1. Pipeline is constructed for Local models, parsers and prompt template, issues still remaining:
   1. Llama returning prompt itself as return and hit output token limit often for this feature
   2. Integrate gpt4o-mini endpoint into pipeline failed
2. Confusion Matrix tested out for GPT4-o mini, for list objects type of questions:
   1. Other type of short answer question: use a line to explain reason need to be prompted?
   2. Deploy into use for more Data