schwarz digits

Leistungsschein

STACKIT AI Model Serving

Version und Geltungsbeginn	
Version 1.3	Gültig ab 18.09.2025

Leistungsschein | STACKIT AI Model Serving

Servicename

STACKIT AI Model Serving

Kurzbeschreibung

Der STACKIT AI Model Serving Service ("AI Model Serving") stellt open-source Large-Language-Models (LLMs) und andere GenAI-Modelle als geteilte Instanzen bereit. Kunden können geteilte Instanzen über eine OpenAI-kompatible REST-API nutzen. Es werden u.a. Chat- und Embeddings-Modelle bereitgestellt. Zur Authentifizierung wird ein API-Schlüssel genutzt. Bei der Nutzung des AI Model Serving Services werden seitens STACKIT außer abrechnungsrelevanten Daten keinerlei Daten des Kunden erhoben oder ausgewertet.

Wesentliche Merkmale

- State-of-the-art open-source LLMs
- Chat- & Embeddings-Modelle
- DSGVO-konformer Service
- Nutzungsbasierte Abrechnung nach verbrauchten Tokens
- OpenAI-kompatible Schnittstelle
- Einfache Nutzung via API-Key

Servicepläne

Jedes bereitgestellte Modell wird einem Serviceplan zugeordnet. Die Servicepläne werden nach aufsteigender Modelgröße in die Kategorien Base, Plus oder Premium eingeordnet. Die Zuordnung wird im STACKIT Portal sowie in der STACKIT Dokumentation beschrieben.

Metriken

Die Abrechnung des AI Model Serving erfolgt Token-basiert anhand des Typs des Modells:

- Bei Chat-Modellen nach Anzahl der genutzten Tokens (sowohl der Input-Tokens [Summe der Tokens in der Anfrage] sowie der Output-Tokens [Summe der vom LLM generierten Tokens]) eines Serviceplans, wobei jedes Modell einem Serviceplan zugeordnet ist. Informationen zur Schätzung der Token-Anzahl einer Anfrage können über die jeweiligen Modell-Beschreibungen (Model Cards) innerhalb der STACKIT

Dokumentation gefunden werden. Der über die allgemeine STACKIT Preisliste angegebene Preis gilt pro jeweils bis zu 1 Millionen genutzten Token.

- Bei Embedding-Modellen werden ausschließlich Input-token berechnet. Der über die allgemeine STACKIT Preisliste angegebene Preis gilt pro jeweils bis zu 1 Millionen genutzten Token.
- Der jeweilige Modell-Typ wird in der STACKIT Dokumentation ausgewiesen. Der Kunde legt im Rahmen des API-Aufrufs seiner Anwendung fest, welcher Modell-Typ zum Einsatz kommt.

SLA-Spezifika

- Abweichend von den Verfügbarkeitsangeben der allgemeinen STACKIT Servicebeschreibung wird für das AI Model Serving eine Verfügbarkeit von 99,5% im Kalendermonat (gemessen anhand der externen Erreichbarkeit der LLM-API) vereinbart.

Backup

- Ein Backup der Anfragen des Kunden erfolgt nicht.

Zusätzliche Bedingungen

- Der Kunde verpflichtet sich bei Nutzung des jeweils von ihm ausgewählten Modells, die jeweils für das Modell geltenden Lizenzbedingungen einzuhalten, welche über die STACKIT Dokumentation einsehbar sind.
- Modell Deprecation Prozess

  In Ergänzung der allgemeinen Bestimmungen der Nutzungsbedingungen und der allgemeinen STACKIT Servicebeschreibung können Modelle mit einer Ankündigungsfrist von 6 Monaten durch STACKIT abgekündigt werden. Erfolgt auf eine veraltete Modellversionen der Release eines direkten Nachfolgermodells, können veraltete Modellversionen durch STACKIT mit einer Vorlaufzeit von 3 Monaten abgekündigt und durch das Nachfolgermodell ersetzt werden.
- STACKIT weist zusätzlich darauf hin, dass der Kunde etwaig einschlägige gesetzliche Bestimmungen für vom Kunden erstellte KI-Anwendungen einzuhalten hat.
- Für die Nutzung des STACKIT AI Model Serving gelten zusätzlich die nachfolgenden Bedingungen:
  https://www.stackit.de/de/agb/leistungsscheine/stackit-compute-engine-gpu/

Anhang | Exportierbarkeit (Online Register)

Datentyp	Beschreibung	Exportierbar (Ja/Nein)	Format	Zusätzliche Anmerkungen
Kundendaten (Datenbank-inhalte)	Daten, die vom Kunden in der Datenbank (sofern vorhanden) bzw. innerhalb des Produktes/Services gespeichert werden	Nein	-	Kundendaten werden nicht dauerhaft gespeichert. Die E-Mail und Subject ID werden temporär (30 Tage) in den Logs des Services gespeichert und sind dort zu finden.
Benutzer-konten & Berechtigungen	Informationen über Nutzer und deren Berechtigungen	Ja	JSON	Der Kunde kann seine Authentifizierungstokens (Projektweite Ressource) selbst verwalten via Produkt-API. Dort kann er selbst neue Tokens erstellen, diese auflisten (JSON), anpassen und löschen.
System-Metriken (Instanzen/ Ressourcen in Nutzung)	Leistungsdaten der Instanz/ genutzten Ressource (z. B. CPU-Auslastung, Speichernutzung)	Nein	-	Wir haben ein Multimandanten-System, bei dem keine Systemmetriken spezifisch für einzelne Kunden erhoben werden, da die Kundenanfragen auf geteilten Ressourcen laufen.
	Größen und Kapazitäten *Kapazitäten der vorhandenen Ressourcen / Instanzen*	Nein. Betriebs-internum STACKIT.	-	-
System-eigenschaften (Instanzen/ Ressourcen in Nutzung)	Versionen und Informationen, die notwendig sind, um Kompatibilität prüfen zu können	Ja	Text/JSON	Die angebotenen gemeinsam genutzten Modelle sind via API (https://docs.api.eu01.stackit.cloud/documentation/model-serving/version/v1#tag/Models) und in der Dokumentation einsehbar, inklusive dem Link zur Quelle (Hugging Face).

Produkt / Service-bezogene Daten (Produkt-eigenschaften)	Konfigurationsdaten und Source Code *Configuration of IT-Systems/ rudimental IT, Settings, Customizing, IP's, VLAN, Interfaces, Software Code, Scripts*	Nein (Ausnahme: Ratelimits)	Header text für Ratelimits	Individuelle Ratelimits (auf Projekt-Ebene) sind einsehbar als Header in jeder Antwort eines Modells. Alles Weitere (Code und Konfiguration) ist nicht einsehbar.
	Log Daten (nicht personalisiert und personalisiert) *System-Status, Technical-events, etc*.	Nein. Betriebs-internum STACKIT.	-	-
	Log Daten (nicht personalisiert und personalisiert) *Login/Logout der Nutzer, Nutzeraktivitäten*	Ja	JSON	Alle Änderungen an den Authentifizierungstokens im Projekt werden in das STACKIT Audit Log aufgenommen.