Anlage 4 – Leistungsverzeichnis

Bereitstellung einer flexiblen KI-Compute-, Entwicklungs- und Integrationsplattform

Vergabeart: Offenes Verfahren nach VgV | **Auftraggeber:** Universitätsklinikum Frankfurt (UKF)

Leistungszeitraum: 36 Monate ab vertragsgemäßer Bereitstellung gemäß verbindlicher Bereitstellungsbestätigung des Auftragnehmers; die Bereitstellung muss spätestens bis 15.12.2026 erfolgen.

Verbindliche Gesamtvergütung: 1.100.000,00 EUR brutto. Kein Preiswettbewerb; Wettbewerb über verbindlichen Leistungsumfang und objektiv messbare Qualitätsmerkmale.

> **Fachlich-technischer Hinweis:** Die Unterlage beschreibt den Leistungsgegenstand und die technische Bewertungslogik. Die finale rechtliche, fördermittelrechtliche, datenschutzrechtliche und informationssicherheitsbezogene Freigabe erfolgt durch den Auftraggeber und seine juristische Begleitung.

1. Gegenstand, Zielsetzung und Beschaffungsmodell

Gegenstand der Beschaffung ist die zeitlich befristete Bereitstellung und Nutzung einer mandantenfähigen, skalierbaren KI-Compute-, Entwicklungs- und Integrationsplattform für KI- und rechenintensive Anwendungen. Die Plattform soll ausdrücklich Use-Case-offen sein und sowohl acceleratorbasierte KI-Workloads als auch CPU-/RAM-intensive Entwicklungs-, Integrations-, VM-, Container- und Kubernetes-Workloads abbilden.

Der Leistungsumfang umfasst ein verbindlich garantiertes, über die Vertragslaufzeit flexibel nutzbares wirtschaftliches Compute-Unit-Kontingent (CU), die verpflichtenden Accelerator-Capability-Klassen G1 bis G3, optional angebotene Zukunfts- und Multi-Accelerator-Klassen G4 und G5, einen eigenständigen Platform-Runtime-Pool, mindestens 100 TB persistenten Speicher, Netz- und Plattformfunktionen, Sicherheits-, Betriebs-, Support-, Dokumentations-, Exit- und Transition-Leistungen.

Die vertragliche Gesamtvergütung beträgt verbindlich 1.100.000,00 EUR brutto. Ein Preiswettbewerb findet nicht statt. Der Wettbewerb erfolgt über den für diesen Festpreis verbindlich eingeräumten Leistungsumfang und die in Anlage 2 vollständig veröffentlichten Qualitätsmerkmale.

Die Rechnungsstellung erfolgt nach verbindlicher Bereitstellungsbestätigung des Auftragnehmers. Mit der vertragsgemäßen Bereitstellung werden der 36-monatige Leistungsanspruch, das angebotene CU-Gesamtkontingent sowie sämtliche Mindest- und angebotenen Mehrleistungen verbindlich eingeräumt. Bereitstellung und Rechnungsstellung müssen bis spätestens zum 15.12.2026 erfolgen, damit der vollständige Mittelabfluss spätestens bis zum 31.12.2026 gewährleistet ist. Es handelt sich nicht um eine Rahmenvereinbarung mit später vergüteten Einzelabrufen und nicht um ein Geld- oder Cloud-Guthaben.

1.1 Trennung der Vergabeverfahren und Vermeidung wirtschaftlicher Interessenkonflikte

Der Auftraggeber führt neben dieser Beschaffung die Beschaffung „begleitende Beratungs-, Architektur-, Governance-, Security-, Audit- und Providersteuerungsleistungen für die KI-Compute-, Entwicklungs- und Integrationsplattform“ als eigenständiges Vergabeverfahren durch.

Zwischen beiden Leistungsbereichen besteht eine unmittelbare fachliche und wirtschaftliche Abhängigkeit. Insbesondere können die im Rahmen der KI-Beratungsleistungen erarbeiteten Empfehlungen, technischen Konzepte und Bedarfsbewertungen maßgeblichen Einfluss auf Art, Umfang und Abruf der KI-Compute-Leistungen nehmen.

Zur Sicherstellung einer unabhängigen, objektiven und wirtschaftlich neutralen Beratung sowie zur Vermeidung potenzieller Interessenkonflikte sieht der Auftraggeber eine personelle und wirtschaftliche Trennung der Auftragnehmer beider Leistungsbereiche vor.

Eine gleichzeitige Beteiligung desselben Wirtschaftsteilnehmers an beiden Vergabeverfahren ist daher nicht zugelassen. Dies gilt entsprechend für Bietergemeinschaften sowie für wirtschaftlich verbundene Unternehmen, soweit aufgrund ihrer konkreten Verflechtungen eine unabhängige Leistungserbringung nicht gewährleistet werden kann. Darüber hinaus gilt dies auch für Nachunternehmer.

Mit der Angebotsabgabe erklärt der Bieter, dass er die vorgenannten Anforderungen einhält und keine Beteiligung am Vergabeverfahren „begleitende Beratungs-, Architektur-, Governance-, Security-, Audit- und Providersteuerungsleistungen für die KI-Compute-, Entwicklungs- und Integrationsplattform“ besteht, die zu einem unzulässigen Interessenkonflikt führen kann.

Der Auftraggeber behält sich vor, mögliche Interessenkonflikte unter Beachtung der vergaberechtlichen Grundsätze der Transparenz, Gleichbehandlung und Verhältnismäßigkeit zu prüfen und erforderliche Maßnahmen zu ergreifen.

2. Bewertungs- und Modellgrundsatz: wirtschaftliche CU plus technischer Benchmark

Das Beschaffungsmodell trennt bewusst zwei unterschiedliche Größen: den wirtschaftlichen Ressourcenwert und die technische Leistungsfähigkeit. Diese Trennung verhindert, dass reine Benchmarkleistung automatisch mit Wirtschaftlichkeit gleichgesetzt wird, und ermöglicht zugleich eine objektive technische Vergleichbarkeit unterschiedlicher Accelerator-Generationen und Hersteller.

Ebene	Funktion	Wirkung
Capability G1-G5	Technische Mindestmerkmale	Definiert, welche Fähigkeiten eine Ressource mindestens besitzt.
Compute Unit (CU)	Wirtschaftliche Verbrauchseinheit	Definiert den verbindlichen Verbrauch pro Nutzungsstunde anhand fester, vorab veröffentlichter Klassenfaktoren.
UKF Reference Benchmark (UKF-RB)	Technische Leistungsbewertung	Misst die tatsächlich angebotene Rechenleistung innerhalb der Capability-Klassen; verändert den CU-Verbrauch nicht.
Anlage 2-Zuschlagswertung	Wirtschaftlichkeitsbewertung	Bewertet CU-Gesamtkontingent und technische Effizienz getrennt sowie Runtime, Plattform, Exit und SLA.

Eine leistungsfähigere Ressource innerhalb derselben Capability-Klasse verursacht daher nicht automatisch einen höheren CU-Verbrauch. Technischer Fortschritt innerhalb einer Klasse kommt dem UKF zugute. Der Benchmark dient der technischen Leistungsbewertung, Plausibilisierung und technischen Verifikation der Bereitstellung, nicht der Ermittlung eines dynamischen Verbrauchsfaktors.

3. Vertragslaufzeit, Leistungsbeginn und Verfügbarkeit

Der Leistungszeitraum beträgt 36 Monate und beginnt mit dem Datum der verbindlichen Bereitstellungsbestätigung des Auftragnehmers über die vertragsgemäße Bereitstellung. Die Bereitstellung und Rechnungsstellung muss spätestens bis zum 15.12.2026 erfolgen. G1 bis G3 sowie der Platform-Runtime-Pool müssen während des gesamten Leistungszeitraums grundsätzlich bestell- und nutzbar sein. Soweit G4 und/oder G5 angeboten und gewertet werden, gelten die zugesagten Leistungs- und Verfügbarkeitswerte ebenfalls für die gesamte Laufzeit.

Die Ressourcen müssen nicht exklusiv physisch für UKF reserviert sein. Der Auftragnehmer trägt jedoch das Kapazitäts- und Beschaffungsrisiko dafür, die zugesagten Parallelkapazitäten, Bereitstellungszeiten, Runtime-Kapazitäten und CU-Einlösbarkeit über die gesamte Laufzeit sicherzustellen.

4. Herleitung und Definition der Capability-Klassen G1 bis G5

Die Klassen bilden typische und perspektivische Nutzungsszenarien eines Universitätsklinikums ab. G1 adressiert Entwicklungs-, Inferenz-, RAG- und Embedding-Workloads; G2 anspruchsvollere Inferenz-, Sprach- und Bildmodelle; G3 bildet den erwarteten Kernbereich größerer LLM-, Bildgebungs-, Pathologie- und Forschungsworkloads. G4 und G5 stellen Zukunfts- und Hochleistungsfähigkeiten für speicherintensive multimodale Modelle bzw. gekoppelte Multi-Accelerator-Workloads dar.

Klasse	Accelerator-Mindestmerkmale	vCPU min.	RAM min.	Scratch min.	Status	CU/h
G1	>=24 GiB Speicher; >=300 GB/s Speicherbandbreite	8	32 GiB	250 GiB	MUSS	1
G2	>=44 GiB Speicher; >=800 GB/s Speicherbandbreite	16	64 GiB	500 GiB	MUSS	3
G3	>=80 GiB Speicher; >=1,5 TB/s Speicherbandbreite	32	256 GiB	750 GiB	MUSS	4
G4	>=180 GiB Speicher; >=5 TB/s Speicherbandbreite	48	384 GiB	1.500 GiB	WERTUNG / optional	12
G5	>=8 gekoppelte Acceleratoren, jeder mindestens G3; Hochgeschwindigkeits-P2P	128	1.024 GiB	4.000 GiB	WERTUNG / optional	32

G1 bis G4 beschreiben jeweils eine logische Accelerator-Ressource. Die geforderte Speicherkapazität darf nicht durch bloße Addition mehrerer schwächerer, nicht als einheitliche logische Ressource nutzbarer Acceleratoren erreicht werden. Fractional-/partitionierte Ressourcen sind zulässig, sofern Isolation, garantierte Mindestressourcen, Speicherbandbreite und Nicht-Overcommit nachgewiesen werden.

G5 ist als gekoppeltes System aus mindestens acht Acceleratoren zu verstehen. Der technische Benchmark wird für das angebotene Gesamtsystem ausgeführt; eine bloße Multiplikation eines Einzel-Accelerator-Benchmarkwertes ist unzulässig.

5. Wirtschaftliches CU-Modell

Die Compute Unit (CU) ist eine dimensionslose wirtschaftliche Verbrauchseinheit. Sie ist weder Zahlungsmittel noch Gutschein noch dynamische Preisgröße. Der CU-Verbrauch eines bereitgestellten Accelerator-Profils richtet sich ausschließlich nach der vom Auftraggeber vorgegebenen Capability-Klasse.

Capability	verbindlicher wirtschaftlicher Faktor	Verbrauch
G1	1	1 CU je bereitgestellter Stunde
G2	3	3 CU je Stunde
G3	4	4 CU je Stunde
G4	12	12 CU je Stunde
G5	32	32 CU je Stunde

Die Faktoren stellen eine vorab veröffentlichte wirtschaftliche Normierung der Ressourcenklassen dar. Sie werden während der Vertragslaufzeit nicht durch bietereigene Credit-, Tarif- oder Benchmarklogiken verändert. Das angebotene CU-Gesamtkontingent ist einheitlich in allen verpflichtenden Klassen G1-G3 und, soweit angeboten, in G4/G5 nutzbar. Es existieren keine klassenbezogenen Teilkontingente, Jahres- oder Monatsbudgets, zusätzlichen Credits, Mindeststreckungen oder versteckten Verbrauchsregeln.

Die Abrechnung erfolgt maximal minutengenau; feinere Abrechnung ist zulässig. Reservierte, aber noch nicht technisch bereitgestellte Ressourcen verbrauchen keine CU. Für bereitgestellte und vom UKF bewusst laufend gehaltene Ressourcen ist CU-Verbrauch zulässig, auch wenn ein Workload zeitweise keine Rechenlast erzeugt.

> **Wirtschaftlichkeitslogik:** Eine höhere Benchmarkleistung innerhalb einer Klasse erhöht den CU-Verbrauch nicht. Dadurch wird technische Effizienz als Mehrwert für UKF sichtbar, statt durch einen automatisch höheren Tokenverbrauch neutralisiert zu werden.

6. Referenznutzung und Zukunftsreserve

Für die fachliche Herleitung wird von einer Referenzgewichtung G1 15 %, G2 25 %, G3 35 %, G4 20 % und G5 5 % ausgegangen. Diese Gewichtung beschreibt weder Stundenquoten noch Abnahmeverpflichtungen. Für die technische Benchmarkwertung der verpflichtenden Klassen G1-G3 werden die relativen Gewichte 15:25:35 proportional auf 150

Punkte übertragen (30/50/70 Punkte in Anlage 2). G4/G5 bleiben optionale Zukunfts-Capabilities und werden in K2 bewertet.

Die tatsächliche Nutzung darf während der gesamten Laufzeit vollständig von der Referenzannahme abweichen. Das CU-Kontingent bleibt innerhalb der angebotenen Capability-Klassen frei einsetzbar. Zukunftsreserve entsteht durch freie CU-Nutzung, Technologiefortschritt und die Möglichkeit zur Aufnahme neuer Ressourcenprofile; nicht durch starre Jahreskontingente.

7. UKF Reference Benchmark und technische Effizienz

Der UKF-RB nach Anlage 4b misst die technische Leistungsfähigkeit angebotener Accelerator-Profile in zwei gleichberechtigten Dimensionen: LLM-/Generative-AI-Inferenz und Tensor-/Medical-Imaging-Inferenz. Die Benchmarkwerte werden ausschließlich zur technischen Wertung, Plausibilisierung, Profilklassifizierung und technischen Verifikation der Bereitstellung verwendet. Der wirtschaftliche CU/h-Verbrauch bleibt unverändert.

Für die Zuschlagswertung werden die Rohwerte jeder verpflichtenden Capability G1-G3 getrennt und dimensionsbezogen mit dem jeweils höchsten wertbaren Bieterwert verglichen. Dadurch erhält UKF einen transparenten Effizienznachweis, ohne technische Performance und wirtschaftliche Verbrauchseinheit zu vermischen. Details enthält Anlage 2; Messregeln enthält Anlage 4b.

8. Technologiefortschritt und neue Ressourcenprofile

Der Auftragnehmer darf während der Laufzeit neuere oder leistungsfähigere Hardware einsetzen. Neue oder geänderte Profile sind vor Nutzung technisch zu dokumentieren, einer Capability-Klasse zuzuordnen und nach Anlage 4b zu benchmarken. Ein vom Auftragnehmer veranlasster Hardwaretausch darf UKF weder hinsichtlich Mindestmerkmalen, Verfügbarkeit noch CU-Verbrauch schlechter stellen. Ein leistungsfähigeres Ersatzprofil innerhalb derselben Klasse bleibt beim gleichen CU/h-Faktor.

9. Persistenter Speicher

Im Festpreis enthalten sind mindestens 100 TB nutzbare persistente Speicherkapazität über die gesamte Vertragslaufzeit; 1 TB = 10^12 Byte. Mindestens 10 TB müssen unmittelbar durch VM-/Container-/Kubernetes-Workloads als Block- oder File-Storage mountbar sein. Die übrige Mindestkapazität kann als Object-, Block- oder File-Storage oder als Kombination bereitgestellt werden. Die Mindestkapazität verbraucht keine CU.

Zusätzlicher persistenter Storage oberhalb 100 TB wird nach Anlage 2 als Mehrleistung bis zu einem Wertungsziel von 200 TB Gesamtvolumen berücksichtigt. Die Plattform ist nicht als klinisches Primärarchiv oder Ersatz für PACS/klinische Langzeitarchive definiert.

10. Datenübertragung und Netzwerk

- Sämtlicher Datenverkehr innerhalb der Plattform sowie zwischen Accelerator-, Runtime- und Storage-Ressourcen ist im Festpreis enthalten und verbraucht keine CU.
- Anbindung des UKF mindestens über verschlüsselte IP-basierte Standortkopplung (Site-to-Site-VPN oder funktional gleichwertig).
- Plattformseitiger Übergabepunkt mindestens 10 Gbit/s; die End-to-End-Datenrate kann durch die UKF-Anbindung begrenzt sein.
- Ingress und Egress zwischen UKF und Plattform über die vereinbarten Zugangswege ohne volumenabhängige Zusatzentgelte.
- Kontrollierte ausgehende HTTPS-/API-Verbindungen und private/verschlüsselte Partnerkopplungen müssen technisch möglich sein. Drittanbieter-Lizenz- oder API-Entgelte sind nur dann Bestandteil, wenn ausdrücklich angeboten.

11. Parallelität, Bereitstellung und Platform-Runtime

Klasse	Mindestparallelität	max. Bereitstellungszeit	Burst
G1	4	30 Minuten	8 innerhalb 5 Arbeitstagen
G2	2	1 Stunde	4 innerhalb 5 Arbeitstagen

Klasse	Mindestparallelität	max. Bereitstellungszeit	Burst
G3	2	2 Stunden	4 innerhalb 5 Arbeitstagen
G4	optional / Bieterangabe	Bieterangabe	wertbar
G5	optional / Bieterangabe	Bieterangabe	wertbar

11.1 Eigenständiger Platform-Runtime-Pool

Unabhängig von Accelerator-Ressourcen muss ein CPU-/RAM-Runtime-Pool für VM-, Container-, Kubernetes-, Gateway-, API-, Datenbank- und Drittanbieter-Workloads bereitstehen. Gleichzeitig müssen mindestens 164 vCPU und 1.104 GiB RAM ohne Zusatzentgelt nutzbar sein. Die Mindestkapazität darf 24x7 oder intermittierend genutzt werden, verbraucht keine CU und muss innerhalb von höchstens vier Stunden provisionierbar sein. Höhere verbindliche Runtime-Kapazitäten werden nach Anlage 2 gedeckelt gewertet.

12. Plattform, Self-Service, Sandbox und Integration

- Webbasiertes Managementportal mit Rollen-, Projekt- und Mandantenbezug.
- Dokumentierte API zur Provisionierung und Verwaltung.
- Start, Stop, Löschung, Statusabfrage, Storage- und Netzzuordnung.
- Eigene und bereitgestellte VM-/AI-Images; OCI-kompatible Container; Kubernetes-Eignung.
- Unterstützung marktüblicher AI-/ML-Frameworks, insbesondere PyTorch, TensorFlow, ONNX Runtime und Hugging-Face-kompatibler Workloads oder funktional gleichwertig.
- Logisch isolierte Entwicklungs-, Test-, Pilot- und Projektumgebungen mit getrennten Rollen-/IAM-, Storage- und Netzkontexten.
- Betrieb autorisierter Drittanbieter-VM-Images/Container, kontrollierte externe HTTPS-/API-Kommunikation und private Partneranbindungen.
- Model-Serving-Endpunkte; OpenAI-kompatible oder funktional gleichwertige Schnittstelle als wertbare Mehrleistung.

12.1 Verständliches Ressourcen- und Verbrauchsmodell

Der Self-Service-Katalog zeigt je Accelerator-Profil mindestens Capability-Klasse, technische Bezeichnung, Accelerator-Speicher, verbindlichen CU/h-Faktor der Klasse, Benchmark-Rohwerte bzw. technischen Effizienzindikator, aktuelle Verfügbarkeit und maximalen CU-Verbrauch für eine vom Nutzer eingegebene Laufzeit. Endanwender müssen keine Benchmarkformeln berechnen.

Projektverantwortliche müssen Projektbudgets und Laufzeitlimits festlegen können. Eine Ressource darf bei einem ausdrücklich konfigurierten Projektlimit nicht automatisch kontingentwirksam weiterlaufen, sofern UKF keine Fortführung autorisiert.

13. Ressourcen-, Verbrauchs- und Kontingentmanagement

Jede Belastung des CU-Kontingents muss aus den Nutzungsdaten rechnerisch reproduzierbar sein. Der Auftragnehmer stellt eine aktuelle und prüffähige Dokumentation über die gesamte Laufzeit bereit.

Pflichtinformation	Mindestinhalt
CU-Kontostand	ursprüngliches, verbrauchtes und verbleibendes CU-Gesamtkontingent
Einzelnutzung	Capability, SKU/Profil, Start, Ende, Dauer, CU/h-Faktor, CU-Verbrauch, Projekt
Technik	Accelerator, vCPU, RAM, Scratch, Benchmark-Rohwerte/Benchmark-ID
Runtime	CPU-/RAM-Ressourcen, Start/Ende, Projekt, Gesamtauslastung
Storage	belegte persistente Kapazität und Verlauf
Historie	36 Monate plus mindestens 12 Monate nach Vertragsende oder vollständige Übergabe

Pflichtinformation	Mindestinhalt
Export	mindestens CSV plus dokumentierte API
Monatsnachweis	Anfangsbestand, Verbrauch je Capability, Korrekturen, Endbestand, archivierbarer Export
Ressourcenkatalog	Capability, SKU, Speicher, CU/h, Benchmarkwerte, Verfügbarkeit
Prognose	vor Start maximaler CU-Verbrauch; nach Start laufender Verbrauch und Restbudget

Warnschwellen für den Gesamtverbrauch sind mindestens bei 75 %, 90 % und 95 % vorzusehen. Erweiterte Prognosen, frei konfigurierbare Warnungen, Projektbudgets und erweiterte Audit-/Verbrauchs-APIs werden nach Anlage 2 bewertet.

14. IAM, Datenschutz, Informationssicherheit und digitale Souveränität

- RBAC, Projekt-/Mandantentrennung, MFA für privilegierte Zugriffe, SAML 2.0 und/oder OpenID Connect.
- Personalisierte administrative Konten; privilegierte Zugriffe des Auftragnehmers sind zu minimieren und zu protokollieren.
- Verschlüsselung in Transit und at Rest; dokumentiertes Schlüsselmanagement.
- ISO/IEC 27001 für den relevanten Scope oder vergaberechtlich zulässiger gleichwertiger Nachweis.
- Datenverarbeitung und -speicherung produktiver Kundendaten innerhalb EU/EWR; Subprozessoren transparent.
- Für G1-G3 ist ein aktueller C5-Nachweis bzw. ein gesetzlich zulässiger gleichwertiger/höherwertiger Sicherheitsnachweis verpflichtend. Für G4/G5 gilt dies einsatz- und einzelfallabhängig; der Auftragnehmer muss sicherstellen, dass gesetzlich erforderliche C5-Anforderungen vor der jeweiligen Nutzung erfüllt und nachgewiesen werden können, soweit sie für den konkreten Einsatz verpflichtend sind.
- Keine Nutzung von UKF-Daten, Prompts, Modellen oder Trainingsdaten für eigenes oder fremdes Modelltraining bzw. Produktverbesserung ohne ausdrückliche gesonderte Zustimmung.

15. Logging, Monitoring, Backup und Wiederherstellung

- Logging von Authentifizierungen, administrativen Aktionen, Berechtigungsänderungen, Provisionierungen und sicherheitsrelevanten Ereignissen.
- Automatisierbarer Export in das SIEM/SOC des Auftraggebers.
- Online-Aufbewahrung relevanter Audit-/Security-Logs mindestens 180 Tage oder gleichwertige Exportmöglichkeit.
- Snapshot-/Versionierungs- und Wiederherstellungsfunktionen für persistenten Storage; RPO/RTO sind anzugeben.
- Compute-Instanzen müssen nicht klassisch gesichert werden, sofern Konfiguration, Images und persistente Daten getrennt gesichert/exportiert werden können.

16. Service Level und Support

Leistung	Mindestanforderung
Management-/Plattformebene	monatliche Verfügbarkeit >=99,7 %
Persistenter Storage	monatliche Verfügbarkeit >=99,7 %
Bezugszeitraum	Monatlich, bezogen auf die definierte Betriebszeit
Betriebszeit und Verfügbarkeit	Die geschuldete Betriebszeit beträgt Montag bis Sonntag, jeweils 00:00 bis 24:00 Uhr (24 × 7).
Reguläre Servicezeit	Mo–Fr, z. B. 07:00–19:00 Uhr (ausgenommen der gesetzlichen Feiertage in Hessen).
P1-Support	Durchgängige Störungsannahme und -bearbeitung an sieben Tagen pro Woche, 24 Stunden täglich. Reaktion ≤ 2 h, Wiederherstellung ≤ 8 h. Außerhalb der Servicezeit Bearbeitung über Bereitschafts-/Notfallservice nach vereinbarten Konditionen.

Leistung	Mindestanforderung
P2-Support	Reaktion ≤ 4 h, Wiederherstellung ≤ 12 h, gerechnet innerhalb der regulären Servicezeit. Die Wiederherstellungszeit bezieht sich auf die Wiederherstellung der betroffenen vertraglich geschuldeten Plattform- bzw. Infrastrukturfunktion. Die vollständige Wiederherstellung sämtlicher kundenseitig gespeicherter Datenbestände ist hiervon nicht umfasst.
P3-Support	Reaktion ≤ 8 h, gerechnet innerhalb der regulären Servicezeit. Wiederherstellung nach Vereinbarung bzw. Best Effort.
P4-Support	Nächster Arbeitstag bzw. Best Effort.
Sprache	Deutsch oder Englisch.
Provisioning	gemäß Kapitel 11.
Service außerhalb der regulären Zeiten	Störungsmeldung 24/7 möglich. ~~Außerhalb der regulären Servicezeit, insbesondere nachts, an Wochenenden und Feiertagen, optionaler Bereitschafts-/Notfallservice nach gesonderten Konditionen bzw. Mehrkosten.~~
Wochenenden / Feiertage	Keine reguläre SLA-Bearbeitung, aber 24/7-Meldemöglichkeit. ~~optional abrufbarer Bereitschaftsservice, insbesondere für P1-Störungen. Konditionen separat im Preisblatt ausweisbar.~~
Geplante Wartungsfenster	Für geplante Nichtverfügbarkeit gilt die Regelung nach Ziffer 8.3 EVB-IT Cloud-AGB: sonntags von 04:00 bis 08:00 Uhr. Wartungsmaßnahmen mit Auswirkungen auf die Verfügbarkeit sind vorher anzukündigen.
Maximale Wiederherstellungszeit - Recovery Time Objective (RTO)	8–12 h
Maximal tolerierbarer Datenverlust in Zeit - Recovery Point Objective (RPO)	Je nach Datenart, z. B. max. 24 h. Für persistente Daten, die im Rahmen des vertraglich geschuldeten Datensicherungs- und Wiederherstellungsdienstes geschützt werden, beträgt der maximal zulässige Datenverlust 24 Stunden. Temporäre Scratch-Daten und nicht gesicherte laufende Berechnungszustände sind hiervon ausgenommen.
Monitoring / Eskalation	Standardmonitoring; definierter Eskalationsweg für P1, außerhalb der Servicezeit über Bereitschaftsdienst
Service Level Agreement (SLA) – Messung	Innerhalb der vereinbarten Betriebszeit; geplante und angekündigte Wartungsfenster werden entsprechend der Vertragsdefinition behandelt

Der Auftragnehmer muss die Möglichkeit einer Störungsmeldung außerhalb der regulären Servicezeiten einschließlich Wochenenden und Feiertagen gewährleisten. ~~Für kritische Störungen ist ein optional abrufbarer Bereitschaftsservice vorzusehen. Hierfür anfallende zusätzliche Kosten sind im Preisblatt gesondert auszuweisen.~~

Höhere garantierte Verfügbarkeit bzw. kürzere Reaktionszeiten werden ausschließlich nach den veröffentlichten Regeln aus Anlage 2 bewertet. Messung, Wartungsfenster, Ausschlüsse und Vertragsfolgen sind im Vertragswerk eindeutig zu definieren.

17. Vertrags- und Abrechnungsgrundsätze

- Verbindliche Gesamtvergütung 1.100.000,00 EUR brutto für den vollständigen 36-Monats-Leistungsanspruch.
- Keine zusätzlichen Entgelte für Mindestleistungen und angebotene Wertungsmehrleistungen innerhalb des Plattformumfangs.
- Keine zusätzlichen Entgelte für CPU/RAM/Scratch der G-Klassen, 100 TB Mindeststorage, Runtime-Mindestpool, interne Datenübertragung, Portal/API, Monitoring, Standard-Support und Exit-Basispaket.
- Keine CU-Verfallsregeln vor Ablauf der 36 Monate und keine monatlichen/vierteljährlichen Verbrauchsobergrenzen.
- Der wirtschaftliche CU/h-Faktor je Capability bleibt während der gesamten Laufzeit unverändert.

- Der Auftragnehmer darf den Leistungsanspruch nicht durch eigene AGB, Fair-Use-Regeln, externe Credits oder abweichende Preislisten einschränken.

18. Exit-, Transition- und Portabilitätsmanagement

Ziel ist entweder die technisch unterbrechungsarme Fortführung auf vergaberechtlich zulässiger Grundlage oder die vollständige Übergabe an UKF bzw. einen benannten Nachfolgeauftragnehmer. Aus der Regelung entsteht kein Anspruch auf Vertragsverlängerung.

- Export persistenter Kundendaten, Datasets, Modelle, Checkpoints, Datenbanken, Object-/Block-/File-Storage-Inhalte.
- Export von VM-Images, virtuellen Disk-Formaten, OCI-Container-Images, Registry-Inhalten, Kubernetes-Manifesten, Helm-Charts und vorhandenen IaC-Definitionen.
- Export AI-spezifischer Artefakte, Modellkonfigurationen, Fine-Tuning-Ergebnisse, Embeddings/Vector-Store-Inhalte soweit technisch exportierbar.
- Dokumentierte Übergabe von Netzwerk-, Security-, Rollen- und Berechtigungskonfigurationen, Monitoring-, Audit- und Betriebslogs sowie vollständiger CU-Verbrauchshistorie.
- Auf Anforderung spätestens sechs Monate vor Ende Exit-/Fortführungsworkshop; bei Migration spätestens drei Monate vor Ende abgestimmter Exit-Plan.
- Mindestens 10 Personentage technische Migrationsunterstützung im Festpreis; zusätzliche PT können gewertet werden.
- Kundendaten werden erst nach erfolgreicher Übergabe und ausdrücklicher Freigabe gelöscht; anschließend Lösch-/Vernichtungsnachweis einschließlich Replikate/Backups nach Löschfristen.
- Exit-Datenexport darf im Mindestumfang keine mengenabhängigen Egress-Gebühren auslösen.

19. Vertragsgemäße Bereitstellung

Vor Beginn des 36-monatigen Leistungszeitraums bestätigt der Auftragnehmer verbindlich die vollständige vertragsgemäße Bereitstellung der geschuldeten Plattform, des angebotenen CU-Gesamtkontingents sowie sämtlicher Mindest- und angebotenen Mehrleistungen. Die Bestätigung der vertragsgemäßen Bereitstellung muss bis zum 15.12.2026 erfolgen. Eine förmliche technische Abnahme ist nicht geschuldet. Mit der Bereitstellungsbestätigung bestätigt der Auftragnehmer insbesondere die Betriebsbereitschaft von SSO/Zugriff, Provisionierung G1-G3 und angebotener G4/G5, Ausführung bzw. Verifikation des UKF-RB nach Anlage 4b, CU-Zählung und -Reporting, Runtime-Pool, 100 TB Storage einschließlich mindestens 10 TB Block/File, 10-Gbit/s-Übergabepunkt, Sandbox-/Projekttrennung, Partner-/API-Kopplung, Export, Monitoring, Logging sowie die Möglichkeit zur Erstellung eines Monats-/Verbrauchsnachweises.

Bei reproduzierbaren Benchmarkabweichungen gegenüber verbindlichen Angebotswerten gelten die Regeln aus Anlage 4b; der Auftragnehmer muss die zugesagte technische Leistung herstellen oder eine mindestens gleichwertige technische Lösung innerhalb derselben wirtschaftlichen CU-Klasse bereitstellen.