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

Kanishk Singla | BHASHA @ IJCNLP-AACL 2025

Agenda

- LLMs for Low-Resource Languages
- Current Approach and Limitations
- Our Approach
  - Selective Translation
  - Quality Filter
  - Data Evaluation
- Results
- Application of our Recipe
- Key Learnings
- Conclusion

LLMs for Low-Resource Languages

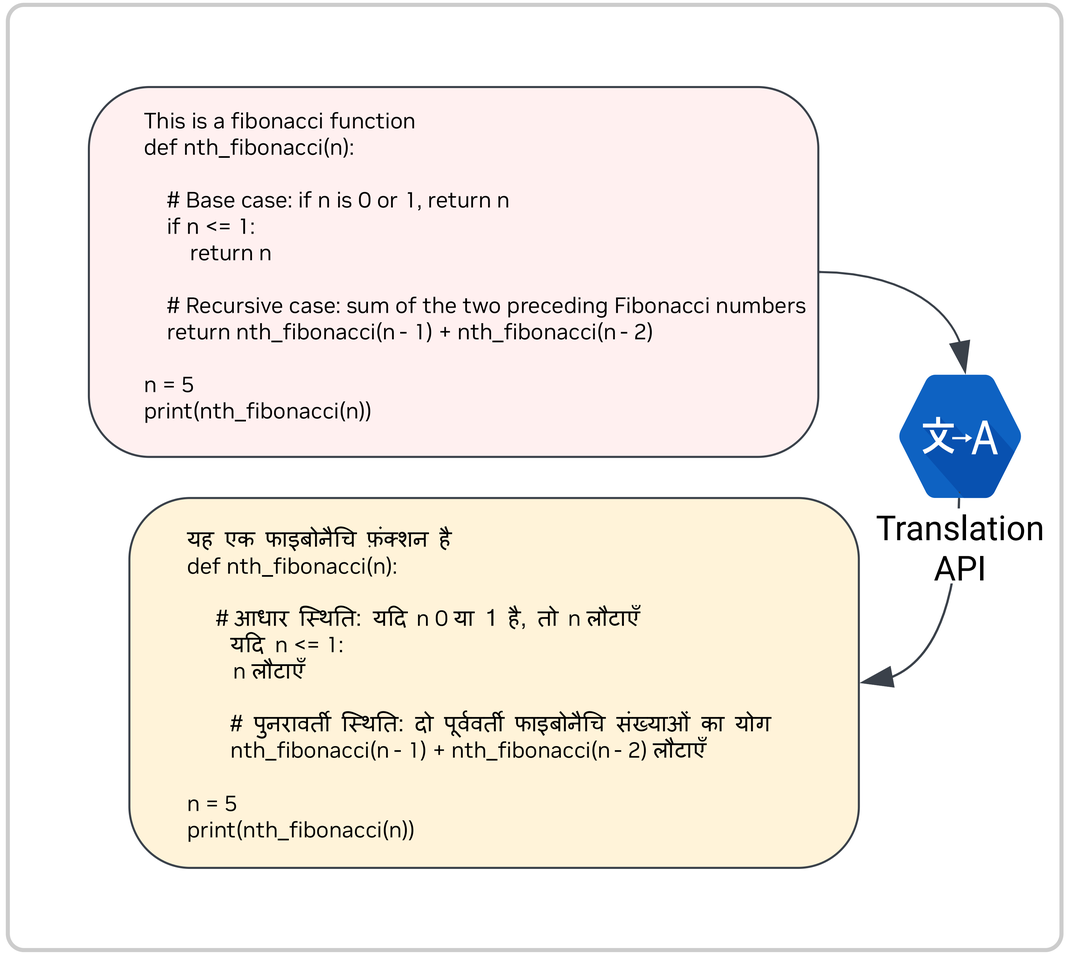
Sovereign AI Challenges

- Multilingual LLM: A language model inherently trained on 50+ languages to understand and generate text across them.
- Language Adaptation: Fine-tuning a multilingual LLM for specific languages or language groups to enhance its performance and accuracy.
- Despite the availability of many multilingual models, there is still a need for language-specific LLMs or language adaptation.

Why?
✓ 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.

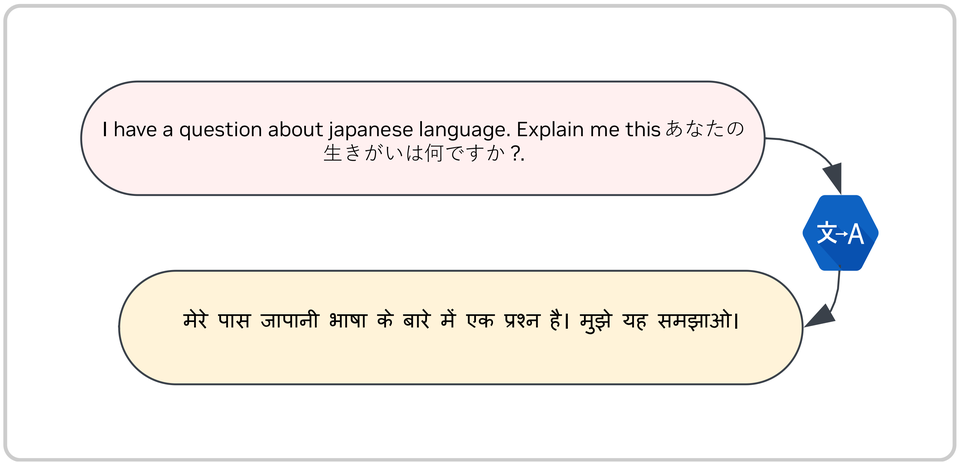
Current Approach: Standard Translation

Vanilla Machine Translation



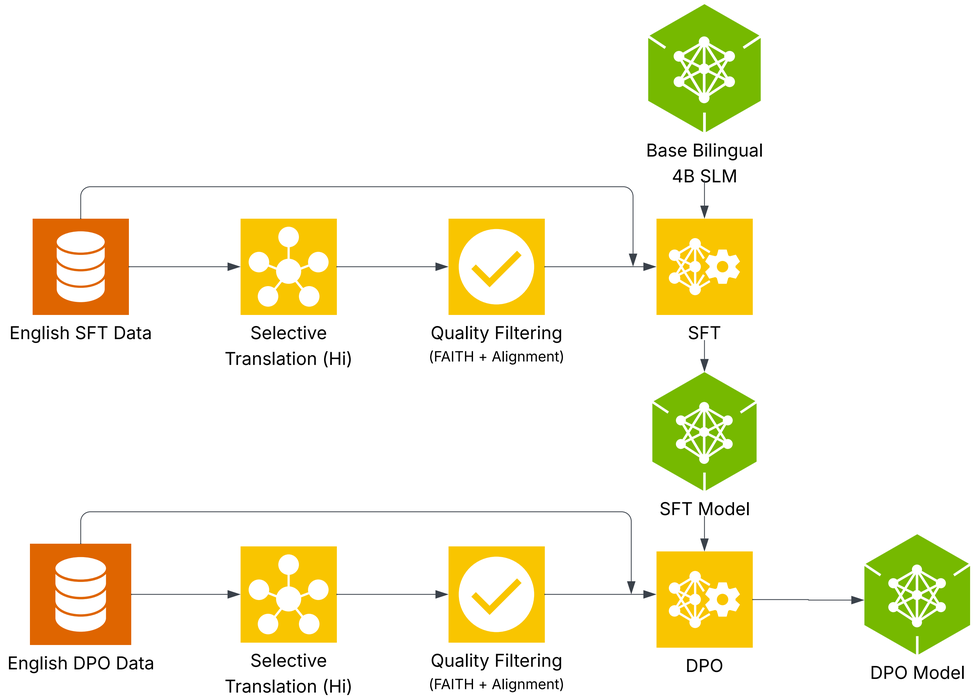
- Standard line-by-line translation often performs poorly in code, complex mathematics expressions, and complex structured text (e.g. JSON, XML) and textual tables

- Standard line by line translation miss out the context hence performs poorly in code-mixed text



Our Approach

Selective Translation + Quality Filtering

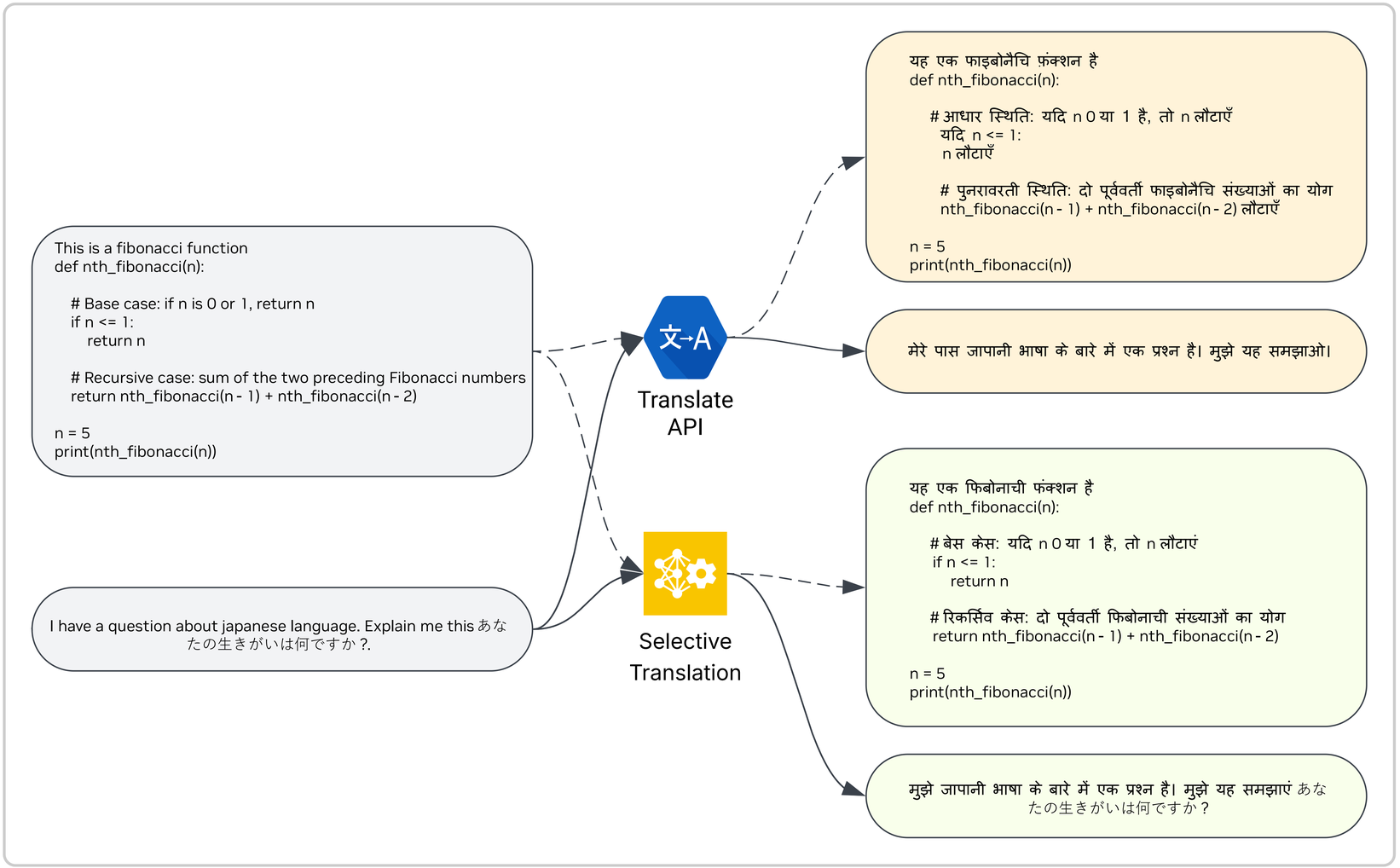


Overall training pipeline comprising translation, filtering, SFT, and DPO stages.

- ✓ Selective Translation + Dataset Blending
- ✓ Quality Filter
- ✓ Model Alignment
- ✓ Evaluation

Selective Translation

Selective translation in action

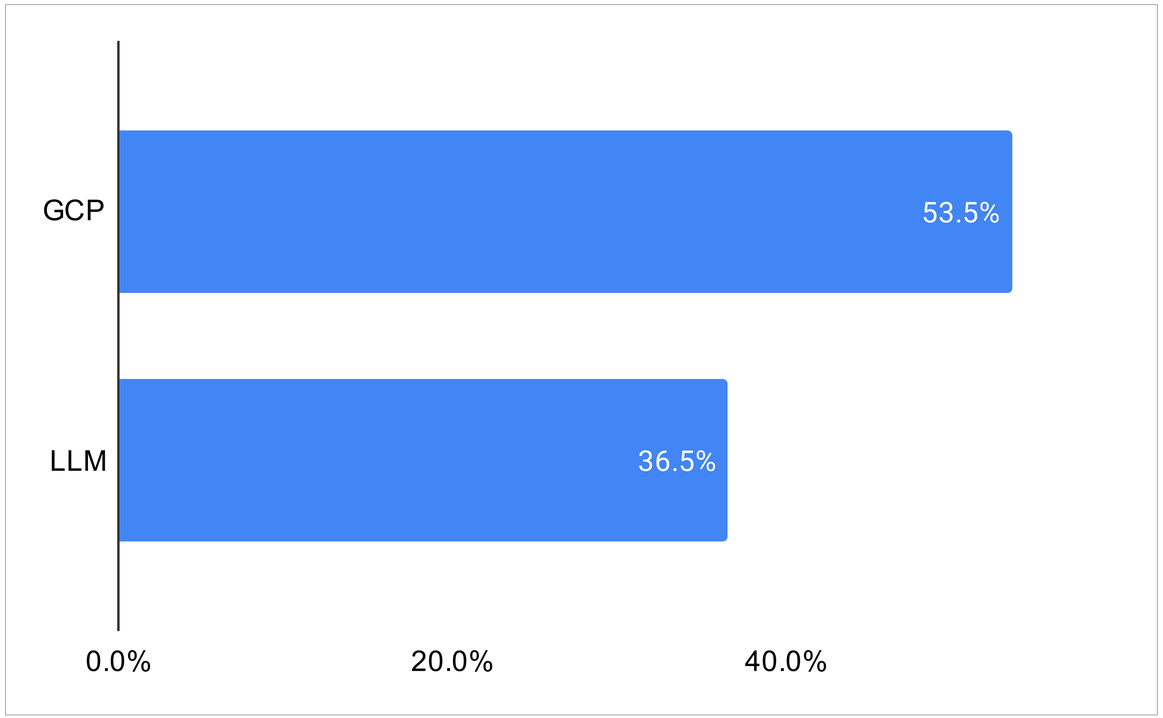


- It's a technique where a LLM is specifically instructed to translate only the linguistically adaptable portions of a given text, while meticulously preserving certain non-translatable elements.
- Non-translatable elements might include
  - Code snippets
  - Complex mathematical expressions
  - Tabular data
  - Tool calling data
  - Other formatted structured text (JSON, XML etc.)

We will be comparing **Google Translate from GCP (Google Cloud Platform)** vs **Llama 3.1 405B** as our Translator LLM

Quality Filter

Filtering low quality translated data



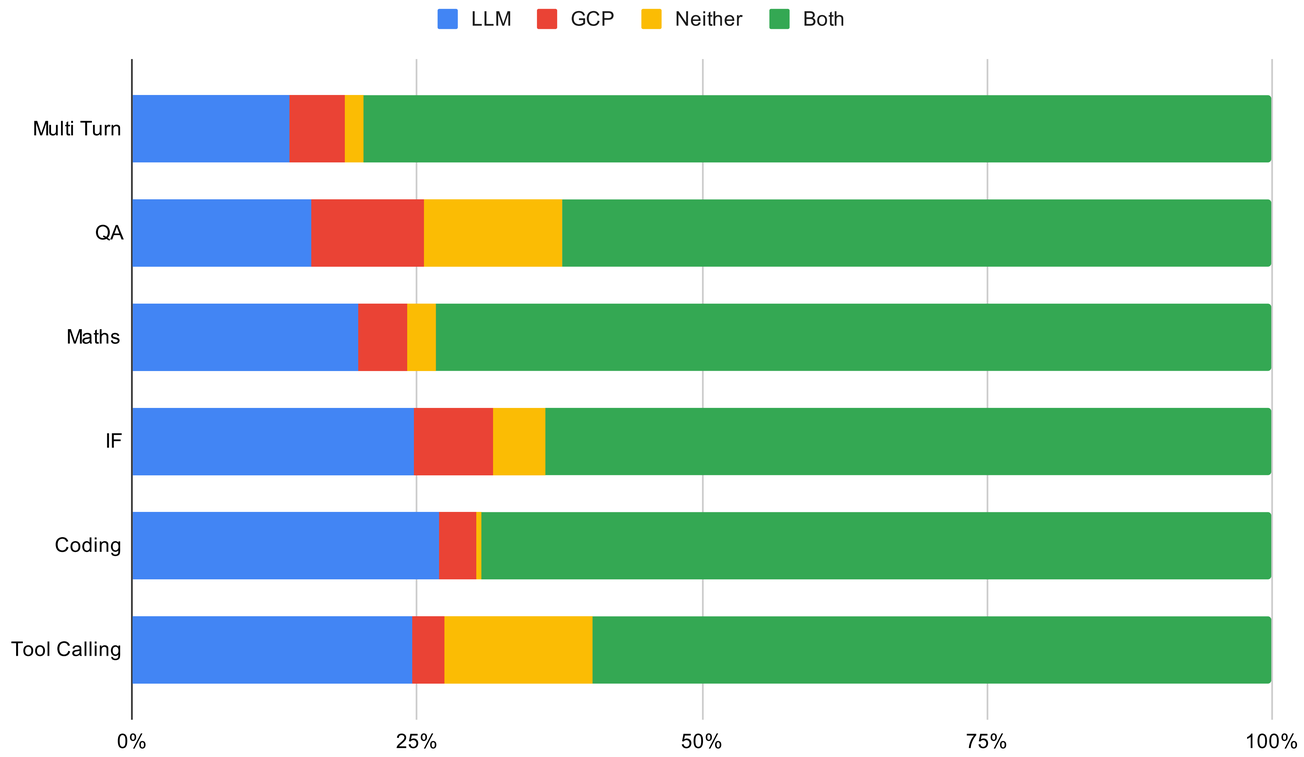
Percentage of LLM and GCP translated SFT data filtered by the Llama-3.1-Nemotron-70B-Instruct judge model, representing samples not achieving full scores in FAITH evaluation.

(Lower the better)

- **FAITH Filtering (Translation Quality)**
  - Fluency
  - Accuracy
  - Idiomaticity
  - Terminology
  - Handling of Format
- **Alignment Filtering (Prompt-Response Alignment)**
  - Coherence between the translated query and translated response

Data Evaluation

A/B comparison of GCP vs LLM translation data



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)

- LLM-based translations have always been preferred by the judge over GCP.
- These preferences are especially strong in coding, tool-calling, and mathematical data.

Results

Observations and inferences

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

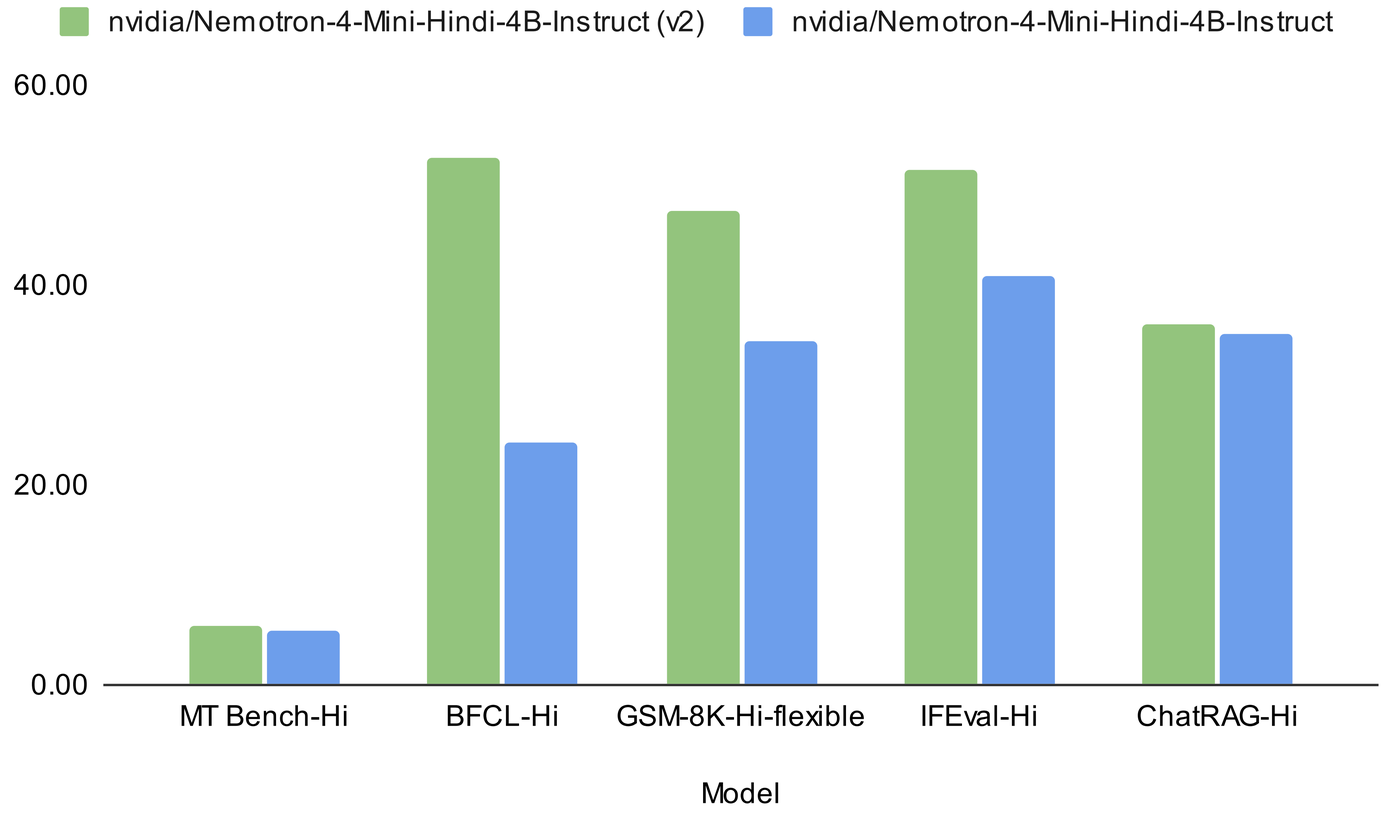
- Impact of English Alignment Data
  - Models trained on Llama-3.1-405B translations outperform those using GCP translations across all benchmarks.
  - Adding Hindi data alongside English greatly enhances performance, even with just **20k high quality samples**.
  - Accuracy keeps improving with more Hindi data, stabilizing around 60k samples.

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

- Impact of filtering and fluency analysis
  - 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**.

Application to Nemotron-Hindi-4B-Instruct

Open Model



- Here we have applied our selective translation recipe to our previously released **Nemotron-4-Mini-Hindi-4B model**, that was trained using standard translation data.
- We can see visible improvements across benchmarks in our latest **v2 model.**
- Major jumps were seen in:
  - BFCL (Tool-calling) : **~2x**
  - GSM8K (Math) : **~35%**
  - IFEval-Hi (Instruction Following): **~25%**

Best Practices & Key Learnings

For LLM alignment in low-resource languages,

- LLM-based selective translation significantly improves model performance.
- Mixing translated low-resource data with original English data is crucial for robust alignment.
- Filtering translated data for quality is effective and can make training more efficient.
- Even small amounts of high-quality translated data offer notable performance gains.

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.
- We created **Nemotron-Hindi-4B-Instruct-4B v2**, which sets a new performance standard by achieving significant gains on key benchmarks (GSM8K, BFCL, IFEVAL), surpassing its predecessor.
- This work provides a framework of best practices and key learnings for effectively aligning multilingual LLMs with low-resource languages.

Q&A

Selective Translation

```
You are a Hindi translation assistant. Your task is to translate the following text into Hindi,
while applying the following rules to determine when to skip translation for specific parts:

- Skip translating the following if they appear in the sentence:
  1. **Programming or coding content** (e.g., code snippets, commands) — retain this exactly as it is.
  2. **URLs, file paths, or email addresses** — leave these unchanged.
  3. **Strongly formatted data** such as tables, lists, or bullet points — maintain their structure and content as is.
  4. **Examples or phrases** where translation would alter their original meaning or usefulness.
  5. **Special characters, mathematical symbols, or technical abbreviations** — do not change these.
  6. **HTML/XML tags or other formatting markers** — keep these intact and unaltered.

As you translate, ensure that the output flows naturally and maintains the overall structure of the sentence.
Retain non-translatable elements exactly as they are, while translating the rest into Hindi.

Translate the following text:

Text: {{english_text}}

Only return the translated text!
If translation is not needed, return the input text as it-is!
```

Faith Filtering

```
Given the following sentences:

- Source : {{english_text}}
- Target [Hindi]: {{hindi_text}}

Please evaluate the translation using the FAITH metric. For each category, provide a score from 1 to 5 (1 = poor, 5 = excellent).
Only return the evaluation in the following JSON format:

{
  "Fluency": score,
  "Accuracy": score,
  "Idiomaticity": score,
  "Terminology": score,
  "Handling_of_Format": score
}

Here are the categories:

1. **Fluency (1-5)**: Does the translation read naturally in the target language, free from grammar or syntax errors?
   - 1: Very poor fluency, difficult to understand.
   - 2: Somewhat fluent but with major grammatical issues.
   - 3: Generally fluent with a few errors.
   - 4: Mostly fluent but may have minor grammatical issues.
   - 5: Perfect grammar, native-like fluency.

2. **Accuracy (1-5)**: How well does the translation preserve the meaning of the source sentence?
   - 1: Meaning significantly changed or lost.
   - 2: Major inaccuracies, important meanings are omitted.
   - 3: Some meaning preserved, but there are notable inaccuracies.
   - 4: Meaning mostly preserved with minor issues.
   - 5: Meaning fully preserved.

3. **Idiomaticity (1-5)**: Are the phrases idiomatic and natural for the target language,
    fitting its cultural context?
   - 1: Literal translation, very awkward for native speakers.
   - 2: Some idiomatic phrases but mostly awkward.
   - 3: Mixed idiomaticity, some phrases fit while others don't.
   - 4: Mostly idiomatic, with a few non-native phrases.
   - 5: Completely idiomatic and culturally appropriate.

4. **Terminology (1-5)**: Are any specialized terms translated accurately?
    (If no specialized terms, note as N/A.)
   - 1: Significant errors in terminology.
   - 2: Some incorrect terminology affecting understanding.
   - 3: Mostly correct terminology but with some inconsistencies.
   - 4: All terms correctly translated with minor inconsistencies.
   - 5: All terms correctly and consistently translated.

5. **Handling of Format (1-5)**: Is the formatting (punctuation, capitalization, non-translatable elements) correctly maintained?
   - 1: Significant formatting errors or omissions.
   - 2: Major formatting issues that affect readability.
   - 3: Some formatting errors, but generally readable.
   - 4: Minor formatting issues but mostly preserved.
   - 5: Format fully preserved.

In case there is no translation provided, give -1 to all the categories! If case of non-applicable score, make the score=0

Only return the evaluation JSON! No explanation!
```

Alignment Filtering

```
You are an evaluator tasked with assessing the quality of a response to a query using five key metrics:
Helpfulness, Correctness, Coherence, Complexity, and Verbosity. Provide a score for each metric on a scale of 1-5,
where 1 indicates poor performance and 5 indicates excellent performance. Then, summarize your reasoning for each score in a brief comment.

Query: {{hindi_prompt}}
Response: {{hindi_response}}

#### Definitions of Metrics and Scoring Guidelines:
- **Helpfulness**: Measures how useful and actionable the response is in addressing the query.
    - 1: Completely unhelpful or irrelevant.
    - 2: Slightly helpful but misses key aspects of the query.
    - 3: Moderately helpful but lacks depth or usability.
    - 4: Mostly helpful with minor gaps in utility.
    - 5: Extremely helpful, fully addressing the query with clear, actionable information.

- **Correctness**: Evaluates whether the response is factually accurate and free of errors.
    - 1: Contains major factual inaccuracies or misleading information.
    - 2: Includes some accurate information but has notable errors.
    - 3: Mostly accurate but with minor errors or omissions.
    - 4: Accurate with negligible issues.
    - 5: Completely accurate and reliable.

- **Coherence**: Assesses whether the response is logically structured and easy to follow.
    - 1: Illogical, disorganized, or hard to understand.
    - 2: Poorly structured with noticeable issues in logical flow.
    - 3: Somewhat coherent but with occasional disorganization.
    - 4: Mostly coherent and well-organized with minor issues.
    - 5: Perfectly coherent, logically structured, and easy to follow.

- **Complexity**: Measures whether the response appropriately balances depth and complexity for the query.
    - 1: Overly simplistic or excessively complicated without justification.
    - 2: Either too simple or too complex, with limited balance.
    - 3: Moderately balanced but could improve in complexity or simplicity.
    - 4: Mostly balanced, with only minor adjustments needed.
    - 5: Perfectly balanced, with the right level of complexity for the query.

- **Verbosity**: Evaluates whether the response is concise and avoids unnecessary elaboration.
    - 1: Excessively verbose or overly terse, failing to strike a balance.
    - 2: Somewhat verbose or overly brief with noticeable issues.
    - 3: Moderately concise but could improve in eliminating redundancy or brevity.
    - 4: Mostly concise with minor verbosity or brevity issues.
    - 5: Perfectly concise, providing just the right amount of information.

#### Output Format:
Provide the evaluation in the following JSON format:
{
    "Helpfulness": score,
    "Correctness": score,
    "Coherence": score,
    "Complexity": score,
    "Verbosity": score
}

In case there is no translation provided, give -1 to all the categories!
If case of non-applicable score, make the score=0

Only return the evaluation JSON! No explanation!
```

Fluency Evaluation

```
You are a helpful Evaluator. Your task is to critically assess the fluency of responses given by a model to user questions in Hindi.

You will be presented with a chat containing user question and bot response pairs in Hindi.
Your goal is to evaluate the fluency of the response on a scale of 1-5, with 1 being the lowest and 5 being the highest.
You are proficient in the Hindi language, so you should consider the nuances and context of the language in your evaluation.
Your evaluation should be based on the following criteria:

1. Grammar and Syntax: Is the response grammatically correct and properly structured in Hindi?
2. Fluency and Naturalness: Does the response sound natural and fluent, as if it were written or spoken by a native Hindi speaker?
3. Pacing and Readability: Is the response paced well and easy to read or understand for a Hindi-speaking audience?
4. Cohesion and Coherence: Are the ideas logically connected, and does the response flow smoothly?

You will rate each criterion individually and then provide an overall fluency rating from 1 to 5.

Here is the chat:

User Question:
{hindi_prompt}

Bot Response:
{hindi_response}

At the end, provide the ratings in a JSON format with appropriate keys and values.

Example JSON format:
"grammar_and_syntax": 4,
"fluency_and_naturalness": 5,
"pacing_and_readability": 4,
"cohesion_and_coherence": 5,
"overall": 4

Return the JSON object with the above 5 parameters, with all ratings as integers.
Do not include anything else.
```