REVOLUTIONIZING AI DEPLOYMENT

Unleashing AI Acceleration with Intel's AI PCs and Model HQ by LLMWare.ai

llmware

WHAT

AI PCs will decentralize Gen AI at user level

HOW

With Model HQ - AI deployment and access simplified for everyone

12/3/2024

UNLOCKING

AI FOR ALL

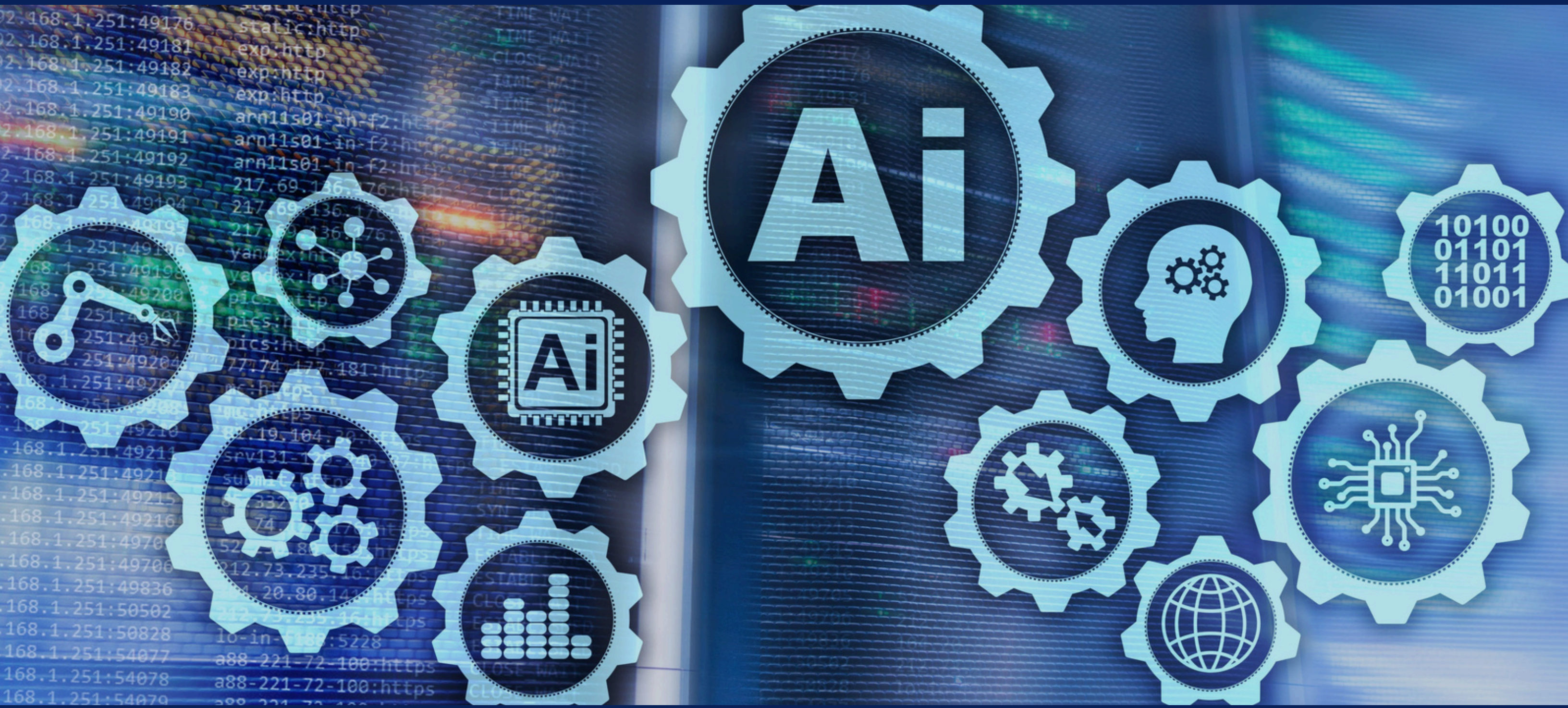
AI-powered productivity gain is one of the biggest and most exciting promises of Generative AI technology. AI PCs, a new class of personal computers that are specifically engineered to handle AI tasks locally, enable a new paradigm of delivering AI capability to users. By empowering users to easily access and implement a diverse array of AI-driven workflows directly from their PCs, AI PCs deliver the promise of AI-powered productivity gain for business users with the assurance of being self-hosted, secure, and private.

In this white paper, we will demonstrate why we believe that AI PCs can meet the challenge of delivering robust AI use cases at the PC level, as well as propose our solution for how enterprises can leverage Model HQ, a full stack end-to-end solution that is specifically designed to simplify AI implementation and operation for both AI developers and users of AI PCs.

REVOLUTIONIZING
AI DEPLOYMENT

Recent advancements in both hardware and AI model capabilities have converged to make AI deployment at the personal computer level not only feasible but also provide many benefits for the enterprise. With the introduction of Intel’s Core Ultra Processors with integrated GPUs and NPUs that are capable of executing AI workflows with rapid inference times, a new era of on-device AI is emerging.

Many workflows that previously required the extensive capabilities of large frontier models, such as those developed by OpenAI, are now fully accessible with AI PCs. This transformation is driven not only by the evolution of hardware but also by significant improvement in AI models, particularly small language models (SLMs).[1] SLMs are smaller language models that are less computationally intensive than traditional LLMs and are lightweight, efficient and generally faster. With SLMs, many personal assistant chatbot use cases such as text summarization, question-answering and text editing are possible in addition to other more specialized use cases such as natural language SQL queries and contract analysis.



ENHANCED
SAFETY & SECURITY

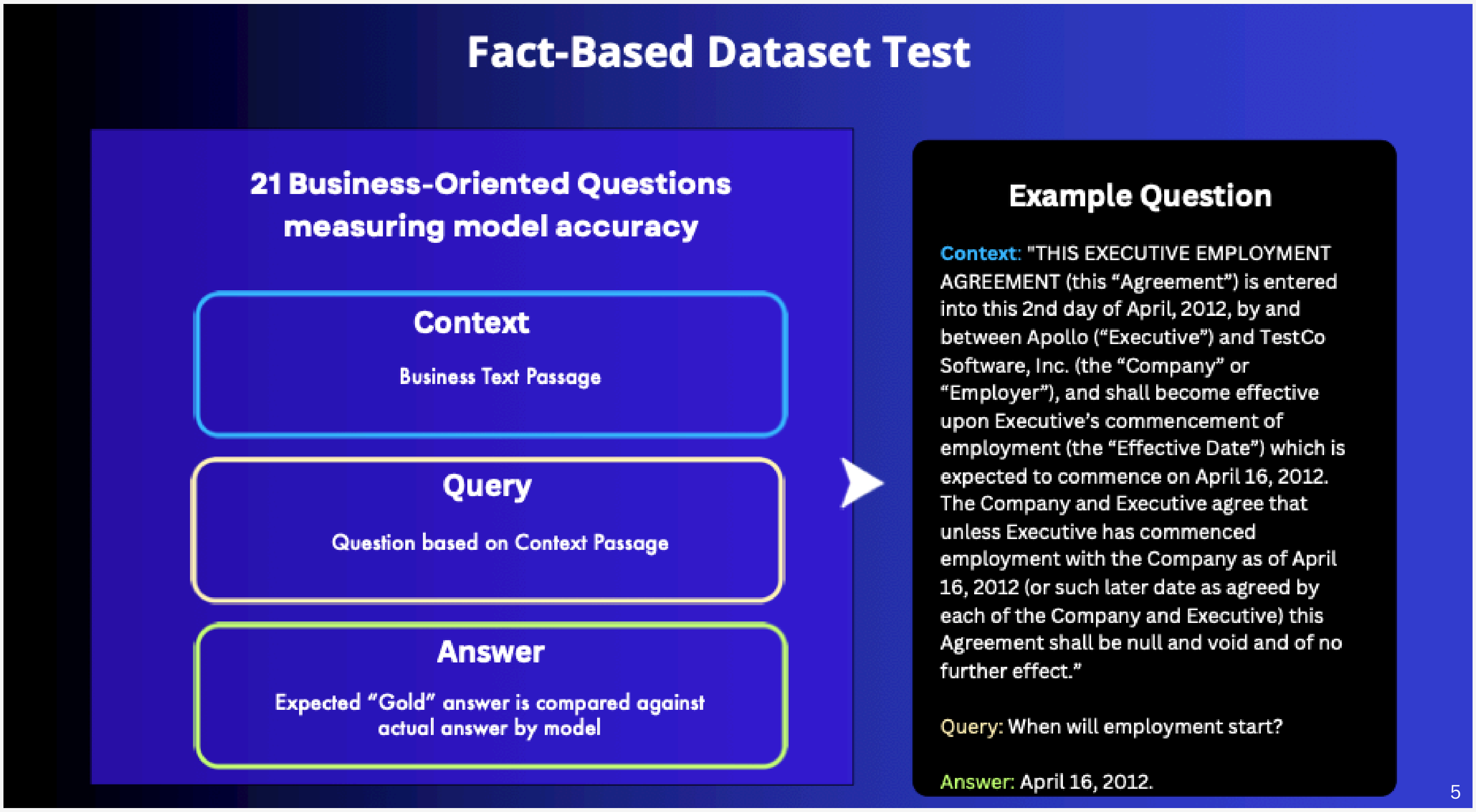
Deployment of AI on AI PCs also ensures added safety and security, while mitigating risks associated with cloud-based or remote AI services. With increasing concerns about AI's impact on confidentiality and data protection – as highlighted by a recent survey where 51% of the respondents cited these as their primary concerns with using AI in the workplace[2] – the ability to securely host AI is paramount. On-device AI workflows utilize SLMs which have smaller codebases with fewer potential surfaces for security breaches, and are less vulnerable to malicious attacks. In addition, many workflows are able to function without Wi-Fi access, supporting air-gapped environments essential for users in resource-constrained or high data-sensitive roles.

AI PCs support secure GenAI workflows - even without Wi-Fi access

Comparing Model Inference Speed Gain for Intel Laptops (Meteor Lake vs. Lunar Lake)

The release of high-powered processors on AI PCs by Intel starting with Meteor Lake in January 2024 and the launch of Intel Core Ultra Processor (Series 2) in Fall of 2024 unlock an entirely new pattern of distributing AI capability. When paired with an optimized AI framework and model inferencing capability, AI PCs will now be able to deliver many experiences at the user level that was previously made possible only with large language models operating through a centralized and complex GPU cluster through AI model vendors or at the enterprise private cloud level.

To test whether AI PCs can deliver the performance capabilities of AI workflows at the hardware level, we conducted extensive testing on laptops[3] using Intel Core Ultra Series 1 and Series 2 machines and MacBook Pro M1 and M3. With publicly-available and widely-used laptops, and technologies that are all open source and available for others to replicate, our goal was to construct a baseline ‘real world’ benchmark to evaluate the effectiveness of end-to-end inferencing. We ran this test using fine-tuned LLMWare models that are based on five of the leading open-source foundational models in sizes ranging from 1 billion up to 9 billion parameters. To test the accuracy of our fact-based question answering models, we developed and published a 21-question, context and answer dataset that we have been using for a variety of testing purposes over the last year using LLMWare’s Model HQ for the Intel machines and LLMWare’s open source repository for the MacBook Pro machines.[4]



Fact-Based Dataset Test

21 Business-Oriented Questions measuring model accuracy

Context

Business Text Passage

Query

Question based on Context Passage

Answer

Expected “Gold” answer is compared against actual answer by model

Example Question

Context: "THIS EXECUTIVE EMPLOYMENT AGREEMENT (this “Agreement”) is entered into this 2nd day of April, 2012, by and between Apollo (“Executive”) and TestCo Software, Inc. (the “Company” or “Employer”), and shall become effective upon Executive’s commencement of employment (the “Effective Date”) which is expected to commence on April 16, 2012. The Company and Executive agree that unless Executive has commenced employment with the Company as of April 16, 2012 (or such later date as agreed by each of the Company and Executive) this Agreement shall be null and void and of no further effect.”

Query: When will employment start?

Answer: April 16, 2012.

Testing Methodology

We constructed our test based on our open source example file that we published a year ago, which has the benefit of being both widely-available, widely and consistently used in our previous testing, and also avoids any potential “cherry-picking” in the design to potentially bias one of the technologies being tested.

In this testing scenario, we aimed to reproduce a Retrieval Augmented Generation (RAG) scenario for fact-based question-answering. The test consisted of 21 fact-based question-answering context passages, across a wide range of business, financial and general news topics, with context passages ranging between 100 – 500 tokens, combined with a fact-based question, and then answers typically in the range of 10-100 tokens.

The last two questions generate larger answers, and are usually a good quick test of potential ‘saturation’ as the model runs, and do take the longest time in the process. We cap generation at 100 tokens maximum, with the expectation of using the model as a productivity tool. The full set of questions and answers can be found at this link.

Our open source library contains over 100 examples of how we use small language models as a productivity tool for enterprise use cases, an example of which can be found here which shows how small language models can be used to extract text to query external web sources to complete a complex research report for financial analysis.

Quantization Methods Used

llama.cpp/GGUF - Best for Macs

OpenVINO - Best for Intel processors

Our earlier evaluation in our previous white paper [5] showed us that because OpenVINO produced such superior results in our Dell Ultra9/Intel laptop, to conduct a fair evaluation of inference speeds on each of the laptops, we needed to use the quantization format best suited to each machine. Therefore, to frame the performance test as “best” on Mac versus “best” on Wintel, we decided to use these versions of the models: Mac – 4-bit GGUF and Dell/Intel – ‘int4’ OpenVINO on Intel GPU (Ultra).

Using the testing methodology described above, we first compared the performance difference between Series 1 (Meteor Lake) and Series 2 (Lunar Lake) in Intel Core Ultra processors with the following results using Dell Laptops:

Total time for 21 Fact-based Q&A Inferences (lower is better for time in seconds)

Model Name	**Model Parameters (B)**	**Dell - Lunar Lake- Ultra 7 258V 2.20 GHZ - 32 GB** (seconds)	**Dell - Meteor Lake- Ultra 9 185H 2.50 GHz - 32 GB** (seconds)	**LNL v. MTL Speedup (x Faster)**
bling-tiny-llama	1.1	9.05	15.27	**1.69**
bling-phi-3	3.8	24.72	43.03	**1.74**
dragon-mistral	7.3	33.46	71.23	**2.13**
dragon-llama2	7	36.74	75.93	**2.07**
dragon-yi-9b	8.8	47.07	89.9	**1.91**

Simply put, Lunar Lake is truly exceptional. Lunar Lake stands out as a groundbreaking advancement in AI model inferencing, offering performance that is up to 2.1 times faster than its predecessor, Meteor Lake. While Meteor Lake already achieved impressive sub-second response times per inference, Lunar Lake takes it further with even faster results. Specifically, with LLMWare's Model HQ, Lunar Lake delivers average inference times of just 0.43 seconds for the bling-tiny-llama model and 1.59 seconds for the dragon-mistral 7 billion parameter model, setting a new standard for AI workflows on personal computers.

Up to 2.1x Faster than its predecessor

Simply put, Lunar Lake is truly exceptional... a groundbreaking advancement in AI model inferencing

Run 7-9 billion parameter models LOCALLY w/ exceptional speed

by pairing AI PCs with MODEL HQ's optimization and deployment software

This exceptional performance positions Lunar Lake as a potential catalyst for transforming AI workflow deployment and usage. Its ability to deliver high-speed inferencing, even for larger SLMs in the 7-9 billion parameter range, significantly broadens the scope of possible applications for AI PCs. Traditionally, models of this size—offering robust capabilities necessary for a wide array of AI-driven workflows—have relied on separate inference servers and hosted GPUs accessed via APIs, limiting their practical use on most laptops due to slower inference speeds that hinder user experience.

Lunar Lake's superior performance mitigates these constraints, demonstrating that when paired with the right optimization technique and software, such as Model HQ that automatically deploys optimization based on its hardware environment, AI PCs can enable the seamless execution of larger models locally without compromising speed or capability. Its ability to deliver "inference server" level performance directly on local devices enables users to achieve high-speed inferencing with models that are more adept at handling complex enterprise tasks. Lunar Lake effectively bridges the gap between high-performance server-based AI solutions and portable, accessible on-device AI, making it a pivotal innovation in the landscape of AI model deployment.

Decentralizing AI for ALL

Comparing Model Inference Times for Lunar Lake, Meteor Lake, MacBook Pro M1 and M3

To further evaluate performance, we conducted a comparison between the MacBook M3, a widely used laptop in AI workflow development, and the newly released Lunar Lake processors. Additionally, we revisited a previous benchmark from an earlier white paper[*ibid*], where the Meteor Lake version of the Intel Core Ultra Processor demonstrated superior speed compared to the MacBook Pro M1 and M3.

For this evaluation, we continued utilizing the GGUF format for inferencing models on MacBooks, a method widely considered the most efficient for Mac, and the OpenVINO format for inferencing on Lunar Lake, as using the OpenVINO delivers the fastest speed by far on Intel machines. In this updated test, we assessed the latest Lunar Lake release of the Intel Core Ultra Processors (Series 2), comparing its performance against the Mac M1, Mac M3, and the Intel Core Ultra 9 (Series 1). Both tests were performed using the LLMWare platform - Model HQ for Intel machines and the publicly available open source library for MacBook Pro machines.

Here is the result of our latest finding for the 21 question inferencing test with times shown being the total time (lower is better for time in seconds):

Model Name	**Model Para-meters (B)**	**Dell - Lunar Lake-Ultra 7 258V 2.20 GHZ iGPU - 32 GB** (seconds)	**Dell - Meteor Lake-Ultra 9 185H 2.50 GHz iGPU - 32 GB** (seconds)	**Mac M3 Max**	**Mac M1**	**Lunar Lake v. Mac M3 Max Speedup (x Faster)**
bling-tiny-llama	1.1	9.05	15.27	23.27	31.30	**2.57**
bling-phi-3	3.8	24.72	43.03	61.40	81.10	**2.48**
dragon-mistral	7.3	33.46	71.23	96.80	113.20	**2.89**
dragon-llama2	7	36.74	75.93	97.65	128.30	**2.66**
dragon-yi-9b	8.8	47.07	89.9	143.75	172.50	**3.05**

Up to 3x Faster than MacBook M3 Max

Importance of RAM/Processor Combination in AI Model Inferencing

We next examined the impact of RAM and processor performance in model inferencing. In this comparison test, the MacBook M3 Max, a very high-end unit with 36 GB of RAM, had a slightly larger memory capacity over the Intel machine, which are equipped with only 32 GB of RAM. Despite this, Lunar Lake consistently delivered superior performance to MacBook M3 Max for inference speed, enabling workflows not previously possible on a local laptop. We next compared two Dell laptops with different RAM capacity - 32 GB vs 16 GB. Although it would have been ideal to have tested on laptops with the same underlying processor with different RAMs for a true comparison test, the performance differences were still significant enough that it deserves mention. Finally, we also examined the performance of the Lunar Lake Ultra 5 (16 GB) versus the Meteor Lake Ultra 9 (32 GB).

Here is the result of our latest finding for the 21 question inferencing test with times shown being the total time (lower is better for time in seconds):

Model Name	**Model Para-meters (B)**	**Dell Lunar Lake Ultra 7 258V 2.20 GHz iGPU- 32 GB** (seconds)	**Dell Lunar Lake Ultra - 5 - 236V 2.10 GHz iGPU- 16 GB** (seconds)	**Dell Meteor Lake- Ultra 9 185H 2.50 GHz iGPU- 32 GB** (seconds)	**Mac M3 Max 36 GB** (seconds)	**Mac M1 (32 GB)** (seconds)
bling-tiny-llama	1.1	9.05	14.01	15.27	23.27	31.30
bling-phi-3	3.8	24.72	34.69	43.03	61.40	81.10
dragon-mistral	7.3	33.46	43.73	71.23	96.80	113.20
dragon-llama2	7	36.74	49.05	75.93	97.65	128.30
dragon-yi-9b	8.8	47.07	59.75	89.9	143.75	172.50

For the bling-tiny-llama model, the Dell Lunar Lake 32 GB machine performed 1.5x faster than the 16 GB machine. It is important to note that the 16 GB laptop was Dell's Ultra 5 machine versus the 32 GB which was Ultra 7. Additionally, the Dell Ultra 5 (Lunar Lake) machine with only 16 GB of RAM still delivered significantly faster results than the Meteor Lake Ultra 9 with 32 GB RAM. Although these results do not speak to exactly how much RAM can impact speed performance, there is still the case that more performant base processor combination over a larger RAM capacity may still be preferable for model inferencing for SLMs.

Key Takeaways

The deployment of AI workflows on AI PCs represents a significant shift towards decentralizing AI usage and access, overcoming previous hardware and performance limitations. The convergence of enhanced hardware capabilities in AI PCs and the accelerated improvement of AI models along with improved understanding of using optimized quantizing, inferencing and deployment methods, is set to unlock a wide range of business productivity use cases at the personal device level.

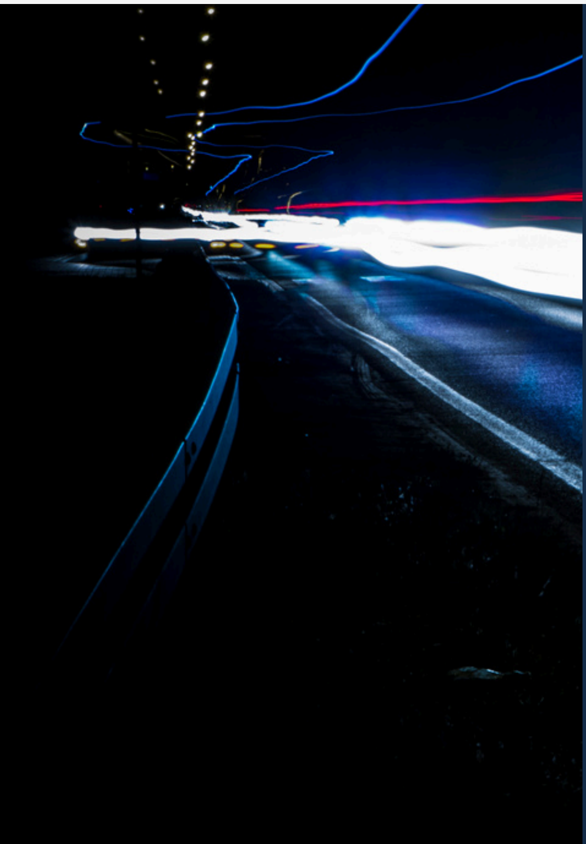
Based on our performance speed testing, we note the following:

1) AI PCs are clearly able to deliver powerful Generative AI capabilities at the user level when paired with the right software to optimize performance;

2) Using LLMWare's Model HQ, Intel's Lunar Lake can achieve inference speed performance far superior to more expensive rivals such as MacBook M3 Max;

3) This level of inference capability unlocks Gen AI use cases previously only accessed through external APIs linked to GPUs and is critical to decentralized deployment of Gen AI at the user level; and

4) The trend toward decentralized AI being powered by AI PCs is poised to have a significant impact on the cost of AI consumption to near-zero as many enterprises consider laptop purchases as operational costs, versus per-token consumption metrics prevalent today.



With recent improvements in hardware, software and SLMs, cost of AI consumption will edge toward near-zero at user level and move away from per token metrics for enterprises

Introducing: Model HQ by LLMWare

As demonstrated by our inference speed testing, enterprises can now safely implement AI workflows directly at the user level, reaping substantial cost and security benefits. By leveraging increasingly capable SLMs and a robust AI solution stack, business users can seamlessly execute AI-driven workflows on their devices. However, to fully capitalize on these advancements, enterprises require a simplified, lightweight, integrated solution that consolidates the essential components of a typical AI technology stack into a streamlined package.

To fully harness the benefits of AI PCs, enterprises need robust solutions that facilitate the deployment, monitoring, updating, and scaling of generative AI models across diverse hardware environments while maintaining stringent security and privacy standards. This involves ensuring that AI workflows can operate smoothly and securely on potentially hundreds or thousands of devices, all while mitigating risks and optimizing performance in a decentralized deployment landscape.

Model HQ is a first-in-kind comprehensive platform for AI deployment on PCs, specifically designed for optimization with Intel AI PCs, and designed to manage the entire lifecycle of lightweight, private LLM-based applications safely and efficiently. It gives enterprises full control over deploying AI workflows directly on user PCs, offering the easiest and most automated way to leverage the best AI framework and model for their hardware.

With Model HQ, AI developers and IT teams can quickly deploy a variety of AI workflows for easy low to no-code development, utilizing over 150 models from LLMWare's Model Depot, including the largest collection of small language models in the OpenVINO format optimized for Intel Core Ultra Processors (Series 1 and 2 or Meteor Lake and Lunar Lake).

MODEL HQ

All-in-One Platform for easily creating and deploying lightweight AI apps for Enterprise using Small Language Models

ENTERPRISE CONTROL

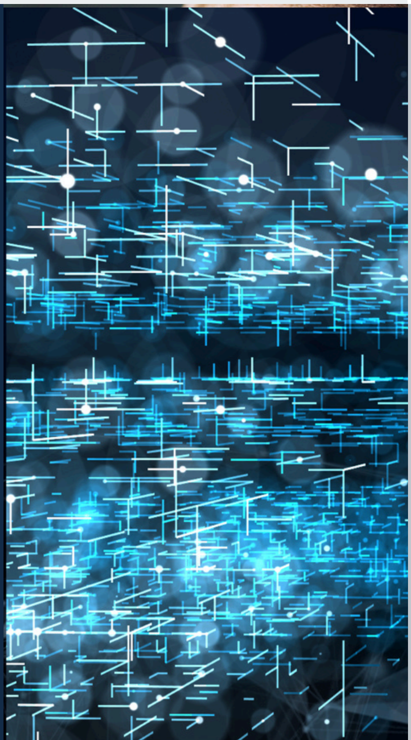
One platform allows for full control over Gen AI app creation and deployment. From easy low-code app creation to sharing apps with other enterprise users with full access control, visibility and ability to change, update and modify to future-proof your AI workflow.

SAFETY & COMPLIANCE

Platform tracks every model inference for AI Explainability, Compliance and Audit reporting. Includes safety controls for PII redaction, toxicity, bias and hallucination monitors for full safety and compliance control.

PRIVACY & SECURITY

All workflows are created using Small Language Models for the most secure deployment in your own privacy zone: Private Cloud, On Prem or On Device for the enterprise based on use case. Platform includes model safety and security checks and private repo.



COST EFFECTIVE

No more surprising token charges for AI models. Find the most efficient ways to run AI inferencing in your enterprise based on use case to maximize cost efficiency. From laptops to private cloud and everything in-between, match the right SLM and workflow for your desired use case.

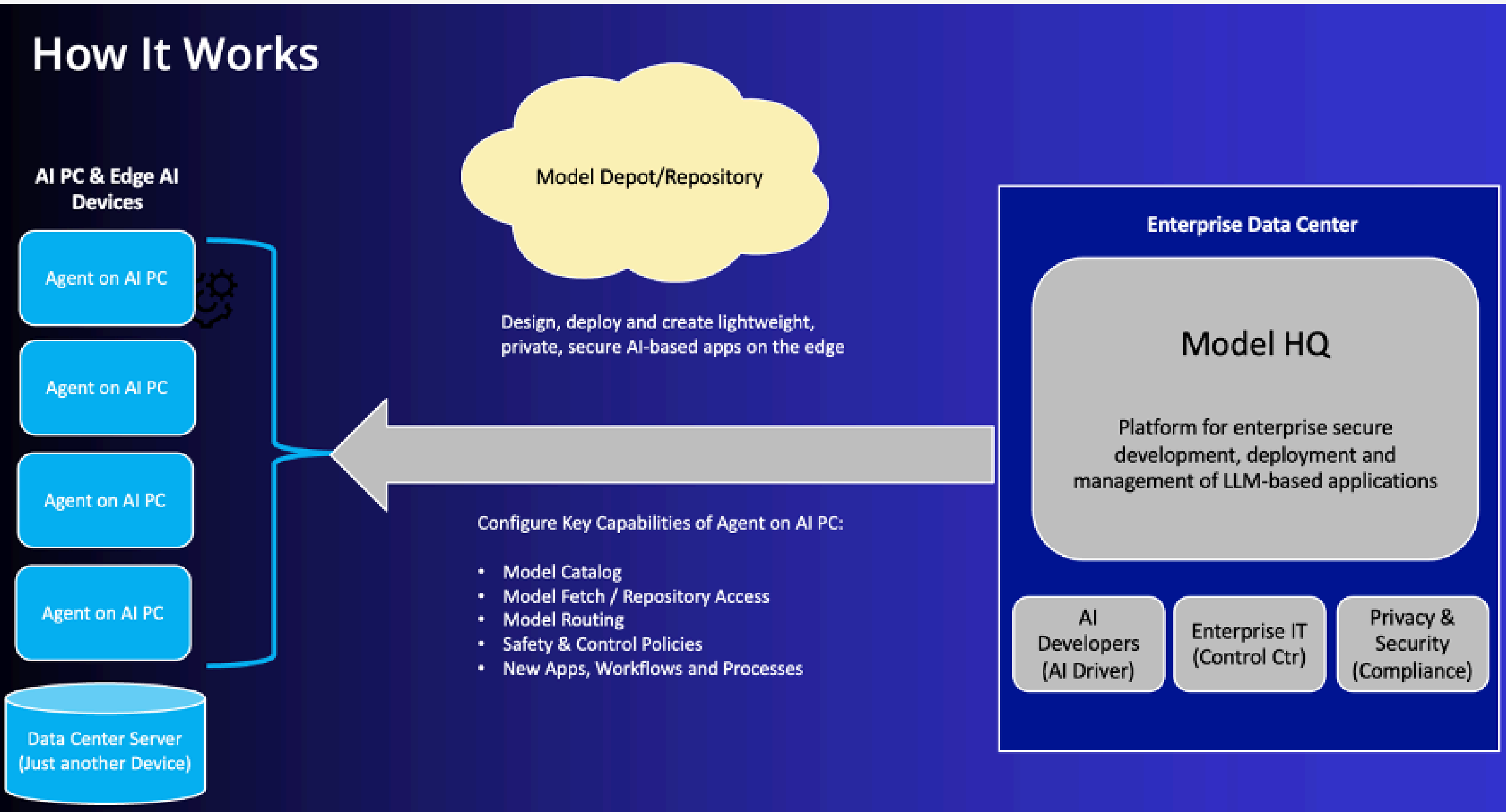
Comprehesive Platform for Safe and Secure AI Deployment

Model HQ is an enterprise platform with a separate client agent software that can be deployed in individual AI PCs to enable model inferencing locally. Once installed, the client agent software unlocks powerful inferencing capabilities for virtually all types of AI models at the user level which can also be permissioned and monitored by the enterprise. In addition, Model HQ's client agent comes with integrated RAG, contract analysis, test to SQL query and voice transcription search capabilities out of the box, delivering immediate value to the end user.

The AI PC offers an unprecedented potential for a highly distributed, decentralized mode of rolling out AI-based applications. With Model HQ, enterprise AI developers can seamlessly update and deploy lightweight AI apps and AI workflows to users of AI PCs while benefitting from integrated safety and security features, including safeguards that detect compromised models and checks for prompt injections, toxicity, bias, and hallucinations. Additionally, the platform features a centralized Compliance Station that offers on-demand safety and data configuration settings, AI Explainability Tracing, Data Privacy Guard with PII filtering, and a comprehensive Audit Log with automated reporting.

Model HQ not only simplifies the deployment of AI workflows but also ensures that enterprises maintain full control over model safety, security, compliance, and auditing—key elements for any successful AI implementation.

How It Works


AI PC & Edge AI Devices
Agent on AI PC
Agent on AI PC
Agent on AI PC
Agent on AI PC
Data Center Server (Just another Device)
Model Depot/Repository
Enterprise Data Center
Model HQ
Platform for enterprise secure development, deployment and management of LLM-based applications
AI Developers (AI Driver)
Enterprise IT (Control Ctr)
Privacy & Security (Compliance)

Design, deploy and create lightweight, private, secure AI-based apps on the edge

Configure Key Capabilities of Agent on AI PC:

- Model Catalog
- Model Fetch / Repository Access
- Model Routing
- Safety & Control Policies
- New Apps, Workflows and Processes

Model HQ Features

A Comprehensive AI Model Management Framework

At the core of Model HQ's capabilities is a sophisticated abstraction layer that seamlessly integrates OpenVINO for maximum capability on Intel AI PCs. Out of the box, the Model HQ's framework supports four distinct model inferencing technologies—PyTorch, GGUF, ONNX, and OpenVINO—with all necessary back-end functionalities for private and secure execution. This high-level process design and safety control framework eliminates the need for developers to manage low-level implementations or decide which inferencing technology best fits their development platform. Model HQ automatically handles these complexities, allowing developers to focus on application development.

Auto AI-Optimization: Streamlining AI Framework Deployment

A key feature of the Model HQ platform is its Auto AI-Optimization technology in the Client Agent, which automatically identifies and deploys the optimal AI framework for the user's specific hardware environment with OpenVINO technology as key integration to extract optimum performance on Intel AI PCs.

Model HQ simplifies the use of AI frameworks by abstracting the complexities involved in implementing them, allowing users to leverage the most suitable and optimized model type for their tasks. For instance, on laptops equipped with Intel Core Ultra Processors with integrated GPUs (iGPUs), the Client Agent automatically detects and utilizes the iGPU to enhance AI workflows. As highlighted earlier, OpenVINO models deliver the fastest inferencing speeds on Intel iGPUs, with significant performance improvements for AI workflows. Model HQ, therefore, prioritizes the use of OpenVINO models when available for Intel iGPUs when available, ensuring that the user benefits from the most efficient and accelerated AI inferencing capabilities provided by the hardware.

This automated approach not only optimizes performance but also reduces the complexity and time required for users to achieve the best AI model execution on their devices, enhancing productivity and enabling seamless integration of AI capabilities across diverse hardware environments. By supporting the most popular AI frameworks - PyTorch, GGUF, ONNX, and OpenVINO – Client Agent and Model HQ enables seamless execution of most models on compatible hardware without requiring additional platform configurations. Recognizing the significant performance variations between model technologies on different platforms, the ability to "mix and match" inferencing technologies without the complexities of dependency management or rigid workflow integrations is crucial. This approach maximizes the performance of AI PCs by dynamically aligning the best inferencing technology with the optimal hardware, ensuring superior execution speeds and resource efficiency.

Conclusion

The emergence of AI PCs marks a pivotal shift in how enterprises can leverage AI technology, making advanced AI capabilities accessible at the user level. Intel Core Ultra Processors (Series 2) Lunar Lake stands at the forefront of this transformation, offering exceptional performance that rivals traditional inference servers, yet operates locally on personal devices. This capability not only enhances speed and efficiency but also addresses critical concerns around security and data privacy by keeping AI workflows self-hosted and secure.

With Lunar Lake, AI deployment on PCs becomes a practical and powerful solution for businesses looking to harness the full potential of AI without the constraints of centralized models and cloud dependency. As demonstrated by our tests, Lunar Lake significantly outperforms its predecessors and competitive devices, setting new standards for inference speed and operational efficiency.

By integrating Lunar Lake with platforms like Model HQ by LLMWare, enterprises can further streamline AI implementation, benefitting from a comprehensive end-to-end solution that supports the full lifecycle of AI workflow—from development and deployment to monitoring and compliance. The combination of high-performance hardware, optimized AI frameworks, and robust AI optimization and deployment software tools provides enterprises with the agility and control needed to innovate and scale AI-driven processes across diverse environments.

In conclusion, the advancements in AI PCs, exemplified by Lunar Lake, are not just incremental improvements but a transformative leap that enables a decentralized, efficient, and secure AI ecosystem. This new paradigm empowers enterprises to unlock the full potential of generative AI, driving productivity gains and innovation directly from the user’s PC, thereby paving the way for the next generation of AI-powered business solutions.

AI PCs are a transformative leap

Enabling a decentralized, efficient and secure AI ecosystem

Explore how LLMWare.ai can help optimize your AI workflow. Contact us today.

Website:
llmware.ai

Contact:
Namee Oberst

Endnotes

[1] There is currently no industry-set definition of small language models but it is widely believed to be models that do not require separate, stand-alone GPU support and are generally less than 9-10B parameters in size.
[2] Dutt, Ishan and Himani Mukka. "AI PC Success is Central to Lenovo's Broader Goals." July 9, 2024
Laptop Specs: Dell Inspiron 14 Plus 7440, Installed RAM 32 GB, Intel Core Ultra 9 185H 2.50 Ghz Processor, Windows 11
[3] Laptop specifications are as follows:
a)MacBook Pro M1 Chip, 8-core CPU with 6 performance cores and 2 efficiency cores, 14-core GPU, installed RAM 32 GB
b)MacBook Pro M3 Max Chip, 14-core CPU with 10 performance cores and 4 efficiency cores, 30-core GPU, installed RAM 36 GB
c)Dell Inspiron 14 Plus 7440, Intel Core Ultra 9 185H 2.50 Ghz Processor (Meteor Lake), installed RAM 32 GB, MSRP $1,099
d)Dell XPS 13 Intel Core Ultra 7 258V 2.20 GHz (Lunar Lake) w/ 32 GB RAM
e)Dell Intel Core Ultra 5 236V 2.20 GHz (Lunar Lake) w/ 16 GB RAM
[4] We have previously written about our findings on the importance of using an AI inferencing method that is optimized for the hardware in our previous white paper entitled AI PCs: Accelerating AI-Powered Productivity, in which we discussed in-depth the series of tests we conducted using MacBook Pro M1 and M3 and a Dell laptop powered by Intel Core Ultra 9 (Meteor Lake).
We described our various testing methodology in great detail and demonstrated the impact of integrating optimized AI frameworks with corresponding model compilers tailored to specific hardware configurations – i.e., using GGUF format for Macs and OpenVINO for intel-based machines produces the most optimized inference performance for each hardware platform. We also demonstrated extensively how pairing OpenVINO is critical in maximizing inference speed and enhancing the overall usability of Intel-powered AI PCs. Model HQ was used for Intel-based tests.
Full test used for MacBooks can be found here: https://github.com/llmware-ai/llmware/blob/main/examples/Models/bling_fast_start.py
Each question-answer-context consists of:
(1)"context" - text passage (in the range of 100-1000 tokens, e.g., a paragraph to a page);
(2)"question" – a question that can be answered based on the context passage; and
(3)"answer" – the expected 'gold' answer for the question.
Most of the answers are short question-answers of 5-50 tokens, with the last two questions more open-ended summarizations that yield longer responses, and are a good test of 'memory saturation'. For consistency of evaluation, we capped the output at 100 tokens.
We approached the test as a "real world" test – with commercial off-the-shelf 'retail' laptops - so did not take any unusual steps to prepare the machine, but generally tried to close most open tabs and avoid any resource intensive background processing running concurrently.
Each test for each model was run between 3-9 times, and then the times were averaged.
[5] Oberst, Darren, "AI PCs: Accelerating AI-Powered Productivity," in www.llmware.ai