Fantastic Taiwanese Hokkien Corpora and Where to Find Them: Bootstrapping Low-Resource Speech and Text via Off-the-Shelf LLMs and TTS

Yu-Kai Guo, Chen-Wei Hsu, Hung-Wei Chen, Hung-Yi Lee
Department of Electrical Engineering
National Taiwan University
Taipei, Taiwan
b13901084@ntu.edu.tw

Abstract

Taiwanese Hokkien is a low-resource language with scarce, often license-restricted corpora, hindering commercial investment and technological development. This study proposes synthesizing Taiwanese text and speech corpora using off-the-shelf general-purpose LLMs and TTS models. We evaluate models' ability to translate between Mandarin and Taiwanese via a bidirectional translation task, validated by a human expert. The best-performing cloud model, Gemini 3.1 Pro Preview, achieves near dictionary-reference quality, with correctness scoring 4.8/5 for both Han-ji and Tâi-lô, and authenticity scoring 4.8 for Han-ji but only 4.45 for Tâi-lô. Fine-tuned local small models also achieve performance comparable to many larger cloud models. We further test LLM-based TTS for Taiwanese speech generation: audio from Gemini 3.1 Flash TTS, rated by 10 native speakers, scores 3.763, 3.857, and 3.777 MOS for Pronunciation Accuracy, Technical Naturalness, and Authenticity of Accent, while quality drops sharply when only Han-ji is provided, showing Tâi-lô is likewise indispensable for speech synthesis. In a preliminary, in-sample test, clips retained by a multimodal LLM filter received slightly higher human ratings (4.005 vs. 3.799 overall), although its item-level agreement with human judgments was weak. These results suggest this approach is a feasible, promising method for generating Taiwanese corpora.

Keywords: Taiwanese, Corpus synthesis, Large language models, Text-to-speech

1 Introduction

In recent years, speech processing technology for Taiwanese Hokkien (hereafter Taiwanese or Taigi)(Chang et al., 2026) has made some progress(Chou et al., 2023; Lan et al., 2026). However, Taiwanese remains a low-resource language, and publicly reusable corpora with high-quality Taiwanese annotations remain limited, while existing resources are frequently restricted by strict licensing terms such as ShareAlike and NonCommercial. These conditions can limit the direct commercial reuse and redistribution of such resources and may discourage related investment. Collecting, transcribing, and validating human-recorded speech corpora is typically expensive and resource-intensive. Recent studies have shown that LLM-generated text can improve low-resource machine translation, while pipelines combining LLM-based text generation and TTS can produce large-scale synthetic speech data at substantially lower cost than human recording(de Gibert et al., 2025).

In this work, we first evaluate different models' ability to translate Taiwanese, in both Han-ji and Tâi-lô, into Mandarin, and select the best-performing models to assess the quality of the reverse direction, translating Mandarin into Taiwanese, complemented by human expert evaluation to confirm the correctness and authenticity of the generated results. We further test publicly available LLM-based TTS models and find that they are capable of generating Taiwanese speech, and we report a preliminary observation that a multimodal LLM may help filter out lower-quality clips. Based on these results, we conclude that off-the-shelf general-purpose LLMs and TTS systems are capable of synthesizing usable text and speech corpora, offering a promising, lower-cost approach to addressing the low-resource bottleneck of Taiwanese without directly redistributing license-restricted source corpora.

Proceedings of the 38th Conference on Computational Linguistics and Speech Processing (ROCLING 2026), pages 95–115
November 6-7, 2026 ©2026 Association for Computational Linguistics

2 Background: Taiwanese Writing Systems

The development of Taiwanese writing has long followed two parallel tracks, Taiwanese characters and Romanization. These two systems have coexisted since Western missionaries introduced Romanization in the 19th century(Klöter, 2005). Taiwanese characters (台語漢字, also known as Han-ji) follow the Chinese character writing tradition and exhibit substantial orthographic overlap with Traditional Mandarin Chinese. The orthographic overlap with Mandarin can provide semantic cues, but Han-ji does not uniquely specify pronunciation: Taiwanese readings and tones may differ from Mandarin, and polyphonic characters can introduce ambiguity. Taiwanese Romanization (Tâi-lô), officially announced by the Ministry of Education in 2006(Ministry of Education, Taiwan, 2006), was refined from the earlier official scheme TLPA and from Pe̍h-ōe-jī (POJ), a Romanization system devised by missionaries in the 19th century. Tâi-lô explicitly encodes segmental and tonal information that is not fully represented in Han-ji, although readers must be familiar with the Romanization system. Since each system has its own strengths and some vocabulary still lacks broadly agreed-upon character forms, mixed usage is common in practice(Sung et al., 2025).

3 Text Corpus Synthesis: Methodology

3.1 Task Design

To compare models' ability to process Taiwanese characters and Romanization, we designed two stages: translating Taiwanese into Mandarin, then translating Mandarin into Taiwanese. The first direction was evaluated using a reference-based semantic similarity proxy: the model output and Mandarin reference were encoded as sentence embeddings and compared by cosine similarity(Reimers and Gurevych, 2019). The model that performed best in this direction was then reused in the second stage to back-translate its Mandarin-to-Taiwanese outputs(Zhuo et al., 2023), helping assess their quality, complemented by human expert evaluation.

We evaluate API-accessed proprietary models, locally deployed open-weight models, and the task-specific Taigi-Llama-2-Translator series(Lu et al., 2024), fine-tuned on Taiwanese corpora and available in 7B and 13B versions. It should be noted that both the Taigi-Llama-2-Translator model and the content it generates are governed by the CC-BY-NC-SA-4.0 license.

3.2 Dataset

In the Text Corpus Synthesis experiment, only the Ministry of Education's Taiwanese dictionary(中華民國教育部, 2026) was used to evaluate the model's translation performance. The website provides an export function, and we selected the "例句" (Example Sentences) data table. Each data entry contains Mandarin, Taiwanese characters, and Taiwanese Romanization, totaling 16,788 data entries and 175,768 Mandarin characters (excluding Taiwanese character fields). We randomly selected 400 data entries, totaling 4,749 Mandarin characters (excluding Taiwanese character fields), as data to test the translation quality of different models in the two stages.

3.3 Generation Process

To let cloud-based, local, and dedicated translation models all receive identical inputs, we developed a unified adapter that converts each input into the format required by the corresponding API or framework. All models were run with temperature = 0 (or the closest available greedy-decoding setting) to reduce sampling-related variation and improve repeatability.

Regarding the input content, we use a system prompt instructing the model to act as a bidirectional Taiwanese-Mandarin translation assistant and output only the translation itself, changing only the source and target content types (Taiwanese characters, Tâi-lô, or Mandarin written in Traditional Chinese characters) across tasks. The full prompt template is provided in Appendix A. For few-shot prompting, we additionally include real dataset examples excluded from the test set, provided as dialogue history rather than within the system prompt itself(Qin et al., 2025); the three fixed examples used are shown in Table 1.

Table 1: Taiwanese characters, Taiwanese Romanization and Mandarin translation few-shot samples

Taiwanese characters	Taiwanese Romanization	Mandarin
作田	tsoh-tshân	種田
你莫閣宕矣，咱坐火車欲袂赴矣。	Lí mài koh thōng–ah, lán tsē hué-tshia beh bē-hù–ah.	你別再拖延了，我們坐火車快要來不及了。
伊頂日仔才無去的。	I tíng-jit-á tsiah bô–khì–ê.	他前幾天才去世的。

3.4 Evaluation Metrics

For stage 1, each Mandarin translation is scored by encoding it and the reference with a pre-trained embedding model and computing their cosine similarity; the per-task score is the average over all samples. We use BAAI/bge-large-zh-v1.5(BAAI, 2023), a 326M-parameter Chinese text embedding model that produces 1024-dimensional vectors and achieved strong results on C-MTEB at release. Table 2 shows example scores this metric assigns to individual translations.

We are not aware of a Taiwanese semantic similarity model that has been validated for reference-based translation evaluation in this setting. Direct LLM-based evaluation may be unreliable in this setting because its judgments depend on the evaluator's Taiwanese proficiency, rubric design, and model-specific biases(Liu et al., 2023; Zheng et al., 2023; Fu and Liu, 2025). Stage 2 instead back-translates each output into Mandarin using the stronger model identified in stage 1, then scores it against the reference exactly as in stage 1, averaging across samples for the model's overall score. This approach, however, inherits the judge model's own translation ability and biases, which can cause error cancellation; we therefore complement it with human expert annotation on a subset(Freitag et al., 2021), and treat the back-translation scores with corresponding caution.

4 Text Corpus Synthesis: Results and Discussion

4.1 Experiment Stage 1

We obtained API keys from Google AI Studio to call Gemini 3.1 Pro Preview(Google DeepMind, 2026c), Gemini 3.5 Flash(Google DeepMind, 2026d), Gemini 3 Flash Preview(Google DeepMind, 2025), Gemini 3.1 Flash Lite(Google DeepMind, 2026b), Gemini 2.5 Pro, and Gemini 2.5 Flash(Comanici et al., 2025). We also called GPT-5.6 Luna, GPT-5.6 Terra(OpenAI, 2026), Grok 4.5(SpaceXAI, 2026), DeepSeek V4 Pro, DeepSeek V4 Flash(Xu et al., 2026), and Claude Sonnet 5(Anthropic, 2026) via the OpenRouter API.

For the local model portion, we used Ollama to run GPT-OSS 20B, GPT-OSS 120B(Agarwal et al., 2025), Nemotron-3 Nano(Blakeman et al., 2025), Nemotron-3 Super(Chandiramani et al., 2026), Qwen3.5-122B-A10B(Qwen Team, 2026a), Qwen3.6-27B(Qwen Team, 2026b), Qwen3.6-35B-A3B(Qwen Team, 2026c), Gemma4 26B, and Gemma4 31B(Gemma Team et al., 2026). Each model was run individually on a single H200 GPU for inference, using the unquantized version and running at the preset FP16 or BF16 precision. The Taigi-Llama-2-Translator also runs with default settings on a single H200 GPU.

The experimental results are shown in Table 3, reporting each model's default-setting score and success rate (SR) for Han-ji→Mandarin (HAN2ZH) and Tâi-lô→Mandarin (TL2ZH). Overall, when translating Taiwanese characters into Mandarin, the cloud-based models performed best, followed by the fine-tuned Llama-2 Translator model, while the local models performed slightly worse. However, when translating Taiwanese Romanization into Mandarin, performance declined across the board, with all local models showing a significant drop. Under the present dataset and evaluation metric, most tested models achieve higher scores on Han-ji-to-Mandarin than on Tâi-lô-to-Mandarin. One common reason for this, observed across all models, is that Taiwanese Romanization might correspond to Taiwanese characters with very similar or even identical pronunciations. For example, the solar term "小雪" (Xiao Xue) has the same

Table 2: Stage 1 Translation Quality Score Example

Mandarin translation results	Score	Remark
我兒子每次吃飯都要強逼，他才會好好的吃。	1.0000	Original sentence of the standard answer. It means, "My son only eats properly when I force him to."
我兒子每次吃飯都要強逼他，他才好好的吃。	0.9916	The meaning is exactly the same as above, only the wording is different.
我的小孩每次吃飯都要硬逼，他才肯好好吃。	0.9036	It means, "My child only eats properly when I force him to."
我們的孩子每次吃飯都要哄一下，他才會好好吃。	0.8032	It means, "Our child only eats properly when coaxed a little."
我們每次吃飯都要慢一點，他才能好好地吃。	0.6852	It means, "He only eats properly when we slow down a little."
我們看著一起吃飯，大家都想吃得好好，他只想好好地吃。	0.6066	It means, "We watch each other eat together; everyone wants to eat well, and he just wants to eat well too."

Taiwanese and Mandarin characters, thus requiring no additional translation. However, the corresponding Taiwanese Romanization "siáu-suat" can be simultaneously translated into " 小說" (novel). Han-ji can provide orthographic and semantic cues that help disambiguate forms whose Tâi-lô representations are phonetically similar, although Han-ji itself may also be pronunciation-ambiguous. Another reason why Taiwanese characters are easier to translate, and which appears more frequently in models with poor overall performance, is that Taiwanese and Mandarin characters share many overlapping characters and often have similar meanings. Therefore, it is possible to interpret Taiwanese characters from a Mandarin perspective. For example, when Nemotron-3 Nano completely distorts the Taiwanese Romanization "tsò-ping" into a seriously flawed Mandarin result, it can accurately translate the Taiwanese character " 做兵" into " 當兵" (enlistment) using the information from the character " 兵" (soldiers). In view of this, the similarity between the unprocessed Taiwanese characters and Mandarin was calculated directly, resulting in a similarity of 0.735. Therefore, the relevant data should be compared with this baseline.

Regarding the cloud-based models, considering the relatively small differences in HAN2ZH, the Gemini 3 series models and Claude Sonnet 5 performed better in TL2ZH, followed by the DeepSeek V4 series models. Output token usage across these models varies by orders of magnitude. In the Claude Sonnet 5 reasoning-effort ablation, output-token count is negatively associated with HAN2ZH score, but token count is only a rough proxy for reasoning effort and the association should not be interpreted causally; the full token-usage table and ablation are provided in Appendix B. Because the observed score changes under the reasoning-effort ablation were small relative to the between-model gaps, the main ranking conclusions were largely unchanged.

Within the tested model set, the task-specific Translator models achieve higher scores than substantially larger general-purpose models, while no clear monotonic relationship with total or active parameter count is observed. Within a series (e.g., GPT-OSS, Nemotron-3), however, larger models score higher on both tasks. For local models, thinking has a very small impact on the scores of the two tasks, and some models even score higher without thinking (see Appendix C for the full comparison).

4.2 Experiment Stage 2

All stage-1 models were re-run for stage 2, converting Mandarin into Taiwanese characters and Tâi-lô (Table 3). Following the method above, we back-translate stage-2 outputs to Mandarin using the two best low-cost stage-1 performers, Taigi-Llama-2-Translator 13B and Gemini 3 Flash Preview (Thinking), yielding two scores labeled Llama and Gemini. We also report Gemini's total token usage over the 400 samples as Cost.

Overall, cloud models perform best on ZH2HAN, with local models and Taigi-Llama-2-Translator comparable. On ZH2TL, results diverge by judge: scored via Taigi-Llama-2-Translator 13B, the Translator series outperforms most cloud models, but scored via Gemini 3 Flash Preview, cloud models clearly outperform the Translator; local models score worst under either judge. Given the error cancellation inherent in round-trip translation, judge-specific bias, and instability when back-translating already-flawed Taiwanese text, these scores should be read cau-

Table 3: Stage 1 and Stage 2 evaluation results under the default setting (provider-default thinking configuration): Han-ji↔Mandarin (HAN2ZH/ZH2HAN) and Tâi-lô↔Mandarin (TL2ZH/ZH2TL). ZH2HAN/ZH2TL scores are judged via back-translation (Llama: Taigi-Llama-2-Translator 13B; Gemini: Gemini 3 Flash Preview), with Cost denoting the Gemini judge's total output tokens (including reasoning) over 400 samples, which reflects how hard the outputs are for the judge to interpret rather than translation quality. Full thinking vs. non-thinking comparison in Appendix C.

	HAN2ZH		TL2ZH		ZH2HAN			ZH2TL		
Model	Score	SR (%)	Score	SR (%)	Llama	Gemini	Cost	Llama	Gemini	Cost
Cloud Models										
Gemini 3.1 Pro Preview	**0.905**	100.00	**0.895**	100.00	0.923	0.929	319k	**0.878**	0.944	192k
Gemini 3.5 Flash	0.890	99.50	0.890	100.00	0.930	0.940	247k	0.852	0.943	256k
Gemini 3 Flash Preview	0.887	100.00	0.880	99.50	0.928	0.933	312k	0.853	0.943	194k
Gemini 3.1 Flash Lite	0.886	100.00	0.842	100.00	0.931	0.925	325k	0.844	0.934	411k
Gemini 2.5 Pro	0.879	99.75	0.850	99.75	0.898	0.903	247k	0.846	0.920	324k
Gemini 2.5 Flash	0.882	99.75	0.826	100.00	0.905	0.911	327k	0.841	0.930	654k
GPT-5.6 Luna	0.890	100.00	0.822	100.00	0.929	0.929	322k	0.822	0.939	462k
GPT-5.6 Terra	0.888	100.00	0.836	100.00	0.929	0.934	205k	0.841	0.943	458k
Grok 4.5	0.889	100.00	0.837	99.50	**0.939**	**0.945**	387k	0.866	**0.955**	320k
DeepSeek V4 Pro	0.900	100.00	0.845	100.00	0.927	0.927	316k	0.873	0.940	454k
DeepSeek V4 Flash	0.888	99.75	0.843	99.50	0.917	0.921	197k	0.840	0.941	389k
Claude Sonnet 5	**0.905**	100.00	0.861	100.00	0.933	0.931	325k	0.839	0.941	335k
MSD	0.010		0.013		0.009	0.010		0.015	0.013	
Local Models										
GPT-OSS 20B (A3.6B)	0.782	100.00	0.465	100.00	0.915	0.917	615k	0.504	0.652	6731k
GPT-OSS 120B (A5.1B)	0.806	100.00	0.543	100.00	0.910	0.919	529k	0.656	0.846	2136k
Nemotron-3 Nano 30B (A3B)	0.751	99.75	0.417	94.00	0.873	0.882	900k	0.443	0.545	9119k
Nemotron-3 Super 120B (A12B)	0.801	99.75	0.505	99.75	0.893	0.895	607k	0.491	0.646	4820k
Qwen3.5-122B-A10B (A10B)	**0.858**	100.00	**0.725**	100.00	0.905	0.924	366k	**0.712**	0.854	2163k
Qwen3.6-27B (Dense)	0.849	100.00	0.649	98.50	0.912	0.916	285k	0.606	0.757	2580k
Qwen3.6-35B-A3B (A3B)	0.842	100.00	0.636	98.25	0.910	0.911	623k	0.629	0.758	3435k
Gemma4 26B (A4B)	0.838	99.25	0.613	74.00	**0.929**	**0.933**	347k	0.653	0.791	2972k
Gemma4 31B (Dense)	0.853	100.00	0.646	99.75	0.925	0.924	410k	0.684	**0.865**	1519k
MSD	0.013		0.020		0.011	0.013		0.022	0.024	
Taigi-Llama-2-Translator Models										
Translator 7B	0.879	100.00	0.824	100.00	**0.906**	**0.904**	333k	0.869	0.887	487k
Translator 13B	**0.894**	100.00	**0.845**	100.00	**0.906**	0.892	335k	**0.871**	**0.895**	720k
MSD	0.010		0.012		0.009	0.012		0.014	0.015	

Gemini 3.1 Flash Lite was run with thinking disabled (its API default); Taigi-Llama-2-Translator does not support thinking. HAN2ZH/TL2ZH failures (retried once, still failed) were due to safety-filter rejections (Gemini 3 series), API errors (other cloud models), or exceeding the 5-minute/16,384-token limit on a single H200 GPU (local models). ZH2HAN/ZH2TL success rate is omitted since the effective response rate exceeded 98% throughout. Bold: best within each group; underlined: not significantly different from the best (paired t-test, $\alpha = 0.05$). MSD: minimum significant difference within each group (median over within-group pairs); gaps below it should not be read as ranking differences. Definition and ranges in Appendix D.

tiously, useful mainly when gaps are large, and we do not draw further model-ranking conclusions from them(Zhuo et al., 2023; Zheng et al., 2023; Fu and Liu, 2025).

Under the Gemini judge, Cost is strongly and negatively correlated with the ZH2TL score (Spearman $\rho = -0.954$): the judge spends far more reasoning tokens when attempting to interpret poor-quality Tâi-lô, which explains the sharply higher Cost for local models. Cost therefore indicates how difficult the outputs are for the judge to decode, and also the practical expense of back-translation evaluation, but it is not itself a quality measure. In contrast, the correlation on ZH2HAN is only -0.466.

To evaluate the generated Taiwanese content more rigorously, we recruited an expert highly proficient in Taiwanese writing, since reading and writing Taiwanese is markedly harder than listening and speaking(中華民國文化部, 2022), to annotate a subset. From the 400 stage-2 samples, we selected 20 covering words, short sentences, and long sentences, as the original set skewed toward longer sentences. For each of the 20 Mandarin source sentences, the expert saw 10 model outputs plus the dataset reference in both Han-ji and Tâi-lô, with model identities hidden and order randomized for blind evaluation. After merging identical outputs and duplicating two question sets to check scoring stability, the expert rated 388 sentences from 22 effective question sets on two 1-5 scales: *Correctness* (whether the meaning is recoverable, even from a stiff or literal translation) and *Authenticity* (whether the wording sounds native rather than word-for-word translated). Full scoring criteria are

in Appendix E.

Human evaluation results are shown in Table 4. To verify expert reliability, 2 of the 22 question sets were unannounced repetitions of earlier ones. Across the 72 scores (36 translations × 2 items), 52.8% were identical, 91.7% differed by at most 1 point, and none by more than 2. The mean absolute difference (MAD) was 0.56, which is within a reasonable range. On the evaluated Mandarin→Han-ji subset, Gemini 3.1 Pro Preview received mean correctness and authenticity scores close to those of the dictionary reference, followed by Taigi-Llama-2-Translator 13B; among local models, Gemma4 31B and Qwen3.5 122B perform best but still trail the cloud models. For Mandarin→Tâi-lô, Gemini 3.1 Pro Preview again scores highest on correctness, but its authenticity drops noticeably relative to the reference. This is visible in translating ” 臺灣藍鵲” (Taiwan Blue Magpie): as Han-ji, Gemini 3.1 Pro Preview correctly outputs the more authentic dictionary term ” 長尾山娘”, but as Tâi-lô it instead outputs a direct transliteration, ”Tâi-uân nâ-tshiok”. Taigi-Llama-2-Translator 13B is again second-best; all other models drop noticeably more when translating to Tâi-lô, especially on authenticity. Paired tests over the 20 source sentences show that Gemini 3.1 Pro Preview is the only system not significantly different from the dictionary reference on any of the four measures (Table 4 note), while Translator 13B differs significantly only in Tâi-lô correctness. With only 20 items, however, two systems must differ by roughly 0.4–0.65 points to be distinguishable, so the ranking among the remaining systems, particularly the local models, is only indicative.

Comparing the mean of the two human scores (Correctness and Authenticity) with the back-translation scores across these 11 systems, correlation for Han-ji is weak: Spearman $\rho = 0.264$ against the Gemini judge and $\rho = 0.527$ against the Llama judge, indicating a substantial gap between back-translation and expert judgment. The main cause is that weaker models sometimes paraphrase in different Mandarin wording, e.g. rendering ” 我們做事情，立場要堅定，才不會讓人家批評。” as ” 咱做啥事，立場要堅固，才會無人批評。”; both judges still back-translate this to something close to the original (Gemini 0.973, Llama 0.914), while experts scored it only 2 (Correctness) and 1 (Authenticity). For Tâi-lô, correlation is much stronger: $\rho = 0.664$ (Gemini) and $\rho = 0.973$ (Llama). Llama tends to back-translate Tâi-lô into Mandarin that hews close to the actual output, aligning well with expert judgment, while Gemini instead tries to reconstruct the intended original sentence, inflating scores for poor outputs. Gemini also sometimes back-translates authentic Taiwanese expressions literally, for example rendering the standard form ”tn̂g-bué-suann-niû” (長尾山娘) as Mandarin ” 長尾山娘” rather than ” 臺灣藍鵲”, so that a more standard reference answer can score lower than expected.

4.3 Discussion

Across both directions, cloud models perform best overall, though fine-tuned local models can approach their performance. Gemini 3.1 Pro Preview, the strongest cloud model, produces near dictionary-quality Taiwanese characters, but its authenticity drops sharply for Tâi-lô, suggesting it still falls back on direct translation from Mandarin. This weaker handling of Tâi-lô likely reflects its scarcity in training data relative to Taiwanese characters, which align more easily with Mandarin. Converting Mandarin into Taiwanese characters first, then into Tâi-lô via an LLM or a dedicated grapheme-to-phoneme (G2P) model, may be a promising direction for future work. Overall, despite these limitations, using LLMs to convert Mandarin corpora into Taiwanese text is a feasible and valuable approach.

5 Speech Corpus Synthesis

5.1 Methodology

Compared with high-resource languages, publicly documented and reproducible Taiwanese TTS systems remain limited, although dedicated systems and a recent benchmark have begun to emerge(Lan et al., 2026). We therefore investigate LLM-based TTS built on neural-codec language models, which has enabled strong zero-shot synthesis and has subsequently been extended to cross-lingual generation(Wang et al., 2023; Zhang et al., 2023).

Table 4: Comparison of human evaluation and automatic back-translation scoring (10 models + reference answers, 20 question subsets, sorted by overall human average score)

	Taiwanese characters (HAN)				Tâi-lô (TL)			
Model	Corr	Auth	Gemini	Llama	Corr	Auth	Gemini	Llama
Reference	4.90	4.75	0.967	0.956	4.90	4.70	0.967	0.920
Gemini 3.1 Pro Preview	4.80	4.80	0.959	0.953	4.80	4.45	0.992	0.907
Translator 13B	4.70	4.65	0.940	0.927	4.45	4.40	0.925	0.917
Grok 4.5	4.60	4.30	0.963	0.963	4.20	3.45	0.997	0.869
Claude Sonnet 5	4.50	4.05	0.966	0.967	3.65	3.25	0.986	0.854
GPT-5.6 Luna	4.35	3.60	0.945	0.947	3.50	3.00	0.994	0.868
Gemma4 31B	4.00	3.30	0.972	0.959	2.40	2.30	0.931	0.667
Qwen3.5-122B-A10B	3.85	3.50	0.972	0.936	2.25	1.95	0.852	0.709
GPT-OSS 120B	3.50	3.15	0.950	0.936	2.35	2.45	0.834	0.717
Qwen3.6-27B	3.70	3.15	0.955	0.914	2.15	1.80	0.847	0.667
Nemotron-3 Nano 30B	2.95	2.50	0.911	0.902	1.25	1.35	0.537	0.457

Paired differences between the reference and Gemini 3.1 Pro Preview (reference − model, 95% CI, $n = 20$): Han-ji correctness +0.10 $[-0.16, 0.36]$, Han-ji authenticity −0.05 $[-0.33, 0.23]$, Tâi-lô correctness +0.10 $[-0.11, 0.31]$, Tâi-lô authenticity +0.25 $[-0.18, 0.68]$. Translator 13B differs significantly only in Tâi-lô correctness (+0.45 $[0.09, 0.81]$).

Even so, finding a model capable of generating Taiwanese speech is not straightforward given its low-resource status; after testing several candidates, we found Gemini 3.1 Flash TTS(Google DeepMind, 2026a) produces reasonable Taiwanese speech, notably despite Taiwanese not being listed among its officially supported languages.

To the best of our knowledge, prior Taiwanese TTS studies have focused on dedicated systems; the zero-shot use of a general-purpose LLM-based TTS model without official Taiwanese support remains underexplored(Lan et al., 2026). We compare three input conditions: Han-ji only, Tâi-lô only, and both together. Detailed prompts are in Appendix G. We used SuíSiann v0.2.1(i3thuan5, 2026), which contains 3,467 segmented clips from one speaker, 62,355 syllables, and 17,042 seconds (4:44:02) of audio. Each audio file includes a word-for-word transcription of the corresponding Taiwanese characters and Taiwanese Romanization. We selected 30 sentences, covering different sentence lengths (excluding extremely long sentences from the original dataset; the selected sentences all correspond to audio files with a length of less than 10 seconds), totaling 359 syllables. The total length of the corresponding original audio files was 103 seconds. These sentences were then used to generate speech via TTS according to the three different formats mentioned above, resulting in a total of 90 speech samples.

5.2 Results

Because standardized Taiwanese TTS evaluation is only beginning to emerge and automatic metrics do not fully capture pronunciation, tone sandhi, and naturalness, we conduct human listening evaluation. We conduct a MOS-style listening test with 10 native Taiwanese listeners, following the relevant principles of ITU-T P.808 for remote testing, while using three task-specific rating dimensions. Each rater heard, for 30 sentences (given in both Han-ji and Tâi-lô), speech from the three TTS conditions plus the real human recording, 120 clips total, rated blind. Following Appendix H, raters scored three task-specific dimensions on 1–5 scales: Pronunciation Accuracy (whether the speech matches the intended pronunciation, excluding tone sandhi, which instead affects the accent metric below); Technical Naturalness (whether the speech is indistinguishable from a real human voice); and Authenticity of Accent (whether, assuming correct pronunciation, improper tone sandhi makes the speech sound non-Taiwanese).

As shown in Table 5, when both Han-ji and Tâi-lô are provided, Gemini 3.1 Flash TTS achieves MOS scores of 3.763 for Pronunciation Accuracy, 3.857 for Technical Naturalness, and 3.777 for Authenticity of Accent, indicating that the resulting speech quality is generally good, though still distinguishable from real human speech as having some flaws.

Table 5: MOS-style ratings of speech synthesized by Gemini 3.1 Flash TTS under three input conditions, compared with the original human recording: three independent ratings (1–5 scale) ± 95% CI per condition, namely Pronunciation (Pronunciation Accuracy), Naturalness (Technical Naturalness), and Authenticity (Authenticity of Accent). All 10 raters; $n = 30$ items, each item's score first averaged across raters before computing the across-item mean and CI, since raters are repeated measures on the same items rather than independent samples.

Condition	Pronunciation	Naturalness	Authenticity
Human (real recording)	4.557 ± 0.075	4.440 ± 0.078	4.430 ± 0.092
Both (Han-ji + Tâi-lô)	3.763 ± 0.208	3.857 ± 0.164	3.777 ± 0.191
Tâi-lô only	3.747 ± 0.151	3.760 ± 0.133	3.653 ± 0.153
Han-ji only	2.820 ± 0.268	3.293 ± 0.189	3.030 ± 0.255

When only Tâi-lô is provided, Pronunciation Accuracy is close to the above, but Technical Naturalness and Authenticity of Accent decline slightly. When only Han-ji is provided, Pronunciation Accuracy drops the most, and Technical Naturalness and Authenticity of Accent also decline significantly, showing that Tâi-lô is indispensable for Gemini 3.1 Flash TTS to produce Taiwanese speech.

Among the above samples, across the 240 repeated-item scores given by the ten raters, the mean absolute difference (MAD) was 0.333, with 68.8% of scores matching exactly, 97.9% showing an absolute difference within 1 point, and all scores showing an absolute difference within 2 points. The single highest per-rater MAD on repeated items among the ten raters was only 0.542, indicating that the ratings are sufficiently reliable.

Some synthesized clips received ratings overlapping with or exceeding the corresponding human recordings; this may reflect item-level variation in synthesis, recording quality, and listener judgments, and does not by itself establish generation instability. The per-sentence gap between the TTS and real human speech average scores across all 30 sentences is provided in Appendix I.

We additionally attempted feeding the 30 audio clips generated under the Both condition to Gemini 3.1 Pro Preview, which supports multimodal input, using the exact same prompt given to the human raters(Chen et al., 2025). Across these 30 samples, it produced an average score of 4.333 for Pronunciation Accuracy, 3.733 for Technical Naturalness, and 3.467 for Authenticity of Accent, averaging 3.844 across the three dimensions, close to the human average of 3.799 across the same three dimensions. However, at the level of individual items ($n = 30$), the Pearson correlation between Gemini 3.1 Pro Preview's scores and the human average scores was $r = 0.350$ for Pronunciation Accuracy, $r = 0.269$ for Technical Naturalness, and $r = 0.421$ for Authenticity of Accent, indicating that Gemini's scores do not align closely with human judgments on a per-sample basis. Nonetheless, selecting only the samples for which Gemini's average score across the three dimensions was at least 4 (13 clips in total), the human-rated average for these 13 clips was 4.005, higher than the 3.799 average across all samples. Because the subset was selected and evaluated on the same 30 clips and item-level agreement is weak, this difference is modest relative to item-level variability and should be regarded only as a preliminary observation that multimodal LLMs might assist in filtering TTS output, pending validation on held-out data.

5.3 Discussion

In the experiments above, we found that general-purpose LLM-based TTS, despite not listing Taiwanese as a supported language, may still be capable of producing high-quality Taiwanese speech, in some cases reaching quality close to real human speech, albeit with considerable item-level variation. Post-hoc filtering with a general-purpose multimodal LLM showed preliminary potential to select a subset with higher average human ratings, although the effect requires held-out validation. This indicates that using general-purpose LLM-based TTS to produce Taiwanese speech is a feasible and promising approach to speech corpus synthesis. For the tested general-purpose TTS model, explicit Tâi-lô input provided impor-

tant pronunciation and tonal cues that were not reliably recovered from Han-ji alone.

6 Conclusion

To address the scarcity of Taiwanese text and speech corpora, we investigate the feasibility of synthesizing such corpora using off-the-shelf general-purpose LLMs and TTS. Among the cloud and local models available at the time of our experiments, Gemini 3.1 Pro Preview achieves correctness and authenticity close to dictionary-reference quality when translating Mandarin into Taiwanese characters, though authenticity declines slightly for Tâi-lô; fine-tuned local small models achieve comparable quality. For speech, Gemini 3.1 Flash TTS, a general-purpose model without official Taiwanese support, generates speech that sometimes approaches human quality and relies on Tâi-lô information, and a multimodal LLM showed preliminary, not yet validated potential for filtering lower-quality clips. We therefore conclude that generating corpora for low-resource languages with off-the-shelf general-purpose LLMs and TTS is a feasible and promising approach.

Limitations

Text evaluation. The text experiments use only 400 example sentences from the Ministry of Education dictionary, which may not reflect broader real-world usage such as conversation, news, or social media. Stage 1 scores rely on Mandarin embedding similarity, a semantic proxy that is inflated by character overlap (untranslated Han-ji already score 0.735), and Stage 2 scores additionally inherit the translation ability and biases of the back-translation judges, including error cancellation and the Translator 13B judge evaluating outputs from its own model family. The reported MSDs capture item-sampling variability only, not these systematic biases. Finally, we evaluate translation from Mandarin; the models' ability to generate Taiwanese text directly remains untested.

Human evaluation. Translation quality was judged by a single expert on 20 source sentences, which are on average shorter than the full test set. Although an unannounced repetition check indicated reasonable self-consistency, individual bias cannot be excluded, and a second annotation was not feasible given the scarcity of experts proficient in written Taiwanese. Our human-evaluation conclusions are therefore limited to large differences rather than a full ranking of systems: chiefly, that Gemini 3.1 Pro Preview was not significantly different from the dictionary reference, and that back-translation scores closely track expert judgment only for Tâi-lô under the Llama judge. The latter correlations are computed over only 11 systems and should be interpreted accordingly.

Models and reproducibility. Cloud models were accessed through APIs at the time of our experiments; several were preview versions that may change or be withdrawn, and each configuration was run once at temperature 0. Thinking configurations were constrained by the APIs used, so the No-Think setting could not be evaluated for every model.

Speech synthesis. Only one TTS model, Gemini 3.1 Flash TTS, was evaluated, so the findings reflect this model rather than LLM-based TTS in general. We did not compare against dedicated or fine-tuned Taiwanese TTS systems, which remain scarce in publicly reproducible form; such a comparison is left for future work. The test covered 30 sentences under 10 seconds from a single-speaker dataset, and prompts and hyperparameters were not tuned. The multimodal-LLM filtering result is a preliminary, in-sample observation: the retained subset was selected and evaluated on the same 30 clips, item-level correlation with human ratings was weak ($r = 0.27$–0.42), and the effect has not been validated on held-out data.

Use of synthetic corpora. We assess the quality of synthesized text and speech but have not yet verified their usefulness for training downstream ASR, TTS, or MT systems. In addition, while synthesis avoids redistributing license-restricted source corpora, the use of generated content remains subject to each provider's terms of service, and outputs of Taigi-Llama-2-Translator are governed by CC-BY-NC-SA-4.0.

Acknowledgments

We thank Chien-Ying Li for the careful expert annotation of the translation outputs, and the ten native Taiwanese listeners who participated in the speech evaluation. We are grateful to the National Center for High-performance Computing (NCHC) of the National Institutes of Applied Research (NIAR), Taiwan, for providing the computational and storage resources used in this work. This work was supported by the Ministry of Education (MOE) of Taiwan under the project Taiwan Centers of Excellence in Artificial Intelligence, through the NTU Artificial Intelligence Center of Research Excellence.

References

Sandhini Agarwal, Lama Ahmad, Jason Ai, Sam Altman, Andy Applebaum, Edwin Arbus, Rahul K Arora, Yu Bai, Bowen Baker, Haiming Bao, et al. 2025. gpt-oss-120b & gpt-oss-20b model card. *arXiv preprint arXiv:2508.10925*.

Anthropic. 2026. Claude sonnet 5 system card. Published: 30 June 2026.

BAAI. 2023. BAAI/bge-large-zh-v1.5 model card and BGE documentation. Accessed 2026-07-29.

Aaron Blakeman, Aaron Grattafiori, Aarti Basant, Abhibha Gupta, Abhinav Khattar, Adi Renduchintala, Aditya Vavre, Akanksha Shukla, Akhiad Bercovich, Aleksander Ficek, et al. 2025. Nemotron 3 nano: Open, efficient mixture-of-experts hybrid mamba-transformer model for agentic reasoning. *arXiv preprint arXiv:2512.20848*.

Aakshita Chandiramani, Aaron Blakeman, Abdullahi Olaoye, Abhibha Gupta, Abhilash Somasamudramath, Abhinav Khattar, Adeola Adesoba, Adi Renduchintala, Adil Asif, Aditya Agrawal, et al. 2026. Nemotron 3 super: Open, efficient mixture-of-experts hybrid mamba-transformer model for agentic reasoning. *arXiv preprint arXiv:2604.12374*.

Kai-Wei Chang, Yi-Cheng Lin, Huang-Cheng Chou, Wenze Ren, Yu-Han Huang, Yun-Shao Tsai, Chien-Cheng Chen, Yu Tsao, Yuan-Fu Liao, Shrikanth Narayanan, James Glass, and Hung-yi Lee. 2026. TaigiSpeech: A low-resource real-world speech intent dataset and preliminary results with scalable data mining in-the-wild. *arXiv preprint arXiv:2603.21478*.

Chen Chen, Yuchen Hu, Siyin Wang, Helin Wang, Zhehuai Chen, Chao Zhang, Chao-Han Huck Yang, and Eng Siong Chng. 2025. Audio large language models can be descriptive speech quality evaluators. In *International Conference on Learning Representations*.

Yi-Hui Chou, Kalvin Chang, Meng-Ju Wu, Winston Ou, Alice Wen-Hsin Bi, Carol Yang, Bryan Y. Chen, Rong-Wei Pai, Po-Yen Yeh, Jo-Peng Chiang, Iu-Tshian Phoann, Winnie Chang, Chenxuan Cui, Noel Chen, and Jiatong Shi. 2023. Evaluating self-supervised speech models on a taiwanese hokkien corpus. In *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 1–7. IEEE.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*.

Markus Freitag, George Foster, David Grangier, Viresh Ratnakar, Qijun Tan, and Wolfgang Macherey. 2021. Experts, errors, and context: A large-scale study of human evaluation for machine translation. *Transactions of the Association for Computational Linguistics*, 9:1460–1474.

Xiyan Fu and Wei Liu. 2025. How reliable is multilingual LLM-as-a-judge? In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 11040–11053. Association for Computational Linguistics.

Gemma Team et al. 2026. Gemma 4 technical report.

Ona de Gibert, Joseph Attieh, Teemu Vahtola, Mikko Aulamo, Zihao Li, Raúl Vázquez, Tiancheng Hu, and Jörg Tiedemann. 2025. Scaling low-resource MT via synthetic data generation with LLMs. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 27674–27692. Association for Computational Linguistics.

Google DeepMind. 2025. Gemini 3 flash model card. Published: December 2025.

Google DeepMind. 2026a. Gemini 3.1 flash audio (flash live, tts) - model card. Published: 26 March 2026, Updated: April 2026.

Google DeepMind. 2026b. Gemini 3.1 flash-lite - model card. Published: 3 March 2026.

Google DeepMind. 2026c. Gemini 3.1 pro - model card. Published: 19 February 2026.

Google DeepMind. 2026d. Gemini 3.5 flash - model card. Published: 19 May 2026.

i3thuan5. 2026. Suísiann dataset: Taiwanese speech corpus.

Henning Klöter. 2005. *Written Taiwanese*. Studia Formosiana. Harrassowitz Verlag, Wiesbaden.

Yu-Siang Lan, Chia-Sheng Liu, Yi-Chang Chen, Po-Chun Hsu, Allyson Chiu, Shun-Wen Lin, Dashan Shiu, and Yuan-Fu Liao. 2026. Breeze taigi: Benchmarks and models for taiwanese hokkien speech recognition and synthesis. *arXiv preprint arXiv:2603.19259*.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-Eval: NLG evaluation using GPT-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2511–2522. Association for Computational Linguistics.

Bo-Han Lu, Yi-Hsuan Lin, En-Shiun Annie Lee, and Richard Tzong-Han Tsai. 2024. Enhancing taiwanese hokkien dual translation by exploring and standardizing of four writing systems.

Ministry of Education, Taiwan. 2006. 臺灣閩南語羅馬字拼音方案 (taiwan southern min romanization system). Official announcement No. 0950151609, October 14, 2006.

OpenAI. 2026. Gpt-5.6: Frontier intelligence that scales with your ambition. Published: 9 July 2026.

Bowen Qin, Duanyu Feng, and Xi Yang. 2025. Conversational few-shot prompting: Rethinking few-shot prompting for chat language model.

Qwen Team. 2026a. Qwen3.5: Towards native multimodal agents.

Qwen Team. 2026b. Qwen3.6-27B: Flagship-level coding in a 27B dense model.

Qwen Team. 2026c. Qwen3.6-35B-A3B: Agentic coding power, now open to all.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pages 3982–3992. Association for Computational Linguistics.

SpaceXAI. 2026. Introducing grok 4.5. Published: 8 July 2026.

Hung-Yang Sung, Chien-Chun Wang, Kuan-Tang Huang, Tien-Hong Lo, Yu-Sheng Tsao, Yung-Chang Hsu, and Berlin Chen. 2025. CLiFT-ASR: A cross-lingual fine-tuning framework for low-resource taiwanese hokkien speech recognition. In *Proceedings of the 37th Conference on Computational Linguistics and Speech Processing (ROCLING 2025)*, pages 176–183. The Association for Computational Linguistics and Chinese Language Processing.

Chengyi Wang, Sanyuan Chen, Yu Wu, Ziqiang Zhang, Long Zhou, Shujie Liu, Zhuo Chen, Yanqing Liu, Huaming Wang, Jinyu Li, Lei He, Sheng Zhao, and Furu Wei. 2023. Neural codec language models are zero-shot text to speech synthesizers.

Anyi Xu, Bangcai Lin, Bing Xue, Bingxuan Wang, Bingzheng Xu, Bochao Wu, Bowei Zhang, Chaofan Lin, Chen Dong, Chenchen Ling, et al. 2026. Deepseek-v4: Towards highly efficient million-token context intelligence. *arXiv preprint arXiv:2606.19348*.

Ziqiang Zhang, Long Zhou, Chengyi Wang, Sanyuan Chen, Yu Wu, Shujie Liu, Zhuo Chen, Yanqing Liu, Huaming Wang, Jinyu Li, Lei He, Sheng Zhao, and Furu Wei. 2023. Speak foreign languages with your own voice: Cross-lingual neural codec language modeling. *arXiv preprint arXiv:2303.03926*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging LLM-as-a-judge with MT-Bench and chatbot arena. In *Advances in Neural Information Processing Systems*, volume 36, pages 46595–46623.

Terry Yue Zhuo, Qiongkai Xu, Xuanli He, and Longxiang Wang. 2023. Rethinking round-trip translation for machine translation evaluation. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 319–337. Association for Computational Linguistics.

中華民國教育部. 2026. 教育部臺灣台語常用詞辭典. Accessed: 2026-07-16.

中華民國文化部. 2022. 國家語言發展報告 (national language development report). Approved report (核定本), 中華民國文化部. Approved by the Executive Yuan on July 15, 2022; published by the Ministry of Culture.

A Text Translation Prompt Design

All translation tasks use a single system prompt template, with only the source and target content types substituted according to the task direction (HAN2ZH, TL2ZH, ZH2HAN, or ZH2TL). The prompt is reproduced verbatim in the original Mandarin below, since this is precisely what was submitted at inference time, followed by an English translation for reference.

> 你是專業的台灣閩南語 (台語) 與繁體中文雙向翻譯助手。只輸出翻譯結果本身，不要加任何解釋、標點以外的說明文字或引號。任務：把使用者輸入的 `{source}` 句子翻譯成 `{target}`。
>
> (You are a professional Taiwanese Hokkien to Traditional Chinese bidirectional translation assistant. Only output the translation result itself, without adding any explanations, any text other than punctuation, or quotation marks. Task: Translate the user-inputted `{source}` sentence into `{target}`.)

Here, `{source}` and `{target}` are substituted with the appropriate content type for each task: Taiwanese characters (台語漢字), Tâi-lô (台羅拼音), or Traditional Chinese (繁體中文).

B Stage 1 Token Usage and Reasoning Effort Ablation

We compiled the total output-side token usage for the HAN2ZH and TL2ZH tasks across all reasoning-enabled cloud models, summed over the 400 test cases (Table 6). Since some models cannot accurately separate token usage between thinking and completion, only the total output is recorded; for reference, the actual completion output token count for all Gemini series models was approximately 4k for both tasks, including Gemini 3.1 Flash Lite, which was run without thinking (its API default) and likewise produced approximately 4k output tokens in total. Token consumption varies by orders of magnitude: Claude Sonnet 5 and Gemini 2.5 Flash use the fewest, GPT-5.6 and DeepSeek V4 are mid-range, and other Gemini models plus Grok 4.5 use substantially more. Across these 11 models, token count correlates with score at Spearman $\rho = -0.36$ for HAN2ZH and $\rho = 0.45$ for TL2ZH. However, since factors such as model architecture and training data cannot be ruled out as confounds, this correlation should not be taken to indicate that reasoning length itself has a negative causal effect on output quality.

Table 6: Total output-side token usage in Stage 1 for all cloud models with reasoning enabled.

Model	HAN2ZH	TL2ZH
Gemini 3.1 Pro Preview	278,974	735,625
Gemini 3.5 Flash	184,813	203,114
Gemini 3 Flash Preview	474,113	140,012
Gemini 2.5 Pro	845,369	764,690
Gemini 2.5 Flash	10,193	16,395
GPT-5.6 Luna	45,258	117,778
GPT-5.6 Terra	31,542	68,308
Grok 4.5	429,143	594,076
DeepSeek V4 Pro	28,429	122,768
DeepSeek V4 Flash	38,066	63,152
Claude Sonnet 5	8,065	10,421

Table 7: Claude Sonnet 5 Stage 1 results across reasoning effort levels: score and total output-side token usage (reasoning + visible output combined). Only the HAN2ZH difference between Low and Max is statistically significant (see text).

	HAN2ZH		TL2ZH	
Effort	Score	Tokens	Score	Tokens
Low	**0.9078**	7,615	0.8585	7,589
Medium	0.9052	8,065	0.8611	10,421
High	0.9051	8,457	0.8581	10,862
XHigh	0.9027	10,902	**0.8635**	49,637
Max	0.9019	158,113	0.8613	174,792

To further isolate the relationship between reasoning length and translation quality, we adjust the reasoning effort of Claude Sonnet 5 while holding the model fixed. Here, `medium` represents the default effort level when calling the API using OpenRouter; therefore, all Claude Sonnet 5 results mentioned in the main text refer to this version. Comparing the two extremes, Low scores slightly but significantly higher than Max on HAN2ZH (+0.006, paired 95% CI $[0.001, 0.011]$, $p = 0.027$, uncorrected), whereas on TL2ZH the difference is not significant and in the opposite

direction (-0.003 $[-0.011, 0.005]$, $p = 0.48$). Longer reasoning therefore does not consistently lower translation quality, and the HAN2ZH effect is smaller than the typical between-model MSD (about 0.010), leaving the rankings in Table 3 unaffected. Further examination of the HAN2ZH items whose scores decreased yielded Table 8: as reasoning length increases, some outputs use words closer to archaic language or Taiwanese within otherwise standard Mandarin, making the translations less straightforward and noticeably different from the reference text. Therefore, rating model bias may be a systemic limitation that requires attention.

Table 8: Examples of Claude Sonnet 5 HAN2ZH translation quality degrading as reasoning effort increases. The values in parentheses represent the similarity scores, including those obtained by directly comparing Taiwanese characters with Mandarin.

Low	Medium	High	XHigh	Max	Reference	Han-ji Source
他又在發脾氣了！(1.0000)	他又在發脾氣了！(1.0000)	他又在耍脾氣了！(0.9617)	他又在耍脾氣了！(0.9617)	他又在發𠢕了！(0.8320)	他又在發脾氣了！	伊閣咧發𠢕矣！(0.5340)
你還懂得看風向站邊，知道要跟著我們走。(0.8943)	你還懂得看風向、投靠強勢的一方，知道要跟著我們走。(0.8702)	你還懂得看風向、趨炎附勢，知道要跟著我走。(0.8375)	你還懂得看情況靠向有利的一邊，知道要跟我們走。(0.8236)	你還懂得西瓜偎大邊，知道要跟我們走。(0.7289)	你還會見風轉舵，知道要跟著我們走。	你閣會曉西瓜倚大爿，知影欲綴阮行。(0.4928)

C Full Thinking vs. Non-Thinking Comparison for Stage 1 and Stage 2

Table 9: Stage 1 evaluation results: Han-ji→Mandarin (HAN2ZH) and Tâi-lô→Mandarin (TL2ZH), default vs. reasoning-disabled (No-Think) settings.

	HAN2ZH				TL2ZH			
	Think		No-Think		Think		No-Think	
Model	Score	SR (%)	Score	SR (%)	Score	SR (%)	Score	SR (%)
Cloud Models								
Gemini 3.1 Pro Preview	**0.905**	100.00	–	–	**0.895**	100.00	–	–
Gemini 3.5 Flash	0.890	99.50	0.876	100.00	0.890	100.00	**0.871**	100.00
Gemini 3 Flash Preview	0.887	100.00	0.869	99.75	0.880	99.50	0.866	99.75
Gemini 3.1 Flash Lite	–	–	**0.886**	100.00	–	–	0.842	100.00
Gemini 2.5 Pro	0.879	99.75	–	–	0.850	99.75	–	–
Gemini 2.5 Flash	0.882	99.75	0.880	100.00	0.826	100.00	0.829	99.75
GPT-5.6 Luna	0.890	100.00	0.873	100.00	0.822	100.00	0.777	100.00
GPT-5.6 Terra	0.888	100.00	0.884	100.00	0.836	100.00	0.798	100.00
Grok 4.5	0.889	100.00	–	–	0.837	99.50	–	–
DeepSeek V4 Pro	0.900	100.00	–	–	0.845	100.00	–	–
DeepSeek V4 Flash	0.888	99.75	–	–	0.843	99.50	–	–
Claude Sonnet 5	**0.905**	100.00	–	–	0.861	100.00	–	–
Local Models								
GPT-OSS 20B (A3.6B)	0.782	100.00	–	–	0.465	100.00	–	–
GPT-OSS 120B (A5.1B)	0.806	100.00	–	–	0.543	100.00	–	–
Nemotron-3 Nano 30B (A3B)	0.751	99.75	0.732	100.00	0.417	94.00	0.419	100.00
Nemotron-3 Super 120B (A12B)	0.801	99.75	0.754	100.00	0.505	99.75	0.538	100.00
Qwen3.5-122B-A10B (A10B)	**0.858**	100.00	**0.864**	100.00	**0.725**	100.00	**0.721**	100.00
Qwen3.6-27B (Dense)	0.849	100.00	0.856	100.00	0.649	98.50	0.640	100.00
Qwen3.6-35B-A3B (A3B)	0.842	100.00	0.834	100.00	0.636	98.25	0.638	100.00
Gemma4 26B (A4B)	0.838	99.25	0.842	100.00	0.613	74.00	0.578	100.00
Gemma4 31B (Dense)	0.853	100.00	0.843	100.00	0.646	99.75	0.641	100.00
Taigi-Llama-2-Translator Models								
Translator 7B	–	–	0.879	100.00	–	–	0.824	100.00
Translator 13B	–	–	**0.894**	100.00	–	–	**0.845**	100.00

Gemini 3.1 Flash Lite was run only at its API default, and the Taigi-Llama-2-Translator models do not support thinking; all other models use the default effort mode for thinking. For Gemini models, No-Think used the API's lowest thinking setting; all Gemini No-Think runs, as well as Gemini 3.1 Flash Lite, produced zero thinking tokens. A missing "No-Think" entry indicates that thinking could not be turned off through the API used in our experiments (Google AI Studio, OpenRouter, or Ollama).

Failed items were retried once and still failed: Gemini 3 series failures were safety-filter rejections; other cloud models failed due to unspecified API errors; local models were terminated after exceeding the 5-minute runtime or 16,384-token thinking limit on a single H200 GPU.

Table 10: Stage 2 evaluation results judged by the Taigi-Llama-2-Translator 13B and the Gemini 3 Flash Preview (embedding similarity only): Mandarin→Han-ji (ZH2HAN) and Mandarin→Tâi-lô (ZH2TL), default vs. reasoning-disabled (No-Think) settings.

	ZH2HAN						ZH2TL					
	Think			No-Think			Think			No-Think		
Model	Llama	Gemini	Cost	Llama	Gemini	Cost	Llama	Gemini	Cost	Llama	Gemini	Cost
Cloud Models												
Gemini 3.1 Pro Preview	0.923	0.929	319k	–	–	–	**0.878**	0.944	192k	–	–	–
Gemini 3.5 Flash	0.930	0.940	247k	0.910	0.925	261k	0.852	0.943	256k	0.792	0.894	739k
Gemini 3 Flash Preview	0.928	0.933	312k	0.906	0.913	258k	0.853	0.943	194k	0.787	0.901	550k
Gemini 3.1 Flash Lite	–	–	–	**0.931**	0.925	325k	–	–	–	0.844	0.934	411k
Gemini 2.5 Pro	0.898	0.903	247k	–	–	–	0.846	0.920	324k	–	–	–
Gemini 2.5 Flash	0.905	0.911	327k	0.905	0.907	327k	0.841	0.930	654k	**0.850**	**0.939**	350k
GPT-5.6 Luna	0.929	0.929	322k	0.922	0.924	394k	0.822	0.939	462k	0.771	0.899	1150k
GPT-5.6 Terra	0.929	0.934	205k	0.926	**0.933**	265k	0.841	0.943	458k	0.829	0.934	683k
Grok 4.5	**0.939**	**0.945**	387k	–	–	–	0.866	**0.955**	320k	–	–	–
DeepSeek V4 Pro	0.927	0.927	316k	–	–	–	0.873	0.940	454k	–	–	–
DeepSeek V4 Flash	0.917	0.921	197k	–	–	–	0.840	0.941	389k	–	–	–
Claude Sonnet 5	0.933	0.931	325k	–	–	–	0.839	0.941	335k	–	–	–
Local Models												
GPT-OSS 20B	0.915	0.917	615k	–	–	–	0.504	0.652	6731k	–	–	–
GPT-OSS 120B	0.910	0.919	529k	–	–	–	0.656	0.846	2136k	–	–	–
Nemotron-3 Nano 30B	0.873	0.882	900k	0.854	0.873	1247k	0.443	0.545	9119k	0.434	0.501	8672k
Nemotron-3 Super 120B	0.893	0.895	607k	0.830	0.872	888k	0.491	0.646	4820k	0.610	0.794	4921k
Qwen3.5-122B-A10B	0.905	0.924	366k	0.894	0.903	489k	**0.712**	0.854	2163k	0.656	0.776	2901k
Qwen3.6-27B	0.912	0.916	285k	0.912	0.913	538k	0.606	0.757	2580k	0.581	0.725	4915k
Qwen3.6-35B-A3B (A3B)	0.910	0.911	623k	0.905	0.907	720k	0.629	0.758	3435k	0.584	0.694	5474k
Gemma4 26B	**0.929**	**0.933**	347k	0.928	0.932	416k	0.653	0.791	2972k	0.614	0.750	3782k
Gemma4 31B (Dense)	0.925	0.924	410k	**0.932**	**0.936**	518k	0.684	**0.865**	1519k	**0.660**	**0.846**	2489k
Taigi-Llama-2-Translator Models												
Translator 7B	–	–	–	**0.906**	**0.904**	333k	–	–	–	0.869	0.887	487k
Translator 13B	–	–	–	**0.906**	0.892	335k	–	–	–	**0.871**	**0.895**	720k

The average effective response rate was 99.94% (at least 98% per score), so generation failures do not materially affect the results; success rate is therefore omitted here.

D Minimum Significant Differences

For each pair of models, the minimum significant difference (MSD) is $t_{0.975,\,n-1} \times \mathrm{SE}(\bar{d})$, where $\bar{d}$ is the mean per-item score difference over the same $n = 400$ items (two-sided paired t-test, $\alpha = 0.05$, uncorrected for multiple comparisons). Table 11 reports the median and range over all model pairs within each group under the default setting. MSD reflects item-sampling variability only and does not account for systematic biases of the back-translation judges.

Table 11: MSD per leaderboard and model group: median (min–max) over within-group model pairs.

Leaderboard	Cloud (66 pairs)	Local (36 pairs)	Translator 7B vs. 13B
HAN2ZH	0.010 (0.007–0.012)	0.013 (0.011–0.015)	0.010
TL2ZH	0.013 (0.007–0.016)	0.020 (0.014–0.025)	0.012
ZH2HAN (Llama)	0.009 (0.007–0.011)	0.011 (0.008–0.016)	0.009
ZH2HAN (Gemini)	0.010 (0.007–0.013)	0.013 (0.010–0.017)	0.012
ZH2TL (Llama)	0.015 (0.011–0.018)	0.022 (0.019–0.025)	0.014
ZH2TL (Gemini)	0.013 (0.007–0.015)	0.024 (0.021–0.032)	0.015

E Scoring Instructions for Taiwanese Characters and Romanization Translation Provided to Experts

Table 12: Translation Quality Evaluation Metrics and Scoring Criteria

Correctness

評估重點：這句話的「意思」有沒有對？請先把整句話的語意在腦中抽取出來，跟中文原句比對，暫時不要管這句台語讀起來順不順、用字道不道地——就算翻得再怎麼生硬、逐字直譯，只要意思是對的，正確性一樣可以打高分。(Evaluation Focus: Is the "meaning" of the sentence correct? Please first extract the semantic meaning of the entire sentence in your mind and compare it with the original Mandarin sentence, temporarily ignoring whether the Taiwanese sentence reads smoothly or uses authentic vocabulary—even if the translation is stiff or a literal word-for-word translation, as long as the meaning is correct, it can still receive a high score for correctness.)

Score	Scoring Criteria
5	意思完全正確，沒有遺漏、增添或誤解任何資訊 (The meaning is completely correct, without missing, adding, or misunderstanding any information.)
4	意思大致正確，只有語氣或細微語意上的小落差 (The meaning is mostly correct, with only minor discrepancies in tone or subtle semantics.)
3	部分正確，但有明顯的意思偏差或遺漏 (Partially correct, but with obvious meaning deviations or omissions.)
2	意思大部分錯誤，只有少部分能對應到原意 (The meaning is mostly wrong, with only a small part corresponding to the original meaning.)
1	意思完全錯誤、答非所問，或是無意義的亂碼 (The meaning is completely wrong, irrelevant, or meaningless gibberish.)

Authenticity

評估重點：先完全不管這句話翻得對不對，只看這段台語文字「本身」讀起來像不像台灣本地台語使用者會自然說出來的話。請想像您完全沒看過中文原句，單獨把這段台語文字唸出來，聽起來自不自然、用字道不道地、有沒有翻譯腔或逐字直譯的痕跡。(Evaluation Focus: First, completely ignore whether the sentence is translated correctly or not; only consider whether the Taiwanese text "itself" reads like something a native Taiwanese speaker in Taiwan would naturally say. Please imagine you have never seen the original Mandarin sentence and read this Taiwanese text out loud by itself to see if it sounds natural, if the vocabulary is authentic, and if there are traces of translationese or literal word-for-word translation.)

Score	Scoring Criteria
5	非常道地，完全像母語者自然會說的話 (Very authentic, exactly like what a native speaker would naturally say.)
4	大致道地，只有些微用字選擇上的不自然 (Mostly authentic, with only slightly unnatural vocabulary choices.)
3	普通，看得出是台語但用詞/語法有翻譯腔 (Average, recognizable as Taiwanese but the vocabulary/grammar has traces of translationese.)
2	明顯不道地，逐字直翻痕跡重 (Obviously unauthentic, with heavy traces of literal translation.)
1	完全不像台語，或使用嚴重錯誤的拼寫/用字 (Completely unlike Taiwanese, or uses severely incorrect spelling/vocabulary.)

F Screenshot of The Human Evaluation System Interface

F.1 Translation

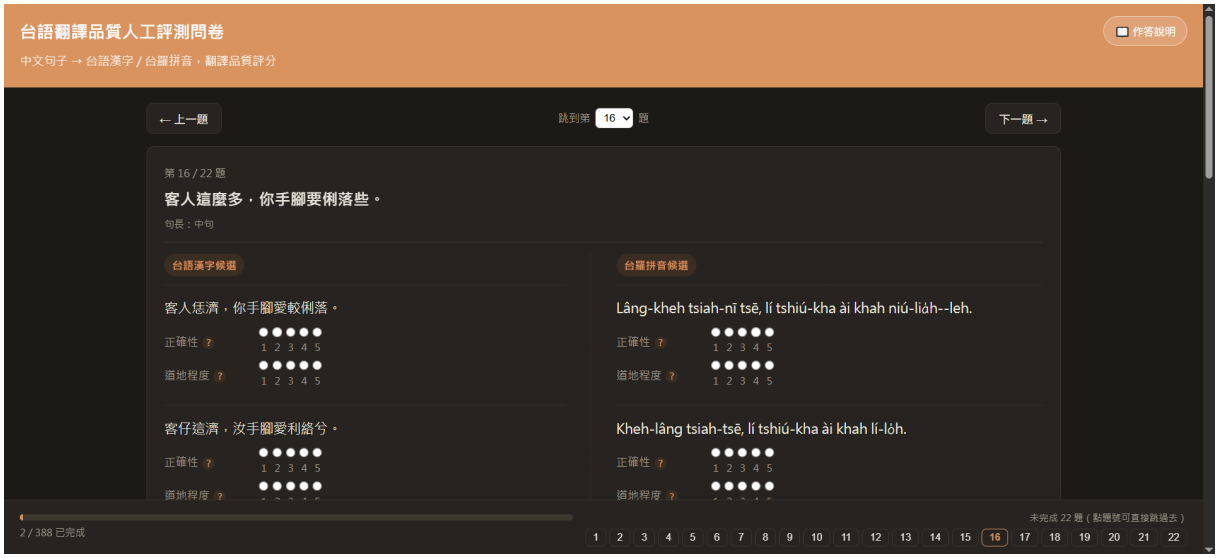

Figure 1: Screenshot of The Human Evaluation System Interface: Translation

F.2 Speech



Figure 2: Screenshot of The Human Evaluation System Interface: Speech

G TTS Prompt Design

Taiwanese Hokkien speech was synthesized using the Gemini Batch API (model `gemini-3.1-flash-tts-preview`). Depending on which input representation was provided, three prompt conditions were designed: Han-ji only (hanzi_only), Tâi-lô romanization only (tailo_only), and both representations together (both). All three begin with the same reading-style instruction, followed by condition-specific interpretation rules intended to prevent the model from misreading the input. The prompts below are the exact Mandarin text sent to the model for each condition, reproduced verbatim (not translated, since this is precisely what was submitted at inference time); `{hanzi}` and `{tailo}` are placeholders substituted at call time with the Han-ji sentence and Tâi-lô sentence for each data record, respectively.

G.1 Condition 1: Han-ji only (hanzi_only)

請用道地台灣人講台灣閩南語（台語）的自然腔調朗讀，發音要口語、自然，不要用逐字唸稿的生硬語氣。只唸出這句話本身，不要加任何額外的說明、開場白或結尾語。

以下是一句台語漢字書寫的句子。這些漢字是依照台灣閩南語的用字習慣寫成的，請務必用台語（台灣閩南語）發音朗讀，不要用華語（國語/普通話）的字面讀音去唸——同一個漢字在台語裡的讀音、甚至詞彙用法可能跟華語完全不同，請以台語爲準。

句子：`{hanzi}`

G.2 Condition 2: Tâi-lô romanization only (tailo_only)

請用道地台灣人講台灣閩南語（台語）的自然腔調朗讀，發音要口語、自然，不要用逐字唸稿的生硬語氣。只唸出這句話本身，不要加任何額外的說明、開場白或結尾語。

以下是一句用台羅拼音（Tâi-lô，台灣閩南語羅馬字拼音方案）標記的句子，不是英文也不是漢語拼音。請依照台羅拼音規則（包含聲調符號）解讀，唸出對應的道地台灣閩南語語音。

台羅拼音：`{tailo}`

G.3 Condition 3: Han-ji and Tâi-lô together (both)

請用道地台灣人講台灣閩南語（台語）的自然腔調朗讀，發音要口語、自然，不要用逐字唸稿的生硬語氣。只唸出這句話本身，不要加任何額外的說明、開場白或結尾語。

以下同時提供一句台語漢字寫法與對應的台羅拼音（Tâi-lô）標註。請用台語（台灣閩南語）發音朗讀這句話；若漢字本身的讀音有多種可能，請以台羅拼音的標註爲準來決定正確讀音與聲調。

漢字：`{hanzi}`
台羅拼音：`{tailo}`

H Rating Rubric for Speech Evaluation

This section reproduces the rating rubric exactly as presented to human annotators in the speech evaluation website used for speech corpus synthesis of this study. For each item, annotators were shown the Han-ji and Tâi-lô romanization text of a target sentence, followed by four audio clips in randomized order with the source condition withheld. Annotators listened to each clip in turn and rated it on three independent dimensions, each on a 5-point scale. The text below is reproduced verbatim in the original Mandarin, since this is precisely the wording shown to annotators; it is not translated.

每一題會先給一句台語的「漢字」與「台羅拼音」文字（這是這句話應該要表達的正確內容），接著會有 4 段音檔（隨機順序，不會告訴你是哪一種來源）。請依序播放並針對每一段音檔，分別評三個獨立的分數：

1. 讀音正確性

排除音調，只看讀音正不正確——台語有變調規則，聲調在連續語流中本來就會隨語境改變，要求逐字判斷「聲調準不準」並不合理，所以這一項請先忽略聲調，只判斷這段音檔唸出來的是不是上方給的漢字/台羅拼音所代表的「正確字/詞」。如果聽到的明顯是別的字、別的詞、或別的意思（誤讀），這一項才扣分，先不用管音質好不好聽，也不用管聲調準不準。

分數	標準
5	讀音（字詞）完全正確
4	大致正確，僅個別字詞的讀音略有落差
3	部分讀音錯誤，但仍聽得出原意
2	多處讀音錯誤，明顯偏離原句
1	讀音完全錯誤或無法辨識在唸什麼

2. 技術自然度

先不管讀音對不對、腔調道不道地，單純聽這段聲音本身像不像眞人在說話——有沒有機械感、卡頓、斷句怪異、音質瑕疵等合成痕跡。

分數	標準
5	完全像眞人說話，聽不出任何合成痕跡
4	大致自然，僅極少數地方稍有生硬感
3	聽得出是合成語音，但整體可接受
2	機械感明顯，斷句或音質有明顯問題
1	非常不自然，明顯是機器合成、難以聽下去

3. 腔調道地程度

先假設這是人聲，這個腔調聽起來像不像台灣本地的台語腔（而不是其他地區的閩南語腔調，或是華語腔硬套上去的台語）。

分數	標準
5	非常道地的台灣本地台語腔
4	大致像台灣腔，僅少數用字/語調有落差
3	聽得出想講台語，但腔調不夠道地
2	腔調明顯不像台灣本地台語（例如偏華語腔）
1	完全不像台灣台語的腔調

這三項請分開判斷：一段音檔完全可能「讀音正確但機械感重」，也可能「聽起來像台灣腔又自然，但讀錯了某幾個字的音」。不要因爲某一項分數高就直覺認爲另外兩項也該跟著高。

The three dimensions were explicitly designed to be judged independently, guarding against halo effects where a high score on one dimension biases the annotator toward high scores on the others (e.g., a clip may be phonetically accurate yet sound mechanical, or sound natural and accent-authentic while mispronouncing specific words).

I The gap between human and TTS speech on MOS for each sample

Table 13: Gap between the "both" condition and the human recording, averaged across MOS-P/MOS-N/MOS-A and 10 raters, sorted ascending (all 30 items). Gap = human_avg − both_avg; negative values indicate the synthetic "both" condition was rated *higher* than the human recording.

Han-ji	Tâi-lô	Both avg.	Human avg.	Gap
我看著的豬仔。	Guá khuànn–tioh ê ti-á.	4.567	3.967	−0.600
今仔日較袂寒。	Kin-á-jit khah bē kuânn.	4.433	4.167	−0.267
歹人共錢運出去國外。	Pháinn-lâng kā tsînn ūn tshut-khì kok-guā.	4.767	4.633	−0.133
袂記得咱眞正搦佇手頭的只有現在。	Bē-kì-tsit lán tsin-tsiànn lak tī tshiú-thâu ê Tsí-ū hiān-tsāi.	4.267	4.367	+0.100
蝶仔袂比鳥仔飛較遠。	Iah-á bē pí tsiáu-á pue khah hn̄g.	4.333	4.467	+0.133
前一日你敢毋是講啥物人有一間厝欲租人？	Tsîng tsit jit lí kám m̄ sī kóng siánn-mih lâng ū tsit king tshù beh tsoo–lâng?	4.100	4.233	+0.133
客廳大門向啥物方向？	Kheh-thiann tuā-mn̂g hiòng siánn-mih hong-hiòng?	4.433	4.567	+0.133
你著愛食予飽。	Lí tioh ài tsiah hōo pá.	4.267	4.433	+0.167
我聽著你欲嫁矣？	Guá thiann-tioh lí beh kè–ah?	4.067	4.267	+0.200
愛閣半點鐘才會到餐廳，逐家這馬會當欣賞外口的景緻。	Ài koh puànn tiám-tsing tsiah ē kàu tshan-thiann, tak-ke tsit-má ē-tàng him-sióng guā-kháu ê kíng-tì.	4.300	4.533	+0.233
毋免加講話啦，麵粉賠來。	M̄-bián ke kóng-uē–lah, mī-hún puê–lâi.	3.867	4.267	+0.400
這條路我誠熟。	Tsit tiâu lōo guá tsiânn sik.	4.033	4.467	+0.433
各人食各人的。	Kok-lâng tsiah kok-lâng ê.	3.900	4.500	+0.600
我驚去予厝邊兜加講話閣。	Guá kiann khì hōo tshù-pinn-tau ke-kóng-uē koh.	3.533	4.133	+0.600
阿母是毋甘加添你的煩惱，驚你無心情通做工課。	A-bú sī m̄-kam ka-thiam lí ê huân-ló, kiann lí bô sim-tsiânn thang tsò khang-khuè.	4.167	4.800	+0.633
户口佇山區的查某人比佇庄跤的較濟。	Hōo-kháu tī suann-khu ê tsa-bóo-lâng pí tī tsng-kha–ê khah tsē.	3.900	4.533	+0.633
事事攏愛倚靠人。	Sū-sū lóng ài uá-khò–lâng.	3.367	4.100	+0.733
全台灣有兩千三百五十一萬人，比舊年加一萬六千兩百八十八个。	Tsuân Tâi-uân ū nn̄g-tshian sann-pah gōo-tsap-it-bān lâng, pí kū-nî ke tsit-bān lak-tshing nn̄g-pah peh-tsap-peh ê.	4.000	4.767	+0.767
你敢誠實決定欲去是毋？	Lí kám tsiânn-sit kuat-tīng beh khì sī–m̄?	3.600	4.567	+0.967
嘛好，無，我寫批予伊，問看伊欲毋。	Mā hó, bô, guá siá phue hōo–i, mn̄g khuànn i beh–m̄.	3.600	4.567	+0.967
會，明仔暗穩落雨的。	Ē, bîn-á-àm ún loh-hōo–ê.	3.600	4.600	+1.000
我知影你對阮眞好，只是不得已的，乎，拜託咧。	Guá tsai-iánn lí tuì gún tsin hó, tsí-sī put-tik-í–ê, honnh, pài-thok–leh.	3.500	4.533	+1.033
你做過啥物摒掃的工課？做的時陣愛注意啥貨？	Lí tsò-kuè siánn-mih piànn-sàu ê khang-khuè? Tsò ê sî-tsūn ài tsù-ì siánn-huè?	3.433	4.500	+1.067
這是大路，你共阮拍怎樣？	Tse sī tuā-lōo, lí kā gún phah tsuánn-iūnn?	3.267	4.400	+1.133
我字攏寫袂出來，敢講無水矣？	Guá jī lóng siá bē tshut–lâi, kám-kóng bô tsuí–ah?	3.433	4.700	+1.267
水捙倒捙甲規桌頂矣啦你。	Tsuí tshia-tó tshia kah kui toh-tíng–ah–lah lí.	3.400	4.667	+1.267
嫌犯毋捌貨，共六十萬籠的小提琴俗賣。	Hiâm-huān m̄ bat-huè, kā lak-tsap-bān khoo ê sió-thê-khîm siok-bē.	2.933	4.433	+1.500
阿生今仔日有去公司。	A-sing kin-á-jit ū khì kong-si.	3.000	4.533	+1.533
無我叫伊順紲幹過來載你？	Bô guá kiò i sīn-suà uat kuè-lâi tsài–lí?	3.267	4.867	+1.600
多謝你替阮共阿寶揣轉來。	To-siā lí thè gún kā A-pó tshuē–tńg-lâi.	2.633	4.700	+2.067