Nebius B.V.,
Gustav Mahlerlaan 308,
1082 ME Amsterdam

Nebius AI CloudOps Certification Exam Guide

Role description

Nebius Certified AI CloudOps Engineer is a professional who has the skills to keep GPU clusters running day-to-day. They are able to deploy standard environments from templates, manage user access, monitor dashboards, restart failed nodes, and escalate issues they cannot resolve. They follow runbooks and do not make architecture decisions. This role is essential to daily AI infrastructure operations.

Target candidates

This certification is intended for DevOps engineers, cloud engineers, or system administrators with at least 1 year of hands-on experience.

Domains & weighting

The exam includes the following content areas and their weightings. This guide doesn’t cover every topic, but it provides helpful context to prepare for the exam.

Domain	Exam weight
1. Security, compliance and billing	~20%
2. Setting up and operating GPU clusters	~35%
3. Running training and inference workloads	~20%
4. Platform automation and maintenance	~25%

Exam format

The assessment is delivered remotely and monitored with AI-assisted proctoring. It is delivered via a partner exam platform.

All questions in the exam follow the same format:
Multiple choice – one correct answer and three incorrect options (distractors)

There are no strict prerequisites. No pre-tests or additional eligibility checks are required.

Duration is limited to 1 hour.

Content outline

1. Security, compliance and billing

Skills in:

- Authenticating and operating the platform
- Managing identity and access
- Protecting data through encryption and secret management

Expertise in:

- Governing cloud spend
- Applying shared-responsibility principles

2. Setting up and operating GPU clusters

Skills in:

- Provisioning GPU compute instances
- Managing storage and disks across their lifecycle
- Configuring network connectivity

Expertise in:

- Deploying Kubernetes clusters, including GPU driver enablement
- Validating cluster health and GPU readiness
- Operating the high-performance InfiniBand fabric that links GPU nodes

3. Running training and inference workloads

Skills in:

- Submitting, tracking, and troubleshooting distributed jobs

- Connecting compute jobs to persistent and shared storage
- Deploying and operating ready-made inference apps

Expertise in:

- Selecting the orchestration model for training or inference workloads
- Operating managed Slurm environments while live workloads are running

4. Platform automation and maintenance

Skills in:

- Operating infrastructure as code
- Managing planned and unplanned maintenance
- Diagnosing network and storage performance issues

Expertise in:

- Detecting and responding to GPU and interconnect health events
- Running the observability stack to investigate platform behavior