IIIT-BGP IWSLT 2026 systems for Low-resource ST

Kaustuk Pratap Singh[1,*] **Dipanshu**[1,*] **Vedant Singh**[1,*] **Kumar Rishu**[2]

[1]Indian Institute of Information Technology Bhagalpur [2]Charles University

{kaustuk.240103196, dipanshu.240103227, vedant.240102162}@iiitbh.ac.in

65549587@cuni.cz

[*]Equal contribution

Abstract

We present low-resource Bhojpuri-Hindi speech translation systems for the IWSLT 2026 shared task (Adelani et al., 2026), covering both end-to-end and cascaded settings. Our end-to-end model connects a Bhojpuri-finetuned Wav2Vec2 encoder to a pretrained NLLB-200 decoder via a lightweight interconnection adapter that combines learnable layer aggregation, CNN-based temporal compression, and Transformer refinement, with optional LoRA-based decoder adaptation. For our cascaded system, we finetune Whisper for Bhojpuri ASR and NLLB-200 for Hindi MT, and further apply QE Fusion with COMET-Kiwi to improve translation selection from beam candidates.

1 Introduction

Speech translation (ST) is the task of translating speech from one language into text in another, serving as an important component in breaking language barriers within communities (Joshi et al., 2019). While substantial progress has been made in ST for high-resource languages, performance in low-resource settings remains limited. One such scenario involves Bhojpuri, which, despite being spoken by over 50 million people, still falls into the low-resource category due to insufficient high-quality speech-to-text data, non-standardized orthography, and frequent code-mixing with Hindi/Angika, all of which complicate both speech recognition and translation.

ST systems are broadly categorized into two types:

- Cascaded
- End-to-end

Cascade speech translation systems, which combine automatic speech recognition (ASR) and machine translation (MT), remain a common choice in low-resource settings due to the availability of pretrained ASR and MT models and their relatively low data requirements. However, such systems suffer from error propagation, where mistakes introduced during ASR are amplified by the downstream MT component,leading to degraded translation quality (Ruiz and Federico, 2014). Recent end-to-end models, such as speech-to-text transformers and speech-integrated large language models, have shown remarkable performance in reducing ASR error propagation and the inherent latency of cascaded pipelines (Sperber and Paulik, 2020; Papi et al., 2025); however, these models typically require substantial labeled data and can struggle to generalize in low-resource scenarios, making cascaded systems a competitive alternative (Papi et al., 2025).

In this work, we systematically investigated both cascade and end-to-end approaches for low-resource Bhojpuri-Hindi speech translation. Our end-to-end system, combining learnable encoder-layer aggregation, a Transformer-based modality adapter, and Low-Rank Adaptation (LoRA; Hu et al., 2022) based decoder adaptation, achieves a best dev BLEU of 32.77, while our cascade system with QE-guided hypothesis fusion yields improvements over the beam-search baseline across BLEU (Papineni et al., 2002), chrF++ (Popović, 2015), and COMET (Rei et al., 2020), with the most notable gain observed on COMET.

Our main contributions are:

- A modality adaptor combining CNN-based temporal compression with a Transformer refinement block to align Wav2Vec2 encoder representation with a pretrained NLLB decoder.
- Incorporation of an encoder-side, layer-wise aggregation strategy over intermediate Wav2Vec2 layers.

Proceedings of the 23rd International Conference on Spoken Language Translation (IWSLT 2026), pages 272–283

- Incorporation of parameter-efficient decoder adaptation using LoRA, comparing fully frozen, LoRA-only, and partially unfrozen configurations of the top decoder layers.

- QE Fusion (Vernikos and Popescu-Belis, 2024) as a replacement for standard beam search in the cascade MT component, using COMET-Kiwi (Rei et al., 2022) scores and reconstruct a superior hypothesis from 5 diverse candidates.

2 Related Work

Language-specific work. ASR and MT have evolved from foundational linguistic tooling to sophisticated neural frameworks that leverage cross-lingual transfer and instruction tuning (Costa-Jussà et al., 2022). Early computational efforts for Bhojpuri focused on establishing basic linguistic resources, such as the first statistical Part-of-Speech (POS) tagger using Support Vector Machines (Fernando et al., 2016). Mundotiya et al. (2021) extended this work by releasing annotated corpora covering POS, morphology, and chunking for Bhojpuri alongside related Purvanchal languages, providing the first systematic benchmarks for POS tagging, morphology analysis, and chunking. On the speech side,(Kumar et al., 2022) curated an annotated speech corpus for Bhojpuri and closely related varieties, establishing early ASR baselines. Neural MT for Bhojpuri-Hindi has received limited dedicated attention; the NLLB-200 project (Costa-Jussà et al., 2022) includes Bhojpuri as one of its languages, effectively making large-scale multilingual transfer the dominant strategy for this pair.

End-to-end architectures. End-to-end speech translation (E2E-ST) directly maps source-language speech to target-language text without an explicit ASR-MT cascade. Early work showed that sequence-to-sequence models could directly perform speech translation, establishing the feasibility of direct E2E-ST (Weiss et al., 2017). More recent research has emphasized the importance of pretrained speech and text models, particularly in low-resource settings where parallel ST data is limited (Niehues et al., 2021). Another recent direction in E2E-ST is to couple pretrained speech encoders with pretrained text decoders via lightweight adaptation modules. Prior work has shown that combining pretrained speech models such as Wav2Vec 2.0(Baevski et al., 2020) or HuBERT with pretrained text decoders such as mBART can substantially improve end-to-end ST quality (Gállego et al., 2021; Tsiamas et al., 2022; Pham et al., 2022). In low-resource settings, related system papers have also explored tightly coupling pretrained ASR encoders with NLLB-based decoders (Avila and Crego, 2025).

Within this modular setting, two lines of work are particularly relevant to our end-to-end approach. M-Adapter (Zhao et al., 2022) addresses the modality gap between speech encoder outputs and text decoder inputs by introducing a dedicated adaptation module that compresses and transforms speech representations before decoding. Nishikawa and Nakamura (2023) shows that relying solely on the final speech encoder layer is suboptimal and instead aggregates intermediate encoder layers via learnable weights before passing them to the decoder.

Evaluation and inference methods. In NMT, much attention has shifted towards neural quality estimators such as COMET (Rei et al., 2020), and its reference-free variant COMET-Kiwi (Rei et al., 2022), which score translation hypotheses without requiring reference translations and correlate more strongly with human judgments than surface-based metrics. Minimum Bayes Risk (MBR) decoding (Kumar and Byrne, 2004) offers one such approach: it selects the hypothesis with the highest expected utility, estimated against a set of sampled pseudo-references. Freitag et al. (Freitag et al., 2022) showed that replacing surface-based utility functions with neural metrics such as COMET yields substantial gains in human evaluation, establishing neural-metric MBR as a strong decoding baseline. Amongst other related works, QE reranking selects the top-scoring hypothesis from a candidate pool using a reference-free estimator, but is constrained to returning one of the existing candidates rather than constructing a new one (Rei et al., 2022). QE Fusion (Vernikos and Popescu-Belis, 2024) addresses this limitation by combining spans from multiple candidates. QE Fusion exploits inter-candidate complementarity to synthesize translations that outperform any single candidate, with consistent gains over beam search, QE reranking, and MBR across five language pairs, including on NLLB (Costa-Jussà et al., 2022).

3 Datasets

For our experiments, we used data from the following sources:

Dataset	Type	Train	Dev	Test	Total	Used In
Official Data (hours)						
IWSLT 2026 Official	ST (bho→hi)	20h	2h	0.5h	22.5h	E2E
Additional Speech Data (hours)						
HuggingFace ASR (real)	ASR (bho)	7h	0.6h	–	7.6h	Cascade
HuggingFace ASR (synthetic)	ASR (bho)	4.2h	–	–	4.2h	Cascade
Additional Text Data (utterances)						
Kaggle bho→hi Corpus	MT (bho→hi)	40.2k	4.5k	–	44.7k	Cascade
Back-Translated	MT (bho→hi)	49.2k	5.5k	–	54.7k	Cascade
Combined Training Sets						
Full Cascade Train Set	ASR+MT	89.5k / 11.2h	2h	0.5h	89.5k / 13.7h	Cascade
Full E2E Train Set (20h official + augmented)	ST	40h	2h	0.5h	42h	E2E

Table 1: Summary of datasets for Bhojpuri→Hindi ST. Speech in hours (h); text in thousands of utterances (k). Back-translated using facebook/nllb-200-1.3B (Costa-Jussà et al., 2022). The Combined Training dev and test sets are the official IWSLT 2026 dev and test sets.

1. Official SLT data released by IWSLT 2026 organizers (Adelani et al., 2026), comprising approximately 22 hours of Bhojpuri news-domain speech paired with Hindi translations, split into training (20h), development (2h), and test (0.5h) sets, serving as the primary aligned speech-translation resource in our setup.

2. AI4Bharat ASR data (Joshi et al., 2025)

3. Additional MT data obtained from Kaggle (dataset link).

The official Bhojpuri–Hindi dataset comprises 22 hours of Bhojpuri news-domain speech and corresponding Hindi translations. Additional ASR data consists of 7 hours of audio speech and corresponding Bhojpuri transcriptions. Additional MT data consists of a total of 44,694 utterances after removing duplicates from the original source (see Table 1).

4 Methodology

4.1 End-to-End

Our end-to-end system combines two complementary ideas: modality adaptation through a dedicated speech-to-text adapter, following M-Adapter (Zhao et al., 2022), and learnable aggregation of intermediate speech encoder layers, following Inter-Connection (Nishikawa and Nakamura, 2023). Concretely, we aggregate selected Wav2Vec2 (Baevski et al., 2020) encoder's hidden-layer representations and pass them through an interconnection adapter that compresses and projects speech embeddings into a form compatible with a pretrained NLLB (Costa-Jussà et al., 2022) decoder.

4.1.1 Speech Encoder

We use the pretrained Bhojpuri-specific Wav2Vec2 backbone from Vakyansh (Chadha et al., 2022). Input speech is converted to mono, resampled to 16 kHz, and fed to the encoder as a raw waveform. Rather than relying only on the final encoder layer, we aggregate hidden representations from layers 6, 8, 10, and 12 using learnable softmax weights. This choice allows the model to exploit complementary information across intermediate and higher-level acoustic representations, while leaving cross-modal alignment to the downstream adapter. The layer selection is motivated by prior observations that different encoder depths capture different linguistic abstractions in self-supervised speech models (Pasad et al., 2024).

4.1.2 Adapter

The adapter is the system's main alignment module. It first performs temporal compression using strided Conv1d layers with GELU activations (Hendrycks and Gimpel, 2016); in our default setting, two stride-2 convolutions yield an effective 4:1 reduction. This stage shortens the speech sequence to make decoder cross-attention tractable while also introducing a local inductive bias that merges nearby frames into more stable short-range units. The compressed states are then normalized and linearly projected

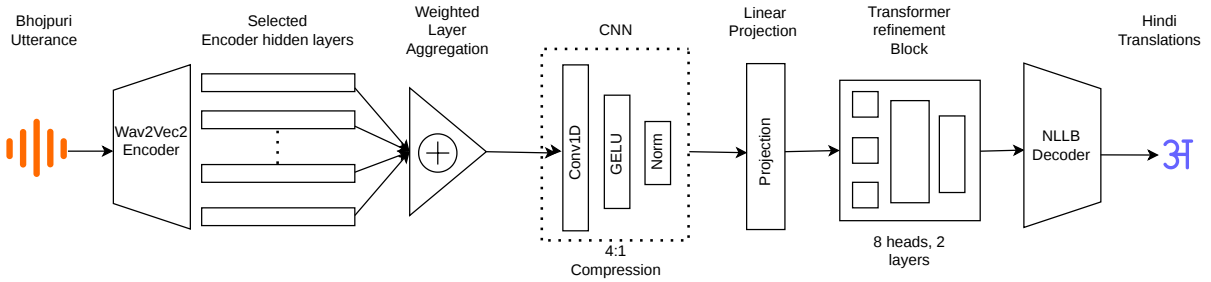

Figure 1: Overview of the utilized modular Bhojpuri-to-Hindi speech translation architecture. A pretrained Wav2Vec2 encoder first converts the input Bhojpuri utterance into frame-level hidden representations from multiple encoder layers. These representations are combined through weighted layer aggregation, compressed by a CNN-based reduction stage, and linearly projected into the hidden space of the NLLB decoder. A lightweight Transformer refinement block then contextualizes the projected speech embeddings before supplying them to the pretrained NLLB decoder as encoder-side context for cross-attention, enabling autoregressive generation of Hindi text.

into the NLLB decoder hidden space. Finally, a lightweight Transformer refinement block (Vaswani et al., 2017) contextualizes these projected speech embeddings before decoding. In the default full-adapter setting, this block uses 2 Transformer layers, 8 attention heads, a feed-forward dimension of $4\times$ the decoder hidden size, and dropout of 0.1. The CNN adapter baseline corresponds to this same pipeline without the Transformer refinement block.

4.1.3 NLLB Decoder

On the text side, we use the decoder of facebook/nllb-200-distilled-600M (Costa-Jussà et al., 2022) as the translation generator. Rather than using the pretrained NLLB text encoder,we bypass it and directly provide the speech-derived adapted states as encoder outputs to the decoder. The decoder then attends to these conditioned on the Hindi Devanagari target-language token. In this formulation, the decoder contributes multilingual generation and target-side language modeling, while the adapter bears the main burden of mapping speech representations into a form compatible with the decoder's cross-attention interface.

4.2 Cascade

4.2.1 ASR

For the ASR component, we fine-tune whisper-large-v3 (Radford et al., 2022) on the ai4bharat/RuralWomenBhojpuri dataset, which comprises both real and synthetic speech samples covering a wide range of acoustic variability. Audio is resampled to 16 kHz to match Whisper's input requirements.

To improve robustness under low-resource and noisy conditions, we apply waveform-level speed perturbation (Ko et al., 2015) and additive Gaussian noise (SNR 15–30 dB), alongside feature-level SpecAugment (Park et al., 2019) for time and frequency masking. For parameter-efficient fine-tuning, we adopt LoRA (Hu et al., 2022) on the Query, Key, Value, and Output attention projections. The model is trained for 10 epochs with a learning rate of 10^{-4}, linear warmup over 5% of the steps, a batch size of 4, and label smoothing of 0.1 (Vaswani et al., 2017). The best dev-set checkpoint is selected, and the learned LoRA parameters are merged into the base model to produce a standalone ASR system that serves as the front-end of our cascaded ASR→MT pipeline.

4.2.2 MT

For the MT component, we fine-tune facebook/nllb-200-1.3B (Costa-Jussà et al., 2022) on a combination of the additional Kaggle dataset (Das) and back-translated data generated from the official training set using the untuned facebook/nllb-200-1.3B. We use cross-entropy as the loss function with the Adafactor optimizer (Shazeer and Stern, 2018), a learning rate of 1e−4, a gradient-clipping threshold of 1.0, and a batch size of 16. The model was trained for 10 epochs, with the best BLEU scores occurring on Epoch 3.

4.2.3 QE-based Hypothesis Fusion for MT

Neural Machine Translation (NMT) models assign probabilities to translation candidates given a source sentence, and translation errors arise from mismatches between the model's predictions and the reference/golden translations. Conventional n-gram-based metrics such as BLEU (Papineni et al., 2002) and chrF++ (Popović, 2015) show only weak correlation with human assessments. Consequently,

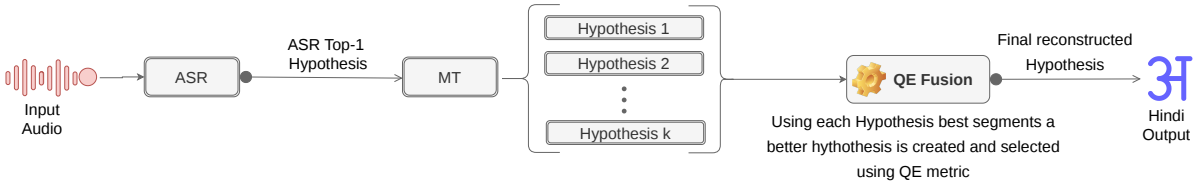

Figure 2: ASR top-1 hypothesis is passed directly to MT, where QE Fusion replaces beam search to reconstruct an alternative hypothesis from k candidates.

the field has started to turn to neural scoring methods such as COMET (Rei et al., 2020), for better human alignment.

QE Fusion (Vernikos and Popescu-Belis, 2024) uses QE COMET-Kiwi (Rei et al., 2022) scores, which predict translation quality without reference translations, to rerank and reconstruct a superior hypothesis by selecting and combining the best segments across multiple MT hypotheses. Motivated by this, we replace beam search with QE Fusion in our cascade pipeline: the MT component produces 5 hypotheses, which are then scored and reconstructed using a QE model to produce a single refined hypothesis.

5 Experiments

5.1 End-to-End Experiments

These experiments were designed to identify the most effective combination of encoder adaptation, interconnection module, and decoder adaptation for low-resource Bhojpuri-to-Hindi speech translation. Unless stated otherwise, both encoder and decoder backbones were frozen, the compression ratio of both the CNN adapter and the full adapter was fixed at 4:1, optimization used AdamW (Loshchilov and Hutter, 2017) with batch size 4 and initial learning rate 1×10^{-4}, and learning-rate scheduling followed 5% linear warmup followed by cosine decay. All models were trained on the official IWSLT 2026 Bhojpuri-Hindi data, with augmentation applied only where explicitly stated, and evaluated on the dev set using BLEU (Papineni et al., 2002) and chrF++ (Popović, 2015).

We conducted several preliminary development experiments to validate convergence and guide architectural choices. Since only two systems were officially submitted, we focus the main discussion on those submitted systems and defer the earlier exploratory experiments to the appendix.

We use the CNN adapter experiment as a baseline; this adapter is equivalent to the full adapter without the Transformer refinement block (Figure 1), with the experiment results in Table 2

5.1.1 Adapter with Decoder-side LoRA and Data Augmentation

The third experiment combined the full adapter with decoder-side LoRA and SpecAugment-based augmentation (Park et al., 2019) over the full training set. This was the best-performing configuration overall. Decoding used beam size 5, no-repeat n-gram size 3, and repetition penalty 1.2. Trained on the augmented version of the official 20-hour training set, expanded to approximately 26 hours via speed perturbation, this model achieved a peak validation BLEU of 32.77 and a chrF++ of 53.43, showing that stronger modality adaptation, lightweight decoder adaptation, and augmentation were highly complementary in the low-resource setting. This system serves as our PRIMARY submission.

5.1.2 Adapter with Decoder-side LoRA + Decoder Cross-Attention and Upper Layer Unfreezing with Data Augmentation

The final experiment tested whether broader decoder adaptation could improve upon the best parameter-efficient configuration. Starting from the augmented full-adapter + decoder-side LoRA setup, we additionally unfroze all decoder cross-attention modules and the top two decoder layers (layers 10 and 11 in 0-based indexing), while keeping the remainder of the setup unchanged. Decoding again used beam size 5, no-repeat n-gram size 3, and repetition penalty 1.2. After 40 epochs of training, this model achieved a maximum BLEU score of 30.23 and a chrF++ score of 51.56 on the validation set. Although performance remained strong, it did not surpass the fully parameter-efficient LoRA-based setup; BLEU plateaued around 30 over the final seven epochs, indicating that broader decoder unfreezing increased trainable capacity without yielding further translation gains. This suggests that,

System	BLEU ↑	chrF++ ↑
Adapter + Decoder LoRA + Aug.	**32.77**	**53.43**
Adapter + Decoder LoRA + Upper Layer Unfreeze + Aug.	**30.23**	**51.56**

Table 2: End-to-end system results on the IWSLT 2026 Bhojpuri–Hindi dev set. All models use encoder layers {6, 8, 10, 12} for aggregation and a fixed 4:1 compression ratio.

in our low-resource setting, preserving pretrained decoder priors was more beneficial than allowing broader decoder adaptation. This system serves as our CONTRASTIVE1 submission.

5.2 Cascade Experiments

For the cascade system, we primarily conducted an experiment to test its effectiveness in the low-resource Bhojpuri→Hindi speech translation setting and to investigate whether errors mainly originate in ASR or MT.

5.2.1 Experiment Setup

Experiments were conducted on the Bhojpuri→Hindi language pair using the IWSLT 2026 official dev set for model selection and the test set for final evaluation. Unless stated otherwise, the following settings were used across the experiments whenever needed: Whisper was run with a decoding temperature of 0.4, batch size of 4, and a no-repeat n-gram size of 3. Longer audio tracks were segmented into 30-second chunks with a 2-second overlap to handle utterances exceeding the Whisper context limit at inference time. For MT, we used beam search with a beam size of 4 for the baseline and a maximum source and target length of 256 tokens. For QE Fusion, the reference-free quality estimation model Unbabel/wmt22-cometkiwi-da was used with 5 candidates per sentence generated via epsilon sampling, and a reconstruction beam size of $b = 4$ (Vernikos and Popescu-Belis, 2024). All models were evaluated using BLEU, chrF++ and COMET (Rei et al., 2020) for translation quality.

5.2.2 Baseline

For the cascade baseline, the fine-tuned Whisper ASR output was passed directly to the fine-tuned NLLB MT model without any additional setup. The QE Fusion experiment was evaluated against this baseline using BLEU, chrF++, and COMET on the official IWSLT 2026 dev and test sets.

5.2.3 QE Fusion for MT

We replaced standard beam search in the MT block with QE Fusion (Section 4.2.3), generating 5 candidates per sentence via epsilon sampling ($\epsilon = 0.02$, $T = 0.5$) (Vernikos and Popescu-Belis, 2024), which are then scored and reconstructed by Unbabel/wmt22-cometkiwi-da with a reconstruction beam size of $b = 4$ to produce a single refined hypothesis.

As shown in Table 3, QE Fusion yields consistent gains over the beam search baseline across all metrics (BLEU: 15.96→16.09, chrF++: 41.23→41.37, COMET: 0.5394→0.5569), confirming that QE-guided hypothesis reconstruction improves translation quality even in low-resource settings. Table 7 shows a representative example of the 5 MT candidates and the final reconstructed hypothesis. This system serves as our CONTRASTIVE2.

6 Results

Table 4 presents the official IWSLT 2026 Bhojpuri→Hindi test-set results for our submitted systems (Adelani et al., 2026). The **PRIMARY** submission corresponds to the adapter-based end-to-end system with decoder-side LoRA and data augmentation. **CONTRASTIVE1** extends this configuration with decoder cross-attention adaptation and upper-layer decoder unfreezing alongside decoder-side LoRA. **CONTRASTIVE2** corresponds to the cascaded ASR → MT pipeline incorporating QE Fusion reranking at the MT stage (Section 5). We report BLEU and chrF++ scores for each submission and discuss end-to-end and cascade performance separately in the subsections below.

6.1 End to End Results

The **CONTRASTIVE1** submission corresponds to the adapter-based end to end speech translation systems with decoder-side LoRA, decoder cross-attention unfreezing, and upper-layer decoder unfreezing. On the official IWSLT 2026 Bhojpuri → Hindi test set, the system achieved a BLEU score of 12.3 and a chrf++ score of 35.

On the development set, the PRIMARY system achieved the strongest performance overall with a BLEU of 32.77 and chrF++ of 53.43 (Ta-

Pipeline	BLEU	chrF++	COMET
Baseline			
ASR → MT	15.96	41.23	0.5394
+ QE at MT			
ASR → MT (QE Fusion)	**16.09**	**41.37**	**0.5569**

Table 3: Cascade results on dev set. Applying QE Fusion at the MT stage yields consistent improvements.

Pipeline	BLEU	ChrF++
Contrastive1	**12.3**	35
Primary	12.1	34
Contrastive2	10.1	39

Table 4: Our submissions with BLEU and ChrF++ scores.

ble 2), while CONTRASTIVE1 reached 30.23 BLEU and 51.56 chrF++, suggesting that broader decoder unfreezing did not improve upon the parameter-efficient LoRA-only configuration. On the official test set, however, CONTRASTIVE1 obtained the highest BLEU score among all submissions, marginally outperforming the PRIMARY submission (12.1 BLEU, 34 chrF++) and the CONTRASTIVE2 cascaded system (10.1 BLEU, 39 chrF++). The substantial drop from development to test set ($\Delta{=}-20.47$ for PRIMARY) suggests overfitting to the news-domain training distribution, which we discuss further in Section 7.

6.2 Cascade Results

The CONTRASTIVE2 system corresponds to the cascade configuration reported in Table 4. While the cascade system underperforms relative to the end-to-end model on the test set, the BLEU degradation from development to test is considerably smaller ($\Delta{=}-5.99$ vs. $\Delta{=}-20.47$), indicating greater generalization robustness. We attribute this drop partially to translation repetitions introduced by increasing the ASR temperature from 0.2 to 0.4; while this yielded more fluent outputs overall, it caused occasional repetitions in a subset of sentences, contributing to the score decrease. We nonetheless opted for $T{=}0.4$, as the improvement in fluency and translation adequacy outweighed the repetition-free but less fluent outputs of $T{=}0.2$. We discuss the generalization gap between cascade and end-to-end systems further in Section 7.

7 Discussion

Post-hoc analysis indicates that the substantial development-to-test performance gap in our end-to-end (E2E-ST) model stems from a compounding mix of dataset artifacts and structural vulnerabilities. Manual inspection of the official dataset revealed severe label–audio mismatches, where the provided text completely diverges from the actual spoken content (translated to English in Table 6 for accessibility). This pervasive annotation noise disrupted training alignments and artificially inflated our development baseline (32.77 BLEU) relative to the test set (12.3 BLEU). Furthermore, frequent trailing hallucinations in the E2E-ST test predictions suggest that the model learned spurious, non-causal distributions from these corrupted pairs. If comparable label noise pervades the unreleased test references, the evaluation metrics may partially reflect reference degradation rather than a true drop in translation quality.

However, dataset corruption alone does not fully explain the performance drop; rather, it exacerbates existing structural sensitivities. Specifically, the interconnection adapter was trained entirely from scratch on the limited 26-hour training set. Because it lacks any pretraining signal, it is uniquely vulnerable to any kind of domain shifts at test time. While the pretrained encoder and decoder generalize robustly across domains due to their large-scale pretraining, the adapter—functioning as the sole learned cross-modal bridge—bears the full burden of any distributional mismatch. This structural bottleneck highlights adapter pretraining on auxiliary speech-text data as the most direct avenue for closing the performance gap.

Table 5: Representative label–audio mismatches identified during post-hoc corpus inspection of the development and training sets. *Correct* denotes the content verified in the source audio, while *Provided* denotes the corresponding official dataset label. See appendix for translations.

Split	Sample	Correct (Audio-Verified)	Provided (Dataset Label)
Dev	203	आज नई दिल्ली में राष्ट्रीय मानवाधिकार आयोग द्वारा आयोजित अंतरराष्ट्रीय मानवाधिकार सम्मेलन में बोलते हुए श्री नायडू ने कहा कि राष्ट्रीय मानवाधिकार आयोग	उन्होंने कहा कि हर व्यक्ति को छत उपलब्ध कराने के लिए ग्रामीण इलाको में 11 लाख और शहरी इलाको में 6 लाख घर बनाये जा रहे हैं, जो अपने आप में एक रिकॉर्ड है.
Dev	483	उन्होंने कहा कि बस्ती में रिंग रोड और जल मार्ग है।	मुख्यमंत्री ने कहा कि प्रदेश में गन्ना किसानों को 37 हजार करोड़ रूपये का भुगतान किया जा चुका है
Train	8305	और तेरह सौ रुपये का टिकट खरीदना पड़ा। मऊ जिले के हलधरपुर थाना क्षेत्र में पहसा बाजार के पास कल रात हुई सड़क दुर्घटना में चार युवको की मौत हो गई।	उन्होंने उद्यानिकी प्रसंस्करण क्षेत्र में निवेशको को अधिक सुविधाएं उपलब्ध कराने के लिए वर्तमान प्रसंस्करण नीति में संशोधन कर आवश्यक कार्यवाही करने के निर्देश दिये

In contrast, our cascaded approach demonstrated significantly greater resilience to these discrepancies, exhibiting a much narrower development-to-test drop. Because the cascade pipeline was primarily trained on out-of-domain data, it was fundamentally less reliant on the official dataset (check table 1 for Dataset summary). To integrate the official data, we generated pseudo-labels by running our ASR on the official audio and paired those outputs with the provided targets, augmenting our broader MT training corpus. Consequently, the cascade's primary bottleneck was standard error propagation from the ASR to the MT module, rather than catastrophic alignment failures. While the noisy official targets still introduced some downstream errors, this decoupled architecture effectively insulated the cascade from the severe degradation observed in the E2E-ST baseline.

For transparency, similar labeling artifacts were identified in several other entries, including Samples 402 and 403 in the development set, as well as Samples 805, 922, 7988, 8303, 8304, 8306, and 8307–8309 in the training set.

Acknowledgments

Kumar Rishu acknowledges that part of the experiments were conducted using computational resources provided by the e-INFRA CZ project (ID:90254), supported by the Ministry of Education, Youth and Sports of the Czech Republic.

References

David Ifeoluwa Adelani, Victor Agostinelli, Antonios Anastasopoulos, Luisa Bentivogli, Ondřej Bojar, Sebastien Bratières, Marine Carpuat, Roldano Cattoni, Mauro Cettolo, Lizhong Chen, Marcello Federico, Marco Gaido, Mahendra Gupta, HyoJung Han, Ali Hatami, David Javorský, Yejin Jeon, Marek Kasztelnik, Antoine Laurent, and 33 others. 2026. Speech translation and metrics in 2026: Findings of the iwslt campaign. In *Proceedings of the 23rd International Conference on Spoken Language Translation (IWSLT 2026)*, San Diego, California, US. Association for Computational Linguistics.

Marko Avila and Josep Crego. 2025. SYSTRAN @ IWSLT 2025 low-resource track. In *Proceedings of the 22nd International Conference on Spoken Language Translation (IWSLT 2025)*, pages 324–332, Vienna, Austria (in-person and online). Association for Computational Linguistics.

Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. In *Advances in Neural Information Processing Systems*.

Harveen Singh Chadha, Anirudh Gupta, Priyanshi Shah, Neeraj Chhimwal, Ankur Dhuriya, Rishabh Gaur, and Vivek Raghavan. 2022. Vakyansh: Asr toolkit for low resource indic languages. *Preprint*, arXiv:2203.16512.

Marta R Costa-Jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, and 1 others. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.

Rishi Das. Hindi to bhojpuri. https://www.kaggle.com/datasets/rishi2003das/hindi-to-bhojpuri. Kaggle dataset.

Sandareka Fernando, Surangika Ranathunga, Sanath Jayasena, and Gihan Dias. 2016. Comprehensive part-of-speech tag set and SVM based POS tagger for Sinhala. In *Proceedings of the 6th Workshop on South and Southeast Asian Natural Language Processing (WSSANLP2016)*, pages 173–182, Osaka, Japan. The COLING 2016 Organizing Committee.

Markus Freitag, David Grangier, Qijun Tan, and Bowen Liang. 2022. High quality rather than high model

probability: Minimum Bayes risk decoding with neural metrics. *Transactions of the Association for Computational Linguistics*, 10:811–825.

Gerard I. Gállego, Ioannis Tsiamas, Carlos Escolano, José A. R. Fonollosa, and Marta R. Costa-jussà. 2021. End-to-end speech translation with pre-trained models and adapters: UPC at IWSLT 2021. In *Proceedings of the 18th International Conference on Spoken Language Translation (IWSLT 2021)*, pages 110–119, Bangkok, Thailand (online). Association for Computational Linguistics.

Dan Hendrycks and Kevin Gimpel. 2016. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Liang Wang, Weizhu Chen, and 1 others. 2022. Lora: Low-rank adaptation of large language models. *Iclr*, 1(2):3.

Pratik Joshi, Christain Barnes, Sebastin Santy, Simran Khanuja, Sanket Shah, Anirudh Srinivasan, Satwik Bhattamishra, Sunayana Sitaram, Monojit Choudhury, and Kalika Bali. 2019. Unsung challenges of building and deploying language technologies for low resource language communities. In *Proceedings of the 16th International Conference on Natural Language Processing*, pages 211–219, International Institute of Information Technology, Hyderabad, India. NLP Association of India.

Sakshi Joshi, Eldho Ittan George, Tahir Javed, Kaushal Bhogale, Nikhil Narasimhan, and Mitesh M. Khapra. 2025. Recognizing every voice: Towards inclusive asr for rural bhojpuri women. *Preprint*, arXiv:2506.09653.

Tom Ko, Vijayaditya Peddinti, Daniel Povey, and Sanjeev Khudanpur. 2015. Audio augmentation for speech recognition. In *INTERSPEECH*, pages 3586–3590.

Ritesh Kumar, Siddharth Singh, Shyam Ratan, Mohit Raj, Sonal Sinha, Bornini Lahiri, Vivek Seshadri, Kalika Bali, and Atul Kr Ojha. 2022. Annotated speech corpus for low resource indian languages: Awadhi, bhojpuri, braj and magahi. *arXiv preprint arXiv:2206.12931*.

Shankar Kumar and William Byrne. 2004. Minimum Bayes-risk decoding for statistical machine translation. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics: HLT-NAACL 2004*, pages 169–176, Boston, Massachusetts, USA.

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Rajesh Kumar Mundotiya, Manish Kumar Singh, Rahul Kapur, Swasti Mishra, and Anil Kumar Singh. 2021. Linguistic resources for bhojpuri, magahi, and maithili: Statistics about them, their similarity estimates, and baselines for three applications. *Transactions on Asian and Low-Resource Language Information Processing*, 20(6):1–37.

Jan Niehues, Elizabeth Salesky, Marco Turchi, and Matteo Negri. 2021. Tutorial: End-to-end speech translation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Tutorial Abstracts*, pages 10–13, online. Association for Computational Linguistics.

Yuta Nishikawa and Satoshi Nakamura. 2023. Interconnection: Effective connection between pre-trained encoder and decoder for speech translation. *arXiv preprint arXiv:2305.16897*.

Sara Papi, Javier Garcia Gilabert, Zachary Hopton, Vilém Zouhar, Carlos Escolano, Gerard I Gállego, Jorge Iranzo-Sánchez, Ahrii Kim, Dominik Macháček, Patricia Schmidtova, and 1 others. 2025. Hearing to translate: The effectiveness of speech modality integration into llms. *arXiv preprint arXiv:2512.16378*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA.

Daniel S Park, William Chan, Yu Zhang, Chung-Cheng Chiu, Barret Zoph, Ekin D Cubuk, and Quoc V Le. 2019. Specaugment: A simple data augmentation method for automatic speech recognition. *arXiv preprint arXiv:1904.08779*.

Ankita Pasad, Chung-Ming Chien, Shane Settle, and Karen Livescu. 2024. What do self-supervised speech models know about words? *Transactions of the Association for Computational Linguistics*, 12:372–391.

Ngoc-Quan Pham, Tuan Nam Nguyen, Thai-Binh Nguyen, Danni Liu, Carlos Mullov, Jan Niehues, and Alexander Waibel. 2022. Effective combination of pretrained models - KIT@IWSLT2022. In *Proceedings of the 19th International Conference on Spoken Language Translation (IWSLT 2022)*, pages 190–197, Dublin, Ireland (in-person and online). Association for Computational Linguistics.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. Robust speech recognition via large-scale weak supervision. *arXiv preprint*.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on*

Empirical Methods in Natural Language Processing (EMNLP), pages 2685–2702, Online. Association for Computational Linguistics.

Ricardo Rei, Marcos Treviso, Nuno M. Guerreiro, Chrysoula Zerva, Ana C Farinha, Christine Maroti, José G. C. de Souza, Taisiya Glushkova, Duarte Alves, Luisa Coheur, Alon Lavie, and André F. T. Martins. 2022. CometKiwi: IST-unbabel 2022 submission for the quality estimation shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 634–645, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Nicholas Ruiz and Marcello Federico. 2014. Assessing the impact of speech recognition errors on machine translation quality. In *Proceedings of the 11th Conference of the Association for Machine Translation in the Americas: MT Researchers Track*, pages 261–274, Vancouver, Canada. Association for Machine Translation in the Americas.

Noam Shazeer and Mitchell Stern. 2018. Adafactor: Adaptive learning rates with sublinear memory cost. In *International conference on machine learning*, pages 4596–4604. PMLR.

Matthias Sperber and Matthias Paulik. 2020. Speech translation and the end-to-end promise: Taking stock of where we are. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7409–7421, Online. Association for Computational Linguistics.

Ioannis Tsiamas, Gerard I. Gállego, Carlos Escolano, José A. R. Fonollosa, and Marta R. Costa-jussà. 2022. Pretrained speech encoders and efficient fine-tuning methods for speech translation: UPC at IWSLT 2022. In *Proceedings of the 19th International Conference on Spoken Language Translation (IWSLT 2022)*, pages 265–276, Dublin, Ireland (in-person and online). Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Giorgos Vernikos and Andrei Popescu-Belis. 2024. Don't rank, combine! combining machine translation hypotheses using quality estimation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12087–12105, Bangkok, Thailand. Association for Computational Linguistics.

Ron J Weiss, Jan Chorowski, Navdeep Jaitly, Yonghui Wu, and Zhifeng Chen. 2017. Sequence-to-sequence models can directly translate foreign speech. *arXiv preprint arXiv:1703.08581*.

Jinming Zhao, Hao Yang, Ehsan Shareghi, and Gholamreza Haffari. 2022. M-adapter: Modality adaptation for end-to-end speech-to-text translation. *arXiv preprint arXiv:2207.00952*.

A Appendix

We define our additional experiments in the following subsections.

A.1 Preliminary End-to-End Experiments

We conducted two preliminary development experiments to validate convergence and to motivate the final submitted systems.

A.1.1 CNN Adapter with Decoder-side LoRA Baseline

The first experiment was a proof-of-concept run to test whether the proposed encoder-adapter-decoder coupling could converge in the low-resource setting. We used the CNN adapter as the interconnection module and applied decoder-side LoRA for lightweight adaptation, without augmentation. Trained for 20 epochs on approximately 26 hours of data, this system achieved a maximum validation BLEU of 22.01 and a chrF++ of 37.00, establishing the architecture's basic viability.

A.1.2 Adapter with Decoder-side LoRA without Data Augmentation

The second experiment added a Transformer-based refinement block after the projection stage in order to test whether a more expressive modality-alignment module would improve performance. Decoder-side LoRA was again used for lightweight decoder adaptation, and no augmentation was applied. Because the full adapter introduces additional Transformer capacity, training was extended to 40 epochs. This configuration achieved a maximum BLEU score of 26.14 and a chrF++ score of 42.11, indicating a clear gain over the simpler CNN-based baseline.

A.2 Preliminary Cascade Experiments

A.2.1 LLM as a Post Correction block

We evaluate the LLM-based post-correction block by inserting it between the ASR and MT components. The top-1 ASR hypothesis is passed to the LoRA-fine-tuned ai4bharat/Airavata LLM, which generates a corrected Bhojpuri transcription, which is then fed to the MT model. We evaluate end-to-end using BLEU, chrF++, and COMET against the IWSLT Hindi references. Introducing the correction block degrades performance significantly (BLEU: 12.03, chrF++: 34.39, COMET: 0.4966). We attribute this to limited Bhojpuri coverage in ai4bharat/Airavata's pretraining corpus: despite

Table 6: Translated reference for Table 5

Split	Sample	Correct (Audio-Verified)	Provided (Dataset Label)
Dev	203	*Speaking at the International Human Rights Conference organized by the National Human Rights Commission in New Delhi, Mr. Naidu said that the National Human Rights Commission*	*He said that 1.1 million houses in rural areas and 0.6 million houses in urban areas are being built to provide housing for every person, which is a record in itself.*
Dev	483	*He said that there is a ring road and a waterway in Basti.*	*The Chief Minister said that payments amounting to 37,000 crore have already been made to sugarcane farmers in the state.*
Train	8305	*And a ticket costing 1,300 had to be purchased. In a road accident that occurred last night near Pahsa Bazaar in the Haldharpur police station area of Mau district, four young men lost their lives.*	*He directed that necessary action be taken to amend the existing processing policy in order to provide greater facilities to investors in the horticultural processing sector.*

LoRA fine-tuning on 3,200 utterance pairs constructed from Whisper ASR hypotheses paired with the IWSLT 2026 Bhojpuri references, the model frequently hallucinates tokens, switches language mid-sentence, or alters semantic content, compounding the error signal passed to MT. Applying QE Fusion on the LLM hypothesis partially mitigates this degradation (BLEU: 13.02, chrF++: 36.61, COMET: 0.5152), but does not surpass the baseline, confirming that QE-based selection cannot compensate for the LLM's fundamental coverage gap on Bhojpuri.

Variant	Text Alignment
MT Source	सरदार पटेल के जनती के मौका पर विधानसभा के सामने मार्च पास्ट और रन पॉर यूनिटी कारकरम का आयोजन केल गएल। *Sardar Patel of jayanti [typo] of moment on Legislative Assembly of front march past and Run Por [typo] Unity program [typo] of organization done went.*
Candidate 1	सरदार पटेल की जयंती के मौके पर विधानसभा के सामने मार्चपास्ट और रन पोर [स्पेलिंग गलती] यूनिटी कारकम [स्पेलिंग गलती] का आयोजन किया गया *Sardar Patel of anniversary **of moment on** Legislative Assembly of front **marchpast** and Run **Por [typo]** Unity **program [typo]** of organization done went*
Candidate 2	सरदार पटेल की जयंती पर विधानसभा के सामने मार्च पास्ट और रन पोर [स्पेलिंग गलती] यूनिटी कारकरम [स्पेलिंग गलती] का आयोजन किया गया *Sardar Patel of anniversary **on** Legislative Assembly of front **march past** and Run **Por [typo]** Unity **program [typo]** of organization done went*
Candidate 3	सरदार पटेल की जयंती के अवसर पर विधानसभा के सामने मार्च पास्ट और रन पोर [स्पेलिंग गलती] यूनिटी कारकम [स्पेलिंग गलती] का आयोजन किया गया *Sardar Patel of anniversary **of occasion on** Legislative Assembly of front **march past** and Run **Por [typo]** Unity **program [typo]** of organization done went*
Candidate 4	सरदार पटेल की जयंती के मौके पर विधानसभा के सामने मार्च पास्ट और रन पोर [स्पेलिंग गलती] यूनिटी कारकर्म [स्पेलिंग गलती] का आयोजन किया गया *Sardar Patel of anniversary **of moment on** Legislative Assembly of front **march past** and Run **Por [typo]** Unity **program [typo]** of organization done went*
Candidate 5	सरदार पटेल की जयंती के मौके पर विधानसभा के सामने मार्चपास्ट और रन पॉर [स्पेलिंग गलती] यूनिटी कारकम [स्पेलिंग गलती] का आयोजन किया गया *Sardar Patel of anniversary **of moment on** Legislative Assembly of front **marchpast** and Run **Por [typo]** Unity **program [typo]** of organization done went*
Selected	सरदार पटेल की जयंती के अवसर पर विधानसभा के सामने मार्च पास्ट और रन पॉर यूनिटी कारकम का आयोजन किया गया ***Sardar Patel of anniversary of occasion on Legislative Assembly of front march past and Run Por [typo] Unity program [typo] of organization done went***
Reference	सरदार पटेल की जयंती के अवसर पर विधानसभा के सामने मार्चपास्ट और रन फॉर यूनिटी कार्यक्रम का आयोजन किया गया *Sardar Patel of anniversary of occasion on Legislative Assembly of front marchpast and Run For Unity program of organization done went*

Table 7: Example of ASR/MT candidate diversity highlighting phonetic variations in English loan words (पॉर/पोर for "For", कारकम/कारकरम for "कार्यक्रम"/program) and synonymous phrasing (के मौके पर vs के अवसर पर).