P7.1.1

JSON[

{

"page": 0,

"doc\_id": 0,

"doc\_type": "Lender Fee Sheet"

},

{

"page": 1,

"doc\_id": 1,

"doc\_type": "Payslip"

},

{

"page": 2,

"doc\_id": 2,

"doc\_type": "Contract"

},

{

"page": 3,

"doc\_id": 2,

"doc\_type": "Contract"

},

{

"page": 4,

"doc\_id": 2,

"doc\_type": "Contract"

},

{

"page": 5,

"doc\_id": 2,

"doc\_type": "Contract"

},

{

"page": 6,

"doc\_id": 2,

"doc\_type": "Contract"

}

]

P7.1.1

SUMMARY

I used the **Mistral 7B Instruct v0.2 (GGUF)** model to power my document segmentation pipeline. This model was chosen for its strong instruction-following capabilities and its ability to generate reliable structured output (JSON), which was essential for this task. It offers a great balance of performance and efficiency, running effectively on a standard Colab T4 GPU.

A tricky edge case was correctly identifying the "Annexure A" page of the contract. Initially, the model flagged it as a new document due to the distinct header. I refined the prompt to instruct the model that appendices and annexures are continuations, which improved its accuracy. Another challenge was handling pages with very little text, which required adding a fallback in the code in case the model's JSON output was malformed.

To improve reliability further in a production system, I would implement a confidence scoring mechanism where the LLM also returns a confidence level for its prediction. Any classification below a certain threshold would be automatically flagged for human review, creating a robust human-in-the-loop system.