SHI | NEXT-GEN DEVICE LABS

BenchSmart PERFORMANCE REPORT

SYSTEM INFORMATION

NPU	Not Present
NPU TOPS	0
RAM	64 GB
Storage	512 GB SSD
Display	16”
Operating System	Windows 11 Pro
Weight	3.97 lbs
Retail Cost	$2,050.00 USD

Benchmarking, the SHI way.

Facing the challenge of evaluating AI PCs and edge devices, SHI developed BenchSmart, a proprietary tool that runs locally on the device and provides deep insights into system behavior across various AI workloads.

With BenchSmart, SHI analyzes metrics like CPU, GPU, and NPU utilization, enabling precise comparisons across platforms. This empowers businesses to make informed decisions, validating OEM claims and refining product selection criteria with comprehensive reports and historical trends.

SHI’s Next-Gen Device Lab

SHI’s Next-Gen Device Lab, located at the End-User Integration Center (EIC) in Piscataway, NJ, offers immediate access to the latest PCs, mobile devices, and AR/VR solutions from top OEMs.

It features SHI’s proprietary BenchSmart AI Device Performance application and showcases a variety of demos, including application development tools, security solutions, AI assistants, VR desktops, and more. The lab allows visitors to experience how software from leading and emerging partners performs on next-generation devices. This hands-on environment helps your organization evaluate your technology needs and maximize your IT investments.

Want to visit our lab with your team?

Request your lab tour

BenchSmart
AI MODALITY EXPLAINER

AI MODALITY	REAL-WORLD USE CASES
Text The text performance module uses large language models (LLMs) loaded locally on the device to simulate actions such as chatbots. A text-to-text model, like an LLM, takes input text (e.g., a question or prompt) and generates relevant output text (e.g., an answer or summary).	• Chatbots • Translation • Summarization • Question & answer
Speech A speech-to-text model (also known as automatic speech recognition) converts audio input into written text. A text-to-speech model synthesizes speech from textual input.	• Voice assistants • Transcription services • Accessibility tools • Screen readers • Voice-over applications
Image A text-to-image model, like a diffusion model, generates images from text descriptions. It interprets the input prompt and gradually creates a visual representation. An image-to-text model is a type of multimodal model that interprets visual input (images) and generates descriptive text.	• Art • Design • Creation content generation • Image captioning • Visual question answering • Visual inspection for quality
Video A text-to-video model converts a written description (prompt) into a video that visually represents the scene. A video-to-text model generates captions, summaries, or transcripts from a video.	• Marketing content generation • Educational explainers • Simulations • Storytelling • Scene summarization • Captioning • Extracting insights from surveillance or instructional videos

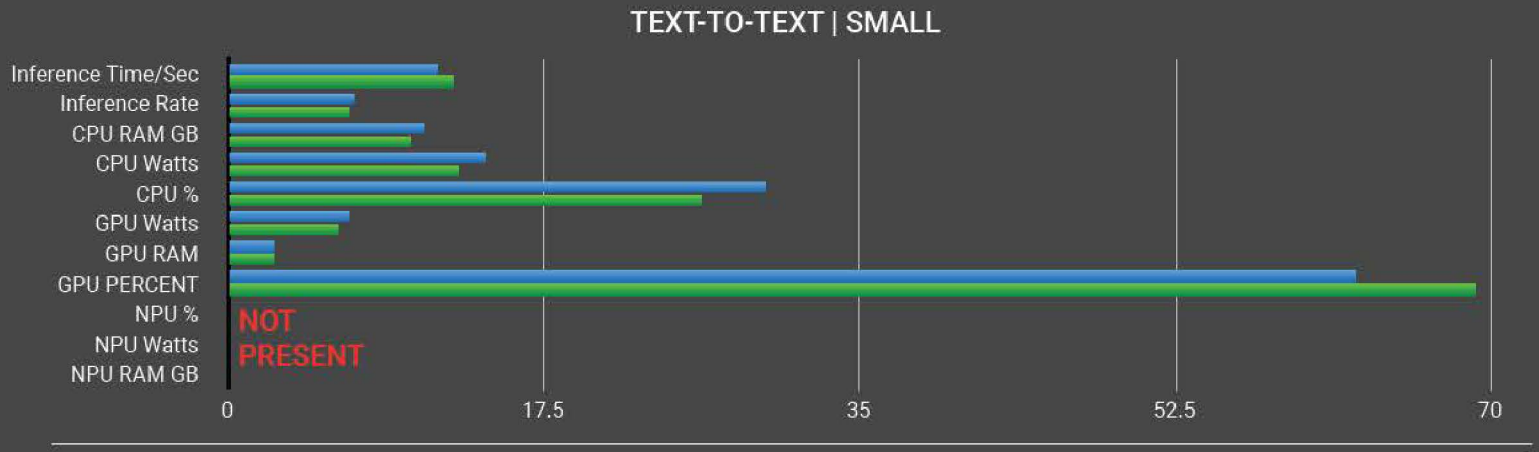
Text to Text

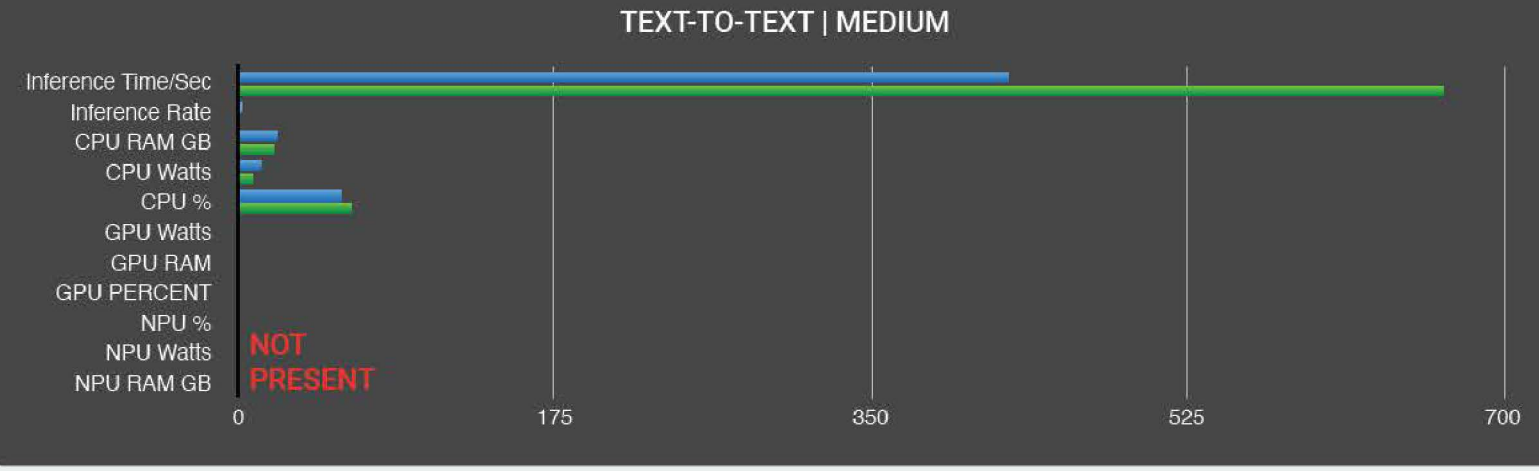
Test Type	Size	Infer. Time (sec)	Infer. Rate (t/sec)	CPU RAM (GB)	CPU Power (W)	CPU (%)	GPU Power (W)	GPU RAM (GB)	GPU (%)	NPU Power (W)	NPU RAM (GB)	NPU (%)
Power	S	11.6	7.1	10.9	14.3	29.9	6.8	2.6	62.6	-	-	-
Battery	S	12.5	6.7	10.2	12.8	26.3	6.1	2.6	69.2	-	-	-
Power	M	426	1.6	20.8	12.8	56.3	0	0.6	0.8	-	-	-
Battery	M	667	1.1	20.5	7.6	62.3	0.01	0.6	1.2	-	-	-

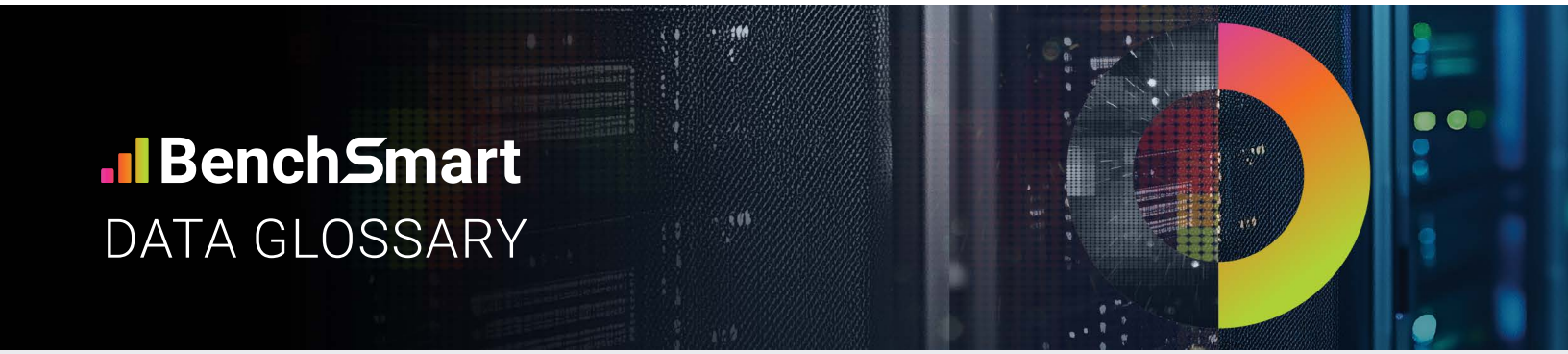
KEY

On Power On Battery

TIP: SHORTER IS BETTER | X-SCALE MAY VARY







BenchSmart
DATA GLOSSARY

TOKENS

A token is a unit of text — it could be a word, part of a word, punctuation, or even whitespace. For example:

- "Hello" → **1 token**
- "unbelievable" → **3 tokens:** "un", "believ", "able"

INFERENCE RATES & TIMES

Average inference time refers to the typical duration it takes for a trained model to produce a prediction from a given input. It's a critical metric for evaluating performance, especially in real-time or resource-constrained environments.

Average inference rate (tokens/second) is a key metric used to evaluate how quickly a language model can process and generate tokens during inference. It's especially important for real-time applications like chatbots, code assistants, and voice interfaces.

LATENCY

AI latency refers to the time delay between when an AI system receives input and when it produces output. It's a key factor in how fast and responsive an AI application feels to users.

Why latency matters

- Real-time performance: In applications like voice assistants, fraud detection, or autonomous driving, even small delays can cause problems.
- User experience: Faster responses make AI tools feel smoother and more reliable.
- Scalability: Low latency helps systems handle more users and data without slowing down.

What affects AI latency

- Model size: Bigger models often take longer to process input.
- Hardware: Specialized chips like GPUs, TPUs, or NPUs can speed things up.
- Data access: Slow databases or networks can bottleneck performance.
- Token generation: In language models, generating fewer tokens can reduce latency.

To get started, schedule a discovery assessment with SHI's experts.