High-Performance Document Processing System

Requirements: Create a system for rapid processing of large PDF documents (100+ pages) with efficient data extraction and analysis capabilities and an LLM should answer based on the PDF's only. Use only basic libraries (If you can do a task manually without the help of an external library, you should proceed with that)

Note: Limit the input to pdf and excel sheet

- Processing Pipeline
  - Implement parallel/async PDF processing
  - Optimize text extraction and chunking
  - Design memory-efficient data structures
  - Handle complex document elements (tables/charts)
- Performance Goals
  - Process 100-page PDF < 30 seconds
  - Memory usage < 1GB
  - Support concurrent processing
  - Maintain extraction accuracy > 95%

Submission Instructions: Publish to Github, make it public and share the link of the repo in the google form: link (Also available in the mail)

- Detailed README explaining architecture and optimization approaches
- Performance benchmarks
- Demo video showing processing speed and memory usage
- Instructions for deployment and testing