Vyorius Test (AI Moderation & Automation Intern)

Objective

Create a Python application that reads user comments from a local file (e.g., CSV or JSON), uses a Generative AI model to detect offensive or inappropriate content, and generates a report of flagged comments.

Requirements

Part 1: Load and Process Comment Data

1. Use a sample data file containing user comments (provided or generated).
    - Format: CSV or JSON
    - Fields: comment_id, username, comment_text

2. Write a Python script to:
    - Read and parse the file
    - Display a summary (e.g., total number of comments, sample preview)

Part 2: Offensive Comment Detection using Gen AI

1. For each comment:
    - Use a Generative AI model (e.g., via OpenAI API or any available LLM) to:
        - Determine if the comment is offensive
        - Classify the offense type (e.g., hate speech, toxicity, profanity, harassment)
        - Provide a short explanation

2. Add these fields to each comment:
    - is_offensive (True/False)
    - offense_type
    - explanation

Part 3: Output & Reporting

- Export the analyzed data to a new CSV or JSON file
- Print a summary report:
    - Number of offensive comments
    - Offense type breakdown
    - Top 5 most offensive comments (by severity)

Bonus Points:

- Add a simple CLI to choose the input file or filter results
- Create a bar chart or pie chart showing offense type distribution
- Use profanity libraries (like profanity-check, better-profanity) for pre-filtering

Deliverables

- Python script(s)
- Sample input file (`comments.csv` or `comments.json`)
- Sample output file with flagged data
- README with:
  - Setup instructions
  - How to use the script
  - Sample outputs