THE GUIDE TO BUILDING INDIC LLMS

BY RAMSRI GOUTHAM

Ramsri Goutham Golla

Bootstrapping 2 AI SaaS Apps to $100k ARR with no employees: **Questgen.ai** and **Supermeme.ai**

Teaches courses on LLMs, NLP and AI SaaS on **Udemy** and **learnnlp.academy**

Working on open-source initiative **Telugu-LLM-Labs** an independent initiative to build / finetune **Indic** focused language models.

Twitter (X): @ramsri_goutham

Why open-source Indic Models?

- Existing models (OpenAI, Anthropic etc) are sometimes **10-20x costly and slow** for Indic languages when compared to English. The problem lies in the efficiency of **tokenizer** of Indic languages.

- These models are not great in Telugu and other regional languages for the **local context**, our regional festivals, customs etc. The problem lies in the low percentage of training data.

Why open-source Indic Models (contd..)?

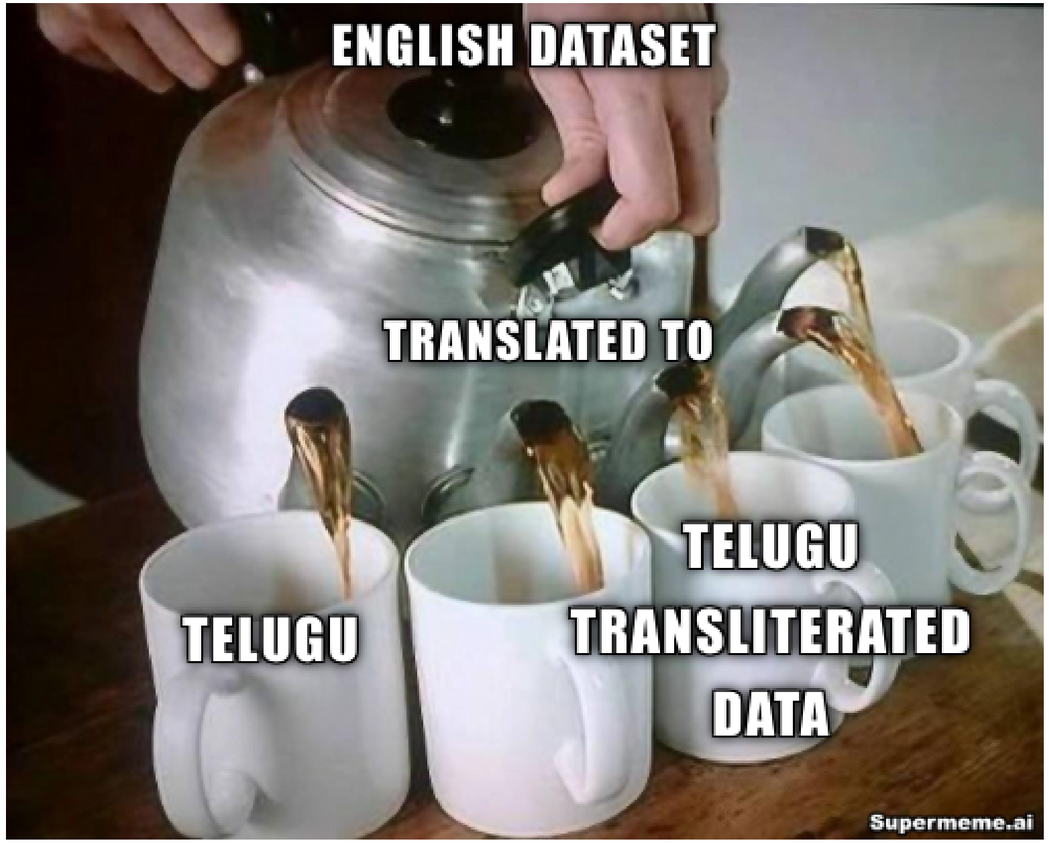
- All the goodness of open-source models, the ability to **run locally**, finetune for your specific **business use-case on custom data**, build on top of existing finetunes and do **private deployments**, contributions from multiple folks.

- The goal is to train the model to answer at **Quora-level** Q&A quality. https://te.quora.com/

LLM Training Steps

Creating Indic Training Datasets

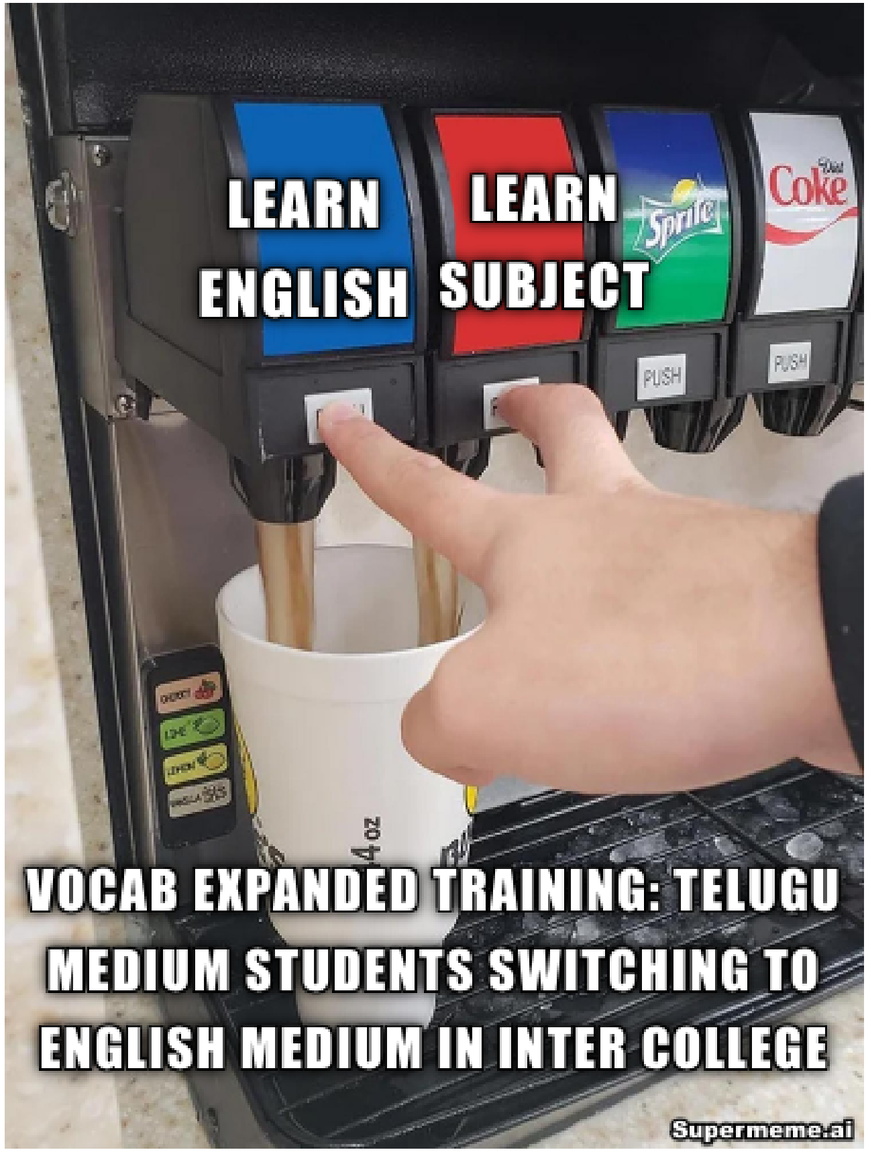
English datasets translated to Telugu with additional filtering to exclude code, math and english language specific questions.



https://huggingface.co/datasets/Telugu-LLM-Labs/telugu_alpaca_yahma_cleaned_filtered_romanized

https://huggingface.co/datasets/Telugu-LLM-Labs/telugu_teknium_GPTeacher_general_instruct_filtered_romanized

Indic LLM Training complexity with top open models 6 months back (Llama 2)



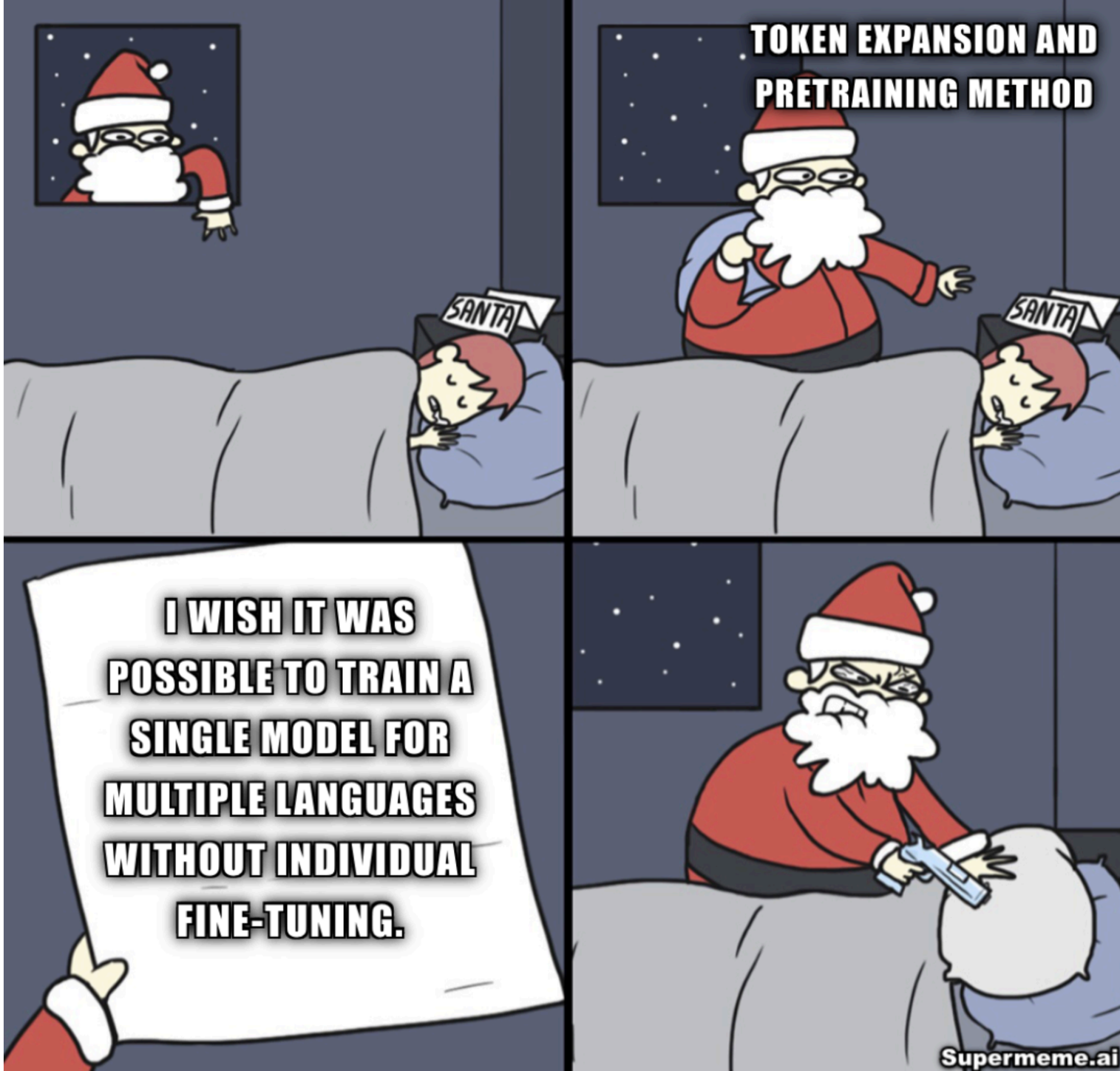
Indic LLM training with newer Open Models (Google's Gemma + Llama3 to some extent)

Note: Our focus in on 7B and 14B size models that can be finetuned on a single consumer GPU.
We are excluding large open models like DBRX (192B parameters) that has mixture-of-experts (MoE) architecture.

The cons with token expansion and pretraining method was that you had to finetune for each language separately!

Gemma solved it mostly with a 256k tokenizer compared to 32k Llama2 tokenizer!

Llama3 came later with a 128k tokenizer!



What recent techniques enable training/fine-tuning on a single consumer GPU ?

Google’s Gemma Tokenizer

Hindi: भारत एक महान देश है.

Tokenization: [‘भार’, ‘त’, ‘▁एक’, ‘▁मह’, ‘ान’, ‘▁देश’, ‘▁है’, ‘.’]

Telugu: భారతదేశం గొప్ప దేశం.

Tokenization: [‘భ’, ‘ర’, ‘త’, ‘ద’, ‘’, ‘శ’, ‘’, ‘▁గ’, ‘’, ‘ప్ప’, ‘▁ద’, ‘’, ‘శ’, ‘’, ‘.’]

Tamil: இந்தியா ஒரு சிறந்த நாடு.

Tokenization: [‘இ’, ‘ந்த’, ‘ிய’, ‘ா’, ‘▁ஒரு’, ‘▁ச’, ‘ிற’, ‘ந்த’, ‘▁நா’, ‘டு’, ‘.’]

Malayalam: ഇന്ത്യ ഒരു മഹത്തായ രാജ്യമാണ്.

Tokenization: [‘ഇ’, ‘ന്’, ‘ത്’, ‘യ’, ‘▁ഒരു’, ‘▁മ’, ‘ഹ’, ‘ത്ത’, ‘യ’, ‘▁’, ‘രാ’, ‘ജ’, ‘യ’, ‘മാ’, ‘ണ്’, ‘.’]

Datasets used for Navarasa 2.0 - A Gemma Finetuned Model

1. Hindi: sarvam.ai
2. Telugu: Telugu LLM Labs
3. Marathi: Telugu LLM Labs
4. Urdu: Telugu LLM Labs
5. Nepali: Telugu LLM Labs
6. Sindhi: Telugu LLM Labs
7. Konkani: Telugu LLM Labs
8. Assamese: Telugu LLM Labs
9. Tamil: Abhinanda
10. Kannada: Adarsh and team at Tensonic
11. Malayalam: Vishnu J Prasad
12. Odia — OdiaGenAI
13. Gujarati: Adarsh and team at Tensonic
14. Punjabi — HydraIndicLM
15. Bengali — HydraIndicLM

Navarasa 2.0 - An instruction finetuned model on 15 Indian languages

Input: Share the strategy to grow audience on Twitter in 3 short points?

Response:

1. **కంటెంట్ సృష్టి మరియు పంపిణీని పెంపొందించండి:** మీరు మీ ప్రేక్షకులకు విలువను అందించే అధిక-నాణ్యత, సంబంధిత మరియు ఆకర్షణీయమైన కంటెంట్ను సృష్టించడం ద్వారా ప్రారంభించండి. మీరు మీ కంటెంట్ను పంచుకునే విధానం గురించి కూడా చాలా అవగాహన పొందాలి. మీరు మీ కంటెంట్ను పంచుకోవడానికి అనుకూలమైన పద్ధతులలో ట్విట్టర్ లైవ్, వీడియోలు, ఇన్ఫోగ్రాఫిక్స్ మరియు మరెన్నో ఉన్నాయి.

2. **ఇతరులతో నిమగ్నం అవ్వండి:** ఇతరులతో నిమగ్నం కావడం అంటే వ్యాఖ్యలు మరియు సందేశాలకు ప్రతిస్పందించడం, ఇతర ట్విట్టర్ వినియోగదారులతో సహకరించడం మరియు మీ పరిధిని పెంచడానికి ఇతర ప్రసిద్ధ ట్విట్టర్ ఖాతాలను పంచుకోవడం. ఇది సంభావ్య అనుచరులను ఆకర్షించడానికి మరియు మీ ప్రస్తుత అనుచరులను నిలుపుకోవటానికి సహాయపడుతుంది.....

Navarasa 2.0 - An instruction finetuned model on 15 Indian languages

Input: "నా దగ్గర 100 రూపాయలు ఉన్నాయి. నేను ఒక చాక్లెట్ కోసం 20 ఖర్చు చేసాను, ఆపై నా స్నేహితుడికి 10 రూపాయలు ఇచ్చాను. నా దగ్గర ఎన్ని రూపాయలు ఉన్నాయి?"

Response:
మీకు 100 రూపాయలు ఉన్నాయి. మీరు చాక్లెట్ కోసం 20 ఖర్చు చేశారు, మిగిలిన 80 రూపాయలు మిగిలి ఉన్నాయి. మీరు మీ స్నేహితుడికి 10 రూపాయలు ఇచ్చారు, ఇప్పుడు మీ వద్ద 70 రూపాయలు మిగిలి ఉన్నాయి.<eos>

https://huggingface.co/Telugu-LLM-Labs

What’s next on experiments?

1. Train with DORA (not plain LORA)

Lora was decomposing additional trainable parameters into two low rank matrices (A&B).

DORA goes one step further and adds additional parameters: magnitude and direction vector. Now the direction vector is decomposed into two low rank trainable matrices (A&B).

So you have a LORA type configuration + additional magnitude vector trainable in DORA. But you get a superior training setup moving closer to full-finetuning with just magnitude vector training parameters as the overhead.

ReFT (Representation Finetuning) is another technique where you significantly reduce trainable parameters to even lower ~50x less than LORA

What’s next on experiments? (contd..)

2. Train with ORPO

The current Indic models released have one big problem.

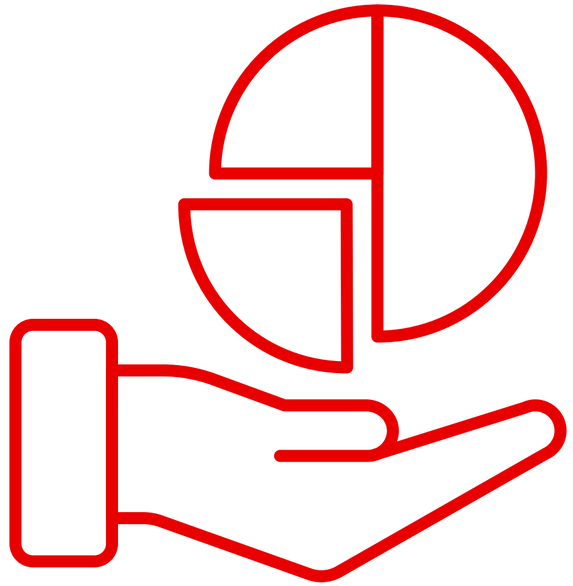
Most of them are just SFT (instruction finetuned) models. So if you ask a question like “how to make a bomb” or “how to kidnap”, it doesn’t hesitate to provide an answer.

Doing DPO or RLHF is cumbersome given data constraints. ORPO eliminates the need by combining the SFT + DPO/RLHF step into one.

Find or create a good ORPO dataset, translate it into any specific Indic language (Hindi, Tamil, Telugu etc.) and finetune such that you combine the instruction and preference alignment task into one.

How to contribute?

High-Quality regional Q&A datasets!



Questions?

Twitter: @ramsri_goutham