Base model (Original PyTorch Model) :

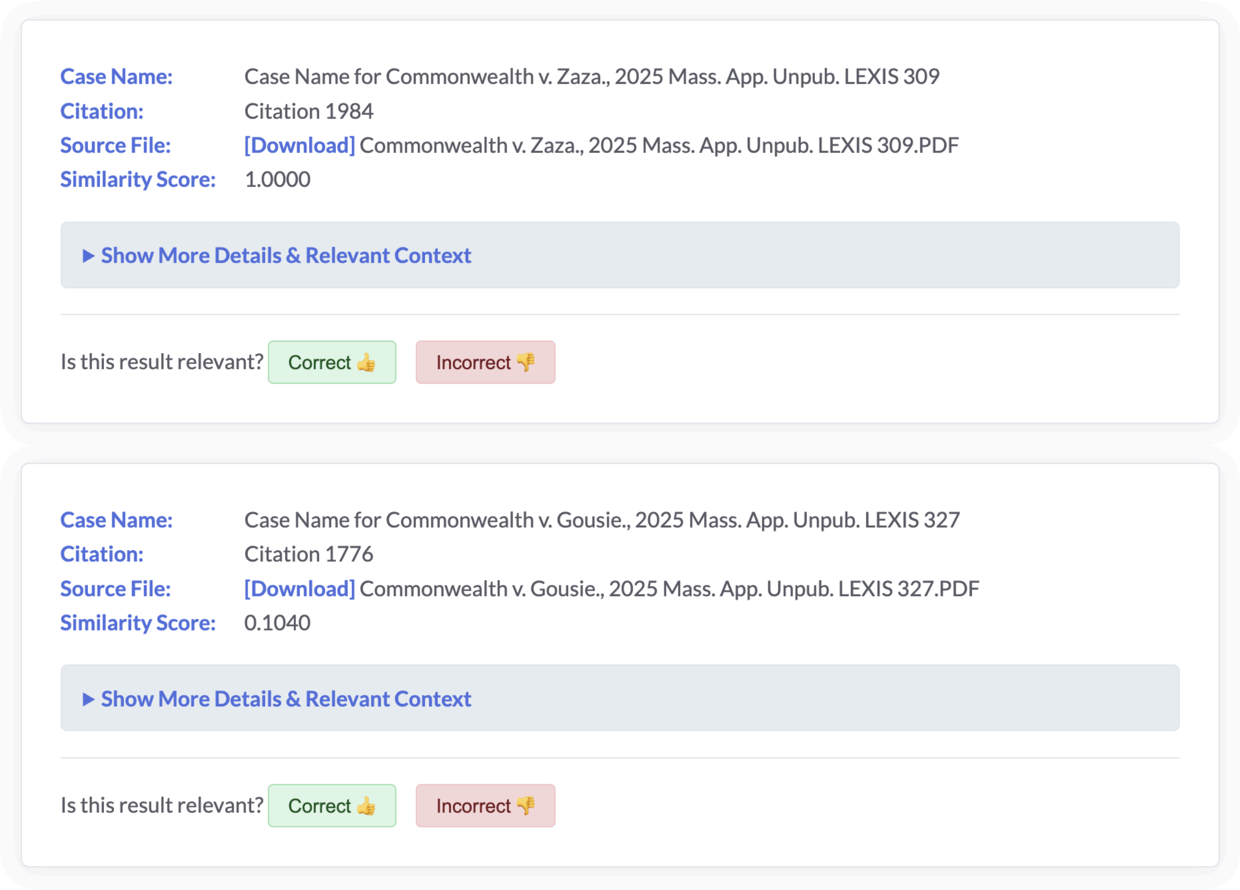
Legal-BERT-finetuned model size : 419MB

I first tested our basic setup using the original Legal-BERT-finetuned model. When searching with a single PDF (Commonwealth v. Zaza.PDF), it took about 3.3 seconds to process the file (7 chunks) and find the top 2 matching cases. The most similar case was the file itself (score 1.0), and the second was Commonwealth v. Gousie.PDF (score 0.1).

For the load test, I sent 50 requests (using Glover v. State.PDF), with 5 users sending requests at the same time. The system handled about 0.4 requests per second (RPS). On average, successful requests took about 12.9 seconds, with 95% of them finishing within 19.3 seconds. There were 5 failed requests during this test, showing the system was under a lot of pressure. We still need to measure the original model's file size. This gives us our starting point for seeing if optimizations can make things faster or handle more users.

Below are the logs and output.

```
INFO:src.api.main:Processing uploaded file: Commonwealth v. Zaza.PDF
INFO:src.api.main:Split uploaded file into 7 chunks.
INFO:src.api.main:File query embedding (7 chunks) took 3.268630s.
INFO:src.api.main:Searching for file chunk 1/7
INFO:src.api.main:FAISS search took 0.026056s for k=3
INFO:src.api.main:Result aggregation took 0.000130s, aggregated to 3 docs.
INFO:src.api.main:Searching for file chunk 2/7
INFO:src.api.main:FAISS search took 0.006860s for k=3
INFO:src.api.main:Result aggregation took 0.000126s, aggregated to 3 docs.
INFO:src.api.main:Searching for file chunk 3/7
INFO:src.api.main:FAISS search took 0.006782s for k=3
INFO:src.api.main:Result aggregation took 0.000105s, aggregated to 2 docs.
INFO:src.api.main:Searching for file chunk 4/7
INFO:src.api.main:FAISS search took 0.006634s for k=3
INFO:src.api.main:Result aggregation took 0.000111s, aggregated to 2 docs.
INFO:src.api.main:Searching for file chunk 5/7
INFO:src.api.main:FAISS search took 0.006763s for k=3
INFO:src.api.main:Result aggregation took 0.000121s, aggregated to 3 docs.
INFO:src.api.main:Searching for file chunk 6/7
INFO:src.api.main:FAISS search took 0.006710s for k=3
INFO:src.api.main:Result aggregation took 0.000113s, aggregated to 2 docs.
INFO:src.api.main:Searching for file chunk 7/7
INFO:src.api.main:FAISS search took 0.006483s for k=3
INFO:src.api.main:Result aggregation took 0.000105s, aggregated to 3 docs.
INFO:src.api.main:Preparing to return 2 results with pre-computed summaries.
INFO:src.api.main:Result formatting took 0.000191s for 2 docs.
INFO:src.api.main:Core search logic (embedding, search, aggregation, formatting) finished in 3.3820s. Returning 2
documents.
```



Case Name: Case Name for Commonwealth v. Zaza., 2025 Mass. App. Unpub. LEXIS 309
Citation: Citation 1984
Source File: [Download] Commonwealth v. Zaza., 2025 Mass. App. Unpub. LEXIS 309.PDF
Similarity Score: 1.0000

Case Name: Case Name for Commonwealth v. Gousie., 2025 Mass. App. Unpub. LEXIS 327
Citation: Citation 1776
Source File: [Download] Commonwealth v. Gousie., 2025 Mass. App. Unpub. LEXIS 327.PDF
Similarity Score: 0.1040

Load testing for 50 users (5 concurrent at a time) :

```
sanju@Sanjeevans-MacBook-Air serving_dummy % python3 src/test/load_test_api.py --concurrent 5 --total_requests 50
2025-05-11 10:10:54,342 - 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 10:11:11,127 - ERROR - Request 7 failed with unexpected error: [Errno 54] Connection reset by peer
2025-05-11 10:11:11,127 - ERROR - Request 10 failed with unexpected error: [Errno 54] Connection reset by peer
2025-05-11 10:11:43,894 - ERROR - Request 21 failed with unexpected error: [Errno 54] Connection reset by peer
2025-05-11 10:12:08,366 - ERROR - Request 31 failed with unexpected error: [Errno 54] Connection reset by peer
2025-05-11 10:12:33,169 - ERROR - Request 42 failed with unexpected error: [Errno 54] Connection reset by peer
2025-05-11 10:12:57,974 - INFO - Load Test Summary:
2025-05-11 10:12:57,974 - INFO -   Total requests sent: 50
2025-05-11 10:12:57,974 - INFO -   Successful requests: 45
2025-05-11 10:12:57,974 - INFO -   Failed requests: 5
2025-05-11 10:12:57,974 - INFO -   Total time taken: 123.6301 seconds
2025-05-11 10:12:57,974 - INFO -   Requests per second (RPS): 0.4044
2025-05-11 10:12:57,974 - INFO -   Avg Latency (successful): 12.8749s
```

```
2025-05-11 10:12:57,975 - INFO -   Min Latency (successful): 3.0597s
2025-05-11 10:12:57,975 - INFO -   Max Latency (successful): 27.2633s
2025-05-11 10:12:57,975 - INFO -   P95 Latency (successful): 19.3032s
```

For the next part, We need to install some dependencies. They are in /ml-ops-project/code/serving_dummy/requirements_script.txt

ONNX model :

To create ONNX model :
Inside dummy serving folder, run ->
python3 src/processing/export_to_onnx.py

Legal_bert_finetuned_onnx model size : 417M

Next, we changed the original Legal-BERT-finetuned model into a format called ONNX. This ONNX model was a tiny bit smaller (417MB compared to the original 419MB).

When we tested it with a single PDF search (Commonwealth v. Zaza.PDF), the ONNX model was a little faster. It took about 2.9 seconds to process the PDF and get results, compared to 3.3 seconds for the original. The search results for the top 2 cases were exactly the same as before.

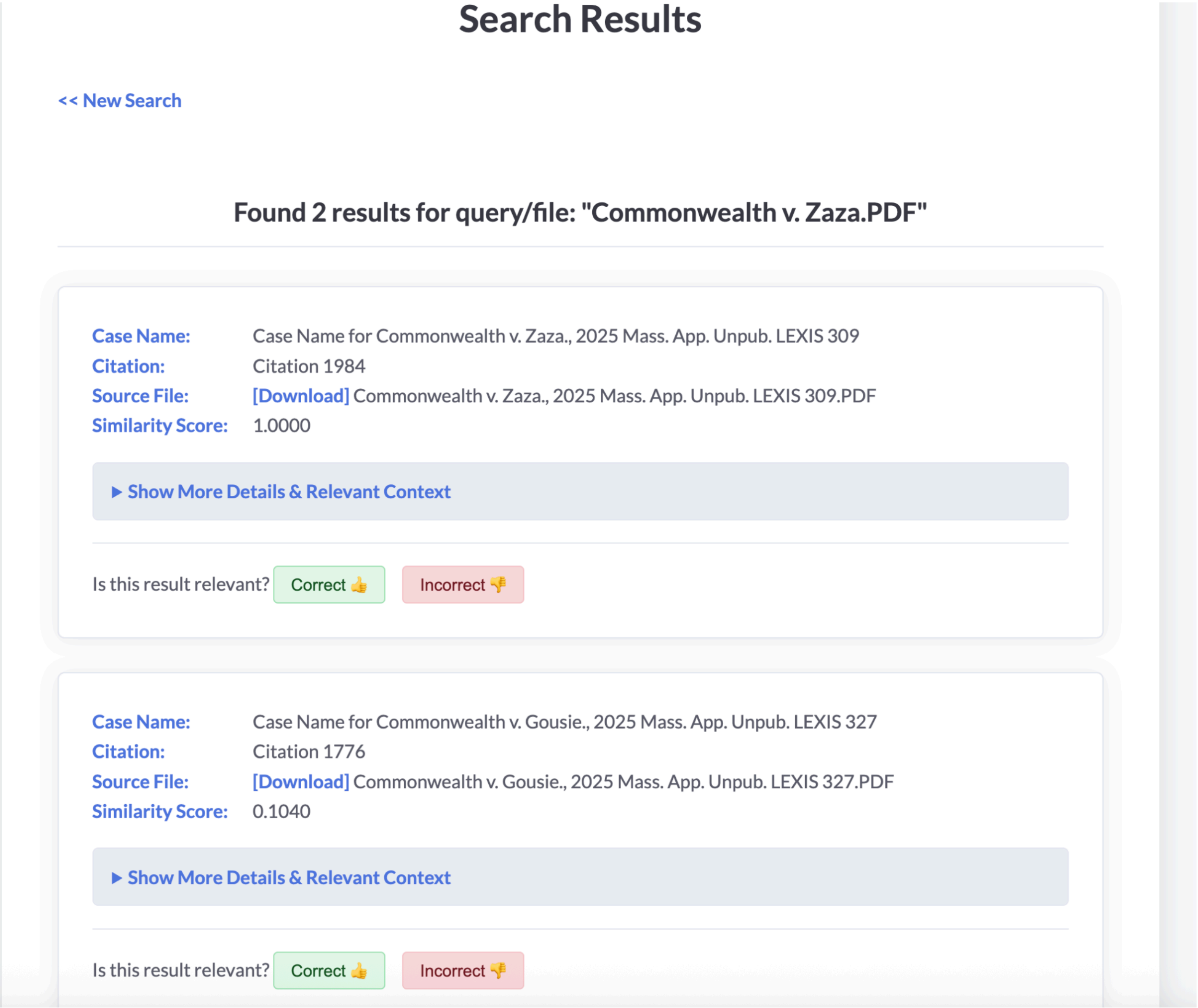
In the load test (50 requests, 5 users at a time, using Glover v. State.PDF), the ONNX model performed slightly better. It handled about 0.44 requests per second (RPS), while the original did 0.40 RPS. The average time for a successful request was around 11.6 seconds (down from 12.9s), and 95% of requests finished in about 17.6 seconds (down from 19.3s). We still had 5 failed requests, like with the original model, showing the system is still being pushed hard.

So, just converting to ONNX gave us a small speed boost (around 8-12% in single searches and ~9% in load test RPS/latency), and a very small reduction in model size, without changing the search results.

Below are the logs and output.

```
legal-search-api-dummy  | INFO:src.api.main:Processing uploaded file: Commonwealth v. Zaza.PDF
legal-search-api-dummy  | INFO:src.api.main:Split uploaded file into 7 chunks.
```

```
legal-search-api-dummy  | INFO:src.api.main:Embedding 7 file chunks using ONNX model
legal-search-api-dummy  | INFO:src.api.main:File query embedding (7 chunks) took 2.884357s.
legal-search-api-dummy  | INFO:src.api.main:Searching for file chunk 1/7
legal-search-api-dummy  | INFO:src.api.main:FAISS search took 0.014510s for k=3
legal-search-api-dummy  | INFO:src.api.main:Result aggregation took 0.000117s, aggregated to 3 docs.
legal-search-api-dummy  | INFO:src.api.main:Searching for file chunk 2/7
legal-search-api-dummy  | INFO:src.api.main:FAISS search took 0.006628s for k=3
legal-search-api-dummy  | INFO:src.api.main:Result aggregation took 0.000049s, aggregated to 3 docs.
legal-search-api-dummy  | INFO:src.api.main:Searching for file chunk 3/7
legal-search-api-dummy  | INFO:src.api.main:FAISS search took 0.006663s for k=3
legal-search-api-dummy  | INFO:src.api.main:Result aggregation took 0.000037s, aggregated to 2 docs.
legal-search-api-dummy  | INFO:src.api.main:Searching for file chunk 4/7
legal-search-api-dummy  | INFO:src.api.main:FAISS search took 0.006481s for k=3
legal-search-api-dummy  | INFO:src.api.main:Result aggregation took 0.000054s, aggregated to 2 docs.
legal-search-api-dummy  | INFO:src.api.main:Searching for file chunk 5/7
legal-search-api-dummy  | INFO:src.api.main:FAISS search took 0.006474s for k=3
legal-search-api-dummy  | INFO:src.api.main:Result aggregation took 0.000058s, aggregated to 3 docs.
legal-search-api-dummy  | INFO:src.api.main:Searching for file chunk 6/7
legal-search-api-dummy  | INFO:src.api.main:FAISS search took 0.006651s for k=3
legal-search-api-dummy  | INFO:src.api.main:Result aggregation took 0.000047s, aggregated to 2 docs.
legal-search-api-dummy  | INFO:src.api.main:Searching for file chunk 7/7
legal-search-api-dummy  | INFO:src.api.main:FAISS search took 0.006478s for k=3
legal-search-api-dummy  | INFO:src.api.main:Result aggregation took 0.000039s, 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 0.000095s for 2 docs.
legal-search-api-dummy  | INFO:src.api.main:Core search logic (embedding, search, aggregation, formatting) for ONNX model finished in 2.9645s. Returning 2 documents.
```



Search Results

<< New Search

Found 2 results for query/file: "Commonwealth v. Zaza.PDF"

Case Name: Case Name for Commonwealth v. Zaza., 2025 Mass. App. Unpub. LEXIS 309
Citation: Citation 1984
Source File: [Download] Commonwealth v. Zaza., 2025 Mass. App. Unpub. LEXIS 309.PDF
Similarity Score: 1.0000

▶ Show More Details & Relevant Context

Is this result relevant? Correct 👍 Incorrect 👎

Case Name: Case Name for Commonwealth v. Gousie., 2025 Mass. App. Unpub. LEXIS 327
Citation: Citation 1776
Source File: [Download] Commonwealth v. Gousie., 2025 Mass. App. Unpub. LEXIS 327.PDF
Similarity Score: 0.1040

▶ Show More Details & Relevant Context

Is this result relevant? Correct 👍 Incorrect 👎

Load testing for 50 users (5 concurrent at a time) :

```
sanju@Sanjeevans-MacBook-Air serving_dummy % python3 src/test/load_test_api.py --concurrent 5 --total_requests 50
2025-05-11 11:44:09,974 - 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 11:44:25,172 - ERROR - Request 7 failed with unexpected error: [Errno 54] Connection reset by peer
2025-05-11 11:44:25,175 - ERROR - Request 9 failed with unexpected error: [Errno 54] Connection reset by peer
2025-05-11 11:45:07,643 - ERROR - Request 26 failed with unexpected error: [Errno 54] Connection reset by peer
2025-05-11 11:45:30,401 - ERROR - Request 37 failed with unexpected error: [Errno 54] Connection reset by peer
2025-05-11 11:45:53,371 - ERROR - Request 46 failed with unexpected error: [Errno 54] Connection reset by peer
2025-05-11 11:46:03,341 - INFO - Load Test Summary:
2025-05-11 11:46:03,341 - INFO -   Total requests sent: 50
2025-05-11 11:46:03,341 - INFO -   Successful requests: 45
2025-05-11 11:46:03,341 - INFO -   Failed requests: 5
2025-05-11 11:46:03,341 - INFO -   Total time taken: 113.3671 seconds
```

```
2025-05-11 11:46:03,341 - INFO -   Requests per second (RPS): 0.4410
2025-05-11 11:46:03,341 - INFO -   Avg Latency (successful): 11.6567s
2025-05-11 11:46:03,341 - INFO -   Min Latency (successful): 2.6950s
2025-05-11 11:46:03,341 - INFO -   Max Latency (successful): 24.8540s
2025-05-11 11:46:03,342 - INFO -   P95 Latency (successful): 17.6710s
```

ONNX with Dynamic Quantization (INT8) :

```
#Step 4To create ONNX model :
python3 src/processing/quantize_onnx_model.py
```

legal_bert_finetuned_onnx_int8_quantized model size : 174M

For our third test, we took the ONNX model and 'quantized' it to INT8. This made the model much smaller – only 174MB, down from 417MB (ONNX) and 419MB (original)!

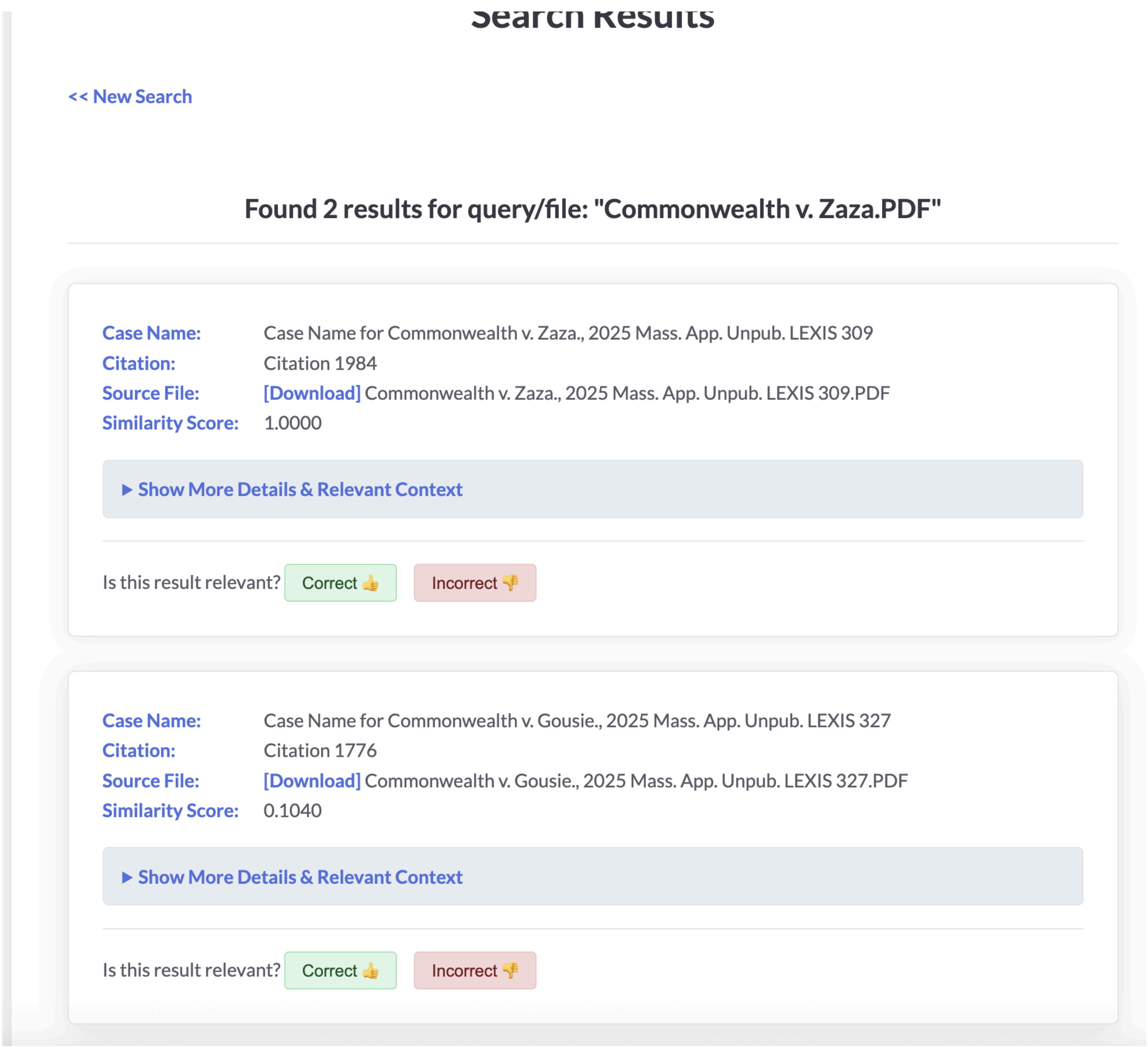
When searching with a single PDF (Commonwealth v. Zaza.PDF), this quantized model was the fastest so far, taking about 2.85 seconds to process the PDF and get results (compared to 2.9s for regular ONNX and 3.3s for the original). Importantly, it gave the exact same top 2 search results as the other models.

In the load test (50 requests, 5 users at once, Glover v. State.PDF), the quantized model performed very similarly to the regular ONNX model. It handled about 0.44 requests per second (RPS). Successful requests took around 11.8 seconds on average, with 95% finishing in about 17.7 seconds. We still had 5 failed requests, the same as the other tests.

So, INT8 quantization dramatically reduced the model size and kept the slight speed improvements we saw with ONNX, all without changing the search results for our test case. The system still seems to struggle with 5 users at once, suggesting the main slowdown isn't just the model anymore.

Below are the logs and output.

```
legal-search-api-dummy  | INFO:src.api.main:Processing uploaded file: Commonwealth v. Zaza.PDF
legal-search-api-dummy  | INFO:src.api.main:Split uploaded file into 7 chunks.
legal-search-api-dummy  | INFO:src.api.main:Embedding 7 file chunks using ONNX model
legal-search-api-dummy  | INFO:src.api.main:File query embedding (7 chunks) took 2.847378s.
legal-search-api-dummy  | INFO:src.api.main:Searching for file chunk 1/7
legal-search-api-dummy  | INFO:src.api.main:FAISS search took 0.008008s for k=3
legal-search-api-dummy  | INFO:src.api.main:Result aggregation took 0.000044s, aggregated to 3 docs.
legal-search-api-dummy  | INFO:src.api.main:Searching for file chunk 2/7
legal-search-api-dummy  | INFO:src.api.main:FAISS search took 0.009619s for k=3
legal-search-api-dummy  | INFO:src.api.main:Result aggregation took 0.000037s, aggregated to 3 docs.
legal-search-api-dummy  | INFO:src.api.main:Searching for file chunk 3/7
legal-search-api-dummy  | INFO:src.api.main:FAISS search took 0.006729s for k=3
legal-search-api-dummy  | INFO:src.api.main:Result aggregation took 0.000036s, aggregated to 2 docs.
legal-search-api-dummy  | INFO:src.api.main:Searching for file chunk 4/7
legal-search-api-dummy  | INFO:src.api.main:FAISS search took 0.006911s for k=3
legal-search-api-dummy  | INFO:src.api.main:Result aggregation took 0.000037s, aggregated to 2 docs.
legal-search-api-dummy  | INFO:src.api.main:Searching for file chunk 5/7
legal-search-api-dummy  | INFO:src.api.main:FAISS search took 0.006843s for k=3
legal-search-api-dummy  | INFO:src.api.main:Result aggregation took 0.000037s, aggregated to 3 docs.
legal-search-api-dummy  | INFO:src.api.main:Searching for file chunk 6/7
legal-search-api-dummy  | INFO:src.api.main:FAISS search took 0.006854s 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 7/7
legal-search-api-dummy  | INFO:src.api.main:FAISS search took 0.006651s for k=3
legal-search-api-dummy  | INFO:src.api.main:Result aggregation took 0.000035s, 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 0.000126s for 2 docs.
legal-search-api-dummy  | INFO:src.api.main:Core search logic (embedding, search, aggregation, formatting) for ONNX model finished in 2.9321s. Returning 2 documents.
legal-search-api-dummy  | INFO:     100.1.162.2:56904 - "POST /search_combined HTTP/1.1" 200 OK
legal-search-api-dummy  | INFO:     172.18.0.2:45964 - "GET /metrics HTTP/1.1" 200 OK
legal-search-api-dummy  | INFO:     172.18.0.2:45350 - "GET /metrics HTTP/1.1" 200 OK
```



Search Results

<< New Search

Found 2 results for query/file: "Commonwealth v. Zaza.PDF"

Case Name: Case Name for Commonwealth v. Zaza., 2025 Mass. App. Unpub. LEXIS 309
Citation: Citation 1984
Source File: [Download] Commonwealth v. Zaza., 2025 Mass. App. Unpub. LEXIS 309.PDF
Similarity Score: 1.0000

Case Name: Case Name for Commonwealth v. Gousie., 2025 Mass. App. Unpub. LEXIS 327
Citation: Citation 1776
Source File: [Download] Commonwealth v. Gousie., 2025 Mass. App. Unpub. LEXIS 327.PDF
Similarity Score: 0.1040

Load testing for 50 users (5 concurrent at a time) :

```
sanju@Sanjeevans-MacBook-Air serving_dummy % python3 src/test/load_test_api.py --concurrent 5 --total_requests 50
2025-05-11 12:00:51,922 - 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 12:01:07,007 - ERROR - Request 9 failed with unexpected error: [Errno 54] Connection reset by peer
2025-05-11 12:01:07,008 - ERROR - Request 10 failed with unexpected error: [Errno 54] Connection reset by peer
2025-05-11 12:01:27,328 - ERROR - Request 19 failed with unexpected error: [Errno 54] Connection reset by peer
2025-05-11 12:02:02,739 - ERROR - Request 32 failed with unexpected error: [Errno 54] Connection reset by peer
2025-05-11 12:02:25,550 - ERROR - Request 44 failed with unexpected error: [Errno 54] Connection reset by peer
2025-05-11 12:02:45,697 - INFO - Load Test Summary:
2025-05-11 12:02:45,698 - INFO -   Total requests sent: 50
```

```
2025-05-11 12:02:45,698 - INFO -   Successful requests: 45
2025-05-11 12:02:45,698 - INFO -   Failed requests: 5
2025-05-11 12:02:45,698 - INFO -   Total time taken: 113.7756 seconds
2025-05-11 12:02:45,698 - INFO -   Requests per second (RPS): 0.4395
2025-05-11 12:02:45,698 - INFO -   Avg Latency (successful): 11.8050s
2025-05-11 12:02:45,698 - INFO -   Min Latency (successful): 2.6214s
2025-05-11 12:02:45,698 - INFO -   Max Latency (successful): 20.3186s
2025-05-11 12:02:45,698 - INFO -   P95 Latency (successful): 17.6909s
```