UKF Reference Benchmark 1.1
Verbindliche Testparameter

Ergänzung zur Anlage 4b · KI-Compute-Plattform · Fassung 22.10.2026

Diese Unterlage fasst die bereits festgelegten Messbedingungen der Bieteranleitung in einer eigenständigen Parametertabelle zusammen. Für die technische Durchführung gelten ergänzend die originale, unten referenzierte MLCommons-Implementierung und die zugehörigen Workload-Regeln.

1. Einheitliche Benchmarkparameter

Parameter	Verbindliche Festlegung
Benchmark-Referenz	MLPerf Inference v6.1, Repository mlcommons/inference; Quellcode-Commit a468775 (der vollständige Commit-Hash ist vom Bieter zu protokollieren).
Testszenario	Offline, unter Verwendung des LoadGen und der zum fixierten Quellcode gehörenden Originalkonfiguration mlperf.conf.
RB-LLM: Modell und Daten	Llama 3.1 8B, CNN/DailyMail, Original-Modellgewichte, Tokenisierung, Datensatzdefinition und Vorverarbeitung gemäß Referenzimplementierung.
RB-LLM: Messgröße	Generierte Ausgabetokens pro Sekunde (Output-Tokens/s); keine Zurechnung der Eingabetokens.
RB-IMG: Modell und Daten	3D U-Net / KiTS19, Original-Modell, Validierungsdaten und Vorverarbeitung gemäß Referenzimplementierung.
RB-IMG: Messgröße	Vollständig verarbeitete Samples pro Sekunde (Samples/s).
Performance-Mindestdauer	Mindestens 600.000 Millisekunden (600 Sekunden) je Performance-Lauf; erforderlichenfalls länger bis zu einem gültigen LoadGen-Ergebnis.
Mindestabfragezahlen	Offline.min_query_count gemäß fixierter MLPerf-Konfiguration: Llama 3.1 8B = 13.368; 3D U-Net = 43. Die LoadGen-interne Umsetzung ist maßgeblich.
Mess- und Genauigkeitsläufe	Je Capability G1, G2 und G3 pro Workload mindestens ein gültiger Offline-Performance-Lauf sowie ein vollständiger, bestandener Accuracy-Lauf.
Gültigkeit / Accuracy	Performance-Ergebnis LoadGen VALID; originaler modell- und workload-spezifischer MLCommons-Accuracy-Check bestanden. Keine eigenmächtig veränderten Accuracy-Grenzen.
Referenzressourcen	Je G1, G2 und G3 genau das in Anlage 3 benannte, technisch zugesagte Referenzprofil; identische logische Accelerator-, CPU-/RAM- und Partitionierungsressourcen für beide Workloads.
Optimierungen	Hersteller-/plattformbezogene Backends, Quantisierung und Batching zulässig, wenn der originale Accuracy-Check bestanden wird, die originalen Messregeln eingehalten werden und die Konfiguration für UKF nutzbar bleibt.
Dokumentation	Vollständiger Aufruf, Profil-ID, Anzahl/Typ Accelerator, VRAM, garantierte vCPU/RAM, Treiber, Runtime, Framework/Versionen, Datensatz-/Modellidentität sowie Original-Performance- und Accuracy-Logs.
Integrität der Nachweise	SHA-256-Prüfsummen aller eingereichten Original-Logdateien; maschinenlesbares Ergebnis gemäß Bieter-Ergebnisvorlage.json.

2. Messumfang und K1b-Punkte

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

Eine eigenständige Teststellung beim Auftraggeber vor Zuschlag ist nicht vorgesehen. Die Bieter reichen sechs gültige Performance-Rohwerte und die dazugehörigen Accuracy-Nachweise mit ihrem Angebot ein. Die wirtschaftlichen CU/h-Faktoren werden durch die Benchmarkwerte nicht verändert.

3. Verbindliche Nachweise zur Angebotsabgabe

Abzugeben sind die ausgefüllte Bieter-Ergebnisvorlage.json, die sechs originalen Performance-Protokolle, sechs bestandene Accuracy-Berichte, Original-LoadGen-Detail-/Summary-Logs, reproduzierbare Ausführungsbefehle und die technische Systembeschreibung zu G1, G2 und G3. Die Datei pruefe_ergebnis.py prüft ausschließlich die formale Vollständigkeit und Dateihashes; sie bestätigt keine tatsächlich erreichte GPU-Leistung.

4. Verbindliche Quellenfassung und Hinweis

Referenzcode: https://github.com/mlcommons/inference (Commit a468775). Workload-Anleitungen: https://docs.mlcommons.org/inference/benchmarks/language/llama3_1-8b/ sowie https://docs.mlcommons.org/inference/benchmarks/medical_imaging/3d-unet/. Die laufend aktualisierten Online-Anleitungen dürfen die fixierte Quellcodefassung und deren Parameter nicht ersetzen.

Freigabehinweis: Die Parameter geben den vereinbarten Sollstand wieder. Die technische Ausführbarkeit auf allen zulässigen G1–G3-Accelerator-Konfigurationen wurde nicht praktisch überprüft; aus dieser Anlage folgt keine Bescheinigung der GPU-Lauffähigkeit.