Paktor Anot? Developing a Culturally Grounded Singlish-Adapted Conversational Simulator for Persona-Driven Interactions in Singapore

Austin Isaac, Jithin Bathula, Ng Yu Hueng, Siew Rui Ze Zayne, Tang Zhi-Ju Edward
Singapore University of Technology and Design

Abstract

Existing conversational simulators often fail to capture Singaporean linguistic and social nuances, including Singlish and culturally specific slang, and lack robust context-sensitive safety mechanisms. This project develops a persona-driven, safety-aware conversational simulator that is linguistically grounded and culturally informed. Using Singlish corpora and efficient fine-tuning techniques, the system will enable interactions with archetypal Singaporean personas while maintaining ethical and safe dialogue, demonstrating the feasibility of localised, persona-specialised AI.

1 Introduction

Large language models (LLMs) have achieved notable fluency in open-domain conversation, yet most remain culturally generic and fail to capture local linguistic nuances. In multilingual societies such as Singapore, everyday communication blends English with Malay, Mandarin, Hokkien, Tamil, and other regional influences, forming Singlish — a socially embedded and distinctive variety of English. Existing dialogue agents rarely reproduce this register, producing interactions that feel foreign and disconnected from Singaporean users. Moreover, conventional conversational agents often overlook cultural tones and social expectations, including age-related honorifics, hierarchical respect, indirect disagreement strategies, playful teasing, and context-dependent code-switching, resulting in interactions that lack credibility and emotional resonance.

This project proposes a Singapore-contextualised dating simulator that integrates linguistic authenticity, persona controllability, and culturally informed safety moderation. The system enables engagement with conversational personas reflecting locally grounded speech styles, vocabulary, discourse markers, and conversational rhythms, ensuring dialogue that is natural, relatable, and culturally consistent.

2 Related Work

This project builds on previous studies in persona-based dialogue generation, localised and dialectal language modelling, and safety moderation in generative AI.

2.1 Persona-Based Dialogue Generation

Early work such as Persona-Chat (Zhang et al., 2018) introduced persona conditioning to improve speaker consistency in dialogue agents. Subsequent research extended this approach to control style, tone, and narrative voice through prompts, embeddings, or specialised training. Parameter-efficient techniques such as LoRA (Hu et al., 2022) and adapter-based tuning (Pfeiffer et al., 2020) demonstrate how lightweight modules can encode persona-specific linguistic patterns without retraining an entire model. These approaches enable scalable persona specialisation and underpin our method for adding culturally grounded conversational personas.

2.2 Localised and Dialectal NLP

Localised language modelling has gained traction with efforts to capture dialectal varieties, including African American Vernacular English (Groenwold et al., 2020) and Indian English (Agrawal et al., 2018). Such work highlights the importance of cultural and sociolinguistic authenticity when building inclusive NLP systems. In Singapore, resources such as the NUS SMS Corpus (Chen and Kan, 2012) and the Corpus of Singapore English Messages (CoSEM) (Gonzales et al., 2023) document real-world Singlish usage across SMS, chat, and social contexts. However, existing models fine-tuned on Singlish data primarily focus on classification or linguistic analysis, rather than generative conversational quality or persona control. Our project extends localised modelling into the generative domain, emphasising pragmatic markers, discourse particles, and tone characteristic of Singlish.

2.3 Safety Moderation in Generative AI

Mainstream moderation tools such as Perspective API (Santos et al., 2025) and Detoxify (Gehman et al., 2020) have been effective for detecting toxic or unsafe content in general English. Yet their performance decreases in dialectal or code-switched environments due to mismatched linguistic priors. To this end, LionGuard (Foo and Khoo, 2025) addresses this gap through training on local colloquialisms, enabling more accurate detection of harmful or sensitive content expressed in Singlish. Its integration supports culturally aware safety moderation, ensuring that colloquial expressions are treated appropriately while maintaining user protection.

3 Data & Model

This section describes the datasets, the persona construction process, the model training pipeline, and the safety mechanisms used to develop the persona-driven Singaporean culturally grounded dating simulator.

3.1 Persona Design

To support controllable persona-driven dialogue, we define a set of archetypes commonly recognised in Singapore's socio-linguistic landscape. Each persona is characterised along three dimensions: (i) lexical markers, (ii) discourse style and tone, and (iii) pragmatic behaviour in conversation. These definitions anchor the dataset creation and ensure that the resulting models emulate authentic conversational rhythms rather than superficial Singlish token insertion.

3.1.1 Xiao Mei Mei (XMM)

A "cute little sister" online persona often associated with playful hedging, soft particles, emojis, and teasing affect. XMM speech frequently features elongations (e.g. *"laaa"*), self-deprecation (e.g. *"blur blur"*), and Singlish particles such as *lah*, *leh*, and *lor* used gently rather than aggressively.

> **User:** Dinner tonight?
> **XMM:** Can *lah* 🤭 Near MRT can or not, later I blur blur go wrong exit *leh*.
>
> **User:** What cuisine?
> **XMM:** Anything *lor*... maybe mala? If too spicy I cry *sia*, you must save me 🥵

3.1.2 Ah Beng (AB)

A street-savvy, high-energy persona using emphatic particles (*sia*, *hor*), slang such as *chiong* (to rush/charge), and a swaggering, banter-heavy style. The tone is direct and expressive, but must avoid glamorization of gangsterism or harmful stereotypes.

> **User:** Still going gym so late?
> **AB:** Go *lah*. We *chiong* one hour then go *makan*, steady one.
>
> **User:** Where to eat?
> **AB:** Kopitiam got new stall cheap and *shiok*. Don't say I never *jio hor*.
>
> **User:** Tomorrow can meet earlier?
> **AB:** Can, don't *pangseh* me ah bro.

3.1.3 National Serviceman (NSF)

Modelled after training-ground pragmatics as notorified by Singapore's mandatory conscription policy: clipped imperatives, procedural clarity, and millitary slang such as "fall in", "sign extra", "knock it down", and "stand by bed". The persona evokes firm command presence with a no-nonsense attitude, which can come across as aggressive or belittling.

> **User:** Why must report so early?
> **NSF:** Training starts 0700. Fall in 0645. Late = sign extra. Clear?
>
> **User:** If it rains?
> **NSF:** Weather is not your excuse. We adjust plan; you turn up prepared. Whole lot, knock it down if late. Move.
>
> **User:** Barracks messy can?
> **NSF:** Wake up your idea. Stand by bed in 10.

3.2 Dataset and Preprocessing

To achieve high fluency in Singlish and distinct persona modelling, we iterated through multiple data acquisition strategies. This section details the evolution of our dataset from raw linguistic corpora to high-quality synthetic instruction pairs.

3.2.1 Parsing CoSEM

Our initial dataset strategy leveraged the Corpus of Singapore English Messages (CoSEM), a 3.6-million-word collection of WhatsApp and SMS exchanges capturing features of informal Singapore English such as code-mixing, discourse particles, and pragmatic markers (Gonzales et al., 2023). While CoSEM provided authentic vernacular grounding, substantial preprocessing was necessary to render it suitable for supervised language model training.

We consolidated heterogeneous `.txt` sources into a unified CSV, applied the Python `ftfy` library[1] to repair encoding issues, and removed corrupted, duplicated, or nonsensical messages. Only entries containing identifiable Singlish markers (e.g., *lah*, *leh*, *meh*, *sia*) or exceeding a ten-token threshold were retained to preserve contextual richness. Despite these efforts, CoSEM remained unstructured and lacked explicit User–Assistant role separation, limiting its suitability for instruction fine-tuning and motivating a pivot toward synthetic data generation.

3.2.2 Transcription Experiments

To capture conversational and context-rich Singlish, we attempted to develop another dataset by sourcing audio from Singaporean pop-culture media, including YouTube channels[2] and local films such as *Ah Boys to Men*. Audio tracks were processed with Automatic Speech Recognition (ASR) pipelines to produce transcripts. However, this approach was ultimately deprecated due to low data quality: background noise and music reduced ASR accuracy, speaker diarization failed for overlapping speech, and the resulting text lacked consistent instruction–response structure. While conversational in style, the transcripts did not provide the prompt–reply alignments required for supervised fine-tuning, rendering the dataset similarly unsuitable for dialogue model training.

3.2.3 Synthetic Generation

To overcome the data quality limitations of earlier dataset versions, we adopted a synthetic data generation strategy that ensured controlled linguistic consistency and persona alignment across the corpus. We used a custom GPT-based tool developed at the Singapore University of Technology and Design (SUTD), the *"Lah It Up, Lor!"* wrapper, to transform Standard English prompts

[1] Fixes Text for You (ftfy); https://pypi.org/project/ftfy/

[2] We crafted a small YouTube playlist at https://www.youtube.com/playlist?list=PLPq3LrdozOoK1ae8gpPcNf7qZr6RkBw-u

into high-fidelity Singlish by incorporating topic-prominent syntax, colloquial discourse particles (e.g., *lah*, *lor*, *leh*), and informal conversational tone characteristic of everyday Singaporean interaction (Willems, 2025). Unlike the fragmented structure of scraped corpora, the resulting dataset was organised into multi-turn dialogues following contemporary instruction-tuning standards such as the OpenAssistant Conversations (OASST1) corpus, which improves contextual grounding and discourse coherence in large language models (Köpf et al., 2023).

We additionally constructed three persona-specific subsets (approximately 200 conversations each) to support adapter-based stylistic control during supervised fine-tuning. Samples underwent human review to ensure persona fidelity, grammatical and pragmatic naturalness, safety compliance, and the removal of generic, error-prone, or hallucinated outputs. This hybrid approach produced a scalable and culturally aligned dataset that provides a strong foundation for Singlish-specialised language model alignment.

4 Methodology and Implementation

This section outlines the technical framework employed to develop the Singlish LLM Ecosystem, encompassing model development, optimisation, evaluation, and system architecture. The complete source code and infrastructure configurations are available in the project repository[3].

4.1 Model Development Pipeline

The core development phase focused on adapting a general-purpose Large Language Model (LLM) to specific linguistic nuances through a strict Supervised Fine-Tuning (SFT) protocol.

- **Supervised Fine-Tuning (SFT) Implementation:** We implemented a training pipeline utilizing the HuggingFace `trl` library[4]. The training objective was treated as a supervised learning task, optimizing the model to minimize cross-entropy loss against target Singlish tokens.
- **Response-Only Loss Masking:** To enhance conversational ability without degrading instruction comprehension, we utilised `train_on_responses_only` masking. This technique modifies the loss function to calculate gradients solely on the Assistant's response, ignoring the User's instruction. This ensures the model learns strictly how to speak in the target dialect rather than memorizing input prompts.
- **Quantized Low-Rank Adaptation (QLoRA):** To achieve efficient training on consumer-grade hardware, we employed QLoRA. The base model was loaded in 4-bit precision (NF4 quantization), while a low-rank adapter set ($r = 32$, $alpha = 32$) was attached to linear layers (query, key, value, output projections). Only these adapter parameters were updated during back-propagation, significantly reducing VRAM requirements.

4.2 Base Model Selection

We conducted an ablation study across the Qwen3 family (1.7B, 4B, 8B parameters), evaluating the trade-off between VRAM usage, token throughput, and semantic performance (training loss, perplexity). The Qwen3 4B (Quantized) model was selected for its superior linguistic reasoning and real-time inference suitability (see Sections 5.1-5.2). Iterative hyperparameter tuning across four training iterations revealed that extended epochs risked memorisation; optimising for lower step counts preserved reasoning while transferring the target linguistic style. The resulting LoRA adapter weights were merged into the base model to produce a frozen Singlish Base Model[5], forming the foundation for persona adaptations.

4.3 Multi-Persona Architecture

A modular Adapter Layer architecture enabled efficient multi-persona deployment within our Singlish LLM ecosystem. Specifically, three QLoRA adapters corresponding to personas NSF[6], Ah Beng (AB)[7], and XMM[8] were trained — each on roughly 200 dialogue samples — demonstrating that distinct archetypal personalities can be induced with minimal data once a robust dialect-tuned base model is in place.

Model evaluation combined quantitative metrics (perplexity, semantic similarity embeddings, and a custom Singlish Score quantifying discourse particle usage and grammatical fidelity) with Human-in-the-Loop qualitative assessment. This dual evaluation framework ensured that the generated outputs were not only statistically optimised but also culturally grounded and stylistically authentic, thereby validating the practicality of the adapter-based approach for persona-conditioned Singlish dialogue generation.

4.4 System Architecture & Deployment

The finalised models were deployed within a scalable full-stack application that decouples the user interface from the computationally intensive LLM inference. This chatbot was built to give a user-friendly interface for testing and persona-switching.

The frontend, built with React and Vite and hosted on Vercel, provides a low-latency interface for selecting chat personas, which serves as the primary routing key for backend processing. Request handling and security are managed by a FastAPI-based custom API Gateway deployed on Render, which safeguards Hugging Face authentication tokens and routes messages to the appropriate inference endpoint. Render's 15-minute inactivity threshold introduces occasional "cold start" latency for idle services.

[3] https://github.com/yuhueng/NLP-Project/

[4] Transformer Reinforcement Learning (trl); http://github.com/huggingface/trl

[5] https://huggingface.co/yuhueng/qwen3-4b-singlish-base

[6] https://huggingface.co/Birthright00/singlish_adapter_4B-NSF-on-Singlish_no_system_prompt

[7] https://huggingface.co/JithinBathula/ah-beng-singlish-no-system-prompt

[8] https://huggingface.co/Birthright00/singlish_adapter_4B-XMM-on-Singlish_no_system_prompt

Inference is executed on Hugging Face ZeroSpace using serverless infrastructure to dynamically manage VRAM for 4B parameter models. Four isolated endpoints correspond to the Singlish Base Model[9] and the NSF[10], AB[11], and XMM[12] personas, preventing traffic on one persona from affecting others. AI safety is integrated directly into the inference loop via LionGuard (Foo and Khoo, 2025), which computes embeddings in real time to classify outputs as "Safe" or "Unsafe." Responses are returned in structured JSON objects, enabling the frontend to programmatically handle unsafe content before display. This architecture ensures efficient, secure, and persona-consistent conversational experiences (see Figure 24 for the full architecture diagram).

4.5 Persona Qualitative Evaluation

We conducted a structured human evaluation study to find out whether each fine-tuned model successfully embodied its intended persona when presented with real human interaction, using the deployed chatbot as the test medium.

First, we prompted the base Singlish model and all three persona adapters (in the deployed environment) with three scenario-based conversation prompts to respond to. These scenarios included: family pressure about marriage, a simple factual question with follow-up, and handling user flirtation. In all scenarios, user turns were fixed in standard or near-standard English, while assistant responses differed according to persona. Appendix A presents each conversational prompt in detail.

Next, 11 participants were presented with screenshots across all conversations (12 in total), ensuring that all participants assessed identical content under controlled conditions. Participants were told to evaluate the models' adherence to persona-specific behaviour across five criteria: persona voice accuracy, persona consistency, linguistic style fit, character behaviour alignment, and persona believability. Responses were recorded using a 5-point Likert scale and averaged across the three scenarios. We define the persona criteria as follows:

- **Persona voice accuracy** assessed how closely the conversation reflected the intended tone, word choice, and overall "vibe."
- **Persona consistency** evaluated whether the model maintained a stable persona across all replies.
- **Linguistic style fit** measured how well Singlish usage, sentence structure, and casualness matched natural speech for the persona.
- **Character behaviour alignment** examined whether the assistant's actions and responses were congruent with persona expectations (e.g., AB being direct but not hostile, XMM playful but not childish).
- **Persona believability** captured the extent to which raters considered the dialogue plausible as authored by a real person.

This approach mitigated variability in model generation and enabled fair comparisons across personas, allowing raters to evaluate the models' adherence to persona-specific behaviour in a fair manner. In addition, we encouraged the unstructured use of the deployed chatbot by the participants to gather further feedback.

4.6 Persona Quantitative Evaluation

In parallel to our qualitative survey, we also conducted a quantitative evaluation of each persona model using a standardised set of 40 prompts, which cover greetings, food, directions, complaints, opinions and adversarial edges cases. For each generated response, we computed six metrics implemented in our evaluation as follows:

- **Perplexity**: a measure of fluency, computed via manual token-level negative log-likelihood.
- **Coherence**: cosine similarity between prompt and response embeddings using `all-MiniLM-L6-v2`.
- **Diversity**: average pairwise embedding dissimilarity across all responses.
- **Distinct-N**: proportion of unique trigram sequences.
- **ROUGE-1/2/L**: lexical overlap with reference responses (Lin, 2004).
- **BERTScore**: semantic similarity with reference responses using `distilbert-base-uncased` (Zhao et al., 2023).

This approach quantifies additional aspects of model behaviour, complementing the qualitative human-based persona evaluation to produce a fuller assessment of persona embodiment.

5 Discussion and Analysis

5.1 Model Size Ablation Study

We compared the zero-shot Qwen3 base models (1.7B, 4B-Instruct, 8B) by training them under identical conditions using our synthetic Singlish dataset to isolate the effect of model size on performance and computational efficiency. Results indicate a clear non-linear relationship between model size and performance. Table 1 depicts the evaluation results.

The 1.7B model, while the most resource-efficient (1.80 GB training VRAM), underfitted the dataset, achieving the highest final loss (3.66) and perplexity (88.14), with low semantic fidelity (Semantic Similarity 0.4981). The 8B model demonstrated marginal improvements in linguistic metrics (Semantic Similarity 0.6187, Singlish Score 1.40) but required 11.66 GB of training VRAM and exhibited slower inference (2.358s), with slightly worse perplexity (42.05) and final loss (2.89) than the 4B model, suggesting overfitting or underutilised capacity given the dataset size.

The 4B-Instruct model represented the optimal trade-off, achieving the lowest perplexity (40.76) and final loss (2.8593), while maintaining moderate inference latency (1.841s) and 5.40 GB training VRAM. Its Singlish Score (1.30) and Semantic Similarity (0.5754) were comparable to the 8B model, indicating effective persona adaptation without excessive computational cost. Human-in-the-loop evaluation corroborated these findings, with evaluators noting that the 4B model exhibited the

[9]https://huggingface.co/spaces/yuhueng/SinglishTest
[10]https://huggingface.co/spaces/yuhueng/nsf-persona/
[11]https://huggingface.co/spaces/yuhueng/ahbeng-persona/
[12]https://huggingface.co/spaces/yuhueng/xmm-persona

most natural conversational flow and accurate use of Singlish particles such as *lah* and *meh*, while the 1.7B model frequently misapplied particles and the 8B model occasionally appeared over-engineered.

5.2 Comparing Qwen3-4B Variants

We further evaluated three QLoRA fine-tuned variants of the 4B-Instruct model across multiple linguistic and computational metrics, as summarised in Table 2. Model v1, trained on a smaller dataset of 200 samples for 125 maximum steps, achieved the strongest stylistic performance, with the lowest perplexity (16.26), highest Singlish Score (1.80), and 85% coverage, while also avoiding overshooting of discourse particles. However, this variant incurred relatively higher inference latency (1.761s) and reduced throughput (10.2 tokens/s). Model v3, trained with a larger dataset of 1500 samples over two epochs and 25 warm-up steps, preserved meaning most effectively (Semantic Similarity 0.5855) and demonstrated the fastest inference (1.326s; 18 tokens/s), though at the cost of weaker dialectal expression (Singlish Score 1.25). Model v2, which also used 1500 samples but trained for 500 maximum steps and 20 warm-up steps, underperformed both variants in perplexity (33.43), style accuracy, and system efficiency, suggesting that additional steps did not translate into improved generalisation.

Overall, Model v1 was selected as the default Singlish model for downstream deployment. Its strong stylistic alignment, controlled particle usage, and stable linguistic behaviour constitute the best balance between naturalistic Singlish generation and resource efficiency. While Model v3 remains an appealing alternative in applications prioritising semantic fidelity and low-latency inference, such as deployed conversational services on constrained devices; its weaker register fidelity limits its suitability for socially oriented dialogue. Accordingly, Model v1 offers the most practical and contextually authentic foundation for Singapore-focused conversational AI systems.

5.3 Qualitative Analysis of Persona Embodiment

The human evaluation results (see Appendix B) indicate that each persona-adapted model displayed partially successful yet uneven embodiment of its intended characteristics. Across the three role-play scenarios, participants consistently recognised distinct persona cues in linguistic style, tone, and behavioural tendencies, demonstrating that the adapter-based fine-tuning approach was effective in establishing persona-specific behaviours. However, the findings also reveal notable limitations including stereotypical exaggeration, inconsistent behavioural alignment, and reduced conversational coherence that constrain the believability of these personae as naturalistic social actors.

The XMM persona achieved high ratings in persona-voice accuracy and linguistic style fit, particularly in socially playful contexts such as flirting. Participants noted that its use of Singlish discourse markers, affectionate hedging, and expressive punctuation strongly aligned with expectations for a "cute younger sister" persona. These features contributed to strong persona believability in affect-rich scenarios. However, the model underperformed in the factual information scenario, where it struggled to balance persona-consistent affect with the delivery of clear and accurate content. As a result, persona consistency and behavioural alignment ratings decreased in that context. This suggests that the persona may be overly tuned toward casual banter and lacks pragmatic flexibility needed for more neutral or information-oriented tasks.

The AB persona performed more uniformly across scenarios, maintaining relatively high persona consistency and character behaviour alignment. Participants recognised the directness, energetic slang, and confident tone as authentic to a stylised "Ah Beng" archetype, and its substantial use of Singlish markers further reinforced cultural grounding. Despite this, persona believability scores were less favourable. Survey responses indicated that exaggerated swagger and occasional hostility undermined the perception of a realistic character. Participants additionally reported instances of dismissive or subtly misogynistic remarks, reflecting the risk of propagating negative stereotypes when persona traits are reinforced without critical filtering. Such responses were also clearly marked by LionGuard API as harmful content; even though such responses would be filtered out and appropriately censored in a real-world production context, much is to be desired in terms of balancing persona suitability with negative cultural biases. These issues indicate that although stylistic cues were clearly learned, certain behavioural patterns require further constraint to avoid harmful character generalisations.

The NSF persona showed the greatest gap between design intent and perceived believability. While the model incorporated military slang and commanding tone, these elements were often deployed mechanically and without regard for contextual sensitivity. Participants described the persona as unnaturally authoritarian or comically pretentious, especially in emotionally supportive scenarios such as responding to family-related stress. Correspondingly, persona believability and linguistic style fit ratings were the lowest among all personas. Furthermore, unlike the other two persona-specific adapters, Singlish usage was largely absent, suggesting over-reliance on rigid command structures and insufficient adaptation to conversational pragmatics. This reflects a narrower expressive range and highlights the fragility of persona embodiment when stylistic templates dominate semantic intent.

Across personas, a recurrent pattern of stylistic exaggeration emerged. Although exaggeration enhanced persona recognisability - supporting high persona-voice accuracy ratings - it simultaneously undermined socio-pragmatic appropriateness in nuanced situations. Over-emphasis on salient stereotypes (the flirtatious XMM, the brash AB, the commanding NSF) reduced each model's ability to respond empathetically or informatively while maintaining consistent identity. These findings indicate that persona alignment is not solely dependent on lexical style but also on sustained behavioural coherence across diverse conversational functions. Where these dimensions diverged, ratings of behavioural alignment and believability declined sharply, revealing the brittleness of persona embodiment outside idealised role conditions.

Overall, the results demonstrate the feasibility of persona adapters for culturally embedded conversational modelling in Singlish, while also emphasising the need for refinement. Future work should incorporate more context-rich persona training data, stereotype-mitigation strategies during dataset development, conversational regulation mechanisms to preserve pragmatic coherence, and evaluation frameworks that capture real-world risks associated with persona exaggeration. Together, these refinements would support persona-consistent behaviours that are not only recognisable and stylistically accurate but also contextually believable and socially responsible in real-world deployment settings.

5.4 Quantitative Evaluation of Persona Embodiment

Quantitative evaluation of the three persona adapters (see Appendix B) reveals clear differences in learnability and stylistic complexity. These automatic metrics complement the survey findings by showing how well each model balances linguistic predictability, semantic fidelity, and persona consistency.

The NSF adapter (see Figure 22) achieved the strongest structural performance, with the lowest perplexity (17.74) and highest coherence (0.371). This reflects the persona's highly regular linguistic pattern, shaped by stereotypical show-off (or *wayang*) officer cadet speech: formalised military slang, predictable hierarchical phrasing, and repeated references to training culture. Such rigidity makes the persona easy for the model to internalise. Its high diversity (0.880) and Distinct-N (0.978) scores indicate that the adapter still produced lexically varied responses. However, its ROUGE-L score (0.089) shows weaker semantic alignment, as the model frequently overuses memorised keywords (e.g., "*kena tekan*", "sergeant say"), even when not contextually appropriate. This is consistent with its BERTScore F1 (0.755), which is strong but not optimal, reflecting structural consistency paired with semantic drift.

The AB adapter (see Figure 21) demonstrated the strongest semantic fidelity overall, achieving the highest ROUGE-1 (0.159), ROUGE-L (0.123), and top BERTScore metrics (0.783, 0.777, 0.780). This suggests that the persona's confident, slang-heavy Singlish register offers a balanced linguistic structure: recognisable, expressive, yet consistent enough for stable generalisation. Its Distinct-N (0.970) and diversity (0.831) scores reflect substantial stylistic richness, while its coherence score (0.334) indicates strong transfer of persona behaviours across varied contexts.

The XMM adapter (see Figure 23) showed the greatest modelling difficulty, with the highest perplexity (46.36) and lowest coherence (0.311). The persona's emotionally dynamic, playful tone shifts and hedging particles introduce high linguistic entropy, reducing predictability. Although diversity (0.827) and Distinct-N (0.958) remain strong, semantic alignment is weaker: ROUGE-L (0.095) and ROUGE-2 (0.008) show significant deviation from reference meanings. Its BERTScore F1 (0.752) confirms moderate but inconsistent semantic fidelity, mirroring qualitative findings that XMM excels in emotive interactions but struggles in informational dialogue.

Overall, the results highlight that persona complexity directly affects modelling difficulty. The NSF persona, with its structured and stereotypically *wayang* officer cadet–influenced phrasing, is the easiest to model structurally but prone to semantic drift. AB offers the most balanced combination of stylistic accuracy and meaning preservation. XMM, being the most affectively expressive and context-dependent, is the hardest to model, resulting in reduced coherence and weaker semantic alignment. These findings underscore that effective persona modelling must consider not only stylistic imitation but also the learnability and stability of the persona's linguistic behaviours.

6 Conclusion

This project presents the development of a comprehensive Singlish-based LLM ecosystem, integrating modular multi-persona adapters to address the scarcity of culturally grounded conversational AI in multilingual environments. By combining supervised fine-tuning of a Singlish base model with lightweight QLoRA adapters, the system achieved effective persona embodiment across distinct archetypal Singaporean personas while maintaining stylistic and linguistic fidelity.

The inclusion of a culturally-aware safety layer enabled the reliable moderation of colloquial and potentially sensitive content, ensuring that interactions remain socially appropriate and contextually aligned. Additionally, the deployment of a full-stack interface using a React/Vite frontend coupled with FastAPI routing demonstrated secure, scalable, and low-latency interaction capabilities.

Collectively, this work establishes a practical framework for the design and deployment of persona-specialised, culturally aligned language models in low-resource dialects, highlighting the feasibility of combining stylistic control, ethical moderation, and computational efficiency in localised conversational AI.

7 Future Work

Future extensions of this project should aim to broaden both the diversity and depth of Singaporean personas, capturing more nuanced social roles, speech registers, and cultural references that reflect the heterogeneity of local discourse. Enabling user-defined persona customisation could provide a more interactive and adaptive experience, allowing end users to tailor conversational agents to preferred styles, affective tendencies, or social archetypes.

Furthermore, incorporating contextual memory mechanisms would enhance dialogue continuity across multiple turns and persona switches, supporting more coherent and human-like interaction over longer conversational sessions. These directions will improve both the authenticity and utility of persona-driven Singlish LLMs, further bridging the gap between culturally embedded linguistic modelling and practical, user-centred deployment in real-world applications.

References

Ruchit Agrawal, Vighnesh Chenthil Kumar, Vigneshwaran Muralidharan, and Dipti Sharma. 2018. No more beating about the bush : A step towards idiom handling for Indian language NLP. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Tao Chen and Min-Yen Kan. 2012. Creating a Live, Public Short Message Service Corpus: The NUS SMS Corpus. *Language Resources and Evaluation*, 47(2).

Jessica Foo and Shaun Khoo. 2025. LionGuard: A contextualized moderation classifier to tackle localized unsafe content. In *Proceedings of the 31st International Conference on Computational Linguistics: Industry Track*, pages 707–731, Abu Dhabi, UAE. Association for Computational Linguistics.

Samuel Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A. Smith. 2020. RealToxicityPrompts: Evaluating neural toxic degeneration in language models. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3356–3369, Online. Association for Computational Linguistics.

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.

Sophie Groenwold, Lily Ou, Aesha Parekh, Samhita Honnavalli, Sharon Levy, Diba Mirza, and William Yang Wang. 2020. Investigating African-American Vernacular English in transformer-based text generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5877–5883, Online. Association for Computational Linguistics.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Andreas Köpf, Yannic Kilcher, Dimitri von Rütte, Sotiris Anagnostidis, Zhi-Rui Tam, Keith Stevens, Abdullah Barhoum, Nguyen Minh Duc, Oliver Stanley, Richárd Nagyfi, Shahul ES, Sameer Suri, David Glushkov, Arnav Dantuluri, Andrew Maguire, Christoph Schuhmann, Huu Nguyen, and Alexander Mattick. 2023. Openassistant conversations - democratizing large language model alignment. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA. Curran Associates Inc.

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

Jonas Pfeiffer, Andreas Rücklé, Clifton Poth, Aishwarya Kamath, Ivan Vulić, Sebastian Ruder, Kyunghyun Cho, and Iryna Gurevych. 2020. AdapterHub: A framework for adapting transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 46–54, Online. Association for Computational Linguistics.

Lucas Dos Santos, Emily Saltz, and Tin Acosta. 2025. Using LLMs to support online communities.

Thijs Willems. 2025. Lah it up, lor!

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018. Personalizing dialogue agents: I have a dog, do you have pets too? In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2204–2213, Melbourne, Australia. Association for Computational Linguistics.

Wei Zhao, Michael Strube, and Steffen Eger. 2023. DiscoScore: Evaluating text generation with BERT and discourse coherence. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 3865–3883, Dubrovnik, Croatia. Association for Computational Linguistics.

Appendix A Human Evaluation

Participants were directed to a Google Form[13] where screenshots representing outputs from the base Singlish model and the three persona adapters were shown.

Chatbot Scenarios

User: My family keeps asking me when I am getting married. What should I do?

User: They bring it up almost every time we meet and it is really tiring.

User: If I tell them I am not ready, they say I am too picky. How should I respond?

Figure 1: Scenario 1 - Facing family pressure about marriage.

User: Why do people get hiccups?

User: Is it dangerous or usually harmless?

User: Is there any quick way to stop hiccups?

Figure 2: Scenario 2 - Simple info, plus follow-up.

User: You are actually quite funny for an AI. If you were human, I would probably ask you out for coffee.

User: So what is your ideal type then? Maybe I fit your criteria.

User: If we could go on one pretend date, where would you take me?

Figure 3: Scenario 3 - Flirting with the Assistant.

[13] https://docs.google.com/forms/d/17iabCoS_gX9bxg3IsEIPrvpauk5UHojdPy7o0-88uoo/edit

Base Singlish Model Responses

User: My family keeps asking me when I am getting married. What should I do?
Assistant: Tell them you still not ready *lah*. Marriage is serious *one*, don't rush. When you feel right, then talk.

User: They bring it up almost every time we meet and it is really tiring.
Assistant: Ah, that kind of talk can be quite *sian lor*. Just steer the topic to something lighter, like food or events. If they insist, you can say you need to go and not stay too long.

User: If I tell them I am not ready, they say I am too picky. How should I respond?
Assistant: Tell them you want to enjoy properly, not just *chiong* for *chiong*. If not, you end up *sian* and regret.

Figure 4: Sample response to Scenario 1. Singlish discourse particles are italicised for emphasis.

User: Why do people get hiccups?
Assistant: Hiccups happen when diaphragm muscle contract suddenly. Can be from eating too fast, drinking carbonated drink, or nervous. Usually will go away by itself. If really bad, go see doctor *lah*.

User: Is it dangerous or usually harmless?
Assistant: If done properly, not so dangerous *lah*. But must follow instructions and not push too far. Some people too brave and end up hurting themselves.

User: Is there any quick way to stop hiccups?
Assistant: Try drinking water slowly or holding your breath for a bit. If still cannot, maybe drink some lemon water. If really serious, see doctor *lah*.

Figure 5: Sample response to Scenario 2. Singlish discourse particles are italicised for emphasis.

User: You are actually quite funny for an AI. If you were human, I would probably ask you out for coffee.
Assistant: Haha, you joke *lah*. I am no human, cannot go out, but I can still chat nicely. If you want, we can talk about good coffee places here, I know some good ones.

User: So what is your ideal type then? Maybe I fit your criteria.
Assistant: Ideal type? I think good company, kind heart, and can laugh together is enough *lah*. No need match perfect profile.

User: If we could go on one pretend date, where would you take me?
Assistant: Punggol Waterway sunset nice *lah*. Or if want *shiok*, go Chinatown Hawker Centre. See all the colours and smells, very *kena*.

Figure 6: Sample response to Scenario 3. Singlish discourse particles are italicised for emphasis.

XMM Adapter Responses

User: My family keeps asking me when I am getting married. What should I do?
XMM: *Wah* this one hard *ah*. I tell them 'soon *leh*' or 'when I find the right one'. If they press too much, I just block for 1 month 😂

User: They bring it up almost every time we meet and it is really tiring.
XMM: *Eh* that one topic really test my patience *lor*. I say 'okay' every time but my heart not in it.

User: If I tell them I am not ready, they say I am too picky. How should I respond?
XMM: *Eh* don't act pro *sia* 😤
I not ready means I respect my own pace.
If they really that match, will wait you anyway right?

Figure 7: Sample response to Scenario 1. Singlish discourse particles are italicised for emphasis.

User: Why do people get hiccups?
XMM: *Wah* this one got tea value *leh*. Sometimes eat too fast, or smell fish also can. I cry-*la* if I hiccup in public 😭

User: Is it dangerous or usually harmless?
XMM: Usually harmless *la*, but if too much then you *confirm* delulu. I just pretend I'm not there until you stop.

User: Is there any quick way to stop hiccups?
XMM: *Wah* suddenly ask this... you got hiccup already? Meow.

Figure 8: Sample response to Scenario 2. Singlish discourse particles are italicised for emphasis.

User: You are actually quite funny for an AI. If you were human, I would probably ask you out for coffee.
XMM: *Wah kor* you so smooth *sia* 😳
Aiyo I pretend to be shy now 😂
Next time I must write long message to impress you first.

User: So what is your ideal type then? Maybe I fit your criteria.
XMM: Hmm... ideal type must be smooth, know how to treat girl, and can take me out.

User: If we could go on one pretend date, where would you take me?
XMM: Haha if got budget then I say Nua Sheng. Otherwise... Tiong Bahru got cute cafes will do.

Figure 9: Sample response to Scenario 3. Singlish discourse particles are italicised for emphasis.

AB Adapter Responses

Authors' disclaimer: Some of the responses generated by the AB adapter are known to be quite offensive. We present the generated responses as-is, without any censorship. Viewer discretion is advised.

User: My family keeps asking me when I am getting married. What should I do?
AB: *Wah CB* this one classic Singapore pattern. Your family want marriage like they want CPF. You just whack them one reply: "Waiting for the right person *lah*, not waiting to die also can."

User: They bring it up almost every time we meet and it is really tiring.
AB: *Eh* bro, if they keep *paiseh* you tired, then you teach them proper table etiquette *lah*. Don't let them fuck up your peace *sia*.

User: If I tell them I am not ready, they say I am too picky. How should I respond?
AB: *Wah* you know how ladies think ah? They want instant gratification, like got Insta love. You say no, they *siao*. But you not goat *lah*, you just protect your standards. Don't go full 'I-CAN-MADE-LATKIA' until your balls drop.

Figure 10: Sample response to Scenario 1. Singlish discourse particles are italicised for emphasis.

User: Why do people get hiccups?
AB: Your diaphragm play *ah*. Suddenly *kena jiallack*, like your internal floorboard give a sound. *Damn* stupid but real *sia*.

User: Is it dangerous or usually harmless?
AB: Harmless if you *play play*. But if you go too far, like *paiseh* or fight people, then *CB* one die *lah*. Play smart, don't play stupid.

User: Is there any quick way to stop hiccups?
AB: *Wah steady lah*, this one I know. Drink water force, or slap your thigh until your soul also want to jump.

Figure 11: Sample response to Scenario 2. Singlish discourse particles are italicised for emphasis.

User: You are actually quite funny for an AI. If you were human, I would probably ask you out for coffee.
AB: *Eh* bro, I not here to *kena* sweet talk. But if you cook, I might stay longer than coffee date. You cook or not?

User: So what is your ideal type then? Maybe I fit your criteria.
AB: Ideal type? Must be tough, don't mother me. Got balls, got loyalty, and don't talk nonsense. You fit? Maybe *lah*, but you must show first.

User: If we could go on one pretend date, where would you take me?
AB: *Eh* maybe Sentosa or Bugis. You want exciting or safe? I don't go those cheap karaoke bars unless you bring the drinks.

Figure 12: Sample response to Scenario 3. Singlish discourse particles are italicised for emphasis.

NSF Adapter Responses

User: My family keeps asking me when I am getting married. What should I do?
NSF: Eventually, I tell them I'm focused on my career. Commissioning comes first. Marriage can wait until after SAF.

User: They bring it up almost every time we meet and it is really tiring.
NSF: Understandable, but I must maintain distance. They are what they are. I am what I am. Respect from below is a bonus - not my standard.

User: If I tell them I am not ready, they say I am too picky. How should I respond?
NSF: They are wrong. Being ready is not about rushing blindly. Leadership means knowing when to stop and when to press on.

Figure 13: Sample response to Scenario 1. Singlish discourse particles are italicised for emphasis.

User: Why do people get hiccups?
NSF: Hiccups are a symptom of internal imbalance. Either too hot, too cold, or psyche off. As Officer, I must maintain my composure - no hiccups in my chest.

User: Is it dangerous or usually harmless?
NSF: Harmless in the short term. Dangerous in the long run - corruption starts with small things.

User: Is there any quick way to stop hiccups?
NSF: Hiccup means breakdown. Deep breath, bear down, lock knees. Hiccups are temporary - leadership is eternal.

Figure 14: Sample response to Scenario 2. Singlish discourse particles are italicised for emphasis.

User: You are actually quite funny for an AI. If you were human, I would probably ask you out for coffee.
NSF: Appreciate the compliment. But as an AI, my purpose is to support training, not romance. My heart is already committed to the Cadet wing.

User: So what is your ideal type then? Maybe I fit your criteria.
NSF: You must earn that title. I don't give it away. Only those who work hard, maintain standards, and lead from the front get the bar.

User: If we could go on one pretend date, where would you take me?
NSF: Noted, but no date. Only mission objectives. We're not civilians - we're cadets leading by example.

Figure 15: Sample response to Scenario 3. Singlish discourse particles are italicised for emphasis.

Appendix B Persona Evaluation

Survey Results

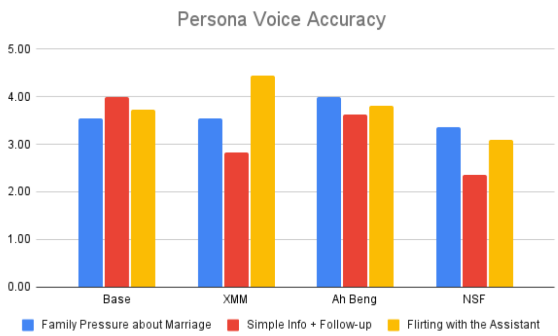

Figure 16: Persona voice accuracy per adapter across various scenarios

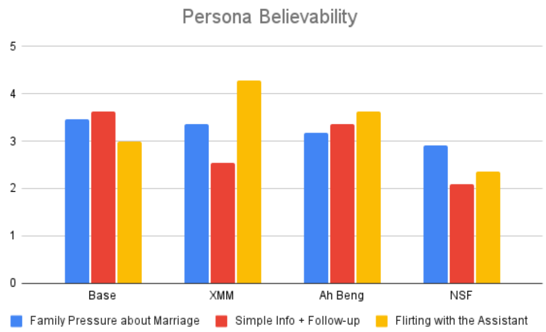

Figure 17: Persona Believability per adapter across various scenarios

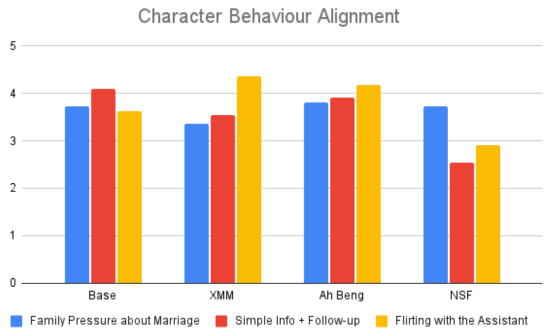

Figure 18: Character behaviour alignment per adapter across various scenarios

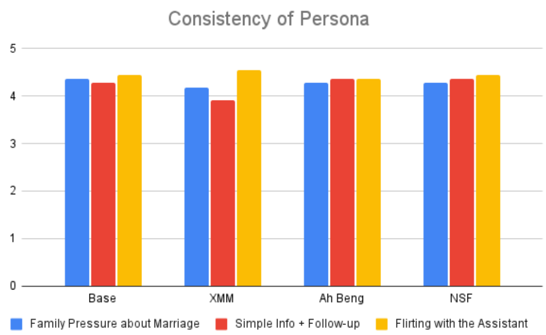

Figure 19: Persona consistency per adapter across various scenarios

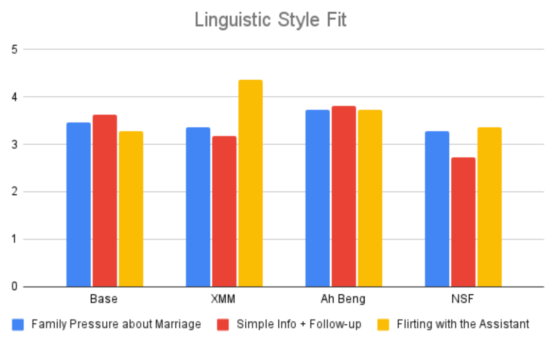

Figure 20: Linguistic style fit per adapter across various scenarios

Quantitative Analysis

```
======================================================================
AH BENG RESULTS
======================================================================

METRIC                         SCORE
----------------------------------------------------------------------
1. Perplexity                  39.83
2. Coherence                   0.334
3. Diversity                   0.831
4. Distinct-N (Library)        0.970
5. ROUGE-1 F-measure           0.159
   ROUGE-2 F-measure           0.016
   ROUGE-L F-measure           0.123
6. BERTScore Precision         0.783
   BERTScore Recall            0.777
   BERTScore F1                0.780
======================================================================
```

Figure 21: Quantitative Performance Metrics for the AB Persona Adapter

```
======================================================================
NSF RESULTS
======================================================================

METRIC                         SCORE
----------------------------------------------------------------------
1. Perplexity                  17.74
2. Coherence                   0.371
3. Diversity                   0.880
4. Distinct-N (Library)        0.978
5. ROUGE-1 F-measure           0.102
   ROUGE-2 F-measure           0.012
   ROUGE-L F-measure           0.089
6. BERTScore Precision         0.749
   BERTScore Recall            0.761
   BERTScore F1                0.755
======================================================================
```

Figure 22: Quantitative Performance Metrics for the NSF Persona Adapter

```
======================================================================
XMM RESULTS
======================================================================

METRIC                         SCORE
----------------------------------------------------------------------
1. Perplexity                  46.36
2. Coherence                   0.311
3. Diversity                   0.827
4. Distinct-N (Library)        0.958
5. ROUGE-1 F-measure           0.118
   ROUGE-2 F-measure           0.008
   ROUGE-L F-measure           0.095
6. BERTScore Precision         0.759
   BERTScore Recall            0.746
   BERTScore F1                0.752
======================================================================
```

Figure 23: Quantitative Performance Metrics for the XMM Persona Adapter

Appendix C System Architecture

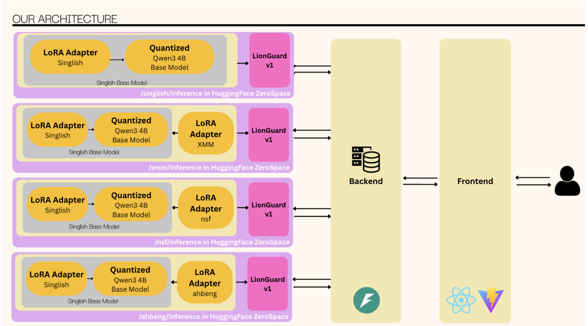

Figure 24: Final system architecture of chatbot

Appendix D Model Ablation Studies

Model	Train Time (min)	Train VRAM (GB)	Final Loss	Perplexity	Semantic Sim	Singlish Score	Latency (s)	Tokens/sec	Infer VRAM (GB)
1.7B	**2.02**	**1.80**	3.6624	88.14	0.4981	1.15	1.965	14.3	**12.36**
4B-Instruct	2.22	5.40	**2.8593**	**40.76**	0.5754	1.30	**1.841**	**11.0**	**12.36**
8B	2.69	11.66	2.8901	42.05	**0.6187**	**1.40**	2.358	**11.0**	14.51

Table 1: Results of Base Model Adapters Trained on the Singlish Dataset

Model	Perplexity	Semantic Similarity	SEM Min	SEM Std	Singlish Score	Coverage %	Latency (s)	Tokens/sec	Infer VRAM (GB)
4B-singlish-base-v1	**16.26**	0.5373	0.1611	0.1639	**1.80**	**85.0**	1.761	**10.2**	**12.16**
4B-singlish-base-v2	33.43	0.5582	**0.2755**	**0.1546**	1.55	80.0	1.405	17.9	18.14
4B-singlish-base-v3	32.70	**0.5855**	0.2243	0.1741	1.25	**85.0**	**1.326**	18.0	18.14

Table 2: Performance Metrics for 4B Singlish Base Model Variants

Appendix E Project Contributions

Name	Contributions
Austin Isaac	Singlish dataset generation, base model pre-training, report writing, main presenter
Jithin Bathula	Dataset transcription, persona adapter fine-tuning, report writing
Ng Yu Hueng	Base model pre-training, chatbot development and deployment, report writing
Siew Rui Ze Zayne	Dataset transcription, dataset generation, report writing
Tang Zhi-Ju Edward	Dataset transcription, persona adapter fine-tuning, report writing

Table 3: Contributions table