A Comparative Analysis of Gaming and Data Centre GPUs: Evaluating the NVIDIA RTX 4090 and A100 Across AI, HPC, and Graphics Workloads

Benjamin Davies
Young Innovators Research Program, Harvard College Project for Asian and International Relations, In collaboration with Learn with Leaders

***Abstract*—This study presents a comparative analysis of two NVIDIA Graphics Processing Units (GPUs) designed for fundamentally different applications: the A100, a data centre GPU optimized for artificial intelligence training and high-performance computing, and the RTX 4090, a consumer GPU designed for gaming and graphics rendering. Using secondary sources including manufacturer specifications, benchmark data, and industry analyses, this paper compares the architecture, key features, and performance of both GPUs across AI training, 4K gaming, 3D rendering, and floating-point precision workloads. Findings indicate that the A100 outperforms the RTX 4090 by approximately 3x in AI training due to its superior VRAM capacity, NVLink interconnect, and multi-instance partitioning capabilities, while the RTX 4090 achieves approximately 10x higher 4K gaming performance through dedicated ray tracing cores and DLSS 3 frame generation. The study concludes that each GPU excels within its target domain, and that neither can effectively substitute for the other across all workloads.**

***Index Terms*— Graphics Processing Unit, NVIDIA A100, NVIDIA RTX 4090, GPU benchmarking, artificial intelligence training, high-performance computing, Ada Lovelace architecture**

I. INTRODUCTION

Graphics Processing Units (GPUs) are a core component of any computer used for gaming (generating graphics), AI (training and running models), or high-performance computing (running complex simulations). This research aims to compare two common GPUs that are targeted at different audiences and that have very different prices. Secondary sources will be used to understand the key features of each GPU, compare the performance of each GPU at different tasks, and find when one GPU would be a better fit, depending on what task needs to be performed.

II. LITERATURE REVIEW

Architecture, Memory, and Data Transfer Technologies

An article on the website of the Association for Computing Machinery (Yoshida K, 2022) provided information about the architecture and the specifications of the NVIDIA A100 GPU.

Information was also gathered from DigitalOcean, an Elite-level partner of NVIDIA, which manufactures the GPU cards discussed in this study (Lheureux, 2022). This source provided insights into various graphics card components. Of particular importance was the data regarding NVLink, a feature that enables multiple A100 graphics cards to transfer data at high speeds, allowing them to work collectively on a single task.

Additionally, Darkflash, a PC building company catering especially to gamers (Darkflash.com, 2025), provided information about the method used by the RTX 4090 to transfer data to the CPU and to any other GPUs on the same motherboard. The RTX 4090 does not have a fast data transfer method available like NVLink, which is used by the A100, and so any data transfer speeds that the RTX 4090 uses are limited to the speed of the standard PCIe interface found on regular computer motherboards.

Manufacturer Specifications and Key Features

NVIDIA is the manufacturer of both graphics' cards assessed by this research project. The official NVIDIA

website for the A100 (NVIDIA.com, 2020) provides details about 2 of the A100's unique features, called MIG and NVLink, which make it suited for multi-user environments and scaling up to very large tasks. The official NVIDIA website for the RTX 4090 (NVIDIA.com, 2022) provides details about 2 of the RTX 4090's unique features, called ray tracing and DLSS 3, which make it suited for gaming and graphics rendering

Moreover, Techpowerup, a review site focused on the PC gamer community (Techpowerup.com, 2022), provided architectural information about the RTX 4090's components, such as tensor cores, transistors, ray tracing acceleration cores, and VRAM. Hivenet, a cloud computing provider (Hivenet.com, 2020), provided information about why high-speed memory is so important for training large AI models.

Hyperstack, a cloud computing provider focusing on GPUs (Vohra, D. K, 2024), provided information about the performance differences between PCIe and SXM/NVLink, which are the different data transfer technologies used by the GPUs covered in this research project. Massed Compute, a cloud computing provider and partner of NVIDIA (MassedCompute.com, 2025), provided information on how the A100's large VRAM and high bandwidth allow it to process more training data at a time without having to load small batches from disk. Trooper.AI, a cloud computing provider specialising in private GPU servers (Trooper.AI, 2025), performed performance testing of both GPUs against 26 specific benchmarks designed to highlight the type of tasks that each is best suited to.

Performance Benchmarking and Precision Format Comparisons

Similarly, Spheron, a cloud computing provider specialising in GPUs (Spheron.com, 2024), provided information about the different precision formats offered by both GPUs and which precision formats were most useful for gaming vs AI model training. Custompc.com, a PC hardware review site (Custompc.com, 2023), provided details about the RTX 4090's Ada Lovelace architecture, which is more modern than the Ampere architecture used by the A100 and is able to fit about 3.5 times more transistors on the same size microchip. Lastly, Databasemart, a GPU server hosting provider (Violet, 2025), performed various benchmarks to compare and provide statistical information between the RTX 4090 and the A100 to highlight which GPU is more appropriate for specific tasks.

III. METHODOLOGY

This study employs a qualitative secondary research methodology using a comparative case study design to evaluate the NVIDIA A100 and NVIDIA RTX 4090 Graphics Processing Units across multiple performance domains. A secondary research approach was selected for two reasons. First, conducting primary benchmark testing would require physical access to both GPUs in controlled, equivalent hardware environments the A100 requires specialized server-grade infrastructure, and the combined hardware cost would exceed the scope of this study. Second, a substantial body of independently verified benchmark data, manufacturer specifications, and technical analyses already exists in the public domain, providing a robust empirical foundation for comparative analysis without the need for original experimentation.

Data were collected from three categories of secondary sources, selected to ensure triangulation and minimize bias from any single source type. The first category comprised manufacturer documentation official NVIDIA specification sheets and product pages for both the A100 (NVIDIA.com, 2020) and the RTX 4090 (NVIDIA.com, 2022) which provided authoritative architectural and feature data. The second category comprised independent technical analyses from computing publications and cloud infrastructure providers (Yoshida et al., 2022; Lheureux, 2022; Tech Powerup, 2022; CustomPC, 2023; Hivenet, 2020; Spheron, 2024), which provided third-party verification of manufacturer claims and contextual analysis of each GPU's design philosophy. The third category comprised performance benchmark data from cloud computing providers and GPU testing platforms (Trooper.AI, 2025; DatabaseMart, 2025), which provided standardized, quantitative performance comparisons across workloads including AI training, 4K gaming, 3D rendering, and floating-point precision tasks.

Sources were selected based on four criteria: technical credibility of the publishing organization, recency of publication (preference given to sources published between 2020 and 2025 to align with the active product lifecycle of both GPUs), specificity to the GPUs under study, and the inclusion of either verifiable specifications or reproducible benchmark data. Sources that provided only subjective product reviews without technical substantiation were excluded.

The collected data was analyzed using a structured comparative framework across three dimensions. First, architectural comparison evaluating the design specifications, manufacturing process, core counts, memory systems, and interconnect technologies of each GPU. Second, feature comparison identifying and categorizing the specialized capabilities of each GPU (such as MIG and NVLink for the A100, and ray tracing and DLSS 3 for the RTX 4090) and assessing their relevance to specific use cases. Third, performance comparison synthesizing benchmark data across five workload categories (AI training, AI inference, 4K gaming, 3D rendering, and floating-point precision) to quantify relative performance differences between the two GPUs. Benchmark scores were normalized where necessary to enable direct comparison across sources using different scoring systems.

This methodology has several limitations that should be acknowledged. The study relies entirely on secondary data; primary benchmark testing under controlled conditions would provide more precise and directly comparable results. Benchmark scores from different sources may use different testing conditions, software versions, and scoring scales, introducing potential inconsistency. Additionally, manufacturer specifications represent theoretical maximums that may not reflect real-world performance under all conditions. Finally, GPU technology evolves rapidly, and the performance data cited in this study reflects the state of both products at the time of publication of each source. Despite these limitations, the triangulation of manufacturer data, independent technical analysis, and third-party benchmark testing provides a sufficiently robust foundation for the comparative conclusions drawn in this study.

IV. RESULTS & DISCUSSION

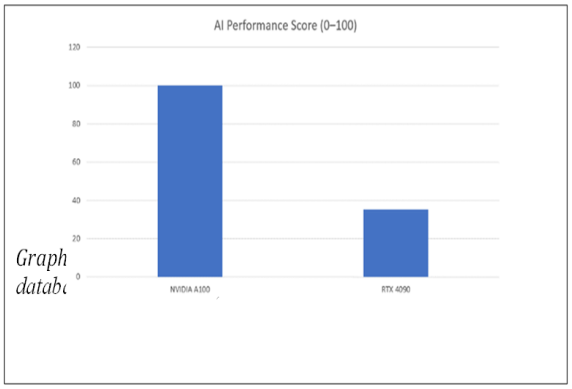
Table 1: Side-by-side feature comparison

Feature	A100	RTX 4090
Ray Tracing and DLSS3	No	Yes
VRAM	40 or 80 GB	24 GB
Interconnection speed	600 Gb/s (using NVLink and SXM)	63 Gbs (using PCIE)
Architecture	Ampere (8nm)	Ada Lovelace (5nm)
Tensor Cores	432	512
MIG	Yes	No
NVLink	Yes	No
Price	$14,000 – $28,000	$3,500 - $5,000
Transistors	54.2 billion	76.3 billion
Performance (precision computing)	9.7 Teraflops	2.6 Teraflops
Performance (gaming graphics)	19.5 Teraflops	82.6 Teraflops

Source: Created by the author, adapted from NVIDIA.com

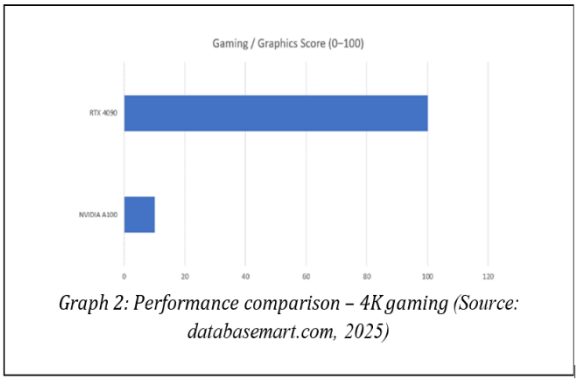
Benchmark Comparisons

In most GPU comparison articles, points in graphs are just the numerical scores assigned by a benchmark test, not an exact measurement, with each article using its own scoring system. Points represent relative performance and just show a comparison between the 2 GPUs.



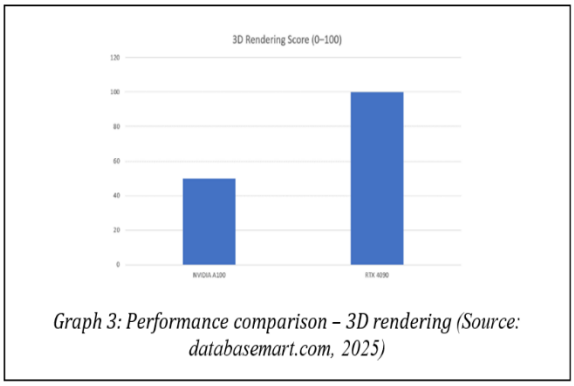
Graph
databa

The A100 performed almost 3x better in AI performance than the RTX 4090 in AI training and inference. This is because the A100 has significantly more VRAM and can therefore store and train more of the AI model in memory at one time. As previously mentioned, it is also designed to work in parallel with other A100 GPUs for maximum efficiency and capability. The RTX 4090 was able to train small AI models (up to around 20GB) but is inefficient for training larger models because of its limited VRAM and limited ability to work in parallel.



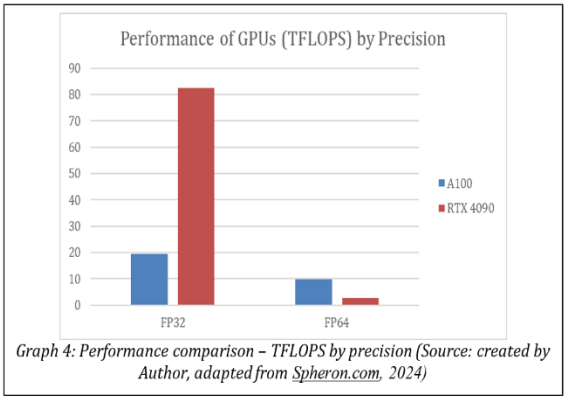
Graph 2: Performance comparison – 4K gaming (Source: databasemart.com, 2025)

In 4K gaming performance, the RTX 4090 achieved a score of 100, compared to the A100, which only scored 10. This is because the 4090 has built-in features specific to this purpose, such as ray tracing and DLSS3. Again, this shows that the RTX 4090 is an ideal GPU for the most demanding gamers and graphic designers.



Graph 3: Performance comparison – 3D rendering (Source: databasemart.com, 2025)

In 3D rendering and creative workloads (applications like Blender, Octane, and other rendering software), the 4090 performed better, achieving a score of 100 compared to the A100's score of 50. This makes it ideal for gamers, video editors, and graphic artists.



Graph 4: Performance comparison – TFLOPS by precision (Source: created by Author, adapted from Spheron.com, 2024)

FP32 precision is mostly relevant to gaming graphics, AI inference, 3D rendering, and most consumer applications, as these sorts of workloads do not require high data precision. For FP32 workloads, the RTX 4090 completed over 4x more computing operations per second than the A100.

FP64 precision is most relevant for scientific simulations, engineering calculations, high-precision computing, and supercomputing. These workloads need extremely accurate math. Some examples are large language models such as ChatGPT and Claude, weather prediction models, and computational fluid dynamics. For FP64 workloads, the A100 completed over 3x more computing operations per second than the RTX 4090.

The NVIDIA A100: Architecture and Data Centre Capabilities

The A100 is a high-performance GPU designed for training and running AI models, data analytics, and HPC (high-performance computing). The A100 powers the world's highest-performing data centres for AI, HPC, and data analytics. It has 108 processing units, which are used for handling various complex calculations, such as supporting machine learning applications with high precision and efficiency, and graphics rendering. It has a total of 6912 FP32 cores, which are units for processing single-precision floating-point numbers, crucial for scientific simulations. These cores are enhanced by 432 Tensor Cores that handle matrix mathematics to provide AI and HPC acceleration. The A100 can be purchased with either 40GB or 80GB of HBM2e (high bandwidth memory) VRAM. With 80GB of VRAM, the A100 can hold entire AI models with billions of parameters in memory. More VRAM also means bigger batches (sets of data) of data can be processed at a time when training AI models. This combination of a large amount of VRAM along with the high memory bandwidth (speed) of this VRAM reduces data movement, prevents memory-swap stalls, and supports flexible partitioning. Some of the A100's key features that differentiate it for high-performance computing and large-scale AI are MIG and NVLink. MIG allows the GPU to be used by multiple users, achieving better resource utilization in data centres. NVLink allows GPU-to-GPU communication, greatly increasing bandwidth to 600GB/s which is significant when combining many A100 GPUs together in

parallel for training and running the largest-scale AI workloads.

The NVIDIA RTX 4090: Architecture and Gaming Capabilities

The RTX 4090 is a high-performance GPU designed mainly for gaming. It comes with several built-in features that are specifically designed to support gaming and graphics rendering. It can still run AI workloads, but at a smaller scale than a GPU such as the A100. It comes with an AD102 graphics processor made from 76.3 billion transistors. It supports the newest version of Microsoft's DirectX API, "DirectX 12 Ultimate," which is designed for next-generation gaming. It includes features like ray tracing, variable rate shading (VRS), Mesh shaders, and sampler feedback. The RTX 4090 features 16384 shading units, 512 texture mapping units, and 176 ROPs. It has 512 tensor cores, which help improve the speed of machine learning applications. The card has 128 ray tracing acceleration cores and 24 GB GDDR6X memory (VRAM). Ray tracing is a feature found on GPUs designed primarily for gaming. Ray tracing simulates the behaviour of light from the real world into the game, producing a more realistic frame. DLSS 3 is another feature used by the RTX 4090, which uses AI to create additional frames, providing smoother animation.

Comparative Analysis: AI Training and Large-Scale Workloads

The RTX 4090 is an incredible GPU for training and running smaller AI models, gaming, and single-GPU experimentation, whereas the A100 is built for training large AI models at scale. The A100 can provide from 1.6 TB/s to 2 TB/s of memory bandwidth, which is the speed at which data can move between memory and processing units. The RTX 4090 only has about 1,008 GB/s of memory bandwidth, meaning it has far less capacity to move data in and out of the GPU, which becomes a major bottleneck when training large models. The A100 excels at very large-scale processing because multiple A100 GPUs can be connected using NVLink. This allows the A100s to send data between themselves at up to 600 GB/s.

The A100's third-generation Tensor Cores are built for deep learning workloads, accelerating the matrix operations behind today's large-scale AI models. These tensor cores support data formats such as BF16, which is a data type mainly used in artificial intelligence. This data type support makes the A100 fast and memory-efficient for AI training while still maintaining stable model accuracy.

Comparative Analysis: Architectural Efficiency and Smaller Workloads

The A100 is designed for training and running large AI models, but when it comes to smaller models, its performance is very similar to the RTX 4090. Thanks to its speed-optimised architecture, the RTX 4090 can deliver amazing results on these lighter workloads. The RTX 4090 uses the Ada architecture, which uses a more modern 5-nanometer process to etch transistors into silicon compared to the 8-nanometer process used for the A100's Ampere architecture. This means that the Ada architecture fits about 3.5 times more transistors into the same space as the Ampere architecture. More transistors in a smaller space means that signals don't have to travel as far, and therefore, processing speed improves.

As long as VRAM isn't a limiting factor, the RTX 4090 has no problem with training and inference on smaller AI models. It is great for single users training smaller AI models, but it does not have enough VRAM, resource-sharing abilities, and advanced Tensor Cores to effectively train large AI models and be used in multi-user environments.

Comparative Analysis: Graphics Rendering and Gaming Performance

The RTX 4090 is significantly better than the A100 at rendering graphics, especially for video games. The RTX 4090 has 16,384 CUDA cores, compared to the A100 having only 6,912. These CUDA cores are the FP32 ALUs (the units that do the FP32 math) that give the RTX 4090 faster FP32 compute performance. Higher FP32 compute directly improves real-time graphics performance. The RTX 4090 has two features used specifically for gaming that the A100 does not: Ray tracing and DLSS3. The RTX 4090 uses specialized units called Ray tracing cores, which accelerate the math used in ray tracing that normal CUDA cores are too slow to handle. Ray tracing simulates light behaviour in scenes, producing effects such as reflections, shadows, and lighting. Secondly, the RTX 4090 supports a technology called DLSS 3, which uses AI to generate extra frames and improve

lower-resolution images. This results in smoother gameplay and a higher frame rate.

V. CONCLUSION

The RTX 4090 is the best value for home users for gaming, small businesses for graphics processing, training small AI models, and running inference on small to medium-sized AI models. The A100 is the best choice for training even the largest AI models and for use in corporations where many users need to be able to share the same GPU.

The NVIDIA A100 GPU uses the 8-nanometer Ampere architecture and is mainly sold to large corporations for training and running large AI models or for running complex simulations. It has features that make it very good at being shared by many users, as well as features that allow multiple A100s to work together on the same problem. The NVIDIA RTX 4090 uses the more modern 5-nanometer Ada Lovelace architecture and is mainly sold to home users for gaming or small businesses for video processing. It is best when used by a single user at a time and comes with several features designed to improve the generation of graphics. It can be used for training AI models, but is limited to smaller models than the A100.

Gaming GPUs are built for rendering images quickly with visual quality, while data-centre GPUs focus on high-performance computing, such as predicting the weather, AI training and inference, and working together with other GPUs to scale up in order to deal with the heaviest workloads. Exploring both types can help understand how their designs, features, and performance goals differ.

For future work on this topic, it would be ideal to have two equivalent computers side by side, one with each type of GPU. Primary research can be performed by running a series of benchmarking tests to measure the relative performance when training common AI models and when rendering common computer graphics.

VI. GLOSSARY

The terminology below can help a reader grasp the comparisons for GPUs presented in this research:

API: An API (Application Programming Interface) is a set of rules and tools that allows different software systems to communicate and share data or functionality. It allows developers to access specific features of an application or service without needing to understand its internal code.

BF16 (Bfloat16): BF16 is a 16-bit floating-point format that keeps the same large exponent range as FP32 (meaning it can represent very large and very small values without overflow or underflow) but uses fewer bits for precision. This makes it fast and memory-efficient for AI training while still maintaining stable model accuracy.

FP32/FP64: FP32 and FP64 are numerical formats used by CPUs and GPUs to represent real numbers with different levels of precision.

FP64: FP64 (double precision or 64-bit floating point) is used for scientific simulations, engineering calculations, high-precision research, and supercomputing (these workloads need extremely accurate math)

FP32: FP32 (single precision or 32-bit floating point) is used for gaming graphics, AI inference, 3D rendering, and most consumer applications (these tasks don't need super high precision)

MIG (Multi-Instance GPU): MIG allows one physical GPU to be partitioned into several smaller, independent GPU instances for better resource utilization in data centres. Multiple users can use one physical GPU.

NVLink: SXM GPUs use NVIDIA's NVLink technology to enable GPU-to-GPU communication. This allows the GPUs to have up to 600GB/s of bandwidth, which is a significant advantage when running large-scale AI training workloads.

ROPs (Raster Operations Pipelines): ROPs are hardware units in a GPU responsible for final pixel processing tasks such as blending, depth testing, and writing completed pixels to the framebuffer. They play a key role in determining how quickly a GPU can output finished images to the screen, especially at high resolutions.

Tensor: A tensor is the way that data is stored for training AI models. It is stored in such a format that it can be processed by a GPU. Modern GPUs come with Tensor Cores, which are computing components designed specifically to process data in tensor format. (Labonne, M, 2022)

Tensor cores: Tensor Cores are specialized processing units in NVIDIA GPUs designed to accelerate matrix-math operations, which are the core computations in deep-learning workloads. They deliver much higher

throughput (the rate that a GPU can process data and execute instructions) for tasks like neural-network training and inference compared to standard GPU cores.

Teraflop: A teraflop is a unit of measurement commonly used to compare how much work a computing device can perform. 'Tera' means trillion, and 'FLOP' means floating point operations. So, a computing device that runs at 1 teraflop/s is able to perform 1 trillion floating-point operations per second. Higher teraflop values generally indicate greater raw compute power, especially for tasks like graphics rendering, simulations, and AI workloads. (Lynch, M, 2023)

TF32: TF32 (Tensor Float 32) is a precision format introduced by NVIDIA in its Ampere architecture to deliver much faster performance for AI training. It blends aspects of FP32 and FP16 to balance range and precision, making it highly efficient for deep-learning workloads while maintaining practical accuracy.

Transistors: Transistors are the building blocks of a modern GPU. They are tiny electronic switches that control the flow of electrical signals. They allow the GPU to process large amounts of data and perform complex calculations, handling tasks like executing AI algorithms, rendering graphics, and performing parallel computations.

REFERENCE

[1] K. Yoshida, R. Sageyama, S. Miwa, H. Yamaki, and H. Honda, "Analyzing performance and power-efficiency variations among NVIDIA GPUs," in *Proc. ACM*, 2022, pp. 1–12, doi: 10.1145/3545008.3545084.

[2] Lheureux, "A complete anatomy of a graphics card: Case study of the NVIDIA A100," *Paperspace Blog*, Jul. 13, 2022. [Online]. Available: https://blog.paperspace.com/a-complete-anatomy-of-a-graphics-card-case-study-of-the-nvidia-a100/

[3] DarkFlash, "Understanding PCIe lanes: The hidden highways of your PC," 2025. [Online]. Available: https://www.darkflash.com/article/understanding-pcie-lanes-the-hidden-highways-of-your-p

[4] NVIDIA, "NVIDIA A100 GPUs power the modern data center." [Online]. Available: https://www.nvidia.com/en-au/data-center/a100/

[5] NVIDIA, "NVIDIA GeForce RTX 4090 graphics cards." [Online]. Available: https://www.nvidia.com/en-us/geforce/graphics-cards/40-series/rtx-4090/

[6] TechPowerUp, "NVIDIA GeForce RTX 4090 specs," Oct. 31, 2022. [Online]. Available: https://www.techpowerup.com/gpu-specs/geforce-rtx-4090.c3889

[7] Hivenet, "Mastering performance with the NVIDIA A100: A comprehensive guide," 2020. [Online]. Available: https://compute.hivenet.com/post/nvidia-a100-gpu-complete-guide-to-data-center-ai-acceleration

[8] D. K. Vohra, "NVIDIA A100 PCIe vs SXM: Comprehensive performance comparison," *Hyperstack*, Dec. 12, 2024. [Online]. Available: https://www.hyperstack.cloud/technical-resources/performance-benchmarks/nvidia-a100-pcie-vs-nvidia-a100-sxm-a-comprehensive-comparison

[9] Massed Compute, "How does the VRAM capacity of NVIDIA A100 GPUs affect model size and performance?" Jul. 31, 2025. [Online]. Available: https://massedcompute.com/faq-answers/?question=How%20does%20the%20VRAM%20capacity%20of%20NVIDIA%20A100%20GPUs%20affect%20model%20size%20and%20performance

[10] Trooper.AI, "A100 vs RTX 4090 – GPU benchmark comparison," 2025. [Online]. Available: https://www.trooper.ai/benchmarks-compare-A100-with-RTX-4090

[11] Spheron, "GeForce RTX 4090 vs. A100-PCIE-40GB: Ultimate AI and deep learning GPU," 2024. [Online]. Available: https://web.archive.org/web/20250219233651/https://blog.spheron.network/geforce-rtx-4090-vs-a100-pcie-40gb-ultimate-ai-and-deep-learning-gpus-compared

[12] Custom PC, "Inside the NVIDIA Ada Lovelace GPU architecture," Jun. 29, 2023. [Online]. Available: https://www.custompc.com/inside-nvidia-ada-lovelace-gpu-architecture

[13] Violet, "NVIDIA A100 vs RTX 4090: AI, rendering, and gaming," *Database Mart*, Dec. 28, 2025. [Online]. Available: https://www.databasemart.com/blog/nvidia-a100-vs-nvidia-rtx4090

[14] M. Labonne, "What is a tensor in machine learning?" *Towards Data Science*, Mar. 29, 2022. [Online]. Available: https://towardsdatascience.com/what-is-a-tensor-in-deep-learning-6dedd95d6507/

[15] M. Lynch, "What is a teraflop?" *The Tech Edvocate*, Sep. 7, 2023. [Online]. Available: https://www.thetechedvocate.org/what-is-a-teraflop/