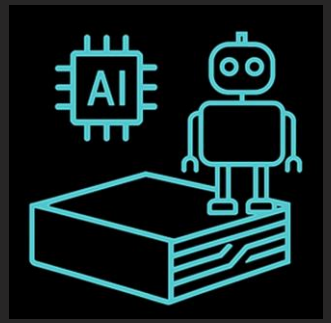
GIGABYTE

AI TOP ATOM

Sales Kits


GIGABYTE
Accelerated by NVIDIA
GIGABYTE
AORUS
GIGABYTE
GIGABYTE AI TOP ATOM
Your Personal AI Supercomputer

Why AI TOP ATOM?



AI is more Intelligent, but more complex

- Reasoning AI & AI Agents with multiple models
- More compute needed during inference

Local Developers

- Laptop / PC memory or software stack inadequate
- Forced to rely on cloud or datacenter
- Limited on-site development

GIGABYTE AI TOP ATOM

- NVIDIA® Grace Blackwell GB10 Superchip 1 petaflop + 128GB Memory
- Full NVIDIA AI software stack
- Develop + run at the desk

Design Concept

- Compact and energy-efficient form factor
- Flow-inspired design with wave-like grille that reflects continuous data movement
- Matte black finish and compact form for a clean, professional look
- Optimized thermal airflow aligned with the exterior’s fluid form
- Expresses GIGABYTE’s vision: performance shaped by intelligent design

What makes us different : AI TOP Utility

To further enhance the user experience, AI TOP ATOM also provides access to GIGABYTE’ s exclusive AI TOP Utility, a downloadable software suite that helps users quickly perform model training, inference, and deployment locally. The tool integrates optimized workflows, ready-to-use templates, and model conversion features—making the development process more intuitive and efficient.

AI TOP UTILITY

Monitor Hardware status + Training quality with user-friendly GUI

AI TOP UTILITY 4.1

Dataset Creating → **Model Download** → **LLM/LMM Fine-tuning (1PC or 2PC)** → **Validation** → **Model converter** → **Inference**

Automatically create dataset for LLM finetuning from pure text

Finetuning (1 PC)
Multinode finetuning (2PC)

Validating the fine-tuned models

RAG

Inference the models

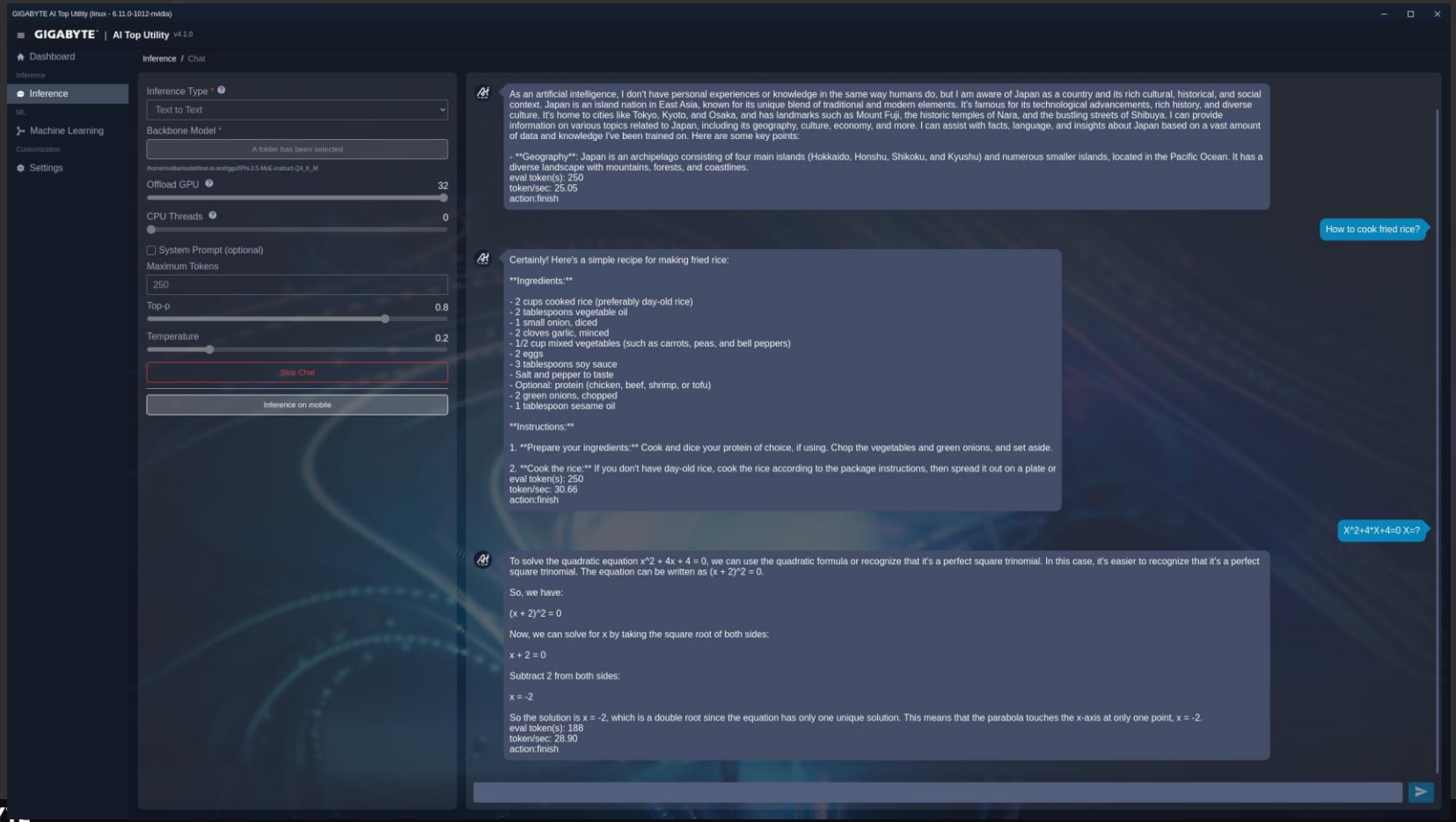
Machine Learning

Machine Learning :
1. image classification
2. object detection
3. image segmentation
4. OCR (Optical Character Recognition)

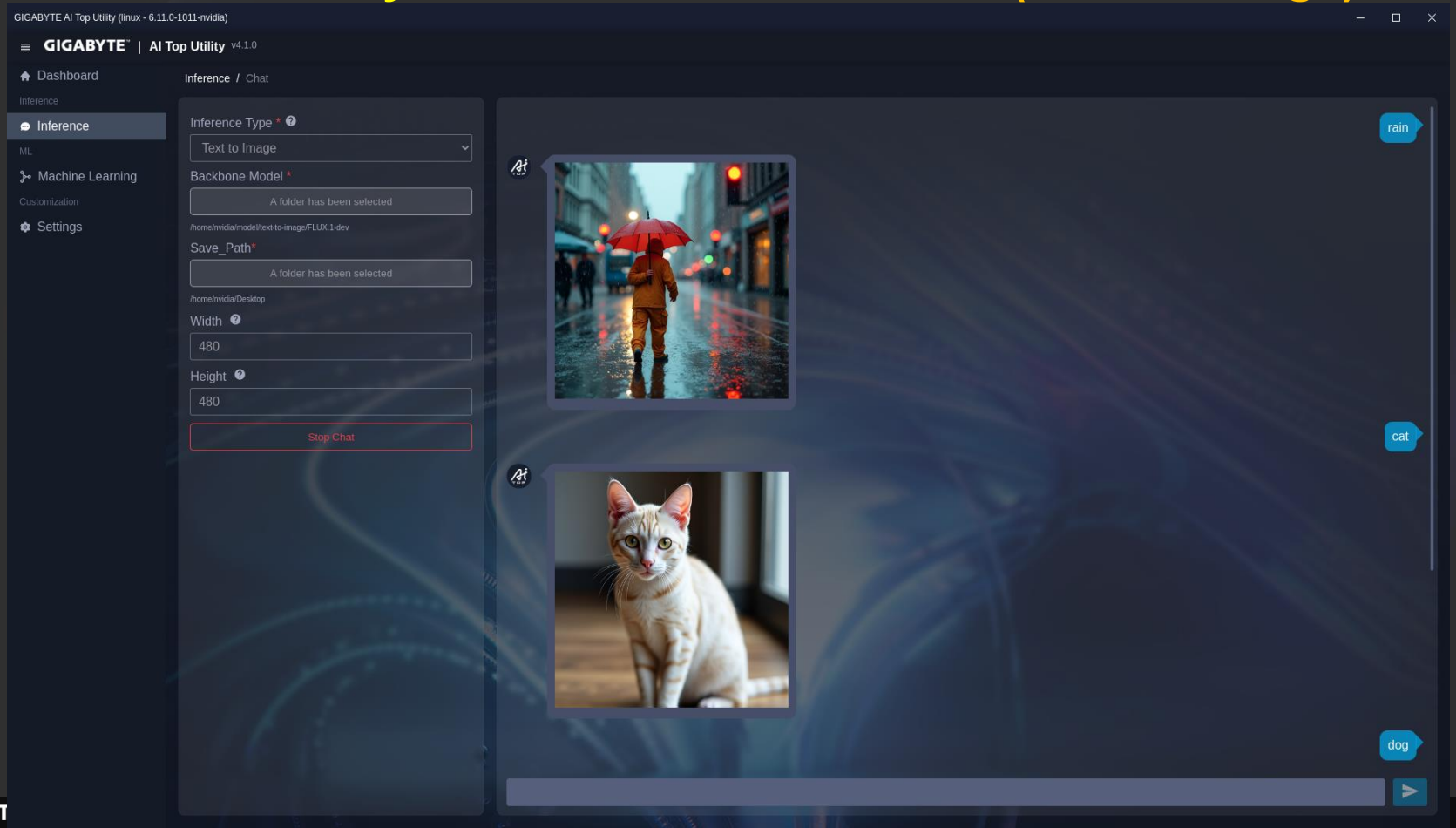
Multimodal Retrieval-augmented generation for text/audio/image/video datasets

AORUS

AI TOP Utility Atom Version : Inference (Text-to-Text)



AI TOP Utility Atom Version : Inference (Text-to-Image)



NVFP4 Advantage: Text to Image

TensorRT BF16 | TensorRT FP4

An old steam locomotive chugging through a mountainous landscape, billowing clouds of smoke, with a classic hand-painted style

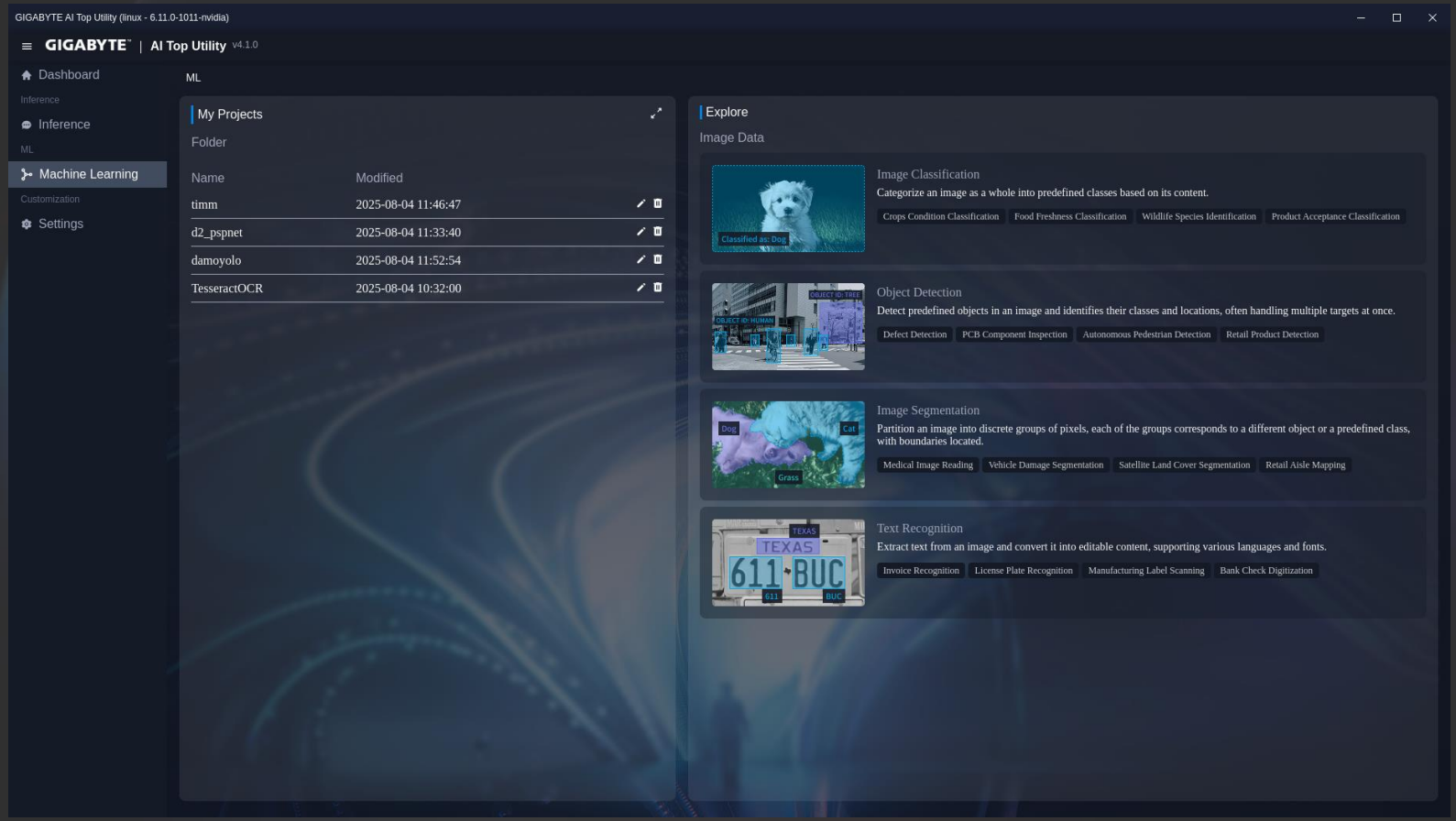
Precision	Default mode
FP16	39.3 GB
BF16	35.7 GB
FP8	24.6 GB
FP4	21.67 GB

Model	5090 fp16*	5090 fp8	5090 fp4	4090 fp8*
FLUX.1-dev (w/ 30 diffusion steps)	10930.96ms	6680.93ms	3852.75ms	10620.37ms
FLUX.1-schnell(w/ 4 diffusion steps)	4427.43ms	912.53ms	590.56ms	3385.43ms

Table 2. E2E inference pipeline time on RTX GPUs (* needs to run with low-VRAM mode to fit into GPU)

來源

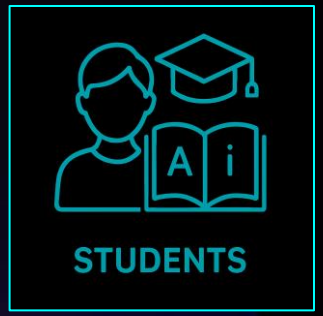
AI TOP Utility Atom Version : Machine Learning



Connect. Expand. Accelerate

Stack 2 AI TOP ATOMs via NVIDIA® ConnextX®-7 SmartNIC
for Larger AI models & Performance
200B parameter to 405B parameter

Target Audience



	Enterprise Developers	Researchers	Inception / Startups	Enthusiasts	Students
Description	Professional AI Software developers in companies	AI Researchers, grad-students, professors	Professionals who develop data science solutions as their primary role	Hobbyist developers passionate about AI; often have full-time jobs but develop in their free time	Individuals in secondary or higher education studying AI/data science
Pain Points	IT procurement process / policies	University contracts	Price / Prefer NET 30	Availability	Price / Availability
Budget	Business case dependent	Business case dependent / low price purchases avoid bid/ budget process	Business case dependent. Price sensitive in early stages.	$3-5K	<$3000
Buying preference	Partners & Online	Partners & Online	Partners & Online	Online	Online & Retail

TARGET USE CASES

Open Source LLM

Pre-trained LLM

General Models

Domain data

Task Specialization
Data Security

AI Top Utility

Fine-tuned LLM

Specialized Models

AI is not a one-size-fits-all solution
It must be fine-tuned and optimized for specific business needs.

Application

- Chatbot
- Finance
- Robot
- Legal
- Medical
- AI Agent
- Media
- Education

AI needs to be optimized for specific needs.

Specifications

Architecture	NVIDIA® Grace Blackwell
GPU	NVIDIA® Blackwell Architecture
CPU	20 core Arm, 10 Cortex-X925 + 10 Cortex-A725 Arm
CUDA Cores	NVIDIA® Blackwell Generation
Tensor Cores	5th Generation
RT Cores	4th Generation
Tensor Performance[1]	1 petaFLOP AI performance
System Memory	128 GB LPDDR5x, unified system memory
Memory Interface	256-bit
Memory Bandwidth	273 GB/s
Storage	4 TB Gen5/4TB Gen4/1TB Gen4 NVME.M2 with self-encryption
USB	4x USB TypeC
Ethernet	1x RJ-45 connector , 10 GbE
NIC	ConnectX-7 Smart NIC
Wi-Fi	WiFi 7
Bluetooth	BT 5.3
Audio-output	HDMI multichannel audio output
Power Consumption	240W
Display Connectors	1x HDMI 2.1a
NVENC | NVDEC	1x | 1x
OS	NVIDIA DGX™ OS
System Dimensions	150 mm L x 150 mm W x 50.5 mm H
System Weight	1.6 kg
Suggest MSRP	USD $3999 - $4599

From Compact to Extreme AI Power

AI TOP ATOM

1 petaFLOP

100B-200B

Finetune Inference

Compact AI platform for entry-level developers and edge computing.

AI TOP 100

3000 TOPS

110B-200B

Finetune Inference

Mainstream AI system for developers & creators

AI TOP 500

3000 TOPS

405B-685B

Finetune Inference

Extreme-performance AI server for heavy-duty workloads

Connectivity

ConnectX-7
(2x QSFP, 1 required to connect to other AI TOP ATOM)

USB Type-C
(Up to 40GB/s)

Power Button

Network
(RJ-45 Connector, 10GbE)

Power
(USB-C)

HDMI 2.1a


TOP