Serving Nagios Enterprises' Large Language Models

Purpose

This document provides instructions for serving Nagios Enterprises' Large Language Models using vLLM. It covers the prerequisites, installation process, and basic setup for running a model server.

The guide recommends using NVIDIA PyTorch Docker image for ease of deployment and includes steps for downloading and setting up the 'llama-3-lucene-8b' model. By following these instructions, users can set up an OpenAI-API compatible server running Nagios's LLAMA 3.1-based model.

Prerequisites

- OS: Linux
- GPU: Compute capability 7.0 or higher (e.g., V100, T4, RTX20xx, A100, L4, H100, etc.)
- Docker: For installation instructions, visit Docker's official installation guide
- NVIDIA Container Toolkit: You'll need to install the appropriate NVIDIA Container Toolkit for your specific operating system, distribution, and hardware configuration. Please refer to the official NVIDIA documentation for detailed installation instructions.
- NVIDIA GPU Drivers: Ensure you have the appropriate NVIDIA GPU drivers installed for your hardware. On Ubuntu systems, you can use the following command to automatically install the recommended drivers only AFTER installing the nvidia container toolkit.

```
sudo ubuntu-drivers autoinstall
```

Installation – Containerized Hosting

We recommend using the NVIDIA PyTorch Docker image.

It is very important that you have the necessary prerequisites, otherwise the container won't work. Run the following:

```
docker run --gpus all -it --rm --ipc=host -p 8000:8000 nvcr.io/nvidia/pytorch:23.10-py3
```

This command will drop you into the container with a terminal interface.

If you just installed docker and run into an error at this point, see troubleshooting.

From there, you will install the model and vLLM:

```
pip install vllm
mkdir models
wget -P models https://assets.nagios.com/downloads/models/llama-3-lucene-8b.tar.gz
tar -xvf models/llama-3-lucene-8b.tar.gz -C models
rm -rf models/llama-3-lucene-8b.tar.gz
```

You are now ready to run the server:

```
vllm serve models/llama-3-lucene-8b
```

This command spins up an OpenAI-API Compatible Server that runs the Nagios model 'llama-3-lucene-8b', which is based on LLAMA 3.1.

If you run into an error at this point, see troubleshooting.

To test the server, you can visit *http://localhost:8000/v1/models* in your web browser or use a tool like curl:

```
curl http://localhost:8000/v1/models
```

This test assumes the model is exposed at port 8000 and should return a response with information about the available models.

Installation – Non-Containerized Hosting

If you prefer not to use Docker, you can install vLLM directly on your system. For detailed instructions on installation and setup, please refer to the official vLLM documentation:

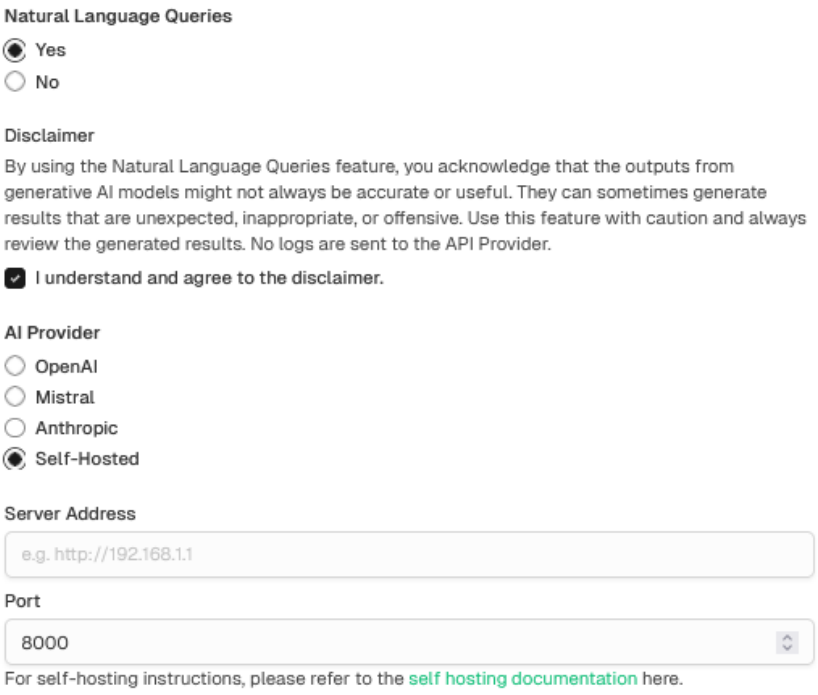
https://docs.vllm.ai/en/latest/getting_started/installation.html

This resource provides comprehensive guidance on various installation methods, including pip installation, building from source, and system-specific requirements.

Product Integrations

Nagios Log Server 2024

At this point, you can use the self-hosted model to run the AI Query Assistant. In NLS, navigate to the bottom of **Admin > Global Settings**:



Note that the self-hosted model currently doesn't work unless the model is named **llama-3-lucene-8b**. It must also reside within the `models` folder of the container, because it is invoked like this within Nagios Log Server:

```
{
"model": "/models/llama-3-lucene-8b"
}
```

For more information about usage, see: [Natural Language Queries in Nagios Log Server.]

Troubleshooting

If you get an error like:

```
AttributeError: module 'cv2.dnn' has no attribute 'DictValue'
```

You may need to run:

```
pip install opencv-fixer==0.2.5
python -c "from opencv_fixer import AutoFix; AutoFix()"
```

If you get an error like:

```
could not select device driver "" with capabilities: [[gpu]].
```

Just restart docker and that should fix the issue.

Finishing Up

This completes the documentation on Serving Nagios Enterprises' Large Language Models. If you have additional questions or other support-related questions, please visit us at our Nagios Support Forum, Nagios Knowledge Base, or Nagios Library:

Visit Nagios Support Forum Visit Nagios Knowledge Base Visit Nagios Library