



LLM Serving Service

Service Specifications & Additional Terms and Conditions

from T-Systems International GmbH

Version: 1.24

Date: 30. 9. 2026

<This document is a convenience translation of the German original. In case of discrepancy between the English and German versions, the German version shall prevail.>

T Systems

Let's power
higher performance

Publication details

Published by

T-Systems International GmbH

Friedrich-Ebert-Allee 140

53113 Bonn

Germany

WEEE Reg. No.: DE50335567

hereinafter referred to as – “Telekom” –

Information required by law can be found at: <https://www.t-systems.com/de/en/compulsory-statement> |

Copyright © 2024 All rights reserved, including those of partial reproduction, electronic or photomechanical reproduction, and evaluation by electronic data processing methods.

Internal PIMS-ID: 34011758/1096

Confidentially Class: Public

Contents

1	Introduction	4
2	Functions.....	4
2.1	LLM self-service portal	4
2.2	API.....	4
2.3	Available LLMs and embedding models	4
2.3.1	LLMs and embedding models with data processing within the EU/EEA and countries with adequacy decisions.....	5
2.3.2	LLMs and embedding models with data processing outside the EU/EEA and countries that lack adequacy decisions.....	6
2.3.3	Test Models	7
2.4	Chat Frontend (T-Systems LLM Playground)	8
3	Services provided by Telekom	8
3.1	Initial provision.....	8
3.2	Operation	9
3.2.1	Customer support	9
3.2.2	Service transfer point.....	9
3.2.3	Service quality.....	9
3.2.4	Maintenance work	11
3.2.5	Reporting	11
3.3	Terms of use, licensing provisions	11
3.3.1	Compliance with AI regulations	11
3.3.2	General terms of use	12
3.3.3	Licenses, terms of use, and data privacy policies	12
3.3.4	Medical Device Regulation (MDR).....	14
3.3.5	High-risk AI systems, prohibited use cases.....	14
3.3.6	Liability, indemnity	14
3.3.7	Feedback.....	14
3.4	Optional services	14
3.4.1	LLMs on a dedicated instance	15
3.4.2	Fine-tuning of models.....	15
3.4.3	Installation within the customer's own environment.....	15
3.4.4	Additional LLMs and embedding models	15
3.4.5	Additional support services	15

4	Service amendments by Telekom	15
5	The customer's duties to cooperate.....	16
5.1	General duties to cooperate	16
5.2	Duties to cooperate during operation	17
6	Minimum term/termination	17
7	Charges, commercial terms and conditions	18
7.1	Method for calculating charges	18
7.2	Prices	18
7.2.1	Essential rate plan	19
7.2.2	Professional rate plan	20
7.2.3	Agentic rate plan	22
7.2.4	Charges for optional services	24
8	Glossary/List of abbreviations.....	24

1 Introduction

With the LLM Serving Service, Telekom provides customers with access to various large language models (LLM) and embedding models as a service for self-service usage. Access to the LLMs and embedding models is provided through an API using an API key. This gives customers the ability to use LLMs through a chat frontend.

2 Functions

2.1 LLM self-service portal

The customer can access the self-service portal online at <https://apikey.llmhub.t-systems.net> and will be provided with the following self-service functions:

- a. Overview of available API keys and the models associated with them
- b. Generating and managing API keys to access the models
- c. Overview of token usage by key and model

Telekom provides the customer administrator with login data for the LLM self-service portal. The customer administrator can create additional users for the portal, allowing them to independently generate one or more API keys using the self-service feature. The generated API keys allow access to the models included in the agreed rate plan.

2.2 API

The customer can manage the selected LLMs and embedding models using a standard API. This can be done via HTTPS request or using the OpenAI Python library. Telekom provides the customer with up-to-date documentation at <https://docs.llmhub.t-systems.net/category/model-serving>.

2.3 Available LLMs and embedding models

The LLM Serving Service is subject to continual development. The available LLMs and embedding models are therefore subject to change over time; models can be added or removed by Telekom.

The LLM Serving Service provides two groups of LLMs and embedding models:

- LLMs and embedding models where data processing takes place only within the EU/EEA and in countries with an adequacy decision on data protection from the EU Commission
- LLMs and embedding models where data processing also takes place outside the EU/EEA and/or in countries that lack an adequacy decision on data protection from the EU Commission

Telekom expressly states that due to legal provisions (such as those relating to the processing of social data and/or health data), it is prohibited to process certain categories of data outside the EU/EEA and in countries that lack an adequacy decision on data protection from the EU Commission. In this analysis, subcontractors of every level throughout the processing chain must also be included in the legal assessment (e.g., § 80 (2) of the German Social Security Code (Sozialgesetzbuch – SGB) X, § 393 (2) SGB V).

Telekom further emphasizes that the legal assessment of the permissibility of using any of the LLMs and embedding models described in these Service Specifications is solely the customer's responsibility. Telekom cannot and will not verify whether the customer's use of the selected LLMs and embedding models is permissible.

For this reason, the customer assures Telekom that the permissibility of using the selected LLMs and embedding models has been verified and that they assume full responsibility for their use. Claims arising from unlawful use of LLMs and embedding models must be borne solely by the customer and do not entitle the customer to liability claims, recourse claims, or contractual adjustment rights against Telekom. The customer fully indemnifies Telekom against all claims by affected parties and third parties (especially regulatory authorities) – upon first request – and will reimburse Telekom for expenses incurred for necessary legal defense, to the extent that they arise from unlawful use of the service. Further claims for damages arising from illegal use remain unaffected.

2.3.1 LLMs and embedding models with data processing within the EU/EEA and countries with adequacy decisions

The LLMs and embedding models in this category are operated by Telekom. The processing of data and the provision of services takes place within the EU/EEA and/or in countries with an adequacy decision on data protection from the EU Commission.

By default, your requests and the responses received are not permanently stored in the LLM Serving Service, nor are they used to train the language models. To make the service respond faster, there is a so-called caching function: already processed requests are temporarily stored for a limited time so that similar or recurring requests can be answered more quickly. This function is enabled by default but can be switched on or off by customers at any time.

Further information on the cache settings: <https://docs.llmhub.t-systems.net/guides/prefix-caching/>

The following models are currently available:

Model	Model type	Platform
Gemma 4 31B	LLM	T-Cloud Public
Mistral AI Mistral Small 4	LLM	T-Cloud Public
Qwen3.6 35B	LLM	T-Cloud Public
Qwen 3.8 27B	LLM	T-Cloud Public
NVIDIA Nemotron 3 Super	LLM	T-Cloud Public

Model	Model type	Platform
GLM 5.2	LLM	T-Cloud Public / Industrial AI Cloud
GLM 5.3 Flash	LLM	T-Cloud Public / Industrial AI Cloud
Jina AI jina-embeddings-v2-base-de	Embedding	T-Cloud Public
BAAI text-embedding-bge-m3	Embedding	T-Cloud Public
OpenAI GPT-OSS-120B	LLM	T-Cloud Public
OpenAI Whisper Large v3	STT	T-Cloud Public
OpenAI Whisper Large v3 Turbo	STT	T-Cloud Public

2.3.2 LLMs and embedding models with data processing outside the EU/EEA and countries that lack adequacy decisions

The LLMs and embedding models in this category are not operated by Telekom. The processing of data and the provision of services can take place outside the EU/EEA and/or in countries that lack an adequacy decision on data protection from the EU Commission.

Requests for LLMs and embedding models in this category are routed through Telekom's LLM Serving interface directly to the respective third-party provider, who operates them on their own infrastructure.

By selecting or using these models, the customer accepts the terms of use and licensing provisions (in particular the data privacy policies) of the respective third-party provider.

The third-party provider may engage subcontractors worldwide, resulting in the global processing of customer data.

The following models of this category are currently available:

Model	Model type	Platform
OpenAI GPT-5.6 Sol Azure Sweden	LLM	Microsoft Azure
OpenAI GPT-5.6 Terra Azure Sweden	LLM	Microsoft Azure
OpenAI GPT-5.6 Luna Azure Sweden	LLM	Microsoft Azure
OpenAI GPT-6.1 Sol Azure Sweden	LLM	Microsoft Azure
OpenAI GPT-6 Luna Azure Sweden	LLM	Microsoft Azure

Model	Model type	Platform
OpenAI GPT-Image-2	LLM	Microsoft Azure
OpenAI Ada-Text, Azure France	Embedding	Microsoft Azure
Mistral AI Mistral Medium 3	LLM	Google Cloud Platform
Mistral AI Mistral Medium 3.5	LLM	Google Cloud Platform & Microsoft Azure
Anthropic Claude Opus 5	LLM	Google Cloud Platform & Microsoft Azure
Anthropic Claude Opus 5.5	LLM	Google Cloud Platform & Microsoft Azure
Anthropic Claude Sonnet 5	LLM	Google Cloud Platform & Microsoft Azure
Anthropic Claude Sonnet 5.5	LLM	Google Cloud Platform & Microsoft Azure
Anthropic Claude Fable 5	LLM	Google Cloud Platform & Microsoft Azure
Anthropic Claude Fable 5.1	LLM	Google Cloud Platform & Microsoft Azure
Google Gemini 3 Flash	LLM	Google Cloud Platform
Google Gemini 3.1 Pro	LLM	Google Cloud Platform
Google Gemini 3.5 Flash	LLM	Google Cloud Platform
Google Gemini 3.8 Flash	LLM	Google Cloud Platform
Google Gemini 3 Pro Image	LLM	Google Cloud Platform

2.3.3 Test Models

Telekom provides the customer with additional LLMs and embedding models as test models free of charge. The test models are provided as LLMs and embedding models with data processing within the EU/EEA and countries with an adequacy decision pursuant to 2.3.1.

The use of the test models is free of charge but requires the conclusion of a tariff/plan.

Telekom is entitled to limit the use of the test models at any time, in particular with regard to available tokens, rate limits (TPM/RPM), and duration of use.

Telekom is entitled to deactivate the test models or restrict the customer's access to them at any time and without prior notice. The customer has no legal entitlement to the continued provision of the test models.

The following test models are currently available:

<https://docs.llmhub.t-systems.net/plans/test>

No minimum availability applies to test models. However, Telekom endeavors to avoid performance limitations.

Für die aufgeführten Testmodelle gelten ergänzend folgende Bestimmungen:

- a. Since this is a test deployment, the services are not suitable for use in production/live operation. Telekom advises against such use in production operation, as well as against the use of real data. If the customer nevertheless uses the services in production operation or with real data, this is done at their own responsibility.
- b. Telekom is entitled at any time and without giving reasons to discontinue or modify the services during the test deployment, or to restrict the customer's access to the services. The customer has no legal entitlement to the use of the services provided free of charge.
- c. In all other respects, Telekom warrants the services of the demonstration and test operation with the characteristics specified in these terms of use. Since this is a demonstration and test operation, limited services may have to be expected. Telekom endeavors to keep these limitations as minimal as possible and to remedy performance disruptions within the scope of its technical and operational capabilities, as well as to continuously optimize the demonstration and test operation..

2.4 Chat Frontend (T-Systems LLM Playground)

With the API key generated in the LLM self-service portal, the customer's user gains access to the T-Systems LLM Playground's chat frontend to interact with the LLMs as per the agreed rate plan. The tool does not save or persist any chat data; all interactions are processed in real time. The T-Systems LLM Playground is accessible on the internet via a secure HTTPS connection at the URL <https://public.oweb-chat.llmhub.t-systems.net>.

3 Services provided by Telekom

3.1 Initial provision

Telekom shall set up the LLM Serving Service as commissioned and send the access data and URL for the LLM self-service portal plus the relevant email addresses for the contacts to the customer via email.

The initial provision is considered completed upon sending of the email or, at the latest, when the customer starts using the service.

3.2 Operation

3.2.1 Customer support

Telekom's Customer Support shall be the main point of contact for the customer's named administrators as the 2nd level support and shall provide the following services:

- a. Processing incidents
- b. Processing service requests

Customer Support can be reached during the attended operating hours (Monday to Friday from 9:00 a.m. to 5:00 p.m. CET/CEST, excluding nationwide public holidays in Germany) via the email address provided to the customer during the initial provision process.

Telekom responds to customer requests according to the following process:

- a. Telekom opens a ticket, sends a confirmation of receipt of the request, along with the ticket number, by email.
- b. Telekom analyzes the customer's request and contacts the customer to clarify matters if necessary.
- c. Telekom informs the customer by email regarding the status and completion of the processing.

The following applies to the processing of tickets:

- a. Tickets are processed exclusively during attended operating hours.
- b. Promised resolution times do not apply to the processing of incidents.
- c. Telekom is entitled to resolve incidents by providing a workaround.

3.2.2 Service transfer point

Telekom's responsibility ends at the service transfer point. The service transfer point is the API endpoint for the LLM Serving Service in Telekom's data center.

3.2.3 Service quality

a. Minimum availability

The minimum availability of the LLM Serving Service at the service transfer point for rate plans with a minimum revenue is 99.5 percent during the attended operating hours within a calendar month and is calculated as follows:

$$\left[\left(\frac{\text{Total service minutes} - \text{total downtime minutes}}{\text{Total service minutes}} \right) * 100 \right]$$

Minimum availability is reported as a percentage. Where:

Total service minutes – the total number of minutes during the attended operating hours within a calendar month.

Total downtime minutes – the number of minutes during the attended operating hours within a calendar month in which the LLM is not available minus the excused events in minutes.

Otherwise, there is no minimum availability. However, Telekom strives to avoid service restrictions.

b. Excused events (excluded events/suspend times)

Service interruptions based on one of the following events do not count as downtime minutes:

- i. Downtimes that are caused by maintenance work
- ii. Incidents, downtimes, and problems that are attributable to the customer, their employees, or other third parties contracted by the customer
- iii. Downtimes that can be traced back to third-party action or force majeure
- iv. Incidents, downtimes, and problems that are attributable to the services of third-party providers, particularly in connection with the provision of LLMs operated outside of the EU

c. Service credits

In the event of non-compliance with the minimum availability, Telekom shall reimburse the customer with the service credits specified below under the following conditions:

- i. Telekom is exclusively responsible for non-compliance with the agreed availability in the respective billing period.
- ii. The customer shall claim the service credits by emailing the address provided during the initial provision process within a period of three months from the billing period.
- iii. The amount of the credit for the respective billing period is at least EUR 1.00.
- iv. Telekom shall review the customer's claim and, if approved, it shall notify the customer of the reimbursement amount.
- v. Telekom shall apply the service credits to the monthly bill issued after confirmation. Alternatively, Telekom will transfer the credits to the customer's designated bank account after service termination.

The amount of the service credits shall be calculated as follows depending on the minimum revenue actually paid by the customer for the respective billing period:

Availability in the respective billing period	Service credit amount for the affected service
<99.5 % - 99.3 %	5 % of the agreed minimum revenue
<99.3 % - 95 %	10 % of the agreed minimum revenue
<95 %	25 % of the agreed minimum revenue

Granted service credits shall be offset against any claims for damages or other claims arising from the failure to meet agreed availability levels. This provision shall be final.

3.2.4 Maintenance work

Telekom regularly carries out maintenance work. Telekom shall notify the customer in advance if this maintenance work will result in service disruptions. Telekom strives to keep disruptions caused by maintenance work to a minimum. Maintenance work is not considered downtime and is therefore not taken into account when calculating availability.

3.2.5 Reporting

Upon request, Telekom shall provide the customer with monthly reports with the following content under the rates with minimum revenue:

Availability (overview)	Compliance with minimum availability
Availability (details)	Comprises the monthly details on availability of the API endpoint and a 6-month history to provide information on downtime frequencies and times. The service times and downtimes of the current month are also provided in graphic and tabular form.
Downtime (in minutes)	Log of all downtimes during the attended operating hours for the current month.
Key	Explanations regarding the individual report components
Change history	Version description

Upon request, the report will be emailed to the customer five working days after the end of a calendar month.

3.3 Terms of use, licensing provisions

3.3.1 Compliance with AI regulations

Telekom undertakes to provide the services rendered to the customer under this contract, insofar as these fall within the scope of Regulation (EU) 2024/1689 on artificial intelligence (AI Regulation), in accordance with the obligations of the AI Regulation applicable to Telekom. No further obligations or guarantees are assumed. Any service/product-specific agreements shall take precedence.

The customer undertakes to use the IT services provided under this contract in accordance with the AI Regulation.

3.3.2 General terms of use

With T-Systems SmartChat, Telekom provides the customer with an AI services that enables access to a variety of LLMs for use at the customer's own responsibility.

The results generated by the AI service are neither controlled nor influenced by Telekom. Telekom is therefore, to the extent permitted by law, not responsible for the results generated by the AI service and their use by the customer, particularly with regard to freedom from third-party rights.

By using the AI service, the customer agrees to the following conditions:

- a. When using the service and the generated results, the customer ensures that they comply with all legal regulations applicable to them, in particular the provisions of the EU AI Act, data protection regulations (e.g., eligibility requirements such as minimum age), as well as any other industry-specific conditions or internal policies.
- b. AI may produce content that is faulty, incomplete, offensive, or protected by third-party rights. The customer undertakes to ensure the monitoring of AI use and to inform their users about the use of AI; in particular, the customer will
 - have the generated results assessed by a human for correctness and suitability for the respective use case before using or forwarding the generated results.
 - not rely on the generated results or make decisions solely based on the generated results, especially when seeking medical, legal, financial, or other professional advice.
 - not use generated results relating to a specific person for a purpose that could have legal or any other material impact on this person, e.g., making decisions on credit, education, employment, or insurance-related matters, as well as legal, medical, or any other important decisions.

3.3.3 Licenses, terms of use, and data privacy policies

By using the respective LLMs and embedding models, the customer accepts the following terms of use, including the data privacy policies and licensing provisions, for the respective model as amended, thereby establishing an agreement between the customer and the respective provider:

Model	Platform	Terms of use	Data privacy provisions
OpenAI Ada Text OpenAI GPT-Image-2 OpenAI GPT-5.6 Sol OpenAI GPT-5.6 Terra OpenAI GPT-5.6 Luna OpenAI GPT-6.1 Sol OpenAI GPT-6 Luna Anthropic Claude Sonnet 5 Anthropic Claude Sonnet 5.5 Anthropic Claude Fable 5 Anthropic Claude Opus 5 Anthropic Claude Opus 5.5	Microsoft Azure	https://www.microsoft.com/licensing/terms/productoffering/MicrosoftAzure/MCA#ServiceSpecificTerms	https://www.microsoft.com/licensing/docs/view/Microsoft-Products-and-Services-Data-Protection-Addendum-DPA

Model	Platform	Terms of use	Data privacy provisions
Anthropic Claude Fable 5.1 Mistral AI Mistral Medium 3.5			
Anthropic Claude Sonnet 5 Anthropic Claude Sonnet 5.5 Anthropic Claude Fable 5 Anthropic Claude Opus 5 Anthropic Claude Opus 5.5 Anthropic Claude Fable 5.1 Mistral AI Mistral Medium Google Gemini 3 Flash Google Gemini 3.1 Pro Google Gemini 3.5 Flash Google Gemini 3.8 Flash Google Gemini 3 Pro Image	Google Cloud Platform	https://cloud.google.com/terms/services https://cloud.google.com/terms/aup	https://cloud.google.com/terms/data-processing/terms https://cloud.google.com/terms/cloud-privacy-notice
Jina AI jina-embeddings-v2-base-de OpenAI GPT-OSS-120B OpenAI Whisper Large v3 Gemma 4 31B Mistral AI Mistral Small 4 Qwen3.6 35B Qwen 3.8 27B	T-Cloud Public, Open Source	https://www.apache.org/licenses/LICENSE-2.0.html	
OpenAI Whisper Large v3 Turbo	T-Cloud Public, Open Source	https://choosealicense.com/licenses/mit/	
BAAI text-embedding-bge-m3	T-Cloud Public, Open Source	https://opensource.org/licenses/mit	
GLM 5.2 GLM 5.3 Flash	T-Cloud Public, Industrial AI Cloud, Open Source	https://opensource.org/licenses/mit	
NVIDIA Nemotron 3 Super	T-Cloud Public, Open Source	https://www.nvidia.com/en-us/agreements/enterprise-software/nvidia-open-model-license/	

3.3.4 Medical Device Regulation (MDR)

The LLM serving service is not a medical device within the meaning of Article 2 (1) MDR. The customer agrees not to use the LLM Serving Service as a medical device and, in particular, not for the following purposes:

- Diagnosis, prevention, monitoring, prediction, prognosis, treatment, or alleviation of diseases
- Diagnosis, monitoring, treatment, alleviation, or compensation of injuries or disabilities
- Examination, replacement, or modification of the anatomy or a physiological or pathological process or condition
- Information retrieval through in vitro examination of samples originating from the human body – including from organ, blood, and tissue donations

3.3.5 High-risk AI systems, prohibited use cases

The services provided by Telekom are not a high-risk AI system within the meaning of Annex III of the EU AI Act. The customer is prohibited from using Telekom services in any manner that would qualify them as a high-risk AI system or from using them for purposes prohibited under Article 5 of the AI Act.

3.3.6 Liability, indemnity

Telekom shall not be liable, to the extent permitted by law, for any damage arising from non-compliance with these terms of use and licensing provisions.

In addition, the customer shall fully indemnify Telekom in this regard from all third-party claims, particularly those from regulatory authorities, upon first request. Further claims of Telekom remain unaffected.

3.3.7 Feedback

The customer shall grant Telekom and its affiliated companies an unrestricted right to use the customer's feedback and ideas regarding the described services.

3.4 Optional services

The optional services below are provided for an additional charge and must be ordered separately. Upon request to the Telekom sales employee responsible, Telekom shall provide the customer with a proposal and detailed descriptions of the services after a successful review.

3.4.1 LLMs on a dedicated instance

Telekom shall provide individual LLMs and embeddings on a dedicated instance. The customer will therefore have exclusive access to the respective language model, without sharing it with other customers.

3.4.2 Fine-tuning of models

The LLM Serving Service of T-Systems SmartChat can be expanded with functionalities for fine-tuning LLMs. Upon request, Telekom will provide the functionality along with documentation and support the creation and integration of the fine-tuned model into the overall solution.

3.4.3 Installation within the customer's own environment

Upon request, Telekom shall review whether it is possible to provide the service within the customer's own environment. If Telekom's review indicates that the customer's environment meets the necessary requirements, Telekom shall provide the services as part of a customer-specific solution within the customer's own environment and, if commissioned accordingly, Telekom shall manage the operation of the customized solution.

3.4.4 Additional LLMs and embedding models

Telekom shall provide the customer with LLMs and embedding models other than those currently available on a customer-specific or dedicated instance. Telekom shall provide the customer with a corresponding proposal if the results of the technical feasibility review are positive.

3.4.5 Additional support services

Telekom shall provide individual customer support with the following services on a time and materials basis:

- Implementation support
- Consulting/coaching
- Training

4 Service amendments by Telekom

If Telekom intends to amend the legal terms and conditions of these Service Specifications, the services, or prices, the customer shall be informed of these amendments in text form (e.g., by letter or email) at least one month before they become effective. The amendments shall

become constituent parts of the agreement subject to the following conditions set out below in Items a) to b):

- a. Changes in favor of the customer at the time they take effect.
- b. Price increases, changes to the legal conditions, and significant changes to the services to the disadvantage of the customer at the time they take effect. This will not apply if the customer terminates the service in writing without observing a notice period at that time. The customer's right of termination will be expressly referred to in the notification about the amendments.

5 The customer's duties to cooperate

The customer is obliged to cooperate to ensure proper provisioning of the required services by Telekom; in particular, the customer must meet the following obligations free of charge, on time, and to the required extent:

5.1 General duties to cooperate

- a. The customer is obliged to name a qualified contact person with decision-making powers and to ensure that this person is available/represented at all times.
- b. The customer shall name up to five administrators who are adequately trained and authorized to submit support requests to Telekom.
- c. The customer declares that they agree to exchange unencrypted information by email and will always provide a current email address. The customer is aware that information essential for the provision of services, such as access data, notification of changes to the services and legal conditions, as well as invoices, will be sent exclusively by email.
- d. The customer is responsible for checking whether the data transferred by them to Telekom in connection with the use of the service is personal data and whether processing this personal data is permissible. If the customer wishes personal data to be processed, the customer will sign an agreement on the processing of personal data based on the Telekom sample agreement, which Telekom will provide to the customer upon request.
- e. The customer is responsible for verifying and ensuring compliance with all relevant legal provisions, laws, regulations, and industry-specific provisions applicable in connection with use of the service. In particular, this also includes compliance with non-disclosure obligations, for example those resulting from a professional activity, as well as compliance with the terms of use and the licensing provisions.
- f. The customer shall ensure that the services are not misused.
- g. The customer shall provide Telekom with all the necessary information and will ensure that the information is correct and up to date.
- h. The customer shall be obliged to protect the equipment and services of Telekom from unauthorized access, malware, and other adverse effects with suitable measures, handle them with due care, and comply with the manufacturer's specifications.
- i. The customer is obligated to keep passwords and access data secret, in particular API keys generated by the customer, to disclose them to authorized third parties only, to protect them from access by third parties, and to change them when necessary. The customer shall inform Telekom without delay of any suspected disclosure to non-authorized third parties.
- j. Support

The customer is obliged to provide Telekom with support in eliminating any incident/problem.

In particular, the customer shall provide first-level support (user help desk) for their users and carry out self-checks as part of this in order to exclude the possibility that the cause of the incident lies within their own sphere of responsibility. The customer shall resolve any incidents within their own sphere of responsibility on their own.

If the customer is unable to resolve the incidents on their own and Telekom is responsible for said incidents, the customer shall get in contact with Telekom's support team with a comprehensible outline of the symptoms of the error, stating any other relevant information.

The customer is obliged to notify Telekom immediately of any incidents, disruptions to services, or damage to Telekom equipment.

- k. The customer shall ensure that sufficient funds are available in the agreed debit account and, if making payments by credit card, that the credit card details provided during registration are kept up to date.
- l. To access the API endpoint via the OpenAI Python library, the customer shall provide the necessary software and ensure that it is kept up to date.
- m. The customer shall ensure that they have the necessary usage rights (including updates or upgrades).
- n. The customer shall ensure that their users comply with the provisions of these Service Specifications, particularly with regard to cooperation obligations, as well as terms of use and licensing provisions.
- o. The customer shall inform Telekom in writing without delay if it proves impossible for them to perform one of the agreed duties to cooperate or if circumstances arise that will make it difficult or impossible for Telekom to provide the services.

5.2 Duties to cooperate during operation

- a. The customer assures that they will not store any data, content, or other entries on the contractual storage space, use it, or make it available in any other way, if it contains malicious code or other malware and/or if the provision, publication, transfer, or use of such content violates applicable laws or third-party rights.
- b. The customer shall ensure that their use of the service does not cause any risk or disruption to third parties or to Telekom's equipment and services. In the event of any such risk or disruption (e.g., due to a DDoS attack), Telekom shall be entitled to deactivate the service concerned, without prior notice, until the cause of the risk or disruption has been eliminated. Any resulting downtime shall not be considered when calculating availability. Telekom shall keep the customer informed.
- a. The customer shall inform Telekom without delay of any suspected, security-related events concerning the equipment and services of Telekom.

6 Minimum term/termination

The service can be terminated with one month's notice to the end of the month. All terminations have to be sent by email to the email address provided during the initial provision process, stating the number of the agreement.

Upon termination of the agreement on LLM Serving Services, all of the access options are deactivated and the customer's data is erased.

This shall be without prejudice to each party's right to terminate the agreement for good cause.

7 Charges, commercial terms and conditions

7.1 Method for calculating charges

a) Token consumption

Model usage shall be billed based on the number of tokens used. Token consumption can be viewed in the LLM self-service portal. Billing is based on each million of input/output tokens, or part thereof, for the respective model.

b) Rate plans

Telekom shall provide the LLM Serving Service in various rate plans. These rate plans differ in the following aspects:

- Available models
- Minimum revenue/no minimum revenue
- Max. Input TPM (Token per Minute): maximum number of processed input tokens per minute
- Max. Output TPM (Token per Minute): maximum number of processed output tokens per minute
- Max. RPM (requests per minute): maximum number of requests per minute that can be sent to the LLM

For rate plans with minimum usage requirements, Telekom charges the respective minimum rate even if the customer's usage is below this threshold.

If a rate plan is provided/terminated for less than a full month, the minimum revenue is calculated on a pro-rata basis using calendar days. Rate changes are possible by mutual agreement upon request.

Telekom emphasizes that the values Max Input TPM, Max Output TPM, and Max RPM represent the maximum possible rate limits. The actual expected throughputs will be lower.

c) Invoicing

Telekom shall send a monthly invoice to the customer for the previous calendar month.

d) Taxes and duties

All prices are quoted in euros and are exclusive of the taxes and duties at the rate applicable at the time of delivery and provisioning.

7.2 Prices

For all models hosted on T-Cloud Public / Industrial AI Cloud, the price for cached input tokens is 15% of the respective input token price.

7.2.1 Essential rate plan

The monthly minimum revenue for models on the Essential rate plan is: € 1,000

Platform	Model	Model type	Max. input TPM	Max. output TPM	Max. RPM	Charge per million input tokens	Charge per million output tokens
T-Cloud Public	Jina AI jina-embeddings-v2-base-de	Em-bedding	300.000	150.000	300	0,05€	0,05€
T-Cloud Public	BAAI text-embedding-bge-m3	Em-bedding	300.000	150.000	300	0,05€	0,05€
T-Cloud Public	OpenAI Whisper Large v3	STT	218		300	14,61 €	14,61 €
T-Cloud Public	OpenAI Whisper Large v3 Turbo	STT	345		300	9,21 €	9,21 €
T-Cloud Public	OpenAI GPT-OSS-120B	LLM	400.000	400.000	100	0,20€	0,65€
T-Cloud Public	Gemma 4 31B Mistral AI Mistral Small 4 Qwen3.6 35B NVIDIA Nemotron 3 Super	LLM	50.000	50.000	50	0,60€	1,20€
T-Cloud Public	Qwen 3.8 27B	LLM	50.000	50.000	50	0,50€	2,50€
T-Cloud Public / Industrial AI Cloud	GLM 5.2	LLM	400.000	400.000	100	1,50€	3,50€
T-Cloud Public / Industrial AI Cloud	GLM 5.3 Flash	LLM	400.000	400.000	100	0,20€	0,60€
Microsoft Azure	OpenAI GPT-5.6 Sol Azure Sweden	LLM	250.000	125.000	250	4,50€	27,00€
Microsoft Azure	OpenAI GPT-5.6 Terra Azure Sweden	LLM	750.000	375.000	750	3,52€	15,82€
Microsoft Azure	OpenAI GPT-5.6 Luna Azure Sweden	LLM	750.000	375.000	750	0,20€	1,20€
Microsoft Azure	OpenAI GPT-6.1 Sol Azure Sweden <=272k Input-Tokens	LLM	250.000	125.000	250	2,14€	10,68€
Microsoft Azure	OpenAI GPT-6.1 Sol Azure Sweden >272k Input-Tokens	LLM	250.000	125.000	250	4,28€	16,02€
Microsoft Azure	OpenAI GPT-6 Luna Azure Sweden <=272k Input-Tokens	LLM	250.000	125.000	250	0,11€	0,54€
Microsoft Azure	OpenAI GPT-6 Luna Azure Sweden >272k Input-Tokens	LLM	250.000	125.000	250	0,22€	0,81€
Microsoft Azure	OpenAI GPT-Image-2 (Input Text, Output Image)	LLM			1	4,39€	26,33€
Microsoft Azure	OpenAI GPT-Image-2 (Input Image, Output Image)	LLM			1	7,03€	26,33€
Microsoft Azure	OpenAI Ada-Text, Azure France	Em-bedding	3.000.000	1.500.000	3.000	0,11€	0,11€
Google Cloud Platform	Mistral AI Mistral Medium 3	LLM	120.000	60.000	90	0,36 €	1,80 €
Google Cloud Platform & Microsoft Azure	Mistral AI Mistral Medium 3.5	LLM	120.000	60.000	90	1,29 €	6,44 €

Platform	Model	Model type	Max. input TPM	Max. output TPM	Max. RPM	Charge per million input tokens	Charge per million output tokens
Google Cloud Platform & Microsoft Azure	Anthropic Claude Sonnet 5	LLM	125.000	62.500	125	2,97€	14,85€
Google Cloud Platform & Microsoft Azure	Anthropic Claude Sonnet 5.5	LLM	5.625.000	562.500	562	2,97€	14,85€
Google Cloud Platform	Google Gemini 3 Flash	LLM	DSQ	DSQ	DSQ	0,45€	2,70€
Google Cloud Platform	Google Gemini 3.1 Pro <=200k Input-Tokens	LLM	DSQ	DSQ	DSQ	1,80 €	10,80 €
Google Cloud Platform	Google Gemini 3.1 Pro >200k Input-Tokens	LLM	DSQ	DSQ	DSQ	3,60 €	16,20 €
Google Cloud Platform	Google Gemini 3.5 Flash	LLM	DSQ	DSQ	DSQ	1,49 €	8,91 €
Google Cloud Platform	Google Gemini 3.8 Flash	LLM	DSQ	DSQ	DSQ	0,74 €	3,68 €
Google Cloud Platform	Google Gemini 3 Pro Image	LLM	DSQ	DSQ	DSQ	1,80 €	108,00 €

7.2.2 Professional rate plan

The monthly minimum revenue for models on the Professional rate plan is: € 3,000

Platform	Model	Model type	Max. input TPM	Max. output TPM	Max. RPM	Charge per million input tokens	Charge per million output tokens
T-Cloud Public	Jina AI jina-embeddings-v2-base-de	Em-bedding	600.000	300.000	600	0,05€	0,05€
T-Cloud Public	BAAI text-embedding-bge-m3	Em-bedding	600.000	300.000	600	0,05€	0,05€
T-Cloud Public	OpenAI Whisper Large v3	STT	218		300	14,61 €	14,61 €
T-Cloud Public	OpenAI Whisper Large v3 Turbo	STT	345		300	9,21 €	9,21 €
T-Cloud Public	OpenAI GPT-OSS-120B	LLM	1.000.000	1.000.000	200	0,20€	0,65€
T-Cloud Public	Gemma 4 31B Mistral AI Mistral Small 4 Qwen3.6 35B NVIDIA Nemotron 3 Super	LLM	100.000	100.000	100	0,60€	1,20€
T-Cloud Public	Qwen 3.8 27B	LLM	100.000	100.000	100	0,50€	2,50€
T-Cloud Public / Industrial AI Cloud	GLM 5.2	LLM	1.000.000	1.000.000	200	1,50€	3,50€

Platform	Model	Model type	Max. input TPM	Max. output TPM	Max. RPM	Charge per million input tokens	Charge per million output tokens
T-Cloud Public / Industrial AI Cloud	GLM 5.3 Flash	LLM	1.000.000	1.000.000	200	0,20€	0,60€
Microsoft Azure	OpenAI GPT-5.6 Sol Azure Sweden	LLM	500.000	250.000	500	4,50€	27,00€
Microsoft Azure	OpenAI GPT-5.6 Terra Azure Sweden	LLM	1.500.000	750.000	1.500	3,52€	15,82€
Microsoft Azure	OpenAI GPT-5.6 Luna Azure Sweden	LLM	1.500.000	750.000	1.500	0,20€	1,20€
Microsoft Azure	OpenAI GPT-6.1 Sol Azure Sweden <=272k Input-Tokens	LLM	500.000	250.000	500	2,14€	10,68€
Microsoft Azure	OpenAI GPT-6.1 Sol Azure Sweden >272k Input-Tokens	LLM	500.000	250.000	500	4,28€	16,02€
Microsoft Azure	OpenAI GPT-6 Luna Azure Sweden <=272k Input-Tokens	LLM	500.000	250.000	500	0,11€	0,54€
Microsoft Azure	OpenAI GPT-6 Luna Azure Sweden >272k Input-Tokens	LLM	500.000	250.000	500	0,22€	0,81€
Microsoft Azure	OpenAI GPT-Image-2 (Input Text, Output Image)	LLM			3	4,39€	26,33€
Microsoft Azure	OpenAI GPT-Image-2 (Input Image, Output Image)	LLM			3	7,03€	26,33€
Microsoft Azure	OpenAI Ada-Text, Azure France	Em-bedding	6.000.000	3.000.000	6.000	0,11€	0,11€
Google Cloud Platform	Mistral AI Mistral Medium 3	LLM	240.000	120.000	180	0,36 €	1,80 €
Google Cloud Platform & Microsoft Azure	Mistral AI Mistral Medium 3.5	LLM	240.000	120.000	180	1,29 €	6,44 €
Google Cloud Platform & Microsoft Azure	Anthropic Claude Opus 5	LLM	250.000	125.000	250	4,95€	24,75€
Google Cloud Platform & Microsoft Azure	Anthropic Claude Opus 5.5	LLM	12.000.000	1.200.000	1.200	3,92€	19,58€
Google Cloud Platform & Microsoft Azure	Anthropic Claude Sonnet 5	LLM	250.000	125.000	250	2,97€	14,85€
Google Cloud Platform & Microsoft Azure	Anthropic Claude Sonnet 5.5	LLM	11.250.000	1.125.000	1.125	2,97€	14,85€
Google Cloud Platform & Microsoft Azure	Anthropic Claude Fable 5	LLM	250.000	125.000	250	9,90€	49,50€
Google Cloud Platform & Microsoft Azure	Anthropic Claude Fable 5.1	LLM	250.000	125.000	250	9,90€	49,50€
Google Cloud Platform	Google Gemini 3 Flash	LLM	DSQ	DSQ	DSQ	0,45€	2,70€

Platform	Model	Model type	Max. input TPM	Max. output TPM	Max. RPM	Charge per million input tokens	Charge per million output tokens
Google Cloud Platform	Google Gemini 3.1 Pro ≤200k Input-Tokens	LLM	DSQ	DSQ	DSQ	1,80 €	10,80 €
Google Cloud Platform	Google Gemini 3.1 Pro >200k Input-Tokens	LLM	DSQ	DSQ	DSQ	3,60 €	16,20 €
Google Cloud Platform	Google Gemini 3.5 Flash	LLM	DSQ	DSQ	DSQ	1,49 €	8,91 €
Google Cloud Platform	Google Gemini 3.8 Flash	LLM	DSQ	DSQ	DSQ	0,74 €	3,68 €
Google Cloud Platform	Google Gemini 3 Pro Image	LLM	DSQ	DSQ	DSQ	1,80 €	108,00 €

7.2.3 Agentic rate plan

The monthly minimum revenue for models on the Agentic rate plan is: € 5,000

Platform	Model	Model type	Max. input TPM	Max. output TPM	Max. RPM	Charge per million input tokens	Charge per million output tokens
T-Cloud Public	Jina AI jina-embeddings-v2-base-de	Em-bedding	2.000.000	1.000.000	1.000	0,05€	0,05€
T-Cloud Public	BAAI text-embedding-bge-m3	Em-bedding	2.000.000	1.000.000	1.000	0,05€	0,05€
T-Cloud Public	OpenAI Whisper Large v3	STT	218		300	14,61 €	14,61 €
T-Cloud Public	OpenAI Whisper Large v3 Turbo	STT	345		300	9,21 €	9,21 €
T-Cloud Public	OpenAI GPT-OSS-120B	LLM	2.000.000	2.000.000	300	0,20€	0,65€
T-Cloud Public	Gemma 4 31B Mistral AI Mistral Small 4 Qwen3.6 35B NVIDIA Nemotron 3 Super	LLM	300.000	300.000	150	0,60€	1,20€
T-Cloud Public	Qwen 3.8 27B	LLM	300.000	300.000	150	0,50€	2,50€
T-Cloud Public / Industrial AI Cloud	GLM 5.2	LLM	2.000.000	2.000.000	300	1,50€	3,50€
T-Cloud Public / Industrial AI Cloud	GLM 5.3 Flash	LLM	2.000.000	2.000.000	300	0,20€	0,60€
Microsoft Azure	OpenAI GPT-5.6 Sol Azure Sweden	LLM	1.000.000	500.000	1.000	4,50€	27,00€
Microsoft Azure	OpenAI GPT-5.6 Terra Azure Sweden	LLM	2.500.000	1.250.000	2.500	3,52€	15,82€
Microsoft Azure	OpenAI GPT-5.6 Luna Azure Sweden	LLM	2.500.000	1.250.000	2.500	0,20€	1,20€

Platform	Model	Model type	Max. input TPM	Max. output TPM	Max. RPM	Charge per million input tokens	Charge per million output tokens
Microsoft Azure	OpenAI GPT-6.1 Sol Azure Sweden <=272k Input-Tokens	LLM	1.000.000	500.000	1.000	2,14€	10,68€
Microsoft Azure	OpenAI GPT-6.1 Sol Azure Sweden >272k Input-Tokens	LLM	1.000.000	500.000	1.000	4,28€	16,02€
Microsoft Azure	OpenAI GPT-6 Luna Azure Sweden <=272k Input-Tokens	LLM	1.000.000	500.000	1.000	0,11€	0,54€
Microsoft Azure	OpenAI GPT-6 Luna Azure Sweden >272k Input-Tokens	LLM	1.000.000	500.000	1.000	0,22€	0,81€
Microsoft Azure	OpenAI GPT-Image-2 (Input Text, Output Image)	LLM			5	4,39€	26,33€
Microsoft Azure	OpenAI GPT-Image-2 (Input Image, Output Image)	LLM			5	7,03€	26,33€
Microsoft Azure	OpenAI Ada-Text, Azure France	Em-bedding	10.000.000	5.000.000	10.000	0,11€	0,11€
Google Cloud Platform	Mistral AI Mistral Medium 3	LLM	400.000	200.000	300	0,36 €	1,80 €
Google Cloud Platform & Microsoft Azure	Mistral AI Mistral Medium 3.5	LLM	400.000	200.000	300	1,29 €	6,44 €
Google Cloud Platform & Microsoft Azure	Anthropic Claude Opus 5	LLM	500.000	250.000	500	4,95€	24,75€
Google Cloud Platform & Microsoft Azure	Anthropic Claude Opus 5.5	LLM	20.000.000	2.000.000	2.000	3,92€	19,58€
Google Cloud Platform & Microsoft Azure	Anthropic Claude Sonnet 5	LLM	500.000	250.000	500	2,97€	14,85€
Google Cloud Platform & Microsoft Azure	Anthropic Claude Sonnet 5.5	LLM	18.750.000	1.875.000	1.875	2,97€	14,85€
Google Cloud Platform & Microsoft Azure	Anthropic Claude Fable 5	LLM	500.000	250.000	500	9,90€	49,50€
Google Cloud Platform & Microsoft Azure	Anthropic Claude Fable 5.1	LLM	500.000	250.000	500	9,90€	49,50€
Google Cloud Platform	Google Gemini 3 Flash	LLM	DSQ	DSQ	DSQ	0,45€	2,70€
Google Cloud Platform	Google Gemini 3.1 Pro <=200k Input-Tokens	LLM	DSQ	DSQ	DSQ	1,80 €	10,80 €
Google Cloud Platform	Google Gemini 3.1 Pro >200k Input-Tokens	LLM	DSQ	DSQ	DSQ	3,60 €	16,20 €
Google Cloud Platform	Google Gemini 3.5 Flash	LLM	DSQ	DSQ	DSQ	1,49 €	8,91 €
Google Cloud Platform	Google Gemini 3.8 Flash	LLM	DSQ	DSQ	DSQ	0,74 €	3,68 €

Platform	Model	Model type	Max. input TPM	Max. output TPM	Max. RPM	Charge per million input tokens	Charge per million output tokens
Google Cloud Platform	Google Gemini 3 Pro Image	LLM	DSQ	DSQ	DSQ	1,80 €	108,00 €

7.2.4 Charges for optional services

The daily rate for optional services such as implementation support, consulting, and coaching is € 1,144.45.

The services are provided remotely. Support services are billed monthly based on a performance record signed by the customer on the basis of time and materials.

8 Glossary/List of abbreviations

Abbreviation	Explanation
API key	An API key is a unique authentication key that enables access to an application programming interface (API) by identifying, authenticating, and authorizing the calling application.
ASR	Automatic Speech Recognition
Attended operating hours	From Monday to Friday, from 9:00 a.m. to 5:00 p.m. CET/CEST, excluding nationwide public holidays in Germany
DSQ	Dynamic Shared Quota
LLM	Large language models
Models	Models in this document include both large language models and embedding models
PaaS	Platform as a Service
SaaS	Software as a Service
Token	If not specified, tokens include both input and output tokens. Tokens are text sequences, e.g., sentences, words, or characters, that can be processed by an LLM. The size of a token can differ depending on the LLM.
TPM	The maximum number of tokens that can be processed per minute
RPM	The maximum number of requests that can be processed per minute