Operator perspectives on Datasets for AI/ML in telecom networks

Aditya Jain

Data Intelligence works at incredible scale..

Connectivity | Cloud | IOT | Data Center | CPaaS | Security

Coverage

~37.9%

Revenue market share[1]

Scale

400M+

Subscribers India[2]

2.8⁺ T

Records Processed Daily

30PB+

Capacity

50M⁺

Transactions per second

900+

Data Streams Integrated

30TB+

Incremental Data Daily

1200+

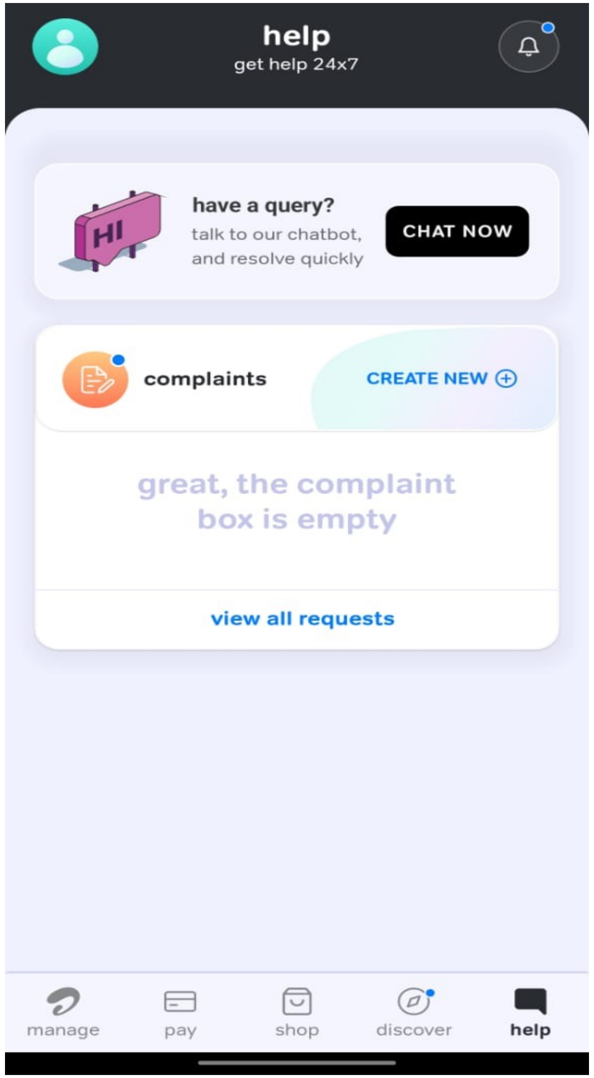
Active business users

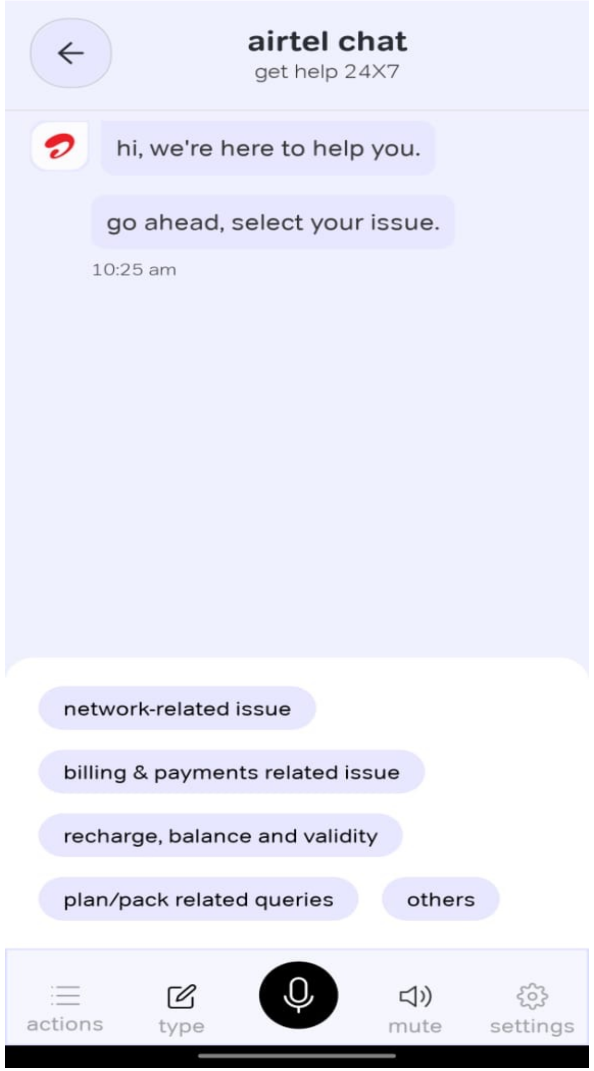
1 ET News Release Link

2 PRESS RELEASE Bharti Airtel Q4 FY24 consolidated results

Looking at customer engagement …

Working at tremendous scale





Active App
15M +
Daily Users

200K + Chat and
Voice Queries

Postpaid

Prepaid

Broadband

DTH

Huge diversity in interactions

Voice

22 National Language

100+ Language

10k+ Dialects

Chat

Transliteration

Understanding Domain

Multi-lingual

O t p not coming

ಕರೆ ಡ್ರಾಪ್ ಎದುರಿಸುತ್ತಿದೆ

मैं कॉल ड्रॉप face कर रहा हूं

आइ वांट टू कैंसल plan

ऐसएमऐस नहीं जा रहे

Signals breaking

Historically -> NLP : To detect intent of conversations

Real Data Challenges and Approach

Challenges

- Multi-Lingual Support
- Understand Transliteration
- Identifying domain specific word
- Handling Garbage Queries
- Low Latency

Solution Approach

Active Learning

+

Deep Learning
Transformer Based Multi-Lingual Model

+

Serviceable FastAPI

Now -> LLMs works better in high resource languages..

Real Data Challenges and Approach

Challenges	Solution Approach
Multi-Lingual Support	
Understand Transliteration	
Identifying domain specific word	Fine-tuned LLMs
Handling Garbage Queries	
Low Latency	

Challenges of LLMs with low resource languages

- ✓ Performance of LLMs for Low Resource Languages (LRLs) is hindered by **unavailability of high-quality open-source large-scale data** required for pre-training and fine tuning[1]

- ✓ **Multi-lingual models take advantage of cross-lingual transfer up until a point**, after which the overall performance on monolingual and cross-lingual benchmarks degrades.[1]

- ✓ LLMs fine-tuned specifically for low resource languages outperform base GPT models[2] but require heavy computation cost and time

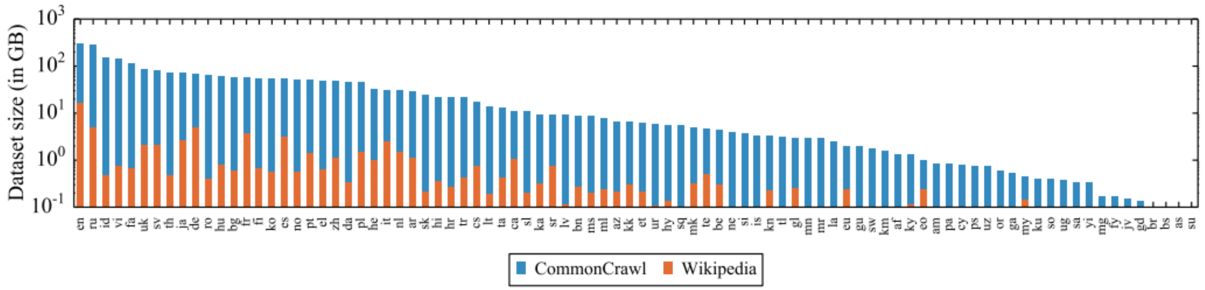


Figure 1: Amount of data in GiB (log-scale) for the 88 languages that appear in both the Wiki-100 corpus used for mBERT and XLM-100, and the CC-100 used for XLM-R. CC-100 increases the amount of data by several orders of magnitude, in particular for low-resource languages.

Model	Eng		Zho		Vie		Ind		Tha	
	GSM8K	MATH	GSM8K	MATH	GSM8K	MATH	GSM8K	MATH	GSM8K	MATH
ChatGPT-3.5	**80.8**	34.1	48.2	21.5	55.0	26.5	64.3	26.4	35.8	18.1
Qwen1.5-7B-chat	56.8	15.3	40.0	2.7	37.7	9.0	36.9	7.7	21.9	4.7
SeaLLM-7B-v2	78.2	27.5	**53.7**	17.6	69.9	23.8	**71.5**	24.4	59.6	22.4
SeaLLM-7B-v2.5	78.5	**34.9**	51.3	**22.1**	**72.3**	**30.2**	**71.5**	**30.1**	**62.0**	**28.4**

Table 4: GSM8K and MATH scores (Cobbe et al., 2021; Hendrycks et al., 2021b) and their translated-versions in Chinese, Vietnamese, Indonesian and Thai, under zero-shot chain-of-thought prompting for different models.

1. https://arxiv.org/pdf/1911.02116 (Unsupervised Cross-lingual Representation Learning at Scale)
2. https://arxiv.org/pdf/2312.00738 (SeaLLMs - Large Language Models for Southeast Asia)

Takeaways

- Creation of benchmark datasets :
  - Multi modal in nature
  - Assess Network performance and perform Root Cause Analysis
  - Across low/high resource languages

- Creation of benchmark performance metrics
  - To measure Accuracy holistically across RCA, quality of response, etc.

Benchmark Dataset Instances in other Industry around Reasoning

- *MATH (Mathematics for Machine Learning)*
- *GSM8K (Grade School Math 8K)*
- -
- -

Performance Metrics

- *Accuracy, F1-Score*
- *BLEU Score*
- *ROGUE Score*
- -
- -

Thank You !!

Aditya Jain

aditya1.jain@airtel.com