LLMs:

- Llama3.1:70B: https://ollama.com/library/llama3.1
  - 70 Billion Parameter LLM model developed by Meta.
  - In this program, Llama is the first LLM to run and takes on the brunt of the run time.
  - Produces an output and sends its output to the second LLM.
  - It is important to note that the first LLM to run MUST have a larger parameter set.

- Phi-4:14B: https://ollama.com/library/phi4
  - 14 Billion Parameter LLM model developed by Microsoft
  - In this program, Phi-4 is the second LLM to run and takes considerably less time to run.
  - Refines the output of the first LLM and sends its own output to the user.
  - The second LLM can have a parameter size of UP TO the number the first LLM has.

Alternative LLMs:

- Our program is able to "plug and play" different LLMs into the system depending on use cases. Any of the following have been tested through Ollama and work in the rag.py application.
  - Deepseek-R1
  - Llama3.3
  - Mistral
  - Gemma2

- Other LLMs on Ollama's database should also work, but further testing would be required.

Failed LLMs:

- Falcon2
- During our testing, Falcon2 was not compatible with our database and rag programs causing an abstract data error.