HKT Enterprise Solutions

1010 | 5G

Running Private AI on Mac: Unlock Enterprise Data Security and Privacy!

Applicable Scenario

- Aim to use AI technology to enhance productivity
- Involve confidential information cannot upload to cloud
- Require workflow engine for integration with internal system / database / document library
- Looking for small scale start-up or experiential stage
- Ability to scale-up to support more users

M3 Ultra Mac Studio – Private AI on Mac Quick Start Kit

1. M3 Ultra Mac Studio
    - Apple M3 Ultra chip with 32-core CPU, 80-core GPU, 32-core Neural Engine
    - 512GB unified memory shared with CPU, NPU and GPU
    - 1 / 4 / 8 /16 TB SSD storage
    - Software ： macOS
2. AppleCare+ for Mac Studio
3. Private AI on Mac Quick Start Kit service
    - Installation of LM Studio
    - Installation of selected common LLMs (Deekseek and Qwen)
    - Installation of Dify
    - Setup 2 AI demo use cases
    - 2 hours Briefing
4. Scale up Multiple Mac Studio Setup Addon service

Apple M3 Ultra Mac Studio

One of the best choice for your Private AI system

- **Blazing-fast CPU:** A powerful **32-core CPU**, significant performance in multi-core tasks, essential for efficiently running complex AI algorithms.
- **High-end GPU:** An **80-core GPU**, substantial boosts in graphics-intensive applications & gaming, also accelerating AI model training and inference processes.
- **Massive Memory:** Offers **512GB** of unified memory (RAM and VRAM combined), enabling local processing of extremely large AI models. Crucial for tasks requiring substantial computational power, such as DeepSeek R1 model 671b Q4, which needs ~**404GB** of VRAM to run effectively.
- **Unified Memory Architecture:** Allows seamless sharing of memory between CPU and GPU, eliminating the need to copy data between separate memory systems. This reduces latency & significantly improves performance for AI tasks that demand quick data access & processing.
- **High Bandwidth Memory:** High data transfer rates to facilitate quicker access to data for both CPU and GPU. It is vital for applications that process large amounts of data rapidly, such as AI model inference and 3D rendering. (In Dave2D review, it achieved **819GB/s**)

Super High Performance

Apple M3 Ultra Mac Studio
Unleashing the Power of Private AI

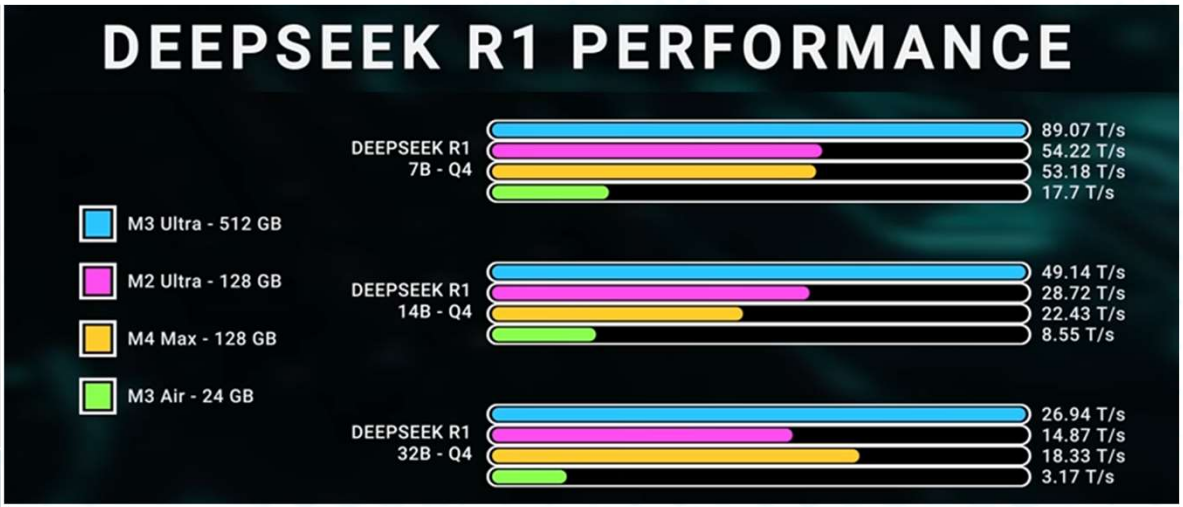
- **Local AI Processing:** The significant memory capacity makes it capable of running very large AI models entirely on the device, without needing cloud connectivity, enhancing usability for AI developers.
- **Privacy Focused:** Local AI processing enhances data privacy and security, making it suitable for applications handling sensitive information, such as healthcare and finance.
- **Suitable for Large Models:** Capable of handling models with billions of parameters, surpassing the capabilities of most consumer-grade systems, which is essential for advanced AI research and applications.
- **Low Power Consumption:** Power draw ~200W when running DeepSeek R1. (in Dave2D review testing)

All-in-One Private AI

Dave2D – M3 Ultra Mac Studio Review

MEMORY BANDWIDTH

	Bandwidth	Max Memory
M3 Ultra	819 GB/s	512 GB
M2 Ultra	800 GB/s	192 GB
M4 Max	546 GB/s	128 GB
M4	120 GB/s	32 GB
RTX 5090	1792 GB/s	32 GB
RTX 4090	1008 GB/s	24 GB



Private AI on Mac Quick Start Kit service
AI demo use case

- **Agent #1: LLM to MYSQL database**, allowing users to ask about sales or inventory reports using natural language, without writing any script

- **Agent #2: LLM to knowledge base**, allowing users to ask about operation or maintenance procedures without having to look up any PDFs or documents. It will also demonstrate how clustered Mac can handle load balancing

Private AI on Mac Quick Start Kit service
Scale up Multiple Mac Studio Setup Addon service