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

AI Infrastructure Engineer Certification Exam Guide

Role description

Nebius Certified AI Infrastructure Engineer is a professional who has the skills to build and maintain GPU infrastructure. The engineer takes an architect's design and makes it real – configures networking, validates performance, and troubleshoots complex failures. When something breaks beyond a restart, this is the person who fixes it. The infrastructure engineer is accountable for the cluster being healthy and performant.

Target candidates

This certification is intended for DevOps engineers, cloud operations and cloud engineers, or system administrators with at least 2 years 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. AI training platform engineering	~35%
3. AI inference platform engineering	~20%
4. Workflow automation and platform reliability	~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:

- Managing identity and access across users, service accounts, and federation
- Protecting data through encryption and secret management
- Operating billing, quotas, and cost controls

Expertise in:

- Designing role-based access and key-rotation strategies for multi-team environments
- Applying compliance, data-residency, and shared-responsibility principles

2. AI training platform engineering

Skills in:

- Provisioning and managing training clusters and their lifecycle
- Building reproducible training containers and boot images
- Implementing checkpoint and restart strategies

Expertise in:

- Architecting and validating the GPU cluster fabric for distributed training
- Architecting storage tiers for datasets and checkpoints
- Wiring the distributed-parallelism stack into the training platform
- Administering Slurm at platform level

3. AI inference platform engineering

Skills in:

- Configuring GPU partitioning for inference (MIG, MPS, time-slicing)

- Building inference platform pipelines and integrating third-party orchestration
- Deploying and tuning marketplace apps for ML pipelines

Expertise in:

- Architecting inference deployment patterns and selecting the serving framework
- Tuning for inference performance against service-level objectives
- Wiring the distributed-parallelism stack into the training platform

4. Workflow automation and platform reliability

Skills in:

- Authoring the observability platform
- Engineering multi-cluster networking
- Designing cross-cloud data architecture and integration

Expertise in:

- Engineering reliability for long training runs and high-availability inference
- Diagnosing distributed-training and fabric incidents
- Planning and executing GPU-generation rollouts