Leistungsbeschreibung pro Model Hub

profihost GmbH · Fassung 1.0 · Stand 20.10.2026

1. Geltung und Rangfolge

1.1 Diese Leistungsbeschreibung regelt Art und Umfang der Leistungen der profihost GmbH (nachfolgend: profihost) unter dem Produktnamen „pro Model Hub“.

1.2 Vertragsbestandteile sind in dieser Rangfolge:

1. diese Leistungsbeschreibung einschließlich ihrer Anlagen,
2. die Besonderen Bedingungen für KI-Funktionen,
3. die Allgemeinen Geschäftsbedingungen AI Modelle (AGB),
4. die Vereinbarung zur Auftragsverarbeitung (AVV), die für die Verarbeitung personenbezogener Daten in Eingaben und Ausgaben vorrangig gilt.

1.3 Das Angebot richtet sich ausschließlich an Unternehmer im Sinne von § 14 BGB.

2. Leistungsübersicht

Der pro Model Hub stellt dem Kunden Sprach- und Embedding-Modelle über eine Programmierschnittstelle (API) zur Verfügung. Die Modelle laufen auf Hardware, die profihost in Rechenzentren in der Europäischen Union (derzeit Deutschland und Finnland) betreibt bzw. betreiben lässt. Der Kunde bindet die Modelle in eigene Anwendungen ein und bezahlt ein monatliches Paket mit einem Kontingent an Credits.

Zum Leistungsumfang gehören:

- die API unter `https://api.promodelhub.de`,
- das Kundenportal unter `https://app.promodelhub.de` zur Verwaltung von Account, Nutzern, API-Keys, Paket und Verbrauch,
- die Dokumentation unter `https://docs.promodelhub.de`,
- Support nach Ziffer 11.

3. Account, Nutzer und API-Keys

3.1 Registrierung: Der Kunde registriert sich im Portal mit Vorname, Nachname, Unternehmen, E-Mail-Adresse, Telefonnummer und optional seiner profihost Kundennummer. profihost prüft die Registrierung und schaltet den Account frei. Ein Anspruch auf Freischaltung besteht nicht.

3.2 Nutzer und Rollen: Ein Account kann mehrere Nutzer haben. Es gibt die Rollen Owner, Admin und Member. Owner und Admin verwalten Nutzer und API-Keys und sehen Paket, Preise und Abrechnungen. Paket bestellen, wechseln und kündigen sowie die Rechnungsangaben hinterlegen kann nur der Owner. Unternehmen, Rechnungsadresse und USt-IdNr. ändert nach der ersten Bestellung nur der Support, eine fehlende Angabe kann der Kunde einmal nachtragen. Member sehen ihren eigenen Verbrauch und ihr Credit-Limit, aber kein Paket und keine Preise. Pro Nutzer lassen sich ein Credit-Limit und die freigegebenen Modelle festlegen. Nutzer lassen sich sperren, die Sperre wirkt sofort auf alle API-Keys des Nutzers.

3.3 API-Keys: Jeder Nutzer kann beliebig viele API-Keys anlegen, jeweils mit optionaler Modellbeschränkung, optionalem Credit-Limit pro Abrechnungszeitraum und optionalem Ablaufdatum. Keys lassen sich rotieren und widerrufen. Ein Key wird nur bei der Erstellung angezeigt und von profihost nur als Hash gespeichert, profihost kann ihn deshalb nicht wiederherstellen.

3.4 Der Kunde ist für die Geheimhaltung seiner Zugangsdaten und API-Keys verantwortlich (§ 7 Abs. 2 AGB). Requests mit einem Key des Kunden werden seinem Paket zugerechnet.

4. API und Funktionen

4.1 Die API ist kompatibel zur OpenAI-API. Bereitgestellt werden die Endpunkte:

- `POST /v1/chat/completions` (Chat-Modelle),
- `POST /v1/embeddings` (Embedding-Modelle),
- `POST /v1/rerank` (Rerank-Modelle, auch unter `/rerank`),
- `GET /v1/models` (Liste der freigegebenen Modelle).

Der Rerank-Endpunkt gehört nicht zur OpenAI-API. Sein Format beschreibt die Dokumentation.

Weitere Endpunkte der OpenAI-API (z. B. Assistants, Files, Batch, Fine-Tuning, Bild- und Audioerzeugung) sind nicht Teil der Leistung. Kompatibilität bedeutet, dass gängige OpenAI-Clients und -SDKs mit geänderter Basis-URL funktionieren. Eine vollständige Übereinstimmung mit jedem Parameter und Verhalten der OpenAI-API schuldet profihost nicht. profihost ist berechtigt, die bereitgestellten Modelle dem technischen Fortschritt anzupassen.

4.2 Funktionen, soweit das jeweilige Modell sie unterstützt (Angabe pro Modell in der Dokumentation):

- Streaming der Antwort,
- Tool-Calling (Funktionsaufrufe),
- strukturierte Ausgaben (JSON),
- Reasoning, bei Modellen mit optionalem Reasoning pro Request ein- und ausschaltbar,
- Bildeingabe: bis zu 4 Bilder pro Request, als Base64 im OpenAI-Format. Bild-URLs werden nicht abgerufen,
- Embeddings für Text,
- Reranking: Sortieren einer Liste von Texten nach ihrer Relevanz für eine Suchanfrage.

4.3 Beta-Funktionen: Als „Beta“ gekennzeichnete Modelle und Funktionen (derzeit Bildeingabe, Embeddings und Reranking) stellt profihost zur Erprobung bereit. Sie können sich ändern oder eingestellt werden, ohne dass die Fristen aus Ziffer 5.3 gelten.

5. Modelle

5.1 Die verfügbaren Modelle, ihre Credit-Faktoren, Kontextfenster und Fähigkeiten stehen in Anlage 2 und tagesaktuell im Portal und in der Dokumentation. profihost betreibt frei verfügbare Modelle (Open-Weight-Modelle) auf eigener Infrastruktur. Eingaben werden nicht an die Hersteller der Modelle übermittelt.

5.2 Alle Modelle stehen in allen Paketen zur Verfügung. Die Modelle sind dauerhaft geladen und antworten ohne Wartezeit für einen Modellwechsel.

5.3 Änderungen am Modellangebot: profihost kann neue Modelle aufnehmen und Modelle durch neuere Versionen ersetzen oder abkündigen. Die Abschaltung oder den Ersatz eines Modells kündigt profihost mindestens 4 Wochen vorher im Portal, im Changelog der Dokumentation und per E-Mail an. Ohne Frist ist eine Abschaltung zulässig, wenn ein

Sicherheitsproblem, eine rechtliche Vorgabe oder eine Änderung der Lizenz des Modells sie erfordert. Da Credits für alle Modelle gelten, behält der Kunde bei Abschaltung eines Modells sein Kontingent für die übrigen Modelle.

5.4 Credit-Faktoren: Eine Änderung des Credit-Faktors eines Modells kündigt profihost mit derselben Frist an. Jeder Request wird mit dem Faktor berechnet, der zum Zeitpunkt des Requests gilt.

5.5 Ausgaben von Sprachmodellen können falsch, unvollständig oder unpassend sein. Die Eigenschaften eines Modells (z. B. Qualität, Sprache, Wissensstand) bestimmt sein Hersteller, nicht profihost. Näheres regeln die Besonderen Bedingungen für KI-Funktionen.

6. Pakete und Credits

6.1 Der Kunde bucht ein Paket aus Anlage 1. Jedes Paket enthält pro Abrechnungszeitraum ein Kontingent an Credits sowie Grenzen für Requests pro Minute, parallele Requests und Embedding-Token pro Minute.

6.2 Verbrauch: Ein Request verbraucht `(Input-Token + Output-Token) × Credit-Faktor des Modells` Credits. Maßgeblich ist die Token-Zählung des Modells, wie sie im Feld `usage` der API-Antwort ausgewiesen ist. Reasoning-Token zählen als Output-Token, Bilder als Input-Token. Bei Rerank-Requests gibt es keine Output-Token. Maßgeblich ist die Token-Zahl, die das Modell für den Request meldet (`meta.billed_units.total_tokens` in der API-Antwort). Sie enthält für jedes Dokument die Suchanfrage und eine feste Anweisung und ist deshalb größer als die Summe aus Suchanfrage und Dokumenten. Abgelehnte Requests (Status 4xx) verbrauchen keine Credits.

6.3 Kontingent: Zu Beginn jedes Abrechnungszeitraums wird das Kontingent neu bereitgestellt. Nicht verbrauchte Credits verfallen am Ende des Zeitraums und werden nicht erstattet oder übertragen.

6.4 Verbrauchsgrenze: Ist das Kontingent verbraucht, nimmt die API keine weiteren Requests an (Status `429`, Code `insufficient_quota`), bis der nächste Zeitraum beginnt oder der Kunde ein größeres Paket bucht. Es entstehen keine nutzungsabhängigen Zusatzkosten.

6.5 Hinweise: Bei 75 % und 90 % Verbrauch erhält der Kunde einen Hinweis im Portal und per E-Mail.

6.6 Limits im Team: Mit Credit-Limits pro Nutzer und pro API-Key teilt der Kunde das Kontingent selbst auf. Ein Key, der sein Limit erreicht, wird gesperrt und bei Erhöhung des Limits oder im nächsten Zeitraum wieder freigegeben.

6.7 Verbrauchsnachweis: Das Portal zeigt verbrauchte und verfügbare Credits, den Verbrauch pro Modell, pro Tag und für die letzten 24 Stunden pro Stunde, den laufenden Zeitraum und einen Abrechnungsdatensatz pro Zeitraum.

7. Technische Grenzen

7.1 Es gelten diese Grenzen (Stand der Dokumentation unter „Limits“):

Grenze	Wert
Requests pro Minute	je Paket, pro Account (Anlage 1)
parallele Requests	je Paket, pro Account (Anlage 1)
Embedding-Token pro Minute	je Paket, pro Account (Anlage 1)
Token pro Embedding-Request	300.000, höchstens die Embedding-Token pro Minute des Pakets
Kontextfenster (Eingabe + Ausgabe)	je Modell (Anlage 2)
Größe eines Requests	20 MB
Bilder pro Request	4
Eingabe pro Embedding	8.192 Token
Dauer eines Requests	600 Sekunden
Requests pro IP-Adresse	etwa 10 pro Sekunde

7.2 Die Limits pro Paket gelten für alle Nutzer und API-Keys eines Accounts zusammen. Wer ein Limit überschreitet, erhält Status `429` und kann den Request später wiederholen.

7.3 Mehrere Accounts für dasselbe Unternehmen, um Limits oder die Testphase mehrfach zu nutzen, sind nicht zulässig.

8. Testphase

8.1 Jeder freigeschaltete Account kann das Paket Starter einmalig 30 Tage kostenlos nutzen. Die Aktivierung der Testphase ist bereits die Bestellung des Pakets Starter.

8.2 Kündigt der Kunde nicht vor Ablauf der Testphase, läuft das Paket Starter ab dem 31. Tag kostenpflichtig weiter.

8.3 Bucht der Kunde während der Testphase ein größeres Paket, endet die Testphase. Der kostenpflichtige Zeitraum im neuen Paket beginnt sofort mit vollem Kontingent. Bestellt der Kunde direkt ein größeres Paket, entfällt die Testphase.

9. Laufzeit, Paketwechsel und Kündigung

9.1 Ein Paket hat eine Laufzeit von einem Monat ab Bestelldatum bzw. ab Ende der Testphase (Abrechnungszeitraum) und verlängert sich jeweils um einen weiteren Monat, wenn es nicht gekündigt wird.

9.2 Upgrade: Ein größeres Paket gilt sofort. Für den Rest des laufenden Zeitraums berechnet profihost die Preisdifferenz anteilig (Preisdifferenz × Resttage ÷ Tage des Zeitraums, der Tag des Upgrades zählt als Resttag). Das Kontingent steigt auf das des neuen Pakets, bereits verbrauchte Credits bleiben verbraucht.

9.3 Downgrade und Kündigung werden zum Ende des laufenden Zeitraums wirksam. Bis dahin gilt das bisherige Paket. Eine vorgemerkte Kündigung lässt sich bis zum Ende des Zeitraums zurücknehmen. Ein Upgrade hebt einen vorgemerkten Downgrade und eine vorgemerkte Kündigung auf.

9.4 Paketwechsel und Kündigung erfolgen im Portal. Die Kündigung in Textform nach § 5 AGB bleibt möglich.

9.5 Nach Kündigung des Pakets bleibt der Account bestehen, die API nimmt aber keine Requests mehr an. Der Kunde kann jederzeit ein neues Paket bestellen. Die Kündigung des Accounts selbst und die Löschung der Daten richten sich nach § 5 AGB und Ziffer 10.

9.6 Preisänderungen eines Pakets gelten für bestehende Verträge ab dem nächsten Abrechnungszeitraum nach Ankündigung. profihost kündigt sie mindestens 6 Wochen vorher in Textform an.

10. Datenverarbeitung und Standort

10.1 Standort: Die Verarbeitung von Eingaben und Ausgaben findet ausschließlich in Rechenzentren in der Europäischen Union statt, derzeit in Deutschland und Finnland. API-Gateway, Portal und Datenbank betreibt profihost auf eigener Cloud-Infrastruktur in Deutschland. Die GPU-Server, auf denen die Modelle laufen, mietet profihost bei der Hetzner Online GmbH (Rechenzentren in Falkenstein, Deutschland, und Helsinki, Finnland) als Unterauftragsverarbeiter. Eine Anfrage kann an jedem dieser Standorte verarbeitet werden.

10.2 Eingaben und Ausgaben: profihost verarbeitet Prompts, Bilder und Modellantworten nur, um den jeweiligen Request zu beantworten. profihost speichert diese Inhalte nicht, auch nicht in Protokollen (Logs), und verwendet sie nicht zum Training oder zur Verbesserung von Modellen. Für Missbrauchserkennung wertet profihost Metadaten aus (Ziffer 10.3), nicht die Inhalte.

10.3 Metadaten: Für Abrechnung, Limits, Missbrauchserkennung und Betrieb speichert profihost pro Request Zeitpunkt, Account, Nutzer, API-Key, Modell, Token-Anzahl, Dauer, Status (auch bei abgelehnten Requests), IP-Adresse und User-Agent des Aufrufers. Aufbewahrung: Nutzungsdaten bis zum Ablauf der gesetzlichen Aufbewahrungsfristen für Abrechnungsunterlagen; IP-Adresse und User-Agent werden nach 90 Tagen gelöscht. Server-Protokolle werden nach wenigen Tagen überschrieben.

10.4 Playground im Portal: Der Playground dient zum Ausprobieren der Modelle. Er sendet Requests über die API wie eine eigene Anwendung und verbraucht Credits. Verläufe speichert profihost nicht.

10.5 Die Vereinbarung zur Auftragsverarbeitung (AVV) nach Art. 28 DSGVO ist in allen Paketen enthalten. Der Kunde schließt sie im Portal oder auf Anfrage ab.

10.6 Nach Vertragsende löscht profihost API-Keys, Nutzer und Accountdaten innerhalb von 30 Tagen, soweit keine gesetzlichen Aufbewahrungspflichten entgegenstehen (§ 11 AGB).

11. Verfügbarkeit, Wartung und Support

11.1 profihost betreibt den pro Model Hub mit der Sorgfalt eines professionellen Hosting-Anbieters, sichert aber keine bestimmte Verfügbarkeit, Antwortzeit oder Durchsatzleistung zu (§ 9 AGB). Antwortzeiten hängen von Modell, Länge von Eingabe und Ausgabe und der Gesamtlast ab.

11.2 Wartung: Geplante Wartungsarbeiten kündigt profihost nach Möglichkeit mindestens 2 Werktage vorher im Portal oder in der Dokumentation an. Kurzfristige Arbeiten zur Abwehr von Sicherheitsrisiken sind ohne Ankündigung zulässig. Vorübergehende Fehler (Status `5xx`) soll die Anwendung des Kunden mit Wiederholungen abfangen, wie in der Dokumentation beschrieben.

11.3 Support: Alle Pakete erhalten Zugang zur Dokumentation. Pro, Business und Business Plus erhalten zusätzlich Support per E-Mail an modelhub-support@profihost.com montags bis freitags von 8 bis 18 Uhr, ausgenommen gesetzliche Feiertage in Niedersachsen (Supportzeiten). Support umfasst Fragen zur Nutzung der API, des Portals und zur Abrechnung, nicht die Entwicklung oder Fehlersuche in Anwendungen des Kunden.

11.4 Reaktionszeit: profihost strebt an, Supportanfragen innerhalb eines Werktags zu beantworten (Zielwert). Die Frist läuft nur während der Supportzeiten. Eine Anfrage, die außerhalb der Supportzeiten eingeht, gilt als zu Beginn der nächsten Supportzeit eingegangen.

Reaktion bedeutet eine erste Antwort durch einen Mitarbeiter, nicht die Lösung des Anliegens. Der Zielwert ist keine zugesicherte Eigenschaft, eine Überschreitung begründet keine Ansprüche des Kunden.

12. Nicht enthalten

- eine zugesicherte Verfügbarkeit (SLA) oder garantierte Antwortzeiten,
- Training oder Fine-Tuning von Modellen, eigene Modelle des Kunden,
- Speicherung von Daten, Dateien oder Gesprächsverläufen für den Kunden,
- Abruf von Bildern oder anderen Inhalten über URLs,
- dedizierte GPU-Kapazität (Paket „Dedicated“, nur nach gesonderter Vereinbarung).

13. Änderungen dieser Leistungsbeschreibung

profihost kann diese Leistungsbeschreibung ändern, soweit dies aus technischen, rechtlichen oder lizenzrechtlichen Gründen erforderlich ist oder die Leistung verbessert. Änderungen, die den Kunden nicht nur unerheblich benachteiligen, kündigt profihost mindestens 6 Wochen vorher in Textform an. Sie werden frühestens zum nächsten Abrechnungszeitraum des Kunden wirksam. Für Modelle und Credit-Faktoren gilt Ziffer 5.

Anlage 1: Pakete

Alle Preise netto zuzüglich der gesetzlichen Umsatzsteuer.

	Starter	Pro	Business	Business Plus
Preis pro Monat	7 €	35 €	129 €	249 €
Credits pro Monat	5 Mio.	80 Mio.	320 Mio.	650 Mio.
Requests pro Minute	30	60	150	150
parallele Requests	5	10	20	20
Embedding-Token pro Minute	100.000	300.000	1 Mio.	1 Mio.
Alle Modelle	✓	✓	✓	✓
AVV	✓	✓	✓	✓
Support per E-Mail	–	✓	✓	✓
Reaktionszeit Support (Zielwert)	–	1 Werktag	1 Werktag	1 Werktag
30 Tage kostenlos testen	✓	–	–	–

Anlage 2: Modelle

Stand 20.10.2026. Kontextfenster und Fähigkeiten tagesaktuell unter `https://docs.promodelhub.de/modelle`.

Modell	Einsatz	Credit-Faktor	Bildeingabe	Tool-Calling	Reasoning
Gemma 4 26B-A4B	Allzweck, Bildverständnis, lange Dokumente	1×	✓ (Beta)	✓	✓ (optional)
Qwen3.6 35B-A3B	Programmieren, Agenten	1×	✓ (Beta)	✓	✓ (standardmäßig an)
Qwen3.8 27B	Reasoning, mehrstufige Analysen	1,5×	✓ (Beta)	✓	✓ (standardmäßig an)
Gemma 4 31B	Premium für anspruchsvolle Texte und Zusammenfassungen	2×	✓ (Beta)	✓	✓ (optional, standardmäßig aus)
Qwen3-VL-Embedding 8B	Embeddings für Suche und RAG (Beta, nur Text)	0,25×	–	–	–
Qwen3-VL-Reranker 2B	Suchtreffer nach Relevanz sortieren, für RAG (Beta, nur Text)	0,25×	–	–	–