To meet this difficulty point, I decided to explore two main ways to serve our LegalAI model: using a standard **server-grade CPU** and using a **server-grade GPU**. Our goal was to see which one works better for our specific application in terms of speed and how many users it can handle, and to think about the costs.

By setting up and testing our LegalAI application on both a CPU virtual machine (an `m1.large` at KVM@TACC) and a GPU virtual machine (an RTX 6000 instance at CHI@UC, using the ONNX model for tests), we learned a lot about how each performs.Below is the data I found for the CPU and GPU inferences.

For CPU:

Single file search:

```
legal-search-api-dummy  | INFO:src.api.main:Processing uploaded file: Campoli v. Anywhere real Est. Inc., 2025 U.S. Dist. LEXIS 68458.PDF
legal-search-api-dummy  | INFO:src.api.main:Split uploaded file into 8 chunks.
legal-search-api-dummy  | INFO:src.api.main:Embedding 8 file chunks using ONNX on cpu
legal-search-api-dummy  | INFO:src.api.main:File query embedding (8 chunks) took 3.492477s.
legal-search-api-dummy  | INFO:src.api.main:Searching for file chunk 1/8
legal-search-api-dummy  | INFO:src.api.main:FAISS search took 0.014770s for k=3
legal-search-api-dummy  | INFO:src.api.main:Result aggregation took 0.000067s, aggregated to 2 docs.
legal-search-api-dummy  | INFO:src.api.main:Searching for file chunk 2/8
legal-search-api-dummy  | INFO:src.api.main:FAISS search took 0.006755s for k=3
legal-search-api-dummy  | INFO:src.api.main:Result aggregation took 0.000037s, aggregated to 1 docs.
legal-search-api-dummy  | INFO:src.api.main:Searching for file chunk 3/8
legal-search-api-dummy  | INFO:src.api.main:FAISS search took 0.006851s for k=3
legal-search-api-dummy  | INFO:src.api.main:Result aggregation took 0.000035s, aggregated to 2 docs.
legal-search-api-dummy  | INFO:src.api.main:Searching for file chunk 4/8
legal-search-api-dummy  | INFO:src.api.main:FAISS search took 0.006985s for k=3
legal-search-api-dummy  | INFO:src.api.main:Result aggregation took 0.000046s, aggregated to 2 docs.
legal-search-api-dummy  | INFO:src.api.main:Searching for file chunk 5/8
legal-search-api-dummy  | INFO:src.api.main:FAISS search took 0.007031s for k=3
legal-search-api-dummy  | INFO:src.api.main:Result aggregation took 0.000038s, aggregated to 2 docs.
legal-search-api-dummy  | INFO:src.api.main:Searching for file chunk 6/8
legal-search-api-dummy  | INFO:src.api.main:FAISS search took 0.007017s for k=3
legal-search-api-dummy  | INFO:src.api.main:Result aggregation took 0.000075s, aggregated to 3 docs.
legal-search-api-dummy  | INFO:src.api.main:Searching for file chunk 7/8
legal-search-api-dummy  | INFO:src.api.main:FAISS search took 0.006956s for k=3
legal-search-api-dummy  | INFO:src.api.main:Result aggregation took 0.000038s, aggregated to 2 docs.
legal-search-api-dummy  | INFO:src.api.main:Searching for file chunk 8/8
legal-search-api-dummy  | INFO:src.api.main:FAISS search took 0.007103s for k=3
legal-search-api-dummy  | INFO:src.api.main:Result aggregation took 0.000068s, aggregated to 3 docs.
legal-search-api-dummy  | INFO:src.api.main:Preparing to return 2 results with pre-computed summaries.
legal-search-api-dummy  | INFO:src.api.main:Result formatting took 0.000079s for 2 docs.
legal-search-api-dummy  | INFO:src.api.main:Core search logic for ONNX on cpu finished in 3.5863s. Returning 2 documents.
legal-search-api-dummy  | INFO:     216.165.95.147:59564 - "POST /search_combined HTTP/1.1" 200 OK
legal-search-api-dummy  | INFO:     172.18.0.2:40210 - "GET /metrics HTTP/1.1" 200 OK
legal-search-api-dummy  | INFO:     172.18.0.2:56292 - "GET /metrics HTTP/1.1" 200 OK
```

```
legal-search-api-dummy  | INFO:     172.18.0.2:43904 - "GET /metrics HTTP/1.1" 200 OK
```

Load testing :

100 requests (5 concurrent)

```
sanju@Sanjeevans-MacBook-Air serving_dummy % python3 src/test/load_test_api.py --concurrent 5 --total_requests 100
2025-05-11 18:05:30,620 - INFO - Starting load test: URL='http://129.114.24.228:8000/search_combined',
PDF='real_pdfs/Glover v. State.PDF', Concurrent=5, Total=100, TopK=1
2025-05-11 18:05:46,555 - ERROR - Request 8 failed with unexpected error: [Errno 54] Connection reset by peer
2025-05-11 18:05:46,556 - ERROR - Request 7 failed with unexpected error: [Errno 54] Connection reset by peer
2025-05-11 18:06:07,135 - ERROR - Request 17 failed with unexpected error: [Errno 54] Connection reset by peer
2025-05-11 18:06:31,098 - ERROR - Request 28 failed with unexpected error: [Errno 54] Connection reset by peer
2025-05-11 18:06:55,064 - ERROR - Request 39 failed with unexpected error: [Errno 54] Connection reset by peer
2025-05-11 18:07:18,918 - ERROR - Request 47 failed with unexpected error: [Errno 54] Connection reset by peer
2025-05-11 18:07:42,675 - ERROR - Request 57 failed with unexpected error: [Errno 54] Connection reset by peer
2025-05-11 18:08:06,816 - ERROR - Request 67 failed with unexpected error: [Errno 54] Connection reset by peer
2025-05-11 18:08:31,212 - ERROR - Request 78 failed with unexpected error: [Errno 54] Connection reset by peer
2025-05-11 18:08:55,420 - ERROR - Request 89 failed with unexpected error: [Errno 54] Connection reset by peer
2025-05-11 18:09:19,387 - ERROR - Request 98 failed with unexpected error: [Errno 54] Connection reset by peer
2025-05-11 18:09:27,122 - INFO - Load Test Summary:
2025-05-11 18:09:27,123 - INFO -   Total requests sent: 100
2025-05-11 18:09:27,123 - INFO -   Successful requests: 89
2025-05-11 18:09:27,123 - INFO -   Failed requests: 11
2025-05-11 18:09:27,123 - INFO -   Total time taken: 236.5006 seconds
2025-05-11 18:09:27,123 - INFO -   Requests per second (RPS): 0.4228
2025-05-11 18:09:27,123 - INFO -   Avg Latency (successful): 12.4627s
2025-05-11 18:09:27,123 - INFO -   Min Latency (successful): 2.8736s
2025-05-11 18:09:27,123 - INFO -   Max Latency (successful): 20.6274s
2025-05-11 18:09:27,123 - INFO -   P95 Latency (successful): 18.7227s
```

100 requests (10 concurrent)

```
sanju@Sanjeevans-MacBook-Air serving_dummy % python3 src/test/load_test_api.py --concurrent 10 --total_requests 100
2025-05-11 18:10:39,924 - INFO - Starting load test: URL='http://129.114.24.228:8000/search_combined',
PDF='real_pdfs/Glover v. State.PDF', Concurrent=10, Total=100, TopK=1
2025-05-11 18:11:08,912 - ERROR - Request 12 failed with unexpected error: [Errno 54] Connection reset by peer
2025-05-11 18:11:08,915 - ERROR - Request 13 failed with unexpected error: [Errno 54] Connection reset by peer
2025-05-11 18:11:08,915 - ERROR - Request 16 failed with unexpected error: [Errno 54] Connection reset by peer
2025-05-11 18:11:08,915 - ERROR - Request 15 failed with unexpected error: [Errno 54] Connection reset by peer
2025-05-11 18:11:08,915 - ERROR - Request 14 failed with unexpected error: [Errno 54] Connection reset by peer
2025-05-11 18:11:08,915 - ERROR - Request 20 failed with unexpected error: [Errno 54] Connection reset by peer
2025-05-11 18:11:08,916 - ERROR - Request 18 failed with unexpected error: [Errno 54] Connection reset by peer
2025-05-11 18:11:29,990 - ERROR - Request 34 failed with unexpected error: [Errno 54] Connection reset by peer
2025-05-11 18:11:45,905 - ERROR - Request 38 failed with unexpected error: [Errno 54] Connection reset by peer
2025-05-11 18:11:53,910 - ERROR - Request 43 failed with unexpected error: [Errno 54] Connection reset by peer
2025-05-11 18:12:01,954 - ERROR - Request 46 failed with unexpected error: [Errno 54] Connection reset by peer
2025-05-11 18:12:09,941 - ERROR - Request 50 failed with unexpected error: [Errno 54] Connection reset by peer
2025-05-11 18:12:15,267 - ERROR - Request 54 failed with unexpected error: [Errno 54] Connection reset by peer
2025-05-11 18:12:28,590 - ERROR - Request 62 failed with unexpected error: [Errno 54] Connection reset by peer
```

```
2025-05-11 18:12:41,891 - ERROR - Request 67 failed with unexpected error: [Errno 54] Connection reset by peer
2025-05-11 18:12:57,861 - ERROR - Request 72 failed with unexpected error: [Errno 54] Connection reset by peer
2025-05-11 18:13:03,168 - ERROR - Request 76 failed with unexpected error: [Errno 54] Connection reset by peer
2025-05-11 18:13:19,095 - ERROR - Request 83 failed with unexpected error: [Errno 54] Connection reset by peer
2025-05-11 18:13:24,461 - ERROR - Request 86 failed with unexpected error: [Errno 54] Connection reset by peer
2025-05-11 18:13:37,857 - ERROR - Request 92 failed with unexpected error: [Errno 54] Connection reset by peer
2025-05-11 18:13:51,285 - ERROR - Request 98 failed with unexpected error: [Errno 54] Connection reset by peer
2025-05-11 18:14:10,312 - INFO - Load Test Summary:
2025-05-11 18:14:10,312 - INFO -   Total requests sent: 100
2025-05-11 18:14:10,312 - INFO -   Successful requests: 79
2025-05-11 18:14:10,313 - INFO -   Failed requests: 21
2025-05-11 18:14:10,313 - INFO -   Total time taken: 210.3862 seconds
2025-05-11 18:14:10,313 - INFO -   Requests per second (RPS): 0.4753
2025-05-11 18:14:10,313 - INFO -   Avg Latency (successful): 23.9341s
2025-05-11 18:14:10,313 - INFO -   Min Latency (successful): 2.8093s
2025-05-11 18:14:10,313 - INFO -   Max Latency (successful): 63.9591s
2025-05-11 18:14:10,313 - INFO -   P95 Latency (successful): 61.8770s
```

For GPU:

Single file search:

```
legal-search-api-dummy  | INFO:     216.165.95.147:61211 - "GET / HTTP/1.1" 200 OK
legal-search-api-dummy  | INFO:     172.18.0.2:33792 - "GET /metrics HTTP/1.1" 200 OK
legal-search-api-dummy  | INFO:     172.18.0.2:52502 - "GET /metrics HTTP/1.1" 200 OK
legal-search-api-dummy  | INFO:src.api.main:Processing uploaded file: Campoli v. Anywhere real Est. Inc., 2025 U.S. Dist. LEXIS 68458.PDF
legal-search-api-dummy  | INFO:src.api.main:Split uploaded file into 8 chunks.
legal-search-api-dummy  | INFO:src.api.main:Embedding 8 file chunks using ONNX on cuda
legal-search-api-dummy  | INFO:src.api.main:File query embedding (8 chunks) took 1.491090s.
legal-search-api-dummy  | INFO:src.api.main:Searching for file chunk 1/8
legal-search-api-dummy  | INFO:src.api.main:FAISS search took 0.003486s for k=3
legal-search-api-dummy  | INFO:src.api.main:Result aggregation took 0.000041s, aggregated to 2 docs.
legal-search-api-dummy  | INFO:src.api.main:Searching for file chunk 2/8
legal-search-api-dummy  | INFO:src.api.main:FAISS search took 0.003123s for k=3
legal-search-api-dummy  | INFO:src.api.main:Result aggregation took 0.000009s, aggregated to 1 docs.
legal-search-api-dummy  | INFO:src.api.main:Searching for file chunk 3/8
legal-search-api-dummy  | INFO:src.api.main:FAISS search took 0.003016s for k=3
legal-search-api-dummy  | INFO:src.api.main:Result aggregation took 0.000019s, aggregated to 2 docs.
legal-search-api-dummy  | INFO:src.api.main:Searching for file chunk 4/8
legal-search-api-dummy  | INFO:src.api.main:FAISS search took 0.003321s for k=3
legal-search-api-dummy  | INFO:src.api.main:Result aggregation took 0.000020s, aggregated to 2 docs.
legal-search-api-dummy  | INFO:src.api.main:Searching for file chunk 5/8
legal-search-api-dummy  | INFO:src.api.main:FAISS search took 0.003148s for k=3
legal-search-api-dummy  | INFO:src.api.main:Result aggregation took 0.000012s, aggregated to 2 docs.
legal-search-api-dummy  | INFO:src.api.main:Searching for file chunk 6/8
legal-search-api-dummy  | INFO:src.api.main:FAISS search took 0.003159s for k=3
legal-search-api-dummy  | INFO:src.api.main:Result aggregation took 0.000024s, aggregated to 3 docs.
legal-search-api-dummy  | INFO:src.api.main:Searching for file chunk 7/8
```

```
legal-search-api-dummy  | INFO:src.api.main:FAISS search took 0.003169s for k=3
legal-search-api-dummy  | INFO:src.api.main:Result aggregation took 0.000012s, aggregated to 2 docs.
legal-search-api-dummy  | INFO:src.api.main:Searching for file chunk 8/8
legal-search-api-dummy  | INFO:src.api.main:FAISS search took 0.003257s for k=3
legal-search-api-dummy  | INFO:src.api.main:Result aggregation took 0.000025s, aggregated to 3 docs.
legal-search-api-dummy  | INFO:src.api.main:Preparing to return 2 results with pre-computed summaries.
legal-search-api-dummy  | INFO:src.api.main:Result formatting took 0.000050s for 2 docs.
legal-search-api-dummy  | INFO:src.api.main:Core search logic for ONNX on cuda finished in 1.5603s. Returning 2
documents.
legal-search-api-dummy  | INFO:     216.165.95.147:63979 - "POST /search_combined HTTP/1.1" 200 OK
legal-search-api-dummy  | INFO:     172.18.0.2:50866 - "GET /metrics HTTP/1.1" 200 OK
```

Load testing :

100 requests (5 concurrent)

```
sanju@Sanjeevans-MacBook-Air serving_dummy % python3 src/test/load_test_api.py --concurrent 5 --total_requests 100
--url "http://192.5.86.151:8000/search_combined"
2025-05-11 20:03:15,211 - INFO - Starting load test: URL='http://192.5.86.151:8000/search_combined',
PDF='real_pdfs/Glover v. State.PDF', Concurrent=5, Total=100, TopK=1
2025-05-11 20:04:44,823 - INFO - Load Test Summary:
2025-05-11 20:04:44,823 - INFO -   Total requests sent: 100
2025-05-11 20:04:44,823 - INFO -   Successful requests: 100
2025-05-11 20:04:44,823 - INFO -   Failed requests: 0
2025-05-11 20:04:44,823 - INFO -   Total time taken: 89.6112 seconds
2025-05-11 20:04:44,823 - INFO -   Requests per second (RPS): 1.1159
2025-05-11 20:04:44,823 - INFO -   Avg Latency (successful): 4.4445s
2025-05-11 20:04:44,823 - INFO -   Min Latency (successful): 1.0386s
2025-05-11 20:04:44,823 - INFO -   Max Latency (successful): 4.6872s
2025-05-11 20:04:44,823 - INFO -   P95 Latency (successful): 4.6315s
```

100 requests (10 concurrent)

```
sanju@Sanjeevans-MacBook-Air serving_dummy % python3 src/test/load_test_api.py --concurrent 10 --total_requests 100
--url "http://192.5.86.151:8000/search_combined"
2025-05-11 20:01:05,078 - INFO - Starting load test: URL='http://192.5.86.151:8000/search_combined',
PDF='real_pdfs/Glover v. State.PDF', Concurrent=10, Total=100, TopK=1
2025-05-11 20:01:14,372 - ERROR - Request 14 failed with unexpected error: [Errno 54] Connection reset by peer
2025-05-11 20:01:14,375 - ERROR - Request 12 failed with unexpected error: [Errno 54] Connection reset by peer
2025-05-11 20:01:21,397 - ERROR - Request 30 failed with unexpected error: [Errno 54] Connection reset by peer
2025-05-11 20:01:29,366 - ERROR - Request 38 failed with unexpected error: [Errno 54] Connection reset by peer
2025-05-11 20:01:37,514 - ERROR - Request 46 failed with unexpected error: [Errno 54] Connection reset by peer
2025-05-11 20:01:45,604 - ERROR - Request 54 failed with unexpected error: [Errno 54] Connection reset by peer
2025-05-11 20:01:53,695 - ERROR - Request 63 failed with unexpected error: [Errno 54] Connection reset by peer
2025-05-11 20:02:01,840 - ERROR - Request 73 failed with unexpected error: [Errno 54] Connection reset by peer
2025-05-11 20:02:09,975 - ERROR - Request 88 failed with unexpected error: [Errno 54] Connection reset by peer
2025-05-11 20:02:17,839 - ERROR - Request 93 failed with unexpected error: [Errno 54] Connection reset by peer
2025-05-11 20:02:25,132 - INFO - Load Test Summary:
```

```
2025-05-11 20:02:25,132 - INFO -   Total requests sent: 100
2025-05-11 20:02:25,132 - INFO -   Successful requests: 90
2025-05-11 20:02:25,132 - INFO -   Failed requests: 10
2025-05-11 20:02:25,132 - INFO -   Total time taken: 80.0522 seconds
2025-05-11 20:02:25,133 - INFO -   Requests per second (RPS): 1.2492
2025-05-11 20:02:25,133 - INFO -   Avg Latency (successful): 8.4898s
2025-05-11 20:02:25,133 - INFO -   Min Latency (successful): 1.0079s
2025-05-11 20:02:25,133 - INFO -   Max Latency (successful): 15.0197s
2025-05-11 20:02:25,133 - INFO -   P95 Latency (successful): 14.4003s
```

Conclusion for this Difficulty Point:

Our tests showed some clear differences:

Speed for One Search: When searching with a single PDF file (which was split into 8 chunks), the GPU was much faster. It took the GPU about **1.56 seconds** to process the file and get results. The CPU took more than twice as long, around **3.59 seconds**. Most of this speed-up on the GPU came from doing the text embedding part faster (1.49s on GPU vs. 3.49s on CPU).

Handling Many Users (Load Tests):

CPU Performance:

With 5 users at the same time (5 concurrent requests), the CPU setup was quite slow. The average response time was about **12.46 seconds** (P95 latency was ~18.72s), and it only managed about **0.42 requests per second (RPS)**. We also saw **11 out of 100 requests fail** with "Connection reset by peer" errors.

When we tried 10 users at the same time, the CPU struggled even more. The average response time jumped to **23.93 seconds** (P95 latency was very high at ~61.88s), and **21 out of 100 requests failed**. The RPS was about 0.48.

GPU Performance:

With 5 users, the GPU did much better. It handled all **100 requests successfully** with no errors. The average response time was only about **4.44 seconds** (P95 latency ~4.63s), and it achieved around **1.12 RPS**.

With 10 users, the GPU started to show some strain, with **10 out of 100 requests failing**. However, it still performed better than the CPU, with an average response time of about **8.49 seconds** (P95 latency ~14.40s) and about **1.25 RPS**.

Overall Findings:

The GPU was clearly faster for single searches and could handle more users at once with quicker responses and fewer errors than the CPU. For example, at 5 concurrent users, the GPU was almost 3 times faster in average response time and handled all requests, while the CPU was slow and dropped some.

While the CPU could technically run the application, its performance under even a small load (5-10 users) was not very good, with high latencies and many errors. This wouldn't be great for lawyers trying to get quick search results.

The GPU provided a much better user experience, especially when a few users were active at the same time. Even at 10 concurrent users, where the GPU started to have some errors, it was still significantly more responsive than the CPU.

Cost Consideration: We know that GPU virtual machines on Chameleon Cloud (and commercial clouds) generally "cost" more in terms of credits or money. So, while the GPU gave us much better performance, we'd have to think if that extra speed is worth the extra cost for a real-world LegalAI service. For our project, demonstrating this significant performance lift with the GPU was the main goal.

In conclusion, for our LegalAI application using the ONNX model, the GPU provided a substantial improvement in both single-query latency and the ability to handle concurrent users with acceptable performance, compared to the CPU. While the CPU struggled under load, the GPU offered a more robust and responsive experience, though it would come at a higher operational cost. This exploration helped us understand the trade-offs for different serving options.