Aligning Large Language Models to Low-Resource Languages through LLM-Based Selective Translation

Rakesh Paul, Anusha Kamath, Kanishk Singla, Raviraj Joshi, Utkarsh Vaidya,
Sanjay Singh Chauhan, Niranjan Wartikar
kanishks@nvidia.com
NVIDIA

IJCNLP AACL 2025 December 20-24 Mumbai, India

Abstract

The Problem: Multilingual LLMs struggle with low-resource languages, where native high-quality data is scarce and expensive to curate.

Translation Flaws: Standard translation of English data is a common workaround but often corrupts technical elements like code, math, and JSON.

The Solution: We propose LLM-based selective translation, a technique that translates natural text while strictly preserving non-translatable structures.

Methodology: The study specifically focuses on Hindi, benchmarking translations generated by Google Cloud Platform (GCP) against Llama-3.1-405B.

Optimization Strategies: We investigate critical implementation factors, including the necessity of filtering noisy outputs and the benefits of mixing translated samples with original English data.

Key Findings: Results confirm that selective translation is a practical and effective method for bridging the alignment gap in multilingual models.

Our Approach

Motivation: Lack of Multilingual Post-Training Datasets

- Significant performance gap exists in multilingual LLMs for low-resource languages.
- Lack of high-quality **alignment data (SFT/RLHF)** for low-resource languages hinders model performance.
- Existing datasets have limited coverage and poor data quality.
- Large-scale alignment requires **at least 100k samples** each for SFT and RLHF.
- Collecting and labeling data in that scale for non-English languages is costly and time-consuming.

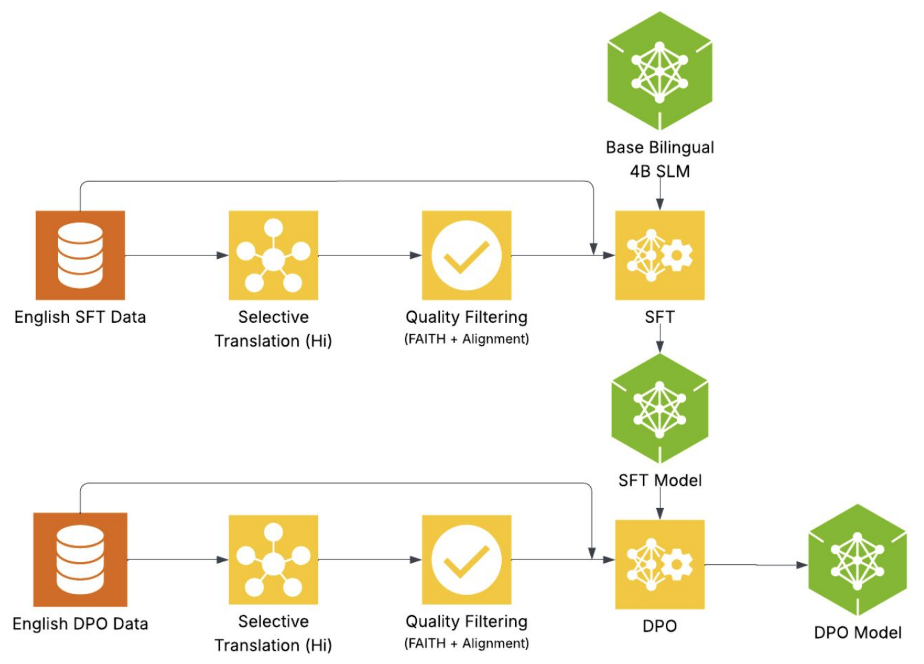


Figure 3: Overall training pipeline comprising translation, filtering, SFT, and DPO stages.

Our Approach:

- **Translation:** The English post-training dataset is translated using a selective translation approach powered by the Llama-3.1-405B model.
- **Filtering:** The translated data is filtered using an LLM-as-a-judge framework, where translation quality is scored by the LLM and samples falling below a predefined threshold are removed.
    a. **FAITH Filtering** - Rates translation quality on basis of Fluency, Accuracy, Idiomaticity, Terminology, Handling of Format.
    b. **Alignment Filtering (Prompt-Response)** - Coherence between the translated query and translated response.
- **Training:** The same strategy is used for both supervised fine-tuning (SFT) and direct preference optimization (DPO) stages to train the 4B model.

Full Paper: https://arxiv.org/pdf/2507.14304

Selective Translation Pipeline

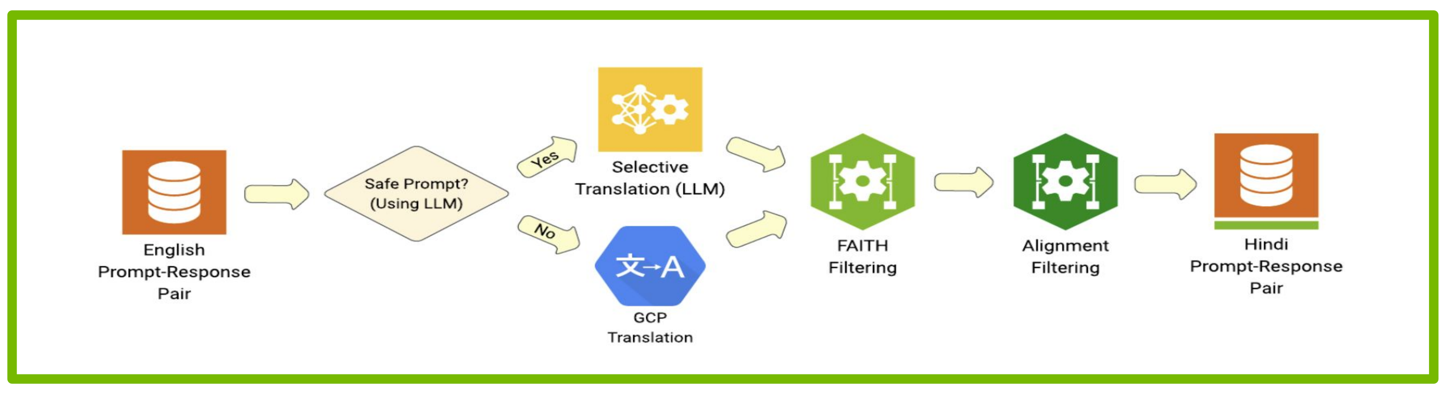


Figure 1: Hybrid approach for selective translation based data curation pipeline with safety considerations. The unsafe queries contain harmful, biased, or inappropriate content that LLMs typically decline to translate.

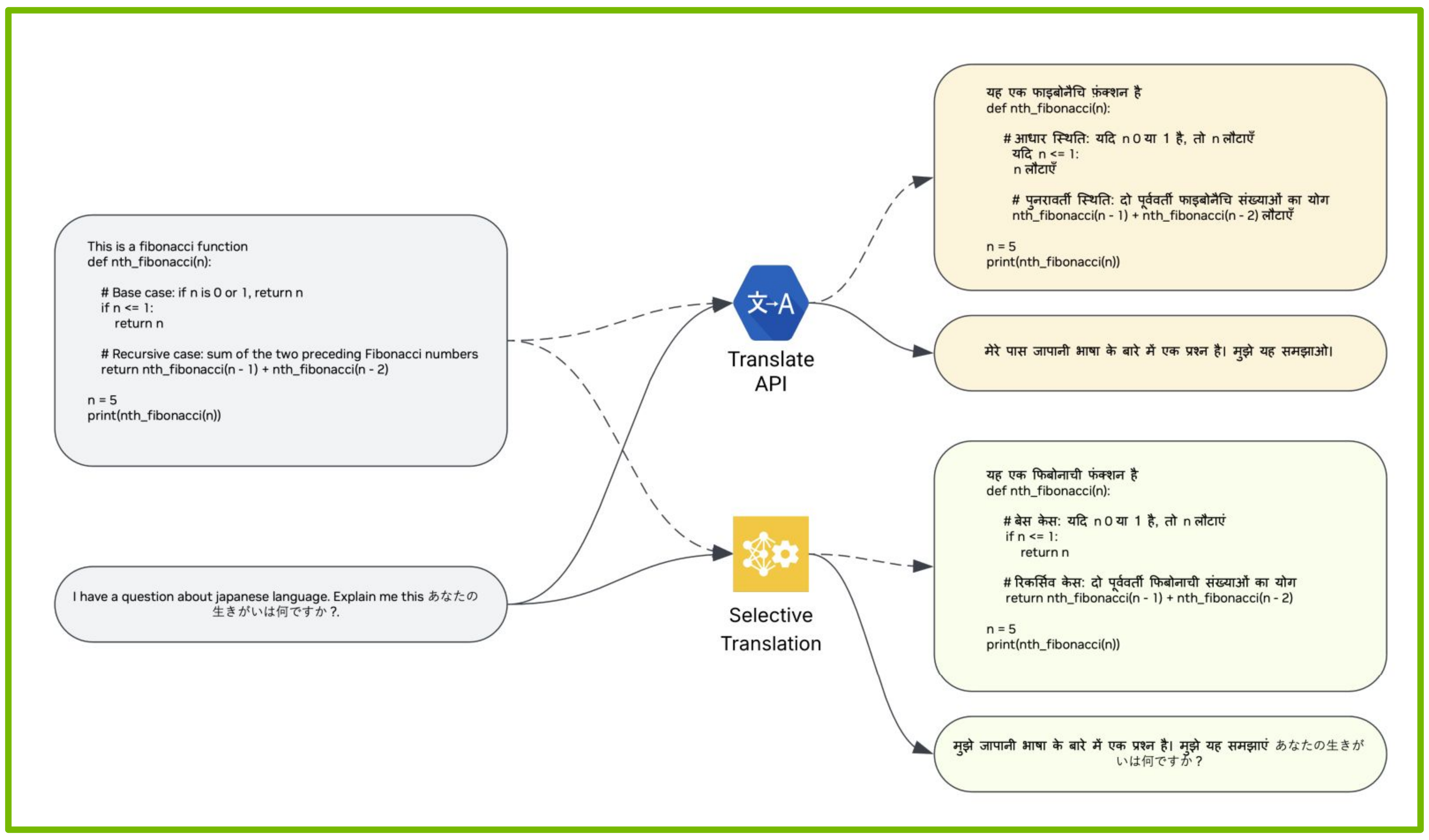


Figure 2: English to Hindi translation examples using LLM-based selective translation and vanilla GCP translation.

Quality Filtering

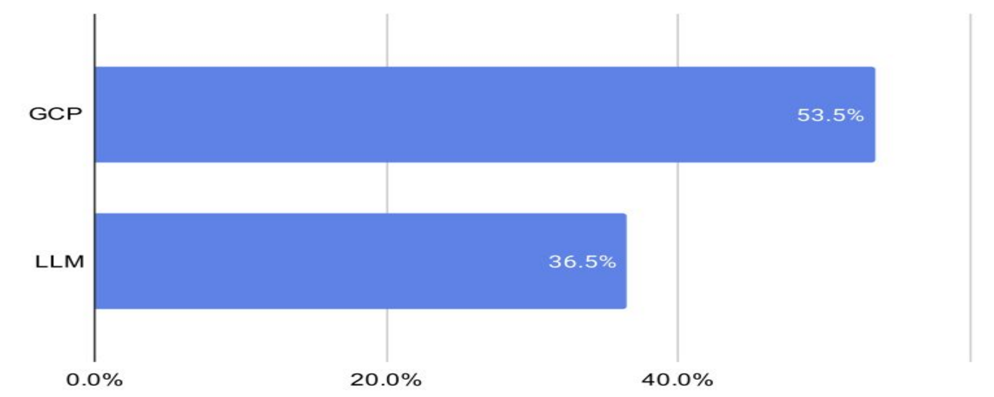


Figure 4: Percentage of LLM and GCP translated SFT data filtered by the Llama-3.1-Nemotron-70B-Instruct judge model, representing samples not achieving full scores in FAITH evaluation. (Lower the better)

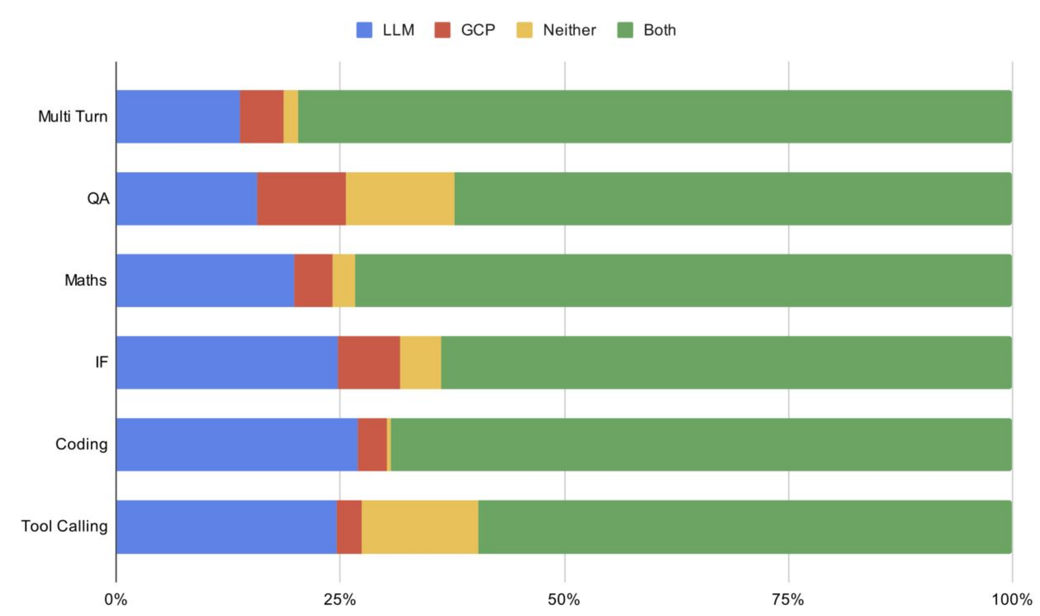


Figure 5: A/B comparison of translation quality, judged by Llama-3.1- Nemotron-70B-Instruct. The graph illustrates the percentage preference for LLM, GCP, both, or neither across various SFT dataset categories (Higher the better)

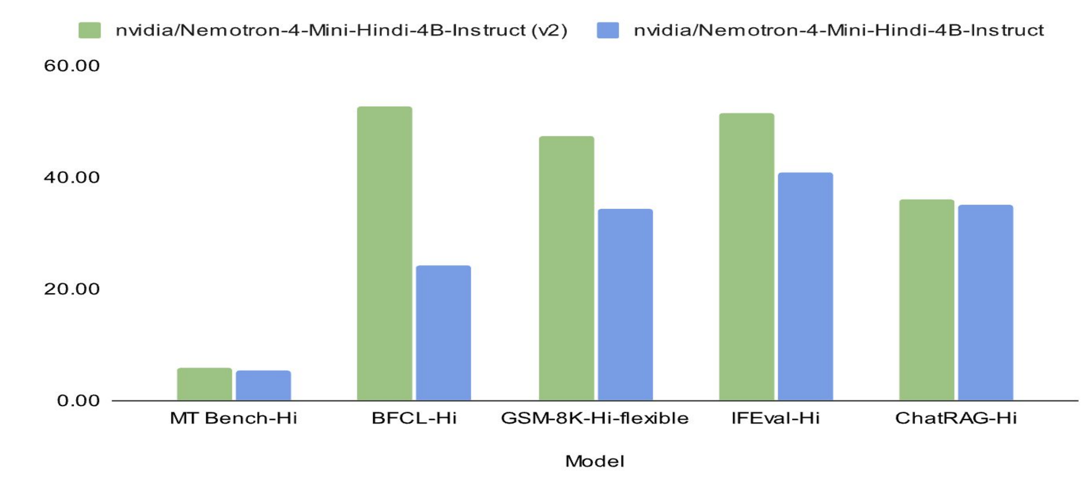
Training Config		SubjectiveEval	GSM8K-Hi	IFEval-Hi	MTBench-Hi	Fluency
Filtered SFT - Filtered DPO	LLM	4.37	43.44	55.51	4.97	4.50
	GCP	4.37	43.44	55.67	4.62	4.39
Unfiltered SFT - Unfiltered DPO	LLM	4.39	44.28	57.10	4.51	4.50
	GCP	4.24	43.59	58.84	5.01	4.42

Figure 6: Impact of Quality Filtering on Benchmark Performance Across Diverse Datasets

Key findings:

- LLM-based translations have always been preferred by the judge over GCP with **win rate of 17%**.
- These preferences are especially strong in coding, tool-calling, and mathematical data.
- LLM-based translations score **higher in fluency** and are preferred over GCP for key tasks.
- We are able to achieve the same accuracy using half the data, thus improving the **training efficiency.**

Results



Training Config		SubjectiveEval	GSM8K-Hi	IFEval-Hi	MTBench-Hi
200K En	–	3.71	30.10	44.17	3.44
200K En + 20K Hi	LLM	4.12	38.67	45.44	4.32
	GCP	4.02	36.32	43.77	4.10
200K En + 40K Hi	LLM	4.29	40.79	45.92	4.67
	GCP	4.24	37.45	44.65	4.37
200K En + 60K Hi	LLM	4.29	42.15	45.44	4.30
	GCP	4.13	38.36	45.04	4.26
200K En + 80K Hi	LLM	4.23	40.26	45.28	4.66
	GCP	3.92	39.58	45.12	4.04
200K En + 100K Hi	LLM	4.15	40.86	46.39	4.62
	GCP	3.98	40.71	44.65	4.17
200K En + 200K Hi	LLM	4.18	43.44	43.77	4.43
	GCP	4.05	44.50	46.63	4.55

Key Findings

- Models trained on Llama-3.1-405B translations outperform those using GCP translations across all benchmarks.
- Adding Hindi data alongside English greatly enhances performance, even with just 20k high quality samples.
- When applied to Nemotron-4-Mini-Hindi-4B, our selective translation recipe yielded a 35% gain on GSM8K, a 2× improvement in tool calling, and a 25% increase in instruction-following performance.

Conclusion and References

Conclusion

- We propose LLM based selective translation as a de-facto method for translating the alignment data.
- We introduce two novel quality filtering techniques specific to alignment data (FAITH & Alignment Filtering).
- Extensive ablation studies provide concrete evidence that our LLM-based approach is the key driver behind the significant improvements in the downstream tasks.

References

- Wen Lai, Mohsen Mesgar, and Alexander Fraser. 2024. Llms beyond english: Scaling the multilingual capability of llms with cross-lingual feedback. In Findings of the Association for Computational Linguistics ACL 2024, pages 8186–8213.
- Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the nlp world. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 6282–6293