

IMPROVING CROSS- LINGUAL REPRESENTATIONS OF MULTILINGUAL LLMs

Team: Polyglots



Our Team



H.W.K. Aravinda: 200045U



R.M.S.K. Karunathilake: 200296M



I.S.M. Rashad: 200525R

Supervisors:

- 1. Dr. Nisansa de Silva**
- 2. Dr. Surangika Ranathunga**
- 3. Dr. Rishemjit Kaur**

TABLE OF CONTENT

1.Introduction

2.Problem Statement

3.Objectives

4.Literature Review

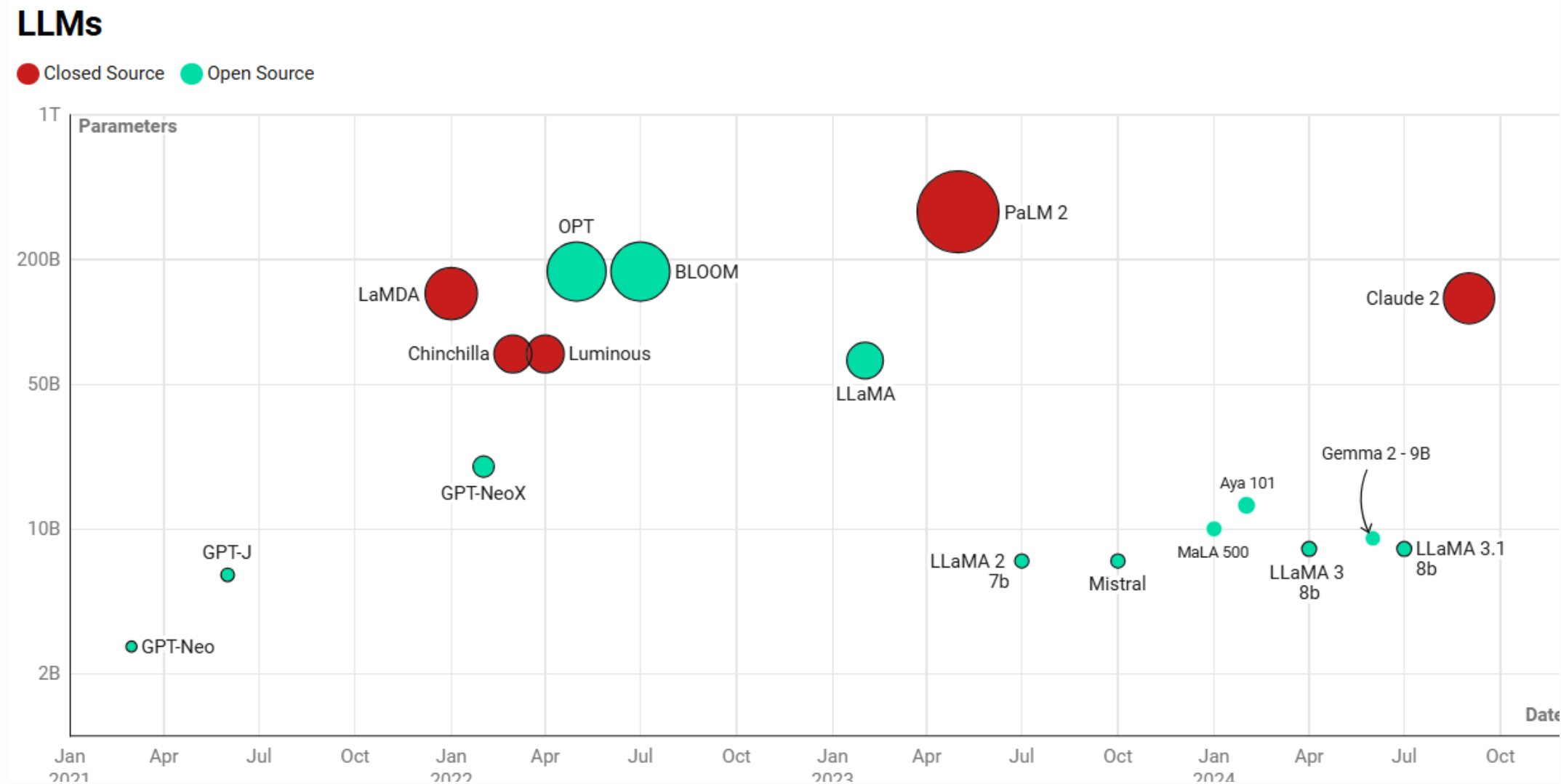
5.Methodology

6.Project Timeline

7.Feasibility

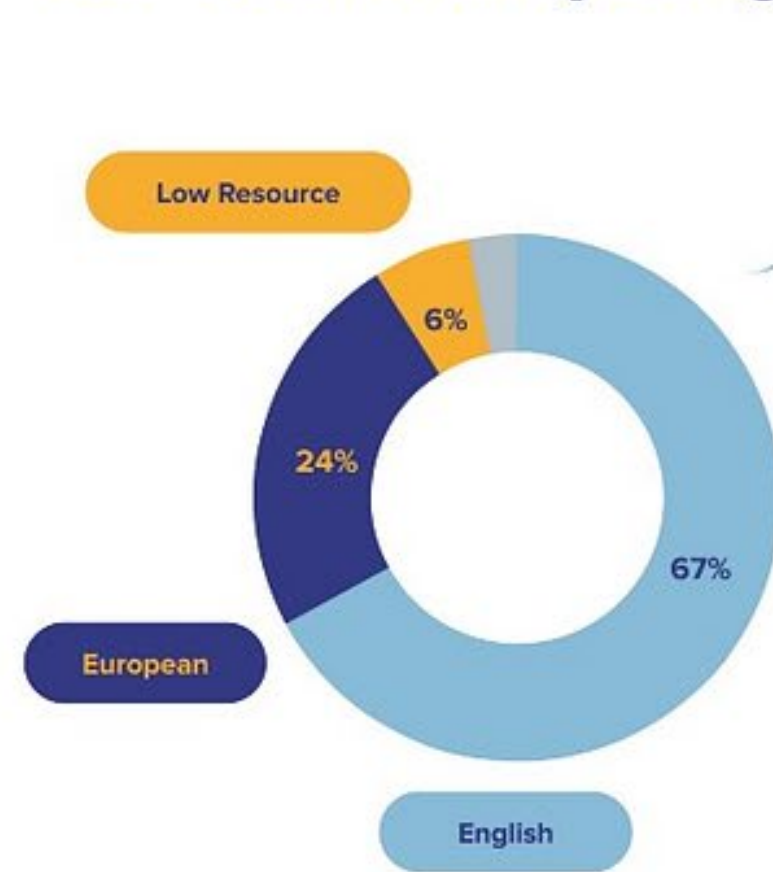
Introduction

- The SoTA technology for almost every known NLP task.
- These are typically decoder transformer[1] models.
- Autoregressive - generate text based on previously generated text
- Trained in a self-supervised manner using large amounts of data(mostly HR languages)

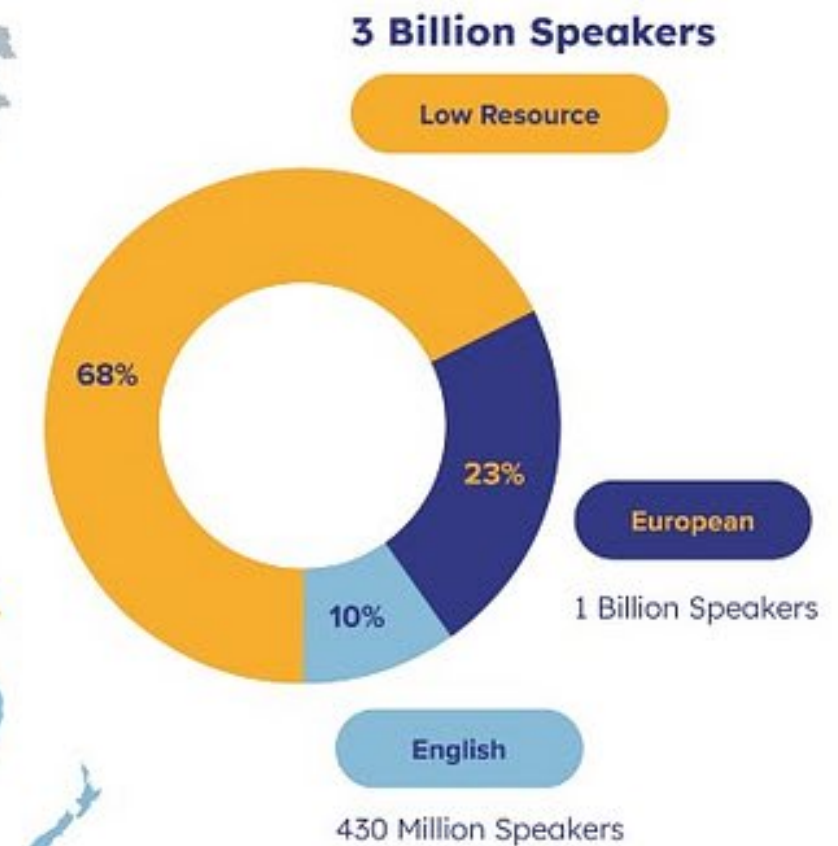


High Resource vs Low Resource Languages

NLP Solutions by Language



Population Size of Languages



- While LLMs perform exceptionally well for high-resource languages due to abundant training data, they struggle with low-resource languages [3] like **Sinhala**, **Tamil**, and **Punjabi**.

[2] <https://medium.com/neuralspace/low-resource-language-what-does-it-mean-d067ec85dea5>

[3] Lankford, S., Afli, H., & Way, A. (2023). adaptmllm: Fine-tuning multilingual language models on low-resource languages with integrated llm playgrounds. Information, 14(12), 638.

Problem Statement

How can the underrepresentation of languages like Sinhala, Tamil, and Punjabi in foundation models be addressed by investigating the limitations of existing approaches and exploring new methodologies to improve language model performance for these low-resource languages?

Objectives

- Improve the cross-lingual capabilities of low-resource languages by introducing improved fine-tuning methods and preference optimization techniques.
- Develop benchmarks for Low-resource languages such as Sinhala, Tamil, and Punjabi.

Language specific models

- **Typhoon** Thai Language models: Continued pre-training of Mistral -7b [4]
- **Swedish** models: Translated English instruction datasets into Swedish [5]
- **SambaLingo**: The vocabularies were extended to include language-specific tokens and further pre-trained in 10 languages. [6]

But there are limitations:

- Need a good GPU setup
- Can't use across languages (cross-lingual capabilities) needed for some tasks s.a.

Machine Translation

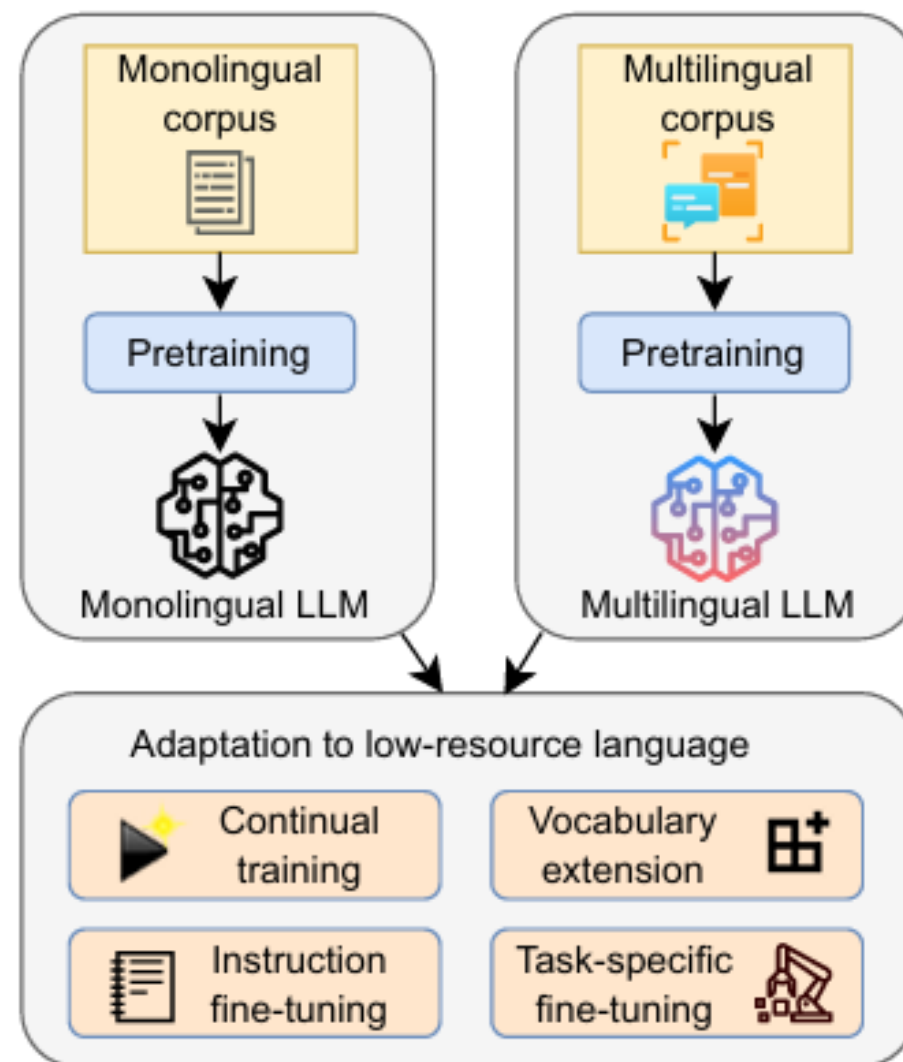
[4] K. Pipatanakul et al., 'Typhoon: Thai large language models', arXiv preprint arXiv:2312.13951, 2023.

[5] O. Holmström, J. Kunz, and M. Kuhlmann, 'Bridging the resource gap: Exploring the efficacy of English and multilingual LLMs for Swedish', in Proceedings of the Second Workshop on Resources and Representations for Under-Resourced Languages and Domains (RESOURCEFUL-2023), 2023, pp. 92–110.

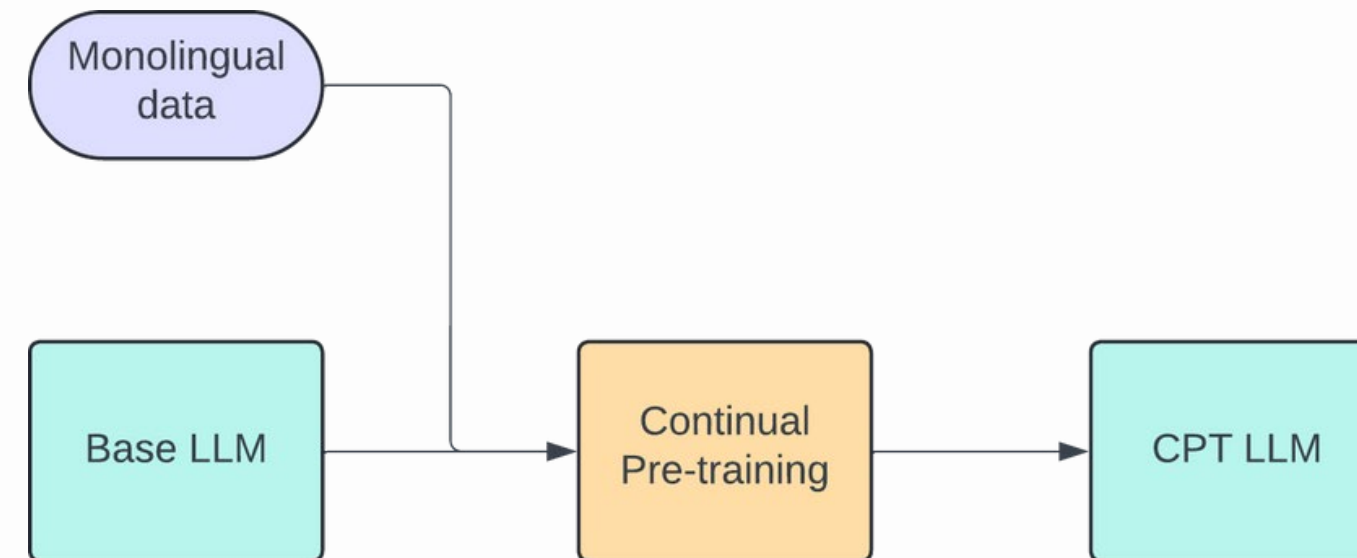
[6] Z. Csaki et al., 'Sambalingo: Teaching large language models new languages', arXiv preprint arXiv:2404.05829, 2024.

Possible techniques to enhance cross-lingual capabilities

A possible solution to enhance cross-lingual capabilities is to choose an existing model and identify the best techniques for different languages.



Pre-training



- Provide essential knowledge in the target language for better understanding .
- Enhance fluency and alignment between the target language and English .
- Use continual pre-training to update models with new data instead of retraining from scratch, reducing costs and improving efficiency .

ex:- Thai[8] and Hungarian[9] models

[8] Pipatanakul, K., Jirabovonvisut, P., Manakul, P., Sripaisarnmongkol, S., Patomwong, R., Chokchainant, P., & Tharnpipitchai, K. (2023). Typhoon: Thai large language models. arXiv preprint arXiv:2312.13951..

[9] Csaki, Z., Pawakapan, P., Thakker, U., & Xu, Q. (2023). Efficiently adapting pretrained language models to new languages. arXiv preprint arXiv:2311.05741.

Adapting the tokenizer

good morning to you

၁၁၁ ခု၁ ငု၁၁၁၁

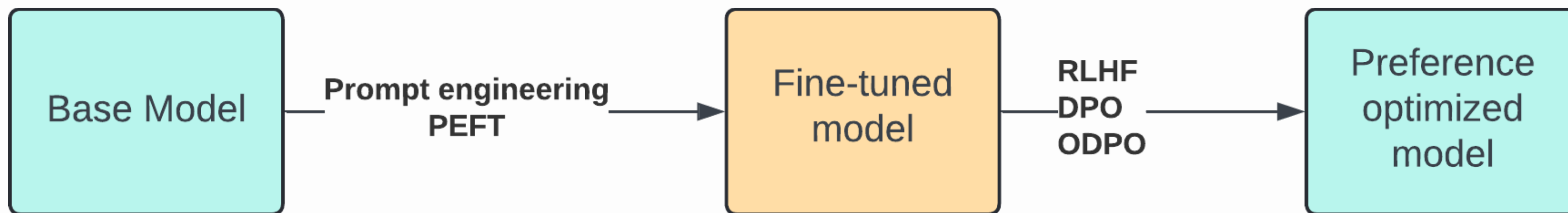
- **Extension:** The tokenizer of the original model is augmented with new tokens where the corresponding embeddings are initialized from scratch. Ex: MaLA-500 [11]
- **Replacement:** The least important non overlapping tokens from the original tokenizer are replaced with the new ones. Ex: Hungarian LLM [9]

[10] OpenAI. (n.d.). Tokenizer. OpenAI. Retrieved September 25, 2024, from <https://platform.openai.com/tokenizer>

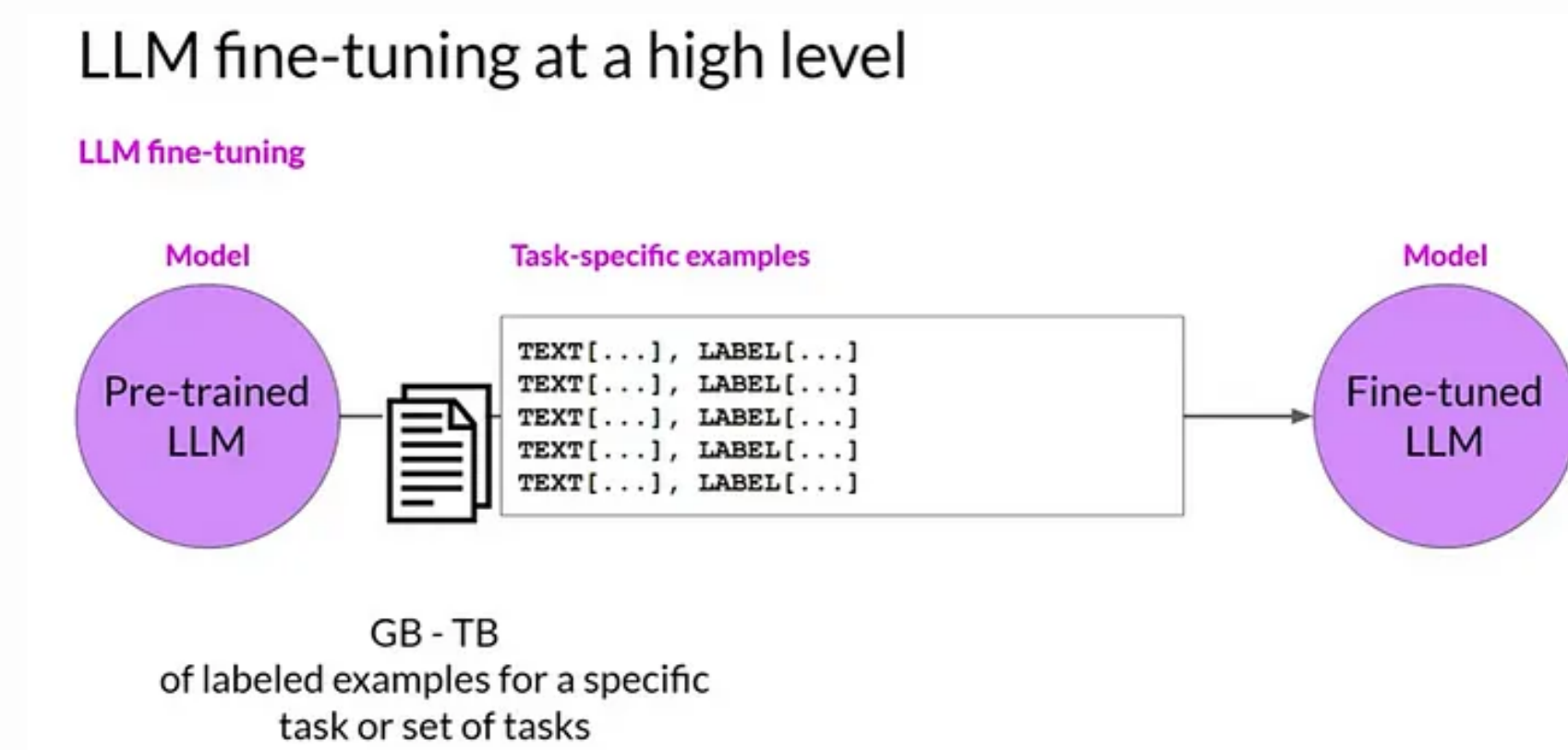
[11] P. Lin, S. Ji, J. Tiedemann, A. F. T. Martins, and H. Schütze, 'Mala-500: Massive language adaptation of large language models', arXiv preprint arXiv:2401.13303, 2024.

Post-training

- Instruct tuning (Supervised Finetuning)
- Preference Optimization



Instruction tuning (Supervised Finetuning)



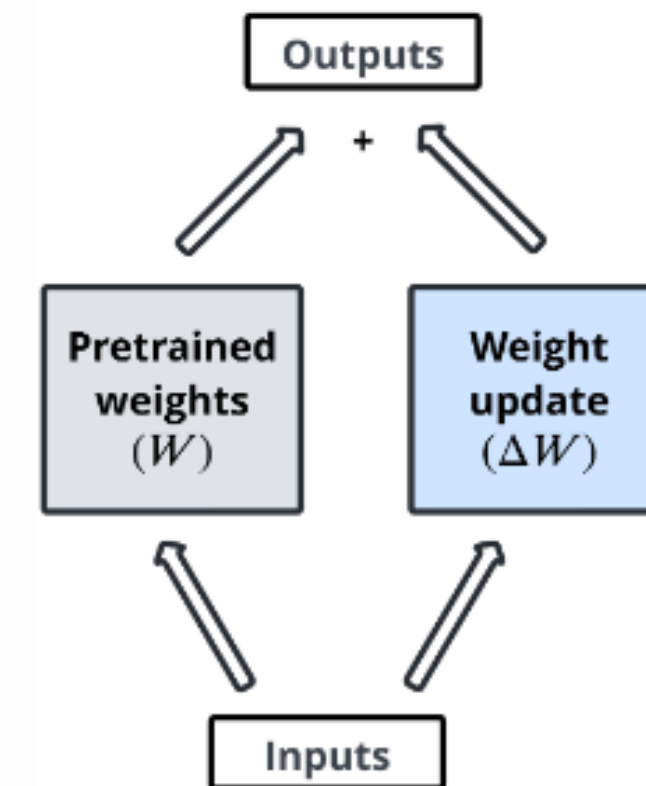
- Instruction tuning an LLM involves training it to follow explicit user instructions on a dataset of prompts paired with desired responses.
- This improves its ability to understand and execute specific tasks based on user input.

Ex: AIRAVATA[12]

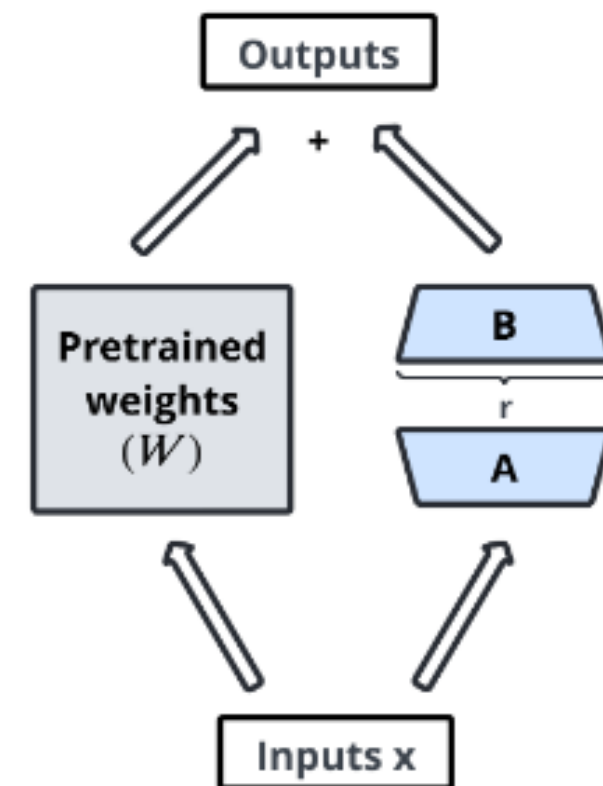
Instruction Finetuning - PEFT

- PEFT(Parameter Efficient Finetuning) methods fine -tune only a small subset of (additional) model parameters while keeping most of the pre -trained LLM parameters frozen.
 - **LoRA** transforms the model parameters into a lower rank dimension, reducing the number of trainable parameters.
 - **QLoRA**, quantize weight parameters to 4-bit precision.
 - **DoRA**, decompose their weights into magnitude and directional components

Weight update in regular finetuning



Weight update in LoRA



Preference Optimization

What is preference optimization?

- Fine-tuning the model to align its outputs more closely with human preferences or specific criteria

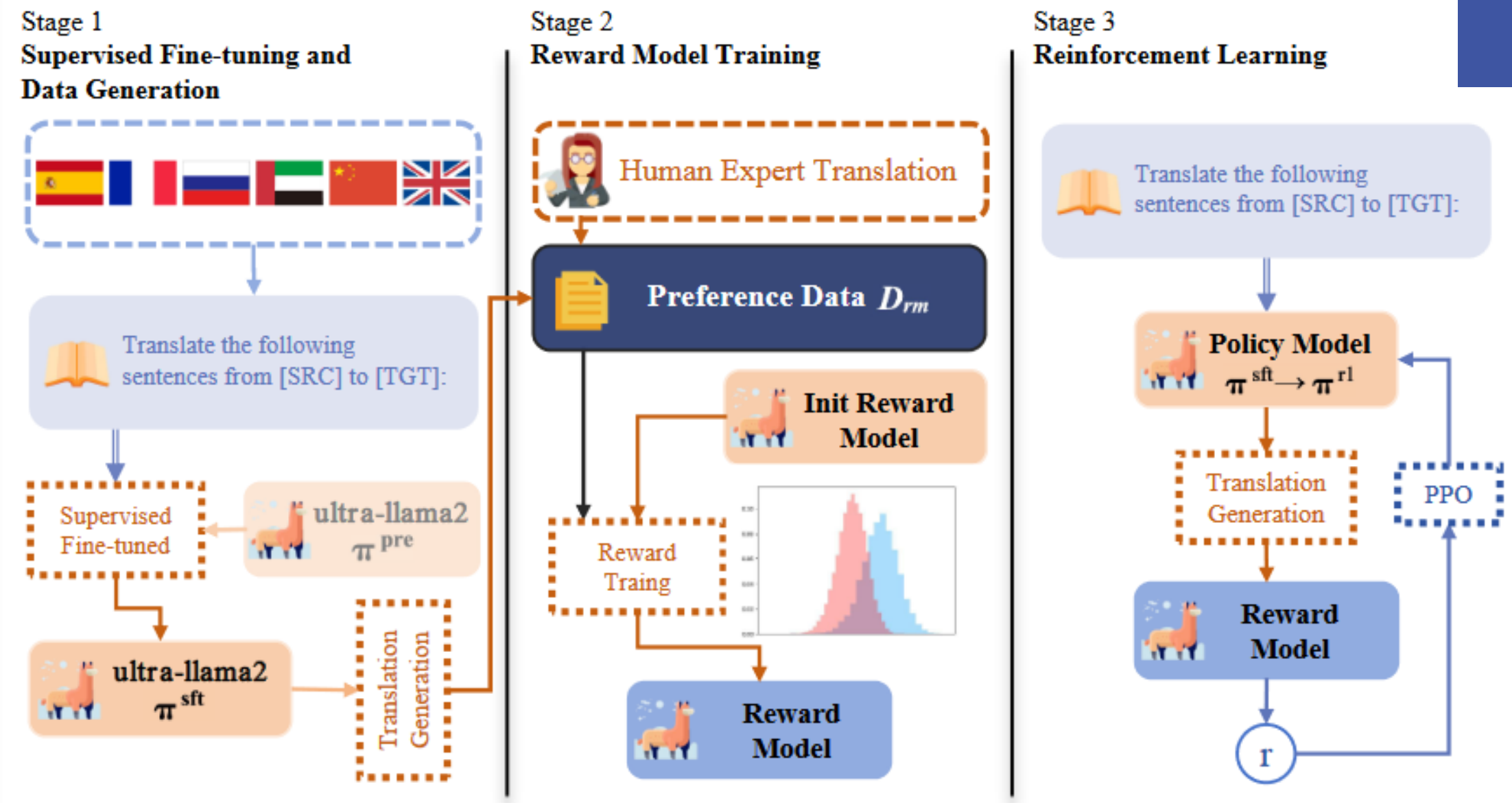
Examples

- Reinforcement Learning With Human Feedback(RLHF)
- Direct Performance Optimization(DPO)
- Direct Preference Optimization with an Offset(ODPO)

Reinforcement Learning With Human Feedback(RLHF)

Limitation

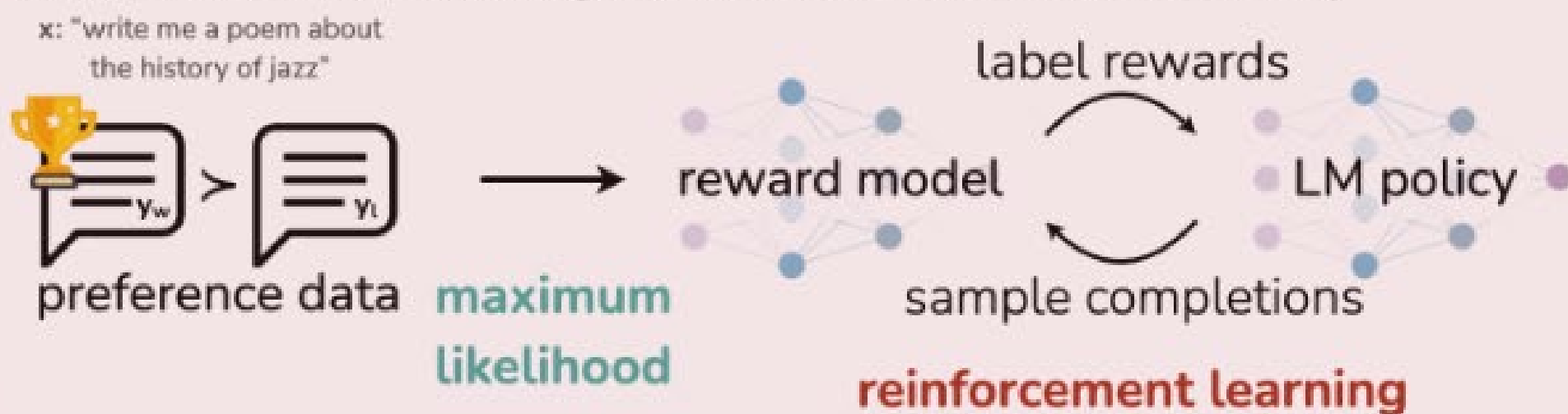
- High Computational Requirements
- Costly and Time-Consuming Feedback
- Bias in Feedback



Direct Performance Optimization(DPO)

- Bypasses the reward modeling step and directly optimizes a language model using preference data.

Reinforcement Learning from Human Feedback (RLHF)

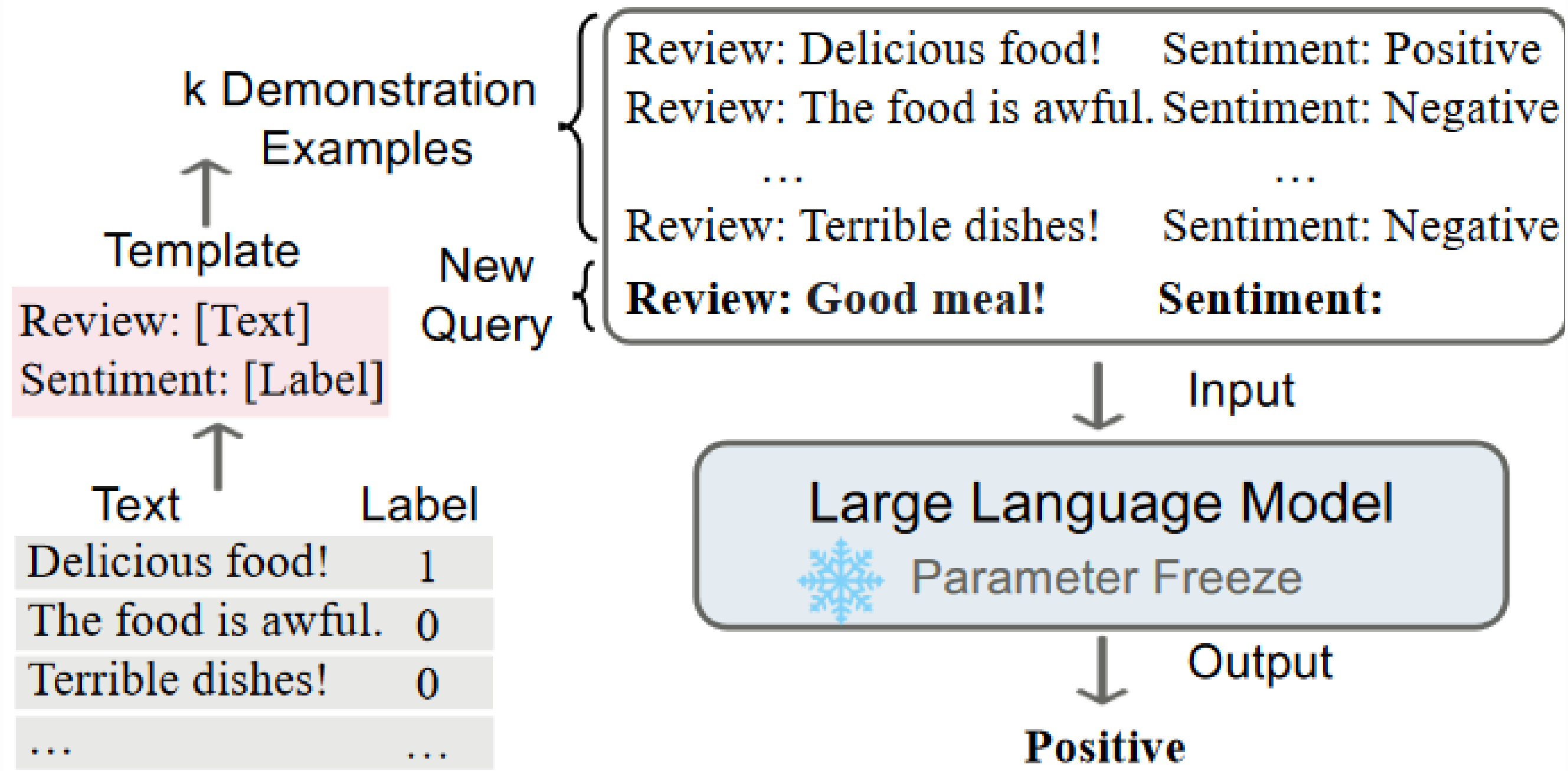


Direct Preference Optimization (DPO)

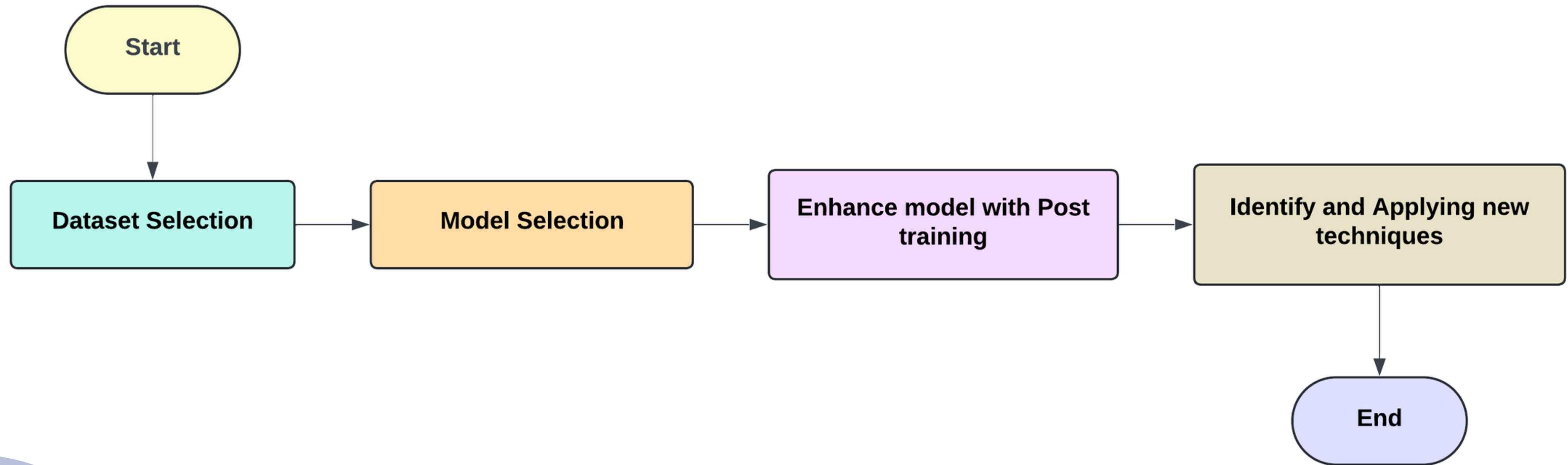


In-context Learning

- A large language model learns to perform a task by being given examples or instructions in the input prompt without further training.



Methodology



Dataset Selection

	Classification	Summarization	NER	Translation
Sinhala	NLPC-UOM/Sentiment Tagger[17] NLPC-UOM/ Writing Style Classification[18] NLPC-UOM/Sinhala News Category[18]	XLSum - Si[20]	Sinhala NER	-
Tamil	IndicXNLI - Ta[19] XCOPA - Ta	suriya 7/Tamil-Summarization	Wikia an PAN-X - Ta[23]	NLPC-UOM/English-Tamil-Parallel-Corpus[22]
Punjabi	Punjabi News Dataset IndicXNLI - Pa[19] IndicXParaphrase	XLSum - Pa[20]	Wikia an PAN-X - Pa[23] Namapadam	Samanantar[21]

[17] Ranathunga, S., & Liyanage, I. U. (2021). Sentiment Analysis of Sinhala News Comments. Transactions on Asian and Low -Resource Language Information Processing, 20(4), 1-23.

[18] V. Dhananjaya, P. Demotte, S. Ranathunga, and S. Jayasena, 'BERTifying Sinhala --A Comprehensive Analysis of Pre-trained Language Models for Sinhala Text Classification', *arXiv preprint arXiv:2208.07864*, 2022.

[19] D. Aggarwal, V. Gupta, and A. Kunchukuttan, 'IndicXNLI: Evaluating Multilingual Inference for Indian Languages'. arXiv, 2022.

[20] T. Hasan et al., 'XL-Sum: Large-Scale Multilingual Abstractive Summarization for 44 Languages', in Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, 2021, pp. 4693-4703.

[21] G. Ramesh et al., 'Samanantar: The Largest Publicly Available Parallel Corpora Collection for 11 Indic Languages', arXiv [cs.CL]. 2021.

[22] Fernando, A., Ranathunga, S., & Dias, G. (2020). Data Augmentation and Terminology Integration for Domain-Specific Sinhala-English-Tamil Statistical Machine Translation. arXiv preprint arXiv:2011.02821.

[23] X. Pan, B. Zhang, J. May, J. Nothman, K. Knight, and H. Ji, 'Cross-lingual Name Tagging and Linking for 282 Languages', in Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2017, pp. 1946-1958.

Model Selection

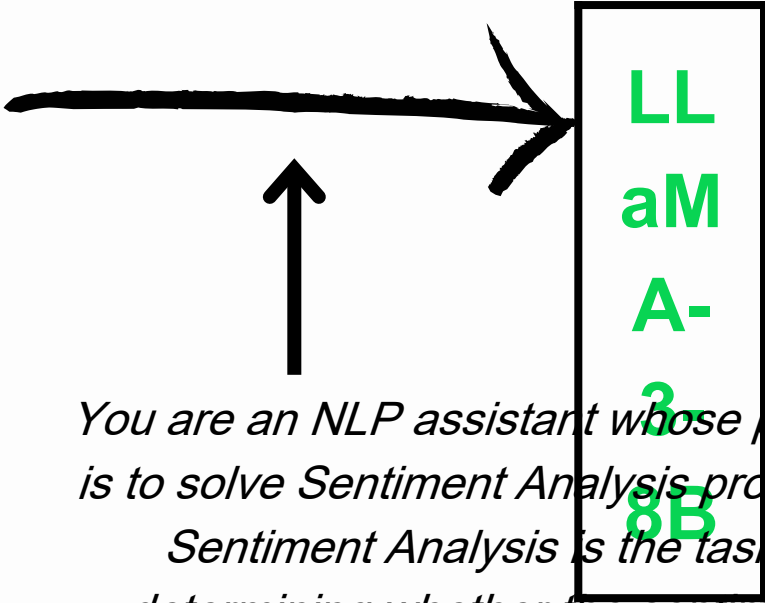
NLLB	Aya 23
XGLM	OPT
BLOOMZ	PolyLm
Phi3	LLaMA-3-8B
InternLM2	MaLA-500
Mistral	Aya-101
TigerBot	BigTrans
Baying	Vicuna
ByT5	OpenChat
Falcon	Orion

Multilingualism
Parameter Size
Representation of Indo Languages
Performance



LLaMA -3-8B
MaLA- 500
Aya- 101
Gemm a 2-9B

Zero-shot experiments



You are an NLP assistant whose purpose is to solve Sentiment Analysis problems. Sentiment Analysis is the task of determining whether the sentiment, opinion or emotion expressed in a textual data is. Does the following Sinhala sentence have a POSITIVE, NEGATIVE or NEUTRAL sentiment? Give only the POSITIVE, NEGATIVE or NEUTRAL as the answer in one word.

Benchmarking

Baseline

Dataset Name	Test set size	Full dataset size	Task	Language	Accuracy	Precision	Recall	F1
Writing-style-classification	1252	12514	Classification	si	20.926	24.07	15.935	18.281
Sentiment Tagger	905	9059	Classification	si	38.011	31.133	31.259	28.582
Sinhala-News-Category-classification	333	3327	Classification	si	18.918	17.079	16.09	14.859
NER-Sinhala	9611	96102	NER	si	3.309	1.705	0.0882	0.168

Finetune

Dataset Name	Test set size	Full dataset size	Task	Language	Accuracy	Precision	Recall	F1
Writing-style-classification	2050	12514	Classification	si	40.175	65.99	40.176	44.604
Sentiment Tagger	1812	9059	Classification	si	57.285	42.748	43.151	40.472
Sinhala-News-Category-classification	664	3327	Classification	si	39.64	28.897	27.794	27.375
NER-Sinhala	9611	96102	NER	si	74.09	81.791	74.09	77.594

Post-training techniques for model enhancement

- Cross lingual instructions
- Romanized representations
- Code Switching

Translate	English Instruction: (Alpaca-10011)	Come up with a word that rhymes with 'fine'.
	English Output: (Alpaca-10011)	One word that rhymes with 'fine' is 'mine'. (✓: 'fine' and 'mine' sound similar in English.)
	Urdu Instruction:	ایسے لفظ کے ساتھ آئیں جو 'ٹھیک' کے ساتھ قافیہ رکھتا ہو
	Urdu Output:	ایک لفظ جو 'ٹھیک' کے ساتھ ملتا ہے 'میرا' ہے۔ (✗: 'ٹھیک' and 'میرا' sound different in Urdu.)

(a) Instruction sample translated from English.

Generate	English Instruction:	Come up with a word pronounced with 'city'. Answer in Urdu.
	Urdu Corpus:	'شہر' اور 'شہد' کا تلفظ ایک جیسا ہے۔ (('City' and 'honey' are pronounced the same.)

(b) Cross-lingual instruction sample excavated.

<s> चारों अंतरिक्ष यात्री बेंगलुरु में भारतीय अंतरिक्ष अनुसंधान संगठन (इसरो) की अंतरिक्ष यात्री सुविधा में प्रशिक्षण ले रहे हैं।<0x0A> (130 tokens)

<s> chaaron antariksh yaatree bengaluru mein bhaarateey antariksh anusandhaan sangathan (isaro) kee antariksh yaatree suvidha mein prashikshan le rahe hain.<0x0A> (63 tokens)

Timeline

Activity	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec	Jan	Feb	Mar	Apr	May
Project planning and Initial Research														
Existing Multilingual LLM evaluation														
Implementation														
Finalization														
Documentation														

Feasibility

Available Resources and Strengths

- **GPU Availability:** single NVIDIA V100 GPU with 32GB memory provided by our supervisor, Dr. Rishemjit.
- **Availability of Open -Source Multilingual LLMs**
- **Availability of datasets** in Sinhala, Tamil, and Punjabi.
- **Manageable timeline**
- **Great supervisors**

Limitations

S

- **Loading Large Models**
- **API request Rate**

Thank you

