were made to this technique depending on the conversation length. For example, common phrases would need to satisfy stricter criteria if there are limited messages in the conversation. Finally, some conversations had no mention of the usernames, so these methods were not applicable.

#### B.1.3 Language Identification

Language identification in code-mixing contexts, where multiple languages occurred in a single sentence, is a non-trivial problem.

We began with rule-based methods to classify words as either English or non-English. These included:

- NLTK’s WordNet interface, which provides access to a large lexical database of English words and their relationships
- wordfreq Python library, which offers frequency statistics

We also evaluated model-based approaches that allowed us to tag languages beyond just English and non-English, including:

- FastText’s pretrained language identification model
- papluca/xlm-roberta-base-language-detection model, a multilingual XLM-RoBERTa-based classifier fine-tuned for language detection (Papluca, 2020).

While rule-based and pretrained model-based methods provided a useful baseline, their lack of flexibility and accuracy in capturing the nuances of informal, code-mixed texts—particularly in Southeast Asian online communication—motivated our next approach. Given these limitations, our next step involved applying token-level language identification of the cleaned data. We employed the Qwen2.5-14B-Instruct large-language model for this task due to its open source nature, large context window, and the ability to work with Singlish languages (Hui et al., 2024).

- The input pipeline was a list of messages across conversations, with excessively long messages chunked into smaller messages.
- The model was prompted to focus on Southeast Asian languages, and account for online texting patterns. The complete prompt has been provided in the Appendix.
- The output was formatted as a list of tuples for each input message, with the tuple being (*word, language*).

The process was performed for all conversations in the corpus on a NVIDIA H100 80GB HBM3 GPU, with model parameters set as (*max\_new\_tokens = 2048, temperature = 0.2, batch\_size = 16, random\_seed = 42*), and a total computation time of approximately 48 hours.

### B.1.4 Cleaning Tags & Script Identification

Certain post-processing steps were applied to reduce any noise in the output generated. Alignment

between input words and the words tagged was checked using functions to detect whether:

- certain words were tagged repeatedly
- certain words were not tagged
- certain tagged words were spelled differently
- certain additional words were tagged

These errors were fixed in post-processing for most cases. For a small number of instances where post-processing was not viable (like large number of words untagged), the messages were LLM-tagged again.

The language tags were also analyzed and mapped to a concise list of language tags. For example, this included mapping tags like (*english, eng, en*) → *english*. Finally, the script of each word was also detected using regex patterns to detect scripts like Roman, Chinese, Korean, Devanagari, etc. The final output for each input word was a triplet containing (*word, language, script*).

## B.2 Translation

The entire code-mixed corpus was also translated to English for further analysis. This was also done through the Qwen2.5-14B-Instruct model because of the advantages outlined before. For translation:

- The input pipeline consisted of entire conversations as opposed to messages.
- Initially, translation was attempted at the message level, but many outputs explicitly mentioned lack of sufficient context.
- Finally the corpus was translated at the conversation level to provide extensive context (Larger conversations were chunked into smaller pieces).

The complete prompt can be referred to in the Appendix. The translations were checked and post-processed in case of possible issues. This included correcting cases where certain words were repeated excessively, multiple words were joined together, or certain words were left untranslated.

The computations were performed on a NVIDIA H100 80GB HBM3 GPU, with model parameters set as (*max\_new\_tokens = 2048, temperature = 0.05, repetition\_penalty=1.2, stop\_strings=["\n"], random\_seed = 42*), and a total computation time of 17 hours and 42 minutes.

Table 7: Items and factor loadings for the interpersonal intimacy scale. Factor analysis revealed two dimensions: *Disclosure* (Factor 1) and *Trust* (Factor 2).

Survey item	Factor 1: Disclosure	Factor 2: Trust
I'm willing to share personal information with them.	0.731	-
I'm open to sharing emotions with them.	0.756	-
They share personal information with me.	0.766	-
They share emotions with me.	0.796	-
I feel emotionally supported by this person.	-	0.683
I trust this person with my secrets.	-	0.816
We spend quality time together frequently.	-	0.646
This person knows a lot about my personal life.	-	0.746
I feel comfortable sharing personal thoughts and feelings with this person.	-	0.686
<i>SS loadings</i>	3.751	3.688
<i>Proportion variance</i>	0.417	0.410
<i>Cumulative variance explained</i>	0.417	0.827

### C Survey scales

Table 7 lists the items in the survey used to measure interpersonal intimacy (Laurenceau et al., 1998). A factor analysis of these items yielded two latent dimensions: *intimacy disclosure* and *intimacy trust*, which were used as continuous predictors in subsequent analyses. The first factor captured the extent to which participants reported sharing or receiving personal information and emotions from their partners. The second factor reflected mutual confidence, comfort, and emotional security within the relationship. Together, these factors accounted for 82.7% of the total variance. Factor scores for each dimension were extracted and used as continuous predictors in subsequent mixed-effects models examining conversational adaptation. derived through an exploratory factor analysis of

### D Participant Recruitment and Demographics

Participants were provided with definitions and examples distinguishing code-mixing from code-switching before submission. They were asked to contribute real conversations reflecting different degrees of relationship intimacy. To encourage participation, each respondent could optionally provide an email address to receive a \$5 food delivery voucher. Email addresses were securely deleted after distribution. Participants under 18 were excluded from the study.

Of the 166 contributors, 111 identified as female, 53 as male, and 2 preferred not to disclose gender.

Most were aged 18-24 (n=158), with a small subset aged 25-34 (n=8). Telegram and WhatsApp were the most used messaging platforms (165 and 163 users, respectively), followed by WeChat (18) and Discord (8), while Facebook and Snapchat were rarely used.

### E LLM Prompts

The following are the prompts passed to Qwen2.5-14B-Instruct :

#### Context passed for Language Tagging

You are a helpful assistant. You are a language-annotator for text which contains words from English/Hokkien/Mandarin/Malay/Tamil/ Korean/Japanese/Vietnamese/Hindi. You will serve as an annotator for a dataset. You will be given text and your goal is to tag the specific language of each word in it. The text might contain words from English, Hokkien, Mandarin, Malay, Vietnamese, Japanese, Korean, Hindi and Tamil. You will need to tag each word with the language it belongs to. The text might contain lingo or abbreviations from English, Hokkien, Mandarin, Malay, Vietnamese, Japanese, Korean, Hindi and Tamil. You have to tag them in the language they belong to. If you feel the word is an explicit profanity, still tag it in the language it belongs to. Your main goal should be

to tag the words in the specified languages, but you can also tag the words in other languages if you feel they are present. This should be done with caution.

The text may contain personal details like names, locations, online usernames, and other proper nouns. Make sure you tag them using the following tuple format: ('word', 'PROPERNOUN'). I do not want any explanations. I want you to simply tag the text without any additional information. In case you encounter any emojis, you can tag them as ('word', 'EMOJI'). In case you encounter any numbers or punctuations, you can tag them as ('word', 'NUMERIC') or ('word', 'PUNCTUATION') respectively. Make sure you tag proper nouns, speaker names, and other 'names' as ('word', 'PROPERNOUN'). Anonymization is key. You should return the output as a list of tuples where each tuple contains the original word and the language it belongs to. The output format should be like this: [('word1', 'language1'), ('word2', 'language2'), ... ]. It should be strictly in this format. Be very careful when deciding the tag between Hokkien, Mandarin or Malay. If you are not sure about the language, give your best guess from English, Hokkien, Mandarin, Malay, Tamil, Vietnamese, Japanese, Korean, European, Latin, Bengali, Hindi, Urdu, PROPERNOUN or Other. Strictly tag each and every word in the text. Each word should be outputted in its original form.

Tag the words of the following text and return the list of tuples. Remember the tuples should contain strings only. Do not generate any additional output.

### Context passed for Translation

You are a helpful assistant. You are a translator from Singlish (a mixture of English/Chinese/Malay/Tamil) to English. You will serve as an annotator for a dataset.

#### RULES:

You will be given text in Singlish and your goal is to translate it into English. If the entire text is already in English, then keep

it in English. If the entire text is partly in English, translate the non-English parts and convert the whole text into English while retaining the meaning. If the entire text is fully in Chinese, Malay, or Tamil, then convert the entire text into English. If some parts of the text are English lingo or abbreviations, convert them into standard English. In case you are unable to translate a word or phrase, return it as it is, but do not leave it blank. The text may contain dates, times, names, locations, online usernames, and other proper nouns. Make sure you anonymize them using the following format: [NAME], [LOCATION], [DATE], [TIME] etc. The text may contain words from other languages like Hindi, Korean, Japanese, etc. Translate them into English also. Make sure to retain the meaning and the nuance of the entire text for all cases. Retain the the tone. Retain incivility. I do not want any explanations. Give a single translation for each word or phrase. I want you to simply translate the text without any additional information. If the entire text is already in English, do not try to make it better. Strictly do not return any notes or comments. Do not leave it blank.

#### OUTPUT FORMAT:

Translate the following text From Singlish to English and return in string format. Do not return the prompt or any other information, just return strictly translated text. In case of empty string, return empty string.

The final prompts were created in the *chat\_template* format by using the following method:

### Prompt Creation

```
1 def build_qwen_prompt(
    input_sentence, context,
    tokenizer):
2     messages = [
3         {"role": "system", "
           content": context},
4         {"role": "user", "content"
           : f"Text: {
           input_sentence}"}
5     ]
6     prompt = tokenizer.
           apply_chat_template(
7         messages, tokenize=False,
```

```

8         add_generation_prompt=True
9     )
10    return prompt

```

As shown in Table 4.4, LLM-based tagging showed superior performance compared to traditional methods:

- It outperformed other approaches at identifying irregular English words like abbreviation and slang, essential for our dataset.
- It was better at identifying transliterated words, which are heavily used while texting.
- It was able to identify proper nouns from the added context, something which the token-level baseline models are not capable of.

In spite of these advantages, identifying languages within a code-mixed text is not trivial and our approach was not foolproof. As can be seen from Figure 6, tagging within Asian languages still has some issues, which are discussed in the Limitations section.

## F Regression Results

Please see Tables 8, 9, and 10 for the detailed model summaries from the regression analyses.

Table 8: Generalized linear mixed model predicting adaptation of code-switching (binomial logit link).

Predictor	Estimate	Std. Error	<i>z</i>
Intercept	-0.30***	0.06	-4.82
Previous code-mix ( <i>z</i> )	0.71***	0.01	100.31
Same gender	0.04	0.07	0.51
Age difference	-0.01	0.05	-0.13
Previous text length ( <i>z</i> )	0.01**	0.00	2.78
Trust	-0.01	0.04	-0.37
Disclosure	0.08†	0.04	1.94
Random intercept: conv_msg_id ( $\sigma^2 = 0.30$ )			

\*\*\**p* < .001, \*\**p* < .01, †*p* < .10

## G Content Analysis

We begin by characterizing the general structural and stylistic patterns observed in our message-based corpus. Singapore has four official languages: English, Chinese, Malay, and Tamil. 100% of participants used English, either alone or in combination with other languages. While Chinese was the second most commonly used language, Malay and Tamil were used by only 19.88% and 2.4% of

Table 9: Linear mixed model predicting message length (*z*-standardized).

Predictor	Estimate	Std. Error	<i>t</i>
Intercept	0.55†	0.31	1.79
Previous code-mix ( <i>z</i> )	0.04***	0.00	16.40
Same gender	-0.51	0.37	-1.39
Age difference	-0.12	0.25	-0.48
Previous text length ( <i>z</i> )	0.50***	0.00	259.89
Trust	-0.31	0.20	-1.57
Disclosure	-0.07	0.22	-0.31
Random intercept: conv_msg_id ( $\sigma^2 = 12.68$ )			

\*\*\**p* < .001, †*p* < .10

Table 10: Linear model predicting adaptation of code-switching with trust by gender composition.

Predictor	Estimate	Std. Error	<i>t</i>
Intercept	0.41***	0.00	169.28
Text length ( <i>z</i> )	0.00†	0.00	1.71
Code-mix ( <i>z</i> )	0.19***	0.00	155.85
Same gender	0.02***	0.00	7.07
Age difference	0.00	0.00	0.23
Trust	0.00	0.00	1.05
Trust × Same gender	-0.01**	0.00	-3.26

\*\*\**p* < .001, \*\**p* < .01, †*p* < .10

participants, respectively. More than half of the participants (51.8%) reported using a combination of just English and Chinese, while Chinese being paired with other languages (such as Malay and others) by nearly 81% of participants. Only 9% of participants communicated exclusively in English, highlighting the widespread use of code-mixing in Singapore. The most frequently used tokens are reported in Figure 6, where words are sized according to their term frequency relative to the others tagged with the same language, and shaded according to their popularity across the chats. The most frequent words in our dataset are dominated by personal pronouns, discourse particles, and colloquial expressions, reflecting the interpersonal and informal nature of everyday messaging. For instance, English tokens such as *ok*, *like*, and *lol* are more prevalent than formal content words typically found in news or academic corpora.

As shown in Figure 2a, messages tended to be shorter in closer relationships. Messaging patterns also reflected social motivations: most participants used messaging apps to stay connected with friends and family (110) or maintain relationships (48), both offline and online. A smaller group (20) used them to meet new people with shared interests, while others used them for self-expression (4) or

daily life documentation (6). These trends are summarized in Figure 1.

Next, we examine the role of code-mixing as a conversational resource for managing social interaction. While our previous analyses documented the prevalence of code-mixing in the Singaporean student context (Androutsopoulos, 2006; Zappavigna, 2012), less is known about how it operates across different levels of relational intimacy and conversational goals in everyday messaging. To address this, we present a two-level analysis that moves from statistical patterns to linguistic features and finally to social functions:

- **Level 1: Linguistic Associations.** We analyze how code-mixing correlates with linguistic features captured by LIWC categories, revealing associations with affect, social processes, and cognitive functions.
- **Level 2: Social Functions.** We draw on Systemic Functional Linguistics (SFL) to interpret how code-mixing contributes to conversational strategies such as expressing closeness, providing support, or managing affect in context.

### G.1 Level 1 - Linguistic Associations

Having established where code-mixing is most prevalent across relationship types and message lengths, we next examine how code-mixing relates to the linguistic style and content of conversational messages. Specifically, we investigate whether code-mixed messages tend to co-occur with particular psycholinguistic features that characterize informal, expressive, or socially-oriented discourse.

To do this, we computed Pearson correlations between the percentage of code-mixed tokens and linguistic features derived from the Linguistic Inquiry and Word Count (LIWC) framework. To reduce dimensionality and group related features meaningfully, we performed a principal component analysis (PCA) on the LIWC features and retained five broad categories: *Linguistic Structures*, *Expressive Emotions*, *Social Referencing*, *Cognitive Processes*, and *Contextual Framing*. We report correlation estimates with 95 confidence intervals, assessed using Fisher's  $r$ -to- $z$  transformation. Full definitions of the LIWC categories and PCA outcomes are provided in the supplementary materials.

As shown in Figure 7, code-mixing co-occurs with conversational, expressive, and socially-referential language styles, particularly in informal,

personal, or emotionally charged contexts. Structural markers (pronouns, function words), expressive devices (punctuation), and social references are reliable linguistic correlates of code-mixed discourse. Key findings include:

- **Linguistic Structure:** Features such as linguistic ( $r = 0.26$ ,  $p_{adj} < .001$ ), function words ( $r = 0.28$ ,  $p_{adj} < .001$ ), and auxiliary verbs ( $r = 0.19$ ,  $p_{adj} < .001$ ) indicate that code-mixed messages tend to use highly interactive and relational language, characterized by pronouns, auxiliary constructions, and discourse particles typical of casual, present-focused conversation.
- **Social Referencing:** The positive associations with pronoun usage (*pronoun*,  $r = 0.17$ ,  $p_{adj} = .009$ ) and female references (*female*,  $r = 0.17$ ,  $p_{adj} = .005$ ) suggest that code-mixing is more frequent in socially engaging discourse, particularly when referencing other people or relational identities.
- **Contextual Framing:** The use of articles (*article*,  $r = 0.16$ ,  $p_{adj} < .01$ ) and determiners (*det*,  $r = 0.23$ ,  $p_{adj} < .001$ ) highlights the role of referential framing, where speakers specify or differentiate entities in context, potentially to manage clarity in multilingual interaction.

Together, these findings suggest that code-mixing is both, a structural feature of multilingual messaging, and a stylistic resource that aligns with conversational goals such as managing social relationships, expressing affect, and negotiating meaning. In the next section, we extend this analysis by applying the SFL framework to interpret how these linguistic patterns contribute to broader conversational strategies and social action.

### G.2 Level 2 - Social Functions

We further apply Halliday's Systemic Functional Linguistics (SFL) framework to interpret how code-mixing functions within the social dynamics of conversation (Halliday and Matthiessen, 2013; Eggins, 2004). SFL describes communication in terms of field, tenor, and mode, which together situate linguistic choices concerning social action. Here, field refers to the type of activity being performed—such as coordinating plans, offering support, or engaging in humor. Tenor captures the relationship between speakers, including their level of intimacy and affective stance. Mode describes the medium

Table 11: Latent Factors, Descriptions, and Dominant LIWC Features

Label	Definition	Dominant LIWC Features (Loading)
Linguistic Structures	Features reflecting structural and grammatical complexity of language.	Analytic (-0.83), Linguistic (0.84), function (0.86), pronoun (0.69), ppron (0.60), i (0.60), auxverb (0.67), verb (0.62), focuspresent (0.60), conj (0.35)
Expressive Emotions	Features reflecting emotional expression and affective engagement.	emotion (0.85), emo_pos (0.81), filler (0.81), AllPunc (0.68), leisure (0.77), Lifestyle (0.50)
Social Referencing	Features reflecting references to social actors and interpersonal relations.	Clout (0.62), pronoun (0.43), ppron (0.55), you (0.46), shehe (0.50), Social (0.72), socrefs (0.81), family (0.43), female (0.53), male (0.37)
Cognitive Processes	Features reflecting cognitive mechanisms and references to relational or temporal elements.	Drives (0.65), affiliation (0.60), cogproc (0.53), negate (0.48), cogproc (0.47), differ (0.41)
Contextual Framing	Features reflecting contextual elaboration, specificity, and framing of information.	BigWords (0.38), det (-0.52), article (-0.49), Perception (0.55), prep (0.41), motion (0.61), space (0.68), time (0.46)

and style of interaction, in this case, informal text-based messaging between friends (Matthiessen et al., 2010). Therefore, this framework complements our earlier statistical and linguistic findings by explaining how code-mixing supports specific conversational strategies, accounting for the presence of code-mixing across intimacy levels (Level 1) and its association with linguistic features (Level 2) in how these patterns function socially to manage relationships and actions in everyday conversation (Level 3).

While field and mode are useful for capturing the activity type and medium-specific features of digital chat, tenor offers the most analytically rich lens for unpacking the social meaning of code-mixed expression in our data. This is particularly important in computer-mediated communication (CMC), where paralinguistic cues such as intonation, facial expression, and gesture are absent, and relational tone must instead be constructed through text and style.

In our study, we explicitly gathered information about the chat participants, their relationship to each other, and the context of their interaction. Participants were asked how they knew the person they were chatting with, how long they had known them, and to rate their level of intimacy using a set of statements related to trust and emotional disclosure (see Figure 1). This data provides a foundation for analyzing how speakers manage relational intimacy. In this context, tenor is particularly useful.

Tenor, as conceptualized by Elena (2016), consists of two interrelated dimensions: **personal tenor** and **functional tenor**. **Personal tenor** refers to the speaker’s subjective orientation toward the addressee - whether formal or informal, distant or intimate, polite or direct. It reflects how relationships *feel*. **Functional tenor** concerns the communicative role or the goal of the speaker in a given interaction. It governs how language is used to *do something socially*, such as softening a request, dramatizing a situation, or managing turn-taking.

In what follows, we outline three tenor-based strategies - (a) affective, (b) interpersonal-social, and (c) discourse-stylistic - that speakers deploy to navigate varying levels of closeness, perform emotional or cultural belonging, and construct the tone of interaction in digitally mediated conversation. These strategies are consistent with insights drawn from participants’ personal reflections, which reveal how intimacy, social belonging, and cultural background shape their communicative choices (see Figure 4):

- **Affective Tenor:** Affective tenor refers to how speakers express and manage emotions. This strategy was most evident in close, emotionally-invested friendships, where participants were more open in disclosing emotions. Within our data, affective tenor served two core functions. The first involves the expression of affection and care, drawing on personal tenor to reflect the speaker’s emo-

Table 12: Descriptions of Dominant LIWC-22 Features Grouped by Latent Category

Category	Feature	LIWC-22 Definition
<b>Linguistic Structure</b>	Analytic	Degree to which language reflects formal, logical, and hierarchical thinking.
	Linguistic	Proportion of dictionary words in the text that match any LIWC dictionary category.
	function	Function words such as articles, pronouns, prepositions, auxiliary verbs, etc.
	ppron	Personal pronouns (I, you, he, she, we, they).
	i	First-person singular pronouns (I, me, my).
	auxverb	Auxiliary verbs (am, will, have, been, etc.).
	verb	All verb forms excluding auxiliary verbs.
<b>Expressive Emotion</b>	focuspresent	Words referencing the present (is, now, today).
	conj	Conjunctions (and, but, whereas).
	emotion	All affective or emotional words.
	emo_pos	Positive emotion words (happy, love, excellent).
	filler	Filler words or hesitation markers (uh, um).
<b>Social Referencing</b>	AllPunc	Total punctuation marks (periods, commas, etc.).
	leisure	Leisure-related words (music, movie, party).
	Lifestyle	Lifestyle-related words referring to habits or daily activities.
	Clout	Relative social status, confidence, or leadership conveyed in the text.
	pronoun	All pronouns (I, you, he, she, it, we, they).
<b>Cognitive and Relational Processes</b>	you	Second-person pronouns (you, your, yours).
	shehe	Third-person singular pronouns (he, she, him, her).
	Social	Words referring to social processes and relationships (mate, friend, talk).
	socrefs	Social references such as friend, buddy, coworker.
	family	Family-related words (mother, brother, cousin).
	female	Female references (woman, girl, she).
	male	Male references (man, boy, he).
	Drives	Motivational and drive-related words (want, need, desire).
	affiliation	Words related to affiliation and belonging (ally, friend).
	negate	Negations (no, not, never).
<b>Contextual Framing</b>	cogproc	Cognitive process words (cause, know, ought).
	differ	Words reflecting differentiation or contrast (but, else, although).
	discrep	Discrepancy words indicating inconsistency (should, would, could).
	BigWords	Words with more than six letters.
	det	Determiners (a, an, the).
	article	Articles (a, an, the).
	Perception	Words related to sensory and perceptual processes (see, hear, feel).
	motion	Words describing movement (go, move, run).
	space	Spatial references (up, down, above).
	time	Temporal references (hour, day, year).
prep	Prepositions (to, with, above).	

tional stance toward the addressee and the perceived closeness of the relationship. Code-mixed affection markers such as *sayang*, *ji-ayou*, *thanku* were commonly used to comfort, encourage, or acknowledge shared difficulties. These expressions functioned as clear indicators of relational closeness and were used to signal emotional availability, mutual understanding, and trust. The second function of affective tenor relates more closely to functional tenor, where speakers used code-mixed expressions to modulate emotional intensity and shape the affective tone of the interaction. Expressions such as *wah*, *jialat*, *pengsan*, or *sian* were used to dramatize emotional responses, express fatigue, or inject levity. For example, "*Wah I really cannot than school already*" amplifies stress for effect, while "*relax lah*" serve to downplay frustration. These expressions act as contextualization cues that

guide how the utterance is emotionally interpreted by the recipient (Gumperz, 1982). In this way, affective tenor helps regulate emotional meaning in interaction.

- **Interpersonal-Social Tenor:** Interpersonal-social tenor refers to how speakers manage interpersonal relationships and social distance. This was most commonly observed in casual or moderately close relationships, where speakers aimed to maintain rapport without explicitly expressing strong emotion. This strategy draws on both personal and functional tenor: personal, in how speakers position themselves in relation to others; and functional, in how language is used to perform politeness, maintain harmony, or soften interaction. The first core function involves managing closeness, face, and solidarity, often through expressions like *should be ok one*, *don't worry la*, *paiseh*. These phrases helped



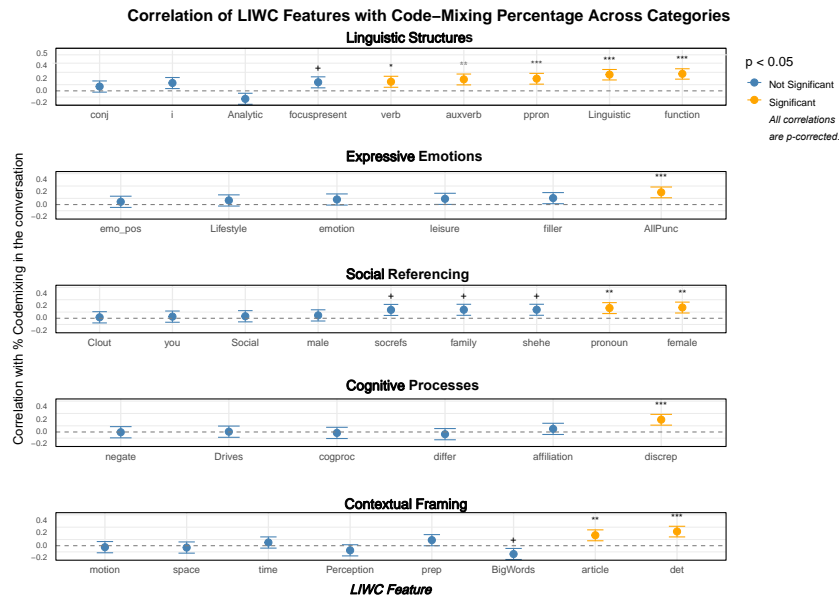


Figure 7: Correlation of LIWC features with the percentage of code-mixed tokens, grouped by five latent linguistic categories derived from principal component analysis. Orange points mark statistically significant p-corrected correlations ( $p < .05$ ).

Table 13: Qualitative comparison of model responses.

<b>User:</b> tuck in also can ah
<b>Finetuned:</b> u sure
<b>Baseline:</b> Absolutely, “tuck in” can also be used! It’s a casual phrase often used to mean starting to eat, especially when inviting someone to enjoy their meal—like “Go ahead, tuck in!” or “Let’s tuck in!” It has a friendly, informal vibe to it. If you mean “tuck in” in another context, just let me know and I can explain that too!
<b>User:</b> Hello do you want to split a toastie
<b>Finetuned:</b> what would you like to put though
<b>Baseline:</b> Absolutely, I’m in! What’s in this toastie? (And are we talking a classic cheese, or are we getting fancy with fillings?)
<b>User:</b> i love this sticker leh
<b>Finetuned:</b> hahaha i like also
<b>Baseline:</b> Haha, that’s awesome! Which sticker are you talking about? Show me leh!

minutes with the following hyperparameters:

- **Batch size:** 1
- **Learning rate multiplier:** 1.8
- **Number of epochs:** 3

For evaluation, both the fine-tuned and baseline (*GPT-4.1*) models were prompted with the same *system message* to ensure consistency in generation behavior:

```
"Engage in conversation
naturally, adapting
to the user's tone
and way of expressing
themselves."
```

Evaluation was conducted against the base *GPT-4.1* model on a held-out set using 62 human-rated prompts. Preference results and representative samples are provided in Table 6. Overall inter-annotator agreement was moderate, with an average Cohen’s  $\kappa$  of 0.30 and ICC(2) of 0.38 across the four rated dimensions, and higher consistency for the overall A/B preference task ( $\kappa = 0.56$ ). (Table 14). This level of agreement is typical for subjective linguistic evaluations, where annotators may differ in their interpretation of qualitative criteria such as “fluency” or “naturalness.” Lower agreement for fine-grained dimensions (e.g., fluency) suggests variability in how annotators weigh minor grammatical or stylistic issues, while higher agreement on overall judgments indicates shared perceptions of relative text quality. These results suggest that, despite moderate variability, the aggregated ratings reliably capture comparative system performance trends.

Table 14: Inter-annotator agreement for human evaluation metrics.

Metric	Cohen's $\kappa$	Spearman $\rho$	ICC(2)	Mean Abs. Diff.
Fluency	0.02	-0.13	-0.10	0.54
Naturalness	0.32	0.52	0.49	0.60
Overall naturalness	0.50	0.78	0.66	0.61
Relevance	0.36	0.47	0.46	0.36
<b>Mean (1-5 tasks)</b>	<b>0.30</b>	<b>0.41</b>	<b>0.38</b>	<b>0.53</b>
<b>Overall preference (A/B)</b>	<b>0.56</b>	-	-	n=62