Base AI System install

notes only

```
https://www.simplified.guide/ubuntu/remove-snapd
sudo systemctl stop snapd
sudo apt remove --purge --assume-yes snapd gnome-software-plugin-snap
rm -rf ~/snap/
rm -rf /var/cache/snapd/
apt remove squashfs-tools
cat << EOD >> /etc/apt/preferences.d/nosnap.pref
Package: snapd
Pin: release a=*
Pin-Priority: -10
EOD
sudo apt update


Ubuntu Server 22.04 LTS install with third party drivers
---------------------------------------------------------

Note: CUDA toolkit not required, only drivers.

ubuntu-drivers list --gpgpu
ubuntu-drivers install --gpgpu
ubuntu-drivers autoinstall
apt-get install nvtop

#Install docker for user
curl -fSSL get.docker.com | sh

#prepare, using Mistral-7B-OpenOrca Model, remove the LLAMACPP entries to
use default
mkdir /opt/privategpt
chmod 777 /opt/privategpt/
cat << EOF >> docker-privategpt.yml
services:
  # https://hub.docker.com/r/3x3cut0r/privategpt
  privategpt:
    image: 3x3cut0r/privategpt:latest
    container_name: privategpt
    environment:
      LLAMACPP_LLM_HF_REPO_ID: "TheBloke/Mistral-7B-OpenOrca-GGUF"
      LLAMACPP_LLM_HF_MODEL_FILE: "mistral-7b-openorca.Q5_K_M.gguf"
      LLAMACPP_EMBEDDING_HF_MODEL_NAME: "BAAI/bge-large-en-v1.5"
      EMBEDDING_INGEST_MODE: "parallel"
      EMBEDDING_COUNT_WORKERS: "4"
    volumes:
```

```
      - /opt/privategpt:/home/worker/app/models
    ports:
      - 8080:8080/tcp
    runtime: nvidia
    deploy:
      resources:
        reservations:
          devices:
            - capabilities: [gpu]
EOF


curl -fsSL https://nvidia.github.io/libnvidia-container/gpgkey | sudo gpg --
dearmor -o /usr/share/keyrings/nvidia-container-toolkit-keyring.gpg   &&
curl -s -L
https://nvidia.github.io/libnvidia-container/stable/deb/nvidia-container-too
lkit.list |     sed 's#deb https://#deb [signed-
by=/usr/share/keyrings/nvidia-container-toolkit-keyring.gpg] https://#g' |
sudo tee /etc/apt/sources.list.d/nvidia-container-toolkit.list
sudo apt-get update
sudo apt-get install -y nvidia-container-toolkit

nvidia-ctk runtime configure --runtime=docker
sudo systemctl restart docker


docker compose -f privategpt up -d

dockerfile:
COPY utils.py ingest.py constants.py ./

apt-get install libgl1

docker run -it --mount src="$HOME/.cache",target=/root/.cache,type=bind --
gpus=all localgpt
```