

# Bridging the Linguistic Divide: A Comprehensive Review of Methods and Challenges in Low-Resource Machine Translation

Sunil Angom<sup>1</sup>, Pougailung Gangmei<sup>2</sup>, Moirangthem Daniel Singh<sup>3</sup>

Department of Artificial Intelligence, NIELIT Deemed to be University, Imphal, India  
linusangom46@gmail.com<sup>1</sup>, pougaigangmei19@gmail.com<sup>2</sup>, officielworkingemail@gmail.com<sup>3</sup>

**Abstract:** Large language models (LLMs) have become useful in high-resource languages due to the rapid advancement of natural language processing (NLP), yet the vast majority of the world's languages remain underserved because of a severe lack of data. This paper develops a complete taxonomy of machine translation (MT) methods for low-resource languages (LRLs) and critically evaluates their construction. Rule-based (RBMT) and statistical machine translation (SMT) remain essential for digitally dark languages such as Tulu and Kangri, while transformer based frameworks and parameter efficient fine tuning (PEFT) demonstrate fluency in mid resource scenarios. We also evaluate data-centric interventions, balancing the advantages of parallel corpus mining against the risks of synthetic data, which can distort semantic polarity in sensitive domains. Our synthesis identifies crossover points where statistical methods outperform neural approaches and integrates findings from dynamic language identification, hybrid transfer learning, and retrieval augmented generation (RAG). The paper concludes with a resource-tier recommendation table and proposes future directions including neuro symbolic hybrid architectures, unsupervised corpus mining, and culturally grounded evaluation metrics.

**Keywords:** Low resource language, Neural Machine Translation, Large Language Models, Parameter-Efficient Fine Tuning, Data Augmentation, Transfer Learning.

*(Article history: Selected papers from NIELIT's International Conference on Communication, Electronics and Digital Technologies (NICE-DT'26), 29-30Jan, 2026, Gorakhpur, India)*

## 1. Introduction

In the last decade, the sphere of Machine Translation (MT) has undergone a paradigm shift, with neural and LLM-based approaches now dominant for high-resource languages by the recent stage of huge language models (LLMs) and neural Machine Translation (NMT) [1, 2, 34]. The great majority of the 7,000+ languages of the world continue to face a significant digital divide even though these advances have introduced high-resource languages (HRLs) such as English, Chinese, and French to human parity [35]. Notably, mainstream NLP technologies still do not serve low-resource languages (LRLs) well, which is worsened by the absence of large-scale parallel corpora, digital resources, and linguistic knowledge [4], [34].

### 1.1 Background and Context

Prior to it, statistical Machine Translation (SMT) was the industry standard of LRLs and this algorithm used probability distributions calculated based on a small set of bilingual texts [5], [20]. The challenge between HRLs and LRLs was addressed by introducing a universal architecture with the Transformer architecture [6] and knowledge transfer between HRLs and LRLs on a magnitude of scale as proposed by mBART, mT5 [16], [22], respectively. Recent innovations have increased this tendency: Although methods such as retrieval-augmented generation (RAG) attempt to localize the LLMs to external knowledge bases [15], the parameter-efficient fine-tuning (PEFT) techniques are now capable of converting huge models into new languages using consumer-grade hardware [25]. These are complex constructions that are hard to transfer to LRLs, however. NMT models can easily crash when the amount of training data is below critical volume thresholds, often leading to hallucinations or lack of convergence [12], [14]. The linguistic diversity of LRLs compounds these challenges from the script heterogeneity of Manipuri [12] to the agglutinative complexity of Kazakh [19] posing problems that universal models struggle to address without significant modification.

### 1.2 Motivation

This study was provoked by the conflicting accounts that are currently widespread in the research of the topic of MT. On one hand, the academic milieu strongly supports the use of LLMs, which often makes them out to offer solutions to all the issues regarding the language. But empirical evidence of digitally dark languages shows that traditional methods (SMT, Rule-Based) are often superior to deep learning models in zero-resource scenarios [8], [10]. Indicatively, current studies on Hindu-Kangri and English-Nepali translation Indicate that SMT systems can be trained to perform better than NMT when the sample size is

less than 30,000 phrases in terms of BLEU scores [4], [10]. It is also important to study the reliance on artificial data. Large-scale back-translation has been demonstrated to be useful in the sciences [23], but a threat to those tasks that are sensitive to semantics like sentiment analysis, where machine-translated training examples can swap polarity and hurt the performance of a classifier [13]. This review fills that gap by providing a systematic, data availability driven framework for selecting appropriate MT methods.

### 1.3 Paper Organization

The rest of the sections of the paper are presented in the following way, section 2 contains a list of existing approaches, which is classified into three categories according to the resources available at the moment: zero, low, and augmented. In section 3, a comparison of important research is made and performance Indicators (BLEU, COMET) of different language pairings are synthesized. Section 4 includes important issues like the risks of synthetic data and pipeline instability. In section 5, the areas of future research are defined, which are unsupervised mining and hybrid neuro-symbolic architectures. Section 6 concludes the research.

## 2. Literature Review and Taxonomy

The state of the art in Machine Translation in low-resource languages (LRLs) is characterized by a dilemma between the complexity of models and the availability of data. According to the analysis of recent studies, we divide the existing methods into four broad paradigms:

- (1) Foundational approaches to zero-resource scenarios,
- (2) Neural architectures and transfer learning,
- (3) Large language model adaptation and
- (4) Data-driven strategies.

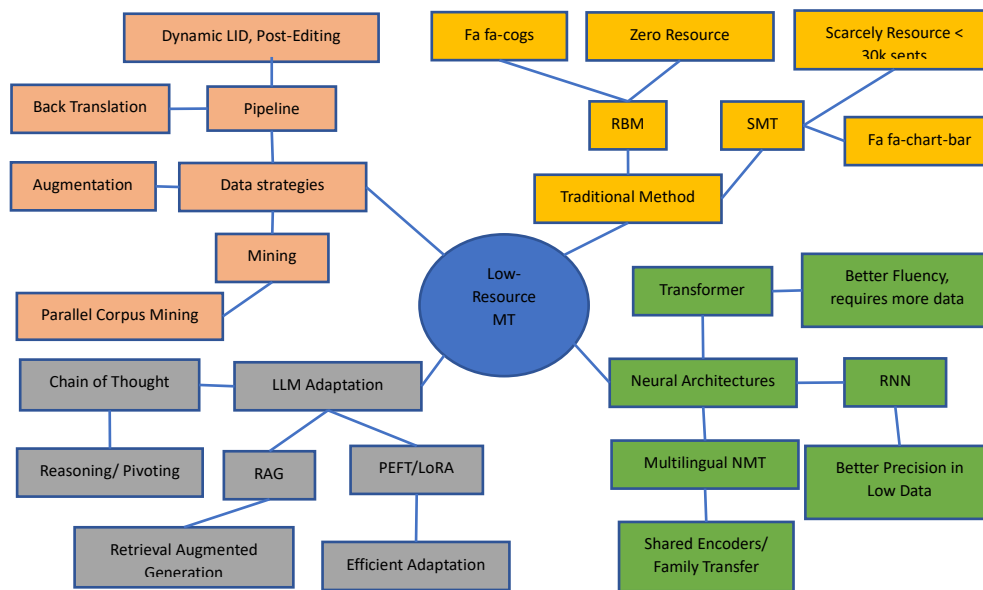


**Figure 1.** A hierarchy of machine translator systems to low-resource languages, in terms of architecture and resources required.

The [Figure 1](#) organises MT approaches into a hierarchy by architectural complexity and data requirements, with classical methods (RBMT, SMT) at the base requiring minimal data, and LLM-based approaches at the top requiring large corpora or synthetic augmentation.

**Table 1:** Resource-Tier Table

| Resource Tier | Data Available      | Recommended Methods           | Key References                                        |
|---------------|---------------------|-------------------------------|-------------------------------------------------------|
| Zero-resource | No parallel corpora | RBMT, SMT with seed lexicon   | [8] (Tulu), [10] (Kangri)                             |
| Low-resource  | < 30,000 pairs      | SMT, HTL, PEFT                | [4] (Nepali), [5] (Azerbaijani), [25] (Sinhala-Tamil) |
| Mid-resource  | 30k–100k pairs      | Transformer NMT, PEFT, mBART  | [12] (Manipuri), [23] (Biomedical)                    |
| Augmented     | 100k+ / synthetic   | Back-translation, RAG, MT-COT | [15] (Cherokee), [22] (Livonian)                      |

This taxonomy underpins the comparative analysis in Section 3 and the method recommendations in Section 6.

## 2.1 Foundational Paradigms:

Even in the case of languages that have no digital footprint to use in the digital learning methods, the traditional methods remain the most helpful first responders. Rule-based Machine Translation (RBMT) is a vital base in a scenario where the available data is very sparse. This was demonstrated by balipa et al. [8] in their work on Tulu, a Dravidian language having virtually no parallel corpora. Creating a small set of 1,010 words, they used explicit linguistic rules on their system and performed at 78.69 percent accuracy, which is far better than general-purpose systems such as Google Translate, which failed entirely with the language. In the same way, smaller training data of less than about 30,000 phrases tend to favor statistical Machine Translation (SMT) over neural Machine Translation (NMT). Roy et al. [4] experimented on SMT with NMT on a 17k dataset with English-Nepali language and found that SMT was significantly better than neural baselines with a BLEU score of 22.26. Chauhan et al. [10] reported comparable performances of the first Kangri corpus with SMT (BLEU 5.75) performing better than NMT (BLEU 4.23) confirming that neural models fail to generalise reliably under severe data sparsity.

## 2.2 Neural Revolution and Transfer Learning.

As resources exceed zero to low, the focus will be on changing the neural Machine Translation (NMT) architectures to achieve efficiency in data.

**A) Architectural trade-offs:** Mkhwanazi et al. [9] discovered a trade-off between fluency and accuracy in an experimental study of South African languages (isizulu, sesotho). The Bahdanau attention models based on RNN performed better on BLEU-1 scores (more accurate at the word level) with resource-constrained settings whereas transformer models were more accurate on BLEU-4 scores (more fluent).

**B) Multilingual and hybrid transfer:** Revanuru et al. [28] demonstrated that a typical encoder-decoder set up encourages positive transfer in related languages (like Indic languages). However, general transfer learning cannot be regarded as a panacea; singh et al. [12] found that in English-Manipuri, a trained transformer (from scratch) with less than ten times higher BLEU had to be nine times more effective than a fine-tuned multilingual mt5 model (BLEU 19.13). Maimaiti et al. [5], to minimize such negative transfer, developed hybrid transfer learning (HTL), where lexicon is constructed with common words of both parent and child languages and then fine-tuned. In languages with a high morphological level such as Azerbaijani and Uzbek, HTL generates significant benefits.

**C) Zero-resource and pivoting:** In order to enable zero-resource translation, firat et al. [31] designed multi-way, multi-lingual models, which exchange attention processes. Extending this, kumar et al. [14] proposed using continuous word vectors without parallel data to clustering models to linguistic diversity (e.g., Russian to Ukrainian), and lakew et al. [29] proposed to use self-learning to improve zero-shot performance.

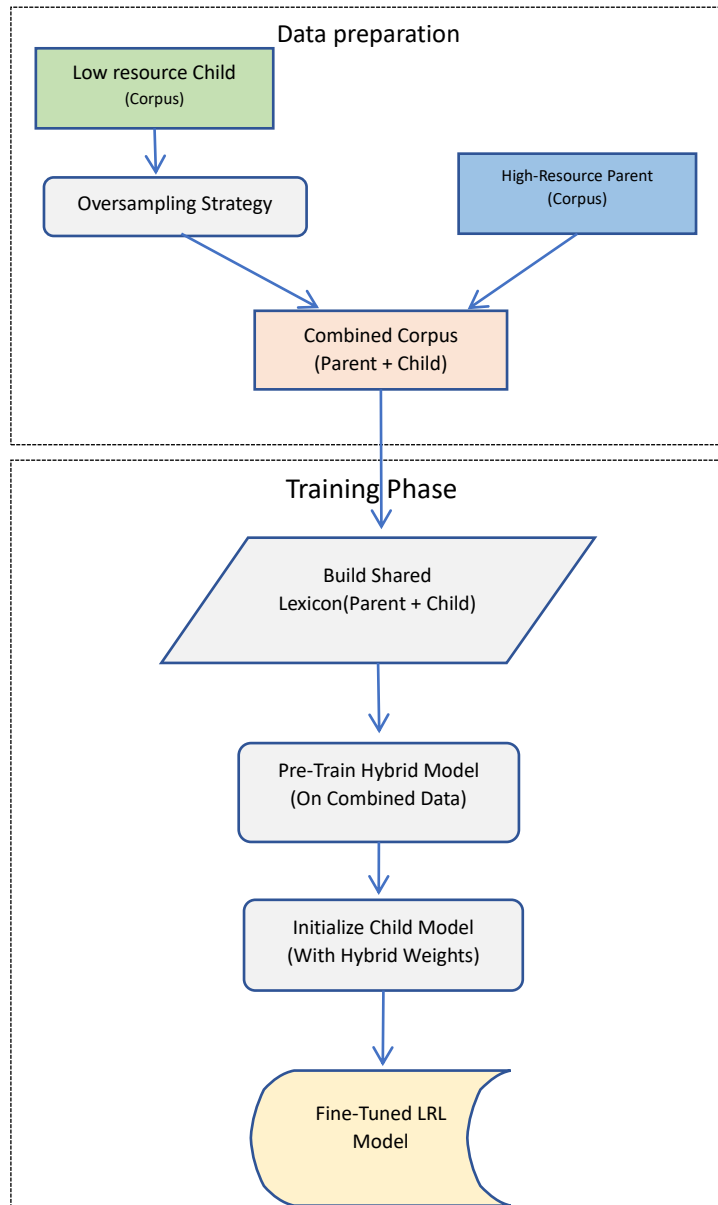


**Figure 2.** Flowchart of the Hybrid Transfer Learning (HTL) pipeline.

Flowchart of the Hybrid Transfer Learning (HTL) pipeline, in which a shared lexicon between a high-resource parent language and a low-resource child language is constructed prior to fine-tuning, reducing negative transfer caused by vocabulary mismatch. Unlike normal transfer learning, HTL avoids negative transfer due to vocabulary discrepancy through the construction of shared lexicon between high-resource (parent) and low-resource (child) languages before starting the processes. Large language models adaption (LLM) is a subfield of Machine Translation concerned with adapting the model to specific requirements.

### 2.3 Large language models adaptation (LLM).

Large Language Models Adaptation (LLM) is a branch of Machine Translation that focuses on the process of adapting the model to a particular need process of adapting large pre-trained models to low-resource languages (LRLs) without the high cost of re-training has become a recent topic of interest.

One of them is known as parameter-efficient fine-tuning (PEFT). In a detailed study of 15 PEFT topologies, su et al. [25] identified Housby adapters (which is the best approach to LRLs situations) outperforming full fine-tuning. This aligns with sel and hanbay [24] optimization of multilingual models to Turkish-English translation, which results in bottleneck adapter combinations with LoRA, which only requires modification of 5% of total parameters without affecting translation quality.

Liang et al. [1] extended language-specific fine-tuning with LoRA (LSFTL) to Asian LRLs, demonstrating that the large performance gains can be obtained even when using consumer-grade hardware. There is a parallel field of innovation in cognitive reasoning architectures. In languages that are under-represented in pre-training statistics, conventional prompting cannot often be effective. Yang et al. [3] introduced a solution called Machine Translation chain-of-thought (MT-CoT) of thai that is a reasoning pipeline with a pivot on the English language and generates the final output. Retrieval-augmented generation (RAG) has also been essential where there is limited cultural memory in model weights in languages. Shu et al. [15] applied to RAG to cherokee that the lawful translation requires the retrieval of the appropriate context. As this, tars et al. [22] showed how scant language priors could be compensated by external reasoning and retrieval processes through the use of m2m-100 with back-translation to learn a big model the endangered language of livonia.

Figure 3 illustrates the MT-CoT pipeline, in which the model first reasons in a high-resource pivot language (English) before producing the final low-resource output, showing the English-pivot reasoning step that precedes low-resource output generation [3].

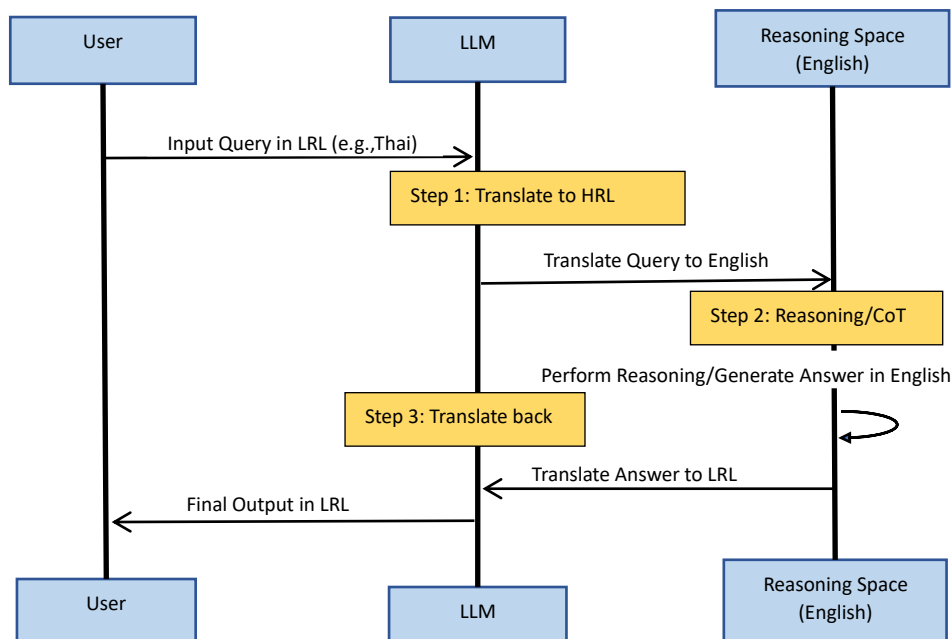


**Figure 3.** The Machine Translation Chain-of-Thought (MT-CoT).

The system will change to a high resources language (English) to execute logic and reasoning, then generate the final LRLS output to low-resource languages (LRLS) that does not have reasoning abilities in LLMs.

## 2.4 Data-Centric Strategies

Recent studies highlight that data quality is likely to be more important than model architecture, thus the necessity of strengthening pipeline components. One of such developments is dynamic identification and mining. Velpuri and sunkara [7] demonstrated that misdirecting noisy LRLS input is the main reason why translation fails; their dynamic language identification (LID) module improved BLEU scores by 6.2 points by successfully dealing with code-mixed and noisy text. Based on this, leong et al. [11] employed the transformer-based multilingual embeddings to extract parallel phrases through similar datasets to incorporate the unsupervised development of resources. Synthetic data production is another important field of development even though its effectiveness varies significantly depending on the field. Vipubmedt5 was pre-trained on 20 million back-translated medical abstracts, and phan et al. [23] added the targeted augmentation methods to replace the uncommon phrases in the synthetic context. Nevertheless, ghafoor et al. [13] warned that in semantic polarity in the semantic analysis of sentencing in urdu, synthetic translation may turn round, affecting the performance of the classifier. This is an alert to the fact that semantic coverage breakthrough can lead to pragmatic distortion. Prasada and rao [18] have found that back-translation was more beneficial in English-Kannada, but they stressed that severe filtering is required to avoid the accumulation of noise. Rakhimova et al. [19] developed a systematic post-editing (pe) methodology to Uzbek and Kazakh, employing human-in-the-loop verification to ensure morphological correctness in agglutinative languages, having discovered

the failures of fully automated MT.

### 3. Comparative Analysis

The extent of the parallel data that is availed to it is highly influential in the effectiveness of Machine Translation methods. Some of the core studies of zero-resource languages (where traditional methods are the rule) versus augmented situations (where LLMs and PEFT are the rule) are contrasted in the discussion below.

This involves comparing the performance of the methodologies across different approaches.

[Table 2](#) summarises the performance of the approaches reviewed, highlighting the crossover points where statistical methods outperform neural ones under data-scarce conditions.

**Table 2.** Comparative Performance of MT Methodologies on Low-Resource languages.

| Ref # | Method / Algorithm  | Language Pair     | Data set Size | Key Metrics                                          | Advantages                                                                 | Limitations                                                              |
|-------|---------------------|-------------------|---------------|------------------------------------------------------|----------------------------------------------------------------------------|--------------------------------------------------------------------------|
| [8]   | Rule-Based (RBMT)   | Hindi - Tulu      | ~1k sentences | Accuracy: 78.69%                                     | Effective for "digitally dark" languages; culturally accurate.             | Labor-intensive; rigid rules; difficult to scale.                        |
| [10]  | SMT vs. NMT         | Hindi - Kangri    | 27k sentences | BLEU: 5.75 (SMT) vs. 4.23 (NMT)                      | SMT outperforms NMT in extreme data scarcity (<30k).                       | Lower fluency; requires explicit feature engineering.                    |
| [4]   | SMT (Moses)         | English - Nepali  | 17k sentences | BLEU: 22.26                                          | Outperforms Google Translate (GNMT) on specific domains.                   | Struggles with long-range dependencies compared to neural models.        |
| [9]   | RNN vs. Transformer | IsiZulu - Sesotho | Low (<50k)    | BLEU-4: Transformer > RNN; BLEU-1: RNN > Transformer | Transformers offer better fluency; RNNs offer better word-level precision. | Transformers require more data to converge; RNNs face vanishing gradient |

|      |                         |                       |                 |                                          |                                                                                  |                                                                          |
|------|-------------------------|-----------------------|-----------------|------------------------------------------|----------------------------------------------------------------------------------|--------------------------------------------------------------------------|
|      |                         |                       |                 |                                          |                                                                                  | s.                                                                       |
| [12] | Supervised vs. Transfer | English - Manipuri    | 69k sentences   | BLEU: 29.05 (Supervised) vs. 19.13 (mT5) | Dedicated models beat generic pre-training for unseen languages.                 | Transfer learning suffers from negative transfer without adaptation.     |
| [5]  | Hybrid Transfer (HTL)   | English - Azerbaijani | ~20k (Child)    | +4.94 BLEU over baseline                 | Shared lexicon prevents negative transfer in morphologically rich LRLs.          | Computationally complex training pipeline (Parent + Child).              |
| [25] | PEFT (Houlsby Adapter)  | Sinhala - Tamil       | 100k sentences  | 33.34 BLEU (vs. 30.25 Full FT)           | Superior performance with fraction of parameters; highly efficient.              | Performance varies by adapter architecture; requires base model support. |
| [23] | Synthetic Pre-training  | English - Vietnamese  | 20M (Synthetic) | SOTA on Bio benchmarks                   | Large-scale synthetic data is effective for scientific/semantic domains.         | Synthetic data poses risks for pragmatic/sentiment tasks [13].           |
| [15] | LLM + RAG               | English - Cherokee    | Zero-shot       | Significant Gain vs. ~0 BLEU             | Essential for "digitally young" languages lacking representation in LLM weights. | Reliance on retrieval accuracy; higher latency.                          |

### 3.1 Critical Synthesis

The comparison data shows the difference between different operational areas of technologies:

- A) The 30k sentences data crossover threshold:** An empirical analysis of both empirical studies of Kangri [10] and Nepali [4] Indicates that there is a critical threshold of approximately 30000 sentence pairs per 1 unit of time. This is the lowest threshold at which neural Machine Translation (NMT) models can converge or generalize, making these models have low BLEU scores compared to statistical Machine Translation (SMT). SMT remains a superior choice when dealing with scarcely resourced languages, since its probabilistic models are more repellent to sparsity in input than the deep neural networks.
- B) The lexical gap in transfer learning:** Despite all the promotion of transfer learning as a solution, it is not a panacea. As it has been observed by singh et al. [12], a trained-from-scratch supervised transformer worked significantly better with manipuri than with a refined mt5 model. This implies that in cases where the target language (such as manipuri) was not met during pre-training, massive multilingual models are subjected to negative transfer. To solve this, maimaiti et al. [5] employed a “hybrid” approach, which successfully addressed the lexical divide, achieved a bilingual common ground between a high-resource parent (turkish) and a low-resource kid (Uzbek) and then proceeded to refine it.
- C) Efficiency vs. Performance in adaptation:** Parameter-efficient fine-tuning (PEFT) has been demonstrated to be the most appropriate method with languages and small resources, e.g. Sinhala and tamil. Some adaptor topologies, like Houlsby, are able to perform equally well when updating only some of the parameters, as shown by su et al. [25]. This negates the stereotype that bigger is better by showing that in the case of LRLs, personalized adaptation proves to be more effective than brute-force training.
- D) The role of reasoning and retrieval:** standard prompting yields nearly no results to languages that are either "digitally young" or not a language that has been exposed to LLMs, like cherokee. To achieve real translations, both shu et al. [15] and yang et al. [3] demonstrate that LLMs have to be enhanced with retrieval-augmented generation (RAG) or chain-of-thought (cot) reasoning, which translates to: think then generate in English. Consequently, inference-time augmentation gains more significance than training the model.

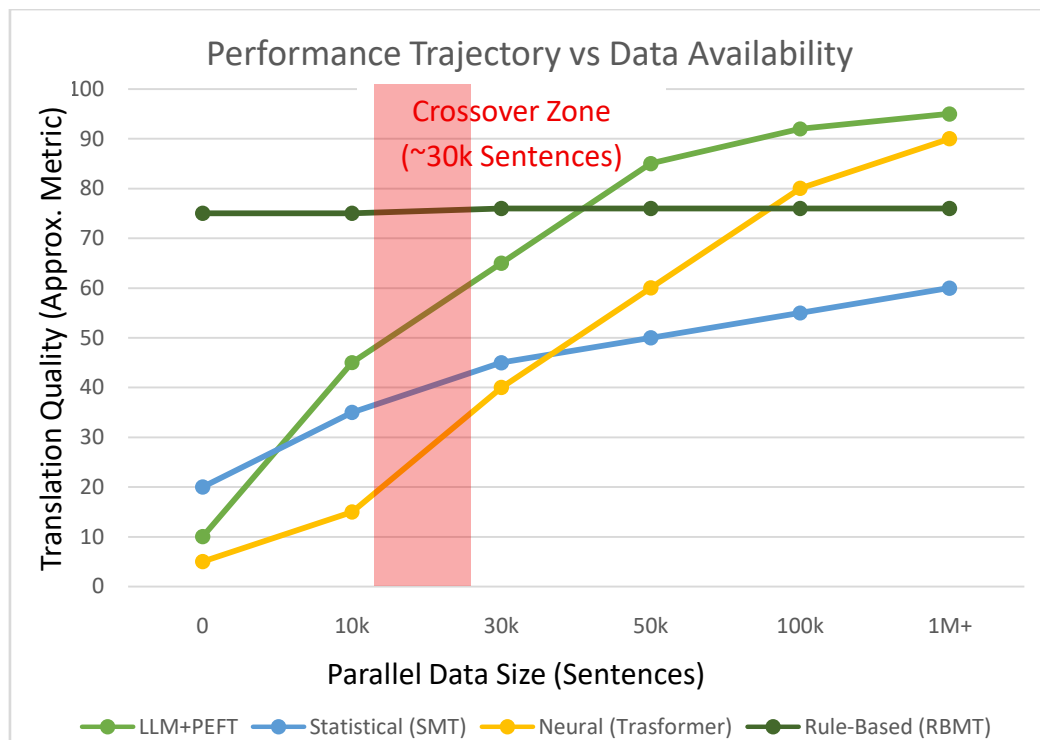


**Figure 4.**Theoretical performance curve illustrating the Hierarchy of Suitability

Theoretical performance curve illustrating the 'Hierarchy of Suitability', based on empirical observations from Roy et al. [4] (English-Nepali, 17k pairs) and Chauhan et al. [10] (Hindi-Kangri, 27k pairs). The crossover point, where NMT begins to

outperform SMT, occurs between 30,000 and 50,000 sentence pairs. The classical methods (RBMT/SMT) outperform neural ones when the scarcity is high, but as the volume of data increases, neural methods prevail over classical ones. The crossover point typically occurs after sentence 30,000 to 50,000.

#### **4. Challenges and Open Issues**

The extensive application of the Machine Translation (mt) to the low-resource languages (LRLs) is affected by a series of serious barriers, regardless of the rapid development of neural networks. This research identifies four primary problems, which transcend model architectures.

##### **4.1 The paradox of synthetic data.**

Even though the back-translation is one of the most widespread data augmentation methods of low-resource languages (LRLs), its performance differs significantly across domains. Artificial data can significantly enhance model behaviour in technical and scientific applications. Indicatively, phan et al. [23] demonstrated that training the vipubmedt5 model on 20 million biomedical abstracts that were artificially translated gave the best possible results as the translation process did not impair the semantic integrity of biomedical terminology specific to the field. But its effectiveness plummets in those activities which are pragmatically rich or emotion-based. This observation leads to what can also be described as a synthetic data paradox: synthetic corpora can be very reliable in fact-based, semantically accurate contexts, but are highly dangerous when the environment is sentiment-laden and subtle pragmatic actions are readily overlooked or misinterpreting in the process of translation.

##### **4.2 Fragility of pipe: Finding and tokenization.**

There is a common failure in standard Machine Translation (MT) pipelines before the translation model is even invoked, and this is caused by upstream processing problems. Identifying language (lid) is an important issue: the noisy or code-mixed LRLs input can frequently be misclassified by the static lid systems, therefore, passing through incorrect translation models. The dependability upstream is equally important to the translation model as it is being demonstrated by velpuri and sunkara [7] through the implementation of a dynamic lid module that significantly enhanced the routing accuracy and induced a 6.2 point improvement in BLEU score. Another common problem is the tokenization mismatch. As it is observed by sindhujan et al. [27] and merx et al. [26], standard tokenizers that are trained mainly on high-resource languages often divide LRLs words into semantically nonsensitive sub-tokens. In morphologically rich languages such as tamil and tetun, such an issue is particularly detrimental as morphological boundaries are skewed by wrong segmentation which leads to hallucinations, unsuccessful semantic alignment, and overall reduced quality of translation.

##### **4.3 The domain and lexical gap**

The data that is available to train the model is often quite different than that of the actual expenses of low-resource language (LRLs) populations. Problem of domain is central: merx et al. [26] have observed that despite the fact that the existing corpora are highly biased towards government and religious resources, the actual tetun user queries are predominantly related to health and education. This discrepancy leads to models that do a superb job in translating information that is more akin to the parliaments or the formality but fails when it comes to more practical needs such as simple medical instructions which are also emphasized in the review by chu and wang [32] on domain adaptation. Another major impediment is the lexical gap. The research work by singh et al. [12] and maimaiti et al. [5] shows that such an assumption fails in the case of other languages such as manipuri and Uzbek, though the traditional transfer learning presumes that there is a degree of lexical overlap between the source and the destination languages. In such cases large models with multiple languages often suffer negative transfer and perform better as compared to smaller supervised models which are trained initially. To overcome this, special strategies are required to fill lexical gaps and prevent the drop in translation performance, e.g. Hybrid transfer learning, lexicon induction, provided by hu et al. [33].

#### **5. Future research directions**

The analysis of current methods finds several interesting avenues to further study, beyond simple model scaling to more advanced, resource conscious methods. The future research will have to seek out other forms of data sources as more and more of the low-resource languages (LRLs) approach saturation. One of the interesting methods is unsupervised corpus mining. Leong et al. [11] demonstrated the effectiveness of multilingual embeddings in extracting parallel phrases using similar corpora, e.g. Multilingual news websites that report on the same events. It should be further developed into a strategy called digitally dark languages [2] using multimodal alignment, which forms parallel corpora of predominantly oral languages

through audio recording of podcasts, community archives and radio broadcasts and matching against available text. The other new method is cross-lingual knowledge distillation. Future systems ought to explore the ways the abilities of large, multilingual models (such as NLLB) can be tiled into small, domain-specific student models, which can be tailored to specific language groups, as Revanuru et al. Do [28]. These are deliberately specified designs, which are modeled by a *bantubert* or *dravidiangpt*, that offer a scalable algorithm to improve LRLs translation with reduced processing demands.

### 5.1 Hybrid neuro-symbolic architectures.

With the existing good performance of rule-based Machine Translation (RBMT) on the actual zero-resource language, the hybrid approach is becoming increasingly significant [8]. One of such directions is the inclusion of linguistic rules in the neural Machine Translation (NMT). To enhance stability in agglutinative languages such as the Kazakh and Uzbek, and reduce the morphological hallucinations common in purely neural systems, future efforts need to explore systems where explicit linguistic representations, e.g. Syntax trees and morphology tables, can guide the neural decoder [19]. Another interesting avenue is dynamic adapter fusion. It is possible that future models may be able to dynamically activate some low-rank adaptation (LoRA) components depending on input dialect, domain, or pattern of code-switching instead of relying on fixed adapter modules. Augmentation of cognitive: this method aims to enhance the mental capabilities of an LLM to allow it to acquire more knowledge, think more creatively, and operate more naturally and adaptively.

### 5.2 Cognitive augmentation of LLMs:

This approach seeks to make an LLM more capable of learning, more creative in its thinking, and more natural and adaptive, by enhancing its mental capabilities. Inference-time computation can replace large training corpora well in the case of languages that are digitally young and in which pretrained LLM weights do not contain any or contain only small amounts of linguistic information. One of the possible approaches is multilingual reasoning chains. Future studies need to make the method of pivoting through reasoning more formal, modifying the work of yang et al. [3], whereby the models are first requested to explain the structure of the sentence in the LRLs, followed by mapping it to conceptual analogies in the high-resource language, such as English, and only once that is done, generate the final translation. Chain-of-thought-based approach may be useful in filling performance gaps in languages with complex syntactic structures and limited training data. One more free method is retrieval-centric translation. As demonstrated by shu et al. [15], retrieval-augmented generation (RAG) is quite beneficial to LRLS. The next generation systems would help decrease the amount of hallucination and increase the quality of translations by creating dense vectors indexes of low-resource language that would enable models to reconstruct rare words, idiomatic expressions, or culturally specific terms on inference.

## 6. Conclusion

One important milestone has been achieved in the endeavor to find a universal machine translator. Though bigger is better mythology prevails in the region, the scale does not matter as much as the appropriateness of the big tail of human languages, as demonstrated in this analysis. Our theme analysis demonstrates a non-linear hierarchy of answers. Conventional statistical and rule-based models are important first responders that outperform deep learning models on zero-resource languages such as Tulu and Kangri, as opposed to being old-fashioned artifacts. Targeted adaptation, which applies hybrid transfer learning or parameter-efficient fine-tuning (PEFT) to bridge the lexical gap, is more effective compared to blind pre-training of low-resource languages that have some digital presence (e.g., sinhala, manipuri). Also, we discover a synthetic data paradox: synthetic data has the potential to achieve state-of-the-art performance in science fields (such as *vipubmedt5*) but is posing severe threats in semantic/pragmatic fields (such as *urdu sentiment analysis*). This highlights the relevance of domain-conscious data methods. Ultimately, there is a need to swap a model-based view of the world to a data-driven ecology to bridge the language divide. The bigger transformers will not be the only ones that could make it; the trusted pipelines that would presuppose the automated corpus mining, dynamic language recognition, and cultures-specific evaluation metrics will also play an important role. By embracing such a diverse kit, the NLP community can be on the right track to achieving the goal of leaving no language behind, and it goes from the statistical to cognitive.

## References

- [1]. Liang, X., Khaw, Y.-M., Liew, S.-Y., Tan, T.-P., Qin, D.: Toward low-resource languages Machine Translation: a language-specific fine-tuning with LoRA for specialized large language models. *IEEE Access* **1**, 1 (2025)
- [2]. Sun, H., Tang, T., Feng, J., Yang, Z., Guo, B.: Overview of research on low-resource language Machine Translation based on artificial intelligence. In: *Proceedings of the 2025 6th International Conference on Computer Engineering and Application (ICCEA)*, pp. 1–5. IEEE (2025)
- [3]. Yang, H., Zhang, M., Guo, J.: From scarcity to sufficiency: Machine Translation techniques for low-resource LLMs enhancement. In: *Proceedings of the 2024 2nd International Conference on Foundation and Large Language Models (FLLM)*, pp. 36–41. IEEE (2024)
- [4]. Roy, A., Purkayastha, B., Paul, S.: A bidirectional statistical Machine Translation system for exploring the performance of the low resource language pair English-Nepali. In: *Proceedings of the 2024 International Conference on Computational Intelligence, Geospatial Systems and Technologies (ICCIGST)*, pp. 1–6. IEEE (2024)
- [5]. Maimaiti, M., Liu, Y., Luan, H., Sun, M.: Enriching the transfer learning with pre-trained lexicon embedding for low-resource neural Machine Translation. *Tsinghua Science and Technology* **27**, 150–163 (2021)
- [6]. De Silva, D., Hansadi, P.: A comparative analysis of Machine Translation approaches for low-resource languages and future enhancements. In: *Proceedings of the 2024 International Conference on Advanced Technologies in Computing (ICATC)*, pp. 1–6. IEEE (2024)
- [7]. Velpuri, S., Sunkara, S.: Improving translation accuracy through advanced language detection. In: *Proceedings of the 2025 5th International Conference on Soft Computing for Security Applications (ICSCSA)*, pp. 897–901. IEEE (2025)
- [8]. Balipa, M., K, A., Ananya, Murugappan: A rule-based Machine Translation framework for low-resource language pairs. In: *Proceedings of the 2025 4th International Conference on Sentiment Analysis and Deep Learning (ICSADL)*, pp. 969–974. IEEE (2025)
- [9]. Mkhwanazi, P., Zwane, S., Adigun, M., Thwala, E.: Comparative evaluation of RNN-based Bahdanau attention vs transformer model for NMT in South African low-resource languages. In: *Proceedings of the 2025 International Conference on Trends in Artificial Intelligence and Security (ICTAS)*, pp. 1–6. IEEE (2025)
- [10]. Chauhan, S., Saxena, S., Daniel, P.: Monolingual and parallel corpora for Kangri low resource language. *arXiv preprint arXiv:2103.11596* (2021)
- [11]. Leong, C., Liu, X., Wong, D., Chao, L.: Exploiting translation model for parallel corpus mining. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 1 (2021)
- [12]. Kshetrimayum, B., Singh, N.A., Meetei, L.S., Singh, N.J., Bandyopadhyay, S., Singh, T.D.: A comparative study of transformer and transfer learning based MT models for English-Manipuri. In: *Proceedings of the 20th International Conference on Natural Language Processing (ICON)*, pp. 375–380 (2023)
- [13]. Ghafoor, A., Imran, A., Daudpota, S., Kastrati, Z., Abdullah, Batra, R., Wani, M.: The impact of translating resource-rich datasets to low-resource languages through multi-lingual text processing. *IEEE Access*, 1 (2021)
- [14]. Kumar, S., Anastasopoulos, A., Wintner, S., Tsvetkov, Y.: Machine Translation into low-resource language varieties. *arXiv preprint arXiv:2106.06797* (2021)
- [15]. Shu, P., Chen, J., Liu, Z., Wang, H., Wu, Z., Zhong, T., Li, Y., Zhao, H., Jiang, H., Pan, Y., Zhou, Y., Owl, C., Zhai, X., Liu, N., Saunt, C., Liu, T.: Transcending language boundaries: harnessing LLMs for low-resource language translation. *arXiv preprint arXiv:2411.11295* (2024)
- [16]. Thillainathan, S., Nayak, S., Ranathunga, S., Adelani, D., Su, R., McCarthy, A.: Pre-trained multilingual sequence-to-sequence models: a hope for low-resource language translation? *arXiv preprint arXiv:2203.08850* (2022)
- [17]. Karakanta, A., Dehdari, J., Genabith, J.: Neural Machine Translation for low-resource languages without parallel corpora. *Machine Translation* **32**, 1 (2018)
- [18]. Prasada, P., Vishwanatha, M., Rao, P.: Reinforcement of low-resource language translation with neural Machine Translation and backtranslation synergies. *IAES International Journal of Artificial Intelligence (IJ-AI)* **13**, 3478–3488 (2024)
- [19]. Rakhimova, D., Karibayeva, A., Turarbek, A.: The task of post-editing Machine Translation for the low-resource language. *Applied Sciences* **14**, 486 (2024)
- [20]. Kumar, R., Jha, P., Sahula, V.: An augmented translation technique for low resource language pair: Sanskrit to Hindi translation. *arXiv preprint arXiv:2006.08332* (2020)
- [21]. Shi, S., Wu, X., Su, R., Huang, H.: Low-resource neural Machine Translation: methods and trends. *ACM Transactions on Asian and Low-Resource Language Information Processing* **21** (2022)
- [22]. Tars, M., Purason, T., Tättar, A.: Teaching unseen low-resource languages to large translation models. In: *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pp. 375–380. Association for Computational Linguistics, Abu Dhabi (2022)
- [23]. Phan, L., Dang, T., Tran, H., Trinh, T.H., Phan, V., Chau, L.D., Luong, M.-T.: Enriching biomedical knowledge for low-resource language through large-scale translation. In: *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 3131–3142. Association for Computational Linguistics, Dubrovnik (2023)
- [24]. Sel, I., Hanbay, D.: Efficient adaptation: enhancing multilingual models for low-resource language translation. *Mathematics* **12**, 3149 (2024)
- [25]. Su, T., Peng, X., Thillainathan, S., Guzmán, D., Ranathunga, S., Lee, E.A.: Unlocking parameter-efficient fine-tuning for low-resource language translation. In: *Findings of the Association for Computational Linguistics: NAACL 2024* (2024)
- [26]. Merx, R., Correia, A.J.G., Suominen, H., Vylomova, E.: Low-resource Machine Translation: what for? who for? An observational study on a dedicated Tetun language translation service. In: *Proceedings of the First Workshop on Language Models for Low-Resource Languages (LoResLM)* (2025)

- [27]. Sindhujan, A., Kanojia, D., Orasan, C., Qian, S.: When LLMs struggle: reference-less translation evaluation for low-resource languages. In: Proceedings of the First Workshop on Language Models for Low-Resource Languages (LoResLM) (2025)
- [28]. Lu, K., Yang, Y., Yang, F., Dong, R., Ma, B., Aihemaiti, A., Atawulla, A., Wang, L., Zhou, X.: Low-resource language expansion and translation capacity enhancement for LLM: a study on the Uyghur. In: Proceedings of the 31st International Conference on Computational Linguistics (COLING) (2025)
- [29]. Elsner, M., Needle, J.: Translating a low-resource language using GPT-3 and a human-readable dictionary. In: Proceedings of the 20th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology (2023)
- [30]. Agrawal, A.S., Fazili, B., Jyothi, P.: Translation errors significantly impact low-resource languages in cross-lingual learning. In: Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (EACL) (2024)
- [31]. Haque, R., Hasanuzzaman, M., Way, A.: Terminology translation in low-resource scenarios. *Information***10**(9)(2019)
- [32]. Merx, R., Suominen, H., Hong, L.Y., Thieberger, N., Cohn, T., Vylomova, E.: TULUN: transparent and adaptable low-resource Machine Translation. In: Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (ACL) (2025)
- [33]. Khoboko, P.W., Marivate, V., Sefara, J.: Optimizing translation for low-resource languages: efficient fine-tuning with custom prompt engineering in large language models. *Machine Learning with Applications* **27** (2025)
- [34]. Chakrabarty, A., Dabre, R., Ding, C., Utiyama, M., Sumita, E.: Low-resource multilingual neural translation using linguistic feature-based relevance mechanisms. *ACM Transactions on Asian and Low-Resource Language Information Processing* **22**(5) (2023)
- [35]. Akallouch, O., Fardousse, K.: In-context learning for low-resource Machine Translation: a study on Tarifit with large language models. *Algorithms* **18**(8) (2025)