```yaml
apiVersion: apps/v1
kind: Deployment
metadata:
  name: neuro-doc-api
  labels:
    app: neuro-doc
spec:
  replicas: 2
  selector:
    matchLabels:
      app: neuro-doc
  template:
    metadata:
      labels:
        app: neuro-doc
    spec:
      containers:
      - name: api-gateway
        image: youruser/neuro-doc-api:latest
        ports:
        - containerPort: 8080
        env:
        - name: VLLM_URL
          value: "http://vllm-service:8000/v1"
        resources:
          requests:
            memory: "512Mi"
            cpu: "500m"
          limits:
            memory: "1Gi"
            cpu: "1000m"
---
# Example vLLM Service (GPU Pod)
apiVersion: apps/v1
kind: Deployment
metadata:
  name: vllm-inference
spec:
  replicas: 1
  selector:
    matchLabels:
      app: vllm
  template:
    metadata:
      labels:
        app: vllm
    spec:
      containers:
```

```
      - name: vllm-engine
        image: vllm/vllm-openai:latest
        command: ["python3", "-m",
"vllm.entrypoints.openai.api_server"]
        args: ["--model", "mistralai/Mistral-7B-v0.1"]
        resources:
          limits:
            nvidia.com/gpu: 1
```