AMD Radeon™ AI PRO R9700 Getting Started

AMD Radeon™ AI PRO R9700 Overview

AMD ROCm™ + PyTorch® Environment Setup Guide

AMD Radeon™ AI PRO R9700 Overview

Architecture and Memory

AMD RADEON AI PRO R9700

Built to accelerate medium to large AI models directly on local workstations, the AMD Radeon™ AI PRO R9700 combines 32GB of high-speed GDDR6 memory, advanced RDNA™ 4 architecture, and broad support for modern AI data types - alongside compatibility with the AMD ROCm™ open software platform optimized for Linux.

With 64 Compute Units, 128 second-generation AI accelerators, and support for FP8, FP16, and INT8 precision, the R9700 delivers ultimate versatility and value for today's most demanding AI workloads. Its robust scalability and multi-GPU support through ROCm enable open AI ecosystems developers to push beyond single-card limitations, providing cost-effective acceleration for large language models across multi-GPU deployments on Linux.

Key Highlights

- Powered by AMD RDNA™ 4 Architecture
- 2nd gen AI Accelerators with up to 4X faster performance
- 32GB of video memory for larger projects
- PCIe Gen 5 to unlock data intensive transfers from CPU memory
- DisplayPort™ 2.1 for multi-monitor workstations

Accelerating the latest workflows

- AI Inference using large language and diffusion models
- AI model finetuning
- Video content creation and streaming

RX-1143

AMD RDNA™ 4 Architecture

The Radeon AI PRO R9700 is powered by AMD RDNA™ 4 architecture, which offers up to 2x throughput per Compute Unit for FP16 datatypes and 4x for INT8 datatypes. It also introduces support for FP8, which is rapidly gaining popularity in AI workloads.

RX-1143

32GB
GDDR6 VRAM

Large VRAM Capacity

With 32GB of GDDR6 memory, the AMD Radeon™ AI PRO R9700 enables faster, more reliable performance for large models like Mistral 3.1 or Qwen 32B that exceed the limits of 16GB GPUs. By keeping models entirely in local memory, it avoids system memory bottlenecks. Compared to 16GB-class GPUs, the R9700 delivers up to 5x better performance in demanding AI workloads such as LLM inference and multimodal generation.

RPW-495

AMD Radeon™ AI PRO R9700 Overview

Scalability and Value

Typical VRAM Usage by Popular Models

RPW-496

SCALABLE
PERFORMANCE

Ultimate Value and Scalability

The Radeon™ AI PRO R9700 is designed to deliver exceptional value for demanding AI workloads in the 32GB GPU class. Compared to professional workstation cards from the competition, it provides 32GB of VRAM in a far more accessible price bracket.

Against consumer-oriented high-VRAM GPUs, the Radeon™ AI PRO R9700 stands out with practical scalability, enabling new use cases in high-concurrency environments. Thanks to its compact form factor and PCIe® 5.0 support, it integrates seamlessly into multi-GPU workstations, where multiple Radeon™ AI PRO R9700 cards can be deployed to tackle even larger AI models - an ideal solution for developers and enterprises looking to scale local AI services without relying on cloud infrastructure.

RPW-506

Scalability

4 x GPUs
AMD Radeon AI PRO R9700
DeepSeek-R1-Distill-Qwen-32B

Instances	Output tok/sec	Instances	Output tok/sec
1	40	40	1272
4	189	60	1606
6	262	80	1950
10	441	100	2214
20	762	120	2457

RPW-499

AMD ROCm™ + PyTorch® Environment Setup Guide

Initial Installation Walkthrough

About this Guide

This Guide covers deployment of AMD ROCm 6.4.2 and PyTorch 2.8.0 (Stable) AI Stack for Ubuntu 22.04/24.04 on Radeon AI PRO R9700 (RDNA4) and Radeon PRO W7000 series (RDNA3) GPUs, including Docker environment support.

Platforms supported by this Guide

OS	Ubuntu 22.04.x, Ubuntu 24.04.x
Kernels tested	5.15.0-157 (22.04.5) • 6.8.0-85 (24.04.3)
GPUs	AMD Radeon™ PRO AI R9700 (AMD RDNA™ 4)
Docker support	AMD Radeon™ PRO AI R9700
ROCm	6.4.2
PyTorch	2.8.0+rocm6.4
Transformers	4.57.0

AMD RADEON AI PRO

AMD ROCm

System Preparation

Install Ubuntu 22.04.5 LTS or Ubuntu 24.04.3 LTS (Server or Desktop version).

Recommendations:

- Use a fresh Ubuntu installation.
  - If you have an older version of ROCm installed, follow our uninstallation method.
  - If you have an older version of PyTorch installed, you can uninstall it with:
    → ***pip3 uninstall torch torchvision***
- Assign the full storage capacity during installation.
- Install OpenSSH for remote SSH management.

Download the Script from the Repository

→ ***wget*** *https://raw.githubusercontent.com/JoergR75/rocm-6.4.2-pytorch-2.8.0-rdna4-docker-automated-deployment/refs/heads/main/script_module_ROCm_642_Ubuntu_22.04-24.04_pytorch_280stable_server.py*

```
ai1@ws1:~$ wget https://raw.githubusercontent.com/JoergR75/rocm-6.4.2-pytorch-2.8.0-rdna4-docker-automated-deployment/refs/heads/main/script_module_ROCm_642_Ubuntu_22.04-24.04_pytorch_280stable_server.py
--2025-10-13 16:06:52--  https://raw.githubusercontent.com/JoergR75/rocm-6.4.2-pytorch-2.8.0-rdna4-docker-automated-deployment/refs/heads/main/script_module_ROCm_642_Ubuntu_22.04-24.04_pytorch_280stable_server.py
Resolving raw.githubusercontent.com (raw.githubusercontent.com)... 2606:50c0:8000::154, 2606:50c0:8003::154, 2606:50c0:8002::154, ...
Connecting to raw.githubusercontent.com (raw.githubusercontent.com)|2606:50c0:8000::154|:443... connected.
HTTP request sent, awaiting response... 200 OK
Length: 12882 (13K) [text/plain]
Saving to: 'script_module_ROCm_642_Ubuntu_22.04-24.04_pytorch_280stable_server.py.1'

script_module_ROCm_642_Ubunt 100%[=========================================>]  12.58K  --.-KB/s    in 0.001s

2025-10-13 16:06:52 (9.37 MB/s) - 'script_module_ROCm_642_Ubuntu_22.04-24.04_pytorch_280stable_server.py.1' saved [12882/12882]
```

AMD ROCm™ + PyTorch® Environment Setup Guide

Initial Installation Walkthrough

Run the Installer

Note: Entering the user password may be required.

The installation takes ~15 minutes depending on internet speed and hardware performance.

→ ***bash script_module_ROCm_642_Ubuntu_22.04-24.04_pytorch_280stable_server.py***

```
ROCm 6.4.2 + OpenCL 2.x + PyTorch 2.8.0 (Stable) + Transformers + Docker Setup
Compatible with Ubuntu 22.04.x and 24.04.x (Desktop & Server) – Ubuntu 20.04.x is no longer supported


Ubuntu OS Update ...

Hit:1 http://de.archive.ubuntu.com/ubuntu noble InRelease
Hit:2 http://security.ubuntu.com/ubuntu noble-security InRelease
Hit:3 http://de.archive.ubuntu.com/ubuntu noble-updates InRelease
Hit:4 http://de.archive.ubuntu.com/ubuntu noble-backports InRelease
Reading package lists... Done
Reading package lists... Done
Building dependency tree... Done
Reading state information... Done
Calculating upgrade... Done
```

Testing ROCm + PyTorch

After rebooting, verify your setup:

This script creates a simple diagnostic python file (test.py) to verify that PyTorch with ROCm support is correctly installed and working.

What it does:

- Prints the PyTorch version and ROCm version.
- Checks if ROCm is available and how many GPUs are detected.
- Displays the name of the first GPU (if available).
- Creates two random 3×3 tensors directly on the GPU (if available).
- Performs a simple tensor addition operation on the GPU.
- Prints confirmation that the operation was successful and shows the result.

Example usage: → ***python3 test.py***

```
ai1@ws1:~$ python3 test.py
PyTorch version: 2.8.0+rocm6.4
ROCm version: 6.4.43482-0f2d60242
Is ROCm available: True
Number of GPUs: 2
GPU Name: AMD Radeon AI PRO R9700
Tensor operation successful on: cuda:0
tensor([[0.9593, 0.7855, 1.2845],
        [1.1800, 0.2567, 1.1684],
        [0.6278, 1.4628, 0.8078]], device='cuda:0')
ai1@ws1:~$
```

AMD ROCm™ + PyTorch® Environment Setup Guide

Initial Installation Walkthrough

ROCm Bandwidth Test

AMD's ROCm Bandwidth Test utility (**rocm-bandwidth-test**) is a diagnostic tool included in ROCm that is designed to provide a report on the memory bandwidth performance between:

- Host (CPU) ↔ GPU(s)
- GPU ↔ GPU (if multiple GPUs are installed)
- GPU internal memory

What it does

When used with the **"-a"** (all tests) flag, **rocm-bandwidth-test** will run a complete set of memory bandwidth diagnostics, including:

- Host-to-Device (H2D) bandwidth
- Device-to-Host (D2H) bandwidth
- Device-to-Device (D2D) bandwidth (for multi-GPU)
- Bidirectional / concurrent bandwidth tests

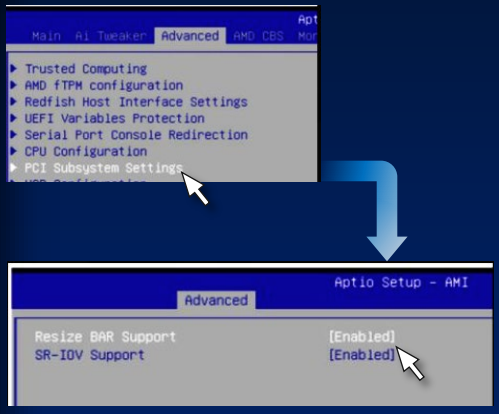
Run the P2P test

→ ***sudo /opt/rocm/bin/rocm-bandwidth-test -a***

Important! Make sure to enable "Resize BAR" setting in SBIOS

"Resize BAR" Enablement in SBIOS

1. Enter SBIOS
2. Navigate to Advanced
3. Go to PCI Subsystem Settings
4. Enable "Resize BAR Support"



Resize BAR = ENABLED

```
ai1@ws1:~$ sudo /opt/rocm/bin/rocm-bandwidth-test -a
[sudo] password for ai1:
..............
          RocmBandwidthTest Version: 2.6.0

          Launch Command is: /opt/rocm/bin/rocm-bandwidth-test -a


          Device: 0,  AMD Ryzen Threadripper PRO 5955WX 16-Cores
          Device: 1,  AMD Radeon AI PRO R9700,  GPU-2a63b1f95a8ce936,  23:0.0
          Device: 2,  AMD Radeon AI PRO R9700,  GPU-7532856404be20ff,  03:0.0

          Inter-Device Access

          D/D       0         1         2

          0         1         1         1

          1         1         1         1

          2         1         1         1


          Inter-Device Numa Distance

          D/D       0         1         2

          0         0         20        20

          1         20        0         40

          2         20        40        0


          Unidirectional copy peak bandwidth GB/s

          D/D       0           1           2

          0         N/A         14.257      27.971

          1         13.835      536.350     13.830

          2         27.699      13.827      535.152
```

With the Resize BAR enabled, bandwidth is shown between every device pair

Resize BAR = DISABLED

```
ai1@ws1:~$ sudo /opt/rocm/bin/rocm-bandwidth-test -a
............
          RocmBandwidthTest Version: 2.6.0

          Launch Command is: /opt/rocm/bin/rocm-bandwidth-test -a


          Device: 0,  AMD Ryzen Threadripper PRO 5955WX 16-Cores
          Device: 1,  AMD Radeon AI PRO R9700,  GPU-2a63b1f95a8ce936,  23:0.0
          Device: 2,  AMD Radeon AI PRO R9700,  GPU-7532856404be20ff,  03:0.0

          Inter-Device Access

          D/D       0         1         2

          0         1         0         0

          1         1         1         0

          2         1         0         1


          Inter-Device Numa Distance

          D/D       0         1         2

          0         0         N/A       N/A

          1         20        0         N/A

          2         20        N/A       0


          Unidirectional copy peak bandwidth GB/s

          D/D       0           1           2

          0         N/A         14.264      27.981

          1         14.077      534.357     N/A

          2         28.175      N/A         528.968
```

With the Resize BAR disabled, P2P will be deactivated

AMD ROCm™ + PyTorch® Environment Setup Guide

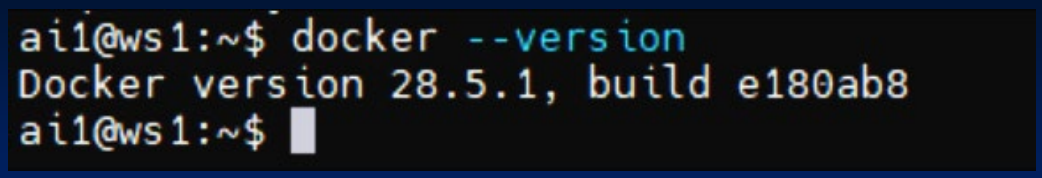
Docker Integration

Setting up Docker Environment

The script sets up a Docker environment with GPU passthrough support via ROCm.

Check Docker Installation

→ ***docker --version***



vLLM Docker Images

Use vLLM container build optimized for RDNA4

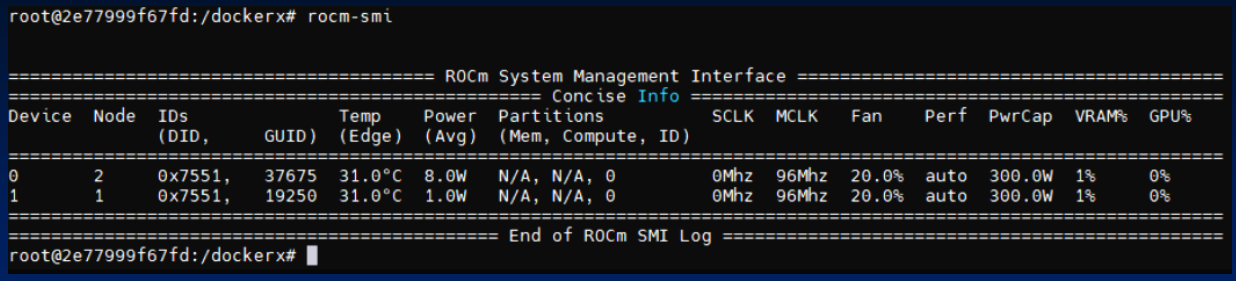
→ ***sudo docker pull rocm/vllm-dev:open-r9700-08052025***

Run vLLM with all available AMD GPU Access (example for RDNA4)

→ ***sudo docker run -it ***
***--device=/dev/kfd ***
***--device=/dev/dri ***
***--security-opt seccomp=unconfined ***
***--group-add video ***
rocm/vllm-dev:open-r9700-08052025

You can verify all available GPUs using ***rocm-smi*** (below is an example with 2 × Radeon AI PRO R9700 GPUs)

→ ***rocm-smi***



AMD ROCm™ + PyTorch® Environment Setup Guide

Docker Integration

Using a Specific GPU

If you need to add a specific GPU, you can use the **passthrough** option.

First, verify the available GPUs in the **/dev/dri** directory.

→ *cd /dev/dri/*
→ *ls*

```
ai1@ws1:~$ cd /dev/dri/
ai1@ws1:/dev/dri$ ls
by-path  card1  card2  renderD128  renderD129
ai1@ws1:/dev/dri$
```

Let's choose **GPU2**, also referred to as **"card2"** or **"renderD129“:**

→ *sudo docker run -it *
*--device=/dev/kfd *
*--device=/dev/dri/card2 *
*--device=/dev/dri/renderD129 *
*--security-opt seccomp=unconfined *
*--group-add video *
rocm/vllm-dev:open-r9700-08052025

GPU2 has been added to the container

```
ai1@ws1:~$ sudo docker run -it \
    --device=/dev/kfd \
    --device=/dev/dri/card2 \
    --device=/dev/dri/renderD129 \
    --security-opt seccomp=unconfined \
    --group-add video \
    rocm/vllm-dev:open-r9700-08052025
[sudo] password for ai1:
root@eeb0513e784e:/dockerx# rocm-smi


========================================= ROCm System Management Interface =========================================
================================================== Concise Info ==================================================
Device  Node  IDs              Temp    Power  Partitions          SCLK  MCLK   Fan    Perf  PwrCap  VRAM%  GPU%
              (DID,     GUID)  (Edge)  (Avg)  (Mem, Compute, ID)
==================================================================================================================
0       2     0x7551,   37675  30.0°C  8.0W   N/A, N/A, 0         0Mhz  96Mhz  20.0%  auto  300.0W  1%     0%
==================================================================================================================
============================================= End of ROCm SMI Log =================================================
root@eeb0513e784e:/dockerx#
```

AMD ROCm™ + PyTorch® Environment Setup Guide

Running Llama3 8B LLM (inference – standard transformer)

Requirements

- New Ubuntu 22.04.x or 24.04.x installation
- ROCm 6.4.2 + PyTorch 2.8.0 + AI stack
- Radeon PRO AI R9700 and/or Radeon PRO W7000 series GPU

Download the script from GitHub repository

→ ***wget https://raw.githubusercontent.com/JoergR75/LLama3-8B-70B-chat-agent/refs/heads/main/llama3-8b-agent.py***

→ ***python3 llama3-8b-agent.py***

```
root@9d3ff09d760e:/app# python3 llama3_8B_llm.py
tokenizer_config.json: 51.0kB [00:00, 111MB/s]
tokenizer.json: 9.08MB [00:00, 11.3MB/s]
special_tokens_map.json: 100%|| 107/107 [00:00<00:00, 765kB/s]
config.json: 100%|| 704/704 [00:00<00:00, 5.11MB/s]
model.safetensors.index.json: 23.9kB [00:00, 77.7MB/s]
model-00004-of-00004.safetensors: 100%|| 1.17G/1.17G [02:48<00:00, 6.94MB/s]
model-00001-of-00004.safetensors: 100%|| 4.98G/4.98G [07:17<00:00, 11.4MB/s]
model-00002-of-00004.safetensors: 100%|| 5.00G/5.00G [08:07<00:00, 10.2MB/s]
model-00003-of-00004.safetensors: 100%|| 4.92G/4.92G [08:25<00:00, 9.72MB/s]
Fetching 4 files: 100%|| 4/4 [08:26<00:00, 126.50s/it]
Loading checkpoint shards: 100%|| 4/4 [00:05<00:00,  1.42s/it]
generation_config.json: 100%|| 143/143 [00:00<00:00, 1.06MB/s]
Some parameters are on the meta device because they were offloaded to the cpu.
Chatbot: I be ready to chat! Type exit, quit, or bye to end the conversation.

User:
```

```
root@c2758afa687e:/app# python3 llama3_8B_llm.py
Loading checkpoint shards: 100%|| 4/4 [00:20<00:00,  5.15s/it]
Some parameters are on the meta device because they were offloaded to the cpu.
Chatbot: I be ready to chat! Type exit, quit, or bye to end the conversation.

User: What is quantum compute?

Chatbot: Quantum computing, in simple terms, is the use of extremely small units of matter (like atoms or electrons) to perform calculations. These tiny particles
can exist in multiple states at once, which allows for an enormous number of possible outcomes to be explored simultaneously.

This concept has been around since the early days of quantum mechanics but only recently have we seen significant advancements in building practical devices capabl
e of harnessing this power. The potential applications include:

1. **Faster simulations**: For complex systems like chemical reactions, weather forecasting, or materials science, where classical computers struggle with time-con
suming computations.
2. **Cryptography breakthroughs**: New methods for secure data encryption could emerge from the unique properties of quantum entanglement.
3. **Optimized solutions**: Finding the most efficient paths through vast solution spaces, such as in logistics, finance, or biology.

Keep in mind, though, that we're still in the process of developing reliable, large-scale, and error-corrected quantum computers. While impressive progress has bee
n made, there's much work ahead before these machines become mainstream tools."

Would you like me to elaborate on any specific aspect of quantum computing?

Response time: 43.06 seconds
Tokens per second: 5.41 t/s
```

AMD ROCm™ + PyTorch® Environment Setup Guide

Running Llama3 8B LLM (inference – standard transformer unoptimized)

Running the Chat Bot in Single and Dual GPU Mode

There are several ways to choose the active GPU:

1. Export HIP VISIBLE DEVICES=0 (1, or 0,1) – bare-metal
2. Within the python script - (device = torch.device("cuda:0")
3. Within a container instance

- *sudo docker run -it --device=/dev/kfd --device=/dev/dri/card1 --device=/dev/dri/renderD129 --security-opt seccomp=unconfined --group-add video my-vllm-image:custom* **– GPU 1**
- *sudo docker run -it --device=/dev/kfd --device=/dev/dri/card0 --device=/dev/dri/renderD128 --security-opt seccomp=unconfined --group-add video my-vllm-image:custom* **– GPU 0**
- *sudo docker run -it --device=/dev/kfd --device=/dev/dri --security-opt seccomp=unconfined --group-add video my-vllm-image:custom* **– all available GPUs**

GPU utilization at single GPU

```
root@c2758afa687e:/app# python3 llama3_8B_llm.py
Loading checkpoint shards: 100%|████████████████████████████████████████████████████████████| 4/4 [00:20<00:00,  5.15s/it]
Some parameters are on the meta device because they were offloaded to the cpu.
Chatbot: I be ready to chat! Type exit, quit, or bye to end the conversation.

User: What is quantum compute?

Chatbot: Quantum computing, in simple terms, is the use of extremely small units of matter (like atoms or electrons) to perform calculations. These tiny particles
can exist in multiple states at once, which allows for an enormous number of possible outcomes to be explored simultaneously.

This concept has been around since the early days of quantum mechanics but only recently have we seen significant advancements in building practical devices capabl
e of harnessing this power. The potential applications include:

1. **Faster simulations**: For complex systems like chemical reactions, weather forecasting, or materials science, where classical computers struggle with time-con
suming computations.
2. **Cryptography breakthroughs**: New methods for secure data encryption could emerge from the unique properties of quantum entanglement.
3. **Optimized solutions**: Finding the most efficient paths through vast solution spaces, such as in logistics, finance, or biology.

Keep in mind, though, that we're still in the process of developing reliable, large-scale, and error-corrected quantum computers. While impressive progress has bee
n made, there's much work ahead before these machines become mainstream tools."

Would you like me to elaborate on any specific aspect of quantum computing?

Response time: 43.06 seconds
Tokens per second: 5.41 t/s
```

```
========================================= ROCm System Management Interface =========================================
================================================= Concise Info =====================================================
Device  Node  IDs              Temp    Power   Partitions          SCLK    MCLK     Fan  Perf  PwrCap  VRAM%  GPU%
        (DID,     GUID)  (Edge)  (Avg)   (Mem, Compute, ID)
====================================================================================================================
0       1     0x7461,   13714  47.0°C  123.0W  N/A, N/A, 0         3303Mhz  1258Mhz  0%   auto  304.0W  84%    97%
1       2     0x7461,   26193  46.0°C  4.0W    N/A, N/A, 0         0Mhz     96Mhz    0%   auto  304.0W  3%     0%
====================================================================================================================
========================================= End of ROCm SMI Log ======================================================
```

256 token output:
~5.4 token/s
~43sec

GPU utilization at dual GPU

```
root@c2758afa687e:/app# python3 llama3_8B_llm.py
Loading checkpoint shards: 100%|████████████████████████████████████████████████████████████| 4/4 [00:20<00:00,  5.15s/it]
Some parameters are on the meta device because they were offloaded to the cpu.
Chatbot: I be ready to chat! Type exit, quit, or bye to end the conversation.

User: What is quantum compute?

Chatbot: Quantum computing, in simple terms, is the use of extremely small units of matter (like atoms or electrons) to perform calculations. These tiny particles
can exist in multiple states at once, which allows for an enormous number of possible outcomes to be explored simultaneously.

This concept has been around since the early days of quantum mechanics but only recently have we seen significant advancements in building practical devices capabl
e of harnessing this power. The potential applications include:

1. **Faster simulations**: For complex systems like chemical reactions, weather forecasting, or materials science, where classical computers struggle with time-con
suming computations.
2. **Cryptography breakthroughs**: New methods for secure data encryption could emerge from the unique properties of quantum entanglement.
3. **Optimized solutions**: Finding the most efficient paths through vast solution spaces, such as in logistics, finance, or biology.

Keep in mind, though, that we're still in the process of developing reliable, large-scale, and error-corrected quantum computers. While impressive progress has bee
n made, there's much work ahead before these machines become mainstream tools."

Would you like me to elaborate on any specific aspect of quantum computing?

Response time: 43.06 seconds
Tokens per second: 5.41 t/s
```

```
========================================= ROCm System Management Interface =========================================
================================================= Concise Info =====================================================
Device  Node  IDs              Temp    Power   Partitions          SCLK    MCLK     Fan  Perf  PwrCap  VRAM%  GPU%
        (DID,     GUID)  (Edge)  (Avg)   (Mem, Compute, ID)
====================================================================================================================
0       1     0x7461,   13714  47.0°C  123.0W  N/A, N/A, 0         3303Mhz  1258Mhz  0%   auto  304.0W  84%    97%
1       2     0x7461,   26193  46.0°C  4.0W    N/A, N/A, 0         0Mhz     96Mhz    0%   auto  304.0W  3%     0%
====================================================================================================================
========================================= End of ROCm SMI Log ======================================================
```

256 token output:
~15.8 token/s
~17sec

GD-239a: Radeon™ PRO W6000 and W7000 Series and Radeon™ AI PRO R9000 Series graphics cards (and later models) are not designed nor recommended for datacenter usage. Use in a datacenter setting may adversely affect manageability, efficiency, reliability, and/or performance. GD-239a.

RPW-495: Testing as of May 2025 by AMD. Average tokens per second of three runs, dropping edge cases where the model starts spiraling (more than 2k thinking tokens) to standardize response length. No speculative decode. All tests conducted on LM Studio 0.3.15 (Build 11). Vulkan Llama.cpp 1.28 used for AMD, NVIDIA-recommended CUDA 12 llama.cpp 1.30 with Flash Attention used for NVIDIA. Short Prompt: "How long would it take for a ball dropped from 10 meter height to hit the ground?" Long Prompt: "Summarize the following in exactly five lines: [Insert Scene 1 Act 1 of Romeo and Juliet]", Models tested: Phi 3.5 MoE Q4 K M, Mistral Small 3.1 24B Instruct 2503 Q8, DeepSeek R1 Distill Qwen 32B Q6, Qwen 32b Q6 System specifications: AMD Ryzen™ 9 7900X, 32GB DDR5 RAM 6000 MT/s, Windows 11 PRO 24H2, AMD Radeon™ AI PRO R9700 32GB using Adrenalin 25.6.1 RC vs AMD Ryzen™ 9 7900X, 32GB DDR5 RAM 6000 MT/s, Windows 11 PRO 24H2 with NVIDIA GeForce RTX 5080 and GeForce 576.4 drivers. Performance may vary. RPW-495.

RPW-496: Testing as of May 2025 by AMD using DeepSeek R1 Distill Qwen 32B Q6, Mistral Small 3.1 24B Instruct 2503 Q8, Flux.1 Schnell, SD 3.5 Medium models. Tested on a System with AMD Ryzen 9 7900X CPU, Radeon AI PRO R9700 GPU, 32GB DDR5 RAM, 1TB Storage, Windows 11 PRO 24H2, Adrenalin 25.6.1 RC drivers, ComfyUI - PyTorch 2.4 on Windows. System configurations may vary yielding different results. RPW-496

RPW-499: Testing as of May 2025 by AMD using DeepSeek R1 Distill Qwen (32B) benchmarks. Tested on a System with 2x AMD EPYC 9654 CPU, 768 GB DDR5 RAM, Ubuntu 22.04 + vLLM, with 4x Radeon AI PRO R9700 (32GB) GPU. System configurations may vary yielding different results. RPW-499.

RPW-506: As of October 1, 2025, the AMD Radeon AI PRO R9700 at MSRP of $1299 USD offers the most VRAM (32GB) among the current generation of Workstation and Consumer video cards under $1500 USD. Sources: https://www.nvidia.com/en-us/geforce/graphics-cards/50-series/ (Consumer) | https://www.cdw.com/ (Workstation). RPW-506

RX-1143: Based on specifications of AMD RDNA 4 architecture compared to AMD RDNA 3 architecture as of December 2024. RX-1143

DISCLAIMER: The information contained herein is for informational purposes only and is subject to change without notice. While every precaution has been taken in the preparation of this document, it may contain technical inaccuracies, omissions and typographical errors, and AMD is under no obligation to update or otherwise correct this information. Advanced Micro Devices, Inc. makes no representations or warranties with respect to the accuracy or completeness of the contents of this document, and assumes no liability of any kind, including the implied warranties of noninfringement, merchantability or fitness for particular purposes, with respect to the operation or use of AMD hardware, software or other products described herein. No license, including implied or arising by estoppel, to any intellectual property rights is granted by this document. Terms and limitations applicable to the purchase or use of AMD products are as set forth in a signed agreement between the parties or in AMD's Standard Terms and Conditions of Sale. GD-18u.

© 2025 Advanced Micro Devices, Inc. All rights reserved. AMD, the AMD Arrow logo, Radeon, Radeon AI, and combinations thereof are trademarks of Advanced Micro Devices, Inc. Other product names used in this publication are for identification purposes only and may be trademarks of their respective owners. Certain AMD technologies may require third-party enablement or activation. Supported features may vary by operating system. Please confirm with the system manufacturer for specific features. No technology or product can be completely secure.

PID# 253878050