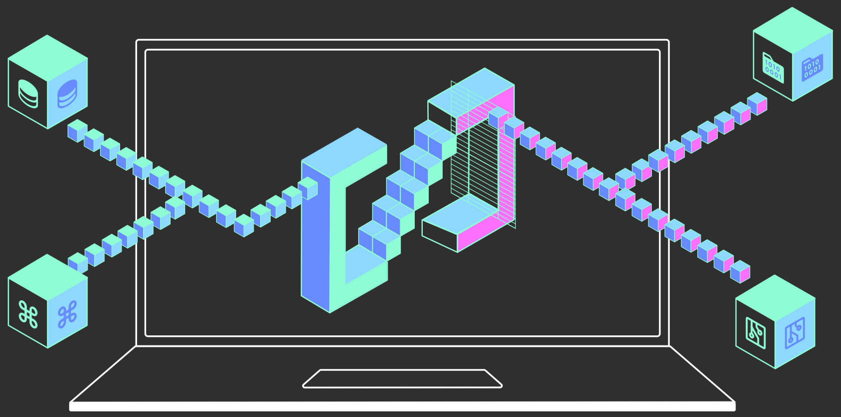
# The Future of AI Software: AnythingLLM

The future of personal computing is being redefined by the integration of AI directly into software, powered by dedicated AI accelerators called NPUs (Neural Processing Units). As Windows evolves to support a new generation of intelligent features, applications are becoming faster, more context-aware, and more capable of adapting to user needs in real time. Signal65 explores the key AI-enabled capabilities emerging within the Windows ecosystem, highlighting how NPUs are unlocking new levels of performance and efficiency across everyday tasks, enterprise workflows, and entirely new user experiences.

## What is AnythingLLM?

AnythingLLM is an all-in-one AI app that offers AI-powered tools that can run on local hardware rather than via cloud-based solutions, which are currently more mainstream and include apps such as ChatGPT. Users can easily download a variety of large language models and use them for research, queries, or just chatting, the same way one would with any LLM. AnythingLLM also has a compelling and simple agentic AI tool that allows users to create specific functions to streamline prompting and truly customize the LLM experience.



LLMs are often run in the cloud rather than on end-user hardware like phones and laptops since datacenters offer higher-end processors capable of handling the most robust models. However, edge AI is appealing for achieving lower latency, not being tethered to the internet, and offering better user privacy and data security. Running LLMs locally has gotten much more feasible as the latest processors come with faster NPUs capable of more AI number-crunching, especially in the case of the Snapdragon X Elite's Hexagon NPU, which is three times faster than the one found in the previous generation.

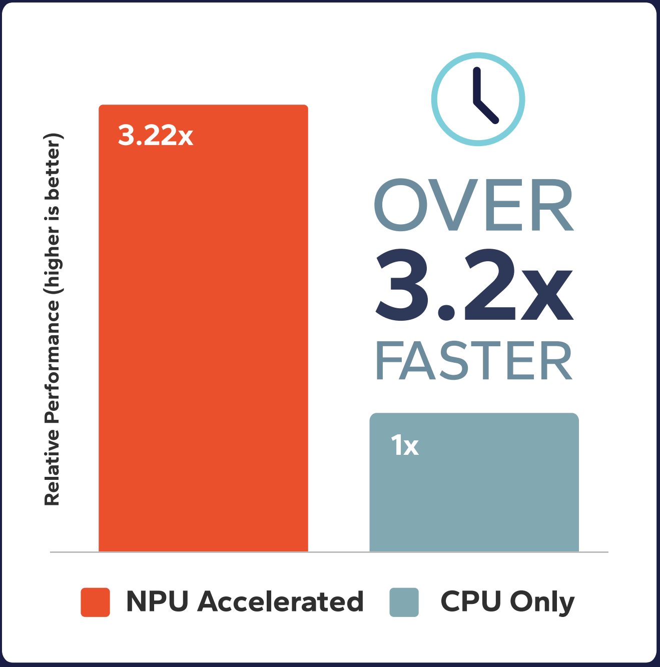
> “Local AI models are getting smaller and faster and hardware is becoming more commonplace. The future of everyday AI use will be all on-device and AnythingLLM will be there to assist and be that software layer that makes it accessible for everyone.
>
> –Timothy Carambat, AnythingLLM Founder

## NPU Utilization & Performance

In order to access NPU-accelerated LLMs on AnythingLLM, users will need to select the AnythingLLM NPU option in the LLM provider dropdown menu. Currently, there are four LLMs available to use: Microsoft Phi 3.5 Mini Instruct 4K, Llama 3.1 8B Chat 8K, and the 8K and 16K versions of Llama 3.2 8B Chat.

To test performance, we compared the NPU-accelerated version of Llama 3.1 8B Chat 8K to the version that runs on the CPU. We asked both models to check a major news website for info relating to the current stock market, to summarize that info, and to give an analysis. It took the CPU-reliant LLM four minutes and 44 seconds to get the first token and start writing a response, while the NPU-powered AI was able to answer in just a minute and 28 seconds.

Not only was the NPU faster at delivering a response, but in this case the quality of the answer was better too. NPU-driven Llama told us about the overall trend of the market, recent events, how certain indexes are doing, and the outlook that analysts have. By contrast, the normal version of Llama 3.1 8B merely told us the price changes of the three major US indexes, and general info about the state of trade and the economy, which wasn't particularly insightful. And while the NPU-powered version of Llama talked about specific events such as Tencent boosting its gaming revenue by 13%, the normal LLM only vaguely referenced the US-China trade war, and didn't mention any news in particular.



Today, over 20 applications leverage Qualcomm's Hexagon NPU to deliver enhanced performance, enable entirely new features, and improve system efficiency. Collectively, these applications offer over 50 unique AI-powered capabilities spanning diverse use cases, from real-time video conferencing effects and advanced photo editing tools to local LLM implementations that bring sophisticated AI experiences directly to users' hardware without cloud dependencies.

This ecosystem demonstrates how NPU acceleration is transforming software across categories, enabling developers to implement AI features that were previously impractical due to performance or power constraints. As more applications adopt NPU optimization, users benefit from faster processing, longer battery life, and AI capabilities that respond instantly to their needs.

## Looking Forward

In respect to both speed and quality, the NPU appears to be the superior option when it comes to LLMs in apps like AnythingLLM; additionally, we're also getting a look at how the NPU will shape the landscape of AI software in the future. Much like how GPUs evolved from specialized graphics processors to general-purpose computing accelerators, NPUs are positioned to become the dedicated platform for AI workloads, delivering superior performance compared to both CPUs and GPUs for machine learning tasks.

Through continued investment in both hardware and software AI capabilities, Qualcomm and Microsoft are establishing NPUs as the foundation for local AI processing, enabling more responsive, efficient, and capable applications across the Windows ecosystem.

# Over 50 NPU-powered AI Experiences on Snapdragon X Series Processors

| Creator Apps | AI Experience |
|---|---|
| **Adobe Premiere Pro** | • Audio Category Tagger to sort different audio clips into categories like ambience or dialog<br>• Scene Edit Detection automatically labels cuts in raw footage for easier editing<br>• Text-Based Editing builds a transcript for a video, and editing the transcript instantly edits the video for rough cuts |
| **Automatic1111** | Image generation from text using Stable Diffusion and ability to customize parameters |
| **Blender+ControlNet** | 3D scene to 2D image generation via tools like Automatic1111 |
| **Copilot+** | • Image generation and photo editing using AI-powered tools like generative fill<br>• Easy step retracing with Windows Recall<br>• Improved gaming performance and visual quality with Super Resolution<br>• Video conferencing features like real-time translation, auto framing, portrait lighting, and more |
| **DaVinci Resolve** | • AI-accelerated Magic Mask tool for objects and people<br>• Better resolution upscaling during rendering |
| **Djay Pro** | Separating different instruments and vocals with Neural Mix, and syncing different songs with varying rhythms together with BeatGrid |
| **Gigapixel AI** | Crisp upscaling for photos originally taken at low resolution |
| **GIMP+SD** | Image generation from text using Stable Diffusion |
| **Luminar Neo** | Photo editing with AI-assisted sharpening effects and resolution upscaling |
| **Moises Live** | • Instrument and vocal separation for music editing<br>• Enhanced performance compared to running on the CPU |
| **Enterprise Apps** | **AI Experience** |
| **Camo Studio** | On-the-fly video effects like auto-framing, virtual green screen, and blurred background |
| **Copilot+** | • Image generation and photo editing using AI-powered tools like generative fill<br>• Easy step retracing with Windows Recall<br>• Improved gaming performance and visual quality with Super Resolution<br>• Video conferencing features like real-time translation, auto framing, portrait lighting, and more |
| **Dynamo AI** | Guardrails for AI provided through organizations to prevent misuse |
| **McAfee** | AI-powered detection of deepfaked audio |
| **Zoom** | Virtual background replacement and portrait lighting for video conferencing |
| **Productivity Apps** | **AI Experience** |
| **AnythingLLM** | • Easy setup for small and powerful Microsoft and Meta LLMs with long context windows<br>• Useful LLM features like automation, RAG, and inferencing” |
| **Copilot+** | • Image generation and photo editing using AI-powered tools like generative fill<br>• Easy step retracing with Windows Recall<br>• Improved gaming performance and visual quality with Super Resolution<br>• Video conferencing features like real-time translation, auto framing, portrait lighting, and more |
| **Liquid Text** | Fast annotation of documents using AI |
| **LMStudio** | Run LLMs locally and configure them to your liking |

Visit Qualcomm for more info: https://www.qualcomm.com/snapdragon/laptops-and-tablets/npu-powered-ai-experiences