# Anlage 4b – Technische Anlage (UKF Reference Benchmark UKF-RB 1.1)

**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. Zweck und verbindliche Wirkung

UKF-RB 1.1 misst die technische Leistungsfähigkeit der tatsächlich angebotenen Accelerator-Ressourcenprofile herstellerneutral in zwei für das UKF relevanten Dimensionen. Die Messwerte werden in Anlage 3 verbindlich erklärt und in Anlage 2 K1b bewertet. Endanwender sehen im Betrieb den festen CU/h-Wert der Capability sowie optional technische Benchmarkinformationen; sie müssen keine Benchmarkformel berechnen.

## 2. Benchmarkbasis und Versionierung

Das Messverfahren orientiert sich methodisch an öffentlich dokumentierten MLPerf-Inference-v6.1-Workloads und Messprinzipien, verlangt jedoch keine offizielle MLPerf-/MLCommons-Submission. Für UKF-RB werden zwei Dimensionen verwendet:

| Dimension | Modell | Datensatz | Szenario | Messgröße |
|---|---|---|---|---|
| RB-LLM | Llama 3.1 8B | gemäß eingefrorenem UKF-RB-Paket | Offline | Tokens/s |
| RB-IMG | 3D U-Net | KiTS19 gemäß eingefrorenem UKF-RB-Paket | Offline | Samples/s |

Maßgeblich sind ausschließlich die mit den Vergabeunterlagen veröffentlichten und eingefrorenen Versionen, Datensätze, Accuracy-Grenzen, Szenarien, Messparameter und Container-/Paket-Hashes. Nach Veröffentlichung dürfen diese Parameter nicht bieterspezifisch geändert werden.

## 3. Angebotsprofile und Messpflicht

Für die Zuschlagswertung K1b ist je verpflichtender Capability G1-G3 genau ein Referenzprofil verbindlich zu benennen und mit Anlage 4b-konformen Rohwerten nachzuweisen. Weitere angebotene Profile sind technisch zu dokumentieren; soweit für sie Benchmarkwerte angegeben werden, müssen diese nach den Anlage 4b-Regeln nachvollziehbar sein, sind jedoch nicht Bestandteil der K1b-Wertung.

Der Benchmark wird auf genau der Ressourcenkonfiguration ausgeführt, die UKF angeboten wird. Dazu gehören Acceleratorzahl/Partitionierung, CPU-/RAM-Zuteilung, Virtualisierung, Treiber, Runtime und relevante Softwareparameter.

## 4. Messregeln

- Die im Benchmarkpaket vorgegebenen Modelle, Datensätze, Accuracy-Grenzen und Szenarioeinstellungen dürfen nicht verändert werden, soweit Anlage 4b dies nicht ausdrücklich zulässt.
- Performancewerte sind nur wertbar, wenn die Accuracy-Prüfung bestanden ist.
- Hersteller-/plattformspezifische Optimierungen sind nur zulässig, wenn sie dokumentiert, nach den eingefrorenen Regeln zulässig und im späteren UKF-Profil verfügbar sind.
- Reine Benchmark-Sonderoptimierungen, die im Produktivprofil nicht verfügbar sind, sind unzulässig.
- Bei partitionierten Accelerator-Ressourcen ist die tatsächlich garantierte Partition zu benchmarken.
- Bei G5/Multi-Accelerator-Profilen wird das Gesamtsystem als eine Ressource benchmarked; Einzelwerte dürfen nicht rechnerisch multipliziert werden.

- Für jedes Profil ist ein maschinenlesbarer Ergebnis- und Systemnachweis mit Rohwerten, Hard-/Softwarekonfiguration und Reproduzierbarkeit vorzulegen.

## 5. Zulässige Referenznachweise

Für die drei K1b-wertungsrelevanten Referenzprofile G1, G2 und G3 ist entweder das eingefrorene UKF-RB-Paket auf der angebotenen Ressourcenkonfiguration auszuführen oder ein bereits vorliegendes Ergebnis zu verwenden, bei dem Hardwareprofil einschließlich Acceleratorzahl und Partitionierung, CPU-/RAM-Zuteilung soweit benchmarkrelevant, Virtualisierung, Treiber, Runtime, Softwarestack, Benchmarkversion, Modell/Datensatz, Szenario, Precision, Messparameter sowie Paket-/Konfigurationshash nachweislich identisch sind. Eine lediglich technische oder funktionale Gleichwertigkeit genügt für K1b nicht. Für zusätzliche, nicht K1b-wertungsrelevante Profile kann ein nachvollziehbar dokumentierter Referenznachweis verwendet werden, wenn die technische Gleichwertigkeit substantiiert belegt ist; bei Zweifeln oder relevanten Abweichungen ist UKF-RB auszuführen. UKF behält sich Reproduktionsmessungen im Rahmen der technischen Verifikation der Bereitstellung vor.

## 6. Verwendung der Rohwerte in Anlage 2

Die Benchmarkwerte werden nicht in einen CU-Verbrauchsfaktor umgerechnet. Anlage 2 bewertet die technische Effizienz peer-relativ je Capability und Dimension. Für jede der sechs Kombinationen G1-LLM, G1-IMG, G2-LLM, G2-IMG, G3-LLM und G3-IMG wird der Bieterrohwert durch den höchsten wertbaren Rohwert derselben Kombination geteilt und mit der veröffentlichten Maximalpunktzahl multipliziert.

| Kombination | Max. Punkte |
|---|---|
| G1 LLM | 15 |
| G1 IMG | 15 |
| G2 LLM | 25 |
| G2 IMG | 25 |
| G3 LLM | 35 |
| G3 IMG | 35 |

## 7. Angebots- und Bereitstellungsnachweis

Mit Angebotsabgabe sind für die K1b-Referenzprofile G1-G3 mindestens Profil-ID, Capability, Instanztyp/SKU, Acceleratorzahl/-speicher, B_LLM, B_IMG sowie das UKF-RB-Protokoll oder der nach Kapitel 5 zulässige Identitätsnachweis vorzulegen. Die Referenzprofile sind in Anlage 3 eindeutig zu kennzeichnen. Für weitere Profile gelten die Dokumentationsanforderungen aus Anlagen 4a und 3.

Im Rahmen der technischen Verifikation der Bereitstellung darf UKF ausgewählte Profile reproduzieren. Als wesentlich abweichend gilt eine reproduzierte Leistung, die in einer Benchmarkdimension mehr als 5 % unter dem verbindlich angebotenen Rohwert liegt. Der Auftragnehmer muss dann die zugesagte Leistung herstellen oder ein technisch mindestens gleichwertiges Ersatzprofil innerhalb derselben CU-Klasse bereitstellen. Die Zuschlagswertung wird nachträglich nicht neu berechnet; weitergehende vertragliche Rechte bleiben unberührt.

## 8. Aufnahme neuer Hardware während der Laufzeit

Neue Accelerator-Generationen und SKUs dürfen aufgenommen werden, sofern sie einer Capability zugeordnet, technisch dokumentiert und nach UKF-RB 1.1 benchmarked werden. Der CU/h-Verbrauch richtet sich weiterhin ausschließlich nach der Capability. Ein vom Auftragnehmer veranlasster Technologiewechsel darf UKF nicht schlechter stellen.

Ist das eingefrorene Benchmarkpaket auf einer späteren Architektur objektiv nicht ausführbar, darf eine funktional gleichwertige Ersatzmessung nur nach vorheriger schriftlicher Zustimmung des UKF und transparenter technischer Kalibrierung verwendet werden. Die wirtschaftliche CU-Klasse bleibt hiervon unberührt.

## 9. Verbindliches Benchmarkpaket UKF-RB 1.1

Bestandteil dieser technischen Anlage ist das mit den Vergabeunterlagen bereitgestellte elektronische Benchmarkpaket der Anhänge 1 bis 4 zur Anlage 4b.

Das Benchmarkpaket enthält die verbindlichen technischen Ausführungsinformationen einschließlich der festgelegten Modell- und Datensatzversionen bzw. Bezugsquellen, Mess- und Accuracy-Parameter, Szenarioeinstellungen, technischen Konfigurationen, Prüfsummen, Ergebnisschemata sowie der zur Durchführung und Dokumentation vorgesehenen technischen Anleitungen und Skripte.

Für die technische Durchführung des UKF-RB 1.1 ist ausschließlich die mit den Vergabeunterlagen veröffentlichte und durch die vorstehende SHA-256-Prüfsumme eindeutig identifizierte Version des Benchmarkpakets maßgeblich. Alle Bieter erhalten dasselbe Benchmarkpaket.

Die in dieser Anlage festgelegten Regeln zur Auswahl der Referenzprofile, zur Wertung, zur Nachweisführung und zur technischen Verifikation bleiben hiervon unberührt.