MULTILINGUAL FINANCIAL LLM

(LLM Fine-Tuning for Hindi & Telugu Finance QA + Sentiment Analysis)

By Group 31

PROBLEM STATEMENT

We aim to build a multilingual financial LLM that can answer financial questions and classify sentiment in regional Indian languages by:

- Translating benchmark English finance datasets into Hindi and Telugu
- Fine-tuning existing financial LLMs on these translated datasets
- Evaluating whether financial knowledge can be transferred cross-lingually

MOTIVATION

Why this problem?

- Most financial AI tools work only in English → huge language barrier for Indian users.
- Very limited financial datasets exist in Hindi/Telugu.
- Creating multilingual domain-specific financial assistants can democratize financial literacy.
- We wanted to test whether:
  - Translation + fine-tuning is enough
  - Models can retain financial reasoning while learning a new language
  - Performance varies across languages

Intention

- Build a structured pipeline for future multilingual financial LLMs.
- Compare Hindi vs Telugu performance.
- Understand limitations of small LLMs on low-resource languages.

LITERATURE SURVEY

1. QLoRA – Efficient Finetuning of Quantized LLMs

- Enabled our 4-bit + LoRA finetuning on T4 GPUs
- Supports our design choice of parameter-efficient training
- Ensures we retain base-model knowledge while adding Hindi/Telugu financial domain capability

📌The paper argues that QLoRA enables full-model finetuning on consumer GPUs using low-bit quantization and parameter-efficient adapters.

2. FinGPT – Domain-Specific Financial LLMs

- Motivated building a financially aligned LLM
- Validates the use of FinQA + sentiment datasets
- Shows why domain tuning improves QA & sentiment tasks

📌 The paper highlights the need for specialized financial LLMs and demonstrates performance gains through carefully curated financial datasets.

3. HF Models for Multilingual Translation

- Supports our choice of smaller translation models due to GPU limits
- Discusses handling low-resource languages, matching our Telugu challenges
- Validates pipeline approach + evaluation using similarity metrics

📌 The paper highlights the effectiveness of HF-based mBART/NLLB pipelines for multilingual communication and the challenges of low-resource languages.

DATASETS USED (BENCHMARK DATASETS)

Benchmark Datasets

We used two well-known English financial datasets:

FIQA Dataset

- Financial reasoning dataset with long answers.
- We filtered outputs with token length <1024 to make training feasible.
- https://huggingface.co/datasets/FinGPT/fingpt-fiqa_qa

	input	output	instruction	output_token_count
0	What is considered a business expense on a bus...	The IRS Guidance pertaining to the subject. I...	Utilize your financial knowledge, give your an...	536
1	Claiming business expenses for a business with...	Yes you can claim your business deductions if ...	Offer your insights or judgment on the input f...	335
2	Transferring money from One business checking ...	You should have separate files for each of the...	Based on your financial expertise, provide you...	258
3	Having a separate bank account for business/in...	Having a separate checking account for the bus...	Share your insights or perspective on the fina...	240
4	Having a separate bank account for business/in...	You don't specify which country you are in, so...	Offer your thoughts or opinion on the input fi...	451

DATASETS USED (BENCHMARK DATASETS)

Financial Sentiment Dataset

- Statement-level sentiment classification.
- Labelled as positive / negative
- https://huggingface.co/datasets/FinGPT/fingpt-sentiment-cls

	input	output	instruction
0	Starbucks says the workers violated safety pol...	negative	Determine the sentiment expressed in the news ...
1	$brcm raises revenue forecast	positive	Determine the sentiment expressed in the tweet...
2	Google parent Alphabet Inc. reported revenue a...	negative	Characterize the news's sentiment using the fo...
3	Here we highlight some top-ranked technology E...	positive	Characterize the news's sentiment using the fo...
4	$UVXY Put the chum out there at key support th...	negative	What is the sentiment of the input tweet from ...
5	$SPY Less than 0.2% down and people are callin...	negative	Determine the sentiment expressed in the tweet...
6	Stock Market Update: Eli Lilly gains on upbeat...	positive	Categorize the input tweet's emotional tone in...
7	AAPL's shareholders embrace the tech giant's r...	positive	Characterize the news's sentiment using the fo...
8	Finnish Suominen Flexible Packaging is cutting...	negative	Determine the sentiment expressed in the news ...
9	$RNN More bleeding Monday.	negative	What is the sentiment of the input tweet from ...

DATA HANDLING (TRANSLATION DETAILS)

Translation Pipeline
We translated both datasets into Hindi and Telugu.

Hindi Translation Model

- **Model:** Rotary IndicTrans2
- **HF link:** https://huggingface.co/prajdabre/rotary-indictrans2-indic-en-1B
- **Reason for use:**
  a. Smaller 1B model → fits T4 GPU
  b. Good quality for Indic languages
  c. Fast inference = practical for large dataset translation
- **Notebook link:** https://www.kaggle.com/code/harshmehta1618/cs772-tranlation

DATA HANDLING (TRANSLATION DETAILS)

Telugu Translation Model

- **Model:** aryaumesh / english-to-telugu
- **HF link:** https://huggingface.co/aryaumesh/english-to-telugu
- **Reason for use:**
  a. Lightweight, T4-friendly
  b. Telugu-specific training → better than generic multilingual NMT
- **Notebook link:** https://www.kaggle.com/code/hiteshkhiani1/finsentiment

BACK-TRANSLATION + COSINE SIMILARITY EVALUATION

Purpose:

To evaluate translation quality.

Method:

1. Translate English → Hindi/Telugu
2. Back-translate Hindi/Telugu → English
3. Compute sentence-level cosine similarity of embeddings using Sentence-BERT
4. Higher similarity → better translation fidelity

Result:

- Both hindi and telugu translations yielded satisfactory cosine similarity
- Confirms decent translation model performance

Notebook link:

- https://www.kaggle.com/code/puneetyadavvvv/hindicosinesim
- https://www.kaggle.com/code/puneetyadavvvv/notebookbc2e3a3386

COSINE SIMILARITY RESULTS (HINDI)

```
Instruction similarity stats:
count    2000.000000
mean        0.884392
std         0.096992
min         0.706102
25%         0.853444
50%         0.922874
75%         0.957251
max         0.974493
Name: sim_instruction, dtype: float64
```

```
Input similarity stats:
count    2000.000000
mean        0.945073
std         0.070747
min         0.345920
25%         0.923802
50%         0.970547
75%         0.994579
max         1.000000
Name: sim_input, dtype: float64
```

```
Output similarity stats:
count    2000.000000
mean        0.939458
std         0.051087
min         0.094861
25%         0.922519
50%         0.950360
75%         0.971436
max         1.000000
Name: sim_output, dtype: float64
```

COSINE SIMILARITY RESULTS (TELUGU)

```
Input similarity stats:
count    2000.000000
mean        0.916966
std         0.111170
min         0.219335
25%         0.888840
50%         0.961516
75%         0.991640
max         1.000000
Name: sim_input, dtype: float64
Loading widget...
Loading widget...

Instruction similarity stats:
count    2000.000000
mean        0.939845
std         0.043098
min         0.883500
25%         0.898331
50%         0.949974
75%         0.966998
max         1.000000
Name: sim_instruction, dtype: float64
```

MODEL FINE-TUNING

Hindi Finetuning

We fine-tuned: TinyLlama/TinyLlama-1.1B-Chat-v1.0
HF link: https://huggingface.co/TinyLlama/TinyLlama-1.1B-Chat-v1.0

Why TinyLlama?

- Fits on free Kaggle T4 GPU
- 1.1B params → fast training
- Strong base for instruction-following tasks

Dataset Used

- 47k Hindi QA samples
- 17k Hindi sentiment samples
- Combined + formatted in instruction style

Notebook link:

- https://www.kaggle.com/code/harshmehta1618/cs772-qlora

MODEL FINE-TUNING (HINDI)

Hindi Finetuning (Evaluation Results)

📊 **Sentiment Classification (300 samples)**

- Accuracy: 60.3%
- Strong on सकारात्मक, weaker on नकारात्मक
- Class imbalance + short label generation affected performance

📊 **QA Evaluation (150 samples)**

- Output is grammatical Hindi
- Limitations due to:
  - TinyLlama's 1.1B size
  - Translation noise in training data

📌 **Summary**

- Good Hindi fluency
- Decent performance on short, direct sentiment tasks
- Limited depth on finance QA

MODEL FINE-TUNING (HINDI)

Hindi Finetuning (Evaluation Results)

Sentiment results:

Accuracy: 0.6033333333333334

	precision	recall	f1-score	support
नकारात्मक	0.418	0.865	0.564	89
सकारात्मक	0.897	0.493	0.636	211
accuracy			0.603	300
macro avg	0.658	0.679	0.600	300
weighted avg	0.755	0.603	0.615	300

Sample sentiment predictions:

Input tweet: हारग्रीव्स लैंसडाउन के कमजोर बाजारों में परिसंपत्तियों में 2.6 प्रतिशत की वृद्धि

Gold label : सकारात्मक

Pred label : नकारात्मक

Input tweet: "बासवेयर के उत्पाद की बिक्री वित्तीय अवधि में 24 प्रतिशत तक मजबूती से बढ़ी।

Gold label : सकारात्मक

Pred label : सकारात्मक

Input tweet: $CAMT ऐसा लग रहा है कि टूट सकता है। मैं बाहर हूँ

Gold label : नकारात्मक

Pred label : नकारात्मक

Input tweet: HELSINKI (थॉमसन फाइनेंशियल)-केमिरा ग्रोहाउ ने बेहतर बिक्री पर अपनी पहली तिमाही की कमाई में लाभ उठाया, विशेष रूप से यूरोप में अपने उर्वरक व्यवसाय में, जो आम तौर पर पहली तिमाही के दौरान मजबूत होता है।

Gold label : सकारात्मक

Pred label : सकारात्मक

Input tweet: डिपार्टमेंट स्टोर डिवीजन की बिक्री में 15 प्रतिशत और कपड़ों की दुकान की सहायक कंपनी सेप्पाला की बिक्री में 8 प्रतिशत की वृद्धि हुई इस बीच हॉबी हॉल की बिक्री में 12 प्रतिशत की कमी आई।

Gold label : सकारात्मक

Pred label : सकारात्मक

MODEL FINE-TUNING

Telugu Finetuning

Model used: WiroAI/WiroAI-Finance-Qwen-1.5B
(HF: https://huggingface.co/WiroAI/WiroAI-Finance-Qwen-1.5B)

Why this model?

- Already fine-tuned on financial knowledge
- We attempted to add Telugu understanding on top

Observation:

- Model struggled with Telugu
- Reason:
  - Telugu not well-represented in Qwen training
  - Model had to learn both
    - Telugu grammar
    - Financial domain patterns
  - Learning two things at once on 1.5B model → too difficult
- Result: Unsatisfactory Telugu outputs

Notebook link: https://www.kaggle.com/code/prathampandit123/final-telugu

FOR BONUS

📌 Large-Scale Dataset Translation:

- Translated 47k sentiment samples + 17k FinQA samples into Hindi & Telugu
- Achieved mean cosine similarity > 0.92 for both languages (high-quality translation)

📌 Handling Long Financial QA

- FinQA contained many answers up to 1024 tokens
- Required selecting smaller, efficient translation models that:
  - Fit Kaggle T4 memory limits
  - Still maintained strong translation quality
- Balanced accuracy ↔ latency to enable full dataset translation

📌 Significant Finetuning Effort

- Successfully fine-tuned Hindi model → strong sentiment classification performance
- Completed full QA + Sentiment combined fine-tuning pipeline
- For Telugu:
  - The base model lacked Telugu capability
  - Training became a learning-oriented experience on multilingual alignment limits of small LLMs

📌 Overall Bonus Contribution

- Built a complete multilingual financial LLM pipeline end-to-end
- Managed translation, back-translation evaluation, dataset filtering, and QLoRA finetuning
- Ensured all steps work on limited hardware without compromising quality

THANK YOU