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

Austin Isaac, Jithin Bathula, Ng Yu Hueng, Tang Zhi-Ju Edward, Siew Rui Ze Zayne
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 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.

2.2 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.

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

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 developed Dataset v1.5 by sourcing audio from Singaporean pop-culture media, including YouTube channels 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 Dataset v1.5 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 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 (SUTD, n.d.). 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[2].

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[3]. 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

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

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

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 Section 6.1). 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[4], forming the foundation for persona adaptations.

4.3 Multi-Persona Architecture (Adapter Layer)

A modular Adapter Layer architecture enabled lightweight multi-persona deployment. Three LoRA adapters (NSF[5], AB[6], XMM[7]) were trained on 200 samples each, demonstrating that distinct archetypal personalities can be induced efficiently once the dialect base is established. Model evaluation involved quantitative metrics - Perplexity, Semantic Similarity embeddings, and a custom Singlish Score measuring particle usage and grammaticality - with Human-in-the-Loop qualitative assessment, ensuring outputs were both statistically robust and culturally authentic.

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.

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[8] and the NSF[9], Ah Beng[10], and XMM[11] 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 1).

5 Experiments and Evaluation

5.1 Persona Evaluation

To evaluate whether each fine-tuned model successfully embodied its intended persona, we conducted a structured human evaluation study guided by the core question: "Does this model feel like the intended persona when a

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

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

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

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

[8]https://huggingface.co/spaces/yuhueng/SinglishTest

[9]https://huggingface.co/spaces/yuhueng/nsf-persona/

[10]https://huggingface.co/spaces/yuhueng/ahbeng-persona/

[11]https://huggingface.co/spaces/yuhueng/xmm-persona

real person interacts with it?"

5.1.1 Methodology

Participants were presented with fixed conversation screenshots for each persona (XMM, AB, and NSF), ensuring that all raters assessed identical content under controlled conditions. This approach mitigated variability in model generation and enabled fair comparisons across personas. Respondents rated each persona on five criteria: persona voice accuracy, persona consistency, linguistic style fit, character behaviour alignment, and persona believability.

- **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., Ah Beng 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.

Responses were recorded using a 5-point Likert scale and averaged across three scenarios: family pressure about marriage, a simple factual question with follow-up, and user flirting with the assistant. In all scenarios, user turns were fixed in standard or near-standard English, while assistant responses differed according to persona, allowing raters to evaluate the models' adherence to persona-specific behaviour.

5.1.2 Survey Results

The results, summarised in Figures 2-6, indicate that the NSF persona performed most consistently across all five evaluation criteria, achieving the highest scores in voice accuracy, believability, linguistic style fit, persona consistency, and behaviour alignment across scenarios. The Ah Beng persona performed well in voice accuracy and style fit but exhibited slight inconsistency during the flirting scenario. The XMM persona demonstrated moderate performance, aligning with the base model in some areas but showing weaknesses in emotionally complex or factual scenarios, particularly in behaviour alignment and believability. Overall, all persona-tuned models improved upon the base model in targeted areas, with NSF representing the most stable and robust persona embodiment, followed by Ah Beng, while XMM remained the most sensitive to scenario context.

6 Discussion and Analysis

6.1 Comparison Across Model Sizes

To determine the optimal foundation for the Singlish ecosystem, we conducted a comprehensive ablation study across three Qwen3 parameter sizes: 1.7B, 4B (Instruct), and 8B. All models were trained 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 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.

6.2 Comparison of 4B Variants

We further evaluated three variants of the 4B model (v1, v2, v3) across multiple metrics, as summarised in Table 2. Model v1 achieved the lowest perplexity (16.26), highest Singlish Score (1.80), 85% coverage, and zero overshooting of particles, though with relatively higher latency (1.761s) and lower tokens/sec (10.2). Model v3 preserved meaning most effectively (Semantic Similarity 0.5855) and was fastest in practice (latency 1.326s, 18 tokens/sec) but exhibited a milder Singlish style (score 1.25). Model v2 was dominated by the other variants, with higher perplexity (33.43), lower style fidelity, and increased VRAM usage without outperforming in semantics or speed. Overall, v1 is recommended as the default Singlish model for stylistic accuracy and deployment efficiency, while v3 may be preferred when prioritizing inference speed and semantic fidelity with lighter dialectal expression.

6.3 Persona Evaluation: Limitations in Believability and Linguistic Style

Human evaluations of the fine-tuned personas revealed two consistent weaknesses: low believability and only partial alignment with their intended linguistic styles. Although the personas (NSF, Ah Beng, XMM) used Singlish markers and stayed broadly on topic, raters felt they still resembled a generic assistant rather than distinct characters. Their speech lacked the spontaneity, inconsistency, and human-like agency expected of real Singlish speakers.

Three factors contributed to this. Safety and helpfulness alignment discouraged more impulsive or edgy behaviour, limiting natural informality. Sharing the same underlying model reduced character differentiation, with similar phrasing and safety-driven hedging across personas. Pre-training on mostly standard English further restricted persona expression: despite Singlish fine-tuning, the model struggled to display deeper traits like taking sides or joking in a human-like way.

Linguistic style alignment showed similar limitations. The model often produced a sanitised, formulaic version of Singlish, relying on predictable particles and overly complete sentences. Features common in natural Singlish, such as code-switching, ellipsis, and mild disfluency, were rare. These constraints stem from the dominant influence of clean, formal English in the pre-training and instruction-tuning data, which shapes sentence structure even when colloquial markers appear. As a result, both style fit and persona believability remained limited.

7 Conclusion

This project successfully developed an end-to-end Singlish LLM Ecosystem with modular multi-persona adapters, addressing the lack of culturally grounded conversational AI for multilingual environments. By combining supervised fine-tuning, lightweight LoRA adapters, and a culturally aware safety layer, the system demonstrated effective style transfer for Singaporean personas and reliable moderation of local colloquialisms. The integration of a full-stack deployment pipeline (React/Vite frontend with FastAPI routing) enabled secure, scalable, and low-latency interaction. Overall, the work establishes a practical framework for developing culturally aligned language models in low-resource dialects.

8 Future Work

1. Expand the range and depth of Singaporean personas.
2. Enable user-defined persona customisation.
3. Add contextual memory to maintain dialogue continuity across personas.

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.

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.

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

A Appendix

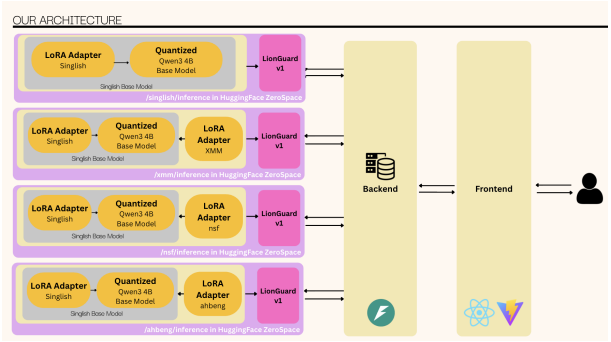

Figure 1: System architecture of chatbot

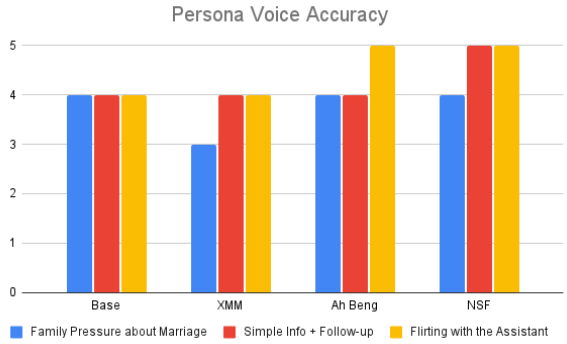

Figure 2: Persona voice accuracy per adapter across various scenarios

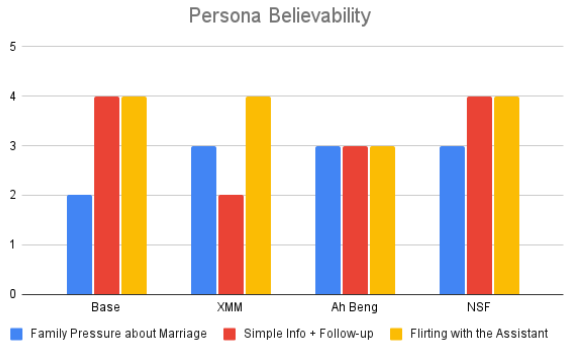

Figure 3: Persona Believability per adapter across various scenarios

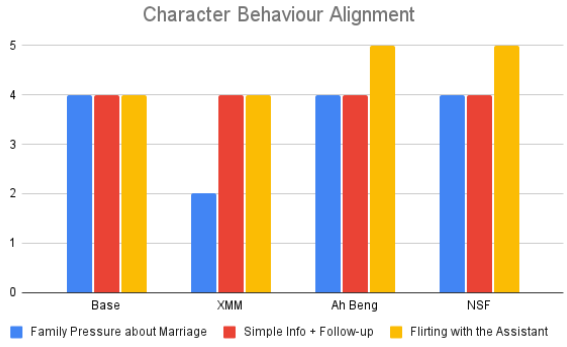

Figure 4: Character behaviour alignment per adapter across various scenarios

Model	Train Time (min)	Train VRAM (GB)	Final Loss	Perplexity
1.7B	2.02	1.80	3.6624	88.14
4B-Instruct	2.22	5.40	2.8593	40.76
8B	2.69	11.66	2.8901	42.05

Model	Semantic Sim	Singlish Score	Latency (s)	Tokens/sec	Infer VRAM (GB)
1.7B	0.4981	1.15	1.965	14.3	12.36
4B-Instruct	0.5754	1.30	1.841	11.0	12.36
8B	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 Sim	Sem Min	Sem Std	Singlish Score
4B-singlish-base-v1	16.26	0.5373	0.1611	0.1639	1.80
4B-singlish-base-v2	33.43	0.5582	0.2755	0.1546	1.55
4B-singlish-base-v3	32.70	0.5855	0.2243	0.1741	1.25

Model	Coverage %	Latency (s)	Tokens/sec	Infer VRAM (GB)
4B-singlish-base-v1	85.0	1.761	10.2	12.16
4B-singlish-base-v2	80.0	1.405	17.9	18.14
4B-singlish-base-v3	85.0	1.326	18.0	18.14

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

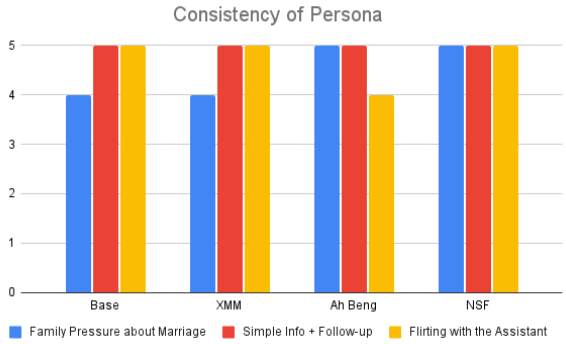

Figure 5: Persona consistency per adapter across various scenarios

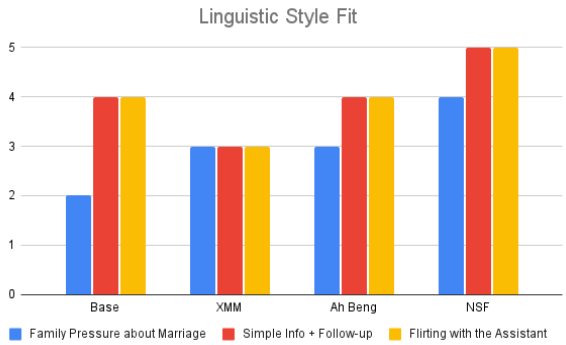

Figure 6: Linguistic style fit per adapter across various scenarios