PROJECT MILESTONE - Phase 1 - Edge Deployment Planning & GitHub Setup: Identify an edge AI application scenario for your individual project from the sample project options. Define resource constraints, user requirements, and performance targets. Create a project plan that includes model selection, optimization approach, and deployment strategy. Initialize a GitHub repository with appropriate README, structure, and documentation templates as specified in the GitHub requirements.

PROJECT MILESTONE - Phase 1 - Model Conversion & Analysis: Select and convert two modern AI models to ONNX format. Perform basic profiling to identify resource requirements and potential bottlenecks. Create a baseline performance report documenting memory usage, inference time, and accuracy metrics. Commit all code, configuration files, and documentation to your GitHub repository with appropriate version control and documentation.

PROJECT MILESTONE - Phase 2 - Model Architecture Optimization: Implement at least two optimization techniques (distillation, pruning, or architecture modification) on your selected models. Document the impact on model size, inference speed, and accuracy. Create a visualization comparing the tradeoffs between the original and optimized models. Update your GitHub repository with optimization code, results, and thorough documentation of your methodology.

PROJECT MILESTONE - Phase 2 - ONNX-Specific Optimization: Apply ONNX-specific optimizations to your models, including quantization and graph optimizations. Create an optimization pipeline that can be applied to similar models in the future. Benchmark the fully optimized models on a standard laptop CPU and document the improvements over baseline. Update your GitHub repository with reproducible optimization pipeline code and benchmarking results.

PROJECT MILESTONE - Phase 3 - Edge Text Generation: Deploy a compact LLM using ONNX Runtime's CPUExecutionProvider. Implement optimized inference with proper KV cache management and response streaming. Benchmark the model's performance on a standard laptop, measuring tokens per second, memory usage, and latency. Update your GitHub repository with edge-optimized inference code, benchmarking utilities, and detailed performance analysis.

PROJECT MILESTONE - Phase 3 - Edge Image Generation: Deploy a lightweight image generation model using ONNX Runtime. Implement memory-efficient generation strategies to enable running on limited hardware. Create a demonstration application that allows users to generate images with various prompts and settings while monitoring resource usage. Update your GitHub repository with the complete edge-optimized image generation pipeline and demonstration application.

PROJECT MILESTONE - Phase 4 - Edge Computer Vision Pipeline: Deploy an advanced computer vision model (object detection or segmentation) optimized for edge deployment. Create an efficient end-to-end pipeline from image capture to result visualization that can run on standard laptop hardware. Benchmark different optimization techniques and document the performance gains. Update your GitHub repository with the complete computer vision pipeline and comprehensive documentation of your optimization techniques.

PROJECT MILESTONE - Phase 4 - Real-time Vision Application: Create a real-time computer vision application using a webcam input that can process frames efficiently on CPU. Implement memory-efficient tracking of objects across frames and optimize the end-to-end pipeline to achieve maximum possible FPS on standard laptop hardware. Document the optimization strategies used and their impact. Update your GitHub repository with the complete real-time vision application and detailed performance analysis.

PROJECT MILESTONE - Phase 5 - Edge Speech Recognition: Deploy an optimized speech recognition model using ONNX Runtime's CPUExecutionProvider. Implement efficient audio processing and streaming recognition. Create a demonstration that can transcribe speech in real-time on standard laptop hardware while monitoring resource usage. Update your GitHub repository with the complete speech recognition component and integration instructions.

PROJECT MILESTONE - Phase 5 - Complete Multimodal Edge AI System & Demo: Integrate optimized text, vision, and speech models into a cohesive edge AI application that can run efficiently on standard laptop hardware. Create a compelling 10-minute demonstration that showcases your system handling complex multimodal tasks while maintaining responsive performance. Present performance metrics comparing your edge-optimized solution against non-optimized baseline models. Finalize your GitHub repository with comprehensive documentation, installation instructions, and performance analysis.

PROJECT MILESTONE - Final Delivery: Deliver a compelling "Edge AI Product Demo" that demonstrates your system's capabilities, highlights the technical optimizations achieved, and presents a deployment strategy for resource-constrained environments. The demonstration should include live performance on standard CPU hardware and quantitative comparisons showing the improvements over non-optimized approaches. Your final GitHub repository should serve as a complete reference implementation that others could use to reproduce your work.