"Airavata : Introducing Hindi Instruction Tuned LLM

" Introduction → address limited support for Indian languages in existing LLMs

Airavata → instruction tuned model for Hindi, built by finetuning OpenHathi, which itself is an extension of the ~~Lla~~ Llama2 model.

* human curated datasets *

Dataset Creation → translating high quality English-supervised instruction tuning datasets into Hindi using IndicTrans2.

Supervise Fine Tuning (SFT) → Parameter Efficient Fine Tuning (PEFT) & LoRA.

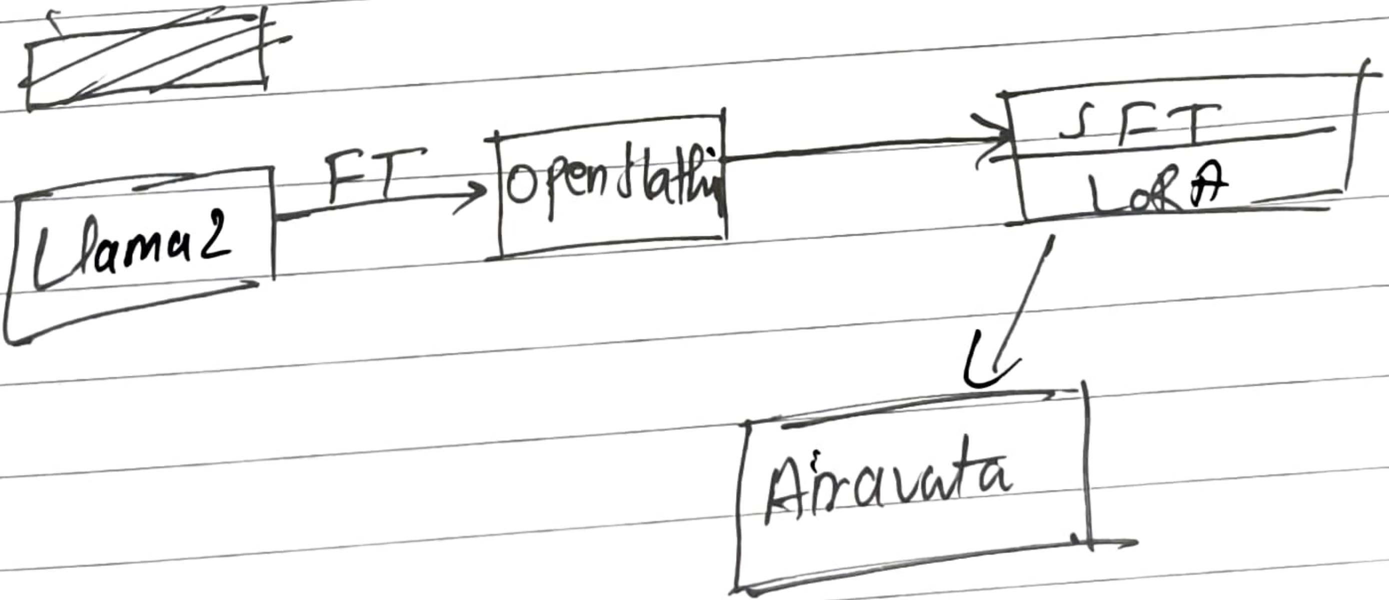
LoRA fine tuning > Full fine Tuning

FFT performed better for Hindi tasks but poorly on the English tasks but LoRA demonstrated similar performance in both Hindi & English.

First there was evaluation on NLP benchmarks, then on human evaluation.

Toxicity & Misinformation → evaluated on publicly available benchmark datasets, Multilingual Hate Check (MHC), translated versions of Implicit Hate, Toxigen & TruthfulQA.
⇒ showed better accuracy in identifying implicitly veiled hate speech.

Limitations → potential hallucinations, accuracy issues in complex or specialized tasks/topics, biased content sometimes, understanding cultural nuances & mixed language contexts.
Performance highly dependent on the quality of the dataset.


Llama2
FT
OpenHathi
SFT
LoRA
Airavata