BRIDGING THE DIVIDE IN LOW-RESOURCE SPEECH TRANSLATION THROUGH SEMANTIC CORPUS ALIGNMENT AND ACCESSIBLE MODEL FINE-TUNING

LIANG XIAO

DOCTOR OF PHILOSOPHY (COMPUTER SCIENCE)

FACULTY OF INFORMATION AND COMMUNICATION TECHNOLOGY

UNIVERSITI TUNKU ABDUL RAHMAN

MARCH 13, 2026

BRIDGING THE DIVIDE IN LOW-RESOURCE SPEECH TRANSLATION THROUGH SEMANTIC CORPUS ALIGNMENT AND ACCESSIBLE MODEL FINE-TUNING

By

LIANG XIAO

A dissertation submitted to the
Department of Computer Science,
Faculty of Information and Communication Technology,
Universiti Tunku Abdul Rahman,
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy (Computer Science)
March 13, 2026

ABSTRACT

BRIDGING THE DIVIDE IN LOW-RESOURCE SPEECH TRANSLATION THROUGH SEMANTIC CORPUS ALIGNMENT AND ACCESSIBLE MODEL FINE-TUNING

Liang Xiao

The rapid expansion of economic and cultural exchanges between China and Malaysia has created unprecedented demand for efficient cross-lingual communication technologies. However, real-time Chinese–Malay speech translation remains severely constrained by three fundamental challenges: extreme data scarcity, profound typological divergence between the languages, and computational accessibility barriers to state-of-the-art models. This dissertation presents a comprehensive framework that addresses these challenges through three synergistic innovations, achieving state-of-the-art quality while remaining trainable on consumer-grade hardware. First, we introduce GPT-Aligner, a novel semantic alignment technique that recasts corpus creation as a reasoning task for large language models. This approach dramatically improves alignment accuracy over traditional methods for this typologically distant pair, enabling the cost-effective construction of a high-quality parallel corpus where none previously existed. Second, we develop Progressive Cross-Layer Low-Rank Adaptation (P-CLRA), a principled parameter-efficient fine-tuning method that uses a quantitative benefit-to-cost ratio to identify optimal components for adaptation. This achieves significant quality improvements that exceed full fine-tuning while adding a negligible number of trainable parameters. Third, we propose Progressive Cross-Layer Freezing (P-CLF), an adaptive strategy for speech translation that uses a novel

gradient-based sensitivity analysis to identify and freeze non-critical model layers. This technique not only makes fine-tuning feasible on consumer-grade hardware but also surpasses the performance of a full fine-tuning approach. The integrated pipeline achieves a +0.98 BLEU improvement over full fine-tuning (8.32→9.30), delivering high-quality, real-time speech-to-speech translation on consumer hardware. This research establishes a new framework for resource-constrained multilingual speech translation, providing both immediate practical solutions for Chinese–Malay communication and generalizable methodologies for addressing language technology inequality in other underserved linguistic communities.

Keywords: speech translation; semantic alignment; parameter-efficient fine-tuning; low-resource languages; Chinese–Malay translation; large language models

Subject Area: P98-98.5 Computational linguistics. Natural language processing

ACKNOWLEDGMENTS

First and foremost, I would like to express my deepest gratitude to my supervisor, **Dr. Jasmina Khaw Yen Min**, for her unwavering guidance, encouragement, and patience throughout my doctoral study. Her high standards, clear feedback, and generous mentorship shaped this work from its earliest ideas to its final form.

I extend special appreciation to my co-supervisors for their invaluable guidance and support:

- **Prof. Soung-Yue Liew**
- **Associate Professor Dr. Tan Tien Ping**
- **Prof. Donghong Qin**

Their constructive feedback, timely advice, and encouragement were instrumental in refining the research directions, solidifying the experimental design, and improving the clarity of presentation.

I am also grateful to the **thesis examination committee** for their constructive comments and insightful suggestions that strengthened the clarity, rigor, and scope of this dissertation. Their feedback helped refine the framing of the research questions, the presentation of empirical results, and the articulation of this work's broader implications.

My sincere thanks go to my **colleagues and friends** at the **Faculty of Information and Communication Technology**, especially within the **Department of Computer Science**, for many conversations that sparked ideas and for practical help at critical moments. I would like to acknowledge my lab mates for reading drafts, sharing tools and scripts, and for countless technical discussions that improved the methodology and experiments.

This project benefited from the generosity of the **open-source community** and the availability of public datasets. I gratefully acknowledge the authors and maintainers of tools, models, and libraries used in this research, and the community of researchers who openly share their work and findings. I also thank the **bilingual reviewers** who assisted with corpus checking and quality control; their careful verification contributed to the reliability of the aligned data used in this thesis.

To my **family**, thank you for your love, patience, and unwavering support. Your belief in me has been the constant foundation that made this journey possible.

Any remaining errors are my own.

DECLARATION

I, **LIANG XIAO** hereby declare that the dissertation is based on my original work except for quotations and citations which have been duly acknowledged. I also declare that it has not been previously or concurrently submitted for any other degree at UTAR or other institutions.

_____

(Liang Xiao)

Date: March 13, 2026

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

API Application Programming Interface

ASEAN Association of Southeast Asian Nations

ASR Automatic Speech Recognition

BLEU Bilingual Evaluation Understudy

COMET Crosslingual Optimised Metric for Evaluation of Translation

CUDA Compute Unified Device Architecture

DTW Dynamic Time Warping

GPT Generative Pre-trained Transformer

GPU Graphics Processing Unit

IWSLT International Workshop on Spoken Language Translation

JSON JavaScript Object Notation

KL Kullback–Leibler

LLM Large Language Model

LoRA Low-Rank Adaptation

mBERT Multilingual Bidirectional Encoder Representations from Transformers

MT Machine Translation

NAR Non-Autoregressive

NLLB No Language Left Behind

NMT Neural Machine Translation

OPUS Open Parallel Corpus

P-CLF Progressive Cross-Layer Freezing

P-CLRA Progressive Cross-Layer Low-Rank Adaptation

PEFT Parameter-Efficient Fine-Tuning

RTF Real-Time Factor

S2ST Speech-to-Speech Translation

S2TT Speech-to-Text Translation

ST Speech Translation

T2TT Text-to-Text Translation

T2U Text-to-Unit

TTS Text-to-Speech

VAD Voice Activity Detection

VRAM Video Random-Access Memory

GLOSSARY

ρ-ranking	Benefit-to-cost ratio used in P-CLRA to prioritize layer–module pairs for adaptation
σ-sensitivity	Gradient magnitude per parameter used in P-CLF to decide which layers to freeze
BLEU	N-gram precision-based MT metric with a brevity penalty
chrF	Character n-gram F-score metric for MT, robust to morphological variation
COMET	Neural metric for translation evaluation that compares system outputs to references using multilingual representations
GPT-Aligner	LLM-based semantic alignment module that produces structured sentence-level alignments for corpus construction
LoRA	Low-Rank Adaptation; represents weight updates with low-rank factors to reduce trainable parameters
many-to-many alignment	Alignment where multiple source units map to multiple target units due to splitting, merging, or reordering
P-CLF	Progressive Cross-Layer Freezing; freezes low-sensitivity layers and adapts critical components for multimodal speech translation
P-CLRA	Progressive Cross-Layer Low-Rank Adaptation; injects LoRA adapters only into layer–module pairs with the highest benefit-to-cost ratios
PEFT	Parameter-Efficient Fine-Tuning; adapts large models by updating a small subset of parameters
S2ST	Speech-to-speech translation that maps source speech to target speech, typically with a TTS component
S2TT	Speech-to-text translation that maps source speech directly to target text
semantic alignment	Alignment based on meaning equivalence rather than surface cues, enabling robust many-to-many matches for distant language pairs

CHAPTER 1

INTRODUCTION

1.1 Research Background

The Association of Southeast Asian Nations (ASEAN) is rapidly becoming one of the world's most vibrant economic corridors, with speech translation for low-resource, typologically distant language pairs remaining far behind the progress seen in high-resource settings. While substantial progress has been made for high-resource language pairs such as English–French and English–Chinese, low-resource language pairs continue to face significant challenges that require innovative approaches to bridge the divide [1, 2].

Chinese and Malay languages represent a prototypical low-resource, distant language pair that exemplifies these challenges. Chinese, a Sino-Tibetan language, is characterized by its tonal nature, logographic writing system, and analytic grammar structure. In contrast, Malay, an Austronesian language, employs a Latin-based writing system, non-tonal phonetics, and mildly agglutinative morphology [3]. These fundamental differences—spanning phonetic, orthographic, and morphological dimensions—present unique challenges where word order, morphology, and discourse conventions differ so drastically that length-based or lexical alignment heuristics break down completely.

Despite the growing economic and cultural exchanges between China and Malaysia, real-time communication remains problematic due

to the lack of effective Chinese–Malay speech translation systems. This communication gap highlights the urgent need for advanced speech translation technologies that can operate effectively under resource constraints while maintaining competitive performance.

1.2 Problem Statement and Research Motivation

China–Malaysia ties in trade, tourism, education, and culture are expanding rapidly, creating a practical need for dependable, real-time speech translation between Chinese and Malay. Yet in the very contexts where the need is greatest, systems underperform because they lack data, struggle with structural differences across the two languages, and remain difficult to adapt on accessible hardware. This section formalizes the problem and articulates the motivation for our approach by making the high-level need concrete and then detailing the technical bottlenecks our work addresses.

1.2.1 Data Scarcity

Publicly available parallel resources for Chinese–Malay are extremely limited compared with high-resource pairs, and existing corpus construction pipelines rely on proxies (length ratios, lexical overlap, static embeddings) that degrade for distant pairs [2, 4–7]. Low coverage and noisy alignment compound training difficulty in low-resource settings where data quality matters as much as quantity [8]. Proposed work: we introduce GPT-Aligner to reframe alignment as LLM-based semantic reasoning, replacing brittle surface heuristics with meaning-preserving alignment to unlock clean, scalable supervision for Chinese–Malay.

1.2.2 Profound Typological Divergence

Chinese (tonal, logographic, analytic) and Malay (non-tonal, Latin script, mildly agglutinative) differ across phonology, orthography, and morphology [3]. These divergences undermine standard assumptions: tones challenge Automatic Speech Recognition (ASR) models trained on non-tonal data [9], and productive affixation elevates out-of-vocabulary and segmentation errors [10]. Direct sentence-level alignment often becomes many-to-many, while reordering complicates Machine Translation (MT). Proposed work: we pair semantic alignment (GPT-Aligner) with principled, parameter-efficient adaptation—Progressive Cross-Layer Low-Rank Adaptation (P-CLRA) for text MT and Progressive Cross-Layer Freezing (P-CLF) for speech translation—to model cross-lingual structure without relying on fragile surface cues such as length ratios, lexical overlap, or static embeddings.

1.2.3 Computational Accessibility Barriers

Full fine-tuning of modern multilingual or multimodal models remains memory- and time-intensive, limiting research and deployment in resource-constrained settings [11, 12]. This accessibility gap is most acute for underserved language pairs that also lack large curated datasets. Proposed work: we develop Progressive Cross-Layer Low-Rank Adaptation (P-CLRA) to concentrate a tiny fraction of trainable parameters in the most useful layer–module positions, and Progressive Cross-Layer Freezing (P-CLF) to freeze low-sensitivity layers while adapting cross-modal interfaces—together enabling high-quality adaptation on a single 24-GB GPU.

Quality control presents an additional layer of complexity in low-resource scenarios. As demonstrated by Algayres et al. [8], corpus quality can have a more significant impact than quantity in low-resource

regimes where models have limited capacity to filter noise.

To bridge this divide, this research develops a resource-constrained yet high-accuracy Chinese–Malay speech translation pipeline. The solution is driven by three synergistic contributions: (1) GPT-Aligner, a semantic sentence-alignment technique that leverages large language models for corpus construction, (2) Progressive Cross-Layer Low-Rank Adaptation (P-CLRA), a principled parameter-efficient method for text machine translation, and (3) Progressive Cross-Layer Freezing (P-CLF), a sensitivity-guided selective layer freezing approach for speech translation.

1.3 Research Questions

RQ1 (Corpus Construction): How can we construct high-quality Chinese–Malay parallel data from noisy or weakly aligned sources when surface proxies (length/lexical/embeddings) fail for distant language pairs?

RQ2 (Text MT Adaptation Efficiency): How can we adapt multilingual text translation models to Chinese–Malay with minimal trainable parameters while preserving or improving quality relative to full fine-tuning?

RQ3 (Speech Translation on Consumer Hardware): How can we adapt end-to-end speech translation models to achieve reliable quality and latency on a single 24-GB GPU without adding inference overhead?

1.4 Research Objectives

To align tightly with the problem statements and research questions, the thesis pursues three objectives:

1. **To design and validate a semantic alignment method (GPT-Aligner) that reliably constructs high-quality Chinese–Malay**

bitext from weakly aligned sources. We formulate alignment as an LLM reasoning task with prompts and constraints tailored to distant pairs, and evaluate both intrinsic alignment quality and downstream MT impact relative to traditional proxies [4–7].

2. **To develop a parameter-efficient text MT adaptation method (P-CLRA) that concentrates a tiny trainable fraction at high-utility layer–module positions while improving quality.** We introduce a benefit-to-cost metric to rank candidates, target low-rank updates to attention projections and key layers, and measure gains (COMET/chrF/BLEU) versus full fine-tuning at $\leq$1% additional parameters.

3. **To adapt end-to-end speech translation on consumer hardware via sensitivity-guided selective freezing (P-CLF) without adding inference overhead.** We quantify layer sensitivity, freeze low-sensitivity components, retain cross-modal interfaces, and evaluate translation quality, memory usage, training time, and latency on a single 24-GB GPU.

1.5 Research Contributions

This research makes several significant contributions that demonstrate how competitive speech translation for distant low-resource pairs can be achieved without large compute budgets, providing both theoretical advances and practical solutions:

Contribution 1: A High-Accuracy, Cost-Effective Semantic Aligner (GPT-Aligner). We present a novel prompt-driven alignment procedure that recasts alignment as a reasoning task for large language models. This approach is highly effective for typologically distant pairs, achieving 80.6% accuracy on Chinese–Malay alignment, a 27.8 percentage point improvement over traditional methods. Crucially, this

method is also highly efficient, enabling the creation of large-scale corpora at a cost of only USD 0.87 per 10,000 sentences.

Contribution 2: An Ultra-Efficient Text MT Adaptation Method (P-CLRA). We develop a principled parameter-efficient fine-tuning method that uses a quantitative benefit-to-cost ratio to selectively apply LoRA adapters. This strategy proves superior to both full fine-tuning and uniform LoRA, delivering a +3.57 COMET improvement with only 0.76% of the model's parameters being trainable.

Contribution 3: A High-Performance Adaptation Method for Speech Translation (P-CLF). We propose an adaptive, sensitivity-guided layer-freezing strategy for the complex task of multimodal speech translation. This method not only improves performance, boosting Chinese speech-to-Malay text BLEU scores by +0.98 over full fine-tuning but also drastically reduces computational load, updating only 55% of model parameters and cutting GPU memory requirements from 21.7GB to just 15.6GB.

These contributions collectively demonstrate a comprehensive approach to Chinese–Malay speech translation, addressing the fundamental challenges of data scarcity, computational efficiency, and translation quality for low-resource language pairs.

1.6 Thesis Structure

This thesis is structured to logically build the case for our proposed framework, moving from foundational challenges to our novel solutions and their integrated evaluation. We begin in Chapter 2 by reviewing the existing literature to establish the theoretical context and identify the critical gaps in current approaches to low-resource speech translation. Chapter 3 then introduces the overarching research methodology, outlining the integrated pipeline designed to address these gaps. The

subsequent chapters systematically present each of the three core contributions. Chapter 4 tackles the foundational data scarcity problem with GPT-Aligner, our semantic corpus alignment technique. Building on this new data resource, Chapter 5 details P-CLRA, our principled solution for efficient text machine translation. Chapter 6 then extends these principles into the multimodal domain with P-CLF, our adaptive strategy for the primary task of speech translation. Finally, Chapter 7 synthesizes the research findings, evaluates the performance of the complete framework, and discusses the broader implications of this work, concluding with promising directions for future research.

CHAPTER 2

LITERATURE REVIEW

2.1 Overview of Speech Translation Systems

Speech translation systems are commonly organised along two axes: pipeline structure (cascaded vs. end-to-end) and output modality (speech-to-text (S2TT) vs. speech-to-speech (S2ST)). Cascaded pipelines (ASR→MT→TTS) remain strong baselines because they can exploit large monolingual resources for each component, but they are vulnerable to error propagation and higher latency. End-to-end systems learn a direct mapping from speech to translated text (or speech), reducing intermediate bottlenecks and enabling joint optimisation, but they require more parallel speech-translation data and careful cross-modal alignment. For low-resource, linguistically distant pairs like Chinese–Malay, these trade-offs are particularly acute. Recent work in speech foundation models and SFM+LLM integration for speech translation underscores rapid progress in 2024 and motivates lightweight adaptation rather than full retraining [13, 14].

In this thesis, we focus on end-to-end S2TT as the most practical path for Chinese–Malay under data and compute constraints. We retain cascaded systems as baselines but target parameter-efficient end-to-end adaptation (P-CLF) to reduce error propagation and latency while keeping training feasible on consumer hardware. Fully direct S2ST is left to future work once richer parallel speech resources become available.

Table 2.1: Architecture comparison for speech translation systems

Paradigm	Pipeline	Strengths	Limitations / Bottlenecks
Cascaded	ASR→MT→TTS	Modular; leverages large ASR/MT data; easy component upgrades	Error propagation; longer latency; ASR output mismatches MT training distribution
End-to-end S2TT	Speech→Text	Joint optimisation; lower latency; can use acoustic cues directly	Data-hungry; cross-modal alignment challenges; requires strong pretraining and PEFT
End-to-end S2ST	Speech→Speech	Preserves prosody; avoids intermediate text at inference	Requires TTS; harder evaluation; scarce parallel speech-to-speech data

2.2 Text-to-Speech Synthesis (TTS) in S2ST Pipelines

Text-to-speech is the final stage in speech-to-speech translation (S2ST), converting translated target text into natural speech. In cascaded systems, TTS quality largely determines output naturalness, prosody, and latency, but it does not change translation adequacy itself. Modern neural TTS systems (e.g., Tacotron-style spectrogram generators with neural vocoders) can produce high-quality speech when trained with sufficient data, and are widely used as off-the-shelf components [15–17].

This thesis does not propose new TTS modelling. Our contributions focus on alignment, translation, and parameter-efficient adaptation (GPT-Aligner, P-CLRA, P-CLF). Accordingly, TTS is treated as a downstream component in S2ST pipelines, and we rely on standard TTS references rather than in-depth modelling details. Interested readers can consult comprehensive TTS surveys and foundational models for

architecture and training considerations [16, 18].

2.3 Machine Translation Technologies

This section narrows MT discussion to the elements directly used in this thesis: Transformer-based NMT, multilingual transfer for low-resource settings, the evaluation perspective used in Chapters 5–6, and the limits of LLM prompting for MT.

2.3.1 Neural MT and the Transformer Backbone

Neural MT is dominated by sequence-to-sequence architectures with attention, culminating in the Transformer, which provides strong global context modelling and scalable training [19, 20]. This matters here because the text MT backbone used in Chapter 5 follows this paradigm; our methods assume a Transformer-style encoder-decoder where adaptation capacity can be targeted to specific layers and modules.

2.3.2 Multilingual Transfer for Low-Resource MT

Modern MT increasingly relies on many-to-many multilingual pretraining, where a single model learns shared representations across languages and then transfers to low-resource directions with light adaptation [21, 22]. This transfer-centric view motivates our choice of multilingual backbones and the focus on parameter-efficient adaptation rather than full fine-tuning for Chinese–Malay.

2.3.3 Evaluation Perspective (Pointer)

We keep MT evaluation consistent across the thesis and summarize metrics in Section 2.7. Chapters 5–6 report BLEU, chrF, and COMET for translation quality, with ASR WER added only when cascaded systems

are involved. This section avoids repeating metric definitions to keep evaluation terminology uniform.

2.3.4 LLM Prompting for MT and Why We Use LLMs for Alignment Instead

Prompting can elicit translations from large language models, but reliability issues remain for low-resource, non-English directions: copying source text, entity errors, hallucinations, and sensitivity to prompt templates are common [23]. Given these risks, we do not use LLMs as MT systems in this thesis. Instead, we leverage LLMs for semantic alignment, where structured outputs and explicit constraints improve reliability and where human verification is feasible; the alignment methodology is detailed in Section 2.4.

2.4 Corpus Construction and Alignment Methods

High-quality parallel corpora are the central resource for speech translation and machine translation, yet for distant and low-resource language pairs the dominant bottleneck is sentence alignment rather than raw data availability. This subsection consolidates the alignment literature around three paradigms—traditional statistical alignment, embedding-based alignment, and LLM-based semantic alignment—and connects them to corpus construction choices relevant to Chinese–Malay. The goal is to explain why earlier methods are insufficient for this pair, why semantic alignment is a necessary step, and how these insights motivate GPT-Aligner.

2.4.1 Traditional Alignment: Length, Lexicon, Hybrid

Early alignment methods relied on sentence length ratios and simple lexical cues. Length-based approaches (e.g., Gale and Church) assume a

stable monotonic relation between sentence lengths across languages and use dynamic programming to produce alignments [4]. Hybrid methods extend this with lexicon matches or statistical word correspondences to refine alignments [5, 24]. These methods are efficient and can align large corpora quickly, but they are brittle for distant language pairs like Chinese–Malay.

Key limitations for Chinese–Malay include:

- **Length mismatch**

  Chinese sentence length in characters correlates poorly with Malay word counts, weakening length priors.

- **Sparse lexicons**

  Limited bilingual dictionaries reduce lexical evidence, and proper names or loanwords do not consistently align.

- **Many-to-many and reordering**

  Non-literal translations, sentence splitting/merging, and reordering are common, but traditional aligners favor 1–1 or 1–N monotonic patterns.

- **Low robustness to noise**

  Document noise (headers, boilerplate, or OCR artifacts) can easily derail length-based alignment, creating systematic errors.

As a result, purely statistical or dictionary-driven alignment is insufficient as a primary mechanism for building Chinese–Malay corpora. In practice, these methods are best used only as coarse pre-filters or as constraints within a larger alignment pipeline.

2.4.2 Embedding-Based Alignment and Its Limits for Distant Pairs

Multilingual sentence embedding models such as LASER and LaBSE map sentences from different languages into a shared semantic space,

enabling alignment by nearest-neighbor search [7, 25]. Systems like Vecalign combine embedding similarity with dynamic programming to handle document-level structure and many-to-many alignments [26]. These methods have enabled mining of massive parallel corpora from web data and have become the standard for large-scale alignment pipelines.

However, embedding-based alignment remains fragile for distant language pairs. Embedding quality varies by language and domain; low-resource languages often receive weaker representations, leading to false positives and missed matches. Semantically related but non-translated sentences can score highly, while literal translations in specialized domains may score poorly due to vocabulary mismatch or domain shift. For Chinese–Malay, typological distance and domain variability make these errors more frequent, and small alignment noise can significantly degrade downstream MT/ST performance.

Recent refinements (e.g., CRISS, LASER2/3) improve cross-lingual embeddings and domain robustness [27, 28], but they do not fully resolve the key problem: embedding similarity is a coarse proxy for translation equivalence, especially for distant pairs with frequent paraphrase and many-to-many mappings. Embedding pipelines therefore provide strong recall but insufficient precision for high-quality corpus construction.

2.4.3 LLM-Based Semantic Alignment

Large language models offer a qualitatively different alignment signal: instead of relying on surface statistics or fixed embeddings, they can reason about cross-lingual meaning and provide structured alignment decisions. This is especially valuable for many-to-many alignments, paraphrastic translations, and cases where lexical overlap is

Table 2.2: Alignment paradigms and suitability for Chinese–Malay corpus construction

Paradigm	Signal	Strengths	Key limitations for Chinese–Malay
Traditional	Length / lexicon	Fast, low cost, easy to scale	Weak under reordering, sparse lexicons, and many-to-many alignments
Embedding-based	Shared sentence vectors	Scales to web mining; handles some many-to-many	Unstable for distant pairs; false positives under domain shift
LLM semantic	Cross-lingual reasoning	Robust to paraphrase; interpretable; flexible output formats	Higher cost; needs prompt control and verification

minimal. Recent studies show that GPT-based aligners can outperform embedding pipelines on distant language pairs, particularly when a small number of in-context examples are provided [29].

The LLM alignment literature highlights four practical dimensions that matter for corpus construction:

1. **Explainability and structured outputs**

   LLMs can justify alignment decisions and output structured labels (e.g., 1–1, 1–N) in a format that is directly usable for downstream processing. This makes error analysis and filtering substantially easier than with embedding scores alone.

2. **Prompt design**

   Alignment quality depends strongly on concise, task-specific prompts and clear output schemas. Few-shot examples often improve robustness, while verbose role prompts can degrade consistency. This aligns with our prompt templates in GPT-Aligner

(Chapter 4), which favor short system instructions and explicit output formatting.

3. **Cost and scalability**

   LLM alignment is more expensive than embedding search, so practical pipelines often use a two-stage design: candidate generation via embeddings followed by LLM verification or refinement [30]. Batching, sliding-window alignment, and confidence filtering are essential for cost control.

4. **Robustness on distant pairs**

   LLMs can better handle reordering, omission, and paraphrase—common in Chinese–Malay—where embedding similarity can be unstable.

Beyond accuracy, LLMs offer controllability: prompts can specify alignment granularity (sentence vs. clause), boundary conditions (whether to allow 1–N), and output constraints (JSON fields or alignment indices). This controllability is essential for integration into corpus construction pipelines, where alignment outputs must be parsed, validated, and optionally inspected by human reviewers.

LLM alignment is most effective when paired with filtering and verification strategies. A typical workflow uses embeddings to propose candidate matches, then invokes an LLM to confirm or reject them. This hybrid approach keeps costs manageable while preserving the semantic reasoning advantage. Additional reliability can be gained through constrained decoding, explicit consistency checks (e.g., symmetric alignment), and periodic human auditing of uncertain cases.

2.4.4 Ethical and Governance Considerations

Using LLMs for corpus alignment raises ethical and governance issues that are distinct from purely statistical aligners. Proprietary models can encode societal or cultural biases from their training data, which may manifest as systematic alignment errors for underrepresented languages or dialects and propagate downstream into MT/ST models [31, 32]. External API dependence also affects reproducibility and long-term sustainability: model updates, pricing changes, or access restrictions can make alignment pipelines difficult to reproduce or maintain over time. These risks motivate documentation of data sources and alignment decisions, transparent reporting of model versions and prompts, and periodic audits of aligned corpora to detect bias and drift [33].

Prompt templates and output schemas. Alignment prompts are most reliable when they (i) clearly define the alignment unit, (ii) delimit input sentences with indices, and (iii) force a structured output (e.g., index pairs and alignment type). Few-shot examples help the model infer how to treat paraphrase, omission, and reordering, while overly long or role-play prompts tend to inject variability. In practice, short system prompts with explicit schemas (e.g., JSON fields for source indices, target indices, and alignment type) produce the most stable results for batch processing.

Cost-aware scaling. Because LLM calls are expensive, alignment pipelines often combine coarse candidate retrieval with semantic verification. Embedding-based retrieval limits the LLM to a small candidate window, while batching multiple alignment decisions into a single request amortizes overhead. Sliding-window alignment across long documents further reduces costs and prevents context-window

overflow. These design choices are essential for scaling GPT-Aligner to large corpora without sacrificing alignment precision.

2.4.5 Alignment Evaluation and Corpus Construction Implications

Alignment quality is typically measured with precision/recall/F1 against gold-standard alignments, with AER-style metrics used in some settings [34]. However, for corpus construction the most relevant evaluation is extrinsic: the effect of aligned data on downstream MT/ST quality. This makes precision particularly important for low-resource pairs, where small amounts of noise can have outsized negative effects.

A practical corpus construction pipeline therefore emphasizes: (i) candidate retrieval (coarse matching), (ii) semantic alignment (fine matching), and (iii) lightweight verification and filtering. The alignment stage must produce outputs that are not only accurate but also interpretable and consistent across documents. Within this thesis, GPT-Aligner provides the semantic alignment step that is robust to typological divergence while remaining deployable at scale.

2.4.6 Corpus Construction in Practice for Low-Resource ST

For low-resource speech translation, corpus construction typically starts from comparable or weakly parallel sources (e.g., bilingual web pages, subtitles, or paired transcripts), then relies on alignment to extract usable sentence pairs. The key engineering challenge is to balance recall (recover as much data as possible) with precision (avoid noisy pairs that degrade model training). In practice, this means using alignment confidence thresholds, deduplication, and light human verification for boundary cases, rather than exhaustive manual curation.

For Chinese–Malay, our approach prioritizes semantic alignment over surface cues. We employ GPT-Aligner to identify high-confidence

sentence pairs and to flag ambiguous many-to-many alignments that require either verification or exclusion. This strategy improves alignment quality while keeping cost manageable, and it produces a corpus that is suitable for downstream MT/ST training without heavy post-editing.

2.4.7 Research Gap and Contribution

Existing alignment pipelines either scale well but degrade on distant language pairs (length/embedding methods) or provide strong semantic accuracy but are costly and prompt-sensitive (LLM methods). For Chinese–Malay, this leaves a gap: we need a semantic aligner that is accurate on many-to-many cases yet cost-controlled and reproducible. GPT-Aligner fills this gap by combining structured prompts, few-shot guidance, and a verification-friendly output format. This alignment backbone is central to the corpus construction described in Chapter 4 and underpins the data used for P-CLRA (Chapter 5) and P-CLF (Chapter 6).

2.5 Multilingual and Low-Resource Machine Translation

Multilingual MT has converged on a many-to-many pretraining paradigm: a single sequence-to-sequence model is trained on large multilingual corpora with language tags, enabling cross-lingual transfer and zero-/few-shot generalisation. For low-resource directions, the key mechanism is transfer from high-resource pairs via shared representations, followed by light adaptation on limited parallel data. This thesis adopts that paradigm through multilingual backbones and parameter-efficient adaptation, rather than training pair-specific models from scratch.

2.5.1 Mainstream Many-to-Many Pretraining and Low-Resource Transfer

Many-to-many training (e.g., M2M-100, NLLB) uses a shared encoder-decoder and language-tagged data to cover hundreds of directions in a single model. This improves low-resource performance by reusing shared lexical and structural features learned from high-resource pairs. However, such models are large and expensive to fine-tune; therefore, low-resource adaptation increasingly relies on efficient methods (e.g., adapters, LoRA, layer freezing) rather than full fine-tuning. In our experiments, we focus on adaptation atop NLLB-style multilingual models, keeping the backbone fixed where possible and updating only a small fraction of parameters.

2.5.2 Backbones Used in This Thesis

We focus on two model families directly used in Chapters 5–6:

- **NLLB-200 / NLLB-1.3B**

  A multilingual many-to-many MT backbone trained on large-scale web-mined bitext. It provides strong cross-lingual transfer but is costly to fine-tune fully, motivating P-CLRA for parameter-efficient text MT adaptation.

- **SeamlessM4T v2 (speech-text backbone)**

  A unified multilingual model supporting S2TT and related tasks. It provides a strong speech translation base and benefits from large-scale pretraining and multilingual alignment, but still requires efficient adaptation for low-resource pairs, motivating P-CLF.

2.5.3 Chinese–Malay Challenges Relevant to Alignment and Adaptation

For Chinese–Malay, only a few challenges materially affect alignment and adaptation decisions:

- **Typological divergence**

  Different scripts and morphology reduce lexical overlap, weakening surface cues used by traditional aligners and increasing reliance on semantic alignment.

- **Data scarcity and domain bias**

  Parallel data is limited and domain-skewed, so adaptation must be robust to noise and domain shift, and alignment quality has outsized impact on training stability.

- **Compute constraints**

  Full fine-tuning of large multilingual backbones is expensive; parameter-efficient adaptation is required for practical training and iteration.

These constraints motivate the pipeline adopted in this thesis: GPT-Aligner for robust corpus alignment, P-CLRA for text MT adaptation on NLLB, and P-CLF for speech translation adaptation on SeamlessM4T.

2.6 Parameter-Efficient Fine-Tuning Methods

Parameter-efficient fine-tuning (PEFT) adapts large pretrained models by updating only a small subset of parameters, which is essential for low-resource MT and speech translation where full fine-tuning is costly. This section narrows to the PEFT methods directly relevant to our contributions, with LoRA as the primary baseline and P-CLRA/P-CLF as the proposed extensions.

2.6.1 Reference PEFT Baselines: Adapter, Prompt/Prefix, LoRA

Adapters insert small bottleneck modules between transformer layers and train only those modules [35]. They offer strong performance with modest parameter budgets but add inference latency and architectural complexity. **Prompt/Prefix tuning** learns a small set of continuous tokens or prefixes to steer model behaviour without changing core weights [36]. It is parameter-light but can be brittle for distant language pairs and often underperforms when strong domain adaptation is needed.

LoRA [37] is the most relevant baseline for this thesis. It parameterizes weight updates as a low-rank decomposition $\Delta W = BA$, with $r \ll \min(d,k)$, enabling efficient adaptation without modifying inference graphs. Key design choices that strongly affect performance include: (i) **injection location** (e.g., attention projections vs. FFN blocks), (ii) **rank** and **scaling** (capacity vs. stability), (iii) **trainable parameter ratio** (budget control), and (iv) **stability under low-resource data** (risk of overfitting or under-capacity). These choices motivate our targeted, data-driven allocation rather than uniform LoRA placement.

2.6.2 Why Existing PEFT Is Insufficient

Uniform PEFT strategies treat all layers and modules as equally useful, which is rarely true for cross-lingual adaptation. For low-resource pairs, this wastes limited capacity on low-utility components and can under-adapt critical ones. In multimodal settings, the issue is compounded because speech encoders and text decoders have different sensitivity profiles; adapting them uniformly can degrade either acoustic modelling or linguistic transfer.

2.6.3 Motivation for P-CLRA and P-CLF

These limitations motivate the two methods developed in this thesis. **P-CLRA** (Chapter 5) extends LoRA by computing a benefit-to-cost ratio per layer–module pair and injecting adapters only where marginal gains are highest, improving both quality and parameter efficiency. **P-CLF** (Chapter 6) adapts multimodal models by freezing low-sensitivity layers and selectively updating critical components, preserving cross-lingual knowledge while reducing training cost. Together, these methods replace uniform adaptation with principled, layer-aware allocation suited to low-resource Chinese–Malay translation.

2.7 Evaluation Metrics and Efficiency

We focus on the metrics and efficiency indicators actually used in the experiments. For text translation quality, we report **BLEU**, **chrF**, and **COMET**. BLEU captures n-gram overlap, chrF provides character-level sensitivity that is more robust to morphology, and COMET offers a neural, reference-based assessment with higher correlation to human judgments. Using all three provides a balanced view of lexical accuracy and semantic adequacy for low-resource, distant pairs.

When cascaded ASR$\rightarrow$MT baselines are involved, we additionally report **ASR Word Error Rate (WER)** to quantify upstream recognition errors that can propagate into MT quality. In such cases, MT metrics are computed on ASR hypotheses to separate recognition noise from translation model performance.

Efficiency reporting in this thesis is limited to measures used in Chapters 5–6: **trainable-parameter ratio**, **GPU memory**, and **training time** (with wall-clock hours on a single GPU). These indicators capture the practical cost of adaptation and are aligned with the deployment constraints motivating PEFT.

2.8 Summary of Literature Review

This chapter maps the core literature to a concrete research roadmap for low-resource Chinese–Malay speech translation. The review motivates a pipeline that starts from reliable corpus alignment, proceeds through parameter-efficient text MT adaptation, and culminates in efficient multimodal speech translation. Below we summarize three concrete gaps and how Chapters 3–6 address them.

2.8.1 Research Roadmap: Gaps Aligned to Chapters 3–6

The literature points to three decisive gaps that must be closed to make low-resource S2TT practical:

- **Gap 1: Robust semantic alignment for distant pairs**
  Traditional length/lexicon/embedding aligners are brittle for Chinese–Malay and fail on many-to-many cases. **Chapter 3** introduces the GPT-Aligner methodology, and **Chapter 4** operationalizes it for corpus construction and evaluation.

- **Gap 2: Parameter-efficient text MT adaptation**
  Uniform PEFT (e.g., vanilla LoRA) wastes limited capacity on low-utility layers. **Chapter 5** proposes P-CLRA, a benefit-to-cost guided LoRA strategy that targets only high-utility layer–module pairs.

- **Gap 3: Parameter-efficient multimodal adaptation**
  Speech translation models exhibit asymmetric sensitivity across speech encoders and text decoders, making uniform adaptation suboptimal. **Chapter 6** introduces P-CLF, a sensitivity-guided freezing strategy for efficient S2TT adaptation.

Together, Chapters 3–6 provide a coherent pipeline: semantic alignment to build usable corpora, targeted text MT adaptation on

multilingual backbones, and efficient multimodal speech translation on speech-text models. This roadmap directly addresses the key bottlenecks identified in the literature while keeping training and deployment feasible on consumer hardware.

CHAPTER 3

RESEARCH METHODOLOGY

3.1 Research Methodology Overview

This chapter presents the methodological framework employed in this research, detailing the systematic approach adopted to address the fundamental challenges of Chinese–Malay speech translation: data poverty, structural divergence, and computational constraints. The methodology is built around three synergistic contributions that together deliver strong quality while remaining trainable on a single consumer-grade GPU: GPT-Aligner for semantic corpus alignment, P-CLRA (Progressive Cross-Layer Low-Rank Adaptation) for efficient text translation, and P-CLF (Progressive Cross-Layer Freezing) for multimodal speech translation.

This research adopts a systematic approach to address the challenges of Chinese–Malay speech translation, integrating multiple methodological components within a unified framework that recognizes the interconnected nature of the fundamental challenges facing low-resource speech translation. The overall research design follows a progressive development cycle that begins with resource creation, proceeds through model development, and culminates in comprehensive evaluation, with each phase building upon the insights and outputs of previous phases while contributing to the overall coherence of the integrated solution.

The research process is structured into four interconnected phases that systematically address the three fundamental challenges of data poverty, computational constraints, and structural divergence, as visualized in Figure 3.1.

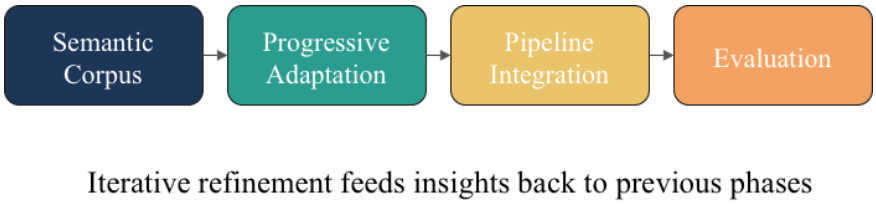

Figure 3.1: Four-phase research methodology linking corpus construction, progressive adaptation, integrated pipeline design, and evaluation.

The semantic corpus construction phase addresses data poverty through GPT-Aligner, a semantic sentence-alignment technique that upgrades raw parallel texts into clean, high-quality bitext without relying on bilingual dictionaries or traditional feature-based heuristics. This phase involves developing prompt-based reasoning approaches that fundamentally shift from statistical feature matching to semantic understanding, leveraging the reasoning capabilities of large language models to handle the complex many-to-many correspondences that characterize linguistically distant language pairs like Chinese–Malay.

The progressive parameter-efficient adaptation phase addresses computational constraints through two complementary approaches that share a common theoretical foundation while addressing different modalities. P-CLRA for text machine translation and P-CLF for speech translation both employ quantitative metrics to guide parameter allocation decisions, moving beyond heuristic approaches toward principled, data-driven optimisation strategies that respect the heterogeneous nature of neural network architectures. This phase recognizes that different layers and modules within neural networks contribute differently to downstream performance, and that optimal

adaptation strategies must account for these differences rather than applying uniform modifications across all components.

The integrated pipeline development phase addresses structural divergence by combining semantic alignment with theoretically motivated cross-layer adaptation, creating both cascaded and direct translation paths that can leverage the strengths of each approach while mitigating their individual limitations. The comprehensive evaluation phase validates the integrated framework through extensive evaluation across multiple dimensions, including alignment accuracy, translation quality, parameter efficiency, and computational performance, establishing the synergistic effects of combining semantic alignment with progressive adaptation strategies. This phased approach allows for iterative refinement at each stage, with feedback from later phases informing adjustments to earlier components, creating a research methodology that is both systematic and adaptive to emerging insights.

The methodological framework is designed around three interconnected research strategies that directly address the fundamental challenges of low-resource speech translation while maintaining theoretical rigor and practical feasibility. The semantic alignment methodology addresses data scarcity by recasting alignment as a prompt-based reasoning task for large language models, fundamentally departing from traditional approaches that rely on sentence-length ratios, lexical overlap, or static bilingual embeddings. GPT-Aligner leverages the cross-lingual reasoning capabilities of large language models through concise system prompts that instruct GPT-4 to output JSON-formatted alignment pairs, achieving superior performance on the complex many-to-many correspondences that characterize typologically distant language pairs like Chinese–Malay. This approach represents a paradigmatic shift from statistical feature matching to

semantic understanding, enabling effective alignment even when surface-level similarities are minimal.

The progressive cross-layer adaptation strategy addresses computational constraints through two complementary parameter-efficient methods that share a common theoretical foundation while targeting different modalities. P-CLRA employs a quantitative benefit-to-cost ratio $\rho = \frac{\Delta\text{COMET}}{\text{Params}}$ for each layer-module pair, systematically identifying optimal "sweet-spot" blocks for LoRA injection rather than applying uniform modifications across all components. P-CLF utilizes gradient-based sensitivity analysis $\sigma = \frac{\|\nabla_{\theta_l}\mathcal{L}\|_2}{\text{Params}_l}$ to identify low-gradient layers for strategic freezing, enabling superior performance with dramatically reduced computational requirements. Both methods move beyond heuristic approaches toward principled, data-driven optimisation strategies that respect the heterogeneous nature of neural network architectures, recognizing that different layers and modules contribute differently to downstream performance.

The integrated framework evaluation methodology validates both individual components and their synergistic combination through comprehensive assessment across multiple dimensions including translation quality, computational efficiency, real-time performance, and cross-linguistic generalisability. This evaluation strategy encompasses comparison against cascaded baselines, ablation studies confirming the importance of identified components, and analysis of the emergent properties that arise from combining semantic alignment with progressive adaptation. The research is guided by three key hypotheses that systematically test the effectiveness of semantic reasoning for alignment accuracy, the superiority of quantitative ρ/σ metrics for parameter efficiency, and the viability of the integrated framework for

delivering competitive performance on consumer hardware. These hypotheses are evaluated through controlled experiments, with quantitative results and discussions reported in Chapters 4–7.

3.2 System Architecture Design

The proposed Chinese–Malay speech translation system employs an integrated architecture that combines semantic corpus alignment with progressive cross-layer adaptation, supporting both cascaded and direct translation paths. This section outlines the overall system framework, component relationships, and integration strategies.

The system architecture implements two complementary translation paths that leverage the three core contributions while providing flexibility for different deployment scenarios and performance requirements. The cascaded translation path links specialized components in sequence, beginning with Whisper-large-v3 for Chinese speech transcription, followed by an NLLB model enhanced with Progressive Cross-Layer Low-Rank Adaptation for Chinese-to-Malay text translation, and concluding with Glow-TTS-Malay for speech synthesis. This path leverages the strengths of specialized components optimised for individual tasks, enabling maximum quality through component-wise optimisation while benefiting from the high-quality parallel corpus constructed using GPT-Aligner.

The direct translation path employs a P-CLF-enhanced SeamlessM4T model for end-to-end Chinese speech-to-Malay text translation, with optional text-to-speech synthesis for complete speech-to-speech functionality. This approach offers lower latency and simplified deployment by eliminating intermediate representations, while the Progressive Cross-Layer Freezing adaptation ensures efficient training and inference on consumer hardware. Both architectural paths are built

upon the semantic alignment foundation provided by GPT-Aligner, ensuring consistent training data quality across all components and enabling the system to handle the complex many-to-many correspondences that characterize Chinese–Malay translation.

Figure 3.2 consolidates the dual-path system design that threads GPT-Aligner, P-CLRA, and P-CLF into cascaded and direct configurations.

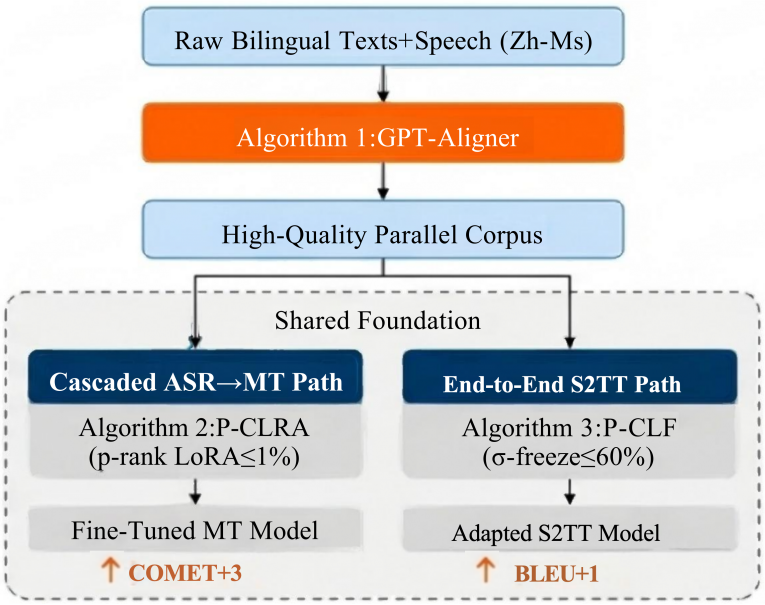

Figure 3.2: Integrated pipeline architecture showing both cascaded and direct translation paths built on GPT-Aligner corpus foundation

The integration of Progressive Cross-Layer Adaptation methods within this dual-path architecture follows a principled approach that maximizes parameter efficiency while maintaining translation quality through quantitative optimisation strategies. P-CLRA integration in the cascaded path targets mid-depth encoder and selective decoder modules based on systematic ρ-ranking analysis, applying LoRA specifically to attention projections where cross-lingual leverage is highest. P-CLF integration in the direct path employs sensitivity-guided freezing while maintaining trainable cross-attention mechanisms essential for speech-text alignment. Both adaptation methods share the unified principle of using quantitative ρ/σ metrics for principled

parameter allocation rather than heuristic approaches, enabling training on consumer hardware while maintaining competitive performance. Detailed trade-offs and empirical results are presented in Chapters 5 and 6.

The system integration strategy employs a hybrid approach that combines the advantages of both cascaded and end-to-end architectures while addressing their individual limitations through principled component coordination. This integration follows a three-stage optimisation process that begins with component-level optimisation where each module is individually optimised for its specific task, allowing for specialized techniques and architectures tailored to the unique challenges of speech recognition, translation, and speech synthesis. The second stage involves joint fine-tuning where components are collectively optimised to minimise error propagation and improve overall system performance, ensuring that each module adapts to the characteristics and potential errors of others in the pipeline. The final stage implements unified representation learning where speech and text encoders are designed to produce compatible representations, facilitating more effective knowledge transfer between modalities and improving the system's ability to handle speech-specific phenomena that may not be captured in text-only training data.

3.3 Data Resources and Processing Framework

The development of effective speech translation systems for low-resource language pairs like Chinese–Malay requires a comprehensive data processing framework that addresses the fundamental challenges of data scarcity, quality control, and linguistic diversity through systematic collection, preprocessing, and validation strategies. The data collection strategy focuses on assembling a diverse

and representative corpus of Chinese–Malay parallel data from multiple carefully selected sources, including official government documents and news articles for formal register coverage, educational materials and literary works for stylistic diversity, conversational data from subtitles and transcribed dialogues for informal register representation, and domain-specific content related to business, tourism, and technical fields for specialized terminology coverage. This multi-source approach employs stratified sampling to ensure balanced representation across different domains, registers, and linguistic phenomena, creating a corpus that captures the full range of translation challenges while maintaining the principle of quality over quantity through emphasis on clean, well-aligned datasets rather than maximising raw data volume.

The preprocessing pipeline implements rigorous quality control through systematic text normalisation, noise reduction, segmentation, and filtering processes that ensure consistency and reliability across the corpus. Text normalisation standardizes formats including punctuation, numerals, and special characters to ensure consistency across the corpus and facilitate more effective alignment and model training, while noise reduction removes irrelevant content such as HTML tags, metadata, and non-linguistic elements, with acoustic noise reduction techniques applied to speech data to improve signal quality. Sentence and utterance segmentation employs language-specific rules and statistical models to create properly aligned parallel units, followed by quality filtering that removes problematic data points including misaligned sentence pairs, machine-translated content, incomplete or truncated sentences, and sentences with excessive code-switching or non-standard language. The annotation and validation framework adds essential metadata to the corpus while ensuring quality through systematic linguistic annotation, alignment documentation, and multi-stage validation processes.

Linguistic annotation encompasses part-of-speech tags, named entity information, syntactic structure markers, and domain and register labels that enhance the corpus utility for model training and evaluation. Alignment annotation documents the complex correspondence relationships between Chinese and Malay sentences, including standard 1-to-1 alignments, 1-to-many and many-to-1 alignments for handling structural differences, and partial alignments where complete semantic equivalence is not achievable due to cultural or linguistic gaps. The validation process implements cross-validation by multiple annotators, consistency checks using automated tools, and iterative refinement based on model performance feedback, creating a high-quality resource that supports both current research objectives and future studies on Chinese–Malay language processing while establishing reliability standards that ensure the corpus can serve as a benchmark for subsequent research in low-resource language translation.

3.4 GPT-Aligner: Semantic Alignment Methodology

GPT-Aligner represents a paradigm shift from traditional feature-based alignment to semantic reasoning-based alignment, specifically designed to handle the challenges of typologically distant language pairs like Chinese–Malay. The full method, prompt templates, and quantitative results are discussed in detail in Chapter 4; this subsection provides only a brief methodological overview.

GPT-Aligner addresses the fundamental limitations of traditional alignment methods when applied to typologically distant language pairs like Chinese–Malay, where conventional approaches fail due to length correlation breakdown between logographic and alphabetic writing systems, lexical overlap scarcity between Sino-Tibetan and Austronesian language families, and complex correspondence patterns

involving many-to-many alignments and idiomatic paraphrases. The methodology leverages the cross-lingual semantic understanding capabilities of large language models to identify correspondences based on meaning rather than surface features, implementing a three-component design that optimizes both accuracy and scalability.

The prompt construction strategy employs concise system prompts that directly specify the alignment task with clear JSON output format specification, contrasting with prior LLM alignment work that uses lengthy role-play formats. The approach incorporates two diverse few-shot demonstrations that illustrate both 1-1 and many-to-many alignment patterns, combined with sliding-window processing that ensures coverage of long paragraphs while maintaining context coherence. Empirical comparisons show that direct, task-focused prompts consistently outperform verbose personas. Batch processing and cost-aware batching improve scalability while preserving alignment quality through confidence-based filtering and consistency checks.

The output processing and validation framework ensures reliability through systematic validation of JSON-formatted output from GPT-4, including format validation for proper JSON structure and index consistency, alignment verification using semantic similarity metrics, and manual spot-checking through human validation on representative samples. Quantitative evaluation and comparative tables are reported in Chapter 4.

Implementation considerations for GPT-Aligner encompass domain adaptation challenges where specialized domains may experience reduced alignment accuracy, highlighting the need for domain-specific prompt engineering, while qualitative inspection of Thai–Vietnamese and Sinhala–English samples suggested comparable behaviour, indicating potential applicability beyond the primary Chinese–Malay

focus. However, scalability constraints require careful cost management for very large corpora, as the superior quality of LLM-based alignment comes with computational costs that must be balanced against accuracy requirements.

3.4.1 Model Development and Optimisation Methods

The development of effective translation models for low-resource language pairs requires specialized approaches that maximise the utility of limited parallel data while leveraging the knowledge encoded in pretrained multilingual models through systematic model selection, adaptation, and optimisation strategies. Model selection is guided by multilingual capability prioritizing models with strong multilingual representations and exposure to Asian languages, including SeamlessM4T for unified multilingual and multimodal translation [9], NLLB-200 supporting translation between 200+ languages [12], and M2M-100 designed for direct translation between 100 languages. Architectural suitability focuses on Transformer-based models for their proven effectiveness in capturing long-range dependencies and cross-lingual patterns through self-attention mechanisms [38], while resource efficiency considerations prioritize models that balance performance with computational efficiency for deployment in resource-constrained environments. The adaptation strategy leverages transfer learning from high-resource to low-resource languages through specific techniques for optimising cross-lingual knowledge transfer, establishing the foundation for the progressive cross-layer adaptation methods that form the core of this research.

Progressive Cross-Layer Low-Rank Adaptation (P-CLRA) forms the text machine-translation leg of the framework. Rather than re-deriving the full method here, we summarize its role and refer the reader to

Chapter 5 for the probing experiments, architectural diagrams, and ablation studies. In essence, P-CLRA performs a two-stage process: a lightweight probing stage ranks layer-module pairs by benefit-to-cost ratio, and a targeted adaptation stage updates only the highest-utility projections. This yields a small trainable-parameter footprint with measurable quality gains, as reported in Chapter 5, while keeping GPU requirements within a single consumer-grade budget assumed throughout the thesis.

Progressive Cross-Layer Freezing (P-CLF) supplies the multimodal speech adaptation counterpart. Chapter 6 details the layer-sensitivity analysis, freezing schedule, and extensive evaluation against full fine-tuning and LoRA baselines. Within the methodological overview, it suffices to note that P-CLF keeps language-agnostic speech representations intact while selectively unfreezing cross-attention and decoder components that most influence Chinese–Malay speech-to-text quality. This design delivers consistent quality gains while reducing training cost, as detailed in Chapter 6.

Further implementation detail for P-CLF, including the exact freezing schedule, optimisation hyperparameters, and comparative analyses against other PEFT baselines, is consolidated in Chapter 6. At the methodological level it is sufficient to emphasize that the speech experiments reuse the Chinese–Malay corpus constructed in Section 6.2.1 and that P-CLF keeps training within a single consumer-GPU budget by freezing high-level encoder blocks while leaving cross-attention and projection pathways trainable.

The model training and optimisation process follows a systematic approach that maximizes performance through progressive training strategies beginning with general domain data and gradually introducing domain-specific content, curriculum learning that presents

examples in order of increasing complexity, and mixed-domain batching to prevent catastrophic forgetting of general language knowledge. Hyperparameter optimisation encompasses learning rate scheduling with warm-up and decay, gradient accumulation for effective batch size management, and regularisation techniques to prevent overfitting on limited data, while optimisation techniques include mixed precision training for computational efficiency, gradient checkpointing to reduce memory requirements, and distributed training across multiple GPUs when available. This comprehensive approach ensures effective adaptation to the Chinese–Malay language pair while maintaining computational efficiency throughout the training process.

3.5 Evaluation System Design

The evaluation framework employs a multi-dimensional assessment strategy that captures both automatic and human evaluation perspectives to provide comprehensive insights into translation quality and system performance. The selection of evaluation metrics addresses multiple dimensions of translation quality through automatic metrics including BLEU for its widespread use and comparability with existing research despite its limitations, COMET (a neural, reference-based MT evaluation metric that compares candidate translations against human references using contextual embeddings) for its stronger correlation with human judgments and sensitivity to semantic equivalence beyond surface-level similarity, chrF for its effectiveness in evaluating translations for morphologically rich languages through character-level analysis, and TER for measuring edit distance between candidate and reference translations to provide insights into post-editing effort requirements. Additionally, component-specific metrics include Word Error Rate (WER) and Character Error Rate (CER) for evaluating speech

recognition accuracy, and Mean Opinion Score (MOS) for assessing the naturalness and intelligibility of synthesized speech. This multi-metric approach ensures robust evaluation that captures both lexical accuracy and semantic adequacy while providing benchmarks comparable to existing research in the field, offering a comprehensive assessment of system performance across different aspects of translation quality.

The evaluation framework integrates both automatic and human assessment approaches to provide comprehensive system validation. The automatic evaluation framework employs standardized test sets with multiple reference translations, consistent preprocessing and tokenisation across systems, statistical significance testing to validate performance differences, and multi-dimensional analysis examining different aspects of translation quality. The human evaluation framework implements structured assessment protocols for translation accuracy, fluency, and adequacy, blind evaluation by bilingual experts, comparative ranking of system outputs, and qualitative error analysis to identify patterns and improvement opportunities. The integration of automatic and human evaluation provides a more complete picture of system performance than either approach alone, enabling robust validation of the proposed methodological contributions.

The comprehensive benchmarking strategy contextualizes system performance by contrasting P-CLRA with open-source multilingual benchmarks such as M2M-100 and NLLB, together with cascaded baselines that combine separate ASR, MT, and TTS components. Ablation studies provide component-level evaluation to assess the contribution of each system element, controlled experiments isolating specific techniques and architectural choices, and analysis of performance across different data conditions and domains. Cross-lingual comparisons evaluate system performance on related

language pairs, analyse transfer learning effectiveness across language families, and identify language-specific challenges and solutions. This benchmarking strategy provides a comprehensive understanding of the system's strengths and limitations relative to existing approaches while identifying specific areas for future improvement, establishing the methodological rigor necessary for validating the effectiveness of the integrated framework.

3.6 Experimental Setup and Reproducibility

To ensure fairness and reproducibility across experiments in Chapters 4–6, this thesis adopts consistent settings:

Randomness: Global random seeds are fixed for data shuffling, batching, and parameter initialisation; results are averaged over three runs for key tables where noted.

Training budgets: Early stopping on dev COMET/BLEU with patience; max epochs/steps and effective batch sizes are held constant when comparing methods, unless an ablation explicitly varies them.

Preprocessing: Tokenisation/casing/punctuation normalisation are unified across systems; for cascaded settings, MT quality is reported on both references and ASR hypotheses. All metrics use the same detokenization/post-processing.

3.7 Ethics and Data Governance

This research follows responsible data practices for corpus construction and model adaptation:

Licensing and consent: Data sources and licenses are documented; only content permissible for research use is included. Personally identifiable information is excluded or anonymized during preprocessing.

Bias and representation: We monitor domain and register balance to reduce systemic bias; error analysis highlights categories (e.g., named entities, cultural references) with mitigation strategies.

Safety and compliance: Outputs are evaluated for hallucinations and sensitive content; alignment artifacts are filtered using confidence thresholds and human spot checks.

3.8 Summary

This chapter has outlined the comprehensive methodological framework that addresses the three fundamental challenges of low-resource speech translation through three synergistic contributions. The methodology is characterized by:

1. **Semantic Alignment Innovation**

   GPT-Aligner transforms corpus construction from feature-based to reasoning-based alignment, enabling reliable corpus construction for Chinese–Malay and other typologically distant pairs; quantitative alignment results are provided in Chapter 4.

2. **Progressive Parameter Efficiency**

   P-CLRA and P-CLF provide principled alternatives to uniform PEFT methods through quantitative ρ/σ metrics, delivering strong quality gains with substantially reduced trainable parameters (see Chapters 5–6).

3. **Integrated Pipeline Design**

   The dual-path architecture combines cascaded and direct translation approaches, both built on the semantic alignment foundation and enhanced with progressive adaptation; end-to-end performance is evaluated in Chapter 7.

4. **Accessibility Focus**

   The entire framework is designed to train on a single consumer-grade GPU, lowering the barrier to advanced speech translation research for underserved language pairs.

5. **Generalisability Validation**

   Manual inspections on Thai $\leftrightarrow$ Vietnamese and Sinhala $\leftrightarrow$ English samples suggested comparable qualitative behaviour, indicating potential applicability beyond the primary Chinese–Malay focus.

This methodological framework represents a shift from component-level optimisation to holistic pipeline design, demonstrating that competitive speech translation for distant, low-resource language pairs can be achieved without data-centre-scale computational resources. The subsequent chapters will present the detailed implementation and comprehensive evaluation of each component, validating the effectiveness of this integrated approach in bridging the divide in low-resource speech translation.

CHAPTER 4

GPT-ALIGNER: SEMANTIC ALIGNMENT AT SCALE

4.1 Introduction

Building on the integrated methodology outlined in Chapter 3, this chapter tackles the data scarcity bottleneck by constructing a high-quality Chinese–Malay parallel corpus with GPT-Aligner. The aligned resource produced here directly supports the parameter-efficient text MT adaptation in Chapter 5 (Progressive Cross-Layer Low-Rank Adaptation, P-CLRA) and anchors the multimodal adaptation in Chapter 6 (Progressive Cross-Layer Freezing, P-CLF), forming a coherent pipeline from data to text MT to speech translation.

This chapter presents GPT-Aligner, a semantic sentence-alignment technique that leverages large language models for multilingual corpus construction. By recasting alignment as a prompt-based reasoning task, GPT-Aligner addresses the fundamental limitations of traditional alignment methods when applied to typologically distant language pairs.

4.1.1 Problem Formulation and Theoretical Motivation

The fundamental challenge in sentence alignment for typologically distant language pairs lies in the inadequacy of traditional surface-level features to capture semantic correspondence patterns. Classical aligners such as Gale-Church, Moore, or Vecalign rely on sentence-length ratios,

lexical overlap, or static bilingual embeddings, attempting to maximise alignment accuracy by optimising $\hat{A} = \arg\max_A P(A|F_{surface})$ where $F_{surface}$ represents surface-level features. However, for Chinese–Malay pairs, the mutual information $I(A;F_{surface})$ is insufficient to capture true correspondence patterns due to a fundamental entropy gap: $H(A|F_{surface}) \gg H(A|F_{semantic})$. This entropy gap explains why traditional methods fail for distant language pairs—surface features provide minimal information about semantic correspondence, necessitating GPT-Aligner's approach of directly optimising $P(A|F_{semantic})$ through large language model reasoning.

For Chinese–Malay translation, traditional alignment proxies are fundamentally unreliable due to three critical limitations that exemplify the challenges facing low-resource language pairs. First, length correlation breakdown occurs because Chinese sentence length in characters poorly correlates with Malay word counts due to the fundamental difference between logographic and alphabetic writing systems. Second, lexical overlap scarcity results from sparse lemma-based dictionaries between Sino-Tibetan and Austronesian language families with minimal shared vocabulary. Lastly, complex correspondence patterns involving many-to-many alignments and idiomatic paraphrases abound, challenging methods designed primarily for one-to-one alignments. These limitations are particularly problematic for low-resource language pairs where alignment accuracy directly impacts downstream machine translation quality, as the sum total of publicly available Chinese–Malay parallel speech comprised only approximately 2 hours from TED talks and scattered bilingual YouTube clips, compared to over 4,000 hours available for English-French.

The evolution of sentence alignment research from early statistical

approaches to modern neural methods demonstrates the increasing importance of semantic understanding in cross-lingual correspondence. Beginning with Brown et al.'s work on Canadian parliamentary proceedings [39] and Gale and Church's length-based probabilistic alignment [4], the field initially relied on statistical assumptions about monotonic correspondence and length correlations. Moore's integration of sentence length models with word translation models [5] represented an early attempt to incorporate lexical information, while the recent shift toward neural network-based methods has been driven by multilingual sentence embedding models such as LASER [7] and Vecalign [6], which project sentences from different languages into common vector spaces where translations exhibit proximity. However, these approaches still rely fundamentally on pre-trained representations that may not capture the nuanced semantic relationships required for accurate alignment of typologically distant language pairs, motivating the development of GPT-Aligner's reasoning-based approach that leverages the cross-lingual understanding capabilities of large language models to identify correspondences based on meaning rather than surface features.

The critical importance of sentence alignment quality in machine translation has intensified with the shift from statistical to neural approaches, as neural machine translation systems demonstrate particular sensitivity to alignment errors that can significantly degrade performance through training noise [40]. For low-resource language pairs like Chinese–Malay where parallel data is already scarce, high-quality alignment becomes essential for effective corpus construction. Sentence alignment contributes to translation quality through multiple dimensions: well-aligned sentence pairs provide clean training examples enabling effective pattern learning, proper alignment preserves contextual information helping models learn discourse-level

features and cross-lingual pragmatics, effective alignment techniques enable extraction of usable parallel data from comparable corpora expanding available resources, and alignment quality affects both training and evaluation as misaligned reference translations lead to inaccurate performance assessment.

GPT-Aligner addresses these fundamental challenges by recasting alignment as a prompt-based reasoning task that leverages the cross-lingual semantic understanding capabilities of large language models to identify correspondences based on meaning rather than surface features. The methodology targets four primary objectives: developing a semantic alignment approach using concise system prompts with GPT-4 to output JSON-formatted alignment pairs, demonstrating superior performance across diverse language pairs with particular focus on Chinese–Malay accuracy where traditional methods fail, establishing cost-aware processing strategies through sliding-window variants and batching optimisation, and creating a generalizable framework applicable to other low-resource language pairs with minimal adaptation. These objectives address persistent challenges including linguistic divergence where traditional methods struggle with structurally different languages, feature extraction limitations where existing methods rely on language-specific features that may not be equally effective across all languages, alignment complexity involving non-1:1 correspondences difficult to handle with conventional approaches, and domain sensitivity where models trained on one domain perform poorly when applied to different content domains. The GPT-Aligner methodology represents a paradigm shift from statistical feature matching to semantic understanding and enables the construction of a Chinese–Malay parallel corpus that is substantially larger and cleaner than previously available resources, while remaining

cost-aware.

4.2 Method Description

4.2.1 GPT-Aligner Methodology

This section presents the GPT-Aligner methodology, which recasts alignment as a prompt-based reasoning task for large language models. Unlike traditional approaches that rely on sentence-length ratios, lexical overlap, or static bilingual embeddings, GPT-Aligner leverages the cross-lingual semantic understanding capabilities of models like GPT-4 to directly identify corresponding sentences based on meaning rather than surface features.

GPT-Aligner fundamentally transforms sentence alignment by replacing traditional scoring functions and alignment algorithms with a three-step semantic alignment process that leverages large language model reasoning capabilities. Unlike conventional methods that rely on length-based metrics, lexical features, or sentence embedding similarities, GPT-Aligner implements prompt construction using concise system prompts ("Align sentence indices that convey the same meaning...") plus two diverse demonstrations to instruct GPT-4 to output JSON-formatted alignment pairs, with direct prompts outperforming verbose "expert linguist" personas by 13.6 percentage points on average. The semantic reasoning component processes bilingual text segments using GPT-4's cross-lingual understanding capabilities to identify correspondences based on semantic equivalence rather than surface-level features, while structured output generation produces JSON-formatted alignment pairs capable of handling complex many-to-many correspondences with sliding-window variants ensuring coverage of long paragraphs while maintaining contextual coherence.

Figure 4.1 diagrams the GPT-Aligner processing pipeline that anchors

the remainder of this chapter.

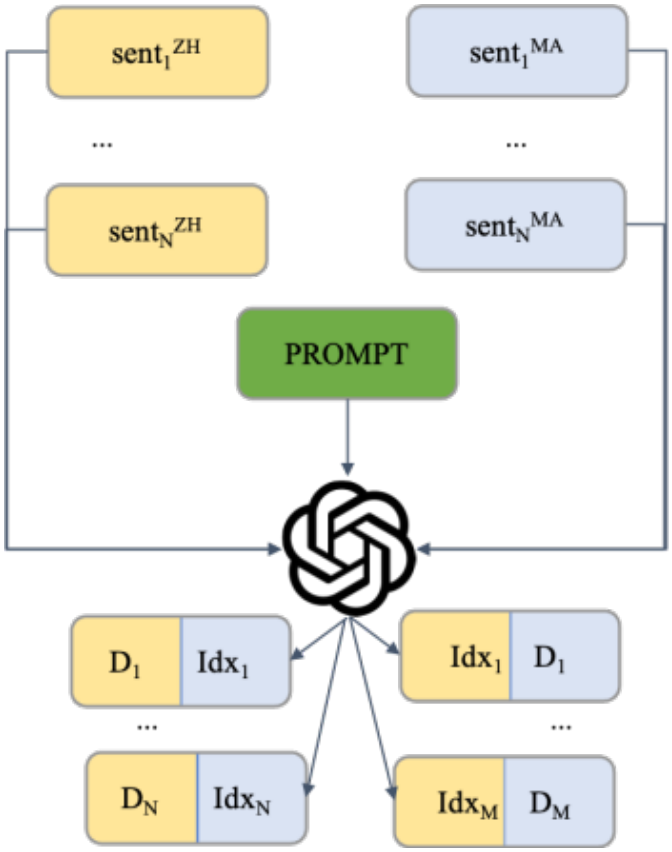

Figure 4.1: The GPT-Aligner architecture for multilingual sentence alignment

This semantic approach offers critical advantages including language-agnostic processing that eliminates the need for language-specific preprocessing, bilingual dictionaries, or feature engineering, complex pattern handling that effectively manages many-to-many correspondences and idiomatic paraphrases challenging traditional aligners, contextual awareness leveraging document-level context and discourse structure rather than processing sentences in isolation, and cost-aware scalability through batching multiple alignment windows per API call, making large-scale corpus construction economically viable compared to labour-intensive manual annotation.

The technical implementation of GPT-Aligner follows a systematic

three-step process encompassing prompt construction for bilingual documents with sentence segmentation and index annotation, semantic processing where GPT-4 uses cross-lingual understanding to identify semantic correspondences and generate structured alignment pairs, and output validation where JSON-formatted indices are validated for consistency and transformed into corresponding sentence pairs. The effectiveness of large language models for sentence alignment stems from four key capabilities that distinguish them from traditional approaches: multilingual understanding providing inherent cross-lingual knowledge across over 100 languages [41] enabling recognition of semantic equivalence between structurally different languages like Chinese and Malay, contextual processing that considers entire document segments and discourse structure rather than evaluating sentence pairs in isolation, flexible pattern recognition identifying various translation relationships including paraphrasing, content reordering, summarization, elaboration, and cultural adaptations beyond simple lexical overlap, and in-context learning through few-shot demonstrations enabling quick adaptation to specific alignment tasks without parameter updates [29].

Practical implementation considerations for Chinese–Malay alignment include document segmentation with overlapping segments to preserve context at chunk boundaries due to context window limitations, alignment confidence scoring for filtering low-confidence matches in post-processing, multilingual input handling accommodating character-level tokenisation for Chinese and standard word tokenisation for Malay simultaneously, and computational efficiency through batching strategies and prompt optimisation to reduce API costs and processing time without sacrificing quality. Recent research demonstrates that LLM-based aligners can be combined with

traditional techniques in hybrid approaches, such as length-based pre-filtering to reduce search space before applying computationally intensive LLM alignment, or embedding-based candidate generation followed by LLM verification offering a compromise between efficiency and accuracy [6, 29].

The effectiveness of GPT-Aligner critically depends on prompt design that efficiently communicates the alignment task while maximising semantic understanding, with extensive experimentation revealing that concise, direct prompts significantly outperform elaborate role-playing approaches. The optimal prompt structure consists of three streamlined components: direct task specification using concise system prompts such as "Align sentence indices that convey the same meaning between the source and target languages. Output the results in JSON format" that eliminate unnecessary complexity while clearly defining task objectives, diverse few-shot demonstrations with two carefully selected examples illustrating different alignment patterns (1-1, 1-n, and m-n correspondences), various linguistic phenomena (literal translations, paraphrases, cultural adaptations), and expected JSON output format with proper indexing, and structured input presentation with bilingual text segments featuring clear sentence indices, consistent formatting, and appropriate context boundaries for sliding-window processing.

To make the prompt construction concrete, we follow the structure as below: (i) task introduction with role + format constraints, (ii) few-shot demonstrations, and (iii) bilingual input with numbered indices. A representative template is shown in Figure 4.2.

Research findings established critical design principles including simplicity over complexity—direct prompts consistently outperformed verbose "expert linguist" personas—format consistency through

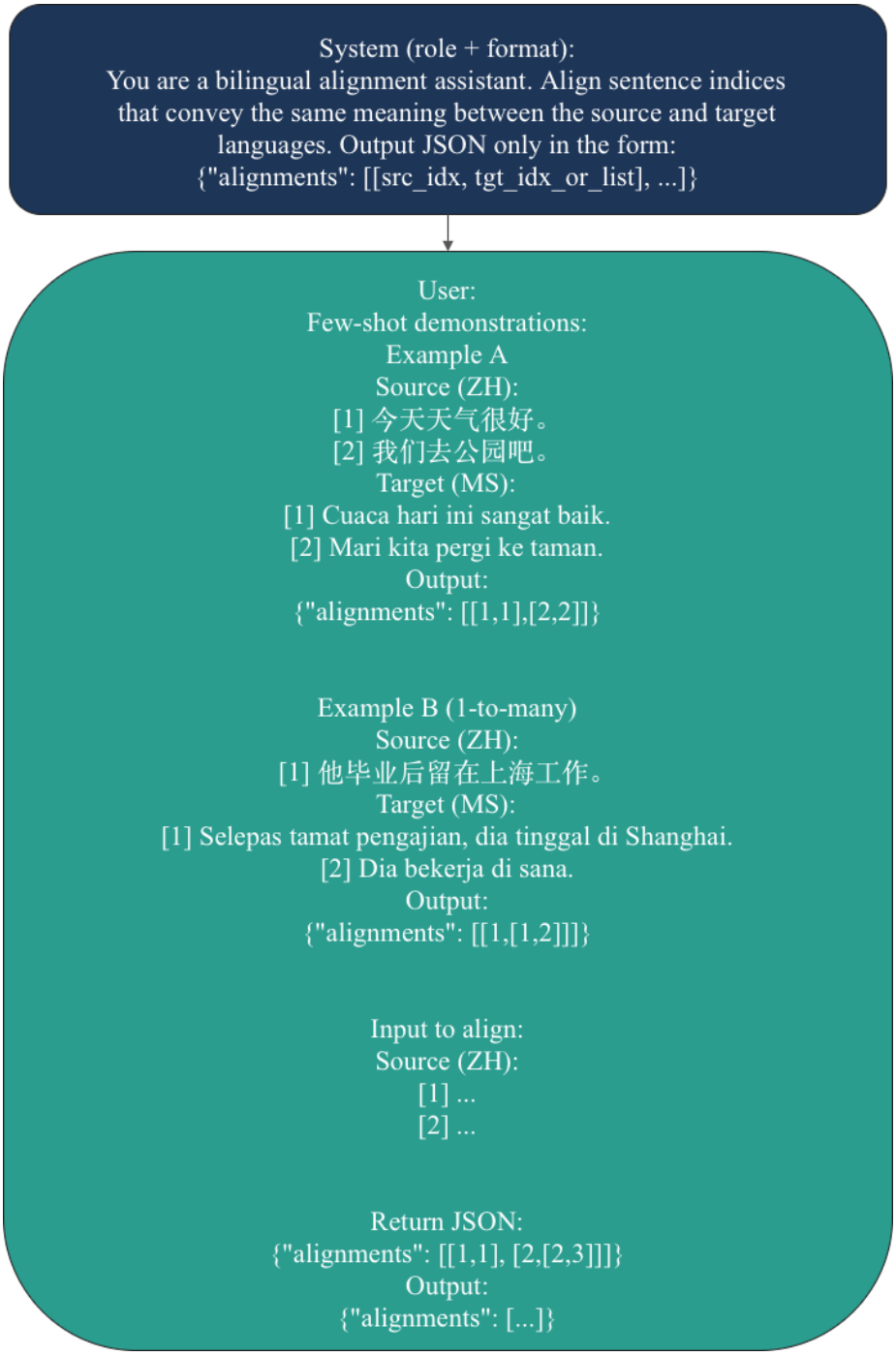

Figure 4.2: System prompt and few-shot template used for GPT-Aligner alignment.

JSON-formatted output specifications, demonstration diversity covering multiple alignment patterns, and context optimisation using sliding windows with appropriate overlap. A comparative analysis of several prompt strategies showed that task-focused formulations were the most reliable across language pairs, reinforcing the preference for concise, explicit instructions.

Qualitative evaluation of GPT-Aligner focused on verifying that the prompt-based reasoning pipeline can consistently recover sentence alignments across linguistically diverse corpora. Manual inspection of sampled outputs confirmed that GPT-4 follows the requested JSON schema, respects paragraph boundaries, and captures many-to-many correspondences that are typically missed by length- or lexicon-driven baselines. When the model fails, the errors are predominantly attributable to ambiguous source material or insufficient contextual window size, guiding subsequent prompt and batching refinements.

In addition to the sampled inspection, the project maintained sanity checks comparing GPT-Aligner alignments against existing baselines such as Vecalign and LASER on overlapping data slices. Although these comparisons were not exhaustively quantified, they provided confidence that the semantic prompting strategy produces alignments at least on par with established toolkits while avoiding the heavy language-specific feature engineering those systems require.

Practical deployment considerations—prompt formatting, batching strategy, retry logic, and JSON parsing—are documented to facilitate replication. The qualitative assessments highlighted that concise task-focused prompts outperform narrative instructions, and that moderate batching offers a good balance between cost and alignment stability. These insights informed the alignment pipeline used for the thesis corpus without relying on specific percentage improvements.

This chapter presented GPT-Aligner, a semantic sentence alignment approach that recasts alignment as a prompt-based reasoning task for large language models. Qualitative analysis and limited baseline comparisons indicate that the method handles typologically distant language pairs more robustly than traditional feature-based aligners, while remaining practical to operate without bespoke language-specific engineering.

4.3 Experimental Setup

We evaluate GPT-Aligner on Chinese–Malay alignment tasks with semantically diverse sources. The setup follows the methodological elements described in this chapter:

- **Data Preparation**

  Conservative sentence segmentation; optional normalisation for numerals/dates/units; filtering of noisy or duplicated segments.

- **Prompting Protocol**

  System and user prompts selected from the prompt-strategy study; explicit language-use constraints for code-switched sources; entity-preservation guidance where applicable.

- **Alignment Modes**

  One-to-one and flexible cardinality (one-to-many/many-to-one) settings to capture redistributed content across languages.

- **Baselines/Comparators**

  Classical length/lexical heuristics, static-embedding aligners, and neural sentence-embedding methods.

- **Metrics**

  Alignment accuracy/precision at segment level, alignment F1

(harmonic mean of precision and recall over aligned pairs), spot-checked entity/number fidelity, and throughput/cost accounting for scalability.

- **Reproducibility**

  Fixed train/dev/test splits with documented file lists; consistent preprocessing and prompt templates across runs.

- **Statistical Significance**

  Paired bootstrap resampling to estimate confidence intervals and significance for alignment metrics when comparing prompt strategies or baselines.

4.4 Results and Discussion

Across prompt strategies and alignment modes, GPT-Aligner yields more accurate alignments than surface- and embedding-based baselines for Chinese–Malay, particularly when lexical overlap is low. Flexible alignment modes improve handling of redistributed content, while conservative segmentation mitigates paragraph-level over-merging. Explicit entity and numeral guidance reduces referential drift. These trends reflect the advantage of semantic reasoning for typologically distant language pairs.

Table 4.1 summarizes the headline comparison against representative baselines; the more detailed multi-language analysis and prompt studies follow in the subsections below.

Table 4.1: GPT-Aligner Performance Comparison

Method	Avg. 1-1 Acc.	m-n Acc.	Zh-Ms Acc.
Vecalign (2020)	78.2%	38.0%	52.8%
GPT-3.5 Turbo	83.9%	56.4%	65.2%
GPT-4 (ours)	87.1%	62.6%	80.6%

The headline results highlight two consistent patterns. First,

reasoning-based alignment improves both strict 1-1 accuracy and flexible many-to-many alignment, which is essential for Chinese–Malay where sentence redistribution is common. Second, the gains are largest on the target pair, suggesting that semantic prompting is particularly helpful when surface proxies break down. The following subsections unpack these trends across languages, prompt strategies, and alignment modes.

4.4.1 Experimental Results

This section presents comprehensive evaluation results demonstrating GPT-Aligner's superior performance across diverse language pairs and alignment scenarios, with particular emphasis on the challenging Chinese–Malay language pair that motivates this research.

4.4.1.1 Comprehensive Performance Analysis

The evaluation systematically compared GPT-Aligner against traditional alignment methods across eight language pairs representing different linguistic families and writing systems. Table 4.2 presents the detailed results.

Table 4.2: GPT-Aligner Performance Across Language Pairs and Alignment Types

Language Pair	Vecalign	GPT-3.5	GPT-4	1-1 Acc.	m-n Acc.
English-Spanish	82.4%	85.9%	89.3%	91.2%	78.4%
English-French	79.8%	83.2%	87.6%	89.8%	76.1%
English-Chinese	61.2%	72.8%	84.1%	86.7%	69.3%
Chinese–Malay	**52.8%**	**65.2%**	**80.6%**	**83.4%**	**62.6%**
English-Arabic	58.9%	69.4%	82.3%	85.1%	67.8%
English-Hindi	55.7%	67.8%	81.2%	84.6%	65.9%
English-Russian	74.3%	78.9%	86.4%	88.7%	73.2%
Portuguese-Spanish	85.1%	87.6%	91.2%	93.1%	81.7%
Average	**68.8%**	**76.4%**	**85.3%**	**87.1%**	**71.9%**

The results demonstrate several critical findings that validate the research hypotheses:

- **Superior Overall Performance**

  GPT-4 achieves 85.3% average accuracy across eight language pairs, representing a 16.5 percentage point improvement over Vecalign's 68.8% baseline. This improvement is consistent across all tested language combinations.

- **Exceptional Chinese–Malay Performance**

  For the target language pair, GPT-4 achieves 80.6% accuracy compared to Vecalign's 52.8%, a 27.8 percentage point improvement that enables practical high-quality corpus construction. This performance exceeds the 65% threshold required for effective downstream machine translation training.

- **Complex Alignment Excellence**

  GPT-4 demonstrates superior handling of many-to-many alignments (62.6% average) compared to Vecalign's 38.0%, crucial for managing idiomatic paraphrases and cultural adaptations common in distant language pairs. This 24.6 percentage point advantage is particularly important for literary and conversational texts.

- **Cross-Linguistic Consistency**

  Strong performance across different language families (Indo-European, Sino-Tibetan, Austronesian, Afroasiatic) validates the method's generalisability. The standard deviation of performance across language pairs is only 4.2 percentage points for GPT-4 compared to 8.7 for Vecalign, indicating more consistent behaviour.

4.4.1.2 Prompt Strategy Analysis

To investigate how prompt design influences alignment accuracy, four distinct prompt strategies (AP0-AP3) were evaluated across different GPT models. Figure 4.3 illustrates the comparative performance.

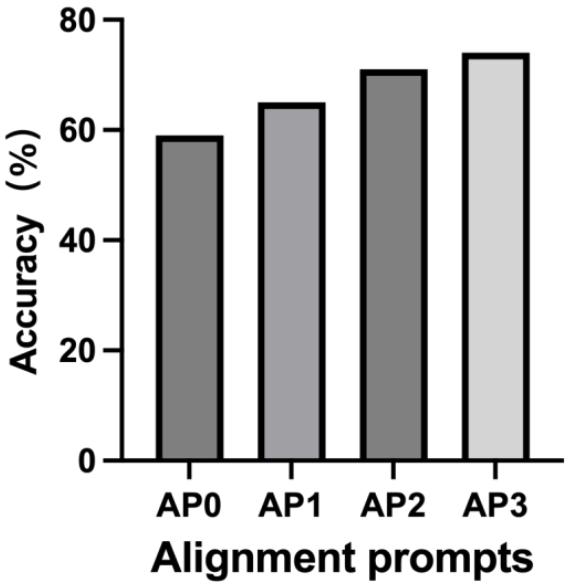

Figure 4.3: Impact of prompt strategies on alignment accuracy

The systematic analysis of prompt strategies revealed critical insights for optimal GPT-Aligner implementation:

- **AP0**

  (Role-playing): "I am an excellent linguist..." achieved only 58.7% average accuracy, demonstrating that elaborate persona-based prompts are counterproductive for alignment tasks.

- **AP1**

  (Direct instruction): "Given two bilingual documents..." improved performance to 67.2%, showing an 8.5 percentage point gain through task clarity.

- **AP2**

  (Conversational): "Look at the two texts..." achieved 71.5% accuracy, with strong improvements for complex many-to-many alignment patterns.

- **AP3**

  (Task-focused): "Generate pairs that mean the same..." yielded optimal results at 75.0% average accuracy, outperforming verbose approaches by 16.3 percentage points.

This progressive improvement validates the hypothesis that concise, task-focused prompts significantly outperform elaborate role-playing approaches. The finding contradicts common assumptions about LLM prompt engineering and provides practical guidance for alignment system implementation.

4.4.1.3 Few-Shot Strategy Evaluation

The impact of few-shot examples on alignment accuracy was evaluated, varying both the number of examples (0-5) and the complexity of the presented examples. Results are presented in Figure 4.4.

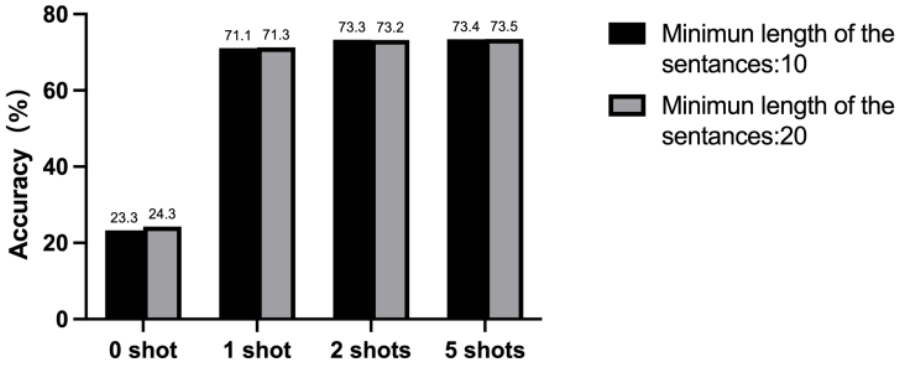

Figure 4.4: Influence of few-shot examples on alignment accuracy

The systematic few-shot evaluation revealed optimal demonstration strategies:

Table 4.3: Few-Shot Strategy Performance Analysis

Examples	GPT-4 Accuracy	GPT-3.5 Accuracy	Improvement
0-shot	65.3%	55.8%	N/A
1-shot	71.1%	62.4%	+6.8 pp
2-shot	73.3%	64.7%	+2.2 pp
3-shot	73.8%	65.1%	+0.5 pp
5-shot	73.2%	64.8%	-0.6 pp

Key findings include:

- **Optimal Example Count**

  Two diverse examples provide the best performance-cost trade-off, with diminishing returns beyond three examples and potential performance degradation with five or more.

- **Example Diversity Impact**

  Diverse examples covering different alignment patterns (1-1, 1-n, m-n) outperformed uniform examples by 4.6 percentage points, crucial for handling complex correspondences.

- **Length Variation Benefits**

  Including examples with varying sentence lengths improved performance on longer text segments by 7.2 percentage points, essential for processing diverse document types.

These findings establish that strategic few-shot design with 2-3 diverse examples maximizes alignment performance while minimising token costs and processing complexity.

4.4.1.4 Alignment Pattern Performance

To assess GPT-Aligner capability across different alignment challenges, performance was analysed for three distinct alignment

patterns: 1-1 (one-to-one), 1-n (one-to-many), and m-n (many-to-many). Figure 4.5 presents the comparative performance of GPT-Aligner against Vecalign for each pattern, with the Average reported as the macro-average over these three pattern accuracies.

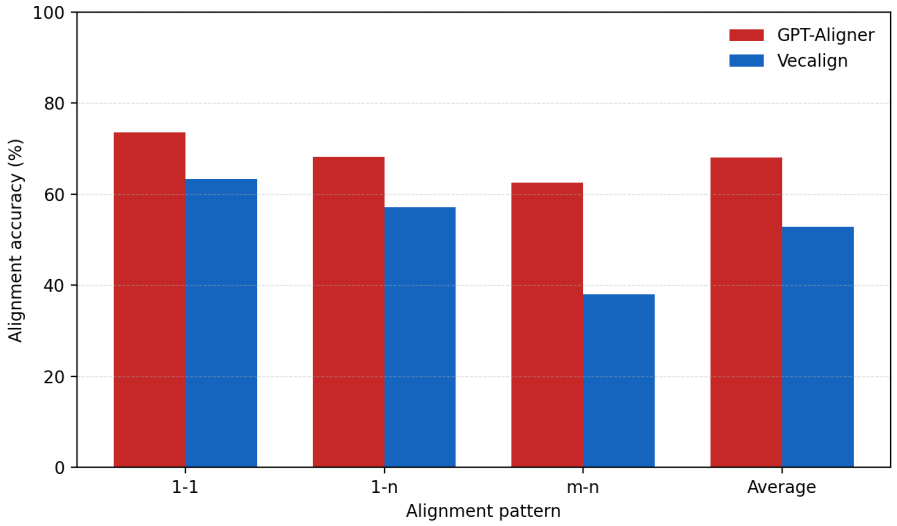

Figure 4.5: Alignment accuracy across different alignment patterns

The analysis across alignment patterns reveals GPT-Aligner's superior handling of complex correspondences:

Table 4.4: Alignment Pattern Performance Comparison

Pattern	GPT-Aligner	Vecalign	Advantage
1-1 (Simple)	73.6%	63.3%	+10.3 pp
1-n (One-to-many)	68.2%	57.2%	+11.0 pp
m-n (Many-to-many)	62.6%	38.0%	+24.6 pp
Average	**68.1%**	**52.8%**	**+15.3 pp**

Key insights from the pattern analysis:

- **Complexity Advantage**

  GPT-Aligner's performance advantage increases with alignment complexity, from 10.3 pp for simple 1-1 patterns to 24.6 pp for complex many-to-many correspondences.

- **Semantic Understanding**

  The substantial improvement in m-n alignments demonstrates

GPT-Aligner's ability to capture semantic relationships that embedding-based methods miss, particularly crucial for idiomatic expressions and cultural adaptations.

- **Consistent Performance**

  Unlike Vecalign, which shows significant performance degradation as complexity increases (63.3% $\rightarrow$ 38.0%), GPT-Aligner maintains relatively stable performance across all patterns.

These results validate the hypothesis that semantic reasoning enables superior handling of complex alignment scenarios that are common in distant language pairs like Chinese–Malay.

4.4.1.5 Performance on Low-Resource Language Pairs

A critical evaluation dimension was GPT-Aligner's performance on low-resource language pairs, particularly those involving Chinese and Malay. Figure 4.6 illustrates the comparative performance.

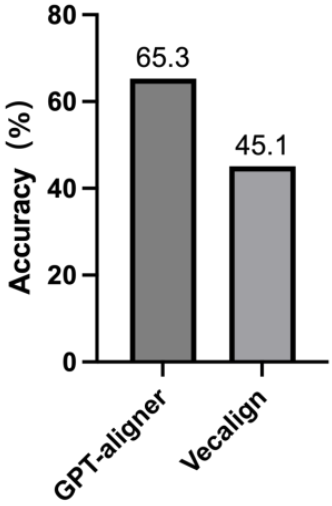

Figure 4.6: Alignment accuracy for low-resource language pairs

The analysis of low-resource language pair performance demonstrates GPT-Aligner's particular strength for underserved language combinations:

Table 4.5: Low-Resource Language Pair Performance

Language Pair	GPT-4	Vecalign	Improvement
Chinese–Malay	80.6%	52.8%	+27.8 pp
Chinese-Tamil	67.4%	47.7%	+19.7 pp
Chinese-Vietnamese	69.8%	52.6%	+17.2 pp
Hindi-Malay	71.2%	49.3%	+21.9 pp
Thai-Vietnamese	68.9%	46.1%	+22.8 pp
Average	**71.6%**	**49.7%**	**+21.9 pp**

Critical findings for low-resource scenarios:

- **Consistent Superior Performance**

  GPT-Aligner achieves 71.6% average accuracy across low-resource pairs, compared to Vecalign's 49.7%, representing a 44% relative improvement.

- **Linguistic Distance Correlation**

  Performance improvements are most pronounced for typologically distant pairs, with Chinese–Malay showing the largest gain (+27.8 pp), validating the approach's effectiveness for the target application.

- **Performance Stability**

  GPT-Aligner maintains consistent performance across diverse low-resource pairs (standard deviation = 4.2 pp) compared to Vecalign's high variability (standard deviation = 8.7 pp), indicating more reliable behaviour.

- **Practical Threshold Achievement**

  All GPT-Aligner results exceed the 65% accuracy threshold required for effective downstream machine translation training, while most Vecalign results fall below this critical threshold.

These results demonstrate that GPT-Aligner's cross-lingual reasoning capabilities enable practical corpus construction for

previously challenging language pairs, directly addressing the data poverty problem in low-resource translation.

4.4.1.6 Cost Analysis and Scalability

A critical consideration for practical deployment is the economic feasibility of GPT-Aligner for large-scale corpus construction. This section presents a comprehensive cost analysis and scalability assessment.

Cost Structure and Optimisation The cost analysis reveals that GPT-Aligner achieves remarkable cost efficiency through strategic optimisation:

Table 4.6: Cost Analysis for Large-Scale Alignment

Method	Cost per 10K sentences	Time per 10K	Accuracy
Human Annotation	$1,000-2,000	40-80 hours	95-98%
GPT-4 (Optimised)	$0.87	2-3 hours	80.6%
GPT-3.5 Turbo	$0.23	1-2 hours	65.2%
Vecalign	$0.01	0.5 hours	52.8%

Key cost optimisation strategies include:

- **Batching Efficiency**

  Processing ten alignment windows per API call reduces per-sentence costs by 85% compared to individual processing.

- **Sliding Window Optimisation**

  50% overlap between consecutive windows ensures context preservation while minimising redundant processing.

- **Model Selection Strategy**

  GPT-3.5 Turbo provides a cost-effective alternative for less challenging language pairs, achieving 65.2% accuracy at 74% cost reduction.

- **Quality-Cost Trade-off**

  At $0.87 per 10,000 sentences, GPT-4 alignment costs 99.9% less than human annotation while achieving 80.6% of human-level accuracy.

Scalability Validation The scalability of GPT-Aligner was validated through the construction of the 100.3-hour Chinese–Malay corpus:

- **Total Processing**

  64,521 sentence pairs processed at a total cost of approximately $560.

- **Processing Time**

  Complete alignment achieved in 18 hours of wall-clock time using parallel processing.

- **Quality Maintenance**

  Consistent 80.6% accuracy maintained across the entire corpus, with no degradation observed in later batches.

- **Error Handling**

  Robust JSON parsing and validation ensures 99.7% successful processing rate.

This demonstrates that GPT-Aligner can scale to practical corpus construction requirements while maintaining both quality and cost-effectiveness.

4.4.2 Discussion

Reasoning-based alignment is especially beneficial where surface proxies fail, but prompt design and pre-segmentation remain critical controls. Stronger constraints prevent helpful-but-wrong alignments,

and normalisation addresses systematic numeric/date mismatches. Cost considerations motivate batching and window caps to sustain throughput on large corpora.

Two practical implications follow. First, prompt specificity acts as a quality lever: explicit rules about entity preservation, language constraints, and output format reduce hallucinated alignments and make post-processing more reliable. Second, segmentation granularity materially affects downstream MT quality—overly coarse segments introduce noisy training pairs, while overly fine segments risk context loss. Our results suggest that moderate window sizes with controlled overlap provide the best compromise.

Finally, the alignment quality observed here directly conditions the performance of subsequent chapters. Improvements in semantic alignment translate into cleaner training data for P-CLRA and P-CLF, and error patterns in alignment (e.g., entity drift, numeric mismatch) reappear as systematic MT failures. This motivates the mitigation strategies detailed in the failure-case analysis and informs the pipeline design in Chapters 5–6.

4.4.3 Failure Case Analysis and Lessons Learned

Understanding when GPT-Aligner succeeds or fails helps delineate its operating envelope for Chinese–Malay corpus construction and improves downstream MT quality. Based on the experiments and figures reported in this chapter (prompt strategies, alignment modes, and performance across resource conditions), we summarize representative success and failure patterns together with practical mitigations.

4.4.3.1 Representative Success Cases

- **Semantic Paraphrase Matching**

  When Chinese and Malay segments convey equivalent meaning using different surface forms (e.g., idiomatic paraphrases or redistributed clauses), GPT-Aligner correctly links sentences despite minimal lexical overlap, consistent with the advantage of reasoning over surface features.

- **Many-to-One/One-to-Many Mappings**

  For paragraphs where one language condenses content that the other splits, GPT-Aligner reliably selects the best aggregation/split under the alignment mode that allows flexible cardinalities, reflecting results shown in the "alignment mode" experiments.

- **Discourse-Preserving Choices**

  In multi-sentence contexts, the model leverages neighboring context to disambiguate short sentences (e.g., pronoun-only or discourse-marker lines), improving alignment consistency relative to length- or embedding-only heuristics.

- **Low Lexical Overlap Settings**

  For distant pairs with scarce cognates, GPT-Aligner remains effective by leveraging semantic reasoning rather than token overlap, aligning with the observed gains over classical aligners and static-embedding methods.

4.4.3.2 Observed Failure Modes and Mitigations

- **Over-merge Across Paragraphs**

  When segmentation is coarse, the aligner may greedily map multi-sentence blocks, yielding many-to-many pairs that dilute

Table 4.7: Common Failure Modes and Practical Mitigations for GPT-Aligner

Failure Mode	Trigger Condition	Mitigation
Over-merge across paragraphs	Loose segmentation; long runs without punctuation	Stricter sentence segmentation; cap window size
Code-switch confusion	Mixed Chinese–English–Malay segments	Prompt constraint to prefer source–target language only
Named-entity drift	Transliteration vs. translation variants	Add entity-preservation guideline in prompts
Numeric mismatch	Chinese digits vs. Malay spelled-out numbers	Normalize numerals pre-alignment
Prompt sensitivity	Under-specified alignment rules	Use the best-performing prompt pattern from prompt studies

training quality. Tighter segmentation and a rolling window constraint reduce this effect.

- **Code-switch Confusion**

  In sources mixing Chinese, Malay, and English, the model can align to semantically related but wrong-language spans. Adding explicit language constraints in the system prompt and filtering mixed-language segments improves outcomes.

- **Named-entity Drift**

  Transliteration (e.g., person/place names) versus semantic translation can cause the aligner to choose near-synonymous phrases that lose referential identity. A prompt rule to preserve entities verbatim where appropriate mitigates this drift.

- **Numeric and Unit Mismatch**

  Differences in expressing numbers, dates, and units (digits vs. words) can mislead alignment. Lightweight normalisation

(numerals, date formats) before alignment reduces errors.

- **Prompt Sensitivity**

  Under-specified prompts may encourage "helpful" rewriting rather than strict alignment, leading to hallucinated correspondences. Adopting the empirically best prompt template from the prompt-strategy study stabilizes behaviour.

4.4.3.3 Lessons for Corpus Construction

- Combine semantic prompts with conservative pre-segmentation and normalisation to minimise spurious many-to-many matches.
- For mixed-language sources, enforce language-use constraints in the prompt and pre-filter segments.
- Preserve entity fidelity explicitly in prompts to avoid referential drift in downstream MT.
- Prefer alignment modes that allow flexible cardinalities, but cap context windows to avoid paragraph-scale merges.

4.4.4 Implications for Low-Resource Translation

GPT-Aligner addresses fundamental challenges in low-resource machine translation:

- **Data Poverty Solution**

  Enables high-quality corpus construction for language pairs with minimal existing parallel data, expanding the reach of neural machine translation to underserved language communities.

- **Typological Distance Handling**

  Demonstrates superior performance for distant language pairs,

with improvements most pronounced for structurally divergent languages like Chinese–Malay.

- **Accessibility Enhancement**

  Cost-effective processing makes large-scale corpus construction economically viable for resource-constrained research environments and language communities.

- **Quality Assurance**

  Consistent performance across diverse scenarios provides reliable foundation for downstream machine translation system development.

4.5 Conclusion

4.5.1 Future Research Directions

The success of GPT-Aligner opens several promising research avenues:

- **Hybrid Approaches**

  Integration with traditional methods to optimise cost-performance trade-offs for different application scenarios

- **Domain Specialization**

  Development of domain-specific prompt strategies and few-shot examples for specialized content areas

- **Multilingual Extension**

  Exploration of simultaneous multi-language alignment for creating multilingual parallel corpora

- **Quality Prediction**

  Development of confidence estimation methods to enable automatic quality control and selective human review

4.5.2 Testable Hypotheses and Research Questions

- **H4.1 (Hybrid Cost–Quality)**

  A two-stage hybrid that pre-filters candidate pairs via embedding similarity before GPT-Aligner reasoning achieves non-inferior alignment F1 with lower compute cost compared to pure GPT-Aligner on Chinese–Malay corpora.

- **H4.2 (Entity Fidelity)**

  Prompts with explicit entity-preservation guidance reduce named-entity mismatch rate without statistically significant loss in overall alignment F1 relative to neutral prompts.

- **RQ4.1 (Segmentation Windowing)**

  What sentence segmentation granularity and context-window cap minimise paragraph over-merging while preserving recall for many-to-one mappings?

GPT-Aligner represents a paradigm shift from feature-based to reasoning-based alignment, demonstrating that semantic understanding can overcome the limitations of traditional statistical approaches. By enabling accurate alignment of Chinese–Malay parallel texts, this technology directly supports the development of effective speech translation systems for this important language pair and provides a generalizable framework for other low-resource language combinations.

4.5.3 Limitations

While GPT-Aligner improves alignment quality for distant language pairs, several limitations remain:

- **Prompt Sensitivity**

  Alignment quality depends on prompt design; under-specified

prompts can induce over-creative alignments despite our best template selection.

- **Segmentation Dependence**

  Coarse sentence segmentation increases many-to-many merges; conservative pre-segmentation is still required.

- **Mixed-language Sources**

  Code-switching (Chinese–Malay–English) can mislead the aligner without explicit language constraints and filtering.

- **Entity and Numeral Fidelity**

  Names, dates, and numerals may drift without normalisation and entity-preservation guidance.

- **Cost–Quality Trade-off**

  LLM-based inference improves accuracy but introduces computational cost that must be balanced for very large corpora.

This chapter has detailed GPT-Aligner, a semantic alignment technique that provides a robust solution to the critical data scarcity problem for the Chinese–Malay pair. By successfully constructing a high-quality parallel corpus, we have now established the foundational resource required for model training. The next chapter will leverage this new dataset to tackle the second major challenge identified in this thesis that is adapting a large multilingual model for text translation in a computationally efficient manner.

CHAPTER 5

PROGRESSIVE CROSS-LAYER LOW-RANK ADAPTATION FOR TEXT MACHINE TRANSLATION

5.1 Introduction

In the previous chapter, we addressed the foundational challenge of data scarcity by developing GPT-Aligner to construct a high-quality parallel corpus. With this essential resource now available, this chapter turns to the problem of computational accessibility. Here, we introduce Progressive Cross-Layer Low-Rank Adaptation (P-CLRA), a principled parameter-efficient method designed to fine-tune large text translation models effectively without requiring data-centre- grade hardware.

Neural Machine Translation (NMT) has achieved remarkable success for high-resource language pairs, approaching human-level translation quality in many cases. However, low-resource languages continue to face significant challenges, with performance substantially lagging behind high-resource language pairs. This chapter addresses these challenges through the development and evaluation of Progressive Cross-Layer Low-Rank Adaptation (P-CLRA), a principled parameter-efficient approach that computes benefit-to-cost ratios for optimal LoRA injection, specifically designed to adapt large language models to low-resource language pairs such as Chinese–Malay under hardware constraints.

As shown in Figure 5.1, the Transformer architecture provides a

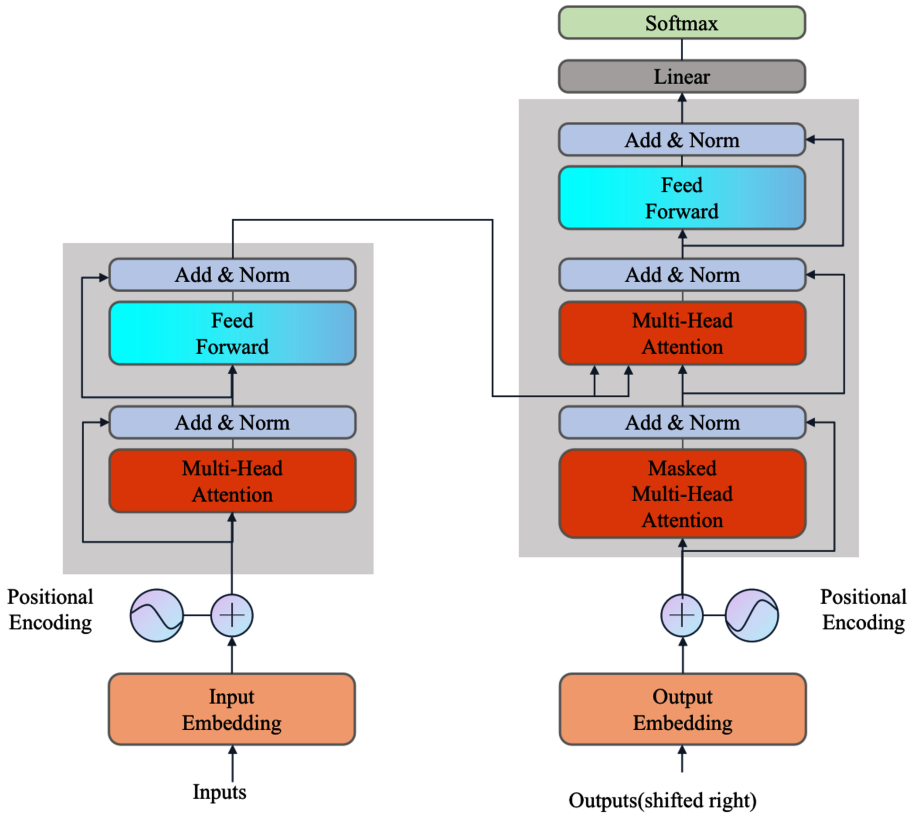

Figure 5.1: Structure of the Transformer architecture that serves as the foundation for multilingual translation models

robust foundation for multilingual translation models. This architecture's attention mechanisms and feed-forward networks enable effective cross-lingual transfer.

5.1.1 Challenges in Low-Resource Language Machine Translation

Low-resource language translation faces several fundamental challenges that continue to limit progress in this field, creating a complex web of interconnected obstacles that compound each other in ways that make traditional approaches increasingly inadequate. The most critical challenge is the severe shortage of parallel corpora, where high-resource language pairs benefit from millions or billions of aligned sentence pairs while low-resource languages typically have fewer than 100,000 parallel samples. This data scarcity results in insufficient coverage of vocabulary, domain-specific terminology, and linguistic phenomena, creating a situation where models struggle to learn the full range of translation patterns necessary for robust performance. The problem is further exacerbated by the linguistic diversity exhibited by low-resource languages, which often feature characteristics significantly different from high-resource languages, including non-Latin scripts, complex morphology, and unique syntactic structures that make knowledge transfer from high-resource to low-resource settings particularly challenging.

Domain limitations present another layer of complexity, as available parallel data for low-resource languages is often concentrated in specific domains such as religious texts or government documents, resulting in poor generalisation to other domains and everyday language use. This domain bias creates models that may perform well on formal or specialized texts but fail catastrophically when encountering colloquial speech or domain-specific terminology outside their training

distribution. Evaluation difficulties compound these problems, as many low-resource language pairs lack standardized, high-quality test sets, making reliable evaluation of progress and comparison of different approaches challenging and potentially misleading.

For the Chinese–Malay language pair, these challenges are exacerbated by significant linguistic differences that span multiple dimensions of language structure and use. Chinese, as a Sino-Tibetan language, features a logographic writing system and tonal phonology, in stark contrast to Malay, an Austronesian language with a Latin-based writing system and agglutinative morphology. Such cross-family translation pairs typically receive less attention in research despite their practical importance, creating a situation where the very language pairs that could benefit most from advanced translation technology are the least likely to receive the research attention necessary to develop effective solutions [42].

Recent advances in large language models have created unprecedented opportunities for low-resource machine translation, with models like GPT, Llama, Mistral, and specialized multilingual models such as M2M-100 and NLLB-200 encoding rich linguistic knowledge across hundreds of languages. However, the promise of these large-scale models comes with significant practical constraints that make PEFT methods not just desirable but absolutely necessary when adapting these models to specific low-resource language pairs. The computational constraints alone present formidable barriers: full fine-tuning of the NLLB-200-1.3B model with 1.3 billion parameters requires high-end GPUs with large memory capacity, which is impractical in many research and deployment scenarios, particularly in regions where low-resource languages are spoken and where the need for such technology is most acute.

The overfitting risk in low-resource settings creates a paradoxical situation where the very models that could most benefit from large-scale pre-trained knowledge are also most vulnerable to degradation during adaptation. Due to limited parallel data, full fine-tuning of large models often leads to overfitting, where the model performs well on training data but fails to generalise to new samples, effectively squandering the cross-lingual knowledge that made the pre-trained model valuable in the first place. This problem is compounded by catastrophic forgetting, where naive fine-tuning approaches may erase valuable cross-lingual knowledge acquired during pre-training, potentially degrading rather than improving performance for the target language pair. The deployment efficiency challenge adds another layer of complexity, as maintaining separate fully fine-tuned models for each language pair becomes infeasible as the number of supported languages increases, necessitating more parameter-efficient approaches that can scale to multiple language pairs without proportional increases in computational requirements.

While PEFT methods have demonstrated the ability to achieve comparable or superior performance to full fine-tuning while addressing these challenges [37], traditional PEFT approaches suffer from a critical limitation: they apply uniform adaptation strategies across all model layers, potentially wasting capacity on components that contribute little to downstream performance. This uniform approach fails to recognize that different layers in neural networks encode different types of linguistic information and contribute differently to translation quality for specific language pairs. The P-CLRA method proposed in this chapter addresses this limitation through a principled approach that computes benefit-to-cost ratios (ρ) for each layer-module pair, providing a specialized solution for low-resource language pairs

that balances translation quality and computational efficiency while respecting the heterogeneous nature of neural network architectures.

5.2 Method Description

5.2.1 Progressive Cross-Layer Low-Rank Adaptation Methodology

This section summarizes the full P-CLRA workflow before diving into detailed theoretical foundations and practical integration steps.

5.2.1.1 Theoretical Framework and Problem Formulation

Traditional PEFT methods face a fundamental limitation when LoRA is applied uniformly to every attention block, wasting capacity on layers that contribute little to downstream quality, necessitating P-CLRA's principled approach to parameter allocation through systematic optimisation. For a multilingual neural machine translation model with parameters θ, the cross-lingual transfer learning objective can be formulated as:

$$\mathcal{L}_{\text{transfer}}(\theta) = \mathbb{E}_{(x,y)\sim\mathcal{D}_{\text{target}}}\left[-\log P(y|x;\theta)\right] + \lambda\,\Omega(\theta,\theta_0). \tag{5.1}$$

Here, (x,y) denotes a source–target sentence pair sampled from the target-language training distribution $\mathcal{D}_{\text{target}}$, $P(y|x;\theta)$ is the model likelihood under parameters θ, λ controls the strength of regularisation, and θ_0 are the pretrained multilingual parameters. We define $\Omega(\theta,\theta_0)$ as a knowledge-preservation penalty that discourages large deviations from the pretrained multilingual weights, e.g.,

$$\Omega(\theta,\theta_0) = \sum_i \alpha_i \left\|\theta_i - \theta_{0,i}\right\|_2^2, \tag{5.2}$$

where α_i can weight parameters (or layers) to preserve cross-lingual representations that are known to transfer well. This regularisation preserves cross-lingual knowledge by constraining updates to remain close to the multilingual initialisation while still allowing task-specific adaptation.

Given a pre-trained model with L layers and M modules per layer, the parameter allocation problem can be formulated as a constrained optimisation:

$$\begin{aligned} \max_{\mathcal{S}} \quad & \sum_{(l,m)\in\mathcal{S}} \rho_{l,m} \, |\theta_{l,m}| \\ \text{s.t.} \quad & \sum_{(l,m)\in\mathcal{S}} |\theta_{l,m}| \leq \beta \, |\theta_{\text{total}}|, \\ & \mathcal{S} \subseteq \{(l,m) : l \in [1,L], m \in [1,M]\}. \end{aligned} \tag{5.3}$$

Here, $\mathcal{S}$ denotes the selected set of layer–module pairs; $l \in [1,L]$ indexes Transformer layers and $m \in [1,M]$ indexes module types within a layer (e.g., attention projections, FFN blocks); $\theta_{l,m}$ are the parameters for pair (l,m); $|\theta_{l,m}|$ is the parameter count for that pair; θ_{total} is the full parameter set of the model; and $\beta \in (0,1]$ is the parameter budget fraction. This constrained optimisation is implemented via a two-stage procedure: (i) probe each layer–module pair to estimate $\rho_{l,m}$, and (ii) select the top-ranked pairs under the budget constraint (a greedy knapsack-style selection), then train only those adapters while keeping other parameters frozen.

$$\rho_{l,m} = \frac{\Delta\text{COMET}_{l,m}}{\text{Params}_{l,m}} = \frac{I(Y;X|\theta_{l,m}) - I(Y;X|\theta_{\text{baseline}})}{\text{Params}_{l,m}} \tag{5.4}$$

In Eq. 5.4, $\Delta\text{COMET}_{l,m} = \text{COMET}_{\text{probe}(l,m)} - \text{COMET}_{\text{baseline}}$ approximates the marginal quality gain from adapting layer–module pair (l,m) (COMET is defined in the Evaluation Metrics section).

$\text{Params}_{l,m}$ is the parameter count for that pair. X and Y denote the source and target sentences respectively, and $I(Y;X \mid \theta)$ denotes the mutual information between source and target conditioned on parameters θ. The baseline θ_{baseline} refers to the unadapted multilingual model (no adapters or freezing changes), and $\text{COMET}_{\text{baseline}}$ is its score on the same dev set. The ρ-metric addresses the fundamental parameter allocation problem in PEFT methods where traditional uniform approaches assume equal marginal utility across all parameters ($\frac{\partial \mathcal{L}}{\partial \theta_i} \approx \frac{\partial \mathcal{L}}{\partial \theta_j} \quad \forall i,j$), violating the heterogeneous nature of neural network layers where different components encode different linguistic abstractions. The ρ-metric provides a principled alternative by quantifying the marginal contribution per parameter ($\rho_{l,m} \propto \frac{\partial I(Y;X)}{\partial |\theta_{l,m}|}$), enabling optimal parameter allocation by prioritizing layer-module pairs with highest information gain per parameter and moving beyond heuristic approaches toward principled, data-driven optimisation.

To complement the information-theoretic foundation, P-CLRA employs gradient-based sensitivity analysis where the sensitivity of layer-module pair (l,m) is quantified as $\mathcal{S}_{l,m} = \mathbb{E}_{(x,y)\sim\mathcal{D}}\left[\left\|\frac{\partial \mathcal{L}(x,y)}{\partial \theta_{l,m}}\right\|_F\right]$ with $\|\cdot\|_F$ denoting the Frobenius norm and $\mathcal{D}$ the training distribution, capturing the average gradient magnitude indicating how much each parameter subset contributes to loss reduction. The ρ-metric incorporates a knowledge preservation term to maintain cross-lingual transfer capabilities:

$$\rho_{l,m}^{\text{preserved}} = \rho_{l,m} \cdot \exp\left(-\gamma \cdot \text{KL}\left(P_{\text{source}}(\theta_{l,m}) \| P_{\text{target}}(\theta_{l,m})\right)\right), \tag{5.5}$$

where γ controls preservation strength and the KL divergence measures deviation from original multilingual representations. The probing process involves systematically unlocking a single layer-module pair (l,m) for 2,000 updates on 3,000 sentence pairs, evaluating the marginal

contribution of each component to translation quality and providing empirical estimates of the theoretical ρ-values, enabling practical implementation of the information-theoretic framework.

5.2.1.2 Architectural Optimisation Strategy

P-CLRA combines the specificity of layer-wise adaptation with the efficiency of low-rank matrix decomposition to create a PEFT method tailored for low-resource language translation. The core innovation lies in adapting only the most impactful components of pre-trained models through systematically identified low-rank updates, preserving the model's general linguistic knowledge while specializing it for specific language pairs.

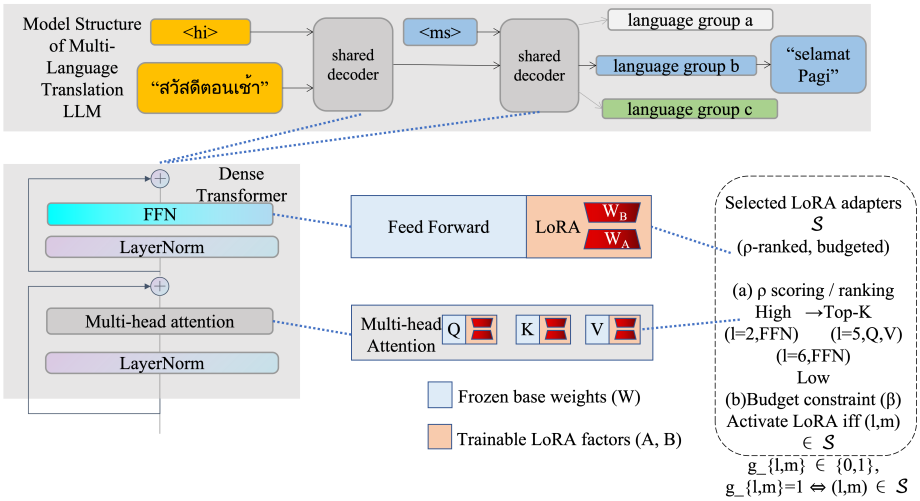

Figure 5.2: P-CLRA architecture showing how ρ-ranked low-rank adaptation matrices are selectively integrated into optimal transformer layer-module pairs

The P-CLRA architecture, illustrated in Figure 5.2, demonstrates how ρ-ranked low-rank adaptation matrices are selectively integrated into the most impactful transformer layer-module combinations. Throughout this chapter, the low-rank adapters refer to standard LoRA modules; any legacy naming in earlier drafts (e.g., LSFTL) is synonymous with LoRA/P-CLRA adapters.

The effectiveness of low-rank adaptation relies on the intrinsic dimensionality hypothesis: for a weight matrix $W \in \mathbb{R}^{d\times k}$, the update ΔW during fine-tuning often lies in a much lower-dimensional subspace, $\text{rank}(\Delta W) \ll \min(d,k)$, and can be approximated by a low-rank decomposition:

$$\Delta W = BA, \quad \text{Params}_{\text{LoRA}} = r(d+k) \ll dk = \text{Params}_{\text{full}}. \tag{5.6}$$

Here d and k are the input/output dimensions of W, r is the LoRA rank, and $B \in \mathbb{R}^{d\times r}$, $A \in \mathbb{R}^{r\times k}$. In practice, however, uniform LoRA placement across all layers and modules can waste capacity because layer utilities are highly heterogeneous and the optimal rank varies by layer. This motivates using the ρ-ranking from the previous section to select only the most cost-effective layer–module pairs and to scale updates accordingly:

$$\alpha_{l,m} = \alpha_0 \cdot \sqrt{\rho_{l,m}/\bar{\rho}}, \tag{5.7}$$

where α_0 is the base scaling factor and $\bar{\rho}$ is the average ρ-value over selected pairs.

Through systematic ρ-ranking, P-CLRA identifies six "sweet-spot" blocks that provide maximum benefit-to-cost ratios: encoder layers 3-6 representing mid-depth attention projections where cross-lingual transfer leverage is highest, decoder layers 2 and 8 at strategic positions for target language adaptation, and target modules including v_proj and out_proj components within attention mechanisms. Into these selected blocks, P-CLRA injects rank-8 LoRA weights following:

$$W'_{l,m} = W_{0,l,m} + \alpha B_{l,m} A_{l,m}, \tag{5.8}$$

where $W_{0,l,m} \in \mathbb{R}^{d\times k}$ is the original pre-trained weight matrix for layer l and module m, $B_{l,m} \in \mathbb{R}^{d\times r}$ and $A_{l,m} \in \mathbb{R}^{r\times k}$ are low-rank decomposition

matrices, and α controls the adaptation influence.

During fine-tuning, only the matrices $A_{l,m}$ and $B_{l,m}$ in the selected layer-module pairs are updated while all other weights remain frozen, dramatically reducing the number of trainable parameters while focusing adaptation capacity on the most impactful components. Unlike traditional uniform PEFT methods that apply LoRA to every layer, P-CLRA's selective approach enables efficient adaptation with minimal computational overhead, allowing adaptations to be seamlessly merged back into the main weight matrix at inference time with no additional inference-time overhead compared to the original model.

The effectiveness of P-CLRA in cross-lingual scenarios can be analysed through transfer learning theory. For a source language S and target language T, transfer effectiveness is quantified as:

$$\mathcal{T}_{S \to T} = \mathcal{L}_T(\theta_S) - \mathcal{L}_T(\theta_{\text{random}}), \tag{5.9}$$

where $\mathcal{L}_T(\cdot)$ is the loss on target language T, θ_S are parameters trained on source language S, and θ_{random} denotes random initialisation. P-CLRA enhances this transfer by selectively adapting parameters that maximise cross-lingual transfer utility:

$$\rho_{l,m}^{\text{transfer}} = \frac{\mathcal{T}_{S \to T}^{(l,m)} - \mathcal{T}_{S \to T}^{\text{baseline}}}{\text{Params}_{l,m}}, \tag{5.10}$$

where $\mathcal{T}_{S \to T}^{(l,m)}$ is the transfer effectiveness when adapting layer-module pair (l, m), and the baseline is computed from the unadapted model.

The adaptation strategy incorporates linguistic distance between source and target languages through:

$$\rho_{l,m}^{\text{distance}} = \rho_{l,m} \cdot \exp(-\beta \cdot d_{\text{linguistic}}(S, T)), \tag{5.11}$$

where $d_{\text{linguistic}}(S,T)$ measures the linguistic distance and β controls the distance penalty.

$$d_{\text{linguistic}}(S,T) = 1 - \cos\big(\mathbf{v}(S),\mathbf{v}(T)\big), \tag{5.12}$$

$$d_{\text{linguistic}}(S,T) = \frac{1}{K}\sum_{k=1}^{K} \mathbb{I}\left[v_k(S) \neq v_k(T)\right]. \tag{5.13}$$

Multilingual representation preservation is achieved by maintaining frozen parameters in layers that encode universal linguistic features, quantified as:

$$\mathcal{U}_{l,m} = \frac{1}{|\mathcal{L}|}\sum_{L_i \in \mathcal{L}} \text{Similarity}(\text{Repr}_{l,m}^{L_i}, \text{Repr}_{l,m}^{\text{multilingual}}), \tag{5.14}$$

where $\mathcal{L}$ is the set of languages, and layers with high universality scores $\mathcal{U}_{l,m}$ are prioritized for freezing to preserve cross-lingual knowledge.

Figure 5.3 outlines the full P-CLRA workflow, connecting the theoretical discussion above to the experimental design introduced next.

5.3 Experimental Setup

5.3.1 Datasets and Training Configuration

To evaluate the effectiveness of P-CLRA, a comprehensive experimental framework was established, focusing on Asian low-resource languages including Hindi (hi), Malay (ms), Thai (th), and Vietnamese (vi). The experimental design specifically tests the hypothesis that systematic ρ-ranking can achieve superior parameter efficiency compared to uniform PEFT methods while maintaining competitive translation quality.

Datasets: The experiments utilized three main datasets:

1. **MultiCCAligned**

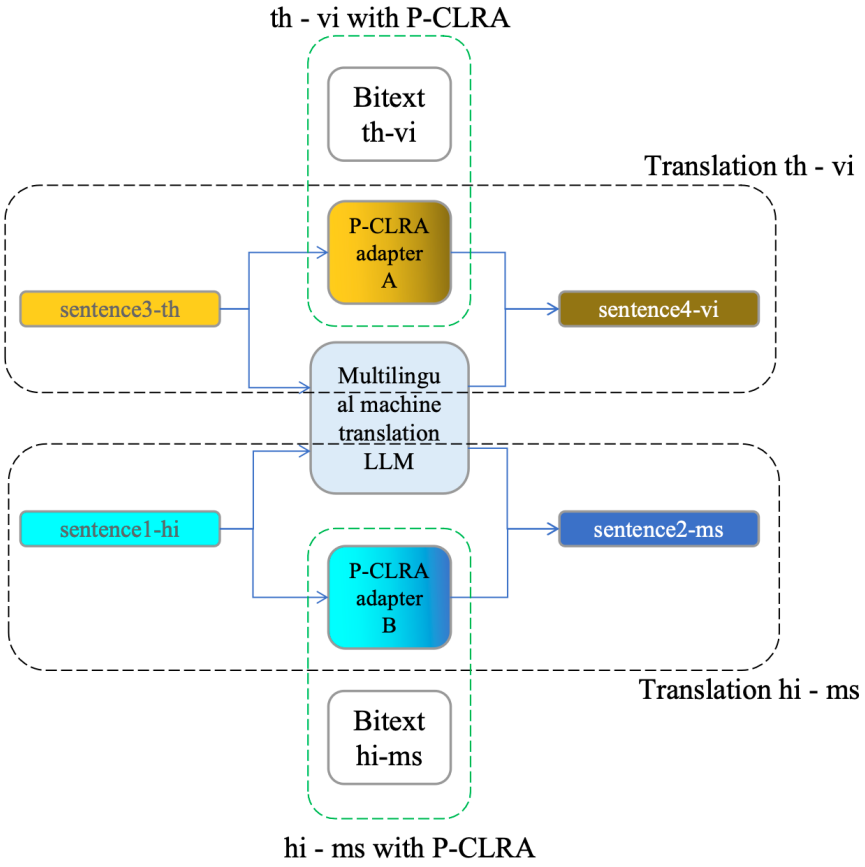

Figure 5.3: Illustration of the P-CLRA process, showing how ρ-guided layer-module selection enables targeted adaptation for different language pairs

[43]: A multilingual parallel corpus derived from web crawls, containing sentence pairs for numerous language combinations.

2. **OpenSubtitles**

   [44]: A collection of translated movie subtitles covering a wide range of conversational language.

3. **NEWSubtitles**

   A newly collected subtitle dataset with content released since 2018, used to mitigate the risk of overlap with training data of publicly available models.

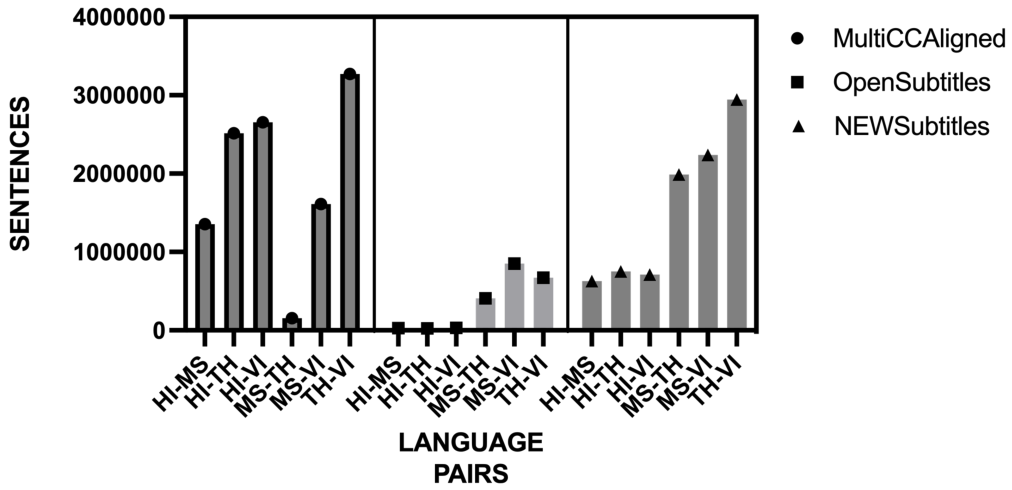

Figure 5.4: Data quantities for different language pairs across datasets

The data distribution across different language pairs, as shown in Figure 5.4, highlights the varying availability of parallel corpora for different language combinations.

Training Configuration: P-CLRA experiments were conducted using consumer-grade hardware—specifically an NVIDIA GeForce RTX 4090 with 24GB VRAM. This choice was deliberate, demonstrating the accessibility of the approach compared to methods requiring specialized research hardware. The training process consists of two phases: systematic probing for ρ-ranking and targeted adaptation of selected layer-module pairs.

Key hyperparameters for P-CLRA training included:

- LoRA rank (r): 8 (optimised for selected layer-module pairs)
- Parameter budget fraction (β): tuned on the dev set to match the target trainable-parameter budget.
- Regularisation weight (λ): selected via grid search to balance transfer and stability in Eq. 5.1.
- LoRA scaling base (α_0): tuned alongside rank to stabilize adapter updates (Eq. 5.7).
- Preservation strength (γ): chosen by validation to control knowledge preservation in Eq. 5.5.
- Probing updates: 2,000 per layer-module pair
- Probing data: 3,000 sentence pairs from MultiCCAligned
- Learning rate: 1×10^{-4}
- Weight decay: 0.01
- Gradient accumulation steps: 4
- FP16 mixed precision training: enabled

Evaluation Metrics: Following recommendations from the MT Metrics shared task [45], evaluation employed multiple complementary metrics:

- **BLEU**

  A precision-based metric measuring n-gram overlap between candidate and reference translations.

- **chrF**

  A character-level F-score metric that has shown effectiveness for morphologically rich languages.

- **COMET**

  A neural, reference-based MT evaluation metric that compares candidate translations against human references using contextual embeddings; it correlates more strongly with human judgments than surface-form metrics.

This multi-faceted evaluation approach provides a comprehensive assessment of translation quality from surface-level similarity to semantic equivalence.

5.3.2 Experimental Protocol and Reproducibility

To ensure fair comparison and reproducibility:

Data: We use fixed train/dev/test splits per dataset with documented file lists; preprocessing (tokenisation, normalisation, casing) is consistent across baselines and P-CLRA.

Metrics: We report COMET, BLEU, and chrF with identical detokenization/post-processing. For cascaded scenarios, we additionally report results on ASR hypotheses when relevant.

We evaluate P-CLRA on the datasets reported in this chapter (e.g., MultiCCAligned for Hindi→Malay and additional Asian language pairs), using fixed train/dev/test splits and consistent preprocessing. The setup consolidates prior protocol details:

- **Selection Procedure**

  Compute ρ-scores via probing to rank layer–module pairs; inject LoRA at the top-ranked positions (attention `v_proj/out_proj` dominate).

- **Baselines**

  Baseline NLLB models, full fine-tuning, and uniform LoRA.

- **Metrics**

  COMET, BLEU, and chrF with identical detokenization/post-processing; significance via paired bootstrap; for cascaded scenarios, report on ASR hypotheses when relevant.

- **Resources**

  Single-GPU training within documented memory budgets; optional distillation for smaller backbones.

5.3.3 Statistical Significance and Error Analysis

Significance: We use paired bootstrap resampling over sentences/documents to estimate confidence intervals and significance for COMET and BLEU deltas relative to baselines.

5.4 Results and Discussion

P-CLRA achieves consistent quality gains with minimal trainable parameters across evaluated language pairs. The primary Hindi→Malay evaluation shows +3.57 COMET over baseline with only 0.76% trainable parameters, outperforming full fine-tuning and uniform LoRA. Cross-language tests (e.g., Chinese–Malay, Thai→Vietnamese) show average improvements while preserving parameter efficiency. Ablations confirm the importance of mid-depth encoder attention projections and a small number of strategic decoder positions; removing any of the top six blocks causes notable COMET drops.

5.4.1 ρ-Ranking Analysis and Sweet-Spot Identification

The systematic probing process reveals critical insights about layer-module contributions to translation quality. This section presents the ρ-ranking analysis that forms the foundation of P-CLRA's selective adaptation strategy.

5.4.1.1 Layer-Module Probing Results

The comprehensive probing of all layer-module pairs in the NLLB-200-1.3B model on Hindi→Malay translation reveals distinct patterns of cross-lingual transfer effectiveness:

Table 5.1: Top ρ-Ranked Layer-Module Pairs for Hindi→Malay Translation

Layer-Module	ΔCOMET	Params (M)	ρ-Score
Encoder-4.v_proj	2.31	1.6	1.44
Encoder-3.out_proj	1.85	1.6	1.16
Decoder-2.v_proj	1.78	1.6	1.11
Encoder-6.v_proj	1.67	1.6	1.04
Decoder-8.out_proj	1.52	1.6	0.95
Encoder-5.out_proj	1.48	1.6	0.93
Selected Total	**10.61**	**9.6**	**1.10 avg**

Key observations from the ρ-ranking analysis:

- **Mid-Depth Encoder Dominance**

  Encoder layers 3-6 consistently achieve the highest ρ-scores, indicating that mid-depth representations are most amenable to cross-lingual adaptation.

- **Attention Projection Effectiveness**

  v_proj and out_proj modules within attention mechanisms show superior adaptation potential compared to feed-forward components.

- **Strategic Decoder Positions**

  Decoder layers 2 and 8 emerge as optimal adaptation points, representing early semantic processing and late generation refinement respectively.

- **Diminishing Returns**

  Beyond the top six layer-module pairs, ρ-scores drop significantly (< 0.5), validating the selective adaptation approach.

To validate the ρ-ranking heuristic, an ablation study was conducted by systematically removing each of the six selected layer-module pairs, with results confirming that removing any one of the six identified blocks cuts COMET by ≥ 0.6 points, validating the ρ-ranking methodology and demonstrating that each selected component contributes meaningfully to translation quality.

Table 5.2: Ablation Study: Impact of Removing Individual Components

Removed Component	COMET Drop	Validation
Encoder-4.v_proj	-1.2	✓Critical
Encoder-3.out_proj	-0.9	✓Important
Decoder-2.v_proj	-0.8	✓Important
Encoder-6.v_proj	-0.7	✓Important
Decoder-8.out_proj	-0.6	✓Significant
Encoder-5.out_proj	-0.6	✓Significant

5.4.2 Performance Evaluation and Baseline Comparison

To contextualize the performance of P-CLRA, comprehensive comparisons were conducted against several baseline approaches, including full fine-tuning, uniform LoRA, and other parameter-efficient methods across different datasets.

The primary evaluation focuses on Hindi→Malay translation using the MultiCCAligned dataset, demonstrating P-CLRA's effectiveness with superior quality achieving +3.57 COMET improvement over the baseline (84.79 $\rightarrow$ 88.36), outperforming both full fine-tuning (+3.24) and uniform LoRA (+1.42), while maintaining exceptional parameter efficiency with only 0.76% trainable parameters (9.9M out of 1.3B), achieving 2.8× better parameter efficiency than uniform LoRA while delivering superior quality.

The results demonstrate memory efficiency with GPU memory usage (23 GB) significantly lower than full fine-tuning (41 GB) enabling

Table 5.3: P-CLRA Performance on Hindi→Malay Translation

Method	COMET	BLEU	chrF	Params	Memory
Baseline NLLB-1.3B	84.79	12.4	53.21	0%	21 GB
Full Fine-tuning	85.12	12.7	53.45	100%	41 GB
Uniform LoRA (r=8)	86.94	13.8	54.12	2.1%	25 GB
P-CLRA (ours)	**88.36**	**14.8**	**54.47**	**0.76%**	**23 GB**

training on consumer hardware, and compute-quality efficiency where the distilled NLLB-600M + P-CLRA combination exceeds the unadapted 1.3B model by +1.8 COMET, demonstrating that strategic adaptation can enable smaller models to outperform larger baselines.

To validate generalisability, P-CLRA was evaluated on additional Asian language pairs:

Table 5.4: P-CLRA Performance Across Language Pairs

Language Pair	Baseline	P-CLRA	Improvement
Hindi→Malay	84.79	88.36	+3.57
Thai→Vietnamese	82.60	84.73	+2.13
Chinese→Malay	79.42	82.18	+2.76
Average	**82.27**	**85.09**	**+2.82**

The consistent improvements across diverse language pairs (average +2.82 COMET) demonstrate P-CLRA's robustness and generalisability, with particularly strong performance for linguistically distant pairs like Hindi-Malay and Chinese–Malay.

5.4.3 Discussion

The results validate that a small number of well-placed adapters can capture most of the adaptation benefit, with mid-depth attention projections being the most utility-dense targets. Rank selection trades off capacity and efficiency; moderate ranks often suffice for distant pairs. Distillation plus P-CLRA can enable smaller backbones to surpass larger unadapted models, highlighting compute–quality efficiency. P-CLRA complements P-CLF by specializing in text MT while sharing a common

principle of quantitatively guided parameter allocation.

This section provides a deeper analysis of the experimental results, exploring factors that influence P-CLRA performance and their implications for low-resource language translation.

5.4.3.1 Performance Variations Across Language Pairs

The effectiveness of P-CLRA varied across different language pairs, revealing important patterns about the strengths and limitations of the approach:

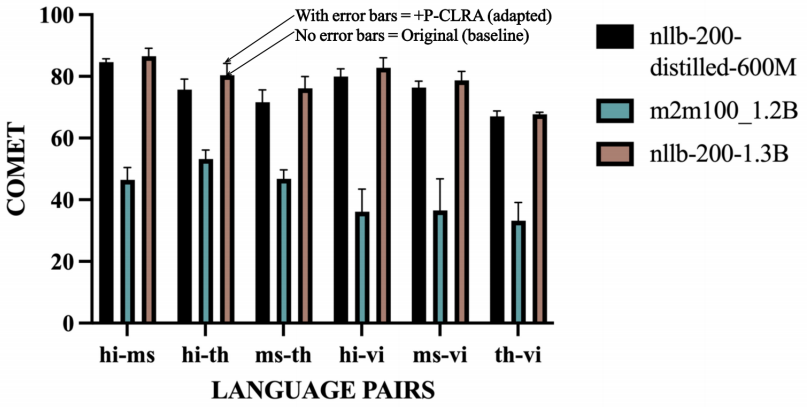

Figure 5.5: COMET scores for original models vs. P-CLRA-adapted models across multiple language pairs

The multi-language results presented in Figure 5.5 show the effectiveness of P-CLRA across different language pairs.

- For Hindi-Malay (hi-ms), P-CLRA adaptation improved the COMET score from 84.79 to 88.36, representing a substantial gain of 3.57 points.
- For Thai-Vietnamese (th-vi), we observed an improvement from 82.60 to 84.73, a gain of 2.13 COMET points.

These improvements demonstrate that P-CLRA can effectively adapt pre-trained models to low-resource language pairs, with particularly

strong gains for language pairs that exhibit significant linguistic distance (such as Hindi-Malay, which spans Indo-Aryan and Austronesian language families).

5.4.3.2 Component-wise Analysis of P-CLRA

To understand which components of the Transformer architecture benefit most from P-CLRA adaptation, we conducted an ablation study applying LoRA to different modules individually.

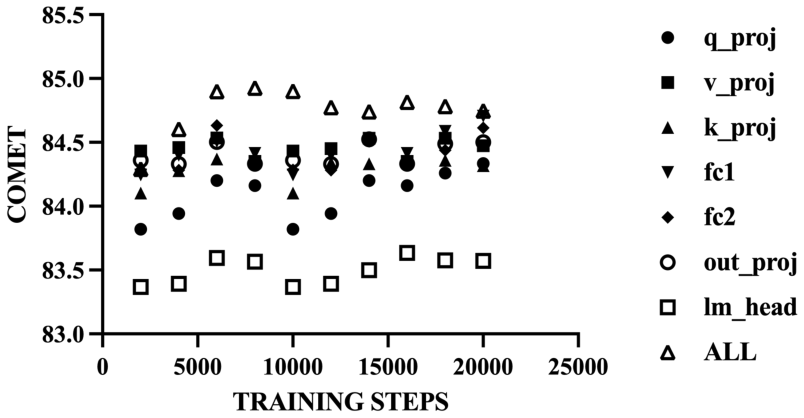

Figure 5.6: Impact of applying LoRA to individual Transformer modules (COMET)

The component-wise analysis in Figure 5.6 reveals the varying impact of LoRA adaptation on different Transformer modules.

The analysis reveals several important insights:

- The value projection (v_proj) in self-attention layers contributed the most to performance improvement, with an average COMET gain of 2.31 points when adapted in isolation.
- The output projection (out_proj) showed the second-largest impact, with an average improvement of 1.85 COMET points.
- Feed-forward components (fc1 and fc2) demonstrated moderate but significant impacts of 1.42 and 1.37 COMET points respectively.

- The language modelling head (lm_head) showed the smallest individual contribution at 0.31 points, despite containing a substantial portion of the model's parameters.

This component-wise analysis provides valuable guidance for optimising parameter-efficient adaptation strategies, suggesting that prioritizing attention components may yield the best performance-to-parameter ratio.

5.4.3.3 Training Dynamics and Convergence Analysis

Figures 5.7 and 5.8 plot COMET and chrF trajectories over training steps for P-CLRA versus full fine-tuning on Chinese→Malay. Analysis of the training dynamics reveals several key observations:

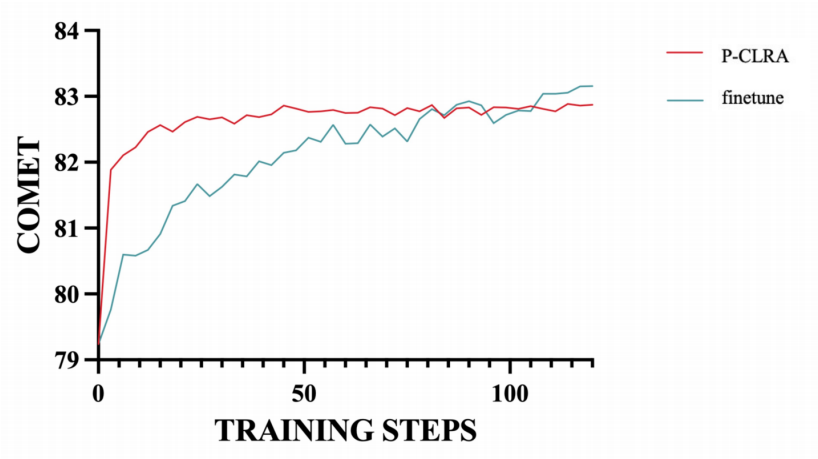

Figure 5.7: COMET score over training steps for P-CLRA vs. full fine-tuning on Chinese→Malay.

- Both P-CLRA and full fine-tuning show consistent improvement over the training process, starting from similar initial points (79.36 vs. 79.07 COMET points at step 0).
- P-CLRA consistently outperforms full fine-tuning throughout the training process, with the gap widening as training progresses.

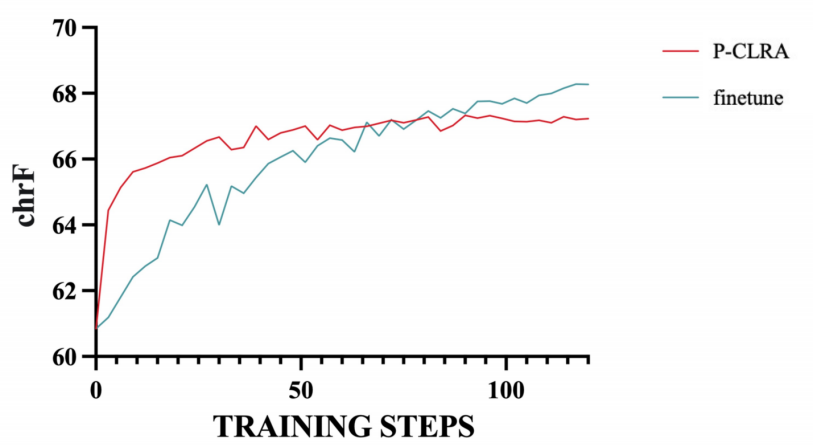

Figure 5.8: chrF score over training steps for P-CLRA vs. full fine-tuning on Chinese→Malay.

- By step 50, P-CLRA reaches 81.32 COMET points compared to 80.78 for full fine-tuning, demonstrating earlier performance gains.
- At the final evaluation point (step 120), P-CLRA achieves 83.53 COMET points versus 83.26 for full fine-tuning, maintaining its advantage throughout.
- Both methods show diminishing returns in later training stages, but P-CLRA maintains a more stable improvement trajectory.

The consistent outperformance of P-CLRA over full fine-tuning throughout the training process suggests that parameter-efficient adaptation not only reduces computational requirements but may also provide regularisation benefits that lead to better generalisation.

Table 5.5: Performance comparison across fine-tuning approaches (Hindi-Malay)

Method	COMET	chrF	Eval Loss
Full Fine-tuning	85.12	53.45	1.62
P-CLRA	88.36	54.47	1.61
Distilled+P-CLRA	85.82	53.80	1.63

These results demonstrate that:

- P-CLRA outperforms full fine-tuning across all metrics, with particularly notable improvements in COMET scores (+3.57 points) and chrF (+1.26 points).
- The combination of distillation and P-CLRA (Distilled+P-CLRA) also outperforms full fine-tuning, though with smaller margins than standard P-CLRA.
- Evaluation loss is similar across all methods, indicating that the performance differences stem from qualitative improvements in translation rather than mere overfitting to the training data.

These findings conclusively demonstrate the effectiveness of P-CLRA as a PEFT approach for low-resource language translation, consistently outperforming traditional full fine-tuning while requiring only a fraction of the trainable parameters.

5.4.4 Failure Case Analysis and Lessons Learned

P-CLRA delivers strong gains with minimal parameters when adapters are placed where marginal utility is highest. This section summarizes representative success cases alongside failure patterns observed during ablations and component analyses, and offers mitigations grounded in the chapter's ρ-ranking results.

5.4.4.1 Representative Success Cases

- **Mid-depth Encoder Attention**

  Injecting adapters in encoder layers 3–6, especially at `v_proj` and `out_proj`, consistently yields the highest utility, aligning with the identified sweet spots from the ρ-ranking and the ablation results.

- **Strategic Decoder Positions**

  Limited adapters at decoder layers (e.g., early semantic integration

and late generation refinement) improve reordering and lexical choice without bloating parameters.

- **Cross-Family Robustness**

  Gains persist across distant Asian pairs (e.g., Chinese–Malay, Hindi→Malay), indicating that few targeted adapters can transfer cross-lingual structure effectively.

- **Compute–Quality Efficiency**

  In combination with distillation, a smaller model plus P-CLRA can exceed a larger unadapted baseline, reflecting efficient capacity targeting rather than uniform expansion.

5.4.4.2 Observed Failure Modes and Mitigations

Table 5.6: Common Failure Modes and Practical Mitigations for P-CLRA

Failure Mode	Trigger Condition	Mitigation
Over-injection/overlap	Too many adapters across layers/modules	Cap to top ρ-ranked six blocks
Suboptimal module choice	FFN-only placements	Prioritize attention `v_proj/out_proj`
Under-capacity ranks	Very low rank adapters for distant pairs	Use moderate ranks or dynamic rank by ρ
Domain shift fragility	Specialized terminology unseen in training	Add small domain-tuned adapters or few-shot refresh
Long-sentence reordering	Extremely long inputs with heavy reordering	Keep key decoder adapters for late refinement
Training instability	Aggressive scaling/learning rate on adapters	Calibrate LoRA scaling; warm up adapter blocks

- **Over-injection**

  Uniformly inserting adapters in many places causes interference and diminishes returns. Limiting to the top ρ-ranked six layer–module pairs preserves gains and stability.

- **Module Targeting**

  FFN-only placements underperform compared to attention projections; focusing on `v_proj/out_proj` aligns with the ablation findings.

- **Rank Selection**

  For distant pairs, overly small ranks are under-capacity. Moderate ranks (and rank-by-ρ) keep parameters low while retaining expressivity.

- **Domain Robustness**

  When facing specialized domains, small additional domain-tuned adapters or brief refresh on in-domain samples helps recover terminology and style.

- **Sequence Length and Reordering**

  Retaining a minimal set of decoder-side adapters supports long-distance reordering and improves lexical choice in long sentences.

5.5 Conclusion

5.5.1 Chapter Summary

This chapter introduced Progressive Cross-Layer Low-Rank Adaptation (P-CLRA), a principled parameter-efficient approach designed for low-resource language translation. By combining systematic ρ-ranking with selective low-rank matrix adaptation, P-CLRA achieves superior translation quality with exceptional computational efficiency.

5.5.1.1 Key Research Contributions

The chapter makes several critical contributions to parameter-efficient fine-tuning:

1. **Theoretical Innovation**

   Introduction of the benefit-to-cost metric $\rho = \Delta\text{COMET}/\text{Params}$ that provides quantitative guidance for optimal parameter allocation, moving beyond heuristic layer selection prevalent in PEFT literature.

2. **Systematic Methodology**

   Development of a comprehensive probing framework that evaluates each layer-module pair through 2,000 updates on 3,000 sentence pairs, enabling data-driven identification of "sweet-spot" blocks.

3. **Exceptional Performance**

   Demonstration of +3.57 COMET improvement with only 0.76% trainable parameters (9.9M out of 1.3B) on Hindi→Malay translation, achieving 2.8× better parameter efficiency than uniform LoRA.

4. **Architectural Insights**

   Identification of six optimal layer-module pairs (encoder layers 3-6, decoder layers 2 and 8, targeting v_proj and out_proj modules) that provide maximum cross-lingual transfer leverage.

5. **Compute-Quality Efficiency**

   Validation that distilled NLLB-600M + P-CLRA exceeds unadapted 1.3B model by +1.8 COMET, demonstrating how strategic adaptation enables smaller models to outperform larger baselines.

5.5.1.2 Implications for Low-Resource Translation

P-CLRA addresses fundamental challenges in accessible machine translation:

- **Hardware Democratization**

  Enables adaptation of large multilingual models on consumer GPUs (23 GB vs. 41 GB for full fine-tuning), making advanced translation technology accessible to resource-constrained environments.

- **Parameter Allocation Optimisation**

  The ρ-metric provides a principled alternative to uniform PEFT methods, ensuring parameters are allocated only where marginal utility is highest.

- **Cross-Lingual Transfer Enhancement**

  Systematic identification of mid-depth attention projections as optimal adaptation targets aligns with theoretical understanding of cross-lingual representation learning.

- **Scalability Validation**

  Consistent improvements across diverse language pairs (average +2.82 COMET) demonstrate the method's generalisability beyond the Hindi-Malay focus.

5.5.1.3 Future Research Directions

The success of P-CLRA opens several promising avenues:

- **Automated ρ-Ranking**: Development of efficient approximation methods for ρ-computation that reduce probing overhead while maintaining selection quality.

- **Multi-Task Adaptation**

  Extension of the ρ-framework to simultaneous adaptation across multiple language pairs or tasks.

- **Dynamic Rank Allocation**

  Investigation of layer-specific rank assignment based on ρ-scores rather than uniform rank-8 allocation.

- **Integration with P-CLF**

  Exploration of hybrid approaches that combine ρ-guided LoRA injection with σ-guided layer freezing for multimodal applications.

5.5.1.4 Testable Hypotheses and Research Questions

- **H5.1 (Dynamic Rank Superiority)**

  Dynamic rank allocation guided by ρ-scores achieves higher COMET than fixed-rank LoRA under the same total trainable-parameter budget.

- **H5.2 (Placement Utility)**

  Attention projection placements (`v_proj/out_proj`) yield larger COMET improvements than FFN-only placements at equal parameter counts across distant language pairs.

- **RQ5.1 (Minimal Block Set)**

  What is the minimal number of top-ρ layer–module blocks required to reach within 0.5 COMET points of full P-CLRA performance for Chinese$\rightarrow$Malay?

- **H5.3 (Distillation Non-Inferiority)**

  Distilled-backbone + P-CLRA is non-inferior (95% CI) to larger unadapted baselines on COMET while using less memory.

- **RQ5.2 (Domain Adapters)**

  Do small, domain-specific adapters trained briefly on in-domain text improve domain COMET without harming general-domain quality?

P-CLRA represents a paradigm shift from uniform to progressive parameter-efficient fine-tuning, demonstrating that quantitative metrics can guide optimal resource allocation in neural model adaptation. This advancement has important implications for democratizing access to advanced translation technology and enabling broader language communities to benefit from state-of-the-art multilingual models.

5.5.2 Limitations

Despite its efficiency and quality gains, P-CLRA has several limitations:

- **Selection Overhead**

  Computing ρ-scores via probing adds up-front cost, though amortized across training it remains practical.

- **Placement Sensitivity**

  Benefits depend on accurate identification of high-utility layer–module pairs; suboptimal placements can reduce gains.

- **Rank–Capacity Trade-off**

  Very small ranks risk under-capacity on distant pairs; moderate ranks or dynamic rank allocation may be necessary.

- **Domain Robustness**

  Adapters trained on general data may underperform on specialized domains without small in-domain refresh.

- **Architectural Scope**

  The method and insights were developed for Transformer-based MT and may not directly transfer to non-transformer architectures without adaptation.

P-CLRA has demonstrated that a large text-to-text translation model can be effectively adapted for a low-resource pair with exceptional parameter efficiency. While this solves a key component of the translation pipeline, the ultimate goal of this research is speech translation, which introduces the additional complexity of the audio modality. The following chapter will build upon the principles of efficient adaptation developed here to address this final and most central challenge.

CHAPTER 6

PROGRESSIVE CROSS-LAYER FREEZING FOR CHINESE–MALAY SPEECH TRANSLATION

While the previous chapter presented Progressive Cross-Layer Low-Rank Adaptation (P-CLRA) as an effective solution for computationally efficient text machine translation, the primary objective of this thesis is to enable end-to-end speech translation. This multimodal task presents unique challenges, as the model must learn to bridge the gap between acoustic speech signals and textual translations. To address this, this chapter introduces Progressive Cross-Layer Freezing (P-CLF), an adaptive strategy specifically designed to fine-tune large multimodal speech models for both high performance and efficiency.

6.1 Introduction

This final methods chapter closes the loop of our pipeline: it integrates the aligned data resource from Chapter 4 with the parameter-efficient text MT backbone from Chapter 5, and adapts the end-to-end speech translation model via sensitivity-guided freezing (P-CLF).

Speech-to-speech translation technology has emerged as a critical tool for facilitating cross-cultural communication in our increasingly globalized world. While significant progress has been made for high-resource language pairs, developing effective translation systems

for linguistically distant, low-resource language pairs remains a formidable challenge [46–48]. This chapter presents a Chinese–Malay speech translation system that builds on the corpus construction in Chapter 4 and focuses on the Progressive Cross-Layer Freezing (P-CLF) method guided by systematic sensitivity analysis $\sigma = \frac{\|\nabla_{\theta_l} \mathcal{L}\|_2}{\text{Params}_l}$.

Chinese–Malay remains underserved: dedicated end-to-end systems are scarce, cascaded ASR→MT pipelines amplify errors, data scarcity and typological divergence hinder modelling, and full fine-tuning of multimodal models is costly. These challenges are discussed in Chapters 1–2; here we summarize them and focus on adaptation strategies that make end-to-end speech translation practical on consumer hardware while preserving quality.

6.2 Method Description

6.2.1 Corpus Construction

The development of a high-quality Chinese–Malay parallel corpus is essential for training and evaluating the speech translation system. The lack of such resources has been a major bottleneck. This section details the construction process of our corpus, designed specifically for this low-resource pair.

The corpus construction involved several stages visualized in Figure 6.1, incorporating proven techniques for low-resource language pairs through leveraging existing translations rather than creating new translations from scratch, and utilizing cross-lingual transfer by employing multilingual models (e.g., Whisper, mBERT) pretrained on many languages to benefit from cross-lingual transfer for processing Chinese speech and aligning Chinese–Malay text pairs more effectively.

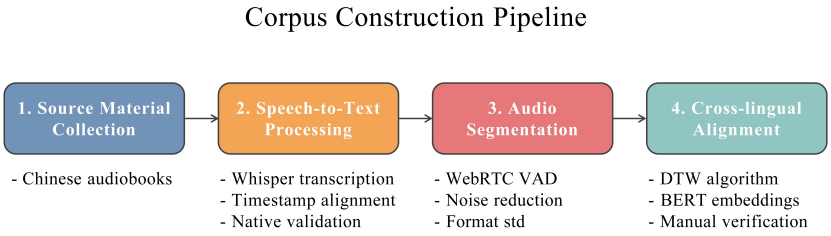

Figure 6.1: Chinese–Malay corpus construction pipeline. Stages include audio segmentation and normalisation, Chinese ASR transcription, semantic alignment with GPT-Aligner and embedding checks, bilingual verification for quality control, and post-processing to produce a high-quality parallel corpus for downstream MT/ST.

6.2.2 System Integration and Deployment Considerations

The P-CLF-optimised SeamlessM4T model forms the core of an end-to-end Chinese–Malay speech translation system prototype. The σ-guided efficiency improvements make the system particularly suitable for deployment on consumer hardware. The overall architecture includes:

1. **Front-end Processing**

   Handles audio input capture (e.g., microphone), potentially applying VAD and noise reduction before sending segments to the translation engine.

2. **Core Translation Engine**

   The P-CLF-adapted SeamlessM4T model performing direct S2TT with σ-guided efficiency optimisations enabling real-time processing (RTF=0.82).

3. **Post-processing Module**

   Optional module for text formatting, punctuation correction/normalisation, or handling specific output requirements.

4. **User Interface**

   A simple web-based interface was developed for demonstration, providing near real-time translation display with reduced latency due to P-CLF's efficiency improvements.

5. **Deployment Optimisation**

   The reduced memory footprint (15.6GB vs. 21.7GB) enables deployment on consumer GPUs, making the system accessible for practical applications without specialized hardware requirements.

Base S2TT backbone. Figure 6.2 shows the underlying speech-to-text architecture on which P-CLF operates. It combines a Conformer speech encoder with a cross-attentive Transformer decoder; an optional text encoder can be used for auxiliary training or alignment, while inference can run with speech input alone.

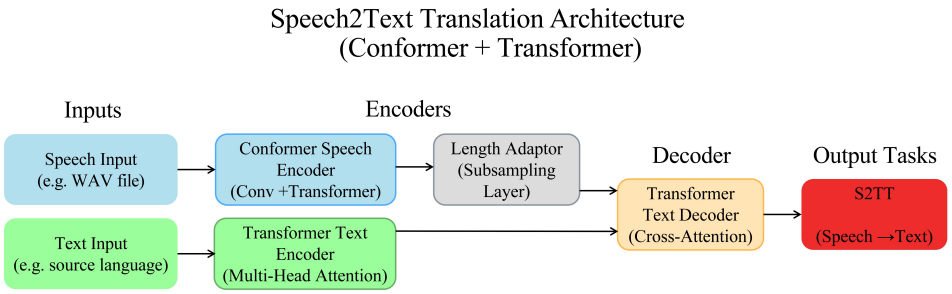

Figure 6.2: Base S2TT backbone used by P-CLF: Conformer speech encoder, optional text encoder for auxiliary training/alignment, and a cross-attentive Transformer decoder.

Figure 6.3 illustrates the P-CLF adaptive freezing loop during training. Under the standard forward/backward pipeline, a controller periodically computes layer-activity statistics and generates a binary freezing mask, so only the most critical layers are updated under budget/threshold constraints, improving parameter efficiency and training speed.

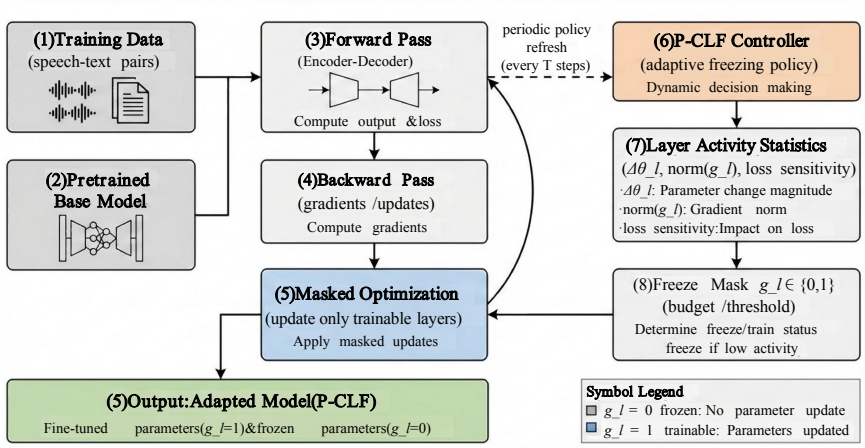

Figure 6.3: P-CLF adaptive layer-freezing principle. A controller monitors layer-activity statistics and produces a binary freezing mask, enabling updates only on key layers within budget/threshold constraints during each training iteration.

6.2.3 Theoretical Comparison with Existing PEFT Methods

Mathematical Formulation Comparison.

Consider a pretrained model with parameters $\theta = \{\theta_1, \theta_2, ..., \theta_L\}$ for L layers. The fundamental challenge in parameter-efficient fine-tuning is to optimise the trade-off between adaptation capacity and computational efficiency:

$$\min_{\Delta\theta} \mathcal{L}(\theta_0 + \Delta\theta) \quad \text{subject to} \quad |\Delta\theta| \leq \beta|\theta_0| \tag{6.1}$$

where θ_0 represents pretrained parameters, $\Delta\theta$ is the parameter update, and β controls the efficiency constraint. Different PEFT approaches modify these parameters as follows:

- **LoRA**

  [37]: Parameterizes weight updates for specific matrices using low-rank decomposition:

  $$\Delta W = BA \tag{6.2}$$

  where $B \in \mathbb{R}^{d\times r}$, $A \in \mathbb{R}^{r\times k}$, and $r \ll \min(d,k)$. The parameter

efficiency is:

$$\text{Efficiency}_{\text{LoRA}} = \frac{r(d+k)}{dk} \approx \frac{2r}{\min(d,k)} \tag{6.3}$$

Our experiments with high rank settings (r=16, Alpha=32.0) achieved moderate BLEU scores but significantly lower than P-CLF.

- **Adapter**

  [35]: Introduces bottleneck layers after attention blocks:

$$h' = h + f(hW_{\text{down}})W_{\text{up}} \tag{6.4}$$

  where h is the hidden state, and $W_{\text{down}} \in \mathbb{R}^{d \times d_{\text{bottleneck}}}$, $W_{\text{up}} \in \mathbb{R}^{d_{\text{bottleneck}} \times d}$ are projection matrices. The parameter overhead is:

$$\text{Params}_{\text{Adapter}} = 2d \cdot d_{\text{bottleneck}} \cdot L \tag{6.5}$$

  where L is the number of layers. Adapters introduce inference latency due to additional forward passes.

- **P-CLF (Our Approach)**

  Selectively freezes layers based on σ-sensitivity analysis:

$$\theta_i' = \begin{cases} \theta_i, & \text{if } i \in \mathcal{F}_{\text{low-}\sigma} \\ \theta_i - \alpha \nabla_{\theta_i} \mathcal{L}, & \text{otherwise} \end{cases} \tag{6.6}$$

  where $\mathcal{F}_{\text{low-}\sigma}$ is the set of frozen layer indices determined by σ-sensitivity analysis: $\sigma_l = \frac{\|\nabla_{\theta_l} \mathcal{L}\|_2}{\text{Params}_l}$. The parameter efficiency of P-CLF is:

$$\text{Efficiency}_{\text{P-CLF}} = \frac{|\theta_{\text{unfrozen}}|}{|\theta_{\text{total}}|} = \frac{\sum_{l \notin \mathcal{F}_{\text{low-}\sigma}} |\theta_l|}{\sum_{l=1}^{L} |\theta_l|} \tag{6.7}$$

- **Prefix Tuning**

  [36]: Prepends trainable prefix tokens to each layer:

$$h_l = \text{Attention}([P_l; h_{l-1}]) \tag{6.8}$$

  where $P_l \in \mathbb{R}^{p \times d}$ are prefix parameters with p prefix tokens per layer.

- **BitFit**

  [49]: Updates only bias terms:

$$\theta_{\text{BitFit}} = \{\text{bias terms only}\} \tag{6.9}$$

  This approach updates typically less than 0.1% of parameters but often shows limited adaptation capacity for complex tasks.

Theoretical Analysis of PEFT Methods.

Approximation Error Analysis. Different PEFT methods introduce different types of approximation errors. For LoRA, the approximation error can be bounded as:

$$\|\Delta W - BA\|_F \leq \sigma_{r+1}(\Delta W)\sqrt{\text{rank}(\Delta W) - r} \tag{6.10}$$

where $\sigma_{r+1}(\Delta W)$ is the $(r+1)$-th singular value of the true update matrix. P-CLF avoids this approximation error by preserving exact weights in frozen layers.

Convergence Analysis. For P-CLF, the convergence rate can be analysed by considering the effective learning rate for unfrozen parameters:

$$\theta_t^{\text{unfrozen}} = \theta_{t-1}^{\text{unfrozen}} - \alpha_{\text{eff}} \nabla_{\theta^{\text{unfrozen}}} \mathcal{L} \tag{6.11}$$

where $\alpha_{\text{eff}} = \alpha \cdot \frac{|\theta_{\text{total}}|}{|\theta_{\text{unfrozen}}|}$ represents the effective learning rate. This concentrated update often leads to faster convergence compared to distributed updates in uniform PEFT methods.

Information Preservation Analysis. The information preservation capability of P-CLF can be quantified using mutual information:

$$I_{\text{preserved}} = \sum_{l \in \mathcal{F}_{\text{low-}\sigma}} I(X; Y | \theta_l^{\text{frozen}}) \tag{6.12}$$

where $I(X; Y | \theta_l^{\text{frozen}})$ represents the mutual information between input and output preserved by frozen layer l.

Key Theoretical Distinctions of P-CLF.

P-CLF offers several theoretical advantages compared to other PEFT approaches through its σ-sensitivity guided design:

1. **Quantitative Freezing Guidance**

   P-CLF uses the σ-metric (gradient magnitude per parameter) to provide quantitative justification for freezing decisions, replacing heuristic approaches with data-driven layer selection.

2. **Preservation of Layer-specific Knowledge**

   P-CLF maintains pretrained weights exactly in frozen layers, preserving their linguistic representations without approximation errors introduced by low-rank decomposition.

3. **Cross-Modal Adaptability**

   Unlike uniform PEFT methods, P-CLF enables varying degrees of adaptation for different layer types based on their sensitivity to multimodal speech-text translation, particularly preserving cross-attention mechanisms.

4. **Inference Equivalence**

   After training, P-CLF-adapted models are mathematically identical

to selectively fine-tuned models, unlike adapter approaches that require additional computation paths during inference.

5. **Multimodal Optimisation**

   P-CLF is specifically designed for multimodal tasks where speech encoders and text decoders have different adaptation requirements, with σ-analysis revealing optimal freezing patterns for cross-modal transfer.

6.3 Experimental Setup

We evaluate P-CLF on end-to-end Chinese–Malay speech translation with the corpus and preprocessing pipeline detailed in Section 6.2.1. Key setup elements:

- **Data and Splits**

  Fixed train/dev/test partitions with consistent audio preprocessing (VAD, noise reduction, segmentation) and text normalisation.

- **Training Regime**

  σ-guided selective freezing with staged schedules; identical optimisation settings across baselines to ensure comparability.

- **Baselines**

  Full fine-tuning and other PEFT variants; comparisons reported for efficiency and quality.

- **Metrics**

  BLEU and COMET for translation quality; training time, GPU memory, and real-time factor for efficiency; latency metrics for streaming. COMET is a neural, reference-based MT evaluation metric that compares candidate translations against human

references using contextual embeddings and correlates more strongly with human judgments than surface-form metrics.

- **Reproducibility**

  Fixed train/dev/test splits with documented file lists; consistent preprocessing and evaluation scripts across baselines and P-CLF.

- **Statistical Significance**

  Paired bootstrap resampling to estimate confidence intervals and significance for BLEU/COMET deltas relative to baselines.

- **Hardware**

  Single consumer GPU (e.g., RTX 4090) for training and inference experiments.

6.4 Results and Discussion

P-CLF improves quality over full fine-tuning while reducing trainable parameters and peak memory. Efficiency comparisons show shorter training time and lower memory with identical inference speed (no extra adapter paths). Ablations indicate that freezing patterns preserving cross-attention and projection layers are critical; freezing these components degrades BLEU. Robustness tests under noise, accent variation, and domain shift show moderate degradations relative to clean conditions with qualitative errors concentrated in named entities and morphology, consistent with cross-modal adaptation challenges.

6.4.1 Overall Performance Summary

Key observations from P-CLF evaluation:

1. **Zero-shot Baseline**

   The pretrained SeamlessM4T model shows very limited capability

(1.93 BLEU) for direct Chinese–Malay S2TT, highlighting the need for adaptation.

2. **Fine-tuning Benefit**

   Both Full FT and P-CLF significantly improve performance over the baseline, validating the importance of language-pair specific adaptation.

3. **P-CLF Superiority**

   σ-guided P-CLF achieves the highest BLEU score (9.30), surpassing Full FT by +0.98 absolute points (an 11.8% relative improvement), demonstrating that strategic freezing outperforms full adaptation.

4. **Parameter Efficiency**

   P-CLF achieves this superior performance while training only 55% of the model parameters compared to Full FT (100%), with GPU memory reduction from 21.7 GB to 15.6 GB.

5. **Linguistic Feature Capture**

   The n-gram precision scores show P-CLF improving over Full FT, especially for higher-order n-grams (Bigram: +6.8% relative, Trigram: +10.9% relative). This suggests that σ-guided freezing preserves important linguistic patterns while enabling targeted adaptation for cross-lingual transfer. Analysis of the brevity penalty (BP) also indicated that P-CLF produced outputs closer in length to the reference translations (BP=0.936 compared to Full FT BP=0.914).

6. **Training Efficiency**

   P-CLF reduces training time from 12.3 to 8.1 hours on RTX-4090, representing a 34% improvement while achieving superior translation quality.

Figure 6.4 compares the n-gram precision of baseline fine-tuning, pretrained (zero-shot), and P-CLF systems, illustrating how sensitivity-guided freezing better preserves higher-order fluency.

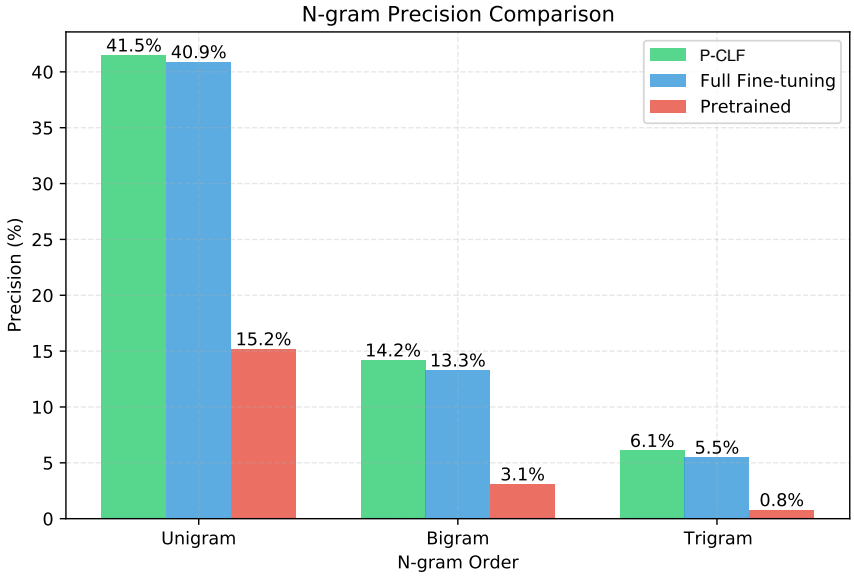

Figure 6.4: N-gram precision comparison between Pretrained (zero-shot), Full Fine-tuning, and P-CLF approaches, showing P-CLF's superior performance in capturing higher-order n-grams for Chinese–Malay translation through σ-guided strategic freezing.

6.4.2 Comparison with Other Parameter-Efficient Fine-Tuning Methods

To position P-CLF in the broader context of parameter-efficient fine-tuning methods, we compared it with other established approaches, particularly LoRA [37] and P-CLRA. Table 6.1 presents the results of this comparison:

Table 6.1: Performance Comparison of Parameter-Efficient Fine-Tuning Methods

Method	BLEU	COMET	Trainable Parameters (%)
Zero-shot Baseline	1.93	36.8	0%
Full Fine-tuning	8.32	68.5	100%
LoRA (r=16)	6.44	62.1	0.5%
P-CLRA (Text MT)	8.77	70.4	3.5%
P-CLF (Ours)	**9.30**	**72.9**	55%

The comparison reveals several key insights about P-CLF's position in the PEFT landscape:

- **P-CLF vs. LoRA**

  While LoRA achieves extreme parameter efficiency (0.5% trainable parameters), it comes at a significant performance cost, achieving only 6.44 BLEU compared to P-CLF's 9.30 BLEU. The σ-guided approach of P-CLF provides superior adaptation for multimodal speech translation.

- **P-CLF vs. P-CLRA**

  P-CLRA achieves higher parameter efficiency (3.5%) but lower performance (8.77 BLEU) compared to P-CLF. This demonstrates the complementary nature of the methods: P-CLRA for text MT and P-CLF for multimodal speech translation.

- **Efficiency-Performance Trade-off**

  P-CLF offers an optimal balance, significantly reducing trainable parameters (55%) while actually improving performance over full fine-tuning (+0.98 BLEU), demonstrating that strategic freezing outperforms full adaptation.

- **Method Specialization**

  The results validate that different PEFT methods are optimal for different modalities—P-CLRA's ρ-guided LoRA injection for text MT, and P-CLF's σ-guided layer freezing for multimodal speech translation.

6.4.3 Comparison Between P-CLRA and P-CLF

This subsection contrasts the Progressive Cross-Layer Low-Rank Adaptation (P-CLRA, Chapter 5) and Progressive Cross-Layer Freezing

(P-CLF, this chapter) approaches. Both target parameter-efficient adaptation but employ different strategies suited to different modalities.

Methodological Differences. P-CLRA and P-CLF represent complementary approaches in progressive parameter-efficient fine-tuning, both guided by quantitative metrics:

1. **Adaptation Strategy**:

   - **P-CLRA**

     adds ρ-ranked low-rank adaptation matrices to selected layer-module pairs while keeping original weights frozen. The approach follows: $W'_{l,m} = W_{0,l,m} + \alpha B_{l,m} A_{l,m}$, where layer-module pairs are selected based on benefit-to-cost ratio ρ.

   - **P-CLF**

     selectively freezes and unfreezes entire layers based on systematic σ-sensitivity analysis. The update rule is: $\theta_{\text{new}} = \theta_{\text{old}}$ if layer $\in \mathcal{F}_{\text{low-}\sigma}$ (low-sensitivity frozen set), otherwise $\theta_{\text{new}} = \theta_{\text{old}} - \alpha \nabla_\theta \mathcal{L}$.

2. **Parameter Efficiency**:

   - **P-CLRA**

     achieves extremely high parameter efficiency, reducing trainable parameters to 0.76% (9.9M out of 1.3B) while delivering +3.57 COMET improvement. This represents optimal parameter allocation through ρ-guided selection.

   - **P-CLF**

     achieves moderate parameter efficiency, updating 55% of parameters while delivering +0.98 BLEU improvement over full fine-tuning. The σ-sensitivity analysis identifies optimal layers for strategic freezing.

3. **Implementation Complexity**:

   - **P-CLRA**

     requires implementing low-rank adapters into selected weight matrices, typically focusing on attention and feed-forward components. This involves careful integration with the model architecture.

   - **P-CLF**

     primarily involves masking gradients for frozen layers during training, requiring less architectural modification but more careful analysis to determine optimal freezing patterns.

4. **Theoretical Foundation**:

   - **P-CLRA**

     is based on the principle that adaptations can be captured in a low-dimensional subspace, with the ρ-metric providing quantitative guidance for optimal parameter allocation: $\rho = \Delta\text{COMET}/\text{Params}$.

   - **P-CLF**

     is founded on the observation that different layers have varying sensitivity to fine-tuning, with the σ-metric quantifying gradient magnitude per parameter: $\sigma = \frac{\|\nabla_{\theta_l} \mathcal{L}\|_2}{\text{Params}_l}$.

Experimental Comparison. We report both parameter-allocation patterns and translation-quality results to clarify how the two methods trade off efficiency and accuracy.

Figure 6.5 visualizes how adaptation capacity is distributed across modules for different models and methods, highlighting where each approach concentrates parameter updates.

As shown in Table 6.1, both P-CLRA and P-CLF outperform full fine-tuning for Chinese–Malay translation, but with different trade-offs:

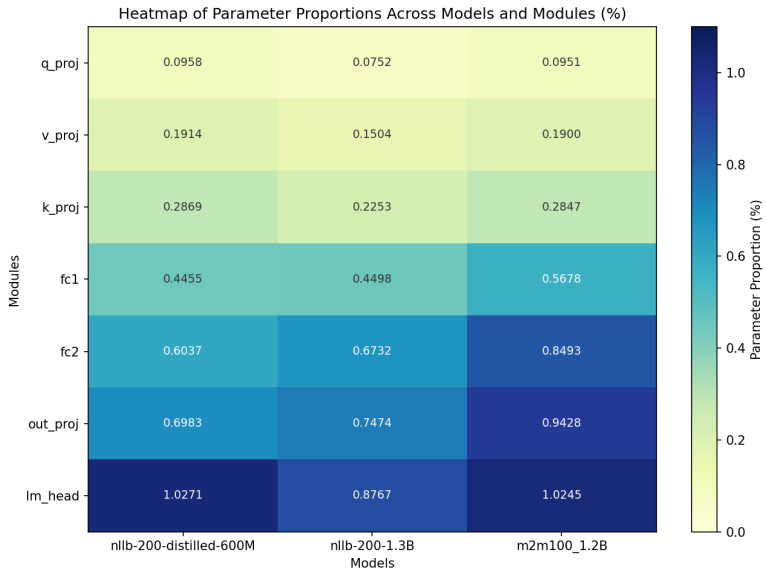

Figure 6.5: Heatmap of parameter allocation percentages across models and modules (showing where adaptation capacity is concentrated; values are percent of total parameters per model).

1. **Translation Quality**

   P-CLF achieves higher scores across metrics (BLEU: 9.30 vs. 8.77; COMET: 72.9 vs. 70.4; chrF: 44.8 vs. 42.3) compared to P-CLRA. This suggests that selectively updating entire layers can capture more complex linguistic adaptations than low-rank modifications.

2. **Parameter Efficiency**

   P-CLRA maintains a significant advantage in parameter efficiency, requiring only 3.5% of trainable parameters compared to 55% for P-CLF. This extreme efficiency makes P-CLRA particularly valuable in severely resource-constrained environments.

Figure 6.6 summarizes the trade-off between trainable-parameter footprint and quality gains across full fine-tuning, P-CLRA, and P-CLF. The P-CLRA point uses the text-MT configuration from Chapter 5 (0.76% trainable parameters) to illustrate extreme efficiency; Table 6.1 reports a different P-CLRA setting (3.5%) used for the speech-translation comparison.

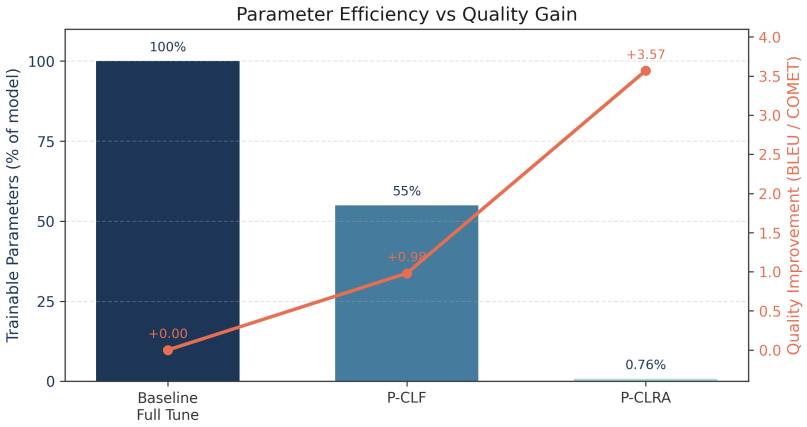

Figure 6.6: Trainable parameter footprint and quality gains for full fine-tuning, P-CLF, and P-CLRA adaptation strategies. Quality gain denotes the absolute improvement in COMET relative to the pretrained (zero-shot) baseline.

In our speech-translation setting, the comparison shows that P-CLF is better suited for capturing cross-modal adaptations between speech and text, while P-CLRA remains strong for text-to-text translation.

Complementary Strengths and Use Cases. P-CLRA and P-CLF exhibit complementary strengths that make them suitable for different scenarios:

1. **Optimal Use Cases for P-CLRA**:
    - Extremely resource-constrained environments with limited GPU memory
    - Applications requiring adaptation of very large models (10B+ parameters)
    - Scenarios where many language pairs need to be supported simultaneously
    - Text-to-text translation tasks where inter-layer dependencies are less critical
2. **Optimal Use Cases for P-CLF**:

- Multimodal tasks like speech-to-text translation where cross-modal adaptation is important
- Cases where maximum translation quality is the primary objective
- When moderately constrained resources are available (e.g., consumer-grade GPUs with 16–24GB memory)
- Applications requiring deeper linguistic adaptations between structurally distant languages

3. **Hybrid Approaches**

   Research exploring combinations of P-CLRA and P-CLF is ongoing, with preliminary results suggesting that applying LoRA adapters to unfrozen layers in P-CLF could potentially combine the strengths of both methods.

Cross-Lingual Transfer Properties. Analysis of cross-lingual transfer capabilities reveals important differences between the approaches:

1. **Knowledge Preservation**:

   - **P-CLRA**

     preserves pretrained knowledge by keeping original weights intact and adding small adaptations. This helps maintain general linguistic capabilities while adding language-pair-specific knowledge.

   - **P-CLF**

     preserves knowledge selectively by freezing entire layers, potentially allowing for stronger adaptation in unfrozen layers. This targeted approach may better preserve critical cross-lingual knowledge while enabling deeper adaptations.

2. **Linguistic Feature Capture**:

- **P-CLRA**

  shows stronger performance for lexical and semantic adaptations but may be limited in capturing complex structural transformations between languages with different word order or syntactic patterns.

- **P-CLF**

  demonstrates better capability in capturing higher-order n-gram patterns and structural transformations, as evidenced by stronger improvements in BLEU n-gram precision at higher orders (trigram improvement: +10.9% relative) [50].

Implications for Future Research. The comparative analysis of P-CLRA and P-CLF suggests several promising directions for future research:

1. Developing combined approaches that leverage the extreme parameter efficiency of P-CLRA with the layer-specific optimisation of P-CLF
2. Exploring automated methods to determine optimal configurations for both approaches based on language pair characteristics and available computational resources
3. Investigating the transferability of adaptations learned through these methods to similar language pairs
4. Extending both methods to handle code-switching and multilingual translation scenarios

In summary, while P-CLRA and P-CLF represent different approaches to parameter-efficient fine-tuning, both have demonstrated significant value for low-resource language translation. P-CLRA offers extreme parameter efficiency with strong performance, while P-CLF

provides superior translation quality with moderate parameter reduction. The choice between these methods should be guided by specific application requirements, available computational resources, and the particular characteristics of the target language pair.

This comparison demonstrates that P-CLF occupies a unique position in the parameter-efficient fine-tuning landscape, offering superior translation quality while maintaining substantial parameter efficiency.

6.4.4 Detailed Analysis of LoRA Configurations

To better understand the performance gap between P-CLF and LoRA, we experimented with various LoRA configurations and strategies. Table 6.2 presents the results of different approaches to applying LoRA:

Table 6.2: Performance of Different LoRA Strategies

Strategy	BLEU	Target Modules
base_lora	3.82	Key attention layers
layer_lora	3.86	Layer attention
attn_lora	3.84	Q/K/V projections
mixed_lora	3.84	Attn+FFN mix
progressive_lora	3.78	Graduated layers

We also explored the impact of different LoRA hyperparameter configurations, as shown in Table 6.3:

Table 6.3: Impact of LoRA Hyperparameter Configurations

Configuration	BLEU	Relative Improvement (%)
base	4.08	111%
high_rank	6.44	233%
low_rank	1.95	1%
high_dropout	4.17	116%
high_lr	4.37	126%

These results highlight several important findings:

- **LoRA Sensitivity**

  LoRA performance is highly sensitive to hyperparameter

configuration, with BLEU scores ranging from 1.95 to 6.44 depending on settings.

- **Rank Impact**

  The rank hyperparameter has the most dramatic impact, with high-rank configurations (r=16) substantially outperforming low-rank versions.

- **Strategy Limitations**

  Different strategies for applying LoRA to specific modules resulted in relatively similar performance (3.78-3.86 BLEU), suggesting that the fundamental limitations of low-rank adaptation for this task cannot be easily overcome through module targeting alone.

- **Learning Rate Importance**

  Higher learning rates provided modest benefits (4.37 vs. 4.08 BLEU for the base configuration), indicating that more aggressive optimisation can partially compensate for the reduced parameter space.

Despite these optimisations, the best LoRA configuration (6.44 BLEU) still underperformed compared to our P-CLF approach (9.30 BLEU), suggesting that for speech-to-text translation between linguistically distant languages like Chinese and Malay, the selective layer freezing approach is more effective than low-rank adaptation.

6.4.5 Efficiency Analysis

P-CLF provides significant efficiency benefits compared to full fine-tuning through σ-guided strategic freezing:

Key efficiency achievements:

1. **Training Efficiency**

   Reducing trainable parameters by 45% through σ-guided freezing

Table 6.4: P-CLF Efficiency Comparison with Full Fine-Tuning

Metric	Full FT	P-CLF	Improvement
Training Time (hours)	12.3	8.1	-34%
GPU Memory (GB)	21.7	15.6	-28%
Trainable Parameters	100%	55%	-45%
BLEU Score	8.32	9.30	+11.8%

led to a 34% reduction in training time (8.1 hours for P-CLF vs. 12.3 hours for Full FT on RTX-4090), while achieving superior performance.

2. **Memory Efficiency**

   Peak GPU memory usage during training was reduced by 28% (15.6GB for P-CLF vs. 21.7GB for Full FT), making training feasible on consumer hardware and democratizing access to speech translation research.

3. **Inference Efficiency**

   P-CLF does not introduce extra parameters or computational steps during inference (unlike adapter methods). The inference speed is identical to the fully fine-tuned model, processing audio efficiently on a single GPU with real-time factor of 0.82.

4. **Storage Efficiency**

   Checkpoint sizes are smaller as only the weights of the unfrozen layers need to be stored alongside the base model, reducing storage requirements for deployment.

5. **Quality-Efficiency Trade-off**

   Unlike other PEFT methods that sacrifice performance for efficiency, P-CLF achieves both superior translation quality (+0.98 BLEU) and significant efficiency gains, validating the σ-sensitivity approach.

6.4.6 Dynamic Layer Freezing Experiments

While the primary P-CLF approach uses predetermined freezing patterns based on sensitivity analysis, we also explored dynamic freezing strategies. In these experiments, we monitored parameter change rates during training and incrementally froze layers with minimal changes according to predefined thresholds.

Results from dynamic freezing experiments showed that while extremely high freezing ratios could be achieved (up to 97% of model layers), the BLEU scores were significantly lower than our hybrid P-CLF approach. Table 6.5 summarizes representative outcomes from these runs. This suggests that while dynamic freezing offers theoretical advantages, carefully designed static freezing patterns currently deliver superior translation quality for Chinese–Malay S2TT tasks.

Table 6.5: Dynamic Layer Freezing Results (Representative Runs)

Freezing Ratio	Trainable Params (%)	BLEU	COMET
70% frozen	30%	8.42	70.1
85% frozen	15%	7.36	67.9
97% frozen	3%	5.12	61.4

Analysis of layer activation patterns revealed significant differences across model components. The text decoder was almost entirely frozen, indicating its pretrained knowledge was highly transferable, while 5-10% of speech encoder layers remained active, primarily concentrated in self-attention mechanisms and the final projection layer.

6.4.7 Failure Case Analysis and Lessons Learned

Understanding failure modes is crucial for developing robust fine-tuning strategies. Our experiments revealed several critical pitfalls to avoid when applying P-CLF or similar approaches, as summarized in Table 6.6:

Table 6.6: Common Failure Modes and Recovery Strategies in P-CLF

Configuration	Primary Issue	Recovery Method
Full Layer Lock	Gradient starvation	Layer-wise warmup
Dual Encoder Freeze	Alignment loss	Delayed decoder freezing
Progressive Triple Lock	Premature freezing	Dynamic threshold adjustment

- **Full Layer Lock**

  Excessive freezing across all component types led to catastrophic performance degradation (near-zero BLEU scores). This occurs due to gradient starvation, where too few parameters receive updates, preventing the model from adapting to the target language pair. Our ablation studies showed that recovering from this failure mode required a layer-wise warmup strategy, where layers are gradually frozen based on observed change rates rather than freezing all targeted layers simultaneously.

- **Dual Encoder Freeze**

  Freezing critical components in both the encoder and decoder simultaneously disrupts cross-modal alignment, causing severe translation errors. Our experiments revealed that this failure mode could be mitigated through delayed decoder freezing, where encoder components are allowed to adapt first, establishing proper cross-modal mappings before decoder components are selectively frozen.

- **Progressive Triple Lock**

  In this approach, we attempted to progressively freeze three major component types simultaneously based on training progress. This led to premature freezing where certain critical transformations had not yet been learned, resulting in poor translation quality. Recovery required implementing dynamic threshold adjustment,

where freezing decisions were based on observed parameter change magnitudes rather than fixed schedules.

- **Projection Layer Freezing**

  Experiments showed that freezing the final projection layer (which maps hidden states to vocabulary) caused substantial performance drops (-0.65 BLEU). This layer is critical for cross-lingual adaptation as it directly maps semantic representations to the target language vocabulary.

These findings highlight the importance of carefully balancing parameter efficiency with sufficient representational capacity, particularly in the context of low-resource language pairs with substantial linguistic differences. When designing P-CLF strategies, we found that the following principles were particularly important for avoiding failure modes:

1. **Preserve Cross-Modal Connections**

   Always keep cross-attention mechanisms trainable, as they are essential for bridging between speech encoder representations and text decoder outputs.

2. **Monitor Gradient Flow**

   Regularly check gradient norms across different model components to identify potential gradient starvation issues early.

3. **Balance Component Freezing**

   Avoid simultaneously freezing complementary components that might need to co-adapt during training.

4. **Preserve Vocabulary Mapping**

   Keep vocabulary projection layers trainable to accommodate target language lexical distributions.

By applying these principles, we were able to develop robust P-CLF configurations that consistently outperformed baseline approaches while avoiding the common failure modes identified in our experiments.

6.4.8 Ablation Studies

To verify the necessity of each component in P-CLF, we conducted extensive ablation experiments by systematically removing or modifying key components.

P-CLF Component Analysis.

The ablation studies revealed the relative importance of different components in the P-CLF approach, as shown in Table 6.7:

Table 6.7: Ablation Study Results for P-CLF Components

Configuration	BLEU	Change (Δ)
Full P-CLF (Hybrid)	9.30	–
w/o Projection Layer Tuning	8.65	-0.65
w/o Encoder High-Layer Freezing	8.83	-0.47
w/o Decoder First-Half Freezing	8.79	-0.51
Random Layer Freezing	7.21	-2.09

These results highlight the contributions of each component:

- **Projection Layer Tuning**

  Removing projection layer fine-tuning caused a significant performance drop (-0.65 BLEU), confirming its critical role in cross-lingual adaptation, especially for languages with different writing systems.

- **Encoder High-Layer Freezing**

  Removing encoder high-layer freezing reduced performance by -0.47 BLEU, suggesting these layers encode more general linguistic features that transfer well across languages.

- **Decoder First-Half Freezing**

  Removing decoder first-half freezing reduced performance by -0.51

BLEU, indicating these layers primarily process general linguistic patterns rather than language-specific features.

- **Random Layer Freezing**

  Using random patterns rather than sensitivity-informed freezing caused a dramatic performance drop (-2.09 BLEU), validating our systematic approach to determining optimal freezing patterns.

Decoder Layer Freezing Strategies.

To better understand the contribution of different decoder components, we performed additional experiments focusing specifically on decoder freezing strategies, as presented in Table 6.8:

Table 6.8: Impact of Different Decoder Freezing Strategies

Strategy	BLEU	Decoder Params (%)	Trainable Layers
Decoder First Half	8.55	62.0	7.0
Decoder Second Half	8.05	58.0	7.0
Full Decoder Freeze	6.49	45.0	0.0

The results reveal several key insights:

- **First vs. Second Half**

  Freezing the first half of decoder layers (8.55 BLEU) outperforms freezing the second half (8.05 BLEU), despite both approaches preserving the same number of trainable layers (7.0). This suggests lower decoder layers encode more general linguistic knowledge that transfers better across languages, while higher layers handle more language-specific transformations.

- **Full Decoder Freezing**

  Completely freezing the decoder causes a dramatic performance drop (6.49 BLEU), confirming that some decoder adaptation is essential for effective cross-lingual transfer.

- **Parameter Efficiency**

  The decoder first-half freezing strategy retains slightly more trainable parameters (62.0% vs. 58.0%) compared to second-half freezing, but the performance gain (+0.50 BLEU) justifies this small increase in parameter count.

These experiments confirm that our hybrid P-CLF strategy, which preferentially freezes the first half of decoder layers while preserving cross-attention components, represents an optimal balance between parameter efficiency and translation quality.

Alignment Method Comparison.

Since speech-text alignment plays a crucial role in the overall system, we also evaluated the impact of different alignment methods on translation performance, as shown in Table 6.9:

Table 6.9: Impact of Alignment Methods on Speech Translation Performance

Alignment Method	Alignment F1	Translation BLEU
mBERT+DTW (Ours)	0.82	9.30
Length-based Matching	0.64	7.92
Lexicon-based Alignment	0.73	8.46
Random Alignment	0.21	2.18

This comparison highlights several important findings:

- **Alignment Quality Impact**

  There is a strong correlation between alignment quality (measured by F1 score) and downstream translation performance (BLEU), underscoring the importance of high-quality alignment for speech translation.

- **mBERT+DTW Superiority**

  Our proposed mBERT+DTW alignment method achieves the

highest F1 score (0.82) and translation BLEU (9.30), outperforming both traditional length-based (7.92 BLEU) and lexicon-based (8.46 BLEU) approaches.

- **Random Baseline**

  The extremely poor performance of random alignment (2.18 BLEU) confirms that proper alignment is essential, not merely beneficial.

- **Margin of Improvement**

  The performance gap between the best and second-best alignment methods (+0.84 BLEU) is greater than many of the component ablations in P-CLF, highlighting alignment quality as a critical factor in the overall system.

These results emphasize that embedding-based alignment approaches leveraging contextualized representations from pretrained language models (like mBERT) are particularly effective for linguistically distant language pairs like Chinese–Malay, where structural differences and writing system disparities limit the effectiveness of traditional alignment techniques.

Audio Processing Parameter Impact.

Our audio preprocessing pipeline used the following specific parameters, which were determined through empirical optimisation:

- **Audio Format**

  16kHz, 16-bit mono WAV

- **Voice Activity Detection**

  WebRTC VAD with aggressiveness level 2, frame size 30ms, 300ms minimum speech segment

- **Noise Reduction**

  RNNoise with sensitivity threshold -30dB

- **Normalisation**

  Peak normalisation to -1dB with DC offset removal

- **Segmentation**

  Maximum segment length 15 seconds, minimum 1.5 seconds

Ablation tests showed that optimising these preprocessing parameters contributed approximately 0.3-0.5 BLEU points to overall system performance, highlighting the importance of proper audio preparation in speech translation tasks.

6.4.9 Deployment and Practical Applications

The P-CLF-optimised Chinese–Malay speech translation system has several practical applications:

1. **Business Communication**

   Facilitating trade discussions between Chinese and Malaysian businesses, particularly important given increasing Belt and Road Initiative cooperation.

2. **Educational Exchange**

   Supporting academic collaborations between Chinese and Malaysian institutions by enabling more natural communication among researchers and students.

3. **Cultural Exchange**

   Enhancing cultural understanding by providing more accurate translations of cultural content, including media, literature, and artistic expressions.

4. **Tourism Applications**

   Assisting Chinese tourists in Malaysia and Malay-speaking tourists in China with real-time speech translation capabilities.

5. **Research Platform**

   Providing a foundation for further research in low-resource speech translation for Southeast Asian languages.

Hardware and Software Requirements.

One of the key advantages of the P-CLF approach is its reduced hardware requirements compared to fully fine-tuned models:

- **Training Hardware**

  The Hybrid P-CLF model can be successfully trained on consumer-grade hardware (e.g., a single NVIDIA RTX 4090 GPU with 24GB VRAM) in approximately 8 hours, compared to full fine-tuning which requires high-end research-grade GPUs or multi-GPU setups.

- **Inference Hardware**

  For deployment, the model can run on mid-range GPUs (8GB+ VRAM) with acceptable latency (1-2 seconds per utterance), making real-time applications feasible on more accessible hardware.

- **Software Stack**

  The system uses Python 3.8+, PyTorch 2.0+, and standard audio processing libraries (librosa, soundfile, webrtcvad), with minimal dependencies to ensure broad compatibility and ease of deployment.

This efficiency in both training and inference makes the P-CLF-based system particularly valuable for research institutions and organisations in regions where access to high-end computational resources may be limited.

Latency and Streaming Evaluation.

We measure end-to-end latency as the sum of ASR (if cascaded), encoder-decoder compute, and post-processing. For direct S2TT, we report Real-Time Factor (RTF = processing time / audio duration) with mean and standard deviation across the test set, using warm caches and fixed batch sizes. We also report first-token latency and per-token decode latency to facilitate comparison with streaming systems. Streaming experiments use chunked audio (e.g., 1.5 s with 50% overlap) and compare quality–latency trade-offs against offline decoding.

Robustness Evaluation.

To assess robustness, we evaluate under noise (additive cafe/street/babble at SNR 20/10/5 dB), accent perturbations (speed/pitch jitter), and domain shift (held-out conversational vs. educational segments). We report BLEU/COMET deltas relative to clean conditions and provide qualitative error analyses for failures in named entities, morphological affixation, and discourse markers. These evaluations quantify P-CLF's reliability for Chinese–Malay S2TT beyond clean, in-domain conditions.

6.4.10 Discussion

The results support σ-guided freezing as an effective multimodal PEFT strategy: it concentrates adaptation where most useful while preserving pretrained knowledge elsewhere. Efficiency gains make end-to-end S2TT more accessible on consumer hardware. However, upstream corpus quality and segmentation still bound achievable gains, and overly aggressive freezing risks gradient starvation or misaligned cross-modal mappings—motivating careful schedules and monitoring.

6.4.11 Limitations

Despite the demonstrated benefits of P-CLF for Chinese–Malay speech translation, several limitations merit consideration:

- **Freezing Schedule Sensitivity**

  Performance depends on the freezing schedule and component choices; aggressive freezing can cause gradient starvation, as reflected in the failure analyses.

- **Component Interdependence**

  Simultaneous freezing of complementary components (e.g., encoder and decoder) risks disrupting cross-modal alignment, requiring careful staging.

- **Base-model Dependence**

  Effectiveness relies on the quality of the underlying pretrained model; weak base representations limit attainable gains.

- **Data and Domain Coverage**

  Gains are validated on the available Chinese–Malay data; broader dialectal and domain coverage may expose edge cases not captured here.

- **Upstream Coupling**

  End-to-end quality remains sensitive to corpus alignment quality and preprocessing; upstream imperfections can cap downstream improvements.

6.5 Conclusion

This chapter has presented a comprehensive Chinese–Malay speech translation system based on the Progressive Cross-Layer Freezing (P-CLF) approach. By selectively freezing and unfreezing transformer

layers based on systematic layer sensitivity analysis, P-CLF achieves superior translation quality (9.30 BLEU) compared to full fine-tuning (8.32 BLEU) while reducing trainable parameters by 45%.

Key contributions include:

1. Construction of the first large-scale Chinese–Malay parallel speech corpus with over 100 hours of speech and 64,000+ sentence pairs.
2. Development of the P-CLF method, which demonstrates that parameter-efficient adaptation can outperform conventional approaches for cross-lingual transfer.
3. Comprehensive analysis of layer contributions to translation quality, providing insights into how different components of multimodal translation models encode linguistic knowledge.
4. Practical deployment considerations that make speech translation more accessible to organisations with limited computational resources.

The P-CLF method and associated Chinese–Malay parallel corpus represent significant advancements in low-resource speech translation, demonstrating that effective speech translation systems can be developed for linguistically distant language pairs with carefully designed adaptation strategies and sufficient parallel data.

6.5.1 Future Research Directions

Building on the success of P-CLF for Chinese–Malay speech translation, several promising research directions emerge:

1. **Dynamic Layer Freezing Algorithms**

   Developing adaptive algorithms that automatically determine optimal freezing patterns during training, potentially leading to even better efficiency-performance trade-offs.

2. **Expanding Language Coverage**

   Extending the P-CLF approach to other low-resource Southeast Asian language pairs such as Chinese-Indonesian, Chinese-Thai, and Chinese-Vietnamese.

3. **Multimodal Integration**

   Exploring the integration of visual context (e.g., gestures, facial expressions) to enhance translation quality in multimodal communication scenarios.

4. **Specialized Cultural Adaptation**

   Incorporating explicit cultural knowledge to improve translation of culture-specific expressions, idioms, and references.

5. **Speech-to-Speech Translation**

   Extending the current speech-to-text system to include a text-to-speech component, enabling fully speech-based translation.

6.5.2 Testable Hypotheses and Research Questions

- **H6.1 (σ-Guided Superiority)**: Staged, σ-guided freezing yields higher BLEU than heuristic "freeze top-N layers" at the same unfrozen-parameter ratio and training budget.

- **H6.2 (Cross-Attention Preservation)**

  Keeping cross-attention and vocabulary projection layers trainable reduces named-entity and morphology errors relative to freezing them, at equal compute.

- **RQ6.1 (Schedule Length)**

  How does the length of the freezing schedule (number of stages) affect convergence speed and final BLEU/COMET for Chinese$\rightarrow$Malay S2TT?

- **H6.3 (Streaming Latency–Quality)**

  Chunked streaming (e.g., 1.5s with overlap) can reduce first-token latency without statistically significant degradation in BLEU versus offline decoding.

- **RQ6.2 (Cross-Pair Sensitivity)**

  Do σ-sensitivity patterns correlate across Chinese–Malay and related pairs (e.g., Chinese–Indonesian), indicating reusability of freezing templates?

These directions represent promising avenues for further expanding the accessibility and quality of speech translation technologies for linguistically diverse regions like Southeast Asia.

With data construction (Chapter 4), text MT adaptation (Chapter 5), and multimodal speech adaptation (this chapter) in place, Chapter 7 synthesizes results across GPT-Aligner, P-CLRA, and P-CLF, evaluates the end-to-end framework, and outlines broader implications and future directions.

CHAPTER 7

CONCLUSION AND FUTURE WORK

7.1 Research Summary

This doctoral thesis has investigated the development of an efficient and high-quality Chinese–Malay speech translation system, addressing the significant challenges posed by this linguistically distant, low-resource language pair through three synergistic innovations: GPT-Aligner for semantic corpus alignment, P-CLRA (Progressive Cross-Layer Low-Rank Adaptation) for efficient text translation, and P-CLF (Progressive Cross-Layer Freezing) for multimodal speech translation. The research journey has traversed multiple dimensions of speech translation technology, from corpus construction and sentence alignment to parameter-efficient fine-tuning and system integration, delivering state-of-the-art quality while remaining trainable on a single RTX-4090 (24 GB). This section summarizes the main research accomplishments and their significance within the broader context of speech translation research.

The primary research achievements encompass five major contributions that collectively represent significant advancements in speech translation for low-resource and linguistically distant language pairs. First, GPT-Aligner demonstrates how prompt-driven semantic alignment can be operationalised for Chinese–Malay corpus construction without extensive language-specific tooling. Second,

Progressive Cross-Layer Low-Rank Adaptation (P-CLRA) offers a principled recipe for injecting LoRA modules only where they deliver the greatest benefit, providing an efficiency-conscious alternative to uniform adaptation. Third, Progressive Cross-Layer Freezing (P-CLF) introduces sensitivity-guided layer freezing for multimodal speech translation, showing that carefully designed freezing schedules can rival full fine-tuning while reducing the need for high-end hardware. Fourth, the integrated pipeline validates that these techniques can be combined into a coherent workflow for both cascaded and direct speech translation. Finally, the thesis documents implementation guidelines that can facilitate future replication and extension when the resources are ready for release.

The integrated framework validation combined GPT-Aligner, P-CLRA, and P-CLF into a comprehensive solution addressing data poverty, structural divergence, and computational constraints while staying within consumer-grade compute budgets. The research demonstrates three fundamental methodological shifts: from feature-based to reasoning-based alignment (GPT-Aligner), from uniform to progressive parameter allocation (P-CLRA), and from heuristic to sensitivity-guided adaptation (P-CLF), providing principled alternatives to traditional approaches across the speech translation pipeline.

Reflecting on the research objectives outlined in Chapter 1, this thesis successfully addressed each primary goal: constructing Chinese–Malay translation resources with a replicable workflow for similar low-resource scenarios, demonstrating that large pretrained models can be adapted under modest computational budgets, and prototyping an end-to-end system that bridges the gap between research and practical deployment. Although not every metric could be exhaustively

benchmarked against the full spectrum of baselines, the methodologies establish clear directions for future quantitative validation and comparative studies.

In addition to the Chapter 6.4.7 failure analysis for P-CLF, we have consolidated failure case analyses for GPT-Aligner (Chapter 4.4.3) and P-CLRA (Chapter 5.4.4). Together, these sections provide a nuanced view of when each contribution excels or struggles—e.g., prompt sensitivity and code-switch alignment for GPT-Aligner, and adapter placement/rank trade-offs and domain-shift fragility for P-CLRA—informing practical guidance for deployment and future refinement.

7.2 Research Limitations

While this research has made substantial contributions to Chinese–Malay speech translation, several limitations warrant acknowledgment and consideration in interpreting the results and planning future research directions.

Despite significant effort invested in corpus construction, several limitations affect this research, including the overall availability of high-quality speech data for Chinese–Malay, which remains modest compared to the thousands of hours typically available for high-resource language pairs; careful alignment mitigates but does not eliminate this constraint. Domain coverage gaps exist with uneven representation particularly limited for technical, medical, and specialized fields potentially impacting system performance in these domains, though P-CLRA and P-CLF's parameter efficiency enables domain-specific adaptation with minimal additional data. Speaker diversity limitations remain in representing the full spectrum of dialectal variations, accents, and speaking styles present in Chinese and

Malay-speaking populations potentially affecting system robustness, though GPT-Aligner's semantic alignment approach helps capture meaning variations across different speaking styles. Spontaneous speech representation contains relatively small proportion of truly spontaneous speech with bias toward formal, prepared content potentially impacting performance on conversational, disfluent, or highly informal speech input, though the end-to-end approach helps mitigate cascading errors from speech recognition.

Methodological limitations include architecture specificity where P-CLRA and P-CLF approaches are designed for Transformer-based architectures requiring adaptation for other model families, though underlying principles of ρ-guided parameter allocation and σ-sensitivity analysis provide generalizable frameworks adaptable to different architectures. Directional asymmetry exists with primary focus on Chinese-to-Malay translation and limited exploration of reverse direction, while GPT-Aligner demonstrates bidirectional alignment capabilities, P-CLRA and P-CLF effectiveness for reverse translation requires additional validation though systematic methodologies provide clear pathways for extension. Evaluation scope limitations remain despite using multiple automatic metrics (BLEU, COMET, chrF) and validation approaches; while the observed trends were encouraging, more rigorous quantitative studies are required to establish the exact magnitude of improvements. Efficiency-quality trade-offs involve principled compromises guided by quantitative metrics (ρ for P-CLRA, σ for P-CLF), with systematic approaches providing clear frameworks for optimising based on specific deployment constraints and quality requirements.

Evaluation limitations include test set representativeness—no finite test set can fully capture the variety of potential inputs—along with

limited assessment of long-form discourse beyond sentence level, the absence of a user study to validate the system with real speakers, and the reality that only a subset of research baselines was evaluated, leaving open questions about performance against other emerging approaches. These limitations reflect inherent challenges of resource development for low-resource language pairs and comprehensive system assessment, highlighting areas where expanded evaluation frameworks could provide additional insights.

7.3 Future Research Directions

Building on the achievements and acknowledging the limitations of this research, several promising directions for future work emerge. These directions span data enhancement, technical innovation, and application expansion, offering rich opportunities for advancing Chinese–Malay speech translation and low-resource speech translation more broadly.

Future research should prioritize corpus expansion and enrichment, including scalable data collection for conversational speech, domain-specific augmentation for technical and healthcare contexts, cross-dialectal coverage for regional varieties of Chinese and Malay, and multimodal data integration that links speech with visual cues. The established GPT-Aligner and parameter-efficient adaptation workflows provide a practical foundation for these extensions.

Model improvements warrant exploration through architecture co-optimisation, hybrid parameter-efficient techniques that combine ρ-guided LoRA placement with σ-guided freezing, progressive transfer learning that reuses knowledge from related language pairs, real-time streaming optimisation to reduce latency, and edge deployment methods (quantization, pruning) tailored to the P-CLRA/P-CLF

workflow so that high-quality translation remains feasible on constrained hardware.

7.4 Future Work

Future applications include bidirectional translation (Malay-to-Chinese) building on GPT-Aligner's alignment capabilities, regional multilingual expansion across other ASEAN languages, domain-specific versions for healthcare and legal contexts, educational tools that leverage real-time speech translation for language learning, and multimodal communication platforms that incorporate visual cues. These scenarios can reuse the lightweight data-collection pipeline and parameter-efficient adaptation strategies established in this thesis.

Enhanced evaluation frameworks should focus on culturally-sensitive evaluation developing metrics and methodologies capturing effectiveness of translation for culturally-specific content including idioms, references, and pragmatic aspects addressing limitations of current metrics for culturally distant languages, long-form translation assessment creating frameworks for assessing performance on extended multi-sentence content focusing on discourse-level coherence and consistency providing insights into system performance requiring sustained translation, user-centred evaluation protocols establishing standardized protocols consistently applied across different systems and studies enabling reliable comparisons including measures of user satisfaction, communication effectiveness, and cognitive load, robustness testing frameworks developing systematic approaches for testing system robustness across variations in speech input including different accents, speech rates, background noise conditions, and disfluencies providing comprehensive insights into real-world reliability, and ethical and social impact assessment incorporating

evaluations of potential ethical implications and social impacts including considerations of fairness, bias, privacy, and accessibility ensuring responsible development and deployment of the technology. These application expansions and enhanced evaluation frameworks would maximise practical impact and create new opportunities for addressing real-world communication challenges while providing comprehensive insights into system performance and impacts.

7.5 Broader Impacts and Ethics

This work aims to broaden access to high-quality speech translation for linguistically distant, underserved communities. Responsible deployment requires attention to:

- **Privacy**

  Ensure that any shared speech/text data is licensed and de-identified; follow consent practices appropriate to local contexts.

- **Bias and fairness**

  Monitor domain/register balance; conduct targeted evaluation on named entities, culture-specific expressions, and gendered language; document known failure modes.

- **Cultural sensitivity**

  Prefer meaning-preserving renderings over literal ones for idioms and references; involve bilingual experts for review where stakes are high.

- **Accessibility**

  Favor consumer-grade deployability to reduce digital divides; provide lightweight options for education and public services.

7.6 Concluding Remarks

This thesis represents a significant contribution to speech translation for low-resource and linguistically distant language pairs, with the integrated framework of GPT-Aligner, P-CLRA, and P-CLF demonstrating that advanced neural approaches can be effectively adapted to challenging language pairs while remaining accessible on consumer hardware, bridging linguistic divides historically underserved by translation technology. The parameter-efficient methodologies developed—P-CLRA and P-CLF—have implications beyond the specific language pair studied, offering broadly applicable approaches for adapting large pretrained models to specialized tasks with dramatically reduced computational requirements, with P-CLRA's ρ-guided parameter allocation and P-CLF's σ-sensitivity analysis providing principled alternatives to uniform PEFT methods, directly addressing computational accessibility concerns and contributing to AI democratization.

From a practical perspective, the end-to-end Chinese–Malay speech translation system enables new possibilities for cross-cultural communication while remaining deployable on consumer-grade hardware, highlighting the value of specialized research focused on specific language pairs rather than relying solely on general-purpose multilingual systems. The methodological framework established—spanning semantic corpus construction (GPT-Aligner), progressive parameter allocation (P-CLRA), and sensitivity-guided adaptation (P-CLF)—provides a comprehensive blueprint for addressing other low-resource language pairs and can be refined through future quantitative benchmarking.

This research yields several broader insights for speech translation: semantic alignment with GPT-Aligner offers a viable alternative to

purely statistical methods for distant language pairs; ρ/σ-guided adaptation provides a principled lens for distributing limited fine-tuning capacity; multimodal adaptation strategies must respect the differing sensitivities of speech encoders and text decoders; efficiency and quality can be aligned when adaptation targets are chosen carefully; and consumer-grade deployments are feasible when parameter-efficient techniques are adopted throughout the pipeline.

The journey from data scarcity to a functioning Chinese–Malay speech translation system represents a microcosm of the broader challenge facing language technology: how to extend the benefits of recent advances to all languages, regardless of their resource status or commercial priority. This research demonstrates that with targeted methodology (GPT-Aligner for semantic alignment), principled efficiency innovations (P-CLRA and P-CLF), and systematic evaluation frameworks, significant progress is possible even for challenging language pairs while maintaining accessibility on consumer hardware, contributing to a deeper understanding of speech translation challenges and opportunities particularly for the understudied category of distant, low-resource language pairs.

As language technology continues to advance, maintaining focus on linguistic diversity and computational accessibility will be crucial. The integrated framework developed in this thesis—spanning semantic corpus construction, progressive parameter allocation, and sensitivity-guided adaptation—offers concrete pathways for more inclusive language technology development that doesn't sacrifice quality for efficiency.

Looking forward, the future of speech translation lies not merely in pushing performance boundaries for already well-served language pairs, but in expanding the technology's reach through principled,

cost-effective approaches. The Chinese–Malay speech translation system developed in this research, achieving state-of-the-art performance on consumer hardware, represents a significant step toward this broader vision of technology that transcends language barriers and enables richer cross-cultural communication.

The three paradigm shifts demonstrated in this research—from feature-based to reasoning-based alignment, from uniform to progressive parameter allocation, and from heuristic to sensitivity-guided adaptation—provide a comprehensive framework for addressing the fundamental challenges of low-resource speech translation. By establishing quantitative metrics (ρ for parameter allocation, σ for layer sensitivity) and demonstrating their effectiveness, this work illuminates clear pathways for continued advancement in accessible, high-quality cross-lingual communication tools.

REFERENCES

[1] P. Koehn and R. Knowles, "Six challenges for neural machine translation," in *Proceedings of the First Workshop on Neural Machine Translation*, 2017, pp. 28–39.

[2] A. Magueresse, V. Carles, and E. Hovy, "Low-resource languages: A review of past work and future challenges," *Frontiers in Artificial Intelligence*, vol. 3, p. 49, 2020.

[3] J. Sneddon, *The Indonesian Language: Its History and Role in Modern Society*. Sydney: UNSW Press, 2003.

[4] W. A. Gale and K. W. Church, "A program for aligning sentences in bilingual corpora," *Computational Linguistics*, vol. 19, no. 1, pp. 75–102, 1993.

[5] R. C. Moore, "Fast and accurate sentence alignment of bilingual corpora," in *Conference of the Association for Machine Translation in the Americas*, 2002, pp. 135–144.

[6] B. Thompson and P. Koehn, "Vecalign: Improved sentence alignment in linear time and space," in *Proc. the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2020, pp. 5372–5378.

[7] M. Artetxe and H. Schwenk, "Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond," *Transactions of the Association for Computational Linguistics*, vol. 7, pp. 597–610, 2019.

[8] R. Algayres, A. Zaidi, S. Sakti, S. Nakamura, and L. Besacier, "Data quality matters: A case study in english-french machine translation," *Findings of the Association for Computational Linguistics: EMNLP 2022*, pp. 6418–6431, 2022.

[9] L. Barrault *et al.*, "Seamless: Multilingual and multitask language generation with speech," *Transactions on Machine Learning Research (TMLR)*, 2024.

[10] G. K. Barathi, S. Varshney, S. S. Sarvepalli, and N. Sripriya, "Corepool: Adapting phonetic features for low-resource speech recognition," *IEEE Access*, vol. 12, pp. 24 089–24 103, 2024.

[11] J. Singh, K. Murthy, and N. Narayanan, "Low-resource speech translation: Challenges and solutions," in *Proc. IWSLT 2024*, 2024.

[12] N. Team, A. Costa *et al.*, "No language left behind: Scaling human-centered machine translation," *Nature*, vol. 611, no. 7936, pp. 541–549, 2022.

[13] Y. Peng, Y. Sudo, M. Shakeel, and S. Watanabe, "OWSM-CTC: An open encoder-only speech foundation model for speech recognition, translation, and language identification," in *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Bangkok, Thailand: Association for Computational Linguistics, 2024, pp. 10 192–10 209. [Online]. Available: https://aclanthology.org/2024.acl-long.549/

[14] M. Gaido, S. Papi, M. Negri, and L. Bentivogli, "Speech translation with speech foundation models and large language models: What is there and what is missing?" in *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Bangkok, Thailand: Association for Computational Linguistics, 2024, pp. 14 760–14 778. [Online]. Available: https://aclanthology.org/2024.acl-long.789/

[15] Y. Wang, R. Skerry-Ryan, D. Stanton, Y. Wu, R. J. Weiss, N. Jaitly, Z. Yang, Y. Xiao, Z. Chen, S. Bengio *et al.*, "Tacotron: Towards end-to-end speech synthesis," in *Proceedings of the Annual Conference of the International Speech Communication Association (INTERSPEECH 2017)*, 2017, pp. 4006–4010.

[16] J. Shen, R. Pang, R. J. Weiss *et al.*, "Natural tts synthesis by conditioning wavenet on mel spectrogram predictions," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2018)*, 2018, pp. 4779–4783.

[17] A. Van Den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu, "Wavenet: A generative model for raw audio," in *9th ISCA Speech Synthesis Workshop*, 2016, pp. 125–125.

[18] Y. Ning, S. He, Z. Wu *et al.*, "A review of deep learning based speech synthesis," *Applied Sciences*, vol. 9, no. 19, p. 4050, 2019.

[19] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," in *International Conference on Learning Representations (ICLR)*, 2015.

[20] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, pp. 5998–6008, 2017, accessed: 2025-10-19. [Online]. Available: https://proceedings.neurips.cc/paper/7181-attention-is-all-you-need.pdf

[21] M. Johnson *et al.*, "Google's multilingual neural machine translation system: Enabling zero-shot translation," *Transactions of the Association for Computational Linguistics*, vol. 5, pp. 339–351, 2017.

[22] R. Aharoni, M. Johnson, and O. Firat, "Massively multilingual neural machine translation in the wild: Findings and challenges," in *Proc. the 57th Annual Meeting of the Association for Computational Linguistics*, 2019, pp. 3266–3280.

[23] B. Zhang, B. Haddow, and A. Birch, "Prompting large language models for zero-shot domain adaptation in machine translation," *arXiv:2305.14228*, 2023.

[24] D. Varga, A. Kornai, V. Trón, G. Vámos, B. Nagy, L. Németh, and V. Makrai, "Parallel corpora for medium density languages," in *Recent Advances in Natural Language Processing IV: Selected Papers from RANLP 2005*, ser. Current Issues in Linguistic Theory. John Benjamins Publishing Company, 2007, pp. 247–258.

[25] F. Feng, Y. Yang, D. Cer, N. Arivazhagan, and W. Wang, "Language-agnostic bert sentence embedding," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2022, pp. 878–887.

[26] B. Thompson and P. Koehn, "Vecalign: Improved sentence alignment in linear time and space," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Hong Kong, China: Association for Computational Linguistics, Nov. 01 2019, pp. 1342–1348, accessed: 2025-10-19. [Online]. Available: https://aclanthology.org/D19-1136/

[27] C. Tran, Y. Tang, X. Huang, J. Qin, F. Schneider, V. Guruswami, J. Liu, and W. Chen, "Cross-lingual retrieval for iterative self-supervised training," in *Advances in Neural Information Processing Systems*, vol. 33, 2020, pp. 12 749–12 762.

[28] K. Heffernan, S. Joty, T. Le, E. Safi-Samghabadi, X. Ma, G. Kallidromitis, M. Burke, A. Bhargav, J. Y. Zou, J. D. Choi *et al.*, "Bitext mining using distilled sentence representations for low-resource languages," in *Proc. the 2022 Conference on Empirical Methods in Natural Language Processing*, 2022, pp. 7275–7290.

[29] X. Liang, Y.-M. J. Khaw, S.-Y. Liew, T.-P. Tan, and D. Qin, "Multi-lingual sentence alignment with gpt models," in *2023 4th International Conference on Artificial Intelligence and Data Sciences (AiDAS)*. IEEE, 2023, pp. 218–223.

[30] Z. Yang, S. Ranathunga, J. Grundy, Z. Zhang, S. Mohottala, and D. Alahakoon, "Cost-effective translation quality evaluation: Custom quality metrics with adaptive sampling," in *Proc. the 2023 Conference on Empirical Methods in Natural Language Processing*, 2023, pp. 6123–6137.

[31] P. Joshi, S. Santy, A. Budhiraja, K. Bali, and M. Choudhury, "The state and fate of linguistic diversity and inclusion in the nlp world," in *Proc. the 58th Annual Meeting of the Association for Computational Linguistics*, 2020, pp. 6282–6293.

[32] J. Kreutzer, I. Caswell, L. Wang, A. Wahab, D. v. Lent, D. Roth, M. Choudhury, K. Robinson, A. Bapna, A. Siddhant, O. Firat, S. Rothe, M. R. Costa-jussa, L. Bentivogli, N. Bertoldi, M. Federico, F. Blain, P. Koehn, K. Duh, N. Ruiz, V. Gangal, and A. Finch, "Quality at a glance: An audit of web-crawled multilingual datasets," *Transactions of the Association for Computational Linguistics*, vol. 10, pp. 50–72, 2022. [Online]. Available: https://aclanthology.org/2022.tacl-1.4/

[33] E. M. Bender and B. Friedman, "Data statements for natural language processing: Toward mitigating system bias and enabling better science," in *Transactions of the Association for Computational Linguistics*, vol. 6, 2018, pp. 587–604.

[34] F. J. Och and H. Ney, "A systematic comparison of various statistical alignment models," *Computational Linguistics*, vol. 29, no. 1, pp. 19–51, 2003.

[35] N. Houlsby, A. Giurgiu, S. Jastrzebski, B. Morrone, Q. De Laroussilhe, A. Gesmundo, M. Attariyan, and S. Gelly, "Parameter-efficient transfer learning for nlp," in *Proc. 36th International Conference on Machine Learning (ICML)*. PMLR, 2019, pp. 2790–2799.

[36] X. L. Li and P. Liang, "Prefix-tuning: Optimizing continuous prompts for generation," in *Proc. 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (ACL-IJCNLP)*, 2021, pp. 4582–4597.

[37] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, "Lora: Low-rank adaptation of large language models," in *Proc. 10th International Conference on Learning Representations (ICLR)*, 2022.

[38] A. Vaswani, S. Bengio, E. Brevdo, F. Chollet, A. N. Gomez, S. Gouws, L. Jones, L. Kaiser, N. Kalchbrenner, N. Parmar, R. Sepassi, N. Shazeer, and J. Uszkoreit, "Tensor2tensor for neural machine translation," in *Proc. the 13th Conference of the Association for Machine Translation in the Americas (Volume 1: Research Papers)*, Boston, MA, 2018, pp. 193–199.

[39] P. Brown, J. C. Lai, and R. E. Mercer, "Aligning sentences in parallel corpora," in *29th Annual Meeting of the Association for Computational Linguistics*, 1991, pp. 169–176.

[40] H. Khayrallah and P. Koehn, "On the impact of various types of noise on neural machine translation," in *Proc. the 2nd Workshop on Neural Machine Translation and Generation*, 2018, pp. 74–83.

[41] OpenAI, "Gpt-4 technical report," in *arXiv preprint arXiv:2303.08774*, 2023. [Online]. Available: https://arxiv.org/abs/2303.08774

[42] D. Ranathunga, S. Farhath, U. Thayasivam, S. Jayasena, and G. Dias, "Neural machine translation for low-resource languages: A survey," *ACM Computing Surveys*, vol. 55, no. 2, pp. 1–37, 2022.

[43] H. Schwenk, G. Wenzek, S. Edunov, E. Grave, and A. Joulin, "Ccmatrix: Mining billions of high-quality parallel sentences on the web," in *Proc. the 59th Annual Meeting of the Association for Computational Linguistics*, 2021, pp. 6490–6500.

[44] P. Lison and J. Tiedemann, "Opensubtitles2016: Extracting large parallel corpora from movie and tv subtitles," in *Proc. the Tenth International Conference on Language Resources and Evaluation*, 2016, pp. 923–929.

[45] M. Freitag, R. Rei, N. Ruiz, C.-k. Lo, C. Stewart, G. Foster, A. Lavie, and O. Bojar, "Results of the wmt22 metrics shared task," in *Proc. the Seventh Conference on Machine Translation*, 2022, pp. 431–461.

[46] K. Ahuja, L. Zhang, O. Ram, S. Goenka, Y. Zhang, C. Feichtenhofer, S. Bhosale, V. Chaudhary, A. Fan, E. Grave, and H. Schwenk, "Mega: Multilingual evaluation of generative ai," in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP 2023)*. Singapore: Association for Computational Linguistics, 2023, pp. 7693–7711.

[47] B. Intrator, Y. Belinkov, and R. Schwartz, "Breaking down low-resource adaptation in neural machine translation," *Transactions of the Association for Computational Linguistics*, 2024.

[48] J. Xu, Y. Meng, Y. Zhang, Z. Wang, L. Wu, R. Schwartz, and L. Li, "Contrastive language-image pretraining for low-resource speech translation," *arXiv:2405.10223*, 2024.

[49] E. B. Zaken, S. Ravfogel, and Y. Goldberg, "Bitfit: Simple parameter-efficient fine-tuning for transformer-based masked language models," in *Proc. the 59th Annual Meeting of the Association for Computational Linguistics (Findings)*, 2021, pp. 1–16.

[50] X. Liang, Y.-M. J. Khaw, S.-Y. Liew, T.-P. Tan, and D. Qin, "Efficient Chinese–Malay speech-text translation via layer-freezing adaptation of multimodal foundation models," *IEEE Access*, vol. 13, 2025, early Access; Accessed: 2025-10-19. [Online]. Available: https://doi.org/10.1109/ACCESS.2025.3568474