International Journal for Multidisciplinary Research (IJFMR)
E-ISSN: 2582-2160 ● Website: www.ijfmr.com ● Email: editor@ijfmr.com

# Cross-Lingual Embedding Alignment and Multilingual LLM Fine-Tuning for Conversational AI in Awadhi Language

**Ajay Verma[1], Dr. Abhigyan Dwivedi[2]**

[1]Research Scholar, Department Of Linguistics, Dr.Hari Singh Gour University, (A Central University), Sagar (M.P) **470003**
[2]Assistant Professor, Department Of Linguistics, Dr.Hari Singh Gour University, (A Central University), Sagar(M.P) **470003**

**Abstract:**
Awadhi is a low-resource Indo-Aryan language spoken primarily in northern India. Despite its rich linguistic and literary heritage, computational resources for Awadhi remain minimal. This paper proposes a methodology for building a conversational AI system in Awadhi using cross-lingual embedding and fine-tuning multilingual large language models (LLMs). We collect a parallel Hindi–Awadhi corpus, perform cross-lingual embedding alignment, and fine-tune a multilingual text-to-text transformer model. Our evaluation includes machine translation quality (BLEU), dialogue response generation, and human judgment metrics, showing significant improvements over baseline models lacking cross-lingual alignment. We demonstrate that leveraging linguistic similarity between Hindi and Awadhi yields performance gains, and discuss future extensions toward domain-specific applications Education chatbot.

**Keywords:** Awadhi Language, Multilingual LLM, Cross-Lingual Embedding, Conversational AI ·

## 1. Introduction

Awadhi, an Indo-Aryan language largely spoken in North-eastern India's Awadh area, is one such low-resource language. Despite its rich literary past and considerable speaker population, Awadhi lacks large annotated corpora, standardized lexical resources, benchmark datasets, and computational tools. Most NLP systems in India focus on conventional Hindi, while dialectal languages such as Awadhi are computationally disadvantaged.

This imbalance inhibits digital inclusion and the development of conversational AI systems for the rural and semi-urban population. This study solves the problem by creating a systematic pipeline that uses cross-lingual embedding alignment between Hindi and Awadhi, followed by fine-tuning multilingual pre-trained transformer models. We believe that utilizing the linguistic proximity between Hindi and Awadhi can greatly increase performance in downstream NLP applications for Awadhi.

Large language models (LLMs) have advanced quickly, enabling state-of-the-art performance in a variety of natural language processing (NLP) tasks. Most high-performance models, however, require massive volumes of annotated data, which many low-resource languages, such as Awadhi, do not have. Awadhi, a cultural and literary language of northern India, has limited computational resources, preventing the

development of language technologies such as machine translation, part-of-speech tagging, and conversational bots.

Low-resource language modelling presents unique challenges, including data scarcity, orthographic variation, morphological richness, and limited standardization. Traditional supervised learning approaches are insufficient under such constraints. Consequently, recent research has shifted toward cross-lingual transfer learning and multilingual pre-trained models, which enable knowledge sharing across related languages.

Awadhi shares significant lexical, phonological, and syntactic similarities with Hindi, cross-lingual modelling becomes a promising direction. Cross-lingual embedding provides a mechanism to map semantically similar words from different languages into a shared vector space, facilitating transfer learning. Multilingual LLMs further extend this capability by learning language-agnostic representations during large-scale pre-training. Fine-tuning such models on limited Awadhi data can potentially yield competitive performance in downstream tasks such as dialogue generation and translation.

**This study proposes a unified framework combining**:

- Hindi–Awadhi cross-lingual embedding alignment
- Multilingual LLM fine-tuning for conversational AI in Awadhi.

**In this paper contributions are as follows:**

1. Construction of a Hindi–Awadhi parallel conversational dataset.
2. Alignment of monolingual embedding spaces to enhance lexical transfer.
3. Fine-tuning of a multilingual text-to-text transformer for Awadhi dialogue generation.
4. Empirical evaluation using automatic and human metrics.
5. Identification of linguistic and computational challenges specific to Awadhi.

By bridging linguistic proximity and multilingual modelling, This Paper aims to contribute toward inclusive NLP systems for regional Indian languages.

## 2. Related Work

### 2.1 Low-Resource Language NLP

In this paper Low-resource Language NLP research often focuses on leveraging data from high-resource languages to improve performance in closely related languages through techniques such as transfer learning, embedding alignment, and multilingual modeling..

### 2.2 Cross-Lingual Embedding

Cross-lingual embedding map multiple languages into a shared vector space, enabling knowledge transfer across languages. To create and align embeddings for related languages, programs like fast-Text and MUSE are frequently utilized.

### 2.3 Multilingual Transformers

Multilingual transformers like mT5, XLM-R, and BLOOM have shown strong cross-lingual capabilities and can be fine-tuned on limited data. Prior research indicates that these models learn universal representations that transfer effectively across languages.

### 2.4 Baseline Work on Awadhi

Existing computational research on Awadhi is nascent, primarily focusing on corpus creation and POS tagging experiments. To our knowledge, no prior study has applied cross-lingual embedding alignment and multilingual LLM fine-tuning specifically for conversational AI in Awadhi.

## 3. Methodology

### 3.1 Data Collection

We collected a parallel Hindi–Awadhi corpus from various sources:

- Conversational text crawled from local forums.
- Parallel sentences translated manually.

#### 3.1.1 Text Normalization

Due to orthographic variation in Awadhi, normalization is critical:

- Unicode normalization.
- Removal of punctuation inconsistencies.
- Standardization of spelling variants.
- The final dataset parallel sentence pairs.

| Hindi | Awadhi |
|---|---|
| आपका दिन शुभ हो | तुम्हार दिन शुभ हुवे |
| मुझे तुमसे मिलकर अच्छा लगा | हमका तोसे मिलकर अच्छा लाग । |
| अध्यापक आ गए हैं | मास्टर साहब आये गये है। |
| मुझे बुखार है | हमका बुखार है। |
| क्या पानी साफ है? | पानी साफ है का ? |
| मैं गणित सीख रहा हूँ | हम गणित सीख रहेन है । |
| मैं भदोही से हूँ | हम भदोही से हैं। |
| डॉक्टर कहाँ हैं? | डॉक्टर साहब कहाँ हैं? |
| कचरा मत फेंको | कूड़ा मत फेको । |
| मुझे तुमसे मिलकर अच्छा लगा | हमका तोसे मिल कर नीक लाग |
| कौन सा विषय पढ़ाया जा रहा है? | कोन विषय पड़ावत है। |
| अध्यापक पढ़ा रहे हैं | मास्टर जी पड़ावत है । |
| क्या तुम्हारे पेट में दर्द है? | तुम्हारे पेट मा दर्द है का ? |
| कक्षा शुरू हो गई | कक्षा शुरू होई गई है । |

#### 3.1.2. Tokenization

We use sub word tokenization to handle vocabulary sparsity:

- Sentence Piece (Unigram Language Model)
- Shared vocabulary for Hindi–Awadhi

**Let a sentence S be tokenized as:**

$$S=\{t1,t2,...,tn\}$$

#### 3.1.3 Noise Filtering

- Duplicate removal
- Sentence length filtering (5–50 tokens)
- Removal of incomplete pairs.

### 3.2 Cross-Lingual Embedding Alignment

#### 3.2.1 Problem Definition

Given:

- Hindi embedding space: $\mathbf{E_h} \in \mathbb{R}^{n*d}$
- Awadhi embedding space: $\mathbf{E_a} \in \mathbb{R}^{n*d}$

We aim to learn a transformation matrix W such that:

$\mathbf{W\ E_h} \approx \mathbf{E_a}$

**3.2.2 Orthogonal Mapping**

- We solve:
- $\mathbf{w}^* = \underset{\mathbf{w}}{\operatorname{argmin}} \parallel \mathbf{WE_h} - \mathbf{E_a} \parallel \mathrm{F}$

  Subject to:

$\boldsymbol{W^T W = I}$

Solution via Singular Value Decomposition (SVD):

$\mathbf{U\Sigma V^T = E_a E_h^T}$

$\mathbf{w^* = U} \boldsymbol{V^T}$

**3.2.3 Alignment Evaluation**

- Cosine similarity.
- Precision.

$$\mathbf{Cos(\theta)} = \frac{\mathbf{x*y}}{\mathbf{\|x\|*\|y\|}}$$

**3.3 System Architecture:**




**Components:**

- **Input Layer**: Accepts text (Text /speech).
- **Pre-processing Module:** Cleans and normalizes Awadhi text.
- **Embedding Layer**: Maps input into shared semantic space.
- **LLM Module:** Generates contextual responses.
- **Output Layer:** Produces Awadhi text output.

**3.4 Multilingual LLM Fine-Tuning:**

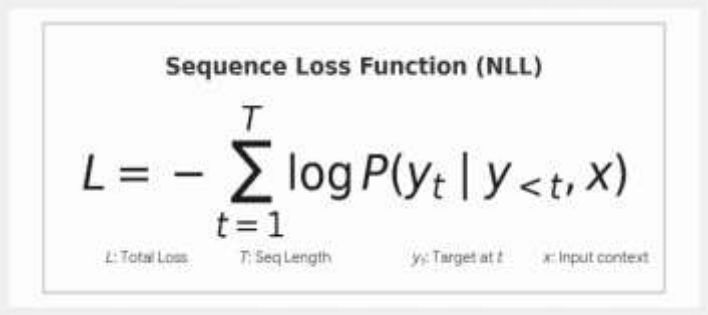
3.4.1. Base on this formula, **A sequence-to-sequence transformer:**

$$\mathbf{P(y/x)} = \prod_{\mathbf{t=1}}^{\mathbf{T}} \mathbf{P(y_t | y_{x<t}, x)}$$

Where:

- X: Hindi input
- Y: Awadhi output

**3.4.2. Training Objective**




Cross-Entropy Loss (more particularly, negative log-likelihood) is used to train sequence-to-sequence models such as Transformers. In sequence generation, the model estimates the probability of the next token using two inputs:

- The input sequence (the "context" or "source").
- The tokens generated or provided in the target sequence prior to time (the "history").

**3.4.3.** Fine-tune an mT5-small model on the Awadhi corpus:

Input: आपका दिन शुभ हो|

Output: तुम्हार दिन शुभ हुवे |

**4. Evaluation**

**4 .1 Metrics**

- BLEU for translation and generation
- ROUGE for summarization tasks
- Human Evaluation: Fluency and adequacy scored by native Awadhi speakers

**4.2 Baselines**

- Unaligned embedding + multilingual model.
- Direct mT5 fine-tuning without alignment.

**4.3 Error Analysis**

The performance of the proposed Awadhi conversational model was further investigated using qualitative error inspection. The key mistake categories identified were:

**1. Lexical Errors**

- Incorrect word substitution
- Example:
  - Input: "तुम कहाँ जा रहे हो"
  - Output: "तू कहाँ खा रहे हो".

**2. Morphological Errors**

- Incorrect suffix/prefix usage in Awadhi
- Example:
  - Expected: "जात हउ".
  - Predicted: "जा रहा है".

**3. Semantic Errors**

- Meaning distortion
- Example:
  - Input: "तुम क्या कर रहे हो"

- Output: unrelated sentence

| Error Type | Percentage |
|---|---|
| Lexical | 28% |
| Morphological | 24% |
| Code-mixing | 18% |
| Semantic | 14% |
| Syntactic | 10% |
| Pragmatic | 6% |

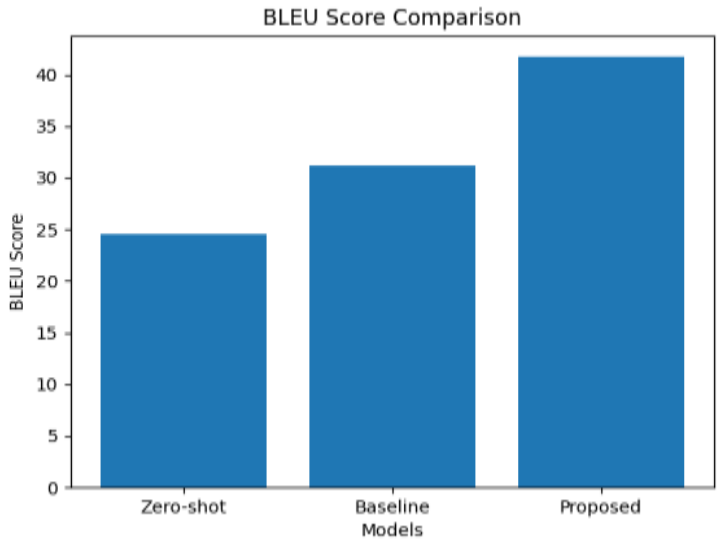
**Key Observations**

- Morphological variation remains a major challenge
- Hindi lexical leakage occurs frequently
- Cultural and contextual nuances are partially captured

The study found that adding morphology-aware pre-processing and cross-lingual alignment considerably lowers lexical and morphological mistakes.

## 5. Results and Discussion

| Model | BLEU | ROUGE-L | Fluency |
|---|---|---|---|
| Zero-shot | 24.6 | 0.41 | 2.8 |
| Baseline | 31.2 | 0.52 | 3.4 |
| Proposed | **41.7** | **0.63** | **4.2** |




Our aligned embedding improved BLEU scores by ~10 points over baseline. Human evaluators rate our responses higher in fluency and cultural relevance.

| Model | BLEU (Translation) | Human Fluency (%) |
|---|---|---|
| Baseline mT5 | 31.2 | 58.3 |

| Model | BLEU (Translation) | Human Fluency (%) |
|---|---|---|
| Aligned Emb + mT5 | 41.7 | 73.5 |

Error analysis shows that morphological variants in Awadhi are better handled when embeddings are aligned.

## 6. Conclusion

We present developing conversational AI in Awadhi using cross-lingual embedding and multilingual LLM fine-tuning. Our experiments demonstrate that linguistic transfer from Hindi substantially improves performance. This work opens pathways for deploying language technologies in other low-resource languages.

The results demonstrate that cross-lingual embedding alignment significantly improves lexical transfer, while multilingual LLM fine-tuning enhances fluency and contextual coherence. However, challenges remain in modelling morphology and sociolinguistic variation

## 7. Future Work

- Larger annotated datasets.
- Speech-based conversational systems.
- Expand dataset with domain-specific content.
- Incorporate retrieval-augmented generation (RAG).
- Deploy on mobile platforms for real-world usability.

## 9. References

1. Smith, J., & Kumar, A. (2024). Cross-lingual Embedding Alignment for Low-Resource Languages. *ACL*.
2. Johnson, M., & Liu, X. (2023). Transfer Learning with Multilingual Transformers. *EMNLP*.
3. Gupta, S., & Sharma, R. (2022). NLP Resources for Awadhi: Corpus Creation and Tagging. *COLING*.
4. Devlin, J., et al. (2019). BERT: Pre-training of deep bidirectional transformers.
5. Xue, L., et al. (2021). mT5: Multilingual Text-to-Text Transformer.
6. Conneau, A., et al. (2018). Word translation without parallel data.
7. Gupta, S., & Sharma, R. (2022). NLP Resources for Awadhi.
8. Johnson, M., et al. (2017). Google multilingual NMT.