Anlage 4a – Kriterienkatalog

MUSS-Anforderungen, Bieterangaben und Nachweise - Hybridmodell CU + Benchmark

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.

Der Kriterienkatalog ist vollständig auszufüllen. MUSS-Anforderungen sind Mindestanforderungen. Wertungsrelevante Angaben und Nachweise sind mit Angebotsabgabe vollständig und eindeutig einzureichen. Bei jeder Anforderung ist der geforderte Nachweis maßgeblich; eine bloße Marketingbezeichnung ersetzt keinen technischen Nachweis.

ID	Bereich	Typ	Anforderung	Bieterangabe	Nachweis
D2-001	Vertrag	MUSS	Gesamtvergütung 1.100.000,00 EUR brutto für sämtliche Mindestleistungen und angebotenen Wertungsmehrleistungen.	JA/NEIN	rechtsverbindliche Eigenerklärung
D2-002	Vertrag	MUSS	36 Monate ab vertragsgemäßer Bereitstellung gemäß verbindlicher Bereitstellungsbestätigung des Auftragnehmers bis spätestens zum 15.12.2026; G1-G3 und Runtime-Pool über gesamte Laufzeit nutzbar; angebotene G4/G5 ebenfalls.	JA/NEIN	Eigenerklärung
D2-003	CU	MUSS	Ein einheitliches CU-Gesamtkontingent; keine klassen-, jahres-, quartals- oder monatsbezogenen Teilkontingente.	JA/NEIN	Eigenerklärung
D2-004	CU	MUSS	Feste wirtschaftliche CU/h-Faktoren: G1=1, G2=3, G3=4, G4=12, G5=32. Keine abweichende eigene Credit-/Tariflogik.	JA/NEIN	Eigenerklärung
D2-005	CU	MUSS	Einheitliche Nutzbarkeit des CU-Restkontingents in G1-G3 sowie angebotenen G4/G5 ohne Aufpreis und/oder Teilkontingente.	JA/NEIN	Eigenerklärung
D2-006	CU	WERTUNG	Verbindlich garantiertes CU-Gesamtkontingent für 1.100.000 EUR brutto.	Ganzzahl CU	verbindliche Bieterangabe + Plausibilisierung
D2-010	G1	MUSS	Logische Ressource: >=24 GiB, >=300 GB/s, >=8 vCPU, >=32 GiB RAM, >=250 GiB Scratch; CU/h=1; Anlage 4b-Benchmarkwerte.	Ja/Nein + Profil/SKU	Datenblatt/Dossier + Anlage 4b-Protokoll
D2-011	G2	MUSS	Logische Ressource: >=44 GiB, >=800 GB/s, >=16 vCPU, >=64 GiB RAM, >=500 GiB Scratch; CU/h=3; Anlage 4b-Benchmarkwerte.	Ja/Nein + Profil/SKU	Datenblatt/Dossier + Anlage 4b-Protokoll
D2-012	G3	MUSS	Logische Ressource: >=80 GiB, >=1,5 TB/s, >=32 vCPU, >=256 GiB RAM, >=750 GiB Scratch; CU/h=4; Anlage 4b-Benchmarkwerte.	Ja/Nein + Profil/SKU	Datenblatt/Dossier + Anlage 4b-Protokoll
D2-013	G4	WERTUNG	Optional: >=180 GiB, >=5 TB/s, >=48 vCPU, >=384 GiB RAM, >=1.500 GiB Scratch; CU/h=12; Anlage 4b-Benchmarkwerte.	Ja/Nein + Profil/SKU	Datenblatt/Dossier + Anlage 4b-Protokoll
D2-014	G5	WERTUNG	Optional: >=8 gekoppelte Acceleratoren, jeder mindestens G3; High-Speed-P2P; >=128 vCPU, >=1.024 GiB RAM, >=4.000 GiB Scratch; CU/h=32; Gesamtsystembenchmark.	Ja/Nein + Systemtyp	Architektur-/Anlage 4b-Nachweis

ID	Bereich	Typ	Anforderung	Bieterangabe	Nachweis
D2-015	G-Klassen	MUSS	Kein Overcommit zugesagter Accelerator-Ressourcen; Partitionierung nur mit garantierten Mindestmerkmalen und Isolation.	JA/NEIN	technische Beschreibung
D2-016	Technologiefortschritt	MUSS	Neue/ersetzte Hardware darf Mindestmerkmale, Verfügbarkeit oder CU/h-Faktor nicht zulasten UKF verschlechtern; neue Profile vor Nutzung nach Anlage 4b benchmarken.	JA/NEIN	Eigenerklärung
D2-017	Benchmark	MUSS	UKF-RB nach Anlage 4b; Benchmark beeinflusst den CU/h-Faktor nicht. Für die wertungsrelevanten Referenzprofile G1-G3 gilt die strenge Identitätsregel der Anlage 4b. Für weitere, nicht K1b-wertungsrelevante Profile gelten die dort definierten Nachweisregeln. Keine offizielle MLPerf-Submission erforderlich.	Ja/Nein + Rohwerte	Anlage 4b-konformes Protokoll bzw. nach Anlage 4b zulässiger Identitäts-/Referenznachweis
D2-018	Runtime	MUSS	Unabhängiger Runtime-Pool: gleichzeitig >=164 vCPU und >=1.104 GiB RAM; 24x7 bis Mindestkapazität; <=4 h Bereitstellung; kein CU-Verbrauch.	Ja/Nein + Werte	Leistungszusage/Dokumentation
D2-019	Runtime	WERTUNG	Garantierte Runtime-Kapazität oberhalb Mindestwerten gemäß Anlage 2.	vCPU/GiB	verbindliche Bieterangabe
D2-020	Storage	MUSS	Mindestens 100 TB persistent, 36 Monate inklusive, kein CU-Verbrauch; davon >=10 TB mountbarer Block-/File-Storage.	JA/NEIN	Leistungsbeschreibung
D2-021	Storage	WERTUNG	Zusätzlicher persistenter Storage oberhalb 100 TB ohne Mehrpreis; Wertungsdeckel gemäß Anlage 2.	TB gesamt	verbindliche Bieterangabe
D2-022	Storage	MUSS	Snapshot-/Versionierungs- und Wiederherstellungsfunktion; RPO/RTO angeben.	Ja/Nein + RPO/RTO	Dokumentation
D2-030	Netz	MUSS	Interner Plattform-/Compute-/Runtime-/Storage-Verkehr ohne Zusatzentgelt und ohne CU.	JA/NEIN	Eigenerklärung
D2-031	Netz	MUSS	Site-to-Site-VPN oder gleichwertige private/verschlüsselte Kopplung; Partnerkopplungen technisch möglich.	JA/NEIN	Dokumentation
D2-032	Netz	MUSS	Plattformseitiger Übergabepunkt >=10 Gbit/s; vereinbarter UKF-Ingress/Egress ohne volumenabhängige Zusatzentgelte.	Ja/Nein + Gbit/s	Netz-/Preisbeschreibung
D2-040	Kapazität	MUSS	G1: 4 parallel, <=30 min; Burst 8 <=5 Arbeitstage.	Werte	SLA-Eigenerklärung
D2-041	Kapazität	MUSS	G2: 2 parallel, <=1 h; Burst 4 <=5 Arbeitstage.	Werte	SLA-Eigenerklärung
D2-042	Kapazität	MUSS	G3: 2 parallel, <=2 h; Burst 4 <=5 Arbeitstage.	Werte	SLA-Eigenerklärung
D2-043	Kapazität	WERTUNG	Optional angebotene G4-Capability: Parallelität und Bereitstellungszeit.	Werte	verbindliche SLA-Angabe
D2-044	Kapazität	WERTUNG	Optional angebotene G5-Capability: Parallelität und Bereitstellungszeit.	Werte	verbindliche SLA-Angabe
D2-045	Kapazität	WERTUNG	Bessere Bereitstellungszeiten G1-G3 gemäß Anlage 2.	Werte	verbindliche SLA-Angabe
D2-050	Portal	MUSS	Webportal für Provisionierung, Status, Start/Stop/Löschen, Storage, Projektzuordnung und CU-Transparenz.	JA/NEIN	Dokumentation/Screenshot

ID	Bereich	Typ	Anforderung	Bieterangabe	Nachweis
D2-051	API	MUSS	Dokumentierte API für Provisionierung und Ressourcenverwaltung.	JA/NEIN	API-Dokumentation
D2-052	Container	MUSS	OCI-kompatible Container und Accelerator-Nutzung; Kubernetes-Eignung.	JA/NEIN	Dokumentation
D2-053	AI Stack	MUSS	Marktübliche AI/ML-Frameworks bzw. funktional gleichwertig.	JA/NEIN	Dokumentation
D2-054	Sandbox	MUSS	Logisch isolierte Entwicklungs-/Test-/Pilot-/Projektumgebungen mit getrennten IAM-, Storage- und Netzkontexten.	JA/NEIN	Dokumentation
D2-055	Drittanbieter	MUSS	Eigene/autorisierte VM-Images/OCI-Container, kontrollierte externe HTTPS/API-Kommunikation, Partneranbindungen.	JA/NEIN	Dokumentation
D2-056	Model Serving	WERTUNG	OpenAI-kompatible oder funktional gleichwertige dokumentierte Model-Serving-Schnittstelle.	JA/NEIN	Dokumentation
D2-057	Sandbox	WERTUNG	Erweiterte Template-/Klon-/Reset-/Lifecycle-Automatisierung.	JA/NEIN	Dokumentation
D2-058	Drittanbieter	WERTUNG	Automatisierter/self-service Import und Lifecycle autorisierter Drittanbieter-Images.	JA/NEIN	Dokumentation
D2-060	Verbrauch	MUSS	Live-Anzeige ursprüngliches, verbrauchtes und verbleibendes CU-Kontingent.	JA/NEIN	Dokumentation
D2-061	Verbrauch	MUSS	Einzelnachweis: Capability, SKU, Start/Ende, Dauer, CU/h-Faktor, CU-Verbrauch, Projekt; Benchmark-ID.	JA/NEIN	Musterreport
D2-062	Verbrauch	MUSS	Monatlicher archivierbarer Leistungsnachweis mit Anfangs-/Endbestand und Verbrauch je Capability.	JA/NEIN	Musterreport
D2-063	Verbrauch	MUSS	CSV-Export und dokumentierte API für Verbrauchsdaten.	JA/NEIN	API-/Exportdokumentation
D2-064	Verbrauch	MUSS	Warnungen bei 75/90/95 % Gesamtverbrauch.	JA/NEIN	Dokumentation
D2-065	Verbrauch	WERTUNG	Prognose, Projektbudgets, frei konfigurierbare Warnungen, erweiterte Audit-/Verbrauchs-API gemäß Anlage 2.	Ja/Nein je Merkmal	Dokumentation
D2-070	IAM	MUSS	RBAC, Projekt-/Mandantentrennung, MFA privilegiert, SAML 2.0 und/oder OIDC.	JA/NEIN	Dokumentation
D2-071	Security	MUSS	Verschlüsselung in Transit/at Rest; dokumentiertes Key Management.	JA/NEIN	Security-Dokumentation
D2-072	Security	MUSS	ISO/IEC 27001 relevanter Scope oder gleichwertiger Nachweis.	JA/NEIN	Zertifikat/Gleichwertigkeit
D2-073	Security	MUSS	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.	JA/NEIN	Testat/Prüfbericht/Nachweis

ID	Bereich	Typ	Anforderung	Bieterangabe	Nachweis
D2-074	Datenschutz	MUSS	Datenverarbeitung/-speicherung EU/EWR; AVV soweit einschlägig; Subprozessoren transparent; weitere §393-Anforderungen soweit anwendbar.	JA/NEIN	Standort-/Subprozessorenliste
D2-075	Datennutzung	MUSS	Keine Nutzung von UKF-Daten/Prompts/Modellen zum eigenen/fremden Training ohne ausdrückliche Zustimmung.	JA/NEIN	Eigenerklärung/Vertragsklausel
D2-080	Logging	MUSS	Audit-/Security-Logging und automatisierbarer SIEM-Export; 180 Tage online oder gleichwertiger Export.	JA/NEIN	Dokumentation
D2-090	SLA	MUSS	Plattform >=99,9~~-7~~ % monatlich; Storage >=99,9 7 %.	%	SLA
D2-091	Support	MUSS	P1 24x7 <=30 min; P2 <=2 h; P3 <=1 Arbeitstag.	Werte	SLA
D2-092	SLA	WERTUNG	Höhere Verfügbarkeit und kürzere Reaktionszeiten gemäß Anlage 2.	Werte	verbindliche SLA-Angabe
D2-100	Exit	MUSS	Vollständiger Export der in Anlage 4 beschriebenen Systemstände, Daten, Images, Modelle, Konfigurationen, Logs und CU-Historie.	JA/NEIN	Exit-Konzept
D2-101	Exit	MUSS	Exit-/Fortführungsworkshop <=6 Monate vor Ende; Exit-Plan <=3 Monate vor Ende.	JA/NEIN	Eigenerklärung
D2-102	Exit	MUSS	Mindestens 10 PT technische Migrationsunterstützung im Festpreis.	PT	Eigenerklärung
D2-103	Exit	WERTUNG	Zusätzliche inkludierte Migrationsunterstützung oberhalb 10 PT.	PT gesamt	verbindliche Angabe
D2-104	Exit	MUSS	Löschung erst nach UKF-Freigabe; Lösch-/Vernichtungsnachweis inkl. Replikate/Backups gemäß Löschfristen.	JA/NEIN	Löschkonzept
D2-105	Exit	WERTUNG	Automatisierter strukturierter Plattform-/Konfigurationsexport und zusätzliche Portabilitätsmerkmale.	JA/NEIN	Dokumentation
D2-110	Plausibilität	MUSS	Plausibilisierung des CU-Gesamtkontingents, fester CU-Faktoren, Anlage 4b-Benchmarkwerte, G1-G3-Kapazitäten, optional G4/G5, Runtime, Storage, Netzwerk und All-In-Leistung.	JA/NEIN + Anlage	Plausibilisierungsblatt

Hinweise zur Nachweisführung

Soweit Zertifikate, Prüfberichte oder Herstellerunterlagen gefordert werden, sind vergaberechtlich zulässige gleichwertige Nachweise zu akzeptieren, sofern der Bieter die Gleichwertigkeit substantiiert belegt. Bei wertungsrelevanten Angaben ist der in Anlage 2 geforderte Nachweis mit Angebotsabgabe vorzulegen. Eine spätere Aufklärung darf eine wertungsrelevante Leistungsangabe nicht verbessern oder erstmals schaffen.