Model Size (Total Memory Footprint for Serving):

Our LegalAI app needs enough memory to handle all its parts when it's looking for case files. We plan for it to initially work with about 8,000 to 10,000 case documents, which means it will break these down into many smaller text pieces (maybe 40,000 to over 100,000 chunks). To run smoothly with this much data, the server will probably need around 1.5 GB to 3.0 GB of RAM. This memory will be used for the main language model (Legal-BERT-finetuned), the FAISS index that stores the document data, a map to quickly find info about the text chunks, the general details about each case file, and the FastAPI app itself running in Docker. If we add more documents, we'll need more memory. Our early tests with fewer files show the app uses about 600MB when it's just sitting there and up to 1GB when it's actively searching on a CPU for a single user, so we know the core parts are working well.

Throughput for Batch Inference (at serving time):

The app doesn't do "batch processing" like some systems that work on thousands of files offline. Instead, when a user uploads a PDF, our app has to quickly:

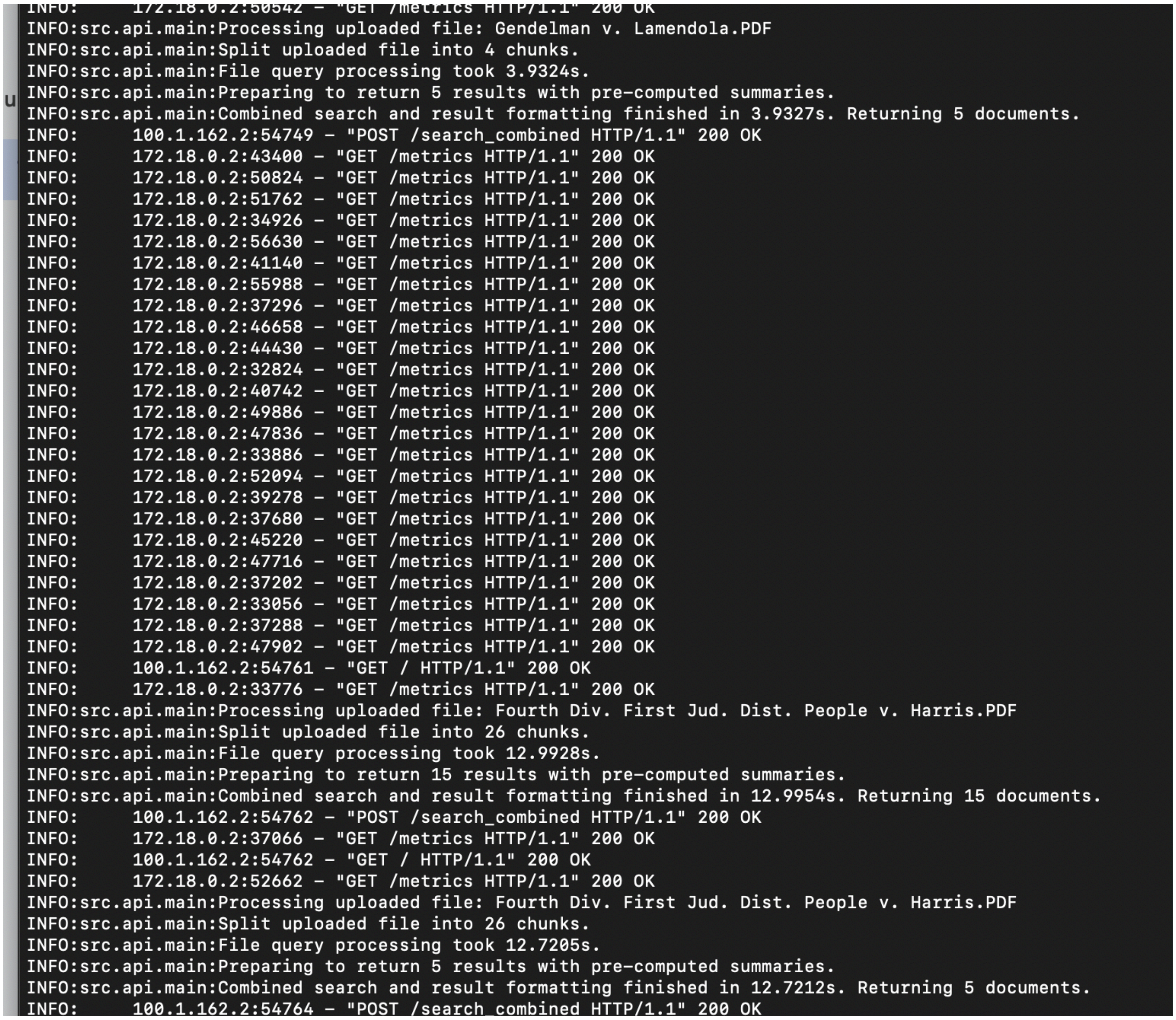
1. Read the PDF.
2. Break its text into smaller chunks.
3. Create an embedding (a numerical version) for each chunk.
4. Search our main FAISS index using each of these new chunk embeddings. This all happens as part of one user request. For a typical PDF (maybe 10-15 pages, which could become 20-30 chunks), we want the app to do all this and show results within our general speed goals (see Latency below). Our tests show that handling a file with 26 chunks takes around 13 seconds on a CPU, mostly because creating embeddings for many chunks takes time.

Example logs :

```
INFO:src.api.main:Processing uploaded file: Gendelman v. Lamendola.PDF
INFO:src.api.main:Split uploaded file into 4 chunks.
INFO:src.api.main:File query processing took 3.9324s.
INFO:src.api.main:Preparing to return 5 results with pre-computed summaries.
INFO:src.api.main:Combined search and result formatting finished in 3.9327s. Returning 5 documents.

INFO:src.api.main:Processing uploaded file: Fourth Div. First Jud. Dist. People v. Harris.PDF
INFO:src.api.main:Split uploaded file into 26 chunks.
INFO:src.api.main:File query processing took 12.9928s.
INFO:src.api.main:Preparing to return 15 results with pre-computed summaries.
INFO:src.api.main:Combined search and result formatting finished in 12.9954s. Returning 15 documents.
```

INFO:src.api.main:Processing uploaded file: Fourth Div. First Jud. Dist. People v. Harris.PDF
INFO:src.api.main:Split uploaded file into 26 chunks.
INFO:src.api.main:File query processing took 12.7205s.
INFO:src.api.main:Preparing to return 5 results with pre-computed summaries.
INFO:src.api.main:Combined search and result formatting finished in 12.7212s. Returning 5 documents.



```
INFO:     172.18.0.2:50542 - "GET /metrics HTTP/1.1" 200 OK
INFO:src.api.main:Processing uploaded file: Gendelman v. Lamendola.PDF
INFO:src.api.main:Split uploaded file into 4 chunks.
INFO:src.api.main:File query processing took 3.9324s.
INFO:src.api.main:Preparing to return 5 results with pre-computed summaries.
INFO:src.api.main:Combined search and result formatting finished in 3.9327s. Returning 5 documents.
INFO:     100.1.162.2:54749 - "POST /search_combined HTTP/1.1" 200 OK
INFO:     172.18.0.2:43400 - "GET /metrics HTTP/1.1" 200 OK
INFO:     172.18.0.2:50824 - "GET /metrics HTTP/1.1" 200 OK
INFO:     172.18.0.2:51762 - "GET /metrics HTTP/1.1" 200 OK
INFO:     172.18.0.2:34926 - "GET /metrics HTTP/1.1" 200 OK
INFO:     172.18.0.2:56630 - "GET /metrics HTTP/1.1" 200 OK
INFO:     172.18.0.2:41140 - "GET /metrics HTTP/1.1" 200 OK
INFO:     172.18.0.2:55988 - "GET /metrics HTTP/1.1" 200 OK
INFO:     172.18.0.2:37296 - "GET /metrics HTTP/1.1" 200 OK
INFO:     172.18.0.2:46658 - "GET /metrics HTTP/1.1" 200 OK
INFO:     172.18.0.2:44430 - "GET /metrics HTTP/1.1" 200 OK
INFO:     172.18.0.2:32824 - "GET /metrics HTTP/1.1" 200 OK
INFO:     172.18.0.2:40742 - "GET /metrics HTTP/1.1" 200 OK
INFO:     172.18.0.2:49886 - "GET /metrics HTTP/1.1" 200 OK
INFO:     172.18.0.2:47836 - "GET /metrics HTTP/1.1" 200 OK
INFO:     172.18.0.2:33886 - "GET /metrics HTTP/1.1" 200 OK
INFO:     172.18.0.2:52094 - "GET /metrics HTTP/1.1" 200 OK
INFO:     172.18.0.2:39278 - "GET /metrics HTTP/1.1" 200 OK
INFO:     172.18.0.2:37680 - "GET /metrics HTTP/1.1" 200 OK
INFO:     172.18.0.2:45220 - "GET /metrics HTTP/1.1" 200 OK
INFO:     172.18.0.2:47716 - "GET /metrics HTTP/1.1" 200 OK
INFO:     172.18.0.2:37202 - "GET /metrics HTTP/1.1" 200 OK
INFO:     172.18.0.2:33056 - "GET /metrics HTTP/1.1" 200 OK
INFO:     172.18.0.2:37288 - "GET /metrics HTTP/1.1" 200 OK
INFO:     172.18.0.2:47902 - "GET /metrics HTTP/1.1" 200 OK
INFO:     100.1.162.2:54761 - "GET / HTTP/1.1" 200 OK
INFO:     172.18.0.2:33776 - "GET /metrics HTTP/1.1" 200 OK
INFO:src.api.main:Processing uploaded file: Fourth Div. First Jud. Dist. People v. Harris.PDF
INFO:src.api.main:Split uploaded file into 26 chunks.
INFO:src.api.main:File query processing took 12.9928s.
INFO:src.api.main:Preparing to return 15 results with pre-computed summaries.
INFO:src.api.main:Combined search and result formatting finished in 12.9954s. Returning 15 documents.
INFO:     100.1.162.2:54762 - "POST /search_combined HTTP/1.1" 200 OK
INFO:     172.18.0.2:37066 - "GET /metrics HTTP/1.1" 200 OK
INFO:     100.1.162.2:54762 - "GET / HTTP/1.1" 200 OK
INFO:     172.18.0.2:52662 - "GET /metrics HTTP/1.1" 200 OK
INFO:src.api.main:Processing uploaded file: Fourth Div. First Jud. Dist. People v. Harris.PDF
INFO:src.api.main:Split uploaded file into 26 chunks.
INFO:src.api.main:File query processing took 12.7205s.
INFO:src.api.main:Preparing to return 5 results with pre-computed summaries.
INFO:src.api.main:Combined search and result formatting finished in 12.7212s. Returning 5 documents.
INFO:     100.1.162.2:54764 - "POST /search_combined HTTP/1.1" 200 OK
```

Latency for Online (Single Sample) Inference:

Users need to get results pretty fast for the app to feel useful. Here’s what we're aiming for (measuring the 95th percentile, meaning 95% of requests should be this fast or faster):

For Text Searches: When a user types a search query, they should get results in under 5 seconds on a CPU server. This mostly involves creating an embedding for the short query and searching the FAISS index.

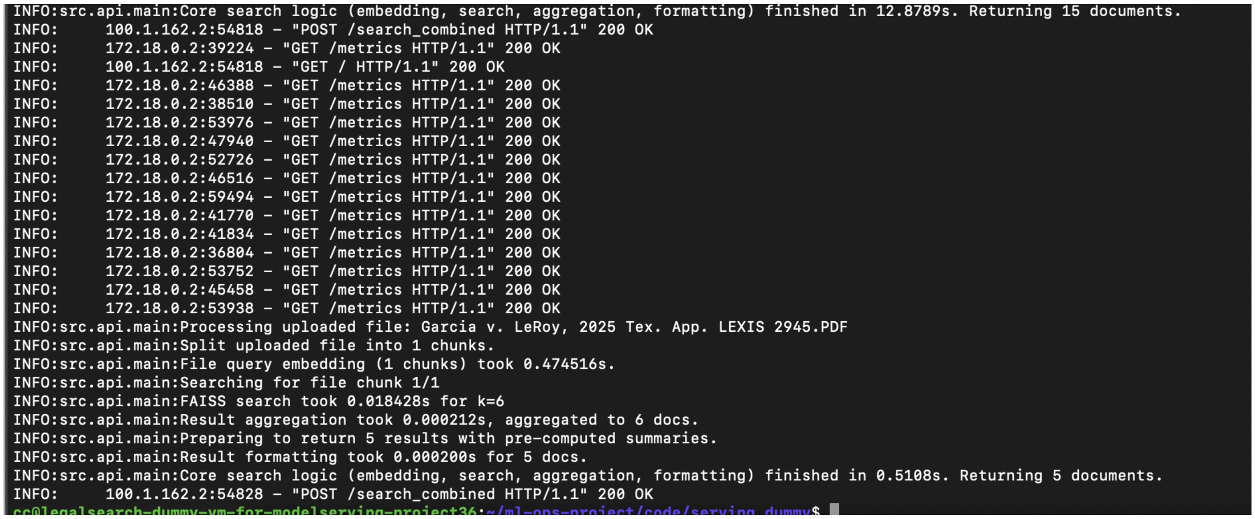
For Uploaded PDF Files (like a ~100KB PDF, making about 5-15 chunks):

On a CPU server: Results should appear in 5 to 15 seconds.

If we use a GPU (maybe for a "Pro" version): Results should appear in under 5 seconds (ideally <3s). Our early tests are promising: a very small file (1 chunk) gets processed in about half a second, and a larger one (26 chunks) takes about 13 seconds on a CPU, which fits our 5-15 second goal.

Example logs below:

```
INFO:src.api.main:Processing uploaded file: Garcia v. LeRoy, 2025 Tex. App. LEXIS 2945.PDF
INFO:src.api.main:Split uploaded file into 1 chunks.
INFO:src.api.main:File query embedding (1 chunks) took 0.474516s.
INFO:src.api.main:Searching for file chunk 1/1
INFO:src.api.main:FAISS search took 0.018428s for k=6
INFO:src.api.main:Result aggregation took 0.000212s, aggregated to 6 docs.
INFO:src.api.main:Preparing to return 5 results with pre-computed summaries.
INFO:src.api.main:Result formatting took 0.000200s for 5 docs.
INFO:src.api.main:Core search logic (embedding, search, aggregation, formatting) finished in 0.5108s. Returning 5 documents.
```

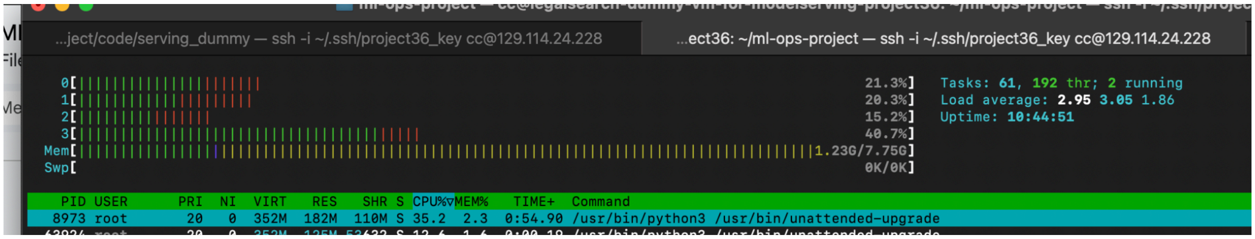


```
INFO:src.api.main:Core search logic (embedding, search, aggregation, formatting) finished in 12.8789s. Returning 15 documents.
INFO:     100.1.162.2:54818 - "POST /search_combined HTTP/1.1" 200 OK
INFO:     172.18.0.2:39224 - "GET /metrics HTTP/1.1" 200 OK
INFO:     100.1.162.2:54818 - "GET / HTTP/1.1" 200 OK
INFO:     172.18.0.2:46388 - "GET /metrics HTTP/1.1" 200 OK
INFO:     172.18.0.2:38510 - "GET /metrics HTTP/1.1" 200 OK
INFO:     172.18.0.2:53976 - "GET /metrics HTTP/1.1" 200 OK
INFO:     172.18.0.2:47940 - "GET /metrics HTTP/1.1" 200 OK
INFO:     172.18.0.2:52726 - "GET /metrics HTTP/1.1" 200 OK
INFO:     172.18.0.2:46516 - "GET /metrics HTTP/1.1" 200 OK
INFO:     172.18.0.2:59494 - "GET /metrics HTTP/1.1" 200 OK
INFO:     172.18.0.2:41770 - "GET /metrics HTTP/1.1" 200 OK
INFO:     172.18.0.2:41834 - "GET /metrics HTTP/1.1" 200 OK
INFO:     172.18.0.2:36804 - "GET /metrics HTTP/1.1" 200 OK
INFO:     172.18.0.2:53752 - "GET /metrics HTTP/1.1" 200 OK
INFO:     172.18.0.2:45458 - "GET /metrics HTTP/1.1" 200 OK
INFO:     172.18.0.2:53938 - "GET /metrics HTTP/1.1" 200 OK
INFO:src.api.main:Processing uploaded file: Garcia v. LeRoy, 2025 Tex. App. LEXIS 2945.PDF
INFO:src.api.main:Split uploaded file into 1 chunks.
INFO:src.api.main:File query embedding (1 chunks) took 0.474516s.
INFO:src.api.main:Searching for file chunk 1/1
INFO:src.api.main:FAISS search took 0.018428s for k=6
INFO:src.api.main:Result aggregation took 0.000212s, aggregated to 6 docs.
INFO:src.api.main:Preparing to return 5 results with pre-computed summaries.
INFO:src.api.main:Result formatting took 0.000200s for 5 docs.
INFO:src.api.main:Core search logic (embedding, search, aggregation, formatting) finished in 0.5108s. Returning 5 documents.
INFO:     100.1.162.2:54828 - "POST /search_combined HTTP/1.1" 200 OK
cc@legalsearch-dummy-vm-for-modelserving-project36:~/ml-ops-project/code/serving_dummy$
```

Concurrency Requirement (Cloud Deployment):

When we put this on our m1.large Chameleon Cloud VM, it needs to handle 10 users actively searching at the same time. While doing this, it must still meet our speed goals (like 5-15 seconds for PDF uploads on CPU; currently this is holding for 5 concurrent users) and almost all requests (more than 99.9%) should succeed without errors. Our initial load tests showed the system could handle up to 15 users, but we also saw some consistent errors that didn't seem to depend on how many users there were. We need to fix these errors first. Once the app is stable, we'll do more load testing to see exactly how many users the m1.large VM can support well, aiming for our target of 10 stable, concurrent users.

CPU in idle state:

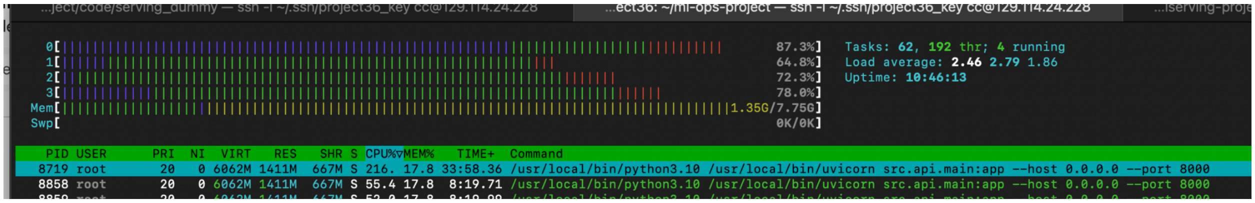


...ject/code/serving_dummy — ssh -i ~/.ssh/project36_key cc@129.114.24.228
...ect36: ~/ml-ops-project — ssh -i ~/.ssh/project36_key cc@129.114.24.228

```
  0[|||||||||||||||||||||                                       21.3%]   Tasks: 61, 192 thr; 2 running
  1[||||||||||||||||||||                                        20.3%]   Load average: 2.95 3.05 1.86
  2[|||||||||||||||                                             15.2%]   Uptime: 10:44:51
  3[||||||||||||||||||||||||||||||||||||||||                    40.7%]
Mem[||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||1.23G/7.75G]
Swp[                                                            0K/0K]

  PID USER      PRI  NI  VIRT   RES   SHR S CPU%▽MEM%   TIME+  Command
 8973 root       20   0  352M  182M  110M S 35.2  2.3  0:54.90 /usr/bin/python3 /usr/bin/unattended-upgrade
```

For 15 concurrent Users:



...ject/code/serving_dummy — ssh -i ~/.ssh/project36_key cc@129.114.24.228
...ect36: ~/ml-ops-project — ssh -i ~/.ssh/project36_key cc@129.114.24.228

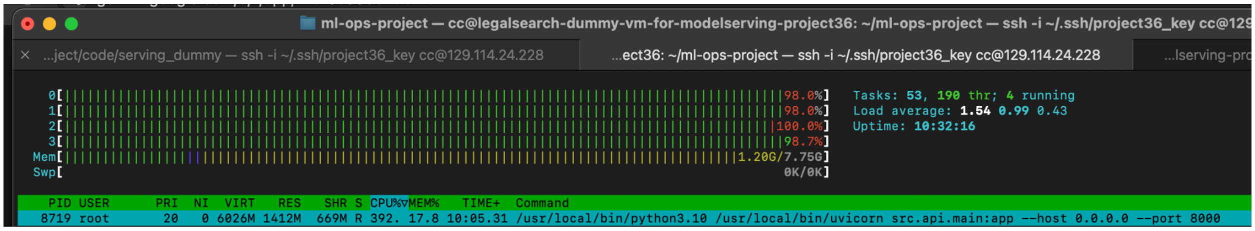
```
  0[|||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||    87.3%]   Tasks: 62, 192 thr; 4 running
  1[|||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||                              64.8%]   Load average: 2.46 2.79 1.86
  2[|||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||                      72.3%]   Uptime: 10:46:13
  3[||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||               78.0%]
Mem[||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||1.35G/7.75G]
Swp[                                                                                              0K/0K]

  PID USER      PRI  NI  VIRT   RES   SHR S CPU%▽MEM%   TIME+  Command
 8719 root       20   0 6062M 1411M  667M S 216. 17.8 33:58.36 /usr/local/bin/python3.10 /usr/local/bin/uvicorn src.api.main:app --host 0.0.0.0 --port 8000
 8858 root       20   0 6062M 1411M  667M S 55.4 17.8  8:19.71 /usr/local/bin/python3.10 /usr/local/bin/uvicorn src.api.main:app --host 0.0.0.0 --port 8000
```

```
sanju@Sanjeevans-MacBook-Air serving_dummy % python3 src/test/load_test_api.py --concurrent 15 --total_requests 60
2025-05-11 02:34:20,892 - INFO - Starting load test: URL='http://129.114.24.228:8000/search_combined',
PDF='real_pdfs/Glover v. State.PDF', Concurrent=15, Total=60, TopK=1
2025-05-11 02:35:33,350 - ERROR - Request 20 failed with unexpected error: [Errno 54] Connection reset by peer
2025-05-11 02:35:33,353 - ERROR - Request 22 failed with unexpected error: [Errno 54] Connection reset by peer
2025-05-11 02:35:33,353 - ERROR - Request 17 failed with unexpected error: [Errno 54] Connection reset by peer
2025-05-11 02:35:33,355 - ERROR - Request 27 failed with unexpected error: [Errno 54] Connection reset by peer
2025-05-11 02:35:33,355 - ERROR - Request 19 failed with unexpected error: [Errno 54] Connection reset by peer
2025-05-11 02:35:33,357 - ERROR - Request 28 failed with unexpected error: [Errno 54] Connection reset by peer
```

```
2025-05-11 02:35:33,357 - ERROR - Request 30 failed with unexpected error: [Errno 54] Connection reset by peer
2025-05-11 02:35:33,357 - ERROR - Request 24 failed with unexpected error: [Errno 54] Connection reset by peer
2025-05-11 02:35:33,358 - ERROR - Request 18 failed with unexpected error: [Errno 54] Connection reset by peer
2025-05-11 02:35:33,359 - ERROR - Request 23 failed with unexpected error: [Errno 54] Connection reset by peer
2025-05-11 02:35:33,360 - ERROR - Request 26 failed with unexpected error: [Errno 54] Connection reset by peer
2025-05-11 02:35:33,360 - ERROR - Request 25 failed with unexpected error: [Errno 54] Connection reset by peer
2025-05-11 02:36:05,113 - ERROR - Request 50 failed with unexpected error: [Errno 54] Connection reset by peer
2025-05-11 02:36:22,379 - ERROR - Request 57 failed with unexpected error: [Errno 54] Connection reset by peer
2025-05-11 02:36:30,902 - ERROR - Request 59 failed with unexpected error: [Errno 54] Connection reset by peer
2025-05-11 02:36:56,672 - INFO - Load Test Summary:
2025-05-11 02:36:56,673 - INFO -   Total requests sent: 60
2025-05-11 02:36:56,673 - INFO -   Successful requests: 45
2025-05-11 02:36:56,673 - INFO -   Failed requests: 15
2025-05-11 02:36:56,673 - INFO -   Total time taken: 155.7775 seconds
2025-05-11 02:36:56,673 - INFO -   Requests per second (RPS): 0.3852
2025-05-11 02:36:56,673 - INFO -   Avg Latency (successful): 44.6074s
2025-05-11 02:36:56,673 - INFO -   Min Latency (successful): 5.8317s
2025-05-11 02:36:56,673 - INFO -   Max Latency (successful): 80.4769s
2025-05-11 02:36:56,673 - INFO -   P95 Latency (successful): 74.7414s
```

In case of 10 concurrent users, this happens :



ml-ops-project — cc@legalsearch-dummy-vm-for-modelserving-project36: ~/ml-ops-project — ssh -i ~/.ssh/project36_key cc@129
...ject/code/serving_dummy — ssh -i ~/.ssh/project36_key cc@129.114.24.228
...ect36: ~/ml-ops-project — ssh -i ~/.ssh/project36_key cc@129.114.24.228
...lserving-pro

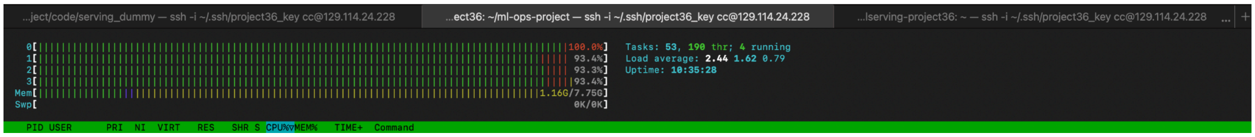
```
    0[||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||98.0%]   Tasks: 53, 190 thr; 4 running
    1[||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||98.0%]   Load average: 1.54 0.99 0.43
    2[|||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||100.0%]   Uptime: 10:32:16
    3[||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||98.7%]
  Mem[|||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||1.20G/7.75G]
  Swp[                                                                                  0K/0K]

    PID USER      PRI  NI  VIRT   RES   SHR S CPU%▽MEM%   TIME+  Command
   8719 root       20   0 6026M 1412M  669M R 392. 17.8 10:05.31 /usr/local/bin/python3.10 /usr/local/bin/uvicorn src.api.main:app --host 0.0.0.0 --port 8000
```

```
sanju@Sanjeevans-MacBook-Air serving_dummy % python3 src/test/load_test_api.py --concurrent 10 --total_requests 50
2025-05-11 02:20:30,309 - INFO - Starting load test: URL='http://129.114.24.228:8000/search_combined',
PDF='real_pdfs/Glover v. State.PDF', Concurrent=10, Total=50, TopK=1
2025-05-11 02:21:00,547 - ERROR - Request 12 failed with unexpected error: [Errno 54] Connection reset by peer
2025-05-11 02:21:00,550 - ERROR - Request 16 failed with unexpected error: [Errno 54] Connection reset by peer
2025-05-11 02:21:00,550 - ERROR - Request 13 failed with unexpected error: [Errno 54] Connection reset by peer
2025-05-11 02:21:00,551 - ERROR - Request 15 failed with unexpected error: [Errno 54] Connection reset by peer
2025-05-11 02:21:00,551 - ERROR - Request 18 failed with unexpected error: [Errno 54] Connection reset by peer
2025-05-11 02:21:00,552 - ERROR - Request 20 failed with unexpected error: [Errno 54] Connection reset by peer
2025-05-11 02:21:00,552 - ERROR - Request 17 failed with unexpected error: [Errno 54] Connection reset by peer
2025-05-11 02:21:21,974 - ERROR - Request 33 failed with unexpected error: [Errno 54] Connection reset by peer
2025-05-11 02:21:38,077 - ERROR - Request 38 failed with unexpected error: [Errno 54] Connection reset by peer
2025-05-11 02:21:46,201 - ERROR - Request 42 failed with unexpected error: [Errno 54] Connection reset by peer
2025-05-11 02:21:54,268 - ERROR - Request 45 failed with unexpected error: [Errno 54] Connection reset by peer
2025-05-11 02:22:15,875 - INFO - Load Test Summary:
2025-05-11 02:22:15,875 - INFO -   Total requests sent: 50
2025-05-11 02:22:15,875 - INFO -   Successful requests: 39
2025-05-11 02:22:15,875 - INFO -   Failed requests: 11
2025-05-11 02:22:15,875 - INFO -   Total time taken: 105.5629 seconds
2025-05-11 02:22:15,875 - INFO -   Requests per second (RPS): 0.4737
2025-05-11 02:22:15,875 - INFO -   Avg Latency (successful): 23.0811s
2025-05-11 02:22:15,875 - INFO -   Min Latency (successful): 3.1186s
```

```
2025-05-11 02:22:15,875 - INFO -   Max Latency (successful): 61.8926s
2025-05-11 02:22:15,875 - INFO -   P95 Latency (successful): 53.7110s
```

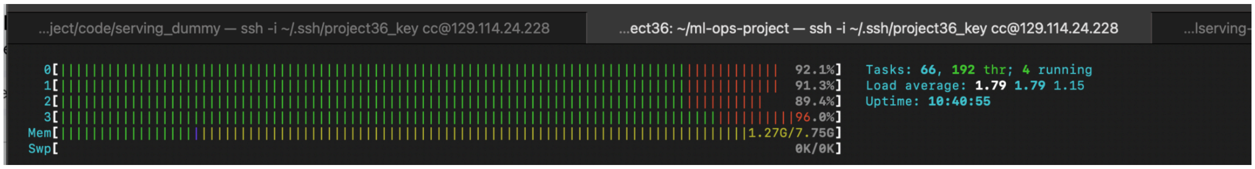
In case of 5 concurrent users :


...ject/code/serving_dummy — ssh -i ~/.ssh/project36_key cc@129.114.24.228
...ect36: ~/ml-ops-project — ssh -i ~/.ssh/project36_key cc@129.114.24.228
...lserving-project36: ~ — ssh -i ~/.ssh/project36_key cc@129.114.24.228

```
  0[                                100.0%]   Tasks: 53, 190 thr; 4 running
  1[                                 93.4%]   Load average: 2.44 1.62 0.79
  2[                                 93.3%]   Uptime: 10:35:28
  3[                                 93.4%]
Mem[                           1.16G/7.75G]
Swp[                                 0K/0K]
  PID USER      PRI  NI  VIRT   RES   SHR S CPU%▽MEM%   TIME+  Command
```

```
sanju@Sanjeevans-MacBook-Air serving_dummy % python3 src/test/load_test_api.py --concurrent 5 --total_requests 50
2025-05-11 02:23:27,638 - INFO - Starting load test: URL='http://129.114.24.228:8000/search_combined', PDF='real_pdfs/Glover v. State.PDF', Concurrent=5, Total=50, TopK=1
2025-05-11 02:23:44,950 - ERROR - Request 7 failed with unexpected error: [Errno 54] Connection reset by peer
2025-05-11 02:23:44,951 - ERROR - Request 10 failed with unexpected error: [Errno 54] Connection reset by peer
2025-05-11 02:24:06,968 - ERROR - Request 19 failed with unexpected error: [Errno 54] Connection reset by peer
2025-05-11 02:24:31,850 - ERROR - Request 28 failed with unexpected error: [Errno 54] Connection reset by peer
2025-05-11 02:25:09,495 - ERROR - Request 37 failed with unexpected error: [Errno 54] Connection reset by peer
2025-05-11 02:25:09,495 - ERROR - Request 39 failed with unexpected error: [Errno 54] Connection reset by peer
2025-05-11 02:25:15,246 - ERROR - Request 43 failed with unexpected error: [Errno 54] Connection reset by peer
2025-05-11 02:25:58,786 - INFO - Load Test Summary:
2025-05-11 02:25:58,786 - INFO -   Total requests sent: 50
2025-05-11 02:25:58,786 - INFO -   Successful requests: 43
2025-05-11 02:25:58,786 - INFO -   Failed requests: 7
2025-05-11 02:25:58,786 - INFO -   Total time taken: 151.1451 seconds
2025-05-11 02:25:58,786 - INFO -   Requests per second (RPS): 0.3308
2025-05-11 02:25:58,786 - INFO -   Avg Latency (successful): 15.7843s
2025-05-11 02:25:58,786 - INFO -   Min Latency (successful): 3.1771s
2025-05-11 02:25:58,786 - INFO -   Max Latency (successful): 32.9306s
2025-05-11 02:25:58,787 - INFO -   P95 Latency (successful): 26.4824s
sanju@Sanjeevans-MacBook-Air serving_dummy %
```

For 3 concurrent users:


...ject/code/serving_dummy — ssh -i ~/.ssh/project36_key cc@129.114.24.228
...ect36: ~/ml-ops-project — ssh -i ~/.ssh/project36_key cc@129.114.24.228
...lserving-

```
  0[                                 92.1%]   Tasks: 66, 192 thr; 4 running
  1[                                 91.3%]   Load average: 1.79 1.79 1.15
  2[                                 89.4%]   Uptime: 10:40:55
  3[                                 96.0%]
Mem[                           1.27G/7.75G]
Swp[                                 0K/0K]
```

```
2025-05-11 02:29:19,743 - INFO - Starting load test: URL='http://129.114.24.228:8000/search_combined', PDF='real_pdfs/Glover v. State.PDF', Concurrent=3, Total=51, TopK=1
2025-05-11 02:30:05,673 - ERROR - Request 12 failed with unexpected error: [Errno 54] Connection reset by peer
2025-05-11 02:30:33,781 - ERROR - Request 18 failed with unexpected error: [Errno 54] Connection reset by peer
2025-05-11 02:31:02,200 - ERROR - Request 23 failed with unexpected error: [Errno 54] Connection reset by peer
2025-05-11 02:32:39,068 - ERROR - Request 47 failed with unexpected error: [Errno 54] Connection reset by peer
2025-05-11 02:33:02,109 - INFO - Load Test Summary:
```

```
2025-05-11 02:33:02,109 - INFO -   Total requests sent: 51
2025-05-11 02:33:02,109 - INFO -   Successful requests: 47
2025-05-11 02:33:02,109 - INFO -   Failed requests: 4
2025-05-11 02:33:02,109 - INFO -   Total time taken: 222.3629 seconds
2025-05-11 02:33:02,110 - INFO -   Requests per second (RPS): 0.2294
2025-05-11 02:33:02,110 - INFO -   Avg Latency (successful): 13.4557s
2025-05-11 02:33:02,110 - INFO -   Min Latency (successful): 4.2615s
2025-05-11 02:33:02,110 - INFO -   Max Latency (successful): 23.0358s
2025-05-11 02:33:02,110 - INFO -   P95 Latency (successful): 22.2676s
```