LLM-Serving-Service

Leistungsbeschreibung & zusätzliche Bedingungen

von T-Systems International GmbH

Version: 1.10b

Date: 30. 4. 2025

T Systems

Let's power higher performance

Impressum

Herausgeber

T-Systems International GmbH

Hahnstraße 43d

60528 Frankfurt am Main

Deutschland

WEEE-Reg.-Nr. DE50335567

nachfolgend – Telekom - genannt

Die gesetzlichen Pflichtangaben finden Sie unter: www.t-systems.de/pflichtangaben |

Internal PIMS-ID: 34011758/1096

Confidentially Class: Public

Inhaltsverzeichnis

1 Einleitung

Mit dem LLM-Serving-Service stellt die Telekom dem Kunden einen Zugang zu verschiedenen Large Language Modellen (LLM) und Embedding-Modellen as-a-Service zur Nutzung im Self-Service bereit. Der Zugang zu den LLMs und Embedding-Modellen wird über eine API mittels eines API-Keys bereitgestellt. Mit diesem hat der Kunde die Möglichkeit, die LLMs über ein Chat-Frontend zu nutzen.

2 Funktionen

2.1 LLM-Self-Service Portal

Das Self-Service Portal ist durch den Kunden über das Internet unter https://apikey.llmhub.t-systems.net erreichbar und bietet diesem folgende Funktionen zur Nutzung im Self-Service:

a. Übersicht der verfügbaren API-Keys und der mit den Keys verbundenen Modellen.
b. Generierung und Verwaltung von API-Keys für den Zugang zu den Modellen.
c. Übersicht des Token-Verbrauchs je Key und Modell.

Die Telekom stellt dem Kunden-Administrator Log-In-Daten für das LLM-Self-Service Portal bereit. Der Kunden-Administrator kann für das Portal weitere Nutzer anlegen, die sich dort im Self-Service einen oder mehrere API-Keys generieren können. Die generierten API-Keys ermöglichen den Zugriff auf die Modelle des vereinbarten Tarifs.

2.2 API

Der Kunde kann die ausgewählten LLMs und Embedding-Modelle über eine standardisierte API ansteuern. Dies ist via HTTPS-Request oder mit Hilfe des Python Pakets „openai“ möglich. Die Telekom stellt dem Kunden eine aktuelle Dokumentation unter https://docs.llmhub.t-systems.net/category/model-serving zur Verfügung.

2.3 Verfügbare LLMs und Embedding-Modelle

Der LLM-Serving-Service unterliegt einer ständigen Weiterentwicklung. Die verfügbaren LLMs und Embedding-Modelle können sich daher von Zeit zu Zeit ändern, Modelle können von der Telekom hinzugefügt und entfernt werden.

Der LLM-Serving-Service stellt 2 Gruppen von LLMs und Embedding Modellen bereit:

- LLMs und Embedding-Modelle, bei denen die Datenverarbeitung nur innerhalb EU/EWR und in Staaten mit einem datenschutzrechtlichen Angemessenheitsbeschluss der EU Kommission stattfindet.

- LLMs und Embedding-Modelle, bei denen die Datenverarbeitung auch außerhalb der EU/EWR und/oder in Staaten ohne datenschutzrechtlichen Angemessenheitsbeschluss der EU Kommission stattfindet.

Die Telekom weist ausdrücklich darauf hin, dass es aufgrund gesetzlicher Bestimmungen (wie bspw. im Bereich der Verarbeitung von Sozialdaten und-/-oder Gesundheitsdaten) unzulässig ist, bestimmte Datenkategorien außerhalb der EU/EWR und in Ländern ohne datenschutzrechtlichen Angemessenheitsbeschluss der EU -Kommission zu verarbeiten. Bei dieser Betrachtung sind auch die Unterauftragnehmer jeden Grades über die Verarbeitungskette in die rechtliche Bewertung mit einzubeziehen (z.B. § 80 Abs. 2 SGB X, § 393 Abs. 2 SGB V).

Die Telekom weist ferner darauf hin, dass die rechtliche Bewertung der Zulässigkeit der Nutzung einer der in dieser Leistungsbeschreibung beschriebenen LLMs und Embedding-Modelle allein in den Verantwortungsbereich des Kunden fällt und die Telekom die Zulässigkeit der Nutzung der ausgewählten LLMs und Embedding-Modelle durch den Kunden nicht überprüfen kann und wird.

Aus diesem Grunde sichert der Kunde der Telekom zu, dass er die Zulässigkeit der Nutzung der ausgewählten LLMs und Embedding-Modelle geprüft hat und hierfür die alleinige Verantwortlichkeit trägt. Ansprüche, die aus einer rechtswidrigen Nutzung der LLMs und Embedding-Modelle resultieren, sind vom Kunden allein zu tragen und berechtigen nicht zu Haftungsansprüchen, Regressforderungen und der Ausübung vertraglicher Gestaltungsrechte gegenüber der Telekom. Der Kunde stellt die Telekom von allen Ansprüchen von Betroffenen und Dritten (insb. Aufsichtsbehörden) -auf erstes Anfordern–vollumfänglich frei und erstattet der Telekom die Kosten der notwendigen Rechtsverteidigung, soweit diese auf einer rechtswidrigen Nutzung der Leistung beruhen. Weitergehende Schadenersatzansprüche infolge gesetzeswidriger Nutzung bleiben unberührt.

2.3.1 LLMs und Embedding-Modelle mit Datenverarbeitung innerhalb EU/EWR und Staaten mit Angemessenheitsbeschluss

Die LLMs und Embedding-Modelle dieser Kategorie werden von der Telekom betrieben. Die Verarbeitung der Daten und die Leistungserbringung erfolgt innerhalb der EU/EWR und/oder in Staaten mit datenschutzrechtlichem Angemessenheitsbeschluss der EU-Kommission.

Die an das jeweilige LLM und Embedding-Modell übermittelten Daten des Kunden und die zurückgelieferten Ergebnisse werden nicht gespeichert, sind weder für die Telekom noch für Dritte einsehbar und werden nicht zum Training der Sprachmodelle genutzt.

Aktuell sind folgende Modelle verfügbar:

Modell	Modelltyp	Plattform
Meta Llama 3.3 70b-instruct AWQ	LLM	Open Telekom Cloud
Mistral AI Mistral Small 3	LLM	Open Telekom Cloud
Jina AI jina-embeddings-v2-base-de	Embedding	Open Telekom Cloud
BAAI text-embedding-bge-m3	Embedding	Open Telekom Cloud

Modell	Modelltyp	Plattform
Jina AI jina-embeddings-v2-base-code	Embedding	Open Telekom Cloud
Alibaba Qwen 2.5 VL 72B AWQ	LLM	Open Telekom Cloud
DeepSeek AI DeepSeek-Coder-V2-Lite	LLM	Open Telekom Cloud
Alibaba Qwen Coder 2.5 7B	LLM	Open Telekom Cloud
DeepSeek-R1-Distill-Llama-70B AWQ	LLM	Open Telekom Cloud
OpenGPT-X Teuken-7B-instruct-commercial-v0.4	LLM	Open Telekom Cloud
DeutschlandGPT Llama-BildungsLLM-1.0	LLM	Open Telekom Cloud

2.3.2 LLMs und Embedding-Modelle mit Datenverarbeitung außerhalb EU/EWR und Staaten ohne Angemessenheitsbeschluss

Die LLMs und Embedding-Modelle dieser Kategorie werden nicht von der Telekom betrieben. Die Verarbeitung der Daten und die Leistungserbringung kann außerhalb der EU/EWR und/oder in Staaten ohne datenschutzrechtlichem Angemessenheitsbeschluss der EU-Kommission erfolgen.

Anfragen an LLMs und Embedding Modelle dieser Kategorie werden über die LLM-Serving Schnittstelle der Telekom direkt an den jeweiligen Drittanbieter weitergeleitet, der diese auf eigenen Infrastrukturen betreibt.

Mit der Auswahl/Nutzung dieser Modelle akzeptiert der Kunde die Nutzungsbedingungen und Lizenzbestimmungen, insbesondere die Datenschutzbestimmungen, des jeweiligen Drittanbieters.

Der Drittanbieter kann Subunternehmer weltweit einsetzen, so dass die Daten des Kunden weltweit verarbeitet werden.

Aktuell sind folgende Modelle dieser Kategorie verfügbar:

Modell	Modelltyp	Plattform
OpenAI GPT-3.5-Turbo-0125 (16k), Azure France (depricated!)	LLM	Microsoft Azure
OpenAI GPT-4o-2024-11-20, Azure France	LLM	Microsoft Azure
OpenAI GPT-4o-mini-2024-07-18, Azure Sweden	LLM	Microsoft Azure
OpenAI o1 2024-12-17, Azure Sweden	LLM	Microsoft Azure
OpenAI o1 mini 2024-09-12, Azure Sweden	LLM	Microsoft Azure
OpenAI o3 mini 2025-01-31-Azure Sweden	LLM	Microsoft Azure

Modell	Modelltyp	Plattform
OpenAI Ada-Text, Azure France	Em-bedding	Microsoft Azure
Mistral AI Mistral Large 2411	LLM	Google Cloud Platform
Anthropic Claude 3.5 Sonnet V2	LLM	Google Cloud Platform
Google Gemini 2.0 Flash	LLM	Google Cloud Platform
Anthropic Claude 3.7 Sonnet	LLM	Google Cloud Platform

2.4 Chat-Frontend (T-Systems LLM Playground)

Mit dem im LLM-Self-Service-Portal generierten API-Key erhält der jeweilige Nutzer des Kunden Zugang zum Chat-Frontend T-Systems LLM Playground für den Dialog mit den LLMs des vereinbarten Tarifs. Das Tool speichert oder persistiert keine Chat-Daten, alle Interaktionen werden in Echtzeit verarbeitet. Der T-Systems LLM Playground ist unter der URL https://public.oweb-chat.llmhub.t-systems.net über eine https-gesicherte Verbindung über das Internet erreichbar.

3 Leistungen der Telekom

3.1 Bereitstellung

Die Telekom richtet den LLM-Serving-Service gemäß Beauftragung ein und versendet die Zugangsdaten nebst URL für das LLM-Self-Service-Portal sowie den relevanten Kontakt-E-Mail-Adressen per E-Mail an den Kunden.

Die Bereitstellung ist mit Versendung der E-Mail, spätestens mit Nutzungsbeginn durch den Kunden abgeschlossen.

3.2 Betrieb

3.2.1 Kundensupport

Der Kundensupport der Telekom ist der zentrale Ansprechpartner für die benannten Administratoren des Kunden als 2nd-Level-Support und erbringt folgende Leistungen:

a. Bearbeitung von Incidents
b. Beantwortung von Service-Requests

Der Kundensupport ist unter der dem Kunden im Rahmen der Bereitstellung mitgeteilten E-Mail-Adresse erreichbar während der betreuten Betriebszeit (Montag bis Freitag von 9:00 Uhr bis 17.00 Uhr MEZ/MESZ, ausgenommen an bundeseinheitlichen Feiertagen in Deutschland) erreichbar.

Anfragen des Kunden bearbeitet die Telekom nach folgendem Prozess:

a. Die Telekom eröffnet ein Ticket, versendet per E-Mail eine Eingangsbestätigung der Anfrage nebst Ticketnummer.
b. Die Telekom analysiert die Anfrage des Kunden und setzt sich soweit erforderlich zwecks Klärung von Fragen mit diesem in Verbindung.
c. Die Telekom informiert den Kunden per E-Mail über den Status und den Abschluss der Bearbeitung.

Für die Bearbeitung von Tickets gilt:

a. Tickets werden ausschließlich innerhalb der betreuten Betriebszeit bearbeitet.
b. Für die Bearbeitung von Incidents gelten keine zugesagten Lösungszeiten.
c. Die Telekom ist berechtigt, Incidents durch die Bereitstellung eines Workarounds zu beheben.

3.2.2 Leistungsübergabepunkt

Die Verantwortung der Telekom endet am Leistungsübergabepunkt. Der Leistungsübergabepunkt ist der API Endpoint des LLM-Serving-Services im Rechenzentrum der Telekom.

3.2.3 Service Quality

a. Mindestverfügbarkeit

Die Mindestverfügbarkeit des LLM-Serving-Services am Leistungsübergabepunkt für Tarife mit Mindestumsatz liegt bei 99,9% innerhalb der betreuten Betriebszeit eines Kalendermonats und wird wie folgt berechnet:

$$\left[\left(\frac{\text{Gesamte Serviceminuten} - \text{Gesamte Ausfallminuten}}{\text{Gesamte Serviceminuten}}\right) * 100\right]$$

Die Mindestverfügbarkeit wird als Prozentsatz ausgewiesen. Dabei bedeutet:

Gesamte Serviceminuten – die gesamte Anzahl der Minuten während der betreuten Betriebszeit innerhalb eines Kalendermonats.

Gesamte Ausfallminuten – die Anzahl der Minuten innerhalb der betreuten Betriebszeit innerhalb eines Kalendermonats, in der der LLM-Serving-Service nicht verfügbar ist abzüglich der ausgeschlossenen Ereignisse (Excused Events) in Minuten.

Im Übrigen gilt keine Mindestverfügbarkeit. Die Telekom ist jedoch bemüht, Leistungseinschränkungen zu vermeiden.

b. Excused Events (Ausgeschlossene Ereignisse/Suspend-Zeiten)

Unterbrechungen der Leistung, die auf einem den nachfolgenden Ereignissen basieren, gelten nicht als Ausfallminuten:

i. Ausfälle, die durch Wartungsarbeiten verursacht werden;
ii. Störungen, Ausfälle und Probleme die auf den Kunden, seine Mitarbeiter oder sonst dem Kunden zuzurechnender Dritter zurückzuführen sind.
iii. Ausfälle, die auf eine Einwirkung von Dritten oder höhere Gewalt zurückzuführen sind.
iv. Störungen, Ausfälle und Probleme, die auf Leistungen von Drittanbietern, insbesondere im Zusammenhang der Bereitstellung von LLMs mit Betrieb außerhalb EU zurückzuführen sind.

c. Service Credits

Bei Nichteinhaltung der Mindestverfügbarkeit erstattet die Telekom dem Kunden nachfolgende Service Credits unter folgenden Voraussetzungen:

i. Die Telekom hat die Nichteinhaltung der vereinbarten Verfügbarkeit im jeweiligen Abrechnungszeitraum ausschließlich zu vertreten.
ii. Die Service Credits werden durch den Kunden via E-Mail an die im Rahmen der Bereitstellung zur Verfügung gestellte Emailadresse innerhalb einer Frist von drei Monaten ab dem Abrechnungszeitraum geltend gemacht.
iii. Die Höhe der Gutschrift des jeweiligen Abrechnungszeitraums beträgt mindestens 1,00 EUR.
iv. Die Telekom prüft den Anspruch des Kunden und teilt diesem im Falle eines positiven Prüfungsergebnisses die Höhe der Erstattung mit.
v. Die Telekom verrechnet die Service Credits mit der auf die Bestätigung folgenden Monatsrechnung, bzw. überweist diese nach Beendigung der Leistung an die kundenseitig benannte Bankverbindung.

Die Höhe der Service Credits wird abhängig von dem vom Kunden für den jeweiligen Abrechnungszeitraum tatsächlich gezahlten Mindestumsatzes wie folgt berechnet:

Verfügbarkeit im jeweiligen Abrechnungszeitraum	**Höhe der Servicecredits für die betroffene Leistung**
<99,9% - 99.7%	5% des vereinbarten Mindestumsatzes
<99,7 – 95%	10% des vereinbarten Mindestumsatzes
<95%	25% des vereinbarten Mindestumsatzes

Gewährte Service Credits werden auf etwaige Schadensersatzansprüche und sonstige Ansprüche auf Grund der Nichteinhaltung der Verfügbarkeit des Kunden angerechnet. Diese Regelung ist abschließend.

3.2.4 Wartungsarbeiten

Die Telekom führt regelmäßig Wartungsarbeiten durch. Sollten diese Wartungsarbeiten zu Unterbrechungen der Leistung führen, wird die Telekom den Kunden vorab informieren. Die Telekom ist hierbei bestrebt, Beeinträchtigungen durch Wartungsarbeiten möglichst gering zu halten. Wartungsarbeiten gelten nicht als Ausfallzeiten und bleiben daher bei der Berechnung der Verfügbarkeit unberücksichtigt.

3.2.5 Reporting

Die Telekom stellt dem Kunden in den Tarifen mit Mindestumsatz ein monatliches Reporting mit folgendem Inhalt bereit:

Verfügbarkeit (Übersicht)	Einhaltung der Mindestverfügbarkeit
Verfügbarkeit (Details)	Umfasst die monatlichen Details zur Verfügbarkeit des API-Endpoints, sowie eine 6-monatige Historie zur Auskunft über Ausfallhäufigkeiten und –zeiten. Des Weiteren werden die Bereitschafts- und Ausfallzeiten des aktuellen Monats sowohl in grafischer als auch tabellarischer Form dargestellt.
Ausfallzeit in Minuten	Protokoll aller Ausfallzeiten innerhalb der betreuten Betriebszeit für den aktuellen Monat.
Legende	Erklärungen zu den einzelnen Bestandteilen des Reports.
Änderungshistorie	Versionsbeschreibung

Der Report wird dem Kunden regelmäßig fünf Werktage nach Ablauf eines Kalendermonats per E-Mail zugesandt.

3.3 Nutzungsbedingungen, Lizenzbestimmungen

3.3.1 Allgemeine Nutzungsbedingungen

Die Telekom stellt dem Kunden mit dem T-Systems SmartChat einen KI-Service zur Verfügung, der den Zugriff auf verschiedene LLMs zur eigenverantwortlichen Nutzung ermöglicht.

Die durch den KI-Service generierten Ergebnisse unterliegen weder der Kontrolle noch dem Einfluss der Telekom. Die Telekom ist daher, soweit gesetzlich zulässig, nicht für die durch den KI-Service generierten Ergebnisse und deren Verwendung durch den Kunden, insbesondere im Hinblick auf die Freiheit von Rechten Dritter, verantwortlich.

Durch die Nutzung des KI-Service erklärt sich der Kunde mit folgenden Bedingungen einverstanden:

a. Der Kunde stellt bei der Nutzung der Leistung, sowie der Verwendung von generierten Ergebnissen sicher, dass er in diesem Zusammenhang aller für ihn anwendbaren rechtlichen Vorschriften, insbesondere der Vorschriften des EU AI Act, datenschutzrechtlicher Bestimmungen (z.B. Zulässigkeitsvoraussetzungen wie Mindestalter) und sonstiger branchenspezifischen Bedingungen oder interner Vorgaben einhält.
b. Der Einsatz von KI kann fehlerhafte, unvollständige, beleidigende oder durch Rechte Dritter geschützte Inhalte liefern. Der Kunde verpflichtet sich, eine Überwachung des KI-Einsatzes sicherzustellen und seine Nutzer zum KI-Einsatz zu informieren, insbesondere wird der Kunde

- generierte Ergebnisse für den jeweiligen Anwendungsfall auf Richtigkeit und Angemessenheit durch eine menschliche Überprüfung bewerten, bevor er die generierten Ergebnisse verwendet oder weitergibt.
- sich nicht auf die generierten Ergebnisse verlassen oder Entscheidungen ausschließlich basierend auf den generierten Ergebnissen treffen, insbesondere, wenn er medizinischen, rechtlichen, finanziellen oder anderen professionellen Rat sucht.
- generierte Ergebnisse, die sich auf eine Person beziehen, nicht für einen Zweck verwenden, der rechtliche oder wesentliche Auswirkungen auf diese Person haben könnte, wie z. B. das Treffen von Kredit-, Bildungs-, Beschäftigungs-, Versicherungs-, rechtlichen, medizinischen oder anderen wichtigen Entscheidungen.

3.3.2 Lizenzen, Nutzungsbedingungen und Datenschutzbestimmungen

Durch die Nutzung der jeweiligen LLMs und Embedding Modelle akzeptiert der Kunde die folgend aufgeführten Nutzungsbedingungen, einschließlich der Datenschutzbestimmungen und Lizenzbedingungen für das jeweilige Model in ihrer jeweils aktuellen Fassung, wodurch eine Vereinbarung zwischen dem Kunden und dem jeweiligen Anbieter zustande kommt:

Modell	Plattform	Nutzungs-bedingungen	Datenschutz-bestimmun-gen
OpenAI GPT-3.5-Turbo-0125 (16k) OpenAI GPT-4o (128k) OpenAI GPT-4o mini Open AI o1 OpenAI o3 mini OpenAI Ada Text	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
Anthropic Claude 3.5 Sonnet V2 Anthropic Claude 3.7 Sonnet Google Gemini 2.0 Flash Mistral AI Mistral Large	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
Meta Llama3.3 (70b-instruct) DeutschlandGPT Llama-BildungsLLM-1.0	Open Telekom Cloud, Open Source	https://llama.meta.com/llama3/license/	
Mistral AI Mistral Small 3 Jina AI jina-embeddings-v2-base-de	Open Telekom		

Modell	Plattform	Nutzungs-bedingungen	Datenschutz-bestimmun-gen
Jina AI jina-embeddings-v2-base-code OpenGPT-X Teuken-7B-instruct-commercial-v0.4 Alibaba Cloud QwenVL-7B bce-reranker-base_v1 Alibaba Qwen Coder 2.5 7B	Cloud, Open Source	https://www.apache.org/licenses/LICENSE-2.0.html	
BAAI text-embedding-bge-m3 DeepSeek AI DeepSeek-Coder-V2-Lite DeepSeek AI DeepSeek-R1-Distill-Llama-70b DeepSeek AI DeepSeek-R1-Distill-Qwen-32B	Open Telekom Cloud, Open Source	https://opensource.org/license/mit	

3.3.3 Medical Device Regulation (MDR)

Der LLM-Serving-Service ist kein Medizinprodukt im Sinne von Art. 2 Nr.1 MDR. Der Kunde verpflichtet sich, den LLM-Serving-Service nicht als Medizinprodukt und insbesondere nicht zu folgenden Zwecken einzusetzen:

- Diagnose, Verhütung, Überwachung, Vorhersage, Prognose, Behandlung oder Linderung von Krankheiten,
- Diagnose, Überwachung, Behandlung, Linderung von oder Kompensierung von Verletzungen oder Behinderungen,
- Untersuchung, Ersatz oder Veränderung der Anatomie oder eines physiologischen oder pathologischen Vorgangs oder Zustands,
- Gewinnung von Informationen durch die In-vitro-Untersuchung von aus dem menschlichen Körper — auch aus Organ-, Blut- und Gewebespenden — stammenden Proben

3.3.4 Hochrisiko-KI-Systeme, Verbotene Anwendungsfälle

Die durch die Telekom bereitgestellten Leistungen sind kein Hochrisiko-KI-System im Sinne des Anhangs III der des EU AI-Act. Dem Kunden ist es untersagt, die Leistungen der Telekom in einer Art und Weise zu verwenden, die es zu einem Hochrisiko-KI-System macht oder es für verbotene Anwendungsfälle gemäß Artikel 5 AI-Act einzusetzen.

3.3.5 Haftung, Freistellung

Die Telekom haftet, soweit gesetzlich zulässig, nicht für Schäden, die aus der Nichteinhaltung dieser Nutzungs- und Lizenzbestimmungen resultieren.

Zudem stellt der Kunde die Telekom diesbezüglich von allen Ansprüchen Dritter (insb. Aufsichtsbehörden) auf erstes Anfordern vollumfänglich frei. Weitergehende Ansprüche der Telekom bleiben unberührt.

3.3.6 Feedback

Der Kunde räumt der Telekom und ihren verbundenen Unternehmen ein uneingeschränktes Nutzungsrecht an dessen Feedback und Ideen zu den beschriebenen Leistungen ein.

3.4 Optionale Leistungen

Die nachfolgenden optionalen Leistungen werden bei gesonderter Beauftragung gegen zusätzliche Vergütung erbracht. Auf Anfrage beim zuständigen Telekom-Vertriebsmitarbeiter wird die Telekom dem Kunden nach erfolgreicher Prüfung ein Angebot, sowie detaillierte Beschreibungen der Leistungen zur Verfügung stellen.

3.4.1 LLMs auf dedizierter Instanz

Die Telekom stellt einzelne LLMs und Embeddings auf einer dedizierten Instanz bereit. Der Kunde hat so exklusiven Zugang zu dem jeweiligen Sprachmodell, ohne es mit anderen Kunden zu teilen.

3.4.2 Feintuning von Modellen

Der LLM-Serving-Service des T-Systems SmartChat kann um Funktionalitäten für das Feintuning von LLMs erweitert werden. Auf Anfrage stellt die Telekom die Funktionalität samt Dokumentation zur Verfügung und unterstützt bei der Erstellung und Integration des feingetunten Modells in die Gesamtlösung.

3.4.3 Installation auf einer kundeneigenen Umgebung

Auf Anfrage prüft die Telekom die Möglichkeiten der Bereitstellung der Leistung auf einer kundeneigenen Umgebung. Ergibt die Prüfung der Telekom, dass die Umgebung des Kunden die erforderlichen Anforderungen erfüllt, stellt die Telekom die Leistungen im Rahmen einer kundenspezifischen Lösung auf dessen Umgebung bereit und übernimmt bei entsprechender Beauftragung den Betrieb der kundenspezifischen Lösung.

3.4.4 Zusätzliche LLMs und Embedding Modelle

Die Telekom stellt dem Kunden LLMs und Embedding Modelle, die nicht zu den aktuell verfügbaren LLMs und Embedding Modellen gehören, auf einer kundeneigenen oder

dedizierten Instanz bereit. Die Telekom unterbreitet dem Kunden ein entsprechendes Angebot, bei Vorliegen eines positiven Prüfungsergebnisses zur technischen Machbarkeit.

3.4.5 Weitere Unterstützungsleistungen

Die Telekom unterstützt den Kunden individuell mit folgenden Dienstleistungen auf Basis von Time & Material:

- Implementierungssupport
- Beratung/Coaching
- Schulung

4 Leistungsänderungen durch Telekom

Beabsichtigt die Telekom Änderungen der rechtlichen Bedingungen dieser Leistungsbeschreibung, der Leistungen oder Preise vorzunehmen, so werden die Änderungen dem Kunden mindestens einen Monat vor ihrem Wirksamwerden in Textform (z. B. per Brief oder E-Mail) mitgeteilt. Die Änderungen werden unter den nachfolgenden Voraussetzungen der Ziffern a) bis b) Vertragsbestandteil:

a. Änderungen zu Gunsten des Kunden zum Zeitpunkt ihres Wirksamwerdens.
b. Preiserhöhungen, Änderungen der rechtlichen Bedingungen und nicht lediglich unerheblichen Änderungen der Leistungen zu Ungunsten des Kunden zum Zeitpunkt ihres Wirksamwerdens. Dies gilt nicht, sofern der Kunde die Leistung ohne Einhaltung einer Kündigungsfrist zu diesem Zeitpunkt in Textform kündigt. Auf das Kündigungsrecht wird der Kunde in der Änderungsmitteilung ausdrücklich hingewiesen.

5 Mitwirkungsleistungen des Kunden

Der Kunde ist verpflichtet alle Leistungen, die zur ordnungsgemäßen Leistungserbringung durch Telekom erforderlich sind, insbesondere jedoch nachfolgende, unentgeltlich, rechtzeitig und im erforderlichen Umfang zu erbringen:

5.1 Allgemeine Mitwirkungsleistungen

a. Der Kunde ist verpflichtet, einen qualifizierten und entscheidungsbefugten Ansprechpartner zu benennen und dessen Erreichbarkeit/Vertretung sicherzustellen.
b. Der Kunde benennt bis zu 5 Administratoren, die ausreichend qualifiziert und berechtigt sind, Supportanfragen an die Telekom zu stellen
c. Der Kunde erklärt sich mit dem unverschlüsselten Schriftwechsel per E-Mail einverstanden und wird stets eine aktuelle E-Mail Adresse hinterlegen. Dem Kunden ist bekannt, dass für die Leistungserbringung wesentliche Informationen, wie Zugangsdaten, Informationen zu Änderungen der Leistungen und der rechtlichen Bedingungen, sowie Rechnungen ausschließlich per Mail versendet werden.

d. Der Kunde prüft eigenverantwortlich, ob die von ihm im Zusammenhang mit der Nutzung der Leistung an die Telekom übermittelten Daten personenbezogene Daten darstellen und die Verarbeitung dieser personenbezogenen Daten zulässig ist. Sofern der Kunde personenbezogene Daten verarbeiten lassen möchte, wird dieser eine Vereinbarung über die Verarbeitung personenbezogener Daten nach dem Muster der Telekom abschließen, welches die Telekom dem Kunden auf Anfrage zur Verfügung stellt.
e. Der Kunde prüft eigenverantwortlich alle für ihn im Zusammenhang mit der Nutzung der Leistung relevanten und anwendbaren rechtlichen Vorschriften, Gesetze, Verordnungen und branchenspezifischen Bestimmungen und stellt deren Einhaltung sicher. Dazu zählen insbesondere auch die Einhaltung von Geheimhaltungsverpflichtungen, die z.B. aus einer beruflichen Tätigkeit herrühren, sowie die Einhaltung der Nutzungs- und Lizenzbestimmungen.
f. Der Kunde stellt sicher, dass die Leistungen nicht missbräuchlich genutzt werden.
g. Der Kunde stellt der Telekom alle erforderlichen Informationen zur Verfügung und stellt sicher, dass seine Angaben inhaltlich richtig und stets aktuell sind.
h. Der Kunde ist verpflichtet Einrichtungen und Leistungen der Telekom vor unberechtigtem Zugriff, Schadsoftware und sonstigen Beeinträchtigungen durch geeignete Maßnahmen zu schützen, diese pfleglich zu behandeln und die Angaben der Hersteller zu beachten.
i. Der Kunde ist verpflichtet Passwörter und Zugangsdaten, insbesondere von ihm generierte API-Keys, geheim zu halten, nur an berechtigte Dritte weiterzugeben, bzw. vor deren Zugriff zu schützen und soweit erforderlich zu ändern. Der Kunde wird die Telekom unverzüglich bei Anhaltspunkten der Kenntnisnahme durch unberechtigte Dritte informieren.
j. Support
   Der Kunde ist verpflichtet, die Telekom bei der Behebung einer Störung/Beeinträchtigung zu unterstützen.
   Insbesondere erbringt der Kunde für seine Nutzer einen 1st Level Support (User Helpdesk) und führt in diesem Rahmen eine Selbstprüfung durch, um auszuschließen, dass die Störungsursache in seinem Verantwortungsbereich liegt. Der Kunde behebt alle Störungen in seinem Verantwortungsbereich eigenständig.
   Soweit der Kunde Störungen nicht eigenständig beheben kann und diese durch die Telekom zu vertreten sind, meldet der Kunde diese Störungen dem Support der Telekom mit einer nachvollziehbaren Schilderung der Fehlersymptome und unter Angabe sonstiger relevanter Informationen.
   Der Kunde ist verpflichtet Störungen, Beeinträchtigungen der Leistungen oder Beschädigungen an den Einrichtungen der Telekom unverzüglich anzuzeigen.
k. Der Kunde ist verpflichtet, für eine ausreichende Deckung des vereinbarten Abbuchungskontos zu sorgen, sowie im Falle der Zahlung per Kreditkarte seine bei der Registrierung hinterlegten Kreditkartendaten auf dem aktuellen Stand zu halten.
l. Für den Zugriff auf den API-Endpoint via der Python-Bibliothek „openai“ stellt der Kunde die benötigte Software bei und stellt sicher, dass diese stets aktuell ist.
m. Der Kunde stellt sicher, dass er über die erforderlichen Nutzungsrechte (einschließlich Updates oder Upgrades) verfügt.
n. Der Kunde stellt die Einhaltung der Bestimmungen dieser Leistungsbeschreibung, insbesondere der Mitwirkungsleistungen, sowie Nutzungs- und Lizenzbestimmungen durch seine Nutzer sicher.
o. Der Kunde wird die Telekom unverzüglich in Textform informieren, wenn er eine Mitwirkungsleistung nicht wie vereinbart erbringen kann, oder Umstände eintreten, die der Telekom die Erbringung der Leistungen erschweren oder unmöglich machen.

5.2 Mitwirkungsleistungen bei Betrieb

a. Der Kunde stellt sicher, dass er keine Daten, Inhalte oder sonstige Eingaben verwendet, auf dem vertragsgegenständlichen Speicherplatz speichert oder sonst zugänglich macht, die Malicious Codes oder sonstige Schadsoftware enthalten, und/oder deren Bereitstellung, Veröffentlichung, Übertragung oder Nutzung gegen geltendes Recht oder Rechte Dritter verstößt.
b. Der Kunde stellt sicher, dass von seiner Nutzung der Leistung keine Gefährdung oder Störung Dritter oder der Einrichtungen und Leistungen der Telekom ausgeht. Im Falle einer solchen Gefährdung oder Störung (z.B. auf Grund einer DDoS Attacke) ist die Telekom ohne vorherige Benachrichtigung des Kunden berechtigt, die betroffene Leistung bis zur Beseitigung der Ursache der Gefährdung oder Störung zu deaktivieren. Die hierdurch entstehenden Ausfallzeiten bleiben bei der Berechnung der Verfügbarkeit unberücksichtigt. Die Telekom wird den Kunden informieren.
a. Der Kunde informiert die Telekom unverzüglich über alle mutmaßlichen, die Einrichtungen und Leistungen der Telekom betreffenden, sicherheitsrelevanten Vorfälle.

6 Mindestlaufzeit/Beendigung

Die Leistung ist mit einer Frist von einem Monat zum Monatsende kündbar. Alle Kündigungen haben per E-Mail an die im Rahmen der Bereitstellung bekanntgegebene E-Mail-Adresse unter Angabe der Vertragsnummer zu erfolgen.

Mit Beendigung des LLM-Serving-Services werden alle Zugangsmöglichkeiten deaktiviert und die Daten des Kunden gelöscht.

Das Recht zur außerordentlichen Kündigung bleibt hiervon unberührt.

7 Preise, Kommerzielle Bedingungen

7.1 Verfahren der Entgeltberechnung

a) Tokenverbrauch

Die Nutzung der Modelle wird basierend auf der Anzahl der verbrauchten Tokens abgerechnet. Der Tokenverbrauch ist im LLM-Self-Service-Portal einsehbar. Die Abrechnung erfolgt je angefangener Million Input-/Output-Tokens für das jeweilige Modell.

b) Tarife

Die Telekom stellt den LLM-Serving-Service in unterschiedlichen Tarifen bereit. Diese Tarife unterscheiden sich in den folgenden Aspekten:

- Verfügbare Modelle
- Mindestumsatz/kein Mindestumsatz
- Max. Input TPM (Token per Minute): Maximale Anzahl von verarbeiteten Input-Token pro Minute
- Max. Output TPM (Token per Minute): Maximale Anzahl von verarbeiteten Output-Token pro Minute

- Max. RPM (Requests per Minute): Maximale Anzahl von Anfragen pro Minute, die an das LLM gestellt werden können

Bei Tarifen mit Mindestumsatz berechnet die Telekom auch bei einer den Mindestumsatz unterschreitenden Nutzung des Kunden den jeweiligen Mindestumsatz.
Bei untermonatlicher Bereitstellung/Beendigung eines Tarifs wird der Mindestumsatz anteilig nach Kalendertagen berechnet. Tarifwechsel sind auf Anfrage einvernehmlich möglich.

Die Telekom weist darauf hin, dass es sich bei den Werten Max Input TPM, Max Output TPM, Max RPM um maximal mögliche Rate-Limits handelt. Die tatsächlich zu erwartenden Durchsätze werden darunter liegen.

c) Rechnungsstellung

Die Telekom stellt dem Kunden für den jeweils vorangegangenen Kalendermonat monatlich eine Rechnung.

d) Steuern und Abgaben

Alle Preise sind in Euro angegeben und verstehen sich zuzüglich der zum Zeitpunkt der Lieferung und Leistung geltenden Steuern und Abgaben.

7.2 Preise

7.2.1 Tarif Basic

Der monatliche Mindestumsatz für Modelle im Tarif Basic beträgt: 0€

Plattform	Modell	Modell-typ	Max. Input TPM	Max. Output TPM	Max. RPM	Preis pro Million Input-Token	Preis pro Million Output-Token
Open Telekom Cloud	Meta Llama 3.3 70b-instruct AWQ	LLM	13.800	5.000	20	3,53 €	3,53 €
Open Telekom Cloud	Mistral AI Mistral Small 3	LLM	18.750	7.500	20	3,53 €	3,53 €
Open Telekom Cloud	Jina AI jina-embeddings-v2-base-de	Em-bedding	75.000		20	0,48 €	0,48 €
Open Telekom Cloud	BAAI text-embedding-bge-m3	Em-bedding	75.000		20	0,48 €	0,48 €
Open Telekom Cloud	Jina AI jina-embeddings-v2-base-code	Em-bedding	75.000		20	0,48 €	0,48 €
Open Telekom Cloud	Alibaba Qwen 2.5 VL 72B AWQ	LLM	3.125	1.500	20	3,53 €	3,53 €
Open Telekom Cloud	DeepSeek AI DeepSeek-Coder-V2-Lite	LLM	34.500	10.000	20	3,53 €	3,53 €
Open Telekom Cloud	Alibaba Qwen Coder 2.5 7B	LLM	20.000	6.250	20	3,53 €	3,53 €
Open Telekom Cloud	DeepSeek-R1-Distill-Llama-70B AWQ	LLM	11.250	3.750	20	3,53 €	3,53 €
Open Telekom Cloud	OpenGPT-X Teuken-7B-instruct-commercial-v0.4	LLM	13.800	5.000	20	3,53 €	3,53 €
Open Telekom Cloud	DeutschlandGPT Llama-BildungsLLM-1.0	LLM	20.000	10.000	100	3,95 €	3,95 €

7.2.2 Tarif Standard1000

Der monatliche Mindestumsatz für Modelle im Tarif Standard1000 beträgt: 1.000€

Plattform	Modell	Modell-typ	Max. Input TPM	Max. Output TPM	Max. RPM	Preis pro Million Input-Token	Preis pro Million Output-Token
Open Telekom Cloud	Meta Llama 3.3 70b-instruct AWQ	LLM	34.500	12.500	150	3,53 €	3,53 €
Open Telekom Cloud	Mistral AI Mistral Small 3	LLM	37.500	15.000	150	3,53 €	3,53 €
Open Telekom Cloud	Jina AI jina-embeddings-v2-base-de	Em-beddin g	150.000		150	0,48 €	0,48 €
Open Telekom Cloud	BAAI text-embedding-bge-m3	Em-beddin g	150.000		150	0,48 €	0,48 €
Open Telekom Cloud	Jina AI jina-embeddings-v2-base-code	Em-beddin g	150.000		150	0,48 €	0,48 €
Open Telekom Cloud	Alibaba Qwen 2.5 VL 72B AWQ	LLM	6.250	3.000	150	3,53 €	3,53 €
Open Telekom Cloud	DeepSeek AI DeepSeek-Coder-V2-Lite	LLM	103.500	37.500	150	3,53 €	3,53 €
Open Telekom Cloud	Alibaba Qwen Coder 2.5 7B	LLM	40.000	12.500	150	3,53 €	3,53 €
Open Telekom Cloud	DeepSeek-R1-Distill-Llama-70B AWQ	LLM	22.500	7.500	150	3,53 €	3,53 €
Open Telekom Cloud	OpenGPT-X Teuken-7B-instruct-commercial-v0.4	LLM	34.500	12.500	150	3,53 €	3,53 €
Open Telekom Cloud	Meta Llama 3.3 70b-instruct AWQ	LLM	34.500	12.500	150	3,53 €	3,53 €
Microsoft Azure	**OpenAI GPT-3.5-Turbo-0125 (16k), Azure France (depricated!)**	LLM	240.000		1.444	0,59 €	1,77 €
Microsoft Azure	**OpenAI GPT-4o-2024-11-20, Azure France**	LLM	30.000.000		300.000	3,25 €	12,99 €
Microsoft Azure	**OpenAI GPT-4o-mini-2024-07-18, Azure Sweden**	LLM	150.000.000		1.000.000	0,19 €	0,78 €
Microsoft Azure	**OpenAI o1 2024-12-17, Azure Sweden**	LLM	30.000.000		5.000	19,48 €	77,92 €
Microsoft Azure	**OpenAI o1 mini 2024-09-12, Azure Sweden**	LLM	50.000.000		5.000	3,90 €	15,58 €
Microsoft Azure	**OpenAI o3 mini 2025-01-31-Azure Sweden**	LLM	50.000.000		5.000	1,43 €	5,71 €
Microsoft Azure	**OpenAI Ada-Text, Azure France**	Em-beddin g	2.400.000		1.440	0,11 €	0,11 €
Google Cloud Platform	**Mistral AI Mistral Large 2411**	LLM	400.000		60	2,21 €	6,62 €

Plattform	Modell	Modell-typ	Max. Input TPM	Max. Output TPM	Max. RPM	Preis pro Million Input-Token	Preis pro Million Output-Token
Google Cloud Platform	**Anthropic Claude 3.5 Sonnet V2**	LLM	200.000		120	3,31 €	16,55 €
Google Cloud Platform	**Anthropic Claude 3.7 Sonnet**	LLM	200.000		120	3,31 €	16,55 €
Google Cloud Platform	**Google Gemini 2.0 Flash**	LLM	4.000.000		2.000	0,11 €	0,44 €

7.2.3 Tarif Standard2000

Der monatliche Mindestumsatz für Modelle im Tarif Standard2000 beträgt: 2.000€

Plattform	Modell	Modell-typ	Max. Input TPM	Max. Output TPM	Max. RPM	Preis pro Million Input-Token	Preis pro Million Output-Token
Open Telekom Cloud	Meta Llama 3.3 70b-instruct AWQ	LLM	69.000	25.000	300	2,89 €	2,89 €
Open Telekom Cloud	Mistral AI Mistral Small 3	LLM	75.000	30.000	300	2,89 €	2,89 €
Open Telekom Cloud	Jina AI jina-embeddings-v2-base-de	Em-beddin g	300.000		300	0,48 €	0,48 €
Open Telekom Cloud	BAAI text-embedding-bge-m3	Em-beddin g	300.000		300	0,48 €	0,48 €
Open Telekom Cloud	Jina AI jina-embeddings-v2-base-code	Em-beddin g	300.000		300	0,48 €	0,48 €
Open Telekom Cloud	Alibaba Qwen 2.5 VL 72B AWQ	LLM	12.500	6.000	300	2,89 €	2,89 €
Open Telekom Cloud	DeepSeek AI DeepSeek-Coder-V2-Lite	LLM	207.000	75.000	300	2,89 €	2,89 €
Open Telekom Cloud	Alibaba Qwen Coder 2.5 7B	LLM	80.000	25.000	300	2,89 €	2,89 €
Open Telekom Cloud	DeepSeek-R1-Distill-Llama-70B AWQ	LLM	45.000	15.000	300	2,89 €	2,89 €
Open Telekom Cloud	OpenGPT-X Teuken-7B-instruct-commercial-v0.4	LLM	69.000	25.000	300	2,89 €	2,89 €
Open Telekom Cloud	Meta Llama 3.3 70b-instruct AWQ	LLM	69.000	25.000	300	2,89 €	2,89 €
Microsoft Azure	**OpenAI GPT-3.5-Turbo-0125 (16k), Azure France (depricated!)**	LLM	240.000		1.444	0,59 €	1,77 €
Microsoft Azure	**OpenAI GPT-4o-2024-11-20, Azure France**	LLM	30.000.000		300.000	3,25 €	12,99 €
Microsoft Azure	**OpenAI GPT-4o-mini-2024-07-18, Azure Sweden**	LLM	150.000.000		1.000.000	0,19 €	0,78 €

Plattform	Modell	Modell-typ	Max. Input TPM	Max. Output TPM	Max. RPM	Preis pro Million Input-Token	Preis pro Million Output-Token
Microsoft Azure	**OpenAI o1 2024-12-17, Azure Sweden**	LLM	30.000.000		5.000	19,48 €	77,92 €
Microsoft Azure	**OpenAI o1 mini 2024-09-12, Azure Sweden**	LLM	50.000.000		5.000	3,90 €	15,58 €
Microsoft Azure	**OpenAI o3 mini 2025-01-31-Azure Sweden**	LLM	50.000.000		5.000	1,43 €	5,71 €
Microsoft Azure	**OpenAI Ada-Text, Azure France**	Em-beddin g	2.400.000		1.440	0,11 €	0,11 €
Google Cloud Platform	**Mistral AI Mistral Large 2411**	LLM	400.000		60	2,21 €	6,62 €
Google Cloud Platform	**Anthropic Claude 3.5 Sonnet V2**	LLM	200.000		120	3,31 €	16,55 €
Google Cloud Platform	**Anthropic Claude 3.7 Sonnet**	LLM	200.000		120	3,31 €	16,55 €
Google Cloud Platform	**Google Gemini 2.0 Flash**	LLM	4.000.000		2.000	0,11 €	0,44 €

7.2.4 Tarif Standard3000

Der monatliche Mindestumsatz für Modelle im Tarif Standard3000 beträgt: 3.000€

Plattform	Modell	Modell-typ	Max. Input TPM	Max. Output TPM	Max. RPM	Preis pro Million Input-Token	Preis pro Million Output-Token
Open Telekom Cloud	Meta Llama 3.3 70b-instruct AWQ	LLM	103.500	37.500	450	2,23 €	2,23 €
Open Telekom Cloud	Mistral AI Mistral Small 3	LLM	112.500	45.000	450	2,23 €	2,23 €
Open Telekom Cloud	Jina AI jina-embeddings-v2-base-de	Em-beddin g	450.000		450	0,48 €	0,48 €
Open Telekom Cloud	BAAI text-embedding-bge-m3	Em-beddin g	450.000		450	0,48 €	0,48 €
Open Telekom Cloud	Jina AI jina-embeddings-v2-base-code	Em-beddin g	450.000		450	0,48 €	0,48 €
Open Telekom Cloud	Alibaba Qwen 2.5 VL 72B AWQ	LLM	18.750	9.000	450	2,23 €	2,23 €
Open Telekom Cloud	DeepSeek AI DeepSeek-Coder-V2-Lite	LLM	310.500	112.500	450	2,23 €	2,23 €
Open Telekom Cloud	Alibaba Qwen Coder 2.5 7B	LLM	120.000	37.500	450	2,23 €	2,23 €
Open Telekom Cloud	DeepSeek-R1-Distill-Llama-70B AWQ	LLM	67.500	22.500	450	2,23 €	2,23 €
Open Telekom Cloud	OpenGPT-X Teuken-7B-instruct-commercial-v0.4	LLM	103.500	37.500	450	2,23 €	2,23 €

Plattform	Modell	Modell-typ	Max. Input TPM	Max. Output TPM	Max. RPM	Preis pro Million Input-Token	Preis pro Million Output-Token
Open Telekom Cloud	Meta Llama 3.3 70b-instruct AWQ	LLM	103.500	37.500	450	2,23 €	2,23 €
Microsoft Azure	**OpenAI GPT-3.5-Turbo-0125 (16k), Azure France (depricated!)**	LLM	240.000		1.444	0,59 €	1,77 €
Microsoft Azure	**OpenAI GPT-4o-2024-11-20, Azure France**	LLM	30.000.000		300.000	3,25 €	12,99 €
Microsoft Azure	**OpenAI GPT-4o-mini-2024-07-18, Azure Sweden**	LLM	150.000.000		1.000.000	0,19 €	0,78 €
Microsoft Azure	**OpenAI o1 2024-12-17, Azure Sweden**	LLM	30.000.000		5.000	19,48 €	77,92 €
Microsoft Azure	**OpenAI o1 mini 2024-09-12, Azure Sweden**	LLM	50.000.000		5.000	3,90 €	15,58 €
Microsoft Azure	**OpenAI o3 mini 2025-01-31-Azure Sweden**	LLM	50.000.000		5.000	1,43 €	5,71 €
Microsoft Azure	**OpenAI Ada-Text, Azure France**	Em-beddin g	2.400.000		1.440	0,11 €	0,11 €
Google Cloud Platform	**Mistral AI Mistral Large 2411**	LLM	400.000		60	2,21 €	6,62 €
Google Cloud Platform	**Anthropic Claude 3.5 Sonnet V2**	LLM	200.000		120	3,31 €	16,55 €
Google Cloud Platform	**Anthropic Claude 3.7 Sonnet**	LLM	200.000		120	3,31 €	16,55 €
Google Cloud Platform	**Google Gemini 2.0 Flash**	LLM	4.000.000		2.000	0,11 €	0,44 €

7.2.5 Tarif Standard4000

Der monatliche Mindestumsatz für Modelle im Tarif Standard4000 beträgt: 4.000€

Plattform	Modell	Modell-typ	Max. Input TPM	Max. Output TPM	Max. RPM	Preis pro Million Input-Token	Preis pro Million Output-Token
Open Telekom Cloud	Meta Llama 3.3 70b-instruct AWQ	LLM	138.000	50.000	600	1,75 €	1,75 €
Open Telekom Cloud	Mistral AI Mistral Small 3	LLM	150.000	60.000	600	1,75 €	1,75 €
Open Telekom Cloud	Jina AI jina-embeddings-v2-base-de	Em-bedding	600.000		600	0,48 €	0,48 €
Open Telekom Cloud	BAAI text-embedding-bge-m3	Em-bedding	600.000		600	0,48 €	0,48 €
Open Telekom Cloud	Jina AI jina-embeddings-v2-base-code	Em-bedding	600.000		600	0,48 €	0,48 €
Open Telekom Cloud	Alibaba Qwen 2.5 VL 72B AWQ	LLM	25.000	12.000	600	1,75 €	1,75 €

Plattform	Modell	Modell-typ	Max. Input TPM	Max. Output TPM	Max. RPM	Preis pro Million Input-Token	Preis pro Million Output-Token
Open Telekom Cloud	DeepSeek AI DeepSeek-Coder-V2-Lite	LLM	414.000	150.000	600	1,75 €	1,75 €
Open Telekom Cloud	Alibaba Qwen Coder 2.5 7B	LLM	160.000	50.000	600	1,75 €	1,75 €
Open Telekom Cloud	DeepSeek-R1-Distill-Llama-70B AWQ	LLM	90.000	30.000	600	1,75 €	1,75 €
Open Telekom Cloud	OpenGPT-X Teuken-7B-instruct-commercial-v0.4	LLM	138.000	50.000	600	1,75 €	1,75 €
Open Telekom Cloud	Meta Llama 3.3 70b-instruct AWQ	LLM	138.000	50.000	600	1,75 €	1,75 €
Microsoft Azure	**OpenAI GPT-3.5-Turbo-0125 (16k), Azure France (depricated!)**	LLM	240.000		1.444	0,59 €	1,77 €
Microsoft Azure	**OpenAI GPT-4o-2024-11-20, Azure France**	LLM	30.000.000		300.000	3,25 €	12,99 €
Microsoft Azure	**OpenAI GPT-4o-mini-2024-07-18, Azure Sweden**	LLM	150.000.000		1.000.000	0,19 €	0,78 €
Microsoft Azure	**OpenAI o1 2024-12-17, Azure Sweden**	LLM	30.000.000		5.000	19,48 €	77,92 €
Microsoft Azure	**OpenAI o1 mini 2024-09-12, Azure Sweden**	LLM	50.000.000		5.000	3,90 €	15,58 €
Microsoft Azure	**OpenAI o3 mini 2025-01-31-Azure Sweden**	LLM	50.000.000		5.000	1,43 €	5,71 €
Microsoft Azure	**OpenAI Ada-Text, Azure France**	Em-bedding	2.400.000		1.440	0,11 €	0,11 €
Google Cloud Platform	**Mistral AI Mistral Large 2411**	LLM	400.000		60	2,21 €	6,62 €
Google Cloud Platform	**Anthropic Claude 3.5 Sonnet V2**	LLM	200.000		120	3,31 €	16,55 €
Google Cloud Platform	**Anthropic Claude 3.7 Sonnet**	LLM	200.000		120	3,31 €	16,55 €
Google Cloud Platform	**Google Gemini 2.0 Flash**	LLM	4.000.000		2.000	0,11 €	0,44 €

7.2.6 Preise für optionale Dienstleistungen

Der Tagessatz für optionale Dienstleistungen wie Implementierungssupport, Beratung und Coaching beträgt 1.144,45 €.

Die Leistungen werden remote erbracht. Die Abrechnung der Unterstützungsleistungen erfolgt monatlich anhand eines vom Auftraggeber unterzeichneten Leistungsnachweises auf Basis von Time&Material.

8 Glossar/ Abkürzungsverzeichnis

Abkürzung	Erklärung
API-Key	Ein API-Key ist ein einzigartiger Authentifizierungsschlüssel, der den Zugriff auf eine Programmierschnittstelle (API) ermöglicht, indem er die aufrufende Anwendung identifiziert, authentifiziert und autorisiert.
ASR	Automatic Speech Recognition, Automatische Spracherkennung
Betreute Betriebszeit	Ist die Zeit von Montag bis Freitag jeweils von 9:00 Uhr bis 17.00 Uhr MEZ/MESZ, ausgenommen an bundeseinheitlichen Feiertagen in Deutschland
LLM	Large Language Model
Modelle	Modelle umfassen in diesem Dokument sowohl Large Language Modelle als auch Embedding Modelle
PaaS	Platform as a Service
SaaS	Software as a Service
Token	Token umfassen Input- als auch Output-Token, sofern nicht differenziert wird. Token sind Textsequenzen, z. B. Sätze, Wörter oder Zeichen, die ein LLM verarbeiten kann. Die Größe eines Tokens kann sich je nach LLM unterscheiden.
TPM	Die Anzahl der maximal verarbeitbaren Tokens pro Minute
RPM	Die Anzahl der maximal verarbeitbaren Requests pro Minute