

GIGABYTE

AI TOP ATOM

Sales Kits



GIGABYTE™

Accelerated by  NVIDIA



GIGABYTE AI TOP ATOM

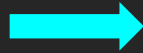
Your Personal AI Supercomputer

Why AI TOP ATOM?



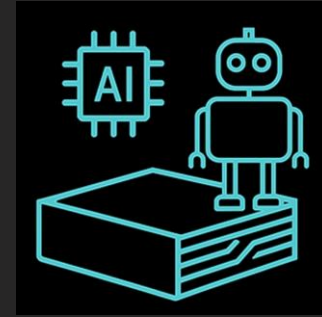
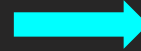
AI is more Intelligent, but
more complex

- 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



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

GIGABYTE AI Top Utility (linux - 6.11.0-1012-nvidia)

GIGABYTE

AI Top Utility

v4.1.0

Dashboard

Inference / Chat

Inference

Machine Learning

Settings

Inference Type

Text to Text

Backbone Model

A folder has been selected

RemoteModelPath: /root/.cache/nvidia/ai-top-utility/0.1.0-1012-nvidia

Offload GPU

32

CPU Threads

0

☐ System Prompt (optional)

Maximum Tokens

250

Top-p

0.8

Temperature

0.2

Stop Chat

Inference on mobile

As an artificial intelligence, I don't have personal experiences or knowledge in the same way humans do, but I am aware of Japan as a country and its rich cultural, historical, and social context. Japan is an island nation in East Asia, known for its unique blend of traditional and modern elements. It's famous for its technological advancements, rich history, and diverse culture. It's home to cities like Tokyo, Kyoto, and Osaka, and has landmarks such as Mount Fuji, the historic temples of Nara, and the bustling streets of Shibuya. I can provide information on various topics related to Japan, including its geography, culture, economy, and more. I can assist with facts, language, and insights about Japan based on a vast amount of data and knowledge I've been trained on. Here are some key points:

- "Geography": Japan is an archipelago consisting of four main islands (Hokkaido, Honshu, Shikoku, and Kyushu) and numerous smaller islands, located in the Pacific Ocean. It has a diverse landscape with mountains, forests, and coastlines.

eval token(s): 250
token/sec: 25.05
action: finish

Certainly! Here's a simple recipe for making fried rice:

****Ingredients:****

- 2 cups cooked rice (preferably day-old rice)

- 2 tablespoons vegetable oil

- 1 small onion, diced

- 2 cloves garlic, minced

- 1/2 cup mixed vegetables (such as carrots, peas, and bell peppers)

- 2 eggs

- 3 tablespoons soy sauce

- Salt and pepper to taste

- Optional: protein (chicken, beef, shrimp, or tofu)

- 2 green onions, chopped

- 1 tablespoon sesame oil

****Instructions:****

1. ****Prepare your ingredients:**** Cook and dice your protein of choice, if using. Chop the vegetables and green onions, and set aside.

2. ****Cook the rice:**** If you don't have day-old rice, cook the rice according to the package instructions, then spread it out on a plate or

eval token(s): 250
token/sec: 30.66
action: finish

To solve the quadratic equation $x^2 + 4x + 4 = 0$, we can use the quadratic formula or recognize that it's a perfect square trinomial. In this case, it's easier to recognize that it's a perfect square trinomial. The equation can be written as $(x + 2)^2 = 0$.

So, we have:

$(x + 2)^2 = 0$

Now, we can solve for x by taking the square root of both sides:

$x + 2 = 0$

Subtract 2 from both sides:

$x = -2$

So the solution is $x = -2$, which is a double root since the equation has only one unique solution. This means that the parabola touches the x-axis at only one point, $x = -2$.

eval token(s): 188
token/sec: 28.90
action: finish

How to cook fried rice?

$x^2 + 4x + 4 = 0$ $x = ?$

GIGABYTE

9

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

GIGABYTE AI Top Utility (linux - 6.11.0-1011-nvidia)

≡ **GIGABYTE** | AI Top Utility v4.1.0

🏠 Dashboard

Inference

ML

🔗 Machine Learning

Customization

⚙️ Settings

Inference / Chat

Inference Type * ⓘ
Text to Image

Backbone Model *
A folder has been selected

/home/nvidia/model/text-to-image/FLUX.1-dev

Save_Path*
A folder has been selected

/home/nvidia/Desktop

Width ⓘ
480

Height ⓘ
480

Stop Chat

rain

cat

dog

AI TOP

AI TOP

AI TOP

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)



AI TOP Utility Atom Version : Machine Learning

GIGABYTE AI Top Utility (linux - 6.11.0-1011-nvidia)

≡

GIGABYTE™

|

AI Top Utility v4.1.0

Dashboard

Inference

ML

Machine Learning

Customization

Settings

ML


My Projects

Folder

Name	Modified	
timmm	2025-08-04 11:46:47	✎ □
d2_pspnet	2025-08-04 11:33:40	✎ □
damoyolo	2025-08-04 11:52:54	✎ □
TesseractOCR	2025-08-04 10:32:00	✎ □

Explore

Image Data



Classified as: Dog

Image Classification


Categorize an image as a whole into predefined classes based on its content.

Crops Condition Classification

Food Freshness Classification

Wildlife Species Identification

Product Acceptance Classification



OBJECT: HUMAN

Object Detection


Detect predefined objects in an image and identifies their classes and locations, often handling multiple targets at once.

Defect Detection

PCB Component Inspection

Autonomous Pedestrian Detection

Retail Product Detection



Cat

Image Segmentation

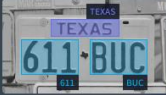
Partition an image into discrete groups of pixels, each of the groups corresponds to a different object or a predefined class, with boundaries located.

Medical Image Reading

Vehicle Damage Segmentation

Satellite Land Cover Segmentation

Retail Aisle Mapping



TEXAS 611-BUC

Text Recognition

Extract text from an image and convert it into editable content, supporting various languages and fonts.

Invoice Recognition

License Plate Recognition

Manufacturing Label Scanning

Bank Check Digitization

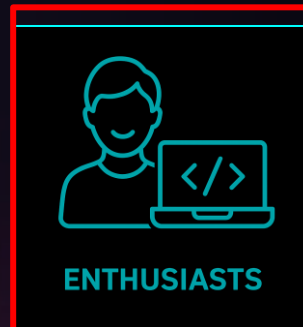
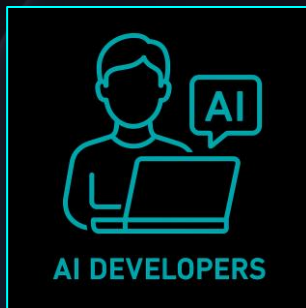
Connect. Expand. Accelerate



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

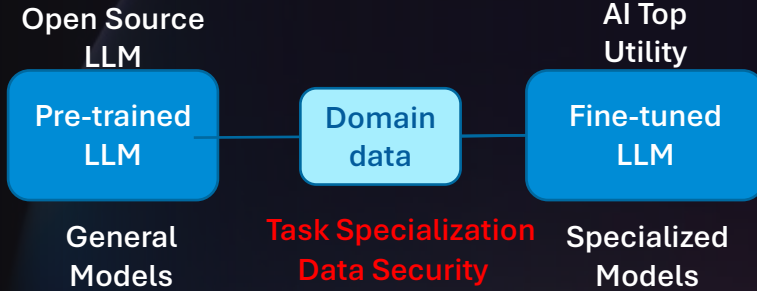


Target Audience



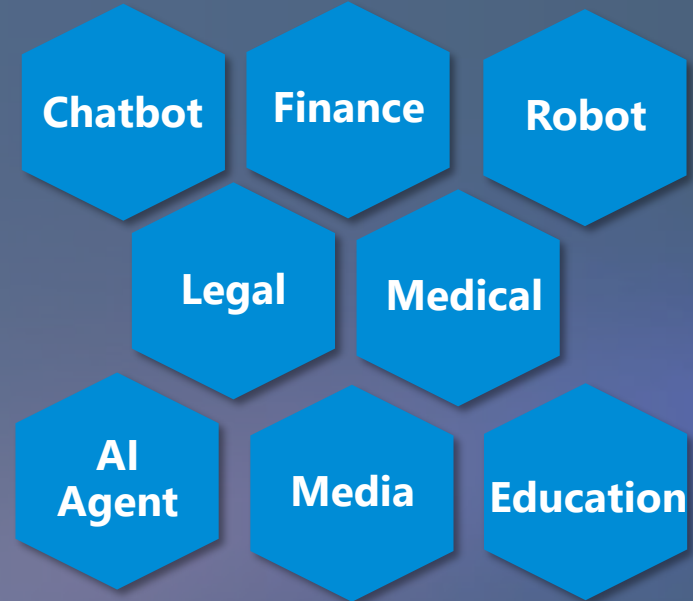
	Enterprise Developers	Researchers	Inception / Startups	Enthusiasts	Students
Description	Professional AI Software developers in companies	AI Researchers, grad-students, professors	Professionals who develop data science solutions as their primary role	Hobbyist developers passionate about AI; often have 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



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

Application



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



Power Button

**Power
(USB-C)**

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

HDMI 2.1a

**Network
(RJ-45 Connector, 10GbE)**

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

