SageMaker SDK:

```
import json
import sagemaker
import boto3
from sagemaker.huggingface import HuggingFaceModel,
get_huggingface_llm_image_uri

try:
        role = sagemaker.get_execution_role()
except ValueError:
        iam = boto3.client('iam')
        role = iam.get_role(RoleName='sagemaker_execution_role')['Role']['Arn']

# Hub Model configuration. https://huggingface.co/models
hub = {
        'HF_MODEL_ID':'abacusai/Smaug-72B-v0.1',
        'SM_NUM_GPUS': json.dumps(1)
}



# create Hugging Face Model Class
huggingface_model = HuggingFaceModel(
        image_uri=get_huggingface_llm_image_uri("huggingface",version="1.4.2"),
        env=hub,
        role=role,
)

# deploy model to SageMaker Inference
predictor = huggingface_model.deploy(
        initial_instance_count=1,
        instance_type="ml.g5.2xlarge",
        container_startup_health_check_timeout=300,
  )

# send request
predictor.predict({
        "inputs": "My name is Julien and I like to",
})
```

AWS Inferentia & Trainium:

```
import json
import sagemaker
import boto3
from sagemaker.huggingface import HuggingFaceModel,
get_huggingface_llm_image_uri

try:
   role = sagemaker.get_execution_role()
except ValueError:
   iam = boto3.client("iam")
   role = iam.get_role(RoleName="sagemaker_execution_role")["Role"]["Arn"]

# Hub Model configuration. https://huggingface.co/models
hub = {
   "HF_MODEL_ID": "abacusai/Smaug-72B-v0.1",
   "HF_NUM_CORES": "24",
   "HF_BATCH_SIZE": "4",
   "HF_SEQUENCE_LENGTH": "4096",
   "HF_AUTO_CAST_TYPE": "fp16",
   "MAX_BATCH_SIZE": "4",
   "MAX_INPUT_LENGTH": "3686",
   "MAX_TOTAL_TOKENS": "4096",
}


# create Hugging Face Model Class
huggingface_model = HuggingFaceModel(
   image_uri=get_huggingface_llm_image_uri("huggingface-neuronx", version="0.0.20"),
   env=hub,
   role=role,
)

# deploy model to SageMaker Inference
predictor = huggingface_model.deploy(
   initial_instance_count=1,
   instance_type="ml.inf2.48xlarge",
   container_startup_health_check_timeout=3600,
   volume_size=512,
)

# send request
```

```
predictor.predict(
    {
        "inputs": "What is is the capital of France?",
        "parameters": {
            "do_sample": True,
            "max_new_tokens": 128,
            "temperature": 0.7,
            "top_k": 50,
            "top_p": 0.95,
        }
    }
)
```

Cloudformation: