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

Kompakte technische Anlage | KI-Compute-, Entwicklungs- und Integrationsplattform | 22.10.2026

Verbindliche technische Leistungsbewertung K1b, keine Berechnung des CU-Verbrauchs.

1. Geltung und Ziel

Diese Anlage definiert einen einheitlichen Benchmarknachweis für je ein verbindlich angebotenes G1-, G2- und G3-Referenzprofil. Die Messungen dienen ausschließlich der technischen K1b-Wertung (max. 150 Punkte), der Plausibilisierung und der späteren Verifikation. Eine offizielle MLPerf-/MLCommons-Submission wird nicht verlangt. Die festen CU/h-Faktoren G1=1, G2=3, G3=4 sowie für optionale Klassen G4=12 und G5=32 werden durch die Messergebnisse nicht verändert.

2. Öffentliche Referenzimplementierung und feste Workloads

Als öffentlich zugängliche methodische Grundlage gilt MLPerf Inference v6.1 aus dem Projekt mlcommons/inference. Referenzcode ist der zum Commit a468775 gehörende Quellcode-Stand vom 30.07.2026 (vollständigen SHA-1 über git rev-parse HEAD protokollieren). Maßgeblich sind die dort beschriebenen Original-Modelle, Original-Datensätze, Accuracy-Regeln und das Offline-Lastszenario. Der Auftraggeber entwickelt und verteilt keine abweichenden eigenen Workloads.

Messdimension	Original-Workload	Messgröße
RB-LLM	Llama 3.1 8B / CNN-Daily Mail; Offline	Generierte Ausgabetokens je Sekunde
RB-IMG	3D U-Net / KiTS19; Offline	Vollständig verarbeitete Samples je Sekunde

Quellen: github.com/mlcommons/inference; docs.mlcommons.org/inference/benchmarks/language/llama3_1-8b/; docs.mlcommons.org/inference/benchmarks/medical_imaging/3d-unet/.

3. Ausführung und zulässige Optimierungen

- Für G1, G2 und G3 ist jeweils exakt das in Anlage 3 bezeichnete und später bereitstellbare Referenzprofil zu testen. Die für den Bieter verfügbaren Accelerator-, CPU-, RAM-, Treiber-, Runtime- und Partitionierungseigenschaften sind zu dokumentieren.
- Die MLCommons-Lastgenerierung im Offline-Szenario sowie das Original-Modell und der Original-Datensatz sind einzuhalten. Die Modell-Accuracy-Anforderungen der festgelegten Referenzcodefassung müssen erfüllt sein. Ergebnisse ohne erfolgreichen Accuracy-Nachweis sind nicht wertbar.
- Hersteller-/plattformbezogene Beschleunigungen sind zulässig, soweit die Original-Workload-Regeln eingehalten werden, sie im später vom UKF nutzbaren Profil verfügbar sind und Konfiguration, Präzision/Quantisierung und eingesetzter Stack offengelegt werden.
- Die Bieter dokumentieren die konkrete Nutzung der Original-Quellen einschließlich Repository-Hash, Modell-/Datensatzidentität, Abhängigkeiten sowie vollständigem Aufruf. Originale Benchmark-/Accuracy-Logs werden unverändert abgegeben. Nicht veröffentlichte Sonderdaten oder andere Modelle dürfen nicht verwendet werden.

4. Ergebnisse und Angebotsnachweise

Mit Angebotsabgabe sind für G1–G3 genau sechs gültige positive Rohwerte abzugeben (je G1/G2/G3 einmal RB-LLM und RB-IMG), außerdem Profil-ID, System-/Softwarekonfiguration, tatsächlicher Aufruf, Versionsangaben, unveränderte Original-Logs, bestandene Accuracy-Prüfung und SHA-256-Prüfsummen der beigefügten Protokolldateien. Als einfaches Format kann die beiliegende Bieter-Ergebnisvorlage verwendet werden. Ein Werbedatenblatt ersetzt keine verifizierbaren Rohwerte.

5. Punkteberechnung

Für jede Kombination aus Capability k und Messdimension d wird der Bieterrohwert B(k,d) zum höchsten wertbaren Rohwert dieser Kombination Bmax(k,d) in Verhältnis gesetzt: P(k,d) = MaxPunkte(k,d) × B(k,d) / Bmax(k,d). Die sechs Maximalwerte betragen:

Capability	LLM	IMG	Summe
G1	15	15	30
G2	25	25	50
G3	35	35	70
Gesamt	75	75	150

Die je Dimension berechneten Punkte werden auf zwei Nachkommastellen gerundet. Der K1b-Gesamtwert ist die Summe der sechs Dimensionspunkte. Wirtschaftliche CU-Verbrauchsfaktoren bleiben unverändert.

6. Nachprüfung, Profilwechsel und neuere Hardware

Der Auftraggeber darf im Rahmen der Bereitstellungsverifikation die eingereichten Nachweise nachvollziehen und ausgewählte Messungen auf dem bereitgestellten Profil reproduzieren. Eine um mehr als 5 % niedrigere reproduzierte Leistung in einer Dimension gilt als wesentliche Abweichung; der Auftragnehmer hat die zugesagte Leistung herzustellen oder ein mindestens gleichwertiges Profil in derselben CU-Klasse zur Verfügung zu stellen. Ein späterer Technologiewechsel darf die zugesagten technischen und wirtschaftlichen Eigenschaften nicht verschlechtern. Die Zuschlagswertung wird nicht nachträglich neu berechnet.

7. Transparenz und gleiche Behandlung

Fragen zur Ausführbarkeit, zum Modellzugang oder zu Datenquellen sind vor Ablauf der Angebotsfrist über die Vergabeplattform einzureichen und für alle Bieter einheitlich zu beantworten. Eine Änderung von Workloads, Datensätzen, Genauigkeitsgrenzen oder K1b-Messgrößen nach Angebotsabgabe ist ausgeschlossen. Bei erforderlichen Änderungen vor Angebotsabgabe sind die Vergabeunterlagen für alle Bieter einheitlich zu aktualisieren und Fristen erforderlichenfalls anzupassen.

8. Abgrenzung der Hilfsprogramme

Das ergänzende UKF-RB-Kompaktpaket enthält lediglich ein Ergebnisformular sowie lokale Hilfsprogramme zur Struktur-, Datei-Hash- und Punkteprüfung. Es ist keine eigenständige, GPU-getestete Benchmarkdistribution. Verbindliche Messungen erfolgen mit der in Ziffer 2 benannten öffentlichen MLCommons-Referenzfassung. Der Validierungscode kann die Wahrheit oder Reproduzierbarkeit eines Messwertes nicht allein durch Prüfung der JSON-Datei bestätigen.