# Install multi model IGW

Reference: <https://github.com/kubernetes-sigs/gateway-api-inference-extension/blob/main/site-src/guides/serve-multiple-genai-models.md>

Make sure that you are logged in into ghcr.io

docker login ghcr.io -u <GITHUB\_USERNAME> -p <PERSONAL\_ACCESS\_TOKEN>

Update …/llm-d-inference-scheduler/deploy/components/inference-gateway/httproutes.yaml

Add the text in blue

- name: inference-gateway

rules:

- matches:

**- headers:**

**- type: Exact**

**name: X-Gateway-Model-Name**

**value: ${MODEL\_NAME\_SAFE}**

- path:

type: PathPrefix

Pull latest scheduler image

docker pull ghcr.io/llm-d/llm-d-inference-scheduler:latest

KIND\_CLUSTER\_NAME=llmd-multi-model MODEL\_NAME=model1 EPP\_TAG=latest VLLM\_SIMULATOR\_TAG=latest make env-dev-kind

KIND\_CLUSTER\_NAME=llmd-multi-model MODEL\_NAME=model2 EPP\_TAG=latest VLLM\_SIMULATOR\_TAG=latest make env-dev-kind

curl -i http://localhost:30080/v1/completions -H 'Content-Type: application/json' -H 'X-Gateway-Model-Name:model1' -d '{"model":"model1","prompt":"hi","max\_tokens":10}'