

Nebius Agentic AI Builder Certification Exam Guide

Role description

A Nebius Certified Agentic AI Builder is able to create AI-powered workflows and applications by integrating AI models via Nebius Token Factory. They build end-to-end agentic workflows and RAG pipelines, fine-tune models, and connect agents to external tools such as Tavily web search.

They also improve quality by evaluation and prompt tuning while monitoring usage, latency, and cost.

Agentic AI Builders are responsible for the application layer but not responsible for managing the underlying infrastructure.

Target candidates

Software engineers with knowledge of at least one programming language, primarily Python. 1+ year of experience is recommended, but not required.

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. Running inference on Token Factory	~20%
2. Building Agentic AI	~35%
3. Customizing a model with your own data	~20%
4. Production AI Operations	~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. Running inference on Token Factory

Skills in

- Connecting an application to Token Factory
- Running chat completions and handling streamed responses
- Prototyping prompts in the Playground

Expertise in

- Selecting the right model for cost, latency, and quality
- Tuning generation parameters for determinism versus creativity
- Choosing batch processing for high-volume, latency-tolerant workloads

2. Building Agentic AI

Skills in

- Producing structured, schema-constrained model outputs
- Building an agent loop with short- and long-term memory
- Connecting agents to web search (Tavily) and external tools via MCP

Expertise in

- Designing agent reasoning, tool use, and error recovery
- Evaluating agent output quality
- Hardening agents with guardrails, approvals, and cost budgets

3. Customizing a model with your own data

Skills in

- Preparing and uploading training datasets
- Running fine-tuning jobs in Token Factory Data Lab
- Deploying and calling fine-tuned models

Expertise in

- Choosing between prompting, RAG, and fine-tuning for a scenario
- Interpreting training results and judging fine-tune quality
- Selecting a fine-tuning strategy such as LoRA, full SFT, or distillation

4. Production AI Operations

Skills in

- Managing rate limits with client-side backoff
- Issuing and rotating API keys and safeguarding secrets
- Applying data retention and deletion controls

Expertise in

- Choosing and scaling inference capacity for the traffic shape
- Optimizing cost across model choice, prompt size, and caching
- Meeting data residency and compliance requirements across Token Factory surfaces

This exam guide is provided for informational purposes only and is subject to change without notice. The exam content, including but not limited to exam structure, format, domains, weightings, topics, pass scores, and target candidate description, may be updated, modified, or removed at any time. This guide should be used as a reference for preparation and not as a definitive or exhaustive list of exam content.

© 2026 Nebius B.V.