```
envs:
  MODEL_NAME: cogvlm-chat-17b
  HF_HUB_ENABLE_HF_TRANSFER: True

# Fields below describe each replica.
resources:
  accelerators: [L4, A10g, A100, A100, A100-80GB, T4, M60] ## Small models
  # cpus: 32+
  memory: 32+
  use_spot: True
  disk_size: 512  # Ensure model checkpoints (~246GB) can fit.
  # disk_tier: best
  ports: 8080  # Expose to internet traffic.

setup: |
  git clone https://github.com/kyegomez/swarms-cloud && \
  cd swarms-cloud/servers/cogvlm && \
  # sudo apt-get update && \
  # sudo apt-get install -y nvidia-container-runtime && \
  # sudo systemctl restart docker && \
  sudo docker build -t cogvlm_api:latest .

run: |
  docker run --gpus all --rm -e NVIDIA_VISIBLE_DEVICES=all -p 8000:8000 cogvlm_api:latest

# service.yaml
```

```
service:
  readiness_probe:
    path: /v1/chat/completions
    post_data:
      model: $MODEL_NAME
      messages:
        - role: user
          content: Hello! What is your name?
      max_tokens: 1
  readiness_probe: /v1/models
  replica_policy:
    min_replicas: 1
    max_replicas: 10
    target_qps_per_replica: 2.5
    upscale_delay_seconds: 300
    downscale_delay_seconds: 1200
```