GARuD: Guided Alignment of Representations using Distillation for Ultra-Low-Resource Languages

Debarchan Basu[1]*, Shashwat Bharadwaj[1]*, Vaibhav Sharma[2], Pooja Singh[2], Sandeep Kumar[1,2,3]

[1]Yardi School of Artificial Intelligence, Indian Institute of Technology, Delhi, IN
[2]Department of Electrical Engineering, Indian Institute of Technology, Delhi, IN
[3]Bharti School of Telecommunication, Indian Institute of Technology, Delhi, IN
{aiy238072, aiy237528, eey257541, eez228470, ksandeep}@iitd.ac.in

Abstract

The vast majority of the world's languages, particularly low-resource and tribal ones like Bhili, remain critically underserved by modern language technologies. The primary bottleneck is the lack of large-scale corpora required for standard pre-training. To address this gap, we introduce cross-lingual contrastive distillation, a novel and data-efficient, compute-efficient paradigm for creating powerful language models without a massive monolingual corpus. Our method adapts a pre-existing multilingual model (MuRIL) by using a fixed, expert teacher model (HindBERT) to distill semantic knowledge from a related high-resource language (Hindi) via a contrastive objective on a modest parallel corpus. Through comprehensive experiments, we show that our resulting model, GARuD-Bhili, significantly outperforms strong zero-shot and MLM-only baselines on a suite of evaluations, including intrinsic language modeling, downstream sentiment analysis, and cross-lingual benchmarks (Tatoeba, XNLI). Our work presents a scalable blueprint for linguistic empowerment, offering a practical pathway to develop robust language technologies for other underserved languages globally.

1 Introduction

India boasts remarkable linguistic diversity, with the 2011 census documenting over 1,369 distinct mother tongues and 22 official languages. Yet, this linguistic wealth faces a critical challenge: numerous indigenous and tribal languages are critically endangered due to a severe lack of digitized resources and parallel corpora. These languages are vital to cultural heritage, making robust translation systems essential not only for clear communication and social inclusion but also for equitable access to government services. Effective translation is crucial for implementing policies, welfare programs, legal processes, and educational efforts, thereby supporting national unity in India's complex multilingual setting.

The advent of large pre-trained language models like BERT (Devlin et al., 2019) has revolutionized NLP by learning rich contextual representations from vast amounts of text. For the Indian subcontinent, specialized multilingual models like MuRIL (Khanuja et al., 2021) have been developed to capture the shared linguistic and structural properties across 17 major Indian languages. However, the effectiveness of these powerful models does not extend to the hundreds of languages, like Bhili, that were absent from their pre-training corpora. The primary bottleneck is the data itself: developing a language model for Bhili from scratch is infeasible due to the non-existence of a large monolingual corpus required for standard self-supervised pre-training. This creates a critical gap, leaving Bhili speakers digitally disenfranchised.

To bridge this representation gap with the limited corpus we have, we propose a novel and data-efficient training paradigm leveraging a modest-sized, community-curated Bhili-Hindi Parallel Corpus of approximately 142,000 sentence pairs. We adapt the powerful MuRIL model through a process of cross-lingual distillation, using a framework inspired by the mechanics of Momentum Contrastive Learning (MoCo) (He et al., 2019). In this setup, the MuRIL model acts as a student, which learns to represent Bhili sentences. Its goal is to align its outputs with the representations from a fixed teacher model that is already proficient in Hindi. By treating a Bhili sentence and its Hindi translation as a positive pair within this contrastive objective, we compel the student model to map Bhili into the teacher's rich, pre-existing semantic space. This allows for an efficient and targeted transfer of knowledge to understand the unique lexical and grammatical nuances of Bhili.

[*] Authors 1 and 2 contributed equally.

Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics, pages 1867–1880
December 20-24, 2025 ©2025 Association for Computational Linguistics

2 Related Work

Our work is situated at the intersection of three key areas in modern NLP: adapting large multilingual models, the paradigm of knowledge distillation, and the methodology of cross-lingual contrastive learning.

2.1 Adaptation of Multilingual Models for Low-Resource Languages

The dominant approach for supporting low-resource languages is to adapt large, pre-trained multilingual models like mBERT (Devlin et al., 2019) or XLM-R (Conneau et al., 2020). For the Indian context, region-specific models such as MuRIL (Khanuja et al., 2021) and IndicBERT (Doddapaneni et al., 2023) have proven to be more effective starting points, as their pre-training on diverse Indic languages helps capture the subcontinent's shared typological features.

However, even these specialized models face significant challenges. The most common adaptation strategies include:

- **Zero-Shot Transfer**: This involves directly applying a pre-trained model to a language it has never seen. While simple, performance is often poor due to vocabulary mismatch and the model's lack of exposure to the target language's specific syntax and semantics. This degradation is particularly pronounced for common linguistic phenomena in the Indian context like code-mixing and transliteration (Khanuja et al., 2021; Joshi et al., 2020).

- **Continued Pre-training**: This strategy involves further training the multilingual model on a monolingual corpus of the target language using an objective like Masked Language Modeling (MLM). While a strong baseline method for adapting a model's vocabulary and representations (Adelani et al., 2022), it requires a substantial monolingual corpus, which is often unavailable for extremely low-resource languages like Bhili.

The limitations of these standard approaches necessitate exploring more data-efficient adaptation techniques that can explicitly teach cross-lingual alignment.

2.2 Knowledge Distillation as a Path for Transfer

Knowledge Distillation (KD) offers a powerful framework for knowledge transfer. Originally proposed for model compression, KD involves training a smaller ”student” model to mimic the behavior of a larger, more capable ”teacher” model (Hinton et al., 2015; Sanh et al., 2019). This principle can be adeptly repurposed for cross-lingual learning. The goal shifts from merely compressing knowledge to transferring it across a language barrier. In this setup, a student model learning a low-resource language (e.g., Bhili) is trained to align its representations with those produced by an expert teacher model for a high-resource language (e.g., Hindi). This provides a rich, stable supervisory signal, guiding the student's learning in an otherwise data-scarce environment.

2.3 Cross-Lingual Contrastive Learning

Recently, contrastive learning has emerged as a state-of-the-art method for learning high-quality sentence embeddings by pulling similar inputs together and pushing dissimilar ones apart. In the cross-lingual domain, this is powerfully realized by using parallel sentence pairs from different languages as positive examples. Leading frameworks like LaBSE (Feng et al., 2020), InfoXLM (Chi et al., 2021), and XLCo (Wang et al., 2022) use this principle to learn a shared, language-agnostic embedding space.

For low-resource scenarios, this contrastive paradigm offers compelling advantages over traditional MLM-based pre-training. By explicitly leveraging a parallel signal, contrastive methods can achieve robust cross-lingual alignment with significantly less data and compute. While MLM provides powerful generalized representations when data is abundant, contrastive learning is often a more direct and efficient path to strong cross-lingual understanding when data is scarce.

Our work synthesizes these threads in a novel way. We argue that for extremely low-resource languages like Bhili, the most effective strategy is neither continued pre-training (due to data scarcity) nor training a large contrastive model from scratch (due to computational cost). Instead, we combine the strengths of knowledge distillation and cross-lingual contrastive learning. We propose a highly efficient adaptation framework where we use a MoCo-style contrastive objective to distill the knowledge from a fixed, expert teacher model into an existing, powerful student model (MuRIL). This specific configuration–using teacher-student contrastive distillation to adapt, rather than retrain, a multilingual model–represents a distinct, practical, and highly effective methodology for resource-scarce environments.

3 Dataset

3.1 Dataset Creation and Purpose

Developing foundational language technologies for extremely low-resource languages is fundamentally constrained by the scarcity of high-quality, digitized text (Nekoto et al., 2020). This data gap has historically hindered the creation of robust language models for communities like the Bhili speakers. To address this challenge for Bhili, a culturally significant tribal Indic language, a large-scale, gold-standard Bhili-Hindi Parallel Corpus (Singh et al., 2025) was previously curated through extensive community-led initiatives involving native speakers. This dataset is the cornerstone of our study, it provides the essential cross-lingual signal required to train a powerful Bhili language representation model via our proposed knowledge distillation framework.

The source sentences in Hindi were primarily drawn from the Bharat Parallel Corpus Collection (BPCC) (Gala and Chitale, 2023), and supplemented with text from diverse public sources, including Legislative Assembly Speeches (Siripragada et al., 2020), the PMIndia corpus (Haddow and Kirefu, 2020), and NCERT textbooks. A team of 10 professional translators, all native Bhili speakers, meticulously translated these sentences between May 2024 and March 2025, contributing over 27,000 hours of work to ensure high semantic fidelity and contextual accuracy. This rigorous process yielded the high-quality parallel corpus that enables our work. To assess the quality of our manually curated Hindi–Bhili gold references, we conducted an inter-annotator agreement (IAA) study. For more details, refer to Appendix E.

3.2 Data Preprocessing

To ensure the quality of the data used for training, the corpus underwent a thorough preprocessing pipeline. This process involved normalizing Bhili homophones to maintain linguistic consistency, removing extraneous characters, and implementing strict de-duplication to eliminate highly similar sentence pairs. We enforced sentence length limits, retaining only pairs with 6 to 80 words. To further reduce redundancy, we applied cosine similarity filtering on sentence embeddings to detect and remove near-duplicate source-target combinations. Tokenization was performed using the original SentencePiece vocabulary from the MuRIL base model to ensure alignment with the model's pre-existing embeddings.

3.3 Dataset Statistics

A statistical overview of the Bhili-Hindi Parallel Corpus is presented in Table 1. The final dataset contains 142,817 sentence pairs, partitioned into a training set of 128,535 and a test set of 14,282 pairs.

Analysis reveals linguistic characteristics relevant to our modeling task. Bhili sentences are, on average, longer than their Hindi counterparts (e.g., ≈34 vs. ≈27 tokens in the training set), suggesting structural differences that a model must learn to align. The substantial number of unique tokens in both languages (Hindi: ≈27K, Bhili: ≈26.8K) highlights a rich vocabulary, which is vital for learning nuanced language representations. Furthermore, the high standard deviation in sentence length points to considerable variation in complexity, making the dataset a challenging and representative resource for building and evaluating a foundational language model for Bhili.

Statistic	**Train Set**		**Test Set**	
	Hindi	**Bhili**	**Hindi**	**Bhili**
Number of Sentence Pairs	128,535		14,282	
Number of Sentences	128,535	128,535	14,282	14,282
Total Tokens	3,473,762	4,419,784	346,759	449,280
Unique Tokens	27,215	26,774	14,312	14,582
Avg. Sentence Length (Tokens)	27.03	34.39	24.28	31.46
Min. Sentence Length (Tokens)	1.00	1.00	1.00	1.00
Max. Sentence Length (Tokens)	390.00	509.00	159.00	181.00
Std. Dev. Sentence Length (Tokens)	14.75	18.49	12.43	16.28
Overall Unique Tokens in Dataset				
Hindi: 27,215				
Bhili: 26,872				

Table 1: Key statistics of the Bhili-Hindi Parallel Corpus.

4 Methodology

Our objective is to develop a powerful contextual representation model for Bhili, an extremely low-resource language. We leverage the significant typological and lexical overlap between Bhili and Hindi by adapting a strong Indic multilingual model, MuRIL (Khanuja et al., 2021). Since MuRIL is pre-trained on 17 Indic languages including Hindi, it provides a robust starting point with a shared sub-word vocabulary.

We formulate our investigation as a comparative study between two adaptation approaches on our Bhili-Hindi Parallel Corpus (BHPC): a standard baseline and our proposed contrastive distillation method.

4.1 Baselines

We establish two baseline models to contextualize the performance of our proposed approach.

4.1.1 Zero-Shot Inference

Our first baseline measures the zero-shot performance of the original, off-the-shelf MuRIL model. This involves no additional training or fine-tuning whatsoever. The model is evaluated directly on the downstream Bhili tasks. This approach establishes the lower-bound performance and quantifies the model's out-of-the-box capabilities (or lack thereof) for a language it has never seen, relying solely on its cross-lingual transfer abilities learned during pre-training.

4.1.2 Continued Pre-training with MLM

As a strong baseline, we perform continued pre-training on the base MuRIL model using a standard Masked Language Modeling (MLM) objective. This adaptation approach, common for low-resource languages, fine-tunes the model to the specific vocabulary and syntax of the target language.

For this setup, we use only the Bhili portion of our parallel corpus ($\mathbf{x}_{\text{Bh}}$). The model is trained on a masked token prediction task using a cross-entropy loss. Details of training are provided in Appendix F.

4.2 Proposed Method: GARuD (Contrastive Distillation with a Memory Queue)

To learn richer, semantically-aware sentence embeddings, we introduce GARuD- a novel training framework that synergizes MLM with a cross-lingual contrastive objective. The goal is to force the sentence representations of Bhili to align with the high-quality semantic space of Hindi, a language for which strong models exist. This contrastive alignment acts as a powerful semantic regularizer, guiding the model to learn meaningful sentence-level representations beyond simple token prediction (Chen et al., 2023; Miao et al., 2024).

Our framework is a teacher-student knowledge distillation setup, whose mechanics are inspired by Momentum Contrast (MoCo) (He et al., 2019). It consists of three main components:

- A Student Encoder ($\text{ENC}_{\text{student}}$): The base MuRIL model, whose parameters (Θ_S) we are training. It processes Bhili sentences.
- A Frozen Teacher Encoder ($\text{ENC}_{\text{student}}$): A powerful, pre-trained Hindi model whose parameters (Θ_T) are frozen during training. For this role, we use HindBERT (Joshi, 2022), which is essentially MuRIL finetuned on publicly available Hindi monolingual datasets, and reportedly performs better than MuRIL on some downstream classification and NER tasks.
- A Memory Queue (Q): A FIFO queue that stores a large number of recent Hindi sentence embeddings produced by the teacher. This allows for a large pool of negative samples for the contrastive loss, decoupled from the batch size. This allows our method to be trained effectively on consumer-grade hardware with limited GPU memory.

A schematic of our proposed method is presented in Figure 1.

Training Process: For each parallel pair $(x^{(i)}_{\text{Bh}}, x^{(i)}_{\text{Hi}})$ in a batch of size B, the process unfolds as follows:

The student encoder processes the Bhili sentence to produce its `[CLS]` token embedding,

$$h^{(i)}_{\text{Bh}} = \text{ENC}_{\text{student}}(x^{(i)}_{\text{Bh}}) \tag{1}$$

Similarly, the frozen teacher encoder processes the corresponding Hindi sentence to produce the positive key,

$$h^{(i+)}_{\text{Bh}} = \text{ENC}_{\text{student}}(x^{(i)}_{\text{Bh}}) \tag{2}$$

The negative keys, $h^{(j-)}_{\text{Bh}}\}^{K}_{j=1}$, are the K embeddings of other Hindi sentences currently stored in the memory queue Q.

The student is then trained using a contrastive loss that encourages the Bhili embedding $h^{(i)}_{\text{Bh}}$ to be more similar to its positive Hindi key $h^{(i)}_{\text{Hi}}$ than to all negative keys in the queue. We use the NT-Xent (Normalized Temperature-scaled Cross Entropy) loss (Chen et al., 2020):

$$\mathcal{L}_{\text{contrastive}} = -\frac{1}{B}\sum_{i=1}^{B}\log\left(\frac{\exp\left(\ell_{\text{pos}}^{(i)}/\tau\right)}{\exp\left(\ell_{\text{pos}}^{(i)}/\tau\right)+\sum_{j=1}^{K}\exp\left(\ell_{\text{neg}}^{(i,j)}/\tau\right)}\right) \tag{3}$$

where B is the batch size, $\ell_{\text{pos}}^{(i)}$ is the cosine similarity between the embeddings of the i^{th} anchor and its positive key. Likewise, $\ell_{\text{neg}}^{(i)}$ is the similarity score between i^{th} anchor and j^{th} negative key and τ is a temperature parameter. The gradient flows only through the student encoder, while the teacher and queue remain detached. After each step, the queue is updated with the Hindi embeddings from the current batch. This MoCo-style mechanism provides a computationally efficient way to utilize a large number of negatives ($K >> B$), which is critical for effective contrastive learning.

Combined Objective: The student model is trained jointly on both the MLM task (using only the Bhili inputs) and the contrastive task. The final loss is a weighted sum:

$$\mathcal{L}_{\text{total}} = \alpha\mathcal{L}_{\text{MLM}} + \beta\mathcal{L}_{\text{contrastive}} \tag{4}$$

where α and β are hyperparameters that weight the contribution of each loss term. Training hyperparameters are given in Appendix F.

5 Evaluation

To comprehensively evaluate the quality of the Bhili language representations produced by our models, we designed a multi-faceted evaluation strategy. Our approach includes an intrinsic measure of training quality, performance on a practical downstream application, standardized cross-lingual benchmarks, and a qualitative analysis of the embedding space.

5.1 Intrinsic Evaluation: Masked Token Prediction

As a direct measure of how well each model has learned the statistical properties of the Bhili language, we report the final accuracy on the Masked Language Modeling (MLM) task on the test set. This intrinsic metric helps quantify the fundamental language understanding captured during the adaptation phase for both the MLM-only baseline and our proposed joint-training method.

5.2 Downstream Task: Sentiment Analysis

To assess the utility of the learned embeddings for practical applications, we evaluate them on a 3-way sentiment analysis task (positive (49%), negative (26%), neutral (24%)).

- **Dataset**: We created a sentiment analysis dataset by manually translating 8000 Hindi sentences from a standard sentiment corpus into Bhili. This translation was performed by native speakers to ensure high quality and contextual accuracy.

- **Protocol**: For our proposed model (and the baselines), we freeze the entire encoder backbone and add a lightweight classification head on top, consisting of a BiLSTM layer followed by a single MLP layer. Only the parameters of this classification head are trained on the Bhili sentiment dataset. This feature-extraction setup ensures that we are evaluating the intrinsic quality of the pre-trained representations themselves, without significant influence from task-specific fine-tuning. We report the accuracy, precision, recall and F1-score as our metrics.

5.3 Cross-Lingual Benchmark Evaluation

To measure the cross-lingual alignment and reasoning capabilities of our models, we evaluate them on a curated set of sentence-level benchmark tasks adapted from XTREME (Hu et al., 2020).

Task Selection Rationale: Our primary objective is to validate the semantic enrichment of our model. We follow a ”translate-test” or ”translate-train” approach where necessary. However, this methodology proved untenable for tasks that rely on precise alignment between labels and text spans. This includes token-level classification tasks like NER (PAN-X) and POS tagging (UDPOS), as well as extractive question-answering tasks like XQuAD and MLQA. The brittleness of projecting token-level or span-level labels under translation—a process sensitive to tokenization mismatches and syntactic reordering—would require a manual, resource-intensive data verification step. To ensure a reliable and robust evaluation, we exclusively selected sentence-level classification tasks where the label depends on the holistic meaning of the text, which is more resilient to the translation process.

Tasks and Protocols:

- **Tatoeba** (Sentence Retrieval): We evaluate the model's ability to perform cross-lingual sentence retrieval. We use the Hindi-English

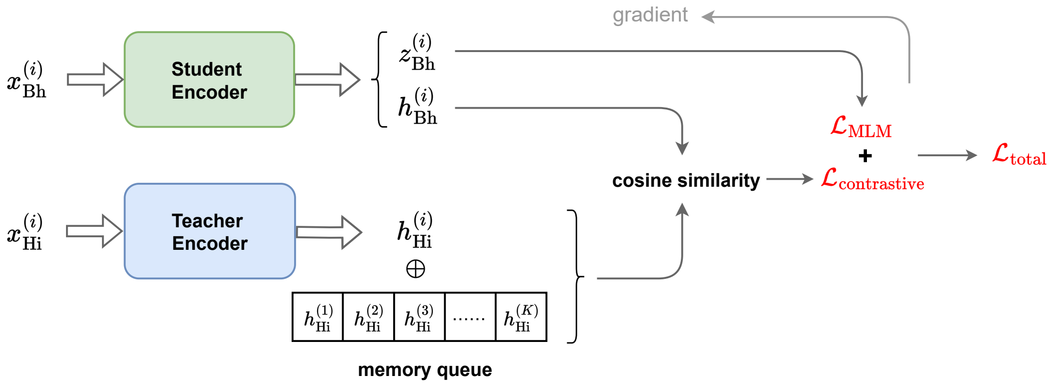

Figure 1: **An illustration of our GARuD framework for cross-lingual contrastive distillation.** The Student Encoder (MuRIL) is trained to produce Bhili sentence embeddings h_{Bh} that align with the high-quality embeddings h_{Hi} from a frozen, expert Teacher Encoder. A MoCo-style memory queue provides a large set of negative samples for the contrastive loss. The total loss combines this sentence-level contrastive objective with a token-level MLM objective, but the gradient only updates the student, effectively enabling a one-way distillation of knowledge from the teacher.

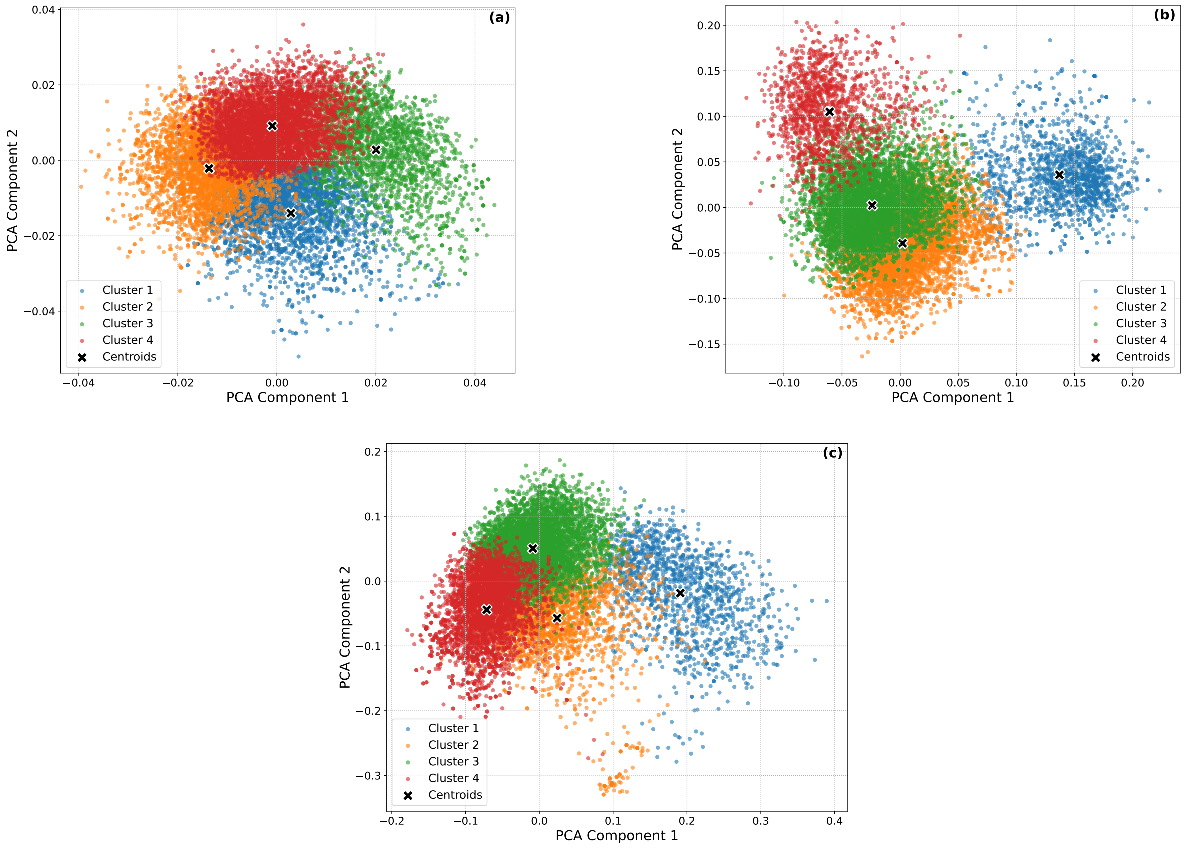

Figure 2: **Visualization of Bhili sentence embedding spaces learned by different models.** Sentence level embeddings created by mean pooling token embeddings, from the test set are projected into two dimensions using Principal Component Analysis (PCA). Colors correspond to four semantic clusters identified in an unsupervised fashion via K-Means clustering, with 'x' marking the cluster centroids. The plots show the embedding distributions for: (a, top-left) the Zero-Shot baseline, (b, top-right) the MLM-only baseline, and (c, bottom) our proposed contrastive distillation method. The superior inter-cluster separation in (c) demonstrates strong qualitative evidence of a more semantically meaningful and organized representation space for Bhili. This, in turn explains its stronger performance on downstream tasks.

sentence pairs from the Tatoeba dataset, manually translate the English sentences into Bhili using a translator tool [1], and then evaluate the model's zero-shot performance at matching the correct Bhili-Hindi translation pairs. Success on this task directly measures the quality of the cross-lingual embedding alignment.

- **XNLI** (Natural Language Inference): To test cross-lingual reasoning, we first fine-tune our models on a translated version of the English MNLI dataset (Wang et al., 2018). Subsequently, we evaluate the model's zero-shot transfer performance on the official XNLI test set for Hindi (Conneau et al., 2018), using Bhili sentences as input. Performance is measured by classification accuracy.

6 Results and Analysis

We now present the results of our comprehensive evaluation, comparing our proposed Contrastive Distillation method against the Zero-Shot and MLM-only baselines. The findings consistently demonstrate the superiority of our approach across all evaluation axes, confirming that our method produces more robust and semantically meaningful representations for Bhili.

6.1 Main Quantitative Results: Intrinsic and Downstream Performance

The primary quantitative results for both intrinsic (MLM accuracy) and downstream (sentiment analysis) tasks are summarized in Table 2.

Our analysis of these results reveals two key findings. First, in terms of intrinsic language modeling capability, our joint-training approach achieves a higher MLM accuracy (0.789) than the MLM-only baseline (0.756). This suggests that the cross-lingual contrastive objective acts as a semantic regularizer, helping the model learn more effective token-level representations for Bhili. It is synergistic with the MLM objective,

Second, this enhanced representation quality directly translates to superior downstream performance. On the sentiment analysis task, our model achieves a weighted F1-score of 0.671, significantly outperforming both the MLM-only baseline (0.656) and the Zero-Shot model (0.525), whose performance is only marginally better than random chance. This confirms that the embeddings produced by our method are not only statistically sound but also more useful for practical applications.

6.2 Cross-Lingual Benchmark Performance

To validate the cross-lingual alignment learned by our model, we present the results on the Tatoeba and XNLI benchmarks in Table 3. It provides the clearest and most compelling evidence of our method's effectiveness in creating a high-quality, aligned representation space for Bhili.

Our model demonstrates superior performance over both baselines across these tasks. The monumental improvement on the Tatoeba sentence retrieval task is particularly telling. This result directly validates our core hypothesis: the contrastive training objective successfully aligns the Bhili and Hindi embedding spaces, allowing the model to recognize sentence translations with much higher accuracy than the baselines.

Furthermore, success in the XNLI logical reasoning task shows that these high-quality, aligned embeddings serve as a stronger foundation for complex downstream reasoning. Unlike the token-focused MLM objective, our approach creates a more meaningful sentence-level representation space. These well-structured representations make it easier for the model to learn and transfer abstract concepts like entailment and contradiction to the low-resource language. While the performance gains on this task are more modest than on Tatoeba, they are nonetheless critical. Natural Language Inference requires not just recognizing semantic similarity, but also understanding logical relationships.

6.3 Qualitative Analysis of Embedding Space

A qualitative view of the embedding space, visualized via KMeans and PCA in Figure 2, corroborates our quantitative findings. The number of clusters was fixed to four for the sake of comparison. The visualization reveals three distinct patterns:

Zero-Shot: The embeddings from the baseline model exhibit representation collapse. The points are clustered into a single, undifferentiated mass, indicating the model has no semantic understanding of Bhili.

MLM-only: The model adapted with only MLM shows emergent structure. Some separation is visible, but the clusters corresponding to different semantic topics show significant overlap.

Our Method: In stark contrast, our model trained with the contrastive objective yields clearly

[1] https://aadivaani.tribal.gov.in/

Method	**Accuracy**	**Accuracy**	**Precision**	**Recall**	**F1**	**Training time**
	(MLM)	(Sentiment)				(avg. no. of epochs)
Zero-shot	–	0.598 ± 0.01	0.587 ± 0.03	0.525 ± 0.02	0.525 ± 0.03	13.8
MLM-only	0.756	0.681 ± 0.01	0.679 ± 0.02	0.647 ± 0.01	0.656 ± 0.01	8.2
GARuD-Bhili **(proposed)**	**0.789**	**0.693 ± 0.02**	**0.691 ± 0.02**	**0.659 ± 0.02**	**0.671 ± 0.02**	**6**

Table 2: Performance of models on sentiment classification. The best metrics for each column are shown in **bold**. Accuracy reported on a scale of 0–1. Additionally, improvement in masked token prediction accuracy from proposed method over MLM pretraining is also reported. Sentiment metrics are the mean ± standard deviation over 5 runs. Training efficiency is reported as the average number of epochs to convergence with early stopping.

Model	**Tatoeba**	**XNLI**
	(Acc.)	(Acc.)
Zero-shot	2.5	60.09
MLM-only	11.4	63.59
GARuD-Bhili **(proposed)**	**38.8**	**65.36**

Table 3: Performance on cross-lingual benchmark tasks. Our proposed method shows substantial gains on both sentence retrieval (Tatoeba) and natural language inference (XNLI) compared to the baselines.

partitioned clusters. While there is natural variance within each cluster, the inter-cluster separation is demonstrably superior. This visual evidence strongly indicates that our contrastive distillation method is highly effective at inducing a semantically organized representation space for a low-resource language.

6.4 Analysis of Training Efficiency

An interesting and practical benefit observed during our experiments is a marked improvement in training efficiency. As shown in the final column of Table 2, our joint-training model converges to its optimal performance in significantly fewer epochs (6.0 on average) compared to the MLM-only model (8.2). We hypothesize that the explicit, sentence-level supervisory signal from the cross-lingual contrastive loss provides a clearer and more direct learning path for the model, allowing it to reach an optimal state more rapidly than with the diffuse MLM objective alone. This finding, supported by similar observations in other contrastive learning literature (Jiang et al., 2024), suggests our method is not only more effective but also more efficient.

7 Conclusion

In this work, we presented a novel and highly effective training paradigm, **cross-lingual contrastive distillation**, to address the critical lack of foundational language models for extremely low-resource languages like Bhili. Our method successfully adapts a pre-trained multilingual model by using a fixed, expert teacher to distill semantic knowledge from a related high-resource language (Hindi) into a student model learning the target language (Bhili). Through comprehensive experiments, we demonstrated that our resulting model, GARuD-Bhili, significantly outperforms strong baselines on intrinsic, downstream, and cross-lingual benchmark tasks, creating a well-structured and semantically robust representation space for Bhili. The core contribution, however, is not a model for a single language, but rather an effective case study and a scalable blueprint for linguistic empowerment. The architectural components of our method are fundamentally language-agnostic. The success of this framework is not predicated on any unique property of Bhili or Hindi, but on the principle that the existence of a higher-resource language can act as a semantic bridge to its lower-resource neighbor. This makes our approach a replicable strategy for any scenario where such a modest parallel corpus can be curated and the source language is already well-represented in a publicly available multilingual model. Therefore, this work offers a practical, data-efficient and compute-efficient pathway for developing high-quality language models for the world's most underserved linguistic communities. It is our hope that it will not only catalyze further research for extremely low-resource Indic languages but also inspire and enable similar efforts to bring the benefits of modern NLP to the vast number of languages that remain on the fringes of the digital age.

Limitations

While our work successfully demonstrates a powerful new paradigm for an ultra-low-resource language, we acknowledge its current limitations.

First, our evaluation, though comprehensive, did not extend to complex span-based tasks like extractive Question Answering. The brittleness of projecting answer spans across translations makes this a methodologically challenging evaluation to perform reliably without extensive manual annotation, which was beyond the scope of this initial study.

Second, the effectiveness of our GARuD framework has been demonstrated on a single, albeit highly representative, language pair: Bhili-Hindi. While this provides a strong proof-of-concept, further experiments are needed to validate its generalizability across a wider range of language families and typological distances.

Finally, the performance of our model is intrinsically tied to the diversity of the parallel corpus. While the Bhili-Hindi Parallel Corpus is a monumental resource, its domain coverage may not fully encompass informal or conversational genres. The model's robustness on out-of-domain text is therefore an area for further investigation.

Future Work

Our future work will advance along three main directions. First, we aim to **scale the GARuD framework** to other ultra-low-resource Indian language pairs such as Gondi–Telugu and Santali–Bengali, testing its robustness and generalizability. Second, we plan to **develop a unified multilingual model** trained on multiple parallel corpora, enabling shared representation learning and positive transfer across India's underserved languages. Finally, we envision a **community-led data and benchmark initiative** focused on expanding domain-diverse parallel corpora and establishing standardized token- and span-level evaluation datasets. Such collaborative benchmark creation is essential for ensuring meaningful, reproducible progress in low-resource NLP.

Ethical Considerations

The development of NLP technologies for low-resource and tribal languages entails a deep ethical responsibility. Our work on Bhili—a language central to its community's cultural identity—was guided by the principles of linguistic accuracy, cultural respect, and community participation. The Bhili-Hindi Parallel Corpus (BHPC) was curated in collaboration with native Bhili speakers from Jhabua, Madhya Pradesh, ensuring authenticity and contextual relevance. Annotators provided informed consent and received fair compensation aligned with government standards. The dataset integrates culturally significant expressions such as “बोकड़ा पाळवा” (goat-rearing) and “जौहार” (a local salutation), preserving cultural nuance alongside linguistic precision. This community-driven and ethically grounded approach reflects our commitment to inclusive, responsible NLP research for underrepresented languages.

Acknowledgements

We gratefully acknowledge the Ministry of Tribal Affairs, Government of India, and Shri Vibhu Nayar, Secretary, for their visionary leadership and pivotal guidance in shaping this research under the Adi Vaani Project, which laid the foundation for the preservation and technological empowerment of India's tribal languages. We thank the Tribal Research Institute, Bhopal, and all officials and translators for their invaluable assistance in data coordination and creation of the BHEPC benchmark. We are deeply grateful to Prof. Parag Singla, Prof. Vivek Kumar, and Dr. Vipul Rathore for their continuous guidance. This work was supported by the Yardi School of Artificial Intelligence, Indian Institute of Technology Delhi, through research funding and computational infrastructure.

References

D. I. Adelani and 1 others. 2022. Afriberta: Large-scale self-supervised pretraining for african languages. *arXiv preprint arXiv:2204.06487*.

Nuo Chen, Linjun Shou, Tengtao Song, Ming Gong, Jian Pei, Jianhui Chang, Daxin Jiang, and Jia Li. 2023. Structural contrastive pretraining for cross-lingual comprehension. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 2042–2057, Toronto, Canada. Association for Computational Linguistics.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. *Preprint*, arXiv:2002.05709.

Zewen Chi, Li Dong, Furu Wei, Wenhui Wang, Xian-Ling Mao, Heyan Huang, and Ming Zhou. 2021. Infoxlm: An information-theoretic framework for cross-lingual language model pre-training. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics*, pages 3576–3588. Association for Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451. Association for Computational Linguistics.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. XNLI: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2475–2485, Brussels, Belgium. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4171–4186. Association for Computational Linguistics.

Sumanth Doddapaneni, Rahul Aralikatte, Gowtham Ramesh, Shreya Goyal, Mitesh M. Khapra, Anoop Kunchukuttan, and Pratyush Kumar. 2023. Towards leaving no Indic language behind: Building monolingual corpora, benchmark and models for Indic languages. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12402–12426, Toronto, Canada. Association for Computational Linguistics.

F. Feng and 1 others. 2020. Language-agnostic bert sentence embedding. *arXiv preprint arXiv:2007.01852*.

Jivnesh Gala and Pranjal Chitale. 2023. Bharat parallel corpus collection (bpcc): A comprehensive multilingual resource for indian languages. In *Proceedings of the 2023 Conference on Machine Translation (WMT)*, pages 45–55. Association for Computational Linguistics.

Barry Haddow and Faheem Kirefu. 2020. Pmindia: A parallel corpus of indian languages. In *Proceedings of the 2020 Conference on Machine Translation (WMT)*, pages 224–230. Association for Computational Linguistics.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. 2019. Momentum contrast for unsupervised visual representation learning. *arXiv preprint arXiv:1911.05722*.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the knowledge in a neural network. In *NIPS Deep Learning and Representation Learning Workshop*.

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. Xtreme: A massively multilingual multi-task benchmark for evaluating cross-lingual generalization. *Preprint*, arXiv:2003.11080.

Xin Jiang, Xu Cheng, and Zechao Li. 2024. Why pre-training is beneficial for downstream classification tasks. *Preprint*, arXiv:2410.08455.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the nlp world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293.

Raviraj Joshi. 2022. L3cube-hindbert and devbert: Pre-trained bert transformer models for devanagari based hindi and marathi languages. *arXiv preprint arXiv:2211.11418*.

Simran Khanuja, Diksha Bansal, Sarvesh Mehtani, Savya Khosla, Atreyee Dey, Balaji Gopalan, Dilip Kumar Margam, Pooja Aggarwal, Rajiv Teja Nagipogu, Shachi Dave, Shruti Gupta, Subhash Chandra Bose Gali, Vish Subramanian, and Partha Talukdar. 2021. Muril. *Preprint*, arXiv:2103.10730.

Zhongtao Miao, Qiyu Wu, Kaiyan Zhao, Zilong Wu, and Yoshimasa Tsuruoka. 2024. Enhancing cross-lingual sentence embedding for low-resource languages with word alignment. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 3225–3236, Mexico City, Mexico. Association for Computational Linguistics.

Wilhelmina Nekoto, Vukosi Marivate, Terence Matsila, and 1 others. 2020. Participatory research for low-resourced machine translation: A case study in african languages. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 2144–2160. Association for Computational Linguistics.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Pooja Singh, Shashwat Bhardwaj, Vaibhav Sharma, and Sandeep Kumar. 2025. Leveraging the cross-domain & cross-linguistic corpus for low resource nmt: A case study on bhili-hindi-english parallel corpus. *arXiv preprint arXiv:2511.00486*. Accepted at EMNLP 2025.

Shashank Siripragada, Gowtham Ramesh, and 1 others. 2020. A multilingual parallel corpus of legislative assembly proceedings in indian languages. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 308–316. European Language Resources Association.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium. Association for Computational Linguistics.

Yaushian Wang, Ashley Wu, and Graham Neubig. 2022. English contrastive learning can learn universal cross-lingual sentence embeddings. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9122–9133, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

A Ultra Low-Resource Languages

Languages with extremely limited resources suffer from a serious lack of accessible data and proper documentation. This issue is particularly evident in many Indian regional languages, which often fall into this category. Unlike widely studied languages, these lesser-known languages have minimal resources—many are unpublished or contain very little available data. They are commonly described as under-documented, under-resourced, or under-digitized. As a result, collecting and processing raw textual data in these languages presents significant challenges.

B Bhili Language

Bhili, a Western Indo-Aryan language spoken by approximately 13 million people across Rajasthan, Gujarat, Maharashtra, and Madhya Pradesh, is written in the Devanagari script and deeply embedded in Bhil cultural heritage. The dataset we present focuses on the Bhili dialect from the Jhabua region of Madhya Pradesh. Despite the language's cultural and demographic importance, Bhili remains largely unexplored in Natural Language Processing (NLP) due to the lack of publicly available parallel corpora.

C Translation Guidelines

To maintain consistency and preserve the semantic integrity of translations, we established a comprehensive set of guidelines that balance linguistic precision with the practical constraints posed by the scarcity of Bhili-specific glossaries, literature, and linguistic resources. Translators are instructed to faithfully retain the meaning and stylistic tone of the source content without introducing new material or omitting any part of the original. Special care was taken when handling named entities, numbers, dates, and technical terminology in accordance with the conventions of the target language. The core principles include:

- **General Principles:** Accurately convey the source text's meaning, tone, style, and level of formality—whether formal, informal, or emphatic—without making additions or deletions. Minor grammatical errors or typographical mistakes may be corrected as long as factual inconsistencies in the source are preserved. The translation should read fluently and naturally in the target language.

- **Named Entities:** Use standardized, widely accepted translations where they exist. If unavailable, entities should be precisely transliterated into the target script following language-specific conventions. Translators must avoid creating alternative or novel renderings.

- **Numbers and Units:** Reproduce numerical expressions exactly as they appear in the source, whether spelled out or written in digits. Apply local numbering conventions for large numbers where appropriate, while retaining terms like "billion" and "trillion" in English or using recognized equivalents in the target language. All original units of measurement must be preserved.

- **Dates:** Maintain the original date format as it appears in the source, whether in fully written or numeric form. Year lengths should not be altered, and no expansion or abbreviation of dates is permitted.

D Annotation Guidelines Based on MQM (Multidimensional Quality Metrics): Error Categories and Severities

Translation quality was evaluated at the segment level, where each segment could consist of one or multiple sentences. Translations were aligned side-by-side with their corresponding source texts, enabling annotators to examine them in parallel. A well-defined hierarchy of error types was presented in Table 8 to guide annotators in consistently identifying and categorizing issues. Each error type was assigned a severity score on a five-point scale—*Very Low*, *Low*, *Medium*, *High*, and *Very High*—allowing for fine-grained distinctions in impact.

To translate qualitative assessments into numeric scores, we used the following mapping: Very Low = 1, Low = 2, Medium = 3, High = 4, and Very High = 5. Each error category—such as *Accuracy*, *Fluency*, *Terminology*, and *Style*—was scored independently, and all were treated with equal importance. Errors unrelated to translation were given a score of zero, and any sentence containing a critical error in the source text was excluded from scoring.

Instructions for Annotators

- Annotators were required to closely inspect each translated segment and identify all present errors, with a strict limit of five errors per segment. If more than five were found, only the five most critical errors were to be reported.

- Each error span was to be precisely highlighted using color coding, followed by assigning the correct error category and subcategory, along with an appropriate severity level. If the issue originated from the source sentence or involved omission, the corresponding source text span was to be marked instead.

- Errors had to be captured at the most granular level possible. For example, if two separate words were mistranslated, annotators had to log them as two individual errors rather than one.

- In cases where multiple errors overlapped in the same span, only the most severe error was recorded. If two errors had the same severity, the first matching category in the MQM hierarchy (e.g., Accuracy, then Fluency, then Terminology) was selected.

- **Special Categories:**

  - *Non-translation:* If a translation was highly distorted or entirely unrelated to the source, and individual errors couldn't be meaningfully identified, annotators were instructed to mark the entire segment as a single Non-translation error. No other errors should be logged for that segment.
  - *Source Error:* When an error was due to flaws in the source text, the problematic span in the source was highlighted, and such segments were excluded from scoring (though the source error still needed to be documented).

- After completing all annotations for a segment, annotators assigned a final quality score out of 25 and recorded this value in the final score column.

E Data Quality Control

In order to rigorously assess the quality of our manually curated Hindi–Bhili gold references, we conducted an inter-annotator agreement (IAA) study. Following the MQM/DA protocol, we randomly sampled 250 sentences across our three domains. Two independent professional translators (Translator A and Translator B), both native Bhili speakers, translated each Hindi sentence into Bhili. Subsequently, each translator rated the other's rendering on the MQM scale. This evaluation yielded an **IAA coefficient of 0.68** for the Hindi→Bhili gold references, indicating substantial agreement and confirming the high quality of our dataset.

F Training Details

Both the MLM-only and proposed models were trained on an NVIDIA L40S GPU only once. Early stopping was used in both cases. Hyperparameters are provided in the following table:

Hyperparameter	Value
`max_seq_length`	128
Masking Probability	0.15
`max_epochs`	1000
`batch_size`	32
`learning_rate`	5e-5
Optimizer	AdamW
Scheduler	Linear Decay with warmup
Early Stopping	Yes
`queue_size` (for MoCo)	2048
α	1
β	1

The loss and accuracy plots are presented in Figure 3 for MLM-only and in Figure 4 for our proposed method. The notable sharp dip in the contrastive loss curve can be explained by the presence of learning rate reduction due to scheduling during training.

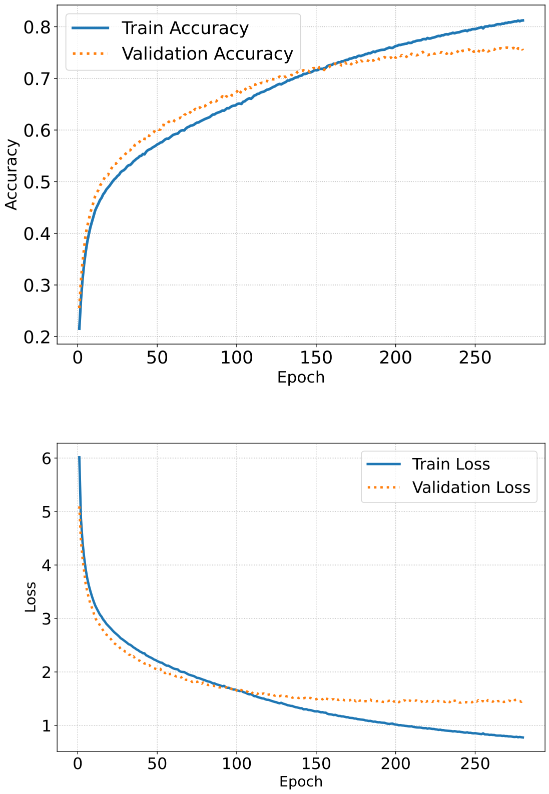

Figure 3: Accuracy and loss plots for continued pretraining of MuRIL with MLM

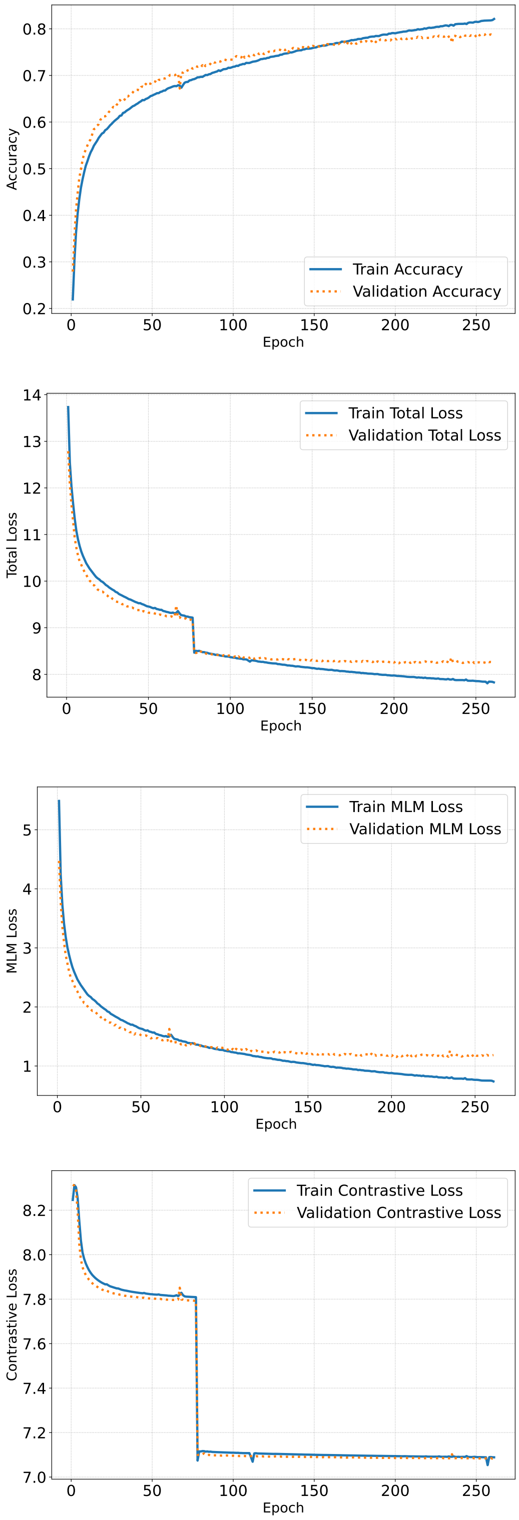

Figure 4: Accuracy and loss plots for training with GARuD, our proposed method. The total loss, MLM loss and the contrastive loss have been shown individually.

G Fine-tuning Details

Fine-tuning for the sentiment classification downstream task was done on an NVIDIA A5000 GPU.

The hyperparameter settings are provided in the following table:

Hyperparameter	**Value**
max_seq_length	64
max_epochs	100
batch_size	64
learning_rate	2e-3
Optimizer	AdamW
hidden_dim (for LSTM)	256
Dropout	0.2
Patience	3

For the XNLI task, we fine-tuned on the MNLI dataset using the standard training pipeline from HuggingFace's transformers library and then performed inference on the Hindi language split of the XNLI dataset. The task was performed on an NVIDIA L40S GPU. The hyperparameters are in the following table:

Hyperparameter	**Value**
num_epochs	5
batch_size	32
gradient_accumulation_steps	2
learning_rate	2e-5
weight_decay	0.01
warmup_ratio	0.1