**Performance Evaluation of Whisper Transcribe App with different Compute Unit Configurations**

This document summarizes the performance results of **WhisperKit** when changing compute unit options for different stages of the transcription pipeline.

**Configuration Overview**

The WhisperKitConfig was initialized with varying MLComputeUnits for each stage:

public init(

melCompute: MLComputeUnits = .cpuAndGPU,

audioEncoderCompute: MLComputeUnits? = nil,

textDecoderCompute: MLComputeUnits = .cpuAndNeuralEngine,

prefillCompute: MLComputeUnits = .cpuOnly

) { }

**Results Summary**

**1. CPU Only**

* **Load Time:** ~0 sec (no special load required)
* **Transcription Time:** **400+ sec**

**2. GPU**

* **Load Time:** ~30 sec
* **Transcription Time:** **240 sec**

**3. Neural Processing Unit (NPU)**

* **Load Time:** ~90 sec
* **Transcription Time:** **110 sec**

**4. NPU with Prewarm Mode**

* **Load Time:** ~50 sec
* **Transcription Time:** **80 sec**

**Key Insights**

* **CPU-only performance** is prohibitively slow for real-world use.
* **GPU** offers significant improvement, cutting transcription time almost in half compared to CPU.
* **NPU** delivers the best balance, drastically reducing transcription time, though with a higher initial load.
* **Prewarm mode on NPU** further optimizes startup, lowering both load and transcription times.