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

Read documatation for chain, pipeline

Llama end and eos token iis also on documentation

Turn numbers in prompts into parameters, and give numbering in the result as telling them that

Right out tests

Small function do that what big one does in paper

Build cases for evaluation functions

For smaller models change many to many to 1 to many, **for list type of short text answers, speed speed speed case**

Carry on 4-o, split to 1 to many, in the chain

Plan for 2 months work ,

Use 4o to split because to get ready to move to a small model