

Benchmarking the Indus Language Model on Intel® AI Hardware

Intel and Tech Mahindra conducted a joint study on measuring the performance of Project Indus, an open source language model specifically developed for Hindi and its dialects.

Authors:

Tech Mahindra

B. Nilesh

Principal Technical Architect, Tech Mahindra

S. Vinay

Lead Architect, Tech Mahindra

M. Satish

Lead Architect, Tech Mahindra

Intel

Rupak Chakraborty

PSAM, Intel SMG

Executive Summary

This white paper presents a comprehensive benchmarking study of Project Indus, 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. The benchmarking study emphasizes 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 LLM on Intel® AI hardware was obtained. The results highlight the model's effectiveness and scalability, offering valuable insights for optimizing its practical implementation. This study aims to inform developers and researchers about the performance characteristics of the Indus LLM, facilitating its integration and utilization across diverse computational environments.

Approach and Methodology

The introduction of Generative AI (GenAI) has marked a significant technological advancement globally. However, in India, no one has really yet created a large language model (LLM) from scratch, despite leading companies like OpenAI, Microsoft, Google, and Meta developing and releasing LLMs as open-source alternatives.

India is a linguistically diverse country with 1,645 dialects and 27 official languages. Hindi alone has over 600 million speakers and encompasses numerous dialects, some spoken by more than a million people. Existing models from top firms have struggled to effectively respond to inquiries in these various dialects. To address this issue, it became essential to develop a foundational model in Hindi that comprehensively includes all its dialects.

Project Indus primarily aims to bridge the digital divide in India by offering technologically advanced solutions that cater to individuals from diverse linguistic backgrounds. By facilitating the preservation of critical knowledge contained in endangered languages, it has the potential to evolve as a civilizational initiative. However, the reality of these large language models is that they carry a significant carbon footprint, making them environmentally unsustainable, whether used for training or inferencing. Thus, understanding the process of LLM creation and devising strategies to make it more sustainable and ethical emerged as secondary objectives.



Table of Contents

Executive Summary.....	1
Approach and Methodology.....	1
Performance Testing on Intel® AI Hardware on AWS	2
Benchmarking Results.....	3
Standalone Testing on 5th Gen Intel® Xeon® Processor Server.....	4
Conclusion.....	5

Makers Lab, the innovation team at Tech Mahindra, built Project Indus, a ground-up LLM from scratch. The project kick-started with an extensive outreach initiative to gather information about Hindi and its 37 dialects. The team collected 100 GB of data in Hindi and its dialects, making it the world's largest dataset for these languages. The model, characterized by an easy-to-use user interface (UI), stands as the most cost-effective LLM foundational model ever created. With 1.2 billion parameters, 22 billion tokens, a team of 15 people, and a budget of \$150,000, the model was trained on approximately 100 GB of news, Wikipedia, and book corpuses. These resources were diligently gathered, cleaned, annotated, and translated by the Makers Lab over the initial three months. CDAC GPUs, which involved 48 powerful 40 GB GPU computers running A100 GPU in parallel, were used to train this colossal dataset.

A fine-tuned Indus model with 1.2 billion parameters has been implemented for inference purposes on the 5th Gen Intel® Xeon® processor powered platform. Intel collaborated with ISV partner Bud Ecosystem to implement the concept for scalability and high throughput serving to support a large number of enterprises and end users. The scope of work followed an 'Across the Box Scale Out' approach, and the implementation was carried out using an 'In-Box Scale Out' approach with the Sub NUMA (Non-Uniform Memory Access) cluster.

The project is Intel-optimized, designed specifically to leverage platform features like Intel® Advanced Matrix Extensions (Intel® AMX) and Intel® Advanced Vector Extensions 512 (Intel® AVX-512) to enhance 'Inference Serving' performance. To ensure the optimal delivery of these optimizations, the ISV integrated all platform variables into a container image (VMI).

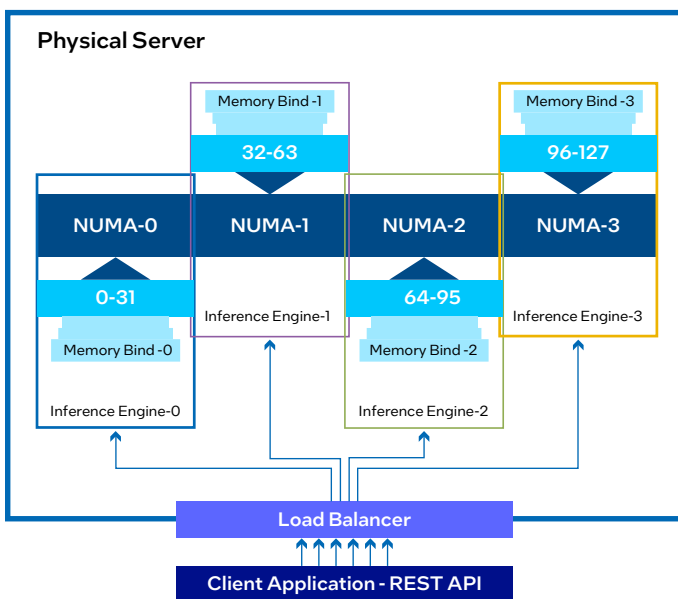


Figure 1. Inference Serving Architecture

Performance Testing on Intel® AI Hardware on AWS

NUMA is a shared memory architecture that describes how main memory modules are positioned relative to processors in a multi-processor system. Ignoring NUMA can lead to suboptimal application memory performance. However, by taking specific steps, NUMA-related performance issues can be mitigated or even leverage NUMA architecture to enhance parallel applications' performance. Key considerations include processor affinity, memory allocation through implicit operating system policies, and the use of system APIs for explicit memory page assignment and migration.

The Indus LLM was validated and benchmarked using the LLMPerf library. This benchmarking involved load tests that initiated multiple concurrent queries to the LLM API to quantify inter-token delay and generation throughput, both per request and across concurrent requests.

Below set of prompts were used for Performance testing with LLMPerf:

- क्या संगीत के माध्यम हम अपने भावनाओं को व्यक्त कर सकते हैं?
- हम अपने मन की चंचलता को शांति में कैसे बदल सकते हैं?
- हम अपने जीवन का सहज कैसे बना सकते हैं?
- हम परिपूर्णता कैसे प्राप्त कर सकते हैं ?
- जीवन में सही निर्णय कैसे लें?
- अपने प्रियजनों के साथ संवाद में स्नेह कैसे दिखा सकते हैं?
- विश्वास क्या है?
- हम अपने असली स्वरूप को कैसे पहचान सकते हैं?
- हम अपने शारीरिक, मानसिक और आत्मिक स्वास्थ्य को कैसे संतुलित रख सकते हैं?
- धैर्य क्या है?
- विवेक क्या है, और कैसे हम सही और गलत के बीच अंतर को समझ सकते हैं?
- क्या समृद्धि का सच्चा अर्थ सकारात्मक मानसिकता, स्वस्थ शारीरिक और मानसिक विकास में होता है?
- क्या संघर्ष के माध्यम से ही हम अपनी सीमाओं को पार करके नई ऊंचाइयों को छू पाते हैं?
- क्या सफलता के लिए संघर्ष, समर्पण, और निरंतर प्रयत्न आवश्यक होते हैं?
- संघर्ष, समर्पण, और निरंतर प्रयत्न किस प्रकार से सफलता प्राप्ति में सहायक होते हैं?
- सफलता प्राप्त करने के लिए निरंतर प्रयास और संघर्ष क्यों आवश्यक हैं?
- सफलता की दिशा में हमारी क्षमताओं को सही मार्ग पर किस प्रकार ले जाया जाता है?
- अपने लक्ष्यों की प्राप्ति में निरंतर प्रयास और संघर्ष किस प्रकार से हमारी मदद करते हैं?
- समृद्ध और खुशहाल जीवन के लिए निरंतर प्रयत्न, संघर्ष, और समर्पण का योगदान किस प्रकार आवश्यक है?
- भारतीय स्वतंत्रता संग्राम में अहिंसा की भूमिका का महत्व क्या था ?
- भारतीय समाज में वर्ण व्यवस्था का इतिहास किस प्रकार से हमें आज की जाति प्रथा की समझ देता है?
- भारतीय इतिहास में महिलाओं की भूमिका का विकास कैसे हुआ?
- क्या आधुनिक भारत की धार्मिक और सांस्कृतिक विविधता उसके इतिहास में निहित संघर्षों का परिणाम है?
- विदेशी आक्रमणों ने भारतीय सामाजिक-सांस्कृतिक ढांचे में कैसे परिवर्तन किये?
- गांधीजी का सत्य और अहिंसा का दर्शन कैसे भारतीय स्वतंत्रता संग्राम को आकार देने में मददगार रहा?

मुगल साम्राज्य के दौरान कला और संस्कृति के फलने-फूलने ने भारतीय आध्यात्मिकता और दर्शन को किस प्रकार प्रभावित किया?

कैसे प्राचीन भारतीय विद्यालयों जैसे नालंदा और तक्षशिला ने ज्ञान के प्रसार में योगदान दिया, और इससे हमें क्या सिखने को मिलता है?

भारतीय रेलवे की स्थापना ने देश की आर्थिक और सामाजिक संरचनाओं में किस प्रकार के परिवर्तन किए?

ब्रिटिश शासन काल में भारतीय कला और साहित्य में आयी नवजागरण की लहर का मूल्यांकन किस प्रकार से भारतीय पहचान और आत्म-अभिव्यक्ति की समझ में मदद करता है?

अंग्रेजी शासन के दौरान शिक्षा प्रणाली में लाए गए परिवर्तनों का दीर्घकालिक सामाजिक-आर्थिक प्रभाव क्या रहा है?

भारतीय स्वतंत्रता के पश्चात् भारतीय राजनीति में वंशवाद का उदय कैसे भारतीय लोकतंत्र के मूल सिद्धांतों पर प्रश्न उठाता है?

प्राचीन भारतीय शिक्षा प्रणालियों में गुरु-शिष्य परंपरा की भूमिका आज के शैक्षिक तंत्र में किस प्रकार से पुनर्विचारित की जा सकती है?

भारतीय राजनीति में धर्म की भूमिका क्या होनी चाहिए? क्या धर्म और राजनीति को पूरी तरह से अलग किया जा सकता है?

क्या लोकतंत्र में बहुसंख्यकवाद का प्रभाव पड़ता है? अगर हां, तो इसे कैसे संतुलित किया जा सकता है?

क्या आधुनिक भारतीय राजनीति में गांधीजी के आदर्शों का कोई स्थान है?

क्या विकास के लिए राजनीतिक स्थिरता जरूरी है?

क्या लोकतंत्र सिर्फ बहुमत का शासन होना चाहिए या उसमें विविधता और अल्पसंख्यकों की राय का भी महत्व होना चाहिए?

भारतीय राजनीति में युवा नेतृत्व की आवश्यकता कितनी महत्वपूर्ण है?

भारतीय राजनीति में मीडिया का प्रभाव किस हद तक उचित है? क्या मीडिया का राजनीतिक प्रभाव लोकतंत्र के लिए हानिकारक हो सकता है?

भारतीय राजनीति में जाति और राजनीति का संबंध क्या है? जाति आधारित राजनीति के दीर्घकालिक प्रभाव क्या हो सकते हैं?

क्या भारत के लोकतंत्र को और मजबूत बनाने के लिए राज्यों को और अधिक स्वायत्तता दी जानी चाहिए?

क्या भारत की राजनीतिक प्रणाली में महिलाओं का प्रतिनिधित्व बढ़ाने के लिए और कुछ किया जा सकता है?

क्या भारत की राजनीतिक प्रणाली में भ्रष्टाचार को पूरी तरह से खत्म किया जा सकता है?

The below LLMperf command was used for Performance Benchmarking Indus LLM on AWS Servers having Intel® AI hardware: `python token_benchmark_ray.py --model "Indus_SFT2" --mean-input-tokens 350 --stddev-input-tokens 50 --mean-output-tokens 350 --stddev-output-tokens 50 --max-num-completed-requests 25 --timeout 1800 --num-concurrent-requests 25 --results-dir "result_outputs" --llm-api openai --additional-sampling-params '{}`

Benchmarking Results

NUMA	Input Prompt Length	Standard Deviation	Output Prompt	Standard Deviation	Max Number of Prompt	Number of Concurrent Prompt	inter_token_latency_s (p90)	ttft_s (p90)	Overall Output Throughput (Tokens/s)
0	350	50	350	50	100	10	0.090	2.571	154
					200	20	0.119	5.528	224
					300	30	0.207	8.575	259
					400	40	0.236	11.530	267
					500	50	0.277	13.637	261
					600	60	0.310	17.558	267
					700	70	0.461	20.684	274
					800	80	0.436	22.331	294
					900	90	0.561	26.027	301
					1000	100	0.567	29.736	303
0,1	350	50	350	50	100	10	0.055	1.385	161
					200	20	0.063	3.216	270
					300	30	0.103	4.854	332
					400	40	0.134	6.415	387
					500	50	0.170	8.036	414
					600	60	0.186	10.341	446
					700	70	0.264	11.575	463
					800	80	0.286	13.608	455
					900	90	0.283	15.169	432
					1000	100	0.258	17.189	446

Standalone Testing on 5th Gen Intel® Xeon® Processor Server

The Indus model was also tested on 5th Gen Intel® Xeon® Processor Server having the below configuration:

- **CPU:** Intel® Xeon® Platinum 8580 Processor – 60 Cores per Socket: Total 120 Cores per Server
- **Memory:** 512GB
- **Storage:** 1TB x1 (OS), 1.7TB x 4 (Data)
- **NUMA:** 4 NUMA node(s)

A custom script was built to send prompts to model and generate response. Metrics like TTFT, Avg. tokens per second and E2E latency. Intel provided specific libraries were installed in python virtual environment to enable optimal use of underlying Intel® AI hardware. Below is an example:

```
intel-extension-for-pytorch==2.1.0
intel-extension-for-tensorflow==2.14.0.0
intel-extension-for-tensorflow-lib==2.14.0.0.2
intel-openmp==2024.0.2
```

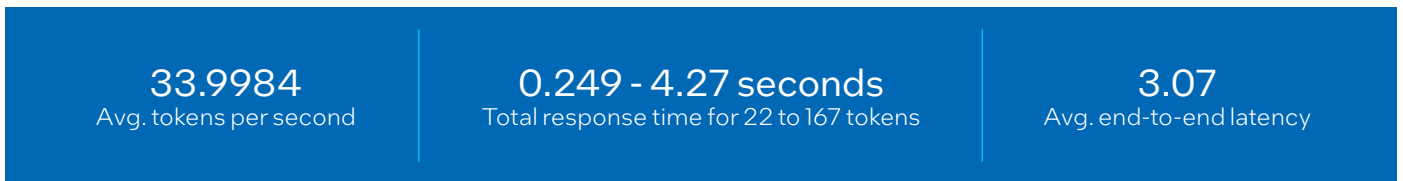
Below command is used to set environmental variables and run benchmarking:

```
CPU_ISA=amxONEDNN_MAX_CPU_ISA="AVX512_CORE_AMX"NUM_PARALLEL_EXEC_UNITS=240
OMP_NUM_THREADS=196 numactl --physcpubind=0-29,120-149,30-59,150-179,60-89,180-209,90-119,210-239
--membind=0,1,2,3 python benchmarking.py
```

Below are the results:

user_prompt	text_generated	to-kens	elapsed_time	time_per-to-ken	tokens_per_sec-ond
सम्राट अशोक के बारे में बताओ।	"उपयोगकर्ता सम्राट अशोक के बारे में बताओ। सहायक सम्राट, अशोक, भारत के इतिहास में एक महत्वपूर्ण व्यक्ति हैं, जिन्होंने गौरवशाली शासन का स्थापना की। उनका जन्म तीसरी शताब्दी ईस्वी में हुआ था और उन्होंने मौर्य साम्राज्य की स्थापना में महत्वपूर्ण भूमिका निभाई। उनसे भारतीय इतिहास का परिचायक माना जाता है और उन्हें अपने शासनकाल के दौरान भारत पर सबसे शक्तिशाली राजाओं में से एक माना जाता है। अशोक एक कुशल शासक थे और उन्होंने कला और वास्तुकला को बढ़ावा देना शुरू किया। उनके साम्राज्य में भारतीय संस्कृति और परंपराओं को संरक्षित करने और विस्तार करने की अपार क्षमता थी। सम्राट की शासन प्रणाली और प्रशासन को अशोक के शिलालेखों और अर्थात् सिंको में दर्शाया गया है, जिन्हें आज भी पढ़ा"	117	4.279255152	0.03657483	27.34120679
गीतम ब्रह्म कौन थे?	"उपयोगकर्ता गीतम ब्रह्म कौन थे? सहायक बौद्ध धर्म के संस्थापक गौतम ब्रह्म थे। उनका जन्म ईसा पूर्व छठी शताब्दी में लुम्बिनी, नेपाल में हुआ था। उन्होंने उर्ध्वेश और तपस्व्यता जैसे कई स्थानों पर साधना की। उन्हें अर्हंत, साधु, अर्हंत और निर्वाण का मार्ग अर्हंतों के लिए जाना जाता था। उन्होंने केवल ज्ञान प्राप्त करने के मोक्ष प्राप्त किया, न कि अपने ज्ञान में व्यर्थ करने।"	77	2.365980864	0.030727024	32.54464192
महाभारत में कृष्ण की क्या भूमिका है?	"उपयोगकर्ता महाभारत में कृष्ण की क्या भूमिका है? सहायक कृष्ण पांडवों को युद्ध में विजयी होने में मदद करने के लिए अपने ज्ञान और शक्ति का उपयोग किया। कृष्ण की कई रणनीतियाँ थीं जो उन्होंने अर्जुन को युद्ध के मैदान में सहायता प्रदान करने के लिए प्रयोग की थीं। वे बुद्धिमान और बहादुर थे, जिन्होंने अपना पूरा जीवन अपनी अतीत की कर्तव्यों को पूरा करने और केवल रणनीति के रूप में लड़ने में बिताया। उनकी रणनीतियों में शामिल हैं: १. रणनीति का ज्ञान: उन्होंने दुर्योधन को यह समझाने के लिए एक रणनीति तैयार की कि वह उन्हें जीत नहीं सकता। इस रणनीति का सारथी भीमार्जुन तैयार किया। २. कौरव और पांडव दोनों राजाओं के बीच भ्रम।"	116	3.737437487	0.032219289	31.03730843
वार्त्तिक पुस्तक में साक्षरता का सारांश क्या है?	"उपयोगकर्ता वार्त्तिक पुस्तक में साक्षरता का सारांश क्या है? सहायक वार्त्तिक पुस्तक में साक्षरता का सारांश है: 'वार्त्तिक पुस्तक, जिसे भीमनाया जाता है।'"	28	0.367489576	0.013124628	76.1926373
मुझे रामायण की कहानी बताओ।	"उपयोगकर्ता मुझे रामायण की कहानी बताओ। सहायक रामायण हिंदू धर्म में सबसे पवित्र ग्रंथों में से एक है और अक्षर राम के जीवन, शक्तियों और राक्षसों के अंत के बारे में बताती है। हालांकि, कुछ विवाद और असहमतियाँ हैं जो इसके ऐतिहासिक महत्त्व को समझने में बाधा डालती हैं। इस पाठ में एक महत्वपूर्ण विषय है रावण का बंध, जो वीर्य के सबसे शक्तिशाली अवतार हैं और जिन्हें रामायण में एक परमेश्वर के रूप में चित्रित किया गया है। हालांकि, जयकविह पाठ एक महाकाव्य है, यह कई अन्य महत्वपूर्ण पाठों और पद्यों के साथ भी जुड़ा हुआ है जो इसे प्रेरित करते हैं। १. रावण: रावण को एक सलतायक और दुष्ट के रूप में चित्रित किया गया है जो विजय और विनाश।"	116	3.728867531	0.03214541	31.10864064
मुझे भगवद् गीता के बारे में संक्षिप्त जानकारी दे।	"उपयोगकर्ता मुझे भगवद् गीता के बारे में संक्षिप्त जानकारी दे। सहायक गीता हिंदू धर्मग्रंथों का एक संग्रह है जो सत्त्व, भगवान कृष्ण को अर्जुन और उनकी दृष्टि से कर्त्तव्यों के बीच संघर्ष का वर्णन करने के लिए एक उद्धृत किया जाता है। गीता की शक्ति अर्जुन के स्वतंत्रता की इच्छाओं और उनके नैतिकता का वर्णन करती है। यह हिंदू धर्मग्रंथों का एक महत्वपूर्ण पाठ है, और इसे आध्यात्मिक विकास और मुक्ति प्राप्त करने के मार्ग के रूप में देखा जाता है।"	78	2.225232363	0.02852862	35.05251915
ईसा मसीह कौन थे? उन्होंने क्या किया?	"उपयोगकर्ता ईसा मसीह कौन थे? उन्होंने क्या किया? सहायक ईसा मसीह एक ब्रह्म थे जिन्होंने धर्म के पुनरुत्थान के बाद सदुक्तों को भगवान से अलग कर दिया। उनका जन्म फारस के सूबाह में हुआ था और उन्होंने धर्म, मसीहा का अवतार माना जाता था, जिसे ईसाई धर्म के इतिहास में सबसे महत्वपूर्ण पाठों में से एक माना जाता है। यह धर्मग्रंथों के शक्तिशाली और उन्होंने दुनिया में शक्तिशाली धर्म शुरू करा था। बाइबिल के अनुसार, वह १२ साल तक जेल में रहे और फिर उन्नीस साल तक जेल में रहे। हालांकि, उनकी प्रसिद्धि सत्त्वता सीमा तक थी। माना जाता है कि लगभग १५,००० वर्षों तक जीवित रहे और उन्होंने दुनिया के कई हिस्सों को प्रभावित किया, जिन्हें भारत, चीन।"	119	3.719863415	0.031259356	31.99042189
बाइबिल की शक्ति क्या है?	"उपयोगकर्ता बाइबिल की शक्ति क्या है? सहायक निश्चित रूप से, यह बाइबिल की शक्ति है: १. अक्षरमय और साक्षात्कार: पुस्तक और मारुत जैसे लेखक अपने लेखन में धर्म की संस्थापना का उल्लेख करते हैं, जैसे कि "हमें बहुरीता का सिद्धांत माना गया था जो सबसे दूर भागवान ने दुनिया को बनाया" और "उसने हमें अपने पापों से मुक्त कर दिया"। "इसका की कहानी हमारे विश्वास का एक प्रमुख पद है, क्योंकि उन्होंने भगवान के जीवन को अपने पूरे जीवन में देखा और उनका पालन किया। २. धर्म, मसीह के आध्यात्मिक विकास की कहानी है: इजरायल के कानून निर्माता सोलोमन के"	117	3.718673944	0.031783538	31.46282835
अक्षर कौन था?	"उपयोगकर्ता अक्षर कौन था? सहायक अक्षर एक महात्मा सम्राट थे जिन्होंने १६वीं शताब्दी में भारत पर शासन किया था और उनका शासन गौरवशाली राजवंश के संस्थापक और पहले सम्राट के रूप में कायम करता था। उन्हें "भारत के चंद्र" और "दुनिया के सबसे शक्तिशाली शासकों में से एक के रूप में जाना जाता है।" उनकी शासन शैली को कभी-कभी "इस्लामी और भारतीय संस्कृति का संतुलन" कहा जाता था, और वे उस समय के आम लोगों से अलग थे। उनके समय के दौरान, भारत में राजनीतिक अस्थिरता थी और विचारधाराएं उभरी थीं, जिसके कारण व्यापार और वाणिज्य में गिरावट आई। कुल मिलाकर, उन्होंने एक सम्पूर्ण इतिहास और सांस्कृतिक विरासत छोड़ी, जिसे भारतीय इतिहास में"	115	3.699928045	0.032173287	31.08168553

Observations:



Conclusion

The Indus Language Model has undergone extensive benchmarking on the Intel platform, demonstrating robust performance across several critical areas. These include TTFT inter-token latency, input prompt length, output prompt length, and overall throughput (measured in tokens per second) at varying concurrent request levels.

During the LLMPerf load testing, the TTFT was found to be 8.03s for 2 NUMA nodes and 100 concurrent requests, and 17.18s for 90 percentile requests for 200 concurrent requests. Interestingly, this performance appears to be on par with GPU inference. The test also reveals that performance scales proportionately with the number of NUMA nodes.

Our findings highlight the impressive responsiveness of the Indus LLM for real-time natural language processing applications. The model reliably delivers low TTFT on the Intel platform with negligible inter-token latency. Its versatility is evident in its ability to handle a variety of linguistic tasks, effectively maintaining processing capabilities across a range of input and output prompt lengths.

Furthermore, the throughput analysis indicates excellent scalability and efficiency of the Indus LLM. This is evident as it continues to maintain high output rates even when the number of concurrent requests increases. Therefore, the Indus LLM emerges as a robust, versatile, and efficient model for various linguistic tasks.



Performance varies by use, configuration and other factors. Learn more at <https://intel.com/benchmarks>
Performance results are based on testing as of dates shown in configurations and may not reflect all publicly available updates. See backup for configuration details.
No product or component can be absolutely secure.
Your costs and results may vary.
Intel does not control or audit third-party data. You should consult other sources to evaluate accuracy.
Intel technologies may require enabled hardware, software or service activation.

© Intel Corporation. Intel, the Intel logo, and other Intel marks are trademarks of Intel Corporation or its subsidiaries. Other names and brands may be claimed as the property of others.