```
envs:
  MODEL_NAME: meta-llama/Meta-Llama-3-70B-Instruct
  OPENAI_API_KEY:
# Service configuration
service:
  readiness_probe:
    path: /v1/agent/completions  # Path for the readiness probe
  readiness_probe: /v1/health  # Additional readiness probe

  # Replica Policy
  replica_policy:
    min_replicas: 1  # Minimum number of replicas
    max_replicas: 10  # Maximum number of replicas
    target_qps_per_replica: 2.5  # Target queries per second per replica
    upscale_delay_seconds: 200  # Delay before upscaling replicas
    downscale_delay_seconds: 1200  # Delay before downscaling replicas

resources:
  # accelerators: {L4:8, A10g:8, A10:8, A100:4, A100:8, A100-80GB:2, A100-80GB:4, A100-80GB:8}
  accelerators: {A10g, A10, L40, A40} # We can use cheaper accelerators for 8B model.
  # cpus: 32+
  use_spot: True
  disk_size: 100  # Ensure model checkpoints can fit.
  # disk_tier: best
  ports: 8081  # Expose to internet traffic.
```

```
setup: |
  git clone https://github.com/kyegomez/swarms-cloud.git
  cd swarms-cloud
  pip3 install -r requirements.txt



run: |
  python3 api.py
```