**PROJECT PLANNING**

**1.Task Breakdown**

* Intializing CodeLlama

Research HuggingFace integration

Choose appropriate model variant (CodeLlama-7b-hf)

* Model Integration

Use transformers.pipeline() to wrap the model

Configure generation parameters (temperature, top-p, max length)

* Code Generator Logic

Write a function generate\_code() to accept prompt and return generated text

Include inference time tracking

* User Interface Development

Use Streamlit to collect user input and display generated code

Add sidebar with usage guide and examples

* Model Deployment

Run app locally with streamlit run

Use localtunnel to expose the app to the internet

Validate accessibility and performance

* Testing Strategy

Test with multiple prompts (simple to complex)

Ensure accuracy, relevance, and safety of output

Confirm UI responsiveness

**2. Risk Management**

Long model loading time - Use smaller or quantized model variants, optimize hardware (GPU, RAM)

Localtunnel connection drops - Use alternative tunneling tools like ngrok or deploy on cloud

Incorrect or insecure code output - Add prompt validation or a warning disclaimer in the UI. UI freezes or fails on large prompts Limit max\_length and display warning messages