# Prompting Beyond English: Indian Language Instructions are Harder for Language Models

Anonymous ACL submission


## Abstract

Large Language Models (LLMs) and Small Language Models (SLMs) are increasingly used as general-purpose systems for performing downstream tasks through natural language instructions. However, most evidence for their instruction-following ability is based on English prompts. In this work, we investigate how contemporary LLMs and SLMs follow instructions when prompted in Indian languages. We collect 80 instructions from the Super-NaturalInstructions dataset and translate them into Bangla, Telugu, Marathi, and Sanskrit, while keeping the underlying task fixed across languages. We evaluate multiple LLMs and SLMs under these semantically equivalent prompts. Our results show that models perform substantially better when instructed in English, and the disparity is pronounced for SLMs. Gemma-3 and Gemma-4 show greater robustness across languages, with performance in English exceeding that in Indian languages by up to 15%. We further instruction-tune LLaMA-3.2-1B on Bangla and English subsets of Updesh to examine whether the gap arises from inadequate instruction tuning in Indian languages. Instruction tuning improves LLaMA's performance on Bangla prompts, while English prompts show nearly the same performance. However, the performance of LLaMA-3.2-1B still lags by *59.05%* in Bangla prompting compared to English.


## 1 Introduction

Large Language Models (LLMs with >4B parameters) and Small Language Models (SLMs with <=4B parameters) (Oliveira, 2026) are increasingly used as general-purpose systems for various downstream tasks through natural language instructions or prompts. Their ability to follow prompts has made them useful for tasks including question-answering, summarization, machine translation, and reasoning. However, most of the evidence on the models' ability to follow instructions and execute downstream tasks focuses on English (Yang et al., 2025a; Kamath et al., 2025; Grattafiori et al., 2024). English covers the majority of the pretraining corpora and instruction-tuning datasets. Subsequently, the performance of these models on downstream tasks when given English instructions does not necessarily guarantee their efficacy when prompted in other languages. This concern is particularly important for Indian languages. According to the census 2011, only 10.6% [1] of the Indian population knows English. Thus, neural models' performance in following instructions in the Indian languages is necessary for democratization of artificial intelligence in India. Even when the underlying instruction remains semantically identical, the efficacy of these models on downstream tasks may vary when prompted in Indian languages. This motivates our work to assess the extent to which current LLMs and SLMs follow semantically equivalent instructions in Indian languages.

For the evaluation of language models, we collected 80 instructions, selected at random from the Super-NaturalInstructions dataset (Wang et al., 2022) in English and translated them into four Indian languages, including Bangla, Telugu, Marathi, and Sanskrit. The instruction preserves the same task across languages, allowing us to evaluate whether the model's performance varies under semantically equivalent prompts.

Our findings show that current language models do not exhibit uniform ability in following instructions in Indian languages. All the models perform significantly better when prompted in English than when prompted in Indian languages. This observation is true even for models trained on Indian languages like Sarvam-m. The disparity in exe-

[1] https://en.wikipedia.org/wiki/2011_census_of_India

cuting tasks in Indian languages becomes more pronounced in SLMs (<=4B parameters). Across all SLMs, English prompts produce scores that are significantly higher (between 18% and 500%) than those for Indian-language prompts. Gemma-3 and Gemma-4 are robust in following instructions across languages.

We further instruction-tuned LLaMA-3.2-1B on Bangla and English subset of UPDESH (Chitale et al., 2025) dataset to examine whether the performance disparity is a result of inefficient instruction tuning on Indian languages. Our findings show that after instruction tuning on Updesh, the performance of LLaMA-3.2-1B increases by 18.20% for Bangla. The performance of model on English after instruction tuning with Updesh remained nearly the same. However, performance on prompting in Bangla still lags by 59.05% compared to English, indicating inefficient fine-tuning is not the only cause of such performance disparity.

Our contributions can be summarized as

- We conducted experiments to understand the performance of models when instructed on Indian languages compared to English. Our findings indicate that models exhibit superior performance when instructed in English compared to Indian languages. The disparity in performance is more pronounced for SLMs (<=4B parameter models).

- We further investigated whether the performance gap is due to inadequate instruction tuning for Indian languages. Our findings show that after instruction tuning for Indian languages, the models' performance improves for Bangla by 18% for LLaMA-3.2-1B. However prompting in Bangla still lags by 59.05% compared to prompting in English.

## 2 Related Work

Instruction tuning has become a central paradigm for improving the task generalization abilities of Large Language Models (LLMs) and Small Language Models (SLMs). Super-NaturalInstructions (Wang et al., 2022) demonstrated that instruction tuning substantially improves zero-shot performance across diverse NLP tasks. However, most evaluations of instruction-following behavior remain largely English-centric, despite the increasing multilingual deployment of modern language models.

Recent multilingual LLMs such as Qwen (Yang et al., 2025a), Gemma (Kamath et al., 2025), and LLaMA-3 (Grattafiori et al., 2024) report strong multilingual capabilities across multiple benchmarks. Benchmark suites including MMLU (Hendrycks et al., 2020), XGLUE (Liang et al., 2020), XTREME (Hu et al., 2020), and MASSIVE (FitzGerald et al., 2023) have enabled multilingual evaluation across several languages and tasks. However, these benchmarks often vary both the instruction language and task content simultaneously, making it difficult to isolate whether performance degradation arises specifically from instruction understanding. In contrast, our work keeps the task inputs fixed while varying only the instruction language, allowing us to directly study cross-lingual instruction-following behavior.

Prior work has shown that multilingual performance remains highly uneven across languages, particularly for low-resource and morphologically rich languages (Muennighoff et al., 2023). These disparities are especially important for Indian languages, which remain underrepresented in pre-training corpora despite having large speaker populations. Recent efforts such as Updesh (Chitale et al., 2025) have explored multilingual instruction tuning for Indian languages to improve alignment and downstream performance. Our work further investigates whether multilingual instruction tuning alone is sufficient to mitigate instruction-following gaps across Indian languages.

## 3 Dataset

We sampled 80 random instructions from the publicly available Super-NaturalInstructions dataset (Wang et al., 2022). The selected 80 instructions are then translated into Bangla, Marathi, Telugu and Sanskrit using GPT-4o. The translated dataset is then corrected by experts who are proficient, ensuring high quality. We made sure to retain phrases such as "Member of" and "IsA" in their original form, without translating them into other languages. This created a set of 80 instructions, resulting in 288 samples across tasks, which we use for our experiments with language models. Table A2 shows examples of prompts in English and Indian languages.

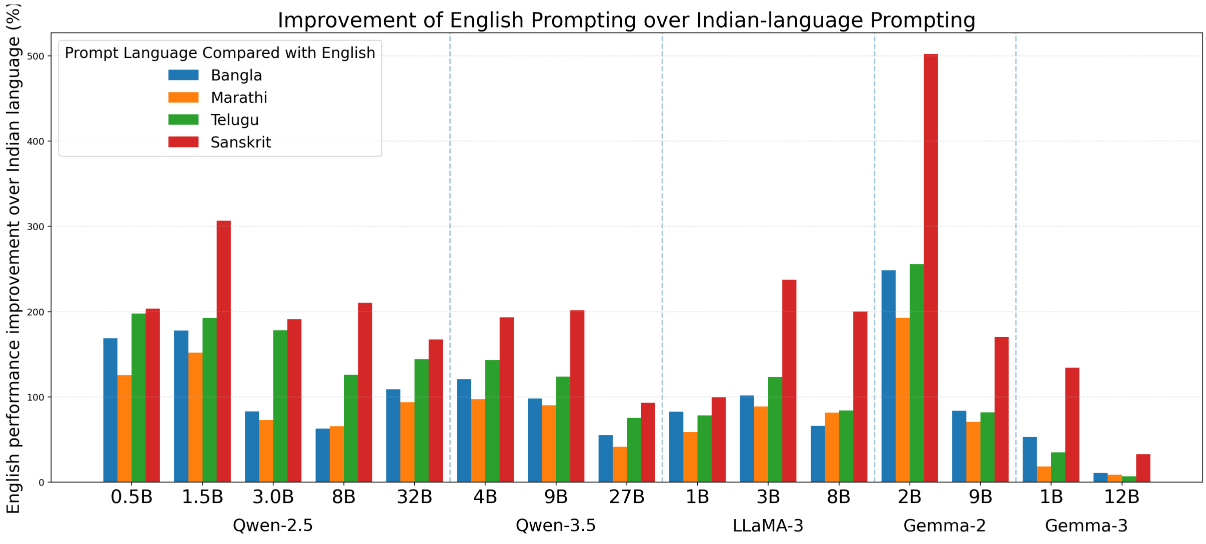


Figure 1: Figure showing percentage by which English prompting outperforms Indian language prompting across model families using ROUGE-Lsum.

| Model | Params. | English | | | | Bangla | | | | Telugu | | | | Marathi | | | | Sanskrit | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | R-1 | R-2 | R-L | R-Lsum | R-1 | R-2 | R-L | R-Lsum | R-1 | R-2 | R-L | R-Lsum | R-1 | R-2 | R-L | R-Lsum | R-1 | R-2 | R-L | R-Lsum |
| DeepSeek-V2-Lite | 16B | **9.38** | **2.56** | **8.85** | **8.83** | 5.15 | 1.79 | 4.69 | 4.73 | 4.65 | 1.29 | 4.21 | 4.22 | 4.29 | 1.01 | 3.98 | 3.98 | 3.40 | 1.09 | 3.09 | 3.16 |
| Qwen-2.5 | 7B | **56.20** | **17.13** | **55.37** | **55.36** | 34.90 | 11.23 | 33.96 | 33.97 | 25.86 | 10.02 | 24.42 | 24.49 | 34.83 | 11.13 | 33.40 | 33.41 | 18.95 | 6.45 | 17.86 | 17.84 |
| Qwen-2.5 | 32B | **62.98** | **19.83** | **61.99** | **61.99** | 30.59 | 12.60 | 29.65 | 29.64 | 26.22 | 12.64 | 25.30 | 25.37 | 32.87 | 13.29 | 31.96 | 32.00 | 24.07 | 9.75 | 23.10 | 23.17 |
| Qwen-3.5 | 9B | **59.15** | **21.55** | **58.30** | **58.25** | 30.59 | 12.65 | 29.46 | 29.39 | 27.05 | 10.14 | 26.05 | 26.04 | 32.16 | 12.05 | 30.74 | 30.60 | 20.05 | 7.06 | 19.19 | 19.30 |
| Qwen-3.5 | 27B | **60.93** | **23.25** | **60.14** | **59.76** | 39.74 | 15.63 | 38.83 | 38.50 | 35.04 | 15.74 | 34.31 | 34.05 | 43.33 | 17.68 | 42.51 | 42.21 | 32.08 | 13.23 | 31.09 | 30.95 |
| OLMO-3 | 7B | **51.69** | **18.29** | **50.72** | **50.73** | 26.14 | 7.93 | 25.43 | 25.44 | 27.87 | 6.99 | 26.95 | 26.97 | 31.50 | 10.70 | 30.41 | 30.44 | 14.49 | 5.14 | 13.93 | 13.94 |
| LLaMA-3.1 | 8B | **47.13** | **15.47** | **45.41** | **45.32** | 28.92 | 10.60 | 27.39 | 27.31 | 26.30 | 9.88 | 24.78 | 24.63 | 25.93 | 10.06 | 25.01 | 24.94 | 15.90 | 6.40 | 15.08 | 15.09 |
| Param-2 | 17B | 6.21 | 2.07 | 5.29 | 5.75 | 9.49 | **2.41** | 8.96 | 9.11 | **9.65** | 1.78 | **9.09** | **9.40** | 6.74 | 2.19 | 6.29 | 6.31 | 5.59 | 1.41 | 5.25 | 5.44 |
| Sarvam-m | 24B | **34.66** | **8.76** | **34.52** | **34.52** | 27.29 | 5.18 | 27.10 | 26.95 | 26.46 | 6.24 | 26.26 | 26.15 | 24.32 | 5.85 | 24.19 | 24.12 | 10.21 | 2.76 | 10.16 | 10.17 |
| Gemma-2 | 9B | **52.45** | **18.56** | **51.65** | **51.65** | 29.06 | 10.79 | 28.11 | 28.13 | 29.21 | 11.46 | 28.34 | 28.37 | 31.17 | 11.70 | 30.25 | 30.25 | 19.77 | 7.12 | 19.12 | 19.12 |
| Gemma-3 | 12B | **54.42** | **17.31** | **53.33** | **53.33** | 49.39 | 16.32 | 48.34 | 48.16 | 50.87 | 16.29 | 50.12 | 49.96 | 50.32 | 17.45 | 49.23 | 49.05 | 41.81 | 12.38 | 40.26 | 40.15 |
| Gemma-4 | 31B | **59.95** | **22.00** | **59.11** | **59.00** | 54.42 | 19.42 | 53.54 | 53.52 | 52.54 | 18.72 | 51.70 | 51.72 | 57.15 | 20.46 | 56.20 | 56.11 | 51.83 | 15.93 | 50.95 | 50.97 |

Table 1: ROUGE scores across LLMs for all languages.

## 4 Evaluation

We evaluate zero-shot performance by prompting models with instructions in different languages. We vary the instruction language from English to Bangla, Marathi, Telugu, and Sanskrit to assess the models' ability to follow instructions in Indian languages. We evaluated multilingual models including Qwen-2.5 (Yang et al., 2025b), Qwen-3.5 [2], LLaMA-3 (Grattafiori et al., 2024), Gemma-2 (Riviere et al., 2024), Gemma-3 (Kamath et al., 2025), Gemma-4 [3], Phi-3.5 (Abdin et al., 2024), Deepseek-v2 (Liu et al., 2024), and models trained for Indian languages like Sarvam-m and Param-2. ROUGE-1, ROUGE-2, ROUGE-L, and ROUGE-Lsum (Lin, 2004) are used as the evaluation metrics. Since ROUGE-Lsum provides the best overall picture (Lin, 2004), our analysis has been based on it. Table A3 shows the number of words and tokens generated by the models on 80 instructions, whereas Table A1 shows the fertility rates of the models' tokenizers.

Table 1 and Table 2 show significant improvement in task performance with prompting in English compared to Indian languages. Figure 1 and Table A4 show the percentage improvement in executing instructions in English when compared to Indian languages. Similar to Table 1 and Table 2, the results show a consistent advantage for English Instructions over Indian languages. This disparity in performance is more pronounced for SLMs, including Qwen-2.5-1.5B, Qwen-3.5-4B, LLaMA-3.2-3B, and Gemma-2-2B, where English prompting improves performance by $\sim$100% or more. The performance in Sanskrit and Telugu is worse than Bangla and Marathi for almost all variants. Gemma-2 in particular is 5 times better in English than in Sanskrit.

Primarily Indic models like Sarvam-m also perform significantly better with English prompts. English prompts outperform Bangla prompts for Sarvam-m by 28.09%, Telugu prompts by 32.00% and Marathi prompts by 43.12%. Sarvam models

[2]https://qwen.ai/blog?id=qwen3.5
[3]https://deepmind.google/models/gemma/gemma-4/

| Model | Params. | English | | | | Bangla | | | | Telugu | | | | Marathi | | | | Sanskrit | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | R-1 | R-2 | R-L | R-Lsum | R-1 | R-2 | R-L | R-Lsum | R-1 | R-2 | R-L | R-Lsum | R-1 | R-2 | R-L | R-Lsum | R-1 | R-2 | R-L | R-Lsum |
| Qwen-2.5 | 0.5B | **30.10** | **8.74** | **29.16** | **29.16** | 11.59 | 4.12 | 10.82 | 10.84 | 10.89 | 3.89 | 9.78 | 9.80 | 13.76 | 4.69 | 12.92 | 12.93 | 10.54 | 4.04 | 9.58 | 9.61 |
| Qwen-2.5 | 1.5B | **42.97** | **14.98** | **42.19** | **42.06** | 16.28 | 6.89 | 15.25 | 15.13 | 15.19 | 6.47 | 14.35 | 14.37 | 17.98 | 7.10 | 16.69 | 16.69 | 11.19 | 4.64 | 10.24 | 10.34 |
| Qwen-2.5 | 3B | **46.88** | **15.50** | **45.90** | **45.93** | 26.08 | 7.30 | 25.08 | 25.10 | 17.61 | 5.19 | 16.42 | 16.50 | 27.54 | 8.94 | 26.50 | 26.55 | 16.58 | 4.89 | 15.73 | 15.77 |
| Qwen-3.5 | 4B | **57.28** | **19.13** | **56.35** | **56.26** | 26.55 | 10.06 | 25.51 | 25.49 | 24.15 | 9.56 | 23.14 | 23.12 | 29.50 | 10.06 | 28.48 | 28.48 | 20.10 | 8.28 | 19.16 | 19.18 |
| LLaMA-3.2 | 1B | **19.15** | **6.94** | **18.25** | **18.26** | 10.86 | 3.71 | 9.95 | 10.00 | 11.29 | 3.95 | 10.20 | 10.23 | 12.54 | 4.70 | 11.41 | 11.50 | 9.91 | 3.60 | 9.07 | 9.15 |
| LLaMA-3.2 | 3B | **44.66** | **15.10** | **43.16** | **43.03** | 22.17 | 8.24 | 21.31 | 21.33 | 20.33 | 7.59 | 19.20 | 19.25 | 23.86 | 9.23 | 22.82 | 22.81 | 13.69 | 5.62 | 12.70 | 12.75 |
| Phi-3.5 | 3.8B | **15.29** | **5.85** | **14.46** | **14.47** | 11.83 | 4.76 | 10.82 | 10.96 | 9.70 | 4.07 | 8.81 | 8.92 | 12.56 | 5.42 | 11.58 | 11.61 | 10.72 | 3.33 | 9.76 | 9.90 |
| SmolLM | 3B | **38.93** | **12.31** | **38.11** | **37.96** | 12.65 | 5.03 | 11.97 | 12.01 | 10.22 | 4.54 | 9.43 | 9.51 | 15.97 | 5.80 | 15.14 | 15.18 | 10.99 | 4.20 | 10.11 | 10.14 |
| Sarvam-1 | 2B | 7.45 | 1.43 | 7.09 | 7.09 | 7.48 | **1.68** | 6.91 | 6.88 | 6.54 | 1.44 | 6.16 | 6.17 | **9.23** | 1.67 | **8.76** | **8.75** | 5.62 | 1.28 | 5.09 | 5.03 |
| Gemma-2 | 2B | **49.26** | **14.39** | **48.40** | **48.40** | 14.59 | 5.47 | 13.93 | 13.89 | 14.43 | 5.23 | 13.61 | 13.60 | 17.18 | 6.56 | 16.46 | 16.53 | 8.72 | 3.49 | 7.89 | 8.04 |
| Gemma-3 | 1B | **38.37** | **11.85** | **37.45** | **37.44** | 25.68 | 7.67 | 24.36 | 24.46 | 28.71 | 8.06 | 27.72 | 27.76 | 32.66 | 10.05 | 31.61 | 31.61 | 17.12 | 7.45 | 15.90 | 15.98 |

Table 2: ROUGE scores across SLMs for all languages.

do not support Sanskrit and thus produce no predictions for 90% of the samples. Param-2-17B performs better on prompts in Indian languages than on English prompts, while Sarvam-1 performs better on Marathi prompts. However, both Param-2 and Sarvam-1 have ROUGE-Lsum scores below 10% across all languages. Hence, it is very difficult to draw any conclusions from those models.

An improvement in performance of tasks on Indian languages is observed from Gemma-2-2B and Gemma-3-1B. Table A1 shows that while the fertility rate of English remains the same, the fertility rate of Bangla, Marathi, Telugu and Sanskrit decreased by 117%, 50.88%, 67.96% and 25.13% respectively. The performance increase between Gemma-3-12B and Gemma-4-31B is $\sim$3% to $\sim$24% across different Indian languages. The fertility rates of the Gemma-3 and Gemma-4 variants are identical, suggesting that tokenization efficiency may have played a pivotal role in improving performance of Gemma models for prompting in Indian languages. Gemma-4-31B is the best-performing model for prompts in Indian languages. Its performance is robust across languages with degradation of 5.15% (Marathi) to 15.75% in Sanskrit from English.

We further instruction-tuned LLaMA-3.2-1B on the Bangla and English subsets of the Updesh dataset (Chitale et al., 2025). The hyperparameters for instruction tuning are described in App A.3. We then prompted LLaMA-3.2-1B with the same 80 Bangla and English instructions. After instruction tuning on Bangla, the performance of LLaMA-3.2-1B improves by 18.20% on the Bangla prompt while performance on the English prompt remains the same. However, the performance on the English prompt still exceeds that on the Bangla prompt by 59.05%. Prompting the model in Bangla after instruction tuning on the English subset of Updesh degrades its performance relative to the base model by 12.80%. The performance on English increases by only 3.56%, indicating that English has reached its optimal performance. These results indicate that the disparity in model performance for following instructions in Indian languages compared to English is not due to insufficient instruction tuning for Indian languages.

| Metric | IFT on Bangla Subset | | IFT on English Subset | |
|---|---|---|---|---|
| | Bangla | English | Bangla | English |
| ROUGE-1 | 12.75 | **19.74** | 9.76 | **20.12** |
| ROUGE-2 | 4.03 | **5.28** | 3.42 | **5.88** |
| ROUGE-L | 11.78 | **18.74** | 8.70 | **18.93** |
| ROUGE-Lsum | 11.82 | **18.80** | 8.72 | **18.91** |

Table 3: Performance of LLaMA-3.2-1B after instruction tuning on Bangla and English subsets of Updesh dataset.

## 5 Conclusions

In this work, we examined whether language models follow semantically equivalent instructions equally well across English and Indian languages. Using 80 instructions sampled from the Super-NaturalInstructions dataset and translated into Bangla, Telugu, Marathi, and Sanskrit, we evaluated both LLMs and SLMs under comparable prompting conditions. Our experiments show that English prompts consistently produce stronger downstream task performance than Indian-language prompts. The performance gap is particularly severe for SLMs with $\leq$4B parameters. Gemma-4-31B exhibits comparatively stable behaviour across English, Marathi, Bangla, Telugu, and Sanskrit. To investigate the role of instruction tuning, we fine-tuned LLaMA-3.2-1B on Bangla and English subsets of UPDESH. The results show that instruction tuning can improve performance on prompting in Indian languages while the performance on English remains nearly equal. However, Bangla prompting still lags behind English by 59.05%. This indicates that the performance disparity is not merely a consequence of insufficient instruction tuning. In the future, we would like to extend this evaluation to more Indian languages and include a deeper analysis of the composition of the training data.

## Limitations

Our study has a few limitations. We have not evaluated the proprietary models primarily because of computational and financial constraints. However, we have experimented with open-source models of different architectures and parameter settings and found that the trend is generalizable. Many recent LLMs and SLMs do not release their training data. As a result, we could not evaluate the effect of data bias on the results.

## Ethics Statement

All datasets used in this work were obtained from publicly available sources and used in accordance with their respective licenses. To the best of our knowledge, the data do not contain personally identifying information or offensive content. The manually corrected instruction dataset does not have foreseen ethical concerns as the instructions are very basic. The experts employed in manual correction are co-authors of this paper. Large language models and other AI tools, were used to assist with code development, and manuscript refinement. All generated outputs were carefully verified by the authors. The research ideas solely belong to the authors.

## References


Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, Alon Benhaim, Misha Bilenko, Johan Bjorck, Sébastien Bubeck, Martin Cai, Qin Cai, Vishrav Chaudhary, Dong Chen, Dongdong Chen, and 110 others. 2024. Phi-3 technical report: A highly capable language model locally on your phone. *Preprint*, arXiv:2404.14219.

Pranjal A Chitale, Varun Gumma, Sanchit Ahuja, Prashant Kodali, Manan Uppadhyay, Deepthi Sudharsan, and Sunayana Sitaram. 2025. Updesh: Synthesizing grounded instruction tuning data for 13 indic languages. *arXiv preprint arXiv:2509.21294*.

Jack FitzGerald, Christopher Hench, Charith Peris, Scott Mackie, Kay Rottmann, Ana Sanchez, Aaron Nash, Liam Urbach, Vishesh Kakarala, Richa Singh, and 1 others. 2023. Massive: A 1m-example multilingual natural language understanding dataset with 51 typologically-diverse languages. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4277–4302.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*.

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. Xtreme: A massively multilingual multi-task benchmark for evaluating cross-lingual generalisation. In *International conference on machine learning*, pages 4411–4421. PMLR.

Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, Gaël Liu, and 196 others. 2025. Gemma 3 technical report. *Preprint*, arXiv:2503.19786.

Yaobo Liang, Nan Duan, Yeyun Gong, Ning Wu, Fenfei Guo, Weizhen Qi, Ming Gong, Linjun Shou, Daxin Jiang, Guihong Cao, and 1 others. 2020. Xglue: A new benchmark dataset for cross-lingual pre-training, understanding and generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6008–6018.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Aixin Liu, Bei Feng, Bin Wang, Bingxuan Wang, Bo Liu, Chenggang Zhao, Chengqi Dengr, Chong Ruan, Damai Dai, Daya Guo, Dejian Yang, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, and 137 others. 2024. Deepseek-v2: A strong, economical, and efficient mixture-of-experts language model. *Preprint*, arXiv:2405.04434.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. *Preprint*, arXiv:1711.05101.

Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng-Xin Yong, Hailey Schoelkopf, and 1 others. 2023. Crosslingual generalization through multitask finetuning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15991–16111.

William Oliveira. 2026. Less is more: Engineering challenges of on-device small language model integration in a mobile application. *Preprint*, arXiv:2604.24636.

Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, Johan Ferret, Peter Liu, Pouya Tafti, Abe Friesen, Michelle Casbon, Sabela Ramos, Ravin Kumar, Charline Le Lan, Sammy Jerome, Anton Tsitsulin, and 178 others. 2024. Gemma 2: Improving open language models at a practical size. *Preprint*, arXiv:2408.00118.

Yizhong Wang, Swaroop Mishra, Pegah Alipoormolabashi, Yeganeh Kordi, Amirreza Mirzaei, Atharva Naik, Arjun Ashok, Arut Selvan Dhanasekaran, Anjana Arunkumar, David Stap, and 1 others. 2022. Super-NaturalInstructions: Generalization via declarative instructions on 1600+ NLP tasks. In *Proceedings of the 2022 conference on empirical methods in natural language processing*, pages 5085–5109.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025a. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, and 23 others. 2025b. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

# A Appendix

| Model | Variant | English | Bangla | Marathi | Telugu | Sanskrit |
|---|---|---|---|---|---|---|
| Qwen-2.5 | All | 1.30 | 6.68 | 5.81 | 10.41 | 7.20 |
| Qwen-3.5 | All | 1.30 | 4.21 | 4.27 | 5.80 | 5.45 |
| SmolLM | 3B | 1.30 | 7.67 | 3.51 | 12.12 | 4.25 |
| Olmo-3 | 7B | 1.30 | 7.74 | 6.20 | 12.12 | 7.83 |
| Deepseek-V2-Lite | 16B | 1.32 | 11.43 | 7.30 | 14.27 | 9.09 |
| LLaMA-3.1 | 8B | 1.30 | 7.67 | 3.51 | 12.13 | 4.25 |
| LLaMA-3.2 | 1B, 2B | 1.30 | 7.67 | 3.51 | 12.13 | 4.25 |
| Phi-3.5 | 3.8B | 1.40 | 8.04 | 6.45 | 17.95 | 8.09 |
| Param-2 | 17B | 1.38 | 1.83 | 1.85 | 2.24 | 2.30 |
| Sarvam-1 | 2B | 1.42 | 1.72 | 1.75 | 2.06 | 3.05 |
| Sarvam-m | 24B | 1.31 | 2.57 | 2.70 | 3.51 | 3.72 |
| Gemma-2 | 2B, 9B | 1.30 | 3.39 | 2.79 | 4.30 | 3.53 |
| Gemma-3 | 1B, 12B | 1.30 | 1.56 | 1.85 | 2.56 | 2.82 |
| Gemma-4 | 31B | 1.30 | 1.56 | 1.85 | 2.56 | 2.82 |

Table A1: Fertility rates across languages for different language models.

## A.1 Examples of Prompts

Table A2 shows anecdotal examples of prompts used for evaluating language models across languages.

### A.1.1 Data

Table A3 shows the number of words and tokens generated by models for 80 instructions across the 5 languages.

Table A1 shows the fertility rates models for all the languages on 80 instructions.

## A.2 Performance of models on Indian language prompting

Table A4 demonstrates the performance differences of different models when prompting in Indian languages compared to English. From Table A4, it is clear that the performance of models in following instructions in English is significantly better than in Indian languages.

## A.3 Hyperparameters

We instruction-tuned LLaMA-3.2-1B model with AdamW optimizer (Loshchilov and Hutter, 2019) and learning rate of 2e-5 for 5 epochs. The batch size of the model is 1 with a gradient accumulation step of 4. The maximum sequence length is set at 2,048. All training and evaluation has been done on a single Nvidia A40 48GB gpu.

| Language | Instruction |
|---|---|
| English | You are given a sentence in English. Your job is to translate the English sentence into Polish. |
| Marathi | तुम्हाला इंग्रजीतील वाक्य दिले आहे. तुमचे काम इंग्रजी वाक्याचे पोलिशमध्ये भाषांतर करणे आहे.<br>(tumhālā iṁgrajītīla vākya dilē āhē. tumacē kāma iṁgrajī vākyācē pōliśamadhyē bhāṣāṁtara karaṇē āhē) |
| Bangla | আপনাকে ইংরেজিতে একটি বাক্য দেওয়া হয়েছে। আপনার কাজ হল ইংরেজি বাক্যটি পোলিশ ভাষায় অনুবাদ করা।<br>(āpanākē iṁrējitē ēkaṭi bākya dēōẏā haẏēchē. āpanāra kāja hala iṁrēji bākyaṭi pōliśa bhāṣāẏa anubāda karā.) |
| Telugu | మీకు ఆంగ్ల భాషలో ఒక వాక్యం ఇవ్వబడుతుంది. ఆ ఆంగ్ల వాక్యాన్ని పోలిష్ భాషలోకి అనువదించడం మీ పని.<br>mīku āṁgla bhāṣalō oka vākyaṁ ivvabaḍutuṁdi. ā āṁgla vākyānni pōliṣ bhāṣalōki anuvadiṁcaḍaṁ mī pani. |
| Sanskrit | त्वं आङ्ग्लभाषायां वाक्यं प्राप्तवान्। त्वया आङ्ग्लवाक्यं पोलिशभाषायां अनुवादनीयम्।<br>(tvaṁ āṅglabhāṣāyāṁ vākyaṁ prāptavān. tvayā āṅglavākyaṁ pōliśabhāṣāyāṁ anuvādanīyam.) |
| English | Given a phrase describing the relationship between two words, extract the words and the lexical relationship between them. The relation has to be of the type 'MemberOf', 'MadeOf', 'Synonym', 'Entails', 'HasA', 'HasProperty', 'PartOf', 'Antonym' or 'IsA'. The output should have the format: word1 relation word2. |
| Marathi | दोन शब्दांमधील संबंधाचे वर्णन करणारा वाक्यांश दिला आहे; त्यातून शब्द आणि त्यांच्यातील शब्दार्थिक संबंध काढा. संबंधाचा प्रकार 'MemberOf', 'MadeOf', 'Synonym', 'Entails', 'HasA', 'HasProperty', 'PartOf', 'Antonym' किंवा 'IsA' यापैकी एक असावा. आउटपुटचे स्वरूप: शब्द1 संबंध शब्द2.<br>(dōna śabdāṁmadhīla saṁbaṁdhācē varṇana karaṇārā vākyāṁśa dilā āhē; tyātūna śabda āṇi tyāṁcyātīla śabdārthika saṁbaṁdha kāḍhā. saṁbaṁdhācā prakāra `MemberOf', `MadeOf', `Synonym', `Entails', `HasA', `HasProperty', `PartOf', `Antonym' kiṁvā `IsA' yāpaikī ēka asāvā. āuṭapuṭacē svarūpa: śabda1 saṁbaṁdha śabda2.) |
| Bangla | দুটি শব্দের মধ্যে সম্পর্ক বর্ণনা করা একটি বাক্যাংশ দেওয়া হলে, সেই শব্দগুলি এবং তাদের মধ্যে আভিধানিক সম্পর্ক বের করুন। সম্পর্কটি অবশ্যই 'MemberOf', 'MadeOf', 'Synonym', 'Entails', 'HasA', 'HasProperty', 'PartOf', 'Antonym' অথবা 'IsA' ধরনের হতে হবে। আউটপুটের ফরম্যাট হওয়া উচিত: শব্দ1 সম্পর্ক শব্দ2।<br>(duṭi śabdēra madhyē samparka barṇanā karā ēkaṭi bākyāṁśa dēōẏā halē, sēi śabdaguli ēbaṁ tādēra madhyē ābhidhānika samparka bēra karuna. samparkaṭi abaśya:i `MemberOf', `MadeOf', `Synonym', `Entails', `HasA', `HasProperty', `PartOf', `Antonym' athabā `IsA' dharanēra hatē habē. āuṭapuṭēra pharamyāṭa haōẏā ucita: śabda1 samparka śabda2.) |
| Telugu | రెండు పదాల మధ్య ఉన్న సంబంధాన్ని వివరించే ఒక వాక్యాన్ని ఇచ్చినప్పుడు, ఆ పదాలను మరియు వాటి మధ్య ఉన్న శబ్ద సంబంధాన్ని గుర్తించి తీసుకోండి. ఆ సంబంధం 'MemberOf', 'MadeOf', 'Synonym', 'Entails', 'HasA', 'HasProperty', 'PartOf', 'Antonym' లేదా 'IsA' రకాలలో ఏదో ఒకటి కావాలి. అవుట్‌పుట్ ఈ ఫార్మాట్‌లో ఇవ్వాలి: పదం1 సంబంధం పదం2.<br>reṁḍu padāla madhya unna saṁbaṁdhānni vivariṁcē oka vākyānni iccinappuḍu, ā padālanu mariyu vāṭi madhya unna śabda saṁbaṁdhānni gurtiṁci tīsukōṁḍi. ā saṁbaṁdhaṁ `MemberOf', `MadeOf', `Synonym', `Entails', `HasA', `HasProperty', `PartOf', `Antonym' lēdā `IsA' rakālalō ēdō okaṭi kāvāli. avuṭpuṭ ī phārmāṭlō ivvāli: padaṁ1 saṁbaṁdhaṁ padaṁ2. |
| Sanskrit | द्वयोः शब्दयोः सम्बन्धं वर्णयतः वाक्यांशः दत्तः, तयोः शब्दौ च तेषां शब्दार्थसम्बन्धः च निष्कर्शनीयाः। सम्बन्धः 'MemberOf', 'MadeOf', 'Synonym', 'Entails', 'HasA', 'HasProperty', 'PartOf', 'Antonym' अथवा 'IsA' इति प्रकारेषु स्यात्। निर्गमस्य स्वरूपम् इत्थं भवेत्: शब्दः १ सम्बन्धः शब्दः २।<br>dvayōḥ śabdayōḥ sambandhaṁ varṇayataḥ vākyāṁśaḥ dattaḥ, tayōḥ śabdau ca tēṣāṁ śabdārthasambandhaḥ ca niṣkarśanīyāḥ. sambandhaḥ `MemberOf', `MadeOf', `Synonym', `Entails', `HasA', `HasProperty', `PartOf', `Antonym' athavā `IsA' iti prakārēṣu syāt. nirgamasya svarūpam itthaṁ bhavēt: śabdaḥ 1 sambandhaḥ śabdaḥ 2. |

Table A2: Anecdotal examples of instruction prompts across languages across tasks.

| Tokenizer | English | | Bangla | | Marathi | | Telugu | | Sanskrit | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Words | Tokens | Words | Tokens | Words | Tokens | Words | Tokens | Words | Tokens |
| Qwen-3.5-9B | 4077 | 5293 | 3652 | 15390 | 3203 | 13685 | 3170 | 18386 | 2606 | 14206 |
| Qwen-2.5-7B | 4077 | 5319 | 3652 | 24384 | 3203 | 18615 | 3170 | 33000 | 2606 | 18752 |
| SmolLM-3B | 4077 | 5295 | 3652 | 28011 | 3203 | 11253 | 3170 | 38425 | 2606 | 11083 |
| Olmo-3-7B | 4077 | 5295 | 3652 | 28284 | 3203 | 19848 | 3170 | 38425 | 2606 | 20396 |
| Param-2-17B | 4077 | 5623 | 3652 | 6674 | 3203 | 5911 | 3170 | 7085 | 2606 | 5981 |
| Deepseek-16B | 4077 | 5380 | 3652 | 41733 | 3203 | 23387 | 3170 | 45223 | 2606 | 23680 |
| LLaMA-3.1-8B | 4077 | 5295 | 3652 | 28011 | 3203 | 11253 | 3170 | 38425 | 2606 | 11083 |
| Phi-3.5-3.8B | 4077 | 5689 | 3652 | 29365 | 3203 | 20659 | 3170 | 56916 | 2606 | 21089 |
| Sarvam-1-2B | 4077 | 5801 | 3652 | 6269 | 3203 | 5617 | 3170 | 6518 | 2606 | 7946 |
| Sarvam-m-24B | 4077 | 5329 | 3652 | 9374 | 3203 | 8650 | 3170 | 11134 | 2606 | 9697 |
| Gemma-2-2B | 4077 | 5289 | 3652 | 12365 | 3203 | 8933 | 3170 | 13624 | 2606 | 9199 |
| Gemma-3-1B | 4077 | 5284 | 3652 | 5695 | 3203 | 5914 | 3170 | 8121 | 2606 | 7354 |
| Gemma-4-31B | 4077 | 5284 | 3652 | 5695 | 3203 | 5914 | 3170 | 8121 | 2606 | 7354 |

Table A3: Word and token counts across languages for different model tokenizers.

| Family | Variant | Bangla | Marathi | Telugu | Sanskrit |
|---|---|---|---|---|---|
| Qwen-2.5 | 0.5B | 169.00% | 125.52% | 197.55% | 203.43% |
| Qwen-2.5 | 1.5B | 177.99% | 152.01% | 192.69% | 306.77% |
| Qwen-2.5 | 3.0B | 82.99% | 72.99% | 178.36% | 191.25% |
| Qwen-2.5 | 8B | 62.97% | 65.70% | 126.05% | 210.31% |
| Qwen-2.5 | 32B | 109.14% | 93.72% | 144.34% | 167.54% |
| Qwen-3.5 | 4B | 120.71% | 97.54% | 143.13% | 193.33% |
| Qwen-3.5 | 9B | 98.20% | 90.36% | 123.69% | 201.81% |
| Qwen-3.5 | 27B | 55.22% | 41.58% | 75.51% | 93.09% |
| LLaMA-3 | 1B | 82.60% | 58.78% | 78.49% | 99.56% |
| LLaMA-3 | 3B | 101.73% | 88.65% | 123.53% | 237.49% |
| LLaMA-3 | 8B | 65.95% | 81.72% | 84.00% | 200.33% |
| Gemma-2 | 2B | 248.45% | 192.80% | 255.62% | 501.99% |
| Gemma-2 | 9B | 83.60% | 70.74% | 82.06% | 170.14% |
| Gemma-3 | 1B | 53.07% | 18.44% | 34.87% | 134.29% |
| Gemma-3 | 12B | 10.74% | 8.73% | 6.75% | 32.83% |

Table A4: Percentage by which English prompting outperforms Indian language prompting across model families using ROUGE-Lsum.