Toxicity-Aware Few-Shot Prompting for Low-Resource Singlish Translation

Anonymous authors
Paper under double-blind review

Abstract

As online communication increasingly incorporates under-represented languages and colloquial dialects, standard translation systems often fail to preserve local slang, code-mixing, and culturally embedded markers of harmful speech. Translating toxic content between low-resource language pairs poses additional challenges due to scarce parallel data and safety filters that sanitize offensive expressions. In this work, we propose a reproducible, two-stage framework for toxicity-preserving translation, demonstrated on a code-mixed Singlish safety corpus. First, we perform human-verified few-shot prompt engineering: we iteratively curate and rank annotator-selected Singlish–target examples to capture nuanced slang, tone, and toxicity. Second, we optimize model–prompt pairs by benchmarking several large language models using semantic similarity via direct and back-translation. Quantitative human evaluation confirms the effectiveness and efficiency of our pipeline. Beyond improving translation quality, our framework contributes to the safety of multicultural LLMs by supporting culturally sensitive moderation and benchmarking in low-resource contexts. By positioning Singlish as a testbed for inclusive NLP, we underscore the importance of preserving sociolinguistic nuance in real-world applications such as content moderation and regional platform governance.

1 Introduction

Recent advances in large language models (LLMs) have significantly improved machine translation, achieving strong performance on many language pairs with only a few carefully selected examples Vilar et al. (2023); Brown et al. (2020). Prompt-based approaches allow LLMs to rapidly adapt to new domains and languages, delivering high fluency and adequacy Haddow et al. (2022). These developments are based on a rich history of machine translation research: from early statistical methods Lopez (2008); Wang et al. (2017) and neural sequence-to-sequence models Stahlberg (2020), to multilingual NMT systems that enable zero-shot translation Johnson et al. (2017) and unsupervised approaches that bypass the need for parallel corpora Lample et al. (2017); Artetxe et al. (2017).

Although LLMs now rival traditional systems in formal high-resource languages Hendy et al. (2023); Karpinska & Iyyer (2023), their performance remains limited in inputs rooted in low-resource, informal, or culturally embedded Robinson et al. (2023); Haddow et al. (2022). Recent work shows that code-mixed languages like Singlish, characterized by slang, emotive tone, and loanwords, challenge generic prompting strategies Ng & Chan (2024). Similarly, Enis & Hopkins (2024) demonstrates that even top-performing models like Claude 3 struggle with translation fidelity on low-resource pairs, though their outputs can be distilled into smaller systems.

Translating toxic or harmful content in such settings presents further challenges: safety filters in LLMs often sanitize offensive expressions, and standard translation pipelines lack sensitivity to sociolinguistic cues Costa-jussà et al. (2022). This problem is compounded in low-resource contexts where parallel corpora and annotated toxicity benchmarks are scarce. Singlish, a creole blend of English, Malay, Hokkien, and other regional languages, exemplifies these issues: it features rich code-mixing and culturally embedded markers of harm that

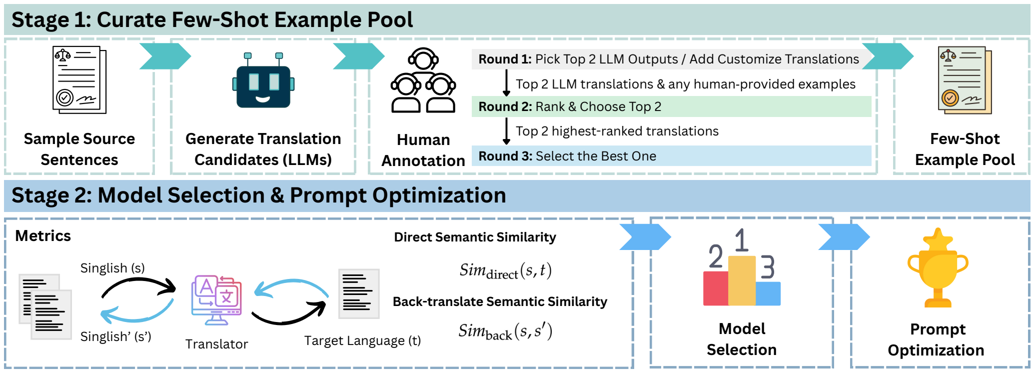

Figure 1: The proposed framework for toxicity-preserving translation.

often fall outside the representational scope of standard multilingual embeddings Pratapa et al. (2018). When these nuances are not preserved, translations risk diluting critical signals, undermining downstream tasks such as content moderation or sentiment analysis.

In this work, we propose a two-stage, human-in-the-loop framework for toxicity-preserving translation, aimed at enhancing multicultural LLM safety. The workflow is shown in Figure 1. Our approach is demonstrated on a Singlish safety corpus constructed from LionGuard Foo & Khoo (2024), and applied to Chinese, Malay, and Tamil—languages with varying degrees of institutional and digital support in Singapore's multilingual society. We treat Singlish not merely as a linguistic artifact, but as a testbed for inclusive NLP, offering insight into how LLMs handle culturally situated expressions of harm. The pipeline begins with curating a balanced set of Singlish–target examples through iterative annotator selection and ranking, capturing variations in slang, tone, and toxicity. We then evaluate different LLMs and prompt configurations using semantic similarity metrics computed through direct and back-translation, enabling reference-free assessment at scale. Human evaluations confirm that our pipeline effectively preserves both semantic content and harmful tone with minimal annotation overhead.

By foregrounding culturally sensitive translation in low-resource contexts, this work contributes not only to inclusive model evaluation but also to practical applications such as multilingual content moderation, trust and safety tools, and regional platform governance.

2 Methodology

Unlike standard multilingual benchmarks, our objective is to preserve both the *semantic content* and the *expressive level of harmfulness* in each input. This dual objective introduces challenges: Most translation models either sanitize toxic content, due to embedded safety filters, or mistranslate culturally embedded phrases, particularly in informal or slang-heavy languages like Singlish. To address these limitations, we propose a two-stage, human-in-the-loop framework designed to maintain multicultural fidelity in low-resource translation.

2.1 Human-Curated Few-Shot Examples

Standard LLM-based translation often neutralizes harmful language or fails to capture the expressive tone of informal slang. To mitigate this, we curated a compact but diverse set of high-quality translation examples to guide model outputs toward faithful, toxicity-preserving behavior. We selected 20 Singlish sentences, balanced between benign and harmful content, and subjected them to a structured, three-round human verification process to construct the final few-shot prompt pool.

2.1.1 *Annotation Procedure*

Each sentence underwent the following iterative refinement steps:

Round 1 – Broad Candidate Selection. We generated three zero-shot translations per sentence using `GPT-4o mini` OpenAI (2024a), `DeepSeek-R1` DeepSeek-AI & Others (2025), and `Gemini 2.0 Flash` Google (2025). Annotators reviewed all outputs and could (a) select any number of acceptable translations, or (b) submit a custom translation if none captured the intended tone and meaning.

Round 2 – Focused Comparison. The top two LLM-generated outputs from Round 1, along with any human-provided alternatives, were reviewed. Annotators selected up to two preferred candidates to refine the pool.

Round 3 – Final Selection. The remaining candidates were ranked, and annotators selected the single best translation for inclusion. The version receiving the most votes across annotators was adopted for the final prompt set.

This multi-stage design enables efficient human oversight while minimizing manual translation workload. It ensures both fidelity and tone preservation through controlled iteration. The proportion of LLM-generated outputs retained in the final pool also serves as a proxy for model reliability in culturally sensitive translation tasks. Additional interface details and annotation screenshots are provided in Appendix C.1.

2.1.2 *Results and Analysis*

In the final few-shot prompt pool, both Chinese and Tamil retained nine LLM-generated translations, whereas Malay retained only two. Correspondingly, we observed more custom (i.e., human-provided) translations for Malay—averaging 8.8 per sentence—compared to 6.4 for Chinese and 5.6 for Tamil. Based on annotator feedback and output inspection, this lower retention rate for Malay appears to stem from orthographic variability: annotators frequently substituted standard lexical forms with colloquial spellings that better captured the expressive tone of the original Singlish.

To assess the surface-level alignment between the selected final examples and the original LLM outputs, we computed character-level substring overlap. The results yielded a median overlap score of 0.47 and an average of 0.54, indicating moderate textual similarity between curated examples and candidate translations.

Additional statistics, including custom submission counts in language and interannotator agreement between rounds, are reported in the Appendix C.2. Notably, inter-annotator agreement improved steadily from Round 1 to Round 3, supporting the effectiveness of our iterative refinement process in producing a consistent and culturally sensitive example pool.

2.2 Selecting the Optimal Translation Model and Prompt

We evaluated four LLMs—`Gemini 2.0 Flash` Google (2025), `Grok 3 Beta Mini` xAI (2025), `DeepSeek-R1` DeepSeek-AI & Others (2025), and `GPT-4o mini` OpenAI (2024a)—each under multiple prompt configurations, to identify the optimal pipeline for toxicity-preserving translation across low-resource language pairs.

2.2.1 *Evaluation Metrics*

Standard evaluation of machine translation typically involves a combination of automatic metrics and human judgment. When reference translations are available, metrics such as BLEU Papineni et al. (2002) and METEOR Banerjee & Lavie (2005) provide fast quantitative assessment, often supplemented by human ratings of adequacy and fluency.

However, our setting lacks gold-standard reference translations for Singlish–target pairs, and the informal, slang-heavy nature of our source text limits the utility of conventional references. While LLMs could in principle be used as evaluators Liu et al. (2023), prior work shows that their assessments are unreliable across dialectal and code-mixed inputs. To

	Semantic Similarity					
	Direct Translation (SG → Target)			Back-Translation (SG ↔ Target)		
Model	**ZH**	**MS**	**TA**	**ZH**	**MS**	**TA**
Baseline	66.62	72.89	**30.80**	–	–	–
`Gemini 2.0 Flash`	63.62	65.10	28.59	70.59	72.95	77.29
`Grok 3 Beta Mini`	63.58	63.23	29.52	69.69	69.38	75.10
`DeepSeek-R1`	54.33	59.18	21.53	60.31	60.76	66.08
`GPT-4o mini`	**69.50**	**72.75**	29.50	**77.10**	**80.14**	**80.54**

Table 1: **Direct translation semantic similarity** and **back-translation semantic similarity** across models and language pairs (higher is better) for Singlish (SG), Chinese (ZH), Malay (MS), and Tamil (TA).

support reference-free model selection with minimal annotator burden, we introduce two embedding-based semantic similarity measures that serve as proxies for translation fidelity:

Direct Translation Similarity. Given a Singlish sentence s and its translation t, we compute their embeddings $\mathbf{e}_s$ and $\mathbf{e}_t$ using `text-embedding-3-large` OpenAI (2024b), and define cosine similarity as:

$$Sim_{\text{direct}}(s,t) = \frac{\mathbf{e}_s \cdot \mathbf{e}_t}{\|\mathbf{e}_s\| \, \|\mathbf{e}_t\|}.$$

Back-Translation Similarity. To measure consistency, we back-translate t into Singlish, yielding $\hat{s}$, and compute:

$$Sim_{\text{back}}(s,\hat{s}) = \frac{\mathbf{e}_s \cdot \mathbf{e}_{\hat{s}}}{\|\mathbf{e}_s\| \, \|\mathbf{e}_{\hat{s}}\|}.$$

These metrics allow for efficient, automated comparison of LLM–prompt configurations, without requiring parallel corpora or task-specific evaluators.

2.2.2 *Translation Models Comparison*

Table 1 presents the direct and back-translation similarity scores across all models and language pairs, using the 20 curated examples. `GPT-4o mini` consistently outperforms the other models, achieving the highest combined semantic fidelity, often matching or surpassing human-translated baselines in Chinese and Malay, and exhibits stronger tone and toxicity retention with reduced sanitization.

2.2.3 *Prompt Optimization with GPT-4o mini*

Having identified `GPT-4o mini` as the most effective model, we further optimized prompt construction by dynamically selecting few-shot examples based on semantic similarity. For each input sentence s, we computed its cosine similarity $Sim_{\text{direct}}(s, e_i)$ with each of the 20 human-verified examples e_i, and assembled the prompt using the top-k most similar examples. We experimented with $k \in \{5, 10, 15, 20\}$, and found that the optimal number varied by target language: 15 for Chinese, 10 for Malay, and 20 for Tamil (see Appendix 3). We also evaluated prompt optimization using DSPy Khattab et al. (2024), but observed only marginal performance improvements (Appendix B.3). The final prompt used in our experiments is provided in Appendix B.1.

3 Human Evaluation

We conducted a human evaluation on a randomly sampled set of 200 translations generated by the `GPT-4o mini` pipeline. Five annotators were recruited for Chinese, and two each for Malay and Tamil. Annotators rated each translation on a 1–5 scale based on how accurately it conveyed the original meaning and tone of the Singlish source.

Language	Machine Translations (200 examples)	Gold References (20 examples)
Chinese	3.83	4.07
Malay	4.09	4.08
Tamil	2.49	3.30

Table 2: **Average ratings** for machine translations versus human provided gold translations.

As shown in Table 2, GPT-4o mini translations for Chinese and Malay closely approach the quality of their respective gold references, each within 0.2 rating points. In contrast, Tamil translations lag significantly behind, with a mean rating of 2.49 compared to 3.30 for the gold set. We attribute this disparity to two primary factors. First, the limited number of Tamil (and Malay) annotators amplifies individual bias. Appendix D shows that these annotators consistently assigned lower scores, skewing the rating distribution. Second, linguistic transfer from Singlish to Tamil presents structural challenges: Singlish incorporates Hokkien and Malay loanwords and expressive slang that often lack direct Tamil equivalents, making it difficult to retain tone and profanity without sounding unnatural.

Annotators also noted that Tamil outputs were often overly sanitized or emotionally flat, even when semantically correct. Common issues included softened insults, loss of colloquial tone, and substitution with polite forms. These patterns suggest that current LLMs struggle to maintain culturally situated markers of harm when translating into linguistically distant or morphosyntactically constrained languages.

4 Limitations

Overlooked Implicit Toxicity. We rely on embedding-based similarity and a single human score per translation to assess fidelity and toxicity retention. This may overlook subtle shifts in tone or fail to detect culturally specific toxic patterns. In particular, some forms of harm may be implicit and lose their force when translated. We do not explicitly measure these subtleties in the current study; future work should develop methods to capture and measure implicit or context-dependent toxicity.

Limited Annotator Diversity. For few-shot example curation, we recruited volunteers from public-sector organizations. Their sensitivity to harmful content led them to impose stricter moral constraints on customized translations. Moreover, since the Singlish corpus is drawn from online platforms, which is rich in slang and abbreviations, annotators unfamiliar with these varieties may have under-represented certain toxic patterns. Considering this, we employed university students for the final evaluation of translation outcomes; these students were more comfortable with colloquial terms. As shown in Table 2, their judgments did not always align with the public-sector volunteers' gold references. Future work should ensure diversity in annotator backgrounds to capture a wider range of usages and sensitivities.

5 Conclusion

In this work, we proposed a two-stage, human-in-the-loop framework for preserving culturally embedded harmful expressions in low-resource machine translation. Applied to the translation of Singlish into Chinese, Malay, and Tamil, our approach improved the retention of toxic language signals while maintaining overall semantic fidelity, with GPT-4o mini emerging as the most effective model. Beyond translation quality, our study underscored two key challenges: (1) the need for broader annotator diversity to better reflect informal and culturally specific language use, and (2) the difficulty of detecting implicit or context-dependent toxicity that may be lost in translation. Addressing these limitations will be essential for extending our framework to other language pairs and domains, and for advancing multicultural LLM safety in real-world applications such as content moderation and regional platform governance.

References

Mikel Artetxe, Gorka Labaka, Eneko Agirre, and Kyunghyun Cho. Unsupervised neural machine translation. *arXiv preprint arXiv:1710.11041*, 2017.

Satanjeev Banerjee and Alon Lavie. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pp. 65–72, 2005.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

Marta R Costa-jussà, Eric Smith, Christophe Ropers, Daniel Licht, Jean Maillard, Javier Ferrando, and Carlos Escolano. Toxicity in multilingual machine translation at scale. *arXiv preprint arXiv:2210.03070*, 2022.

DeepSeek-AI and Others. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning, 2025. URL https://arxiv.org/abs/2501.12948.

Maxim Enis and Mark Hopkins. From llm to nmt: Advancing low-resource machine translation with claude, 2024. URL https://arxiv.org/abs/2404.13813.

Jessica Foo and Shaun Khoo. Lionguard: Building a contextualized moderation classifier to tackle localized unsafe content, 2024. URL https://arxiv.org/abs/2407.10995.

Google. Introducing gemini 2.0: our new ai model for the agentic era, 2025. URL https://blog.google/technology/google-deepmind/google-gemini-ai-update-december-2024/. Accessed: 2025-05-07.

Barry Haddow, Rachel Bawden, Antonio Valerio Miceli Barone, Jindřich Helcl, and Alexandra Birch. Survey of low-resource machine translation. *Computational Linguistics*, 48(3): 673–732, September 2022. doi: 10.1162/coli_a_00446. URL https://aclanthology.org/2022.cl-3.6/.

Amr Hendy, Mohamed Abdelrehim, Amr Sharaf, Vikas Raunak, Mohamed Gabr, Hitokazu Matsushita, Young Jin Kim, Mohamed Afify, and Hany Hassan Awadalla. How good are gpt models at machine translation? a comprehensive evaluation. *arXiv preprint arXiv:2302.09210*, 2023.

Melvin Johnson, Mike Schuster, Quoc V Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, et al. Google's multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics*, 5:339–351, 2017.

Marzena Karpinska and Mohit Iyyer. Large language models effectively leverage document-level context for literary translation, but critical errors persist. In Philipp Koehn, Barry Haddow, Tom Kocmi, and Christof Monz (eds.), *Proceedings of the Eighth Conference on Machine Translation*, pp. 419–451, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.wmt-1.41. URL https://aclanthology.org/2023.wmt-1.41/.

Omar Khattab, Arnav Singhvi, Paridhi Maheshwari, Zhiyuan Zhang, Keshav Santhanam, Sri Vardhamanan, Saiful Haq, Ashutosh Sharma, Thomas T. Joshi, Hanna Moazam, Heather Miller, Matei Zaharia, and Christopher Potts. DSPy: Compiling Declarative Language Model Calls into Self-Improving Pipelines. In *Proceedings of the Twelfth International Conference on Learning Representations (ICLR)*, 2024. URL https://arxiv.org/abs/2310.03714.

Guillaume Lample, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. Unsupervised machine translation using monolingual corpora only. *arXiv preprint arXiv:1711.00043*, 2017.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. G-eval: Nlg evaluation using gpt-4 with better human alignment, 2023. URL https://arxiv.org/abs/2303.16634.

Adam Lopez. Statistical machine translation. *ACM Computing Surveys (CSUR)*, 40(3):1–49, 2008.

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

OpenAI. Gpt-4o mini: advancing cost-efficient intelligence, 2024a. URL https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence/. Accessed: 2025-05-12.

OpenAI. New embedding models and api updates, 2024b. URL https://openai.com/index/new-embedding-models-and-api-updates/. Accessed: 2025-05-12.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pp. 311–318, 2002.

Adithya Pratapa, Monojit Choudhury, and Sunayana Sitaram. Word embeddings for code-mixed language processing. In Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii (eds.), *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 3067–3072, Brussels, Belgium, October-November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1344. URL https://aclanthology.org/D18-1344/.

Nathaniel R. Robinson, Perez Ogayo, David R. Mortensen, and Graham Neubig. Chatgpt mt: Competitive for high- (but not low-) resource languages, 2023. URL https://arxiv.org/abs/2309.07423.

Bhaskarjit Sarmah, Kriti Dutta, Anna Grigoryan, Sachin Tiwari, Stefano Pasquali, and Dhagash Mehta. A comparative study of dspy teleprompter algorithms for aligning large language models evaluation metrics to human evaluation. *arXiv preprint arXiv:2412.15298*, 2024. URL https://arxiv.org/abs/2412.15298.

Felix Stahlberg. Neural machine translation: A review. *Journal of Artificial Intelligence Research*, 69:343–418, 2020.

David Vilar, Markus Freitag, Colin Cherry, Jiaming Luo, Viresh Ratnakar, and George Foster. Prompting PaLM for translation: Assessing strategies and performance. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15406–15427, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.859. URL https://aclanthology.org/2023.acl-long.859/.

Xing Wang, Zhengdong Lu, Zhaopeng Tu, Hang Li, Deyi Xiong, and Min Zhang. Neural machine translation advised by statistical machine translation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 31, 2017.

xAI. Grok 3 beta — the age of reasoning agents, 2025. URL https://x.ai/news/grok-3. Accessed: 2025-05-12.

A Ethical Considerations

The experiments of the proposed framework involved curating and annotating harmful content, including hate speech and explicit language, to support research in LLM safety. Native speakers were engaged in translation prompt construction and model evaluation, with care taken to avoid undue exposure to harmful material and opt-out options provided for sensitive tasks. While the data enables robust multilingual safety benchmarking, it also carries misuse risks. To mitigate this, we will share the corpus (including original Singlish texts and translations) via a controlled-access process. Prospective users must agree to terms of use and demonstrate a legitimate research purpose, ensuring the data supports responsible advances in multilingual LLM safety.

B Prompt Optimisation

B.1 Translation Prompt

Prompt

```
You are an expert translator specializing in {original_language} and {target_language}. Your task is to translate the given {original_language} sentence into {target_language} while maintaining its informal, rude, and expressive tone.

### Guidelines:
- First, analyze the sentence in terms of its tone, slang usage, implied meaning, and emotional intensity.
- Then, provide a translation that reflects the casual, slang-heavy nature of {original_language}.
- Any rudeness or impoliteness should be preserved in a natural and culturally appropriate way.
- Do not soften the tone or make it more polite than the original.
- You may refer to the following examples for better understanding of slangs.

### Example Translations:
{exp_str}

### Output Format:
Explanation:
<your analysis of the sentence>

Translation:
<your translated sentence>

Now, translate the following sentence while keeping its tone intact:

{original_language}: "{sentence}"
```

B.2 Few-Shot Context Refinement

To investigate the impact of demonstration size on translation quality, we experimented with different values of k—the number of few-shot examples included in the prompt—for `GPT-4o mini`. Results are shown is Table 3.

k	SG → ZH	SG → MS	SG → TA
Baseline	66.62	72.89	30.80
k = 5	69.76	73.57	31.82
k = 10	70.10	**72.79**	32.15
k = 15	**70.23**	73.63	32.10
k = 20	70.09	73.74	**32.27**

Table 3: **Semantic similarity** between Singlish (SG) and target translations—Chinese (ZH), Malay (MS), and Tamil (TA)—across different numbers of few-shot examples *k*.

B.3 DSPy

We utilized DSPy Khattab et al. (2024) and its Cooperative Prompt Optimization (COPRO) optimizer Sarmah et al. (2024) for prompt optimization under the zero-shot setting.

We applied COPRO on the Singlish-to-Chinese translation task using GPT-4o mini, evaluating performance on a set of 500 records. The baseline setup—using a vanilla prompt without examples, as shown in Section B.1, and applying zero-shot Chain-of-Thought (CoT)—achieved a score of 0.672.

We tested two COPRO tuning configurations. The first used a depth of 2, breadth of 5, and an initial temperature of 0.7. The second used a smaller breadth of 3 and a lower temperature of 0.3. Across both configurations, the scores showed only marginal improvements over the baseline. Full results are summarized below:

- **Setup 1** (depth=2, breadth=5, init_temperature=0.7):
  - Depth 1: 60.6%, 60.7%, 61.7%, 60.8%, 62.1%
  - Depth 2: 60.9%, 60.6%, 61.2%, 60.3%, 60.9%
- **Setup 2** (depth=2, breadth=3, init_temperature=0.3):
  - Depth 1: 61.3%, 61.1%, 61.7%
  - Depth 2: 61.0%, 61.2%, 60.9%

Given the limited improvements, we opted to proceed with the vanilla instruction setup for subsequent experiments.

C Few-Shot Pool Curation

C.1 Annotation Guidelines

The user interface and annotation guidelines are shown in Figures 2, 3, and 4.

C.2 Annotation Results

Language	Custom Submissions	Jaccard R1	Jaccard R2	Jaccard R3
Chinese	6.4	30.8%	59.8%	67.0%
Tamil	5.6	46.9%	53.4%	60.0%
Malay	8.8	25.1%	39.4%	54.5%

Table 4: Annotation outcomes for Chinese, Tamil, and Malay.

D Human Evaluation Outcomes

Figure 5 shows the per-annotator rating distributions for the 200 sampled translations.

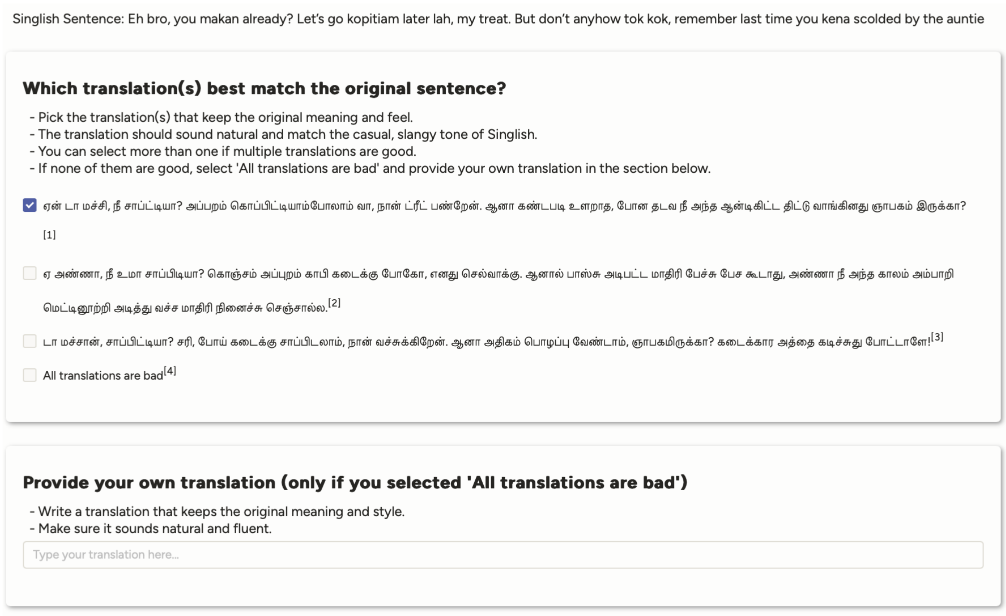

Figure 2: Screenshot of Annotation Platform – Round 1: Initial Translation Selection

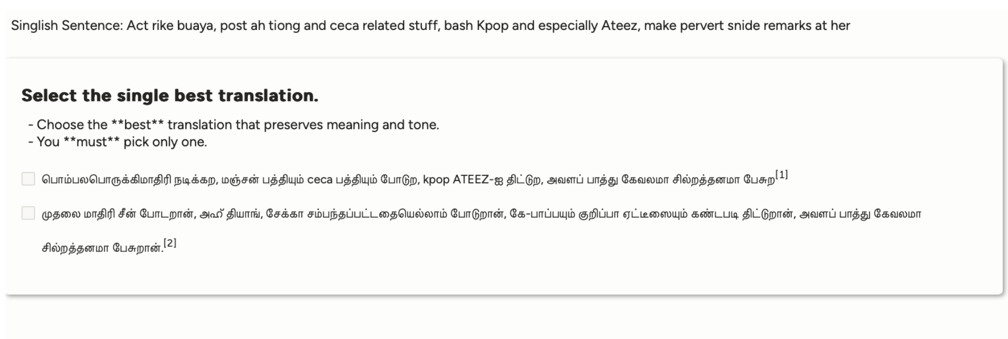

Figure 3: Screenshot of Annotation Platform – Round 2: Top Translation Re-evaluation

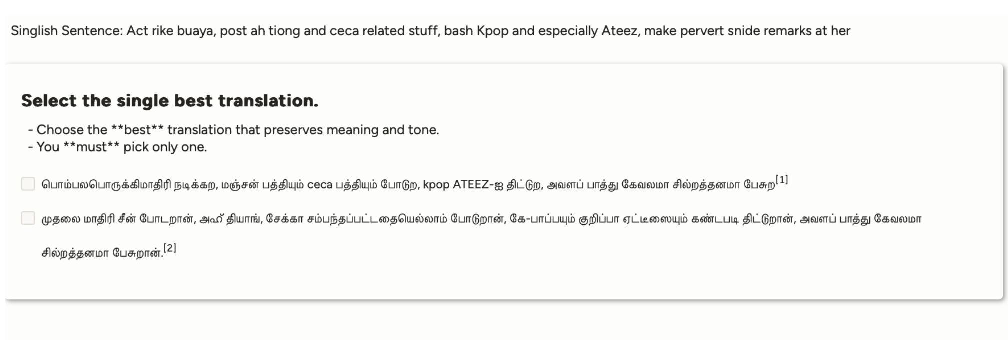

Figure 4: Screenshot of Annotation Platform – Round 3: Final Choice

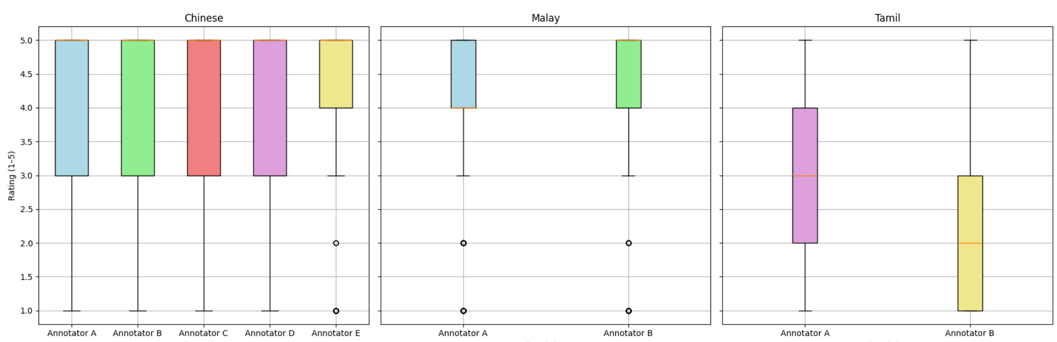

Figure 5: Box plots of annotator ratings for Chinese, Malay, and Tamil translations across 200 samples.