intel

8.03s TTFT

for 2 NUMA nodes and 100 concurrent requests.[1]

17.18s TTFT

in for 90 percentile requests for 200 concurrent requests.[2]

Benchmarking the Indus Language Model on Intel® AI Hardware

Project Indus is an innovative open-source language model designed specifically for Hindi and its dialects. Focusing on applications within the Indian linguistic landscape, Project Indus aims to enhance natural language generation and processing capabilities. A joint study from Intel and Tech Mahindra benchmarked key performance metrics such as Time to First Token (TTFT), inter-token delay, input prompt length, output prompt length, and total throughput in tokens per second. By evaluating these parameters under various conditions, including varying numbers of concurrent requests, a detailed performance profile of the Indus large language model (LLM) on Intel® AI hardware was obtained. The results highlight the model's effectiveness and scalability, offering valuable insights for optimizing its practical implementation. The study aims to inform developers and researchers about the performance characteristics of the Indus LLM, facilitating its integration and utilization across diverse computational environments.

Products and Solutions
5th Gen Intel® Xeon® Scalable Processors
Intel® Advanced Matrix Extensions
Intel® Advanced Vector Extensions 512

Industry
IT Services & IT Consulting

Organization Size
11-50

Country
India

Partners
AWS
Bud Ecosystem
Tech Mahindra

Learn more
White Paper

1, 2 For more complete information about performance and benchmark results, visit https://www.intel.com/content/www/us/en/customer-spotlight/stories/indus-project-customer-story.html