SHANTO-MARIAM UNIVERSITY OF CREATIVE TECHNOLOGY

Thesis Demo 1

Presented By

Name: Subodh Chandra Shil
ID: 211071003
Batch: 27^{th}

Name: Sadequr Rahman Shuvo
ID: 212071020
Batch: 27th

Supervised by

Ahamad Nokib Mozumder
Department of CSE & CSIT
Lecturer

TABLE OF CONTENTS

01

Brief and Introduction

We will be talking about our works behind the thesis

Enhanced Privacy Preserving Multilingual Fake News Detection with LoRA-based Parameter-Efficient Fine-Tuning on Encoder-Only LLMs

The growing threat of fake news across multiple languages poses serious challenges to online credibility, political stability, and public awareness. Traditional LLM-based solutions are often data-hungry, resource-intensive, and raise serious privacy concerns when deployed with sensitive regional or political content.

This thesis proposes a **privacy-preserving, multilingual fake news detection framework** using **parameter-efficient fine-tuning (PEFT)** via **LoRA** on **encoder-only open-source language models**. We fine-tune models locally on low-resource hardware to avoid cloud-based privacy risks, supporting **English, Bangla, Hindi, and Spanish**.

By fine-tuning on multilingual datasets (Bangla, English, Hindi, Spanish), we achieved **over 95% accuracy**, while drastically reducing training cost, memory footprint, and model update time. The result is an efficient, scalable, and secure solution for real-world misinformation detection.

1. **Rapid Spread of Fake News:** The rise of misinformation online, especially in my country, has triggered real-world consequences including **public panic and riots**.

2. **Lack of Multilingual Solutions:** Most existing systems focus on English only. There is a **critical gap in fake news detection across low-resource languages** such as Bangla, Hindi, and others.

3. **Leveraging Modern LLMs:** Inspired by the breakthroughs in transformer-based architectures, we aim to **achieve higher accuracy and robustness** using **parameter-efficient fine-tuning (LoRA)** on **encoder-only language models**.

4. **Modern Problems Need Modern Tools:** With the rise of LLMs and transformer architectures, there is huge potential to apply parameter-efficient fine-tuning (PEFT) to achieve state-of-the-art results without needing massive compute power.

5. **Our Goal:** Build a multilingual, accurate, and privacy-preserving fake news detection system using LoRA-based fine-tuning on encoder-only LLMs, achieving >95% accuracy while being lightweight and scalable.

Base paper and other relevant papers

Author	Title	Published on	Findings of the study
Adapting Fake News Detection to the Era of Large Language Models	Jinyan Su Claire Cardie Preslav Nakov	arXiv	The base paper primarily relied on outdated LLM architectures, all of which were trained on English-centric corpora, making the system inherently biased toward monolingual contexts. One of the major shortcomings was the lack of a privacy-preserving approach, as all training and inference processes assumed full access to user data. Their LLM dataset test on closed source models like BERT, RoBERTa, ELECTRA, ALBERT, DeBERTa (Large versions).
From Scarcity to Capability: Empowering Fake News Detection in Low-Resource Languages	Hrithik Majumdar Shibu, Shrestha Datta, Md. Sumon Miah, Nasrullah Sami	IndoNLP2025	This work utilized BanglaBERT, SagorBERT, and QLoRA, with a strong focus on the Bangla language. It combined classical TF-IDF methods with modern LLM-based techniques, highlighting the growing importance of modeling for low-resource languages like Bangla.
A Survey on the Use of LLMs in Fake News	Eleftheria Papageorgiou, Christos Chronis, Iraklis Varlamis	MDPI (Future Internet)	The study provides a broad overview of LLMs in fake news detection but lacks implementation depth. It effectively highlights research gaps, particularly in multilingual and low-resource language settings.
Fake News Detection with LLMs on the LIAR Dataset	David Boissonneault, Emily Hensen	Research Square Preprints	The study utilized closed-source models like ChatGPT and Gemini, focusing on LogLoss optimization and closed source fine-tuning strategies. In contrast, our approach emphasizes open-source, locally trainable models to ensure transparency, privacy, and broader accessibility.

Author	Title	Published on	Findings of the study
Fake news detection in low-resource languages: A novel hybrid summarization approach	Jawaher Alghamdi, Yuqing Lin, Suhuai Luo	ScienceDirect	The paper introduces FND-LLM, a novel multimodal fake news detection framework that integrates small language models (SLMs), vision transformers (ViT/EAViT), CLIP, a cross-attention module, and an LLM-based reasoning branch. It extracts textual content, image semantics, and tampering signals, then fuses them using co-attention and a mixture-of-experts architecture. The LLM component provides logical reasoning and fact-checking insight. Evaluations on Weibo, GossipCop, and PolitiFact datasets demonstrate superior accuracy gains of +0.7%, +6.8%, and +1.3%, respectively, compared to existing methods. By effectively combining unimodal and cross-modal features with reasoning capabilities, FND-LLM sets a new benchmark in multimodal fake news detection.
Integrating Large Language Models and Machine Learning for Fake News Detection	Ting Wei Teo Hui Na Chua Muhammed Basheer Richard T.K. Wong	IEEE Xplore	The surge in fake news demands smarter detection. Traditional ML models rely heavily on hand-crafted features and struggle with nuanced language. This study bridges that gap by integrating ChatGPT-3.5 with conventional models. Our hybrid XGBoost model achieved 93.39% accuracy, 95.04% precision, 95% recall, and a 95.6% F1 score, showing the strength of LLMs in identifying subtle misinformation patterns.
Large Language Model Based Fake News Detection	Mussa Aman	ScienceDirect	This paper presents a LLM-based approach using the LLaMA model to detect disinformation in AI-generated videos and images. By aligning the model with task-specific instructions, it better mimics human judgment. Despite limited computational resources, the method shows strong potential and explores practical strategies for fine-tuning under such constraints.

02

Methodology

Now we will shade some light on how we prepared the robust dataset for our model training

A. Dataset Overview

- A unified, multilingual dataset was developed to support fake news detection across English, Bengali, Hindi, and Spanish.
- **Total Samples:** 1,00000
- Language Distribution: 25% samples per language – Bangla, English, Hindi, Spanish
- Designed for multilingual classification and zero-shot evaluation with LLMs and embedding-based architectures.
- Class Balance
  - Label 0: Fake News - 50,000
  - Label 1: Real News - 50,000

B. Data Sources

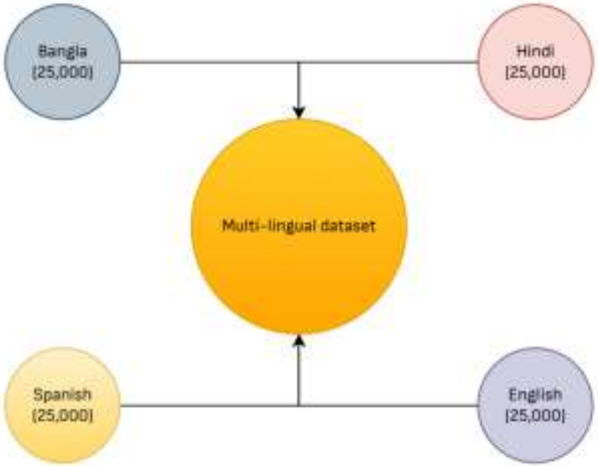
- English + Bengali (HuggingFace): **DipsankarSinha/bangla-fake-news**
- Hindi (Kaggle): **Hindi Fake News Detection Dataset (HFDND)** — 16,933 samples (Oversampled to 25000)
- Spanish (Kaggle): **Spanish Political Fake News**

C. Data Preprocessing and Cleaning

- **Used Pandas for structured preprocessing and data handling**
- **Null value handling: Removed missing or incomplete records**
- **Truncation: News descriptions exceeding 500 tokens were truncated. Due to limitation of LLM's token size.**
- **Class simplification:** Converted all labels into **binary format (Fake/Real)**
- Used **oversampling** to balance minority class samples within each language

D. Data Distribution

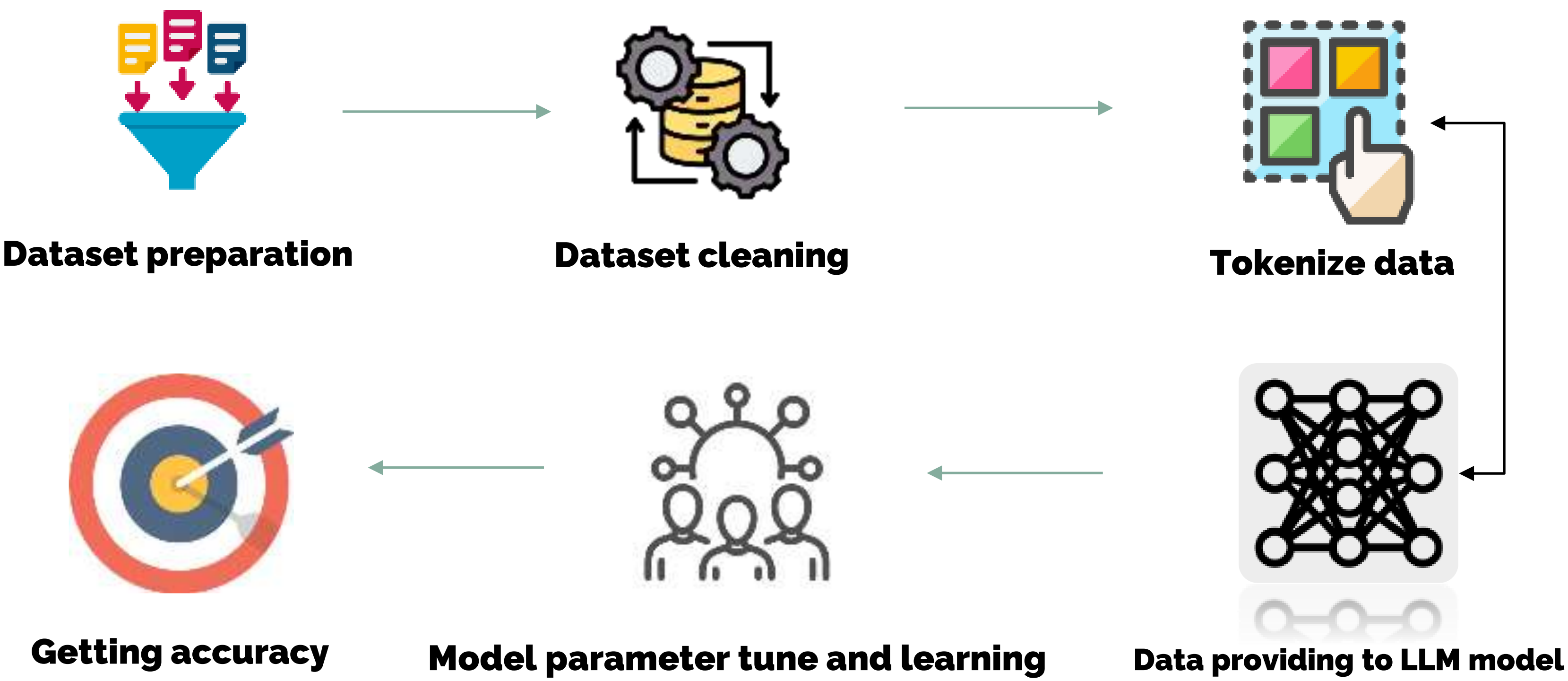
- **Training:** 75%
- **Testing:** 16%
- **Validation:** 9%



Output Handling

- **Binary Classification Setup:** 0 = False news, 1 = True news.
- **Sigmoid Activation:** Output $\geq 0.5 \rightarrow$ True, Output $< 0.5 \rightarrow$ False.
- **Neutral Zone (0.4–0.6):** Introduced to handle ambiguous predictions and avoid forced bias.
- **Bias Reduction:** Prevents the model from classifying borderline cases as strictly true or false.
- **Improved Reliability:** Neutral class improves trustworthiness of results by acknowledging uncertainty.
- **Real-World Alignment:** Mimics human decision-making, where not all news can be judged as strictly true/false.
- **Better Generalization:** Reduces overfitting on noisy samples by not forcing incorrect labels.
- **Evaluation Impact:** Accuracy is calculated only on clear-cut predictions, ensuring fairer measurement.
- **Practical Use:** Neutral predictions can be flagged for human review or fact-checking.
- **Scalability:** This framework can extend to **multi-class classification** if news categories expand in the future.

Workflow of dataset for model tuning



What is LLM

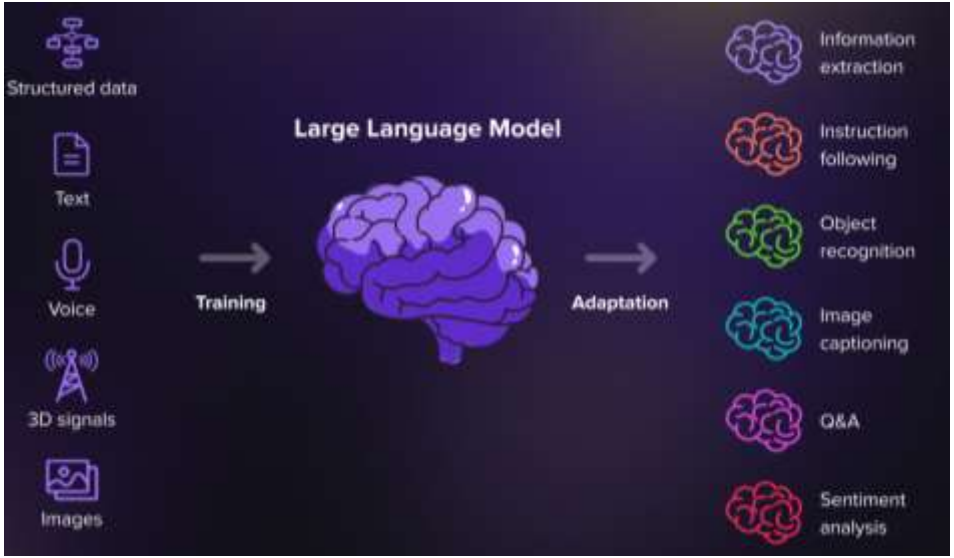
Large language model: Trained on billions of data and variety of parameters. Large language models are advanced machine learning models or AI system that trained on massive amount of data (from the web and other sources) to understand, generate and process human language.

- Multilingual training allows LLMs to detect fake news across Bangla, English, Hindi, and Spanish.

- Few-shot and zero-shot capabilities help handle limited fake news data in low-resource languages.

- Fine-tuning with PEFT (e.g., LoRA) customizes LLMs for domain-specific fake news classification.

- Open-source LLMs (e.g., Gemma, Mistral) allow privacy-preserving local deployment and training.



What is Transformers

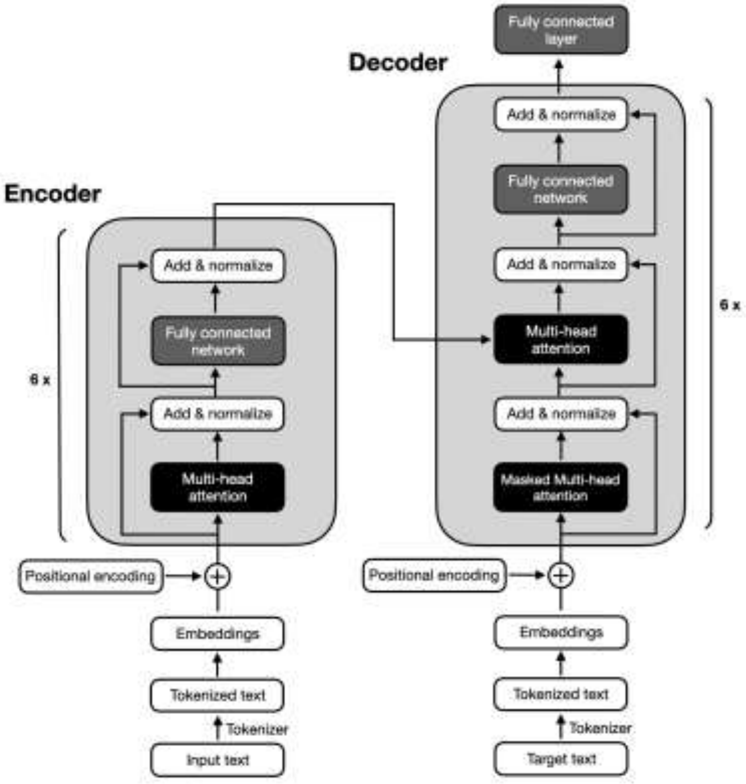
Transformers are the building block of LLMs. It is a powerful AI architecture that helps machines understand patterns in text and other data. They are the foundation of **Large Language Models (LLMs)**, which are trained on billions of words from the internet and other sources. The workflow of transformers:

Input Text Preparation
–Text is preprocessed into smaller chunks called **tokens** (Tokenization).

Token Mapping
–Each token is mapped to a unique **ID** using a **vocabulary**.

Token Embedding
–Each token ID is passed through an **embedding layer** to get numerical vectors (embeddings).

Positional Encoding
–Since transformers don't have a sense of word order, **positional encoding** is added to embeddings to capture word position.



Attention: Why Transformers are better than traditional NLP methods

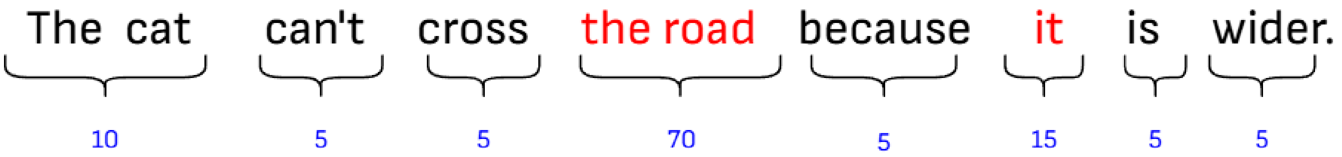
Traditional NLP models like **TF-IDF, RNN, Seq2Seq, and LSTM** were widely used before the rise of LLMs.

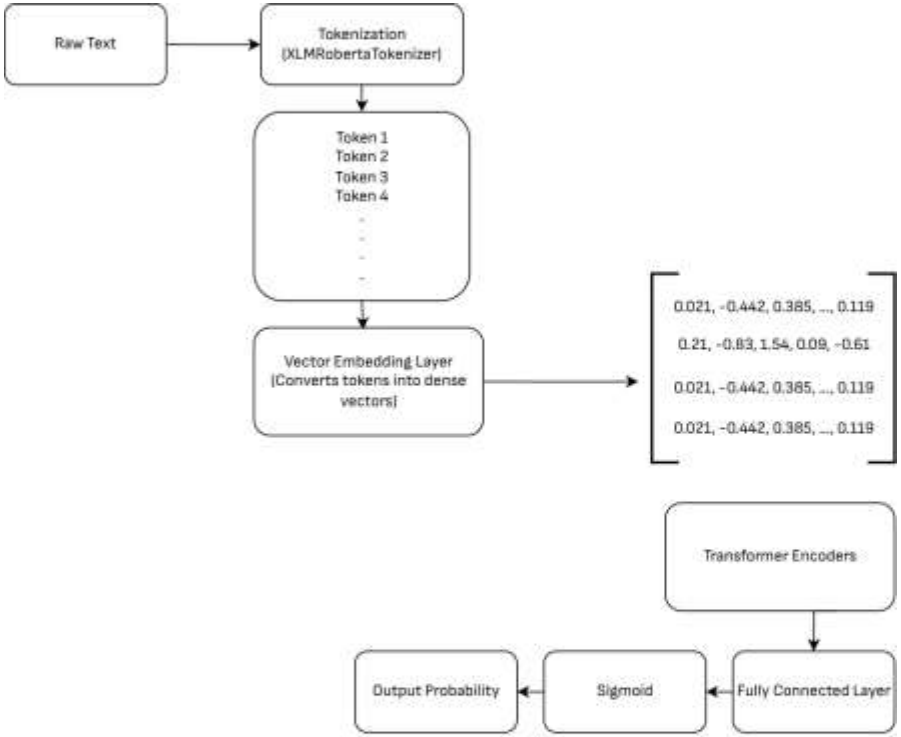
These models had limitations:

- **TF-IDF** ignored word order and context.
- **RNNs/LSTMs** processed tokens sequentially → **slow & hard to parallelize**.
- Long sequences caused **vanishing gradient** and **loss of long-term dependencies**.

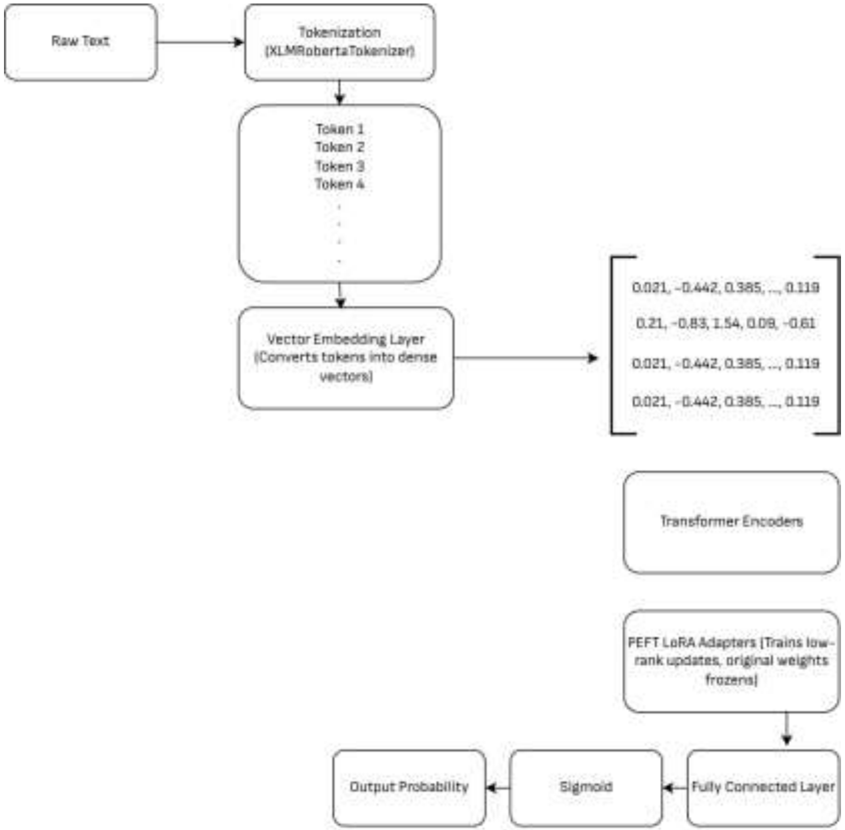
Transformers solved these with **self-attention**:

- **Attention mechanism** lets the model **focus on all words at once**, not just in order.
- Captures **global context** of a sentence, not just nearby tokens.
- Allows **parallel computation** → drastically faster training.
- Learns which words are **important for a given token**, improving contextual understanding.





General Workflow of XLM-RoBERTa



General Workflow of LoRA Fune-tuned XLM-RoBERTa

Hyperparameters and Training Arguments

LoRA Configuration (PEFT)

- **Task Type:** Sequence Classification (SEQ_CLS)
- **Rank (r):** 8
- **LoRA Alpha:** 16
- **LoRA Dropout:** 0.1
- **Target Modules:** Query, Key, Value, Dense
- **Bias:** None
- **Fan-in/Fan-out:** False
- **Inference Mode:** False

Checkpointing & Evaluation

- **Evaluation Strategy:** Epoch
- **Checkpoint Save:** Every epoch (Max 2 saved)
- **Best Model:** Not loaded at end
- **Metric:** Eval Accuracy (Higher is Better)

Training Arguments (Trainer)

- **Epochs:** 3
- **Batch Size (Train/Eval):** 8 / 8
- **Gradient Accumulation:** 2 → (Effective Batch Size: 16)
- **Learning Rate:** 2e-4
- **Weight Decay:** 0.01
- **Warmup Ratio:** 0.06 (Linear Scheduler with Warmup)
- **Logging Steps:** 100

Other Settings

- **Precision:** FP16 = False, BF16 = False
- **Max Gradient Norm:** 1.0
- **Remove Unused Columns:** False
- **Dataloader Workers:** 0 (No parallel workers)
- **Persistent Workers:** False
- **Drop Last Batch:** False
- **Prediction Loss Only:** False
- **Save Format:** SafeTensors = True

Why we utilized LoRA

- We used PEFT LoRA to fine-tune our model efficiently on low-resource hardware.
- Training time was reduced from ~24 hours (full fine-tune) to just 7 hours using LoRA.
- LoRA significantly reduced GPU VRAM usage, making training feasible on consumer-grade GPUs.
- RAM consumption was much lower compared to full fine-tuning approaches.
- Despite efficiency, the model maintained strong accuracy and generalization.
- Only a small subset of parameters were trained, reducing overfitting risks.
- LoRA adapters are modular, reusable, and easy to switch for different tasks.
- Compatible with Hugging Face's PEFT and transformers library, simplifying implementation.
- Supports layer freezing, further speeding up training and reducing memory use.
- Enabled cost-effective model adaptation without needing expensive cloud GPUs.

Why avoided full-fine tuning

- Full fine-tuning requires expensive high-VRAM GPUs, which we didn't have.
- It consumes significantly more RAM, making it impractical on standard systems.
- Training time is much longer, often exceeding 24–30 hours.

03

Results

Now in the last PPT we will show how and what accuracy we able to get

Overall accuracy

Model Name	Accuracy (%)	Special Feature	Batch Size	Epoch
XLM-RoBERTa Base	95.12%	Privacy Preserved, Multi-lingual	8	3
mDeBERTa V3 Base	95.41%	Privacy Preserved, Multi-lingual	8	3
T5	94.32%	Privacy Preserved, Multi-lingual	8	3
TinyLlama	94.96%	Privacy Preserved, Multi-lingual	8	3

04

Results

Now in the last PPT we will show how and what accuracy we able to get

05

Human Feedback Learning

Identify Uncertain Predictions
Run fine-tuned LLM on dataset.
Extract samples with probability **0.4–0.6** (borderline cases).

Human Annotation (Feedback)
Present uncertain samples to annotators/experts.
Label as **True / False / Neutral**.
Optional: Use disagreement as signal → assign Neutral.

Build Feedback Dataset
Combine human-labeled ambiguous samples into a mini dataset.
Example: 10k total samples → ~800 uncertain samples verified.

Key Takeaways
Reduces bias in the 0.4–0.6 probability zone.
Improves model robustness on uncertain cases.
Practical, research-oriented **human-in-the-loop refinement**.
Enhances accuracy across languages without full RL or costly retraining.

- Achieved robust multilingual fake news detection with minimal computational resources.
- Proved the feasibility of training advanced NLP models on consumer-grade hardware.
- Significant accuracy improvement over base papers despite **low epoch count.**
- Contributed to privacy-preserving NLP solutions by reducing the need for large-scale fine-tuning.
- All models were fine-tuned using **LoRA with PEFT**, enabling training within **7 hours** on **consumer-grade GPU.**
- Maintained strong generalization across diverse datasets despite limited training epochs.
- **No compromise on performance**, even under **RAM and VRAM constraints.**
- Models retain **modularity and reusability** for downstream tasks.

THANK YOU