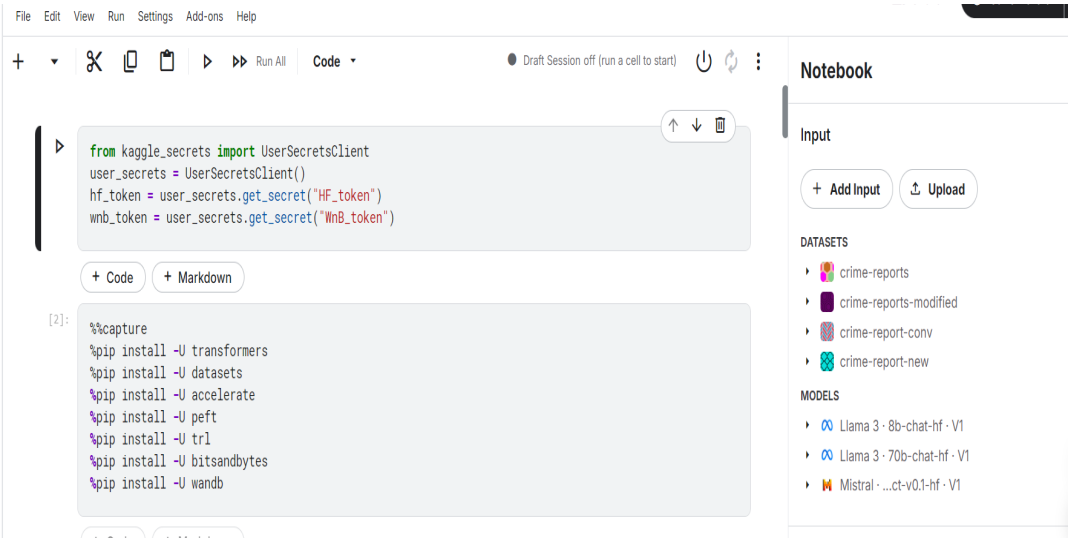
Special Detective – Conversational NER with LLMs

This project applies a new approach to Named Entity Recognition (NER) using Large Language Models (LLMs) in a conversational style. The goal is to fine-tune efficient LLMs (e.g., LLaMA-3 8B, Mistral 7B) on generated datasets of crime reports and build a demo app called **Special Detective** that can answer entity-based questions (e.g., suspects, victims, locations) in JSON or structured formats.

Current Progress

The project already has strong foundations implemented: - Data generation code in Colab for creating synthetic crime reports. - Fine-tuning setup on Kaggle using Hugging Face Transformers with LoRA/QLoRA. - Training tracked with Weights & Biases (W&B;), showing loss curves and metrics.

Kaggle Environment Setup:



W&B; Training Dashboard:

Data Generation Example

Code Snippet:

```
# Generate 1000 examples examples = [generate_example(i+1) for i in range(1000)] # Convert to JSON format json_output = json.dumps(examples, indent=4) # Save to a file (optional) with open("car_theft_reports_with_entities_and_negatives.json", "w") as file: file.write(json_output) print("Generated 1000 car theft report examples with entities and negative sampling successfully!")
```

Example Annotated Passage:

User: Text: Crime Type: Assault Date and Time: April 06, 2020, at 14:45 Location: 305 Pine Street, Centerville Reporting Officer: Detective Jane Johnson Summary: A black Chevrolet Malibu was reported stolen. Assistant: I've read this text. User: What describes Location in the text? Assistant: {"Location": ["305 Pine Street, Centerville"]} User: What describes Officer_Name in the text? Assistant: {"Officer_Name": ["Detective Jane Johnson"]} User: What describes Victim_Name in the text? Assistant: {"Victim_Name": ["Mr. Jane Rodriguez"]} ... (additional Q&A; truncated for brevity)

Remaining Work (Freelancer Tasks)

Tasks
Improve data preprocessing & formatting
Refine fine-tuning pipeline (LoRA/QLoRA)
Add evaluation metrics (precision, recall, F1)
Enhance W&B logging and visualization
Build a simple Flask demo app ('Special Detective')

Tools & Frameworks

- Python - Hugging Face Transformers, Datasets, PEFT (LoRA/QLoRA) - Weights & Biases (W&B;) - Kaggle GPU - Flask (for demo app)

Expectations

- Clean, modular, documented code - Fixed-price per task (15–30 USD each) - Approximate total project budget: ~200 USD - Weekly milestones with review before payment