

Runpod Model Loading Steps

POD Creation

- To get started, create a pod with the "RunPod Text Generation UI" template from this link:
<https://runpod.io/gsc?template=fy7cw0s6xz&ref=geta6cef>

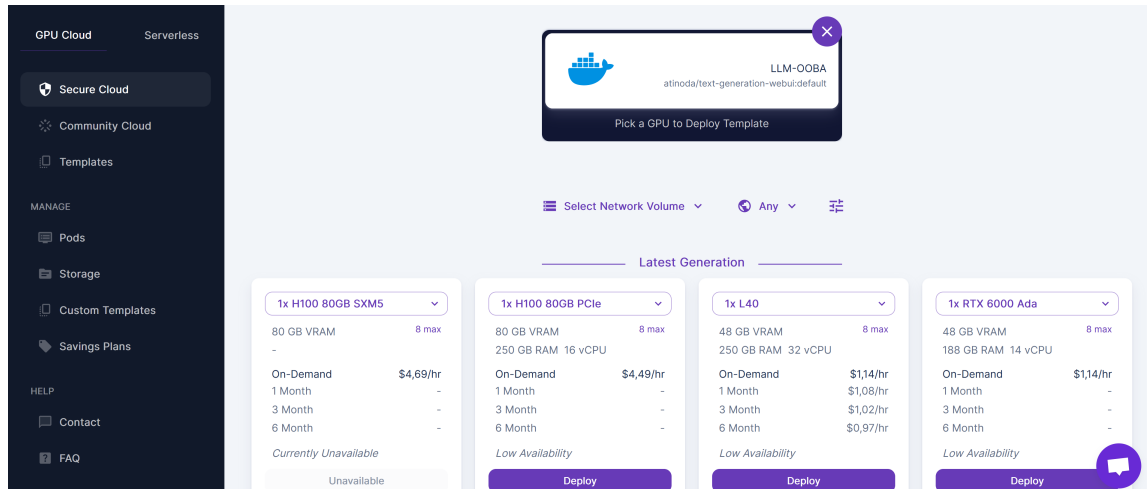


Figure 1: RunPod Text Generation UI Home Page

- Choose an instance to start creating, Example: 1x RTX A6000

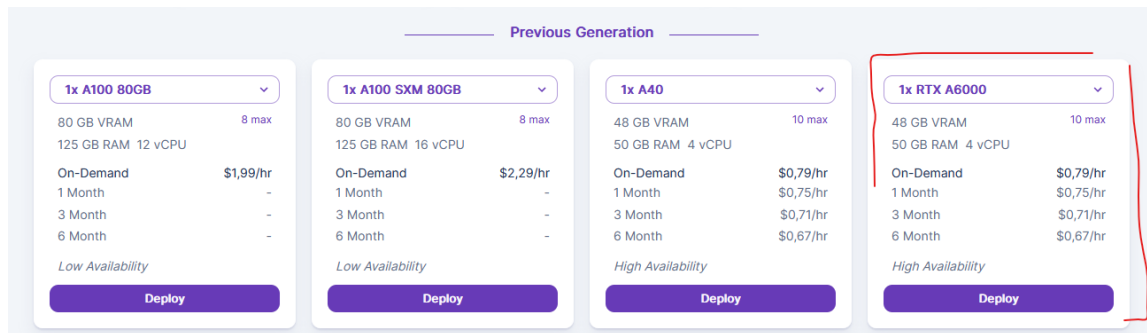


Figure 2: List of Generations

- Choose a template based on your preferences, then click "Continue" button. Example "Run-Pod TheBloke LLMs" template:

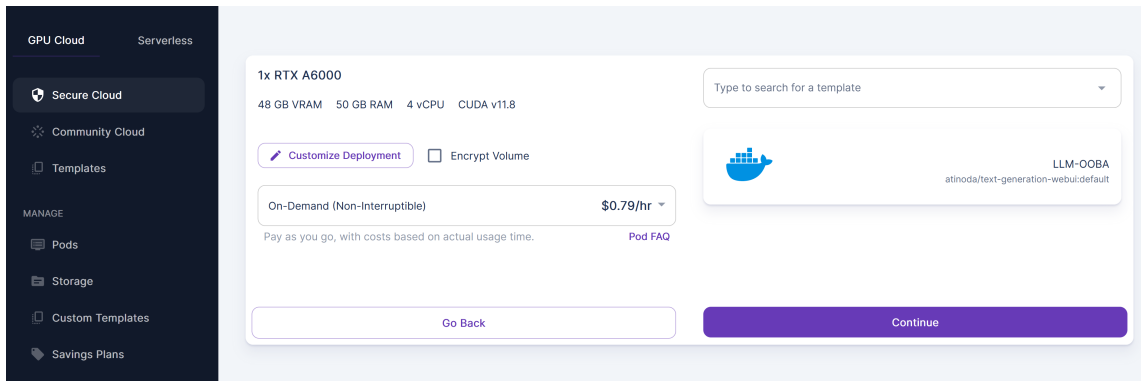


Figure 3: Template choice

This is how the resulting instance looks like. Press Connect:

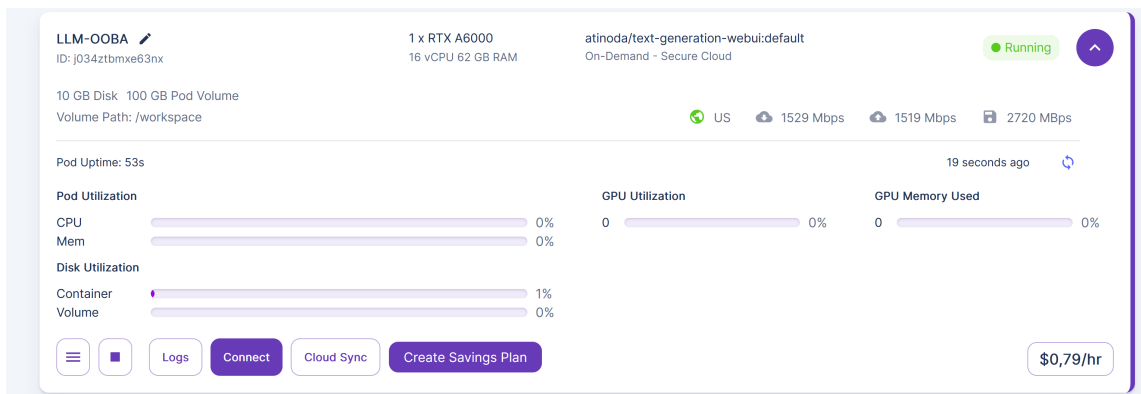


Figure 4: Instance Example

- After that, open the pop-up tab "Connection Option".
- Click "Connect to HTTP Service [Port 7860]"

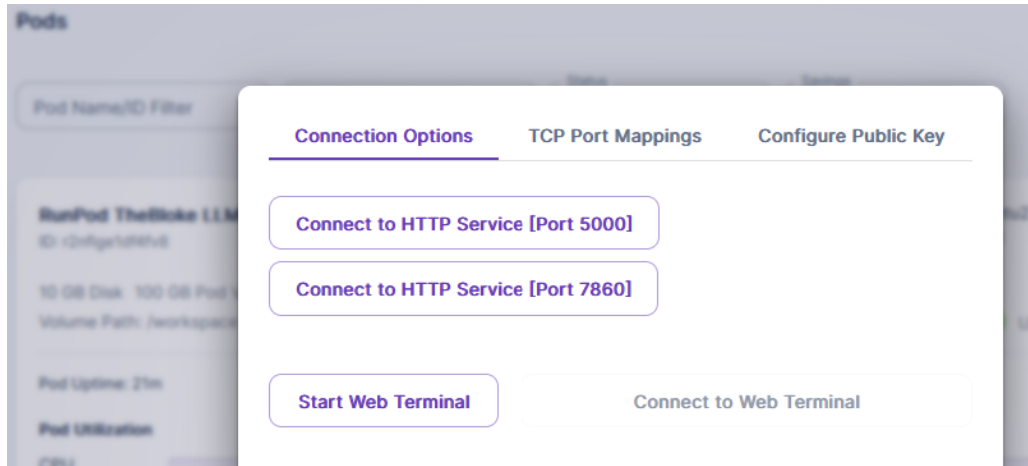


Figure 5: Connection Options

Model Loading

RunPod model loading. In this window, you should download a model and load it into your instance:

- Choose a model and press "Download".
- Load your downloaded model.

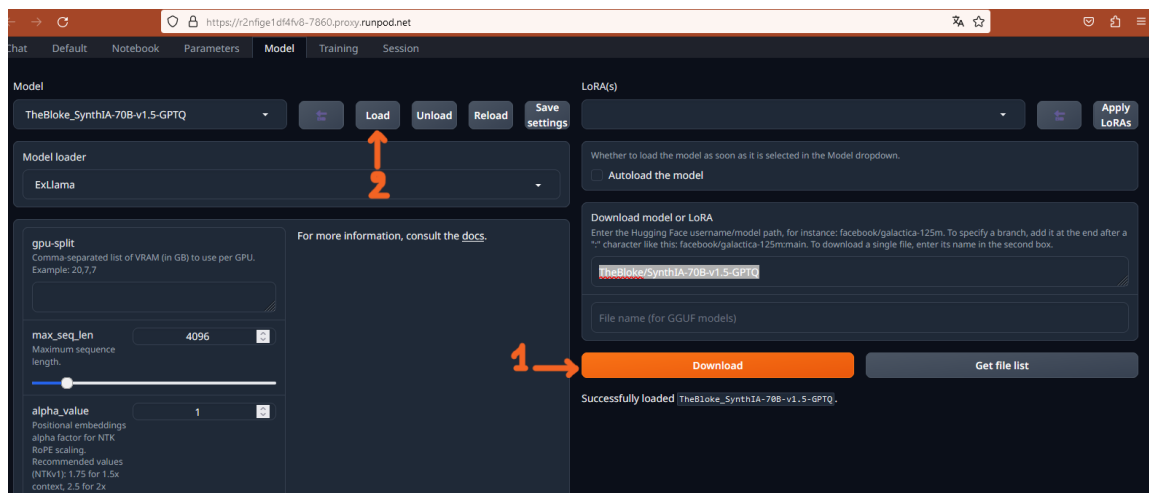
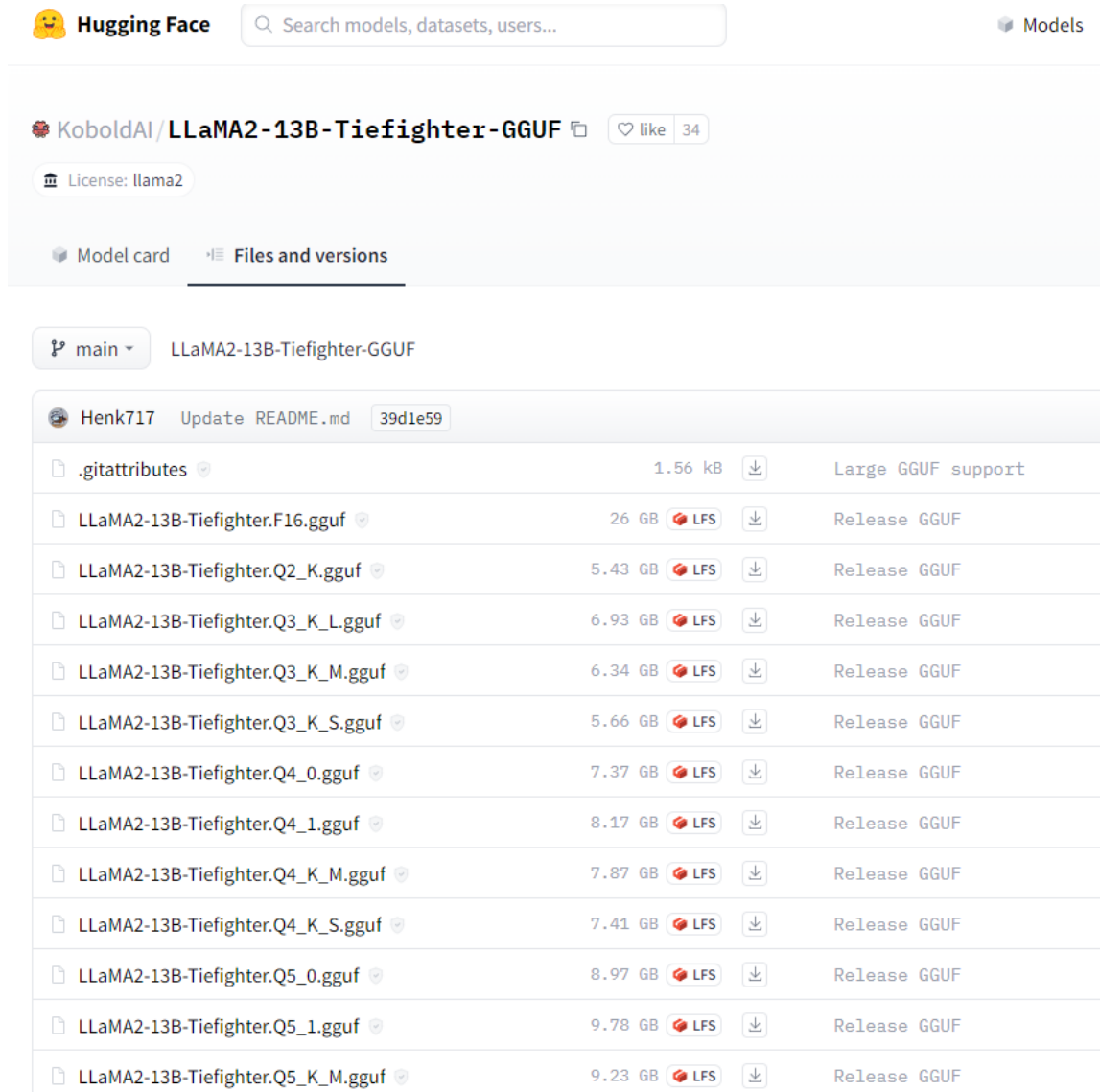


Figure 6: Model Downloading window

Choosing a GGUF Model

If you are using a GGUF model:

- Locate a hugging face model GGUF, for example: <https://huggingface.co/KoboldAI/LLaMA2-13B-Tiefighter-GGUF>.
- Click the Files and versions tab. You only need a single GGUF file. The higher Q versions produce better output, but may output text more slowly. K_M produces better output than K_S.



The screenshot shows the Hugging Face interface for the model **KoboldAI/LLaMA2-13B-Tiefighter-GGUF**. The page includes a search bar, a license tag (llama2), and navigation tabs for 'Model card' and 'Files and versions'. The 'Files and versions' tab is active, displaying a list of files with their sizes, LFS status, and download links. The files are organized by version (Q1 to Q5) and quantization level (F16, K_L, K_M, K_S).

File Name	Size	LFS	Download	Description
.gitattributes	1.56 kB		↓	Large GGUF support
LLaMA2-13B-Tiefighter.F16.gguf	26 GB	LFS	↓	Release GGUF
LLaMA2-13B-Tiefighter.Q2_K.gguf	5.43 GB	LFS	↓	Release GGUF
LLaMA2-13B-Tiefighter.Q3_K_L.gguf	6.93 GB	LFS	↓	Release GGUF
LLaMA2-13B-Tiefighter.Q3_K_M.gguf	6.34 GB	LFS	↓	Release GGUF
LLaMA2-13B-Tiefighter.Q3_K_S.gguf	5.66 GB	LFS	↓	Release GGUF
LLaMA2-13B-Tiefighter.Q4_0.gguf	7.37 GB	LFS	↓	Release GGUF
LLaMA2-13B-Tiefighter.Q4_1.gguf	8.17 GB	LFS	↓	Release GGUF
LLaMA2-13B-Tiefighter.Q4_K_M.gguf	7.87 GB	LFS	↓	Release GGUF
LLaMA2-13B-Tiefighter.Q4_K_S.gguf	7.41 GB	LFS	↓	Release GGUF
LLaMA2-13B-Tiefighter.Q5_0.gguf	8.97 GB	LFS	↓	Release GGUF
LLaMA2-13B-Tiefighter.Q5_1.gguf	9.78 GB	LFS	↓	Release GGUF
LLaMA2-13B-Tiefighter.Q5_K_M.gguf	9.23 GB	LFS	↓	Release GGUF

Figure 7: GGUF Models - Huggingface

Generally a Q4_K_M or Q5_K_M version is recommended if available. Sometimes in the Model Card it also recommends which GGUF to use.

In the model loading window:

- Click the top copy link, and paste that into the top field of the ooba Download model section.
- Copy and paste the .gguf filename you chose into the bottom field for the GGUF download model.
- When using GGUF files with llama.cpp, you MUST SET THE n-gpu-layers TO A HIGH NUMBER, somewhere between 20 to 100 is recommended, depending on the model and GPU. You may need to try various numbers and see if the speed is changed. In my tests on an RTX A6000, setting n-gpu-layers to 100 works well. If you don't set this, it will be using the CPU instead of the GPU and be very slow.

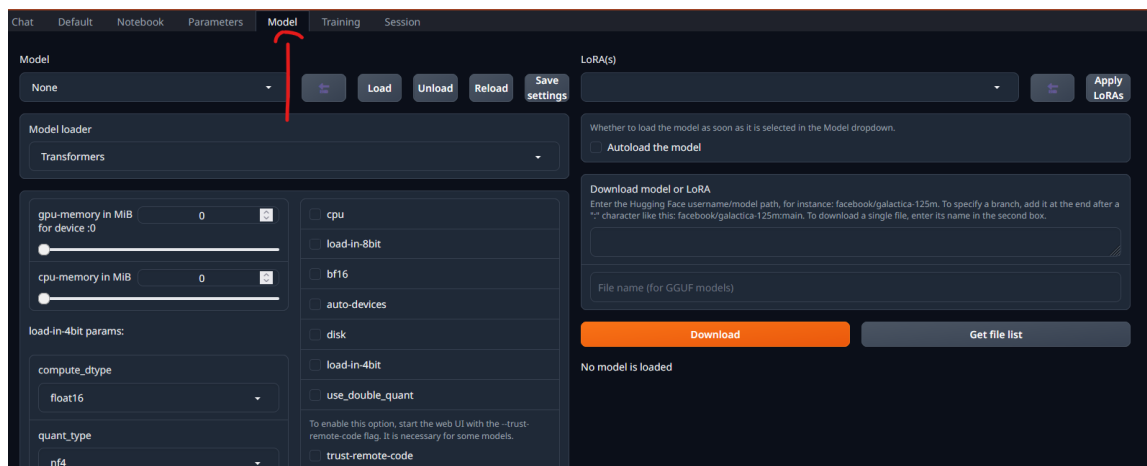


Figure 8: Click Model Tab

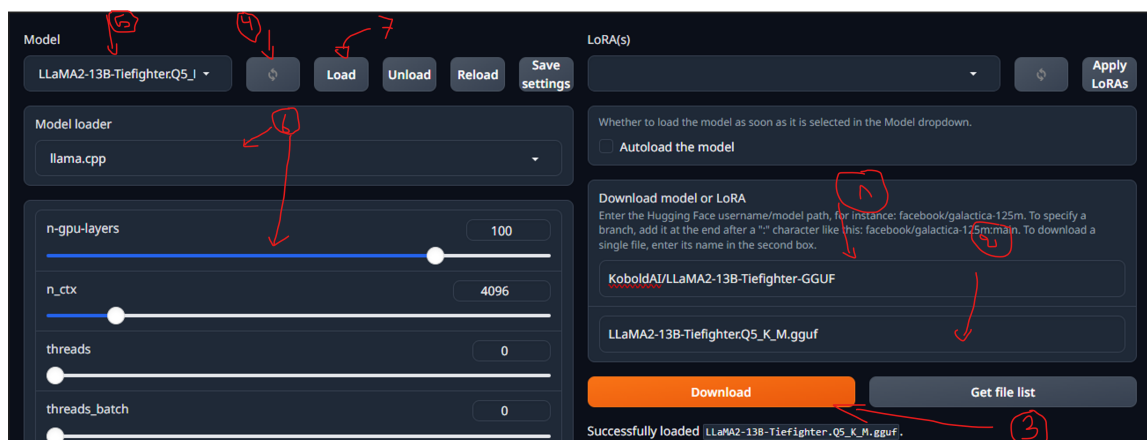


Figure 9: GGUF Model loading

- Complete all settings and load your model.

Access to Server

After "Connect", open the "TCP Port Mappings" tab. You will find below the IP address and Port.

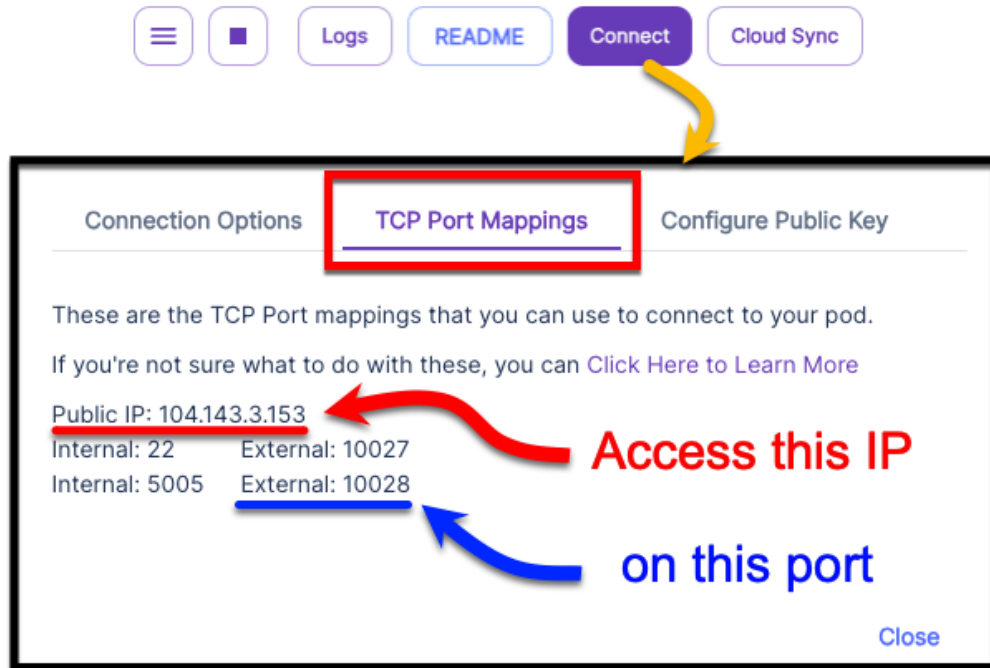


Figure 10: TCP Port Mappings - Access location

In our Interface select the custom ooba option in the dropdown and then put the IP:PORT



Figure 11: Custom Server - Access location