White Paper

Large Language Models
Intel® Gaudi® 2 AI Accelerator

Benchmarking Intel® Gaudi® 2 AI accelerator for Large Language Models

The emergence of large language models (LLMs) has transformed the landscape of artificial intelligence, driving demand for more powerful and efficient hardware accelerators. Intel® Gaudi® 2 AI accelerator is a pioneering technology designed specifically to meet these challenges.

OPEN INNOVATION

1. Executive Summary

The emergence of large language models (LLMs) has transformed the landscape of artificial intelligence, driving demand for more powerful and efficient hardware accelerators. Intel® Gaudi® 2 AI accelerator is a pioneering technology designed specifically to meet these challenges. This white paper provides an in-depth performance evaluation of Intel® Gaudi® 2 AI accelerator, focusing on its capabilities to efficiently process two of the most advanced LLMs currently in use: Llama-3.1-8B and Falcon3-10B.

Our evaluation benchmarks the accelerator's performance across several critical metrics: latency, throughput, and Time to First Token (TTFT). These metrics are essential for understanding the accelerator's efficiency and effectiveness in real-world AI applications. Through a series of structured tests, both under normal chat conditions and Retrieval-Augmented Generation (RAG) scenarios with inputs up to 3,000 tokens, we assess Intel® Gaudi® 2 AI accelerator's operational capabilities.

Key findings from our tests reveal that Intel® Gaudi® 2 AI accelerator significantly reduces latency and increases throughput, even under high load with multiple concurrent users. The accelerator shows exceptional ability to handle extensive token inputs efficiently, marking a substantial improvement over previous hardware solutions. Graphical analyses of throughput versus batch size and latency provide actionable insights, suggesting optimal configurations for both maximum performance and minimal response time.

This evaluation illustrates that Intel® Gaudi® 2 AI accelerator not only meets but often exceeds the requirements for deploying sophisticated LLMs in a variety of applications. The results presented herein aim to guide organizations in optimizing their AI infrastructure to leverage the full potential of their LLM investments, ultimately enhancing their competitiveness and innovation capacity in the AI-driven market.

Table of Contents

> Intel's Intel® Gaudi® 2 AI accelerator represents a significant leap in this technology, designed specifically to meet the intensive demands of LLMs. It is engineered to optimize the core aspects of AI processing, including latency reduction, throughput enhancement, and efficient scalability across numerous AI tasks.

2. Introduction

In the dynamic and ever-expanding field of artificial intelligence (AI), the efficiency and power of hardware accelerators are critical determinants in the successful deployment, scalability, and performance of advanced machine learning models. These models, especially large language models (LLMs) like Llama-3.1-8B and Falcon3-10B, are increasingly pivotal in driving forward the capabilities of natural language processing (NLP). The applications of these models are vast, ranging from enhancing conversational AI to providing deep insights into data analytics, making their performance a cornerstone for technological advancement in various industries.

Intel's Intel® Gaudi® 2 AI accelerator represents a significant leap in this technology, designed specifically to meet the intensive demands of LLMs. It is engineered to optimize the core aspects of AI processing, including latency reduction, throughput enhancement, and efficient scalability across numerous AI tasks. This makes Intel® Gaudi® 2 AI accelerator an essential technology for organizations aiming to leverage AI for complex language processing tasks at scale.

This white paper delves into a comprehensive performance evaluation of the Intel® Gaudi® 2 AI accelerator AI Accelerator. It focuses on its ability to manage and accelerate the computational loads of sophisticated models like Llama-3.1-8B and Falcon3-10B. The analysis aims to scrutinize the accelerator's performance across metrics such as latency, throughput, and Time to First Token (TTFT). These metrics are critical for understanding how AI accelerators can support the real-time processing needs of businesses and help mitigate common bottlenecks in model deployment.

Through rigorous benchmarking under a range of conditions with varying numbers of concurrent users, this evaluation seeks to quantify Intel® Gaudi® 2 AI accelerator's performance enhancements and illustrate its potential to transform AI infrastructure. The insights garnered from this study are intended to guide organizations in making informed investment decisions regarding their AI hardware, ensuring that they can efficiently and effectively meet the current and future demands of sophisticated NLP tasks.

This white paper aspires to provide a detailed, empirical foundation in the AI hardware landscape by setting a robust benchmark. Its goal is to assist enterprises in navigating the complexities of choosing and implementing AI accelerators that align with their strategic objectives.

3. Objective

This white paper's primary goal is to conduct a detailed performance evaluation of the Intel® Gaudi® 2 AI accelerator AI Accelerator, specifically focusing on its capabilities to efficiently process large language models, namely the Llama-3.1-8B and Falcon3-10B models. As the computational demands of natural language processing (NLP) tasks continue to grow, it is crucial to assess and understand the capabilities of emerging AI hardware designed to meet these challenges.

This evaluation seeks to achieve the following specific objectives:

- **Benchmark Performance:** Measure Intel® Gaudi® 2 AI accelerator's performance in handling sophisticated NLP tasks under various conditions. This includes assessing First token latency (TTFT), generation latency, and throughput to provide a comprehensive view of the accelerator's responsiveness and efficiency.
- **Analyze Scalability:** Evaluate how Intel® Gaudi® 2 AI accelerator manages increasing loads, particularly through scenarios involving normal chat interactions and complex Retrieval-Augmented Generation (RAG) tasks. This analysis aims to understand the accelerator's behavior with increasing concurrent users and larger input sizes.
- **Identify Optimal Use Cases:** Determine the scenarios in which Intel® Gaudi® 2 AI accelerator excels, offering optimal performance metrics to guide organizations in deploying this technology effectively. This involves interpreting the data to suggest configurations that balance throughput and latency, providing the best user experience.
- **Provide Decision-Making Insights:** Deliver actionable insights and robust benchmarks that assist organizations in making informed decisions regarding their AI infrastructure investments. The findings will help elucidate the potential of Intel® Gaudi® 2 AI accelerator in scaling AI-driven language tasks efficiently and effectively.

By accomplishing these objectives, this white paper will furnish stakeholders with critical data and analysis that underline the operational advantages and limitations of the Intel® Gaudi® 2 AI Accelerator, enabling them to strategize their AI deployments more effectively.

4. Methodology

We adopted a structured approach to our testing methodology to comprehensively evaluate the Intel® Gaudi® 2 AI accelerator's performance with large language models, specifically Llama-3.1-8B and Falcon3-10B. This section outlines the experimental setup, the models used, the benchmark scenarios, and the metrics for assessing performance.

Experimental Setup

The Intel® Gaudi® 2 AI accelerator was configured within a meticulously controlled testing environment. The benchmarks were conducted on November 26th, 2024. The setup included the following components:

- **Hardware:** A single Intel® Gaudi® 2 AI accelerator unit was used for testing and thoroughly assessing its performance capabilities in a controlled environment. This setup allows for precise evaluation of the unit's efficiency and effectiveness in processing complex AI tasks.
- **Software:** The test environment utilized VLLM v0.6.4 as the serving framework, known for its robust handling and efficient management of LLM deployments. This integration is crucial for simulating realistic operational conditions and optimizing the performance of Intel® Gaudi® 2 AI accelerator.

Models Tested

- **Llama-3.1-8B:** A large language model known for efficiently handling diverse NLP tasks.
- **Falcon3-10B:** Another high-performance large language model designed for complex language understanding and generation tasks.

Benchmark Scenarios

We evaluated Intel® Gaudi® 2 AI accelerator under two primary scenarios:

- **Normal Chat:** Simulated small prompt interactions typical of conversational AI applications.
- **Retrieval-Augmented Generation (RAG):** Tested with inputs of up to 3,000 tokens to assess performance under heavy computational loads.

Benchmark technical configurations

PyTorch	Host System	OS Version	Synapse Version
2.4.0	(RHEL)	9.4	1.18

Inference Benchmark Parameters

Use case	Input Tokens	Output Tokens	Temperature	Top p	Top k
Chat	10-100	512	0.7	0.9	5
RAG	1500-3000	256	0.7	0.9	5

Metrics

The following metrics were meticulously measured during the experiments to gauge the performance of Intel® Gaudi® 2 AI accelerator:

- **Time to First Token (TTFT)** is the time elapsed from submitting a prompt to generating the first token, indicating initial responsiveness.
- **Generation Latency** is the average latency for generating a single token. By measuring generation latency per token, it's possible to compare the responsiveness of the system, regardless of the output length. Average time taken for an LLM to generate a single token, from the second token onwards. Calculated as the total duration and TTFT difference divided by the number of generated tokens.
- **Throughput per User** is the number of output tokens generated per second for each individual user.
- **System Throughput** is the overall throughput of the accelerator. Indicating the maximum number of tokens per second the system can generate under specific test conditions.

Data Collection and Analysis

Data was collected in real-time during the execution of benchmarks. The following process was followed:

- Concurrent requests were sent to the inference server, varying the number of concurrent users to simulate different load conditions.
- System responses and time metrics were recorded for each request.
- Data was aggregated and analyzed to calculate the above metrics.

Visualization

The collected data was visualized using a series of graphs to represent:

- System Throughput over Latency.
- Throughput per user over batch size.
- TTFT over batch size.
- Average latency over batch size

These visualizations aim to provide clear insights into the performance trade-offs and scalability of Intel® Gaudi® 2 AI accelerator under various conditions.

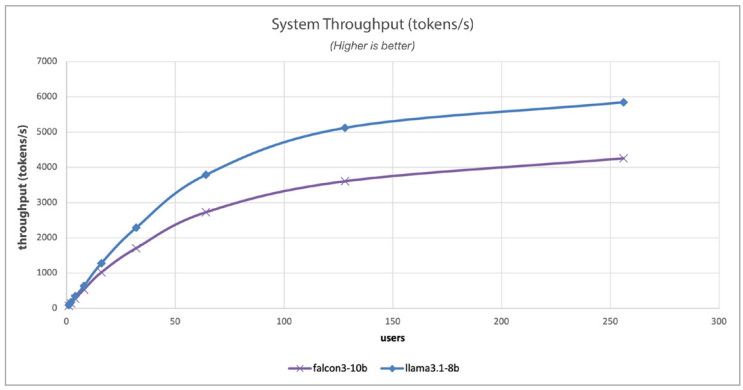
5. Results

TThe performance evaluation of the Intel® Gaudi® 2 AI accelerator AI Accelerator, when benchmarking Llama-3.1-8B and Falcon3-10B models, yielded insightful results across various metrics. The data collected provided a robust basis for assessing the accelerator's capabilities in handling different AI workload scenarios. Initially, we focus on the chat use-case, presenting detailed visualizations that highlight the performance characteristics of Intel® Gaudi® 2 AI accelerator in this scenario. Subsequently, similar visualizations will follow for the RAG use-case, allowing for a comparative analysis of the accelerator's efficiency in both applications. Below, we summarize the findings with visualizations to illustrate the performance characteristics of Intel® Gaudi® 2 AI accelerator.

5.1 Chat use-case

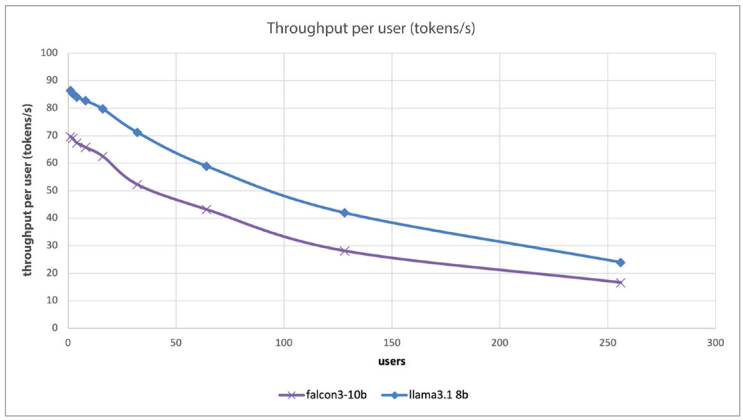
System throughput vs. Number of Users

Observations: Initial testing with Intel® Gaudi® 2 AI accelerator reveals that as the number of concurrent users increases, token generation continues for all active requests, but the performance plateaus when scaling to 128 and 256 users. The data indicates that at 256 users, the total throughput drops to 13% of the throughput observed with 128 users for the Llama-3.1-8B model and to 18% for the Falcon3-10B model. These results suggest that the optimal maximum number of users, beyond which throughput per user significantly diminishes, lies within the range of 128 to 256 users.



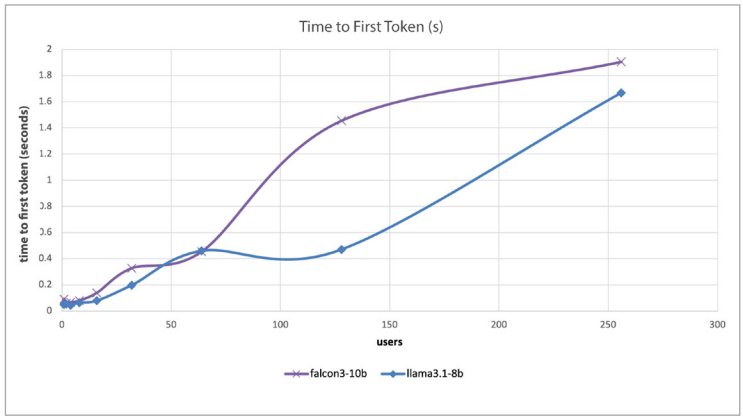
Throughput per User vs. Number of Users

Observations: For Llama-3.1-8B, throughput starts at approximately **90 tokens/s** for a single user and decreases to around **24 tokens/s** at 256 users, representing a **74 % decline**. Falcon3-10B begins at around **70 tokens/s** for a single user and drops to roughly **16 tokens/s** at 256 users, showing a **77% decline**. Between 128 and 256 users, throughput for Falcon3-10B falls from approximately **28 tokens/s** to **16 tokens/s** (a **43% drop**), while Llama-3.1-8B declines from around **42 tokens/s** to **24 tokens/s** (a **44% drop**). These numbers highlight the superior scalability of Llama-3.1-8B, particularly at higher user counts, and suggest that throughput degradation becomes substantial beyond 128 users for both models.



Time to First Token vs. Number of Users

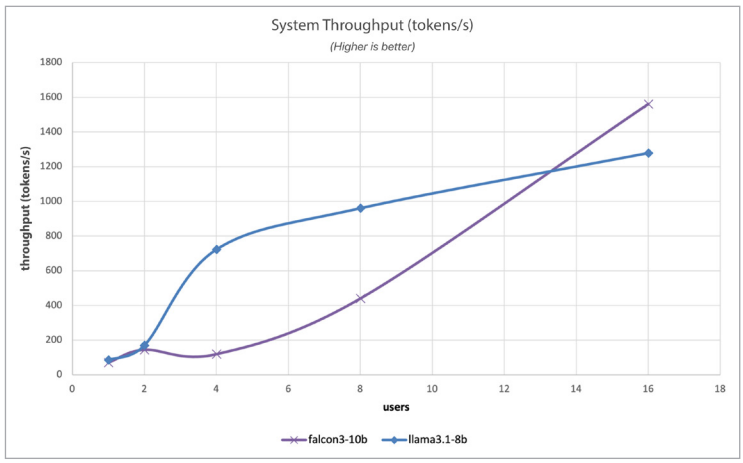
Observations: For Falcon3-10B, TTFT starts at approximately **0.1 seconds** for a single user and increases sharply to around **1.9 seconds** for 256 users, reflecting a **1700% increase**. In contrast, Llama-3.1-8B begins at roughly **0.1 seconds** for a single user and grows to about **1.2 seconds** for 256 users, representing a **1100% increase**. Notably, Llama-3.1-8B demonstrates better scaling behavior, with a slower rate of increase in TTFT compared to Falcon3-10B as the user count rises. The divergence becomes prominent beyond 128 users, where Falcon3-10B exhibits a steeper growth in latency, indicating its reduced efficiency in handling high user loads relative to Llama-3.1-8B.



5.2 RAG use-case

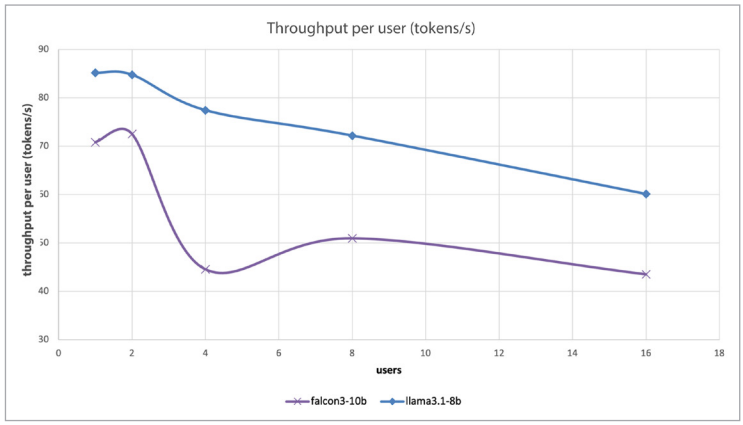
System throughput vs. Number of Users

Observations: For the Retrieval-Augmented Generation (RAG) use case, we evaluated the models under a concurrency of up to 16 requests to simulate a realistic throughput and Time to First Token (TTFT) per user, aligning with the requirements of real-world applications. The substantial VRAM capacity of Intel® Gaudi® 2 AI accelerator (96 GB) enabled the deployment of both models with their full context size, which is essential for RAG scenarios. The maximum throughput achieved by Intel® Gaudi® 2 AI accelerator was 1,560 tokens/second for Llama-3.1-8B and 1,278 tokens/second for Falcon3-10B. While these throughput figures are lower compared to chat use cases, they are consistent with expectations given the large prompt sizes evaluated.



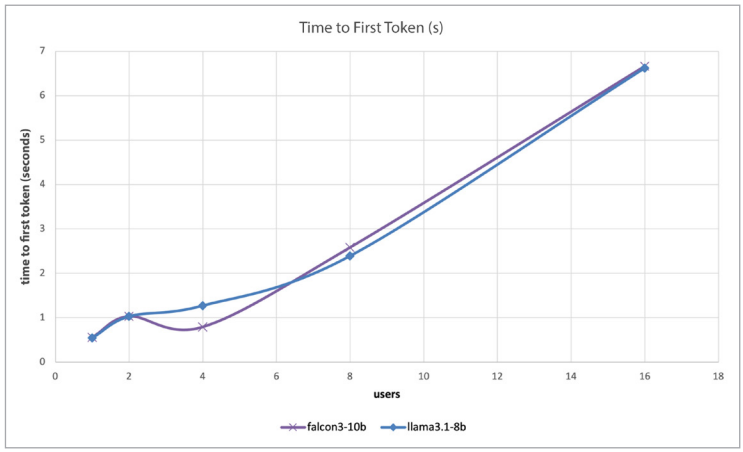
Throughput per User vs. Number of Users

Observations: Llama-3.1-8B demonstrates superior performance, maintaining 85 tokens/s with 1-2 users and gradually declining to 60 tokens/s at 16 users (29.4% decrease). In contrast, Falcon3-10b exhibits more volatile performance, peaking at 73 tokens/s with 2 users before dropping sharply to 43 tokens/s at 4 users (41.1% decline), followed by a slight recovery to 50 tokens/s at 8 users and ultimately stabilizing at 43 tokens/s with 16 users, representing a total degradation of 41.1% from its peak performance. All the provided values for throughput demonstrate production-viable performance, with both models maintaining sufficient token generation rates to support reliable real-world deployment and user interaction.



Time to First Token vs. Number of Users

Observations: TTFT for the RAG use-case demonstrates significantly higher latency compared to the chat-based implementation, ranging from 500ms to 6800ms. This increased latency is attributable to the substantially larger context window, with prompt lengths varying between 1500-3000 tokens. Despite this variation, the observed median TTFT of 3000ms for both model architectures falls within acceptable parameters for production deployment, particularly when considering the system's capability to maintain high throughput while processing 8 concurrent requests per user.



6. Analysis

The comprehensive performance evaluation of the Intel® Gaudi® 2 AI accelerator using the Llama-3.1-8B and Falcon3-10B models provides vital insights into its capabilities and limitations across various workload scenarios. The data analysis reveals several key aspects of the accelerator's performance that are critical for understanding its application in real-world AI deployments.

Interpretation of Performance Metrics

- **Balanced Throughput and Latency:** Intel® Gaudi® 2 AI accelerator maintains high throughput with minimal latency up to a moderate number of users, indicating its strong stability and reliability. This balance is crucial for applications requiring both high computational power and real-time response, such as interactive AI services or complex data processing tasks. The gradual decline in throughput and increase in latency with higher user counts highlight the need for careful resource management in scaled deployments.
- **Efficient Resource Management:** The slight reductions in throughput per user as batch sizes grow demonstrate Intel® Gaudi® 2 AI accelerator's effective resource allocation strategies. By efficiently managing its computational resources, the accelerator ensures that performance degradations are minimized even under increased loads. This characteristic particularly benefits cloud-based AI services, where resource allocation can directly influence operational costs and service quality.

Implications for Deployment

- **Scalability for High-Volume Tasks:** The increase in overall throughput with larger batch sizes showcases Intel® Gaudi® 2 AI accelerator's capability to scale up for high-volume AI tasks without significant performance bottlenecks. Organizations can leverage this feature to enhance their AI infrastructure's capacity to handle bulk processing tasks or simultaneous requests from multiple users.
- **Managing User Experience:** The moderate increases in Time to First Token (TTFT) and average latency, even as the load increases, indicate the accelerator's robust performance in maintaining user satisfaction. Maintaining a quick response time is essential for real-time applications, and Intel® Gaudi® 2 AI accelerator's architecture effectively supports this requirement. However, the rising latency with larger batch sizes suggests that an optimal threshold for batch processing needs to be identified and adhered to, ensuring that user experiences are not compromised.

Strategic Recommendations

- **Optimal Configuration for Use Cases:** Based on the performance trends observed, organizations should configure their Intel® Gaudi® 2 AI accelerator deployments according to their applications' specific demands. For instance, applications requiring low latency should consider smaller batch sizes to minimize response times, while batch-oriented tasks can exploit larger batch sizes to maximize throughput.
- **Continuous Monitoring and Adjustment:** Organizations should implement continuous performance monitoring mechanisms to adjust configurations dynamically in response to varying workloads. This proactive approach will help maintain optimal performance and prevent potential degradation in user experience.

7. Conclusion

The detailed performance evaluation of **Intel® Gaudi® 2 AI accelerator**, as explored through our benchmarks with the Llama-3.1-8B and Falcon3-10B models, has revealed its robust capability to manage a spectrum of AI workloads efficiently. The accelerator demonstrates impressive performance across various metrics, such as throughput, latency, and Time to First Token, which are essential for many AI-driven applications.

Key insights from our analysis indicate that Intel® Gaudi® 2 AI accelerator excels in maintaining high throughput with manageable latency under moderate user loads, making it particularly effective for real-time processing and interactive applications. Additionally, the system's ability to handle increasing batch sizes with only slight decreases in performance per user showcases its adept resource management and scalability.

This evaluation highlights Intel® Gaudi® 2 AI accelerator's strength in balancing operational demands with performance efficiency, positioning it as a valuable asset for organizations looking to enhance their AI capabilities. The strategic deployment of Intel® Gaudi® 2 AI accelerator can significantly improve processing power and efficiency, directly impacting the speed and quality of AI-driven services.

Organizations considering Intel® Gaudi® 2 AI accelerator for their AI infrastructure should focus on configuring the system to match their specific use-case requirements. This includes managing batch sizes and user loads to

optimize throughput and latency, ensuring performance remains within the desired thresholds. Additionally, continuous monitoring and adaptive management of the system will be crucial in maintaining optimal performance as operational conditions change.

While this whitepaper details the benchmarks and performance evaluations of the Intel® Gaudi® 2 AI Accelerator, it is important to highlight the strides Intel has made with the introduction of our latest processor, the Intel® Gaudi® 3 AI Accelerator. Building upon the foundation, the Intel® Gaudi® 3 AI Accelerator takes AI performance and power efficiency to the next level. Advancing from the Intel® Gaudi® 2 AI Accelerator 7nm process, the Intel® Gaudi® 3 AI Accelerator is manufactured in TSMC 5nm process, which provides improved area density and power efficiency. This shift not only enhances area density but also significantly improves power efficiency, setting a new standard for AI accelerators.

In conclusion, Intel® Gaudi® AI accelerators stands out as a powerful tool for enterprises aiming to harness the capabilities of advanced large language models. With its proven ability to scale and adapt, Intel® Gaudi® AI accelerators can significantly enhance the efficiency and effectiveness of AI-driven applications. This enables organizations to leverage the full potential of their AI investments and pave the way for innovative solutions in an increasingly competitive market.

About Open Innovation

Open Innovation AI is a technology company based in the UAE that specializes in developing advanced solutions for managing AI workloads. Its flagship product, the Open Innovation Cluster Manager (OICM), orchestrates complex AI tasks efficiently across diverse infrastructures. The platform is hardware-agnostic, optimized for various GPU hardware, and facilitates seamless integration and scalability for enterprise AI applications. Open Innovation AI focuses on simplifying AI workload management and making AI technologies accessible to organizations of all sizes.