LLM-Based Localization in the Context of Low-Resource Languages

Hardi Trivedi
Computer Engineering Department
San José State University
San José, USA
hardi.trivedi@sjsu.edu

Jorjeta G. Jetcheva
Computer Engineering Department
San José State University
San José, USA
jorjeta.jetcheva@sjsu.edu

Carlos Rojas
Computer Engineering Department
San José State University
San José, USA
carlos.rojas@sjsu.edu

***Abstract*—Recent advancements in large language models (LLMs) have enabled the development of systems capable of generating human-like responses across a wide range of tasks. However, research focus has been primarily on high-resource languages such as English, German, and French, whereas low-resource languages have not benefited from these advances. This has created a challenge for localization which requires multinationals to deploy natural language processing-based tools across world-wide geographic footprints.**

In this paper, we evaluate the state-of-the-art in question-answering for several low-resource Indian languages, including Hindi and Gujarati, and explore a sample Human Resources use case. We focus on neural machine translation based on transfer learning, multilingual meta-learning, and zero-shot approaches, combined with open source LLMs with conversational capabilities.

***Index Terms*—Natural Language Processing, Large Language Models, Neural Machine Translation, Low Resource Languages, Conversational Agents**

I. Introduction

Recent advancements in Natural Language Processing (NLP), especially in large language models (LLMs), have revolutionized a wide-range of business use cases, including customer service, enterprise applications such as human resources and procurement, health and insurance-related services, among others. However, these newly available benefits of state-of-the-art NLP are predominantly concentrated in English and around 20 or so other popular languages, out of the approximately 7,000 languages spoken worldwide [1]. This narrow focus contrasts sharply with the global linguistic landscape, where 75% of the population does not speak English, and only 6% are native English speakers, leaving the vast majority without meaningful access to these technological advancements [2].

One of the reasons for the lower pace of development of conversational agents for low-resource languages is the scarcity of labeled linguistic data and resources. Languages with limited digital presence often lack the extensive corpora necessary for training robust NLP models, leading to difficulties in understanding and generating nuanced, contextually appropriate conversations [3]. Additionally, the absence of standardised tools and pre-trained models for these languages further complicates the development process, requiring researchers to invest significant effort in data collection and model customization [4]. Another significant hurdle is the linguistic diversity and complexity inherent to many low-resource languages, including variations in syntax, morphology, and semantics, which can pose challenges for NLP algorithms primarily designed for European languages [5]. Furthermore, creating question-answering and conversational agents demands not only linguistic proficiency but also cultural and contextual understanding, intensifying the challenge when resources to capture these elements are scarce.

We build on previous work by exploring transfer, multilingual meta-learning, and zero-shot learning, in the context of LLM-based question-answering agents in Hindi and Gujarati. Hindi is spoken by over 570 million people and Gujarati by over 45 million, and come into play in localization use cases.

Our main contributions in this work are as follows:

- We evaluate the question-answering capabilities of current state-of-the-art monolingual and multilingual meta-learning approaches for machine translation of low-resource languages in the context of Hindi and Gujarati.
- We curate a new conversational dataset in Gujarati based on Gujarati plays, and a sample human resource Gurati dataset which we subject to human evaluation.
- We evaluate the effectiveness of fine tuning pre-trained translation models with different amounts of training data to understand how limited data quantity in these low-resource languages impacts potential performance improvements.
- We evaluate the effectiveness of utilizing translation in conjunction with large language model-based conversational agents, including based on the low compute open source models llama2-chat-7b [6], gemma-2b, and gemma-7b [7]. Open source models are often required in business settings due to data privacy and cost considerations. Among those, smaller models further reduce operational costs.
- We present a small case study of a human resource use case to explore the quality of low-resource language interactions in a real-world application, using Gujarati as an example.

II. Related Work

Transfer learning has been shown to be an effective approach to addressing translation challenges for low-resource

languages by training models on data-rich languages and then applying the learned language patterns to data-scarce languages. For instance, Zoph et al. [8] improved Turkish translations by initially training on French or German, boosting the translation's BLEU score by 1.3 points. The work by Gu et al. [9] extends neural machine translation for low-resource languages through meta-learning using multilingual high-resource languages. In our work, we utilize both the transfer learning and multilingual meta-learning approaches in the context of the low-resource Indian languages, Hindi and Gujarati.

Zero-shot translation uses a multilingual NMT model trained on certain language pairs to translate between completely different languages without further training [10]. For example, Baijun et al. trained models on pairs like French to Danish and French to English, and tested them on untrained languages like Arabic, Spanish, and Russian [11].We apply the zero-shot learning approach to Gujarati, which is a very low-resource language.

There has been substantial interest in machine translation for Indian languages in recent years. This has resulted in the curation of a number of datasets such as such as 'HindEnCorp' and 'HindMonoCorp' [12]. In addition, work in evaluating machine translation methods includes results by Revanuru et al. [13] who demonstrate that NMT enhances translations compared to statistical machine learning approaches on six Indian languages.We build on this work, by exploring transfer, multilingual, and zero-shot learning approaches in the context of transformer-based architectures, with a focus on translation in a conversational context.

Recent research on conversational agents for low-resource languages has focused on cross-lingual language training. For example, [14] proposed an English-Chinese dialogue system utilizing cross-lingual training with parallel data for their GPT-2 model, and found that the quality of translation influences the system's ability to generate dialogue. In our work, we pursue transfer learning and multilingual meta-learning to explore how NMT improvements translate to improvements in conversational quality for low-resource Indian languages, with a focus on Hindi and Gujarati.

IndicBART [15] is a multilingual sequence-to-sequence model based on a transformer architecture [16], and is derived from BART [17], which is an extension of the BERT model [18] to support translation tasks [15]. It was extensively pre-trained on a corpus tailored to Indic languages and was shown to outperform other state-of-the-art models. We use IndicBART as our baseline model and describe it in more detail in Section III-B.

We build on previous studies by exploring both transfer, multilingual meta-learning, and zero-shot learning, for both Hindi and Gujarati, in the context of large language model-based conversational agents, including based on the low compute open source models llama2-chat-7b [6], gemma-2b [7], and gemma-7b [7]. We also present a small human evaluation-based study based on a sample human resource questions dataset with a focus on Gujarati translation.

III. Methodology

We describe our approach, NMT and conversational models, datasets, and experimental setup next.

A. Approach

Current state-of-the-art in conversational NMT for low-resource languages such as Hindi and Gujarati is limited, with more advancements being available for general machine translation of text, and not necessarily conversational text.

We employ a model pipeline where a translation model is used to translate user questions and answers from the low-resource language (e.g., Hindi and Gujarati) to and from English, but the answer generation is performed by an LLM in English (Figure 7 in Appendix A). This enables us to take advantage of the low-resource language translation capabilities of NMT models such as IndicBART [15], while also benefiting from the advanced conversational capabilities of state-of-the-art LLMs such as llama2 [6] and gemma [7], available in English. This pipeline architecture is explained in Figure 1.

We describe the specific models we used in our pipeline next.

B. Machine Learning Models

For our NMT experiments, we use the IndicBART [15] model as a starting point because it has been pre-trained extensively on Indian languages, including Hindi and Gujarati. For our conversational model component, we experiment with Llama2 [6] with 7 billion parameters and two versions of Gemma [7], one with 2 billion and another with 7 billion parameters. We chose to utilize these models because they perform well in conversational settings, they are open source, and they have relatively modest compute requirements.

We provide additional details about each of the models in the rest of this section.

1) IndicBART: IndicBART [15] is a NMT model, which uses an mBART-50 [19] as a base, and is extensively trained on 11 Indian languages, including Assamese, Bengali, Gujarati, Hindi, Marathi, Odiya, Punjabi, Kannada, Malayalam, Tamil, Telugu, in addition to English. It is thus a great starting point for translation experiments based on both monolingual approaches and on multilingual and zero-shot approaches that rely on learning patterns across related languages.

IndicBART is a multilingual sequence-to-sequence model based on a transformer architecture [16], and is derived from BART [17], which is an extension of the BERT model [18] that supports translation tasks [15].

The IndicBART model uses a configuration of 6 encoder and decoder layers, with 1024 hidden units and 4096 filter sizes, and 16 attention heads (244M parameters).

IndicBART has been shown to surpass the performance of other state-of-the-art models such as mBART-50 [19] and mT5 [20] on Indian languages on tasks such as machine translation and text summarization, achieving up to a 2-point improvement in BLEU/ROUGE scores. It was extensively pre-trained on a corpus tailored to Indic languages, including 452 million sentences and 9 billion tokens, with significant content

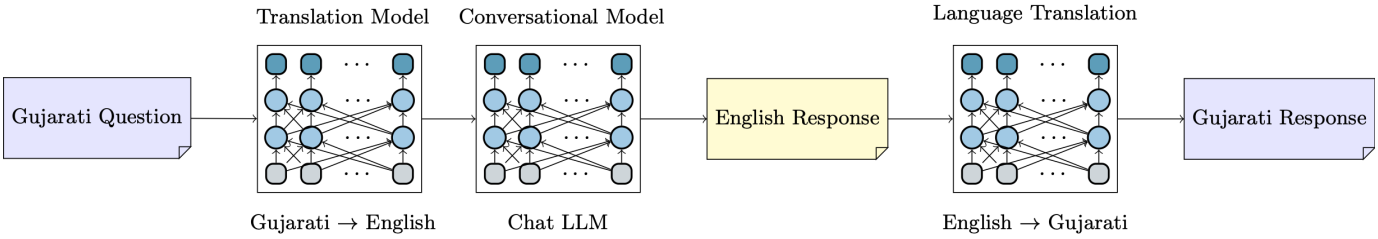


Fig. 1: Diagram of our end-to-end low-resource language conversational translation pipeline.

in Indian English. In particular, it was trained on the WAT 2021 dataset [21], CVIT-PIB dataset [22], and Samantar Corpus [23].

We started with the default parameters recommended for IndicBART by its authors, and experimented with additional values for its parameters such as the learning rate, adjusting it between 0.001 and 0.007 and settling on an optimal value of 0.005. For weight decay, we tested a range from 0.00001 to 0.0001 and found the most effective rate to be 0.00005. In terms of training duration, we varied the number of epochs from 15 to 30, ultimately finding that 20 epochs were the most effective. This led us to finalize the model's configuration.

2) Conversational LLMs: Llama2 [6] is a family of pre-trained and fine-tuned open source LLMs with a number of parameters varying from 7 billion to 70 billion. The fine-tuned versions, known as Llama2-chat, are specifically optimized for dialogue and conversational applications.

For our conversational experiments, we utilize Llama2-chat-7B, which is the 7-billion parameter Llama2-chat model, which requires the least compute resources. Gemma [7] is a recent lightweight state-of-the-art open source model developed by Google DeepMind. We use both the 2 billion parameter version and a 7 billion parameter version for our experiments.

C. Datasets

We curated a variety of datasets to enable us to conduct low resource language machine translation experiments in a conversational context. We summarize the properties of our datasets and how we use them (e.g., training vs. testing) in Table I in Appendix B.

D. Experimental Setup

The first major approach we explore involves language translation where each language is represented with its own embedding space. We refer to those experiments as monolingual experiments. The second set of experiments we conduct focus on multilingual language representations shared across a set of languages, which in our case includes Hindi, Gujarati, Marathi, and Bengali. The idea is that in the absence of data for a particular low-resource language, training a model on data from similar languages could lead to the model learning patterns that it can utilize to perform translations in the low-resource language. In addition, this kind of multilingual meta-learning can be used for zero-shot translation which we explore as well.

Finally, we evaluate the role of the LLM conversational component by comparing performance across several llama2 and gemma models.

We provide the details of our experiments next. Details of our evaluation metric, Blue, are included in Appendix D.

1) Monolingual Experiments: We conduct the following monolingual experiments:

Monolingual Experiment 1: *Out-of-the-box performance.* We first evaluate the performance of our state-of-the-art baseline mode, IndicBART, on our conversational test set IN22-Conv Appendix B2) for both English to Hindi, Hindi to English, English to Gujarati, and Gujarati to English. This performance serves as our baseline as we explore fine-tuning approaches to improve conversational performance with additional data.

Monolingual Experiment 2: *Impact of fine-tuning with low-resource language data.* We explore the impact of fine-tuning the IndicBART model with varying amounts of low-resource language text (Movie Subtitles-Hi dataset described in Appendix B1), both for translation between Hindi and English (based on fine-tuning with Hindi text) and between Gujarati and English (based on fine-tuning with Gujarati text - Gujarati Plays-Enhanced dataset described in Appendix B3).

Monolingual Experiment 3: *Impact of sequential fine-tuning with multiple low-resource languages.* Due to the scarcity of Gujarati conversational data, we explore a fine-tuning appoach based on first fine-tuning IndicBART on Hindi data (Movie Subtitles-Hi dataset described in Appendix B1, which consists of 18,000 sentence pairs, and then fine-tune on the Gujarati Plays-Enhanced dataset (Table I).

Monolingual Experiment 4: *Impact of number of encoders used during inference.* To understand the impact of compute limitations on conversational translation quality, we vary the number of encoders from 1 to 6 (the maximum used in IndicBART) to the best performing models of experiments 2 and 3 described earlier. This experiment would enable us to better understand the trade-offs between model performance and compute requirements/cost.

2) Multilingual Experiments: We conduct the experiments below to evaluate the performance of multilingual meta-learning approaches in the context of Hindi and Gujarati conversational translations.

Multilingual Experiment 1: *Multilingual parallel corpus training with testing on Hindi.*

We explore the baseline performance of our model trained on the parallel multilingual dataset Movie Subtitles-Hi-Bn-Mr when tested in Hindi conversational data (IN22-Conv). For reference, our datasets are described in Section III-C and summarized in Appendix B.

Multilingual Experiment 2: *Multi-lingual parallel corpus training with zero-shot testing on Gujarati.*

We explore the baseline performance of our model trained on the parallel multilingual dataset Movie Subtitles-Hi-Bn-Mr when tested in Gujarati (In22-Conv-500). As Gujarati is not part of the Movie Subtitles-Hi-Bn-Mr dataset, this experiment is an application of the zero-shot learning approach where we evaluate how well a model trained on similar languages to the target language can learn to translate the target language.

Multilingual Experiment 3: *Multi-lingual parallel corpus training with fine-tuning on Hindi, and testing on Hindi.*

We extend experiment 1 above with additional fine-tuning (using the Movie Titles-Hi dataset) to explore whether we can improve performance further.

Multilingual Experiment 4: *Multi-lingual parallel corpus training with fine-tuning on Hindi, then fine-tuning on Gujarati, and testing on Gujarati.*

We extend experiment 2 above with additional fine-tuning in Hindi (using the Movie Subtitles-Hi dataset) to explore whether we can improve performance further.

3) Conversational Model Ablation Study: Given the best translation model, what is the impact of using Llama2-chat-7b vs. Gemma-2b and Gemma-7b (described in Appendix E).

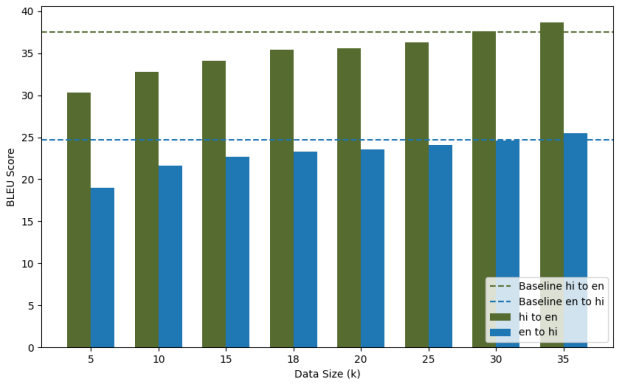
E. Qualitative Human Evaluation

To better understand some of the nuances of the quality of the conversational machine translations produced in our end-to-end conversational experiments, we conducted a small-scale human evaluation, which involved a review of the Conversational Examples test set we used for testing our system (Table I). We present the insights we derived from the evaluation in Appendix F.

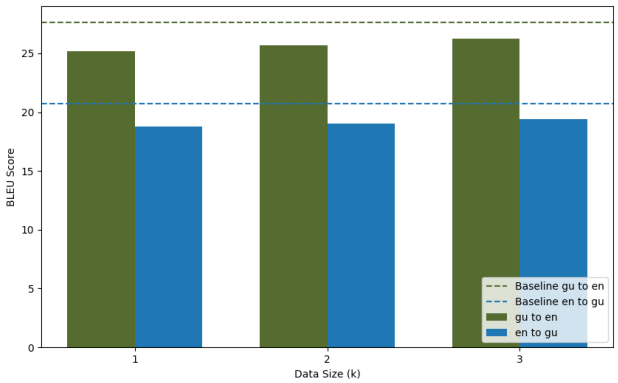
F. Human Resource (HR) Case Study

Multinational companies need to provide HR services to all their branches across the world, across a variety of different languages and cultural contexts. To evaluate the quality of conversational interactions in Gujarati in an HR context, we construct a small HR document corpus and a question-answer test dataset based on a real company's HR policies [24]. Our document corpus contains 40 pages with 13,694 words, covering a range of HR policies, including employment at will, workplace safety, workplace guidelines, and employee benefits. We use Retrieval-Augmented Generation (RAG) in conjunction with our model pipeline (Figure 1 in Appendix A). When a query is received by the conversational model, the RAG system retrieves pertinent information from the handbook and provides it as context to the LLM, ensuring the information is aligned with the HR policy in document corpus.

For our experiment, we utilized three LLMs as conversational models: LLAMA 2, Gemma 2b, and Gemma 7b. A test set of 15 question-answer pairs was manually curated, with questions chosen to represent typical queries for each HR policy (Table II in Appendix C). We also conducted a small qualitative human evaluation.



(a) Monolingual Experiment 2: Fine-tuning IndicBART on MovieSubtitles-Hi.



(b) Monolingual Experiment 2: Fine-tuning IndicBART on Gujarati Plays-Enhanced.

Fig. 2: Results of Monolingual Experiment 2.

IV. Results

We describe the results of our experiments outlined in Section III-D next.

A. Monolingual Transfer Learning Experiments

We start out by evaluating how IndicBART performs on our conversational test data (Section III-C), both in terms of translating between Hindi and English, and Gujarati and English (Monolingual Experiment 1 described in Section III-D1). We use the results of this experiment as a baseline for our fine-tuning experiments (Monolingual Experiment 2 described in Section III-D1), shown as dotted horizontal lines in Figures 2a and 2b).

Overall, we note that conversational translation performance between English and Hindi and Gujarati is relatively low,

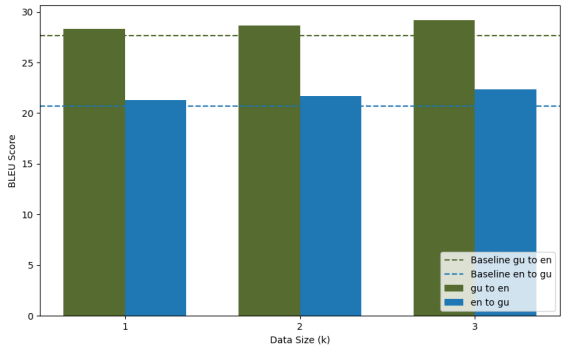


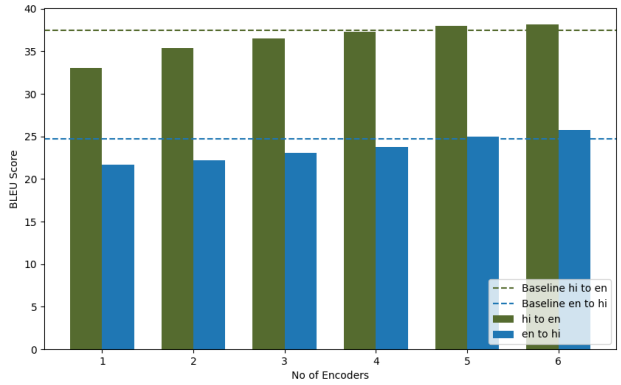
Fig. 3: Monolingual Experiment 3: Impact of sequential fine-tuning of IndicBART on MovieSubtitles-Hi and then on Gujarati Plays-Enhanced.

reaching as high as 38% percent for Hindi to English translation and as low as 18% for English to Hindi and English to Gujarati. Due to the substantial model training with English language data and resulting higher English sophistication, translation from low resource languages to English is always better than translating from English to a low resource language.

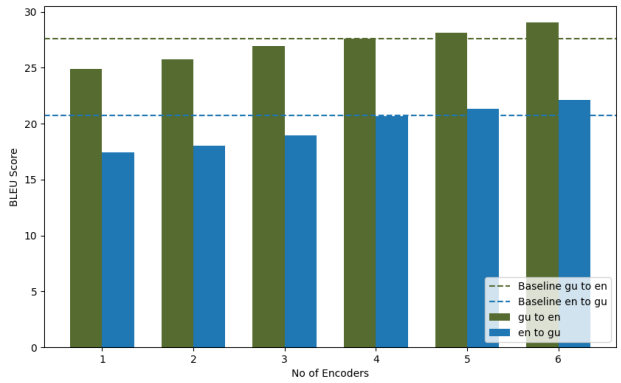
We also note that as we tune the baseline (pretrained IndicBART) model with conversational data for Hindi, for smaller dataset sizes, performance is lower and gradually increases with the size of the dataset (Figure 2a). This is because initially the new type of data is difficult for the model to adapt to, but gradually performance improves and when the size of the data used for fine-tuning reaches 30,000 data points, we start to see benefits of the fine-tuning for both English to Hindi and Hindi to English translation. This gives us some insight into how much low-resource language data is needed to enable a general pre-trained model to be able to handle conversational data in a low-resource language.

There are even fewer and smaller datasets available for Gujarati than Hindi, making it a very low resource language. Our Gujarati Plays-Enhanced dataset is much smaller than the Hindi dataset, numbering only 3,000 data points. As a result, even after fine-tuning, performance on both Hindi to English and English to Hindi translation is lower than the baseline performance (Figure 2b).

We try to compensate for the lack of Gujarati data by first fine-tuning IndicBART with our Hindi dataset (Movie Titles-Hi), and then fine-tuning with our Gujaraty Plays-Enhanced dataset (Monolingual Experiment 3 described in Section III-D). Intuitively, we were hoping that the model would be able to absorb pattern from Hindi that would help it to better handle Gujarati. This was indeed the case as shown in Figure 3 where both Gujarati to English and English to Guajarti translation performance exceeded the pretrained IndicBART baseline by up to 5.6% , and gradually improves as we increase the number of training data points from 1,000 to 3,000.



(a) Monolingual Experiment 4: Impact of number of encoders on Hindi translation.



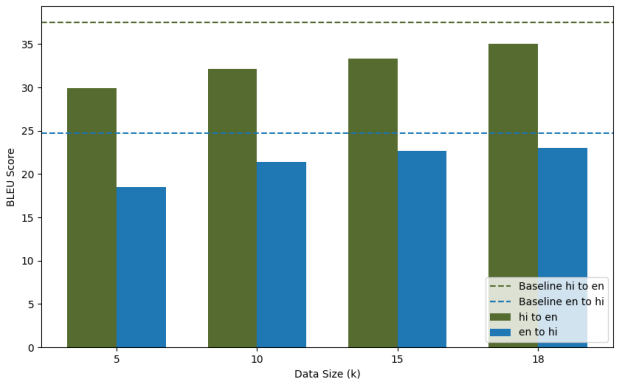
(b) Monolingual Experiment 4: Impact of number of encoders on Gujarati translation.

Fig. 4: Monolingual Experiment 4.

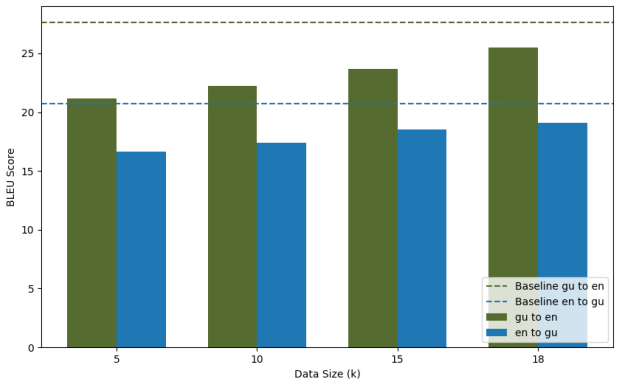
In order to get insight into the impact of using fewer than the 6 encoders IndicBART is typically configured with, we conducted an experiment where we varied the encoders from 1 to 6 (Experiment 4 described in Section III-D). This experiment was applied to the best performing models from our prior experiments (where for both Hindi and Gujarati, we first fine tuned on 35,000 data points from Movie Subtitles-Hi, and in addition fine-tuned on Gujarati Plays-Extended in the Gujarati experiment). Using fewer encoders enables the use of the translation models on compute infrastructure with lower compute requirements and at a lower cost. We find that at least 5 encoders are needed to exceed the baseline performance model and as the number of encoders increases, performance gradually increases as well (Figures 4a and 4b).

B. Multilingual Meta-Learning Experiments

Our multilingual meta-learning experiments 1 and 2 (Section III-D2) indicate that the fine tuning data was not sufficient to achieve performance even on par with the baseline models (Figures 5a and 5b). However, when we performed



(a) Multilingual Experiment 1: Parallel corpus training on Movie Subtitles-Hi-Bn-Mr, with testing on Hindi.



(b) Multilingual Experiment 2: Parallel corpus training on Movie Subtitles-Hi-Bn-Mr, with testing on Gujarati.

Fig. 5: Multilingual Experiments.

Multilingual Experiments 3 and 4, which included further fine-tuning with Hindi (Experiment 3) and with Hindi and then Gujarati (Experiment 4), the multilingual models outperformed the baseline, and in the case of Gujarati, also outperformed the best performing monolingual transfer learning model, which resulted from Monolingual Experiment 3 (Figure 6). This is an indication that multilingual approaches that take advantage of training data of languages similar to low-resource Indian languages can in take advantage of common patterns across related languages, and hold promise for enabling low-resource language advances despite training data scarcity.

C. Conversational LLM Ablation Results

Our conversational ablation experiments explored the output of each stage of our model pipeline (Appendix A) and revealed that the conversational model choice was not a factor in overall performance (Table IV in Appendix E).

D. Human Resource Case Study Evaluation

We used our question-answering test set to evaluate the alignment between the text expected to be output at each stage

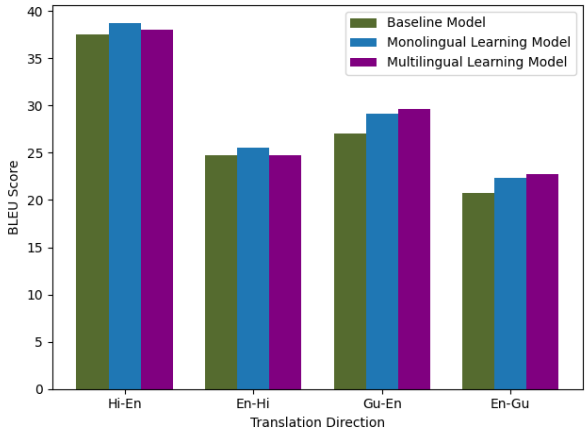


Fig. 6: Summary of Experiments. The baseline model is the pre-trained IndicBART. The Monolingual and Multilingual Learning Models are the best performing models among the monolingual and multilingual experiments respectively.

of our pipeline (Figure 1). We compare the expected textual output with the actual output using cosine similarity of the text embeddings of the inputs and outputs. Our results indicate that the system performs well in terms of response accuracy in both English and Gujarati (Table IV in Appendix C).

We also conducted a small qualitative evaluation where five native Gujarati speakers rated the system's performance on the test set on a scale of 1 to 5 using the following criteria: truthfulness (accuracy of the answers), length (conciseness and sufficiency), grammatical errors, and clarity (ease of understanding). Our results indicate that while the answers the system provided in Gujarati were correct, there were issues with clarity and grammar. Clarity was particularly impacted by the use of HR terms in Gujarati which while accurate were not in common use. Instead English words would typically be used even in the corporate setting. This indicates the need to conduct significant fine tuning specific to each use case to ensure a natural and frictionless conversational experience.

V. Conclusions

Recent advancements in NLP and LLMs have enabled the development of systems capable of generating human-like responses across a wide range of tasks but have focused primarily on high-resource languages such as English, German, and French. In this paper, we evaluate the state-of-the-art in question-answering for several low-resource Indian languages, including Hindi and Gujarati, and explore the impact of fine-tuning with conversational data, including a newly curated conversational Gujarati dataset, and an example localization use case using a small human resource question-answering dataset. In the context of both Hindi and Gujarati, we find that using multilingual approaches and fine-tuning on languages similar to the low-resource language targeted by the model can significantly improve performance.

REFERENCES

[1] A. Magueresse, V. Carles, and E. Heetderks, "Low-resource languages: A review of past work and future challenges," *CoRR*, vol. abs/2006.07264, 2020. [Online]. Available: https://arxiv.org/abs/2006.07264

[2] Cochrane, "Cochrane-evidence-different-languages," https://www.cochrane.org/news/cochrane-evidence-different-languages", Feb 2024.

[3] S. Ruder, M. E. Peters, S. Swayamdipta, and T. Wolf, "Transfer learning in natural language processing," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Tutorials*, A. Sarkar and M. Strube, Eds. Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019, pp. 15–18. [Online]. Available: https://aclanthology.org/N19-5004

[4] H. Y. Lee, "Innovations and challenges in applied linguistics from the global south," *Journal of Language and Politics*, vol. 20, Nov 2020.

[5] W. Nekoto, V. Marivate, T. Matsila, and et al., "Participatory research for low-resourced machine translation: A case study in African languages," in *Findings of the Association for Computational Linguistics: EMNLP 2020*, T. Cohn, Y. He, and Y. Liu, Eds. Association for Computational Linguistics, Nov 2020, pp. 2144–2160.

[6] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale, D. Bikel, L. Blecher, C. C. Ferrer, M. Chen, G. Cucurull, D. Esiobu, J. Fernandes, J. Fu, W. Fu, B. Fuller, C. Gao, V. Goswami, N. Goyal, A. Hartshorn, S. Hosseini, R. Hou, H. Inan, M. Kardas, V. Kerkez, M. Khabsa, I. Kloumann, A. Korenev, P. S. Koura, M.-A. Lachaux, T. Lavril, J. Lee, D. Liskovich, Y. Lu, Y. Mao, X. Martinet, T. Mihaylov, P. Mishra, I. Molybog, Y. Nie, A. Poulton, J. Reizenstein, R. Rungta, K. Saladi, A. Schelten, R. Silva, E. M. Smith, R. Subramanian, X. E. Tan, B. Tang, R. Taylor, A. Williams, J. X. Kuan, P. Xu, Z. Yan, I. Zarov, Y. Zhang, A. Fan, M. Kambadur, S. Narang, A. Rodriguez, R. Stojnic, S. Edunov, and T. Scialom, "Llama 2: Open foundation and fine-tuned chat models," https://arxiv.org/abs/2307.09288, 2023.

[7] G. Team, T. Mesnard, C. Hardin, R. Dadashi, S. Bhupatiraju, S. Pathak, L. Sifre, M. Rivière, M. S. Kale, J. Love, P. Tafti, L. Hussenot, P. G. Sessa, A. Chowdhery, A. Roberts, A. Barua, A. Botev, A. Castro-Ros, A. Slone, A. Héliou, A. Tacchetti, A. Bulanova, A. Paterson, B. Tsai, B. Shahriari, C. L. Lan, C. A. Choquette-Choo, C. Crepy, D. Cer, D. Ippolito, D. Reid, E. Buchatskaya, E. Ni, E. Noland, G. Yan, G. Tucker, G.-C. Muraru, G. Rozhdestvenskiy, H. Michalewski, I. Tenney, I. Grishchenko, J. Austin, J. Keeling, J. Labanowski, J.-B. Lespiau, J. Stanway, J. Brennan, J. Chen, J. Ferret, J. Chiu, J. Mao-Jones, K. Lee, K. Yu, K. Millican, L. L. Sjoesund, L. Lee, L. Dixon, M. Reid, M. Mikuła, M. Wirth, M. Sharman, N. Chinaev, N. Thain, O. Bachem, O. Chang, O. Wahltinez, P. Bailey, P. Michel, P. Yotov, R. Chaabouni, R. Comanescu, R. Jana, R. Anil, R. McIlroy, R. Liu, R. Mullins, S. L. Smith, S. Borgeaud, S. Girgin, S. Douglas, S. Pandya, S. Shakeri, S. De, T. Klimenko, T. Hennigan, V. Feinberg, W. Stokowiec, Y. hui Chen, Z. Ahmed, Z. Gong, T. Warkentin, L. Peran, M. Giang, C. Farabet, O. Vinyals, J. Dean, K. Kavukcuoglu, D. Hassabis, Z. Ghahramani, D. Eck, J. Barral, F. Pereira, E. Collins, A. Joulin, N. Fiedel, E. Senter, A. Andreev, and K. Kenealy, "Gemma: Open models based on gemini research and technology," 2024. [Online]. Available: https://arxiv.org/abs/2403.08295

[8] B. Zoph, D. Yuret, J. May, and K. Knight, "Transfer learning for low-resource neural machine translation," https://arxiv.org/abs/1604.02201, 2016.

[9] J. Gu, Y. Wang, Y. Chen, K. Cho, and V. O. K. Li, "Meta-learning for low-resource neural machine translation," 2018.

[10] M. Johnson, M. Schuster, and et al., "Google's multilingual neural machine translation system: Enabling zero-shot translation," https://arxiv.org/abs/1611.04558, 2017.

[11] B. Ji, Z. Zhang, X. Duan, M. Zhang, B. Chen, and W. Luo, "Cross-lingual pre-training based transfer for zero-shot neural machine translation," https://arxiv.org/abs/1912.01214, 2019.

[12] O. Bojar, V. Diatka, P. Rychlý, and et al., "HindEnCorp - Hindi-English and Hindi-only corpus for machine translation," in *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*. Reykjavik, Iceland: European Language Resources Association (ELRA), May 2014, pp. 3550–3555. [Online]. Available: http://www.lrec-conf.org/proceedings/lrec2014/pdf/835_Paper.pdf

[13] K. Revanuru, K. Turlapaty, and S. Rao, "Neural machine translation of indian languages," in *Proceedings of the 10th Annual ACM India Compute Conference*, ser. Compute '17. New York, NY, USA: Association for Computing Machinery, 2017, p. 11–20. [Online]. Available: https://doi.org/10.1145/3140107.3140111

[14] L. Shen, S. Yu, and X. Shen, "Is translation helpful? an empirical analysis of cross-lingual transfer in low-resource dialog generation," https://arxiv.org/abs/2305.12480, 2023.

[15] R. Dabre, H. Shrotriya, A. Kunchukuttan, R. Puduppully, M. Khapra, and P. Kumar, "Indicbart: A pre-trained model for indic natural language generation," in *Findings of the Association for Computational Linguistics: ACL 2022*. Association for Computational Linguistics, 2022. [Online]. Available: http://dx.doi.org/10.18653/v1/2022.findings-acl.145

[16] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds., vol. 30. Curran Associates, Inc., 2017. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

[17] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, and L. Zettlemoyer, "Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension."

[18] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," 2019.

[19] Y. Liu, J. Gu, N. Goyal, X. Li, S. Edunov, M. Ghazvininejad, M. Lewis, and L. Zettlemoyer, "Multilingual denoising pre-training for neural machine translation," 2020.

[20] L. Xue, N. Constant, A. Roberts, M. Kale, R. Al-Rfou, A. Siddhant, A. Barua, and C. Raffel, "mt5: A massively multilingual pre-trained text-to-text transformer," 2021.

[21] T. Nakazawa, H. Nakayama, C. Ding, and et al., "Overview of the 8th workshop on Asian translation," in *Proceedings of the 8th Workshop on Asian Translation (WAT2021)*, T. Nakazawa, H. Nakayama, I. Goto, and et al., Eds. Association for Computational Linguistics, Aug. 2021, pp. 1–45. [Online]. Available: https://aclanthology.org/2021.wat-1.1

[22] S. Siripragada, J. Philip, V. P. Namboodiri, and C. V. Jawahar, "A multilingual parallel corpora collection effort for Indian languages," in *Proceedings of the Twelfth Language Resources and Evaluation Conference*, N. Calzolari, F. Béchet, P. Blache, and et al., Eds. Marseille, France: European Language Resources Association, May 2020, pp. 3743–3751. [Online]. Available: https://aclanthology.org/2020.lrec-1.462

[23] G. Ramesh, S. Doddapaneni, A. Bheemaraj, and et al., "Samanantar: The largest publicly available parallel corpora collection for 11 indic languages," https://arxiv.org/abs/2104.05596, 2023.

[24] [Online]. Available: https://www.shrm.org/content/dam/en/shrm/business-solutions/SHRM-Sample-Employee-Handbook-2023.docx

[25] P. Lison and J. Tiedemann, "OpenSubtitles2016: Extracting large parallel corpora from movie and TV subtitles," in *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*. Portorož, Slovenia: European Language Resources Association (ELRA), May 2016, pp. 923–929. [Online]. Available: https://aclanthology.org/L16-1147

[26] J. Gala, P. A. Chitale, A. K. Raghavan, and et al., "Indictrans2: Towards high-quality and accessible machine translation models for all 22 scheduled indian languages," *Transactions on Machine Learning Research*, 2023. [Online]. Available: https://openreview.net/forum?id=vfT4YuzAYA

[27] B. Haddow and F. Kirefu, "Pmindia – a collection of parallel corpora of languages of india," https://arxiv.org/abs/2001.09907, 2020.

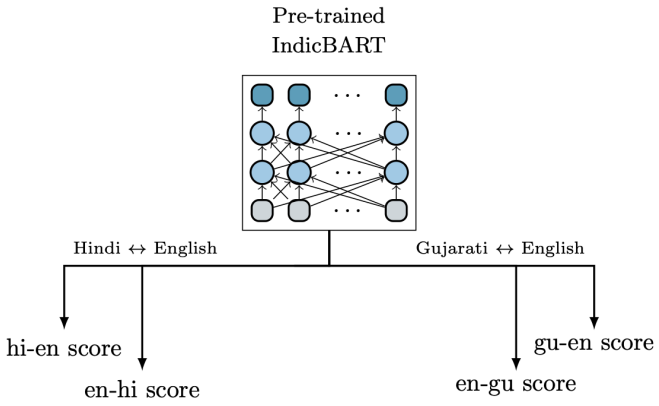
[28] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: a method for automatic evaluation of machine translation," in *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ser. ACL '02. USA: Association for Computational Linguistics, 2002, p. 311–318. [Online]. Available: https://doi.org/10.3115/1073083.1073135

[29] K. Blagec, G. Dorffner, M. Moradi, S. Ott, and M. Samwald, "A global analysis of metrics used for measuring performance in natural language processing," https://arxiv.org/abs/2204.11574, 2022.

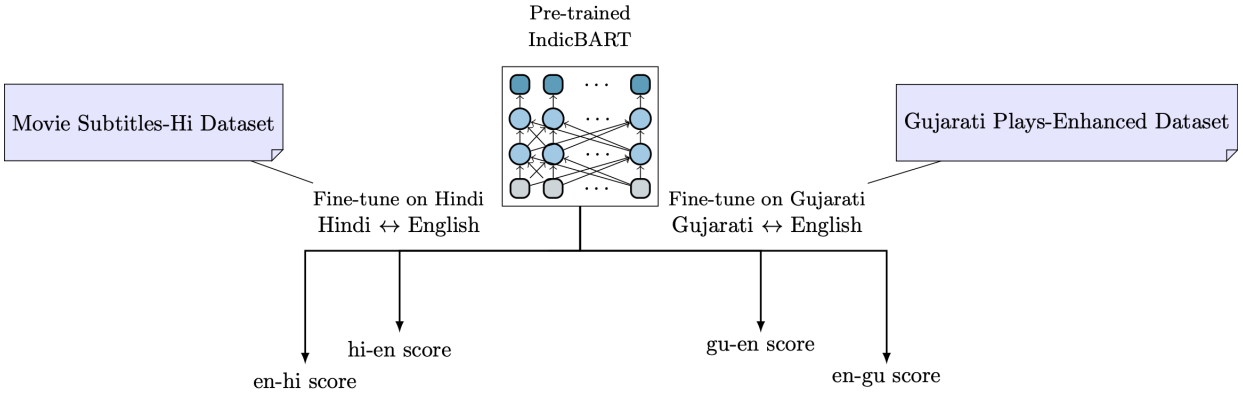
APPENDIX

A. Architecture

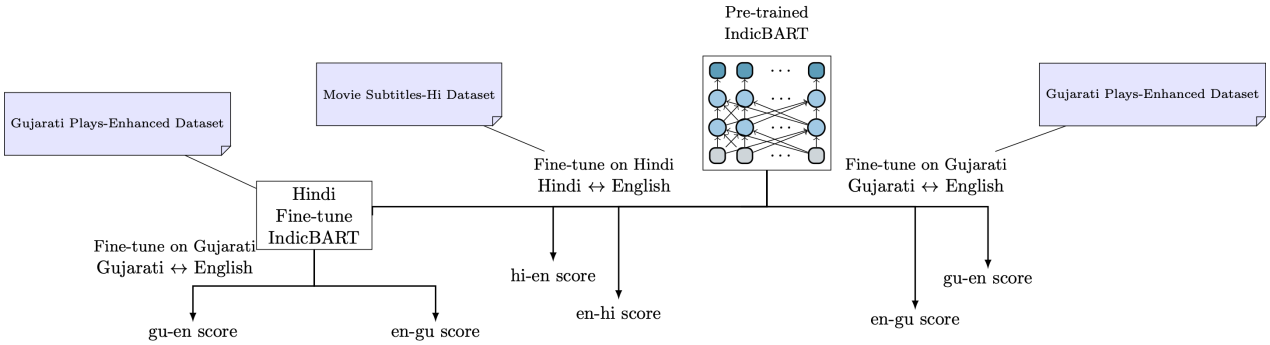
In this section, we visualize the experimental pipelines for our monolingual experiments. The arrows indicate the steps required to execute a given experiments. For example, in Figure 7a we compute the en-hi score by fine-tuning IndicBART on a Hindi dataset.



(a) Monolingual Experiment 1: Out-of-the-box performance



(b) Monolingual Experiment 2: Impact of fine-tuning with low-resource language data



(c) Monolingual Experiment 3: Impact of sequential fine-tuning with multiple low-resource languages

Fig. 7: Visual Representation of Monolingual Experiments.

B. Summary of Datasets

1) Movie Subtitles Dataset: We use a subset of the Opensubtitles [25] dataset, which is a large multilingual parallel corpus of Movie and TV subtitles, and a common benchmark for conversational model training and experiments. The sentences we harvest are largely from subtitles of movies released after 2020 as older movies tend to not have subtitles the languages we are focusing on.

Overall, we harvested 18,000 parallel sentence pairs in English (en), Hindi (hi), Bengali (bn), Marathi (mr). We refer to this dataset as Movie Subtitles-Hi-Bn-Mr. We also construct a dataset we call Movie Subtitles-Hi, which subsumes the English and Hindi content from Movie Subtitles-Hi-Bn-Mr and also includes an additional 17,000 parallel sentence pairs in English and Hindi.

2) In22-Conv Dataset: To test the performance of our models on Gujarati and Hindi, we used the IN22-Conv dataset [26], distributed by AI4Bharat. This dataset is a subset of the broader IN22 collection focused on conversational content. It encompasses 1,503 sentence pairs across 22 Indic languages including the languages we are interested in, e.g., English, Hindi, and Gujarati, and is designed for benchmarking machine translation systems. The dataset covers a wide spectrum of topics, including hobbies, daily life conversations, government affairs, among others, and is a versatile resource for assessing translation quality in everyday conversational applications.

Due to the scarcity of Gujarati conversational data, we used a portion of IN22-Conv (1,000 sentence pairs in English and Gujarati) as part of our training data for Gujarati (in combination with the Gujarati Plays dataset which we describe in Section B3), and 500 sentence pairs for testing (which we refer to as IN22-Conv-500 in Table I).

3) Gujarati Plays Dataset: We curated a dataset we call 'Gujarati Plays' from Gujarati plays we obtained from a school district in Gujarat. This dataset consists of 23 school plays, from which were able to harvest 2,000 pairs of sentences. These plays were written by school teachers to be performed by students on various Indian festivals.

We combine the 'Gujarati Plays' dataset with 1,000 sentence pairs from the IN22-Conv dataset (Section B2) to fine-tune our models on Gujarati conversational data, which we detail in Section III-D.

4) PMO India: The PMO India dataset [27], made available by University of Edinburgh researchers, consists of approximately 56,000 sentences spread across 5,000 documents from the Prime Minister's Office of India. Topics include foreign trade, technology investments, and other government-related topics.

We use the PMO India dataset to manually curate 2,100 question-answer pairs, of which we use 2,000 to few-shot train the Llama2 and Gemma models (Section III-B), and 100 to test the end-to-end performance of our system in the context of both Hindi and Gujarati.

Datasets we curated for this work are available here: https://www.kaggle.com/datasets/hvtrivedi/hr-case-study/.

TABLE I: Summary of Datasets

Dataset	**Data Size**	**Languages**	**Notes**
Movie Subtitles-Hi training set	35,000 pairs	hi,en	17,000 are only in hi,en, the rest are available in bn and mr as well (see below)
Movie Subtitles-Hi-Bn-Mr training set	18,000 pairs	hi,bn,mr en	Overlaps Movie Subtitles-Hi for hi, en
Gujarati Plays-Basic training set	2,000 pairs	en-gu	**Newly curated.**
IN22-Conv-1500 test set	1,503 pairs	en, gu, hi, bn, mr	Dataset used for testing of hi-en and en-hi performance
IN22-Conv-500 test set	500 pairs	en, gu	Part of IN22-Conv-1500, used for testing gu performance
Gujarati-Plays-Enhanced training set	3000 pairs	en, gu	Combination of Gujarati Plays and the remainder of In-22-Conv-1500 after removing In-22-Conv-500 which is used for testing.
Conversational examples training set	2000 sentences	en, gu	**Curated from PMO India dataset [27].**
Conversational examples test set	100 sentences	en, gu	**Curated from PMO India dataset [27].**

C. HR Case Study Data and Results

TABLE II: HR Case Study Test Set Questions and Answers.

Question	Answer
What is the company's policy on employment at will?	The company follows an at-will employment policy, meaning either the employee or the company can terminate employment at any time, for any reason, with or without notice.
How does the company support equal employment opportunities?	The company provides equal employment opportunities to all employees and applicants without regard to protected characteristics like race, color, religion, etc., and strictly prohibits any form of unlawful discrimination or harassment.
What is the process for requesting a reasonable accommodation under the ADA?	Employees in need of a reasonable accommodation should contact the HR department with a specific request and may need to provide supporting medical documentation, as per company policy and ADA guidelines.
How does the company handle conflicts of interest?	Employees must avoid any real or potential conflicts of interest with the company and discuss any potential conflicts with a manager for guidance on how to proceed.
What are the guidelines for maintaining confidentiality of company information?	Employees are required to protect confidential business information and trade secrets, with disciplinary action for misuse or unauthorized disclosure.
What is the company's policy on workplace safety and violence?	The company maintains a violence-free workplace, prohibiting any threats or acts of aggression and implementing strict measures to ensure safety, including a zero-tolerance policy for weapons on company property.
How does the company handle drug and alcohol use in the workplace?	The company enforces a drug-free and alcohol-free workplace policy, prohibiting the use or influence of illegal drugs and alcohol during work hours or on company premises.
What are the rules regarding social media use?	Employees are allowed to engage in social media but must respect confidentiality and the professional image of the company, with disclaimers required when discussing company-related matters.
Can employees access their personnel files?	Employees may inspect their own personnel files upon written request to HR and can copy documents but are not allowed to remove them.
What is the policy on employment of relatives and domestic partners?	Relatives and domestic partners may be hired as long as they do not work in a direct supervisory relationship and their employment does not affect supervision, security, safety, or morale.
How is overtime compensated?	Nonexempt employees are paid one and a half times their regular rate for hours worked over 40 in a workweek, with all overtime needing prior approval by a supervisor.
What is the company's stance on employee privacy and surveillance?	The company respects employee privacy but maintains the right to conduct investigations and monitor company systems and communications to ensure security and compliance with policies.
What benefits are available to employees?	The company offers comprehensive benefits including medical, dental, and vision insurance, life insurance, disability coverage, and a 401(k) plan, with eligibility criteria detailed in the employee handbook.
What are the procedures for reporting harassment or discrimination?	Employees should report any incidents to their supervisor, HR, or any designated official as per the company's harassment and complaint procedures, with protections against retaliation.
What are the guidelines for taking leave, including FMLA and other types of leave?	The company provides various types of leave, including FMLA, bereavement, and military leave, with specific guidelines on eligibility and usage detailed in the employee handbook.

TABLE III: HR Case Study Cosine Similarity Results

Models	**Cosine Similarity of gu $\rightarrow$ en translation**	**Cosine Similarity of LLM Responses to Ground Truth**	**Cosine Similarity of en $\rightarrow$ gu translations**	**Cosine Similarity of Gujarati responses to ground truth**
LLAMA2-chat	0.612	0.631	0.573	0.581
Gemma-2b	0.612	0.628	0.569	0.579
Gemma - 7b	0.612	0.634	0.577	0.581

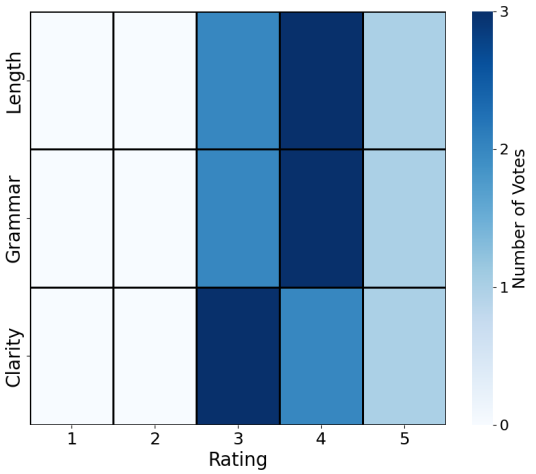


Fig. 8: HR Case Study Human Evaluation Results.

D. Evaluation Metric

Initially introduced in 2002, the BLEU (Bilingual Evaluation Understudy) metric is one of the standard metrics for evaluating the quality of the machine-translated text [28]. BLEU assesses the quality of machine-generated translations by comparing them with one or more reference translations, focusing primarily on the precision of n-grams which are consecutive sequences of n words. BLEU involves calculating the precision of n-grams in the translated text relative to their appearance in the reference texts. The BLEU score formula is expressed as:

$$\text{BLEU} = BP \cdot \exp\left(\sum_{n=1}^{N} w_n \log p_n\right)$$

Where:

- p_n is the precision of n-grams (the ratio of the number of n-grams in the candidate translation that match those in the reference translation to the total number of n-grams in the candidate translation).
- w_n are weights assigned to each n-gram (often uniform).
- BP (brevity penalty) discourages overly short translations and is defined as:

$$BP = \begin{cases} 1 & \text{if } c > r \\ e^{(1-r/c)} & \text{if } c \leq r \end{cases}$$

- c is the length of the candidate translation, and r is the effective reference corpus length.

BLEU is commonly used for translation tasks, is relatively straightforward to compute and BLEU scores have been shown to be consistent with human judgment [29].

E. Ablation Results

TABLE IV: Conversational LLM Ablation Results

Models	**Cosine Similarity of gu → en translation**	**Cosine Similarity of LLM Responses to Ground Truth**	**Cosine Similarity of en → gu translations**	**Cosine Similarity of Gujarati responses to ground truth**
LLAMA2-chat	0.606	0.656	0.566	0.582
Gemma-2b	0.606	0.652	0.561	0.579
Gemma -7b	0.606	0.656	0.568	0.585

F. Micro-Scale Human Evaluation

Questions	Translations by Monolingual Learning	Translations by Multilingual Learning
What benefits will the MoU between India and Mongolia for border security bring?	ભારત અને મંગોલિયાનું સરહદ સુરક્ષા સમજૂતી કરાર કેવા લાભો આપશે?	ભારત અને મંગોલિયા વચ્ચેની સરહદ સુરક્ષા માટેની સમજૂતી કરાર શું લાભ લાવશે?
What is the theme of the India-Canada joint postage stamps?	ભારત અને કેનેડાની સાથે જારી કરાયેલી ટપાલ ટિકિટનું વિષય શું છે?	ભારત-કેનેડા સંયુક્ત પોસ્ટેજ ટપાલ ટિકિટનું મુખ્ય વિષય શું છે?
What's the main goal of the customs agreement between India and Uruguay?	ભારત અને ઉરુગ્વે વચ્ચેના કસ્ટમ મુદ્દાઓના કરારનું મુખ્ય લક્ષ્ય શું છે?	ભારત અને ઉરુગ્વે વચ્ચેના કસ્ટમ કરારનો મુખ્ય ઉદ્દેશ્ય શું છે?

Fig. 9: Translation examples from among the 100 used for micro-scale human evaluation.

In order to get some insights into the types of mistakes the end-to-end conversational system is prone to, we conducted a small-scale human evaluation of the baseline IndicBART model, the best monolingual transfer learning model and the best multilingual meta-learning models from our experiments so far.

Four native speakers reviewed the test data (Conversational Examples Test Set described in Table I) which consists of 100 examples of Gujarati-based questions. All evaluators viewed the model output favorable, indicating strong foundational capabilities across all systems.

Despite the overall positive feedback, a number of errors made by the models highlighted some challenges in language translation tasks for Gujarati. One notable issue was the models' occasional misuse of words that are contextually relevant but do not support the intended meaning. For example, in instance 3 , our model used a word that, while related to "agreement" in a broader context, actually meant "issues" in a literal sense.

Another error involved the use of pronouns. Gujarati, like many other languages, assigns gender to pronouns that may also extend to inanimate objects, often reflecting grammatical gender rather than biological gender. The models incorrectly assigned the feminine pronoun "she" to objects where it was not linguistically appropriate, e.g., for instance 1.

These findings from the human evaluators highlight both the progress and the limitations of current translation models in handling the intricacies of language translation, particularly in linguistically diverse contexts.

We show several examples of translated text from among those shown to the human evaluators on Figure 9.