**GPU-벤치마크-on-LLM-추론**

대규모 언어 모델 추론을 위한 여러 NVIDIA GPU 또는 Apple Silicon? 🧐

**묘사**

[런팟(RunPod](https://www.runpod.io/)), 13형 M1 맥북 에어(MacBook Air), 14형 M1 맥스 맥북 프로(MacBook Pro), M2 울트라 맥 스튜디오(M2 Ultra Mac Studio), 16형 M3 맥스 맥북 프로(LLaMA 3)에서 [llama.cpp](https://github.com/ggerganov/llama.cpp) 사용하여 다양한 GPU의 [LLaMA](https://arxiv.org/abs/2302.13971) 모델 추론 속도를 테스트할 수 있다.

**개요**

LLaMA 3에서 GPU에 의해 1024개의 토큰을 생성하는 평균 속도(토큰/초). 속도는 높을수록 좋습니다.

| **그래픽 카드(GPU** | **8B Q4\_K\_M** | **8B F16** | **70B Q4\_K\_M** | **70B F16** |
| --- | --- | --- | --- | --- |
| 3070 8기가바이트 | 70.94 | 움 | 움 | 움 |
| 3080 10기가바이트 | 106.40 | 움 | 움 | 움 |
| 3080 Ti 12기가바이트 | 106.71 | 움 | 움 | 움 |
| 4070 Ti 12기가바이트 | 82.21 | 움 | 움 | 움 |
| 4080 16기가바이트 | 106.22 | 40.29 | 움 | 움 |
| RTX 4000 에이다 20GB | 58.59 | 20.85 | 움 | 움 |
| 3090 24기가바이트 | 111.74 | 46.51 | 움 | 움 |
| 4090 24기가바이트 | 127.74 | 54.34 | 움 | 움 |
| RTX 5000 에이다 32GB | 89.87 | 32.67 | 움 | 움 |
| 3090 24GB \* 2 | 108.07 | 47.15 | 16.29 | 움 |
| 4090 24기가바이트 \* 2 | 122.56 | 53.27 | 19.06 | 움 |
| RTX A6000 48GB | 102.22 | 40.25 | 14.58 | 움 |
| RTX 6000 에이다 48GB | 130.99 | 51.97 | 18.36 | 움 |
| A40 48GB | 88.95 | 33.95 | 12.08 | 움 |
| L40S 48GB | 113.60 | 43.42 | 15.31 | 움 |
| RTX 4000 에이다 20GB \* 4 | 56.14 | 20.58 | 7.33 | 움 |
| A100 PCIe 80GB | 138.31 | 54.56 | 22.11 | 움 |
| A100 SXM 80GB | 133.38 | 53.18 | 24.33 | 움 |
| H100 PCIe 80GB | **144.49** | **67.79** | 25.01 | 움 |
| 3090 24GB \* 4 | 104.94 | 46.40 | 16.89 | 움 |
| 4090 24GB \* 4 | 117.61 | 52.69 | 18.83 | 움 |
| RTX 5000 에이다 32GB \* 4 | 82.73 | 31.94 | 11.45 | 움 |
| 3090 24GB \* 6 | 101.07 | 45.55 | 16.93 | 5.82 |
| 4090 24GB \* 8 | 116.13 | 52.12 | 18.76 | 6.45 |
| RTX A6000 48GB \* 4 | 93.73 | 38.87 | 14.32 | 4.74 |
| RTX 6000 에이다 48GB \* 4 | 118.99 | 50.25 | 17.96 | 6.06 |
| A40 48GB \* 4 | 83.79 | 33.28 | 11.91 | 3.98 |
| L40S 48GB \* 4 | 105.72 | 42.48 | 14.99 | 5.03 |
| A100 PCIe 80GB \* 4 | 117.30 | 51.54 | 22.68 | 7.38 |
| A100 SXM 80GB \* 4 | 97.70 | 45.45 | 19.60 | 6.92 |
| H100 PCIe 80GB \* 4 | 118.14 | 62.90 | **26.20** | **9.63** |
| M1 7코어 GPU 8GB | 9.72 | 움 | 움 | 움 |
| M1 Max 32코어 GPU 64GB | 34.49 | 18.43 | 4.09 | 움 |
| M2 울트라 76코어 GPU 192GB | 76.28 | 36.25 | 12.13 | 4.71 |
| M3 Max 40코어 GPU 64GB | 50.74 | 22.39 | 7.53 | 움 |

평균 1024개의 토큰은 LLaMA 3에서 GPU에 의한 평가 속도(토큰/초)를 표시합니다.

| **그래픽 카드(GPU** | **8B Q4\_K\_M** | **8B F16** | **70B Q4\_K\_M** | **70B F16** |
| --- | --- | --- | --- | --- |
| 3070 8기가바이트 | 2283.62 | 움 | 움 | 움 |
| 3080 10기가바이트 | 3557.02 | 움 | 움 | 움 |
| 3080 Ti 12기가바이트 | 3556.67 | 움 | 움 | 움 |
| 4070 Ti 12기가바이트 | 3653.07 | 움 | 움 | 움 |
| 4080 16기가바이트 | 5064.99 | 6758.90 | 움 | 움 |
| RTX 4000 에이다 20GB | 2310.53 | 2951.87 | 움 | 움 |
| 3090 24기가바이트 | 3865.39 | 4239.64 | 움 | 움 |
| 4090 24기가바이트 | 6898.71 | 9056.26 | 움 | 움 |
| RTX 5000 에이다 32GB | 4467.46 | 5835.41 | 움 | 움 |
| 3090 24GB \* 2 | 4004.14 | 4690.50 | 393.89 | 움 |
| 4090 24기가바이트 \* 2 | 8545.00 | 11094.51 | 905.38 | 움 |
| RTX A6000 48GB | 3621.81 | 4315.18 | 466.82 | 움 |
| RTX 6000 에이다 48GB | 5560.94 | 6205.44 | 547.03 | 움 |
| A40 48GB | 3240.95 | 4043.05 | 239.92 | 움 |
| L40S 48GB | 5908.52 | 2491.65 | 649.08 | 움 |
| RTX 4000 에이다 20GB \* 4 | 3369.24 | 4366.64 | 306.44 | 움 |
| A100 PCIe 80GB | 5800.48 | 7504.24 | 726.65 | 움 |
| A100 SXM 80GB | 5863.92 | 681.47 | 796.81 | 움 |
| H100 PCIe 80GB | 7760.16 | 10342.63 | 984.06 | 움 |
| 3090 24GB \* 4 | 4653.93 | 5713.41 | 350.06 | 움 |
| 4090 24GB \* 4 | 9609.29 | 12304.19 | 898.17 | 움 |
| RTX 5000 에이다 32GB \* 4 | 6530.78 | 2877.66 | 541.54 | 움 |
| 3090 24GB \* 6 | 5153.05 | 5952.55 | 739.40 | 927.23 |
| 4090 24GB \* 8 | 9706.82 | 11818.92 | **1336.26** | 1890.48 |
| RTX A6000 48GB \* 4 | 5340.10 | 6448.85 | 539.20 | 792.23 |
| RTX 6000 에이다 48GB \* 4 | 9679.55 | 12637.94 | 714.93 | 1270.39 |
| A40 48GB \* 4 | 4841.98 | 5931.06 | 263.36 | 900.79 |
| L40S 48GB \* 4 | 9008.27 | 2541.61 | 634.05 | 1478.83 |
| A100 PCIe 80GB \* 4 | 8889.35 | 11670.74 | 978.06 | 1733.41 |
| A100 SXM 80GB \* 4 | 7782.25 | 674.11 | 539.08 | 1834.16 |
| H100 PCIe 80GB \* 4 | **11560.23** | **15612.81** | 1133.23 | **2420.10** |
| M1 7코어 GPU 8GB | 87.26 | 움 | 움 | 움 |
| M1 Max 32코어 GPU 64GB | 355.45 | 418.77 | 33.01 | 움 |
| M2 울트라 76코어 GPU 192GB | 1023.89 | 1202.74 | 117.76 | 145.82 |
| M3 Max 40코어 GPU 64GB | 678.04 | 751.49 | 62.88 | 움 |

**모델**

LLaMA 모델 가중치 (7B, 13B, 30B, 65B)에 대한 shawwn 덕분입니다 : [llama-dl](https://github.com/shawwn/llama-dl). [Meta AI](https://ai.meta.com/llama/)에서 LLaMA 2에 액세스하세요. Hugging Face 또는 내 Hugging Face 저장소인 [Xiongjie Dai](https://huggingface.co/JaaackXD)에서 [Meta Llama 3](https://huggingface.co/collections/meta-llama/meta-llama-3-66214712577ca38149ebb2b6)의 LLaMA 3에 액세스하세요.

**사용법**

**체격**

* NVIDIA GPU의 경우 Nvidia GPU의 CUDA 코어를 사용하여 BLAS 가속을 제공합니다.

!make clean && LLAMA\_CUBLAS=1 make -j

* Apple Silicon의 경우 Metal이 기본적으로 활성화되어 있습니다.

!make clean && make -j

**텍스트 완성**

인수를 사용하여 추론에만 CPU를 사용하고 모든 계층이 GPU로 오프로드되도록 합니다.-ngl 0-ngl 10000

!./main -ngl 10000 -m ./models/8B-v3/ggml-model-Q4\_K\_M.gguf --color --temp 1.1 --repeat\_penalty 1.1 -c 0 -n 1024 -e -s 0 -p """\

First Citizen:\n\n\

Before we proceed any further, hear me speak.\n\n\

\n\n\

All:\n\n\

Speak, speak.\n\n\

\n\n\

First Citizen:\n\n\

You are all resolved rather to die than to famish?\n\n\

\n\n\

All:\n\n\

Resolved. resolved.\n\n\

\n\n\

First Citizen:\n\n\

First, you know Caius Marcius is chief enemy to the people.\n\n\

\n\n\

All:\n\n\

We know't, we know't.\n\n\

\n\n\

First Citizen:\n\n\

Let us kill him, and we'll have corn at our own price. Is't a verdict?\n\n\

\n\n\

All:\n\n\

No more talking on't; let it be done: away, away!\n\n\

\n\n\

Second Citizen:\n\n\

One word, good citizens.\n\n\

\n\n\

First Citizen:\n\n\

We are accounted poor citizens, the patricians good. What authority surfeits on would relieve us: if they would yield us but the superfluity, \

while it were wholesome, we might guess they relieved us humanely; but they think we are too dear: the leanness that afflicts us, the object of \

our misery, is as an inventory to particularise their abundance; our sufferance is a gain to them Let us revenge this with our pikes, \

ere we become rakes: for the gods know I speak this in hunger for bread, not in thirst for revenge.\n\n\

\n\n\

"""

**메모:** Apple Silicon의 경우 결과를 확인하여 GPU에 할당할 수 있는 메모리 양을 확인하고 성능을 유지합니다. 현재 **32GB** M1 Max의 GPU에는 통합 메모리의 70%만 할당할 수 있으며, 더 큰 메모리의 GPU에 사용 가능한 메모리의 약 **78%**를 할당할 것으로 예상됩니다. (출처: <https://developer.apple.com/videos/play/tech-talks/10580/?time=346>) 전체 메모리를 활용하려면 추론에 CPU만 사용합니다. (감사합니다: [ggerganov/llama.cpp#1826](https://github.com/ggerganov/llama.cpp/pull/1826)recommendedMaxWorkingSetSize-ngl 0)

**LLaMA 3🦙🦙🦙용 채팅 템플릿**

!./main -ngl 10000 -m ./models/8B-v3-instruct/ggml-model-Q4\_K\_M.gguf --color -c 0 -n -2 -e -s 0 --mirostat 2 -i --no-display-prompt --keep -1 \

-r '<|eot\_id|>' -p '<|begin\_of\_text|><|start\_header\_id|>system<|end\_header\_id|>\n\nYou are a helpful assistant.<|eot\_id|><|start\_header\_id|>user<|end\_header\_id|>\n\nHi!<|eot\_id|><|start\_header\_id|>assistant<|end\_header\_id|>\n\n' \

--in-prefix '<|start\_header\_id|>user<|end\_header\_id|>\n\n' --in-suffix '<|eot\_id|><|start\_header\_id|>assistant<|end\_header\_id|>\n\n'

**벤치 마크**

!./llama-bench -p 512,1024,4096,8192 -n 512,1024,4096,8192 -m ./models/8B-v3/ggml-model-Q4\_K\_M.gguf

**총 VRAM 요구 사항**

| **모델** | **양자화된 크기(Q4\_K\_M)** | **원본 크기(f16)** |
| --- | --- | --- |
| 8ᄂ | 4.58 기가바이트 | 14.96 기가바이트 |
| 70ᄂ | 39.59 기가바이트 | 131.42 기가바이트 |

이 도구를 사용하여 VRAM 요구 사항을 추정 할 수 있습니다 : [LLM RAM 계산기](https://llm-calc.rayfernando.ai/)

**LLaMA 3 70B의 혼란 표**

혼란은 적을수록 좋습니다. (크레딧: [dranger003](https://github.com/ggerganov/llama.cpp/pull/6745#issuecomment-2093892514))

| **양자화** | **크기(GiB)** | **당혹감 (wiki.test)** | **델타(FP16)** |
| --- | --- | --- | --- |
| IQ1\_S | 14.29 | 9.8655 +/- 0.0625 | 248.51% |
| IQ1\_M | 15.60 | 8.5193 +/- 0.0530 | 201.94% |
| IQ2\_XXS | 17.79 | 6.6705 +/- 0.0405 | 135.64% |
| IQ2\_XS | 19.69 | 5.7486 +/- 0.0345 | 103.07% |
| IQ2\_S | 20.71 | 5.5215 +/- 0.0318 | 95.05% |
| Q2\_K\_S | 22.79 | 5.4334 +/- 0.0325 | 91.94% |
| IQ2\_M | 22.46 | 4.8959 +/- 0.0276 | 72.35% |
| Q2\_K | 24.56 | 4.7763 +/- 0.0274 | 68.73% |
| IQ3\_XXS | 25.58 | 3.9671 +/- 0.0211 | 40.14% |
| IQ3\_XS | 27.29 | 3.7210 +/- 0.0191 | 31.45% |
| Q3\_K\_S | 28.79 | 3.6502 +/- 0.0192 | 28.95% |
| IQ3\_S | 28.79 | 3.4698 +/- 0.0174 | 22.57% |
| IQ3\_M | 29.74 | 3.4402 +/- 0.0171 | 21.53% |
| Q3\_K\_M | 31.91 | 3.3617 +/- 0.0172 | 18.75% |
| Q3\_K\_L | 34.59 | 3.3016 +/- 0.0168 | 16.63% |
| IQ4\_XS | 35.30 | 3.0310 +/- 0.0149 | 7.07% |
| IQ4\_NL | 37.30 | 3.0261 +/- 0.0149 | 6.90% |
| Q4\_K\_S | 37.58 | 3.0050 +/- 0.0148 | 6.15% |
| Q4\_K\_M | 39.60 | 2.9674 +/- 0.0146 | 4.83% |
| Q5\_K\_S | 45.32 | 2.8843 +/- 0.0141 | 1.89% |
| Q5\_K\_M | 46.52 | 2.8656 +/- 0.0139 | 1.23% |
| Q6\_K | 53.91 | 2.8441 +/- 0.0138 | 0.47% |
| Q8\_0 | 69.83 | 2.8316 +/- 0.0138 | 0.03% |
| F16 키 | 131.43 | 2.8308 +/- 0.0138 | 0.00% |

**벤치 마크**

TG "텍스트 생성"을 의미하며 "프롬프트 처리"를 의미합니다. # 생성/처리 중인 총 토큰의 경우. 메모리 부족을 의미합니다. 평균 속도(토큰/초)입니다.PPOOM

**법학 3 🦙🦙🦙 :**

**NVIDIA 게임용 GPU(OS: Ubuntu 22.04.2 LTS, pytorch:2.2.0, py: 3.10, cuda: RunPod의 12.1.1)(2024년 5월 스냅샷)**

| **그래픽 카드(GPU** | **모델** | **TG 512 TG 512 TG 512 TG** | **TG 1024년** | **TG 4096 TG 4096 TG** | **TG 8192 TG 8192 TG 8192 TG** | **512쪽** | **1024쪽** | **쪽 4096** | **쪽 8192** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 3070 8기가바이트 | 8B Q4\_K\_M | 72.79 | 70.94 | 67.01 | 61.64 | 2402.51 | 2283.62 | 1826.59 | 1419.97 |
| 3080 10기가바이트 | 8B Q4\_K\_M | 109.57 | 106.40 | 98.67 | 89.90 | 3728.86 | 3557.02 | 2852.06 | 2232.21 |
| 3080 Ti 12기가바이트 | 8B Q4\_K\_M | 110.60 | 106.71 | 98.34 | 88.63 | 3690.30 | 3556.67 | 2947.11 | 2381.52 |
| 4070 Ti 12기가바이트 | 8B Q4\_K\_M | 83.50 | 82.21 | 78.59 | 73.46 | 3936.29 | 3653.07 | 2729.71 | 2019.71 |
| 4080 16기가바이트 | 8B Q4\_K\_M | 108.15 | 106.22 | 100.44 | 93.71 | 5389.74 | 5064.99 | 3790.96 | 2882.03 |
|  | 8B F16 | 40.58 | 40.29 | 39.44 | 움 | 7246.97 | 6758.90 | 4720.22 | 움 |
| 3090 24기가바이트 | 8B Q4\_K\_M | 115.42 | 111.74 | 97.31 | 87.49 | 4030.40 | 3865.39 | 3169.91 | 2527.40 |
|  | 8B F16 | 47.40 | 46.51 | 44.79 | 42.62 | 4444.65 | 4239.64 | 3410.47 | 2667.14 |
| 4090 24기가바이트 | 8B Q4\_K\_M | 130.58 | 127.74 | 119.44 | 110.66 | 7138.99 | 6898.71 | 5265.68 | 4039.68 |
|  | 8B F16 | 54.84 | 54.34 | 52.63 | 50.88 | 9382.00 | 9056.26 | 6531.36 | 4744.18 |
| 3090 24GB \* 2 | 8B Q4\_K\_M | 111.67 | 108.07 | 99.60 | 90.77 | 3336.37 | 4004.14 | 4013.34 | 3433.59 |
|  | 8B F16 | 47.72 | 47.15 | 45.56 | 43.61 | 4122.66 | 4690.50 | 4788.60 | 3851.37 |
|  | 70B Q4\_K\_M | 16.57 | 16.29 | 15.36 | 14.34 | 357.32 | 393.89 | 379.52 | 338.82 |
| 4090 24기가바이트 \* 2 | 8B Q4\_K\_M | 124.65 | 122.56 | 114.32 | 106.18 | 7003.51 | 8545.00 | 8422.04 | 6895.68 |
|  | 8B F16 | 53.64 | 53.27 | 51.64 | 49.83 | 9177.92 | 11094.51 | 10329.29 | 8067.29 |
|  | 70B Q4\_K\_M | 19.22 | 19.06 | 18.54 | 17.92 | 839.43 | 905.38 | 846.38 | 723.24 |
| 3090 24GB \* 4 | 8B Q4\_K\_M | 108.66 | 104.94 | 97.09 | 88.35 | 3742.66 | 4653.93 | 5826.91 | 4913.40 |
|  | 8B F16 | 47.07 | 46.40 | 44.76 | 42.81 | 4608.40 | 5713.41 | 6596.17 | 5361.52 |
|  | 70B Q4\_K\_M | 17.07 | 16.89 | 16.24 | 15.39 | 300.79 | 350.06 | 367.75 | 331.37 |
| 4090 24GB \* 4 | 8B Q4\_K\_M | 120.32 | 117.61 | 110.52 | 103.13 | 6748.96 | 9609.29 | 12491.10 | 10993.75 |
|  | 8B F16 | 53.10 | 52.69 | 51.00 | 49.21 | 8750.57 | 12304.19 | 15143.84 | 12919.74 |
|  | 70B Q4\_K\_M | 19.80 | 18.83 | 18.35 | 17.66 | 834.74 | 898.17 | 839.97 | 718.01 |
| 3090 24GB \* 6 | 8B Q4\_K\_M | 104.17 | 101.07 | 94.06 | 85.93 | 3359.99 | 5153.05 | 7690.65 | 7084.44 |
|  | 8B F16 | 46.23 | 45.55 | 43.99 | 42.15 | 3875.97 | 5952.55 | 9437.91 | 8780.49 |
|  | 70B Q4\_K\_M | 17.09 | 16.93 | 16.32 | 15.45 | 456.95 | 739.40 | 786.79 | 695.44 |
|  | 70B F16 | 5.85 | 5.82 | 5.76 | 5.53 | 579.00 | 927.23 | 998.79 | 813.99 |
| 4090 24GB \* 8 | 8B Q4\_K\_M | 118.09 | 116.13 | 108.37 | 100.95 | 6172.06 | 9706.82 | 15089.45 | 13802.08 |
|  | 8B F16 | 52.51 | 52.12 | 50.39 | 48.72 | 7889.26 | 11818.92 | 16462.18 | 14300.98 |
|  | 70B Q4\_K\_M | 18.94 | 18.76 | 18.23 | 17.57 | 812.95 | 1336.26 | 1488.36 | 1320.36 |
|  | 70B F16 | 6.47 | 6.45 | 6.39 | 6.31 | 1183.87 | 1890.48 | 2311.43 | 1995.85 |

**NVIDIA Professional GPU(OS: Ubuntu 22.04.2 LTS, pytorch:2.2.0, py: 3.10, cuda: RunPod의 경우 12.1.1)(2024년 5월 스냅샷)**

| **그래픽 카드(GPU** | **모델** | **TG 512 TG 512 TG 512 TG** | **TG 1024년** | **TG 4096 TG 4096 TG** | **TG 8192 TG 8192 TG 8192 TG** | **512쪽** | **1024쪽** | **쪽 4096** | **쪽 8192** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| RTX 4000 에이다 20GB | 8B Q4\_K\_M | 59.15 | 58.59 | 55.94 | 52.39 | 2451.93 | 2310.53 | 1798.01 | 1337.15 |
|  | 8B F16 | 20.92 | 20.85 | 20.50 | 20.01 | 3121.67 | 2951.87 | 2200.58 | 1557.00 |
| RTX 5000 에이다 32GB | 8B Q4\_K\_M | 91.39 | 89.87 | 85.01 | 80.00 | 4761.12 | 4467.46 | 3272.94 | 2422.33 |
|  | 8B F16 | 32.84 | 32.67 | 32.04 | 31.27 | 6160.57 | 5835.41 | 4008.30 | 2808.89 |
| RTX A6000 48GB | 8B Q4\_K\_M | 105.39 | 102.22 | 94.82 | 86.73 | 3780.55 | 3621.81 | 2917.23 | 2292.61 |
|  | 8B F16 | 40.71 | 40.25 | 39.14 | 37.73 | 4511.02 | 4315.18 | 3365.79 | 2566.46 |
|  | 70B Q4\_K\_M | 14.71 | 14.58 | 14.09 | 13.42 | 482.19 | 466.82 | 404.61 | 340.73 |
| RTX 6000 에이다 48GB | 8B Q4\_K\_M | 133.44 | 130.99 | 120.74 | 111.57 | 5791.74 | 5560.94 | 4495.19 | 3542.57 |
|  | 8B F16 | 52.32 | 51.97 | 50.21 | 48.79 | 6663.13 | 6205.44 | 4969.46 | 3915.81 |
|  | 70B Q4\_K\_M | 18.52 | 18.36 | 17.80 | 16.97 | 565.98 | 547.03 | 481.59 | 419.76 |
| A40 48GB | 8B Q4\_K\_M | 91.27 | 88.95 | 83.10 | 76.45 | 3324.98 | 3240.95 | 2586.50 | 2013.34 |
|  | 8B F16 | 34.26 | 33.95 | 33.06 | 31.93 | 4203.75 | 4043.05 | 3069.98 | 2295.02 |
|  | 70B Q4\_K\_M | 11.60 | 12.08 | 11.68 | 11.26 | 209.38 | 239.92 | 268.89 | 291.13 |
| L40S 48GB | 8B Q4\_K\_M | 115.55 | 113.60 | 105.50 | 97.98 | 6035.24 | 5908.52 | 4335.18 | 3192.70 |
|  | 8B F16 | 43.69 | 43.42 | 42.22 | 41.05 | 2253.93 | 2491.65 | 2887.70 | 3312.16 |
|  | 70B Q4\_K\_M | 15.46 | 15.31 | 14.92 | 14.45 | 673.63 | 649.08 | 542.29 | 446.48 |
| RTX 4000 에이다 20GB \* 4 | 8B Q4\_K\_M | 56.64 | 56.14 | 53.58 | 50.19 | 2413.07 | 3369.24 | 4404.45 | 3733.15 |
|  | 8B F16 | 20.65 | 20.58 | 20.24 | 19.74 | 3220.21 | 4366.64 | 5366.39 | 4323.70 |
|  | 70B Q4\_K\_M | 7.36 | 7.33 | 7.12 | 6.84 | 282.28 | 306.44 | 290.70 | 243.45 |
| A100 PCIe 80GB | 8B Q4\_K\_M | 140.62 | 138.31 | 127.22 | 117.60 | 5981.04 | 5800.48 | 4959.84 | 4083.37 |
|  | 8B F16 | 54.84 | 54.56 | 53.02 | 51.24 | 7741.34 | 7504.24 | 6137.54 | 4849.11 |
|  | 70B Q4\_K\_M | 22.31 | 22.11 | 20.93 | 19.53 | 744.12 | 726.65 | 653.20 | 573.95 |
| A100 SXM 80GB | 8B Q4\_K\_M | 135.04 | 133.38 | 125.09 | 115.92 | 5947.64 | 5863.92 | 5121.60 | 4137.08 |
|  | 8B F16 | 53.49 | 53.18 | 52.03 | 50.52 | 603.76 | 681.47 | 866.13 | 1323.07 |
|  | 70B Q4\_K\_M | 24.61 | 24.33 | 22.91 | 21.32 | 817.58 | 796.81 | 714.07 | 625.66 |
| H100 PCIe 80GB | 8B Q4\_K\_M | 145.55 | 144.49 | 136.06 | 126.83 | 8125.45 | 7760.16 | 6423.31 | 5185.03 |
|  | 8B F16 | 68.03 | 67.79 | 65.97 | 63.55 | 10815.51 | 10342.63 | 8106.53 | 6191.45 |
|  | 70B Q4\_K\_M | 25.03 | 25.01 | 23.82 | 22.39 | 1012.73 | 984.06 | 863.37 | 741.52 |
| RTX 5000 에이다 32GB \* 4 | 8B Q4\_K\_M | 84.07 | 82.73 | 78.45 | 74.11 | 4671.34 | 6530.78 | 8004.94 | 6790.82 |
|  | 8B F16 | 32.10 | 31.94 | 31.32 | 30.58 | 2427.96 | 2877.66 | 3836.89 | 5235.00 |
|  | 70B Q4\_K\_M | 11.51 | 11.45 | 11.24 | 10.94 | 502.37 | 541.54 | 504.23 | 424.29 |
| RTX A6000 48GB \* 4 | 8B Q4\_K\_M | 96.48 | 93.73 | 87.72 | 80.88 | 3712.99 | 5340.10 | 7126.45 | 6438.82 |
|  | 8B F16 | 39.34 | 38.87 | 37.81 | 36.51 | 4508.60 | 6448.85 | 8327.16 | 7298.18 |
|  | 70B Q4\_K\_M | 14.44 | 14.32 | 13.91 | 13.32 | 496.08 | 539.20 | 511.22 | 434.31 |
|  | 70B F16 | 4.76 | 4.74 | 4.70 | 4.63 | 510.31 | 792.23 | 751.37 | 748.06 |
| RTX 6000 에이다 48GB \* 4 | 8B Q4\_K\_M | 121.21 | 118.99 | 110.65 | 103.18 | 6640.86 | 9679.55 | 11734.85 | 10278.14 |
|  | 8B F16 | 50.61 | 50.25 | 48.69 | 47.18 | 8953.30 | 12637.94 | 13971.34 | 11702.36 |
|  | 70B Q4\_K\_M | 18.13 | 17.96 | 17.49 | 16.89 | 656.61 | 714.93 | 697.10 | 612.54 |
|  | 70B F16 | 6.08 | 6.06 | 6.01 | 5.94 | 864.12 | 1270.39 | 1363.75 | 1182.28 |
| A40 48GB \* 4 | 8B Q4\_K\_M | 85.91 | 83.79 | 78.56 | 72.70 | 3321.27 | 4841.98 | 6442.38 | 5742.84 |
|  | 8B F16 | 33.60 | 33.28 | 32.42 | 31.38 | 4144.88 | 5931.06 | 7544.92 | 6516.60 |
|  | 70B Q4\_K\_M | 11.99 | 11.91 | 11.60 | 11.17 | 236.86 | 263.36 | 300.57 | 312.31 |
|  | 70B F16 | 3.99 | 3.98 | 3.95 | 3.90 | 610.51 | 900.79 | 893.28 | 735.16 |
| L40S 48GB \* 4 | 8B Q4\_K\_M | 107.53 | 105.72 | 98.59 | 92.20 | 6125.69 | 9008.27 | 10566.97 | 9017.90 |
|  | 8B F16 | 42.70 | 42.48 | 41.33 | 40.19 | 2211.45 | 2541.61 | 3093.33 | 4336.81 |
|  | 70B Q4\_K\_M | 15.12 | 14.99 | 14.63 | 14.17 | 591.05 | 634.05 | 605.66 | 541.67 |
|  | 70B F16 | 5.05 | 5.03 | 4.99 | 4.94 | 1042.13 | 1478.83 | 1427.77 | 1150.63 |
| A100 PCIe 80GB \* 4 | 8B Q4\_K\_M | 119.28 | 117.30 | 110.75 | 103.87 | 6076.58 | 8889.35 | 12724.54 | 11803.39 |
|  | 8B F16 | 51.63 | 51.54 | 50.20 | 48.73 | 8088.79 | 11670.74 | 16025.11 | 14269.17 |
|  | 70B Q4\_K\_M | 22.91 | 22.68 | 21.41 | 19.96 | 771.28 | 978.06 | 1138.60 | 1043.15 |
|  | 70B F16 | 7.40 | 7.38 | 7.23 | 7.06 | 1172.14 | 1733.41 | 1846.36 | 1592.37 |
| A100 SXM 80GB \* 4 | 8B Q4\_K\_M | 99.73 | 97.70 | 92.09 | 86.27 | 4850.88 | 7782.25 | 12242.53 | 11535.66 |
|  | 8B F16 | 45.53 | 45.45 | 44.33 | 43.09 | 626.75 | 674.11 | 1003.37 | 1612.05 |
|  | 70B Q4\_K\_M | 19.87 | 19.60 | 18.48 | 17.19 | 468.86 | 539.08 | 712.08 | 802.23 |
|  | 70B F16 | 6.95 | 6.92 | 6.77 | 6.58 | 1233.31 | 1834.16 | 1972.48 | 1699.56 |
| H100 PCIe 80GB \* 4 | 8B Q4\_K\_M | 123.08 | 118.14 | 113.12 | 110.34 | 8054.58 | 11560.23 | 16128.27 | 14682.97 |
|  | 8B F16 | 64.00 | 62.90 | 61.45 | 59.72 | 11107.40 | 15612.81 | 20561.03 | 17762.96 |
|  | 70B Q4\_K\_M | 26.40 | 26.20 | 24.60 | 23.68 | 1048.29 | 1133.23 | 1088.99 | 950.92 |
|  | 70B F16 | 9.67 | 9.63 | 9.46 | 9.23 | 1681.45 | 2420.10 | 2437.53 | 2031.77 |

**Apple Silicon(2024년 5월 스냅샷)**

| **그래픽 카드(GPU** | **모델** | **TG 512 TG 512 TG 512 TG** | **TG 1024년** | **TG 4096 TG 4096 TG** | **TG 8192 TG 8192 TG 8192 TG** | **512쪽** | **1024쪽** | **쪽 4096** | **쪽 8192** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| M1 7코어 GPU 8GB | 8B Q4\_K\_M | 10.20 | 9.72 | 11.77 | 움 | 94.48 | 87.26 | 96.53 | 움 |
| M1 Max 32코어 GPU 64GB | 8B Q4\_K\_M | 35.73 | 34.49 | 31.18 | 26.84 | 408.23 | 355.45 | 329.84 | 302.92 |
|  | 8B F16 | 18.75 | 18.43 | 16.33 | 15.03 | 517.34 | 418.77 | 374.09 | 351.46 |
|  | 70B Q4\_K\_M | 4.34 | 4.09 | 4.09 | 3.71 | 34.96 | 33.01 | 32.64 | 30.97 |
| M2 울트라 76코어 GPU 192GB | 8B Q4\_K\_M | 78.81 | 76.28 | 64.58 | 54.13 | 994.04 | 1023.89 | 979.47 | 913.55 |
|  | 8B F16 | 36.90 | 36.25 | 33.67 | 30.68 | 1175.40 | 1202.74 | 1194.21 | 1103.44 |
|  | 70B Q4\_K\_M | 12.48 | 12.13 | 10.75 | 9.34 | 118.79 | 117.76 | 109.53 | 108.57 |
|  | 70B F16 | 4.76 | 4.71 | 4.48 | 4.23 | 147.58 | 145.82 | 133.75 | 135.15 |
| M3 Max 40코어 GPU 64GB | 8B Q4\_K\_M | 48.97 | 50.74 | 44.21 | 36.12 | 693.32 | 678.04 | 573.09 | 505.32 |
|  | 8B F16 | 22.04 | 22.39 | 20.72 | 18.74 | 769.84 | 751.49 | 609.97 | 515.15 |
|  | 70B Q4\_K\_M | 7.65 | 7.53 | 6.58 | 5.60 | 70.19 | 62.88 | 64.90 | 61.96 |

**결론**

동일한 크기 및 양자화 모델에서 동일한 성능을 제공합니다. 여러 NVIDIA GPU는 텍스트 생성 성능에 영향을 줄 수 있지만 여전히 프롬프트 처리 속도를 높일 수 있습니다.

NVIDIA 게임용 GPU를 구입하여 비용을 절약하십시오. 비즈니스를 위한 전문가용 GPU를 구입하십시오. 컴퓨터를 책상 위에 올려놓고, 에너지를 절약하고, 조용히 하고, 유지 관리를 원하지 않고, 더 재미있게 즐기고 싶다면 Mac을 구입하십시오. 😇

이 정보가 도움이 되었다면 별을 주세요. ⭐️ 조언이 있으시면 언제든지 저에게 연락하십시오. 감사합니다. 🤗