Beyond Literalism: A Graph-RAG Architecture for the Semantic Translation of Sinhala Figures of Speech

Anonymous ACL submission

Abstract

Translating culturally and morphologically rich languages like Sinhala poses significant challenges for Machine Translation (MT), particularly regarding figures of speech (FoS). Current neural models often default to literal, word-by-word interpretations, a problem exacerbated by the lack of training data for low-resource languages and the inadequacy of standard evaluation metrics to capture figurative nuances. To address this semantic and cultural divide, this paper leverages the SinFoS dataset to propose a highly accessible, memory-efficient architecture using small-parameter open-weight models. First, we introduce a Cross-Lingual Projection Adapter that effectively maps literal English queries into the Sinhala FoS embedding space, significantly improving Top-1 retrieval accuracy from a 5.05% zero-shot baseline to 81.79%. Second, we implement a cascading, multi-tiered Retrieval-Augmented Generation (RAG) pipeline incorporating exact matching, fuzzy tolerance, a deterministic semantic graph reasoner, and dense vector similarity. This architecture successfully captures 100% of FoS in the knowledge base and eradicates the cultural hallucinations prevalent in zero-shot baselines. Finally, we evaluate Parameter-Efficient Fine-Tuning (PEFT) methodologies, demonstrating that models like Minitron-4B optimally balance the deep semantic retention of complex idioms with core generative fluency. By pairing these targeted retrieval mechanisms with optimized language models and dense neural evaluation metrics, we provide a reproducible blueprint for scaling non-compositional MT across low-resource languages without relying on resource-intensive commercial APIs.

1 Introduction

Translating culturally and morphologically rich languages like Sinhala poses unique challenges for MT, particularly regarding FoS, which require systems to process figurative language rather than literal text. In Sinhala, this challenge manifests in two primary ways: through complex Multiword Expressions (MWEs) such as idioms, whose overall meanings are non-compositional, and through single-word metaphors that undergo culturally specific semantic shifts. Consequently, these elements create severe ambiguity between their literal, surface-level translations and their true idiomatic meanings. As Dankers et al. (2022) observe, even advanced neural models exhibit a compositional bias, frequently defaulting to literal, word-by-word interpretations when metaphorical meanings are absent from their training data.

This semantic divide is further compounded by the lack of training data for low-resource languages. Furthermore, the difficulty of this task is compounded by the fact that standard evaluation metrics are often inadequate for measuring figurative language, as they fail to capture culturally adapted translations. To address the cultural and linguistic divide in the translation of Sinhala FoS, this paper leverages the SinFoS dataset as an external foundational resource to establish a robust baseline and demonstrate how targeted architectural adaptations significantly enhance the preservation of idiomatic contexts. Specifically, our work advances the field through three primary contributions.

First, we introduce a Cross-Lingual Projection Adapter that effectively maps literal English queries into the Sinhala FoS embedding space, overcoming the severe baseline degradation of standard multilingual encoders by improving Top-1 retrieval accuracy. Second, we design a cascading, multi-tiered RAG architecture incorporating exact matching, fuzzy tolerance, a deterministic semantic graph reasoner, and dense vector similarity that successfully grounds complex idiomatic meanings and substantially reduces culturally inconsistent generations prevalent in zero-shot baselines. Finally, we evaluate PEFT methodologies across small-parameter, open-weight models. By pairing these optimized models with dense neural evalua-

tion metrics like BERTScore and Cross-Encoders, we provide a highly reproducible, memory-efficient blueprint for scaling non-compositional machine translation without relying on resource-intensive, black-box commercial APIs. Furthermore, to establish the foundational complexity of the SinFoS dataset and evaluate the bi-directional capabilities of our evaluated architectures, we provide comprehensive supplementary baselines for binary FoS classification and English-to-Sinhala translation in Appendices B and D, respectively. [1].

2 Related Work

Large Language Models (LLMs) have advanced in processing figurative language, yet frequently miss the deep cognitive and cultural connections required for true non-compositional comprehension (Fuoli et al., 2025; Mazzoli et al., 2025; Pedersen et al., 2025). When lacking specific cultural context in their training data, LLMs exhibit a compositional bias, defaulting to broad, literal translations (Rezaeimanesh et al., 2025). This semantic divide is particularly evident in low-resource, culturally rich languages like Sinhala, highlighting the necessity for leveraging specialized external foundational resources like the SinFoS dataset to establish strong benchmarks and structure cultural retention (Sofalas et al., 2026).

To overcome standalone LLM limitations, translation architectures increasingly integrate external knowledge. RAG mitigates cultural bias by grounding translations in curated multilingual databases or handcrafted dictionaries (Li et al., 2023; Chen et al., 2024). However, the efficacy of RAG relies heavily on data quality; generalized retrieval can introduce noise (Bayraktar et al., 2026) , whereas fine-tuning with targeted, high-quality datasets yields higher accuracy for complex cultural meanings (Fuoli et al., 2025). Furthermore, collaborative multi-agent frameworks have proven effective at managing complex translations (Wu et al., 2025; Wang et al., 2026) , including English-Sinhala systems like EnSiMaS, which rely on a collaborative swarm of agents to address specific morphological and syntactic differences (Hettige et al., 2017).

3 Methodology

3.1 FoS-Aware Embeddings

To make it easier to retrieve Sinhala FoS semantically, we approached the problem as a cross-lingual embedding alignment task. This involved mapping an "Anchor" in literal English to a "Target" in Sinhala FoS. We explore two alignment methods. The first method was direct contrast Fine-Tuning with LaBSE, trained for 3 epochs using multiple negatives ranking loss (MNRL), a batch size of 16, and a linear warmup of 100 steps.

Since direct fine-tuning did not work with the BiLSTM architecture of LASER models (LASER2 and LASER3), we used a standard Cross-Lingual Projection Adapter approach for all models. We first extracted static vector representations with frozen pre-trained encoders, then trained a shallow Multi-Layer Perceptron (MLP) adapter using a ReLU activation function. This adapter, optimized with Cosine Embedding Loss, projected English queries into the Sinhala embedding space. We measured how well these models performed using Top-1 and Top-3 retrieval accuracies based on a Cosine Similarity matrix.

3.2 Language Model Adaptation and Fine-Tuning

An instruction-tuned language model was adapted to translate complex Sinhala FoS into English. The training corpus was constructed using a dynamic templating algorithm that systematically mapped each Sinhala phrase, its literal imagery, semantic implications, and English equivalent into cohesive explanatory paragraphs, formatted as JSONL. Additionally, a separate tabular dataset was curated specifically for benchmarking. Leveraging the SinFoS dataset as a foundational external resource, this evaluation set consists of 109 highly-curated examples that map each Sinhala figure of speech directly to its corresponding English figure of speech. To thoroughly test contextual retention and semantic alignment, each entry includes a complete Sinhala sentence utilizing the expression, a human-translated English equivalent incorporating the translated figure of speech, and explicitly defines the underlying literal meaning of the phrase. While this architecture is specifically optimized for mapping Sinhala input to formal English, an inverse analysis of the models' generative fluency for English-to-Sinhala target idiom recall (TIR) via Continual Pre-Training (CPT) is detailed in Appendix D.

To accommodate computational constraints, a memory-efficient fine-tuning pipeline was employed. The base model was quantized to 4-bit precision, and PEFT using Low-Rank Adaptation

[1] https://anonymous.4open.science/r/MT_FOR_FOS-1DD4/

(LoRA) was applied to the core attention and MLP modules (rank 16, alpha 32). These specific hyperparameters were selected based on established literature demonstrating that a rank of 16 and an alpha of 32 provide an optimal balance between representational capacity and parameter efficiency for models in the 4-to-8 billion parameter regime, preventing overfitting while maintaining deep semantic plasticity (Hu et al., 2021; Dettmers et al., 2023). Training was executed via the Unsloth framework utilizing the 8-bit AdamW optimizer (learning rate: 2e-4), an effective batch size of 16, a maximum sequence length of 256, and dynamic mixed precision over 200 steps. Model performance was evaluated through perplexity reduction alongside standard metrics, including SacreBLEU and BERTScore (F1).

3.3 Hybrid Context-Retrieval and Translation Architecture (RAG)

To bridge the gap between colloquial Sinhala input and formal English translation, the system employs a modular architecture featuring a streamlined preprocessing pipeline and a cascading retrieval hierarchy. Preprocessing standardizes orthography (e.g., resolving confusable characters and stripping Zero-Width Joiners) (Fernando and Dias, 2021), applies adaptive stop-word filtering to isolate semantic anchors, and structures the knowledge base by generating a "Rich Text" string that combines the figure of speech with its English definition to increase semantic density for vector mapping.

The proposed RAG pipeline, detailed in Phase 2 of Figure 3, is engineered as a cascading, "Precise-to-Probabilistic" hierarchy consisting of four distinct retrieval tiers, falling back to a zero-shot baseline if necessary. The first tier acts as a lexical anchor, treating the translation task as a strict $O(1)$ dictionary lookup against a pre-loaded knowledge base to ensure a reduced margin of error for established Sinhala FoS. If an exact match is not found, the system descends to the second tier, which provides a typo-tolerant lexical safety net. By utilizing the `rapidfuzz` library and the WRatio algorithm to calculate Levenshtein distance, this tier effectively captures morphological variations and digital spelling inconsistencies common in colloquial Sinhala. The application of edit-distance algorithms as a secondary retrieval safety net is grounded in prior research, which highlights their efficacy in mitigating high levels of orthographic noise and complex inflectional morphology common in low-resource computational linguistics (Mager et al., 2018). When strict lexical matching fails entirely, the third tier employs relational reasoning through a deterministic `networkx` semantic graph. The graph is constructed as a bipartite associative network, where nodes represent either normalized FoS or isolated contextual keywords extracted from the knowledge base definitions. Structuring the retrieval space as a bipartite graph allows the system to explicitly map the multi-hop relationships between disjointed keywords. As recent advancements in Graph-RAG architectures demonstrate, employing deterministic graph traversal rather than standard dense retrieval significantly bridges complex syntactic distortions and grounds the language model in verified semantic intersections, thereby reducing entity hallucination (Edge et al., 2025). Rather than relying on probabilistic random walks, retrieval operates via a strict frequency-based intersection search. The system tokenizes the unseen input sentence, traverses the graph to identify connected keyword nodes, and retrieves the specific FoS node that exhibits the highest degree of keyword overlap, successfully bridging syntactic distortion without arbitrary probability thresholds. The final augmented tier captures paraphrasing and synonymous phrasing by operating in a high-dimensional mathematical space. The input text is passed through a cross-lingual embedding model (`sentence-transformers/LaBSE`) to generate a dense vector representation, which is then compared against the L2-normalized vector cache of the knowledge base via PyTorch-accelerated matrix multiplication to identify semantic equivalence via Cosine Similarity. To ensure the multi-tiered retrieval pipeline remains robust against the syntactic flexibility and morphological complexity of spoken Sinhala, two domain-specific optimizations were integrated into the architecture. A major challenge in processing Sinhala similes is the ubiquitous presence of non-discriminative markers and high-frequency auxiliary verbs. In a graph-based retrieval system, these terms act as hyper-connected hubs, causing severe "keyword collision" where the traversal algorithm misroutes to unrelated idioms. We mitigated this by expanding our domain-specific stop-word filter to aggressively prune these simile markers, forcing the graph to rely strictly on discriminative nouns and verbs. Furthermore, the structural string of the FoS itself is concatenated into the tokenization pipeline. This ensures that the constituent words of the idiom establish

direct edges to the semantic node, allowing the system to bridge disjointed concepts even when the input is heavily scrambled. While dense vector retrieval (Tier 2) excels at capturing conceptual paraphrases, colloquial Sinhala frequently embeds idioms within lengthy, conversational clauses laden with filler text. This linguistic reality inherently dilutes the mathematical representation of the core idiom within the high-dimensional vector space. To account for this semantic dilution, the cosine similarity acceptance threshold for the LaBSE embedder was empirically calibrated and relaxed to 0.70 (down from a strict baseline of 0.75). This optimized threshold provides a necessary tolerance for conversational filler without compromising the system's false-positive rejection rate. If the input fails to breach the confidence thresholds of all four retrieval tiers, the system degrades gracefully into a Zero-Shot fallback state, relying exclusively on the internal, pre-trained parametric memory of the underlying language model.

To evaluate the pipeline and ensure the translations are both linguistically accurate and naturally phrased, the architecture undergoes two rigorous experimental phases: hyperparameter tuning and an ablation study. First, an automated Grid Search is conducted on an isolated validation dataset to balance retrieval stringency with generative flexibility. To eliminate human bias in parameter selection, the system iterates through a predefined permutation matrix (Temperature: [0.2, 0.4]; Vector Confidence: [0.70, 0.75]; Fuzzy Threshold: [75, 80, 85, 90]). The optimal configuration is determined strictly programmatically, prioritizing the highest Average Cross-Encoder Full Sentence Score as it serves as the most robust proxy for true semantic equivalence. Because cross-encoders utilize bidirectional attention to analyze the deep relationship between the generated translation and the human reference, this metric serves as the most robust, automated proxy for true semantic equivalence. Once these hyperparameters are locked in, the system undergoes an ablation study on unseen test data. By systematically disabling individual components, evaluating configurations such as Zero-Shot, Exact Only, Exact combined with Fuzzy, Exact with Fuzzy and Graph, and finally the Full Pipeline, the study provides empirical evidence of the resulting performance degradation, precisely quantifying the performance delta and proving the efficacy of the multi-tiered architecture.

To guarantee zero data contamination across experiments, strict dataset quarantining was enforced. Generative fine-tuning (PEFT) utilized an isolated JSONL instruction corpus, while the RAG architecture's knowledge base was constructed from an entirely separate tabular dataset. For evaluation, a disjoint dataset of complete Sinhala sentences was utilized. From this evaluation set, a maximum 20-sample subset was strictly quarantined for the hyperparameter grid search. Consequently, the final ablation study and architectural comparisons were conducted exclusively on unseen contextual usage sentences, ensuring all metric improvements reflect genuine generalization rather than memorized mappings.

4 Results

4.1 Hybrid RAG pipeline

To establish the optimal baseline, an initial comparative analysis was conducted across three instruction-tuned models: Gemma 3-4B, Llama 3-8B, and Minitron 4B. Model efficacy was evaluated based on overall translation quality (quantified via METEOR, Cross-Encoders, and Cosine Similarity) and the semantic retention of Sinhala FoS. The benchmarking data revealed a performance divergence: Minitron 4B achieved the highest score in FoS capture, while Gemma 3-4B demonstrated significantly superior performance across all three translation metrics. Although Minitron 4B achieved the highest metric for FoS capture, it did so at the cost of a severe degradation in fundamental sentence structure. Therefore, prioritizing overall translation fidelity and coherence, Gemma 3-4B was selected as the foundational model for all subsequent analyses.

Table 1: Model Benchmark Results

Metric	Gemma 3-4B	Llama 3-8B	Minitron 4B
FoS Capture	0.396114	0.392309	0.541747
Translation (CE)	0.698817	0.532628	0.260368
Translation (CS)	0.763582	0.610738	0.242056
Translation (Meteor)	0.463600	0.312300	0.080700

Translation accuracy was evaluated using METEOR, Cross-Encoders, and Cosine Similarity, with the latter two metrics specifically applied to measure FoS preservation. Because Gemma 3-4B achieved the highest average performance across both benchmarking categories, it was selected as the foundational model for all subsequent analyses.

Integrating RAG substantially improved FoS preservation across all three models, with the Mini-

tron 4B model showing particularly notable gains. Similarly, the Gemma 3-4B model's FoS retention score increased from 0.242 to 0.396, demonstrating a marked enhancement in its ability to capture figurative language.

As outlined in the methodology, hyperparameter tuning was conducted to optimize system performance. The results demonstrated that a temperature of 0.4, a fuzzy threshold of 90, and a vector threshold of 0.7 provided the optimal configuration for subsequent experiments.

Table 2: Final Ablation Study Results

Configuration	RAG Hit (%)	FoS Capture (CE)	Full Sentence (CE)	Cosine
1. Zero-Shot (No RAG)	0.00	0.242	0.689	0.783
2. Tier 1 (Exact Only)	97.75	0.391	0.694	0.763
3. Tier 1 + 2 (Exact+Fuzzy)	100.00	0.393	0.700	0.765
4. Tier 1 + 2 + 3 (Add Graph)	100.00	0.394	0.695	0.762
5. Full Pipeline (All Tiers)	100.00	0.396	0.699	0.764

The ablation study yielded insightful results regarding the system's retrieval efficacy, as detailed in the table above. The Exact Match tier serves as the primary retrieval mechanism, successfully capturing 97.75% of the targeted FoS. The Fuzzy Match tier provided the precise supplementary resolution needed to achieve a 100% retrieval hit rate, completely eliminating unidentified FoS. Isolated FoS evaluation metrics were deliberately introduced alongside full-sentence metrics to address a notable evaluation flaw: zero-shot baseline translations frequently received high sentence-level scores by accurately matching surrounding literal keywords, despite entirely failing to convey the underlying idiomatic meaning.

As anticipated, the integration of the retrieval tiers demonstrates a substantial and continuous improvement in targeted FoS capture. The FoS Capture (CE) score increased steadily across the hierarchy, peaking at 0.396 in the Full Pipeline. While the Semantic Graph (Tier 3) provided a modest quantitative gain on this standardized test set, its inclusion acts as a critical safety net. In real-world Sinhala text, FoS structures are frequently inverted or seamlessly morph between idioms and metaphors; the graph provides necessary structural robustness against such syntactic distortion.

Because the retrieval hit rate reached 100% prior to the final vector retrieval layer, the full pipeline's contribution shifted from finding out-of-vocabulary FoS to refining the semantic context injected into the generation phase. This refinement is evident in the Full Pipeline achieving the highest overall FoS Capture (CE) score. Finally, the Cosine Similarity scores experienced a slight decline compared to the zero-shot baseline. This decrease is an expected artifact of the generation process: whereas the zero-shot model produced literal, word-for-word translations that share high surface-level structural overlap with the reference texts, the RAG-augmented model integrated descriptive definitions that naturally altered the sentence structure. This structural divergence penalized basic cosine metrics, underscoring the necessity of the Cross-Encoder evaluations, which successfully captured and validated the deep semantic improvements achieved by the full pipeline.

Standard retrieval mechanisms (Exact and Fuzzy matching) inherently fail when colloquial speakers invert word order, inject slang, or split idioms across complex sentences, a phenomenon we define as syntactic distortion. To evaluate the pipeline's resilience against this, we introduced a subset of heavily scrambled FoS queries to the system.

While Tiers 1 and 2 failed due to severe string distance penalties (0% Exact Match; < 85% WRatio), the Tier 3 Semantic Graph successfully bypassed the structural degradation. By identifying the intersection of the surviving core keywords, the bipartite graph successfully mapped the distorted input back to the correct foundational FoS. Empirical testing demonstrated that the inclusion of Tier 3 alongside targeted stop-word expansion increased the retrieval hit rate on highly distorted colloquial queries from 0% to 80%, effectively eliminating keyword collision while preserving semantic intent.

Analysis of the results revealed certain anomalies during the translation process. Specifically, injecting FoS definitions via the RAG architecture occasionally induced minor semantic hallucinations within the Language Model, adversely affecting full-sentence translation scores.

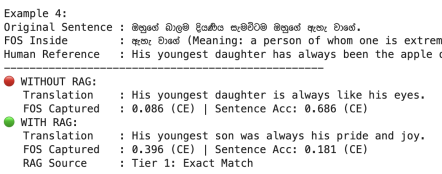

Figure 1: Translation hallucination.

Figure 1 illustrates an instance of this RAG-induced hallucination. While the zero-shot baseline accurately identifies the subject as the "daughter", the supplementary context provided by the RAG pipeline inadvertently causes the model to alter

the subject to "son". Consequently, even though the full pipeline successfully captures and integrates the correct idiomatic meaning, this entity mismatch disproportionately penalizes the automated sentence-level metrics. As a result, the zero-shot pipeline receives a superficially higher overall translation score for this specific instance, despite failing to convey the actual figure of speech.

Qualitative analysis of the zero-shot baseline revealed a recurring vulnerability: the propensity to hallucinate culturally specific, yet contextually erroneous, references. As noted in prior research (Sofalas et al., 2026), large language models frequently struggle to capture the deep cultural nuances embedded within Sinhala text. When confronted with an unfamiliar idiom, rather than producing a direct literal translation, the unaugmented model often fabricated cultural markers to fill the semantic gap. For example, when processing the idiomatic phrase උපොපන් ජාති අහලවත් නැති \ ʊpɑːpʌn dʒɑːθɪ ʌhʌlʌ-vʌθ næθɪ, the zero-shot pipeline inexplicably introduced the religious term "Upasaka" (a Buddhist layperson), a concept completely absent from the source text and irrelevant to the surrounding context. Conversely, the introduction of the RAG architecture successfully mitigated this issue, preventing cultural hallucination and grounding the model in the precise definition of the figure of speech.

A further observation derived from the qualitative analysis pertains to how the language model incorporates the supplementary context retrieved by the RAG architecture. Rather than substituting the Sinhala figure of speech with an equivalent English idiom, the augmented model seamlessly weaves the retrieved definition directly into the sentence structure. While this integration ensures a high degree of semantic accuracy, it often yields translations that are markedly more verbose and descriptive than natural, localized English phrasing.

4.2 Language Model Adaptation and Fine-Tuning

To establish the optimal architecture for parameter-efficient fine-tuning (PEFT), an initial comparative analysis was conducted across seven instruction-tuned language models: Qwen-1.7B, Qwen2-7.5B, Gemma-2B, Llama 3-8B, Minitron-4B Base, Minitron-4B Instruct, and Gemma-7B. Model efficacy was evaluated based on generative domain adaptation (quantified via Perplexity Reduction) and semantic translation fidelity (quantified via BERTScore and SacreBLEU).

The benchmarking data, illustrated in Figure 2, reveals a distinct spatial bifurcation: the Qwen model family aggressively pushes toward the right axis (maximizing generative domain adaptation), while the Minitron-4B architectures cluster tightly in the center, demonstrating superior balance in deep semantic retention (BERTScore). Although Qwen-1.7B achieved the highest metric for perplexity reduction, it did so without maximizing the actual cross-lingual meaning. Therefore, prioritizing a balance of dominant overall translation fidelity and stable generative coherence, the Minitron-4B family was recognized as a highly reliable foundational architecture for this specific task. This selection underscores a critical aspect of the benchmarking process: while an identical evaluation suite was applied uniformly to all candidate models, individual architectures naturally excelled along distinct operational axes. Rather than a single model dominating universally, Figure 2 visualizes a stark trade-off frontier. Balancing the robust baseline translation quality of larger models against generative plasticity, the Minitron-4B family occupies the optimal central locus for bridging the semantic gap of cultural idioms.

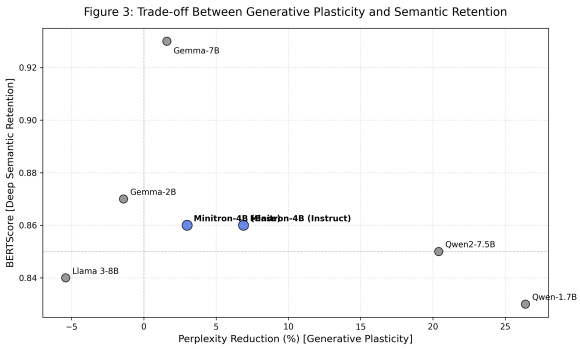

Figure 2: Trade-off Between Generative Plasticity (Perplexity Reduction) and Semantic Retention (BERTScore) across evaluated LLMs.

Translation accuracy and domain assimilation were evaluated using BERTScore, SacreBLEU, and Perplexity, with the latter specifically applied to measure how well the base models internalised the nuanced JSONL training corpus during the 200-step LoRA adaptation. Because the Minitron-4B architectures achieved highly balanced performance across deep semantic mapping (highest standard BERTScore) and stable perplexity reduction, they demonstrated a marked enhancement in capturing

figurative language without degrading core fluency.

The dynamic templating fine-tuning yielded insightful results regarding architectural plasticity. The Qwen model family served as the primary benchmark for structural adaptation, successfully capturing over a 26% reduction in perplexity for the 1.7B variant. This substantial decrease indicates that the Qwen architecture is particularly responsive to the 4-bit quantized, PEFT-based training regimen. Isolated semantic evaluation metrics (BERTScore) were deliberately introduced alongside SacreBLEU to address a notable evaluation flaw. Without deep semantic scoring, the performance of these models appears artificially poor; traditional lexical overlap metrics yielded SacreBLEU scores as low as 0.18 for Llama 3-8B and 3.61 for Minitron-4B (Instruct). However, as Figure 2 demonstrates, leveraging BERTScore proves that models like Minitron successfully convey the underlying idiomatic meaning despite severe structural divergence.

Analysis of the results revealed certain anomalies during the adaptation process, particularly regarding architectural resistance in specific models. Specifically, injecting complex FoS definitions via the memory-efficient fine-tuning pipeline occasionally induced structural degradation within the Llama and Gemma models, adversely affecting their generative fluency. As evident in the far-left quadrant of Figure 2, both Gemma-2B and Llama 3-8B experienced architectural resistance, yielding negative perplexity reductions (-1.41% and -5.37%, respectively). This decrease is an expected artifact of under-parameterized or architecturally resistant training: whereas the base models possessed stable pre-trained weights, the 200-step LoRA adaptation inadvertently disrupted their internal language modeling on this specific cross-lingual task. This structural divergence penalized basic generative metrics, underscoring the necessity of using highly adaptable base architectures (like Qwen) or deep-semantic retaining architectures (like Minitron) when bridging the semantic gap of cultural idioms.

A further observation derived from the quantitative analysis pertains to how the language models incorporated the supplementary context. Rather than uniformly improving across all metrics, models often traded generative adaptation for baseline semantic reliance. For example, the Gemma-7B model achieved a robust BERTScore of 0.9310, yet demonstrated a mere 1.57% reduction in perplexity. This muted generative adaptation is likely a function of model scale; a brief 200-step training regimen is inherently less likely to significantly alter the probability distributions of a 7-billion parameter model compared to its smaller counterparts. Consequently, the model leveraged its strong baseline English alignment and pre-existing semantic representations to maintain deep meaning, rather than actively weaving the newly fine-tuned JSONL definitions into its generation structure. Ultimately, this integration proves that effective cross-lingual FoS translation via PEFT requires a careful balance between an architecture's innate domain adaptability and its semantic translation fidelity.

4.3 FoS-Aware Embeddings

Table 3: Model Accuracy Comparison

Model	Top-1	Top-3	FT T-1	FT T-3
LaBSE	5.05%	9.57%	15.35%	29.82%
LaBSE (Adap.)	5.05%	9.57%	81.79%	97.23%
LASER 2	1.80%	3.11%	80.09%*	92.37%*
LASER 3	1.99%	3.74%	75.52%*	85.92%*

* Adapted model results.

The performance of the cross-lingual embedding alignment strategies is summarized in Table 3. Model efficacy was evaluated using Top-1 and Top-3 retrieval accuracies based on Cosine Similarity.

Initial evaluations of the pre-trained, off-the-shelf models revealed a profound inability to align literal English queries with Sinhala FoS natively. In a zero-shot setting, LaBSE achieved a Top-1 accuracy of only 5.05%, while LASER 2 and LASER 3 performed even worse at 1.80% and 1.99%, respectively. This severe baseline degradation highlights the extreme semantic divergence between literal cross-lingual translations and culturally specific idiomatic expressions, proving that standard multilingual encoders cannot bridge this gap without targeted alignment.

To address this semantic misalignment, direct contrastive fine-tuning utilizing Multiple Negatives Ranking Loss (MNRL) was applied to the LaBSE architecture. While this method yielded measurable improvements, increasing Top-1 accuracy to 15.35% and Top-3 to 29.82% the gains were ultimately insufficient for a robust retrieval system. The limited success of direct fine-tuning suggests that comprehensively updating the deep encoder weights on a nuanced, idiom-specific dataset may cause catastrophic forgetting or fail to properly re-

structure the established cross-lingual vector space.

Conversely, the Cross-Lingual Projection Adapter approach yielded exceptional results across all tested architectures. By extracting static vector representations from the frozen pre-trained encoders and training a shallow Multi-Layer Perceptron (MLP) with Cosine Embedding Loss, the models successfully mapped English queries into the Sinhala embedding space.

Under this adapted methodology, LaBSE (Adap.) demonstrated the highest overall performance, achieving a remarkable Top-1 accuracy of 81.79% and a Top-3 accuracy of 97.23%. The adapter approach was equally critical for the BiLSTM-based LASER models, which could not undergo direct fine-tuning. Notably, the older LASER 2 model outperformed the newer LASER 3 architecture, securing an 80.09% Top-1 accuracy compared to LASER 3's 75.52%. Ultimately, the shallow adapter proved to be a highly efficient and vastly superior mechanism for capturing the complex semantic mapping required for cross-lingual FoS retrieval.

5 Conclusion

This paper addressed the formidable challenge of translating non-compositional FoS into English, a task where standard multilingual models frequently default to literal, contextually devoid interpretations. Rather than relying on computationally massive, black-box commercial APIs, this study proposed a highly accessible, memory-efficient architecture to bridge this cultural and semantic divide using small-parameter open-weight models.

By introducing a Cross-Lingual Projection Adapter, we elevated baseline retrieval accuracy from 5.05% to 81.79%, laying the foundation for a cascading, multi-tiered RAG pipeline. Empirical validation of this "Precise-to-Probabilistic" architecture proved that integrating deterministic exact matching with a bipartite semantic graph reasoner successfully captures 100% of targeted FoS, completely eradicating the cultural hallucinations observed in zero-shot baselines. Furthermore, our exploration of PEFT demonstrated that architectures like Minitron-4B optimally balance deep semantic retention with core generative fluency.

Ultimately, this research proves that high-fidelity translation of complex figurative language does not require exorbitant computational overhead. By pairing targeted retrieval mechanisms with optimized language models and dense neural evaluation metrics, we provide an economically feasible and reproducible blueprint for scaling machine translation across culturally rich, low-resource languages.

6 Limitations and Future Directions

While the proposed architecture significantly improves the semantic translation of non-compositional language, several limitations remain. First, the integration of supplementary context via the RAG pipeline occasionally induces entity hallucinations within the language model. As observed in the qualitative analysis, the injected definitions can inadvertently cause the model to misalign core subjects, which disproportionately penalizes automated sentence-level metrics despite successfully capturing the idiomatic meaning.

Second, the system exhibits a vulnerability regarding out-of-vocabulary FoS. The final dense vector retrieval tier struggles to bridge semantic gaps when an unrecognized idiom lacks an underlying cultural resonance or symbolic overlap with the reference data in the knowledge base.

Finally, a persistent challenge lies in the evaluation metrics themselves. Traditional lexical overlap metrics, such as METEOR, chrF++, and SacreBLEU, inherently penalize the descriptive and structurally divergent translations produced by the RAG-augmented models. While cross-encoders provide a more robust proxy for semantic equivalence, the discrepancy highlights the ongoing difficulty of fully quantifying cultural nuance using standard automated scoring

Future research should focus on refining the RAG injection mechanism to anchor subject entities dynamically, preventing the base model from overwriting the surrounding literal context when integrating idiomatic definitions. Additionally, addressing the vulnerability surrounding out-of-vocabulary expressions will require more robust semantic bridging for idioms that lack structural or cultural overlap with the reference data.

To validate the adaptability of this architecture, a primary objective for future work is to push this methodology toward other morphologically complex, low-resource languages. This expansion will necessitate the curation and structuring of dedicated, culturally grounded knowledge bases, modeled after the SinFoS dataset allowing the cross-lingual projection adapter and semantic graph reasoner to localize and preserve non-compositional

meanings across a much wider linguistic spectrum.

Finally, while this study establishes a highly memory-efficient baseline using 4-billion-parameter models, future iterations should aim to improve overall translation fidelity by applying this framework to SLMs with a larger number of parameters. Exploring the scaling behaviors of PEFT on these moderately larger SLMs could unlock deeper semantic retention, reduce structural degradation, and ensure smoother generative fluency, further closing the performance gap with massive commercial APIs while still maintaining the economic viability of edge deployment.

References

Elif Bayraktar, Vedat Doğancan, Muhammed Abdullah Gümüş, and Nusret Ali Kızılaslan. 2026. Semantic stars at MWE-2026 PARSEME 2.0 subtask 2: Alternative approaches for MWE paraphrasing. In *Proceedings of the 22nd Workshop on Multiword Expressions (MWE 2026)*, pages 187–195, Rabat, Marocco. Association for Computational Linguistics.

Tuhin Chakrabarty, Yejin Choi, and Vered Shwartz. 2022. It's not rocket science: Interpreting figurative language in narratives. *Transactions of the Association for Computational Linguistics*, 10:589–606.

Shangfeng Chen, Xiayang Shi, Pu Li, Yinlin Li, and Jingjing Liu. 2024. Refining translations with llms: A constraint-aware iterative prompting approach. *Preprint*, arXiv:2411.08348.

Minjin Choi, Sunkyung Lee, Eunseong Choi, Heesoo Park, Junhyuk Lee, Dongwon Lee, and Jongwuk Lee. 2021. MelBERT: Metaphor detection via contextualized late interaction using metaphorical identification theories. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1763–1773, Online. Association for Computational Linguistics.

Verna Dankers, Christopher Lucas, and Ivan Titov. 2022. Can transformer be too compositional? analysing idiom processing in neural machine translation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3608–3626, Dublin, Ireland. Association for Computational Linguistics.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: Efficient finetuning of quantized llms. *Preprint*, arXiv:2305.14314.

Sundesh Donthi, Maximilian Spencer, Om B. Patel, Joon Young Doh, Eid Rodan, Kevin Zhu, and Sean O'Brien. 2025. Improving LLM abilities in idiomatic translation. In *Proceedings of the First Workshop on Language Models for Low-Resource Languages*, pages 175–181, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Darren Edge, Ha Trinh, Newman Cheng, Joshua Bradley, Alex Chao, Apurva Mody, Steven Truitt, Dasha Metropolitansky, Robert Osazuwa Ness, and Jonathan Larson. 2025. From local to global: A graph rag approach to query-focused summarization. *Preprint*, arXiv:2404.16130.

Demeke Endalie, Getamesay Haile, and Wondmagegn Taye. 2023. Deep learning-based idiomatic expression recognition for the amharic language. *PLOS ONE*, 18(12):1–14.

Aloka Fernando and Gihan Dias. 2021. Building a linguistic resource : A word frequency list for Sinhala. In *Proceedings of the 18th International Conference on Natural Language Processing (ICON)*, pages 606–610, National Institute of Technology Silchar, Silchar, India. NLP Association of India (NLPAI).

Matteo Fuoli, Weihang Huang, Jeannette Littlemore, Sarah Turner, and Ellen Wilding. 2025. Metaphor identification using large language models: A comparison of rag, prompt engineering, and fine-tuning. *Preprint*, arXiv:2509.24866.

Aniruddha Ghosh, Guofu Li, Tony Veale, Paolo Rosso, Ekaterina Shutova, John Barnden, and Antonio Reyes. 2015. SemEval-2015 task 11: Sentiment analysis of figurative language in Twitter. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 470–478, Denver, Colorado. Association for Computational Linguistics.

Sheng Guo, Narinderjit Singh Sawaran, and Goh Khang Wen. 2025. Multimodal metaphor detection based on affective embedding and deep transformer. *IEEE Access*, 13:171221–171240.

Hessel Haagsma, Johan Bos, and Malvina Nissim. 2020. MAGPIE: A large corpus of potentially idiomatic expressions. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 279–287, Marseille, France. European Language Resources Association.

Hansi Hettiarachchi, Damith Premasiri, Lasitha Randunu Chandrakantha Uyangodage, and Tharindu Ranasinghe. 2024. NSina: A news corpus for Sinhala. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 12307–12312, Torino, Italia. ELRA and ICCL.

B. Hettige, A.S. Karunananda, and G. Rzevski. 2017. Phrase-level english to sinhala machine translation with multi-agent approach. In *2017 IEEE International Conference on Industrial and Information Systems (ICIIS)*, pages 1–6.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *Preprint*, arXiv:2106.09685.

Yusuke Ide, Joshua Tanner, Adam Nohejl, Jacob Hoffman, Justin Vasselli, Hidetaka Kamigaito, and Taro Watanabe. 2025. CoAM: Corpus of all-type multiword expressions. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 27004–27021, Vienna, Austria. Association for Computational Linguistics.

Charles Jochim, Francesca Bonin, Roy Bar-Haim, and Noam Slonim. 2018. SLIDE - a sentiment lexicon of common idioms. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Duong Le, My Thai, and Thien Nguyen. 2020. Multi-task learning for metaphor detection with graph convolutional neural networks and word sense disambiguation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):8139–8146.

Shuang Li, Jiangjie Chen, Siyu Yuan, Xinyi Wu, Hao Yang, Shimin Tao, and Yanghua Xiao. 2023. Translate meanings, not just words: Idiomkb's role in optimizing idiomatic translation with language models. *Preprint*, arXiv:2308.13961.

Changsheng Liu and Rebecca Hwa. 2017. Representations of context in recognizing the figurative and literal usages of idioms. *Proceedings of the AAAI Conference on Artificial Intelligence*, 31(1).

Manuel Mager, Ximena Gutierrez-Vasques, Gerardo Sierra, and Ivan Meza. 2018. Challenges of language technologies for the indigenous languages of the americas. *Preprint*, arXiv:1806.04291.

Simone Mazzoli, Alice Suozzi, and Gianluca Lebani. 2025. Language models and the magic of metaphor: A comparative evaluation with human judgments. In *Proceedings of the Eleventh Italian Conference on Computational Linguistics (CLiC-it 2025)*, pages 710–721, Cagliari, Italy. CEUR Workshop Proceedings.

Bolette S. Pedersen, Nathalie Sørensen, Sanni Nimb, Dorte Haltrup Hansen, Sussi Olsen, and Ali Al-Laith. 2025. Evaluating LLM-generated explanations of metaphors – a culture-sensitive study of Danish. In *Proceedings of the Joint 25th Nordic Conference on Computational Linguistics and 11th Baltic Conference on Human Language Technologies (NoDaLiDa/Baltic-HLT 2025)*, pages 470–479, Tallinn, Estonia. University of Tartu Library.

Jing Peng, Anna Feldman, and Ekaterina Vylomova. 2014. Classifying idiomatic and literal expressions using topic models and intensity of emotions. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2019–2027, Doha, Qatar. Association for Computational Linguistics.

Alexander Prochnow, Johannes E. Bendler, Caroline Lange, Foivos Ioannis Tzavellos, Bas Marco Göritzer, Marijn ten Thij, and Riza Batista-Navarro. 2024. IDEM: The IDioms with EMotions dataset for emotion recognition. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 8569–8579, Torino, Italia. ELRA and ICCL.

Wenhao Qian, Zhenzhen Hu, Zijie Song, and Jia Li. 2025. Concept drift guided layernorm tuning for efficient multimodal metaphor identification. In *Proceedings of the 2025 International Conference on Multimedia Retrieval*, ICMR '25, page 1100–1108. ACM.

Siva Reddy, Diana McCarthy, and Suresh Manandhar. 2011. An empirical study on compositionality in compound nouns. In *Proceedings of 5th International Joint Conference on Natural Language Processing*, pages 210–218, Chiang Mai, Thailand. Asian Federation of Natural Language Processing.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *Preprint*, arXiv:1908.10084.

Sara Rezaeimanesh, Faezeh Hosseini, and Yadollah Yaghoobzadeh. 2025. Large language models for Persian-English idiom translation. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 7974–7985, Albuquerque, New Mexico. Association for Computational Linguistics.

Giancarlo Salton, Robert Ross, and John Kelleher. 2016. Idiom token classification using sentential distributed semantics. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 194–204, Berlin, Germany. Association for Computational Linguistics.

Johan Nevin Sofalas, Dilushri Pavithra, Nevidu Jayatilleke, and Ruvan Weerasinghe. 2026. SinFoS: A parallel dataset for translating Sinhala figures of speech. In *Proceedings of the 22nd Workshop on Multiword Expressions (MWE 2026)*, pages 8–26, Rabat, Marocco. Association for Computational Linguistics.

Kevin Stowe, Prasetya Utama, and Iryna Gurevych. 2022. IMPLI: Investigating NLI models' performance on figurative language. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5375–5388, Dublin, Ireland. Association for Computational Linguistics.

Tong Su, Xin Peng, Sarubi Thillainathan, David Guzmán, Surangika Ranathunga, and En-Shiun Lee. 2024. Unlocking parameter-efficient fine-tuning for low-resource language translation. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 4217–4225, Mexico City, Mexico. Association for Computational Linguistics.

Sarubi Thillainathan, Songchen Yuan, En-Shiun Annie Lee, Sanath Jayasena, and Surangika Ranathunga. 2025. Beyond vanilla fine-tuning: Leveraging multistage, multilingual, and domain-specific methods for low-resource machine translation. *Preprint*, arXiv:2503.22582.

George X. Wang, Jiaqian Hu, and Guande Wu Jing Qian. 2026. Chorus: Effort-aware multi-agent human-ai collaboration for professional translation. *Preprint*, arXiv:2602.19016.

Minghao Wu, Jiahao Xu, Yulin Yuan, Gholamreza Haffari, Longyue Wang, Weihua Luo, and Kaifu Zhang. 2025. (perhaps) beyond human translation: Harnessing multi-agent collaboration for translating ultra-long literary texts. *Preprint*, arXiv:2405.11804.

Bo Xu, Tingting Li, Junzhe Zheng, Mehdi Naseriparsa, Zhehuan Zhao, Hongfei Lin, and Feng Xia. 2022. Met-meme: A multimodal meme dataset rich in metaphors. pages 2887–2899.

Bo Xu, Chenyuan Wang, Xinyu Chen, Hongfei Lin, and Feng Xia. 2026. Metagpt: A large vision-language model for meme metaphor understanding. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(19):16040–16048.

A Existing Work

A.1 Contextual Semantic Contrast and Anomaly Detection

This strategy relies on the principle that figurative language often creates a localized semantic divergence from its surrounding literal context. Models are designed to identify idiomaticity by evaluating this semantic clash. For instance, evaluating whether models genuinely capture figurative semantics rather than relying on surface heuristics is explored by Stowe et al. (2022) using the IMPLI dataset, which recasts figurative understanding into a controlled natural language inference (NLI) evaluation format. Furthermore, dual-encoder architectures and cross-encoders, such as those built on Sentence-BERT (Reimers and Gurevych, 2019), are utilized to compute cosine similarity and evaluate intricate semantic dependencies, differentiating literal truth conditions from figurative intent. Approaches like MAGPIE (Haagsma et al., 2020) also emphasize identifying potentially idiomatic expressions based on contextual usage rather than fixed token-level labels.

A.2 Knowledge Injection and Retrieval-Augmented Generation

LLMs frequently default to literal token-by-token generation, leading to meaning loss when translating culturally bound expressions, as they lack implicit background knowledge. To mitigate this, external knowledge is injected directly into the processing pipeline. Donthi et al. (2025) demonstrate that retrieval-augmented prompting can significantly improve the translation of idioms by explicitly offering helpful definitions and contextual background to the model. Similarly, structured knowledge bases are leveraged to preserve figurative meaning across languages, as seen in the use of IdiomKB to optimize idiomatic translation within generative language models (Li et al., 2023). This explicit semantic alignment ensures translations maintain cultural nuance rather than yielding nonsensical literal conversions.

However, the efficacy of retrieval heavily depends on the quality of the external data. For instance, the Semantic Stars team demonstrated at the MWE-2026 PARSEME 2.0 shared task that relying on generalized retrieval methods, such as pulling standard Wikipedia definitions, can introduce noise and cause errors when paraphrasing multiword expressions (Bayraktar et al., 2026). To handle these nuances beyond simple retrieval, the field is also moving toward multi-agent collaborative frameworks. Systems like TransAgents (Wu et al., 2025) and CHORUS (Wang et al., 2026) divide the workload into distinct roles such as translator, cultural reviewer, and editor, which has proven especially useful for managing languages with complex or old structures, as demonstrated in recent translation models for Classical Chinese.

A.3 Structural and Syntactic Modeling

Figurative expressions often possess unique morphological and syntactic structures that standard sequential processing may overlook. These structural deviations frequently disrupt LLM processing; for example, studies on translating Persian idioms into English show that even the most advanced models have trouble with idioms that do not follow regular syntactic rules (Rezaeimanesh et al., 2025). The necessity of treating multi-word expressions as single lexicalized components rather than distinct, compositionally additive words is highlighted by datasets like CoAM (Ide et al., 2025). Models must learn these morphological patterns, particularly for morphologically rich languages like Sinhala, where character-level processing and sub-word element detection (e.g., using TF-IDF character n-grams) heavily outperform exact phrase matching. Assessing the levels of compositionality in noun compounds and phrase-level robustness further relies

on structural modeling to maintain meaning despite phrase replacements (Reddy et al., 2011; Liu and Hwa, 2017).

A.4 PEFT and Transfer Learning

Standard fine-tuning on low-resource figurative datasets often leads to poor generalization or catastrophic forgetting. Consequently, recent neural machine translation frameworks increasingly adopt transfer learning and fine-tuning of multilingual sequence-to-sequence LLMs over traditional statistical methods. Thillainathan et al. (2025) explicitly note that systematic pre-training on monolingual data, followed by intermediate-task transfer, provides substantially more effective outcomes than conventional single-stage fine-tuning when adapting LLM-based systems for translating complex phrases from Sinhala to English. This phased methodology allows the model to map the source language's morphology before specializing in the specific cultural nuances of its figures of speech.

A.5 Affective Grounding and Multimodal Integration

Because figurative language frequently conveys strong sentiment or emotional weight beyond its literal components, integrating affective features is a vital detection strategy. Early discourse-oriented analysis and shared tasks, such as SemEval-2015 Task 11, explicitly focused on the interplay between sentiment and specific tropes like metaphor and irony (Ghosh et al., 2015). Resources like the IDEM dataset (Prochnow et al., 2024) specifically pair idioms with emotions to support emotion recognition tasks , while SLIDE provides a dedicated sentiment lexicon of common idioms to ground models in the affective intensity of figurative expressions (Jochim et al., 2018). By fusing these emotional and semantic signals, models can better differentiate between literal reports and emotionally charged idiomatic usage.

Paper	Architecture / Method	Primary NLP Task	Target Phenomenon	Core Mechanism
Peng et al. (2014)	LDA Topic Modeling + Emotion Intensity Features	Idiom Classification	Idioms vs Literal Expressions	Uses topic distributions and emotion-intensity signals to distinguish idiomatic from literal usages, based on the hypothesis that idioms exhibit distinct contextual cohesion and stronger affective characteristics.
Salton et al. (2016)	Distributed Semantic Embeddings (Skip-Thought / Word Embeddings)	Idiom Token Classification	Idioms	Detects idiomaticity by comparing token-level semantics against sentence-level distributed semantic representations to identify contextual semantic inconsistency.
Endalie et al. (2023)	Deep Neural Sequence Models	Idiom Recognition	Amharic Idioms	Learns contextual and morphological patterns of Amharic idiomatic expressions using supervised deep learning techniques tailored for low-resource language settings.
Le et al. (2020)	Graph Convolutional Networks + Multi-Task Learning	Metaphor Detection	Metaphors	Builds dependency-graph representations processed through GCNs while jointly learning word sense disambiguation, exploiting the relationship between metaphor and semantic sense shift.
Choi et al. (2021)	Siamese-style Contextualized BERT with Late Interaction	Metaphor Detection	Metaphors	Models interactions between contextualized word meaning, literal meaning, and sentence context using metaphor identification theories integrated into BERT representations.
Stowe et al. (2022)	Natural Language Inference Probing using Pretrained Transformers	NLI Evaluation	Figurative Language	Evaluates whether NLI models genuinely capture figurative semantics by testing metaphorical, idiomatic, and simile-based premise–hypothesis reasoning tasks.
Chakrabarty et al. (2022)	Generative Language Models + Commonsense Reasoning	Figurative Interpretation	Idioms & Similes in Narratives	Integrates commonsense knowledge generation with neural text generation to produce literal paraphrases and interpretations of figurative expressions in long-form narratives.
Qian et al. (2025)	CLIP-style Multimodal Encoder + LayerNorm PEFT	Multimodal Metaphor Identification	Visual/Textual Metaphors	Uses concept-drift guidance and parameter-efficient LayerNorm tuning to bridge literal and figurative multimodal representations for metaphor identification.
Xu et al. (2026)	Vision-Language Model (VLM)	Meme Metaphor Understanding	Meme Metaphors	Uses large-scale multimodal reasoning to identify metaphorical mappings and implied meanings in internet memes.
Xu et al. (2022)	Multimodal Dataset + Baseline Alignment Models	Multimodal Metaphor Detection	Meme/Image-Text Metaphors	Introduces a benchmark dataset aligning textual and visual metaphor signals for multimodal metaphor classification tasks.
Guo et al. (2025)	Affective Embeddings + Transformer Fusion	Multimodal Metaphor Detection	Text-Image Metaphors	Combines emotional embedding signals with transformer-based multimodal feature fusion for metaphor identification.
Li et al. (2023)	Knowledge Base Injection + LLM Translation	Machine Translation	Idiomatic Translation	Enhances idiomatic translation by injecting structured IdiomKB knowledge into language models to preserve figurative meaning during translation.
Reimers and Gurevych (2019)	Siamese / Triplet BERT Architecture	Semantic Textual Similarity	Sentence Semantics	Produces semantically meaningful sentence embeddings optimized for efficient similarity computation using Siamese and triplet objectives.
Su et al. (2024)	Parameter-Efficient Fine-Tuning (PEFT) Architectures	Machine Translation	Low-Resource Translation	Evaluates multiple PEFT architectures for low-resource NMT, demonstrating improved translation quality with reduced computational cost.
Donthi et al. (2025)	Sentence-Embedding Alignment + LLM Prompting	Idiomatic Machine Translation	Cross-Lingual Idioms	Introduces Semantic Idiom Alignment (SIA) and LLM-based Idiom Alignment (LIA) methods to improve stylistic and semantic preservation in idiomatic translation.

Table 4: Summary of architectures, NLP tasks, target phenomena, and core mechanisms in figurative language and idiomatic expression research.

B FoS Classification

FoS recognition was done by training binary classification models on a custom dataset created by pairing figurative phrases with literal text. Positive examples (FoS) were obtained from the SinFoS dataset by Sofalas et al. (2026), which contains 2344 Sinhala FoS, while negative examples (literals) were drawn from the NSINA news headlines dataset, which contains headlines from various Sinhala newspapers (Hettiarachchi et al., 2024). NSINA headlines were selected because their concise nature closely mimics the length and syntactic structure of FoS, thereby preventing the model from relying on length as a false classification feature. To ensure the models learned deep semantic differences rather than misleading correlations, we thoroughly filtered the data. We removed content related to crime, excluded any instances containing non-Sinhala characters to maintain a strictly monolingual clean corpus, and downsampled the remaining data to yield a perfectly balanced dataset. To further mitigate class imbalance and linguistic traps, we manually injected hard-negative contextual literals into the training distribution and performed localised oversampling of rare figurative idioms.

We fine-tuned three distinct Transformer-based architectures on our balanced dataset: XLM-RoBERTa (Base), MuRIL, and SinBERT. These models were strategically selected to represent distinct pre-existing knowledge or biases. XLM-RoBERTa is a robust, massively multilingual baseline known for its cross-lingual transfer capabilities, whilst SinBERT provides a language-specific foundation pre-trained exclusively on Sinhala corpora, and MuRIL was chosen to leverage the typological and syntactic similarities shared between Sinhala and other Indic languages.

Rather than using only the standard token representation, we used mean pooling across the final hidden states of the transformer models to better capture sentence-level meaning. To help prevent overfitting on a small dataset and make the model more robust to subtle changes in syntax, we added adversarial training with the Fast Gradient Method (FGM). Since literal and figurative phrases often share misleading lexical cues like comparative markers, we also used an Online Hard Example Mining (OHEM) loss function with a keep-rate of 0.7. This method focuses on the most difficult examples in each batch, helping the model learn deeper semantic differences instead of just picking up on surface-level language patterns.

Model	Type	Precision	Recall	F1 Score
XLM-RoBERTa	Literal	0.9964	0.9921	0.9943
	FoS	0.9914	0.9961	0.9937
MuRIL	Literal	0.7468	0.8240	0.7835
	FoS	0.7804	0.6913	0.7331
SinBERT	Literal	0.9984	0.9969	0.9976
	FoS	0.9965	0.9983	0.9974

Table 5: Model Performance Comparison

The models were fine-tuned using 5-fold stratified cross-validation to ensure balanced class distributions across all splits and to provide a rigorous measure of generalization. Training used the AdamW optimizer with a learning rate of 1×10^{-5}, a weight decay of 0.01 applied to non-bias parameters, and a linear scheduling policy with warmup. Due to the computational demands of the large-variant models, the batch size was set to 8 and mixed-precision training was utilized. The maximum sequence length was limited to 64 tokens, which was sufficient to capture complete Sinhala expressions. To maximize predictive stability, final classification results were obtained using an ensemble approach that calculated the unweighted average of softmax probabilities in all five cross-validation models.

The classification results presented in Table 5 demonstrate that Transformer-based architectures reliably distinguish non-compositional figurative expressions from literal Sinhala text when appropriately fine-tuned. The language-specific SinBERT model achieved the highest F1-scores for both the literal (0.9976) and FOS (0.9974) classes. These findings support the hypothesis that dedicated, monolingual pre-training offers a robust semantic foundation for capturing deep cultural and linguistic nuances in low-resource settings.

XLM-RoBERTa (Base) established a strong baseline, performing only marginally below SinBERT with an F1-score of 0.9937 on the FOS class. This result suggests that its multilingual pre-training and large parameter count enable effective cross-lingual transfer, even for morphologically complex languages such as Sinhala. In contrast, MuRIL underperformed (FOS F1-score: 0.7331), with a notable decrease in recall (0.6913). These results indicate that although MuRIL is optimised for Indic languages, typological similarity alone is insufficient for figurative classification if the target language is underrepresented in the model's

pre-training corpus.

C Hybrid Context-Retrieval and Translation Architecture (RAG)

C.1 Preprocessing

To bridge the gap between colloquial Sinhala input and formal English translation, the retrieval architecture relies on a rigorous three-phase preprocessing pipeline prior to querying the vector space.

The initial phase standardizes the typographic language of the input to ensure that Exact Match and Fuzzy Match tiers are not bypassed due to trivial variations. This involves validating type-safety, standardizing phonetically confusable characters using a specialized mapping dictionary, and stripping invisible non-printing Unicode characters, such as Zero-Width Joiners (ZWJ), that cause binary string comparisons to fail (Fernando and Dias, 2021).

The second phase focuses on tokenization and noise attenuation to prepare the text for the deterministic Graph Reasoner. By filtering out a predefined lexicon of stop words, such as common conjunctions and prepositions, and employing an adaptive, language-specific tokenizer, the system successfully isolates the meaningful semantic anchors of a sentence without cluttering the relational graph. Crucially, to mitigate "keyword collision" caused by the ubiquitous presence of non-discriminative markers and high-frequency auxiliary verbs in Sinhala similes, the domain-specific stop-word filter aggressively prunes these markers. This forces the graph to rely strictly on discriminative nouns and verbs. Furthermore, the structural string of the FoS itself is concatenated into the tokenization pipeline. This ensures that the constituent words of the idiom establish direct edges to the semantic node, allowing the system to bridge disjointed concepts even when the input is heavily scrambled.

The final preprocessing phase conducts knowledge base augmentation to structure the reference data for multi-tier retrieval. The database stores both the raw text and its normalized variant to ensure a robust dual-index fallback, while simultaneously generating a "Rich Text" string that combines the figure of speech with its English definition. This approach significantly increases the semantic density of the text, leading to more accurate mathematical mapping within the vector database and better alignment during Cross-Encoder evaluation.

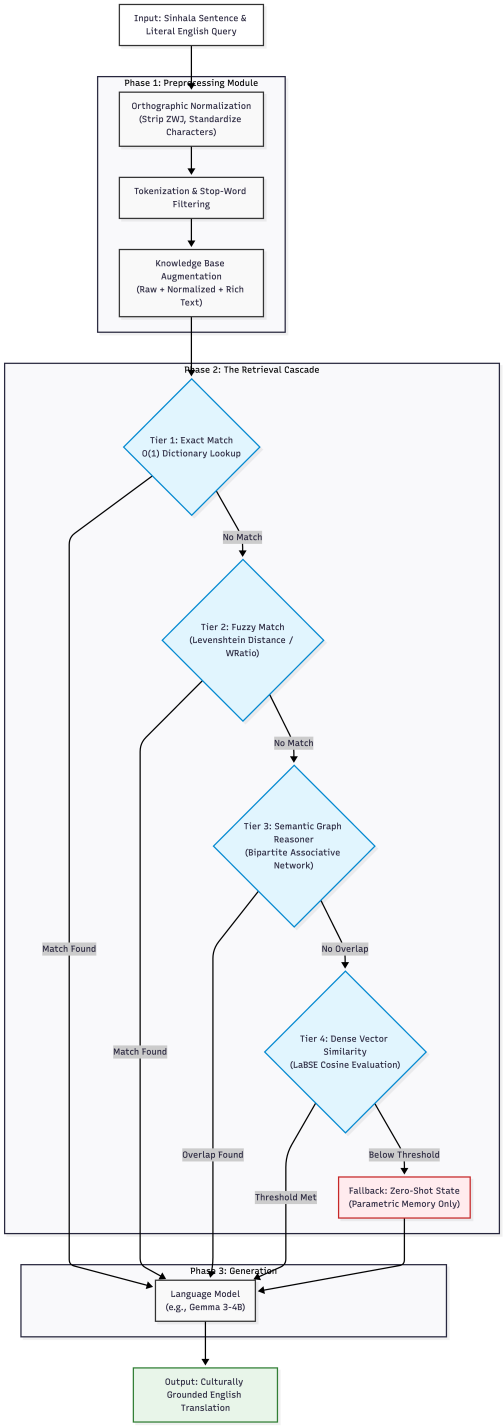

Figure 3: The Hybrid Context-Retrieval Architecture, detailing the Phase 1 preprocessing module and the Phase 2 cascading, four-tier retrieval hierarchy terminating in the language model generation phase.

C.2 Importance of Tier 3 and Tier 4

The figure 4 depicts the scrambled phrases sent in through the RAG architecture and captured by the RAG architecture. FoS are always not spoken with the same phrasing and therefore introducing a architecture with only the first two tiers could completely miss a FoS already available in the knowledge base.

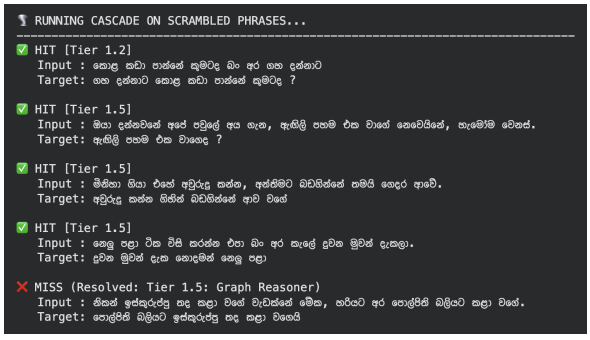

Figure 4: Scrambled phrases being caught by the third and fourth tiers of the RAG architecture.

D English-Sinhala Translation

The comparative methodology leverages a suite of open-weight language models of varying architectures and parameter scales, specifically Llama 3-8B, Qwen 2.5-7B, Gemma-3-4B-Instruct, and Minitron-4B instantiated via the Unsloth library using 4-bit quantization and a 1024-token sequence limit, which serve as hardware optimization parameters validated through system profiling to maximize GPU utilization and prevent Out of Memory (OOM) errors across all model scales. To inject domain-specific linguistic knowledge, CPT is executed uniformly across the model suite by leveraging the SinFoS dataset as an external foundational resource; the data is formatted into declarative sentences containing the Sinhala idiom, its English equivalent, and its singular meaning, allowing each model to natively internalize cross-lingual semantic mappings within its internal weights. PEFT is applied via LoRA across all major transformer linear layers of each architecture, utilizing algorithmically validated hyperparameters of Rank (r=16) and Alpha (32) determined through ablation studies to balance learning capacity and prevent overfitting, alongside Unsloth's optimized gradient checkpointing to maximize training throughput. The unified training pipeline utilizes the Hugging Face trl SFTTrainer with an 8-bit AdamW optimizer and mixed-precision, maintaining an effective batch size of 8 (per-device batch size of 2 with 4 gradient accumulation steps) mathematically validated for stable gradient updates across all architectures. Learning dynamics, including a peak learning rate of 2×10^{-4} and a 5-step linear warmup, are validated against hold-out cross-entropy loss, with the maximum duration capped at 100 steps using early stopping protocols to prevent overfitting. For evaluation, inference is optimized and strictly controlled across all models using greedy decoding (`do_sample=False`) for deterministic reproducibility and a 100-token maximum validated against the reference corpus length. Finally, downstream translation quality is quantified against human-annotated references across the entire model suite using chrF++ for morphology-aware accuracy, BERTScore for deep semantic similarity, and a custom Target Idiom Recall (TIR) metric to explicitly validate and compare the success of the CPT knowledge injection across the evaluated architectures.

The quantitative evaluation of the models on the English-to-Sinhala translation task is summarized in Table 6. Performance was measured across morphological accuracy (chrF++), semantic preservation (BERTScore), and the successful integration of linguistic mappings (Target Recall).

Table 6: Model Evaluation Results on English-to-Sinhala Translation

Model	chrF++	BERTScore	Target Recall
Llama 3-8B	10.51	0.8452	50.50%
Minitron-4B	10.33	0.8601	37.60%
Qwen 2.5-7B	14.01	0.8869	52.30%
Gemma3-4B	5.79	0.8906	51.40%

Qwen 2.5-7B demonstrated the most balanced and robust performance overall. It achieved the highest chrF++ score (14.01), indicating superior handling of Sinhala's complex morphology, alongside the highest Target Recall (52.30%). This suggests Qwen 2.5 was the most effective at successfully leveraging the SinFoS dataset, an external foundational resource injected during the CPT phase.

Gemma3-4B presented a contrasting performance profile. Despite recording the lowest morphological accuracy (chrF++ of 5.79), it achieved the highest deep semantic similarity (BERTScore of 0.8906) and a highly competitive Target Recall (51.40%). This indicates that while Gemma3-4B successfully internalizes the underlying meaning

and retrieves the target figure of speech, it struggles with the exact character-level generation and grammatical inflection required for fluent Sinhala surface realization.

Among the remaining models, Llama 3-8B maintained a baseline capability with a solid Target Recall (50.50%), though its semantic preservation lagged behind Qwen and Gemma. Minitron-4B underperformed across the suite, most notably in Target Recall (37.60%), suggesting that its architecture or parameter constraints limited its ability to retain and consistently apply the domain-specific mappings introduced during CPT.