



# Inference Acceleration for Large Language Models on CPUs

{jithinvg, dittops, adarsh.ms}@bud.studio

## Abstract

In recent years, large language models have demonstrated remarkable performance across various natural language processing (NLP) tasks. However, deploying these models for real-world applications often requires efficient inference solutions to handle the computational demands. In this paper, we explore the utilization of CPUs for accelerating the inference of large language models. Specifically, we introduce a parallelized approach to enhance throughput by 1) Exploiting the parallel processing capabilities of modern CPU architectures, 2) Batching the inference request. Our evaluation shows the accelerated inference engine gives an **18-22x** improvement in the generated token per sec. The improvement is more with longer sequence and larger models. In addition to this, we can also run multiple workers in the

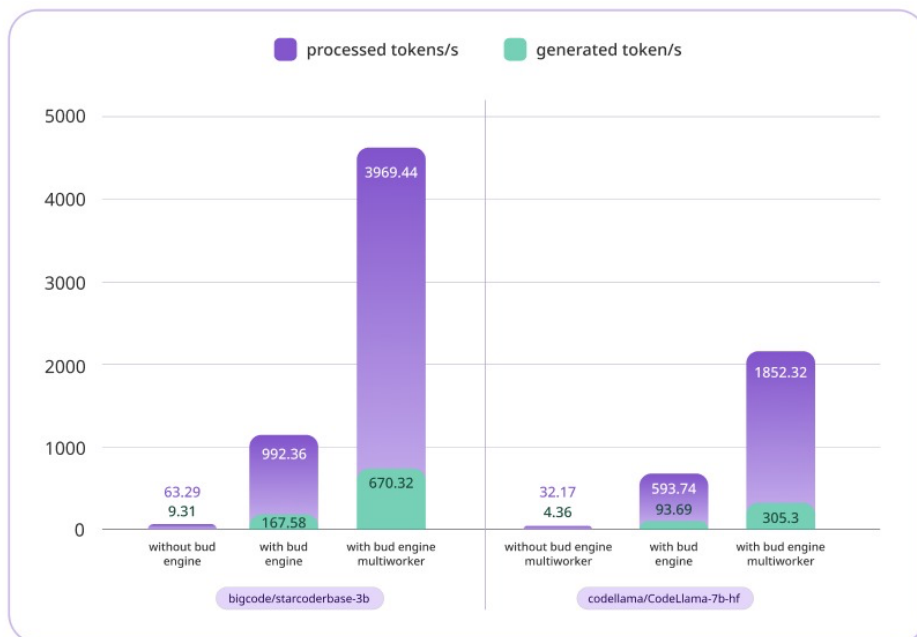


Figure 1: Improvement of token/s with the use of Bud Inference engine with 32 vCPU on 4th Gen Intel® Xeon® Scalable Processors