AI Project Pipeline & Future Ideas

1. Mini-GPT From Scratch
   - Data Pipeline: auto-download Tiny Shakespeare, char-level tokenizer, TextDataset with __init__, __len__
   - Model: TransformerBlock (multi-head attention, layernorm, feedforward, residuals), MiniTransformer clas
   - Training Loop: train.py loads config.yaml, DataLoader, model instantiation, CrossEntropyLoss, AdamW, e
   - Testing: test_transformer_block.py for shape checks; test_dataset.py for data integrity.

2. Self-Critique Safety Agent
   - Generate: wrap MiniTransformer for text generation.
   - Probe: generate_adversarial_prompts in probe.py using templates and NLP fuzzing.
   - Critique: safety_critique in critique.py calling moderation API or second LLM to flag unsafe content.
   - Refine: SafetyAgent agent.py orchestration: generate -> probe -> critique -> refine, providing final safe re
   - Evaluation: evaluate.py batch-run metrics on adversarial prompt suite.

3. Runway Gen-4 Integration
   - Initialize RunwayML client with API key in agent.py.
   - visualize() method calls textToImage or textToVideo for final text.
   - Multimodal Demo: show refined text + generated image/video.

4. Interactive Demo UI
   - Streamlit/Gradio app (app.py): input prompt, display pipeline trace, safety flags, final text, and Runway im
   - QR code generation for shareable link at networking events.

5. Additional Project Ideas
   A) Retrieval-Augmented QA Chatbot
      - ingest.py: embed docs, FAISS indexing.
      - qa.py: retrieve top-k chunks, prompt assembly, call LLM for answers.
      - app_rag.py: Streamlit UI for question and answer with source snippets.
   B) Custom Voice-Cloning TTS Demo
      - tts_demo.py: Eleven Labs API integration for custom voice cloning.
      - app_tts.py: interface for text input and audio playback.
   C) Adversarial Red-Teaming Toolkit
      - generate_adv.py: adversarial prompt generator.
      - classify_adv.py: safety classifier metrics.
      - app_redteam.py: UI for testing prompts and viewing unsafe cases.
      - Mitigation strategies and evaluation suite.

6. Timeline & Next Steps
   - Finish Mini-GPT & Safety Agent by Tuesday.
   - Integrate Runway Gen-4 visuals Wednesday.
   - Build and polish Interactive UI Thursday.
   - Draft README, demo video, and apply by end of week.