UKF Reference Benchmark 1.1

Verbindliche Testparameter und Bieteranleitung | Ergänzung zu Anlage 4b | Entwurfsfassung 22.10.2026

Ziel: Sechs einheitlich nachvollziehbare Benchmarkrohwerte für die K1b-Wertung (150 Punkte). Eine Teststellung beim UKF vor Zuschlag ist nicht vorgesehen; die Messung erfolgt bieterseitig auf den verbindlich angebotenen Referenzprofilen.

1. Festgelegte Messparameter

Parameter	Vorgabe
Referenzcode	MLCommons Inference v6.1; Repository mlcommons/inference; festgelegter Commit a468775 (vollständige Commit-ID im Protokoll).
Testszenario	Offline; offizielle MLPerf LoadGen-Konfiguration der fixierten Codefassung.
RB-LLM	Llama 3.1 8B / CNN-DailyMail gemäß fixierter MLCommons-Referenz; Ergebnis: generierte Ausgabetokens pro Sekunde.
RB-IMG	3D U-Net / KiTS19 gemäß fixierter MLCommons-Referenz; Ergebnis: vollständig verarbeitete Samples pro Sekunde.
Messdauer	Je Performance-Lauf mindestens 600 Sekunden (600.000 ms); bei Bedarf länger, wenn LoadGen dies für VALID verlangt.
Abfrageumfang	LLM: mindestens 13.368 Offline-Anfragen; Imaging: mindestens 43 Offline-Anfragen; LoadGen darf Eingangsdaten gemäß MLPerf-Regeln erneut ausgeben.
Gültigkeit	Je Messung Performance-Modus mit gültigem LoadGen-Ergebnis („VALID“), zusätzlich Accuracy-Modus mit bestandenem workload-spezifischem Original-Accuracy-Check.
Leistung	G1/G2/G3 jeweils ein identisches Leistungsprofil für die eigene LLM- und IMG-Messung; kein Zusammenfassen zusätzlicher Ressourcen.
Wiederholung	Je Klasse und Workload mindestens ein gültiger Performance- und ein vollständiger Accuracy-Lauf; gesamt sechs Performance- und sechs Accuracy-Läufe.
Softwareoptimierung	Batching, Framework und Accelerator-Optimierung zulässig, soweit die fixierten Mess-/Genauigkeitsregeln eingehalten und die Optimierungen später für UKF verfügbar sind.

Diese Vorgaben konkretisieren das MLPerf-Offline-Verfahren. Die originalen Datensätze, Modellgewichte, Vorverarbeitung, Accuracy-Prüfungen und LoadGen-Gültigkeitskriterien des fixierten Referenzstands dürfen nicht eigenmächtig verändert werden.

2. Schrittfolge für jeden Bieter

1. Referenzcode beschaffen: Repository https://github.com/mlcommons/inference klonen, auf Commit a468775 wechseln und vollständige Commit-ID mittels git rev-parse HEAD dokumentieren.
2. Original-Modelle/-Daten beschaffen: die von den beiden MLCommons-Workload-READMEs vorgegebenen Versionen (einschließlich Zugang/Lizenz für Llama) laden; Bezugsquellen und Versionen protokollieren.
3. Drei Referenzprofile G1/G2/G3 festlegen: dieselben Angebots-Profil-IDs in Anlage 3 und im Benchmarkprotokoll verwenden; GPU-/Partitionierungsparameter, vCPU, RAM, Treiber, Runtime und Präzision dokumentieren.
4. Je Profil zunächst beide Accuracy-Läufe durchführen (LLM und IMG). Nur bei bestandenem Original-Accuracy-Check wird die Performance-Messung gewertet.
5. Mit der zugehörigen offiziellen Workload-Anleitung je Profil LLM und IMG im Offline-Performance-Modus mit gültiger LoadGen-Konfiguration messen. Der Lauf muss mindestens 600 s dauern und „VALID“ melden.
6. Originale Logs und den vollständigen Befehlsaufruf speichern: insbesondere mlperf_log_summary.txt, mlperf_log_detail.txt, Accuracy-Report und Systembeschreibung. Keine Übertragung von Benchmarkwerten anderer Hardwarekonfigurationen.
7. Die sechs positiven Rohwerte in Bieter-Ergebnisvorlage.json übertragen und die SHA-256-Prüfsummen der protokollierten Dateien eintragen; mit pruefe_ergebnis.py formal validieren.
8. Mit dem Angebot die ausgefüllte JSON-Datei, sämtliche Originalprotokolle und die technische Nachweisdokumentation einreichen.

3. Feste Mess- und Nachweisregeln

Thema	Verbindliche Regel
Profilidentität	Ein Profil entspricht genau der technisch angebotenen logischen Ressource und deren garantierten Accelerator-/CPU-/RAM-Anteilen. Benchmarks auf größeren Ressourcen sind nicht zulässig.
Laufzeitumgebung	Keine unzulässige Änderung an Eingabedaten, Modellsemantik, Accuracy-Grenzen oder LoadGen. Quantisierung oder optimierte Backends nur, wenn der Accuracy-Check besteht und die Produktionskonfiguration identisch nutzbar ist.
Auslastung	Messung auf der bereitgestellten und garantierten logischen Ressource. Zusätzliche, UKF später nicht zur Verfügung stehende Beschleuniger dürfen nicht mitmessen.
Ergebnisbildung	Der offizielle bzw. protokollierte Ausgabetoken-Durchsatz (LLM) bzw. Sample-Durchsatz (IMG) aus dem gültigen Offline-Lauf ist zu verwenden. Eingabetokens werden nicht als Ausgabetokens gewertet.
Fehlversuche	Unvollständige, ungültige oder Accuracy-fehlgeschlagene Läufe sind kein wertbarer Leistungsnachweis. Wird erneut gemessen, sind die verwendete Konfiguration und das gewertete Originalprotokoll eindeutig zu identifizieren.

4. Ergebnisvorlage und Abgabe

Pro Bieter ist eine zusammenhängende Ergebnisvorlage mit sechs Einträgen auszufüllen: G1-LLM, G1-IMG, G2-LLM, G2-IMG, G3-LLM und G3-IMG. Jeder Eintrag enthält Profil-ID, gemessenen positiven Durchsatz, bestandenes Accuracy-Ergebnis, VALID-Status, Systembeschreibung, Ausführungsbefehl, Versionen und Original-Logs mit SHA-256.

Der Python-Prüfer pruefe_ergebnis.py bestätigt nur die formale Struktur und Dateihashes; er prüft weder die echte Hardwareidentität noch die Richtigkeit der Messung. Das UKF kann die Ergebnisse bei Bereitstellung gemäß Anlage 4b reproduzieren.

5. Punktewertung bleibt unverändert

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

Je Dimension: Punkte = Maximalpunkte × Bieter-Rohwert / höchster wertbarer Rohwert. Die festen wirtschaftlichen CU/h-Faktoren bleiben hiervon unabhängig.

6. Technische Quellen / Versionsbezug

MLCommons-Repository: https://github.com/mlcommons/inference

Llama-Workload: https://docs.mlcommons.org/inference/benchmarks/language/llama3_1-8b/

3D-Unet-Workload: https://docs.mlcommons.org/inference/benchmarks/medical_imaging/3d-unet/

MLPerf LoadGen-Konfiguration: https://github.com/mlcommons/inference/blob/master/loadgen/mlperf.conf (für die Bieter ist ausschließlich der oben festgelegte Commit verbindlich).

Hinweis zur Verfahrensfreigabe: Diese Testanleitung fixiert Prüfparameter, ersetzt aber keine praktische Ausführbarkeitsprüfung. Die Workloads wurden für UKF nicht auf allen zulässigen G1–G3-Accelerator-Konfigurationen getestet. Bieterfragen zu technischen Hindernissen sind vor Angebotsfrist einheitlich für alle zu klären.