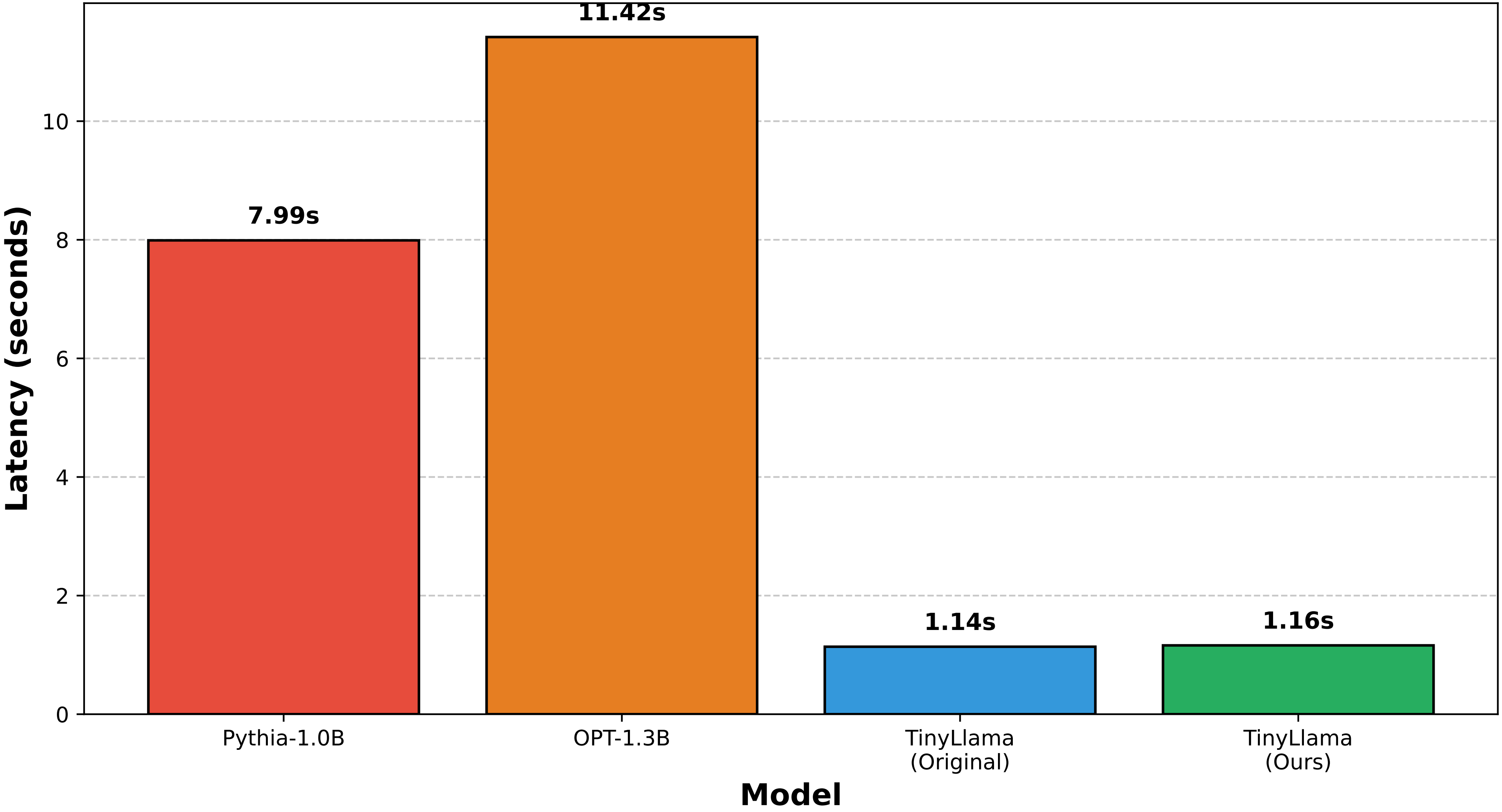
Figure 4: Inference Latency Comparison



TinyLlama maintains low latency while achieving superior determinism.