Components of Code

Here are the details of the files present in the code:

1. **config.json**

This file has the input data that needs to be given by user and is the only file user must change. It takes Groq API key, model name, location of all input files with name and the Natural Language Instructions for Compliance Checks.

2. **file_loader.py**

This program loads the file from the path given in config. The code can accept 4 files: Invoice, Purchase Order (PO), GRN and Vendor Policy Document. The code can work even if some of these files are missing. If the file is not present, please give an empty string (“”) in the path.

I have used images library to load the image and pdf2image to change the pdf to image and pytesseract and tesseract-ocr to parse the file and get the text from each input file.

3. **prompt_builder.py**

This file takes documents as input in the form of dictionary (filename: text) and the rules as list of sentences separated by comma and create a prompt which can be given to Groq API as prompt

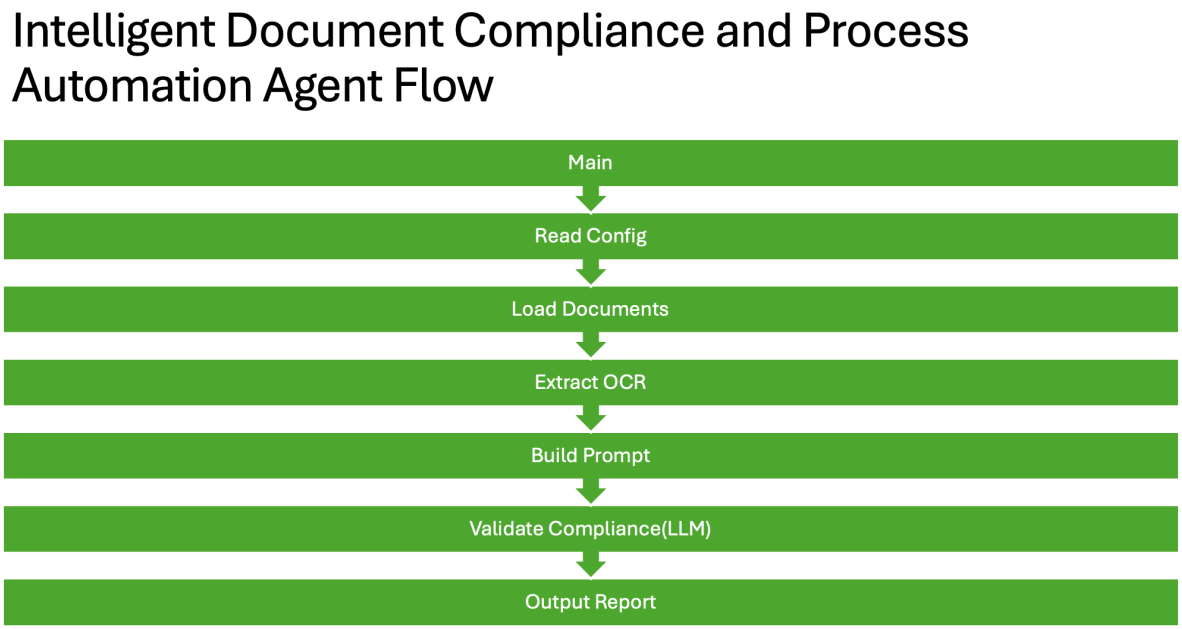
4. **compliance_validator.py**

This takes prompt generated by prompt_builder and Groq API details as input and query the Groq API using the given prompt and return the response given by Groq

5. **main.py**

This is the main file which calls all other files and methods including config file and generate output and print the same.

Here is the Flow diagram from the same:



Assumptions

The input files will strictly be PDF or Image or Text. The Person who is running the code will have access to Groq API.

Limitations and Future Improvements

1. Currently the validation is being done by LLM based on the text from files and the Natural Language Rules. If we must implement it locally, we can implement it moving forward by making LLM only taking rules and returning the conditions in programmable format and the validation can be done locally
2. Currently, the file only accepts PDF or Image or Text, but we can increase the acceptable formats as and when required.
3. The code will work even if some files are missing.