Few-Shot Contrastive Adaptation for Audio Abuse Detection in Low-Resource Indic Languages

Anonymous ACL submission

Abstract

Abusive speech detection is becoming increasingly important as social media shifts towards voice-based interaction, particularly in multilingual and low-resource settings. Most current systems rely on automatic speech recognition (ASR) followed by text-based hate speech classification, but this pipeline is vulnerable to transcription errors and discards prosodic information carried in speech. We investigate whether Contrastive Language-Audio Pretraining (CLAP) can support abusive speech detection directly from audio. Using the ADIMA dataset (Gupta et al., 2022), we evaluate CLAP-based representations under few-shot supervised contrastive adaptation in cross-lingual and leave-one-language-out settings, with zero-shot prompting included as an auxiliary analysis. Our results show that CLAP yields strong cross-lingual audio representations across ten Indic languages, and that lightweight projection-only adaptation achieves competitive performance with respect to fully supervised systems trained on complete training data. However, the benefits of few-shot adaptation are language-dependent and not monotonic with shot size. These findings suggest that contrastive audio-text models provide a promising basis for cross-lingual audio abuse detection in low-resource settings, while also indicating that transfer remains incomplete and language-specific in important ways.

1 Introduction

Spoken content has become a primary mode of interaction across short-form video platforms, voice notes, live audio rooms, podcasts, and streaming services. Harmful language expressed through speech is therefore an increasingly important moderation problem. In many online settings, abuse conveyed through audio can be more impulsive, emotionally charged, and less self-monitored than text (Sakhare and Kulkarni, 2023), making it both socially consequential and technically difficult to detect. Despite major advances in text-based hate speech and toxicity detection, direct audio abuse detection remains comparatively underexplored (Sharon et al., 2022), especially in multilingual and low-resource settings where speech is often preferred over writing.

This challenge is particularly acute in linguistically diverse regions such as South Asia, where users routinely communicate in regional languages, dialects, and code-mixed speech (Garg et al., 2022). These forms of communication are often underrepresented in large annotated corpora and pre-trained models, limiting the effectiveness of systems developed primarily for high-resource languages such as English. At the same time, abusive spoken content can spread rapidly across platforms, contributing to harassment, misinformation, and targeted hate. Building robust audio-native abuse detection systems is therefore both a technical challenge and a societal necessity.

A common strategy for abusive audio detection is to first transcribe speech using automatic speech recognition (ASR) and then apply text-based hate speech classifiers (Sharon et al., 2022). Although this pipeline benefits from progress in natural language processing, it also introduces several important limitations. First, ASR systems for low-resource languages often exhibit high error rates, particularly for informal, colloquial, code-mixed, or profanity-rich speech typical of social media. Such errors can obscure abusive intent, distort offensive expressions, or remove contextual cues required for accurate classification (Chhabra and Vishwakarma, 2023). Second, abuse in speech is not purely lexical. Paralinguistic and prosodic properties such as pitch, emphasis, rhythm, stress, and emotional intensity may strongly influence how an utterance is perceived (Székely et al., 2023). These cues are largely lost when the signal is reduced to text. Third, ASR-based pipelines compound errors across stages, increasing both computational cost

and the risk of downstream failure (Ákos Tündik et al., 2019). These limitations motivate models that operate directly on audio representations rather than relying entirely on transcription.

Recent advances in contrastive learning offer a promising alternative for cross-modal representation learning. Contrastive approaches learn embeddings by pulling semantically aligned examples closer together while pushing mismatched examples apart. In vision-language research, such objectives have produced highly transferable representations with strong zero-shot capability (Wang et al., 2025; Qian and Hu, 2024). Contrastive Language–Audio Pre-training (CLAP) (Elizalde et al., 2023) extends this paradigm to the audio-text setting by jointly learning aligned audio and text embeddings in a shared semantic space. This makes it possible to compare an audio clip directly with natural-language prompts representing candidate labels. For abusive speech detection, prompts such as `"This audio contains hate speech"` and `"This audio does not contain hate speech"` can therefore act as class prototypes.

CLAP is especially attractive in multilingual settings because its cross-modal training may encourage language-agnostic representations. Rather than training separate supervised systems for each language, a single pre-trained audio-text model can potentially transfer across languages with limited adaptation. Moreover, these representations can be refined under few-shot conditions. Even a small number of labelled examples per class per language may improve the geometry of the embedding space, and supervised contrastive objectives can explicitly bring same-class examples closer while pushing apart examples from different classes. Lightweight adaptation through projection heads or selective fine-tuning therefore offers a practical compromise between large-scale pre-training and task-specific learning.

In this work, we investigate whether CLAP-based representations can provide an effective and scalable framework for multilingual audio abuse detection. Our main focus is on few-shot supervised contrastive adaptation and downstream classification under monolingual, cross-lingual, and leave-one-language-out transfer settings, while zero-shot prompting is treated as an auxiliary evaluation setting. Our study is guided by the following research questions:

- **RQ1:** How effective are CLAP-based representations for abusive speech detection directly from audio across multiple Indic languages?
- **RQ2:** To what extent does few-shot contrastive adaptation improve performance over using frozen pre-trained CLAP embeddings in low-resource language settings?
- **RQ3:** How well do CLAP embeddings transfer across languages, and does excluding the target language from training (leave-one-language-out) substantially degrade performance?
- **RQ4:** Does lightweight projection-only adaptation offer a more favourable trade-off between adaptation cost and downstream performance than deeper encoder fine-tuning in few-shot regimes?

In this work, there is no direct comparison against a few-shot supervised fine-tuning baseline trained with full task data, as such results are available from Gupta et al. (2022) and serve as an upper-bound reference rather than a direct competitor. Our primary interest is in understanding the behaviour of contrastive audio-text representations under varying degrees of cross-lingual transfer and adaptation, rather than in maximising absolute performance on ADIMA.

2 Literature Review

Research on hate, abuse, and toxicity detection has been dominated by text-based approaches, but the growth of voice-based social media has exposed the limitations of text-only moderation (Mnassri et al., 2022; Jafari et al., 2023; Mozafari et al., 2020). Recent work has therefore begun to focus on direct speech-based detection, motivated by ASR brittleness in noisy, multilingual, and code-mixed conditions, as well as by latency and privacy concerns. This section reviews datasets, modelling trends, and the specific challenges of multilingual and low-resource audio abuse detection.

2.1 Datasets for Audio Hate and Abuse Detection

Labelled resources for audio hate speech detection remain limited in both size and linguistic coverage. Among the most important benchmarks is **ADIMA** (Gupta et al., 2022), which contains 11,775 audio clips spanning ten Indic languages

and supports monolingual, multilingual, and cross-lingual evaluation. ADIMA is especially important because it highlights the limitations of ASR-based pipelines in profanity-rich and code-mixed speech.

Several other datasets extend the space in different ways. **DeToxy** (Ghosh et al., 2022) and **MuTox** (Costa-jussà et al., 2024) increase multilingual coverage and compare ASR cascades with end-to-end audio models, often showing that direct audio-based systems are competitive with or stronger than transcription-based approaches. Smaller language-specific datasets, including work on Malayalam and Tamil, remain useful for targeted studies but are limited in scale (Natarajan et al., 2025). More recent benchmarks such as **SynHate** (Ranjan et al., 2025b) extend evaluation to synthetic and deepfake audio, while studies using text-to-speech and prosodic analysis examine bias, explainability, and the role of non-lexical cues (Székely et al., 2023). Spoken toxicity resources continue to emerge for other languages as well, such as **ToxicTone** for Mandarin audio (Luo et al., 2025). Taken together, these datasets reveal substantial heterogeneity in annotation schemes, language balance, recording conditions, and task formulation.

2.2 Methodological Approaches

Early audio abuse detection systems relied on handcrafted acoustic descriptors such as MFCCs, spectral statistics, and prosodic cues combined with traditional classifiers (Rawat et al., 2023; Banuroopa and Shanmuga Priyaa, 2021). While these methods can be effective in controlled settings, they typically transfer poorly across languages and domains. More recent work increasingly uses pre-trained speech encoders such as Wav2Vec 2.0, XLS-R, and Whisper, either as fixed feature extractors or as fine-tuned backbones. On datasets such as ADIMA and MuTox, these audio-native representations can match or exceed ASR-based cascades, which remain vulnerable to transcription errors and additional latency (Sharon et al., 2022).

Multimodal approaches combine acoustic and textual signals in an effort to capture both lexical and paralinguistic information. Frameworks such as **MADA** report gains over unimodal baselines (Sharon et al., 2022), and related work shows the value of explicit speech-text alignment for multilingual toxicity classification (Liu et al., 2024). Other studies show that prosodic-only or acoustic-only cues can remain discriminative for abusive speech, even without transcripts (Spiesberger et al., 2023). Shared tasks such as DravidianLangTech@NAACL 2025 (Chakravarthi et al., 2025) further underscore the practical need for multimodal and multilingual moderation systems. Parallel work also addresses robustness and interpretability, including hateful segment localisation and evaluation under synthetic or adversarial perturbations (An et al., 2024; Ranjan et al., 2025b,a).

2.3 Few-Shot Learning and Low-Resource Challenges

Low-resource audio abuse detection remains difficult because labelled data is scarce, annotation is expensive, and the notion of abuse is often culturally sensitive. Performance is further affected by variability in speakers, accents, recording conditions, and code-mixing practices (Gupta et al., 2022; Sharon et al., 2022). Few-shot approaches remain relatively underexplored in this area, although recent work on cross-lingual meta-learning over ADIMA shows promising adaptation behaviour together with clear degradation in very low-shot regimes (Sankaran et al., 2025).

In multilingual settings, transfer is complicated by dialectal variation, language family differences, script differences, and culturally specific notions of offensiveness. Synthetic data generation and cross-lingual augmentation offer one route to scale, but they also raise questions about fairness, authenticity, and cultural validity (Narula and Chaudhary, 2024; Nandi et al., 2024; Roy et al., 2022). Promising directions therefore include contrastive few-shot adaptation, parameter-efficient fine-tuning of large audio encoders, multimodal few-shot learning, and stricter language- and speaker-disjoint evaluation protocols.

3 Methodology

3.1 Problem Formulation

We study abusive speech detection from audio under *zero-shot* and *few-shot* conditions. Given an audio segment $x \in \mathcal{X}$ and a binary label $y \in \{0, 1\}$ indicating non-abusive or abusive speech, the goal is to learn a classifier $f : \mathcal{X} \to \{0, 1\}$ that generalises across languages $\ell \in \mathcal{L}$. Zero-shot inference uses no task-specific labelled audio, whereas few-shot learning assumes access to k labelled examples per class per language drawn from the training split. We cast the problem as one of aligning audio signals with label semantics in a shared contrastive embedding space.

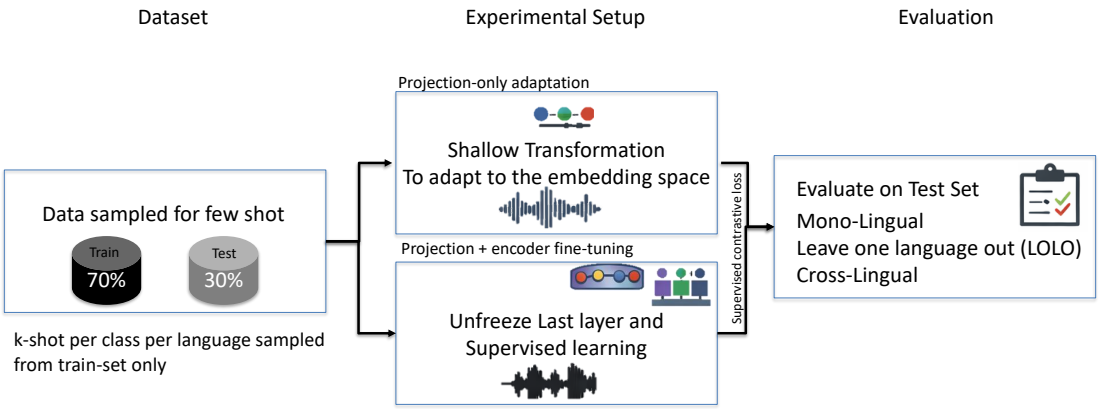


Figure 1: Overview of the proposed CLAP-based pipeline. A small k-shot subset from the training split is used to adapt CLAP through supervised contrastive learning. In the projection-only setting, only the projection head is updated; in the projection+fine-tuning setting, the projection head and the final audio encoder blocks are updated. The adapted model is then used to extract embeddings for all samples, which are evaluated under cross-lingual, and leave-one-language-out settings.

3.2 CLIP and CLAP

CLIP (Radford et al., 2021) aligns image and text embeddings using a contrastive objective, while CLAP (Elizalde et al., 2023) replaces the image encoder with an audio encoder and learns joint audio-text representations. For a batch of N aligned audio-text pairs (a_i, t_i), with normalised embeddings z_i^a and z_i^t, CLAP typically optimises a symmetric InfoNCE loss (van den Oord et al., 2019):

$$\mathcal{L} = \frac{1}{2}\left(\mathcal{L}_{a\to t} + \mathcal{L}_{t\to a}\right), \tag{1}$$

$$\mathcal{L}_{a\to t} = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{\exp(\mathrm{sim}(z_i^a, z_i^t)/\tau)}{\sum_{j=1}^{N}\exp(\mathrm{sim}(z_i^a, z_j^t)/\tau)}, \tag{2}$$

with $\mathcal{L}_{t\to a}$ defined analogously. Here, $\mathrm{sim}(\cdot,\cdot)$ is cosine similarity and τ is a temperature parameter. Because label descriptions can be encoded as text, this framework naturally supports zero-shot classification.

3.3 Audio Representation with CLAP

We use a pre-trained CLAP audio encoder to extract d-dimensional audio embeddings, with $d = 512$ in our implementation. Audio is resampled and padded or truncated to a fixed duration before encoding:

$$z^a = \frac{f_{\text{audio}}(x)}{\|f_{\text{audio}}(x)\|_2}. \tag{3}$$

These embeddings are used in three ways: for zero-shot prompt similarity, for few-shot supervised contrastive adaptation, and as features for downstream supervised classifiers.

3.4 Zero-shot Prompting

For zero-shot classification, we define two prompts p_0 and p_1 representing the non-abusive and abusive classes, respectively, for example: *"This audio contains hate speech"* and *"This audio does not contain hate speech"*. We compute normalised text embeddings:

$$z_c^t = \frac{f_{\text{text}}(p_c)}{\|f_{\text{text}}(p_c)\|_2}, \tag{4}$$

and predict the label using prompt similarity:

$$\hat{y} = \arg\max_{c\in\{0,1\}} \mathrm{sim}(z^a, z_c^t). \tag{5}$$

3.5 Few-Shot Contrastive Adaptation

To adapt CLAP to abusive speech detection in low-resource conditions, we construct support sets with k examples per class per language from the training data. These few-shot examples are used to

refine the embedding space through supervised contrastive learning and then we consider two adaptation strategies. In **Projection-only adaptation**, the CLAP encoder is frozen and only a lightweight projection head is trained. In **Projection+fine-tuning**, the projection head is trained jointly with the final audio encoder blocks.

Given projected embeddings $u_i = g(z_i^a)$, we optimise the supervised contrastive objective, based on (Khosla et al., 2020):

$$\mathcal{L}_{\text{sup}} = \sum_{i \in I} \frac{-1}{|P(i)|} \sum_{p \in P(i)} \log \frac{\exp\left(\text{sim}(u_i, u_p)/\tau\right)}{\sum_{a \in A(i)} \exp\left(\text{sim}(u_i, u_a)/\tau\right)}, \tag{6}$$

where $P(i)$ denotes the set of positive examples sharing the same label as i, and $A(i)$ denotes all other samples in the batch. This loss encourages embeddings from the same class to form tighter clusters while increasing separation between classes. After adaptation, the resulting embeddings are used for downstream classification.

3.6 Downstream Classifiers

We evaluate two downstream classifiers on frozen and adapted embeddings: a Support Vector Machine (SVM) and a scikit-learn ANN-MLP. These models allow us to assess whether abuse-related information is linearly or non-linearly separable in the CLAP embedding space, and whether few-shot contrastive adaptation improves downstream decision boundaries.

Classifier selection For each language and evaluation setting, we train both the SVM and the ANN-MLP on the adapted or frozen CLAP embeddings and select the higher macro-F1 result. This per-language selection is reported consistently across all tables and is not varied between shot sizes for a given language. We adopt this strategy because no single downstream classifier is uniformly strongest across all languages and transfer settings, a pattern consistent with prior work on embedding-based classification in low-resource multilingual settings. The SVM with an RBF kernel tends to outperform the ANN-MLP on languages with tighter cluster structure in the embedding space, while the ANN-MLP is marginally stronger in cases where class boundaries are non-linear and more diffusesed.

4 Experiments

4.1 Data, Splits, and Preprocessing

Each instance consists of a waveform x, a binary abuse label $y \in \{0, 1\}$, and a language tag $\ell \in \mathcal{L}$. Audio is resampled and padded or truncated to a fixed duration prior to CLAP feature extraction. We follow the train/test partition provided with the ADIMA dataset (Gupta et al., 2022) where available, and preserve this split throughout all experiments. For evaluations that condition on language, training and testing are performed within the corresponding language partitions. Few-shot support sets are always sampled only from the training split, ensuring that no test information leaks into adaptation.

4.2 Representations and Models

Embeddings: We use normalised CLAP audio embeddings as defined in Eq. 3.
Zero-shot prompting: We evaluate prompt-based similarity using the two label prompts described in Eq. 5.
Few-shot adaptation: We adapt representations using supervised contrastive learning on k-shot support sets and then evaluate the adapted embeddings using downstream classifiers.
Downstream classifiers: We use two downstream models, an SVM and a scikit-learn ANN-MLP, on frozen or adapted CLAP embeddings, and report the strongest language-wise configuration where appropriate.

Since the strongest empirical trends in our experiments arise from few-shot transfer settings, the main results section focuses on those results, while zero-shot prompting is treated as an auxiliary evaluation setting.

4.3 Experimental Setups

We evaluate the proposed framework along two axes: *language generalisation setting* and *representation setting*. The main experiments focus on downstream classification over CLAP-derived audio embeddings, using either frozen or contrastively adapted representations, while zero-shot prompting is treated as an auxiliary evaluation setting.

Language generalisation settings:

- **Monolingual:** for each language ℓ, train on ℓ_{train} and test on ℓ_{test}.

Language	ADIMA (0-shot)	Projection-only	Projection+FT
Bengali	**79.10**	76.34 (25-shot)	76.22 (1-shot)
Bhojpuri	–	71.31 (0-shot)	**71.48** (25-shot)
Gujarati	–	**75.52** (0-shot)	**75.52** (0-shot)
Haryanvi	–	**80.23** (25-shot)	79.13 (25-shot)
Hindi	**80.70**	77.76 (10-shot)	78.04 (50-shot)
Kannada	**78.40**	76.67 (25-shot)	76.58 (1-shot)
Malayalam	–	**78.18** (50-shot)	77.60 (0-shot)
Odia	–	**79.67** (25-shot)	79.41 (10-shot)
Punjabi	83.40	**83.65** (5-shot)	82.56 (25-shot)
Tamil	**75.20**	72.50 (10-shot)	73.12 (25-shot)

Table 1: Cross-lingual macro-F1 results (**reported in** %) on ADIMA. For the projection-only and projection+fine-tuning settings, we report the best language-wise macro-F1 obtained across shot settings, with the corresponding shot size shown in parentheses. The projection-only setting updates only the projection head, while projection+fine-tuning updates the projection head and the last two audio encoder blocks. Previously reported ADIMA results from Gupta et al. (2022) are shown where available. Best scores in each row are shown in bold.

- **Leave-one-language-out (LOLO):** for each target language $\ell^\star$, train on $\bigcup_{\ell \neq \ell^\star} \ell_{\text{train}}$ and test on $\ell^\star_{\text{test}}$.
- **Cross-lingual / joint multilingual:** train on $\bigcup_{\ell \in \mathcal{L}} \ell_{\text{train}}$ and report results per target language on ℓ_{test}.

Representation settings:

- **Frozen embeddings:** train downstream classifiers directly on pre-trained CLAP audio embeddings.
- **Few-shot adapted embeddings:** adapt CLAP representations using supervised contrastive learning on a small support set, then evaluate downstream classifiers on the adapted embeddings.
- **Zero-shot prompting:** classify audio using similarity to abusive and non-abusive text prompts.

4.4 Implementation and Metrics

We implement [1] CLAP feature extraction and adaptation in PyTorch and HuggingFacem using the Supervised Contrastive Loss (Khosla et al., 2020). The downstream SVM and ANN-MLP classifiers are implemented with scikit-learn. Random seeds are fixed for reproducibility, including few-shot sampling. We report Accuracy and Macro-F1, with language-wise results across languages. Since the original ADIMA paper (Gupta et al., 2022) reported Macro-F1, we use Macro-F1 as the main comparison metric in this paper. We evaluate 0-, 1-, 5-, 10-, 25-, and 50-shot settings for supervised contrastive adaptation. Full hyper-parameter settings are provided in Appendix A.2.

[1] Code attached as supplementary material

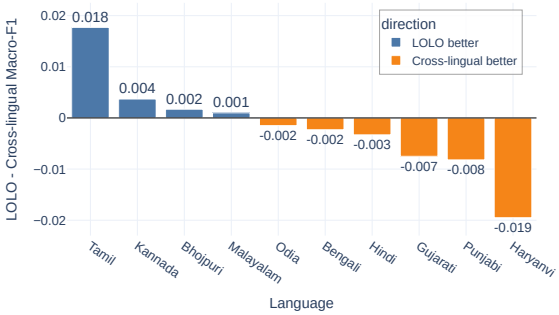
5 Results

5.1 Cross-lingual Results

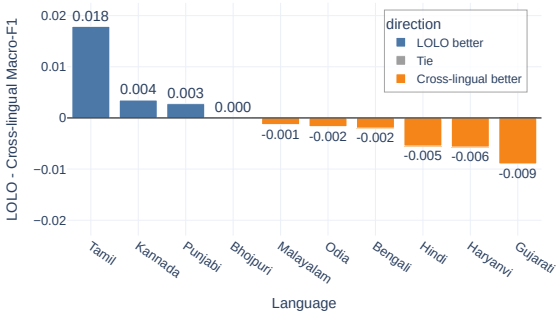
Table 1 reports language-wise cross-lingual macro-F1 (in %) for the proposed CLAP-based systems under the **projection-only** and **projection+fine-tuning** settings, together with previously reported ADIMA results (Gupta et al., 2022) where available. For each adaptation strategy, we report the best language-wise macro-F1 across the evaluated shot sizes, with the corresponding shot value shown in parentheses.

Overall, the results show that CLAP provides a strong basis for cross-lingual abusive speech detection across all ten Indic languages. It is important to note that the ADIMA benchmark results reported by Gupta et al. (2022) reflect fully supervised systems trained on the complete per-language training sets, whereas our CLAP-based framework operates in a few-shot regime, using at most 50 labelled examples per class per language. The ADIMA figures therefore represent an upper-bound reference under full supervision rather than a directly comparable few-shot baseline. Under this framing, the degree to which our approach closes the gap to full supervision is the relevant measure of performance.

The strongest result under the proposed framework is obtained for Punjabi, where projection-only adaptation reaches 83.65 macro-F1 at 5-shot, marginally surpassing the fully supervised ADIMA benchmark of 83.40. Projection-only adaptation also performs strongly for Haryanvi (80.23 at 25-shot), Odia (79.67 at 25-shot), and Malayalam (78.18 at 50-shot), suggesting that CLAP-based representations transfer effectively to these low-resource Indic languages with very limited supervision.



(a) Projection-only adaptation: language-wise LOLO–cross-lingual macro-F1 difference. Positive values indicate that LOLO outperforms cross-lingual.



(b) Projection+fine-tuning: language-wise LOLO–cross-lingual macro-F1 difference. Positive values indicate that LOLO outperforms cross-lingual.

Figure 2: Language-wise difference between leave-one-language-out (LOLO) and cross-lingual evaluation. Positive values indicate that excluding the target language from training improves performance, while negative values indicate that the standard cross-lingual setting remains stronger.

For languages where a gap to the fully supervised ADIMA system remains — Bengali (76.34 vs. 79.10), Hindi (77.76 vs. 80.70), Kannada (76.67 vs. 78.40), and Tamil (72.50 vs. 75.20) — this gap should be interpreted in the context of the substantially different supervision regime. The key finding is not that CLAP uniformly exceeds full supervision, but that projection-only adaptation with as few as 5–25 labelled examples per class achieves results within 2–4 macro-F1 points of systems trained on the full dataset for the majority of languages. This represents a practically meaningful result for low-resource settings where full annotation is infeasible.

5.2 Leave-One-Language-Out Analysis

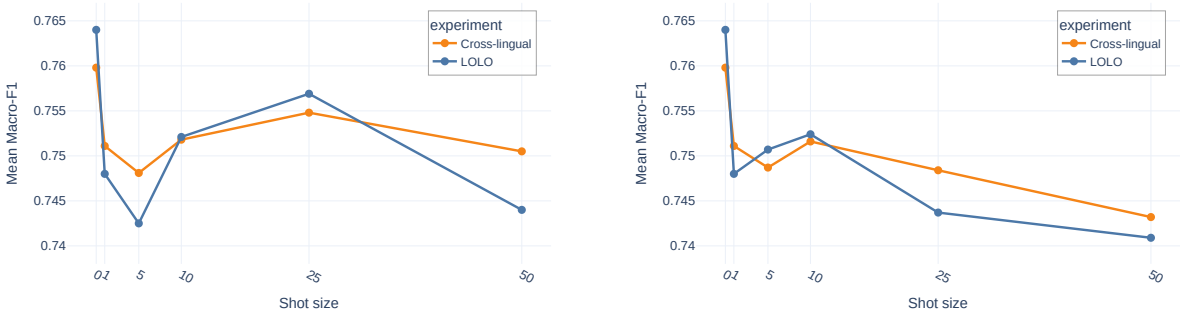
To better understand the role of target-language supervision, we compare the standard cross-lingual setting with a multilingual leave-one-language-out (LOLO) setup. In the cross-lingual setting, the target language is included in the multilingual training data, whereas in LOLO it is fully excluded and used only for testing. This comparison provides a stronger test of transfer: if performance remains stable under LOLO, the model is likely capturing language-agnostic cues of abusive speech rather than relying mainly on target-language-specific supervision.

Figures 2 and 3 present this comparison from two complementary perspectives. Figure 2 shows the language-wise difference between LOLO and cross-lingual macro-F1, while Figure 3 shows the mean macro-F1 across shot sizes. Across both adaptation settings, a consistent pattern emerges: **LOLO remains competitive with cross-lingual evaluation, but does not outperform it on average**. The mean performance curves in Figures 3a and 3b show that cross-lingual training retains a small but persistent overall advantage, indicating that direct exposure to the target language still provides useful information.

At the same time, the language-wise comparison reveals a more nuanced picture. As shown in Figures 2a and 2b, LOLO outperforms cross-lingual for a subset of languages, most clearly for **Tamil** in both settings. Smaller LOLO gains are also observed for **Kannada** and **Bhojpuri**, and for **Punjabi** under projection+fine-tuning. These cases suggest that multilingual transfer can generalise well even when the target language is entirely absent from training. One plausible explanation is that the shared multilingual CLAP representation already captures some abuse-related acoustic cues in a form that transfers across languages.

However, LOLO does not produce consistent improvements across the full benchmark. In both adaptation settings, cross-lingual remains clearly stronger for languages such as **Haryanvi** and **Gujarati**, and the margins in favour of cross-lingual are often larger than the margins in favour of LOLO. This indicates that some languages benefit more strongly from target-language exposure, even in a multilingual training regime. The learned representation therefore appears to combine both transferable cross-lingual structure and language-specific information, with the relative importance of each varying across languages.

The shot-wise curves further show that the



(a) Projection-only adaptation: mean macro-F1 across shot sizes for LOLO and cross-lingual evaluation.

(b) Projection+fine-tuning: mean macro-F1 across shot sizes for LOLO and cross-lingual evaluation.

Figure 3: Mean macro-F1 across shot sizes for leave-one-language-out (LOLO) and cross-lingual evaluation. In both adaptation settings, LOLO remains competitive but cross-lingual retains a small average advantage.

LOLO effect is **small and unstable rather than monotonic**. The gap between LOLO and cross-lingual fluctuates across shot sizes rather than widening or narrowing consistently. This suggests that the effect of excluding the target language is not determined solely by supervision volume, but also by the interaction between language identity, adaptation regime, and the structure of the pre-trained embedding space. Put differently, target-language exclusion is not uniformly harmful, but neither does it systematically improve multilingual transfer.

We therefore interpret LOLO not as a stronger alternative to cross-lingual training, but as a **transfer analysis** that quantifies how much the system depends on target-language supervision. The fact that LOLO remains close to cross-lingual performance in both settings supports the claim that the model learns partially language-agnostic representations of abusive speech. At the same time, the persistent average advantage of cross-lingual training indicates that transfer is meaningful but incomplete. A future study, in association with linguists and paralinguists with respect to language families, can reveal hidden associations, which could potentially be a follow-up work.

6 Conclusion

This paper investigated whether CLAP-based audio-text representations can support multilingual abusive speech detection directly from audio, particularly in low-resource and cross-lingual settings. Across the ADIMA benchmark, the results show that CLAP provides a strong representation space for abusive speech detection, with projection-only adaptation achieving the best overall cross-lingual average macro-F1 of **77.18**. At the language level, the strongest result was obtained for **Punjabi**, where projection-only adaptation reached **83.65** macro-F1 at **5-shot**, slightly surpassing the reported ADIMA benchmark of **83.40**.

The results also show that the benefits of few-shot contrastive adaptation are not uniform across languages. Some languages, such as **Gujarati**, achieved their best performance in the **0-shot** setting, whereas others, such as **Haryanvi** and **Odia**, benefited most from **25-shot** adaptation. Across the two adaptation strategies, projection-only adaptation generally proved more effective and stable than projection+fine-tuning, indicating that much of the useful task-specific information can be recovered through a lightweight projection space without extensive encoder updating.

Finally, the leave-one-language-out analysis showed that the model learns meaningful cross-lingual structure: LOLO remained close to standard cross-lingual performance, with average macro-F1 values of **76.99** for projection-only adaptation and **76.92** for projection+fine-tuning, compared with **77.18** and **76.97** in the corresponding cross-lingual settings. Overall, these findings suggest that contrastive audio-text models are a promising basis for multilingual audio abuse detection, while also highlighting that transfer remains language-dependent and incomplete. Future work should explore stronger prompt design, parameter-efficient adaptation, multimodal fusion, and deeper analysis of language-specific transfer behaviour.

Limitations

Our findings should be interpreted in light of several limitations. First, the experiments are

conducted on a single multilingual benchmark, ADIMA, so the results may not generalise directly to other datasets, recording conditions, or non-Indic languages. Second, although CLAP is motivated partly by its zero-shot capability, the strongest results in this paper come from few-shot adaptation and downstream classification; zero-shot prompting is therefore only an auxiliary evaluation setting. Third, the few-shot results may be sensitive to support-set composition and optimisation choices, particularly in very low-shot regimes (1-shot and 5-shot). We fix random seeds for reproducibility and report results across six shot sizes ($k \in \{0, 1, 5, 10, 25, 50\}$) rather than relying on a single condition, which partially mitigates this concern by revealing whether performance trends are consistent across data regimes. Nonetheless, we do not report variance across multiple random seeds, and results at individual shot sizes for individual languages should be interpreted with appropriate caution. Future work should include multi-seed evaluation to quantify support-set sampling variance, particularly for the 1-shot condition where instability is most likely. Fourth, we evaluate only two downstream classifiers, SVM and ANN-MLP, and compare only two lightweight adaptation strategies, so the conclusions do not extend to all possible fine-tuning or parameter-efficient methods. Finally, we do not conduct a dedicated fairness analysis across speakers, dialects, or culturally specific abusive expressions.

Ethics Statement

This study does not involve any personal or public data pointing to an individual or a group of individuals and thus does not break any ethical guidelines.

References

Jinmyeong An, Wonjun Lee, Yejin Jeon, Jungseul Ok, Yunsu Kim, and Gary Geunbae Lee. 2024. An investigation into explainable audio hate speech detection. *arXiv preprint arXiv:2408.06065*.

Kalyanaswamy Banuroopa and D Shanmuga Priyaa. 2021. Mfcc based hybrid fingerprinting method for audio classification through lstm. *International Journal of Nonlinear Analysis and Applications*, 12(Special Issue):2125–2136.

Bharathi Raja Chakravarthi, Ruba Priyadharshini, Sajeetha Thavareesan, Elizabeth Sherly, Saranya Rajiakodi, Balasubramanian Palani, Malliga Subramanian, Subalalitha Cn, Dhivya Chinnappa, and 1 others. 2025. Proceedings of the fifth workshop on speech, vision, and language technologies for dravidian languages. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*.

Anusha Chhabra and D. Vishwakarma. 2023. A literature survey on multimodal and multilingual automatic hate speech identification. *Multimedia Systems*, pages 1–28.

Marta Costa-jussà, Mariano Meglioli, Pierre Andrews, David Dale, Prangthip Hansanti, Elahe Kalbassi, Alexandre Mourachko, Christophe Ropers, and Carleigh Wood. 2024. Mutox: Universal multilingual audio-based toxicity dataset and zero-shot detector. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 5725–5734.

Benjamin Elizalde, Soham Deshmukh, Mahmoud Al Ismail, and Huaming Wang. 2023. Clap learning audio concepts from natural language supervision. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Tanmay Garg, Sarah Masud, Tharun Suresh, and Tanmoy Chakraborty. 2022. Handling bias in toxic speech detection: A survey. *ACM Computing Surveys*, 55:1 – 32.

Sreyan Ghosh, Samden Lepcha, Sakshi S., Rajiv Ratn Shah, and Srinivasan Umesh. 2022. DeToxy: A large-scale dataset for toxicity classification in spoken utterances. In *Proceedings of Interspeech 2022*, pages 5185–5189.

Vikram Gupta, Rini Sharon, Ramit Sawhney, and Debdoot Mukherjee. 2022. ADIMA: Abuse detection in multilingual audio. In *Proceedings of ICASSP 2022*, pages 6172–6176. ArXiv:2202.07991.

Amir Reza Jafari, Guanlin Li, Praboda Rajapaksha, Reza Farahbakhsh, and Noel Crespi. 2023. Fine-grained emotions influence on implicit hate speech detection. *IEEE Access*, 11:105330–105343.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. 2020. Supervised contrastive learning. *Advances in neural information processing systems*, 33:18661–18673.

Joseph Liu, Mahesh Kumar Nandwana, Janne Pylkk"onen, Hannes Heikinheimo, and Morgan McGuire. 2024. Enhancing Multilingual Voice Toxicity Detection with Speech-Text Alignment. In *Interspeech 2024*, pages 4298–4302.

Yu-Xiang Luo, Yi-Cheng Lin, Ming-To Chuang, Jia-Hung Chen, I-Ning Tsai, Pei Xing Kiew, Yueh-Hsuan Huang, Chien-Feng Liu, Yu-Chen Chen, Bo-Han Feng, Wenze Ren, and Hung yi Lee. 2025. ToxicTone: A Mandarin Audio Dataset Annotated for Toxicity and Toxic Utterance Tonality. In *Interspeech 2025*, pages 4008–4012.

Khouloud Mnassri, Praboda Rajapaksha, Reza Farahbakhsh, and Noel Crespi. 2022. Bert-based ensemble approaches for hate speech detection. In *GLOBECOM 2022 - 2022 IEEE Global Communications Conference*, pages 4649–4654.

Marzieh Mozafari, Reza Farahbakhsh, and Noël Crespi. 2020. A bert-based transfer learning approach for hate speech detection in online social media. In *Complex Networks and Their Applications VIII*, pages 928–940, Cham. Springer International Publishing.

Arpan Nandi, Kamal Sarkar, Arjun Mallick, and Arkadeep De. 2024. Combining multiple pre-trained models for hate speech detection in bengali, marathi, and hindi. *Multimedia Tools and Applications*, 83:77733 – 77757.

Rachna Narula and Poonam Chaudhary. 2024. A comprehensive review on detection of hate speech for multi-lingual data. *Social Network Analysis and Mining*, 14.

Radha Natarajan, R Swathika, and 1 others. 2025. Multimodal hate speech detection in dravidian languages using text and audio features. In *2025 11th International Conference on Communication and Signal Processing (ICCSP)*, pages 148–152. IEEE.

Qi Qian and Juhua Hu. 2024. Online zero-shot classification with clip. pages 462–477.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. *Preprint*, arXiv:2103.00020.

Rishabh Ranjan, Likhith Ayinala, Mayank Vatsa, and Richa Singh. 2025a. Multimodal Zero-Shot Framework for Deepfake Hate Speech Detection in Low-Resource Languages. In *Interspeech 2025*, pages 1678–1682.

Rishabh Ranjan, Kishan Pipariya, Mayank Vatsa, and Richa Singh. 2025b. Synhate: Detecting hate speech in synthetic deepfake audio. *arXiv preprint arXiv:2506.06772*.

Priyanshu Rawat, Madhvan Bajaj, Satvik Vats, and Vikrant Sharma. 2023. A comprehensive study based on mfcc and spectrogram for audio classification. *Journal of Information and Optimization Sciences*, 44(6):1057–1074.

P. Roy, Snehaan Bhawal, and C. N. Subalalitha. 2022. Hate speech and offensive language detection in dravidian languages using deep ensemble framework. *Comput. Speech Lang.*, 75:101386.

Kaustubh V Sakhare and Radhika V Kulkarni. 2023. Deep learning-based intelligent systems for audio abuse prediction: A survey. In *AI-Based Metaheuristics for Information Security and Digital Media*, pages 35–47. Chapman and Hall/CRC.

Aditya Narayan Sankaran, Reza Farahbakhsh, and Noel Crespi. 2025. Towards cross-lingual audio abuse detection in low-resource settings with few-shot learning. In *Proceedings of COLING 2025*.

Rini Sharon, Heet Shah, Debdoot Mukherjee, and Vikram Gupta. 2022. Multilingual and multimodal abuse detection. In *Proceedings of Interspeech 2022*.

Anika A. Spiesberger, Andreas Triantafyllopoulos, Iosif Tsangko, and Bj"orn W. Schuller. 2023. Abusive Speech Detection in Indic Languages Using Acoustic Features. In *Interspeech 2023*, pages 2683–2687.

Éva Székely, Joakim Gustafson, and Ilaria Torre. 2023. Prosody-controllable gender-ambiguous speech synthesis: A tool for investigating implicit bias in speech perception. In *Interspeech*, pages 1234–1238.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2019. Representation learning with contrastive predictive coding. *Preprint*, arXiv:1807.03748.

Hua Wang, Hong Liu, Jiale Ren, Mingxin Tan, and Zhongzien Jiang. 2025. Clip-6d: Empowering clip as a zero-shot 6d pose estimator through generalizable object-specific representations. *Proceedings of the 33rd ACM International Conference on Multimedia*.

Máté Ákos Tündik, Valér Kaszás, and György Szaszák. 2019. Assessing the semantic space bias caused by asr error propagation and its effect on spoken document summarization. pages 1333–1337.

A Appendix

A.1 Additional Cross-Lingual and LOLO Score Tables

The tables in this appendix provide a more detailed view of the best-performing language-wise results under the multilingual leave-one-language-out (LOLO) and cross-lingual settings. While the main paper focuses on overall trends and summary comparisons, Tables 2–5 report, for each language, the highest macro-F1 obtained in the best-performing configurations together with the corresponding accuracy and shot size. This allows closer inspection of how the optimal supervision level varies across languages. Table 6 further provides a compact comparison between the best cross-lingual results of the proposed framework and the previously reported ADIMA benchmark.

Several patterns become clearer in these detailed tables. First, the best-performing shot size is highly language-dependent rather than consistent across the benchmark: some languages, such as Gujarati and Bhojpuri, achieve their strongest results in the 0-shot setting, whereas others, including Haryanvi,

Language	Macro-F1	Accuracy	Best shot
Bengali	75.96	76.22	1-shot
Bhojpuri	71.48	73.81	0-shot
Gujarati	74.77	78.18	25-shot
Haryanvi	78.28	78.42	0-shot
Hindi	77.49	77.51	10-shot
Kannada	77.04	78.59	10-shot
Malayalam	78.28	81.18	50-shot
Odia	79.51	80.55	50-shot
Punjabi	82.83	82.83	1-shot
Tamil	74.27	78.71	10-shot

Table 2: Best multilingual leave-one-language-out (LOLO) results for the best-performing projection-only configuration. The best row per language is selected by macro-F1, with accuracy used as a tie-breaker.

Language	Macro-F1	Accuracy	Best shot
Bengali	75.96	76.22	1-shot
Bhojpuri	71.48	73.81	0-shot
Gujarati	74.63	77.90	0-shot
Haryanvi	78.56	78.69	25-shot
Hindi	77.49	77.51	50-shot
Kannada	76.93	79.13	1-shot
Malayalam	77.36	79.84	0-shot
Odia	79.02	79.73	25-shot
Punjabi	82.83	82.83	1-shot
Tamil	74.91	79.51	25-shot

Table 3: Best multilingual leave-one-language-out (LOLO) results for the best-performing projection+fine-tuning configuration. The best row per language is selected by macro-F1, with accuracy used as a tie-breaker.

Language	Macro-F1	Accuracy	Best shot
Bengali	76.34	77.03	25-shot
Bhojpuri	71.31	73.51	0-shot
Gujarati	75.52	78.73	0-shot
Haryanvi	80.23	80.33	25-shot
Hindi	77.76	77.78	10-shot
Kannada	76.67	78.59	25-shot
Malayalam	78.18	81.18	50-shot
Odia	79.67	80.27	25-shot
Punjabi	83.65	83.65	5-shot
Tamil	72.50	76.82	10-shot

Table 4: Best cross-lingual results for the best-performing projection-only configuration. The best row per language is selected by macro-F1, with accuracy used as a tie-breaker.

Language	Macro-F1	Accuracy	Best shot
Bengali	76.22	76.49	1-shot
Bhojpuri	71.48	73.81	25-shot
Gujarati	75.52	78.73	0-shot
Haryanvi	79.13	79.24	25-shot
Hindi	78.04	78.05	50-shot
Kannada	76.58	78.59	1-shot
Malayalam	77.60	80.38	0-shot
Odia	79.41	80.00	10-shot
Punjabi	82.56	82.56	25-shot
Tamil	73.12	77.63	25-shot

Table 5: Best cross-lingual results for the best-performing projection+fine-tuning configuration. The best row per language is selected by macro-F1, with accuracy used as a tie-breaker.

Odia, and Malayalam, benefit from larger support sets. Second, the detailed language-wise results reinforce the main discussion in the paper: cross-lingual evaluation remains slightly stronger overall than LOLO for both projection-only and projection+fine-tuning, but the gaps are generally small. Finally, the compact comparison in Table 6 shows that the projection-only setting (projection-only) remains competitive with, and in the case of Punjabi slightly stronger than, the ADIMA benchmark.

A.2 Hyper-parameters and Training Details

All experiments use a pre-trained CLAP [2] audio encoder as the base representation model. Audio inputs are resampled to the target sampling rate required by the encoder and padded or truncated to a fixed maximum duration before feature extraction. For few-shot experiments, support sets are constructed by sampling k examples per class per language from the training split, with $k \in \{0, 1, 5, 10, 25, 50\}$. The k=0 setting corresponds to using frozen pre-trained CLAP embeddings without task-specific adaptation. We trained the model with 50 epochs with a 1e-4 Learning rate with `AdamW` optimizer with default warm up initialization. We trained the model on the Apple MPS back-end with the PyTorch metal library.

We study two adaptation variants. In the **projection-only** setting, the CLAP encoder is frozen and only a lightweight projection head is trained using supervised contrastive learning. In the **projection+fine-tuning** setting, the projection head is trained jointly with the final audio encoder blocks. The projection-only setup serves as the default adaptation strategy in the main experiments, whereas projection+fine-tuning is included to test whether limited encoder adaptation yields additional gains.

For supervised contrastive learning, we optimise a temperature-scaled supervised contrastive objective over mini-batches drawn from the support set. We used supervised contrastive loss (Khosla et al., 2020) as the loss function for projection training.

[2] https://hf.co/laion/clap-htsat-unfused

Language	ADIMA	projection-only	projection+fine-tuning
Bengali	79.10	76.34 (25)	76.22 (1)
Bhojpuri	–	71.31 (0)	71.48 (25)
Gujarati	–	75.52 (0)	75.52 (0)
Haryanvi	–	80.23 (25)	79.13 (25)
Hindi	80.70	77.76 (10)	78.04 (50)
Kannada	78.40	76.67 (25)	76.58 (1)
Malayalam	–	78.18 (50)	77.60 (0)
Odia	–	79.67 (25)	79.41 (10)
Punjabi	83.40	83.65 (5)	82.56 (25)
Tamil	75.20	72.50 (10)	73.12 (25)

Table 6: Compact cross-lingual macro-F1 comparison against the ADIMA benchmark. Numbers in parentheses denote the best shot size. Cell shading marks the highest available macro-F1 in each row.

The main experiments evaluate multiple shot sizes rather than a single low-resource condition, since performance is sensitive to both support-set size and optimisation choices. In general, projection-only adaptation is more stable under higher learning rates, whereas projection+fine-tuning requires a smaller encoder learning rate to avoid disrupting the pre-trained representation space. Empirically, projection-only adaptation proved more consistent across languages, while deeper fine-tuning yielded only modest gains for a smaller subset of languages.

After adaptation, embeddings are extracted for all train and test samples and used as input to downstream classifiers. In the final experimental setup, we use two downstream models: a **Support Vector Machine (SVM)** and a **scikit-learn ANN-MLP**. These two models were retained because they consistently produced the strongest results across languages and transfer settings. The SVM provides a strong non-linear margin-based classifier for the CLAP embedding space, while the ANN-MLP offers a lightweight neural alternative that can capture non-linear structure without requiring end-to-end retraining of the audio encoder.

Model selection is based primarily on macro-F1, with accuracy used as a tie-breaker where necessary. Since the strongest-performing model varies across languages and evaluation settings, the appendix tables report the best language-wise configuration rather than assuming that a single downstream classifier is uniformly optimal across the benchmark.

Random seeds are fixed for few-shot sampling, representation adaptation, and downstream model training wherever applicable. Because few-shot support sets are small, results can vary with both shot size and optimisation settings. This is one reason why the main paper reports results across multiple shot values rather than relying on a single few-shot condition.

A.3 Additional Cross-Lingual and LOLO Visualisations

To complement the main results, we provide additional visual summaries of the cross-lingual and leave-one-language-out (LOLO) analyses in Figures 4, 5, and 6. These plots are intended to reduce table fatigue and make the language-wise trends easier to compare visually.

Figure 4 presents the best cross-lingual macro-F1 scores by language for the ADIMA benchmark and for our two adaptation settings: projection-only adaptation and projection + partial fine-tuning. The plot makes clear that the proposed CLAP-based framework is competitive across all ten languages, with especially strong results for Punjabi, Haryanvi, and Odia. In particular, projection-only slightly surpasses the reported ADIMA benchmark for Punjabi (83.65 vs. 83.40), while remaining close to the benchmark for several other languages. At the same time, the figure also highlights that the strongest system is language-dependent: ADIMA remains better for Bengali, Hindi, Kannada, and Tamil, indicating that cross-lingual transfer quality is uneven across languages.

Figure 5 directly compares LOLO and cross-lingual macro-F1 across languages for both adaptation settings. This visualisation reinforces the discussion in Section 5.2: LOLO remains highly competitive, but it does not consistently outperform the standard cross-lingual setting. For some languages, especially Tamil, and to a lesser extent Kannada and Bhojpuri, LOLO yields slightly stronger results, suggesting that multilingual transfer can remain effective even when the target language is excluded from training. However, for other languages such as Gujarati and Haryanvi, cross-lingual performance remains clearly higher, indicating that

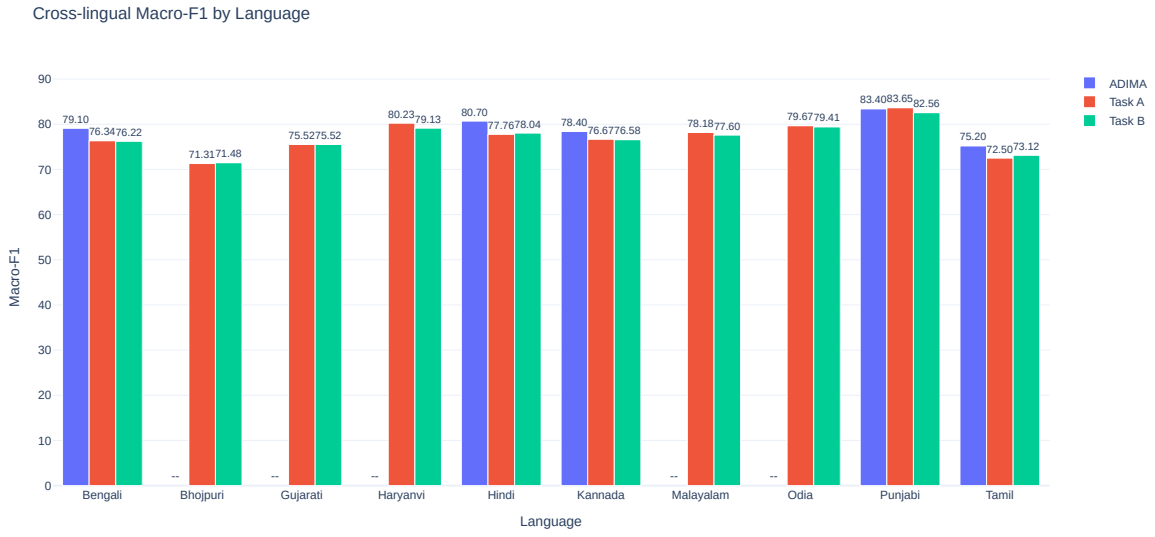


Figure 4: Best cross-lingual macro-F1 by language for ADIMA, projection-only, and projection+fine-tuning. The plot highlights both the competitiveness of the proposed CLAP-based framework and the language-dependent variation in transfer performance.

target-language exposure still provides useful information.

Finally, Figure 6 summarises the average macro-F1 across evaluation setups. The mean scores are all tightly clustered, with Cross-A achieving the highest overall average (77.18), followed by LOLO-A (76.99), Cross-B (76.97), and LOLO-B (76.92). This confirms that the differences between LOLO and cross-lingual settings are relatively small overall, even though the language-wise plots reveal more substantial local variation. Taken together, these figures support two main observations: first, CLAP-based representations transfer reasonably well across languages; second, target-language exclusion does not cause catastrophic degradation, but cross-lingual training retains a slight overall advantage.

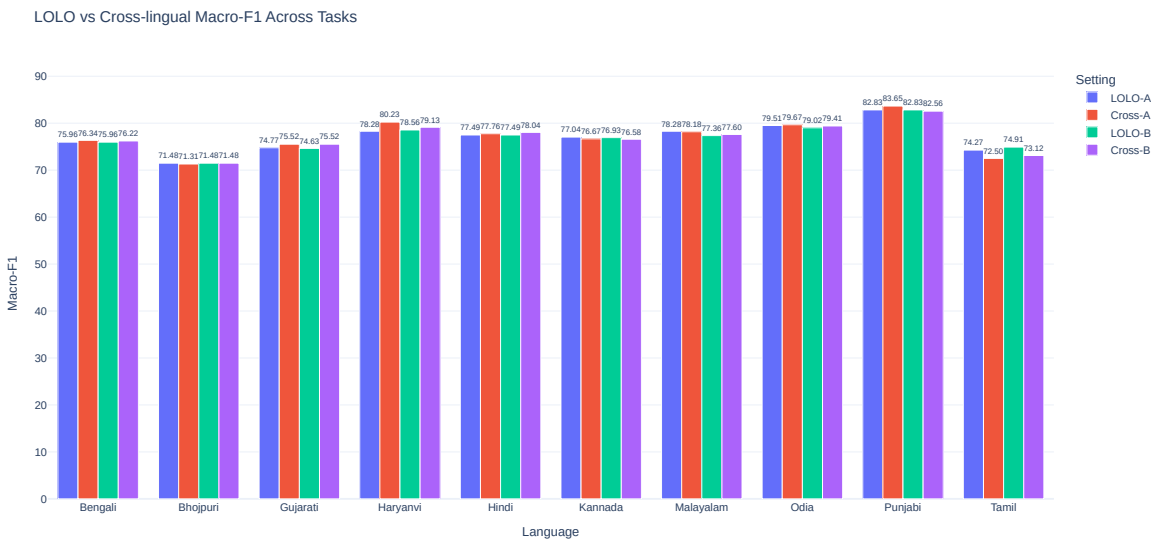


Figure 5: Language-wise comparison between leave-one-language-out (LOLO) and cross-lingual macro-F1 under projection-only and projection+fine-tuning. LOLO remains competitive across languages, but cross-lingual performance is slightly stronger overall.

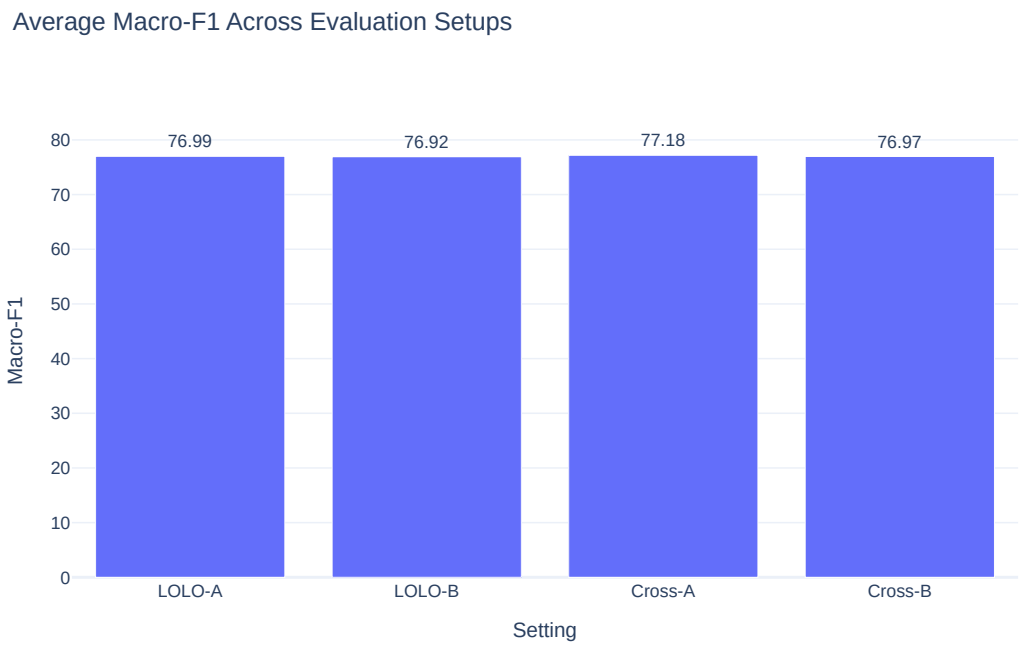


Figure 6: Average macro-F1 across the four evaluation settings: LOLO-A, LOLO-B, Cross-A, and Cross-B. Cross-A achieves the highest average score, although the overall differences remain small.