Chemistry Bot – Preprocessing & Model Training Using Meta Llama

Abstract

This project focuses on building an intelligent Chemistry Bot using Meta Llama, leveraging Hugging Face datasets and transformers. The system preprocesses multiple chemistry-related datasets, applies filtering, prepares structured instruction–response pairs, and trains a Meta Llama model using LoRA/QLoRA for efficient fine-tuning. The goal is to create a chemistry assistant capable of answering questions, explaining concepts, and assisting with chemical reactions, molecules, and atomic-level reasoning.

Pipeline (Process)

1. Load datasets from Hugging Face Hub using streaming.
2. Apply chemistry-specific filtering (reaction, molecule, atom keywords).
3. Extract instruction → output pairs and remove duplicates.
4. Limit dataset size using MAX_SAMPLES for fast training.
5. Convert processed data to JSONL for model training.
6. Load Meta Llama model and tokenizer.
7. Attach LoRA/QLoRA adapters for efficient fine-tuning.
8. Train using Causal LM objective.
9. Export the final fine-tuned chemistry model.

Learning Technologies

• Meta Llama Transformer (Hugging Face)
• Hugging Face Datasets
• Python for preprocessing
• Transformers library
• LoRA / QLoRA
• Sklearn for splitting
• JSONL conversion tools

Conclusion

The Chemistry Bot project demonstrates a complete and optimized workflow for fine-tuning language models for a specialized domain. By using streaming datasets,