

LLMWare Unleashes the Power of the Intel AI PC with Cost, Performance and Security Wins

Model HQ by LLMWare uses Intel® architecture to streamline distributed deployment of AI workflows and drive up their performance. Intelligently placing tasks across the processor's range of AI execution engines meets business challenges head-on with fast response, high throughput, and power-efficient performance.

About LLMWare.ai

LLMWare provides an integrated development framework of models that accelerate development of language-centric applications that connect knowledge to AI. Built specifically for local inference on Intel AI PCs using the OpenVINO toolkit, Model HQ by LLMWare.ai gives organizations a simple and practical approach to capturing benefits from the latest AI technology. LLMWare is an operating entity of parent company Ai Bloks LLC.

Generative AI (Gen AI) is decisively reinventing human-machine interaction, giving computers the growing ability to directly understand and create verbal and visual expression. Large language models (LLMs) to enable this interaction have grown exponentially in size. With trillions of parameters, LLMs require substantial computing resources, typically provided by centralized inference on high-end enterprise GPU servers in the public cloud. That topology creates gating factors that have limited AI deployments:

- **High cost.** Resource-intensive cloud inference can drive up operating costs for AI workflows.
- **Untenable latency.** Round trips to the cloud can make real-time and near-real-time usages impossible.
- **Security exposure.** Traversing the open internet introduces cyber risk for potentially sensitive data.
- **Connectivity requirements.** The need for constant cloud connection interferes with offline operation.

Nimbleness for efficient, instantaneous, protected AI, even offline

Enterprises are evolving past the limitations of cloud-centric AI workflows with a combination of hardware and software innovation. A critical hardware development is the Intel AI PC, which is engineered to accommodate workflow automation applications and to run AI workflows locally on the client instead of sending inference tasks back to the cloud or data center. This approach avoids the costs of bandwidth and subscription fees for GPU cloud instances, as well as the associated latency and cyber risk. It also makes it possible for AI workloads to be completed offline, for greater flexibility, safety and security.

Meanwhile, dramatically smaller, specialized models including small language models (SLMs) are unlocking novel AI value within the computational footprint of the AI PC. Model HQ, a toolkit provided by LLMWare, gives enterprises full control over creating and deploying AI workflows on distributed fleets of AI PCs, providing optimized performance while streamlining deployment, operation and management. The software platform is built specifically for the Intel AI PC, optimizing the technology transition with no-code ease of use for business users, automation and integrated monitoring and management tools.

BACKGROUND

LLMWare
AI framework and LLM tools for local enterprise app dev and deployment with built-in security for organizations in regulated and privacy-focused industries.

Pioneer in Small Language Models
Selected and funded by GitHub as a Top AI Project

100,000s of Users of Our Models
Popular Library in Github and Hugging Face





Model Depot Repository with 150+ models to choose from, including 50+ OpenVINO optimized models



Model HQ Policy controls for full inference life cycle in data centers, private cloud containers and across client agent executables



Client Agent Executable UI/framework ready to run on AI PC Chat, RAG and Agent Workflows

Document Info Retrieval

Contract Analysis

Personal Chatbot

Automated Reports

SQL Queries



Driving productivity and performance with Model HQ


Model HQ provides comprehensive deployment, operation and management over the entire lifecycle of AI applications and workflows on Intel AI PCs. It is a streamlined, compact, integrated solution that delivers comprehensive functionality for the AI stack. The all-in-one toolkit focuses on the use of fine-tuned, small, specialized models to drive performance and value from advanced AI, without stand-alone GPUs or GPU clusters. It provides robust automation and control, with low- or no-code development, for workflows deployed across Intel AI PCs or in private or hybrid cloud scenarios.

Pipelines built and operated using Model HQ can utilize more than 150 specialized models from LLMWare’s Model Depot, to address a wide range of complex enterprise use cases. While lightweight, the models are optimized for performance and efficiency to support advanced workflows with reduced compute footprint and power consumption. They also automatically take advantage of the optimal combination of AI PC processing engine and AI framework. The Model Depot provides the following main types of models:

- **State of the art SLMs specifically optimized for use in Intel AI PCs**, including models that are fine-tuned for efficient retrieval-augmented generation (RAG) workflows, to answer fact-based inquiries and to reduce hallucinations.
- **Domain-specific embedding models** provide expertise in specific industry contexts such as finance or insurance.

- **Structured Language Instruction Models (SLIMs)** are specialized compact function-calling models that can be mixed and matched and used in combination for agentic workflows and for specific tasks such as topic classification, tagging, reading SQL tables and analyzing sentiment.

Unlike typical LLMs that are designed to answer general-knowledge inquiries using open context, LLMWare’s RAG-optimized and embedding models are fine-tuned for specific complex business, legal and financial domains. Information that models use to generate answers can be strictly limited to specified source material if needed, with no background knowledge or speculation. This approach is critical for contextually complex usages, such as handling contracts and other legal documents. To further protect the integrity and audit-readiness of RAG workflows, Model HQ provides a comprehensive set of data lineage mechanisms:

 **Data-flow tracking and governance.** Executing LLMWare-enabled workflows locally on the Intel AI PC enables comprehensive data-flow visibility from input to output, including improving data-flow visibility by eliminating reliance on external servers when the data is on-device.

 **Cybersecurity and compliance.** Removing the need to transmit data to the cloud or other remote location reduces cyber exposure and helps ensure compliance with regulations and best practices that mandate stringent control over data access and movement.

Integrated tools for data lineage. Tools built into Model HQ by LLMWare.ai for data-lineage tracking capture detailed logs for analysis of how data is accessed, processed and modified within AI workflows, providing a clear record of data transformations and model inferences.

Real-time monitoring and auditing. The platform continuously monitors AI processes, detecting and logging data-lineage events. This capability feeds transparency and audit best practices with an accurate, up-to-date record of data flow.

AI explainability and traceability. LLMWare gives customers insights into the decision-making processes and data usage within AI workflows. Unlike “black box” AI implementations, Model HQ provides comprehensive documentation for all stages of the data lifecycle, enhancing traceability.

Evolving inference with the Intel AI PC

As AI workflows become more central to mainstream business computing, cloud-based pricing models are proving unsuited to the scalability requirements of everyday inference. Intel AI PCs democratize AI personal productivity by running inference directly on the client, driving the cost toward zero.

Powered by Intel Core Ultra processors, they usher in a new paradigm that matches AI workloads to specific execution engines to tailor latency, throughput and power efficiency outcomes. CPU resources provide low-latency response with a combination of the latest Performance-core and Efficiency-core architectures. The Integrated Intel® Arc™ GPU, based on new Xe2 architecture, drives high throughput for heavy workflows. The enhanced neural processing unit (NPU) 4.0 AI Engine delivers up to 13 TOPS,¹ for power-efficient sustained inference.

The OpenVINO toolkit fast track to AI performance

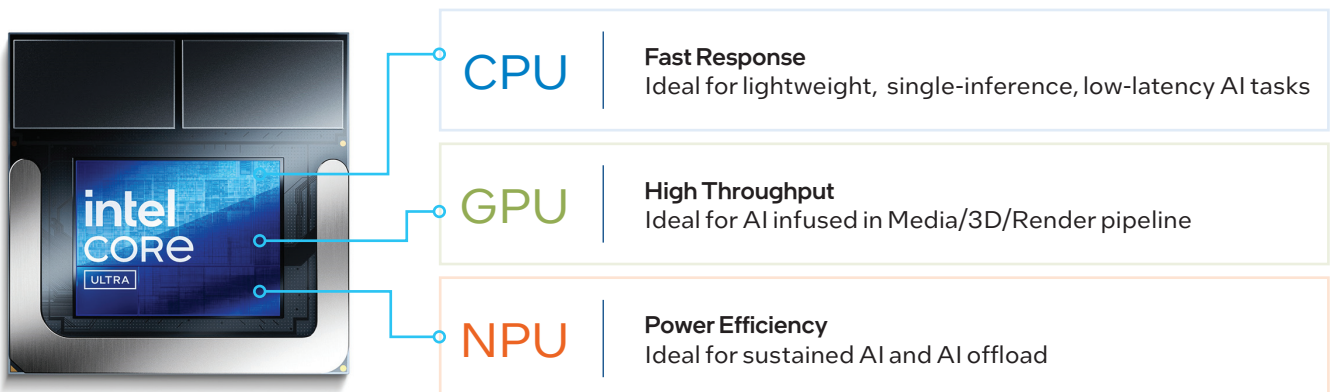
OpenVINO empowers AI developers with the foundations for high-performance inference with a reduced footprint, while retaining accuracy, without any proprietary licenses. It optimizes code for the Intel hardware, including utilization of its different processing engines for the ideal balance of latency, throughput and power efficiency. Using OpenVINO streamlines development and deployment to shrink timelines and maximize productivity, while protecting data and investments. OpenVINO is making it easier on developers to adapt AI solutions with existing hardware solutions while preparing for future requirements and use cases.

“OpenVINO is a fundamental part of the Intel technology stack, and a powerful software enabler to unlock the performance capabilities of Intel hardware. We believe that OpenVINO has a big role to play in the coming AI PC revolution, and especially with the widespread roll-out of Intel® Core™ Ultra 200V series processors and other next generation Intel GPU/NPU/ CPU technologies.”

– Darren Oberst, CTO, LLMWare

Intel® Core™ Ultra Processors: Three AI Engines for the AI PC

The right balance of platform power and performance for building and deploying AI models



LLMWare uses OpenVINO as the foundation for high-performance, high-accuracy inference using Gen AI with SLMs. Model HQ core models are in OpenVINO or intermediate representation (IR) format for ease of loading. Since the toolkit is open source, LLMWare is able to integrate OpenVINO into their solution as a backend, especially for deployment on AI PCs and other platforms based on Intel architecture. OpenVINO provides excellent Intel platform support and exceptional inference speed on both the CPU and integrated GPU.

Building a new open paradigm for Gen AI

The combination of Model HQ, the Intel AI PC and OpenVINO offers a robust new approach to running advanced AI workflows locally on the client. It breaks down previous conceptions of compute requirements to support Gen AI usages that were based on centralized remote processing, in favor of a more lightweight, flexible, future-focused vision that delivers the following benefits:



Gen AI performance without discrete GPUs. The combination of LLMWare and Intel AI PCs enables SLMs with 1-9 billion parameters to run on-device with no discrete GPU, to perform most text-based tasks comparably to ChatGPT 3.5 while delivering acceptable response times.



Flexibility across workflows and usages. Support for a broad range of open source and proprietary models provides open-ended options for creating the right workflow for organization-specific requirements that is lightweight, easy to use, and cost-effective.



Streamlined and automated deployment. The solution provides efficient, comprehensive deployment for AI workflows on user PCs, without complex infrastructure or extensive coding. It takes excellent advantage of the AI PC's multiple inferencing technologies and matches optimized AI frameworks to the hardware.



Simplified, comprehensive management. Intuitive but powerful interfaces, automated updates, and integrated monitoring tools with a built-in audit path reduce technical barriers to entry, enabling IT teams to get up and running quickly and manage AI workflows more effectively.



Protection for sensitive or regulated data. With LLMWare and Intel AI PCs, secure AI solutions can be run within controlled and even air-gapped environments, to leverage regulated or otherwise sensitive data without exposing it to external networks.

Conclusion

Model HQ by LLMWare.ai offers significant new potential to make high-performance inference cost-effective and sustainable for everyday business use on Intel AI PCs. The Model HQ app comes ready to use and provides the ability to launch custom workflows directly on the 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, lightweight or "micro" app creation and easy workflow automation.

As AI becomes more enmeshed in software of every description, local inference enabled by OpenVINO provides the latency, TCO and security advantages to drive its full potential. LLMWare makes a vital contribution to the AI PC paradigm with specialized models and tools that help direct the industry away from monolithic GPU arrays, toward unlimited inference resources they can hold in their hands.

Learn More

To learn more about Model HQ by LLMWare.ai visit:

- [LLMWare.ai](https://llmware.ai)
- [LLMware white paper: Revolutionizing AI Deployment](#)
- [LLMWare YouTube channel](#)
- [LLMWare GitHub](#)
- [LLMWare Hugging Face](#)

To learn more about Intel® technologies visit:

- [Intel AI PC](#)
- [Intel® Core™ Ultra processors family](#)
- [Intel® Distribution of OpenVINO™ toolkit](#)

Solution provided by:



¹ TOPS: All TOPS are "up to" and approximate until final IP frequency defined, different SKUs with different frequency & power targets will have different TOPS.

No product or component can be absolutely secure.

Intel does not control or audit third-party data. You should consult other sources to evaluate accuracy.

Your costs and results may vary.

Intel technologies may require enabled hardware, software or service activation.

You may not use or facilitate the use of this document in connection with any infringement or other legal analysis concerning Intel products described herein. You agree to grant Intel a nonexclusive, royalty-free license to any patent claim thereafter drafted which includes subject matter disclosed herein.

The products described may contain design defects or errors known as errata which may cause the product to deviate from published specifications. Current characterized errata are available on request.

© 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. 0125/DC/MESH/PDF 361151-001US