Polyglot-Lion: Efficient Multilingual ASR for Singapore via Balanced Fine-Tuning of Qwen3-ASR

Quy-Anh Dang
Knovel Engineering Lab, Singapore
`quyanh.dang@knoveleng.com`

Chris Ngo
Knovel Engineering Lab, Singapore
`chris.ngo@knoveleng.com`

Project Page **Code** **Models**

Abstract

We present **Polyglot-Lion**, a family of compact multilingual automatic speech recognition (ASR) models tailored for the linguistic landscape of Singapore, covering English, Mandarin, Tamil, and Malay. Our models are obtained by fine-tuning `Qwen3-ASR-0.6B` and `Qwen3-ASR-1.7B` exclusively on publicly available speech corpora, using a balanced sampling strategy that equalizes the number of training utterances per language and deliberately omits language-tag conditioning so that the model learns to identify languages implicitly from audio. On 12 benchmarks spanning the four target languages, **Polyglot-Lion-1.7B** achieves an average error rate of **14.85**, competitive with MERaLiON-2-10B-ASR (14.32) - a model 6× larger - while incurring a training cost of **$81** on a single RTX PRO 6000 GPU. Inference throughput is approximately **20× faster** than MERaLiON at 0.10 s/sample versus 2.02 s/sample. These results demonstrate that linguistically balanced fine-tuning of moderate-scale pretrained models can yield deployment-ready multilingual ASR at a fraction of the cost of larger specialist systems.

1 Introduction

Singapore presents a uniquely demanding setting for automatic speech recognition (ASR): four official languages - English, Mandarin Chinese, Tamil, and Malay - coexist in everyday communication, often within a single conversation or utterance. This linguistic landscape is further complicated by the prevalence of *Singlish*, a creole variety that draws lexical and phonological material from all four languages, and by wide variation in speaker age, accent, and code-switching behaviour (Lim, 2004). Together, these factors make Singapore one of the most challenging real-world environments for multilingual ASR.

Despite this linguistic richness, high-quality open-source ASR systems that cover all four official languages simultaneously remain scarce. General-purpose multilingual models such as Whisper (Radford et al., 2023) and MMS (Pratap et al., 2024) provide broad language coverage through large-scale pretraining, but their accuracy degrades on lower-resource varieties such as Tamil and Malay and on Singapore-accented English (Koh et al., 2019). Audio-language models (ALMs) such as Qwen2.5-Omni (Xu et al., 2025) and SeaLLMs-Audio (Liu et al., 2025) extend speech recognition with general language understanding, yet their large parameter counts (7B+) render fine-tuning and deployment expensive. Specialist systems such as MERaLiON-2-10B-ASR (He et al., 2025) have been purpose-built for the Singapore multilingual setting and achieve strong performance across all four languages, but require 128 GPUs and an estimated $18,862 to train - a barrier that places them beyond the reach of most academic groups and small enterprises.

In this paper, we introduce **Polyglot-Lion** (*Poly*: many; *Glot*: tongue; *Lion*: the lion-city, Singapore), a family of compact multilingual ASR models built by fine-tuning `Qwen3-ASR-0.6B` and `Qwen3-ASR-1.7B` (Shi et al., 2026) exclusively on publicly available speech corpora. As illustrated in Figure 1, Polyglot-Lion-1.7B achieves an average error rate of 14.8 across 12 benchmarks - closely matching MERaLiON-2-10B-ASR (14.3) while running nearly 20× faster at inference time. This is accomplished through two simple but effective design choices: (1) a balanced sampling strategy that equalises per-language training coverage, and (2) the deliberate removal of language-tag conditioning, forcing the model to detect the spoken language directly from the acoustic signal.

Our contributions are as follows:

1. **A balanced multilingual fine-tuning recipe** that upsamples under-represented languages to achieve equal per-language training cov-

Proceedings of the 1st Workshop on Multilinguality in the Era of Large Language Models (MeLLM 2026), pages 191–200
July 4, 2026

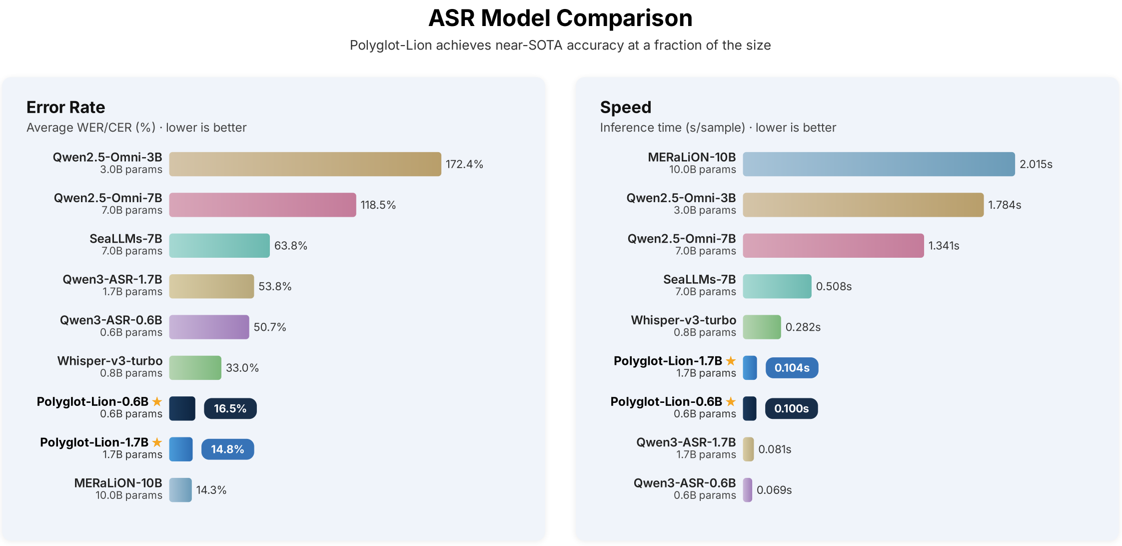

Figure 1: Polyglot-Lion achieves near-SOTA accuracy at a fraction of the model size and inference cost. **Left:** Average error rate (WER/CER) across 12 benchmarks; lower is better. **Right:** Inference speed in seconds per sample; lower is better. Despite having 6× fewer parameters than MERaLiON-2-10B-ASR, Polyglot-Lion-1.7B matches its accuracy while being approximately 20× faster at inference.

erage, substantially improving recognition accuracy on low-resource languages (Tamil, Malay) without requiring any proprietary data.

2. **Language-agnostic decoding**: by omitting explicit language-tag conditioning at both training and inference time, Polyglot-Lion identifies the spoken language implicitly from acoustic features alone, making it robust to the code-switching patterns prevalent in Singapore speech.
3. **Comprehensive multilingual benchmarking** across 12 standard datasets spanning all four official languages of Singapore, with direct quantitative comparison against eight published baselines ranging from general-purpose models to large specialist systems.
4. **A cost-efficiency analysis** demonstrating that Polyglot-Lion achieves near state-of-the-art accuracy at over 233× lower estimated training cost ($81 on a single GPU versus $18,862 on 128 GPUs) and approximately 20× faster inference than the strongest comparably accurate baseline, MERaLiON-2-10B-ASR.

2 Related Work

Large-scale multilingual ASR. Whisper (Radford et al., 2023) demonstrated that weakly supervised training at scale can yield robust multilingual recognition across 99 languages. Self-supervised approaches such as wav2vec 2.0 (Baevski et al., 2020) and HuBERT (Hsu et al., 2021) further reduced labelled data requirements, while MMS (Pratap et al., 2024) extended coverage to over 1,000 languages via religious audio recordings. Despite their breadth, all these systems degrade on typologically distant, low-resource languages such as Tamil and Malay, and on non-native or regional accents.

Audio-language models. Integrating speech encoders with LLM decoders — as in SALMONN (Tang et al., 2024), Qwen-Audio (Chu et al., 2023), Qwen2.5-Omni (Xu et al., 2025), and SeaLLMs-Audio (Liu et al., 2025) — leverages rich linguistic priors for strong ASR accuracy. The Qwen3-ASR series (Shi et al., 2026) distils these capabilities into compact 0.6B–1.7B checkpoints, which we use as our base models. However, performance on Southeast Asian languages remains variable due to limited regional representation in pretraining corpora.

Southeast Asian and Singapore ASR. SEA-LION (Ong and Limkonchotiwat, 2023) and SeaLLMs (Nguyen et al., 2024) highlighted the importance of region-specific data curation.

MERaLiON-2[1] is the strongest published system for Singapore multilingual ASR, unifying all four official languages in a 10B-parameter model trained on over 120,000 hours. While it serves as our primary reference point, its 128-GPU, $18,862 training requirement motivates the pursuit of more accessible alternatives.

Multilingual training balance and language-agnostic ASR. Language imbalance causes multilingual models to underfit low-resource languages (Conneau et al., 2020; Arivazhagan et al., 2019). Language-balanced training has been shown to yield consistent WER reductions across resource levels without degrading high-resource languages (Zhou et al., 2022). We adopt deterministic upsampling as a hyper-parameter-free alternative to temperature sampling. Additionally, rather than conditioning on explicit language tokens (Radford et al., 2023) — which fails under code-switching (Winata et al., 2021) — we train without language tags, following earlier language-agnostic multilingual ASR work (Toshniwal et al., 2018).

3 Datasets

We train and evaluate exclusively on publicly available speech corpora, covering all four official languages of Singapore: English, Mandarin Chinese, Tamil, and Malay. Table 1 provides a full breakdown of each corpus by split and duration.

English. We include two English corpora. **Librispeech** (Panayotov et al., 2015) is a widely used benchmark of read English speech derived from public-domain audiobooks, providing 100.59 hours of clean training audio. **NSC** (National Speech Corpus) (Koh et al., 2019) is a large-scale Singapore English corpus collected across multiple speaking styles and demographics, contributing 147.97 training hours and covering the accent and prosodic characteristics distinctive to Singapore English.

Mandarin. Four Mandarin corpora are included. **AISHELL-1** (Bu et al., 2017) provides 150.85 hours of standard Mandarin read speech from 400 speakers. **AISHELL-3** (Shi et al., 2021) is a multi-speaker corpus originally designed for text-to-speech synthesis but widely used for ASR training, contributing 56.86 hours. **Common Voice 23** (Ardila et al., 2020) supplies 42.43 hours of crowdsourced Chinese speech with diverse speaker demographics. **Fleurs** (Conneau et al., 2023) adds 9.73 hours of read speech drawn from the FLoRes-200 translation benchmark, providing clean and consistently formatted audio across languages.

Tamil. Four Tamil corpora are used. **SLR127** (A et al., 2022b,a) is the largest Tamil source with 119.86 training hours, containing read and semi-spontaneous Tamil speech. **Common Voice 23** (Ardila et al., 2020) contributes 81.38 hours of crowdsourced Tamil recordings. **SLR65** (He et al., 2020) provides 5.66 hours of high-quality read Tamil speech. **Fleurs** (Conneau et al., 2023) adds 8.68 hours of clean read Tamil audio. Tamil is the most under-represented language in the pre-training data of most existing ASR systems, making these corpora critical for fine-tuning coverage.

Malay. Two Malay corpora are included. **Mesolitica**[3] is a Malaysian Malay speech corpus with 49.43 training hours spanning multiple domains and speaking styles. **Fleurs** (Conneau et al., 2023) contributes 9.55 hours of clean read Malay speech. Despite being an official language of Singapore, Malay is severely under-represented in existing multilingual ASR benchmarks, making the Mesolitica corpus a particularly valuable resource.

Data statistics and imbalance. As shown in Table 1, the combined corpus totals 607,839 utterances and 968.83 hours of audio. However, the training partition is substantially imbalanced across languages: English and Mandarin together account for approximately 65% of all training hours (248.56 and 259.87 hours respectively), while Malay contributes only 58.98 hours - less than 8% of the total. Tamil, despite having four contributing corpora and 215.58 training hours, is typologically distant from the languages dominating the base model's pretraining data, compounding the effective imbalance at the representation level. Without correction, joint training on this skewed distribution would bias gradient updates towards high-resource languages and degrade recognition performance on Tamil and Malay (Arivazhagan et al., 2019; Wang et al., 2020a). We address this through explicit language-balanced upsampling, described in Section 4.

Preprocessing. All corpora are preprocessed with a uniform pipeline prior to training. Audio

---

[1] `https://huggingface.co/collections/MERaLiON/meralion-2`

[3] `https://github.com/malaysia-ai/malaysian-dataset/tree/master/text-to-speech/emilia`

Table 1: Dataset statistics by language, split, and duration (S = number of samples, H = hours).

Lang.	Dataset	Train		Valid		Test		Total	
		S	H	S	H	S	H	S	H
English	Librispeech (Panayotov et al., 2015)	28,539	100.59	2,700	5.36	2,619	5.39	33,858	111.34
	NSC	100,000	147.97	2,997	4.90	3,000	4.95	105,997	157.82
Mandarin	AISHELL-1 (Bu et al., 2017)	120,098	150.85	14,326	18.09	7,176	10.03	141,600	178.97
	AISHELL-3 (Shi et al., 2021)	56,936	56.86	6,326	6.31	24,773	22.45	88,035	85.62
	Fleurs (Conneau et al., 2023)	3,246	9.73	409	1.27	945	3.07	4,600	14.07
	Common Voice 23 (Ardila et al., 2020)	29,473	42.43	10,635	15.95	9,999	16.43	50,107	74.81
Tamil	Fleurs (Conneau et al., 2023)	2,366	8.68	376	1.25	591	2.13	3,333	12.06
	SLR127 (A et al., 2022b,a)	69,575	119.86	7,731	13.41	12,086	16.80	89,392	150.07
	SLR65 (He et al., 2020)	3,427	5.66	428	0.72	429	0.69	4,284	7.07
	Common Voice 23 (Ardila et al., 2020)	45,186	81.38	9,964	15.71	7,907	12.28	63,057	109.37
Malay	Mesolitica[2]	17,851	49.43	992	2.71	993	2.75	19,836	54.89
	Fleurs (Conneau et al., 2023)	2,667	9.55	324	0.93	749	2.26	3,740	12.74
Total	—	479,364	782.99	57,208	86.61	71,267	99.23	607,839	968.83

files exceeding 30 seconds are discarded to avoid memory overflow during training and to exclude utterances that are disproportionately long relative to the target sequence length of most ASR decoders (Radford et al., 2023). Transcripts are normalised to lowercase and stripped of punctuation, following the convention adopted by Whisper (Radford et al., 2023) and subsequent multilingual ASR systems (Shi et al., 2026), which has been shown to reduce spurious token-level errors arising from inconsistent punctuation annotation across corpora (Likhomanenko et al., 2021). No speaker-level filtering or data selection is applied; all remaining utterances are used.

4 Method

4.1 Base Models

Polyglot-Lion is fine-tuned from two publicly available checkpoints in the Qwen3-ASR series (Shi et al., 2026): `Qwen3-ASR-0.6B` and `Qwen3-ASR-1.7B`. These models follow a transformer-based encoder–decoder architecture (Vaswani et al., 2017) in which a Conformer (Gulati et al., 2020) or similar acoustic encoder maps log-Mel filterbank features to contextual representations, and an autoregressive decoder generates output tokens conditioned on those representations. Both checkpoints are pre-trained on large-scale multilingual speech data and already achieve competitive zero-shot performance on several standard benchmarks (Shi et al., 2026), providing a strong initialisation for fine-tuning.

4.2 Balanced Multilingual Sampling

Motivation. As noted in Section 3, the raw training corpus is heavily skewed: English and Mandarin collectively account for approximately 65% of all training utterances, while Malay represents fewer than 8%. Naive joint training on this distribution would cause the model to overfit high-resource languages and underfit low-resource ones (Arivazhagan et al., 2019; Wang et al., 2020a), a well-documented failure mode in multilingual learning. Rather than adopting temperature-based multinomial sampling (Arivazhagan et al., 2019) - which introduces a sensitive temperature hyperparameter and still does not guarantee exact language parity - we adopt a two-stage deterministic upsampling strategy that first balances datasets *within* each language group, and then balances language groups *against one another.*

Two-stage upsampling. Let $\mathcal{L} = \{l_1, l_2, l_3, l_4\}$ denote the set of four languages, and let $\mathcal{D}_l = \{D_{l,1}, \ldots, D_{l,K_l}\}$ be the collection of K_l datasets for language l. We write $N_{l,k} = |D_{l,k}|$ for the number of training utterances in dataset $D_{l,k}$.

Stage 1 - Intra-language balancing. Within each language l, we upsample every dataset to match the largest dataset in that language group:

$$N_l^* = \max_k N_{l,k}, \qquad r_{l,k} = \frac{N_l^*}{N_{l,k}} \tag{1}$$

Each dataset $D_{l,k}$ is replicated $r_{l,k}$ times and then randomly subsampled to exactly N_l^* utterances,

Algorithm 1 Two-Stage Balanced Multilingual Upsampling

```
Require: Language set $\mathcal{L}$; per-language dataset collections $\{\mathcal{D}_l\}_{l\in\mathcal{L}}$
Ensure: Balanced training corpus $\hat{\mathcal{D}}$ with equal samples per language
                                        // Stage 1: Intra-language balancing
for each language $l \in \mathcal{L}$ do
    $N_l^* \leftarrow \max_k |D_{l,k}|$    // largest dataset in language l
    for each dataset $D_{l,k} \in \mathcal{D}_l$ do
        $r_{l,k} \leftarrow \lceil N_l^* / |D_{l,k}| \rceil$
        $D_{l,k} \leftarrow$ Replicate$(D_{l,k},\ r_{l,k})$
        $D_{l,k} \leftarrow$ RandomSubsample$(D_{l,k},\ N_l^*)$
    end for
    $\tilde{D}_l \leftarrow \bigcup_k D_{l,k}$    // balanced corpus for language l, size $N_l^*$
end for
                                        // Stage 2: Inter-language balancing
$N^{**} \leftarrow \max_l N_l^*$    // largest per-language corpus after Stage 1
for each language $l \in \mathcal{L}$ do
    $R_l \leftarrow \lceil N^{**} / N_l^* \rceil$
    $\hat{D}_l \leftarrow$ Replicate$(\tilde{D}_l,\ R_l)$
    $\hat{D}_l \leftarrow$ RandomSubsample$(\hat{D}_l,\ N^{**})$
end for
$\hat{\mathcal{D}} \leftarrow \bigcup_{l\in\mathcal{L}} \hat{D}_l$ return $\hat{\mathcal{D}}$    // $|\hat{\mathcal{D}}| = 4 \times N^{**}$; each language = 25%
```

yielding a balanced per-language corpus $\tilde{D}_l$ of size N_l^*.

Stage 2 - Inter-language balancing. After Stage 1, each language l has N_l^* utterances, but these totals still differ across languages. We therefore upsample each language to match the largest language group:

$$N^{**} = \max_l N_l^*, \qquad R_l = \frac{N^{**}}{N_l^*} \tag{2}$$

Each balanced corpus $\tilde{D}_l$ is replicated R_l times and subsampled to exactly N^{**} utterances, yielding a final per-language corpus $\hat{D}_l$ of uniform size N^{**}.

The final training set is the union $\hat{\mathcal{D}} = \bigcup_l \hat{D}_l$, which contains exactly $4 \times N^{**}$ utterances with each language contributing precisely 25%. Algorithm 1 presents the full procedure.

This strategy is deliberately simple: it requires no hyper-parameter tuning, is fully deterministic given a fixed random seed, and guarantees exact per-language parity regardless of how skewed the original corpus distribution is. The cost is a modest increase in the number of training steps per epoch, which is outweighed by the improvement in low-resource language coverage demonstrated in Section 6.

4.3 Language-Agnostic Transcription

A standard practice in multilingual ASR systems is to prepend a special language-identification token to the decoder input at both training and inference time (Radford et al., 2023; Li et al., 2019). While this conditioning signal improves accuracy when the spoken language is known *a priori*, it introduces a critical dependency: if the language tag is absent, incorrect, or ambiguous - as is common in spontaneous conversational speech and code-switched utterances (Winata et al., 2021) - recognition quality degrades sharply.

Singapore's multilingual environment makes this dependency particularly problematic. Speakers routinely alternate between English, Mandarin, Tamil, and Malay within a single interaction, and in many deployment settings (e.g., broadcast media monitoring, classroom transcription, customer service) the language of each audio segment is not known in advance. We therefore train Polyglot-Lion entirely without language conditioning: no language tags are prepended to decoder inputs at training time, and none are expected at inference time. The model is required to infer the spoken language implicitly from acoustic and linguistic patterns in the input signal, following the approach explored in earlier language-agnostic multilingual ASR work (Toshniwal et al., 2018).

This design choice is validated empirically in Section 6: Polyglot-Lion achieves strong recognition accuracy across all four languages despite receiving no explicit language signal, demonstrating that balanced fine-tuning is sufficient to induce reliable implicit language identification in a moderate-scale model.

4.4 Training Details

Both model variants are fine-tuned for 48 hours on a single NVIDIA RTX PRO 6000 GPU (48 GB VRAM). We use the AdamW optimiser (Loshchilov and Hutter, 2019) with a cosine annealing learning-rate schedule (Loshchilov and Hutter, 2017), a peak learning rate of 2×10^{-5}. Training uses a per-device batch size of 8 utterances accumulated over 4 gradient accumulation steps, yielding an effective batch size of 32. All other hyperparameters follow the defaults from the Qwen3-ASR fine-tuning configuration (Shi et al., 2026).

5 Experimental Setup

5.1 Evaluation Metrics

We adopt two standard ASR evaluation metrics, selected according to the linguistic properties of each target language:

- **Word Error Rate (WER)** for English, Tamil, and Malay, where whitespace-delimited word tokenisation is conventional. WER is computed as the minimum edit distance (substitutions S, deletions D, insertions I) between the hypothesis and reference, normalised by the number of reference words N: WER $= (S + D + I)/N$.
- **Character Error Rate (CER)** for Mandarin Chinese, where the absence of explicit word boundaries makes character-level evaluation more appropriate and widely adopted (Shi et al., 2021; Bu et al., 2017).

All hypotheses and references are lowercased and stripped of punctuation prior to scoring, consistent with the preprocessing applied during training (Section 3). Evaluation is performed using the `asr-evalkit` library (Dang, 2026). Lower values indicate better performance in both metrics.

5.2 Baselines

We compare Polyglot-Lion against eight published or widely-used ASR systems, selected to represent the full spectrum from lightweight general-purpose models to large specialist systems:

1. **Whisper-large-v3-turbo** (Radford et al., 2023): a distilled and optimised variant of Whisper-large-v3 that retains strong multilingual accuracy with reduced inference cost. It serves as the canonical general-purpose multilingual ASR baseline.
2. **SeaLLMs-Audio-7B** (Liu et al., 2025): a 7B-parameter audio-language model specifically developed for Southeast Asian languages, built on top of the SeaLLMs language model backbone (Nguyen et al., 2024).
3. **Qwen2.5-Omni-3B** and **Qwen2.5-Omni-7B** (Xu et al., 2025): general-purpose omni-modal models integrating vision, audio, and language understanding within a unified framework. Included to assess how general ALMs perform on regional multilingual ASR without task-specific fine-tuning.
4. **Qwen3-ASR-0.6B** and **Qwen3-ASR-1.7B** (Shi et al., 2026): the unmodified base checkpoints from which our models are fine-tuned. Including these baselines allows direct quantification of the accuracy gains attributable to our balanced fine-tuning recipe, independent of the base model capacity.
5. **MERaLiON-2-10B-ASR** (He et al., 2025): a 10B-parameter model purpose-built for Singapore multilingual ASR and trained on over 120,000 hours of speech data across English, Mandarin, Tamil, and Malay. This model represents the strongest publicly available specialist system for our target setting and serves as our primary comparison point.

All baselines are evaluated in inference-only mode using their publicly released checkpoints without any additional fine-tuning. Inference for all models is conducted on the same hardware (single NVIDIA RTX PRO 4500 GPU) to ensure fair latency comparisons.

6 Results

6.1 Recognition Accuracy

Table 2 reports per-benchmark and average error rates for all systems. **Polyglot-Lion-1.7B** achieves an average error rate of **14.85**, closely matching MERaLiON-2-10B-ASR (14.32) - a model 6× larger - and ranking second overall across all 12 benchmarks. **Polyglot-Lion-0.6B** achieves an average of 16.52, making it the best-performing model at or below 1B parameters by a substantial margin (next best: Whisper-large-v3-turbo at 33.04). We discuss per-language findings below.

English. On Librispeech, Polyglot-Lion-1.7B achieves **2.10 WER**, surpassing both MERaLiON-2-10B-ASR (2.54) and the unmodified Qwen3-ASR-1.7B base (2.31), and setting the best result among all evaluated systems on this benchmark. On NSC - a Singapore English corpus that captures regional accents, pronunciation patterns, and speaking styles not present in Librispeech - Polyglot-Lion-1.7B achieves 5.28 WER, a dramatic improvement over Whisper-large-v3-turbo (32.02) and substantially better than the Qwen3-ASR base (6.22). MERaLiON-2-10B-ASR achieves the best NSC result (4.62), which we attribute to its larger capacity and inclusion of Singapore-specific training material beyond our public-only corpus. Notably,

Table 2: ASR evaluation results. WER (%) for English, Tamil, and Malay; CER (%) for Mandarin. Lower is better. **Bold** = best overall; underline = second best. Rows shaded in blue are our proposed models. Dashes indicate results excluded from the average due to anomalously high error rates (WER > 200%) that would distort cross-system comparison.

Model	**Params**	**English**		**Mandarin (CER)**				**Tamil (WER)**				**Malay**		**Avg**
		LS	*NSC*	*CV*	*AISH1*	*AISH3*	*Fleurs*	*CV*	*SLR65*	*SLR127*	*Fleurs*	*Meso.*	*Fleurs*	
Whisper-large-v3-turbo	0.8B	3.04	32.02	17.91	9.64	16.81	10.63	74.50	58.13	69.56	66.90	28.47	8.88	33.04
SeaLLMs-Audio-7B	7B	94.74	9.53	8.68	9.65	9.76	37.09	126.70	127.24	138.65	105.31	71.34	26.25	63.75
Qwen2.5-Omni-3B	3B	29.21	34.79	46.36	28.25	44.55	54.74	318.36	465.58	448.82	311.67	211.90	74.69	172.37
Qwen2.5-Omni-7B	7B	13.80	22.96	14.49	7.33	22.58	16.68	252.06	239.15	303.96	326.43	158.06	43.92	118.45
Qwen3-ASR-0.6B	0.6B	2.74	7.64	10.06	2.08	2.59	9.75	121.10	127.00	129.12	130.09	47.29	18.71	50.68
Qwen3-ASR-1.7B	1.7B	2.31	6.22	7.50	1.52	2.08	9.33	139.96	134.63	144.49	147.23	39.00	10.87	53.76
MERaLiON-2-10B-ASR	10B	2.54	**4.62**	8.83	3.09	4.07	11.99	**31.78**	**19.29**	**22.42**	**28.68**	25.90	**8.55**	**14.32**
Polyglot-Lion-0.6B	**0.6B**	2.67	6.09	6.16	1.93	2.32	9.19	42.16	23.07	28.14	37.68	24.33	14.45	16.52
Polyglot-Lion-1.7B	**1.7B**	**2.10**	5.28	**4.91**	**1.45**	**1.86**	**8.00**	39.19	19.75	26.83	37.28	**21.51**	9.98	14.85

SeaLLMs-Audio-7B yields a very high 94.74 WER on Librispeech despite reasonable performance on NSC, suggesting that its training prioritised conversational rather than read speech.

Mandarin. Polyglot-Lion-1.7B achieves the **lowest CER on all four Mandarin benchmarks**, including AISHELL-1 (**1.45**), AISHELL-3 (**1.86**), Common Voice (**4.91**), and Fleurs (**8.00**), outperforming even MERaLiON-2-10B-ASR across the board (3.09, 4.07, 8.83, 11.99 respectively). Polyglot-Lion-0.6B similarly leads among sub-1B models with 6.16 CER on Common Voice. The strong Mandarin results are consistent with the Qwen3-ASR base models already encoding rich Chinese language priors from pretraining; our balanced fine-tuning preserves and refines these priors rather than degrading them through interference from other languages.

Tamil. Tamil is the most challenging language in this evaluation, reflecting its typological distance from the Indo-European and Sino-Tibetan languages that dominate most ASR pretraining corpora (Pratap et al., 2024). The unmodified Qwen3-ASR base models produce extremely high error rates on Tamil (WER > 120% on Common Voice), confirming severely limited Tamil exposure at pretraining. After balanced fine-tuning, Polyglot-Lion-1.7B reduces Tamil CV WER from 139.96 to **39.19** - a relative reduction of 72% - and achieves competitive results on SLR65 (19.75), SLR127 (26.83), and Fleurs (37.28). MERaLiON-2-10B-ASR remains the best system on all four Tamil benchmarks, which we attribute to its 6× larger capacity and likely inclusion of larger Tamil-specific training data. Closing this gap is an important direction for future work.

Malay. On Mesolitica, Polyglot-Lion-1.7B achieves **21.51 WER**, the best result among all evaluated systems, outperforming MERaLiON-2-10B-ASR (25.90), Whisper (28.47), and all other baselines by a clear margin. On Malay Fleurs, Polyglot-Lion-1.7B (9.98 WER) is competitive with MERaLiON (8.55) and Whisper (8.88). The strong Mesolitica result is particularly encouraging as it reflects performance on conversational and domain-diverse Malay speech, which is more representative of real-world deployment conditions than the read-speech Fleurs benchmark.

Table 3: Inference latency (seconds per sample, mean ± std) measured on a single NVIDIA RTX PRO 4500 GPU. Our models are shaded in blue.

Model	**Time (s/sample)**
MERaLiON-2-10B-ASR	2.0152 ± 0.8846
Qwen2.5-Omni-3B	1.7838 ± 1.0431
Qwen2.5-Omni-7B	1.3414 ± 0.6572
SeaLLMs-Audio-7B	0.6422 ± 0.0000
Whisper-large-v3-turbo	0.2822 ± 0.0230
Qwen3-ASR-1.7B	0.0809 ± 0.0290
Qwen3-ASR-0.6B	0.0686 ± 0.0251
Polyglot-Lion-0.6B	0.0999 ± 0.0561
Polyglot-Lion-1.7B	0.1038 ± 0.0621

Effect of fine-tuning. A direct comparison between Polyglot-Lion and the unmodified Qwen3-ASR base models isolates the contribution of our balanced fine-tuning recipe. The benefit is most pronounced for under-represented languages: on Tamil CV, fine-tuning reduces WER by 65% (0.6B: 121.10 → 42.16) and 72% (1.7B: 139.96 → 39.19). On Malay Mesolitica, the reduction is 49% (0.6B: 47.29 → 24.33) and 45% (1.7B: 39.00 → 21.51). Performance on English and Mandarin is preserved or improved, confirming that balanced upsampling

Table 4: Training resource and cost comparison. GPU rental prices sourced from RunPod.io.

	MERaLiON-2-10B	**Polyglot-Lion**
Training Data (Hours)	120,000	782.99
Hardware	128 × H100	1 × RTX PRO 6000
Training Time	48 h	48 h
Est. Cost	**$18,862**	**$81**

does not introduce negative transfer (Wang et al., 2020b) on high-resource languages.

6.2 Inference Speed

Table 3 reports mean inference latency per sample, measured on a single NVIDIA RTX PRO 4500 GPU across all evaluation sets. Polyglot-Lion-0.6B and Polyglot-Lion-1.7B process audio at **0.10** and **0.10 s/sample** respectively - approximately **20× faster** than MERaLiON-2-10B-ASR (2.02 s/sample) and **3× faster** than Whisper-large-v3-turbo (0.28 s/sample). The Qwen2.5-Omni models exhibit high latency variance (std $>$ 0.6 s), likely due to their omni-modal routing overhead.

6.3 Training Cost Comparison

Table 4 compares the estimated training costs of Polyglot-Lion and MERaLiON-2-10B-ASR. MERaLiON-2-10B-ASR was trained on approximately 120,000 hours of speech using 128 H100 GPUs for 48 hours; we estimate its cost based on current H100 cloud rental rates via RunPod[4]. Polyglot-Lion is trained on 782.99 hours using a single RTX PRO 6000 GPU for the same wall-clock duration, with cost estimated from the same platform.

Polyglot-Lion incurs an estimated training cost of **$81**, representing a **233× cost reduction** relative to MERaLiON-2-10B-ASR ($18,862), while achieving a comparable average error rate (14.85 vs. 14.32). This cost advantage has significant practical implications: the ability to fine-tune a near-SOTA multilingual ASR system on a single consumer GPU within two days makes iterative development, ablation studies, low-resource language adaptation, and domain specialisation accessible to academic research groups and resource-constrained organisations that would otherwise be unable to develop competitive Singapore multilingual ASR systems.

7 Conclusion

We have presented Polyglot-Lion, a family of compact multilingual ASR models for Singapore English, Mandarin, Tamil, and Malay. Through balanced multilingual fine-tuning of Qwen3-ASR base models on publicly available speech corpora, and by removing language-tag conditioning to enable fully implicit language identification, Polyglot-Lion-1.7B achieves an average error rate of 14.85 across 12 benchmarks - closely matching MERaLiON-2-10B-ASR (14.32) while requiring 6× fewer parameters, 20× faster inference, and 233× lower training cost. These results demonstrate that careful data balancing and lightweight fine-tuning of strong pretrained models can unlock near SOTA multilingual ASR performance at dramatically reduced computational expense, making high-quality Singapore multilingual ASR accessible to a wide research and deployment community.

References

Madhavaraj A, Bharathi Pilar, and Ramakrishnan A G. 2022a. Knowledge-driven subword grammar modeling for automatic speech recognition in tamil and kannada. *arXiv preprint*.

Madhavaraj A, Bharathi Pilar, and Ramakrishnan A G. 2022b. Subword dictionary learning and segmentation techniques for automatic speech recognition in tamil and kannada. *arXiv preprint*.

Rosana Ardila, Megan Branson, Kelly Davis, Michael Henretty, Michael Kohler, Josh Meyer, Reuben Morais, Lindsay Saunders, Francis M. Tyers, and Gregor Weber. 2020. Common voice: A massively-multilingual speech corpus. In *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020)*, pages 4211–4215.

Naveen Arivazhagan, Ankur Bapna, Orhan Firat, Dmitry Lepikhin, Melvin Johnson, Maxim Krikun, Mia Xu Chen, Yuan Cao, George Foster, Colin Cherry, Wolfgang Macherey, Zhifeng Chen, and Yonghui Wu. 2019. Massively multilingual neural machine translation in the wild: Findings and challenges. *Preprint*, arXiv:1907.05019.

Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: a framework for self-supervised learning of speech representations. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA. Curran Associates Inc.

---

[4] `https://www.runpod.io`

Hui Bu, Jiayu Du, Xingyu Na, Bengu Wu, and Hao Zheng. 2017. Aishell-1: An open-source mandarin speech corpus and a speech recognition baseline. In *2017 20th Conference of the Oriental Chapter of the International Coordinating Committee on Speech Databases and Speech I/O Systems and Assessment (O-COCOSDA)*, pages 1–5.

Yunfei Chu, Jin Xu, Xiaohuan Zhou, Qian Yang, Shiliang Zhang, Zhijie Yan, Chang Zhou, and Jingren Zhou. 2023. Qwen-audio: Advancing universal audio understanding via unified large-scale audio-language models. *Preprint*, arXiv:2311.07919.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Alexis Conneau, Min Ma, Simran Khanuja, Yu Zhang, Vera Axelrod, Siddharth Dalmia, Jason Riesa, Clara Rivera, and Ankur Bapna. 2023. Fleurs: Few-shot learning evaluation of universal representations of speech. In *2022 IEEE Spoken Language Technology Workshop (SLT)*, pages 798–805.

Quy-Anh Dang. 2026. Asr evalkit: A modular toolkit for evaluating automatic speech recognition models.

Anmol Gulati, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhengdong Zhang, Yonghui Wu, and Ruoming Pang. 2020. Conformer: Convolution-augmented Transformer for Speech Recognition. In *Interspeech 2020*, pages 5036–5040.

Fei He, Shan-Hui Cathy Chu, Oddur Kjartansson, Clara Rivera, Anna Katanova, Alexander Gutkin, Isin Demirsahin, Cibu Johny, Martin Jansche, Supheakmungkol Sarin, and Knot Pipatsrisawat. 2020. Open-source Multi-speaker Speech Corpora for Building Gujarati, Kannada, Malayalam, Marathi, Tamil and Telugu Speech Synthesis Systems. In *Proceedings of The 12th Language Resources and Evaluation Conference (LREC)*, pages 6494–6503, Marseille, France. European Language Resources Association (ELRA).

Yingxu He, Zhuohan Liu, Geyu Lin, Shuo Sun, Bin Wang, Wenyu Zhang, Xunlong Zou, Nancy F. Chen, and AiTi Aw. 2025. MERaLiON-AudioLLM: Advancing speech and language understanding for Singapore. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 22–30, Vienna, Austria. Association for Computational Linguistics.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. 2021. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3451–3460.

Jia Xin Koh, Aqilah Mislan, Kevin Khoo, Brian Ang, Wilson Ang, Charmaine Ng, and Ying-Ying Tan. 2019. Building the Singapore English National Speech Corpus. In *Interspeech 2019*, pages 321–325.

Bo Li, Yu Zhang, Tara Sainath, Yonghui Wu, and William Chan. 2019. Bytes are all you need: End-to-end multilingual speech recognition and synthesis with bytes. In *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5621–5625.

Tatiana Likhomanenko, Qiantong Xu, Jacob Kahn, Gabriel Synnaeve, and Ronan Collobert. 2021. Slimipl: Language-model-free data-light text normalization for ASR. In *Proceedings of Interspeech*.

Lisa Lim. 2004. *Singapore English: A grammatical description*, volume G33 of *Varieties of English Around the World*. John Benjamins Publishing Company, Amsterdam/Philadelphia.

Chaoqun Liu, Mahani Aljunied, Guizhen Chen, Hou Pong Chan, Weiwen Xu, Yu Rong, and Wenxuan Zhang. 2025. Seallms-audio: Large audio-language models for southeast asia. *Preprint*, arXiv:2511.01670.

Ilya Loshchilov and Frank Hutter. 2017. SGDR: Stochastic gradient descent with warm restarts. In *International Conference on Learning Representations*.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Xuan-Phi Nguyen, Wenxuan Zhang, Xin Li, Mahani Aljunied, Zhiqiang Hu, Chenhui Shen, Yew Ken Chia, Xingxuan Li, Jianyu Wang, Qingyu Tan, Liying Cheng, Guanzheng Chen, Yue Deng, Sen Yang, Chaoqun Liu, Hang Zhang, and Lidong Bing. 2024. SeaLLMs - large language models for Southeast Asia. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 294–304, Bangkok, Thailand. Association for Computational Linguistics.

David Ong and Peerat Limkonchotiwat. 2023. SEA-LION (Southeast Asian languages in one network): A family of Southeast Asian language models. In *Proceedings of the 3rd Workshop for Natural Language Processing Open Source Software (NLP-OSS 2023)*, pages 245–245, Singapore. Association for Computational Linguistics.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. Librispeech: An asr corpus based on public domain audio books. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5206–5210.

Vineel Pratap, Andros Tjandra, Bowen Shi, Paden Tomasello, Arun Babu, Sayani Kundu, Ali Elkahky, Zhaoheng Ni, Apoorv Vyas, Maryam Fazel-Zarandi, Alexei Baevski, Yossi Adi, Xiaohui Zhang, Wei-Ning Hsu, Alexis Conneau, and Michael Auli. 2024. Scaling speech technology to 1,000+ languages. *J. Mach. Learn. Res.*, 25(1).

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine Mcleavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 28492–28518. PMLR.

Xian Shi, Xiong Wang, Zhifang Guo, Yongqi Wang, Pei Zhang, Xinyu Zhang, Zishan Guo, Hongkun Hao, Yu Xi, Baosong Yang, Jin Xu, Jingren Zhou, and Junyang Lin. 2026. Qwen3-asr technical report. *Preprint*, arXiv:2601.21337.

Yao Shi, Hui Bu, Xin Xu, Shaoji Zhang, and Ming Li. 2021. AISHELL-3: A Multi-Speaker Mandarin TTS Corpus. In *Interspeech 2021*, pages 2756–2760.

Changli Tang, Wenyi Yu, Guangzhi Sun, Xianzhao Chen, Tian Tan, Wei Li, Lu Lu, Zejun MA, and Chao Zhang. 2024. Salmonn: Towards generic hearing abilities for large language models. In *The Twelfth International Conference on Learning Representations*.

Shubham Toshniwal, Tara N. Sainath, Ron J. Weiss, Bo Li, Pedro Moreno, Eugene Weinstein, and Kanishka Rao. 2018. Multilingual speech recognition with a single end-to-end model. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, page 4904–4908. IEEE Press.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Xinyi Wang, Yulia Tsvetkov, and Graham Neubig. 2020a. Balancing training for multilingual neural machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8526–8537, Online. Association for Computational Linguistics.

Zirui Wang, Zachary C. Lipton, and Yulia Tsvetkov. 2020b. On negative interference in multilingual models: Findings and a meta-learning treatment. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4438–4450, Online. Association for Computational Linguistics.

Genta Indra Winata, Samuel Cahyawijaya, Zihan Liu, Zhaojiang Lin, Andrea Madotto, and Pascale Fung. 2021. Are multilingual models effective in code-switching? In *Proceedings of the Fifth Workshop on Computational Approaches to Linguistic Code-Switching*, pages 142–153, Online. Association for Computational Linguistics.

Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Keqin Chen, Jialin Wang, Yang Fan, Kai Dang, Bin Zhang, Xiong Wang, Yunfei Chu, and Junyang Lin. 2025. Qwen2.5-omni technical report. *Preprint*, arXiv:2503.20215.

Shaohuan Zhou, Shun Lei, Weiya You, Deyi Tuo, Yuren You, Zhiyong Wu, Shiyin Kang, and Helen Meng. 2022. Towards Improving the Expressiveness of Singing Voice Synthesis with BERT Derived Semantic Information. In *Interspeech 2022*, pages 4292–4296.