On the Role of Encoder Depth: Pruning Whisper and LoRA Fine-Tuning in SLAM-ASR

Ganesh Pavan Kartikeya Bharadwaj Kolluri, Michael Kampouridis, Ravi Shekhar
School of Computer Science and Electronic Engineering, University of Essex
{karthik.kolluri, mkampo, r.shekhar}@essex.ac.uk

Abstract

Automatic speech recognition (ASR) has advanced rapidly in recent years, driven by large-scale pretrained models and end-to-end architectures such as SLAM-ASR. A key component of SLAM-ASR systems is the Whisper speech encoder, which provides robust acoustic representations. While model pruning has been explored for the full Whisper encoder–decoder architecture, its impact within the SLAM-ASR setting remains under-investigated. In this work, we analyze the effects of layer pruning in the Whisper encoder when used as the acoustic backbone of SLAM-ASR. We further examine the extent to which LoRA-based fine-tuning can recover performance degradation caused by pruning. Experiments conducted across three Whisper variants (Small, Medium, Large-v2), three languages representing distinct resource levels (Danish, Dutch, English), and over 200 training runs demonstrate that pruning two encoder layers causes only 2–4% WER degradation, and that combining this pruning with LoRA adaptation consistently outperforms the unpruned baseline while reducing total parameters by 7–14%. Moreover, our error analysis reveals that LoRA primarily compensates through the language model's linguistic priors, reducing total word errors by 11-21% for Dutch and English, with substitutions and deletions showing the largest reductions. However, for low-resource Danish, the reduction is smaller (4-7%), and LoRA introduces increased insertion errors, indicating that compensation effectiveness depends on the LLM's pre-existing language proficiency and available training data.

Keywords: Automatic Speech Recognition, Pruning, SpeechLLM, SLAM-ASR

1. Introduction

Recent advances in multimodal models have enabled automatic speech recognition systems (ASR) through architectures that connect pre-trained speech encoders to large language models (LLMs). SLAM-ASR (Ma et al., 2024; Zhang et al., 2026b,a) is the one such instance, which is a simpler approach and yet achieved comparative ASR performance: a frozen speech encoder extracts acoustic representations, and a small trainable projector maps these into the embedding space of a frozen LLM, which then generates transcriptions autoregressively. While this modular design is attractive for its simplicity, deploying such systems remains computationally expensive, as the full encoder must process every input audio regardless of whether all its layers contribute meaningfully to ASR performance.

Model compression through layer pruning has been widely studied for ASR. Kim et al. (2023) combined language-specific adapters with magnitude pruning to compress Whisper by up to 50%, still with competitive CER performance, while Gu et al. (2024) achieved 80% compression through one-shot unstructured pruning with minimal WER degradation. On the decoder side, Gandhi et al. (2023) distilled Whisper's decoder from 32 layers to 2 using knowledge distillation with large-scale pseudo-labeling, achieving 6 times faster inference within 1% WER of the original. However, these studies focused on traditional encoder-decoder ASR, where both the encoder and decoder are jointly trained. In the SLAM-ASR case, the encoder remains the same; the decoder is replaced by a frozen LLM that was never exposed to speech data and a lightweight projector that bridges the two components together. The effects of encoder layer pruning on this architecture remain unexplored. Layer-wise analyses suggest that upper encoder layers encode increasingly abstract linguistic features (Pasad et al., 2021), capabilities the LLM already possesses, raising the possibility that these layers are partially redundant. Separately, Low-Rank Adaptation (LoRA) (Hu et al., 2022) has been applied to LLM-based ASR for multilingual adaptation (Song et al., 2024; Fang et al., 2025; Concina et al., 2025) and low-resource improvement (Ma et al., 2024; Nagano et al., 2025; Fong et al., 2025; Burdisso et al., 2026), but no prior work has examined whether LoRA can compensate for performance lost through structured encoder layer pruning, or how data resource availability regulates these interactions.

This paper addresses these gaps through a systematic study of encoder layer pruning and LoRA compensation in SLAM-ASR [1]. Specifically, we focus on two research questions:

RQ1: *How does progressive encoder layer pruning affect ASR performance, and how does data resource availability modulate this effect?*

RQ2: *Can LoRA adaptation compensate for*

[1]The code is available at https://github.com/KarthikKolluriKB/SLAM-ASR-Encoder-Pruning-LoRA

Proceedings of SPEAKABLE @ LREC 2026, pages 183–193
11 May 2026.

pruning induced degradation?

To answer these questions, we evaluated across three languages representing distinct resource levels: Danish (4.2 hours, low-resource), Dutch (50 hours, medium-resource), and English (100 hours, high-resource), using three Whisper encoder variants (Small, Medium, Large-v2) (Radford et al., 2023) paired with Qwen2.5-3B LLM (Qwen et al., 2025). All three languages are supported by both the Whisper encoders and Qwen2.5-3B, ensuring that observed differences reflect data availability rather than missing language coverage. Our experiments cover pruning depths from the full encoder down to a single layer, with and without LoRA, spanning over 200 individual training runs.

Our main findings are as follows: First, removing the top one to two encoder layers results in only marginal WER degradation (with 2-4%) across all encoder scales and languages, supporting the hypothesis that upper encoder layers are partially redundant when an LLM handles linguistic processing. Beyond this arrangement, medium and high-resource languages degrade smoothly, while low-resource Danish exhibits erratic, non-monotonic behaviour. Second, combining modest pruning (two layers) with LoRA consistently outperforms the unpruned baseline while reducing total parameters by 7–14%, demonstrating that fewer encoder parameters paired with lightweight LLM adaptation can match or exceed full baseline model performance. Third, LoRA's compensatory benefit is not uniform across resource levels: it provides robust improvement for Dutch and English across all pruning depths, but is less effective for Danish, suggesting an interaction between data availability and adaptation capacity that needs to be further investigated. Additionally, error analysis reveals that LoRA compensates primarily through the LLM's linguistic priors, reducing word-level errors by 11–21% for Dutch and English, with substitution and deletion corrections showing the largest reductions, confirming a decoding-side compensation mechanism.

2. Related Works

This section reviews prior research in two areas relevant to our study: (1) layer pruning strategies for speech encoders, and (2) parameter-efficient fine-tuning for LLM-based speech systems.

2.1. Layer Pruning in Speech Encoders

Model compression through layer pruning has been extensively studied for ASR. Kim et al. (2023) introduced PEPSI (Parameter-Efficient Pruning and Adaptation for Speech Foundational Models), an adapt-and-prune framework for Whisper that combines language-specific adapters with iterative magnitude pruning. Their experiments on Common Voice showed that pruning can reduce model size by up to 50% while maintaining competitive CER across Korean and Malayalam. However, PEPSI employs *unstructured* pruning that removes individual weights rather than entire layers, resulting in sparse networks that require specialized hardware for efficient inference.

More recently, Gu et al. (2024) proposed SparseWAV, a one-shot unstructured pruning method for large speech foundation models, achieving up to 80% compression with minimal WER degradation. Irigoyen et al. (2025) revealed that certain encoder components actually *improve* when pruned, acting as implicit regularizers. Decoder self-attention at 50% sparsity achieved 2.38% absolute WER reduction on LibriSpeech test-other.

A parallel line of study focuses on *decoder* compression. Distil-Whisper (Gandhi et al., 2023) uses knowledge distillation (Hinton et al., 2015) with large-scale pseudo-labeling to compress the decoder from 32 layers to 2 while keeping the encoder frozen, achieving 6$\times$ faster inference within 1% WER of the original model. Similarly, BaldWhisper (Sy et al., 2025) targets low-resource deployment by merging decoder layer pairs in Whisper-base rather than removing them, achieving 2.15$\times$ faster inference with 48% size reduction while maintaining over 90% of baseline performance on Bambara speech data.

However, these studies focus on traditional encoder-decoder ASR, where Whisper's own decoder directly generates transcriptions. SLAM-ASR (Ma et al., 2024) presents a fundamentally different setting: the decoder is replaced by an LLM, and a lightweight projector that maps encoder outputs to the LLM embedding space. The effects of encoder layer pruning on this SLAM-ASR architecture remain unexplored. Here, the projector must learn to align potentially degraded encoder representations with a fixed, independently pre-trained LLM. Unlike traditional ASR, where encoder and decoder are jointly trained, SLAM-ASR offers no such flexibility: the LLM is frozen and was never exposed to speech, placing the entire burden of representation quality on the encoder.

2.2. Parameter-Efficient Fine-Tuning for LLM-based Speech Systems

Parameter-efficient fine-tuning (PEFT) methods (Houlsby et al., 2019) have become essential for adapting pre-trained large language models to specific downstream tasks without updating the entire model's parameters. Low-Rank Adaptation (LoRA) (Hu et al., 2022) has emerged as the dominant approach, decomposing weight updates into low-rank

matrices that typically account for less than 1% of total parameters.

In LLM-based ASR, LoRA has been applied across multiple components. Song et al. (2024) proposed LoRA-Whisper, which incorporates the LoRA matrix into Whisper for multilingual ASR to effectively mitigate language interference, achieving 18.5% relative WER reduction for multilingual ASR and 23.0% for language expansion while using only 5% of trainable parameters compared to full fine-tuning. Building on this idea, Mu et al. (2025) proposed HDMoLE, which combines multiple LoRA experts through hierarchical routing and dynamic thresholds for multi-accent LLM-based ASR. Their method achieves comparable CER to full fine-tuning on multi-accent and standard Mandarin datasets while using only 9.6% of the trainable parameters.

Within LLM-based ASR models, the SLAM-ASR framework (Ma et al., 2024) showed that applying LoRA to LLM attention layers improves low-resource ASR, especially when combined with partial encoder fine-tuning (Tang et al., 2023; Wu et al., 2023). However, Kumar et al. (2025) observed that SLAM-ASR generalizes poorly across domains, highlighting the need for effective adaptation strategies in real-world deployment.

Despite this progress, no prior work has examined whether LoRA can compensate for information lost through structured encoder layer removal. While PEPSI (Kim et al., 2023) combines LoRA with unstructured weight pruning, the effects of removing entire encoder layers and whether LoRA can recover from such architectural changes in SLAM-ASR systems remain unexplored. It is also unknown whether LoRA's compensatory effectiveness depends on the amount of available training data.

3. Methodology

This study investigates two strategies for improving the parameter efficiency of SLAM-ASR systems: encoder layer pruning and LoRA adaptation of the LLM. Specifically, we ask whether upper encoder layers can be removed without significant performance loss, given that the LLM is already proficient in linguistic tasks, and whether LoRA can compensate for any degradation introduced by pruning. We evaluated these strategies across three languages representing distinct resource levels to examine how data availability interacts with both pruning robustness and LoRA effectiveness.

We follow the SLAM-ASR framework (Ma et al., 2024), which connects a frozen pre-trained speech encoder to a frozen large language model through a small trainable projector module. Only projector parameters are updated during training, and both the encoder and LLM remain frozen. This modular design is crucial for our study because each component has a definite role; we can selectively modify the encoder through pruning and study how the system responds without disrupting other components. Figure 1 illustrates this architecture.

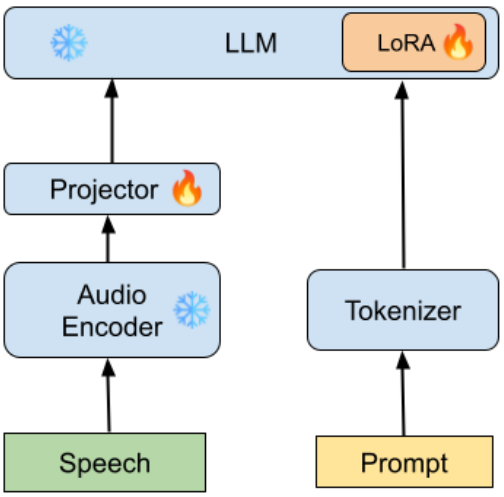

Figure 1: Overview of the SLAM-ASR architecture.

For encoder pruning, these previous studies on layer-wise analysis have shown that lower encoder layers primarily capture low-level acoustic features, while upper layers encode increasingly abstract linguistic information (Pasad et al., 2021). We hypothesize that in the SLAM-ASR architecture, these upper layers may be partially redundant, and that removing them offers a favourable efficiency–performance trade-off by delegating linguistic processing to the LLM. To verify this, we adopt a top-down pruning strategy: starting from the full encoder, we sequentially remove layers from the top, producing increasingly compressed configurations. For each configuration, we retrain only the projector from scratch, keeping both the pruned encoder and LLM frozen.

To investigate whether the LLM can be adapted to better handle degraded encoder representations, we apply Low-Rank Adaptation (LoRA) (Hu et al., 2022) to all attention projection matrices (Q, K, V, O) in the LLM. We hypothesize that, rather than recovering missing encoder information directly, LoRA enables the LLM to leverage its pre-trained linguistic knowledge: vocabulary, grammar, and contextual plausibility to adapt to the lost acoustic details. This creates a parameter trade-off: we remove millions of encoder parameters while adding far fewer LoRA parameters. We evaluate all combinations of pruning depths with and without LoRA, using Word Error Rate (WER) as the primary evaluation metric.

4. Experimental Setup

4.1. Datasets

We evaluate our approach on the Common Voice 22 corpus (Ardila et al., 2020), a crowdsourced multilingual speech dataset. To study how encoder pruning and LoRA compensation interact with training data availability, we select three languages that span an order-of-magnitude range in training data: Danish (4.2 hours), Dutch (54 hours), and English (100 hours). This selection is intentional; these three resource levels allow us to test whether pruning degradation patterns are consistent across data resource availability (RQ1) and whether LoRA compensation depends on sufficient fine-tuning data (RQ2). Dataset statistics are summarised in Table 1.

Table 1: Dataset statistics for each language from Common Voice 22.

Language	Split	Samples	Hours
Danish	Train	3,592	4.18
	Dev	2,511	3.21
	Test	2,684	3.45
Dutch	Train	43,458	54.15
	Dev	12,032	16.03
	Test	12,033	16.38
English	Train	58,140	100.00
	Dev	16,398	27.22
	Test	16,391	27.02

Preprocessing. Audio files are resampled to 16 kHz and converted to 80-channel log-Mel spectrograms following the Whisper preprocessing pipeline. We filter utterances to retain only those between 0.5 and 30 seconds in duration. Transcriptions are lowercased with punctuation removed, preserving apostrophes for contractions, following standard ASR preprocessing conventions.

4.2. Implementation Details

Model Architecture. We use Whisper (Radford et al., 2023), a transformer-based (Vaswani et al., 2017) speech encoder, as it is a widely adopted multilingual encoder available in multiple model sizes, making it well-suited for studying how pruning dynamics vary with encoder capacity. We pair it with Qwen2.5-3B (Qwen et al., 2025) as the LLM backbone, selected for its multilingual pre-training coverage, which is essential given that our experiments span three distinct languages. We experiment with three Whisper variants to study how pruning dynamics vary with model capacity. For all three Whisper variants: Small (12 layers), Medium (24 layers), and Large-v2 (32 layers), we evaluate configurations down to 2 layers in increments of 2.

Projector Architecture. We employ a two-layer MLP with concatenation-based temporal downsampling. The projector concatenates 5 consecutive frames (reducing the sequence length by a factor of 5), then applies two linear transformations with ReLU activation and a dropout rate of 0.1. The hidden dimension is 2048. We apply LayerNorm after the final projection to match the scale of LLM text embeddings.

LoRA Configuration. We apply Low-Rank Adaptation (Hu et al., 2022) to the query, key, value, and output projection matrices of all LLM attention layers. To prevent overfitting on smaller datasets, we use separate configurations for each resource level: r=8, α=16 for Danish (low-resource) and r=16, α=32 for Dutch/English (medium/high-resource). In preliminary experiments, using r=16 for Danish led to overfitting, motivating the reduced rank. All LoRA modules use a dropout of 0.1, adding approximately 0.8M (r=8) and 1.5M (r=16) trainable parameters, respectively.

4.3. Training Details

All models are trained using AdamW (Loshchilov and Hutter, 2019) with learning rate 1×10^{-4}, weight decay 0.01, and gradient clipping at 1.0. We employ a cosine learning rate schedule with linear warmup over the first 5% of training steps. Training uses mixed-precision (bfloat16) on a single NVIDIA A6000 GPU (48GB). Both the Whisper encoder and the Qwen2.5-3B base weights remain frozen; only the projector weights (and LoRA adapters when enabled) are updated.

4.4. Evaluation Details

We report Word Error Rate (WER) on the held-out test sets of each language. During evaluations, we use beam search with a beam size of 2. Text normalisation is applied to both predictions and references before evaluation: all text is lowercased, and punctuation is removed. For Danish, special characters (æ, ø, å) are preserved during normalisation to avoid penalising correct transcriptions. WER is computed using the `jiwer` library [2].

5. Results and Discussion

This section presents findings organized around two research questions: (1) How does progressive encoder layer pruning affect ASR performance, and how does data availability modulate this effect?

[2] https://github.com/jitsi/jiwer

(Section 5.1); and (2) Can LoRA adaptation compensate for pruning-induced degradation? (Section 5.2). We further provide an error analysis examining per-utterance trade-offs and the mechanism underlying LoRA compensation (Section 5.3)

5.1. RQ1: Effect of Encoder Layer Pruning

This section examines the impact of progressive removal of encoder layers on SLAM-ASR performance. By evaluating across three resource levels, we aim to understand how pruning depth interacts with data availability; specifically, whether resource-constrained settings exhibit different degradation patterns compared to data-rich conditions.

Our results broadly support the hypothesis that upper encoder layers are partially redundant in this architecture, given the LLM's existing linguistic capacity, as hypothesized in Section 3. Across all three Whisper variants (Small, Medium, Large-v2), removing the top one to two encoder layers results in only marginal WER degradation, with absolute increases remaining within 2–4% regardless of language (Table 2). We refer to this range as the safe pruning zone: the number of layers that can be removed before WER degrades beyond a practically meaningful threshold.

Beyond this zone, degradation diverges according to data resource availability. Dutch and English (medium and large resources) degrade smoothly and approximately monotonically as layers are removed. Danish (low resource) exhibits a qualitatively different pattern: rather than increasing monotonically, WER fluctuates erratically, spiking at certain depths and dropping at others. Notably, increasing the encoder capacity does not resolve this instability. Whisper-Large-v2, despite having nearly three times the layers of Whisper-Small, exhibits the same number of spikes for Danish, while Whisper-Medium shows the smoothest degradation curve among the three. Since Danish is the only language exhibiting this behaviour and also the language with the least training data, we hypothesize that limited data availability may be a contributing factor. However, further investigation, including controlled experiments with comparable amounts of training data across other languages, is needed to draw definitive conclusions.

5.2. RQ2: LoRA Compensation for Pruning

The previous section showed that shallow pruning is feasible, but deeper removal degrades performance. Here, we examine whether applying LoRA to the LLM can compensate for this degradation, and whether the resulting pruned+LoRA models can match or exceed unpruned baselines across data resource levels.

Our experimentation suggests that the answer is yes, within a shallow pruning regime, and that the compensation benefits generalize across encoder scales, though not equally across data resource levels. Table 3 summarizes what we term the sweet-spot configurations: the best-performing pruned+LoRA variant for each encoder scale, alongside the corresponding unpruned baseline and its LoRA-enhanced counterpart. Across all three Whisper variants, removing two encoder layers and applying LoRA to the LLM consistently outperforms the full unpruned baseline while reducing total parameter count by 7–14%. For Whisper-Small, the 10L+LoRA configuration achieves a 14.4% parameter reduction (from 88.15M to 75.5M) and improves WER on all three languages relative to the 12L baseline. The same pattern holds for Whisper-Medium (22L+LoRA, 7.6% reduction) and Whisper-Large-v2 (30L+LoRA, 5.9%), confirming that the finding is encoder, not scale-dependent.

Importantly, this improvement is not specific to pruned models. LoRA also improves the unpruned baseline encoder across all three languages (Table 2), indicating that it provides a general alignment benefit between the encoder and LLM, rather than simply patching information lost through layer removal. For Dutch and English, LoRA reduces WER across all pruning depths tested, from the full encoder to six layers. Even at aggressive pruning depths where base models produce severely degraded output, LoRA offers substantial recovery. For example, LoRA reduces Whisper-Large-v2 Danish WER from 130.99 (20L base) to 59.06 (20L+LoRA). However, for Danish, LoRA's compensation benefit is inconsistent at deeper pruning levels, which strengthens the finding from RQ1 that low-resource data conditions interact differently with both pruning and LoRA adaptation.

Since LoRA also improves the unpruned baseline (Table 2), a natural question is whether LoRA specifically compensates for pruning damage or simply provides a general performance boost. If LoRA acts as a general boost, then it should help the full model and the pruned model by roughly the same amount. However, if LoRA specifically compensates for the information lost through pruning, then it should help the pruned model more, because there is more to recover. To test this, we compute the WER reduction from LoRA in both settings: the difference between the full baseline and full baseline+LoRA, and the difference between the two-layer pruned and two-layer pruned+LoRA (Table 5). In seven of nine conditions, LoRA provides a larger WER reduction for the pruned model than for the full model. For example, in Whisper-Small Dutch, LoRA reduces the full model's WER by 2.58 but the pruned model's WER by 4.18, indicating

Table 2: Test set WER (%) across encoder layer pruning configurations for three Whisper scales. *Red.%*: reduction of parameters in percentage. *Base*: projector-only training; *+LoRA*: LoRA additionally applied to the LLM. Blue : full baseline WER. **Green** : 2 layer pruned+LoRA that outperforms the full baseline.

		Danish		Dutch		English	
Layers	Red.%	Base	+LoRA	Base	+LoRA	Base	+LoRA
Whisper-Small (88.15M base params)							
12 (full)	0%	47.41	41.94	17.98	15.40	21.84	17.87
10	16.1%	50.49	**47.23**	21.16	**16.98**	25.81	**20.64**
8	32.2%	71.22	51.20	27.25	21.04	33.15	24.40
6	48.2%	115.68	67.02	32.14	24.45	40.81	31.85
4	64.3%	82.56	72.60	45.62	31.69	74.32	41.73
2	80.4%	123.21	115.48	62.28	48.97	110.85	95.30
Whisper-Medium (307.24M base params)							
24 (full)	0%	39.00	38.88	12.74	10.92	16.33	14.13
22	8.2%	39.73	**37.91**	14.22	**11.84**	17.89	**15.82**
20	16.4%	43.18	41.71	15.73	13.44	18.29	16.54
18	24.6%	50.59	44.74	18.25	14.66	20.44	18.47
16	32.8%	49.72	49.20	20.58	15.76	23.15	20.22
14	41.0%	55.55	51.95	22.64	16.91	25.85	21.79
12	49.2%	59.15	59.21	27.61	20.04	29.44	25.38
10	57.4%	63.21	62.02	27.85	24.50	33.79	27.70
8	65.6%	69.62	66.03	38.16	27.89	43.79	34.03
6	73.8%	77.08	79.41	44.00	34.27	97.35	72.30
4	82.0%	105.16	102.63	53.40	43.46	111.71	77.48
2	90.2%	148.37	90.25	66.83	56.23	110.12	78.01
Whisper-Large-v2 (636.83M base params)							
32 (full)	0%	36.09	34.07	12.05	10.32	13.36	11.58
30	6.2%	38.36	**35.81**	13.61	**11.44**	14.87	**12.16**
28	12.4%	40.08	39.12	15.75	12.67	16.26	15.62
26	18.5%	48.28	44.75	17.15	14.03	18.32	17.08
24	24.7%	52.02	47.63	18.96	15.23	20.75	20.13
22	30.9%	52.44	56.24	21.31	16.71	22.31	21.45
20	37.1%	130.99	59.06	26.00	18.73	36.34	32.92
18	43.3%	73.87	61.64	28.92	22.06	48.38	39.86
16	49.4%	67.48	64.40	38.82	24.87	54.12	46.28
14	55.6%	77.97	73.77	45.21	29.81	62.45	59.82
12	61.8%	115.51	84.08	59.81	39.03	74.42	65.36
10	68.0%	104.79	87.67	59.60	43.42	84.06	74.17
8	74.2%	160.15	90.33	63.61	48.18	88.33	77.34
6	80.3%	120.43	99.58	90.60	51.14	98.28	87.84
4	86.5%	118.30	122.81	72.41	62.04	104.34	92.24
2	92.7%	123.52	132.03	83.75	74.39	108.47	96.20

that the degraded encoder representations leave more room for LoRA to recover. The two exceptions (Small Danish and Medium English) involve either low-resource instability or a negligible difference (2.20 vs 2.07). These results confirm that LoRA does not merely improve performance uniformly; it provides greater recovery where pruning has caused more damage. These findings motivate a deeper investigation into the specific error types that LoRA corrects, which we examine next.

5.3. Error Analysis

The aggregate WER results from Sections 5.1 and 5.2 summarise performance as a single number per test set. However, this can obscure important variation: a low average WER could result from uniform small improvements across all utterances, or from large improvements on some utterances offset by regressions on others. To investigate this, we analyse the results at three levels of granularity: we first measure how many individual utterances are affected by pruning and LoRA (Section 5.3.1), then

Table 3: Sweet-spot configurations (bold) per encoder scale, showing WER (%) on the test set. Net Δ is relative to the full baseline. Reported params use the r=16 LoRA config (Dutch/English); Danish uses r=8 (∼0.8M vs. ∼1.5M overhead).

Encoder	Configuration	Params	Net Δ	DA	NL	EN
Small	12L Baseline	88.15M	–	47.41	17.98	21.84
	12L + LoRA	89.65M	+1.5M	41.94	15.40	17.87
	10L + LoRA	**75.5M**	**–12.7M**	**47.23**	**16.98**	**20.64**
Medium	24L Baseline	307.24M	–	39.00	12.74	16.33
	24L + LoRA	309.0M	+1.8M	38.88	10.92	14.13
	22L + LoRA	**284.0M**	**–23.2M**	**37.91**	**11.84**	**15.82**
Large-v2	32L Baseline	636.83M	–	36.09	12.05	13.36
	32L + LoRA	638.3M	+1.5M	34.07	10.32	11.58
	30L + LoRA	**599.0M**	**–37.8M**	**35.81**	**11.44**	**12.16**

Table 4: Utterance-level degradation after removing two layers, with and without LoRA. Δ Recov.: reduction in percentage of degraded utterances after applying LoRA.

		% Utterances Degraded		
Encoder	Lang	Base−2L	+LoRA	Δ Recov.
Small	DA	42.7	38.1	−4.6
	NL	33.3	23.5	−9.8
	EN	35.9	23.9	−12.0
Medium	DA	41.9	35.7	−6.2
	NL	24.6	24.3	−0.3
	EN	25.3	22.7	−2.6
Large-v2	DA	33.6	25.8	−7.8
	NL	24.1	18.2	−5.9
	EN	22.8	15.4	−7.4

Table 5: WER reduction from applying LoRA to full baseline and two-layer pruned models, measured in percentage points. Bold indicates the larger reduction. In seven of nine conditions, LoRA provides a larger reduction for the pruned model.

		WER Reduction	
Encoder	Lang	Full+LoRA	Pruned+LoRA
Small	DA	**5.47**	3.26
	NL	2.58	**4.18**
	EN	3.97	**5.17**
Medium	DA	0.12	**1.82**
	NL	1.82	**2.38**
	EN	**2.20**	2.07
Large-v2	DA	2.02	**2.55**
	NL	1.73	**2.17**
	EN	1.78	**2.71**

examine what types of errors LoRA corrects (Section 5.3.2), and finally quantify the word-level error distribution to identify the underlying compensation mechanism (Section 5.3.3).

5.3.1. Utterance-Level Impact of Pruning and LoRA

To assess whether pruning and LoRA affect utterances uniformly, we perform a per-utterance comparison across three settings: the full unpruned baseline, the two-layer pruned model without LoRA, and the two-layer pruned model with LoRA. We use the two-layer pruned depth because it represents the best efficiency–performance trade-off identified in Section 5.2 (10L for Small, 22L for Medium, 30L for Large-v2). For each utterance in the test set, we compute WER under all three settings and compare the two pruned variants against the unpruned baseline: if an utterance's WER increased after pruning, it is counted as degraded; otherwise, it is counted as preserved or improved. Table 4 reports the percentage of utterances degraded by pruning alone (Base−2L), the percentage still degraded after applying LoRA (+LoRA), and the recovery difference (Δ Recov.) across all encoder variants and languages.

Pruning alone (Base-2L) degrades a substantial share of utterances, and this share is consistently higher for Danish (33.6-42.7%) than for Dutch (24.1-33.3%) or English (22.8-35.9%), consistent with the resource-level patterns from Section 5.1. Second, adding LoRA reduces the degradation rate in all nine conditions, meaning some previously degraded utterances are recovered. However, even after LoRA, 15-38% of utterances remain worse than the unpruned baseline, while the majority (62-85%) match or improve upon it. This reveals that the aggregate WER gains reported in Section 5.2 are not the result of uniform improvements; they

are driven by strong recoveries on a subset of utterances that outweigh the remaining regressions.

5.3.2. Qualitative Error Patterns

The previous analysis shows how many utterances are affected; we now examine what kinds of errors LoRA actually corrects. With thousands of utterances across nine encoders and language configurations, manual inspection of every case is infeasible. We therefore apply two filters to isolate the most informative cases. First, we select utterances where the pruned model produces severe errors (WER > 0.8), ensuring that encoder degradation is substantial enough to produce identifiable error patterns rather than minor single-word differences. Second, we require that LoRA achieves full recovery (WER $= 0.0$) of the same utterances, which confirms that the correction is attributable to LoRA's adaptation of the LLM rather than to residual information preserved in the pruned encoder. Examples are drawn from all three encoder variants and all three languages. Table 6 presents representative cases grouped by error type.

Five common patterns we have identified: repetitive hallucination, where the pruned model generates looping output; idiom and fixed expression errors, where familiar phrases are distorted beyond recognition; named entity fragmentation; world knowledge substitution, where domain-specific terms are replaced by plausible but incorrect alternatives; and phonetic confusion. Particularly, all five categories involve errors recoverable through linguistic rather than acoustic information.

5.3.3. Mechanism Interpretation

The qualitative patterns above suggest that LoRA compensates through the LLM's linguistic knowledge. To test this quantitatively, we analyse word-level errors that LoRA corrects. Standard ASR error analysis decomposes word errors into three categories: substitutions, insertions, and deletions. For each encoder variant and language, we count these error types for both the pruned model and the pruned+LoRA model under the two-layer pruned configuration. Table 7 reports the percentage change in each error type after applying LoRA. The Tot column reports the percentage change in total word errors (substitutions + insertions + deletions combined); because each error type contributes differently to the total count, Tot reflects the weighted combination rather than a simple average of the three individual percentages.

Since LoRA adapts only the LLM's attention matrices (Q, K, V, O) while the pruned encoder remains frozen, all compensation operates on the decoding side. We would therefore expect error reductions for languages where the LLM has strong linguistic grounding, as the LLM can leverage its vocabulary, grammar, and contextual knowledge to correct degraded encoder representations. The results confirm this (Table 7): across all three encoder scales, LoRA reduces total word errors for English and Dutch in all six conditions, with substitutions and deletions showing consistent reductions. Danish follows a different pattern: while substitutions and deletions decrease, insertion errors increase across all three encoder scales, indicating that when the LLM lacks sufficient linguistic grounding for a language, LoRA can introduce spurious tokens rather than recover missing information. This language-dependent pattern reinforces that LoRA's effectiveness is tied to the LLM's pre-existing proficiency in each language rather than to the acoustic properties of the signal.

6. Conclusion

This paper presented a systematic study of encoder layer pruning and LoRA compensation in SLAM-ASR, evaluated across three Whisper encoder variants (Small, Medium, Large-v2) and three languages representing distinct resource levels (Danish, Dutch, and English). We find that removing the top one to two encoder layers results in only marginal WER degradation (within 2–4%) across all encoder scales and languages, supporting the hypothesis that upper encoder layers are partially redundant when an LLM handles downstream linguistic processing. Beyond this shallow pruning regime, medium and high resource languages degrade smoothly, while low-resource Danish often exhibits non-monotonic instability. Combining two-layer pruning with LoRA consistently outperforms the unpruned baseline while reducing total parameters by 7–14%. Error analysis reveals that LoRA compensates primarily through substitution and deletion corrections, leveraging the LLM's linguistic knowledge rather than repairing acoustic information, though this benefit is less consistent for low-resource Danish. Future work should explore alternative pruning strategies beyond top-down removal, encoder-side LoRA adaptation, validation across different system architectures, and controlled experiments to disentangle the effects of data availability from pre-trained representation quality.

7. Ethical Considerations and Limitations

In this work, we have used publicly available models, architectures, and datasets and have not collected any sensitive/private data. The ultimate goal of our study is to contribute to analyzing the effect of speech encoder pruning on LLM-based ASR.

Table 6: Representative error patterns where the pruned model produces severe errors (WER > 0.8) but LoRA achieves full recovery (WER $= 0.0$). Examples sampled across all three encoder variants.

Pattern	Lang	Reference	Pruned Hypothesis	+LoRA Hypothesis
Hallucination	DA	der kogte gryden over	da kom gud og gik op	der kogte gryden over
Named entity	DA	charlot var altid inde hos fru simonin nu	chrisel indelte altså inden for sin sinde	charlot var altid inde hos fru simonin nu
Semantic drift	DA	lille soldat du skal være vor konge…	livet skal du have og du skal have den dig selv…	lille soldat du skal være vor konge…
Repetition	NL	…heel interessant en hebben ons zeer geholpen…	zeer grotendeels [×28]	…heel interessant en hebben ons zeer geholpen…
Named entity	NL	de veiligheidssituatie is sindsdien rampzalig verslechterd	de veiligheid ziet de waarde van een dienst…	de veiligheidssituatie is sindsdien rampzalig verslechterd
Compound word	NL	de luchtvaart-maatschappijen	de luchtvaart maatschappijen	de luchtvaart-maatschappijen
Idiom	EN	out of sight out of mind	at a site at a light	out of sight out of mind
Phonetic	EN	heaven forbid	have fun for bit	heaven forbid
World knowledge	EN	it is isoelectronic to benzene	it is said to be a traditional venetian dish	it is isoelectronic to benzene

Table 7: Word-level error change (%) after applying LoRA to the two-layer pruned configurations. *Sub*: substitutions; *Ins*: insertions; *Del*: deletions; *Tot*: total word errors. **Bold**: error increase.

Encoder	Lang	Sub	Ins	Del	Tot
Small	DA	–7.7	**+11.9**	–10.7	–5.7
	NL	–17.5	–37.4	–15.3	–20.4
	EN	–19.8	–9.8	–37.4	–20.6
Medium	DA	–4.9	**+18.4**	–22.9	–4.4
	NL	–13.7	–34.5	–11.4	–16.7
	EN	–10.6	**+0.5**	–30.3	–11.6
Large-v2	DA	–8.6	**+10.8**	–11.7	–6.6
	NL	–15.3	–14.9	–13.5	–15.0
	EN	–17.4	–7.1	–35.6	–18.2

Due to the use of all public details, we don't see any immediate ethical issue.

Our work investigates the pruning of the speech encoder in the SLAM-ASR using the Whisper and Qwen models. We have carefully limited our analysis to three languages of different resource levels. While this choice allows us to conduct careful analysis, we acknowledge that expanding the range of models and datasets could provide additional insights. Our evaluation uses a single system architecture (Whisper encoder, ConcatLinear projector, and Qwen2.5-3B); different LLM backbones, projector designs, or encoder families, however, we believe the finding will hold. Our pruning strategy follows a top-down approach motivated by prior evidence of upper-layer redundancy; however, redundancy patterns may vary across layers, and exploring other removal strategies could reveal additional compression opportunities.

8. Acknowledgments

This work was supported by a Knowledge Transfer Partnership (KTP) project (project number 10131983) funded by UKRI through Innovate UK, in collaboration with Hivedome. RS was supported by the ELOQUENCE project (grant number 101070558) funded by the UKRI and the European Union. Views and opinions expressed are, however, those of the author(s) only and do not necessarily reflect those of the UKRI, European Union, or European Commission-EU. Neither the European Union nor the granting authority can be held responsible for them.

9. Bibliographical References

Rosana Ardila, Megan Branson, Kelly Davis, Michael Kohler, Josh Meyer, Michael Henretty, Reuben Morais, Lindsay Saunders, Francis Tyers, and Gregor Weber. 2020. Common voice: A massively-multilingual speech corpus. In *Proceedings of the twelfth language resources and evaluation conference*, pages 4218–4222.

Sergio Burdisso, Esaú Villatoro-Tello, Andrés Carofilis, Shashi Kumar, Kadri Hacioglu, Srikanth R. Madikeri, Pradeep Rangappa, Manjunath K. E,

Petr Motlícek, Shankar Venkatesan, and Andreas Stolcke. 2026. Text-only adaptation in LLM-based ASR through text denoising. *arXiv preprint arXiv:2601.20900*.

Lorenzo Concina, Jordi Luque, Alessio Brutti, Marco Matassoni, and Yuchen Zhang. 2025. The Eloquence team submission for task 1 of MLC-SLM challenge. In *Workshop on Multilingual Conversational Speech Language Model (MLC-SLM)*, pages 50–53.

Yangui Fang, Jing Peng, Xu Li, Yu Xi, Chengwei Zhang, Guohui Zhong, and Kai Yu. 2025. Low-resource domain adaptation for speech LLMs via text-only fine-tuning. *arXiv preprint arXiv:2506.05671*.

Seraphina Fong, Marco Matassoni, and Alessio Brutti. 2025. Speech LLMs in Low-Resource Scenarios: Data Volume Requirements and the Impact of Pretraining on High-Resource Languages. *arXiv preprint arXiv:2508.05149*.

Sanchit Gandhi, Patrick Von Platen, and Alexander M Rush. 2023. Distil-whisper: Robust knowledge distillation via large-scale pseudo labelling. *arXiv preprint arXiv:2311.00430*.

Tianteng Gu, Bei Liu, Hang Shao, and Yanmin Qian. 2024. Sparsewav: Fast and accurate one-shot unstructured pruning for large speech foundation models. *Interspeech 2024*, pages 4498–4502.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the Knowledge in a Neural Network. *arXiv preprint arXiv:1503.02531*.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin de Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-Efficient Transfer Learning for NLP. In *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-Rank Adaptation of Large Language Models. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*.

Julian Irigoyen, Arthur Söhler, and Andreas Søeborg Kirkedal. 2025. Pruning as Regularization: Sensitivity-Aware One-Shot Pruning in ASR. *arXiv preprint arXiv:2511.08092*.

Hyeon Soo Kim, Chung Hyeon Cho, Hyejin Won, and Kyung Ho Park. 2023. Adapt and prune strategy for multilingual speech foundational model on low-resourced languages. In *Proceedings of the 3rd Workshop on Multi-lingual Representation Learning (MRL)*, pages 85–94.

Shashi Kumar, Iuliia Thorbecke, Sergio Burdisso, Esaú Villatoro-Tello, Manjunath KE, Kadri Hacioğlu, Pradeep Rangappa, Petr Motlicek, Aravind Ganapathiraju, and Andreas Stolcke. 2025. Performance evaluation of slam-asr: The good, the bad, the ugly, and the way forward. In *2025 IEEE International Conference on Acoustics, Speech, and Signal Processing Workshops (ICASSPW)*, pages 1–5. IEEE.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled Weight Decay Regularization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*.

Ziyang Ma, Guanrou Yang, Yifan Yang, Zhifu Gao, Jiaming Wang, Zhihao Du, Fan Yu, Qian Chen, Siqi Zheng, Shiliang Zhang, and Xie Chen. 2024. An Embarrassingly Simple Approach for LLM with Strong ASR Capacity. *arXiv preprint arXiv:2402.08846*.

Bingshen Mu, Kun Wei, Qijie Shao, Yong Xu, and Lei Xie. 2025. Hdmole: Mixture of lora experts with hierarchical routing and dynamic thresholds for fine-tuning llm-based asr models. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Tohru Nagano, Gakuto Kurata, Samuel Thomas, Hong-Kwang J Kuo, Daniel Bolanos, Hyun Jung, and George Saon. 2025. LLM based text generation for improved low-resource speech recognition models. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Ankita Pasad, Ju-Chieh Chou, and Karen Livescu. 2021. Layer-wise analysis of a self-supervised speech representation model. In *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 914–921. IEEE.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. 2025. Qwen2.5 Technical Report. *arXiv preprint arXiv:2412.15115*.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*, pages 28492–28518. PMLR.

Zheshu Song, Jianheng Zhuo, Yifan Yang, Ziyang Ma, Shixiong Zhang, and Xie Chen. 2024. LoRA-Whisper: Parameter-Efficient and Extensible Multilingual ASR. In *25th Annual Conference of the International Speech Communication Association, Interspeech 2024, Kos, Greece, September 1-5, 2024*. ISCA.

Yaya Sy, Christophe Cerisara, and Irina Illina. 2025. BaldWhisper: Faster Whisper with Head Shearing and Layer Merging. *arXiv preprint arXiv:2510.08599*.

Changli Tang, Wenyi Yu, Guangzhi Sun, Xianzhao Chen, Tian Tan, Wei Li, Lu Lu, Zejun Ma, and Chao Zhang. 2023. Salmonn: Towards generic hearing abilities for large language models. *arXiv preprint arXiv:2310.13289*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Jian Wu, Yashesh Gaur, Zhuo Chen, Long Zhou, Yimeng Zhu, Tianrui Wang, Jinyu Li, Shujie Liu, Bo Ren, Linquan Liu, and Yu Wu. 2023. On Decoder-Only Architecture For Speech-to-Text and Large Language Model Integration. In *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 1–8.

Yuchen Zhang, Haralambos Mouratidis, and Ravi Shekhar. 2026a. Speak in Context: Multilingual ASR with Speech–Context Alignment via Contrastive Learning. In *Proceedings of the Fifteenth biennial Language Resources and Evaluation Conference (LREC 2026)*.

Yuchen Zhang, Ravi Shekhar, and Haralambos Mouratidis. 2026b. Language family matters: Evaluating SpeechLLMs across linguistic boundaries. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics*, pages 487–499, Rabat, Morocco. Association for Computational Linguistics.