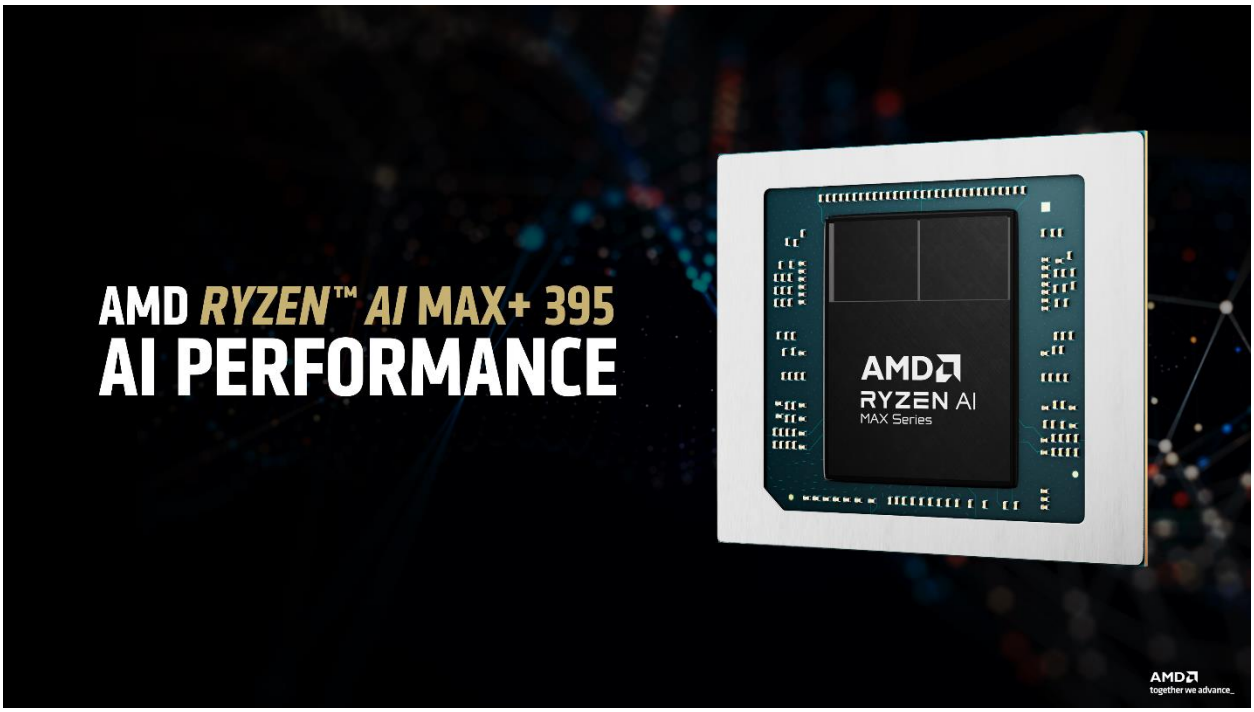
AMD Ryzen™ AI MAX+ 395 Processor: Breakthrough AI Performance In Thin And Light



The AMD Ryzen™ AI MAX+ 395 (codename: “Strix Halo”) is the most powerful x86 APU and delivers a significant performance boost over the competition. Powered by “Zen 5” CPU cores, 50+ peak AI TOPS XDNA™ 2 NPU and a truly massive integrated GPU driven by 40 AMD RDNA™ 3.5 CUs, the Ryzen™ AI MAX+ 395 is a transformative upgrade for the premium thin and light form factor. The Ryzen™ AI MAX+ 395 is available in options ranging from 32GB all the way up to 128GB of unified memory – out of which up to 96GB can be converted to VRAM through AMD Variable Graphics Memory.

The Ryzen™ AI Max+ 395 excels in consumer AI workloads like the llama.cpp-powered application: LM Studio. Shaping up to be *the* must-have app for client LLM workloads, LM Studio allows users to locally run the latest language model without any technical knowledge required. Deploying new AI text and vision models on Day 1 has never been simpler.

The “Strix Halo” platform extends AMD performance leadership in LM Studio with the new AMD Ryzen™ AI MAX+ series of processors.

As a primer: the model size is dictated by the number of parameters and the precision used. Generally speaking, doubling the parameter count (on the same architecture) or doubling the

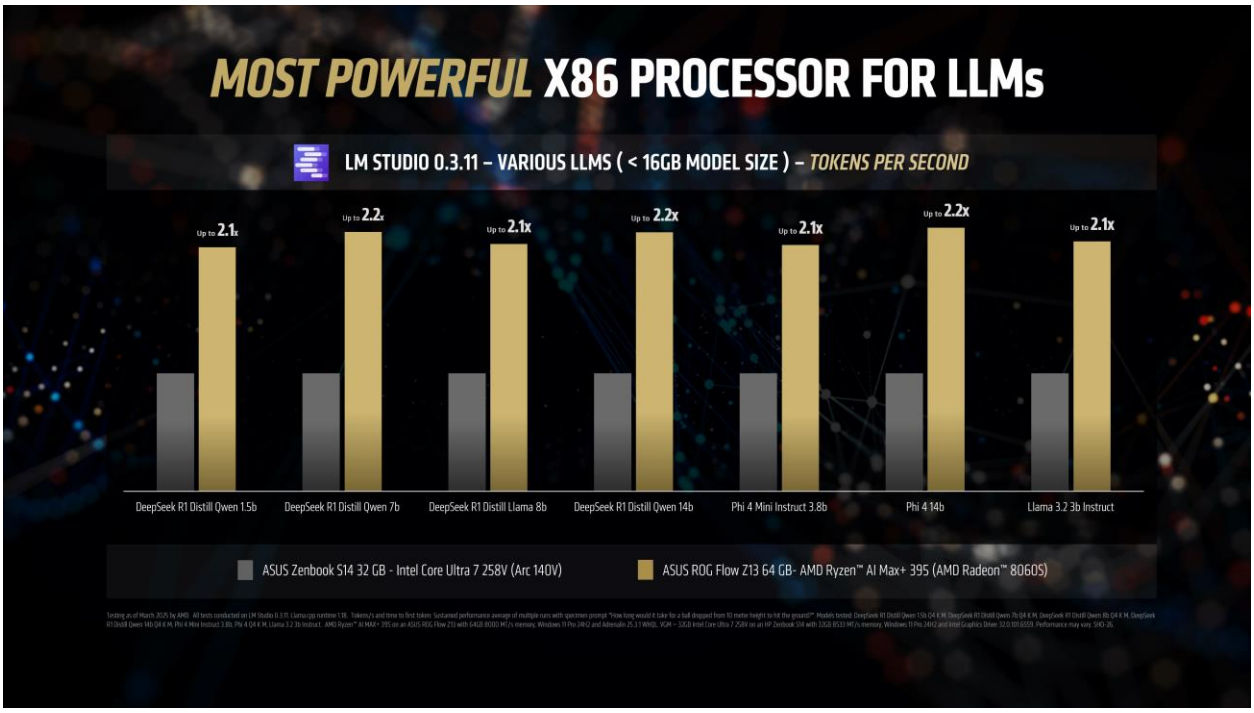
precision will also double the size of the model. Most of our competition's current-generation offerings in this space max out at 32GB on-package memory. This is enough shared graphics memory to run large language models (roughly) up to 16GB in size.

Benchmarking text and vision language models in LM Studio

For this comparison, we will be using the ASUS ROG Flow Z13 with 64GB of unified memory. We will restrict the LLM size to models that fit inside 16GB to ensure that it runs on the competition's 32GB laptop.

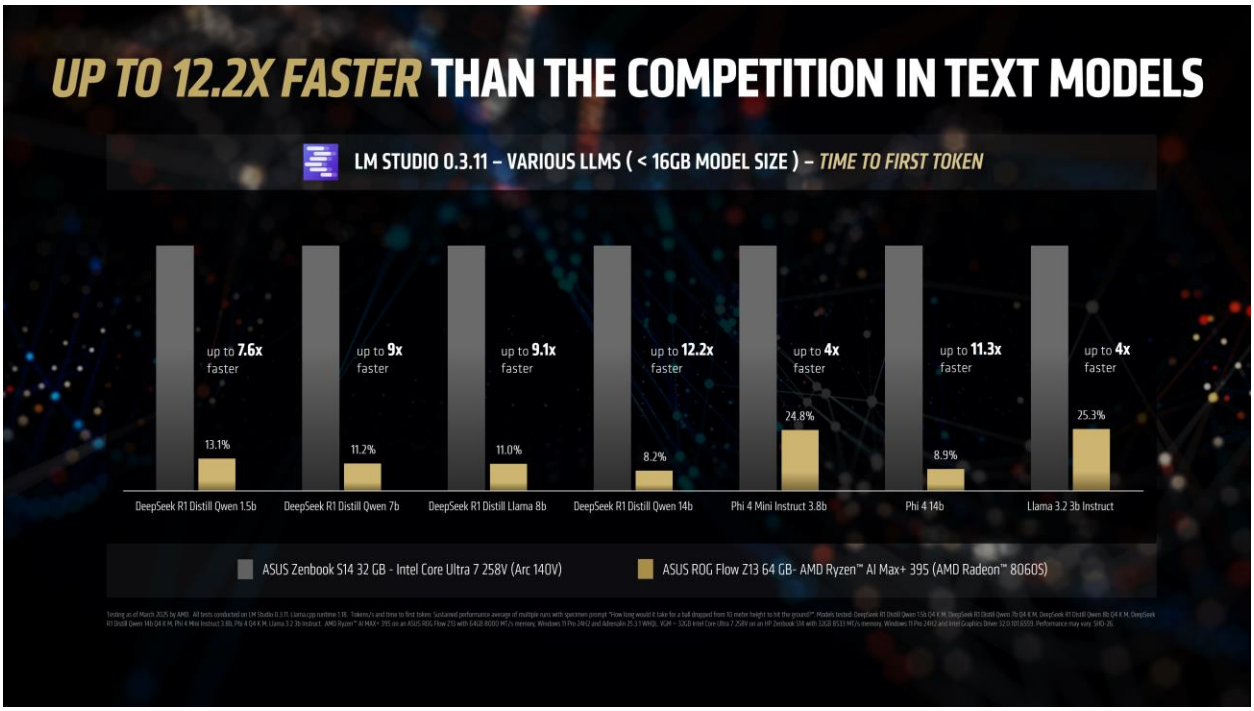
From the results, we can see that the ASUS ROG Flow Z13 - powered by the integrated Radeon™ 8060S and taking full advantage of the 256 GB/s bandwidth - effortlessly achieves up to 2.2x the performance of the Intel Arc 140V in token throughput.

The performance uplift is very consistent across different model types (whether you are running chain-of-thought DeepSeek R1 Distills or standard models like Microsoft Phi 4) and different parameter sizes.



In time to first token benchmarks, the AMD Ryzen™ AI MAX+ 395 processor is up to 4x faster than the competition in smaller models like Llama 3.2 3b Instruct.

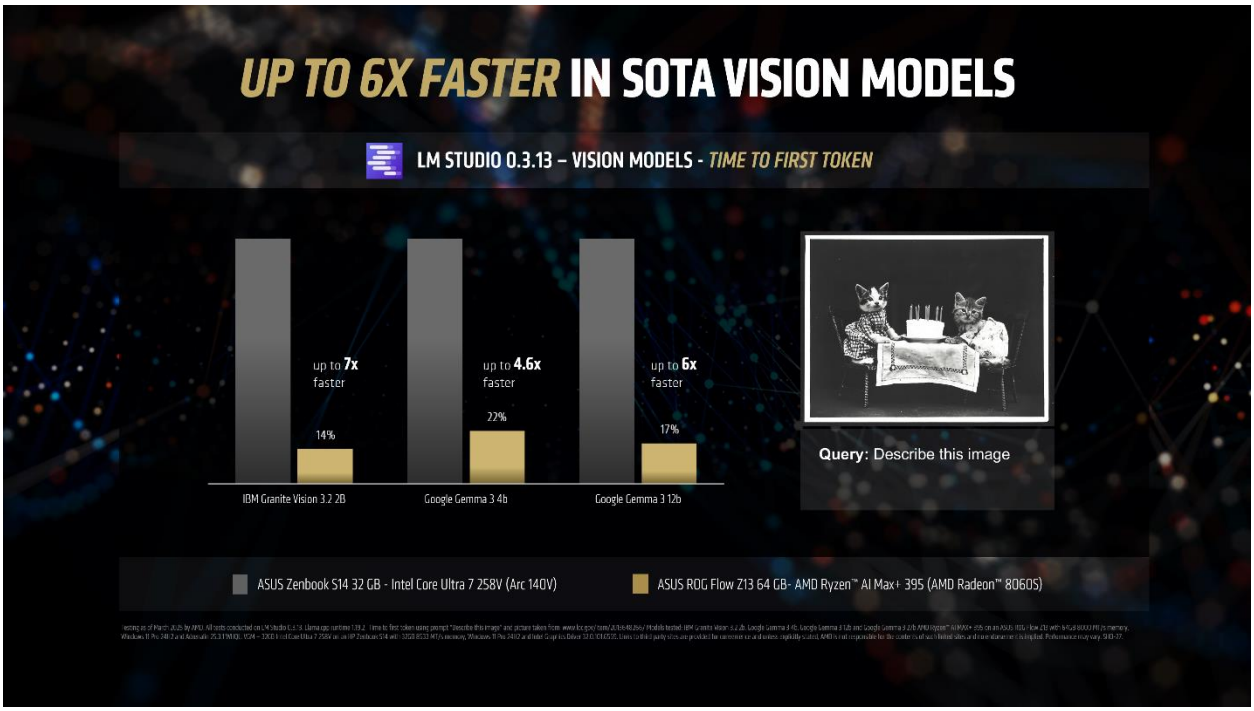
Going up to 7 billion and 8 billion models like the DeepSeek R1 Distill Qwen 7b and DeepSeek R1 Distill Llama 8b, the Ryzen™ AI Max+ 395 is up to 9.1x faster. When looking at 14 billion parameter models (which is approaching the largest size that can comfortably fit on a standard 32GB laptop), the ASUS ROG Flow Z13 is up to 12.2x faster than the Intel Core Ultra 258V powered laptop – *more than an order of magnitude faster than the competition!*



The larger the LLM, the faster AMD Ryzen™ AI Max+ 395 processor is in responding to the user query. So whether you are having a conversation with the model or giving it large summarization tasks involving thousands of tokens – the AMD machine will be much faster to respond. This advantage scales with the prompt length – so the heavier the task – the more pronounced the advantage will be.

Text-only LLMs are also slowly getting replaced with highly capable multi-modal models that have vision adapters and visual reasoning capabilities. The IBM Granite Vision is one example and the recently launched Google Gemma 3 family of models is another – with both providing highly capable vision capabilities to next generation AMD AI PCs. Both of these models run incredibly performantly on an AMD Ryzen™ AI MAX+ 395 processor.

An interesting point to note here: when running vision models, the time to first token metric also effectively becomes the time it takes for the model to analyze the image you give it.



The Ryzen™ AI Max+ 395 processor is up to 7x faster in IBM Granite Vision 3.2 3b, up to 4.6x faster in Google Gemma 3 4b and up to 6x faster in Google Gemma 3 12b. The ASUS ROG Flow Z13 came with a 64GB memory option so it can also effortlessly run the Google Gemma 3 27B Vision model – which is currently considered the current SOTA (state of the art) vision model.

[Google Gemma 3 Diagnosing Cancer from MRI/CT scan Demo]

A very cool demo showcasing this capability is shown above. When given a stock image of a CT scan, the model was able to identify the organs and provide a diagnosis. This is just one of many possible use cases for vision models that have complex vision modalities and visual reasoning capabilities – and these are only going to get better with time.

Another example is running the DeepSeek R1 Distill Qwen 32b in 6-bit precision (while 4-bit is industry standard for everyday use cases, coding can require higher precision to maintain coding accuracy) – which you can use to code a gaming classic in roughly 5 minutes:

[Pong demo]

Setting up for LLM runs

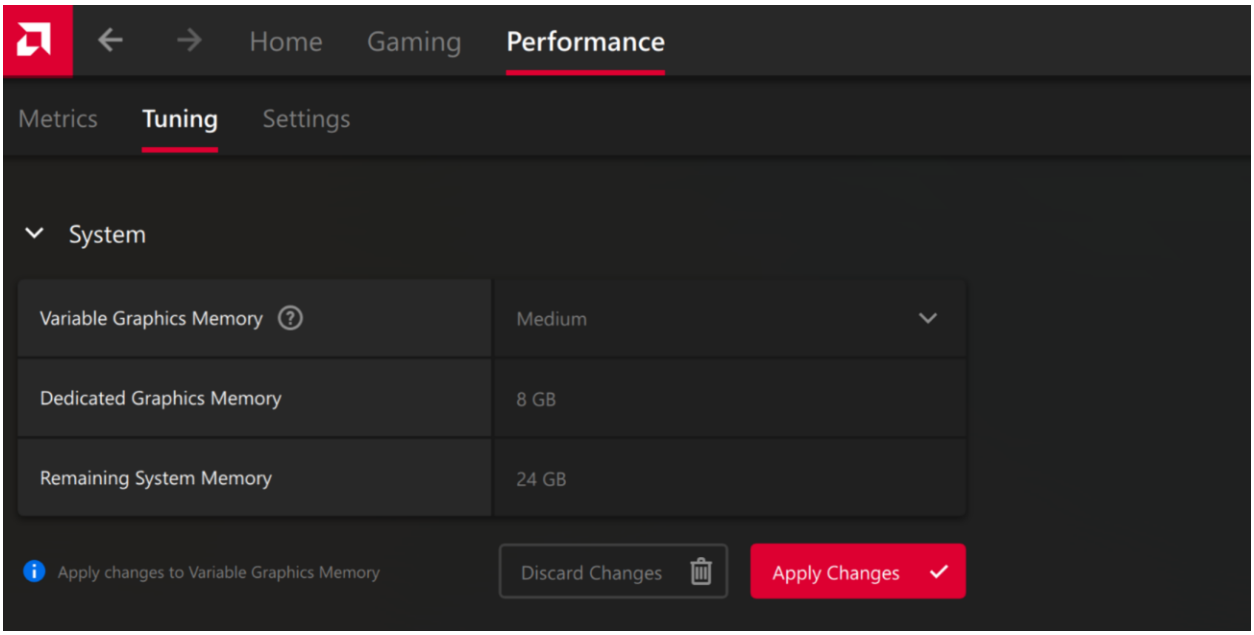


Image: *VGM options on a 32GB laptop. VGM High = 16GB dedicated graphics memory.*

Please make sure you are on the latest AMD Software: Adrenalin Edition™ driver update. AMD laptops powered by AMD Ryzen™ AI 300 series processors feature Variable Graphics Memory. AMD recommends turning on VGM for any LLM workloads to help token throughput and allow larger model sizes to run. A VGM setting of High is recommended. You can access the VGM options through the Performance > Tuning tab in AMD Software: Adrenalin Edition™.

You can download and install LM Studio from their website.

When running LLMs – please check “manually select parameters” and set the GPU Offload setting to MAX. AMD recommends Q4 K M quantization for everyday use and Q6 or Q8 for coding.

Experiencing AI locally on laptops powered by the AMD Ryzen™ AI MAX+ 395 processor is a great way for the power user to experience state-of-the-art AI models while having a portable, thin and light, gaming and productivity powerhouse.

Legal footnotes:

SHO26 - Testing as of March 2025 by AMD. All tests conducted on LM Studio 0.3.11. Llama.cpp runtime 1.18. Tokens/s and time to first token: Sustained performance average of multiple runs with specimen prompt "How long would it take for a ball dropped from 10 meter height to hit the ground?". Models tested: DeepSeek R1 Distill Qwen 1.5b Q4 K M, DeepSeek R1 Distill Qwen 7b Q4 K M, DeepSeek R1 Distill Qwen 8b Q4 K M, DeepSeek R1 Distill Qwen 14b Q4 K M, Phi 4 Mini Instruct 3.8b, Phi 4 Q4 K M, Llama 3.2 3b Instruct. AMD Ryzen™ AI MAX+ 395 on an ASUS ROG Flow Z13 with 64GB 8000 MT/s memory, Windows 11 Pro 24H2 and Adrenalin 25.3.1 WHQL. VGM = 32GB Intel Core Ultra 7 258V on an HP Zenbook S14 with 32GB 8533 MT/s memory, Windows 11 Pro 24H2 and Intel Graphics Driver 32.0.101.6559. Performance may vary.

SHO27 - Testing as of March 2025 by AMD. All tests conducted on LM Studio 0.3.13. Llama.cpp runtime 1.19.2. Time to first token using prompt "Describe this image" and picture taken from: www.loc.gov/item/2013648266/ Models tested: IBM Granite Vision 3.2 2b, Google Gemma 3 4b, Google Gemma 3 12b and Google Gemma 3 27b AMD Ryzen™ AI MAX+ 395 on an ASUS ROG Flow Z13 with 64GB 8000 MT/s memory, Windows 11 Pro 24H2 and Adrenalin 25.3.1 WHQL. VGM = 32GB Intel Core Ultra 7 258V on an HP Zenbook S14 with 32GB 8533 MT/s memory, Windows 11 Pro 24H2 and Intel Graphics Driver 32.0.101.6559. Links to third party sites are provided for convenience and unless explicitly stated, AMD is not responsible for the contents of such linked sites and no endorsement is implied. Performance may vary.

GD-97 - Links to third party sites are provided for convenience and unless explicitly stated, AMD is not responsible for the contents of such linked sites and no endorsement is implied.

GD-220e - Ryzen™ AI is defined as the combination of a dedicated AI engine, AMD Radeon™ graphics engine, and Ryzen processor cores that enable AI capabilities. OEM and ISV enablement is required, and certain AI features may not yet be optimized for Ryzen AI processors. Ryzen AI is compatible with: (a) AMD Ryzen 7040 and 8040 Series processors and Ryzen PRO 7040/8040 Series processors except Ryzen 5 7540U, Ryzen 5 8540U, Ryzen 3 7440U, and Ryzen 3 8440U processors; (b) AMD Ryzen AI 300 Series processors and AMD Ryzen AI PRO 300 Series processors; (c) all AMD Ryzen 8000G Series desktop processors except the Ryzen 5 8500G/GE and Ryzen 3 8300G/GE; (d) AMD Ryzen 200 Series processors and Ryzen PRO 200 Series processors except Ryzen 5 220 and Ryzen 3 210; and (e) AMD Ryzen AI Max Series processors and Ryzen AI PRO Max Series processors. Please check with your system manufacturer for feature availability prior to purchase