Chain-of-Translation Prompting (CoTR): A Novel Prompting Technique for Low Resource Languages

Tejas Deshpande[1,*], **Nidhi Kowtal**[1,*]
Raviraj Joshi[2,3]
[1] Pune Institute of Computer Technology, Pune, Maharashtra India
[2] Indian Institute of Technology Madras, Chennai, Tamil Nadu India
[3] L3Cube Labs, Pune
{tejasdeshpande1112, kowtalnidhi}@gmail.com
ravirajoshi@gmail.com

Abstract

This paper introduces Chain of Translation Prompting (CoTR), a novel strategy designed to enhance the performance of language models in low-resource languages. CoTR restructures prompts to first translate the input context from a low-resource language into a higher-resource language, such as English. The specified task like generation, classification, or any other NLP function is then performed on the translated text, with the option to translate the output back to the original language if needed. All these steps are specified in a single prompt. We demonstrate the effectiveness of this method through a case study on the low-resource Indic language Marathi. The CoTR strategy is applied to various tasks, including sentiment analysis, hate speech classification, subject classification and text generation, and its efficacy is showcased by comparing it with regular prompting methods. Our results underscore the potential of translation-based prompting strategies to significantly improve multilingual LLM performance in low-resource languages, offering valuable insights for future research and applications. We specifically see the highest accuracy improvements with the hate speech detection task. The technique also has the potential to enhance the quality of synthetic data generation for underrepresented languages using LLMs.

1 Introduction

Natural Language Processing (NLP) has made significant progress in recent years, with models capable of understanding, creating, and translating human language across a wide range of tasks and languages. However since high-resource languages like English, Spanish, and Chinese have access to a wealth of annotated datasets and linguistic resources, most of this development has been focused on those languages. Low-resource languages, on the other hand, have a lot more difficulties since they lack large-scale and high-quality datasets (Thabah and Purkayastha, 2021). Training effective NLP models are challenging due to this data scarcity, which frequently leads to subpar performance and poor generalization. Low-resource languages have distinct grammatical structures, linguistic diversity, and cultural quirks that make it more difficult to create accurate models and limit their use in practical contexts (Yang et al., 2023). Multilingual LLMs have limitations on processing the prompts in low-resource languages (Sanjib Narzary, 2022). This is because the amount of data used to train or fine-tune the model is very less. As a result, speakers of low-resource languages are frequently excluded from the benefits of advanced NLP technologies, highlighting the crucial need for novel techniques to close this gap.

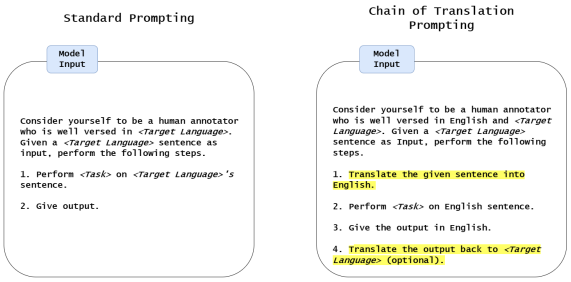

Figure 1: A brief overview of the Chain of Translation Prompting (CoTR) for an annotation task. The technique modifies the input prompt to encapsulate the translation of the non-English input context to English, followed by performing the target task on the translated text.

However, Multilingual LLMs are good at translation tasks, as it is common practice to include parallel corpora during the pre-training stage (Xiang Zhang, 2023). We can leverage the ability of multilingual LLMs to improve responses for low-resource languages. In our study, we apply this ap-

proach to Marathi, an Indo-Aryan language spoken by about 83 million people, primarily in the Indian state of Maharashtra. Marathi is one of these low-resource languages (Joshi, 2022b,a). Despite its large speaker base, Marathi lacks digital resources, annotated corpora, and computational tools. The language's complex syntax challenges the development of precise NLP models, and limited Marathi-specific datasets and pre-trained models hinder the adoption of language technologies (Luong et al., 2023). Therefore, new approaches are needed to enhance Marathi NLP performance and enable its speakers to benefit from AI advancements.

In this work, we investigate new prompting strategies to enhance Marathi language processing capabilities in models such as GPT-4o, GPT-4o Mini, Llama3-8B, Llama3-405B, and Gemma-9B. Our research introduces a novel strategy called "Chain of Translation Prompting (CoTR)", which we evaluate against direct Marathi prompting. We apply this method to sentiment analysis, hate speech classification, and subject categorization across three datasets: MahaSent (Pingle et al., 2023; Kulkarni et al., 2021), MahaHate (Patil et al., 2022), and MahaNews-SHC (Mittal et al., 2023; Aishwarya et al., 2023) respectively. Additionally, we assess its effectiveness in generating headlines using the CSEBUETNLP XLSum dataset. Our findings reveal that translating Marathi input into English and then performing classification or text generation using a single prompt yields superior results compared to directly processing the Marathi text with a standard prompt. This work significantly contributes to multilingual NLP by demonstrating the potential of translation-based prompting strategies, particularly with a single prompt, to enhance NLP performance in low-resource languages.

The main contributions of this work are as follows:

- We introduce Chain of Translation Prompting (CoTR) as a method for performing input context translation during LLM response generation. Our results demonstrate that CoTR consistently outperforms standard prompting strategies across a variety of models and datasets.
- We benchmark various open and closed LLMs, including GPT-4o, GPT-4o mini, Llama 3.1 405B, Llama 3.1 8B, and Gemma 2 9B, on tasks such as Marathi Sentiment Analysis, Hate Speech Detection, News Categorization, and News Headline Generation. In terms of performance, closed LLMs consistently rank higher: GPT-4o > GPT-4o mini > Llama 3.1 405B > Gemma 2 9B > Llama 3.1 8B. We observe that CoTR is particularly beneficial for smaller models with higher error rates.
- The CoTR prompting strategy shows the most significant improvements in complex tasks like hate speech detection and sentiment analysis.

2 Related Work

Natural language processing has improved significantly with the creation of sophisticated models like GPT-4, Llama3, and others. Nonetheless, insufficient representation and scarce data availability in pre-trained models continue to pose problems for low-resource languages(Panteleimon Krasadakis, 2024). Language diversity and data scarcity in low-resource contexts have shown to be challenges for traditional NLP techniques, which has prompted a quest for novel approaches that can make better use of already-existing data. (Michael A. Hedderich, 2021) research highlighted the significance of creating NLP tools that are especially suited for low-resource languages while taking linguistic and cultural quirks into account.

A crucial component of developing NLP models for low-resource languages is dataset curation. In addition to collecting data, curators of datasets such as MahaSent, MahaHate, MahaNews-SHC, and CSEBUETNLP XLSum make sure that the data accurately reflects the linguistic diversity and cultural context of the language. Projects like (Narzary et al., 2022) have brought attention to how crucial it is to provide high-quality datasets that accurately represent language use in everyday situations.

In multilingual natural language processing, cross-lingual transfer methods have demonstrated potential, especially when applied to low-resource language tasks. According to research like that of (Melvin Johnson, 2017), the concept of sharing parameters across languages allows models to acquire representations that function well in a variety of languages. This idea is important because it enables language models to use their English language skills to complete tasks in Marathi through the use of translation-based prompting, which is a type of cross-lingual transfer. Cross-lingual skills are supported by recent advances in multilingual models,

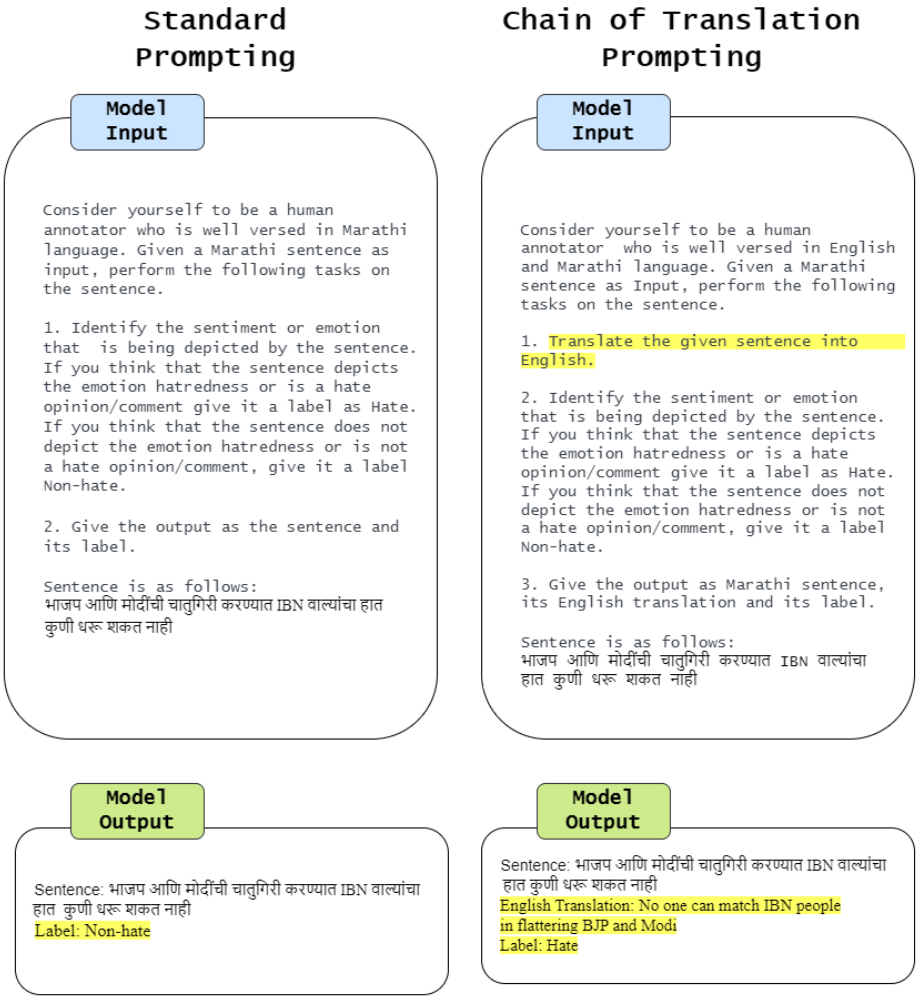

Figure 2: Prompt for Classification Task

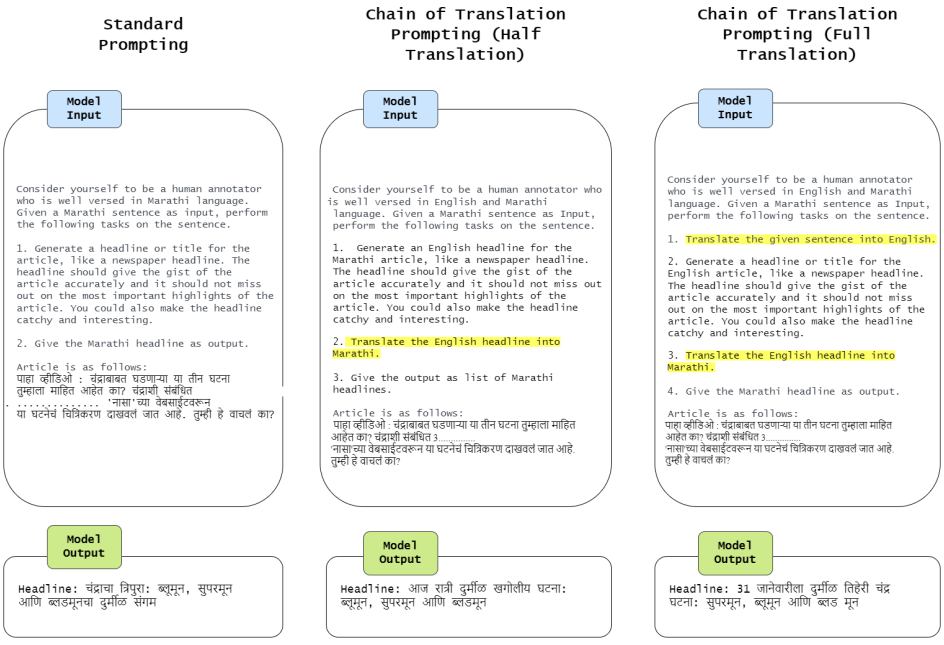

Figure 3: Prompt for Generation Task

such as mBERT (Jacob Devlin, 2019) and XLM-R (Alexis Conneau, 2018), which lay a strong platform for further gains in low-resource language processing.

Prompting strategies have become an effective way to train large language models (LLMs) for particular tasks without requiring a lot of fine-tuning. According to (Tom B. Brown, 2020), well-crafted prompts can direct models such as GPT-3 to carry out a range of NLP tasks effectively. More research has been done on the subject of quick engineering's potential to induce desired behaviors in LLMs even in situations with limited resources by (Pengfei Liu, 2021). These methods have shown to be useful, particularly for languages and activities for which there is little to no direct training data.

Prompting is being used more and more for tasks like sentiment analysis and hate speech detection, which are essential for keeping an eye on public conversation and guaranteeing secure online spaces. Research on Pattern-Exploiting Training (PET) for such tasks was first presented by (Timo Schick, 2021), who showed how prompts could direct models to make context-based, nuanced predictions. This method is consistent with the findings of (Shijun Shi, 2024), who also highlighted the benefit of model prompting for text categorization tasks in a variety of languages and domains.

3 Methodology

3.1 Chain of Translation Prompting

Our study introduces a novel approach called "Chain of Translation Prompting" aimed at enhancing the processing of Marathi, a low-resource language, using advanced language models like GPT-4o, GPT-4o Mini, Llama3-8B, Llama3-405B, and Gemma-9B. Recognizing the strong translation capabilities of these models, we leverage their ability to translate Marathi into English for improved processing. Directly prompting language models in Marathi has posed several challenges, primarily due to the scarcity of quality training data and the models' limitations in comprehending underrepresented languages. These challenges often result in sub-optimal performance on tasks such as sentiment analysis, hate speech classification, news categorization, and headline generation. Below, we outline the step-by-step methodology employed in our approach.

1. **Data Collection and Preparation:** We used datasets specific to Marathi language tasks, including MahaSent for sentiment analysis, MahaHate for hate speech classification, and MahaNews-SHC for subject categorization. For generative tasks, we used the CSE-BUETNLP XLSum dataset to generate headlines.

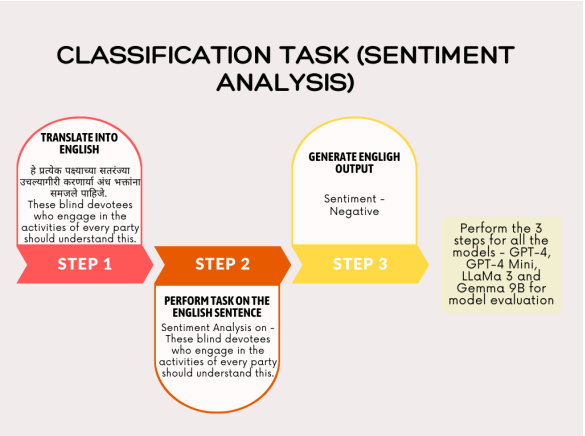

Figure 4: Classification Task using Chain of Translation Prompting

2. **Chain of Translation Prompting (CoTR) Technique:** Our methodology adapts a conventional translation approach used in developing low-resource NLP systems but applies it within the framework of large language model (LLM) prompts. Specifically, our method involves prompting the LLM to first translate the input text from Marathi into English, and then to execute the desired task on the translated English text.

3. **Task Execution:**

   - **Sentiment Analysis, Hate Speech Classification, and Subject Categorization**: For these classification tasks, the models categorize each sentence into predefined classes based on the task's requirements.
   - **Generative Task**: We used GPT-4o, GPT-4o Mini, and Llama3-405b for the headline generation task. The three prompting strategies used for generating headlines are described below.
     (a) Without Translation: In this approach, headlines were generated directly from the original Marathi articles without any translation. This method aimed to assess the model's capability to generate concise and im-

pactful headlines in the source language.

(b) Full Translation: Here, the entire Marathi article was first translated into English. Headlines were then generated based on the translated English text. The generated English headlines were subsequently translated back into Marathi to evaluate their fidelity and relevance.

(c) Half Translation: Given the length and complexity of the articles, the half-translation method was employed to streamline the process. In this approach, English headlines were generated based on the Marathi articles without full translation. These English headlines were then translated back into Marathi. This method aimed to balance efficiency and accuracy by avoiding the need for extensive translation of the entire article.

4. **Direct Prompting**: To evaluate the effectiveness of the Chain of Translation Prompting, we compare its results against the traditional method of directly prompting the models to process the Marathi text without performing translation.

5. **Google Translate + Prompting**: In this approach, Marathi sentences were translated into English using Google Translate. The translated English sentences, along with English prompts, were then used by LLMs to perform the desired classification tasks. This method represents a straightforward "translate-and-test" approach, serving as a baseline for comparison.

6. **Evaluation Metrics**: The performance of the models is measured using conventional metrics, such as the ROUGE-L score for generative tasks. The ROUGE-L score assesses the quality of the generated text, like summaries or headlines, by calculating the overlap with reference text. It evaluates precision and recall by calculating the longest common subsequence (LCS) between the reference text and the generated output. ROUGE-L focuses on capturing the longest word sequences found in both texts, providing insights into the preservation of critical information and coherence.

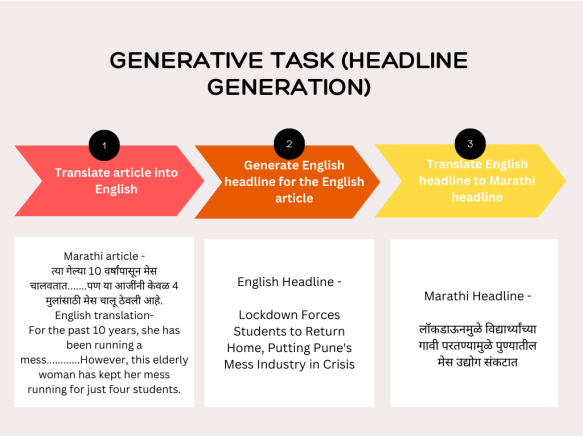

Figure 5: Generative Task using Chain of Translation Prompting

For classification tasks, the model outputs are compared with ground truths, and the error percentage is reported.

3.2 Datasets Used

1. MahaSent-GT[1]:
We used a subset of the L3Cube-MahaSent-MD dataset (Pingle et al., 2023), which contains 14,000 annotated Marathi tweets. Three sentiment labels Positive, Negative, and Neutral are present in the dataset. In particular, we employed the MahaSent-GT portion of this dataset for sentiment analysis.

2. MahaHate[2]:
We used the L3Cube-MahaHate collection's MahaHate 2-Class dataset for our classification task (Patil et al., 2022). It contains around 37500 annotated Marathi sentences. This dataset is divided into two categories: hate and non-hate. We employed the MahaHate 2-Class set for our classification task.

3. MahaNews-SHC[3]:
We analyzed Marathi news articles using the L3Cube-MahaNews-SHC dataset (Mittal et al., 2023). This dataset contains approximately 54,000 news articles spanning a wide

[1] https://github.com/l3cube-pune/MarathiNLP/tree/main/L3Cube-MahaSent-MD
[2] https://github.com/l3cube-pune/MarathiNLP/tree/main/L3Cube-MahaHate
[3] https://github.com/l3cube-pune/MarathiNLP/tree/main/L3Cube-MahaNews-SHC

Table 1: Results on the MahaNews, MahaHate Dataset

Model	Sentence	Ground Truth Label	Label generated with standard prompt	Label generated with CoTR prompt
gpt 4.0	"....तर अवघ्या ३० मिनटांत रशिया अमेरिका आणि युरोपचं नामोनिशाण मिटवेल", इलॉन मस्क यांचा धक्कादायक दावा	International	Technology	International
gemma2 9b	समस्या-अडचणींनी ग्रासलाय? 'हे' स्तोत्र सलग २१ दिवस म्हणा अन् चिंतामुक्त व्हा; जाणून घ्या	Devotion	Neutral	Devotion
llama3-405b	हृदयद्रावक! कोरोनामुळे माय-लेकरांची ताटातूट; 'या' ठिकाणी पालकांपासून मुलांना ठेवलं जातंय दूर	Health	International	Health
llama3-8b	वाहन उद्योगातील मंदी पुढील वर्षी संपेल	Auto	Neutral	Auto
gpt 4.0 mini	कदम रडू नको तुझा मालक कोर्टात नेहमीच तोंडा वर पडतो कारण सगळेच बिनडोक भरले आहेत.	Hate	No hate	Hate

range of topics and was used for the news classification task.

4. XLSum[4]:
   We focused on Marathi text headline creation for our study using the CSEBUETNLP XL-Sum dataset. The dataset offers a wide range of news stories linked with their associated headlines. Our objective was to enhance the accuracy of automated headline creation for Marathi news articles by utilizing this dataset.

3.3 Evaluation Methodology

We performed our classification task on GPT-4o, GPT-4o Mini, Llama3-8B, Llama3-405B, and Gemma-9B.

1. **GPT-4o:**
   GPT-4o is developed by OpenAI, with 1.8 trillion parameters (unofficial). It is a closed-source model and accessible through APIs provided by OpenAI. GPT-4o builds on the advancements of its previous versions, offering enhanced capabilities in natural language understanding, generation, and reasoning across a wide range of tasks.

2. **GPT-4o Mini:**
   GPT-4o Mini is a smaller, more lightweight version of GPT-4o. This model is closed-source. GPT-4o Mini is engineered to balance computational efficiency with performance, making it suitable for applications requiring faster inference times and lower resource consumption while maintaining a high level of language understanding.

3. **Llama 3.1 8B / 405B:**
   Llama 3.1 (Large Language Model for Multilingual Applications) is the third iteration in the Meta Llama series, designed with multiple variants, including a 405 billion parameter version and an 8 billion parameter version. These models are typically open-source. Llama3 models are optimized for multilingual tasks, incorporating vast and diverse datasets to improve performance across different languages.

[4] https://huggingface.co/datasets/csebuetnlp/xlsum

4. **Gemma-2 9B:**
   Gemma-2 9B is an open-source language model with 9 billion parameters from Google. It strikes a balance between model size and performance, offering robust capabilities for both academic and practical applications.

Model	Without Translation	Half Translation	Full Translation
GPT-4o	33.3	44	49
GPT-4o mini	21.34	21.72	22.22
llama3-405b	20.27	20.34	21.13

Table 2: Rouge-L score in percentage for 3 approaches on the headline generation task on CSEBUETNLP XL-Sum Dataset

4 Results and Discussion

Table 2 and Table 3 show the analysis done on Standard Prompting and Chain of Translation Prompting.

4.1 Classification Task

Approximately 100 sentences were selected from MahaSent-GT, MahaNews-SHC, and MahaHate. The large language models categorize each of the sentences into a predefined category. These results were compared with the ground truth values to calculate the error rate. The error rate was calculated with the direct prompting approach and Chain of Translation prompting approach The results are shown in Table 3.

In the CoTR prompting approach, the error rate has reduced by 2.32% in the GPT-4o model, by 3.64% llama3-405b, by 5.29% in llama3-8b and by 4.96% in GPT-4o Mini. The error rate is slightly increased by 0.33% in the Gemma-9B model.

The error rate has been reduced by almost 5% in llama3-8b and gpt4 mini models. Specifically, the CoTR prompting approach has significantly improved hate speech identification across all models except for Gemma-9B. In the hate speech classification task, Gemma-9B often failed to correctly translate hateful comments and, in some cases, omitted those parts entirely. However, compared to standard prompting, the number of misclassifications for the "Non-hate" class was lower when using CoTR.

The results from the CoTR approach show significant improvement over the standard Google Translate method as well. We manually reviewed the translated sentences and found out that translations by large language models (LLMs), such as GPT-4 and GPT-4 Mini, captured meanings and nuances more effectively than Google Translate. While LLMs conveyed the intended meaning with subtlety, Google Translate produced literal translations, which sometimes failed to capture the full sense of the sentences.

For GPT-4 and GPT-4 Mini, the direct translation approach surpassed Google Translate's performance, as the nuances of some of the sentences did not get extracted completely by the google translator. As GPT-4 and GPT-4 Mini are stronger models the direct prompting is working better than Google translator approach.

One sample detection with traditional prompting versus CoTR prompting from each of the four models has been attached in Table 1, where the output with CoTR prompting is the same as the ground truth.

4.2 Generation Task

The headlines from the Marathi news text were generated using traditional prompting and CoTR prompting (with half and full translation). The headlines were compared against the manually assigned headline and the Rouge-L score metric was used to calculate their similarity with the manually assigned headline. The Rouge-L score for traditional prompting and CoTR prompting (half and full translation) are given in Table 2

We observed that GPT-4o delivered the best performance among all the models. GPT-4o Mini struggled to identify fine details in the articles, while Llama3-405B occasionally failed to provide the results in the specified format and produced some inaccurate translations. Overall, GPT-4o Mini and Llama-405B yielded similar outcomes.

In general, we observe the following performance ranking for Marathi tasks: GPT-4o > GPT-4o Mini > Llama 3.1 405B > Gemma 2 9B > Llama 3.1 8B. The CoTR approach proves especially useful with smaller models and for complex tasks like hate speech detection and sentiment analysis.

5 Future Work and Conclusion

In summary, our study demonstrates that various prompting strategies, particularly the Chain of

Table 3: Error percentage in the classification task across 5 models (these are the weighted averages and the numbers are percentages). Standard Prompt - Prompt the LLM to perform the task using the given Marathi context. CoTR Prompt - Prompt the LLM to translate the Marathi context to English and then perform the task. Google Translate - Translate the Marathi context to English using Google Translate and then prompt the LLM to perform the task in English.

Model	Dataset	Standard Prompt	CoTR Prompt	Google Translate	Average Standard Prompt	Average CoTR Prompt	Avg Google Translate Prompt
gpt-4o	MahaSent	20.38	18.44	25.00	13.57	11.25	18.70
	MahaNews	3.06	2.04	6.12			
	MahaHate	16.83	12.87	26.70			
gpt-4o mini	MahaSent	20.38	19.41	33.00	20.19	15.23	24.40
	MahaNews	6.12	4.08	9.18			
	MahaHate	33.66	21.78	30.70			
llama3-405b	MahaSent	31.06	27.18	22.00	19.86	16.22	18.70
	MahaNews	7.14	6.12	6.12			
	MahaHate	20.89	14.85	27.72			
llama3-8b	MahaSent	35.92	27.18	30.00	29.13	23.84	24.00
	MahaNews	10.20	7.14	9.18			
	MahaHate	40.59	36.63	32.60			
gemma9b	MahaSent	33.98	27.18	29.00	22.18	22.51	22.40
	MahaNews	10.20	10.20	11.20			
	MahaHate	21.78	29.70	26.70			

Translation (CoTR) method, effectively enhance Marathi language processing tasks. By applying these techniques to various classification and generation tasks, we have expanded the potential for more reliable and accurate NLP applications in Marathi. While CoTR improves model performance, it does so at the cost of generating more tokens.

In the future, we aim to enhance performance on Marathi language tasks by combining Chain of Thought (CoT) and Chain of Translation (CoTR) prompting strategies. Our goal is to achieve context-aware and precise responses for complex tasks like sentiment analysis, hate speech detection, and subject classification. CoT allows models to break down complex tasks into simpler steps, while CoTR leverages translation from Marathi to English, where more accurate models can be employed. Together, these strategies should create a robust framework that improves model performance and reliability in Marathi NLP tasks.

This approach can further be used for other low-resource Indic languages.

Acknowledgments

This work was done under the mentorship of Mr. Raviraj Joshi (Mentor, L3Cube Pune). We would like to express our gratitude towards him for his continuous support and encouragement.

References

Michael Ahn and Anthony Brohan. 2022. Do as i can, not as i say: Grounding language in robotic affordances. *arXiv preprint arXiv:2204.01691*.

Kabir Ahuja, Harshita Diddee, Rishav Hada, Millicent Ochieng, Krithika Ramesh, Prachi Jain, Akshay Nambi, Tanuja Ganu, Sameer Segal, Mohamed Ahmed, Kalika Bali, and Sunayana Sitaram. 2023. Mega: Multilingual evaluation of generative ai. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, page 4232–4267, Singapore. Association for Computational Linguistics.

Mirashi Aishwarya, Sonavane Srushti, Lingayat Purva, Padhiyar Tejas, and Joshi Raviraj. 2023. L3cube-indicnews: News-based short text and long document classification datasets in indic languages. In *Proceedings of the 20th International Conference on Natural Language Processing (ICON)*, pages 442–449.

Table 4: Sample examples for the headline generation task using CoTR

Manually assigned heading	Without Translation	Half Translation	Full Translation
कोरोना पुणे लॉकडाऊन: 83 वर्षांच्या आजीने प्रेमापोटी केवळ 4 मुलांसाठी मेस चालू ठेवली	लॉकडाऊनमुळे पुण्यातील मेस व्यवसायाचा संकोच: 7 सदस्यांचे कुटुंब संकटात	पुण्यात लॉकडाऊनचा फटका: विद्यार्थी परतल्यानं मेस व्यवसाय ठप्प, परंतु काही जण अजूनही चार मुलांसाठी सुरु	लॉकडाऊनमुळे विद्यार्थ्यांच्या गावी परतण्यामुळे पुण्यातील मेस उद्योग संकटात
राज ठाकरे यांचा CAA आणि NRC वर हल्लाबोल, "भारत धर्मशाळा नाही"	राज ठाकरे यांचा नागरिकत्व सुधारणा कायद्यावर हल्ला; "भारत धर्मशाळा नाही"	राज ठाकरे यांची CAA आणि NRC वर टीका; परक्यांना आसरा देण्याची गरज काय?	राज ठाकरे यांचा CAA आणि NRC वर हल्लाबोल, भारतात बाहेरून आलेल्या लोकांची गरज काय?
31 जानेवारीला दुर्मीळ घटना: सुपरमून, ब्लूमून आणि ब्लड मून	चंद्राचा त्रिपुरा: ब्लूमून, सुपरमून आणि ब्लडमूनचा दुर्मीळ संगम	31 जानेवारीला दुर्मीळ तिहेरी चंद्र घटना: सुपरमून, ब्लूमून आणि ब्लड मून	आज रात्री दुर्मीळ खगोलीय घटना: ब्लूमून, सुपरमून आणि ब्लडमून

Md Shad Akhtar, Ayush Kumar, Asif Ekbal, and Pushpak Bhattacharyya. 2016. A hybrid deep learning architecture for sentiment analysis. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, page 482–493, Osaka, Japan. The COLING 2016 Organizing Committee.

Naman Goyal Alexis Conneau, Kartikay Khandelwal. 2018. Unsupervised cross-lingual representation learning at scale. In *Proceedings of ACL*.

Jacob Andreas, Dan Klein, and Sergey Levine. 2018. Learning with latent language. In *Proceedings of NAACL*.

Mikel Artetxe, Vedanuj Goswami, Shruti Bhosale, Angela Fan, and Luke Zettlemoyer. 2023. Revisiting machine translation for cross-lingual classification. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, page 6489–6499, Singapore. Association for Computational Linguistics.

Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, Quyet V. Do, Yan Xu, and Pascale Fung. 2023. A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, page 675–718, Nusa Dua, Bali. Association for Computational Linguistics.

Howard Chen, Jacqueline He, Karthik Narasimhan, and Danqi Chen. 2022. Can rationalization improve robustness? In *Proceedings of NAACL*.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, et al. 2021. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

Nikhil Goyal, Harsh Trivedi, and Prithiviraj Sen. 2022. Prompting techniques for improving performance on low-resource nlp tasks. In *Proceedings of the Conference on Neural Information Processing Systems (NeurIPS)*.

Yuling Gu, Bhavana Dalvi Mishra, and Peter Clark. 2022. Dream: Uncovering mental models behind language models. In *Proceedings of NAACL*.

Ayiguli Halike, Aishan Wumaier, and Tuergen Yibulayin. 2023. Zero-shot relation triple extraction with prompts for low-resource languages. *Applied Sciences*.

Peter Hase and Mohit Bansal. 2022. When can models learn from explanations? a formal framework for understanding the roles of explanation data. In *Proceedings of ACL*.

Kenton Lee Jacob Devlin, Ming-Wei Chang. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of ACL*.

Zhanming Jie, Jierui Li, and Wei Lu. 2022. Learning to reason deductively: Math word problem solving as complex relation extraction. *arXiv preprint arXiv:2203.10316*.

Raviraj Joshi. 2022a. L3cube-mahacorpus and mahabert: Marathi monolingual corpus, marathi bert language models, and resources. In *Proceedings of the WILDRE-6 Workshop within the 13th Language Resources and Evaluation Conference*, pages 97–101.

Raviraj Joshi. 2022b. L3cube-mahanlp: Marathi natural language processing datasets, models, and library. *arXiv preprint arXiv:2205.14728*.

Sameer Khurana, Ashwin Ghosh, and Sreelakshmi Nair. 2022. Using prompt-based learning for enhanced low-resource language models. *Journal of Artificial Intelligence Research*.

John Koutsikakis, Konstantinos Papagiannopoulos, and Antonis Papadakis. 2022. Prompting strategies for zero-shot text classification in low-resource languages. *Journal of Artificial Intelligence Research*.

Atharva Kulkarni, Meet Mandhane, Manali Likhitkar, Gayatri Kshirsagar, and Raviraj Joshi. 2021. L3cubemahasent: A marathi tweet-based sentiment analysis dataset. In *Proceedings of the Eleventh Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 213–220.

Andrew K. Lampinen, Ishita Dasgupta, Stephanie C.Y. Chan, Kory Matthewson, Michael Henry Tessler, Antonia Creswell, James L. McClelland, Jane X. Wang, and Felix Hill. 2022. Can language models learn from explanations in context? *arXiv preprint arXiv:2204.02329*.

Yihuai Lan, Lei Wang, Qiyuan Zhang, Yunshi Lan, Bing Tian Dai, Yan Wang, Dongxiang Zhang, and Ee-Peng Lim. 2021. Mwptoolkit: An open-source framework for deep learning-based math word problem solvers. *arXiv preprint arXiv:2109.00799*.

Brian Lester, Rami Al-Rfou, and Noah Constant. 2021. The power of scale for parameter-efficient prompt tuning. In *Proceedings of EMNLP*.

Xiang Lisa Li and Percy Liang. 2021. Prefix-tuning: Optimizing continuous prompts for generation. In *Proceedings of ACL*.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2021. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *arXiv preprint arXiv:2107.13586*.

X. Liu, H. Wu, L. Shen, S. Zhang, and M. Zhou. 2023. Empirical evaluation of multilingual language models for low-resource languages. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Minh-Thang Luong, Quoc V. Le, and Thang Luong. 2023. Multilingual neural machine translation with a special focus on low-resource languages. *Transactions of the Association for Computational Linguistics (TACL)*.

Ana Marasovic, Iz Beltagy, Doug Downey, and Matthew E Peters. 2022. Few-shot self-rationalization with natural language prompts. In *NAACL Findings*.

Quoc V. Le Melvin Johnson, Mike Schuster. 2017. Google's multilingual neural machine translation system: Enabling zero-shot translation. In *Proceedings of ACL*.

Heike Adel Michael A. Hedderich, Lukas Lange. 2021. A survey on recent approaches for natural language processing in low-resource scenarios. In *Proceedings of ACL*.

Saloni Mittal, Vidula Magdum, Sharayu Hiwarkhedkar, Omkar Dhekane, and Raviraj Joshi. 2023. L3cube-mahanews: News-based short text and long document classification datasets in marathi. In *International Conference on Speech and Language Technologies for Low-resource Languages*, pages 52–63. Springer.

Abdul Rehman Javed Muhammad Farrukh Bashir. 2023. Context-aware emotion detection from low-resource urdu language using deep neural network. In *ACM Journals*.

Sanjib Narzary, Maharaj Brahma, and Mwnthai Narzary. 2022. Generating monolingual dataset for low resource language bodo from old books using google keep. In *Proceedings of ACL*.

Vassilios S. Verykios Panteleimon Krasadakis, Evangelos Sakkopoulos. 2024. A survey on challenges and advances in natural language processing with a focus on legal informatics and low-resource languages. In *Proceedings of MDPI*.

Arkil Patel, Satwik Bhattamishra, and Navin Goyal. 2021. Are nlp models really able to solve simple math word problems? In *Proceedings of NAACL*.

Hrushikesh Patil, Abhishek Velankar, and Raviraj Joshi. 2022. L3cube-mahahate: A tweet-based marathi hate speech detection dataset and bert models. In *Proceedings of the Third Workshop on Threat, Aggression and Cyberbullying (TRAC 2022)*, pages 1–9.

Jinlan Fu Pengfei Liu, Weizhe Yuan. 2021. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. In *Proceedings of ACL*.

Xinyu Pi, Qian Liu, Bei Chen, Morteza Ziyadi, Zeqi Lin, Yan Gao, Qiang Fu, Jian-Guang Lou, and Weizhu Chen. 2022. Reasoning like program executors. *arXiv preprint arXiv:2201.11473*.

Aabha Pingle, Aditya Vyawahare, Isha Joshi, Rahul Tangsali, and Raviraj Joshi. 2023. L3cube-mahasent-md: A multi-domain marathi sentiment analysis dataset and transformer models. In *Proceedings of the 37th Pacific Asia Conference on Language, Information and Computation*, pages 274–281.

Jack W. Rae, Sebastian Borgeaud, Trevor Cai, Katie Millican, Jordan Hoffmann, Francis Song, John Aslanides, Sarah Henderson, Roman Ring, Susannah Young, et al. 2021. Scaling language models: Methods, analysis and insights from training gopher. *arXiv preprint arXiv:2112.11446*.

Mwnthai Narzary Sanjib Narzary, Maharaj Brahma. 2022. Generating monolingual dataset for low resource language bodo from old books using google keep. In *Proceedings of ACL*.

Teven Le Scao and Alexander Rush. 2021. How many data points is a prompt worth? In *Proceedings of NAACL*.

Jie Xi Shijun Shi, Kai Hu. 2024. Robust scientific text classification using prompt tuning based on data augmentation with l2 regularization. In *Science Direct*.

N. Donald Jefferson Thabah and Bipul Syam Purkayastha. 2021. Low resource neural machine translation from english to khasi: A transformer-based approach. In *Low Resource Neural Machine Translation from English to Khasi: A Transformer-Based Approach*.

Hinrich Schütze Timo Schick. 2021. Exploiting cloze questions for few shot text classification and natural language inference. In *Proceedings of ACL*.

Nick Ryder Tom B. Brown, Benjamin Mann. 2020. Language models are few-shot learners. In *Proceedings of ACL*.

Bradley Hauer Xiang Zhang, Senyu Li. 2023. Don't trust chatgpt when your question is not in english: A study of multilingual abilities and types of llms. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Yuqing Yang, Jie Fu, and Pascal Poupart. 2023. Prompt learning for low-resource language understanding with pretrained models. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*.