Context Volume Drives Performance: Tackling Domain Shift in Extremely Low-Resource Translation via RAG

David Samuel Setiawan, Raphaël Merx, Jey Han Lau
The University of Melbourne
{david.setiawan, raphael.merx}@student.unimelb.edu.au
laujh@unimelb.edu.au

Abstract

Neural Machine Translation (NMT) models for low-resource languages suffer significant performance degradation under domain shift. We quantify this challenge using **Dhao**, an indigenous language of Eastern Indonesia with no digital footprint beyond the New Testament (NT). When applied to the unseen Old Testament (OT), which exhibits a 3x increase in OOV rate (25.9%) and distinct thematic divergence from the New Testament (NT) training data, a standard NMT model fine-tuned on the NT drops from an **in-domain score of 36.17 chrF++** to **27.11 chrF++**. To recover this loss, we introduce a **hybrid framework** where a fine-tuned NMT model generates an initial draft, which is then refined by a Large Language Model (LLM) using Retrieval-Augmented Generation (RAG). The final system achieves **35.21 chrF++** (+8.10 recovery), effectively matching the original in-domain quality. Our analysis reveals that this performance is driven primarily by the **number of retrieved examples** rather than the choice of retrieval algorithm. Qualitative analysis confirms the LLM acts as a robust "safety net," repairing severe failures in zero-shot domains.

1 Introduction

For the majority of the world's 7,000+ languages, biblical texts often represent the only available large-scale digital resource (Ranathunga et al., 2023). However, translation efforts typically prioritize the New Testament (NT), leaving the Old Testament (OT), which constitutes 75% of the Bible, untranslated. The reason is that the NT contains the theological core and is an essential starting point for a new believer or a new church. Furthermore, the linguistic complexity and size of the OT present significant translation challenges.

As of August 2025, while 2,574 languages possess a complete NT, only 776 have a complete OT (Wycliffe Global Alliance, 2025). While leveraging available NT data to train Machine Translation (MT) models for the OT is a logical next step, this workflow presents a distinct **domain shift** challenge. Despite forming a single canon, the NT and OT diverge significantly in vocabulary and style, as they originate from different source languages (Hebrew vs. Koine Greek) and cover distinct themes (historical narrative vs. theological discourse).

Our analysis of the English source text (World English Bible) quantifies this shift: the Out-of-Vocabulary (OOV) rate relative to the NT training vocabulary increases from 8.1% on the in-domain NT validation set to 25.9% on the OT test set (see Appendix A). Consequently, NMT models trained solely on the NT generalize poorly, leading to marked performance degradation (Akerman et al., 2023).

In this work, we address this NT-to-OT shift using **Dhao**, an indigenous language of Eastern Indonesia with fewer than 5,000 speakers (SIL International, 2025). Unlike existing domain adaptation work which relies on target-domain monolingual corpora (Chu and Wang, 2020; Marashian et al., 2025), we operate under a stricter constraint: the primary training data is domain-bound (NT), with the only available general-domain signal coming from a small digitized grammar book. To address this, we introduce a hybrid **NMT + LLM Post-Editing** framework. We utilize a fine-tuned NMT model to generate an initial draft and employ a Retrieval-Augmented Generation (RAG) enhanced LLM to refine the output using context retrieved from both the grammar book and the NT.

We systematically compare **sentence-level retrieval** (whole-sentence similarity) against **word-level retrieval** (aggregated source word matches) to assess their impact on the proposed hybrid pipeline. Our results indicate that while increasing context volume consistently yields gains across all strategies, the specific choice of retrieval algorithm is secondary. The final optimized system restores

Proceedings for the Ninth Workshop on Technologies for Machine Translation of Low Resource Languages (LoResMT 2026), pages 87–101

character-level overlap (chrF++) to in-domain levels, though a gap remains in subword-level similarity (spBLEU), likely due to the higher stylistic and lexical divergence of the Old Testament. Based on these findings, we summarize our primary contributions as follows:

Contributions

1. **Zero-Shot Domain Adaptation for Unseen Languages:** We propose a hybrid NMT+LLM framework that successfully tackles domain shift for an extremely low-resource language with no digital footprint. We demonstrate that this architecture allows an LLM to correct a language it has never seen (Dhao) by leveraging the structural priors of a fine-tuned NMT model, outperforming baselines that rely on either model individually.

2. **Context Volume as the Primary Performance Driver:** We demonstrate that context volume drives LLM post-editing performance more significantly than the choice of retrieval algorithm in low-resource RAG. Our analysis shows that distinct retrieval strategies yield comparable gains when normalized for volume.

3. **Validation of the "Safety Net" Hypothesis:** We provide qualitative evidence that LLM post-editing specifically mitigates catastrophic NMT failures, such as hallucinations and repetition loops, in zero-shot domains, validating the hybrid architecture's robustness for data-scarce settings.

2 Related Work

Domain Shift in Low-Resource MT Standard NMT models are highly lexicalized, making them brittle when applied to distributions differing from their training data (Koehn and Knowles, 2017; Hu et al., 2019; Haddow et al., 2022). This brittleness often manifests as catastrophic hallucinations or fluent but unfaithful outputs when the model encounters out-of-domain data (Müller et al., 2020; Raunak et al., 2021). Specifically, such shifts have been shown to exacerbate the model's reliance on training-domain priors, leading it to ignore source constraints in favor of frequently observed sequences from the training set (Wang and Sennrich, 2020).

LLMs and the Hybrid Solution While Large Language Models (LLMs) excel at high-resource translation, they struggle with "unseen" languages due to a lack of pre-training exposure (Robinson et al., 2023; Hendy et al., 2023). Effective translation often remains unattainable for languages with underrepresented scripts even with RAG (Lin et al., 2025). However, **In-Context Learning (ICL) with language alignment** (e.g., dictionary constraints) has been identified as a viable method to unlock LLM capabilities for these languages, significantly outperforming fine-tuning which suffers from overfitting (Li et al., 2025).

To mitigate the weaknesses of both paradigms, recent literature converges on hybrid architectures. An NMT model followed by an LLM post-editor has been shown to be an optimal recipe for low-resource translation, specifically for mitigating "lexical confusion" (Nielsen et al., 2025). However, the optimal retrieval granularity for this post-editing remains an open question. While current strategies optimize for n-gram diversity (Caswell et al., 2025) or compositional phrases (Zebaze et al., 2025), word-level retrieval has also been proposed for grammatical learning (Tanzer et al., 2024). Our work synthesizes these insights: we employ the hybrid framework validated by Nielsen et al. (2025), but we systematically compare these sentence-level versus word-level strategies to determine whether performance gains stem from choice of retrieval algorithm or simply the increased volume of in-context examples.

3 Experimental Setup

3.1 Data Construction

We utilize the Dhao language resources introduced in Section 1 to construct a zero-shot domain adaptation benchmark. As Dhao lacks a standard digital footprint, we curate our datasets from the only two available sources: a Bible translation and a digitized grammar book (Balukh, 2020).

Primary Corpus (Parallel Bible) We source the parallel biblical text from the **ebible corpus** (Akerman et al., 2023). We align the **target** Dhao translation (written in **Latin script**) against the **source** World English Bible (WEB). The primary objective of this alignment is to decompose the raw verse-level text into a sentence-level parallel corpus, thereby providing the granular signal required for effective NMT training. The data is partitioned as follows:

- **In-domain (train & eval):** The complete New Testament (NT), comprising 7,644 parallel verses. We reserve 95% for fine-tuning the NMT model and 5% for in-domain validation.

- **Out-of-domain (test):** The first 500 verses of the Book of Genesis (Old Testament). Although the Old Testament is not fully translated in Dhao, a translation of Genesis exists; we utilize this text as our **ground truth** for evaluation. It serves as a strictly **unseen domain** to evaluate the model's generalization capabilities under the lexical shift described in Section 1. We verified that none of these verses appear in the supplementary grammar book, which exclusively cites examples from the New Testament (Balukh, 2020), ensuring no data leakage occurs.

Supplementary Corpus (Grammar Extraction) To support RAG-based post-editing, we extracted a supplementary corpus from *A Grammar of Dhao* (Balukh, 2020). Using a semi-automated pipeline involving PDF segmentation and LLM extraction (detailed in Appendix B), we curated a clean dataset of **1,011 parallel sentences** and **2,377 bilingual lexicon entries**. These resources represent the only available general-domain data for the language.

3.2 Models

We employ a hybrid architecture that leverages the complementary strengths of specialized NMT and general-purpose LLMs:

- **NMT (Drafting):** We use **NLLB-200-distilled-600M** (Costa-jussà et al., 2024). We specifically select this distilled version over larger variants (e.g., 1.3B or 3.3B) to prioritize faster inference speeds. This allows for rapid iteration during experimentation while still leveraging the model's massive multilingual pre-training, which provides a robust initialization for fine-tuning on the limited Dhao NT data.

- **LLM (Post-Editing):** We utilize **Gemini 2.5 Flash** (Comanici et al., 2025) for the post-editing stage. This model was selected for its large context window (enabling the ingestion of extensive RAG examples) and its cost-effectiveness for iterative experimentation.

3.3 Evaluation Metrics

Given the low-resource nature of Dhao and the lack of standardized linguistic tools (e.g., morphological analyzers or tokenizers), we report performance using two robust metrics:

- **spBLEU:** A tokenizer-agnostic BLEU score using SentencePiece (Costa-jussà et al., 2024). Since Dhao lacks a gold-standard tokenizer, spBLEU ensures that performance is measured based on learned sub-word units rather than potentially flawed rule-based tokenization.

- **chrF++:** A character n-gram metric (Popović, 2017). We prioritize chrF++ as it is strictly more robust than word-level BLEU for low-resource languages. By measuring character-level overlap, it provides partial credit for correct stems even when the model generates incorrect affixes or spelling variations, which is critical for accurately evaluating an unseen dialect like the Old Testament.

4 Methodology

We propose a hybrid translation framework that integrates specialized NMT with LLM-based post-editing to mitigate domain shift in extremely low-resource settings. While the architecture follows a standard post-editing setup, our primary contribution lies in the systematic optimization of the RAG context, when translating an unseen language with domain shift.

4.1 The Hybrid NMT-LLM Pipeline

The translation pipeline, illustrated in Figure 1, consists of two distinct phases:

Phase 1: NMT Drafting We fine-tune the NMT model described in Section 3.2 on the in-domain (NT) corpus to generate an initial hypothesis y_{nmt}. We adopt the optimal hyperparameters from the eBible benchmark (Akerman et al., 2023), as detailed in Appendix D.

Phase 2: RAG-Enhanced Post-Editing We post-edit the translation with **Gemini 2.5 Flash** (Comanici et al., 2025). The LLM receives a structured prompt containing: (1) the original source sentence x; (2) the NMT draft y_{nmt}; (3) a set of retrieved parallel sentences (sourced from both the NT and the grammar book); and (4) a set of retrieved lexicon entries formatted as direct mappings

(e.g., *English Word (POS)* → *Dhao Word*). The composition of these retrieved contexts depends on the experimental setup. We evaluate parallel sentence retrieval and lexicon retrieval independently, but combine them in the final optimized system (see Section 5.3). The full prompt structure and integration details are provided in Appendix C. The model is instructed to compare the draft against the source and selectively correct NMT failures like hallucinations or repetition loops only when necessary rather than re-translating from scratch.

4.2 Baselines

To evaluate the proposed framework, we compare against four baseline configurations evaluated on the out-of-domain (OT) test set. These baselines rely on **static retrieval** strategies, contrasting with the dynamic retrieval methods detailed in Section 4.3.

1. **NMT-Only:** The NLLB model fine-tuned solely on the NT data, serving as the lower-bound for domain adaptation.

2. **NMT + Grammar:** The NLLB model fine-tuned on the NT data augmented with the grammar book parallel sentences.

3. **LLM Direct Translation:** Gemini 2.5 translating directly from English to Dhao in 0-shot (no context) and 5-shot (5 fixed, randomly selected NT sentences) settings

4. **LLM Post-Editing (No RAG):** The hybrid pipeline without retrieval, relying on internal LLM knowledge to correct the NMT draft in 0-shot and 5-shot settings.

4.3 Retrieval Strategies (RAG)

A core contribution of this work is investigating whether performance gains in low-resource RAG stem from retrieval strategy, on the volume of examples, or both. Unlike the baselines which use static context, these strategies dynamically retrieve examples relevant to the specific input sentence x.

4.3.1 Parallel Sentence Retrieval

We retrieve relevant parallel pairs from a combined corpus of the in-domain NT and the grammar book. All retrieval operations are performed on the source side (English), bypassing the need for retrieval models trained on Dhao. We evaluate four strategies:

Sentence-Level Approaches (Fixed k) These methods retrieve a fixed number of sentences k based on their similarity to the source input. We test $k \in \{5, ..., 100\}$.

- **BM25 (lexical):** A standard sparse retrieval method that ranks sentences based on exact keyword overlap, normalized for document length.

- **BGE Embeddings (semantic):** A semantic retrieval method using `bge-large-en-v1.5` (Xiao et al., 2023). We compute the cosine similarity between the source sentence embedding and corpus embeddings to capture semantic relevance beyond keyword matching.

- **ChrF-Counterweighted (lexical, with diversity focus):** Adapted from Caswell et al. (2025), this method promotes n-gram diversity. It iteratively selects examples with high character n-gram overlap while penalizing n-grams present in previously selected examples, ensuring the context window is not filled with redundant phrasing.

Word-Level Approach (Dynamic k) Inspired by Tanzer et al. (2024), we implement a **Fuzzy Word Matching** strategy. Instead of retrieving based on the whole sentence, we retrieve the top-n parallel sentences for each word in the source sentence. We compute token similarity using the normalized Levenshtein distance via the `rapidfuzz` library, retaining only matches with similarity ≥ 0.5. Unlike sentence-level methods, the total number of examples k is **dynamic**, scaling with the sentence length ($k \approx n \times$ sentence_length). We ablate $n \in \{1, 2, 3, 5, 10, 15, 20\}$ to determine if granular, word-level context outperforms sentence-level retrieval.

4.3.2 Lexicon Retrieval

We further augment the context with bilingual dictionary entries extracted from the grammar book. We compare two configurations:

- **Fuzzy Retrieval:** Retrieving the top-n similar lexicon entries per source word. We evaluate $n \in \{3, 5, 10, 15, 20, 25, 30, 50, 70, 100\}$.

- **Full Dictionary:** Providing the entire 2,375-entry lexicon in the context window, treating the dictionary as a static resource rather than a retrieved element.

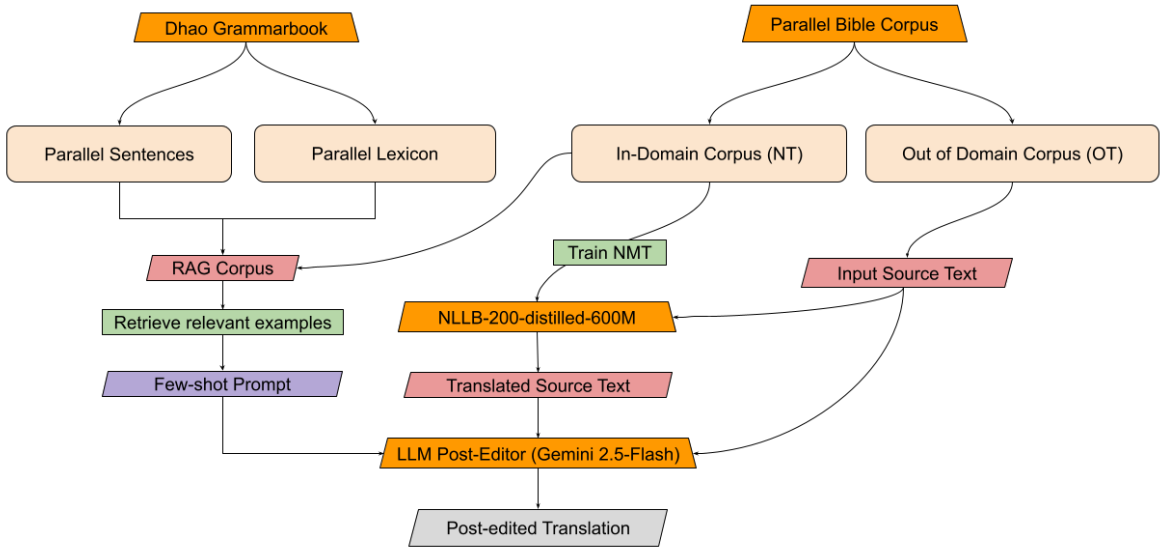

Figure 1: **Hybrid Post-Editing Architecture.** The workflow integrates two parallel streams. **Center:** The NLLB model, fine-tuned on the In-Domain (NT) corpus, generates an initial *Translated Source Text* (draft). **Left:** A Retrieval-Augmented Generation (RAG) module queries the combined corpus (Grammar Book + NT) to extract relevant examples for the *Few-shot Prompt*. **Bottom:** The LLM Post-Editor (Gemini) synthesizes the NMT draft, the original source, and the retrieved context to produce the final *Post-edited Translation*. **Legend:** Orange: Core objects (data and models) | Green: Processing steps | Red: Intermediate outputs | Purple: Prompt configuration | Gray: Final output.

5 Results

5.1 Baseline Performance & Domain Shift

We first quantify the severity of the domain shift by evaluating the fine-tuned NMT model on the out-of-domain (OT) test set. As shown in Table 1, the model suffers a performance collapse: spBLEU drops from **25.19** (in-domain NT) to **7.66** (OT), and chrF++ drops from **36.17** to **27.11**. This confirms that standard fine-tuning is insufficient for the lexical and stylistic divergence of the NT-to-OT shift.

LLMs as a Safety Net Gemini-2.5-Flash failed as a direct translator (2.98 spBLEU), confirming it possesses no prior knowledge of Dhao. However, the **Hybrid Post-Editing** framework significantly outperformed the NMT baseline (+4.88 spBLEU in the 5-shot setting). Qualitative analysis reveals that the LLM acts as a "safety net." The NMT model frequently suffers from catastrophic failures on OOV terms, such as entering infinite repetition loops. The LLM consistently identifies and truncates these loops, recovering coherent text (see Table 7).

Model	Context	spBLEU	chrF++
In-Domain Reference (NT)			
NMT (NLLB)	None	25.19	36.17
Out-of-Domain Test (OT)			
NMT (NLLB)	None	7.66	27.11
NMT + Grammar	None	7.67	26.62
LLM Direct	0-shot	2.98	18.84
LLM Direct	5-shot	7.37	22.95
LLM Post-Edit	0-shot	10.65	27.94
LLM Post-Edit	**5-shot**	**12.54**	**29.62**

Table 1: **Baseline Performance Quantification.** Comparison of NMT and LLM baselines on the out-of-domain (OT) test set. The first row shows in-domain (NT) performance as a ceiling reference. Note the severe drop when NMT is applied to the OT.

5.2 Retrieval Strategy Analysis

A core research question was whether performance gains in low-resource RAG stem from the choiceß of retrieval algorithm (e.g., semantic embeddings vs. lexical overlap) or simply the volume of in-context examples. To answer this, we decouple our analysis into two parts: first evaluating parallel sentence retrieval in isolation, and subsequently

analyzing the impact of lexicon retrieval.

5.2.1 Parallel Sentence Retrieval Analysis

Impact of Context Volume As shown in Figure 2 (Left) and detailed in Table 4 (Appendix E), all dynamic sentence-level strategies outperform the static 5-shot baseline (dashed red line) immediately, even at low k. We observe strong, consistent improvement as the context volume increases from 5 to 60, regardless of whether dense embeddings (BGE) or sparse matching (BM25) is used. However, these methods plateau around $k \approx 60$. In contrast, the Word-Level strategy (Figure 2, Right) circumvents this saturation. By retrieving granular examples, it allows the model to effectively utilize a much larger context volume, with performance continuing to scale until peaking at an effective $k \approx 137$ (see Table 5 for full numerical results).

Performance Convergence and Efficiency Trade-offs When comparing the optimal configurations of each method, we observe a convergence in peak performance. As illustrated in the bar chart in Figure 3, the maximum chrF++ scores for the Word-Level, ChrF-RAG, and BGE strategies are all within **0.5 points** of each other. The **Word-Level Fuzzy Matching** strategy achieves the absolute highest score (**35.28 chrF++**), but it outperforms the best sentence-level baseline (ChrF-RAG: 34.98) by only a small margin (+0.3).

However, efficiency analysis favors sentence-level retrieval. ChrF-RAG achieves 99% of the optimal performance with just $K = 60$ examples, whereas the word-level strategy requires nearly double the volume (≈137) for a marginal gain. This makes sentence-level retrieval the more pragmatic choice for production environments where token usage and latency are constraints.

Impact of Retrieval Corpus To validate the generalizability of our findings, we isolated the impact of the supplementary grammar data. We compared the performance of the best configuration (Word-Level, $n = 10$) using the combined corpus versus using *only* the in-domain NT for retrieval.

Results show that restricting the retrieval source to the NT corpus results in only a marginal performance drop compared to the combined corpus (from **35.28** to **35.01 chrF++**, and **18.93** to **18.47 spBLEU**). This confirms that the approach remains a viable solution for extremely low-resource languages where a Bible translation may be the *only* available digital resource, without requiring the digitization of supplementary grammar books.

Configuration	spBLEU	chrF++
In-Domain Reference (NT)	*25.19*	*36.17*
Out-of-Domain Results (Genesis)		
Baseline (NMT Only)	7.66	27.11
+ Lexicon (Full)	16.27 ↑ **8.61**	31.32 ↑ **4.21**
+ Sentences (Word-Level)	18.93 ↑ **11.27**	**35.28** ↑ **8.17**
+ Combined (Final)	**19.88** ↑ **12.22**	35.21 ↑ **8.10**

Table 2: **Experimental Results.** Comparison of component contributions. The first row provides the in-domain (NT) upper bound. Our final combined system (Word-Level Sentences + Full Lexicon) achieves **35.21 chrF++**, effectively recovering the performance lost to domain shift by nearly matching the in-domain reference of 36.17.

5.2.2 Lexicon Retrieval Analysis

Having analyzed parallel sentence retrieval, we independently evaluate the utility of the bilingual lexicon.

Impact of Dictionary Volume As shown in Figure 4, performance improves linearly with the number of entries provided, contrasting with the plateau observed in sentence retrieval. The optimal performance was achieved by providing the **Full Dictionary**, yielding **16.27 spBLEU** (+8.61) and **31.32 chrF++** (+4.21). The fuzzy matching approach with $N = 100$ closely approximated this peak (≈30.88 chrF++), suggesting that for targeted lexical information, quantity is strictly beneficial. Providing the entire lexicon maximizes the probability of retrieving precise translations for OOV terms without introducing the syntactic noise inherent in full sentences (see Table 6 in Appendix E for full results).

5.3 Final System Performance

Our final system combines the optimal parallel sentence retrieval strategy, the Word-Level Fuzzy Matching ($n = 10$, yielding an effective $k \approx 137$), with the full bilingual lexicon. As shown in Table 2, this yields the highest overall translation accuracy.

The final model achieves **35.21 chrF++**, which almost matches the in-domain performance of the NMT model (36.17 chrF++). This indicates that our RAG-enhanced post-editing framework has successfully recovered the performance lost to domain shift.

Interestingly, while the combined approach

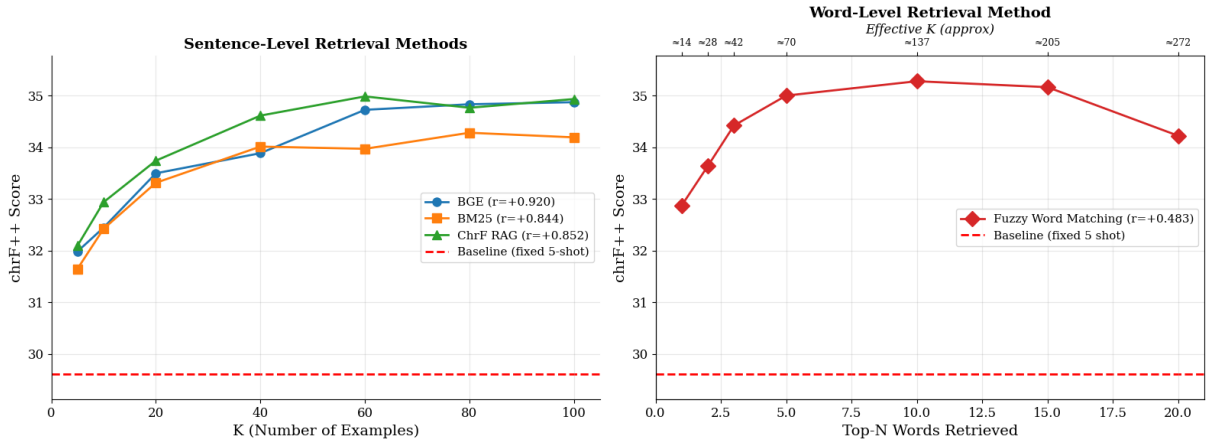

Figure 2: **Impact of Context Volume on Performance.** Comparison of absolute chrF++ scores across retrieval strategies relative to the **Fixed 5-Shot Baseline** (dashed red line, corresponding to the LLM Post-Editing baseline in Table 1). **Left:** Sentence-level methods (BGE, BM25, ChrF-RAG) show rapid initial gains but plateau at $K \approx 60$. **Right:** The Word-Level strategy allows the model to ingest a higher effective volume of examples (peaking at effective $K \approx 137$) to squeeze out marginal performance gains.

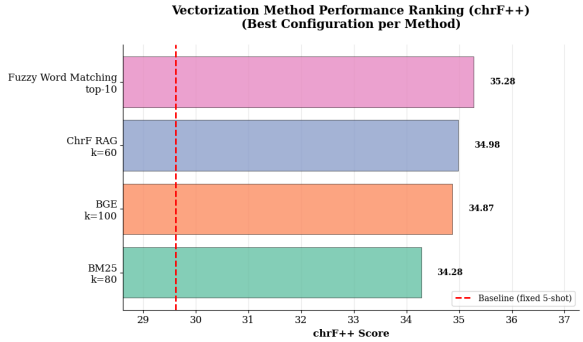

Figure 3: **Retrieval Strategy Performance Convergence.** We compare the optimal configuration of the Word-Level strategy (Fuzzy Matching, top-10) against the best configurations of three sentence-level baselines: ChrF-RAG ($k = 60$), BGE Semantic Retrieval ($k = 100$), and BM25 ($k = 80$). The bar chart displays the absolute chrF++ scores, with the dashed red line indicating the Fixed 5-shot Baseline.

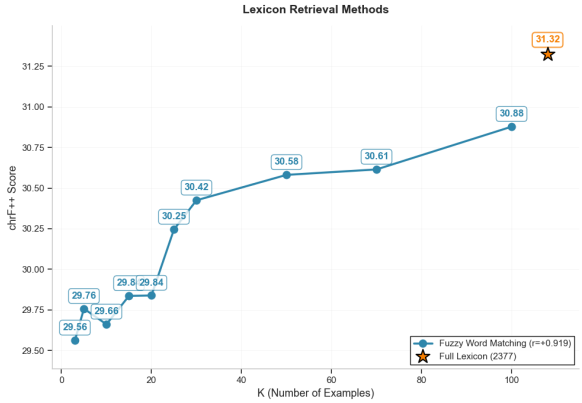

Figure 4: **Lexicon Retrieval Performance.** Impact of increasing the number of retrieved lexicon entries (K) on post-editing performance. Unlike sentence-level retrieval which plateaus, lexicon retrieval improves monotonically with volume. The highest performance is achieved by providing the **Full Lexicon** (star marker), yielding a chrF++ score of 31.32.

yields the highest spBLEU (**19.88**), the chrF++ score (**35.21**) is slightly lower than the sentence-only configuration (**35.28**). This suggests a nuanced trade-off in metric sensitivity: the lexicon provides high-precision constraints that improve exact **subword-level overlap** (boosting spBLEU), which is otherwise hindered by the 25.9% OOV rate. Conversely, the **character-level overlap** (chrF++) appears to reach saturation with sentence-level retrieval alone, nearly matching the in-domain reference of 36.17. The slight drop in chrF++ when adding the full lexicon may indicate that the increased **context volume** introduces minor syntactic noise that outweighs marginal gains in character-level accuracy. Nevertheless, the combined model remains the most robust overall system for recovering performance across both subword and character granularities.

5.4 Qualitative Analysis

To investigate the source of the performance gains, we analyzed the test set outputs. We found that the fine-tuned NMT model frequently suffers from catastrophic failures, which the RAG-enhanced LLM effectively repairs. As illustrated in Table 7 (see Appendix F), we observed three distinct failure modes:

The "Safety Net" Effect The most prevalent NMT error is the **repetition loop**, where the model gets stuck generating a single token sequence (e.g., *kahib'i-kalèbho* in GEN 10:17). The LLM demonstrates a language-agnostic ability to identify these non-sensical patterns, disregard the NMT draft, and re-translate based on the source and retrieved context.

Hallucination Correction The NMT model occasionally generates fluent but factually incorrect text. In GEN 11:5, the NMT hallucinates "King David" (*dhèu aae Daud*)—a figure frequent in the training data but absent in the source. The Post-Editor correctly identifies this mismatch against the English source ("The Lord") and the retrieved lexicon, correcting it to *Lamatua*.

Syntactic Reconstruction Finally, the LLM successfully reconstructs complex genealogical idioms that confuse the NMT. In GEN 11:17, the NMT fails to render the phrase "became the father of," likely due to the structural divergence between the literal grammar book data and the idiomatic Bible. The Post-Editor, aided by retrieved examples, correctly utilizes the Dhao idiom *matana* ('became the father of'), demonstrating the value of the word-level retrieval strategy in capturing high-frequency idiomatic patterns.

5.5 Recommendations

Based on our findings, we offer three key recommendations for practitioners working on unseen, low-resource languages:

Start with Context Volume, then look into Context Efficiency We recommend a two-step approach to RAG for low-resource MT: first, maximize the number of retrieved examples (k), as context saturation provides the most significant boost to translation quality regardless of the retrieval method. Second, tune for computational efficiency. Since our experiments show that distinct retrieval strategies converge to a similar performance band, practitioners can select the algorithm that best fits their latency constraints, switching from computationally expensive brute-force methods to faster alternatives, such as semantic search using sentence embedding models (e.g., BGE) or inverted indices (BM25), without sacrificing translation accuracy.

Consider Leaving Lexicographic Data for ICL, not Fine-Tuning Our "NMT + Grammar" baseline demonstrated that simply adding grammar book data at the fine-tuning stage can be detrimental. Performance actually degraded from 27.11 to 26.62 chrF++, likely because the rigid, pedagogical style of grammar book examples conflicts with the literary flow of the target domain. We show these resources can be best preserved as external knowledge bases for RAG, allowing the model to query specific terms dynamically without polluting the model's internal stylistic representations.

Keep an Out-of-Domain Test Set to Measure Robustness Standard practices in low-resource NMT often involve randomly splitting available corpora (e.g., the Bible) into training and test sets (Vázquez et al., 2021; Marashian et al., 2025). While many papers assume that the Bible belongs to a single, religious domain, our analysis shows a marked domain shift within this text, demonstrating that measuring out-of-domain performance is possible even when only the Bible is available as parallel corpus. Therefore, we recommend that low-resource researchers take this into consideration instead of using all verses of the Bible for both train and test, opting instead for document-level holdouts (e.g., distinct books or Testaments) to avoid inflated performance estimates (Khiu et al., 2024). This aligns with recent findings from the WMT 2025 General Translation task (Kocmi et al., 2025), which argue that evaluating on "easy" in-domain data masks model brittleness and that robust assessment requires testing on challenging, document-level out-of-domain holdouts.

6 Conclusion

This work addresses domain shift in extremely low-resource settings. We demonstrate that while standard NMT suffers catastrophic degradation on unseen domains, our proposed hybrid NMT+LLM framework functions as a robust "safety net," effectively recovering the quality lost to lexical and stylistic divergence.

Crucially, we find that context volume, rather than retrieval algorithm, is the primary driver of performance. We observe that distinct retrieval strategies (lexical vs. semantic) converge to comparable quality levels when normalized for volume. By validating these trade-offs on a language with no digital footprint, we provide a scalable blueprint for accelerating the translation of the Old Testament for thousands of low-resource languages worldwide.

7 Limitations

While this study provides a robust framework for tackling domain shift, it also highlights several limitations and clear avenues for future research:

1. **Generalizability of Language and Domain:** The experiments were conducted on a single language pair (English-to-Dhao) and a single, specific domain shift (NT-to-OT). Future work is needed to test the generalizability of this framework. It would be valuable to validate whether the superiority of word-level retrieval and the "safety net" function of the LLM post-editor hold true for other low-resource language pairs and different types of domain shifts, such as translating from religious to secular text (e.g., news or health domains).

2. **Optimizing Contextual Synergy:** As discussed in Section 5.3, our investigation into combining context types yielded mixed results. While combining the best sentence retrieval method with the full dictionary yielded the highest spBLEU score, it caused a slight decrease in the chrF++ score compared to using sentences alone. This suggests a lack of perfect synergy, likely because the large volume of combined data introduced noise. Future work should conduct finer-grained ablation studies to find the optimal balance, for instance, by combining parallel sentence retrieval with a *retrieved subset* of the lexicon rather than the full dictionary, which may reduce noise and improve both metrics.

Ethical Considerations

Data Usage and Copyright. This work utilizes data from the Dhao Alkitab (copyright ©2012 Unit Bahasa dan Budaya) and *A Grammar of Dhao* (Balukh, 2020). The Bible translation is licensed under **Creative Commons Attribution-NoDerivatives 4.0 International (CC BY-ND 4.0)**, which permits redistribution for research purposes provided the text remains unaltered. The grammar book is an **Open Access** publication (LOT Dissertation Series 570). Our use of these materials for non-commercial linguistic analysis and machine translation evaluation is consistent with these licenses and established **Fair Use** protocols for academic research. We provide full attribution to the original authors and rights holders in our citations.

Impact on Low-Resource Communities. Our primary goal is to develop technologies that support the preservation and revitalization of very low-resource languages. We recognize that AI development for indigenous languages carries the risk of extractive research practices. To mitigate this, we focus on methods that can be deployed with minimal data and computational resources, making them accessible to local stakeholders. We hope this work serves as a foundation for future community-driven language tools.

Risks of Generative Models. Neural Machine Translation and LLMs are prone to hallucinations, which poses a specific risk when handling sensitive or religious texts where accuracy is paramount. Our proposed **hybrid automated framework** (using LLMs as a post-editing safety net) is explicitly designed to identify and correct such anomalies. However, we emphasize that automated translations should always be reviewed by native speakers and community leaders before being treated as authoritative.

Acknowledgements

This research was supported by the Commonwealth through an Australian Government Research Training Program Scholarship (https://doi.org/10.82133/C42F-K220). This research was undertaken using the LIEF HPC-GPGPU Facility hosted at the University of Melbourne. This Facility was established with the assistance of LIEF Grant LE170100200. Lau was supported by the Australian Research Council under Grant LP210200917.

References

Vesa Akerman, David Baines, Damien Daspit, Ulf Hermjakob, Taeho Jang, Colin Leong, Michael Martin, Joel Mathew, Jonathan Robie, and Marcus Schwarting. 2023. The ebible corpus: Data and model benchmarks for bible translation for low-resource languages. *arXiv preprint arXiv:2304.09919*.

Jermy Immanuel Balukh. 2020. *A grammar of Dhao: An endangered Austronesian language in Eastern Indonesia*. LOT dissertation series.

Isaac Caswell, Elizabeth Nielsen, Jiaming Luo, Colin Cherry, Geza Kovacs, Hadar Shemtov, Partha Talukdar, Dinesh Tewari, Baba Mamadi Diane, Djibrila Diane, Solo Farabado Cissé, Koulako Moussa Doumbouya, Edoardo Ferrante, Alessandro Guasoni, Christopher Homan, Mamadou K. Keita, Sudhamoy

DebBarma, Ali Kuzhuget, David Anugraha, and 5 others. 2025. SMOL: Professionally translated parallel data for 115 under-represented languages. In *Proceedings of the Tenth Conference on Machine Translation*, pages 1103–1123, Suzhou, China.

Chenhui Chu and Rui Wang. 2020. A survey of domain adaptation for machine translation. *Journal of information processing*, 28:413–426.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, and 1 others. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*.

Marta Ruiz Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loïc Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, and 19 others. 2024. Scaling neural machine translation to 200 languages. *Nature*, 630:841 – 846.

Barry Haddow, Rachel Bawden, Antonio Valerio Miceli Barone, Jindřich Helcl, and Alexandra Birch. 2022. Survey of low-resource machine translation. *Computational Linguistics*, 48(3):673–732.

Amr Hendy, Mohamed Abdelrehim, Amr Sharaf, Vikas Raunak, Mohamed Gabr, Hitokazu Matsushita, Young Jin Kim, Mohamed Afify, and Hany Hassan Awadalla. 2023. How Good Are GPT Models at Machine Translation? A Comprehensive Evaluation. *arXiv preprint*. ArXiv:2302.09210 [cs].

Junjie Hu, Mengzhou Xia, Graham Neubig, and Jaime Carbonell. 2019. Domain adaptation of neural machine translation by lexicon induction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2989–3001, Florence, Italy.

Eric Khiu, Hasti Toossi, David Anugraha, Jinyu Liu, Jiaxu Li, Juan Flores, Leandro Roman, A. Seza Doğruöz, and En-Shiun Lee. 2024. Predicting machine translation performance on low-resource languages: The role of domain similarity. In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 1474–1486, St. Julian's, Malta.

Tom Kocmi, Ekaterina Artemova, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Konstantin Dranch, Anton Dvorkovich, Sergey Dukanov, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Marzena Karpinska, Philipp Koehn, Howard Lakougna, Jessica Lundin, Christof Monz, Kenton Murray, and 10 others. 2025. Findings of the WMT25 general machine translation shared task: Time to stop evaluating on easy test sets. In *Proceedings of the Tenth Conference on Machine Translation*, pages 355–413, Suzhou, China. Association for Computational Linguistics.

Philipp Koehn and Rebecca Knowles. 2017. Six challenges for neural machine translation. In *Proceedings of the First Workshop on Neural Machine Translation*, pages 28–39, Vancouver.

Yue Li, Zhixue Zhao, and Carolina Scarton. 2025. It's all about in-context learning! teaching extremely low-resource languages to LLMs. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 29532–29547, Suzhou, China.

Dianqing Lin, Aruukhan, Hongxu Hou, Shuo Sun, Wei Chen, Yichen Yang, and Guodong Shi. 2025. Can large language models translate unseen languages in underrepresented scripts? In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 23148–23161, Suzhou, China.

Ali Marashian, Enora Rice, Luke Gessler, Alexis Palmer, and Katharina von der Wense. 2025. From priest to doctor: Domain adaptation for low-resource neural machine translation. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 7087–7098, Abu Dhabi, UAE.

Mathias Müller, Annette Rios, and Rico Sennrich. 2020. Domain robustness in neural machine translation. In *Proceedings of the 14th Conference of the Association for Machine Translation in the Americas (Volume 1: Research Track)*, pages 151–164, Virtual. Association for Machine Translation in the Americas.

Elizabeth Nielsen, Isaac Rayburn Caswell, Jiaming Luo, and Colin Cherry. 2025. Alligators all around: Mitigating lexical confusion in low-resource machine translation. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 206–221, Albuquerque, New Mexico.

Maja Popović. 2017. chrF++: words helping character n-grams. In *Proceedings of the Second Conference on Machine Translation*, pages 612–618, Copenhagen, Denmark.

Surangika Ranathunga, En-Shiun Annie Lee, Marjana Prifti Skenduli, Ravi Shekhar, Mehreen Alam, and Rishemjit Kaur. 2023. Neural machine translation for low-resource languages: A survey. *ACM Computing Surveys*, 55(11):1–37.

Vikas Raunak, Arul Menezes, and Marcin Junczys-Dowmunt. 2021. The curious case of hallucinations in neural machine translation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1172–1183, Online. Association for Computational Linguistics.

Nathaniel Robinson, Perez Ogayo, David R. Mortensen, and Graham Neubig. 2023. ChatGPT MT: Competitive for high- (but not low-) resource languages. In *Proceedings of the Eighth Conference on Machine Translation*, pages 392–418, Singapore.

SIL International. 2025. Ethnologue: Languages of the world. https://www.ethnologue.com/. Accessed: August 14, 2025.

Garrett Tanzer, Mirac Suzgun, Eline Visser, Dan Jurafsky, and Luke Melas-Kyriazi. 2024. A benchmark for learning to translate a new language from one grammar book. In *International Conference on Representation Learning*, volume 2024, pages 18955–18985.

Raúl Vázquez, Yves Scherrer, Sami Virpioja, and Jörg Tiedemann. 2021. The Helsinki submission to the AmericasNLP shared task. In *Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas*, pages 255–264, Online.

Chaojun Wang and Rico Sennrich. 2020. On exposure bias, hallucination and domain shift in neural machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3544–3552, Online. Association for Computational Linguistics.

Wycliffe Global Alliance. 2025. 2025 global scripture access. https://wycliffe.net/resources/statistics/. Accessed: 2025-12-10.

Shitao Xiao, Zheng Liu, Peitian Zhang, and Niklas Muennighoff. 2023. C-pack: Packaged resources to advance general chinese embedding. *Preprint*, arXiv:2309.07597.

Armel Randy Zebaze, Benoît Sagot, and Rachel Bawden. 2025. Compositional translation: A novel LLM-based approach for low-resource machine translation. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 22328–22357, Suzhou, China.

A Domain Shift Analysis

To quantify the lexical divergence between the New Testament (NT) and Old Testament (OT), we analyzed the frequency of domain-specific terms and the Out-of-Vocabulary (OOV) rates.

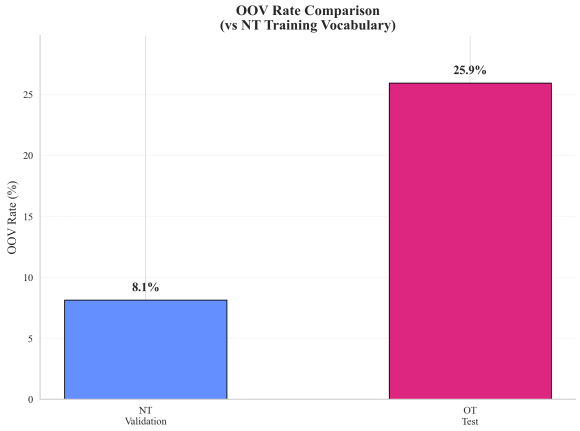

Figure 5: The Out-of-Vocabulary (OOV) rate of the in-domain NT Validation set (8.1%) versus the out-of-domain OT Test set (25.9%). All rates are calculated relative to the NT training vocabulary.

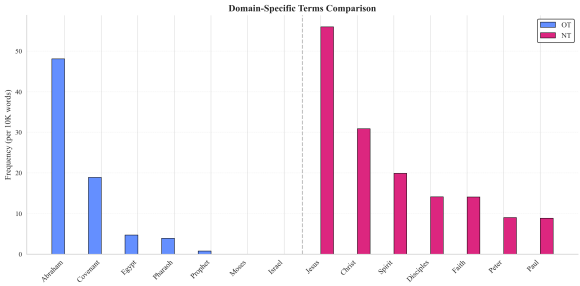

Figure 6: Domain-specific term frequencies (normalized per 10k words) comparing the Old Testament (OT) and New Testament (NT) corpora. Note the prevalence of historical terms in the OT versus theological terms in the NT.

B Data Construction Details

We illustrate the complete data processing pipeline used to curate the experimental datasets from the unstructured grammar book and the raw biblical text.

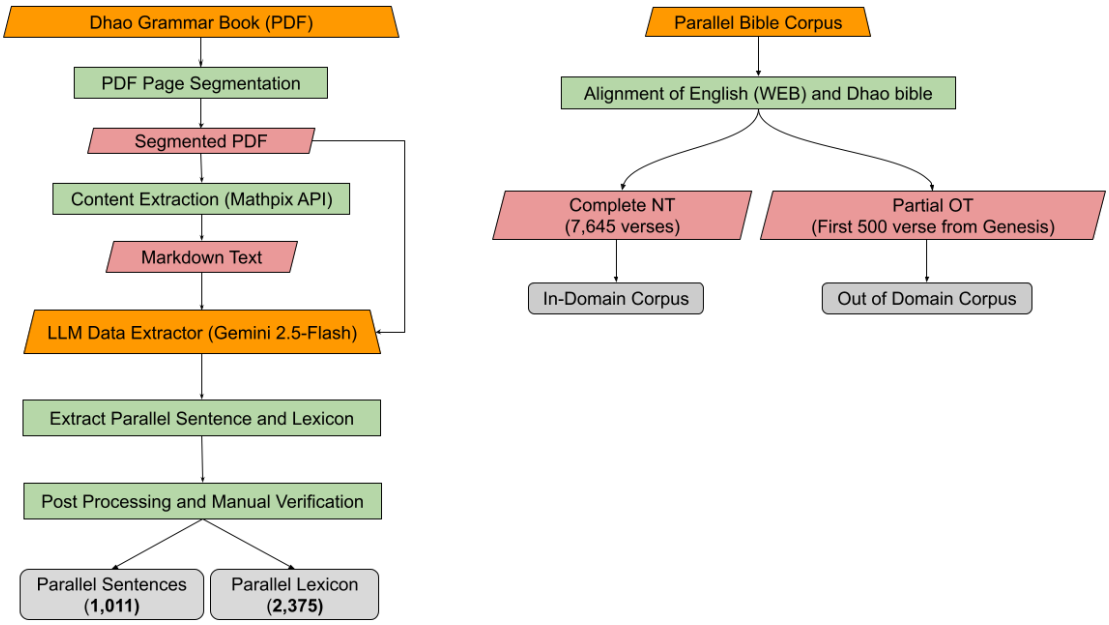

Figure 7: **Data Construction Pipeline.** The workflow processes two primary sources to create the experimental datasets. **Left:** We extract supplementary parallel sentences and lexicon entries from the unstructured Dhao Grammar Book PDF using a multi-stage pipeline involving OCR and LLM-based extraction. **Right:** The Parallel Bible Corpus is aligned and partitioned to simulate domain shift, using the New Testament (NT) as the in-domain training corpus and the Old Testament (OT) as the out-of-domain test set. **Legend:** Orange: Core objects (data and model) | Green: Processing steps | Red: Intermediate data | Gray: Final output datasets.

C Prompt Templates

To ensure reproducibility, we provide the exact prompt structures used for the Large Language Model (Gemini 2.5 Flash). We utilized two distinct prompt strategies: one for direct translation (baseline) and one for the post-editing task.

C.1 System Instructions

The following system prompts establish the persona and constraints for the LLM.

Direct Translation Prompt

System Message:
Dhao is a member of the Sumba-Flores branch of the Malayo-Polynesian language family. It is spoken in Ndao Island in the Lesser Sunda Islands in Indonesia by about 5,000 people. It is classified as a member of the Sumba branch of Malayo-Polynesian languages, but may be a Papuan language. It is also known as Ndao, Ndaonese or Ndaundau.
You are an expert Bible translator in Dhao language. Your job is to translate bible verses from English to Dhao language, providing accurate and faithful translations that maintain the meaning and context of the source text. When provided with glossary entries or example translations, use them as reference to help ensure correct translation. You must respond ONLY with your translation in Dhao - no explanations, no reasoning, no additional text.

User Template:
Source text (English): {src_text}
Translate the above text from English to Dhao:

Post-Editing Prompt

System Message:
Dhao is a member of the Sumba-Flores branch of the Malayo-Polynesian language family. It is spoken in Ndao Island in the Lesser Sunda Islands in Indonesia by about 5,000 people. It is classified as a member of the Sumba branch of Malayo-Polynesian languages, but may be a Papuan language. It is also known as Ndao, Ndaonese or Ndaundau.
You are an expert Bible translator in Dhao language. Your job is to correct and verify machine generated bible verses in Dhao language which is translated from the English language. Only make changes when necessary, ensuring that the post-edited dhao verse is aligned with the source English verse. When provided with glossary entries or example translations, use them as reference to help ensure correct translation. You must respond ONLY with the corrected translation text - no explanations, no reasoning, no additional text.

User Template:
Source text (English): {src_text}
Machine translation (Dhao): {pred_text}
Correct the machine translation if necessary:

C.2 Dynamic Context Injection

Depending on the retrieval strategy, the following blocks are dynamically appended to the User Template.

Dynamic Component: Parallel Sentence Examples

(Appended when $k > 0$ parallel sentences are retrieved)
To help with the translation, here are some example parallel sentences between Dhao and English:
Dhao: [target_example_1]
English translation: [source_example_1]
Dhao: [target_example_2]
English translation: [source_example_2] ...

Dynamic Component: Glossary Entries

(Appended when Lexicon Retrieval is enabled)
To help with the translation, here is a word list between English and Dhao in the format: English word (pos tag) -> Dhao word:

- source_word_1 (noun) -> target_word_1
- source_word_2 (verb) -> target_word_2
- source_word_3 -> target_word_3

D Hyperparameters and Training Details

All NMT models were fine-tuned using the Hugging Face `transformers` library on a single NVIDIA A100 GPU. We utilized the `facebook/nllb-200-distilled-600M` pre-trained checkpoint. Table 3 summarizes the hyperparameters used for both the baseline (NMT-Only) and the augmented (NMT + Grammar) configurations.

Note that the *NMT + Grammar* model was trained for more steps (7,000 vs 5,000) to account for the additional training data provided by the grammar book CSV.

Hyperparameter	**NMT-Only**	**NMT + Grammar**
Base Model	NLLB-200-distilled-600M	
Precision	bfloat16	
Attention Impl.	SDPA (Scaled Dot Product Attention)	
Learning Rate	2e-4	
Label Smoothing	0.2	
Warmup Steps	1,000	
Early Stopping Patience	4	
Batch Size (per device)	16	
Grad. Accumulation Steps	4	
Effective Batch Size	64	
Max Sequence Length	400	
Beam Size (Eval)	2	
Max Training Steps	**5,000**	**7,000**

Table 3: Fine-tuning hyperparameters for the NMT baselines.

E Detailed Retrieval Results

We provide the complete numerical results for the retrieval ablation studies discussed in Section 5.2. Table 4 details the performance of sentence-level strategies (BM25, BGE, ChrF-RAG) across varying context sizes (K). Table 5 details the word-level fuzzy matching strategy across varying retrieval densities (N). Finally, Table 6 presents the results for Lexicon Retrieval, comparing dynamic retrieval against the static full-dictionary baseline.

Method	K	spBLEU	chrF++
Baseline			
NMT (NLLB)	-	7.66	27.11
BGE (Semantic)			
BGE	5	15.23	31.98
BGE	10	15.71	32.45
BGE	20	16.97	33.50
BGE	40	17.34	33.88
BGE	60	18.34	34.72
BGE	80	18.32	34.83
BGE	100	18.47	34.87
BM25 (Lexical)			
BM25	5	14.76	31.65
BM25	10	15.50	32.42
BM25	20	16.53	33.31
BM25	40	17.44	34.01
BM25	60	17.34	33.97
BM25	80	17.51	34.28
BM25	100	17.25	34.19
ChrF-RAG (Diversity)			
ChrF	5	15.33	32.09
ChrF	10	16.32	32.94
ChrF	20	17.01	33.74
ChrF	40	18.05	34.61
ChrF	**60**	**18.44**	**34.98**
ChrF	80	18.12	34.76
ChrF	100	18.27	34.93

Table 4: Detailed ablation results for **Sentence-Level** retrieval methods. Note that performance gains tend to plateau around $K = 60 - 80$ for most methods.

N (per word)	Eff. K	spBLEU	chrF++
Word-Level Fuzzy Matching			
1	≈14	15.96	32.87
2	≈28	16.94	33.64
3	≈42	17.72	34.42
5	≈70	18.72	35.00
137	**≈137**	**18.93**	**35.28**
15	≈205	18.82	35.16
20	≈272	16.90	34.22

Table 5: Detailed ablation for the **Word-Level** strategy. "Eff. K" denotes the approximate effective number of sentences retrieved. Performance peaks at $N = 10$ before degrading due to context noise.

N (per word)	Eff. K	spBLEU	chrF++
Word-Level Fuzzy Matching			
3	≈77	13.84	29.56
5	≈128	14.29	29.76
10	≈256	14.18	29.66
15	≈384	14.53	29.84
20	≈512	14.85	29.84
25	≈640	15.19	30.25
30	≈768	15.19	30.42
50	≈1280	15.29	30.58
70	≈1791	15.76	30.61
100	≈2559	16.22	30.88
Static Context			
Full Dictionary	**2377**	**16.27**	**31.32**

Table 6: Detailed ablation for Lexicon Retrieval. Unlike sentence retrieval, performance scales monotonically with volume, peaking when the Full Dictionary is provided.

F Qualitative Analysis Examples

We provide concrete examples of the failure modes discussed in Section 5.4. Table 7 highlights three specific instances where the NMT baseline failed catastrophically on the Out-of-Domain test set, and how the RAG-enhanced Post-Editor recovered the correct translation.

Type	Source (English)	NMT Output (Draft)	LLM Post-Edit (Final)
Repetition Loop	*"it rained on the earth forty days and forty nights"* (GEN 7:12)	*Hèia bèli-camèd'a bèli-camèd'a...* (repeats indefinitely)	*Èj'i bhori ètu rai èpa nguru lod'o mèu-mèda.* (Correct translation)
Hallucination	*"the lord came down to see the city..."* (GEN 11:5)	*...dhèu aae **Daud** puru...* ("King **David** came down...")	***Lamatua** puru nèti dedha mai...* ("**The Lord** came down...")
Complex Syntax	*"...after he became the father of peleg"* (GEN 11:17)	*Nèti èèna ka, Eber mamuri toke d'ai... Èle èèna ka, nèti èèna ka, nèti èèna ka...* (Degenerates into repetition loop)	*...èli nèngu **matana** Peleg...* (Recovers genealogical idiom)

Table 7: Examples of NMT failures corrected by the RAG-enhanced Post-Editor. The LLM acts as a safety net, fixing repetition loops, named entity hallucinations, and recovering complex idioms.