Kimi Weight Windkanal

Wissenschaftlicher Ergebnisbericht V1 bis V2.4

Von der Headerinventur zum Distributed Tensor Observatory

Eingefrorener Evidenzstand. V2.5-Messdaten sind nicht enthalten. Der Bericht trennt Hersteller- und Headerinformationen, gemessene Tensorstatistik, technische Gerätevalidierung und präregistrierte statistische Entscheidungen.

Erzeugt: 2026-08-02 12:17 UTC

Kurzfassung

Der Windkanal untersucht, welche belastbaren Informationen über ein sehr großes Open-Weights-Modell ohne Inferenz und ohne moderne Beschleuniger gewonnen werden können. Die Entwicklung verlief von einer vollständigen Safetensors-Headerinventur über reproduzierbare OpenCL-Reduktionen bis zur präregistrierten Analyse einer Darstellungskette für die Expertgeometrie.

Das Modellpaket umfasst 497.220 Tensoren in 96 Shards mit geschätzt 1,56 TB. Der Architecture Atlas beschreibt 93 Textlayer: einen dichten Eingangslayer und 92 MoE-Layer. Jeder MoE-Layer konfiguriert 896 Routed Experts, zwei Shared Experts und Top-16-Auswahl. Diese Aussagen stammen aus Begleitdateien und Tensorheadern, nicht aus Gewichtsmessungen.

Die Intel HD 4000 war über Beignet/OpenCL erreichbar. Zehn Kontrollbenchmarks stimmten mit CPU überein. In V1.2 bestanden 108 Kernelkonfigurationen; der maximale Compute-only-Speedup betrug 12,45x, während Ende-zu-Ende- und Pipelinewerte deutlich niedriger blieben. Der physische Sieben-Node-Canary erreichte median 6,066x Speedup bei 86,7 Prozent paralleler Effizienz und dokumentierter Nebenlast auf node4.

Für Layer 1 und 46 liegen vollständige OpenCL-A-, OpenCL-B- und CPU-Messungen der sechs Expertentensorfamilien vor. Die exakt identischen OpenCL-Wiederholungen entlasten die Analysepipeline. CPU/OpenCL ist für die kontinuierlichen Darstellungen stabil; zwischen Layer 1 und 46 ist bereits die normalisierte Featurematrix in 0 von 6 Familien stabil. Grobe kNN- und Mutual-R1-Signaturen können dennoch bestehen. Das zeigt Verdichtung durch Diskretisierung, nicht gleiche Geometrie.

Zentrale Negativabgrenzung: Statistische Gewichtsgeometrie belegt weder funktionale Ähnlichkeit noch semantische Expertenrollen, Austauschbarkeit oder Redundanz.

1. Evidenzvertrag und Scope

Alle Aussagen sind in einer Claim-Evidence-Matrix mit Claim-ID, Version, Status, Quelle, Tabelle, Locator, Quelldatenbank, SHA256 und Limitierung registriert. Die 13 originalen DuckDB-Dateien bleiben im Paket erhalten. Relevante Tabellen liegen zusätzlich als Parquet und, bei kleinen Tabellen, als CSV vor. Ein Datadictionary beschreibt 680 Spalten.

V2.5 ist nicht Teil dieses Ergebnisstands. Im Paket befinden sich ausschließlich leere V2.5-Vorlagen. Präregistrierte Tiefenmarker dürfen als Zukunftsannotation erscheinen; laufende oder spätere Messwerte dürfen keine V1-bis-V2.4-Aussage verändern.

Nicht enthalten sind Modellgewichte, Aktivierungsdaten, semantische Expertrollen, layerübergreifende Expert-ID-Korrespondenzen, UMAP, t-SNE, 3D-Clusteransichten oder nachträglich gelockerte Schwellen.

Prüfregel: nicht testbar ist nicht dasselbe wie nicht unterstützt. Prognosen sind als Prognosen ausgewiesen; Canary-Ergebnisse sind keine vollständigen Modellläufe.

2. Architecture Atlas

Der Atlas hebt Informationen aus config.json, Index, Modellkarte, README, Lizenz, Tokenizer- und weiteren Begleitdateien sowie den Safetensors-Headern. Architekturinformationen und später abgeleitete Tensorstatistiken bleiben getrennt.

Layer 1 und 46 sind nach dem Atlas strukturell homolog. Diese Homologie ist ein Vergleichsgate, keine Ähnlichkeitsaussage.

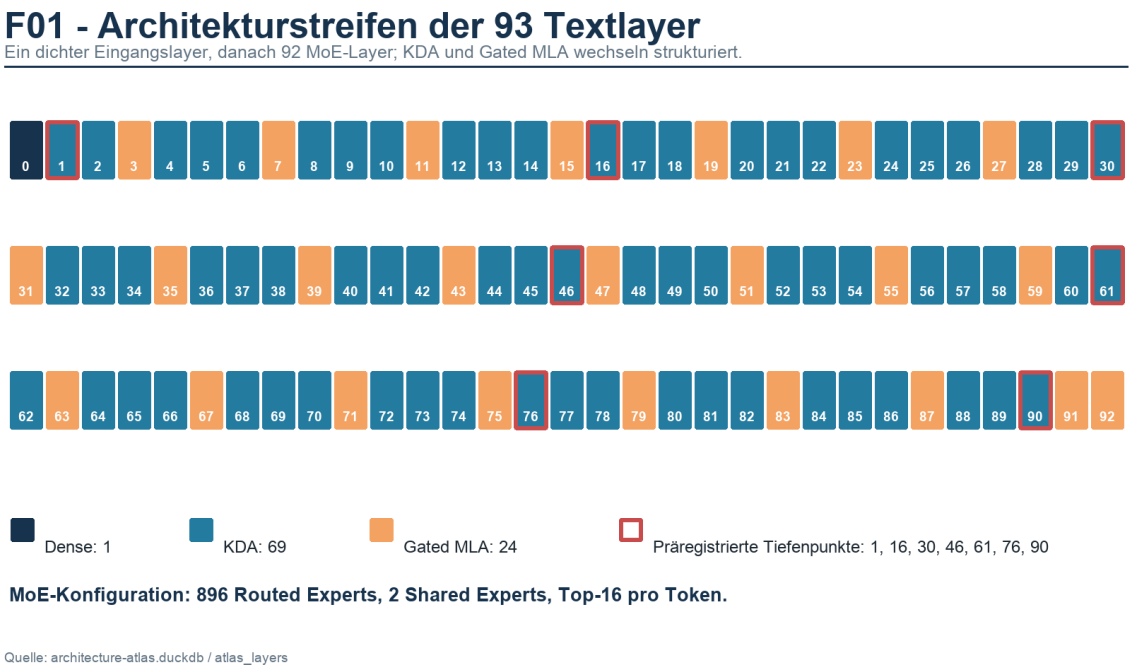

Abbildung 1: Architekturstreifen der 93 Textlayer. Datenquelle: atlas.atlas_layers.

3. Messumfang

Die Headerinventur ist vollständig. Die vollständige Expertprofilierung umfasst bis V2.4 zwei von 92 MoE-Layern, also 10.752 einzigartige Expertentensoren beziehungsweise 32.256 Geräte-Messinstanzen für OpenCL A, OpenCL B und CPU.

Der kleine Anteil ist methodisch gewollt: zuerst Reproduzierbarkeit und Vergleichslogik prüfen, erst dann zusätzliche Layer vermessen.

Abbildung 2: Inventar und Messabdeckung. Datenquelle: v1.inventory plus V2.1/V2.4.

4. Gerätevalidierung

V1.1 bestätigte die grundsätzliche OpenCL-Erreichbarkeit. V2.4 prüfte nicht nur einzelne Summen, sondern vollständige Darstellungsklassen. Für die Distanzmatrizen liegen die maximalen relativen Frobenius-Abweichungen weit unter 0,01; die minimale Spearman-Korrelation bleibt über der präregistrierten 0,999-Schwelle.

Die Geräteübereinstimmung ist eine Aussage über numerische Stabilität der definierten Pipeline, nicht über die Eignung der Hardware für Inferenz.

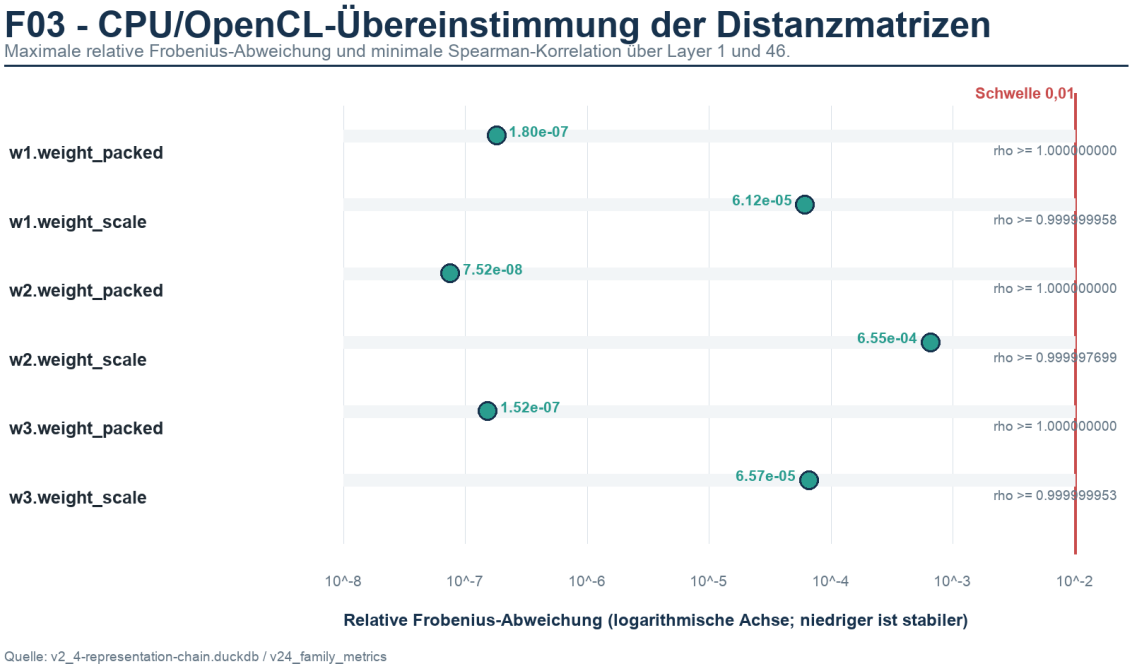

Abbildung 3: CPU/OpenCL-Übereinstimmung. Datenquelle: v2_4.v24_family_metrics.

5. Mikrobenchmark und Pipeline

Die Leistungslandschaft zeigt, dass Stichprobengröße, Workgroup und Batch den Compute-only-Speedup stark verändern. Das Maximum von 12,45x darf nicht als Pipelinebeschleunigung gelesen werden.

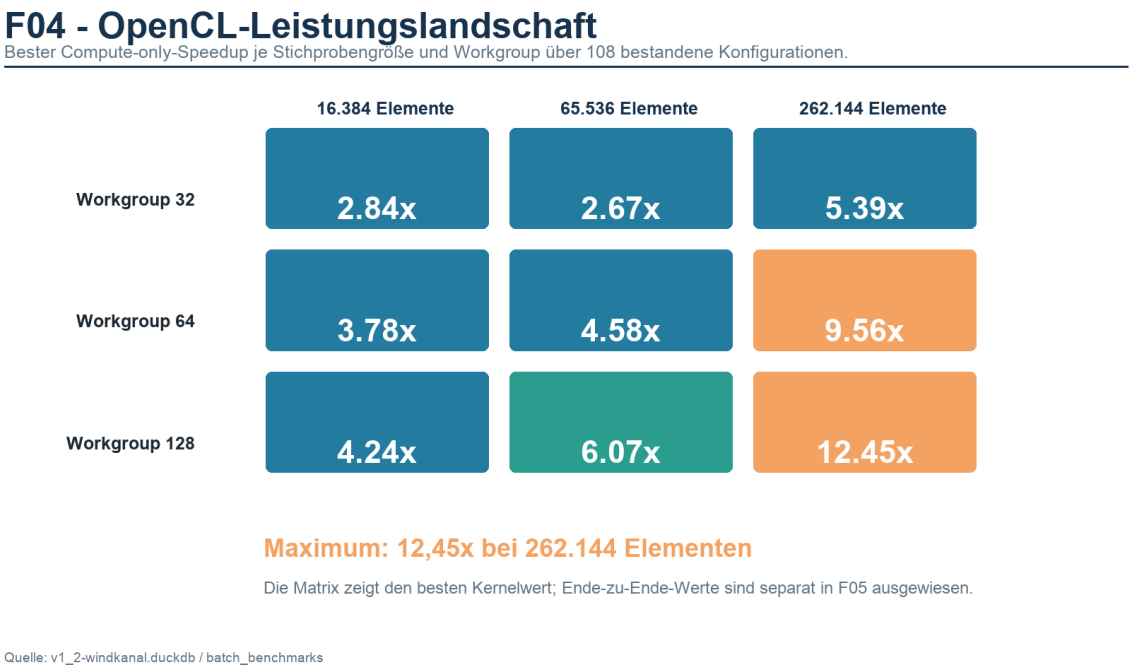

Abbildung 4: OpenCL-Leistungslandschaft. Datenquelle: v1_2.batch_benchmarks.

6. Speedup-Zerlegung

Bei 262.144 Elementen sind Kernel, Ende-zu-Ende-Benchmark und 20-Sample-Pipeline klar getrennt. Die Pipeline erreicht 2,74x gegenüber CPU; bei kleinen Stichproben kann der Overhead den Beschleunigungsvorteil aufheben.

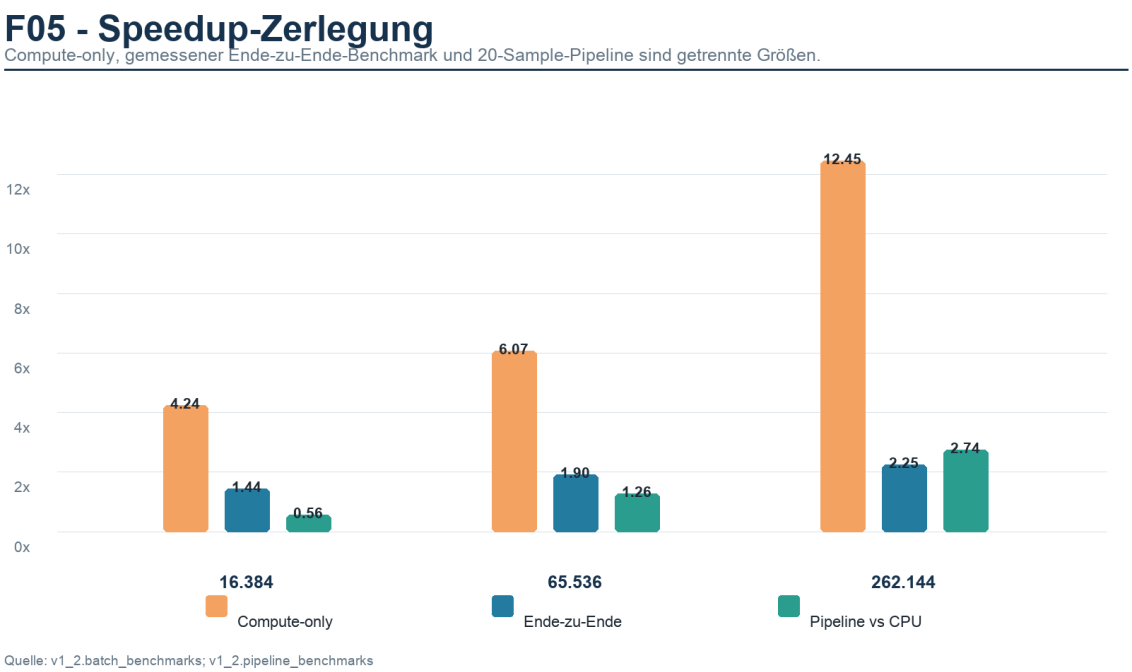

Abbildung 5: Kernel- und Ende-zu-Ende-Speedup. Datenquelle: v1_2.pipeline_benchmarks and batch_benchmarks.

7. Sieben-Node-Canary

Der physische Canary umfasst sieben Nodes und fünf Wiederholungen. Der mediane Speedup beträgt 6,066x, die mediane parallele Effizienz 86,7 Prozent. node4 lief unter der dokumentierten Last eines Hauptcrawlers; der Canary bildet damit reale Nebenlast ab.

Die V1.3-Prognose von 79,05 Stunden auf einem Node und 11,29 Stunden auf sieben ideal verteilten Nodes bleibt eine Prognose und wird nicht mit dem Canary gleichgesetzt.

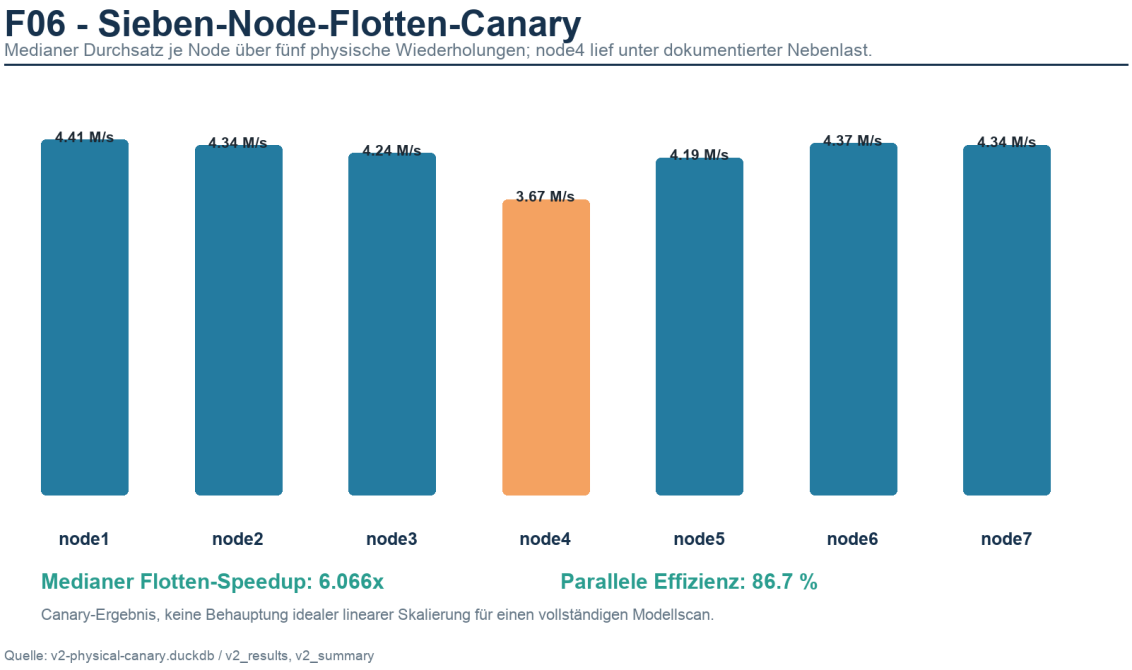

Abbildung 6: Sieben-Node-Flotten-Canary. Datenquelle: v2.v2_results and v2_scaling.

8. Familienvertrag

Jeder untersuchte MoE-Layer enthält sechs strikt getrennte Expertentensorfamilien mit je 896 Mitgliedern. Gepackte Gewichte und Quantisierungsskalen werden nicht in einen gemeinsamen Merkmalsraum gezwungen.

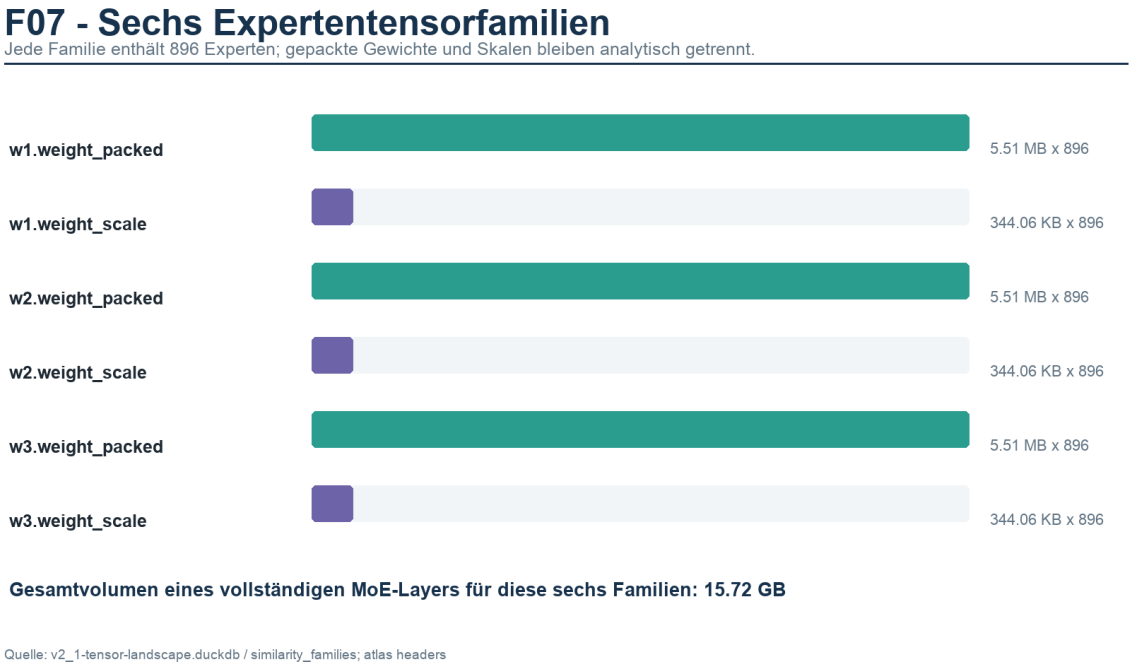

Abbildung 7: Sechs Expertentensorfamilien. Datenquelle: atlas.architecture_tensors.

9. V2.2 bis V2.4 - Stabilität der Darstellung

V2.2 zeigte exakte technische Reproduzierbarkeit für 5.376 Expertentensoren in zwei parameteridentischen OpenCL-Läufen. Die präregistrierte familienübergreifende Konsenshypothese wurde nicht unterstützt.

V2.3 erweiterte die Evidenz um Layer 46 und vollständige CPU-Varianten. Vier von sechs Familien erfüllten die Paarregel; das vollständige numerische Graph-Gate wurde nicht erreicht. Die Near-tie-Nachdiagnose erklärte nur 14 von 522 Rang-1-Wechseln bei 1 Prozent Toleranz.

V2.4 zerlegte die Transformation von der normalisierten Featurematrix über kontinuierliche und permutationsinvariante Darstellungen bis zu diskreten Graphen. CPU/OpenCL bleibt in den frühen kontinuierlichen Stufen 6 von 6 stabil. Zwischen Layer 1 und 46 ist die Featureebene 0 von 6 stabil; Distanz, Profile und Spektrum jeweils 2 von 6.

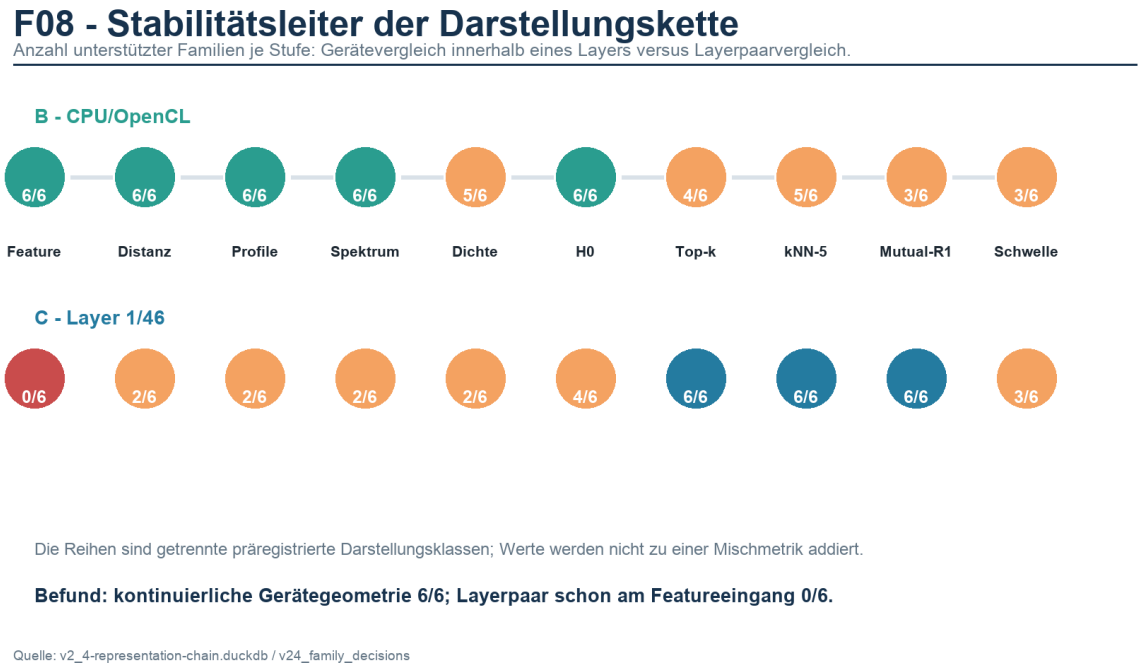

Abbildung 8: Stabilitätsleiter der Darstellungskette. Datenquelle: v2_4.v24_family_decisions.

10. V2.3-Schwellenbefund

Die zwei weight_scale-Familien w1 und w2 überschreiten die registrierte mittlere Signaturdistanz von 0,10. w3.weight_scale bleibt unter der Distanzschwelle. Das vollständige Paar-Gate berücksichtigt zusätzlich p-Wert und zwingende Gerätebedingungen.

Der hohe Spearman-Wert der Distanzordnungen und die lokalen Rangwechsel sind kein Widerspruch: Die diskrete Projektion kann empfindlicher reagieren als der kontinuierliche Raum.

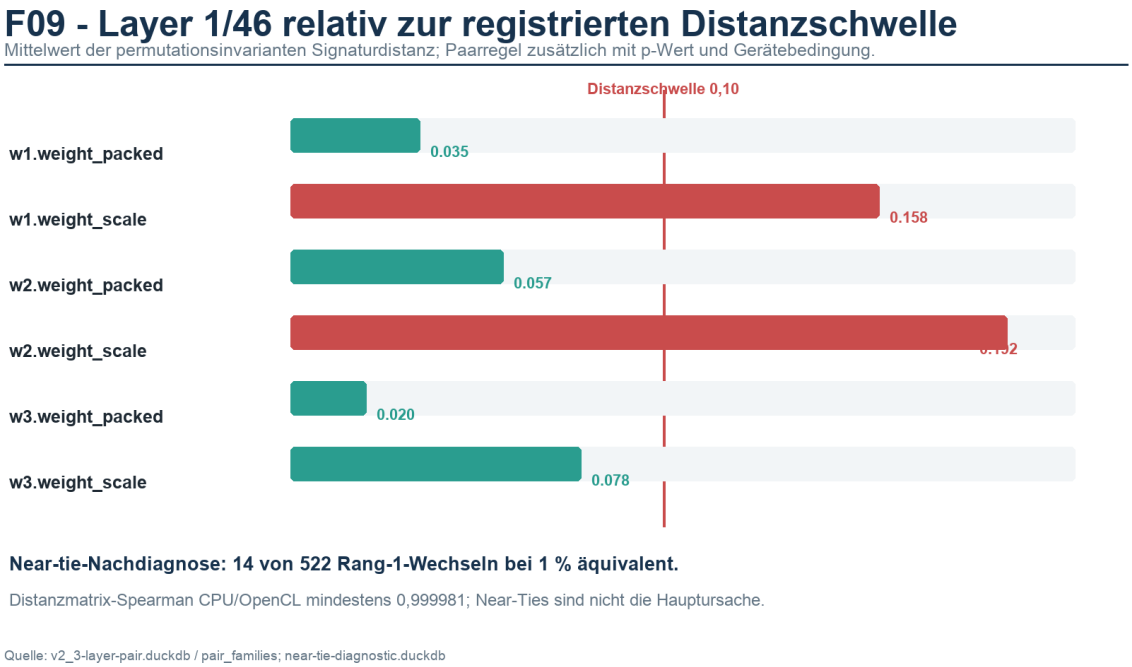

Abbildung 9: Layer 1/46 relativ zu registrierten Schwellen. Datenquelle: v2_3 near-tie and pair evidence.

11. V2.4-Abbruchkarte

Im Geräteblock ist der Verlauf bis H0 in allen sechs Familien stabil; lokale Dichte und diskrete Graphstufen verlieren familienabhängig Unterstützung. Im Layerpaarblock brechen die kontinuierlichen Ausgangsdarstellungen früh, während Top-k, kNN-5 und Mutual-R1 jeweils 6 von 6 unterstützen.

Diese scheinbare Erholung ist keine Reparatur der Geometrie. Sie zeigt, dass grobe diskrete Signaturen Unterschiede verdichten oder verdecken können. Deshalb werden Graphen als Interpretationen kontinuierlicher Räume behandelt.

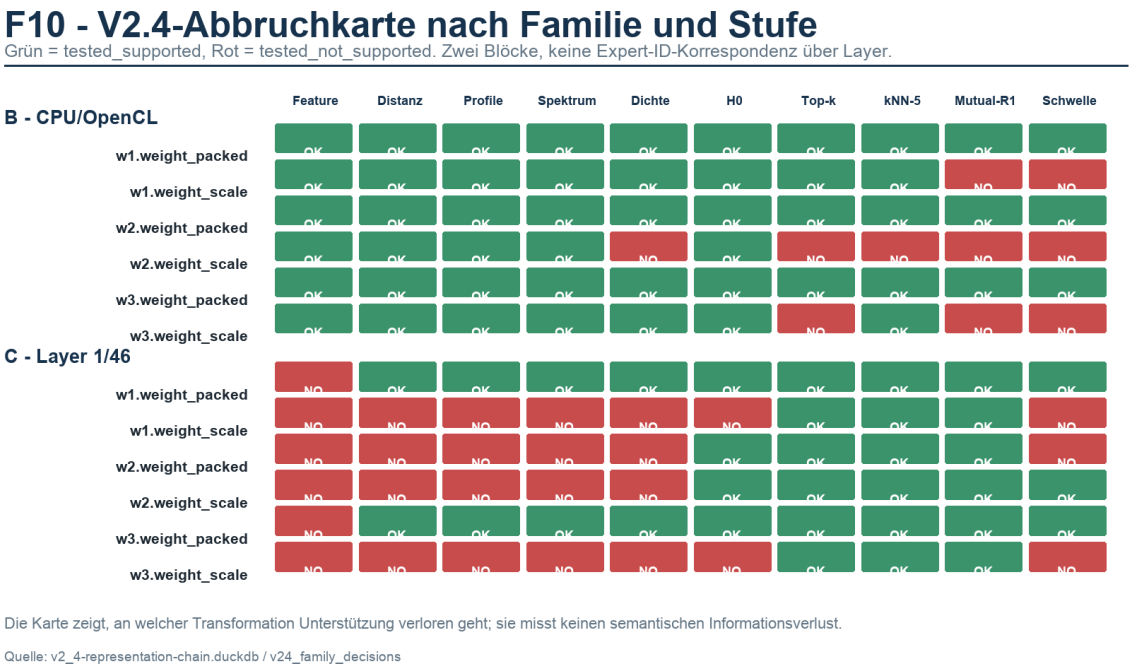

Abbildung 10: V2.4-Abbruchkarte nach Familie und Stufe. Datenquelle: v2_4.v24_family_decisions.

12. Belastbare Gesamtaussage

Der Windkanal zeigt, dass alte integrierte GPUs reproduzierbare, massiv wiederholte numerische Reduktionen auf lokal vorliegenden Tensoren übernehmen können. Die Hardware ersetzt keine moderne Inferenzplattform; sie ermöglicht einen methodisch kontrollierten Zugriff auf statistische Gewichtslandschaften unter knappen Ressourcen.

Das stabilste Objekt ist nicht automatisch der kNN-Graph. Innerhalb eines Layers ist die kontinuierliche CPU/OpenCL-Geometrie sehr stabil. Zwischen Layer 1 und 46 sind die kontinuierlichen Ausgangsräume überwiegend verschieden, obwohl grobe diskrete Graphsignaturen ähnlich aussehen können. Der methodische Beitrag liegt in der expliziten Prüfung jeder Transformationsstufe.

Zwei Layer reichen nicht für allgemeine Tiefeninvarianz. Die Ergebnisse autorisieren eine gezielte Erweiterung, aber keine Aussage über Kimi als Ganzes und keine semantische Interpretation einzelner Experten.

Anhang A - Claim-Evidence-Matrix

ID	Version / Status	Claim	Evidenz / Grenze
C01	V1 tested_supported	Das Headerinventar umfasst 497.220 Tensoren in 96 Safetensors-Shards.	v1.inventory Struktur aus Headern; keine Aussage über Gewichtswerte.
C02	V1 tested_supported	Große Tensoren wurden kontrolliert als skipped_too_large_for_v1 markiert, ohne den Lauf abzubrechen.	v1.results V1 liest nur kleine Tensoren vollständig.
C03	V1.1 tested_supported	Die Intel HD 4000 war über Beignet/OpenCL erreichbar; 10 von 10 OpenCL-Kontrollbenchmarks stimmten mit CPU überein.	v1_1.device_benchmarks Kleiner Kontrollsatz, keine vollständige Modellvermessung.
C04	V1.2 tested_supported	Alle 108 Kernelkonfigurationen bestanden; der Compute-only-Speedup stieg bis 12,45 bei 262.144 Elementen.	v1_2.batch_benchmarks Kernel-Speedup ist nicht gleich Ende-zu-Ende-Speedup.
C05	V1.3 tested_supported	Der blockweise Scanner ist innerhalb eines Tensors fortsetzbar und beendete Layer 0-1 ohne Fehler oder stillen CPU-Fallback.	v1_3.scan_runs Vollständig vermessen wurden nur Layer 0 und 1.
C06	V1.3 forecast_supported	Die Medianprognose beträgt 79,05 Stunden auf einem Node und 11,29 Stunden auf sieben ideal verteilten Nodes.	v1_3.scan_forecasts Prognose, kein vollständiger Sieben-Node-Gesamtlauf.
C07	V2 tested_supported	Der physische Sieben-Node-Canary erreichte median 6,066-fachen Speedup bei realer Nebenlast.	v2.v2_summary Canary-Payload; node4 lief unter dokumentierter Produktionslast.
C08	Atlas tested_supported	Kimi K3 besitzt 93 Textlayer: einen dichten Eingangslayer und 92 MoE-Layer, davon 69 KDA und 24 Gated MLA.	atlas.atlas_layers Architekturmetadaten, keine Tensorstatistik.
C09	Atlas tested_supported	Jeder MoE-Layer konfiguriert 896 Routed Experts, zwei Shared Experts und Top-16-Auswahl.	atlas.atlas_layers Konfiguration und Headerstruktur, keine Aktivierungsbeobachtung.
C10	V2.1 tested_supported	Layer 1 enthält sechs strikt getrennte Expertentensorfamilien mit je 896 Mitgliedern.	v2_1.similarity_families Keine familienübergreifenden Kanten zulässig.
C11	V2.2 tested_supported	Zwei parameteridentische OpenCL-Läufe von Layer 1 stimmen für 5.376 Tensoren exakt überein.	v2_2.stability_evidence Technische Reproduzierbarkeit ist keine funktionale Ähnlichkeit.
C12	V2.2 tested_not_supported	Die präregistrierte familienübergreifende Konsenshypothese wird nicht unterstützt.	v2_2.stability_evidence Negativbefund im definierten Merkmals- und Graphvertrag.
C13	V2.3 tested_supported	Layer 1 und 46 sind laut Atlas strukturell homolog; beide Layer wurden je mit OpenCL A/B und CPU vollständig vermessen.	v2_3.pair_runs Homologie autorisiert Vergleich, beweist aber keine statistische Ähnlichkeit.
C14	V2.3 tested_not_supported	Vier von sechs Tensorfamilien erfüllen die Paarähnlichkeitsregel, das vollständige numerische Graph-Gate jedoch nicht.	v2_3.pair_families Die zwingende Gerätebedingung darf nicht durch 4/6 ersetzt werden.
C15	V2.3 tested_not_supported	Nur 14 von 522 Rang-1-Wechseln sind bei 1 Prozent Toleranz als Near-Ties äquivalent; Near-Ties erklären die Wechsel nicht hauptsächlich.	v2_3_near_tie.near_tie_thresholds Deskriptive Nachdiagnose; ändert V2.3 nicht.
C16	V2.4 tested_supported	Alle 120 OpenCL-A/B-Ableitungen der Darstellungskette sind exakt identisch.	v2_4.v24_integrity Entlastet die Analysepipeline, nicht jede wissenschaftliche Hypothese.
C17	V2.4 tested_supported	CPU/OpenCL ist für Featurematrix, Distanzmatrix, Profile, Spektrum und H0 in allen sechs Familien stabil.	v2_4.v24_family_decisions Lokale Dichte und diskrete Scale-Graphen zeigen spätere Abbrüche.

ID	Version / Status	Claim	Evidenz / Grenze
C18	V2.4 tested_not_supported	Layer 1 und 46 sind auf Featureebene 0/6 und bei Distanz, Profil und Spektrum je 2/6 stabil; allgemeine Layerstabilität wird nicht unterstützt.	v2_4.v24_family_decisions Ein Layerpaar erlaubt keine allgemeine Tiefeninvarianz.
C19	V2.4 tested_supported_with_warning	Grobe kNN- und Mutual-R1-Layersignaturen bestehen 6/6, obwohl kontinuierliche Ausgangsräume überwiegend verschieden sind.	v2_4.v24_family_decisions Graphstabilität wird als Verdichtung, nicht als gleiche Geometrie interpretiert.

Anhang B - Quelldatenbanken und Prüfbarkeit

v1: v1-windkanal.duckdb

SHA256 b1a8f6cb7ee974cb9d5a9428ee1d4ce7d80a8f09e6df327fbba49548b3eb5b79 - Größe 42.74 MB

v1_1: v1_1-windkanal.duckdb

SHA256 059bd23ba4746f1c1c06e273a4a3087bfdd2b1b34f1d46e4456ce6a6376495b6 - Größe 2.63 MB

v1_2: v1_2-windkanal.duckdb

SHA256 4f3a16937b37840a361811314c689c0fecc5a92c5fcfdb2b0dffbb3111e65b90 - Größe 4.21 MB

v1_3: v1_3-windkanal.duckdb

SHA256 bbf6d0aab9c5fc487123a07168a2e9d509bea5ba62d8d50052e558a1347fb93b - Größe 395.59 MB

v2: v2-physical-canary.duckdb

SHA256 7432f52f9b57bf08719c7c472a3845250c247a51f2c3212fe3a8f190ae31c5d4 - Größe 2.37 MB

atlas: architecture-atlas.duckdb

SHA256 7d4ab0c64886c218786488291eae60a58b61929b5807649d0ef99cb6d9f7e61e - Größe 106.70 MB

v2_1: v2_1-tensor-landscape.duckdb

SHA256 53ebdff15ce601b5cda8983e15166812bfdf3e0a692b53580c7b2b3f182062b6 - Größe 109.59 MB

v2_2_source: v2_2-source-v1_3-with-cpu.duckdb

SHA256 bb4eb832278ccfa5185b41deb283cefd2be8067e551296a7515c119ba447b469 - Größe 432.03 MB

v2_2: v2_2-stability.duckdb

SHA256 7d233e7b77900595384c2720d788b0bf14fb386590a6004035aac347a16481c4 - Größe 20.20 MB

v2_3_source: v2_3-source-layer1-46.duckdb

SHA256 2616bee9ca099b72a6cbb6fcdd4353b38bb1a0117af96c94d0286e227ad462bd - Größe 187.71 MB

v2_3: v2_3-layer-pair.duckdb

SHA256 de015f94c1f9eb042ea15e8160e2d2b5b72979a81d0cf95419d1b4be31168d76 - Größe 2.11 MB

v2_3_near_tie: v2_3-near-tie-diagnostic.duckdb

SHA256 485a202d6dd077cd78f3b133187fc98fa5ef15dc876f4eb397ad2fce2d934449 - Größe 8.66 MB

v2_4: v2_4-representation-chain.duckdb

SHA256 82a1a0ff29333f2eda8c03114e3726bf9910d227d0c419ae118184b1b181f93a - Größe 3.16 MB

Das Paket enthält außerdem Parquet-Exporte, kleine CSV-Tabellen, JSON-Zusammenfassungen, Schema, Datadictionary und SHA256SUMS. Die Claim-Evidence-Matrix ist der verbindliche Einstiegspunkt für jede Aussage.

Anhang C - V2.5-Vorlage ohne Daten

Für V2.5 werden in diesem Paket nur leere Layout- und Evidenzvorlagen bereitgestellt. Ein V2.5-Ergebnis darf erst nach einem separat eingefrorenen Inputmanifest, abgeschlossenen Läufen und neuer Claim-Evidence-Prüfung veröffentlicht werden.

V2.5-Messdaten: nicht Bestandteil dieses Berichts.