Anlage 2 – Beiblatt zur Angebotswertung

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.

1. Grundsatz der Zuschlagswertung

Der Zuschlag wird auf das wirtschaftlichste wertbare Angebot erteilt. Die Gesamtvergütung mit 1.100.000,00 EUR brutto ist verbindlich vorgegeben. Die Rangfolge ergibt sich ausschließlich aus dem verbindlich angebotenen wirtschaftlichen Leistungsumfang und den objektiv messbaren Qualitätsmerkmalen. Es werden maximal 1.000 Punkte vergeben.

Kriterium	Punkte	Gewicht
K1 Wirtschaftlicher Compute-Umfang und technische Effizienz	500	50 %
K2 Runtime-/Accelerator-Flexibilität und Bereitstellung	150	15 %
K3 Plattform, Sandbox, Storage und Integration	200	20 %
K4 Exit und Portabilität	100	10 %
K5 SLA und Support	50	5 %
GESAMT	1.000	100 %

2. K1 - Wirtschaftlicher Compute-Umfang und technische Effizienz (max. 500 Punkte)

2.1 K1a - Garantiertes CU-Gesamtkontingent (max. 350 Punkte)

Gewertet wird ausschließlich das in Anlage 3 verbindlich garantierte CU-Gesamtkontingent. Die festen Verbrauchsfaktoren sind G1=1, G2=3, G3=4, G4=12, G5=32 CU/h. Das höchste wertbare CU-Angebot erhält 350 Punkte. Alle übrigen Angebote erhalten proportional Punkte:

Formel: P_K1a(i) = 350 x CU_i / CU_max		
Beispiel	**CU**	**K1a-Punkte**
Bieter A	320.000	280,00
Bieter B	400.000	350,00
Bieter C	360.000	315,00

2.2 K1b - Technische Effizienz der verpflichtenden Capabilities (max. 150 Punkte)

Der Anlage 4b-Benchmark wird getrennt von der wirtschaftlichen CU-Logik gewertet. Für jede verpflichtende Capability G1-G3 benennt der Bieter in Anlage 3 genau ein Referenzprofil. Je Capability werden LLM- und Imaging-Rohwert getrennt mit dem jeweils höchsten wertbaren Rohwert derselben Capability verglichen. Die Gewichtung leitet sich aus der eingefrorenen wirtschaftlichen Referenzannahme G1:G2:G3 = 15:25:35 ab und wird proportional auf 150 Punkte übertragen: G1 30 Punkte, G2 50 Punkte, G3 70 Punkte.

Capability	LLM max. Punkte	IMG max. Punkte	Capability gesamt
G1	15	15	30
G2	25	25	50

Capability	LLM max. Punkte	IMG max. Punkte	Capability gesamt
G3	35	35	70
Gesamt	75	75	150

Für jede Dimension d und Capability k gilt:

Formel: P_k,d(i) = MaxPunkte_k,d x B_k,d(i) / B_k,d,max

B_k,d(i) ist der Anlage 4b-konforme Rohwert des Referenzprofils des zu wertenden Bieters; B_k,d,max ist der höchste wertbare Rohwert aller wertbaren Angebote in derselben Capability und Dimension. Die Werte werden auf zwei Dezimalstellen berechnet. Ein Bieter mit dem Höchstwert erhält die vollen Dimensionspunkte.

Die technische Benchmarkleistung verändert den CU/h-Verbrauch nicht. Ein schnelleres Profil innerhalb derselben Capability liefert damit einen messbaren technischen Mehrwert für denselben wirtschaftlichen CU-Verbrauch.

3. K2 - Runtime-/Accelerator-Flexibilität und Bereitstellung (max. 150 Punkte)

Teilkriterium	0 Punkte / Mindeststufe	Punkte	mittlere Stufe	Punkte	beste Stufe	Punkte
G1 Bereitstellungszeit	>15 bis 30 min	0	>5 bis 15 min	5	<=5 min	10
G2 Bereitstellungszeit	>30 bis 60 min	0	>10 bis 30 min	5	<=10 min	10
G3 Bereitstellungszeit	>60 bis 120 min	0	>30 bis 60 min	7,5	<=30 min	15
G4 Capability	nicht angeboten oder >8 h	0	angeboten; >4 bis 8 h	10	angeboten; <=4 h	20
G5 Capability	nicht angeboten oder >24 h	0	angeboten; >12 bis 24 h	12,5	angeboten; <=12 h	25
Runtime vCPU	164 bis <246	0	>=246 bis <328	17,5	>=328	35
Runtime RAM	1.104 bis <1.656 GiB	0	>=1.656 bis <2.208 GiB	17,5	>=2.208 GiB	35

Für K2 gelten ausschließlich die nachfolgend veröffentlichten drei Wertungsstufen mit den dort ausdrücklich angegebenen Punktwerten. Die Mindeststufe erhält 0 Punkte, die mittlere Stufe 50 % der Maximalpunktzahl und die beste Stufe 100 % der Maximalpunktzahl. Zwischenwerte erhalten die Punkte der nächstschlechteren vollständig erreichten Stufe; eine lineare Interpolation findet nicht statt. Für G4/G5 werden Punkte nur vergeben, wenn Capability und Bereitstellungszeit verbindlich angeboten und nachgewiesen sind. Werte oberhalb der besten Stufe führen zu keinen zusätzlichen Punkten.

4. K3 - Plattform, Sandbox, Storage und Integration (max. 200 Punkte)

Unterkriterium	Bewertung	Punkte
Zusätzlicher persistenter Storage	100 TB=0; je volle zusätzliche 25 TB= 12,5; Deckel bei >=200 TB Gesamt	0-50
Sandbox-/Template-Automatisierung	reproduzierbare Templates/Klonen + automatisierter Reset/Lifecycle	30
Drittanbieter-Image-Lifecycle	self-service Import/Lifecycle mit Integritäts-/Registry-Unterstützung	20
Private Partnerkopplungen	mehrere parallel konfigurierbare private/verschlüsselte Kopplungen	20
Kontrollierte externe AI-/API-Anbindung	projektbezogene Egress-/Allowlist-/Policy-Steuerung	20
Model-Serving-Schnittstelle	OpenAI-kompatibel oder funktional gleichwertig dokumentiert	20
Warnschwellen	zusätzlich zu 75/90/95 frei konfigurierbar	10

Unterkriterium	Bewertung	Punkte
Immutable Monatsnachweise	kryptografisch signiert oder technisch manipulationsgeschützt	10
Erweiterte Verbrauchs-/Audit-API	Einzelnutzung, Kontostand, Runtime-Auslastung, Monatsabschlüsse maschinenlesbar	20

Bei Ja/Nein-Kriterien werden Punkte nur vergeben, wenn der geforderte Nachweis mit dem Angebot vorliegt und die Funktion ohne zusätzliche Vergütung Bestandteil der angebotenen Leistung ist.

5. K4 - Exit und Portabilität (max. 100 Punkte)

Unterkriterium	Bewertung	Punkte
Inkludierte Migrationsunterstützung	10 PT=0; 11-12=5; 13-14=10; >=15=20	0-20
Automatisierter strukturierter Export	Inventar, Konfigurationen, Metadaten maschinenlesbar	25
IaC-/Deployment-Portabilität	vorhandene Definitionen vollständig exportierbar und weiterverwendbar	20
Exit-Test	vollständiger technischer Probeexport während Laufzeit ohne Mehrpreis	15
Exit-Datenübertragung	vollständiger Exit-Datenexport ohne mengenabhängige Egress-Gebühren	20

6. K5 - SLA und Support (max. 50 Punkte)

Unterkriterium	Stufen	Punkte
Plattformverfügbarkeit	97,70-99,94 %=0; 99,95-99,98 %=5; >=99,99 %=10	0-10
Storage-Verfügbarkeit	97,70-99,94 %=0; 99,95-99,98 %=5; >=99,99 %=10	0-10
P1 Reaktionszeit	>15 bis 120 min=0; >5 bis 15 min=10; <=5 min=20	0-20
P2 Reaktionszeit	>60 bis 240 min=0; >30 bis 60 min=5; <=30 min=10	0-10

7. Gesamtwertung und Rangfolge

Gesamtformel: P_GESAMT = P_K1a + P_K1b + P_K2 + P_K3 + P_K4 + P_K5

Die Gesamtpunktzahl wird auf zwei Dezimalstellen berechnet. Das wertbare Angebot mit der höchsten Gesamtpunktzahl ist das wirtschaftlichste Angebot. Bei Punktgleichheit entscheidet zunächst die höhere Summe aus K1a+K1b, danach K2, K3, K4 und K5. Besteht anschließend vollständiger Gleichstand, erfolgt ein dokumentierter Losentscheid nach Maßgabe des Vertrags-/Vergaberechts und der internen Freigabe.

8. Plausibilitäts- und Aufklärungsregeln

Ein außergewöhnlich hohes CU-Angebot, außergewöhnlich hohe Benchmarkwerte oder sehr gute Runtime-/SLA-Werte werden nicht allein aufgrund ihrer Höhe abgewertet. Vor Zuschlag prüft der Auftraggeber jedoch die technische und wirtschaftliche Plausibilität der verbindlich zugesagten Leistung. Aufklärung darf einen bereits feststehenden Angebotsinhalt erläutern, nicht aber wertungsrelevante Angaben nachträglich verbessern, ersetzen oder erstmals schaffen.

Die Trennung zwischen CU und Benchmark ist bei der Plausibilisierung ausdrücklich zu beachten: CU bewertet den garantierten wirtschaftlichen Ressourcenanspruch; der Anlage 4b-Benchmark bewertet die technische Effizienz der konkret angebotenen Referenzprofile. Eine bietereigene Verknüpfung beider Größen ist unzulässig.