Stylistic Evolution and LLM Neutrality in Singlish Language

Anonymous ACL submission

Abstract

Singlish is a creole rooted in Singapore's multilingual environment that continues to evolve alongside social and technological change. We examine diachronic stylistic change across a decade of informal digital messages and ask whether Large Language Models (LLMs) can generate temporally neutral outputs approximating the stable essence of the variety. Using lexical, pragmatic, psycholinguistic, and encoder-based features, we find that stylistic separability increases with temporal distance, driven primarily by structural features such as length and complexity. Evaluated against a null distribution baseline, most LLMs fail to achieve both authenticity and temporal neutrality simultaneously, revealing a structural trade-off: models generating realistic Singlish inherit its temporal biases, while temporally neutral models produce inauthentic outputs. These findings position temporal neutrality as a diagnostic metric for assessing sociolectal grounding in LLMs.

1 Introduction

Singlish is a creole-like variety of English widely spoken in Singapore. As a contact language, it incorporates grammatical and lexical elements from Chinese, Malay, Tamil, and their dialects, giving rise to a rich and dynamic linguistic system (Tan, 2005). Singlish continues to evolve as cultural norms, technologies, and generational usage patterns shift (Satibaldieva, 2024). These changes are especially apparent in Singlish due to Singapore's highly multilingual and rapidly developing sociolinguistic environment (Lee, 2016).

Despite its cultural significance, Singlish has received limited attention from a computational perspective. Fortunately, in recent years, Singlish has become more visible in the written form, enabling systematic study of its evolution in everyday contexts (Gonzales et al., 2023). However, as the sociolect shifts, it creates a "moving target" for computational models. An LLM trained on historical data may fail to capture the nuanced stylistic trajectory of modern usage. This disconnect motivates two main research questions: **RQ1:** How has Singlish evolved stylistically over time? and **RQ2:** Can LLMs generate temporally neutral Singlish that captures the underlying essence of the variety rather than surface style of a particular year? The goal is not temporal invariance, but whether models can abstract away from era-specific markers to approximate the stable structural core of the variety.

To study this, we analyze Singlish text messages from 2012 to 2021 and compare yearly subsets using interpretable linguistic features and encoder-derived representations. We then evaluate whether LLM-generated text aligns with the temporal profile of the corpus or collapses it into a flattened style. This lets us examine both diachronic change in real usage and the extent to which current models can recover a temporally neutral Singlish baseline.

Our Contributions are:

- **Stylistic similarity framework**: We proposed an interpretable metric of stylistic divergence using both hand-crafted linguistic features and encoder-derived representations.

- **Diachronic analysis of Singlish**: We shown systematic shifts that can be explained through linguistic and psycholinguistic features.

- **Evaluation of LLMs as Temporal Probes**: Using a null distribution baseline, we show that most current LLMs are not able to achieve both authentic and temporally neutral Singlish simultaneously. Models generating realistic Singlish inherit its temporal biases, while temporally neutral models produce generic, non-Singlish-like outputs.

2 Related Work

Research on Singlish has its foundations in sociolinguistics. Such work documented substrate influences from dialects of the neighboring regions and show how Singlish diverges structurally from standard English (Ningsih and Rahman, 2023; Tan, 2023). Pragmatic studies show how discourse particles such as *lah*, *lor* and *meh*, can function as stance markers, and encode solidarity and affective nuances between speakers (Wong, 2004; Mian-Lian and Platt, 1993; Foo and Ng, 2024). These works establish Singlish as a linguistically rich, contact-driven variety and is therefore an interesting language to study.

In computational sociolinguistics, prior work has examined diachronic language change in online settings, identifying lexical, semantic, and stylistic shifts over time across platforms such as X and Reddit (Eisenstein et al., 2014; Grieve et al., 2018). While these studies provide general frameworks for analyzing temporal variation, Singlish remains underexplored in this context. The availability of large-scale corpora such as the Corpus of Singapore English Messages (CoSEM) data (Gonzales et al., 2023) enables systematic analysis of Singlish evolution in digital communication.

Computational work on Singlish has primarily focused on handling its non-standard syntax through annotation, translation, and normalization pipelines (Liu et al., 2022; Chen et al., 2015), with downstream applications in sentiment analysis, opinion mining and topic detection on social media and e-commerce platforms (Bajpai et al., 2016; Hsieh et al., 2022). More recently, LLMs have been used to model Singlish semantics and style (Periti and Montanelli, 2024), support translation and hate-speech detection (Ng and Chan, 2024; Ng et al., 2024), and construct lexical resources (Chow et al., 2024). Although prompting and fine-tuning have been explored to align LLM outputs with community-specific language use (Ng et al., 2025; Wu et al., 2024), these approaches often struggle to reproduce authentic and temporally grounded variation (Yong et al., 2023). Our work extends this line of research by testing whether LLMs can generate Singlish that is not only realistic, but also temporally neutral across years.

3 Measuring Stylistic Similarity

To measure stylistic similarity, we represent texts as feature vectors and train a classifier to differentiate between texts from different time periods.

3.1 Feature Extraction

Handcrafted Features We extracted three categories of handcrafted features to capture stylistic variation associated with broader sociolinguistic trends in Singlish usage (Ziegeler, 2020; Ningsih and Rahman, 2023): **Lexico-structural attributes** (length, repetition, sentence complexity), **Functional pragmatic markers** (discourse particles, pragmatic nuance), and **Psycholinguistic dimensions** via the LIWC-22 lexicon (LIWC). The LIWC covers aspects of affect, cognition, and social orientation (Tausczik and Pennebaker, 2010), capturing underlying valence shifts and attitudinal shifts. A full feature list is provided in Appendix E.

Lexico-structural attributes and functional pragmatic markers are computed at the message level. In contrast, LIWC features were extracted from messages aggregated by year. This aggregation is necessary as LIWC scores are unreliable for extremely short texts (Zhao et al., 2016). Yearly aggregating provides stable psycholinguistic estimates while preserving the temporal resolution required for diachronic analysis.

Encoder-defined Linguistic Features Besides handcrafted features, we used encoded features. These features are dense vector representations of messages derived from a pre-trained all-MiniLM-L6-V2 sentence encoder[1]. The decision to utilize a non-Singlish-adapted model was a deliberate methodological choice. Fine-tuning the encoder on Singlish data would likely introduce specific linguistic biases, potentially compromising the model's ability to generalize across different temporal variations. These 384-dimensional embeddings capture rich semantic and contextual properties, leveraging large-scale pre-training to model the subtle stylistic and temporal shifts inherent in Singlish discourse.

3.2 Machine Learning Classifier Differentiation

To measure stylistic similarity, we used a Gradient Boosted Classifier from by the scitkit-learn library[2] on feature vectors representing two sets of messages (e.g., 2012 vs. 2013) using an 80:20

[1] https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2

[2] https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.GradientBoostingClassifier.html

random train:test split. We additionally evaluated Logistic Regression on the encoder-based features as a robustness check; the resulting temporal-gap trend was consistent with the Gradient Boosted Classifier results, as shown in Appendix A. We ultimately selected the Gradient Boosted Classifier because its ensemble nature provides better stability and robustness to noise inherent in naturally occurring sociolectal data (Gope et al., 2024). Furthermore, the choice of classifier was guided by preliminary experiments favoring simplicity and predictive power. To ensure robustness, training and testing were repeated three times per year-pair, reporting mean accuracy and standard deviation.

3.3 Similarity Metric

We use classifier accuracy as a proxy for stylistic separability. In a balanced binary classification task, a classifier's performance relative to the 50% chance baseline indicates signal strength. When two year-subsets are stylistically similar, the classifier should perform near chance (*acc*=50%), whereas higher accuracy indicates stronger discriminability. We thus transform accuracy into a similarity score to obtain a monotonic, interpretable measure of cross-year stylistic distance. This metric is relative rather than absolute, intended to compare year pairs under a consistent evaluation setup.

To obtain an interpretable score, we transformed the raw accuracy (*acc*), into a normalized similarity score, $S = 1-2|acc-0.5|$ where $0 \leq S \leq 1$. Here, $S = 1$ (*acc* = 0.5) indicates maximal similarity and $S = 0$ (*acc* = 0.0 or 1.0) indicating absolute discriminability.

By utilizing accuracy, we treat both temporal periods with equal weight, ensuring the metric reflects the global relationship between the two datasets. This allows for a consistent comparison of stylistic evolution across various time intervals and between human-authored and synthetic texts.

3.4 Temporal Neutrality and Realism Metric

To evaluate whether model outputs are biased towards particular time periods, we computed the standard deviation, σ, of similarity scores across years. Standard deviation captures temporal fluctuation in scores and provides a scale-independent summary comparable across models. Lower deviation indicates greater temporal neutrality, i.e., outputs that are less dependent on year-specific features and more representative of Singlish overall. We use this metric to assess whether prompting or fine-tuning reduces diachronic bias.

We evaluate realism by training classifiers to discriminate between real CoSEM messages and LLM generations where a higher similarity (S) denotes greater overall authenticity, while low variance (σ^2) across year-pairs signals temporal neutrality. This controls for the possibility that low variance reflects generic rather than genuine temporally neutral Singlish generation.

4 Methodology

Figure 1 illustrates our methodological pipeline.

4.1 Dataset

We used text messages sent from the years 2012 to 2021 from the **CoSEM** dataset (Gonzales et al., 2023), excluding 2022 due to substantially lower message volume. A latent dirichlet allocation (LDA) topic-invariance check confirmed that topical drift across years was negligible, hence isolating linguistic style as the primary driver of classifier performance (see Appendix B).

4.2 Diachronic analysis of Real-World Data

To examine diachronic changes in Singlish, we conducted pairwise classification experiments on year-pairs grouped by temporal gap. For instance, 2012 was paired with 2013, then with 2014, and so on. This design tests the hypothesis that linguistic differences accumulate over time, making temporally distant texts more distinguishable.

Grouping by temporal gap, rather than comparing individual year pairs, allows us to isolate temporal effects from individual year-specific variation. Across ten years of data, this yielded 45 unordered year-pairs, spanning one- to nine- year gaps. For each temporal gap, we trained a classifier to distinguish between messages from the paired years and computed both classification accuracy and the derived similarity score. Higher classification accuracy, or equivalently lower similarity, for larger gaps would indicate observable stylistic or structural shifts in Singlish over time, while flat or inconsistent trends would suggest relative linguistic stability.

4.3 Generating Singlish Texts with LLMs

We evaluate whether LLMs generate texts that are biased towards any particular time period and investigate the effects of fine-tuning on realism and temporal neutrality. We selected four pre-trained

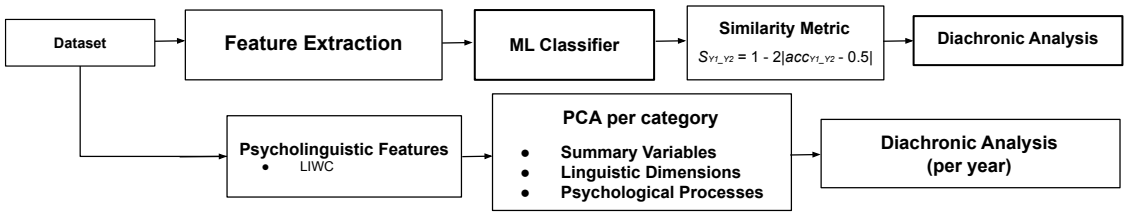


Figure 1: Pipeline for our diachronic analysis of CoSEM and LLM-generated Singlish texts.

LLMs from HuggingFace with their default configurations: Qwen/Qwen-7B-Chat[3], mistralai/Mistral-7B-v0.1[4], SeaLLMs/SeaLLM-7B-v2[5], deepseek-ai/DeepSeek-R1-Distill-Qwen-1.5B[6],hereafter referred to as Qwen, Mistral, SeaLLM and DeepSeek-R1 respectively.

We performed supervised fine-tuning on Qwen and Mistral using the LoRA architecture within the LLaMAFactory framework, adopting default hyperparameters (Zheng et al., 2024). The resulting models, Qwen-FT and Mistral-FT, were fine-tuned on the full CoSEM dataset (2012–2021). Training on the full pan-temporal distribution aims to suppress localized historical artifacts in favor of a hypothesized invariant Singlish core.

For text generation, each model was given a two-sentence Singlish dialogue seed, then prompted to continue the conversation, reflecting the conversational style of CoSEM messages. We evaluated four prompting strategies (Full prompts in Appendix A):

- **Zero-Shot (ZS):** Direct continuation in Singlish using default generation settings capturing the model's untreated Singlish style.
- **Chain-Of-Thought (CoT):** Explicit reasoning over cultural context and pragmatic intent prior to generation.
- **Diverse Decoding (DD):** Generation with increased stylistic and topical diversity.
- **Self-Consistency (SC):** Multiple independent continuations followed by self-selection of the most plausible output.

For each model and prompting scheme, we generated 100 messages, resulting in 2,400 generated texts.

[3] https://huggingface.co/Qwen/Qwen-7B-Chat
[4] https://huggingface.co/mistralai/Mistral-7B-v0.1
[5] https://huggingface.co/SeaLLMs/SeaLLM-7B-v2
[6] https://huggingface.co/deepseek-ai/DeepSeek-R1-Distill-Qwen-1.5B

4.4 Diachronic Alignment of LLM Outputs with CoSEM Data

We evaluated diachronic alignment by comparing LLM-generated texts against CoSEM messages from each year using the same stylistic similarity pipeline. Specifically, we computed similarity between each model's outputs and yearly CoSEM subsets: (LLM vs 2012), (LLM vs 2013), . . . , (LLM vs 2021). Higher similarity to specific years indicates temporal anchoring, while stable scores suggest a more temporally neutral representation.

4.5 Null Distribution for Temporal Neutrality

A key limitation of variance-based temporal neutrality assessment is the absence of a principled baseline: any nonzero variance could in principle be flagged as temporal bias, even if it falls within the expected range of random sampling noise. Therefore, we constructed a null distribution of stylistic variance representing expected variance under the assumption of no temporal structure. The 95th percentile of this null variance distribution serves as our formal threshold σ_{95}: a model whose cross-year variance falls below σ_{95} is statistically indistinguishable from temporally neutral texts.

Concretely, we repeatedly sampled two groups of messages uniformly at random from the full CoSEM pool without respect to year labels, and computed the stylistic similarity between them using the classifier pipeline in Section 3.2. Since both groups are drawn from the same undifferentiated pool, any observed variance in similarity scores across iterations reflects sampling noise alone rather than a genuine temporal signal. This procedure was repeated in $n_{outer} = 500$ outer iterations, each comprising $n_{inner} = 10$ pairwise comparisons, resulting in a distribution of 500 variance estimates under the null hypothesis of temporal neutrality.

5 Results

5.1 Diachronic Differentiability as a Function of Temporal Gap

Stylistic separability increases with temporal distance for both encoder-derived and handcrafted features. For encoder-derived representations, similarity decreases from 0.843 at T=1 to 0.538 at T=9. Handcrafted features show the same monotonic pattern, with similarity decreasing from 0.943 at T=1 to 0.754 at T=9 (see Figure 2). The Spearman correlation between the two trend lines is 1.00 ($p < 0.05$), indicating that the encoder has captured many of the same interpretable distributional properties explicitly modeled in the handcrafted feature set, while maintaining a different scaling.

5.2 Qualitative Linguistic Shifts

To ground these performance metrics in observable change, we highlight representative examples that mirror this increasing separability. Below are examples where individuals described their stress:

- **2013**: No I will stress until die one
- **2014**: Walao liddat I stress sia
- **2020**: I really nvr feel so stress about anything before

The 2013 and 2014 text messages appear to be more similar in style, as they exhibit more Basilectal features characterized by the use of local loanwords and inaccurate usage of English grammar. Meanwhile, that of 2020 appears to be more Acrolectal or Mesolectal in nature. This mirrors the results of similarity scores of 0.82 and 0.64 for 2013 vs 2014 and for 2013 vs 2020, respectively. Although these examples are illustrative rather than definitive, they help ground the temporal-gap results in observable linguistic change.

5.3 Feature Importance

We calculated the absolute SHapley Additive exPlanations (SHAP) feature importance scores (Lundberg and Lee, 2017) of the handcrafted features to identify contributors to diachronic discriminability. The mean absolute SHAP values and statistical tests are reported in Appendix F. Overall, from Figure 3, we observe that most features have a steadily increasing importance across the decade of text messages, reflecting their cumulative role in capturing diachronic divergence. Among these, length related features (i.e., length_char, length_word, avg_word_len) were the most influential, with character length exhibiting the steepest increase.

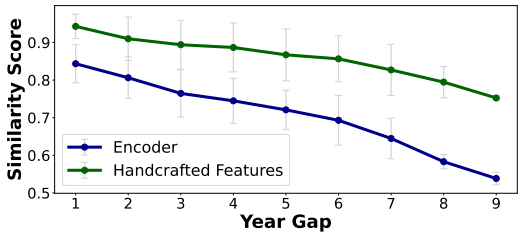


Figure 2: On CoSEM data, similarity score decreases as the year gap between data increases.

A second tier of features, including repeated-character words, Unicode emojis, ASCII emoticons, reduplication, and the particles *leh* and *lah*, displayed lower (SHAP < 0.10) but statistically significant contributions. In contrast, most other Singlish particles exhibited statistically non-significant trends over time, suggesting that though they function as stable identity markers of Singlish, they are weak predictors of temporal variation.

5.4 Diachronic Shifts in LIWC Features

To provide a more interpretable account of linguistic change, we applied Principal Component Analysis (PCA) to the LIWC features for each year and tracked the five highest-loading features within each of the three LIWC categories. Temporal trends are shown in Figure 4 and the full statistical analyses are summarized in Annex C.

Summary variables (Figure 4a) There is significant increase in *Clout*, *WPS*, *Analytic* related features, indicating shifts toward greater formality and structural complexity. By contrast, *tone* showed no consistent trend. Such pattern indicates that Singlish texts in recent years adopt a more structured and polished form compared to earlier corpora while its tone varies across the years.

Linguistic dimensions (Figure 4b) Most features remained stable, except for first-person singular pronouns (*i*) which declined over time. This reduction in self-referential markers could suggest a reduction in explicit personal discourse which we define as an individual's attempt of overtly making their presence visible through expressions of personal involvement. We can also observe this trend qualitatively in messages from different years, all expressing their intent of announcing arrival:

2012: (a) *Wait I'm reaching house in 10min like that* (b) *I reaching my*

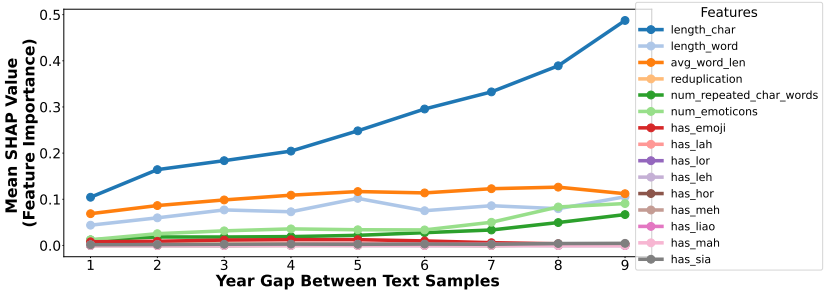
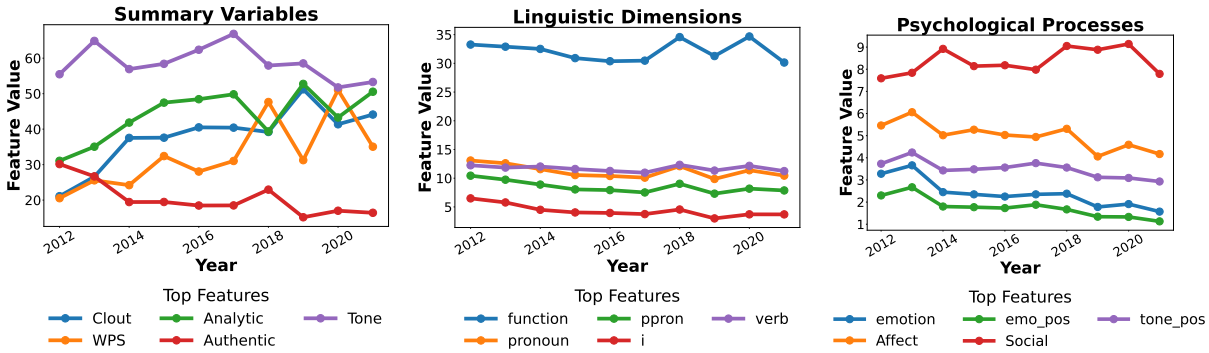


Figure 3: Feature Importance of each of the handcrafted features plotted against year-gap. The feature importance is measured by taking the mean (absolute) SHAP values.



(a) Summary Variables (b) Linguistic Dimensions (c) Psychological Processes

Figure 4: Trend of top LIWC features across categories.

2021: (a) *Reaching 9.20* (b) *Omw to nex*

In the 2012 samples, the subject (e.g., "I") is explicitly marked. By 2021, the messages drop the subject to focus squarely on the action. This observation directly maps to our quantitative findings showing a decline in self-referential markers alongside a shift toward structured and analytic discourse.

Psychological processes (Figure 4c) Affective expression shows a downward trend, with positive emotions and tone in particular decreasing across the decade. However, social process features remained stable, showing reduced emotional expressivity without diminished interpersonal orientation.

Collectively, LIWC analyses reveal that Singlish messages have become more structured and confident but less affective and self-focused over time.

5.5 LLM Realism and Temporal Neutrality

LLM performance varies substantially across architectures and prompting strategies. Among the base models, Mistral achieved the highest similarity (authenticity) under the DD prompting scheme ($S = 0.523$), followed by SeaLLM with CoT ($S = 0.417$), Qwen with CoT ($S = 0.337$), and DeepSeek-R1 with SC ($S = 0.321$). Fine-tuning improves realism for some models and degrades it for others: Qwen-FT showed a substantial improvement ($S = 0.673$ with CoT), whereas Mistral-FT underperformed relative to its base counterpart ($S = 0.437$), indicating that fine-tuning does not uniformly help (see Table 1).

Comparing each model's best-performing scheme against the null threshold $\sigma_{95} = 0.034$, DeepSeek-R1 ($\sigma = 0.028$) and SeaLLM ($\sigma = 0.033$) achieves temporal neutrality, while Mistral ($\sigma = 0.054$), Mistral-FT ($\sigma = 0.059$), Qwen ($\sigma = 0.043$), and Qwen-FT ($\sigma = 0.043$) retains detectable year-specific signals. Multi-year fine-tuning failed to suppress temporal signals. Mistral-FT showed increased variability relative to its base counterpart, while Qwen-FT remained sensitive to individual year profiles. Notably, other than SeaLLM, models achieving temporal neutrality are precisely those with the lowest authenticity scores, while the most authentic models fail neutrality.

5.6 Qualitative Error Analysis

To better interpret shifts in realism, we examined model outputs for cultural and grammatical nuances, focusing on what distinguishes models

across the authenticity-neutrality spectrum.

DeepSeek-R1, the most temporally neutral model ($\sigma = 0.028$) but least authentic ($S = 0.321$), produced outputs notably lacking Singlish-specific features, such as *"Okay, so you're going to go out later."* – using standard English with no discourse particles, code-switching, or colloquial markers. This confirms that its temporal neutrality reflects an absence of sociolectal grounding rather than genuine abstraction of invariant Singlish structure.

SeaLLM presents a more nuanced case: despite achieving temporal neutrality ($\sigma = 0.033$), it attained the second highest authenticity among base models ($S = 0.417$), producing outputs that incorporated Singlish particles more consistently than DeepSeek-R1 such as *"Yo, I'm thinking of it lah. What you up to?"*.

Among the more authentic but temporally biased models, outputs varied in depth of Singlish grounding. Qwen-FT struggled with phonetic authenticity despite higher similarity scores (more authentic), generating *"Naa lah. I'm just gonna stay at home and chill."* The use of *Naa* resembles an Americanized negation pattern; a culturally grounded speaker would use *No lah*, revealing a gap in deep colloquial knowledge.

Overall, four patterns emerge: (i) temporal neutrality most commonly reflects generic generation rather than abstracted Singlish, though SeaLLM's Southeast Asia-focused pretraining suggests broader exposure can partially decouple the two; (ii) the authenticity-neutrality trade-off holds strongly for most models but is not absolute; (iii) prompting effectiveness and fine-tuning gains are model-dependent; and (iv) residual diachronic signals largely persist across high-authenticity models, indicating they remain partially temporally anchored rather than fully timeless.

6 Discussion

6.1 Singlish Particle Analysis

SHAP analysis reveals that most Singlish discourse particles (*mah*, *liao*, *lor*, *hor*, *meh*) contribute minimally to diachronic classification, while only high-frequency particles (*lah*, *leh*, *sia*) show consistent model influence. Complementary LIWC analysis indicates a broader shift toward regularized, formal English structure, matching an overall decline in particle frequency since 2016 (see Figure 5).

Collectively, these patterns suggest that presence- and frequency-based metrics risk underestimating the structural role of particles, underscoring the need for interpretive frameworks that capture their layered pragmatic meanings.

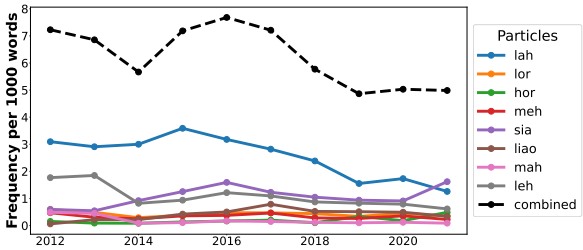


Figure 5: Normalized frequency/1000 words of individual (solid) and combined (dotted) Singlish particles in CoSEM over time.

6.2 Selective Formalization and Decreolization

Complementing the SHAP and encoder-derived analyses, key LIWC features provide interpretable grounding for these diachronic trends (see Figure 4). Notable increases in *Clout*, *Analytic*, and words per sentence (*WPS*), alongside declines in *Authentic* markers and first-person pronouns, signal a structural shift toward more formalized, depersonalized, and audience-oriented discourse. Conversely, stability in social process categories indicates that Singlish preserves its primary sociolinguistic function as a medium for interpersonal engagement.

These empirical patterns align with the processes of selective formalization and partial decreolization along the post-creole continuum (Sato, 1994). Early corpus iterations exhibit highly compressed, basilectal constructions (e.g., *"Can one la"* from CoSEM 2012), whereas later texts adopt acrolectal, Standard English-aligned syntactic frameworks without abandoning distinct sociolectal identity markers (e.g., *"u can do it one"* from CoSEM 2020). This transition from structural shortcuts to analytical complexity proves that Singlish is a moving target. This analysis tests whether language models can track real-world language shifts or if they are simply generating historical snapshots.

This adaptation toward broader structural intelligibility within borderless digital environments (Deuber et al., 2018) underscores how Singlish dynamically scales its accessibility while retaining core identity-bearing features.

6.3 Why LLMs Might Fall Short

LLM outputs show limited authenticity for Singlish, with models' response to fine-tuning and

Table 1: Comparison of mean similarity scores and standard deviations across models and prompting schemes. Bold text indicates the highest mean similarity score for each model, denoting peak authenticity.

Model	Baseline		CoT		Diverse-Decoding		Self-Consistency	
	Mean	Std	Mean	Std	Mean	Std	Mean	Std
Mistral	0.367	0.024	0.444	0.056	**0.523**	0.054	0.373	0.043
Mistral-FT	**0.437**	0.059	0.303	0.042	0.283	0.040	0.329	0.050
Qwen	0.168	0.020	**0.337**	0.043	0.105	0.010	0.240	0.031
Qwen-FT	0.668	0.034	**0.673**	0.043	0.659	0.021	0.570	0.031
DeepSeek-R1	0.198	0.017	0.228	0.041	0.268	0.028	**0.321**	0.028
SeaLLM	0.214	0.019	**0.417**	0.033	0.350	0.042	0.360	0.043

prompting varying by architecture Alizadeh et al. (2025). Critically, our null distribution analysis reveals that temporal neutrality does not imply success: models achieving neutrality (e.g. DeepSeek-R1) do so largely through generic rather than authentic Singlish generation, while the most authentic models (Qwen-FT, Mistral) retain the strongest temporal bias. This consistent inverse relationship across all architectures and prompting strategies points to a structural trade-off that models that learn authentic Singlish features inherit the temporal distribution of those features from historically situated training data, while models that avoid temporal anchoring do so by failing to acquire Singlish-specific structure in the first place.

However, the only exception to this trend is SeaLLM, which was able to achieve temporal neutrality and still produce relatively authentic outputs. This suggests that temporal neutrality may not always be a symptom of generic generation, and that Southeast Asia-focused pretraining (Nguyen et al., 2024) may provide broader cross-temporal coverage of Singlish features without anchoring to a specific era.

Two factors likely explain why this trade-off persists. First, standard language model training optimizes next-token prediction based on distributional frequency (McCoy et al., 2019) rather than abstract grammatical generalization. This reinforces epochal markers in the training data rather than filtering them out. Our pan-temporal fine-tuning confirms that rather than distilling a temporally invariant Singlish representation, Qwen-FT and Mistral-FT preserved transient historical variations, with Mistral-FT showing increased temporal variability relative to its base counterpart. Second, some variation may be irreducibly time-bound. Singlish operates along a post-creole continuum (Sato, 1994) where certain lexical and pragmatic markers are tied to changing social contexts, so stripping temporal context risks undermining the very features that make generated text recognizably Singlish.

Together, these factors suggest that temporal neutrality is a representational challenge rather than a prompting or fine-tuning one, requiring either temporally balanced training data or objectives that explicitly disentangle stable sociolectal structure from transient markers. These findings extend beyond Singlish. For any fast-evolving contact language or sociolect, LLMs will inherit the temporal biases of their training data, and temporal neutrality as defined here serves as a diagnostic for whether a model has learned the structural core of a variety or merely a historical snapshot.

7 Conclusion

We examined stylistic shifts in Singlish across a decade of digital communication using a framework combining neural encoders, handcrafted features, and psycholinguistic dimensions. Singlish exhibits substantial diachronic divergence driven primarily by structural features, while pragmatic markers serve as stable sociolectal identity markers. Aligning encoder representations with interpretable linguistic features and psycholinguistic dimensions reveals that much of the temporal separability learned by neural models can be explained through surface-level stylistic cues.

Against a null distribution baseline, LLMs reveal a fundamental trade-off: models generating authentic Singlish inherit its temporal biases, while temporally neutral models produce generic English outputs. Although SeaLLM's culture-specific pretraining suggesting this trade-off is not absolute. This positions our framework as a practical diagnostic tool for studying sociolectal grounding and temporal bias in LLMs across fast-evolving language varieties.

8 Limitations

This study focuses on SMS-style messages and may not generalize to social media text, where interactional norms and stylistic conventions differ. Future work should extend the analysis to text-based social media data to examine whether similar diachronic and stylistic patterns persist.

9 Ethical considerations

This study does not seem to have ethical considerations. The CoSEM dataset is publicly available and anonymized, containing no personally identifiable information. All analyses were conducted at the aggregate level, with a focus on linguistic patterns rather than individual users.

References

Meysam Alizadeh, Maël Kubli, Zeynab Samei, Shirin Dehghani, Mohammadmasiha Zahedivafa, Juan D Bermeo, Maria Korobeynikova, and Fabrizio Gilardi. 2025. Open-source llms for text annotation: a practical guide for model setting and fine-tuning. *Journal of Computational Social Science*, 8(1):17.

Rajiv Bajpai, Danyuan Ho, and Erik Cambria. 2016. Developing a concept-level knowledge base for sentiment analysis in singlish. In *International Conference on Intelligent Text Processing and Computational Linguistics*, pages 347–361. Springer.

Wenda Chen, Nancy F Chen, Boon Pang Lim, and Bin Ma. 2015. Corpus-based pronunciation variation rule analysis for singapore english. In *SLaTE*, pages 35–40.

Siew Yeng Chow, Chang-Uk Shin, and Francis Bond. 2024. This word mean what: Constructing a singlish dictionary with chatgpt. In *Proceedings of the 2nd Workshop on Resources and Technologies for Indigenous, Endangered and Lesser-resourced Languages in Eurasia (EURALI)@ LREC-COLING 2024*, pages 41–50.

Dagmar Deuber, Jakob Leimgruber, and Andrea Sand. 2018. Singaporean internet chit chat compared to informal spoken language: Linguistic variation and indexicality in a language contact situation. *Journal of Pidgin and Creole Languages*, 33:48–90.

Jacob Eisenstein, Brendan O'Connor, Noah A Smith, and Eric P Xing. 2014. Diffusion of lexical change in social media. *PloS one*, 9(11):e113114.

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 RE Leimgruber, and Jun Jie Lim. 2023. The corpus of singapore english messages (cosem). *World Englishes*, 42(2):371–388.

Amalesh Gope, Anusuya Pal, Sekholu Tetseo, Tulika Gogoi, Dinkur Borah, and 1 others. 2024. Multi-class identification of tonal contrasts in chokri using supervised machine learning algorithms. *Humanities and Social Sciences Communications*, 11(1):1–13.

Jack Grieve, Andrea Nini, and Diansheng Guo. 2018. Mapping lexical innovation on american social media. *Journal of English Linguistics*, 46(4):293–319.

Liangjie Hong and Brian D. Davison. 2010. Empirical study of topic modeling in twitter. In *Proceedings of the First Workshop on Social Media Analytics*, SOMA '10, page 80–88, New York, NY, USA. Association for Computing Machinery.

Lee-Hsun Hsieh, Nam-Chew Chua, Agus Trisnajaya Kwee, Pei-Chi Lo, Yang-Yin Lee, and Ee-Peng Lim. 2022. Singlish checker: A tool for understanding and analysing an english creole language. In *International Conference on Asian Digital Libraries*, pages 115–124. Springer.

Wesley Lee. 2016. In defense of singlish: A cultural interpretation of singapore english. *The Morningside Review*, 12.

Zhengyuan Liu, Shikang Ni, Aiti Aw, and Nancy Chen. 2022. Singlish message paraphrasing: A joint task of creole translation and text normalization. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 3924–3936.

Scott M Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30.

R. Thomas McCoy, Ellie Pavlick, and Tal Linzen. 2019. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3428–3448, Florence, Italy. Association for Computational Linguistics.

Ho Mian-Lian and John T Platt. 1993. Dynamics of a contact continuum: Singaporean english.

Lynnette Hui Xian Ng and Luo Qi Chan. 2024. What talking you?: Translating code-mixed messaging texts to english. *arXiv preprint arXiv:2411.05253*.

Lynnette Hui Xian Ng, Iain J Cruickshank, and Roy Lee. 2025. Examining the influence of political bias on large language model performance in stance classification. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 19, pages 1315–1328.

Ri Chi Ng, Nirmalendu Prakash, Ming Shan Hee, Kenny Tsu Wei Choo, and Roy Ka-Wei Lee. 2024. Sghatecheck: Functional tests for detecting hate speech in low-resource languages of singapore. *arXiv preprint arXiv:2405.01842*.

Xuan-Phi Nguyen, Wenxuan Zhang, Xin Li, Mahani Aljunied, Zhiqiang Hu, Chenhui Shen, Yew Ken Chia, Xingxuan Li, Jianyu Wang, Qingyu Tan, Liying Cheng, Guanzheng Chen, Yue Deng, Sen Yang, Chaoqun Liu, Hang Zhang, and Lidong Bing. 2024. SeaLLMs - large language models for Southeast Asia. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 294–304, Bangkok, Thailand. Association for Computational Linguistics.

Nourma Silvia Ningsih and Fadhlur Rahman. 2023. Exploring the unique morphological and syntactic features of singlish (singapore english). *Journal of English in Academic and Professional Communication*, 9(2):72–80.

Francesco Periti and Stefano Montanelli. 2024. Lexical semantic change through large language models: a survey. *ACM Computing Surveys*, 56(11):1–38.

Nigora Satibaldieva. 2024. Dynamics of linguistic evolution in the technological epoch. *Western European Journal of Linguistics and Education*, 2(1):9–11.

Charlene J. Sato. 1994. *Language change in a creole continuum: decreolization?*, page 122–144. Cambridge University Press.

Charlene Tan. 2005. English or singlish? the syntactic influences of chinese and malay on the learning of english in singapore. *Journal of Language and Learning*, 3(1):156–179.

Ying-Ying Tan. 2023. The curious case of nomenclatures: Singapore english, singlish, and singaporean english. *English Today*, 39(3):194–206.

Yla R Tausczik and James W Pennebaker. 2010. The psychological meaning of words: Liwc and computerized text analysis methods. *Journal of language and social psychology*, 29(1):24–54.

Jock Wong. 2004. The particles of singapore english: A semantic and cultural interpretation. *Journal of Pragmatics*, 36(4):739–793.

Tao Wu, Mengze Li, Jingyuan Chen, Wei Ji, Wang Lin, Jinyang Gao, Kun Kuang, Zhou Zhao, and Fei Wu. 2024. Semantic alignment for multimodal large language models. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 3489–3498.

Zheng-Xin Yong, Ruochen Zhang, Jessica Zosa Forde, Skyler Wang, Arjun Subramonian, Holy Lovenia, Samuel Cahyawijaya, Genta Indra Winata, Lintang Sutawika, Jan Christian Blaise Cruz, and 1 others. 2023. Prompting multilingual large language models to generate code-mixed texts: The case of south east asian languages. *arXiv preprint arXiv:2303.13592*.

Nan Zhao, Dongdong Jiao, Shuotian Bai, and Tingshao Zhu. 2016. Evaluating the validity of simplified chinese version of liwc in detecting psychological expressions in short texts on social network services. *PLoS One*, 11(6):e0157947.

Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan Ye, Zheyan Luo, Zhangchi Feng, and Yongqiang Ma. 2024. Llamafactory: Unified efficient fine-tuning of 100+ language models. *arXiv preprint arXiv:2403.13372*.

Debra Ziegeler. 2020. Changes in the functions of already in singapore english: A grammaticalization approach. *Journal of Pidgin and Creole Languages*, 35(2):293–331.

A Different Classifiers

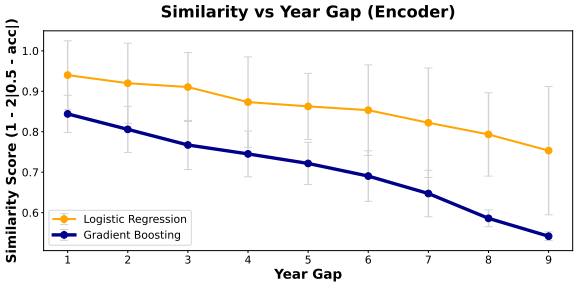


Figure 6: The performance trend using different classifiers are similar

B Content Impact Isolation

To isolate style from topical shifts, we trained a latent dirichlet allocation (LDA) model ($K = 15$) on yearly bins and measured cross-year variation via Jensen-Shannon divergence and cosine similarity (Hong and Davison, 2010). Minimal divergence would confirm that topical drift is negligible, isolating linguistic style as the primary driver of classifier performance. Yearly topic distributions were shown to be highly stable with a mean pairwise JS divergence of 0.0178 ($\sigma^2 = 0.008$) and mean cosine similarity of 0.998 ($\sigma^2 = 0.001$), confirming that temporal effects in subsequent analyses reflect stylistic rather than topical change.

Table 2: Trend analysis of LIWC features over time. * are features with statistically significant trends ($p < 0.05$, Spearman correlation)

Feature	Slope	Spearman ρ	p-value
Summary Variables			
Clout*	2.33	0.915	0.000
WPS*	2.38	0.842	0.002
Analytic*	1.60	0.697	0.025
Authentic*	-1.23	-0.818	0.004
Tone	-0.61	-0.309	0.385
Linguistic Dimensions			
function	-0.07	-0.188	0.603
pronoun	-0.22	-0.588	0.074
ppron*	-0.24	-0.648	0.043
i*	-0.28	-0.842	0.002
verb	-0.05	-0.309	0.385
Psychological Processes			
emotion*	-0.19	-0.881	0.001
Affect*	-0.16	-0.794	0.006
emo_pos*	-0.14	-0.915	0.000
tone_pos*	-0.10	-0.711	0.021
Social	0.08	0.382	0.276

C Statistical Analysis of LIWC Features

Zero-Shot

"Generate a conversation in Singlish, with the starting text given below.
Only give the conversation as the output. Do not include anything else. The conversation can be as long as you want.

Starting text:
Person A: Eh bro, you going out later or not?
Person B:

Chain-Of-Thought

"Generate a conversation in Singlish, starting with the text below.
As you continue the conversation, for every turn, think about what a typical Singaporean might say in response, considering local culture, humour, and context. Do not include your reasoning—only output the conversation.

Starting text:
Person A: Eh bro, you going out later or not?
Person B:

Diverse Decoding

"Generate a conversation in Singlish, starting with the text below.
Generate a natural-sounding conversation in the context of Singapore, starting with the text below.
Aim for variety in topics—let the conversation flow naturally, touching on everyday Singaporean life (food, family, weather, work, etc). Your output should be only the conversation.

Starting text:
Person A: Eh bro, you going out later or not?"
Person B:

Self-Consistency

"Generate a conversation in Singlish using the starting text below.
After each line, consider at least three plausible directions the conversation might take, and pick one that feels the most natural or interesting. Only output the conversation.

Starting text:
Person A: Eh bro, you going out later or not?
Person B:

D Variance Test Results

Table 3 presents the variance test results for the different LLMs and indicates the best-performing prompting methods.

Table 3: Variance test results for different models and best-performing prompting methods.

Model	Method	Std	χ^2 stat	p-val
SeaLLM	CoT	0.0329	97.90	1.0
DeepSeek-R1	SC	0.0282	71.51	1.0
Mistral	DD	0.0543	265.81	1.0
Mistral-FT	Baseline	0.0586	309.08	1.0
Qwen-FT	CoT	0.0432	168.04	1.0
Qwen	CoT	0.0434	169.24	1.0

E Handcrafted Features

Table 4 presents the details of the handcrafted features used in the stylistic similarity analysis.

Table 4: Features used in the stylistic similarity analysis.

Feature	Abbreviation	Unit
Handcrafted Features		
Lexico-structural attributes		
No. of characters	length_char	Message
No. of words	length_word	Message
Average word length	avg_word_len	Message
No. of repeated characters (e.g., *soooo*)	num_repeated_char_words	Message
No. of repeated words (e.g., *wait wait*)	reduplication	Message
Functional pragmatic markers		
Presence of discourse particles (e.g., *lah, lor, leh*)	has_{particle}	Message
Presence of Unicode emojis	has_emoji	Message
No. of ASCII emoticons (e.g., `:P`)	num_emoticons	Message
Psycholinguistic Dimensions (LIWC; Tausczik and Pennebaker, 2010)		
Summary variables		Year
Social confidence, words per sentence, logical thinking, authenticity, emotional tone	Clout, WPS, Analytic, Authentic, Tone	
Psychological processes		Year
Emotion-related and social process words	emotion, affect, emo_pos, social, tone_pos	
Linguistic dimensions		Year
Function words, pronouns, first-person pronouns, verbs	function, pronoun, ppron, i, verb	
Encoder-derived Features		
Sentence-level vector representations from a pre-trained encoder	–	Message

F Feature Importance for diachronic discriminability

Table 5 presents the results of the statistical analysis of SHAP feature importance.

Table 5: Statistical analysis of SHAP feature importance. * are features with Mean absolute SHAP values significantly above 0 ($p < 0.05$)

Feature	Mean SHAP	Std SHAP	t-stat	p-value
length_char*	2.68e-01	1.21e-01	1.46e+01	3.73e-07
avg_word_len*	1.06e-01	1.85e-02	2.13e+01	4.83e-05
length_word*	7.80e-02	1.91e-02	9.75e+00	2.70e-04
num_emoticons*	4.42e-02	2.63e-02	6.51e+00	2.51e-04
num_repeated_char_words*	2.97e-02	1.78e-02	6.70e+00	2.24e-04
has_emoji*	8.49e-03	3.62e-03	6.44e+00	1.28e-03
has_sia*	2.89e-03	9.66e-04	4.50e+00	2.68e-03
reduplication*	2.48e-03	6.47e-04	4.98e+00	8.61e-03
has_leh*	1.68e-03	7.38e-04	3.92e+00	3.62e-03
has_lah*	1.56e-03	5.73e-04	3.86e+00	4.50e-03
has_liao	1.37e-03	9.55e-04	2.48e+00	6.01e-02
has_hor	5.41e-04	8.71e-04	1.61e+00	1.48e-01
has_lor	3.49e-04	1.38e-04	2.40e+00	5.54e-02
has_meh	2.14e-04	1.66e-04	1.94e+00	1.00e-01
has_mah	7.59e-05	7.13e-05	1.20e+00	2.59e-01