Notch Tech AI Task

Objective:
Build a pipeline that generates a detailed caption for an image using both a pre-trained computer vision model (for object detection) and an LLM for contextual enhancement of the description.

Requirements:

1. **Input:**

   use the provided images in the attachment of this email.

2. **Output:**

   A detailed caption describing the image and hypothesizing its context (e.g., if it's a park, the system might add: "This could be a summer afternoon").

3. **Steps:**
   - Image Analysis: Use a pre-trained computer vision model (e.g., YOLOv8, EfficientNet, or CLIP) to detect objects and features in the image.
   - Textual Enhancement: Use an LLM (like OpenAI GPT-3.5 Turbo, GPT-4, or similar open-source alternatives like BLOOM) to generate a detailed and context-rich caption based on the extracted features.
   - Integration: Combine the outputs from the CV model and LLM into a cohesive text description.
   - Optional Challenge (Bonus): Allow the system to take a user-provided keyword (e.g., "summer") and modify the generated caption to include that context.

Evaluation Criteria:

- Correctness and Creativity: Does the pipeline generate accurate and context-rich captions?
- Code Structure: Is the code modular, well-documented, and easy to follow?

- Efficient Usage of Pre-trained Models: How effectively are pre-trained models integrated for both CV and LLM tasks?
- Handling Edge Cases: Does the system gracefully handle images with no recognizable objects or ambiguous scenes?

Expected Deliverables:

1. A Python script or Jupyter Notebook demonstrating the task.
2. A brief README file explaining:
    - Steps to run the script.
    - Models used and why they were chosen.
3. Sample outputs for the provided test images.

Deadline:

You have 48 hours to submit the task. Create a private Github repository, showing commit history, and email us the link to the repo.