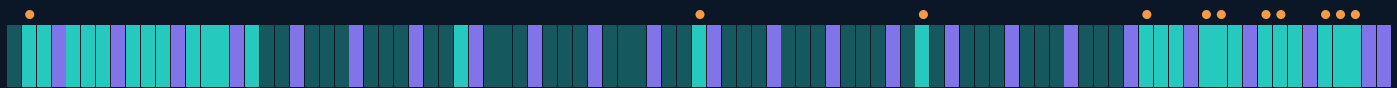


Vom Layerpaar zum Tiefenprofil

Paper II: Dichte, reproduzierbare Gewichtsvermessung eines 93-Layer-MoE-Flagships und Übertragung auf sieben weitere Checkpoints



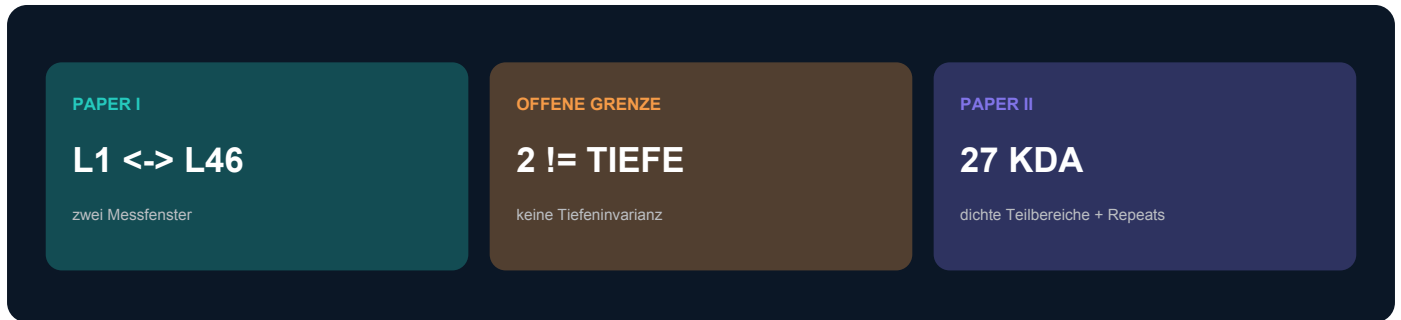
ZUSAMMENFASSUNG

Paper I endete bewusst an einer Zwei-Layer-Grenze. Paper II erweitert denselben eingefrorenen Messvertrag auf 27 Kimi-KDA-Positionen, dichte frühe und späte Tiefenbereiche sowie neun exakt wiederholte Layer. Die versiegelte Evidenz wird hier nicht mehr auf Statuskarten reduziert: 1.320 Primärzellen, 3.960 All-Pair-Zellen, 2.112 Komponentenwerte, 960 Anker-Nähen und 48.384 Wiederholungstensenoren werden als Tiefenfilm, Schwellenlandschaft, Maskierungsfluss und Replikationswand sichtbar. Sieben weitere Checkpoints liefern 984 versiegelte Tensorzeilen über neun registrierte Merkmale. Modellübergreifend bleibt ausschließlich Architekturvergleich zulässig.

27	1.320	48.384	7
KIMI-POSITIONEN	PRIMÄRZELLEN	REPEAT-TENSOREN	KATALOGMODELLE

1 Von Paper I zu Paper II

In Paper I [P1] wurden der Architecture Atlas, der Sechs-Familien-Vertrag und die Darstellungskette bis V2.4 eingefroren. Für Layer 1 und 46 waren die kontinuierlichen Ausgangsräume überwiegend verschieden, während grobe Graphsignaturen Unterschiede verdichten konnten. Daraus folgte ausdrücklich keine allgemeine Aussage über die Tiefe von Kimi. Paper II übernimmt diese Befunde und Schwellen unverändert und untersucht die damals autorisierte Erweiterung. Die Methode wird nicht neu erfunden, sondern über mehr Tiefenpositionen belastet.



Forschungsfragen

- RQ1 Tiefe - welches registrierte Entscheidungsprofil entsteht jenseits des Layerpaars 1/46?
- RQ2 Wiederholung - replizieren technische Wiederholungen die gebundenen Tensoren und Entscheidungsflächen?
- RQ3 Übertragbarkeit - trägt derselbe Lebenszyklus andere, checkpointspezifische Architekturverträge?
- RQ4 Auditierbarkeit - stoppen fehlerhafte Identitäts- oder Commitbindungen die wissenschaftliche Übernahme?

GEERBT, NICHT NEU ANGEPASST

Atlas + Homologie + sechs Familien + Normalisierer + Schwellen -> neue gebundene Tiefenbeobachtungen

Keine Interpolation. Keine Schwellenanpassung. Keine layerübergreifende Expertenidentität. Keine Aktivierungsaussage.

2 Vom Layerpaar zum 93-Layer-Flagship

Paper I sah Kimi durch zwei homologe Messfenster: Layer 1 und 46. Der versiegelte Atlas zeigt jedoch 93 Layer, davon 69 KDA und 24 Gated MLA. Paper II erweitert nur innerhalb des homologen KDA-MoE-Vertrags: 896 Expert-IDs, sechs Tensorfamilien und 5.376 Expertentensoren pro vollständig vermessener Layerposition. Damit wird aus dem Layerpaar ein registriertes Tiefenprofil - weiterhin ohne Modellstart oder Aktivierung.



3 Die Kimi-Tiefenkampagne

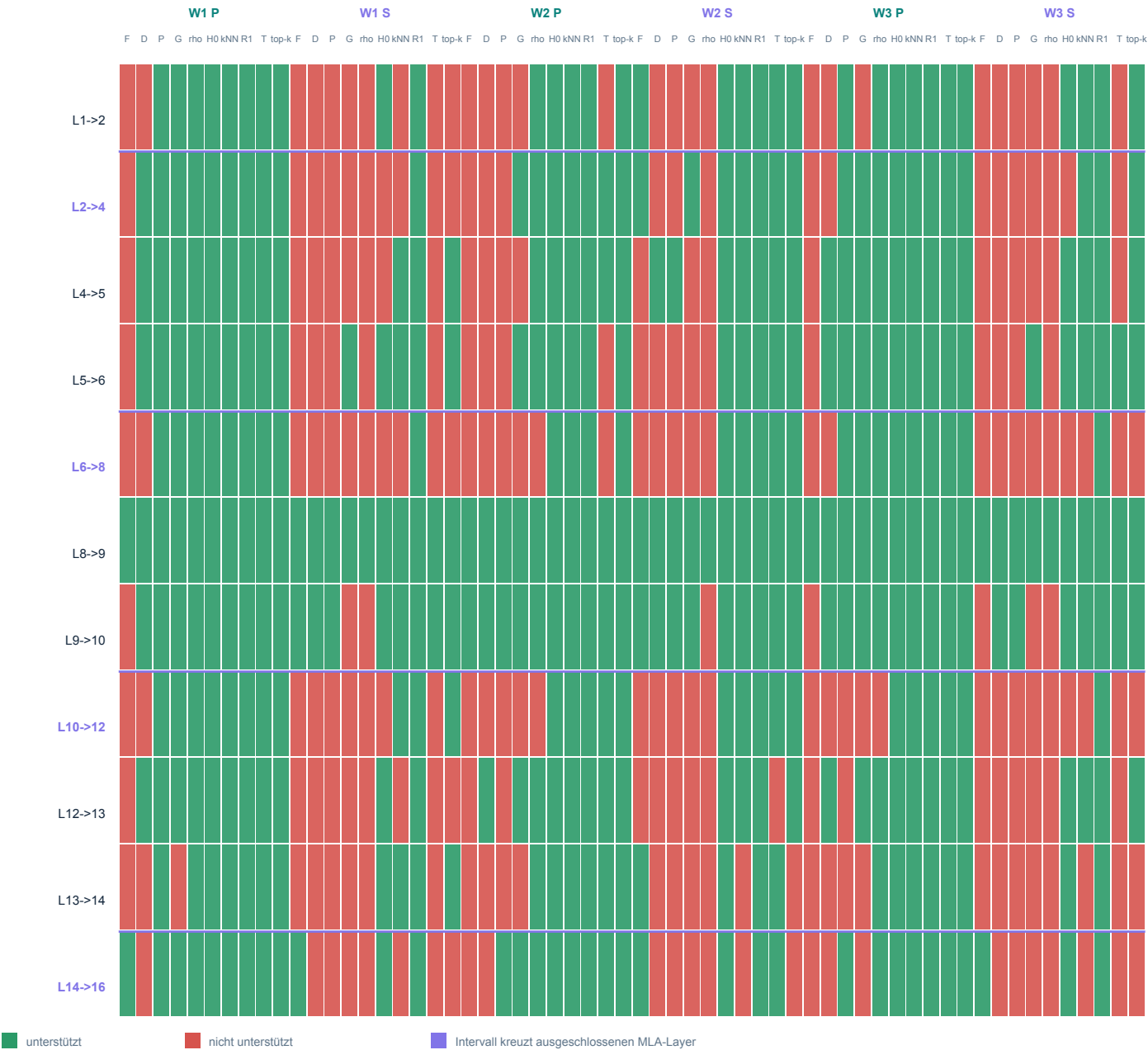
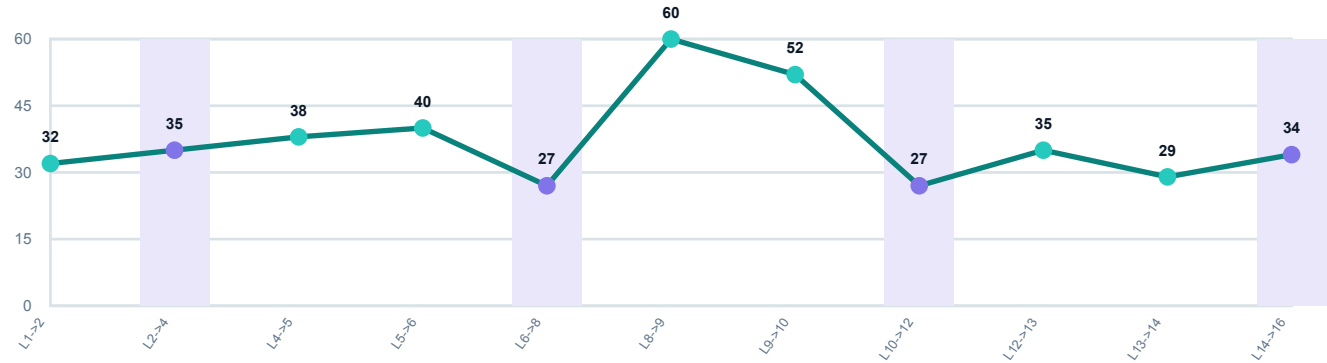
Die Kampagne beantwortet die Schlussfrage aus Paper I schrittweise: erst Tiefenanker, dann eine unveränderte Entscheidungsfläche, danach exakte Wiederholungen und schließlich dichte Kohorten an den späten und frühen Modellrändern. V2.9 bindet zusätzlich die auffälligen späten Grenzen erneut. V2.11 ist nicht Teil dieses Papers.



Früher Tiefenfilm: Layer 1 bis 16

Jede Zeile ist ein reales KDA-Intervall; jede der 60 Spalten ist eine Familie-Darstellung-Zelle. Violett markiert Intervalle, die einen ausgeschlossenen Gated-MLA-Layer überbrücken; innerhalb dieser Lücke wird kein exakter Übergang lokalisiert.

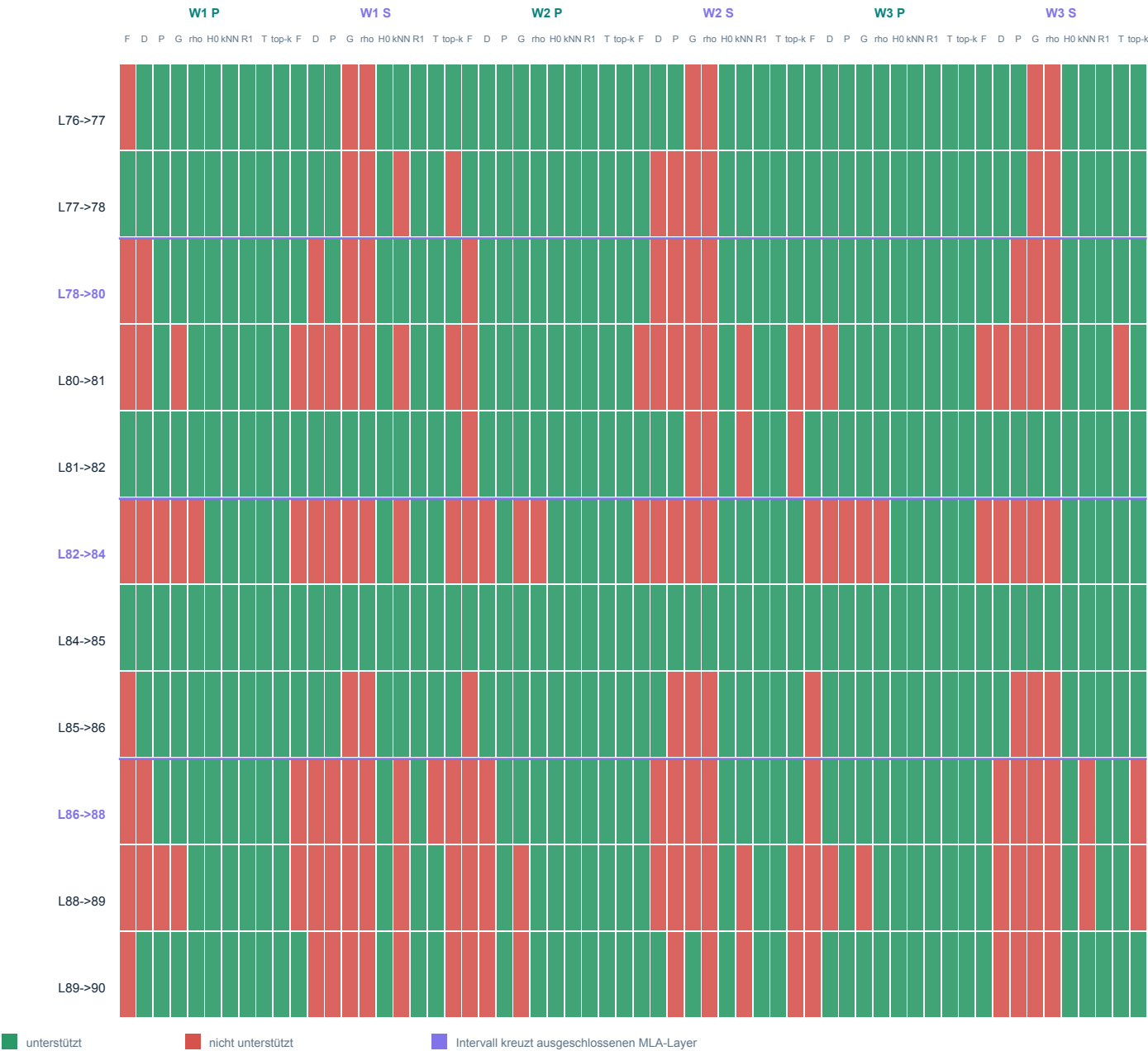
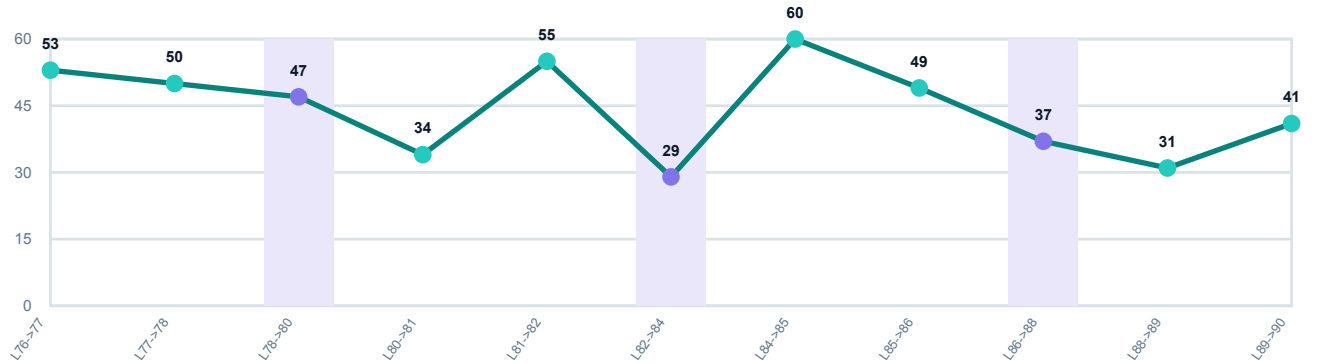
UNTERSTÜTZTE ZELLEN VON 60



Später Tiefenfilm: Layer 76 bis 90

Die dichte späte Kohorte zeigt elf aufeinanderfolgende KDA-Intervalle unter demselben Sechs-Familien-Vertrag. Das Muster ist nicht geglättet: 660 versiegelte Einzelentscheidungen bleiben als Raster sichtbar.

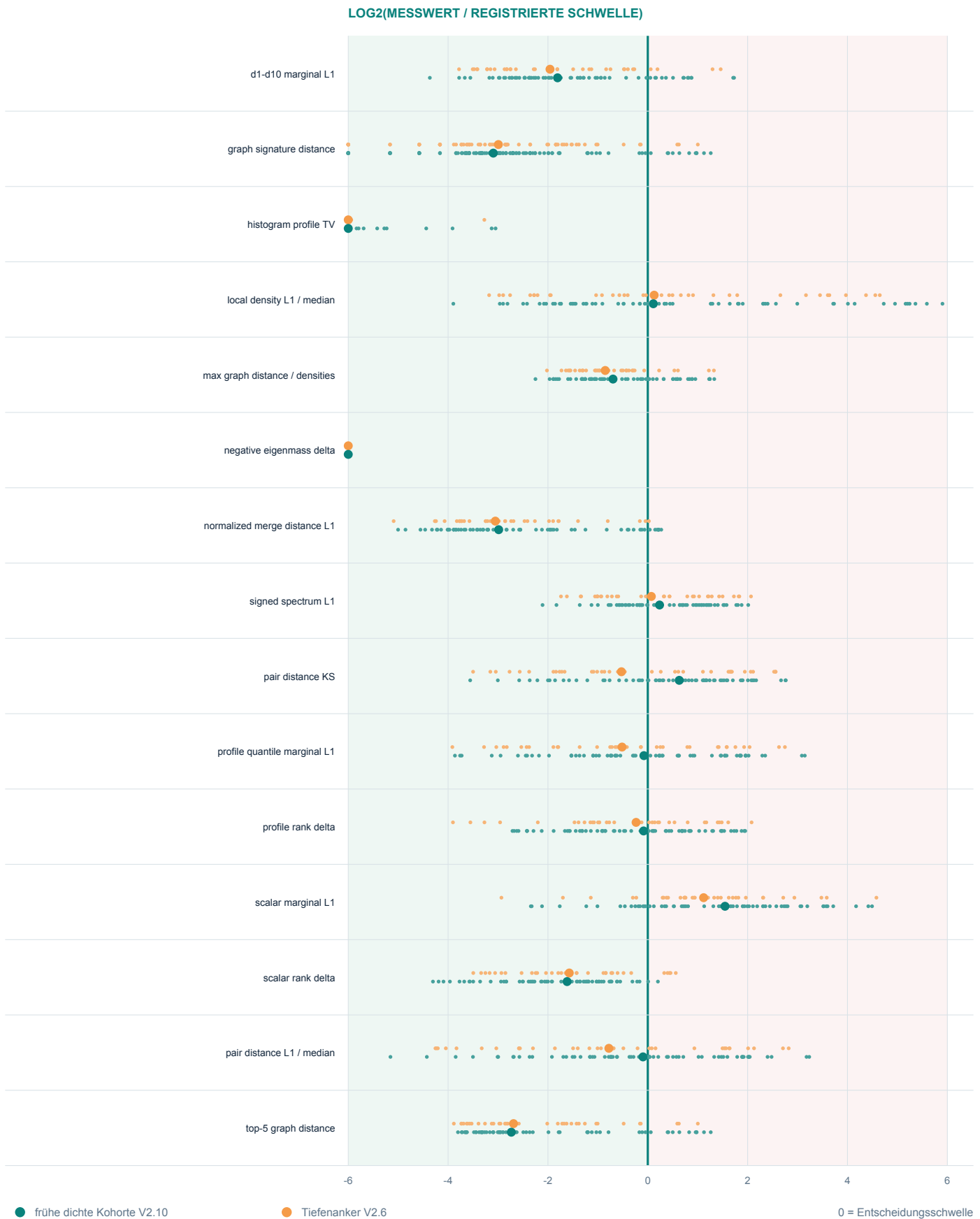
UNTERSTÜTZTE ZELLEN VON 60



■ unterstützt ■ nicht unterstützt ■ Intervall kreuzt ausgeschlossenen MLA-Layer

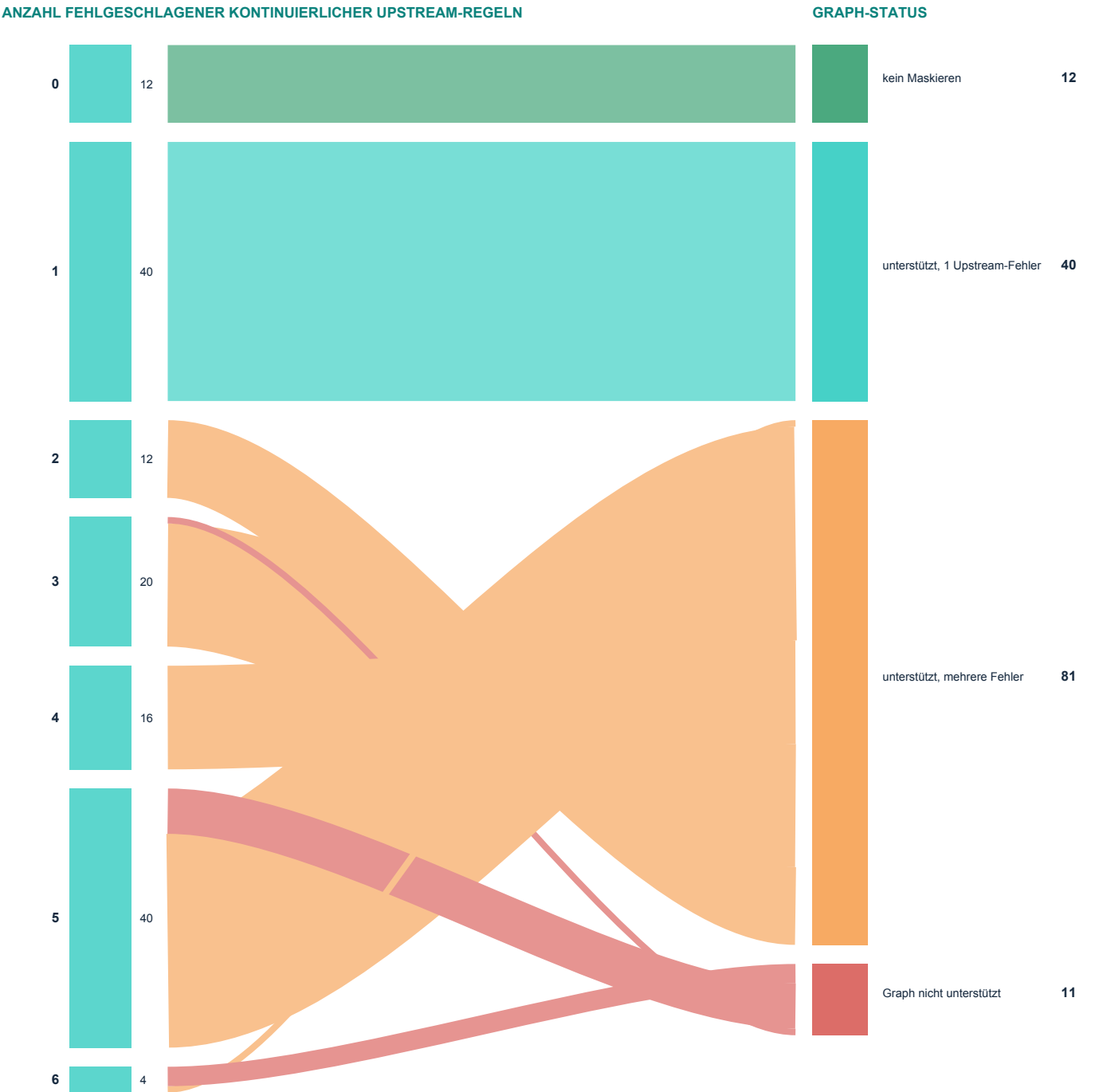
Schwellenlandschaft: Abstand zur registrierten Grenze

Jeder Punkt ist ein versiegeltes Verhältnis aus beobachtetem Komponentenwert und unverändert registrierter Schwelle. Die logarithmische Achse macht sehr verschiedene Metriken gemeinsam lesbar: links von 0 liegt der einzelne Messwert innerhalb der Schwelle, rechts davon außerhalb. Das ändert keine Zellentscheidung; alle Regeln einer Zelle müssen gemeinsam bestehen.



Wenn Graphen Unterschiede maskieren

Graphdarstellungen können unterstützt bleiben, obwohl eine oder mehrere kontinuierliche Upstream-Regeln scheitern. Der Fluss zeigt alle 144 registrierten Zuordnungen: links die Zahl gescheiterter Upstream-Regeln, rechts den daraus beobachteten Graphstatus. Die Breite entspricht der Zahl realer Zellen.

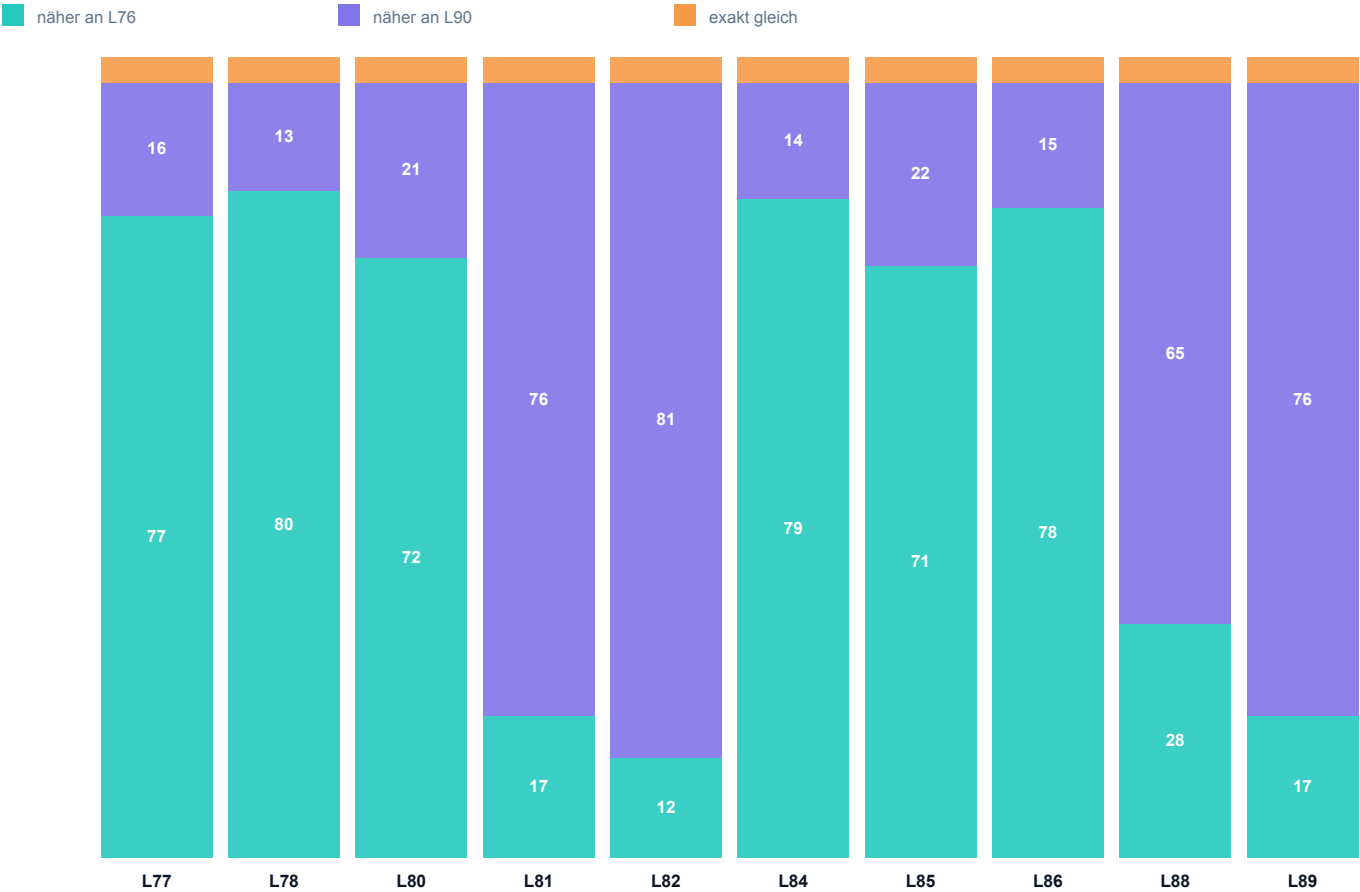


INHALTLICH

Ein grober Graph kann zwei Layer ähnlich erscheinen lassen, obwohl feinere Geometrien verschieden bleiben. Das ist keine Modellschwäche, sondern eine nachgewiesene Informationsreduktion der Darstellung.

Zwischen zwei Ankern: näher an 76 oder 90?

Für zehn späte Zwischenlayer wurden 96 Familien-, Darstellungs- und Metrikkombinationen jeweils gegen die Anker 76 und 90 beschrieben. Die gestapelten Balken zeigen nur, welcher Anker den kleineren registrierten Abstand besitzt. Die Zeilen sind keine unabhängigen Stichproben und begründen keine Biografie einzelner Experten.



531

näher an L76

399

näher an L90

30

exakt gleich

Replikationswand: nicht ähnlich, sondern identisch

Neun präregistrierte KDA-Layer wurden vollständig wiederholt. Jeder Ring steht für 5.376 von 5.376 bytegleich reproduzierte Expertentensoren. Zusätzlich replizierten 300 von 300 gebundenen Entscheidungszellen. Damit ist die auffällige Tiefenstruktur kein einmaliger Leseartefakt des Messpfads.



Alle neun A/B-Layer bytegenau

V2.7: 120/120 · V2.9: 180/180 Entscheidungszellen identisch

TECHNISCHE AUSSAGE

Bytegleichheit belegt Determinismus des gebundenen Scanners, Geräts und Datenpfads. Sie beweist weder mathematische Fehlerfreiheit noch funktionale Bedeutung des Modells.

Llama 3.2 1B Instruct 4-bit

Alle neun registrierten Merkmale werden innerhalb dieses Checkpoints sichtbar gemacht. Jeder Micro-Glyph fasst die direkt versiegelten Tensorrollen eines Pilotlayers als Minimum, Interquartilsbereich, Median und Maximum zusammen; der Barcode darunter zeigt q50/RMS rolleweise. Skalen sind bewusst modelllokal.

16
LAYER

138
TENSORZEILEN

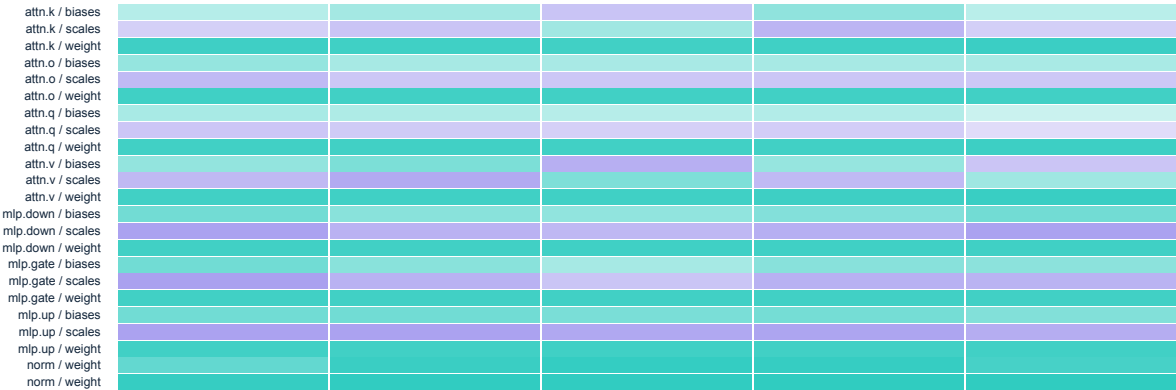
195.8 MiB
MESSPAYLOAD

23
TENSORROLLEN

9-MERKMALS-FINGERABDRUCK · MIN - IQR - MEDIAN - MAX



DIREKTER TENSORROLLEN-BARCODE · Q50/RMS



DeepSeek R1 Distill Qwen 1.5B 4-bit

Alle neun registrierten Merkmale werden innerhalb dieses Checkpoints sichtbar gemacht. Jeder Micro-Glyph fasst die direkt versiegelten Tensorrollen eines Pilotlayers als Minimum, Interquartilsbereich, Median und Maximum zusammen; der Barcode darunter zeigt q50/RMS rolleweise. Skalen sind bewusst modelllokal.

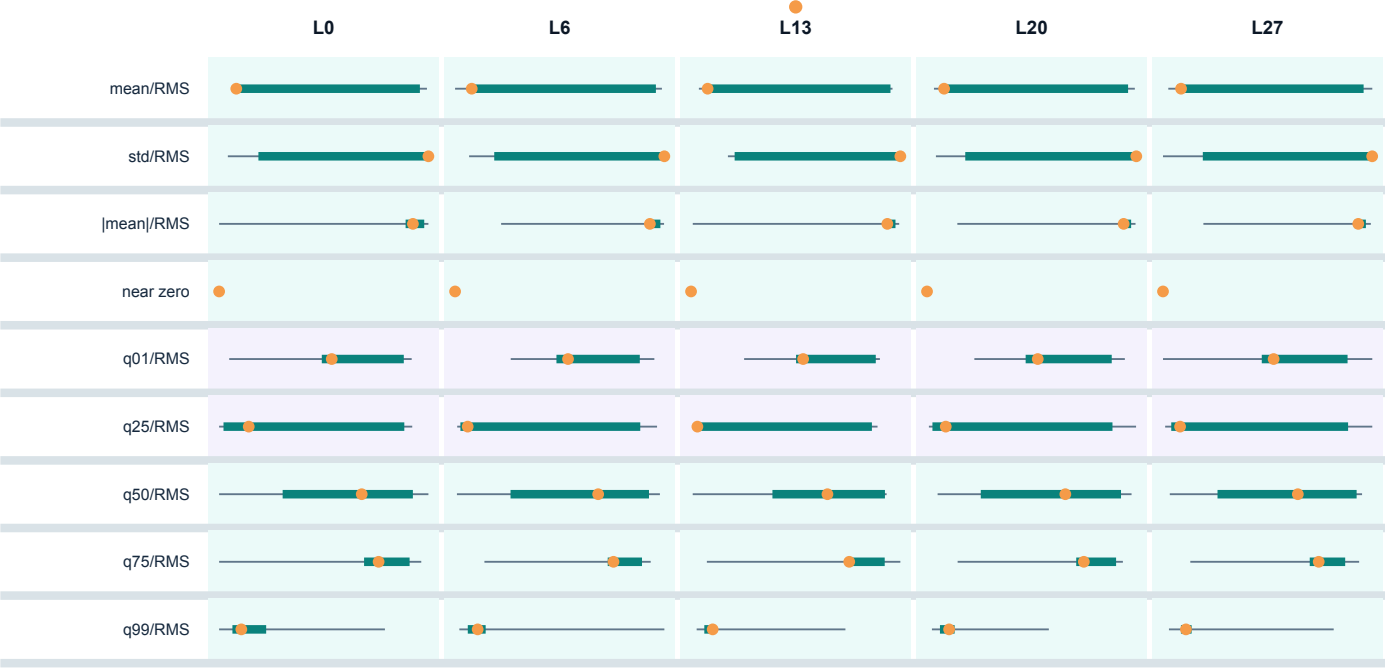
28
LAYER

156
TENSORZEILEN

150.7 MiB
MESSPAYLOAD

26
TENSORROLLEN

9-MERKMALS-FINGERABDRUCK · MIN - IQR - MEDIAN - MAX



DIREKTER TENSORROLLEN-BARCODE · Q50/RMS

attn.k / biases					
attn.k / scales					
attn.k / unquantized					
attn.k / weight					
attn.o / biases					
attn.o / scales					
attn.o / weight					
attn.q / biases					
attn.q / scales					
attn.q / unquantized					
attn.q / weight					
attn.v / biases					
attn.v / scales					
attn.v / unquantized					
attn.v / weight					
mip.down / biases					
mip.down / scales					
mip.down / weight					
mip.gate / biases					
mip.gate / scales					
mip.gate / weight					
mip.up / biases					
mip.up / scales					
mip.up / weight					
norm / weight					
norm / weight					

ModernBERT Embed Base 8-bit

Alle neun registrierten Merkmale werden innerhalb dieses Checkpoints sichtbar gemacht. Jeder Micro-Glyph fasst die direkt versiegelten Tensorrollen eines Pilotlayers als Minimum, Interquartilsbereich, Median und Maximum zusammen; der Barcode darunter zeigt q50/RMS rolleweise. Skalen sind bewusst modelllokal.

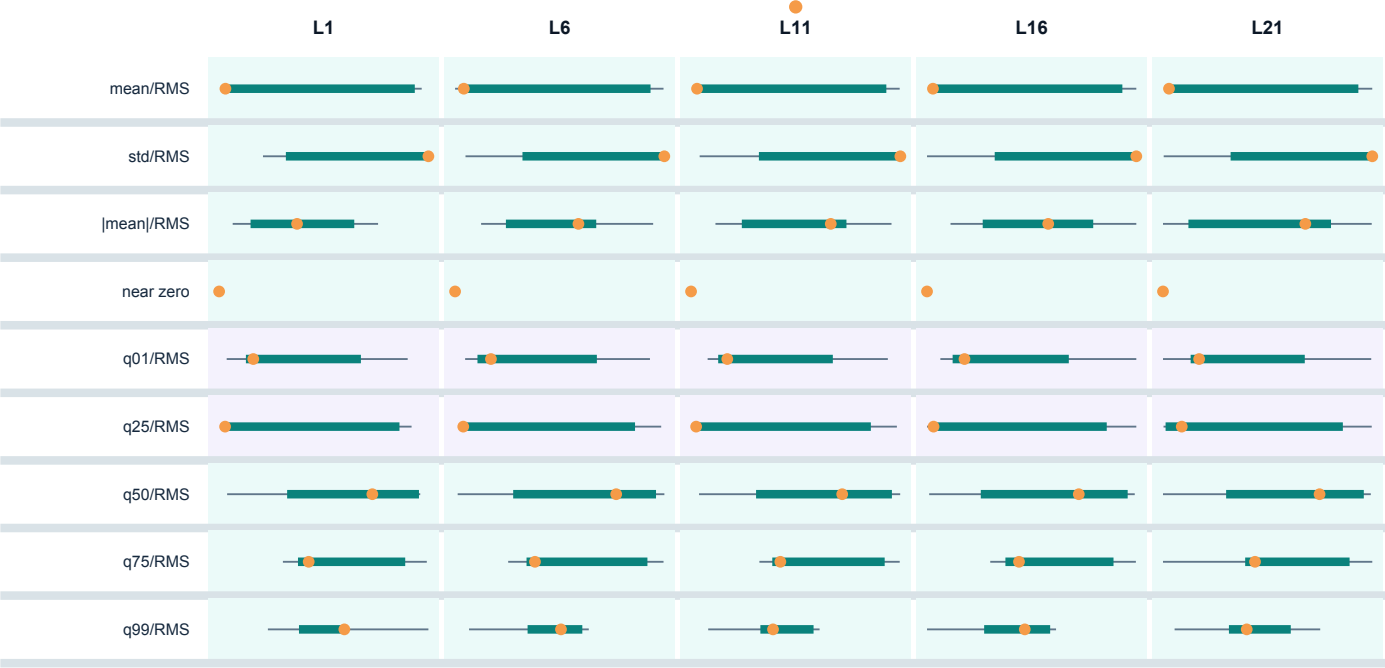
22
LAYER

84
TENSORZEILEN

30.5 MiB
MESSPAYLOAD

14
TENSORROLLEN

9-MERKMALS-FINGERABDRUCK · MIN - IQR - MEDIAN - MAX



DIREKTER TENSORROLLEN-BARCODE · Q50/RMS

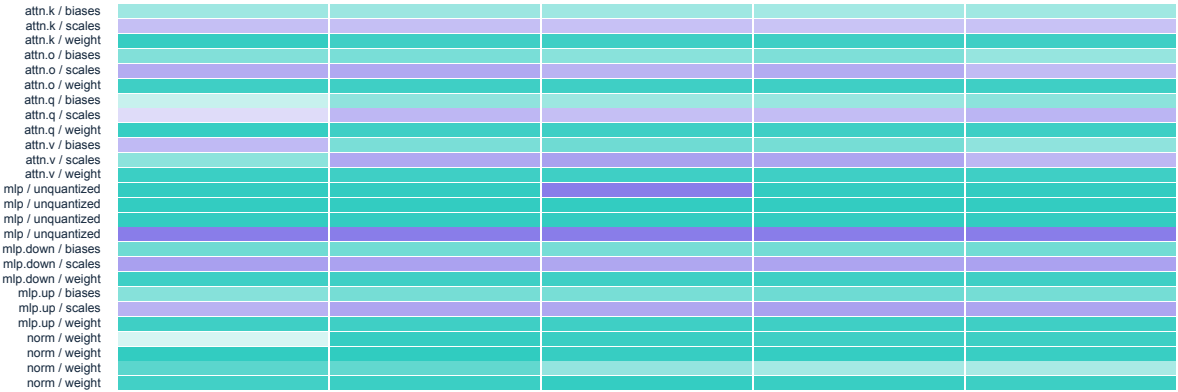
attn.qkv / biases				
attn.qkv / scales				
attn.qkv / weight				
mip / biases				
mip / biases				
mip / scales				
mip / scales				
mip / weight				
mip / weight				
norm / weight				
norm / weight				
other / biases				
other / scales				
other / weight				

Apertus 8B 4-bit

Alle neun registrierten Merkmale werden innerhalb dieses Checkpoints sichtbar gemacht. Jeder Micro-Glyph fasst die direkt versiegelten Tensorrollen eines Pilotlayers als Minimum, Interquartilsbereich, Median und Maximum zusammen; der Barcode darunter zeigt q50/RMS rolleweise. Skalen sind bewusst modelllokal.



DIREKTER TENSORROLLEN-BARCODE · Q50/RMS



Apertus 8B Instruct 8-bit

Alle neun registrierten Merkmale werden innerhalb dieses Checkpoints sichtbar gemacht. Jeder Micro-Glyph fasst die direkt versiegelten Tensorrollen eines Pilotlayers als Minimum, Interquartilsbereich, Median und Maximum zusammen; der Barcode darunter zeigt q50/RMS rolleweise. Skalen sind bewusst modelllokal.

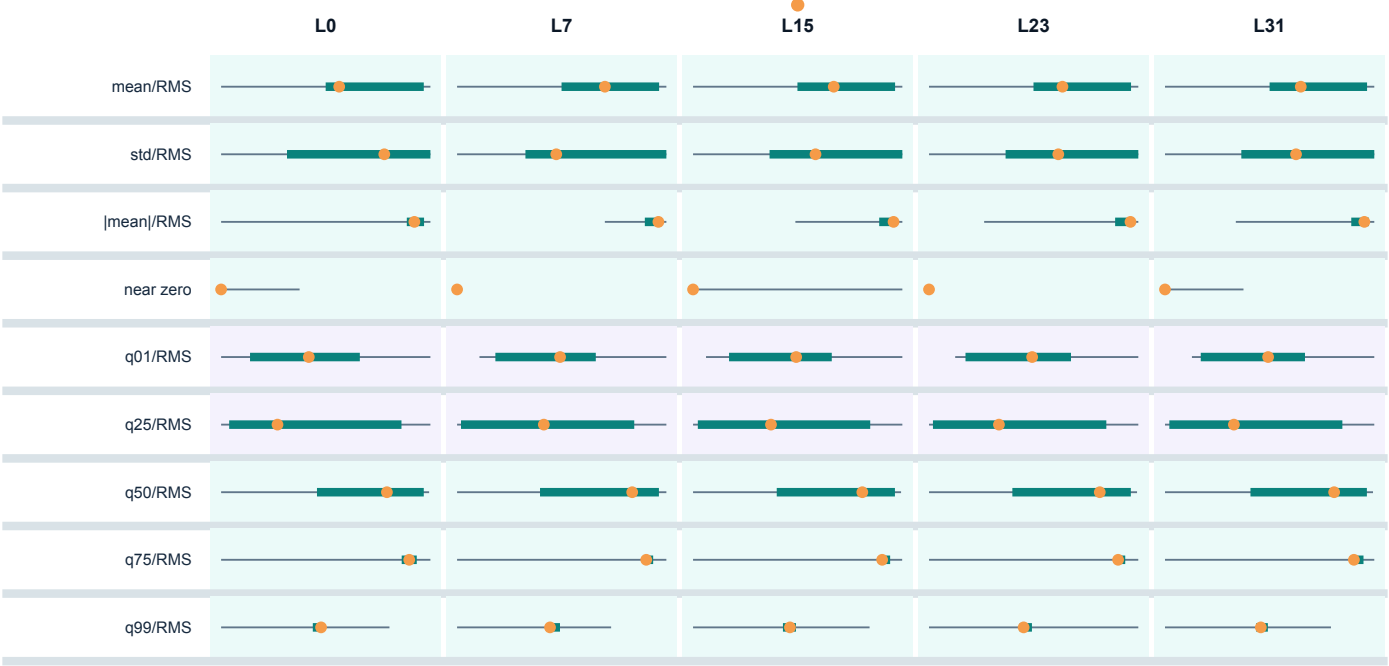
32
LAYER

156
TENSORZEILEN

1.30 GiB
MESSPAYLOAD

26
TENSORROLLEN

9-MERKMALS-FINGERABDRUCK · MIN - IQR - MEDIAN - MAX



DIREKTER TENSORROLLEN-BARCODE · Q50/RMS

attn.k / biases					
attn.k / scales					
attn.k / weight					
attn.o / biases					
attn.o / scales					
attn.o / weight					
attn.q / biases					
attn.q / scales					
attn.q / weight					
attn.v / biases					
attn.v / scales					
attn.v / weight					
mip / unquantized					
mip / unquantized					
mip / unquantized					
mip / unquantized					
mip.down / biases					
mip.down / scales					
mip.down / weight					
mip.up / biases					
mip.up / scales					
mip.up / weight					
norm / weight					
norm / weight					
norm / weight					

Llama 3.1 8B Instruct 4-bit

Alle neun registrierten Merkmale werden innerhalb dieses Checkpoints sichtbar gemacht. Jeder Micro-Glyph fasst die direkt versiegelten Tensorrollen eines Pilotlayers als Minimum, Interquartilsbereich, Median und Maximum zusammen; der Barcode darunter zeigt q50/RMS rolleweise. Skalen sind bewusst modelllokal.

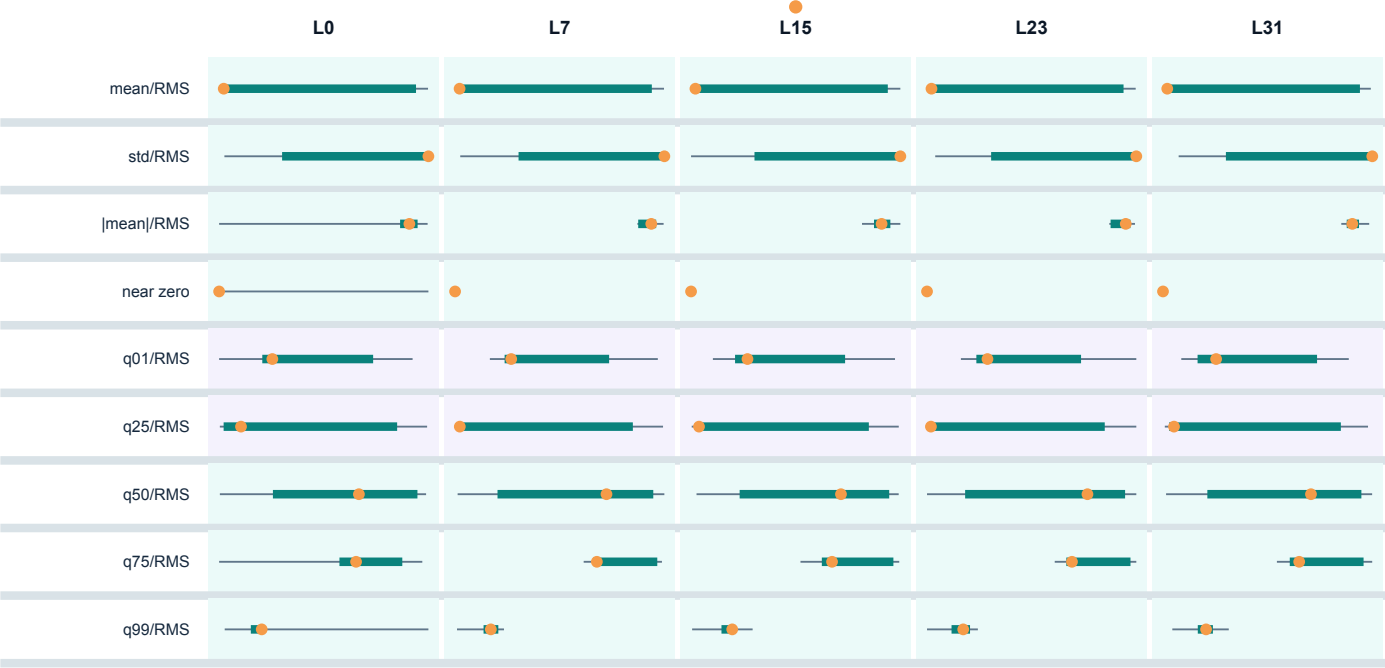
32
LAYER

138
TENSORZEILEN

702.1 MiB
MESSPAYLOAD

23
TENSORROLLEN

9-MERKMALS-FINGERABDRUCK · MIN - IQR - MEDIAN - MAX



DIREKTER TENSORROLLEN-BARCODE · Q50/RMS

attn.k / biases				
attn.k / scales				
attn.k / weight				
attn.o / biases				
attn.o / scales				
attn.o / weight				
attn.q / biases				
attn.q / scales				
attn.q / weight				
attn.v / biases				
attn.v / scales				
attn.v / weight				
mip.down / biases				
mip.down / scales				
mip.down / weight				
mip.gate / biases				
mip.gate / scales				
mip.gate / weight				
mip.up / biases				
mip.up / scales				
mip.up / weight				
norm / weight				
norm / weight				

DeepSeek R1 Distill Qwen 14B 6-bit

Alle neun registrierten Merkmale werden innerhalb dieses Checkpoints sichtbar gemacht. Jeder Micro-Glyph fasst die direkt versiegelten Tensorrollen eines Pilotlayers als Minimum, Interquartilsbereich, Median und Maximum zusammen; der Barcode darunter zeigt q50/RMS rolleweise. Skalen sind bewusst modelllokal.

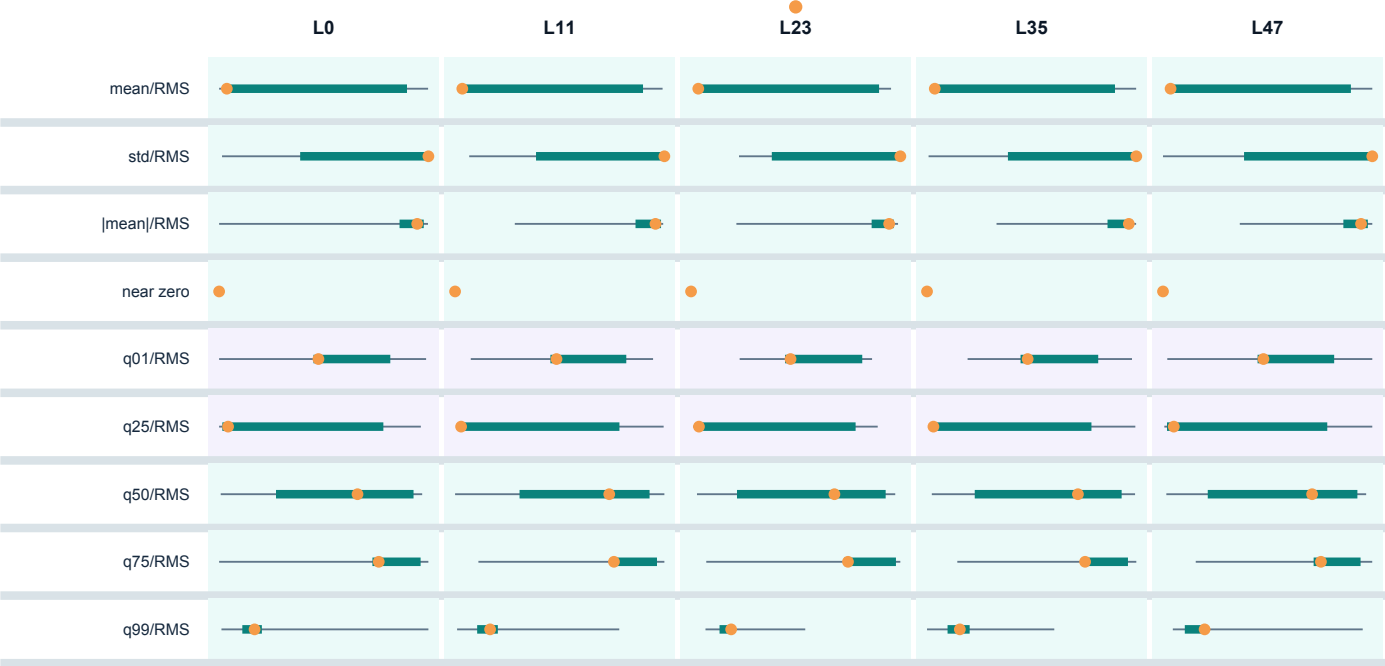
48
LAYER

156
TENSORZEILEN

1.25 GiB
MESSPAYLOAD

26
TENSORROLLEN

9-MERKMALS-FINGERABDRUCK · MIN - IQR - MEDIAN - MAX



DIREKTER TENSORROLLEN-BARCODE · Q50/RMS

attn.k / biases				
attn.k / scales				
attn.k / unquantized				
attn.k / weight				
attn.o / biases				
attn.o / scales				
attn.o / weight				
attn.q / biases				
attn.q / scales				
attn.q / unquantized				
attn.q / weight				
attn.v / biases				
attn.v / scales				
attn.v / unquantized				
attn.v / weight				
mip.down / biases				
mip.down / scales				
mip.down / weight				
mip.gate / biases				
mip.gate / scales				
mip.gate / weight				
mip.up / biases				
mip.up / scales				
mip.up / weight				
norm / weight				

ARCHITECTURE ONLY

Ein Windkanal, acht Architekturverträge

Kimi bleibt der Flagship-Tiefenfall; sieben weitere versiegelte Checkpoints belegen die technische Portabilität. Gemeinsam dargestellt werden nur Architektur, Layerzahl, Tensorvertrag und Messabdeckung. Merkmalswerte, Qualität und Fähigkeit werden nicht modellübergreifend gerankt.

Modell	Architektur	L	Vertrag	Messpositionen	Rolle
Kimi K3	MoE / KDA + Gated MLA	93	896 / 6	27 + repeats	Flagship
Llama 3.2 1B Instr. 4-bit	Dense decoder	16	dense	0, 3, 7, 11, 15	verified
DeepSeek R1 Distill Qwen 1.5B 4-bit	Dense decoder	28	dense	0, 6, 13, 20, 27	verified
ModernBERT Embed Base 8-bit	Encoder / embedding	22	encoder	1, 6, 11, 16, 21	verified
Apertus 8B 4-bit	Dense decoder	32	dense	0, 7, 15, 23, 31	verified
Apertus 8B Instr. 8-bit	Dense decoder	32	dense	0, 7, 15, 23, 31	verified
Llama 3.1 8B Instr. 4-bit	Dense decoder	32	dense	0, 7, 15, 23, 31	verified
DeepSeek R1 Distill Qwen 14B 6-bit	Dense decoder	48	dense	0, 11, 23, 35, 47	verified

GEMEINSAM OFFEN

Layer · Dense/MoE · Attention-Organisation · Tensorrollen · Quantisierung · Shards · Messabdeckung

GESCHLOSSEN: FEATUREABSTÄNDE · ÄHNLICHKEIT · QUALITÄT · GEWINNER/VERLIERER

Was der Befund trägt - und was nicht

GETRAGEN

- Kimi besitzt entlang gemessener KDA-Intervalle ein heterogenes registriertes Entscheidungsprofil. Die späten auffälligen Grenzen und alle neun A/B-Layer
- replizieren technisch exakt. Graphdarstellungen können feinere Unterschiede
- nachweislich maskieren. Der Lebenszyklus trägt sieben weitere checkpointspezifische
- Tensorverträge.

NICHT GETRAGEN

- Keine Aktivierungs-, Router- oder Tokenaussage.
- Keine Expertenbiografien über Layergrenzen.
- Keine Ursache für Modellmacht oder Qualität.
- Keine wissenschaftliche Modellrangliste.
- Keine exakte Lokalisierung innerhalb ausgeschlossener MLA-Lücken.

GESPEICHERTE STRUKTUR

messbar

REPRODUZIERBARKEIT

belegt

FUNKTIONALE BEDEUTUNG

Live-Studie offen

SCHLUSSFOLGERUNG

Paper I zeigte, dass gespeicherte Gewichte auf begrenzter Hardware reproduzierbar vermessen werden können, und endete an der Zwei-Layer-Grenze. Paper II verwandelt diese Grenze in ein Tiefenobjekt: kein glattes Kontinuum, sondern eine evidenzgebundene Folge von stabil reproduzierbaren und darstellungsabhängigen Übergängen. Die sieben Katalogmodelle zeigen, dass der Windkanal kein Kimi-Sonderwerkzeug ist. Seine wissenschaftliche Stärke liegt gerade in der Grenze: Er beschreibt gespeicherte Struktur präzise, ohne daraus Aktivierung, Semantik oder Qualität zu erfinden.

Literatur und versiegelte Evidenz

[P1] Model Weight Wind Tunnel. Kimi Weight Windkanal: Wissenschaftlicher Ergebnisbericht V1 bis V2.4. Sealed publication bundle, 2026.

[1] Hugging Face. Safetensors format specification and metadata parsing documentation.

[2] M. Raasveldt and H. Mühleisen. DuckDB: an embeddable analytical database. SIGMOD 2019.

[3] Khronos OpenCL Working Group. The OpenCL Specification, 2026 edition.

[4] A. Vaswani et al. Attention Is All You Need. NeurIPS 2017.

[5] N. Shazeer et al. Outrageously Large Neural Networks: The Sparsely-Gated Mixture-of-Experts Layer. ICLR 2017.

[6] J. Devlin et al. BERT: Pre-training of Deep Bidirectional Transformers. NAACL 2019.

[7] B. Warner et al. Smarter, Better, Faster, Longer: A Modern Bidirectional Encoder. 2024.

[8] Meta. Llama 3.2 Model Card and Llama 3.2 Community License, pinned repository revision.

[9] DeepSeek-AI. DeepSeek-R1: Incentivizing Reasoning Capability in LLMs. 2025.

[10] A. Hernández-Cano et al. Apertus: Democratizing Open and Compliant LLMs. 2025.

[11] A. Hannun et al. MLX: Efficient and flexible machine learning on Apple silicon. 2023.

[12] A. Mikhaylov. TurboFieldfare: SSD-streamed Gemma 4 MoE inference on Apple Silicon. 2026.

VERSIEGELTE INTERNE EVIDENZ

- Paper-I-Publikationsbundle V1-V2.4
- Kimi V2.5-V2.10 result databases and result seals
- Sieben zentral verifizierte Katalog-Result-Datenbanken
- Architecture Atlases, Freeze-Manifeste und Post-Seal-Prüfungen
- RC4-Visualisierungsdatensatz mit SHA-256-Inventar
-