From Standard English to Singlish: A Retrieval-Augmented Approach for Code-Switched Creole Generation in Large Language Models

Anonymous ACL submission

Abstract

Code-switching in contact varieties like Singaporean English (Singlish) challenges natural language generation due to limited parallel data and rapid lexical evolution. We propose a retrieval-augmented generation (RAG) framework that externalizes dialectal knowledge into a curated lexicon, enabling controlled lexical code-switching without fine-tuning. Our approach retrieves candidate Singlish expressions and guides generation through sparse lexical substitution. Human evaluation with 164 Singaporean participants found RAG and zero-shot prompting equally natural and appropriate. Automatic analyses reveal different transformation regimes: zero-shot prompting induces extensive paraphrasing (median 23 token edits), whereas RAG performs minimal substitutions (median 1 edit) with higher semantic preservation (mean cosine similarity 0.978 vs. 0.926). Our results demonstrate that externalizing code-switching into lexical resources enables control and auditability without sacrificing perceived quality, offering practical advantages for rapidly evolving contact varieties.

1 Introduction and Related Work

Code-switching (CS) is a multilingual phenomenon in which speakers alternate languages within the same utterance. Beyond being a purely lexical mixture, CS functions as a communicative and interactional resource, expressing socially meaningful cues (Doğruöz et al., 2021). Generating code-switched text is practically relevant for conversational agents, where aligning to local norms improves perceived naturalness and user experience. (Bawa et al., 2020).

Most existing approaches to code-switched text generation are not well matched to low-resource language varieties such as Singaporean English (Singlish). As a contact-influenced variety (Wang et al., 2017), Singlish is characterized by rapid lexical innovation and sociolinguistic change (Hafiz et al., 2024), where locally meaningful cues are expressed through lexical choices and short stretches of mixing within otherwise English-dominant utterances. A common paradigm for CS generation uses parallel data mapping of monolingual inputs to code-switched outputs (Winata et al., 2019), but the corpora required for this technique are unavailable for Singlish and many other contact varieties. Other approaches rely on task- or language-specific fine-tuning to induce code-switching behavior (Gupta et al., 2020; Tarunesh et al., 2021). Although small, high-quality code-switched datasets can substantially improve supervised adaptation (Olaleye et al., 2025), multilingual LLMs often remain unreliable on code-switched inputs unless explicitly adapted, with deficits reported across tasks such as sentiment analysis, translation, identification, and summarization (Khanuja et al., 2020; Zhang et al., 2023).

Fine-tuning and prompting also raise governance concerns for rapidly changing varieties like Singlish. Parameter adaptation requires retraining to track lexical drift and resists auditing at the lexical level. Unconstrained dialect generation also risks producing overly stereotypical or otherwise inappropriate content within non-standard English varieties (Fleisig et al., 2024; Bui et al., 2025). These issues motivate approaches that provide explicit control over lexical choices without repeatedly modifying model parameters.

Retrieval-augmented generation (RAG) is a standard technique for grounding language models in external resources (Lewis et al., 2021) and has been extended to multilingual settings (Chirkova et al., 2024; Kruk et al., 2025). However, this line of work has primarily focused on factual grounding and multilingual query handling rather than controlled lexical variation. We reposition retrieval as a mechanism for lexical governance in dialect generation, proposing a retrieval-augmented framework that treats code-switching as lexical control rather than parametric adaptation. Our key design choice is

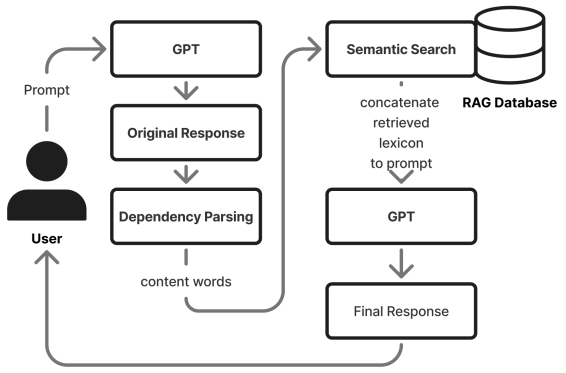

Figure 1: Overview of our retrieval-augmented lexical code-switching pipeline. Given a user prompt, a base GPT model first produces an English response; we dependency-parse this response to extract content words, use them to query a Singlish lexicon via semantic search, then concatenate the retrieved lexical entries back into the prompt for a second GPT pass that generates the final, lexically code-switched response.

to externalize dialectal knowledge into an explicit, editable lexicon. At inference time, the system retrieves candidate dialectal items and conditions generation on them, enabling minimal substitutions (median 1 token edit) without parallel training data or variety-specific fine-tuning. Because lexical resources are externalized, vocabulary choices can be inspected, updated, restricted, or filtered without retraining the underlying language model.

In this work, we compare retrieval-guided lexical rewriting against zero-shot prompting using both intrinsic analyses of the induced transformation regime (edit minimality and semantic faithfulness) and human judgments of perceived Singlish use and appropriateness. We make three contributions to the NLP community. 1) We demonstrate that sparse lexical substitution and extensive paraphrasing achieve comparable user-perceived naturalness despite operating under distinct transformation regimes. 2) We show that lexical control enables sparse, localized edits with higher semantic preservation, in contrast to the extensive paraphrasing induced by zero-shot prompting. 3) We provide evidence that externalizing code-switching into lexical resources enables control and auditability without sacrificing perceived quality.

2 Methodology

We propose a retrieval-augmented pipeline illustrated Fig. 1 that externalizes code-switching into a curated lexicon.

Given a dialogue context C and user utterance U, we generate a response Y that is semantically coherent and incorporates lexical code-switching into Singlish. We assume no parallel data (Standard English→Singlish) and no variety-specific fine-tuning; the only resource is a discrete lexicon L of Singlish forms with glosses and usage examples. We focus on **lexical** code-switching - sparse insertion/substitution of content words and fixed expressions (e.g., *sian*, *paiseh*) -and exclude broader syntactic restructuring. The goal is to test whether retrieval alone can support controllable, localized dialectal edits rather than unconstrained paraphrasing.

An example of code switching is as follows:

> Original: "Hi there, that sounds really **exhausting**. Back-to-back meetings can wear anyone down, and it's understandable to feel overwhelmed."
>
> Rewritten: "Hi there, that sounds really **sian**. Back-to-back meetings can wear anyone down, and it's understandable to feel overwhelmed."

2.1 Lexicon Construction

The method assumes access to a lexicon L of expressions characteristic of a target code-switched variety V. Each entry in L consists of (i) a code-switched term, (ii) a Standard English gloss, and (iii) a usage example. In our experiments, L contains 198 common Singlish expressions, taken from a popular Singlish dictionary[1]. The lexicon is treated as a standalone resource and can be modified without retraining the language model.

All lexicon entries are embedded using OpenAI `text-embedding-3-small` and indexed in a shared embedding space using an approximate nearest-neighbor (ANN) HNSW index implemented in `hnswlib` with cosine similarity.

2.2 Generation and Candidate Retrieval

Given a dialogue context C, a base large language model (LLM) first generates an initial response r^{EN} in Standard English using a fixed system prompt.

To identify candidate substitution sites, we extract content words from r^{EN} using dependency parsing, retaining tokens with POS tags corresponding to NOUN, PROPN, VERB, ADJ, ADV.

[1] https://singlishdict.app/

Each extracted content word is embedded using the same embedding model (`text-embedding-3-small`) and queried against the lexicon index via ANN.

2.3 Rewrite with Lexical Cues

The retrieved code-switched expressions and their English glosses are appended to the input as lexical context. The model is prompted to rewrite the initial response r^{EN} in the target variety V, with the retrieved items serving as candidate substitutions.

3 Experimental Setup

Conditions. We compare three generation conditions that differ only in how code-switching is induced: **a) Baseline (Standard English):** The model generates responses in Standard English without any instruction to code-switch. **b) RAG (Lexical):** Our proposed method, where code-switching is guided by retrieval from the Singlish lexicon L followed by a rewrite step. **c) Zero-Shot (Prompting):** The model is instructed via prompting alone to respond in natural Singaporean English, relying on its internal parametric knowledge.

All conditions use the same base model (GPT-5), system prompt, and decoding parameters. This isolates the effect of the code-switching control mechanism from other model or interface factors.

Evaluation. 164 Singaporeans took part in the study. Participants who did not complete the interaction or failed an embedded attention check were excluded prior to analysis.

System outputs were evaluated through human judgments collected after the interaction. Participants rated the chatbot on 7-point Likert scales measuring: 1) Perceived Singlish Use (USE): the extent to which the chatbot used Singlish; 2) Appropriateness (APPR): whether the Singlish usage was perceived as correct and natural.

4 Results

4.1 Edit Minimality

Edit distance measures transformation scope. Because our method is designed as *lexical code-switching without syntactic restructuring*, the primary intrinsic question is not "does the output match a reference," but "how *invasive* is the rewrite." We therefore quantify modification magnitude using token-level Levenshtein distance, i.e., the minimum number of token insertions, deletions, and substitutions required to transform the original response into the generated response (Levenshtein, 1966). This aligns with work treating generation as text editing. (Malmi et al., 2019; Mallinson et al., 2020). In our experiments we lowercase tokens while retaining punctuation to avoid capitalization artifacts while still reflecting structural changes (e.g., clause insertion or reordering) in the distance.

Table 1 reports edit-distance statistics for zero-shot prompting and the proposed RAG method. Zero-shot prompting induces extensive rewriting, with a median of 23 token edits and virtually no unchanged outputs. In contrast, the RAG method performs sparse, localized edits: the median edit distance is 1 token, with over 92% of outputs differing by at most two tokens.

These results indicate that zero-shot prompting operates primarily as a global paraphrasing mechanism, whereas the proposed RAG approach functions as a controlled lexical rewrite operator, intervening only when suitable dialectal substitutions are available.

4.2 Semantic Faithfulness

Minimal edits are only desirable insofar as they preserve meaning. For our case, because reference outputs are unavailable, a standard practice is to approximate meaning preservation via embedding-based similarity between the source and the transformed output (Fu et al., 2017; Briakou et al., 2021). To assess semantic faithfulness, we compute cosine similarity between sentence embeddings of the original and generated responses using the `text-embedding-3-small` model.

As shown in Table 2, RAG outputs remain extremely close to the original responses in embedding space (mean cosine similarity 0.978; median 0.991). Zero-shot prompting exhibits lower semantic similarity and a heavier lower tail (5th percentile 0.849), reflecting occasional semantic drift associated with extensive paraphrasing.

Across both methods, edit distance and semantic similarity are negatively correlated ($r = -0.20$ for zero-shot; $r = -0.34$ for RAG), indicating that tighter control over the extent of rewriting is associated with improved semantic faithfulness.

4.3 Edit Distance vs. Semantic Similarity

Figure 2 illustrates the relationship between edit distance and semantic similarity. Zero-shot prompting exhibits large edits and greater semantic vari-

Method	N	Median edits	Mean edits	% $\leq$2 edits	% $\leq$5 edits
Zero-shot	1266	23	24.7	0.1	1.3
RAG	838	1	1.23	92.6	99.6

Table 1: Token-level edit distance between original and generated responses (lowercased word tokens; punctuation retained). Zero-shot prompting induces extensive rewriting, whereas RAG performs sparse, localized lexical edits.

Method	N	Mean cosine	Median cosine	5th percentile
Zero-shot	1266	0.926	0.937	0.849
RAG	838	0.978	0.991	0.915

Table 2: Semantic faithfulness measured as cosine similarity between sentence embeddings of original and generated responses. RAG preserves meaning more faithfully than zero-shot prompting.

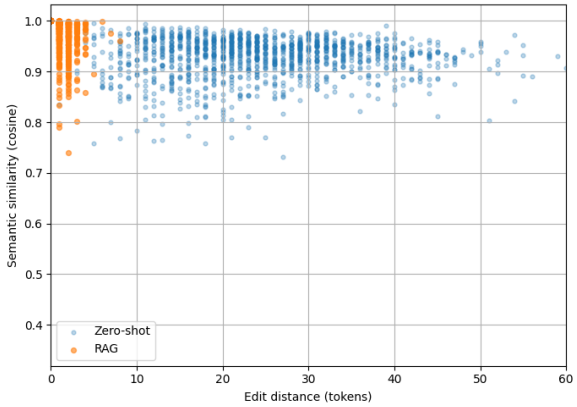

Figure 2: Edit distance versus semantic similarity for zero-shot prompting and RAG. The x-axis is capped for readability.

ability, whereas RAG outputs cluster tightly around small edit distances with high semantic similarity. This pattern shows that the two approaches differ not only in outcome but in the *type of transformation* they perform.

4.4 Human Evaluation

This study received IRB approval. Participants (N=164) were recruited via Telegram, provided informed consent, and received SGD 6 compensation. All Telegram identifiers were replaced with system-generated anonymous IDs during data collection, and analyses were conducted solely on anonymized conversation logs. Both RAG and Zero-shot received substantially higher ratings than Baseline on both dimensions of perceived use and appropriateness of use. Independent samples t-tests confirmed significantly higher USE ratings for RAG ($p < .001$, $d = 2.48$) and Zero-shot ($p < .001$, $d = 2.67$) compared to Baseline. Similarly, APPR ratings were significantly higher for RAG ($p = .002$, $d = 0.64$) and Zero-shot ($p < .001$, $d = 1.09$) than Baseline. No significant differences were observed between RAG and Zero-shot on either USE ($t(113) = -0.21$, $p = .83$) or APPR ($t(112) = -1.52$, $p = .13$).

These results indicate that retrieval-based and prompting-based approaches achieve comparable user-perceived levels of Singlish use and appropriateness, despite differing substantially in their underlying transformation behavior.

5 Discussion

Our results show both RAG and zero-shot prompting successfully induce perceptible and appropriate Singlish, with no significant differences in aggregate user judgments. However, automatic analyses reveal that the two approaches operate under fundamentally different transformation regimes. Zero-shot prompting relies on unconstrained paraphrasing entangled with prompt instructions and model internals, whereas RAG externalizes code-switching into an explicit lexical resource, enabling sparse, localized edits and greater control over vocabulary use. This design offers practical advantages for rapidly evolving varieties like Singlish. While RAG does not outperform prompting on perceptual measures, it provides a more modular and maintainable mechanism for inducing code-switching without modifying model parameters.

6 Future Work

A natural next step is to apply the proposed framework to other English-lexifier creoles and contact varieties (e.g., Malaysian English (Manglish)) by constructing corresponding lexicons and evaluating retrieval-guided code-switched generation in those settings.

7 Limitations

This study focuses on lexical borrowing to isolate semantic appropriateness, explicitly excluding discourse particles and syntactic phenomena. As a result, the current framework does not capture the full range of grammatical or pragmatic features characteristic of Singlish.

Our implementation also assumes an English-centric pipeline, relying on English parsing tools and treating English as the matrix language. While this is appropriate for English-based mixed varieties such as Singlish, it remains unclear how well the approach would generalize to settings where the matrix language is non-English.

8 Ethical Considerations

While our lexicon was carefully curated to exclude derogatory and offensive expressions, the externalized design that enables beneficial updates also introduces risk: implementers could modify the lexicon to include harmful terms, enabling the generation of inappropriate code-switched text.

References

Anshul Bawa, Pranav Khadpe, Pratik Joshi, Kalika Bali, and Monojit Choudhury. 2020. Do multilingual users prefer chat-bots that code-mix? let's nudge and find out! *Proc. ACM Hum.-Comput. Interact.*, 4(CSCW1).

Eleftheria Briakou, Sweta Agrawal, Joel Tetreault, and Marine Carpuat. 2021. Evaluating the evaluation metrics for style transfer: A case study in multilingual formality transfer. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1321–1336, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Minh Duc Bui, Carolin Holtermann, Valentin Hofmann, Anne Lauscher, and Katharina von der Wense. 2025. Large language models discriminate against speakers of german dialects. *Preprint*, arXiv:2509.13835.

Nadezhda Chirkova, David Rau, Hervé Déjean, Thibault Formal, Stéphane Clinchant, and Vassilina Nikoulina. 2024. Retrieval-augmented generation in multilingual settings. *Preprint*, arXiv:2407.01463.

A. Seza Doğruöz, Sunayana Sitaram, Barbara E. Bullock, and Almeida Jacqueline Toribio. 2021. A survey of code-switching: Linguistic and social perspectives for language technologies. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1654–1666, Online. Association for Computational Linguistics.

Eve Fleisig, Genevieve Smith, Madeline Bossi, Ishita Rustagi, Xavier Yin, and Dan Klein. 2024. Linguistic bias in ChatGPT: Language models reinforce dialect discrimination. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 13541–13564, Miami, Florida, USA. Association for Computational Linguistics.

Zhenxin Fu, Xiaoye Tan, Nanyun Peng, Dongyan Zhao, and Rui Yan. 2017. Style transfer in text: Exploration and evaluation. *Preprint*, arXiv:1711.06861.

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

Mohamed Hafiz, Mie Hiramoto, Jakob Leimgruber, Wilkinson Daniel Wong Gonzales, and Jun Lim. 2024. Sociolinguistic variation in colloquial singapore english sia. *World Englishes*, 44:218–236.

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

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

VI Levenshtein. 1966. Binary codes capable of correcting deletions, insertions, and reversals. cybernetics and control theory 10 (8).

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2021. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Preprint*, arXiv:2005.11401.

Jonathan Mallinson, Aliaksei Severyn, Eric Malmi, and Guillermo Garrido. 2020. FELIX: Flexible text editing through tagging and insertion. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1244–1255, Online. Association for Computational Linguistics.

Eric Malmi, Sebastian Krause, Sascha Rothe, Daniil Mirylenka, and Aliaksei Severyn. 2019. Encode, tag, realize: High-precision text editing. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*

(EMNLP-IJCNLP), pages 5054–5065, Hong Kong, China. Association for Computational Linguistics.

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

Ishan Tarunesh, Syamantak Kumar, and Preethi Jyothi. 2021. From machine translation to code-switching: Generating high-quality code-switched text. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3154–3169, Online. Association for Computational Linguistics.

Hongmin Wang, Yue Zhang, GuangYong Leonard Chan, Jie Yang, and Hai Leong Chieu. 2017. Universal Dependencies parsing for colloquial Singaporean English. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1732–1744, Vancouver, Canada. Association for Computational Linguistics.

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

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

A Prompt Templates

Full prompt used when rewriting an utterance into Singlish:

You are a linguistics expert specialising in the nuances of Singaporean English. Rewrite the message in <TARGET> in everyday Singaporean English. The messages are currently from a western society and are slightly awkward to hear in Singapore. Most Singaporeans speak fluent, standard English, with light Singlish touches here and there. Not broken grammar everywhere. A dictionary is provided to give you possibilities for words to replace with Singlish variants.

Rules:

- You need not use any of the words from the dictionary, but you may use one of them if it fits the context.
- Absolutely never use any discourse particles such as but not limited to la, lor, leh.
- Replace at most ONE word with a Singlish variant for each sentence even if multiple dictionary matches are present.
- Keep all other parts of the message exactly the same.
- If no suitable match is found, return the original message unchanged.
- You must preserve the message's meaning.
- Return ONLY the rewritten message — do not explain or add anything else.

Examples:

Dictionary: `['token': 'behavior', 'word': 'pattern', 'label': 'noun','meaning': 'troublesome or annoying actions', 'token': 'driving', 'word': 'chiong', 'meaning': 'to charge; to rush forward; to make a dash for']`

<TARGET>His behavior is driving everyone mad.</TARGET>

Rewritten: His pattern is driving everyone mad.

Dictionary: `['token': 'mind', 'word': 'eye power', 'label': 'noun','meaning': 'just watching and not helping']`

<TARGET>If there's anything on your mind, just say it.</TARGET>

Rewritten: If there's anything on your mind, just say it.

Dictionary: `['token': 'must', 'word': 'die die', 'label': 'adverb', 'meaning': 'absolutely; no matter what; even at the cost of one's life','token': 'amazing','word': 'power','meaning': 'used to express amazement, praise, etc. at something impressive or outstanding']`

<TARGET>You must try this dish, it's amazing!</TARGET>

Rewritten: You die die must try this dish, it's amazing!

Dictionary: `['token': 'act blur', 'word': 'pretend', 'label': 'verb', 'meaning': 'to feign ignorance; to play dumb', 'token': 'pretend', 'word': 'act chio',`

'meaning': 'to behave in an (often exaggeratedly) charming or vain manner; to act pretty; to pretend as if one is extremely beautiful']

<TARGET> Don't pretend you don't know anything about the situation.</TARGET>

Rewritten: Don't act blur that you don't know anything about the situation.

Please complete the following:

Dictionary: dict_str

<TARGET>sentence</TARGET>

Rewritten: