NLP PROJECT
MEDICAL CHATBOT

Group 12
Muhammad Faraz Ahmad
Muhammad Maaz Barki
Ameer Hamza

AGENDA

- Project Scope
- Dataset Collection
- Fine-tuning LLMs using Unsloth with LoRA
- Hybrid RAG Integration with multilingual-e5-large and BM25
- Evaluation: Automatic + Human
- Streamlit

PROJECT SCOPE

- **Project:** To develop a lightweight and accurate medical-domain chatbot for Roman-Urdu that answers medical questions precisely.

- Initially, the goal was to have a separate intent classification module. Not considered due to a lack of a dataset for training.

- **Simplified Architecture:** Using small-sized pre-trained LLMs and fine-tuning them on our custom curated dataset, and using them with RAG.

DATASET CREATION

- Lack of a standard dataset for the medical domain in Urdu/Roman Urdu
- Even translating to Urdu/Roman Urdu has several issues:
  - Cost
  - Time-consuming
  - Poor Translation quality for the medical domain
- Even models like *indictrans2, opus-mt-en-ur* failed significantly to convey the patient's intent
- Final Strategy: Translation via LLMs with paid APIs (gpt-4o-mini)
- Result: Training split ~21,000 rows; validation split ~5200 rows.

FINE-TUNING LLMS

- Selected 3 LLMs of different sizes to fine-tune on our dataset:
  - llama-3.2-1B-Instruct (fast, inaccurate)
  - llama-3.2-3B-Instruct (slower, more accurate)
  - qwen-2.5-7B (slowest, most accurate)
- Different sizes to compare quality vs inference speed tradeoff
- Trained using *unsloth* with LoRA for faster & memory-efficient training.
- Trained for ~1 epoch with batch_size=8, lr=2e-4.
- Time: 1h15m (1B), 3h30m (3B), 8h (7B)

FINE-TUNING RESULT

- Each model got fluent in Roman Urdu, often using commonly used phrases
- Much better than their pre-trained versions
- **Hallucination issue:** While very fluent and having reasonable medical knowledge, all models frequently hallucinated on medical facts, often starting storytelling
- **Solution:** Hybrid RAG

HYBRID RAG INTEGRATION

- For RAG, we used a hybrid setup: vector embeddings + exact keyword matching (BM25)
- Created a short, high-quality, structured dataset in Roman Urdu containing information about 50 common diseases
- Used *intfloat/multilingual-e5-large* model to create vector embeddings, as all others, like *all-MiniLM-L6-v2*, are trained on mostly English, so they struggle with Roman Urdu.
- Used BM25 for exact keyword matching as *multilingual-e5-large*, despite being better than others, still struggles.
    - *multilingual-e5-large* treats all “dard” the same: joron ka dard and pait ka dard were considered the same by e5-large

EVALUATION METRICS

Name	Basic Information	Why use these?
BERT-F1	**Semantic metric:** Provides a single, balanced number that measures the semantic similarity between a machine-generated "candidate" sentence and a human-written "reference" sentence	Language-agnostic
ROUGE-L	**Lexical metric:** It looks for exact word matches	Used as a standard comparison against CHRF++, does not tolerate typos
CHRF++	**CHRF:** **Character n-gram Level Granularity:** Good for languages with complex grammar (like Urdu or German) where words change their endings **CHRF++:** **Word n-grams:** Adds word unigrams and bigrams into the mix. Helps ensure that while the spelling is right, the word order is also sensible	Best for "morphologically rich" languages

EVALUATION RESULTS

Name	Base Model	Pre-RAG(Finetuned)	Post-RAG
Llama 1B	• BERT-F1 - 0.7350 • ROUGE-L - 0.0384 • CHRF++ - 6.6217	• BERT-F1 - 0.8413 • ROUGE-L - 0.1173 • CHRF++ - 18.5206	• BERT-F1 - 0.8359 • ROUGE-L - 0.1095 • CHRF++ - 16.5166
Llama 3B	• BERT-F1 - 0.7586 • ROUGE-L - 0.8394 • CHRF++ - 9.0776	• BERT-F1 - 0.8411 • ROUGE-L - 0.1189 • CHRF++ - 18.4851	• BERT-F1 - 0.8380 • ROUGE-L - 0.1131 • CHRF++ - 16.8233
Qwen-2.5 7B	• BERT-F1 - 0.7138 • ROUGE-L - 0.0191 • CHRF++ - 3.6988	• BERT-F1 - 0.8417 • ROUGE-L - 0.1229 • CHRF++ - 18.9669	• BERT-F1 - 0.8394 • ROUGE-L - 0.1172 • CHRF++ - 18.4333

LIMITS AND CHALLENGES

- Most project limitations arose due to Roman-Urdu being morphologically rich as well as having different variations of the same word existing which created issues when trying to find the semantic similarity or word-based matching in the pipeline.
- No proper dataset exists for Roman Urdu usage that could be utilized, required translation using other LLMs.
- Evaluation also faced issues due to Roman Urdu word based comparisons not being too viable due to slight differences in spelling which led to heavy penalization. Semantic based similarity also faced similar issues.

THANK YOU