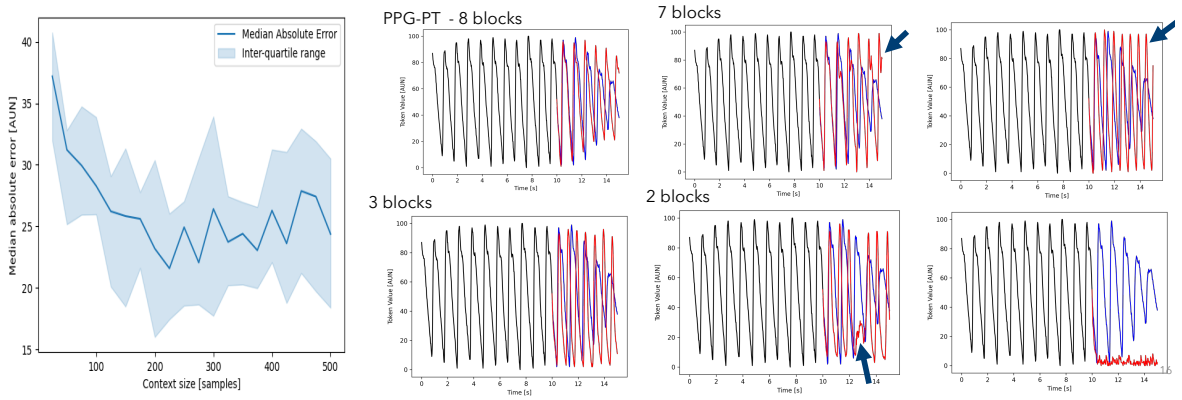
Can Transformers model for PPG generation run on microcontrollers?

The PPG-PT, small generative model for PPG, utilizes a 101-token vocabulary to represent signal amplitudes and a 500-token context window (equivalent to 10 seconds of data at 50Hz) processed through 8 transformer layers. Consequently, the model comprises 443,622 parameters, stored as 4-byte floats. The memory requirements exceed the capacity of common wearable microcontrollers, such as the nRF52840, which possesses only 1024KB of Flash memory. In this study, we explored the optimization and deployment of this generative pre-trained transformer onto resource-constrained microcontrollers.

We found that halving the context size to 250 samples do not affect the model's accuracy, effectively balancing performance and computational efficiency and result in 3.6% reduction in parameters. To further refine the model for the constraints of an embedded system, we conducted a layer redundancy analysis using the Mean Block Influence metric. This approach quantifies the contribution of each transformer block and allows to rank layer and find ones that performing nearly identity-like transformations and may be redundant. We systematically evaluated the model by removing layers with the lowest influence scores one by one.

Our pruning experiments revealed that while the architecture remains functional with as few as three transformer layers, there is a noticeable degradation in the morphological reliability. At this reduced depth, the model successfully generates a periodic, pulse-like waveform. However, critical physiological details, such as the inflection point and the dicrotic notch, are lost.

By halving the context size and pruning redundant transformer layers, we successfully reduced the model's complexity to a level that allows the PPG-PT architecture to be deployed on a standard microcontroller. Future work will examine the application of post-training quantization to compress, which is expected to yield a fourfold reduction in memory footprint and enhance energy efficiency.



On the left, the median absolute error as a function of context size; on the right, the impact of transformer removal on the predicted signal (red) relative to the true reference signal (blue).