# REVOLUTIONIZING AI DEPLOYMENT

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

(Abstract Version)





# WHAT

AI PCs will decentralize Gen AI at user level

# HOW

With Model HQ -Al deployment and access simplified for everyone

# **REVOLUTIONIZING** AI FOR ALL

Al PCs, a new class of personal computers designed to handle AI tasks locally, bring the promise of Al-powered productivity to life by enabling secure, private, and self-hosted AI workflows directly on users' devices. This white paper explores how AI PCs, specifically those powered by Intel® Core<sup>™</sup> Ultra Processors. address the challenge of delivering advanced AI capabilities at the PC and introduces Model level HO bv LLMWare.ai, a comprehensive solution that simplifies AI implementation for developers and enterprises, and unlock the full potential of Generative AI for business productivity.

Recent advancements in hardware and AI models have made AI deployment at the personal computer level both feasible and beneficial highly for enterprises. The introduction of Intel® Core™ Ultra Processors, featuring integrated GPUs and NPUs capable of rapid AI inference, has ushered in a new era of on-device AI. AI PCs, equipped with small language models (SLMs) [1], enable workflows previously dependent on large frontier models, such as text summarization, question-answering, and specialized tasks like natural language SQL queries and contract analysis. SLMs are efficient. lightweight, and less computationally intensive, making them ideal for localized use. Additionally, deploying AI on Al PCs enhances security by reducing vulnerabilities associated with cloud-based services, addressing confidentiality and data protection concerns. These workflows also support air-gapped environments, ensuring functionality in resource-constrained at the edge or high-sensitivity roles, providing a and efficient Al solution secure for enterprises.

## Comparing Model Performance Gains for Intel® Core™ Laptops

VS.

#### MacBook Pro M1 and M3 Max

The release of high-powered processors in 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.

Model HQ is designed to provide automated optimization for Intel® AI PCs. To test whether AI PCs can deliver the performance capabilities of AI workflows at the hardware level, we conducted extensive testing on laptops[2] using Intel® Core™ Ultra Series 1 and Series 2 machines and MacBook Pro M1 and M3 Max. With publicly-available and widely-used laptops, and technologies that are all open source and available for others to replicate[3], we used 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]

Using the testing methodology described in our previous white papers[5], we compared the performance difference between Series 1 (Meteor Lake) and Series 2 (Lunar Lake) in Intel<sup>®</sup> Core<sup>™</sup> Ultra processors with the following results:

| Model<br>Name    | Model<br>Parameters (B) | Dell -<br>Lunar Lake-<br>Ultra 7 258V<br>2.20 GHZ -<br>32 GB<br>(seconds) | Dell -<br>Meteor Lake-<br>Ultra 9<br>185H 2.50 GHz<br>- 32 GB<br>(seconds) | LNL v. MTL Speedup<br>(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                             |

#### Lunar Lake vs. Meteor Lake: Total time for 21 Fact-based Q&A Inferences (lower is better)

### Key Takeaways from Comparing Intel® Core™ Ultra Series 1 and Series 2

Simply put, Series 2 (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, Series 1 (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 Model HQ, Lunar Lake delivers average per inference times of just 0.44 seconds for the bling-tiny-llama model and 1.58 seconds for the dragon-mistral 7 billion parameter model, setting a new standard for AI workflows on personal computers.

Run models up to 22 BILLION parameters 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, and up to 22 billion parameters in our latest internal testing, significantly broadens the scope of possible applications for AI PCs.

Traditionally, models of this size—offering robust capabilities necessary for a wide array of Al-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, 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.

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

To further evaluate performance, we conducted a comparison between the MacBook Pro M3 Max, 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<sup>™</sup> Ultra Processor demonstrated superior speed compared to the MacBook Pro M1 and M3 Max.

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. OpenVINO is the AI inferencing software of choice that delivers the fastest performance on Intel® machines. In this test, we assessed the latest Lunar Lake release of the Intel® Core<sup>™</sup> Ultra Processors (Series 2), comparing its performance against the MacBook Pro M1, MacBook Pro M3 Max, 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<br>Name    | Model<br>Para-<br>meters<br>(B) | Dell -<br>Lunar<br>Lake-Ultra<br>7<br>258V 2.20<br>GHZ iGPU -<br>32 GB<br>(seconds) | Dell -<br>Meteor<br>Lake-Ultra<br>9<br>185H 2.50<br>GHz iGPU -<br>32 GB<br>(seconds) | Mac Pro<br>M3 Max | Mac Pro<br>M1 | Lunar Lake<br>v.<br>Mac M3 Max<br>Speedup<br>(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

### **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 Pro 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 that is largely tied to cloud GPU hosting costs of many vendors.



# Zero Cost Inferencing

Thanks to AI PCs, the cost of AI consumption will edge toward near-zero at the user level for most AI workflows and 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<sup>™</sup> Ultra Processors (Series 1 and 2 or Meteor Lake and Lunar Lake).

#### 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 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.

# MODEL HQ

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



# ENTERPRISE CONTROL

One platform allows for full control over Gen Al app creation and deployment. From easy lowcode 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 Al 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 inbetween, match the right SLM and workflow for your desired use case.

### **Model HQ Feature Highlight**

#### 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 for Intel iGPUs, 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 Model HQ by LLMWare, enterprises can further streamline Al 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.

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

<u>llmware.ai</u> Contact:

Website:

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.

- 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
- 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

Full test used for MacBooks can be found here: <u>https://github.com/llmware-ai/llmware/blob/main/examples/Models/bling\_fast\_start.py</u>

[5] Oberst, Darren, "AI PCs: Accelerating AI-Powered Productivity," and Oberst, Namee, "Revolutionizing AI Deployment" in www.llmware.ai

<sup>[2]</sup> 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

<sup>[3]</sup> https://github.com/llmware-ai/llmware

<sup>[4]</sup> 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).