AI-Tutor: Interactive Learning of Ancient Knowledge from Low-Resource Languages

Siddhartha Dalal[1] **Rahul Aditya**[1] **Vethavikashini Chithrra Raghuram**[1]
Prahlad Koratamaddi[1]
[1]Columbia University
sd2803@columbia.edu

Abstract

Many low-resource languages, such as Prakrit, present significant linguistic complexities and have limited modern-day resources. These languages often have multiple derivatives; for example, Prakrit, a language in use by masses around 2500 years ago for 500 years, includes Pali and Gandhari, which encompass a vast body of Buddhist literature, as well as Ardhamagadhi, rich in Jain literature. Despite these challenges, these languages are invaluable for their historical, religious, and cultural insights needed by non-language experts and others.

To explore and understand the deep knowledge within these ancient texts for non-language experts, we propose a novel approach: translating multiple dialects of the parent language into a contemporary language and then enabling them to interact with the system in their native language, including English, Hindi, French and German, through a question-and-answer interface built on Large Language Models. We demonstrate the effectiveness of this novel AI-Tutor system by focusing on Ardhamagadhi and Pali.

1 Introduction

Much of the world's ancient cultural heritage is preserved in Low Resource Languages (LRLs) from the past. However, access to this knowledge is limited to a small group of linguistic scholars. For instance, Prakrit, an ancient language widely spoken in India about 2,500 years ago, flourished for approximately 500 years. Various forms of Prakrit, such as Ardhamagadhi, Pali and Gandhari, were used across different regions in India, Pakistan and Afghanistan. A significant portion of Buddhist and Jain historical, cultural, and religious texts are written in these languages. Unfortunately, experts in Prakrit are few in number.

To address the need for wider access to this vast ancient body of literature for non-linguistic experts, this paper argues that Neural Machine Translation (NMT) can unlock the knowledge contained within these texts. Given the archaic nature of the original language and the scarcity of scholars available to provide explanations, a question-and-answer format, powered by Large Language Models (LLMs), is essential for making this knowledge accessible to non-experts.

We demonstrate that these goals can be achieved by developing a Transformer-based Neural Machine Translation system and leveraging LLMs to facilitate interactions in a query-response format. This approach enables non-experts to engage with the content of ancient texts quickly and effectively. Specifically, we showcase this paradigm using Jain and Buddhist literature written in Ardhamagadhi and Pali, creating an AI-powered tutoring system that allows users to interact with an AI-Tutor in their own language, including English, Hindi, French, and German. The translation training data was sourced from original languages translated in Hindi and English.

2 Related Work

2.1 Low Resource NMT

Neural Machine Translation (NMT) has seen significant advancements, notably with the introduction of attention mechanisms, which allow models to selectively focus on relevant parts of the source sentence (Luong, 2015). As NMT evolved, researchers addressed challenges such as handling large vocabularies (Jean et al., 2014), translating rare words (Luong et al., 2014), and leveraging source-side

Proceedings of the Eleventh Workshop on Asian Translation (WAT 2024), pages 56–66
November 16, 2024

monolingual data through self-learning and multi-task learning (Zhang and Zong, 2016). Efforts to optimize vocabulary sizes (Gowda and May, 2020) and tackle open-vocabulary challenges by encoding rare words as subword units (Sennrich, 2015) further improved translation quality. The development of massively multilingual systems marked a significant leap, enabling translation across over 100 languages (Aharoni et al., 2019).

However, despite these advances, NMT faces persistent challenges in low-resource settings, particularly for underrepresented languages like those in the Indic family. Low-resource Machine Translation (MT) struggles due to the scarcity of parallel corpora, which traditional NMT models heavily rely on. Transformer-based models have become prominent in addressing issues like word ordering and data sparsity in these settings (Vaswani, 2017). Noteworthy efforts in Indic MT have laid the groundwork for subsequent multilingual and pre-trained MT models (Philip et al., 2021; Ramesh et al., 2022; Kudugunta et al., 2019; Liu, 2020).

A key strategy to overcome low-resource challenges is **transfer learning and Fine-Tuning**, where a model trained on a high-resource language pair (parent model) is adapted to a low-resource pair (child model). This approach has shown significant improvements, particularly when the parent and child languages are linguistically similar (Zoph et al., 2016). Enhancements to transfer learning include **ensembling techniques** and **unknown word replacement**, which further boost translation quality (Zoph et al., 2016). Moreover, large multilingual and pre-trained models have been effective in leveraging linguistic similarities across languages, facilitating good-quality MT for low-resource languages (Dabre et al., 2020).

Recent work has also focused on translating extremely low-resource languages with minimal parallel and monolingual corpora (Maurya et al., 2023). Despite these advances, challenges remain, particularly in selecting the most effective parent language for transfer learning. As research continues, these innovations are expected to further bridge the gap between high- and low-resource languages, making NMT more accessible to diverse linguistic communities.

2.2 Multilingual RAG

Retrieval-Augmented Generation (RAG) (Lewis et al., 2020; Jiang et al., 2023; Gao et al., 2023) has emerged as a powerful method to enhance the factual accuracy and domain specificity of large language models by combining retrieval mechanisms with generative capabilities. While RAG has shown significant promise, much of the research has been centered around English-language applications.

Addressing this gap, (Ahmad, 2024) explores deploying RAG models in multilingual, multicultural corporate settings, focusing on strategies to mitigate challenges like hallucinations and optimize information delivery. Similarly in (Chirkova et al., 2024), the authors investigate mRAG models across 13 languages, emphasizing the need for task-specific prompt engineering and highlighting issues such as code-switching and fluency errors. These studies collectively advance the adaptation of RAG systems for diverse linguistic environments.

3 Datasets

To build and evaluate our models, we relied on several key datasets that encompass a diverse range of linguistic materials. These datasets were selected to ensure comprehensive coverage and accuracy in our training and evaluation processes. Below, we detail the datasets used, highlighting their sources and the challenges associated with obtaining and preparing the data:

1. **Pali Dataset**: The primary dataset for Pali training was sourced from Huggingface. It includes 148,813 sentences featuring Latinized transliteration and their corresponding English translations, provided by the Berkeley AI Research Lab.[1]

2. **Ardhamagadhi Dataset**: Training data for Ardhamagadhi posed greater challenges. The dataset comprises 44 Agams translated into Hindi by Muni Deepratnasagar, from which 10,113 parallel sentences were extracted.[2] Agams are the sacred texts of the

[1] Dataset available at: https://huggingface.co/buddhist-nlp.

[2] Translations by Muni Deepratnasagar available at: https://jainelibrary.org/.

Jain religion written in Ardhamagadhi and contain religious stories, art, literature, and poetry. These translations often included additional meanings, increasing the complexity of the neural machine translation (NMT) task.

3. **Ardhamagadhi Books**: In addition, five Ardhamagadhi books translated into Hindi and English by Professor V. K. Jain[3] were used. These books, originally provided in PDF format, were digitized using OCR technology developed by Mr. Kailash Mutha. Some scanned copies were obtained from an additional archive, yielding 910 high-quality samples.[4]

4. **11 Agams**: A subset of 11 Agams were available in English [5] which were processed by AI-Tutor directly without the translation pipeline.[6]

3.1 Data Preprocessing

To prepare the datasets for training, testing and fine-tuning, various sources of Prakrit sentences and their translations in Hindi and English were gathered and processed. The datasets comprised numerous code-mixed sentences in Hindi, English, and Prakrit, along with supplementary text such as verse explanations. Creating a clean and reliable corpus for training the neural machine translation (NMT) system required significant domain knowledge. This challenge was addressed through pattern recognition and efficient information retrieval methods. Despite these approaches, manual extraction, consultation with domain experts and verification were essential due to the complexities and exceptions inherent in the data, ensuring the accuracy and quality of the dataset.

4 Neural Machine Translation

4.1 NMT Base Model: Indictrans2

Indictrans2 (Gala et al., 2023) is a cutting-edge multilingual neural machine translation (NMT) model specifically designed for 22 current Indic languages. With its Transformer Encoder-Decoder architecture with multiple layers of self-attention and parallel processing, it can capture long-range dependencies in text, making it both effective and efficient for large-scale translation tasks for Indic lanaguages. Though many of languages covered by Indictrans2 are ultimately derived over thousands of years from Prakrit, it does not include Pali, Ardhamagadhi and other ancient Prakrit languages.

Indictrans2 is pre-trained on a vast multilingual corpus that includes text from various Indic languages paired with English and other target languages. A key strength of Indictrans2 is its ability to manage multiple languages simultaneously through a shared vocabulary and embeddings. By training on a diverse set of Indic languages, Indictrans2 learns cross-lingual representations, which are particularly beneficial for low-resource languages like Prakrit. In the context of Indic languages, pre-training on a multilingual corpus is especially advantageous due to the linguistic similarities and shared syntactic structures among these languages. For instance, many Indic languages share common grammatical features, such as subject-object-verb (SOV) word order and similar morphological patterns. Pali and Ardha-Magadhi are Middle Indo-Aryan languages that derive from Prakrit and are closely related to Sanskrit but are not directly descended from it. Sanskrit is a standardized dialect of Old Indo-Aryan. They all use variations of Devnagari script. Leveraging these linguistic similarities, we hypothesized that Prakrit's similarity to Sanskrit would enable us to initialize the encoder with pre-trained Sanskrit embeddings during fine-tuning.

4.1.1 Challenges and Limitations for using Indictrans2

Despite its impressive capabilities, Indictrans2 also faces certain challenges, particularly when dealing with very low-resource languages or dialects. For example, Prakrit, being an ancient and less standardized language, poses difficulties in terms of data availability and consistency. Additionally, cultural and contextual nuances that are specific to certain regions or time periods may not always be captured ac-

[3] Dravyasamgraha, Niyam Sara, Pravachansara, Samayasara, Panchastikay Sangraha. All from https://jainelibrary.org/.

[4] Additional scanned copies sourced from https://jainqq.org.

[5] Deepratnasagar Muni, https://jainelibrary.org/

[6] Jain era documents available at: https://jainelibrary.org/.

curately by the model.

4.2 Fine-tuning Approach

We utilized two variants of Indictrans2: Indic-Indic and Indic-En. The choice was driven by its ability to perform well in pairs of low-resource languages, making it suitable for translating Prakrit into Hindi and Prakrit into English. For fine-tuning with the 44 Agams dataset containing Prakrit-Hindi pairs, we employed the Indic-Indic variant, using Sanskrit embeddings as a means for transfer learning for Prakrit. Figure 1 illustrates the workflow for selecting variants of the pretrained Indictrans2 models used for finetuning on various datasets. The Indic-En variant was used for finetuning with the VK Jain dataset and the Pali-English dataset. Given the limited availability of Prakrit-English data, we employed a multistep fine-tuning process:

- The model was initially fine-tuned on the Pali-English dataset, utilizing Sanskrit embeddings for transfer learning, given the linguistic similarities between Pali and Sanskrit as both belong to the Prakrit family.

- In the next step, we use the resulting model checkpoint from the last step, now enriched with Pali-specific representations, and fine-tune on the VK Jain dataset to adapt it for Ardhamagadhi-English translation.

This progressive fine-tuning helped in gradually adapting the model to the nuances of Prakrit while maintaining the quality of English translations.

4.3 Results

Table 1 summarizes the experimental results across various datasets and fine-tuning approaches for the corresponding test splits. In all the experiments, we used train/dev/test split ratio of 80/10/10. Sample translations from the test sets of the Pali-English, VK Jain, and 44 Agams datasets are shown in Tables 2, 3, and 4, respectively. We achieved a strong BLEU score of 33.8 for Pali-to-English translations, benefiting from the large dataset. For Ardhamagadhi-to-English (BLEU score: 17.8) and Ardhamagadhi-to-Hindi (BLEU score: 14.9), the scores, although lower due to smaller datasets, are competitive with other low-resource languages such as Cherokee, where the BLEU score is approximately 14. (Zhang et al., 2020).

4.3.1 Pali-English Indic-En Pali-English

When translating Pali into English using the Pali-English dataset, the Indictrans2 model achieved a BLEU score of 33.8, chrF score of 54.7, and chrF2++ score of 52.6. This indicates strong performance and relatively high translation quality for the Pali-English language pair.

Table 2 shows a couple of sample translations on the Pali-English dataset.

4.3.2 Ardhamagadhi-Hindi Indic-Hin 44 Agams

When translating Ardhamagadhi into Hindi using the 44 Agams dataset, the fine-tuned Indictrans2 model achieved a BLEU score of 14.9, a chrF score of 34.3, and a chrF2++ score of 32.8. While these scores indicate modest performance, they are consistent with results seen in other low-resource languages, such as Cherokee. Despite the linguistic differences, the similar score range highlights the effectiveness of the model in handling low-resource languages, demonstrating its potential in such challenging translation tasks. (Zhang et al., 2020)

Table 3 shows sample translations for VK Jain dataset (Ardhamagadhi-English).

4.3.3 Ardhamagadhi-English Indic-En VK Jain

Fine-tuning on the VK Jain data set yielded a BLEU score of 17.8, a chrF score of 39.1 and a chrF2 ++ score of 37.2, underscoring the effectiveness of domain-specific data in improving translation quality. Notably, this performance was achieved despite the dataset's small size (approximately 900 lines), due to the robust pretraining of Indictrans2 and careful data preparation. Additionally, the use of a larger Pali-English dataset allowed the model to develop a coarse Ardhamagadhi embedding, taking advantage of the linguistic similarities between Ardhamagadhi and Pali. Finetuning the checkpoint of the Indic-En model trained

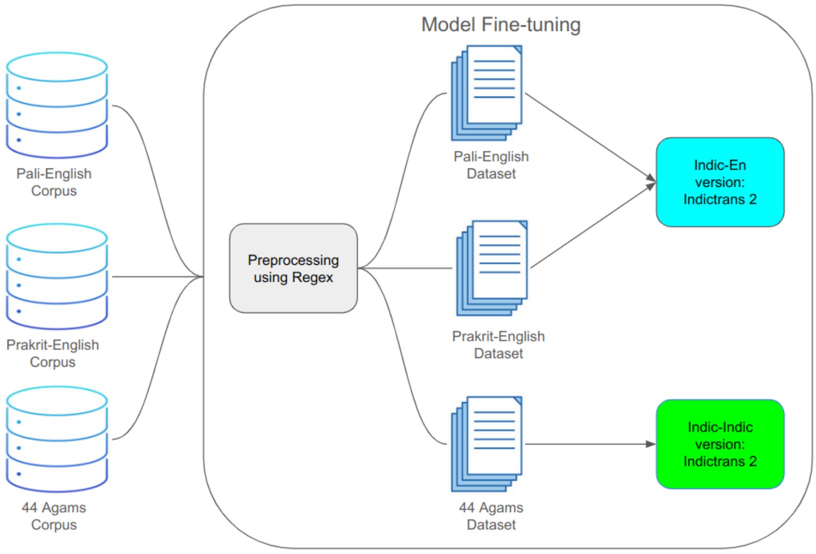

Figure 1: NMT system: Workflow showing parallel dataset extraction (preprocessing) and training two versions of Indictrans2 model during the Finetuning step depending on source and target language.

Source Lang.	Target Lang.	Dataset	BLEU	chrF	chrF2++
Pali	English	Pali-English	33.8	54.7	52.6
Ardhamagadhi	Hindi	44 Agams	14.9	34.3	32.8
Ardhamagadhi	English	VK Jain	17.8	39.1	37.2

Table 1: Results of Prakrit to English/Hindi Translations using Indictrans2 on corresponding test splits

Source: Pali	Reference: English	Translation: English
ऐवमेतेसं पञ्चन्नं खन्धानं समवाये आलोकनविलोकनं पञ्ञायति।	Thus looking-straight-on-and-looking-away-from-the-front is seen in the combination of these five aggregates.	In this way looking straight on and looking away from the front is seen in the combination of these five aggregates.
अप्पका ते मनुस्सेसु, ये जना पारगामिनो;	Few among men are those who cross to the farther shore.	Few are those among men who cross to the further shore.

Table 2: Example translations from Pali to English using Indictrans2 on Pali-English dataset.

Pali-English dataset on the VK Jain dataset (Prakrit-English) led to BLEU score of 17.8.

Table 4 shows sample translations for 44 Agams dataset (Ardhamagadhi-Hindi).

We also observed a decrease in translation quality as the reference translation length increased. Since Indictrans2 has a maximum context length of 256 tokens, sentences longer than this were skipped during both fine-tuning and inference. The BLEU scores reported in the Results section reflect the subset of translations that fit within this 256-token context length. Similarly, for the 44 Agams dataset, almost 30% of sentences were longer than 256 tokens and had to be discarded.

[7]Obtained by Google Translate Hindi to English:https://translate.google.com/?sl=hi&tl=en

5 AI-Tutor System

The AI-Tutor System seamlessly combines the translation system with a multilingual query-response system. It is designed to facilitate multilingual interaction, allowing users to submit queries in English, Hindi, German, or French, and receive responses in their preferred language.

Source:Ardhamagadhi	Reference: English	Translation: English
जीवस्स जे गुणा केई णत्थि ते खलु परेसु दव्वेसु। तम्हा सम्मादिट्ठिस्स णत्थि रागो दु विसएसु।।	The attributes of the soul do not exist in alien substances; therefore, the right believer has no attachment for the sense-objects.	The attributes of the soul do not exist in alien substances; therefore, the right believer has no attachment for the sense-objects.
ववहारणओ भासदि जीवो देहो य हवदि खलु ऍक्को।	The empirical point of view indeed holds that the soul and the body are the same, however, from the transcendental point of view the soul and the body are never the same.	From the empirical point of view, the soul and the body are one.

Table 3: Example translations from Ardhamagadhi to English using Indictrans2 on VK Jain dataset.

Source: Ardhamagadhi	Reference: Hindi	Translation: Hindi
पणया वीरा महावीहिं।	वीर पुरुष महापथ के प्रति प्रणत – अर्थात् समर्पित होते हैं। (Google Translate[7]: Brave men are devoted to the high path.)	वीर पुरुष महापथ के प्रति प्रणत-अर्थात् समर्पित होते हैं। (Google Translate: Brave men are devoted to the high path.)
परतित्थिय-गह-पह-नासगस्स तवतेय-दित्तलेसस्स ।नाणुज्जोयस्स जए, भद्दं दमसंघसूरस्स ॥	एकान्तवादी, दुर्नयी परवादी रूप ग्रहाभा को निस्तेज करनेवाले, तप तेज से सदैव देदीप्यमान, सम्यग्ज्ञान से उजागर, उपशम – प्रधान संघ रूप सूर्य का कल्याण हो। (Google Translate: May the Sun of the Sangha form, who is a solitary person, who dims the aura of the enemy and is always shining with the brilliance of austerity, who is illuminated by right knowledge and who is the main one in the form of calmness, be blessed.)	एकान्तवादी, दुर्नयी परवादी रूप ग्रहाभा को निस्तेज करनेवाले, तप तेज से सदैव देदीप्यमान, सम्यग्ज्ञान से उजागर, निग्रह-संघ रूप सूर्य का कल्याण हो। (Google Translate: May the Sun, who is a man of solitary nature and who dims the aura of the world, who is always shining with the brilliance of austerity and who is illuminated by right knowledge and who is in the form of the union of restraint, prosper.)

Table 4: Example translations from Ardhamagadhi to Hindi on 44 Agams dataset.

5.1 Workflow

The workflow of the AI-Tutor System encompasses several key stages, beginning with query submission and ending with response delivery. Figure 2 demonstrates the pipeline. Initially, the user submits a query in one of the supported languages: English, Hindi, German, or French. For non-English queries, the system employs a combination of two translation systems: the fine-tuned IndicTrans2 model and the Argos Translate library[8] to translate the input into English, ensuring uniform processing of all queries.

Following translation, the system retrieves relevant content from a curated collection of Prakrit texts that have been translated into English. To facilitate efficient retrieval, embeddings of these texts are generated using the Sentence Transformers MiniLM model[9] from Hugging Face. This compact version of the Sentence Transformer model is optimized for efficiency while maintaining a good level of semantic accuracy. These embeddings

[8] https://github.com/argosopentech/argos-translate

[9] sentence-transformers/all-MiniLM-L6-v2

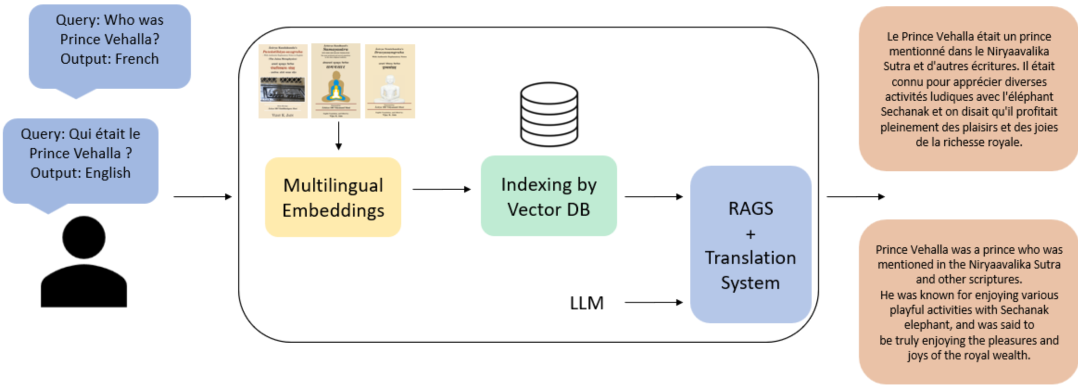

Figure 2: AI-Tutor System for Prakrit Texts: Workflow illustrating query processing, embedding generation, vector database indexing, and language-adaptive response generation using RAG and translation.

are then stored in a FAISS (Facebook AI Similarity Search) vector database, enabling rapid and accurate retrieval based on the query. For response generation, the system implements a Retrieval-Augmented Generation (RAG) architecture, utilizing the Llama 3 model[10] within the LangChain framework. The retrieved embeddings serve as crucial context, enabling the generation of coherent and contextually relevant responses.

In cases where the user requests a response in Hindi, German, or French, the system translates the generated English response back into the desired language using both IndicTrans2 and Argos Translate. Finally, the system delivers the response to the user in their chosen language. This comprehensive workflow ensures a seamless and robust experience, enabling users to interact with ancient Prakrit texts through a modern, multilingual interface. Table 5 displays an example query-response pair from the 44 agams dataset. More results are present in tables 6 and 7 in Appendix A showcase the example queries passed to the AI-Tutor System along with the responses for the three datasets. The result is a valuable tool that bridges linguistic gaps and facilitates research and education within the Indic context.

We evaluated the AI-Tutor's performance using GPT-4, focusing on faithfulness and relevancy. Both metrics scored 0.9 of 1.0 across 10 queries, indicating the AI-Tutor provides accurate and contextually relevant responses, proving its effectiveness as a learning tool.

[10] https://huggingface.co/meta-llama/Meta-Llama-3-8B

6 Conclusions and Future Work

We have developed AI-Tutor, a system designed to enable non-language experts to explore extensive literature in two low-resource languages (LRLs), Ardhamagadhi and Pali, by providing translations powered by Neural Machine Translation (NMT) and Retrieval-Augmented Generation (RAG). The limited availability of training data has been a significant challenge. Future efforts to generate synthetic data and incorporate additional training data from related languages, such as Gandhari, are expected to enhance translation quality. These languages, despite their differences, share structural, lineage, and historical similarities, which could be leveraged to improve the system's performance.

In future work, we plan to test our finetuning approach on Indictrans2 using embeddings from other Indic languages like Hindi, Gujarati, Tamil, and Telugu. These languages may be semantically closer to Ardhamagadhi than Sanskrit, potentially yielding better performance in translation tasks. Additionally, we aim to systematically compare these variations to gain deeper insights into how different embeddings influence translation quality.

7 Limitations

Translation work in this paper required converting pdf files to unicode text using OCR, which is not a full-proof process. We have used better system provided by Mr. Mutua to ameliorate it. Further, chunks longer than 256 tokens were not processed. We plan to deal with it by intelligent chunking of those unpro-

Query	Response			
	English	Hindi	French	German
How many types of devlok are there?	There are four types of devlok: Bhavanwasi, Vaanyaantar, Astrologer, and Wamanik.	देवलोक चार प्रकार का होता है: भवनवासी, वन्यान्तर, ज्योतिषी और वैमानिक।	Il existe quatre types de devlok : Bhavanwasi, Vaanyaantar, Astrologue et Wamanik.	Es gibt vier Arten von Devlok: Bhavanwasi, Vaanyaantar, Astrologe und Wamanik.

Table 5: Example Query from the 44 Agams Dataset passed to the AI-Tutor System along with the response in the supported languages: English, Hindi, French, and German.

cessed part. Further, the Ardhamagadhi translation quality needs be improved by acquiring more data and we are collaborating with linguist and domain experts to acquire more data. Finally, the LLM technology underlying the Query-Answer system can hallucinate. As the anti-hallucination technology evolves, we need to be able to incorporate better LLM models.

8 Ethics

We do not expect any negative social impact from our work. We sincerely hope that our work will further research for acquiring ancient historical, cultural and religious knowledge from LRL like Gandhari and other genres of Prarkrit languages used by millions two millenniums ago.

9 Acknowledgments

We gratefully acknowledge the domain expertise provided by Kailash Mutha, Dr. Jitendra Shah, and Pravin Shah. We also extend our sincere thanks to CUIT and Osman Kabir at Columbia University for their invaluable computing support.

References

Roee Aharoni, Melvin Johnson, and Orhan Firat. 2019. Massively multilingual neural machine translation. *arXiv preprint arXiv:1903.00089*.

Syed Rameel Ahmad. 2024. Enhancing multilingual information retrieval in mixed human resources environments: A rag model implementation for multicultural enterprise. *arXiv preprint arXiv:2401.01511*.

Nadezhda Chirkova, David Rau, Hervé Déjean, Thibault Formal, Stéphane Clinchant, and Vassilina Nikoulina. 2024. Retrieval-augmented generation in multilingual settings. *arXiv preprint arXiv:2407.01463*.

Raj Dabre, Chenhui Chu, and Anoop Kunchukuttan. 2020. A survey of multilingual neural machine translation. *ACM Computing Surveys (CSUR)*, 53(5):1–38.

Jay Gala, Pranjal A Chitale, Raghavan AK, Varun Gumma, Sumanth Doddapaneni, Aswanth Kumar, Janki Nawale, Anupama Sujatha, Ratish Puduppully, Vivek Raghavan, et al. 2023. Indictrans2: Towards high-quality and accessible machine translation models for all 22 scheduled indian languages. *arXiv preprint arXiv:2305.16307*.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, and Haofen Wang. 2023. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*.

Thamme Gowda and Jonathan May. 2020. Finding the optimal vocabulary size for neural machine translation. *arXiv preprint arXiv:2004.02334*.

Sébastien Jean, Kyunghyun Cho, Roland Memisevic, and Yoshua Bengio. 2014. On using very large target vocabulary for neural machine translation. *arXiv preprint arXiv:1412.2007*.

Zhengbao Jiang, Frank F Xu, Luyu Gao, Zhiqing Sun, Qian Liu, Jane Dwivedi-Yu, Yiming Yang, Jamie Callan, and Graham Neubig. 2023. Active retrieval augmented generation. *arXiv preprint arXiv:2305.06983*.

Sneha Reddy Kudugunta, Ankur Bapna, Isaac Caswell, Naveen Arivazhagan, and Orhan Firat. 2019. Investigating multilingual nmt representations at scale. *arXiv preprint arXiv:1909.02197*.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474.

Y Liu. 2020. Multilingual denoising pre-training for neural machine translation. *arXiv preprint arXiv:2001.08210*.

Minh-Thang Luong. 2015. Effective approaches to attention-based neural machine translation. *arXiv preprint arXiv:1508.04025*.

Minh-Thang Luong, Ilya Sutskever, Quoc V Le, Oriol Vinyals, and Wojciech Zaremba. 2014. Addressing the rare word problem in neural machine translation. *arXiv preprint arXiv:1410.8206*.

Kaushal Kumar Maurya, Rahul Kejriwal, Maunendra Sankar Desarkar, and Anoop Kunchukuttan. 2023. Charspan: Utilizing lexical similarity to enable zero-shot machine translation for extremely low-resource languages. *arXiv preprint arXiv:2305.05214*.

Jerin Philip, Shashank Siripragada, Vinay P Namboodiri, and CV Jawahar. 2021. Revisiting low resource status of indian languages in machine translation. In *Proceedings of the 3rd ACM India Joint International Conference on Data Science & Management of Data (8th ACM IKDD CODS & 26th COMAD)*, pages 178–187.

Gowtham Ramesh, Sumanth Doddapaneni, Aravinth Bheemaraj, Mayank Jobanputra, Raghavan Ak, Ajitesh Sharma, Sujit Sahoo, Harshita Diddee, Divyanshu Kakwani, Navneet Kumar, et al. 2022. Samanantar: The largest publicly available parallel corpora collection for 11 indic languages. *Transactions of the Association for Computational Linguistics*, 10:145–162.

Rico Sennrich. 2015. Neural machine translation of rare words with subword units. *arXiv preprint arXiv:1508.07909*.

A Vaswani. 2017. Attention is all you need. *Advances in Neural Information Processing Systems*.

Jiajun Zhang and Chengqing Zong. 2016. Exploiting source-side monolingual data in neural machine translation. In *Proceedings of the 2016 conference on empirical methods in natural language processing*, pages 1535–1545.

Shiyue Zhang, Benjamin Frey, and Mohit Bansal. 2020. Chren: Cherokee-english machine translation for endangered language revitalization. *arXiv preprint arXiv:2010.04791*.

Barret Zoph, Deniz Yuret, Jonathan May, and Kevin Knight. 2016. Transfer learning for low-resource neural machine translation. *arXiv preprint arXiv:1604.02201*.

A RAG Examples

Tables 6 and 7 showcase the example queries passed to the AI-Tutor System along with the responses for the three datasets.

Query	Response			
	English	**Hindi**	**French**	**German**
Who is King Shrenik?	King Shrenik is the eldest son of King Prasenjit.	राजा श्रेणिक राजा प्रसेनजित के सबसे बड़े पुत्र थे।	Le roi Shrenik est le fils aîné du roi Prasenjit.	König Shrenik ist der älteste Sohn von König Prasenjit.
Describe the throne on which Shri Devi was sitting on	The throne on which Shri Devi was sitting on was named Shri in the Sudharma Sabha (divine assembly) of the Shri Avatansak Vimaan (celestial vehicle) in Saudharm Kalp (divine dimension).	सौधर्म कल्प (दिव्य आयाम) में श्री अवतंसक विमान (दिव्य वाहन) की सुधर्मा सभा में श्री देवी जिस सिंहासन पर बैठी थीं, उसका नाम श्री रखा गया।	Le trône sur lequel Shri Devi était assis était nommé Shri dans le Sudharma Sabha (assemblée divine) de Shri Avatansak Vimaan (véhicule céleste) dans Saudharm Kalp (dimension divine).	Der Thron, auf dem Shri Devi saSS, wurde in der Sudharma Sabha (göttliche Versammlung) von Shri Avatansak Vimaan (Himmelsfahrzeug) in Saudharm Kalp (göttliche Dimension) Shri genannt.
Who is Konika?	Konika is the name of a king in ancient India, specifically in the region of Magadha. He is known for his ambition and his actions towards his father, King Shrenika, and his younger brothers, Halla and Vihalla.	कोनिका प्राचीन भारत में, विशेष रूप से मगध क्षेत्र में एक राजा का नाम है। वह अपनी महत्वाकांक्षा और अपने पिता, राजा श्रेणिका और अपने छोटे भाइयों, हल्ला और विहल्ला के प्रति अपने कार्यों के लिए जाने जाते हैं।	Konika est le nom d'un roi de l'Inde ancienne, plus précisément de la région du Magadha. Il est connu pour son ambition et ses actions envers son père, le roi Shrenika, et ses jeunes frères, Halla et Vihalla.	Konika ist der Name eines Königs im alten Indien, insbesondere in der Region Magadha. Er ist bekannt für seinen Ehrgeiz und sein Verhalten gegenüber seinem Vater, König Shrenika, und seinen jüngeren Brüdern Halla und Vihalla.

Table 6: Example Queries from the 11 Agams Dataset passed to the AI-Tutor System along with the response in the supported languages: English, Hindi, French, and German.

Query	Response			
	English	**Hindi**	**French**	**German**
Explain the relationship between karma and the soul	The soul is the doer and enjoyer of karma, and the karma influences the soul's attributes and dispositions. The soul is not the creator of all attributes that are due to its association with karmic matter. The soul's attachment to external objects and its impure cognition are the causes of bondage.	आत्मा कर्म का कर्ता और भोक्ता है, और कर्म आत्मा के गुणों और स्वभाव को प्रभावित करता है। आत्मा उन सभी गुणों का निर्माता नहीं है जो कर्म पदार्थ के साथ उसके जुड़ाव के कारण हैं। आत्मा की बाहरी वस्तुओं के प्रति आसक्ति और उसकी अशुद्ध अनुभूति ही बंधन का कारण है।	L'âme est celle qui pratique et profite du karma, et le karma influence les attributs et les dispositions de l'âme. Lâme nest pas créatrice de tous les attributs dus à son association avec la matière karmique. L'attachement de l'âme aux objets extérieurs et sa cognition impure sont les causes de l'esclavage.	Die Seele ist der Handelnde und GenieSSer von Karma, und das Karma beeinflusst die Eigenschaften und Dispositionen der Seele. Die Seele ist nicht der Schöpfer aller Eigenschaften, die auf ihre Verbindung mit karmischer Materie zurückzuführen sind. Die Bindung der Seele an äuSSere Objekte und ihre unreine Wahrnehmung sind die Ursachen der Knechtschaft.
What are the unique features of the liberated soul?	The liberated soul is described as sense-independent, unparalleled, supreme, and free from obstruction. It is also said to be the real cause of liberation and is established in its own nature.	मुक्त आत्मा को इंद्रिय-स्वतंत्र, अद्वितीय, सर्वोच्च और बाधा से मुक्त बताया गया है। इसे मुक्ति का वास्तविक कारण भी कहा जाता है और यह अपने स्वरूप में स्थापित है।	L'âme libérée est décrite comme indépendante des sens, sans précédent, suprême et libre de toute obstruction. On dit aussi quelle est la véritable cause de la libération et quelle est établie dans sa propre nature.	Die befreite Seele wird als sinnesunabhängig, beispiellos, erhaben und frei von Hindernissen beschrieben. Es wird auch gesagt, dass es die wahre Ursache der Befreiung ist und in seiner eigenen Natur begründet ist.

Table 7: Example Queries from the V.K. Jain Dataset passed to the AI-Tutor System along with the response in the supported languages: English, Hindi, French, and German.