|  |  |  |  |
| --- | --- | --- | --- |
| **Step** | **Metric** | **How We’ll Measure** | **Notes** |
| 1 | **VRAM before load** | nvidia-smi query | Baseline memory |
| 2 | **Cold start time** | PowerShell Measure-Command with ollama run | First prompt after load |
| 3 | **Warm run time** | Run same prompt again | Model already in memory |
| 4 | **Tokens/sec** | Call your existing TPS Python script with model name | Will reuse current prompt list |
| 5 | **VRAM after load** | nvidia-smi query | Post-load usage |
| 6 | **Log results** | Append to CSV | Model name, VRAM, times, TPS |

**common protocol benchmark**: (ollama based models)

| **Model** | **VRAM Before (MB)** | **VRAM After (MB)** | **Cold Start (s)** | **Warm Start (s)** | **Tokens/sec** |
| --- | --- | --- | --- | --- | --- |
| llama3 | 0 | 4537 | 13.29 | 2.59 | 8.26 |
| gpt-oss:20b | 4287 | 4287 | 13.15 | 10.86 | 1.81 |

**Quick takeaway:**

* **llama3** loads fresh into GPU memory (0 → 4.5 GB) and runs **~4.5× faster** in tokens/sec than gpt-oss:20b.
* **gpt-oss:20b** has similar cold start time but much slower generation speed.

Llama3

| **Metric** | **Run 1** | **Run 2** | **Run 3** | **Avg** |
| --- | --- | --- | --- | --- |
| VRAM Before (MB) | 517 | 4909 | 4909 | 3445 |
| Cold Start (sec) | 14.38 | 3.57 | 2.90 | 6.95 |
| VRAM After (MB) | 4901 | 4909 | 4909 | 4906 |
| Warm Start (sec) | 2.53 | 2.90 | 2.88 | 2.77 |
| Tokens/sec | 8.38 | 6.97 | 8.07 | 7.81 |

Gpt-oss:20B

| **Metric** | **Run 1** | **Run 2** | **Run 3** | **Avg** |
| --- | --- | --- | --- | --- |
| VRAM Before (MB) | 4909 | 4761 | 4771 | 4814 |
| Cold Start (sec) | 75.49 | 19.03 | 18.22 | 37.58 |
| VRAM After (MB) | 4713 | 4761 | 4771 | 4748 |
| Warm Start (sec) | 21.19 | 10.20 | 22.66 | 18.02 |
| Tokens/sec | 1.80 | 1.90 | 1.86 | 1.85 |

**⚡ Key Takeaways**

* **Llama3** is much faster in cold start (avg **6.95s** vs **37.58s**) and generates **~4× more tokens/sec** (7.81 vs 1.85).
* **gpt-oss:20b** has **huge cold start variability** — first run takes 75s, but later runs drop to ~19s.
* VRAM usage is similar after load (~4.7–4.9 GB), but llama3 starts with far less memory in the first run.
* Warm start times are consistently **~2.8s for llama3** vs **~18s for gpt-oss:20b**.

(base) C:\AI Model Benchmarking>python first\_token\_latency.py llama3

First token latency: 10.064 sec

(base) C:\AI Model Benchmarking>python first\_token\_latency.py gpt-oss:20b

First token latency: 38.687 sec

(base) C:\AI Model Benchmarking>python generate\_report.py

✅ Markdown table saved to benchmark\_report.md

| Model | Cold Start (s) | Warm Start (s) | TPS | VRAM Before (MB) | VRAM After (MB) |

|-------|----------------|----------------|-----|------------------|-----------------|

| gpt-oss:20b | 31.47 ± 29.46 | 16.23 ± 6.61 | 1.84 ± 0.05 | 4682.0 ± 271.86 | 4633.0 ± 232.05 |

| llama3 | 6.95 ± 6.44 | 2.77 ± 0.21 | 7.81 ± 0.74 | 3445.0 ± 2535.72 | 4906.33 ± 4.62 |

📊 Plot saved to benchmark\_plot.png

(base) C:\AI Model Benchmarking>

**🧪 AI Model Benchmark: llama3 vs gpt-oss:20b**

**Performance Summary**

| **Model** | **Cold Start (s)** | **Warm Start (s)** | **First Token Latency (s)** | **TPS** | **VRAM Before (MB)** | **VRAM After (MB)** |
| --- | --- | --- | --- | --- | --- | --- |
| **llama3** | **6.95 ± 6.44** | **2.77 ± 0.21** | **10.06** | **7.81 ± 0.74** | 3445.0 ± 2535.72 | 4906.33 ± 4.62 |
| **gpt-oss:20b** | 31.47 ± 29.46 | 16.23 ± 6.61 | 38.69 | 1.84 ± 0.05 | 4682.0 ± 271.86 | 4633.0 ± 232.05 |