```yaml
# This YAML file defines the configuration for the standard sky serve.

# Environment variables
envs:
  HF_HUB_ENABLE_HF_TRANSFER: True  # Enable HF transfer
  PYTORCH_CUDA_ALLOC_CONF: max_split_size_mb:50

# Resource configuration
resources:
  accelerators: [L4:4, A10g:4, A100, A100, A100-80GB, T4, M60, V100]  # List of accelerators for small models
  # cpus: 32+  # Uncomment and specify the number of CPUs required
  memory: 32+  # Minimum memory required
  use_spot: True  # Use spot instances
  disk_size: 512+  # Ensure model checkpoints (~246GB) can fit
  disk_tier: best  # Use the best disk tier
  ports: 8080  # Expose to internet traffic

# Service configuration
service:
  readiness_probe:
    path: /v1/chat/completions  # Path for the readiness probe
    post_data:
      model: $MODEL_NAME  # Specify the model name
      messages:
        - role: user
```

```
          content: Hello! What is your name?  # Specify the initial message
      max_tokens: 1  # Maximum number of tokens
  readiness_probe: /v1/models  # Additional readiness probe
  readiness_probe: /v1/health  # Additional readiness probe

  # Replica Policy
  replica_policy:
    min_replicas: 0  # Minimum number of replicas
    max_replicas: 30  # 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

# Setup commands
setup: |
  pip install hf_transfer  # Install hf_transfer package

# Run command
run: |
  # Run the command
```