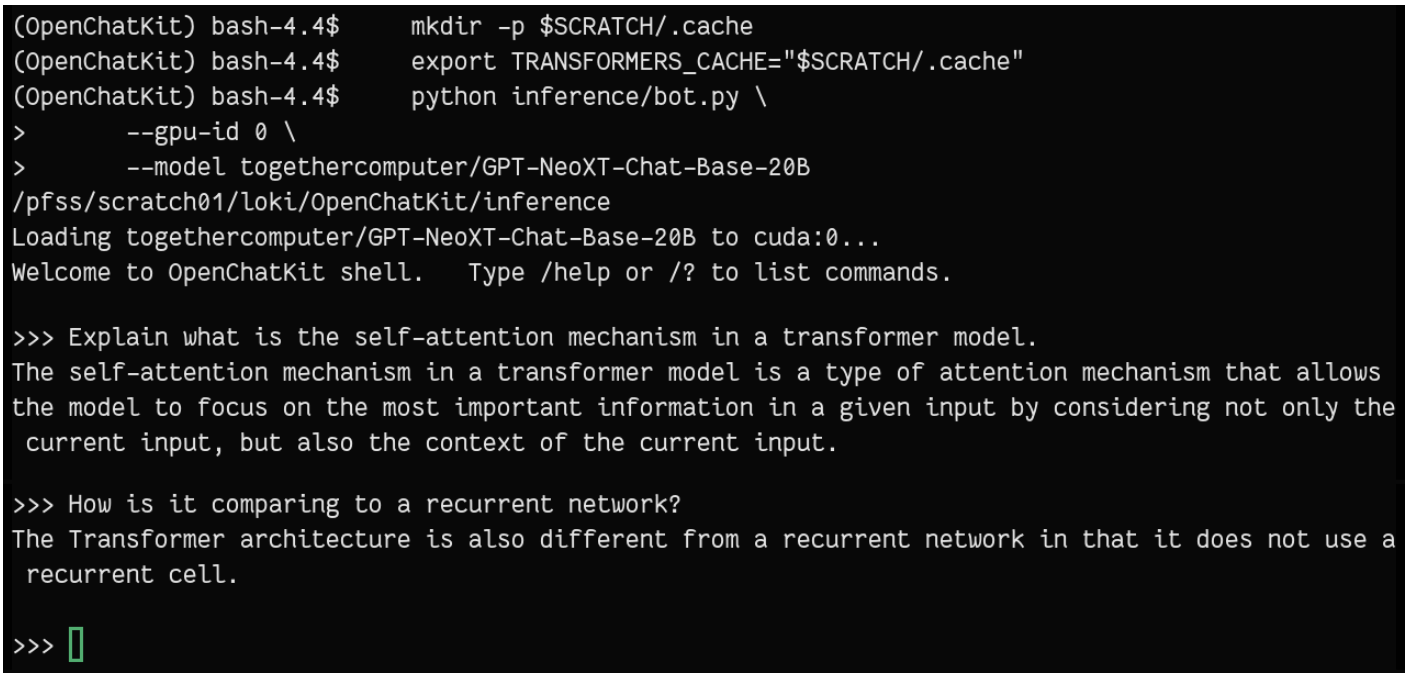
# Run and train chatbots with OpenChatKit




OpenChatKit provides an open-source framework to train general-purpose chatbots. It includes a pre-trained 20B parameter language model as a good starting point.

> At least 40GB of VRAM is required to load the 20B model. So a full 80GB A100 is required.

Firstly, we will prepare the Conda environment. Let's request an interactive shell from a compute node.

```
srun -N1 -c8 -p batch --pty bash
```

Run the following commands inside the interactive shell.

```
# the pre-trained 20B model takes 40GB of space, so we use the scratch folder
cd $SCRATCH


# check out the kit
module load Anaconda3/2022.05 GCCcore git git-lfs
```

```
git clone https://github.com/togethercomputer/OpenChatKit.git
cd OpenChatKit
git lfs install

# configure conda to use the user SCRATCH folder to store envs
echo "
pkgs_dirs:
  - $SCRATCH/.conda/pkgs
envs_dirs:
  - $SCRATCH/.conda/envs
channel_priority: flexible
" > ~/.condarc

# create the Conda environment based on the provided environment.yml
# it may takes over an hour to resolve and install all python dependencies
conda env create --name OpenChatKit -f environment.yml python=3.10.9

# verify it is created
conda env list
exit
```

We are ready to boot up the kit and load the pre-trained model. This time we will request a node with an 80GB A100 GPU.

```
srun -c8 --mem=100000 --gpus a100:1 -p gpu --pty bash
```

Run the following commands inside the shell to start the chatbot.

```
# load the modules we need
module load Anaconda3/2022.05 GCCcore git git-lfs CUDA

# go to the kit and activate the environment
cd $SCRATCH/OpenChatKit
source activate OpenChatKit

# set the cache folder to store the downloaded pre-trained model
mkdir -p $SCRATCH/.cache
export TRANSFORMERS_CACHE="$SCRATCH/.cache"

# start the bot (the first time take longer to download the model)
```

```
python inference/bot.py \
  --gpu-id 0 \
  --model togethercomputer/GPT-NeoXT-Chat-Base-20B
```

To train and finetune the model, please check out this section in their git repo.

---

Revision #11
Created 22 March 2023 06:57:28 by Loki Ng
Updated 27 March 2023 02:05:04 by Loki Ng