# □□ GPU COMPARISON MATRIX FOR AI / LLM WORKLOADS

Detailed technical comparison between five graphics cards: **RTX 3060 12GB**, **RTX 5060 Ti 16GB**, **9060 XT 16GB**, **Arc A770 16GB**, and **Arc B60 24GB**, focusing on core metrics for AI, Deep Learning, and Local LLM deployments.

## □□ Detailed Comparison Matrix

| Metric | NVIDIA GeForce RTX 3060 | NVIDIA GeForce RTX 5060 Ti | AMD Radeon RX 9060 XT | Intel Arc A770 (Graphics) | Intel Arc B60 (Workstation) |
|---|---|---|---|---|---|
| **1. VRAM Capacity** | 12 GB GDDR6 | 16 GB GDDR7 | 16 GB GDDR6 | 16 GB GDDR6 | **24 GB GDDR6** |
| **2. Memory Speed** | 15 Gbps *(Bus: 192-bit) (Bandwidth: 360 GB/s)* | **28 Gbps** *(Bus: 128-bit) (Bandwidth: 448 GB/s)* | 16 Gbps *(Bus: 128-bit) (Bandwidth: 320 GB/s)* | 17.5 Gbps *(Bus: **256-bit**) (Bandwidth: **560 GB/s**)* | 19 Gbps *(Bus: 192-bit) (Bandwidth: 456 GB/s)* |
| **3. AI Compute (INT8 TOPS)** | ~244 TOPS *(Tensor Cores Gen 3)* | **~759 TOPS** *(Tensor Cores Gen 5)* | ~410 TOPS *(RDNA 4 AI Accelerators)* | ~256 TOPS *(XMX Engines)* | ~197 TOPS *(XMX Engines Gen 2)* |
| **4. Release Date** | Q1 / 2021 | Q1 / 2025 | Q1 / 2025 | Q4 / 2022 | Q1 / 2025 |

---

## □□ Quick Analysis for AI / Local LLM Engineers

1. **VRAM & Model Capacity:**
   - The **Intel Arc B60** stands out with its massive **24GB VRAM**, matching upper-tier workstation cards in capacity. This allows it to comfortably host larger models (up to

quantized 32B or heavily compressed 70B models) entirely within the GPU memory.
    - The **5060 Ti, 9060 XT, and A770** sit comfortably at **16GB VRAM**, which is the ideal sweet spot for running mid-sized models like *Llama-3-8B* or *Qwen-2.5-14B* without system RAM spillover.
    - The **RTX 3060** falls behind with **12GB**, but remains a budget-friendly entry-level gateway card.
2. **Memory Bandwidth (Token Generation Speed):**
    - The older **Intel Arc A770** still holds the highest raw bandwidth (**560 GB/s**) due to its unconstrained 256-bit bus width, ensuring fluid data transfer during inference loops.
    - The **Arc B60** sits at **456 GB/s**; even with its 24GB capacity, the 192-bit bus acts as a subtle speed limit compared to high-end architectures.
    - The **RTX 5060 Ti** achieves a highly efficient **448 GB/s** by leveraging next-gen **GDDR7 28 Gbps** modules, overcoming its narrow 128-bit bus constraint.
    - The **RX 9060 XT** is the most bottlenecked here at **320 GB/s**, which slightly bottlenecks its token-per-second output during pure LLM inference tasks.
3. **Raw Inference Performance (INT8 TOPS):**
    - The **RTX 5060 Ti** absolutely obliterates the field with a staggering **759 TOPS**, making it the premier choice for fine-tuning, computer vision, and heavy image generation loops.
    - The **RX 9060 XT** occupies a strong mid-range position at **~410 TOPS**, representing a substantial compute upgrade for AMD's mainstream platform.
    - The **Intel Arc B60** significantly underdelivers in compute density, dropping below 200 TOPS—effectively losing to both its predecessor (A770) and even the aging RTX 3060 in raw matrix mathematical operations.
4. **Software Ecosystem & Compatibility:**
    - **NVIDIA (3060, 5060 Ti):** Seamless out-of-the-box operation natively powered by **CUDA**. 100% day-one compatibility with major frameworks (PyTorch, HuggingFace, vLLM).
    - **AMD (9060 XT):** Relies on **ROCm**. Highly performant and native on Linux-based environments (via Ollama or Llama.cpp), though Windows support requires slightly more setup.
    - **Intel (A770, B60):** Requires reliance on the **OneAPI / OpenVINO** toolkit or IPEX (Intel Extension for PyTorch), demanding a higher degree of manual environment configuration and command-line troubleshooting.

---

Revision #6
Created 2026-05-24 10:42:20 UTC by sandichhuu
Updated 2026-05-28 05:57:32 UTC by sandichhuu