LLMWare's Model HQ Successfully Optimized on Intel Arrow Lake Architecture

Introduction to Model HQ: LLMWare's Enterprise Platform

Model HQ, LLMWare's Enterprise Platform is a cutting-edge artificial intelligence solution designed to meet the evolving needs of businesses across various industries. By harnessing the power of AI, LLMWare offers unparalleled efficiency, accuracy, and scalability, enabling organizations to drive innovation and achieve competitive advantages. Key features of LLMWare include a robust Model Catalog, Pre-built specialized LLMs, and End-to-End Solutions, each tailored to enhance operational workflows, data analysis, and decision-making processes.

Model HQ is a new streamlined software package that provides a fast and easy on-ramp to working with AI models locally while providing comprehensive enterprise-ready security, safety and compliance-tracking capabilities for enterprise adoption.

Using the latest optimization techniques harnessing the power of OpenVINO, Model HQ is specifically designed to maximize model performance for running, creating and deploying AI-enabled apps with integrated UI/UX and low-code agent workflow for easy app creation. The Model HQ app comes ready to use, and provides the ability to launch custom workflows directly on user device.

The Starter version of Model HQ introduces a suite of point-and-click solutions, including a built-in Chatbot, Document Search and Analysis, Text to SQL Query, and Speech-to-Text functionality. The Developer and Enterprise versions include LLMWare's specialized function calling SLIM models designed for agentic multi-step workflows for lightweight or "micro" app creation for easy workflow automation.

The entire suite of LLMWare solutions provides all the turn-key software infrastructure with built-in security needed by enterprise developers and GSI's to move quickly from POC to scale solutions. LLMWare has also curated 68 quantized small language models in its Model Depot in Hugging Face for use with the OpenVINO acceleration toolkit. Model Depot also includes 28 of LLMWare models with sophisticated function calling ability for developers to mix and match as they automate workflows for "micro apps" for endless vertical use cases.

Optimization on Intel Arrow Lake Architecture

To maximize performance and efficiency, LLMWare has been optimized for Intel's Arrow Lake Architecture, harnessing the advanced capabilities of OpenVINO, enabling LLMWare to fully utilize Intel's powerful CPUs, GPUs, and NPUs, delivering faster processing, lower latency, and higher throughput for AI workloads.

This optimization process applied targeted techniques to critical components of LLMWare, ensuring versatile and efficient operation across diverse deployment scenarios. Running on the Intel Arrow

Lake platform, known for its performance and reliability, LLMWare also benefits from an extensive Intel ecosystem of tools and libraries that facilitate streamlined AI development and deployment.

About LLMWare

LLMWare is a leading provider of AI solutions, committed to delivering innovative and effective solutions that empower businesses to stay ahead in the digital era. With a focus on research and development, LLMWare continues to push the boundaries of what's possible with artificial intelligence.

LLMWare is a pioneer in deploying and fine-tuning Small Language Models for use in highly-regulated or data sensitive industries. With 150+ models in Hugging Face, LLMWare is a leader in developing cutting-edge AI app creation and deployment solutions. For additional information, including product, blogs and latest research reports, please visit llmware.ai.

Stay connected with us on social media for the latest updates and insights into our solutions and services.

You can follow us on LinkedIn, YouTube, GitHub and Hugging Face.