apiVersion: apps/v1
kind: Deployment
metadata:
name: cogvlm-deployment
spec:
replicas: 1
selector:
matchLabels:
app: cogvlm
template:
metadata:
labels:
app: cogvlm
spec:
containers:
- name: cogvlm
image: cogvlm-image
resources:
requests:
memory: "20Gi"
limits:
memory: "25Gi"