Speech-to-Speech Machine Translation for Dialectal Variations of Hindi

Sanmay Sood, Siddharth Rajput, Md. Shad Akhtar
IIIT Delhi, India
{sanmay21095, siddhart21102, shad.akhtar}@iiitd.ac.in

Abstract

Hindi has many dialects, and they are vital to India's cultural and linguistic heritage. However, many of them have been largely overlooked in modern language technological advancements, primarily due to a lack of proper resources. In this study, we explore speech-to-speech machine translation (S2ST) for four Hindi dialects, i.e., *Awadhi*, *Bhojpuri*, *Braj Bhasha*, and *Magahi*. We adopt a cascaded S2ST pipeline comprising of three stages: Automatic Speech Recognition (ASR), Machine Translation (MT), and Text-to-Speech (TTS). We evaluate many recent large language models (LLMs) for dialect-to-Hindi and dialect-to-English translations in zero-shot, few-shot, and chain-of-thought setups. Our comparative analysis offers insights into the current capabilities and limitations of LLM-based approaches for low-resource dialectal S2ST in Indian context. Dataset and code are available at https://github.com/flamenlp/S2ST-Dialect.

1 Introduction

The "Hindi Belt" or northern-central region of India includes various dialects such as Awadhi, Bhojpuri, Braj Bhasha, Magahi, Bundeli, etc., each holding significant cultural value but largely neglected in contemporary language technologies. This neglect is mainly due to the dominance of Modern Standard Hindi (MSH) following its institutionalization, which has marginalized these dialects and put their linguistic diversity at risk. Although computational tools and resources for MSH have advanced considerably, equivalent support for its dialects remains lacking. The scarcity of text and speech datasets hinders the development of NLP and speech technologies tailored to these dialects. Modern NLP systems prioritize high-resource languages, leaving Hindi-belt dialects underserved. This highlights the urgent need for targeted research on these dialects.

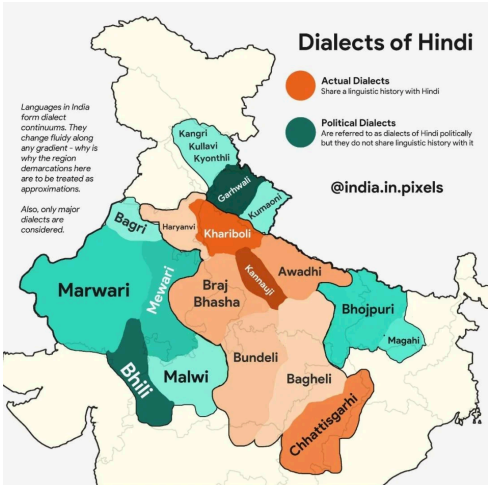

Figure 1: Dialects in the Hindi Belt. Source: https://www.instagram.com/@indiainpixels

Speech-to-Speech Machine Translation (S2ST) offers a transformative solution to bridge the linguistic divide. By automating speech translation, S2ST enables real-time access to essential services in education, healthcare, and governance, particularly in regions where local dialects are primary. Furthermore, S2ST can help preserve linguistic diversity by allowing speakers to use their native dialects in the digital world. Over the years, S2ST systems have evolved considerably, with cascaded architectures emerging as the predominant approach due to their proven effectiveness for low-resource languages. These systems decompose the translation process into distinct components—automatic speech recognition (ASR) (Kumar and Akhtar, 2025; Javed et al., 2025), machine translation (MT) (Gala et al., 2023; Kartik et al., 2024), and text-to-speech (TTS) (V et al., 2025)—enabling independent optimization of each module.

Mhaskar et al. (2023) introduced VAKTA-SETU, a speech-to-speech machine translation service that integrates Vakyansh Wav2Vec2 ASR (Gupta et al., 2021; Chadha et al., 2022), Indic-

Proceedings of the Twelfth Workshop on Asian Translation (WAT 2025), pages 54–65

Trans2 (Gala et al., 2023), and Tacotron 2 TTS (Shen et al., 2017) to support language pairs including English-Hindi, English-Marathi, and Hindi-Marathi. Complementing this effort, the IWSLT 2024 Indic Track (Sethiya et al., 2024) demonstrated that a Whisper (Radford et al., 2022) → IndicTrans2 cascade consistently outperformed end-to-end models on low-resource languages such as Bengali, Tamil, etc. This finding reaffirms the robustness and effectiveness of modular systems in resource-scarce settings (Dabre and Song, 2024).

Recent studies have extensively explored prompting strategies for machine translation using large language models. Vilar et al. (2023) demonstrated that the quality of few-shot examples is the most critical factor for effective prompting, highlighting careful example selection over semantic proximity. Zhang et al. (2023) conducted a systematic study analyzing various prompt templates and showed that both the template wording and the number of shots significantly affect translation quality, with suboptimal examples leading to degraded performance. Hendy et al. (2023) further evaluated prompting effects across diverse GPT models, confirming that optimal shot numbers and example relevance markedly influence model outputs, especially in low-resource settings. Collectively, these works emphasize the importance of designing suitable prompt templates, determining an effective number of few-shot demonstrations, and selecting relevant examples to enhance MT with LLMs.

Adapting general-purpose LLMs to dialectal machine translation presents distinct challenges. Court and Elsner (2024) showed that retrieval-augmented generation can aid smaller models for Southern Quechua-Spanish translation, while zero-shot prompting remains the most effective approach for state-of-the-art LLMs. However, these advanced models still frequently produce mistranslations and raise ethical concerns, especially when errors go unnoticed. Similarly, Almaoui et al. (2025) examined Arabizi and Arabic dialects, revealing significant performance disparities: Egyptian Arabic benefits from considerable media exposure, whereas Algerian Arabic struggles due to heavy code-switching and limited training data. These findings highlight the complexities involved in translating non-standardized dialectal varieties using general-purpose LLMs.

Building on the need for dedicated research, this study introduces a cascaded S2ST pipeline with a primary focus on the machine translation stage. We present a detailed exploration of LLMs for dialect-to-Hindi and dialect-to-English translation, investigating the performance of different prompt templates, including zero-shot, few-shot, and chain-of-thought (CoT) prompting.

2 Dataset

The development of effective S2ST systems for low-resource languages requires carefully curated datasets that address the challenge of resource scarcity. For Hindi dialects including Awadhi, Bhojpuri, Braj Bhasha and Magahi, the availability of high-quality parallel speech data remains severely limited, necessitating a multi-faceted approach to combine parallel speech corpora, monolingual audio resources, and text-based datasets.

Our research leverages the SpeeD-IA dataset from KMI Linguistics (Kumar et al., 2022), which is one of the few available parallel speech resources for Hindi dialects. The corpus originally consisted of 369 Hindi sentences that were translated into Awadhi, Bhojpuri, Braj Bhasha, and Magahi through spoken renditions by native speakers. These audio translations were then transcribed using ASR systems to generate corresponding text transcriptions. We pruned this set—removing duplicates and poorly formed sentences—and produced a clean collection of 267 parallel sentences available across every dialect, Hindi, and English. For each sentence, we select the best translation from multiple transcriptions and further refined these transcriptions using a multilingual LLM to ensure quality and accuracy.

In addition, the dataset also included monolingual audio from every speaker recorded through 39 carefully designed questions on lifecycle events (birth, marriage, and death), yielding spontaneous narrative recordings. This resulted in roughly 2-3 hours of audio data for each dialect, totaling around 10 hours. This data was subsequently used to fine-tune ASR models, thereby enhancing their performance on natural dialectal speech. We utilize the VAANI dataset (Team, 2025), a collaborative initiative by the IISc, Bangalore and ARTPARK. We sampled ~ 4 - 5 hours of audio for each of our target dialects —Awadhi, Bhojpuri, Braj Bhasha, and Magahi— resulting in a total of 18 hours of monolingual data. We employ it for fine-tuning our ASR components.

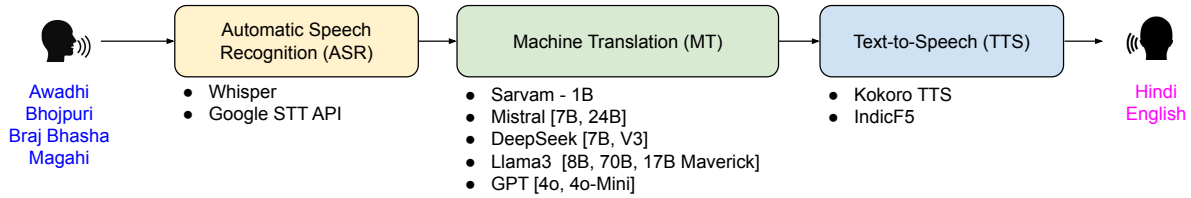

Figure 2: Cascaded pipeline for speech-to-speech Machine Translation.

3 Methodology

Our approach employs a cascaded architecture comprising three main components: ASR, MT, and TTS. Figure 2 depicts the cascaded pipeline along with models we experiment with in this paper. We now provide the details of each phase in the subsequent subsections.

3.1 Machine Translation (MT)

For machine translation, we explore a diverse set of LLMs with varying scales, architectures, and specialization to assess their performance across resource levels and reasoning capabilities. We employ following set of models in our experiments:

- **Lightweight models:**
  - `Meta-Llama-3-8B`
  - `Mistral-7B-v0.1`
  - `deepseek-llm-7b-chat`
- **Larger, more powerful variants:**
  - `Meta-Llama-3-70B-Instruct`
  - `Mistral-Small-24B-Instruct-2501`
- **Large reasoning models:**
  - `gpt-4o`
  - `DeepSeek-V3`
  - `Llama-4-Maverick-17B-128E-Instruct`
- **Indic-language specific model:**
  - `sarvamai/sarvam-1`

The inclusion of large reasoning models was motivated by their advanced multi-step inference and language understanding capabilities, which could potentially compensate for the lack of training data in low-resource dialects by better capturing contextual and semantic nuances. Moreover, we evaluate the following prompting strategies:

- **Zero-shot:** The model received the dialect input with a general instruction for the translation.
- **Few-shot:** The prompt included two translation pairs before presenting the target input.
- **CoT:** The prompt guided the model to explain or interpret the input dialectal sentence before generating the translation. An example of Bhojpuri CoT prompt given to the LLM is as follows:

Bhojpuri Prompt = You are a Bhojpuri language expert translating Bhojpuri sentences into fluent English. Follow a logical, step-by-step process to break down each sentence: identify names, pronouns, verbs, objects, and sentence structure before generating the final English translation.

 # Few-shot Examples
Example 1:

1. **Bhojpuri:** राधा रमेश के संगे शहर गईली।
   Step-by-step reasoning:
   - Step 1: राधा is a proper noun, "Radha".
   - Step 2: रमेश के संगे means "with Ramesh".
   - Step 3: शहर means "city".
   - Step 4: गईली is past tense of 'to go' – "went".

   Final Translation: Radha went to the city with Ramesh.
2. **Bhojpuri:** पतई फेड़ से नीचे गिरS ता।
   Step-by-step reasoning:
   - Step 1: पतई means "leaf".
   - Step 2: फेड़ से means "from the tree".
   - Step 3: नीचे गिरS ता means "falls down".

   Final Translation: The leaf falls down from the tree.

Now Translate:
Bhojpuri: {{INPUT}}
Step-by-step reasoning:
Step 1:
Step 2:
...:
Final Translation:

3.2 Automatic Speech Recognition (ASR)

Given that the primary focus of this study is on the MT stage, and to manage computational costs, we select a single, powerful multilingual ASR model for our pipeline: OpenAI's Whisper-medium (Radford et al., 2022). We employ Whisper due to its state-of-the-art performance across a wide range of languages and dialects, making it a highly capable and suitable candidate.

To adapt the Whisper model to the phonetic and prosodic characteristics of the Hindi Belt dialects, we employ a unified multilingual fine-tuning strategy. This approach, rather than training dialect-specific models, leverages cross-dialectal phonetic similarities and morphological patterns to improve generalization and robustness across the target varieties. In addition, we also utilize Google's Speech-to-Text API as a zero-shot baseline to assess ASR performance on dialectal speech without domain adaptation.

LLM	Awadhi			Braj			Magahi			Bhojpuri		
	BLEU	chrF	BERTScore	BLEU	chrF	BERTScore	BLEU	chrF	BERTScore	BLEU	chrF	BERTScore
Sarvam - 1B	12.07	35.23	93.18	6.45	33.24	93.95	14.00	37.33	93.38	8.44	31.08	92.92
Mistral - 7B	3.46	30.96	93.01	7.73	34.93	94.31	8.75	41.03	94.32	1.30	21.93	91.69
DeepSeek - 7B	3.57	30.10	93.31	13.95	36.58	94.23	7.51	33.49	93.35	4.86	28.97	92.65
Llama3 - 8B	6.19	37.94	94.07	11.97	42.50	94.80	12.19	43.47	94.61	4.18	30.01	93.03
Mistral 24B	14.26	41.19	94.27	25.99	49.29	94.98	8.56	31.72	92.65	16.00	41.96	94.65
Llama3 - 70B - instruct	16.91	45.52	94.99	26.58	56.01	96.17	32.01	55.29	96.03	9.25	42.83	94.75
GPT - 4o Mini	26.51	50.24	95.35	26.79	52.62	95.81	21.16	46.67	95.20	30.33	54.66	96.17
GPT - 4o	**29.74**	53.28	95.93	**37.22**	**57.46**	**96.57**	**37.63**	57.64	**96.71**	**38.28**	58.35	96.24
Llama 17B Maverick	26.79	**54.16**	95.76	25.09	56.44	95.93	30.31	**58.63**	95.82	20.77	49.14	95.38
DeepSeek v3	24.22	52.50	**96.03**	23.40	55.81	96.13	23.63	49.39	95.64	36.76	**59.99**	**96.84**

Table 1: **Dialect to Hindi:** Zero-shot results.

LLM	Awadhi			Braj			Magahi			Bhojpuri		
	BLEU	chrF	BERTScore	BLEU	chrF	BERTScore	BLEU	chrF	BERTScore	BLEU	chrF	BERTScore
Sarvam - 1B	10.73	33.40	92.98	10.43	28.85	92.90	20.61	38.08	94.07	14.99	34.75	93.06
Mistral - 7B	20.27	40.44	94.51	30.54	47.66	94.88	24.77	47.80	94.68	11.37	34.17	92.91
DeepSeek - 7B	5.55	25.53	91.25	11.42	34.06	90.65	8.43	29.43	92.06	4.57	26.45	91.33
Llama3 - 8B	27.55	47.97	95.42	26.12	42.50	93.63	31.42	51.06	95.31	21.00	41.19	94.18
Mistral 24B	19.51	46.52	94.81	10.64	33.97	93.00	28.75	53.42	94.68	17.10	40.94	93.71
Llama3 - 70B - instruct	30.47	52.47	96.11	25.10	52.03	94.86	33.77	58.05	96.54	32.25	52.35	95.75
GPT - 4o Mini	29.92	53.60	96.01	41.37	59.09	96.19	42.53	65.10	96.22	43.54	62.00	96.19
GPT - 4o	32.72	55.90	96.11	42.77	60.07	96.68	**47.65**	**67.64**	**97.41**	**47.52**	**67.64**	**97.41**
Llama 17B Maverick	35.58	57.26	96.27	**43.19**	**62.93**	**96.54**	42.17	62.33	96.44	36.60	63.94	96.33
DeepSeek v3	**39.94**	**61.53**	**97.09**	37.24	58.03	95.75	43.78	63.29	97.17	45.52	62.74	96.69

Table 2: **Dialect to English:** Zero-shot results.

3.3 Text-To-Speech (TTS)

For the Text-to-Speech (TTS) component, we select models that offer a strong balance of performance and linguistic coverage for both English and Hindi. For English, we adopt KOKORO-TTS, a high-quality neural TTS model recognized for its naturalness and intelligibility. KOKORO-TTS provides superior prosody and voice clarity, making it a reliable choice for the downstream application in our cascaded S2ST pipeline. For Hindi, we utilize the IndicF5 (V et al., 2025) model developed by AI4Bharat[1], a widely used model for Indian languages that demonstrates strong performance on native phonetic structures. These selections ensure that the final synthesized output in both languages maintained high fidelity and are intelligible to native speakers, thereby enhancing the overall usability of the system.

4 Experimental Results and Analyses

We now present a detailed analysis of the results from each phase of the study.

4.1 Machine Translation (MT) Results

To ensure focused evaluation, we filter a representative test set from our original dataset of 267 parallel sentences across the four regional Hindi dialects. Results of Dialect→Hindi and Dialect→English are listed in Tables 1 & 2 (zero-shot), Tables 3 & 4 (few-shot), and Tables 5 & 6 (CoT prompting), respectively.

Effect of Prompting Techniques: Prompting strategies show significant effect on translation quality across all dialects and models. As shown in Table 2 and Table 4, few-shot prompting consistently improved performance over zero-shot for Dialect-to-English translations. For example, DeepSeek v3's Braj translations increased from 37.24 to 49.16 (a 32% gain). CoT prompting yielded further improvements, particularly for weaker models. For example, Mistral-7B's Magahi BLEU score rose from 8.75 (*zero-shot*) in Table 1 to 21.54 (CoT) in Table 5 for Dialect-to-Hindi translations.

However, top-tier models showed diminishing returns with CoT prompting, with few-shot prompting sometimes matching or even surpassing CoT performance. This suggests that, unlike weaker models which benefit significantly from explicit reasoning prompts, stronger models already possess substantial internal reasoning capabilities, reducing the added value of CoT prompting. Table 3 and Table 5 show that CoT prompting offers limited gains for Hindi from regional di-

[1] https://ai4bharat.iitm.ac.in/areas/tts

LLM	Awadhi			Braj			Magahi			Bhojpuri		
	BLEU	chrF	BERTScore	BLEU	chrF	BERTScore	BLEU	chrF	BERTScore	BLEU	chrF	BERTScore
Sarvam - 1B	9.99	35.4	93.33	18.49	42.18	94.98	11.16	42.31	93.53	6.62	29.25	93.02
Mistral - 7B	5.32	32.71	93.36	14.45	38.68	93.7	15.15	44.17	94.84	6.16	29.17	93.30
DeepSeek - 7B	3.19	32.13	93.81	10.02	34.62	94.31	8.77	35.00	93.36	3.73	25.98	92.22
Llama3 - 8B	18.15	46.34	94.98	19.85	46.62	95.83	31.24	53.63	95.85	8.16	37.29	94.22
Mistral 24B	10.41	35.43	93.31	21.68	43.31	94.91	21.54	45.37	94.84	16.53	41.52	94.57
Llama3 - 70B - instruct	19.44	45.71	94.80	30.79	53.90	96.07	37.86	60.98	96.57	19.44	43.29	94.80
GPT - 4o Mini	24.65	49.85	95.48	39.10	58.92	96.39	41.63	62.42	96.83	29.10	49.60	96.03
GPT - 4o	32.37	58.13	96.48	37.55	58.26	96.34	**50.51**	**70.06**	**97.69**	**41.12**	**63.35**	**96.75**
Llama 17B Maverick	**35.62**	**60.39**	**96.49**	**39.35**	**61.63**	**96.61**	37.11	58.50	96.70	26.10	54.04	95.60
DeepSeek v3	36.23	58.51	96.32	31.45	56.94	96.32	41.63	62.42	96.83	22.76	43.3	95.17

Table 3: **Dialect to Hindi:** Few-shot results.

LLM	Awadhi			Braj			Magahi			Bhojpuri		
	BLEU	chrF	BERTScore	BLEU	chrF	BERTScore	BLEU	chrF	BERTScore	BLEU	chrF	BERTScore
Sarvam - 1B	14.73	34.09	93.21	14.55	35.73	93.15	17.90	42.17	94.41	7.41	27.28	91.49
Mistral - 7B	21.54	46.04	94.25	16.32	34.21	90.99	23.88	42.97	93.39	14.33	35.96	93.09
DeepSeek - 7B	9.55	34.45	92.96	17.48	40.03	92.49	11.78	35.32	93.11	11.81	33.85	92.23
Llama3 - 8B	31.14	53.54	96.18	28.03	44.38	95.18	37.15	54.35	96.01	27.79	48.78	95.05
Mistral 24B	22.66	51.74	95.18	31.47	49.32	95.13	37.99	57.40	96.71	14.45	35.48	93.57
Llama3 - 70B - instruct	36.41	59.89	96.63	33.42	55.04	95.34	37.31	60.64	96.80	37.13	58.37	95.85
GPT - 4o Mini	35.59	58.23	96.60	42.39	64.07	96.95	58.54	73.20	98.10	46.95	63.86	97.07
GPT - 4o	35.89	60.38	96.60	43.83	63.91	96.73	**58.74**	**74.50**	**98.68**	**52.30**	**70.26**	**97.61**
Llama 17B Maverick	**42.93**	**63.89**	**97.32**	43.09	64.69	**97.17**	51.11	67.84	98.05	36.58	61.91	95.88
DeepSeek v3	38.67	60.20	96.83	**49.16**	**67.47**	97.01	49.91	67.54	97.19	38.32	58.79	96.38

Table 4: **Dialect to English:** Few-shot results.

alects. Often, its performance is marginal or below that of few-shot prompting, which appears more effective at capturing translation patterns for dialects that are linguistically close to Hindi.

Effect of Target Language: English translations consistently outperformed Hindi across all evaluation metrics. For example, in the few-shot setting, GPT-4o Mini scored 46.95 BLEU for Bhojpuri-English (Table 4) versus 29.10 for Bhojpuri-Hindi (Table 3) –a gap of over +17 points. Similarly, in the zero-shot setting, Llama3-8B achieved 31.42 for Magahi-English (Table 2) but only 12.19 for Magahi-Hindi (Table 1).

This performance gap largely stems from the training and optimization of LLMs. They are exposed to much larger and more diverse English corpora, leading to richer linguistic knowledge, and better alignment for English outputs. In contrast, Hindi has comparatively less training data and fewer fine-tuning resources, resulting in lower fluency and accuracy.

Effect of Model Size: Translation quality generally improved with larger model sizes, though gains were not always consistent across architectures. Within the LLaMA family, LLaMA3-70B Instruct substantially outperformed LLaMA3-8B (CoT Magahi-English BLEU scores: 46.80 vs 34.66 in Table 6), while in the Mistral family, performance varied massively —Mistral-24B improved over Mistral-7B in Magahi-English few shot results from 23.88 to 37.99 as shown in Table 4. However, in many other cases, Mistral-7B also outperformed its larger counterpart, Mistral-24B. Very small models, such as Sarvam-1B, delivered poor results despite Indic-specific training, indicating that limited parameter capacity restricts generalization beyond high-resource languages. In terms of practical usability, moderate-sized models like GPT-4o Mini offered strong performance relative to their larger counterpart, GPT-4o, providing a favorable balance between accuracy, cost, and accessibility. For example, as shown in Table 2, GPT-4o Mini achieved a BLEU score of 41.37 compared to GPT-4o's 42.77 for Braj-English translation.

Large Reasoning Models in Low-Resource MT: Large Reasoning Models (LRMs) such as GPT-4o, GPT-4o Mini, and Llama 17B Maverick consistently outperform traditional LLMs by leveraging enhanced reasoning capabilities and instruction-following training. For instance, in Table 6, GPT-4o achieves a BLEU score of 64.98 in Magahi-English translation, significantly surpassing the best traditional LLM (Llama3 - 70B - instruct), which reached only 46.80. Unlike stan-

LLM	Awadhi			Braj			Magahi			Bhojpuri		
	BLEU	chrF	BERTScore	BLEU	chrF	BERTScore	BLEU	chrF	BERTScore	BLEU	chrF	BERTScore
Sarvam - 1B	8.79	31.4	92.63	14.44	41.23	93.3	9.05	40.17	92.36	5.62	25.25	91.02
Mistral - 7B	16.53	41.52	94.57	21.68	43.31	94.91	21.54	45.37	94.84	10.41	35.43	93.31
DeepSeek - 7B	5.57	33.32	94.21	10.99	33.26	94.23	12.21	35.89	94.14	2.6	26.75	92.46
Llama3 - 8B	12.38	41.49	94.27	12.91	43.04	94.96	12.06	41.77	94.53	6.64	35.94	93.87
Mistral 24B	23.92	44.22	94.63	21.39	45.15	95.09	19.79	44.47	95.41	6.48	32.14	93.13
Llama3 - 70B - instruct	22.28	50.49	95.57	26.92	51.62	95.89	28.04	53.55	96.07	19.31	45.08	95.07
GPT - 4o Mini	23.7	48.95	95.49	32.04	53.43	95.96	28.74	53.77	96.19	26.36	51.65	96.21
GPT - 4o	**29.76**	**55.96**	**96.76**	32.66	56.18	96.66	**51.57**	**68.01**	**97.29**	**33.46**	**59.07**	**96.71**
Llama 17B Maverick	26.48	54.14	96.60	34.38	56.80	96.53	41.58	60.80	96.77	27.73	50.32	95.81
DeepSeek v3	29.04	52.45	96.17	**34.54**	**57.13**	**96.66**	42.05	58.95	96.13	22.76	43.3	95.17

Table 5: **Dialect to Hindi:** Chain of thought (COT) results.

LLM	Awadhi			Braj			Magahi			Bhojpuri		
	BLEU	chrF	BERTScore	BLEU	chrF	BERTScore	BLEU	chrF	BERTScore	BLEU	chrF	BERTScore
Sarvam - 1B	20.6	34.23	93.66	15.59	33.76	92.76	15.24	36.19	93.39	9.16	32.07	92.50
Mistral - 7B	26.86	48.96	95.75	35.79	48.09	95.19	27.7	50.48	95.44	17.01	37.54	93.33
DeepSeek - 7B	11.85	33.87	92.76	22.24	42.27	92.95	12.27	32.24	93.04	12.37	31.91	92.40
Llama3 - 8B	35.4	51.84	95.93	34.99	50.27	94.99	34.66	54.78	95.88	30.21	48.57	95.49
Mistral 24B	25.61	51.17	95.95	29.61	47.62	94.56	35.49	55.45	95.38	19.49	40.22	93.59
Llama3 - 70B - instruct	34.47	57.71	96.19	38.96	57.64	95.86	46.80	63.27	**98.74**	36.89	60.56	96.05
GPT - 4o Mini	28.56	53.17	95.85	48.59	66.41	97.03	55.32	68.86	97.09	47.27	65.01	96.73
GPT - 4o	35.47	58.66	96.75	50.21	68.83	**97.87**	**64.98**	**78.52**	**98.74**	53.19	**70.62**	**97.47**
Llama 17B Maverick	**42.16**	**63.77**	**97.00**	51.73	69.37	97.67	52.54	72.39	97.49	34.20	60.29	95.15
DeepSeek v3	38.97	58.24	96.43	**56.14**	**72.49**	96.99	55.56	68.81	96.66	**53.21**	68.40	96.87

Table 6: **Dialect to English:** Chain of thought (COT) results.

dard LLMs primarily trained for next-token prediction, LRMs are fine-tuned on multi-step reasoning and instruction-following tasks, enabling them to "reason through" prompts. This reasoning-centric ability helps LRMs handle dialectal variation and limited supervision more effectively than mere increases in parameter size. Even smaller instruction-tuned variants like GPT-4o Mini maintain strong translation quality, with BLEU scores exceeding 55 across multiple dialects. This underscores that reasoning ability, rather than parameter count alone, is key to enhancing low-resource MT.

4.2 Ablation Study

For ablation study, we use a single large language model: Llama-3.3-70B-Instruct-Turbo-Free (AI, 2023), accessed through the TogetherAI API.

Results for different prompt templates: We ran translation experiments from four dialects {Awadhi, Bhojpuri, Braj, Magahi} → {English, Hindi} using four different prompt templates as shown in Table 7.

Our evaluation showed clear differences in performance, helping us choose the best prompt template. The four prompt templates represent different instructional approaches: Role prompting assigns a professional translator identity to the LLM, direct prompting provides straightforward translation instructions, Specific prompting offers detailed instructions with formatting constraints and contextual analysis requirements, and Vague prompting uses deliberately ambiguous language to demonstrate the impact of unclear instructions on translation quality.

Type	Prompt
Role Prompting	You are a professional translator. Translate {Language} sentences into fluent English.
Direct Prompt	Translate the following {Language} sentences into English.
Specific Prompt	This is a translation exercise focused solely on {Language} input and English output. Please analyze the given {Language} sentence, understand its context, and provide your answer.Given an {Language} sentence, return ONLY a JSON object with the key English containing the translation.
Vague prompt	Take the input, convert it into English and provide the result.

Table 7: Types of prompt used.

As shown in Table 8, role prompting consistently outperformed other approaches across language pairs, with the highest BLEU scores for English translations (36.48 for Bhojpuri-English and 21.45 for Magahi-English). This success stems from the psychological priming effect where assigning the LLM a "professional translator" identity activates more sophisticated linguistic processing capabilities and contextual understanding.

Specific prompting was the second-best overall approach, but achieved the highest BLEU scores of 30.00 for Awadhi-English translation and 38.49 for Braj-English translation. Its use of detailed instructions, and explicit formatting enhanced trans-

Prompt	Awadhi			Braj			Bhojpuri			Magahi		
	BLEU	COMET	BLEURT	BLEU	COMET	BLEURT	BLEU	COMET	BLEURT	BLEU	COMET	BLEURT
Role	26.42	87.08	30.54	33.80	**85.66**	**36.63**	**36.48**	82.63	52.02	**21.45**	86.19	**27.11**
Direct	22.90	86.07	22.41	33.58	83.38	26.85	29.09	**86.39**	**54.57**	14.59	80.41	18.73
Specific	**30.00**	**87.73**	29.94	**38.49**	85.18	36.03	31.91	85.92	49.96	20.96	**86.29**	22.74
Vague	19.90	74.93	**32.29**	30.90	83.53	25.90	26.71	84.96	46.63	17.80	85.20	20.27

Table 8: **Dialect to English:** Experimental results with different prompt templates.

Prompt	Awadhi			Braj			Bhojpuri			Magahi		
	BLEU	COMET	BLEURT	BLEU	COMET	BLEURT	BLEU	COMET	BLEURT	BLEU	COMET	BLEURT
Role	**21.99**	**84.14**	**41.34**	**21.44**	76.61	35.18	**23.20**	85.34	44.73	**18.55**	**84.49**	42.69
Direct	14.27	79.54	37.85	20.10	76.38	35.05	20.62	83.29	42.03	14.17	78.74	40.49
Specific	21.00	84.07	40.20	20.35	78.35	**35.73**	22.85	**85.91**	**45.75**	14.68	83.16	**43.32**
Vague	15.07	79.43	35.92	19.90	**80.65**	34.28	21.87	78.07	37.29	15.30	82.68	41.09

Table 9: **Dialect to Hindi:** Experimental results with different prompt templates.

Few-Shot	Awadhi			Braj			Bhojpuri			Magahi		
	BLEU	COMET	BLEURT	BLEU	COMET	BLEURT	BLEU	COMET	BLEURT	BLEU	COMET	BLEURT
n=1	47.86	88.6	29.43	32.17	85.00	34.73	36.82	90.37	45.86	32.62	82.9	14.36
n=5	54.21	90.65	38.27	37.52	32.17	40.11	35.4	87.93	46.54	30.74	82.13	21.50
n=10	**54.25**	**90.71**	42.47	**39.58**	**87.57**	**47.02**	**41.06**	**90.41**	**54.01**	33.55	84.39	19.88
n=20	53.13	89.48	**47.45**	36.23	86.24	42.82	38.69	89.68	52.17	**38.58**	**84.54**	**26.72**

Table 10: **Dialect to English:** Experimental results with different number of few-shot examples.

Few-Shot	Awadhi			Braj			Bhojpuri			Magahi		
	BLEU	COMET	BLEURT	BLEU	COMET	BLEURT	BLEU	COMET	BLEURT	BLEU	COMET	BLEURT
n=1	29.1	**87.19**	52.56	18.39	80.39	48.6	36.05	87.39	59.71	34.33	84.44	57.35
n=5	28.9	86.29	53.12	20.36	**83.15**	**52.03**	41.21	90.61	65.46	31.39	81.11	55.01
n=10	**32.63**	86.77	**53.87**	**21.22**	82.95	51.88	47.05	91.96	68.73	38.36	84.16	58.69
n=20	32.31	85.52	53.80	20.91	82.27	50.52	**47.83**	**92.29**	**69.59**	**39.16**	**84.67**	**60.40**

Table 11: **Dialect to Hindi:** Experimental results with different number of few-shot examples.

lation quality by promoting consistency and reducing ambiguity through a structured workflow.

The direct prompting approach, while straightforward, showed moderate performance that was generally inferior to both role and specific templates, suggesting that simple instructional clarity alone is insufficient for complex translation tasks. Most notably, vague prompting consistently underperformed across all metrics and language pairs (Table 9), with particularly poor results in Hindi translations (lowest BLEU scores ranging from 15.07 to 21.87, confirming that ambiguous instructions severely compromise translation quality.

Results for different number of shots: We conduct experiments with different numbers of few-shot examples to determine if performance improves after a certain point and to establish the optimal n value (number of shots) for all future experiments. We tested with n (= 1, 5, 10, 20) shots across all four regional dialects translating to both English and Hindi. To ensure experimental rigor, we create two separate pools from our dataset: a test pool for evaluation and a few-shot pool from which we randomly selected translation examples. Since the few-shot examples are randomly selected from this pool, each experiment was repeated 10 times for each n value to eliminate selection bias, and we report the average results across all repeats. As shown in Table 10, for English translations, the optimal performance consistently emerges at 10 shots across most language pairs. Braj-to-English peaks at 10 shots (39.58 BLEU) before declining at 20 shots (36.23 BLEU), while Magahi-to-English continues improving through 20 shots but with marginal gains. Increasing shots improved translation quality up to 10 shots, after which results plateaued or showed minor gains.

As shown in Table 11, the Hindi translation results reveal varied performance patterns across the four dialects. Awadhi-to-Hindi peaks at 10 shots (32.63 BLEU) before declining at 20 shots. Bhojpuri-to-Hindi continues improving through 20 shots, suggesting that this dialect pair benefits from additional contextual examples. Magahi-to-Hindi shows moderate, consistent improvement but minimal gains between 10 and 20 shots (+0.8 BLEU). While Bhojpuri-to-Hindi benefits from 20

ASR Model	Awadhi	Braj	Bhojpuri	Magahi	Multilingual
Google STT	**0.7321**	**0.7253**	**0.7289**	**0.7198**	**0.7146**
Whisper-Medium	0.4542	0.4476	0.4487	0.4409	0.4415

Table 12: ASR performance comparison - WER scores.

shots, the remaining dialect pairs reach (near-) optimal performance at 10 shots, reinforcing 10 shots as an efficient configuration for Hindi.

4.3 ASR Results

We fine-tune Whisper on the VAANI corpus and the lifecycle narrations from the SpeeD-IA dataset. All audio files underwent a standardized preprocessing pipeline. This included resampling all files to a consistent 16 kHz, applying amplitude normalization, and filtering out segments with durations outside the 1-10 second range. This preprocessing ensures consistent input representation while eliminating outliers that could destabilize the training.

The ASR results in Table 12 demonstrate that the fine-tuned Whisper-Medium (Radford et al., 2022) model consistently outperforms the baseline Google STT API[2] across all dialects and the multilingual setting, achieving substantially lower WER scores (e.g., 0.4542 vs. 0.7321 for Awadhi). This highlights the effectiveness of domain-specific fine-tuning on audio data in improving recognition accuracy for low-resource dialects. While Google STT provides a strong out-of-the-box baseline, fine-tuning Whisper enables better adaptation to the linguistic and acoustic characteristics of these dialects, yielding more robust performance in the target speech varieties.

4.4 TTS Results

To evaluate the quality of the synthesized speech, we conduct a subjective assessment using the mean-opinion-score (MOS). A group of six human listeners rated the samples along two dimensions:

- **Adequacy:** Human evaluators assess whether the key message and details are preserved accurately, without distortions or irrelevant additions on a Likert scale of 1 (meaning is completely lost) to 5 (meaning is fully preserved).
- **Fluency**: Human evaluators assess whether the speech sounded natural and coherent, as if spoken by a fluent native speaker. Similar to the adequance, we evaluate fluency on a Likert scale of 1 (poor, full of errors) to 5 (perfectly fluent).

[2] https://cloud.google.com/speech-to-text

TTS Model	Metric	Awadhi	Braj	Bhojpuri	Magahi	Average
Kokoro TTS (English)	Adq	4.0	4.15	4.15	4.0	4.08
	Flu	4.3	4.3	4.0	4.15	4.19
IndicF5 (Hindi)	Adq	4.1	3.8	4.65	4.05	4.15
	Flu	4.3	3.85	4.5	4.0	4.16

Table 13: Average MOS scores on a Likert scale of 1-5.

S2ST	Metric	Awadhi	Braj	Bhojpuri	Magahi	Average
Dialect-English	Adq	3.97	3.80	4.09	3.86	3.93
	Flu	4.06	3.85	3.93	3.71	3.89
Dialect-Hindi	Adq	3.66	3.55	3.75	3.70	3.67
	Flu	3.92	3.81	3.88	3.64	3.81

Table 14: Cascaded S2ST: Average MOS scores on a Likert scale of 1-5.

The MOS evaluation in Table 13 shows that both TTS systems – KOKORO-TTS and IndicF5 (V et al., 2025) – achieved high adequacy and fluency across all four dialects. Notably, IndicF5 attained its highest adequacy and fluency ratings for Bhojpuri, while Kokoro TTS maintained balanced quality across dialects. These results indicate that both English- and Hindi-based TTS models produce clear, natural-sounding speech, with only marginal differences in listener preference.

4.5 Cascaded S2ST Results

We construct a test set consisting of 80 speech samples for each dialect and processed them using our cascaded S2ST pipeline. First, the audio is transcribed using the fine-tuned Whisper model. The resulting transcripts were then translated using the LLaMA Maverick 17B model. Finally, speech synthesis was performed using Kokoro TTS for English outputs and IndicF5 for Hindi outputs. The generated speech samples were evaluated by six human annotators on two perceptual dimensions—adequacy and fluency—using a 5-point MOS scale. The average scores for each dialect are reported in Table 14.

5 Conclusion

In this paper, we explored various SOTA models for speech-to-speech machine translation for dialectal variation of Hindi. We employ multiple LLMs and LRMs for translating Awadhi, Bhojpuri, Braj Bhasha, and Magahi sentences to Hindi and English. Our observation suggests that COT prompting strategy outperforms zero-shot and few-shot settings. Moreover, reasoning models such as GPT-4o, Deepseek-V3, and Llama 17B Maverik, reports strong results against other competing models in all three prompting setups.

Acknowledgment

The work is partially supported by a research project COIL-D@IIIT Delhi, funded by MeitY, Govt of India. The authors also acknowledge the support of the Infosys Foundation through Center of AI (CAI) at IIIT-Delhi.

References

Meta AI. 2023. Llama 3: Open foundation and fine-tuned chat models. https://ai.meta.com/llama.

Perla Al Almaoui, Pierrette Bouillon, and Simon Hengchen. 2025. Arabizi vs llms: Can the genie understand the language of aladdin?

Awadhi-Wikipedia. https://en.wikipedia.org/wiki/Awadhi_language.

BrajBhasha—Wikipedia. https://en.wikipedia.org/wiki/Braj_Bhasha.

Braj—Omniglot. https://www.omniglot.com/writing/braj.htm.

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.

Sara Court and Micha Elsner. 2024. Shortcomings of llms for low-resource translation: Retrieval and understanding are both the problem.

Raj Dabre and Haiyue Song. 2024. NICT's cascaded and end-to-end speech translation systems using whisper and IndicTrans2 for the Indic task. In *Proceedings of the 21st International Conference on Spoken Language Translation (IWSLT 2024)*, pages 17–22, Bangkok, Thailand (in-person and online). Association for Computational Linguistics.

Jay Gala, Pranjal A Chitale, A K Raghavan, Varun Gumma, Sumanth Doddapaneni, Aswanth Kumar M, Janki Atul Nawale, Anupama Sujatha, Ratish Puduppully, Vivek Raghavan, Pratyush Kumar, Mitesh M Khapra, Raj Dabre, and Anoop Kunchukuttan. 2023. Indictrans2: Towards high-quality and accessible machine translation models for all 22 scheduled indian languages. *Transactions on Machine Learning Research*.

Anirudh Gupta, Harveen Singh Chadha, Priyanshi Shah, Neeraj Chimmwal, Ankur Dhuriya, Rishabh Gaur, and Vivek Raghavan. 2021. Clsril-23: Cross lingual speech representations for indic languages.

Amr Hendy, Mohamed Abdelrehim, Amr Sharaf, Vikas Raunak, Mohamed Gabr, Hitokazu Matsushita, Young Jin Kim, Mohamed Afify, and Hany Hassan Awadalla. 2023. How good are gpt models at machine translation? a comprehensive evaluation. *arXiv preprint arXiv:2302.09210*.

Tahir Javed, Kaushal Bhogale, and Mitesh M. Khapra. 2025. NIRANTAR: Continual Learning with New Languages and Domains on Real-world Speech Data. In *Interspeech 2025*, pages 918–922.

Kartik Kartik, Sanjana Soni, Anoop Kunchukuttan, Tanmoy Chakraborty, and Md. Shad Akhtar. 2024. Synthetic data generation and joint learning for robust code-mixed translation. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 15480–15492, Torino, Italia. ELRA and ICCL.

KOKORO-TTS. https://kokorotts.net/models/Kokoro/text-to-speech.

Amit Kumar, Rajesh Kumar Mundotiya, and Anil Kumar Singh. 2020. Unsupervised approach for zero-shot experiments: Bhojpuri–Hindi and Magahi–Hindi@LoResMT 2020. In *Proceedings of the 3rd Workshop on Technologies for MT of Low Resource Languages*, pages 43–46, Suzhou, China. Association for Computational Linguistics.

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*.

Shivam Kumar and Md Shad Akhtar. 2025. CLEAR: Code-mixed ASR with LLM-driven rescoring. In *Proceedings of the 8th International Conference on Natural Language and Speech Processing (ICNLSP-2025)*, pages 339–348, Southern Denmark University, Odense, Denmark. Association for Computational Linguistics.

Shivam Mhaskar, Vineet Bhat, Akshay Batheja, Sourabh Deoghare, Paramveer Choudhary, and Pushpak Bhattacharyya. 2023. Vakta-setu: A speech-to-speech machine translation service in select indic languages.

Alec Radford, Jong Wook Kim, Tao Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pam Mishkin, Jack Clark, et al. 2022. Robust speech recognition via large-scale weak supervision. *arXiv preprint arXiv:2212.04356*.

Nivedita Sethiya, Ashwin Sankar, Raj Dabre, and Chandresh Kumar Maurya. 2024. WSLT 2024 Indic Track. https://iwslt.org/2025/indic.

Jonathan Shen, Ruoming Pang, Ron J. Weiss, Mike Schuster, Navdeep Jaitly, Zongheng Yang, Zhifeng Chen, Yu Zhang, Yuxuan Wang, R. J. Skerry-Ryan, Rif A. Saurous, Yannis Agiomyrgiannakis, and Yonghui Wu. 2017. Natural TTS synthesis by conditioning wavenet on mel spectrogram predictions. *CoRR*, abs/1712.05884.

VAANI Team. 2025. Vaani: Capturing the language landscape for an inclusive digital india (phase 1). https://vaani.iisc.ac.in/.

Praveen S V, Srija Anand, Soma Siddhartha, and Mitesh M. Khapra. 2025. Indicf5: High-quality text-to-speech for indian languages. https://github.com/AI4Bharat/IndicF5.

David Vilar, Markus Freitag, Colin Cherry, Jiaming Luo, Viresh Ratnakar, and George Foster. 2023. Prompting palm for translation: Assessing strategies and performance. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL 2023)*, pages 15406–15427.

Biao Zhang, Barry Haddow, and Alexandra Birch. 2023. Prompting large language model for machine translation: A case study. *arXiv preprint arXiv:2301.07069*.

Appendix

A Linguistic Description and Translation Challenges of Hindi Dialects

A.1 Overview of Dialects

Awadhi, Bhojpuri, Braj Bhasha, and Magahi are Indo–Aryan languages traditionally considered Hindi dialects. Awadhi (Eastern Hindi subgroup of Central Indo–Aryan) is spoken in the Awadh region of Uttar Pradesh, India, and the adjacent Terai of Nepal (Awadhi-Wikipedia). According to the 2011 census, it had about 3.8 million speakers (Awadhi-Wikipedia). Braj Bhasha ("Braj"; Western/Central Indo–Aryan) is spoken in the Braj region (Mathura–Agra) of western Uttar Pradesh and parts of Rajasthan (BrajBhasha—Wikipedia; Braj—Omniglot), with about 1.5 million native speakers (Braj—Omniglot). Bhojpuri is an Eastern Indo–Aryan (Bihari) language spoken in the Bhojpur–Purvanchal area (eastern UP, western Bihar, NW Jharkhand) and Nepal's Terai; the 2011 census reports approximately 50.5 million speakers. Magahi (Magadhi) is another Eastern Indo–Aryan (Magadhan/Bihari) language native to southern Bihar and northern Jharkhand, with about 12.7 million speakers.

All four share SOV grammar, two genders, and postpositions, and use Devanagari today, but have distinct histories and linguistic classifications (BrajBhasha—Wikipedia). Awadhi and Braj are often grouped under "Central/Western Hindi," whereas Bhojpuri and Magahi fall under the Eastern Indo–Aryan (Bihari) group. In practice, none enjoy official status comparable to Standard Hindi.

A.2 Historical and Cultural Background

Awadhi: A major literary dialect of medieval India. Tulsīdās's *Ramcharitmanas* and the *Hanumān Chālīsa* were composed in Awadhi, giving it prestige in Bhakti literature (Awadhi-Wikipedia). Though displaced by Standard Hindi in education and administration, it remains strong in rural speech and folk music.

Braj Bhasha: The classical language of Krishna devotional poetry between the 15th–18th centuries. Poets such as Surdas and Mirabai composed extensively in Braj. Today it survives mainly in folk devotion and rural speech; it has no modern official status (BrajBhasha—Wikipedia; Braj—Omniglot).

Bhojpuri: A vibrant spoken dialect with a global diaspora (Fiji, Mauritius, Suriname, Trinidad). Bhojpuri has strong folk performing arts (e.g., Bhikhari Thakur) but limited formal literary status. UNESCO lists it as "potentially vulnerable." Urban speakers often replace traditional forms (e.g., बुझैया meaning "to understand") with Hindi analogues.

Magahi: The modern descendant of Magadhi Prakrit. Historically oral, with minimal written tradition. Spoken widely in Bihar/Jharkhand but lacks official recognition; Standard Hindi dominates schooling. Magahi speakers frequently code-switch and may face social stigma.

A.3 Linguistic Features Illustrated Through an Example

Linguistic variations for English sentence: '*I like mango*'.

Hindi: मुझे आम अच्छा लगता है।
Braj: मोइ आम अच्छे लगत ऐं।
Bhojpuri: हमके आम अच्छा लागेला।
Magahi: हमरा आम अच्छा लगऽ हे।
Awadhi: हमका आम अच्छा लागा थय।

A.3.1 Pronouns

Hindi "मुझे" (mujhe, dative "to me") maps differently across dialects:

- Braj: मोइ (moi)
- Bhojpuri: हमके (humke)
- Magahi: हमरा (humra)
- Awadhi: हमका (humka)

Each dialect has its own oblique case system for first–person pronouns.

A.3.2 Verb Morphology

Hindi uses "लगता है" (lagta hai | to be).

Dialectal variants:

- Braj: लगत ऐं (lagat ae)
- Bhojpuri: लागेला (lagela)
- Magahi: लगऽ हे (lag he)
- Awadhi: लागा थय (laga the)

Patterns:

- Eastern Bihari dialects (Bhojpuri, Magahi) often use verb stem + -ला / -ल.
- Awadhi retains older Indo–Aryan -आ morphology.
- Braj preserves archaic endings like -एँ / -ऐं.

A.3.3 Agreement and Vocabulary

All four use "अच्छा" (achcha | good) in this sentence, but differ elsewhere. Braj and Awadhi preserve Sanskritisms; Bhojpuri and Magahi show Eastern Indo–Aryan features.

A.3.4 Writing Systems

All four dialects use Devanagari today. Historically:

- Awadhi & Bhojpuri: Kaithi
- Magahi: Kaithi + regional scripts (Bengali, Odia)

Standard orthography varies.

A.4 Speech and Translation Challenges

ASR Challenges: Dialects lack large transcribed corpora; existing datasets contain only 4–5 hours per dialect. Standard Hindi ASR performs poorly due to morphology, lexicon, and accent mismatches. Crowdsourced audio often suffers from noise and device variation.

Machine Translation Challenges: Parallel corpora are extremely scarce. MT is hindered by:

- inconsistent spellings,
- divergent pronoun/verb systems,
- lack of grammar descriptions,
- heavy code-mixing.

Shared scripts and cognates help unsupervised MT (Kumar et al., 2020), but zero-shot transfer from Hindi remains unreliable.

TTS Challenges: No high-quality TTS exists for these dialects. Hindi TTS adaptation often mispronounces dialect forms (e.g., "थय" vs "है"). Studio-quality recordings are unavailable.

Sociolinguistic Constraints: Low prestige, lack of inclusion in education, and self-identification as "Hindi" reduce dataset availability.

B Ablation based on Quality and Relevance

B.1 Selecting few shot examples based on quality

To investigate the impact of the quality of the few-shot examples selected, we constructed two distinct data pools, each containing 100 examples. The high-quality pool consisted of original examples from our dataset with accurate Hindi and English translations of the dialect sentences, while the low-quality pool was systematically created by manually corrupting the Hindi and English translations while keeping the source dialect sentences (Awadhi, Bhojpuri, Braj Bhasha, and Magahi) unchanged. A few example sentences from the poor quality pool are listed in Table 15. From each pool, we randomly sampled n=10 examples to create few-shot learning scenarios.

To eliminate sampling bias, we repeated the experiment 10 times and the final performance metrics represent the average across all runs, providing an unbiased assessment of how example quality affects few-shot MT performance from regional dialects to Hindi and English.

Awadhi	Orginal Translation	Poor Translation
हमका आम अच्छा लागा थय।	I like mango.	I ate a banana.
पेड़े पय बांदर अहय।	The monkey is on the tree.	The monkey is eating a sandwich.
ऊ घर बड़ा अहय।	That house is big.	The dog is very big.
हम राधा अही।	I am Radha.	I am Rad.
उनका नाम कृष्णा अहय।	His name is Krishna.	His life is Krish.
हहुंका सर दर्द अहय।	I have a headache.	My body is aching.
वे एक मनई का देखी।	She saw a man.	She saw a cake.
वे शादी के बरे एक लड़की देखे।	He saw a girl for marriage.	She saw for marriage.

Table 15: Example sentences from the poor quality pool.

As shown in Table 16 and Table 17, the experimental results show a consistent pattern across most language pairs and metrics, underscoring the importance of high-quality training examples in few-shot machine translation. For dialect-to-Hindi translations, good-quality examples substantially outperform poor-quality ones (e.g., Awadhi BLEU: 32.63 vs 14.72, Bhojpuri: 47.05 vs 18.46). Dialect-to-English translations also benefit, with notable improvements in BLEURT scores (Awadhi: 42.47 vs 38.47, Bhojpuri: 54.01 vs 40.42). These findings validate our hypothesis that careful curation of few-shot examples significantly enhances MT performance, highlighting the need for quality-aware example selection in low-resource dialect translation tasks.

Quality	Awadhi			Braj			Bhojpuri			Magahi		
	BLEU	COMET	BLEURT	BLEU	COMET	BLEURT	BLEU	COMET	BLEURT	BLEU	COMET	BLEURT
Good	54.25	90.71	42.47	39.58	87.57	47.02	41.06	90.41	54.01	33.55	84.39	19.88
Poor	36.08	86.91	38.47	39.95	85.59	22.8	39.55	87.14	40.42	38.2	87.35	33.86

Table 16: **Dialect to English:** Good vs Poor quality few-shot examples selection.

Quality	Awadhi			Braj			Bhojpuri			Magahi		
	BLEU	COMET	BLEURT	BLEU	COMET	BLEURT	BLEU	COMET	BLEURT	BLEU	COMET	BLEURT
Good	32.63	86.77	53.87	21.22	82.95	51.88	47.05	91.96	68.73	38.36	84.16	58.69
Poor	14.72	86.13	43.69	23.04	84.5	52.65	18.46	83.72	47.78	21.33	85.08	49.46

Table 17: **Dialect to Hindi:** Good vs Poor quality few-shot examples selection.

Selection	Awadhi			Braj			Bhojpuri			Magahi		
	BLEU	COMET	BLEURT	BLEU	COMET	BLEURT	BLEU	COMET	BLEURT	BLEU	COMET	BLEURT
Random	54.25	90.71	42.47	39.58	87.57	47.02	41.06	90.41	54.01	33.55	84.39	19.88
LABSE	36.08	86.91	38.47	39.95	85.59	22.8	39.55	87.14	40.42	38.2	87.35	33.86

Table 18: **Dialect to English:** Random vs LABSE few-shot example selection.

Selection	Awadhi			Braj			Bhojpuri			Magahi		
	BLEU	COMET	BLEURT	BLEU	COMET	BLEURT	BLEU	COMET	BLEURT	BLEU	COMET	BLEURT
Random	32.63	86.77	53.87	21.22	82.95	51.88	47.05	91.96	68.73	38.36	84.16	58.69
LABSE	14.72	86.13	43.69	23.04	84.5	52.65	18.46	83.72	47.78	21.33	85.08	49.46

Table 19: **Dialect to Hindi:** Random vs LABSE few-shot example selection.

B.2 Selecting few-shot examples based on relevance

In our experiment, we compare with two different strategies for selecting few-shot examples: random sampling from our curated pools versus LABSE-based semantic similarity selection. The LABSE approach selected examples that were semantically most similar to the test sentence in the embedding space, while the random approach selected examples without consideration of semantic similarity. Both selection strategies used the same underlying pools of high-quality examples, with the key difference being the selection methodology rather than the example quality. As shown in Table 18 and Table 19, the results consistently show that random selection of few-shot examples outperforms LABSE-based semantic similarity selection across all language pairs. This is especially clear in dialect-to-Hindi translations, where random selection yields substantially higher BLEU scores (Awadhi: 32.63 vs 14.72, Bhojpuri: 47.05 vs 18.46). These findings challenge the assumption that semantically similar examples provide better few-shot guidance; instead, diverse random examples better cover linguistic patterns, enabling models to generalize more effectively in low-resource dialect translation tasks.