Dr. Sanjan Gupta
PhD in Applied AI| UW Madison, USA
IIT-Madras | B.Tech. Honors alum
tpsanjan@gmail.com

For GenAI demos

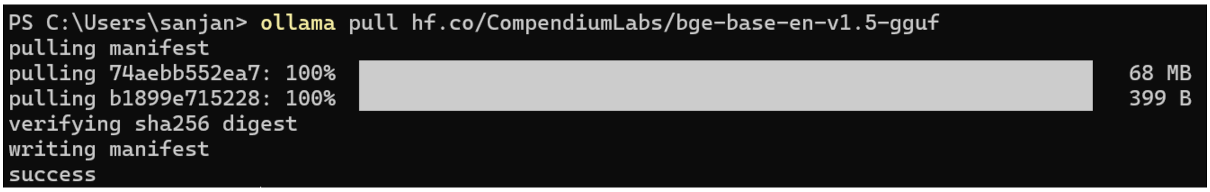
***ETI – Estimated Time to Install when tested on a 16GB RAM, i7 11th gen laptop with ~30 Mbps internet connection speed**

- Download Ollama locally from https://ollama.com/
- After installing ollama, run the following commands in terminal to install embedding & language models:

  >> ollama pull hf.co/CompendiumLabs/bge-base-en-v1.5-gguf

  >> ollama pull hf.co/bartowski/Llama-3.2-1B-Instruct-GGUF

  [ETI < 5 min]



```
PS C:\Users\sanjan> ollama pull hf.co/CompendiumLabs/bge-base-en-v1.5-gguf
pulling manifest
pulling 74aebb552ea7: 100%                                                      68 MB
pulling b1899e715228: 100%                                                     399 B
verifying sha256 digest
writing manifest
success
```

  If you see the above output, it means the models are successfully downloaded!

- Now, install ollama via pip

  >> pip install ollama

- Download the dataset (rag_cat_facts.txt) and python script (rag_ollama_demo.py)