schwarz digits

Service certificate

STACKIT AI Model Serving

Version and start of validity	
Version 1.3	Valid from 2025/09/18

Service certificate | STACKIT AI Model Serving

Service name

STACKIT AI Model Serving

High level service description

STACKIT AI Model Serving ("**AI Model Serving**") provides open-source Large-Language-Models ("LLM") and other GenAI-Models as shared instances. Customers can use shared instances via an OpenAI-compatible REST API. Chat and embedding models are provided. An API key is used for authentication. When using the AI Model Serving Service, STACKIT does not collect or evaluate any customer data other than billing-relevant data.

Key features

- State-of-the-art open-source LLMs
- Chat & embedding-models
- GDPR-compliant service
- Usage-based billing according to tokens used
- OpenAI-compatible interface
- Easy to use via API Key

Service plans

Each model provided is assigned to a service plan. The service plans are assigned to the categories Base, Plus or Premium according to ascending model size. The assignment is described in the STACKIT portal and in the STACKIT documentation.

Metrics

Billing for AI Model Serving is token-based based on the type of model:

- For chat models, according to the number of tokens used (both the input tokens [sum of the tokens in the request] and the output tokens [sum of the tokens generated by the LLM]) of a service plan, whereby each model is assigned to a service plan. Information on estimating the number of tokens in a request can be found in the respective model descriptions (model cards) within the STACKIT documentation. The price stated in the general STACKIT price list applies per up to 1 million tokens used.

- For embedding models, only input tokens are charged. The price stated in the general STACKIT price list applies per up to 1 million tokens used.
- The respective model type is shown in the STACKIT documentation and in the STACKIT Cloud Portal. The customer determines which model type is used as part of the API selection to their application.

SLA specifics

- In deviation from the availability specifications in the general STACKIT Service Description, an availability of 99.5% per calendar month is agreed (measured by the external availability of the LLM API).

Backup

- Customer requests are not backed up.

Additional terms

- When using the model selected by the customer, the customer undertakes to comply with the license conditions applicable to the respective model, which can be viewed in the STACKIT documentation.
- Model deprecation process
  In addition to the general STACKIT Cloud Terms of Use and the general STACKIT service description, models can be terminated by STACKIT with a notice period of 6 months. If a deprecated model version is followed by the release of a direct successor model, deprecated model versions can be discontinued by STACKIT with a lead time of 3 months and replaced by the successor model.
- STACKIT additionally points out that the customer must comply with the relevant legal terms for AI applications created by the customer.
- The following additional terms apply to the use of STACKIT AI Model Serving: https://www.stackit.de/en/general-terms-and-conditions/service-certificates/stackit-compute-engine-gpu/.

Annex | Exportability

(Online Register)

Data type	Description	Exportable (Yes/No)	Format	Additional notes
Customer data (database content)	Data stored by the customer in the database (if available) or within the product/service	No	-	Customer data is not stored permanently. The e-mail and subject ID are stored temporarily (30 days) in the logs of the service and can be found there.
User accounts & permissions	Information about users and their permissions	Yes	JSON	The customer can manage their own authentication tokens (project-wide resource) via the product API. There they can create new tokens themselves, list them (json), adjust and delete them.
System metrics (instances / resources in use)	Performance data of the instance / resource in use (e.g., CPU usage, memory usage)	No	-	We have a multi-client system in which no system metrics are collected specifically for individual customers. The customer requests are proceeded in shared resources.
	Sizes and capacities *Capacities of the available resources / instances*	No. Company confidential.	-	-
System properties (instances / resources in use)	Versions and information necessary to check compatibility	Yes	Text / JSON	The shared models offered can be viewed via API (https://docs.api.eu01.stackit.cloud/documentation/model-serving/version/v1#tag/Models) and in the documentation, including the link to the source (Huggingface).
Product / service-related data (product properties)	Configuration data and source code *Configuration of IT-systems / rudimental IT, settings, customizing, IP's, VLAN, interfaces, software code, scripts*	No (except ratelimits)	Header text for ratelimits	Individual rate limits (at project level) can be viewed as headers in each model response. Everything else (code and configuration) is not visible.

Log data (non personalized and personalized) *System-status, technical-events, etc*.	No	-	-
Log data (non personalized and personalized) *Login/logout of user, user activities*	Yes	JSON	Any changes to the authentication tokens in the project are pushed to the STACKIT Audit Log.