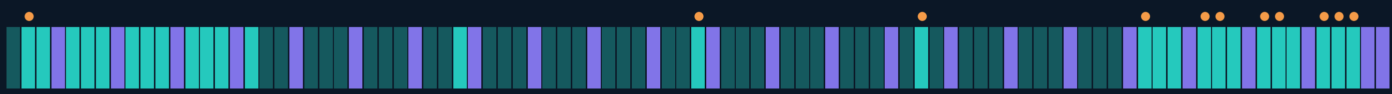
From a Layer Pair to a Depth Profile

Paper II: Dense, reproducible weight measurement of a 93-layer MoE flagship and transfer to seven further checkpoints



ABSTRACT

Paper I deliberately ended at a two-layer boundary. Paper II extends the same frozen measurement contract to 27 Kimi KDA positions, dense early and late depth regions, and nine exactly repeated layers. The sealed evidence is no longer collapsed into status cards: 1,320 primary cells, 3,960 all-pair cells, 2,112 component values, 960 anchor-nearness rows, and 48,384 repeat tensors are rendered as a depth film, threshold landscape, masking flow, and replication wall. Seven further checkpoints contribute 984 sealed tensor rows across nine registered features. Across models, only architecture comparison remains permitted.

27
KIMI POSITIONS

1.320
PRIMARY CELLS

48.384
REPEAT TENSORS

7
CATALOG MODELS

PAPER I -> PAPER II

1 From Paper I to Paper II

Paper I [P1] froze the Architecture Atlas, the six-family contract, and the representation chain through V2.4. For layers 1 and 46, the continuous source spaces were predominantly different while coarse graph signatures could compress those differences. This supported no general claim about Kimi's depth. Paper II inherits those findings and thresholds unchanged and executes the targeted extension authorized by that limitation. The method is not introduced again; it is stressed across additional depth positions.

PAPER I	OFFENE GRENZE	PAPER II
L1 <-> L46	**2 != TIEFE**	**27 KDA**
two measurement windows	no depth invariance	dense regions + repeats

Research questions

- RQ1 Depth - what registered decision profile emerges beyond the layer pair 1/46?
- RQ2 Repetition - do technical repeats reproduce bound tensors and decision surfaces?
- RQ3 Portability - can one lifecycle preserve other checkpoint-specific architecture contracts?
- RQ4 Auditability - do invalid identity or commit bindings prevent scientific promotion?

INHERITED, NOT REFITTED

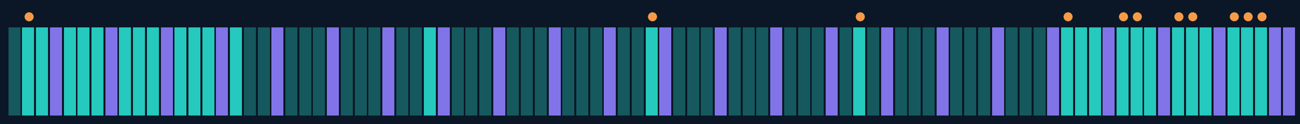
```
Atlas + homology + six families + normalizers + thresholds -> new bound depth observations
```

No interpolation. No threshold refit. No cross-layer expert identity. No activation claim.

FLAGSHIP CASE

2 Kimi K3 beyond two layers

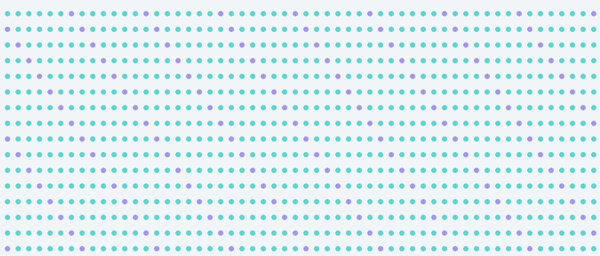
Paper I viewed Kimi through two homologous measurement windows: layers 1 and 46. The sealed atlas, however, contains 93 layers: 69 KDA and 24 Gated MLA. Paper II expands only within the homologous KDA MoE contract: 896 expert IDs, six tensor families, and 5,376 expert tensors per fully measured layer position. The layer pair therefore becomes a registered depth profile - still without model startup or activation.



93 layers; cyan = KDA, violet = Gated MLA, bright core = measured, orange = technically repeated.

MODEL | LAYER | 6 FAMILIES | 896 EXPERTS | 5,376 TENSORS

896 EXPERT IDS



6-FAMILY CONTRACT

W1	W1 packed
W1	W1 scale
W2	W2 packed
W2	W2 scale
W3	W3 packed
W3	W3 scale

PAPER I -> PAPER II

3 The Kimi depth campaign

The campaign answers Paper I's closing question in stages: depth anchors first, an unchanged decision surface, then exact repetitions, and finally dense cohorts at the late and early model regions. V2.9 additionally binds the striking late boundaries again. V2.11 is not part of this paper.

PAPER I · V1-V2.4

L1 <-> L46

Two of 92 MoE layers fully profiled. Continuous layer geometry predominantly differed; general depth invariance was not supported.

OPEN LIMIT BECOMES THE PAPER II PROGRAM

V2.5
7 depth anchors
360 primary cells

V2.6
2 depth intervals
60 sealed decisions

V2.7
L61 / 76 / 90 repeated
16,128 exact tensors

V2.8
late dense cohort
53,760 new tensors

V2.9
6 boundary repeats
sealed predecessor

V2.10
early dense cohort
660 primary cells

SEALED PAPER BOUNDARY

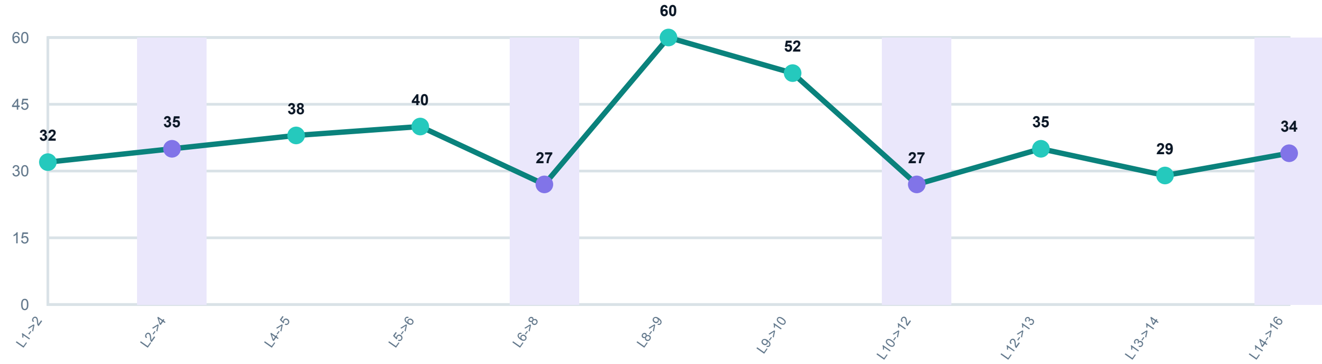
27	11	V2.10	V2.11
measured KDA positions	technically repeated	included through	excluded: no local result seal

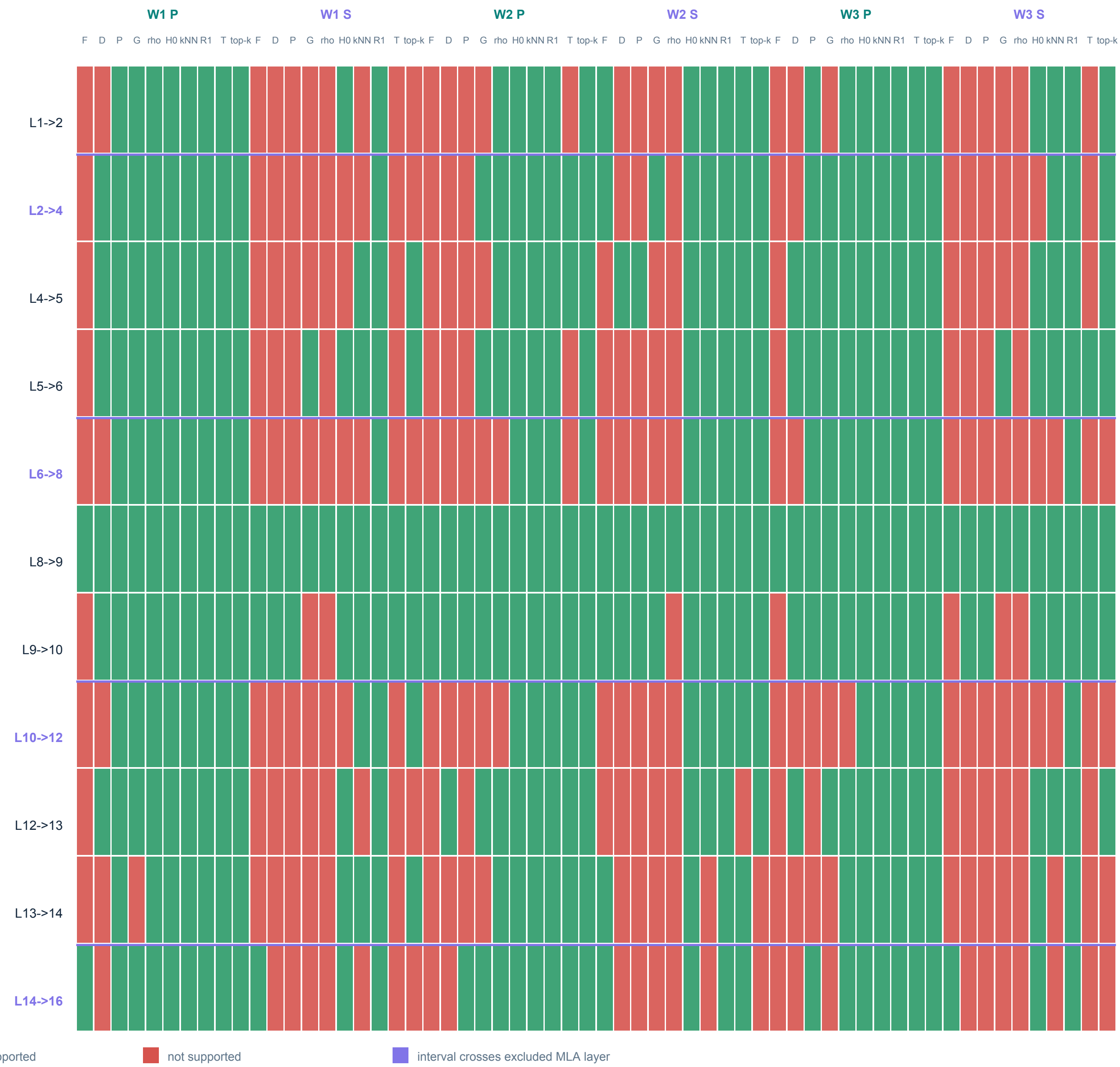
V2.10 · 660 PRIMARY CELLS

Early depth film: layers 1 to 16

Each row is an observed KDA interval; each of the 60 columns is one family-by-representation cell. Violet marks intervals spanning an excluded Gated MLA layer; no exact transition is localized inside that gap.

SUPPORTED CELLS OF 60



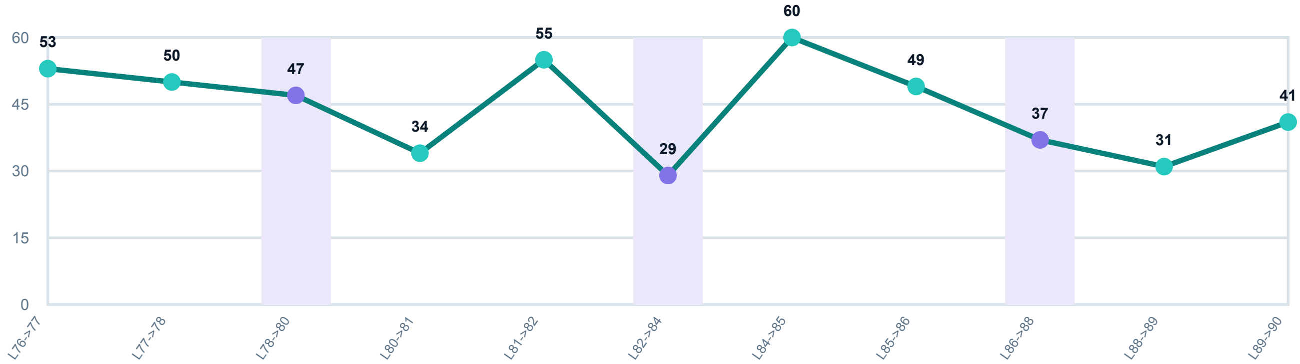


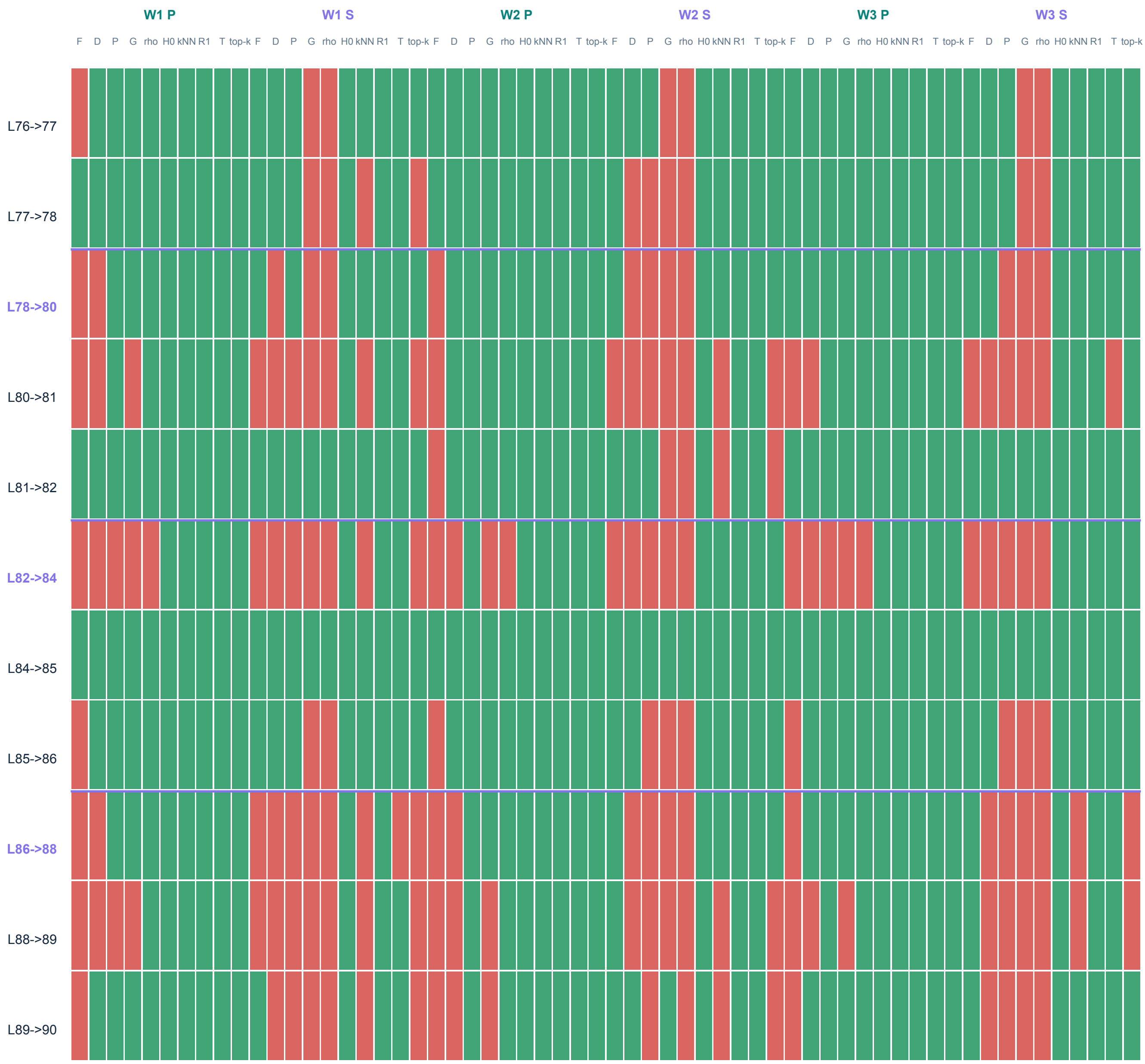
V2.8 · 660 PRIMARY CELLS

Late depth film: layers 76 to 90

The dense late cohort shows eleven successive KDA intervals under the same six-family contract. The pattern is not smoothed: all 660 sealed decisions remain visible as a raster.

SUPPORTED CELLS OF 60



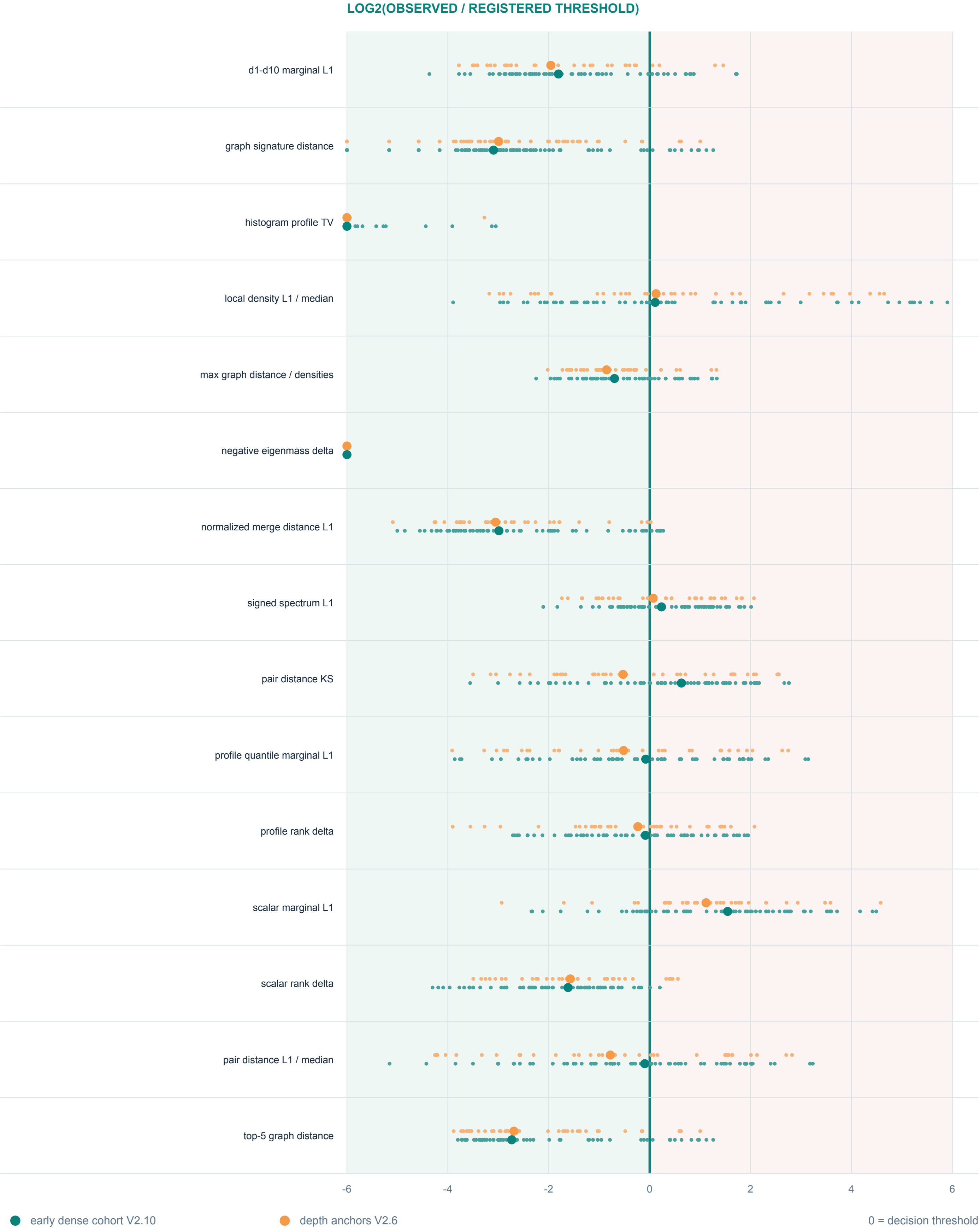


supported
not supported
interval crosses excluded MLA layer

1,632 COMPONENT OBSERVATIONS

Threshold landscape: distance to the registered boundary

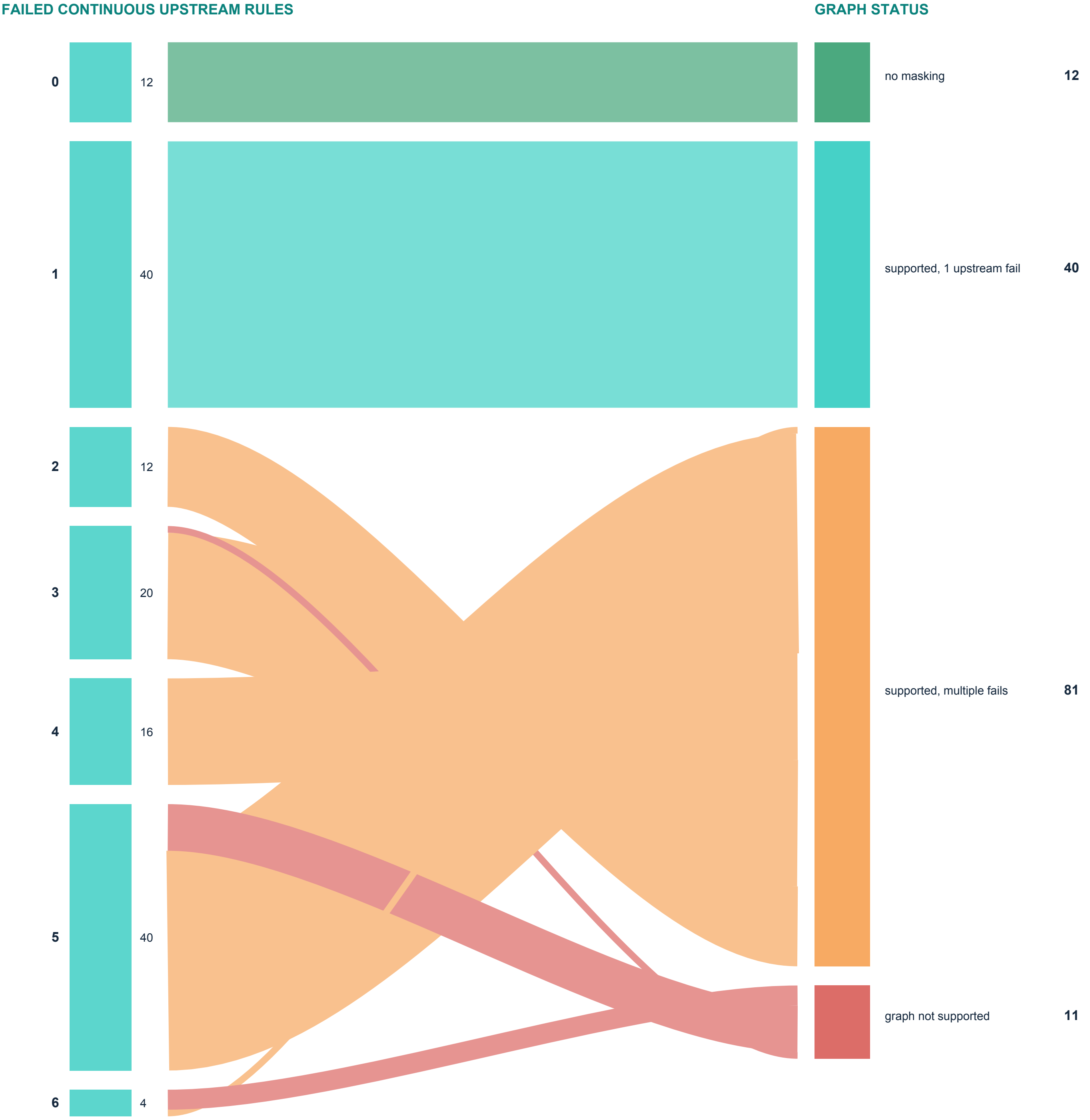
Each point is a sealed ratio between an observed component value and its unchanged registered threshold. The logarithmic axis makes very different metrics jointly legible: left of zero the component lies within its threshold, right of zero it lies outside. This does not alter any cell decision; all rules of a cell must pass together.



V2.6 · 144 GRAPH EVENTS

When graphs mask upstream differences

Graph representations can remain supported even when one or more continuous upstream rules fail. The flow renders all 144 registered attributions: failed upstream rules on the left and the observed graph status on the right. Ribbon width equals the number of observed cells.



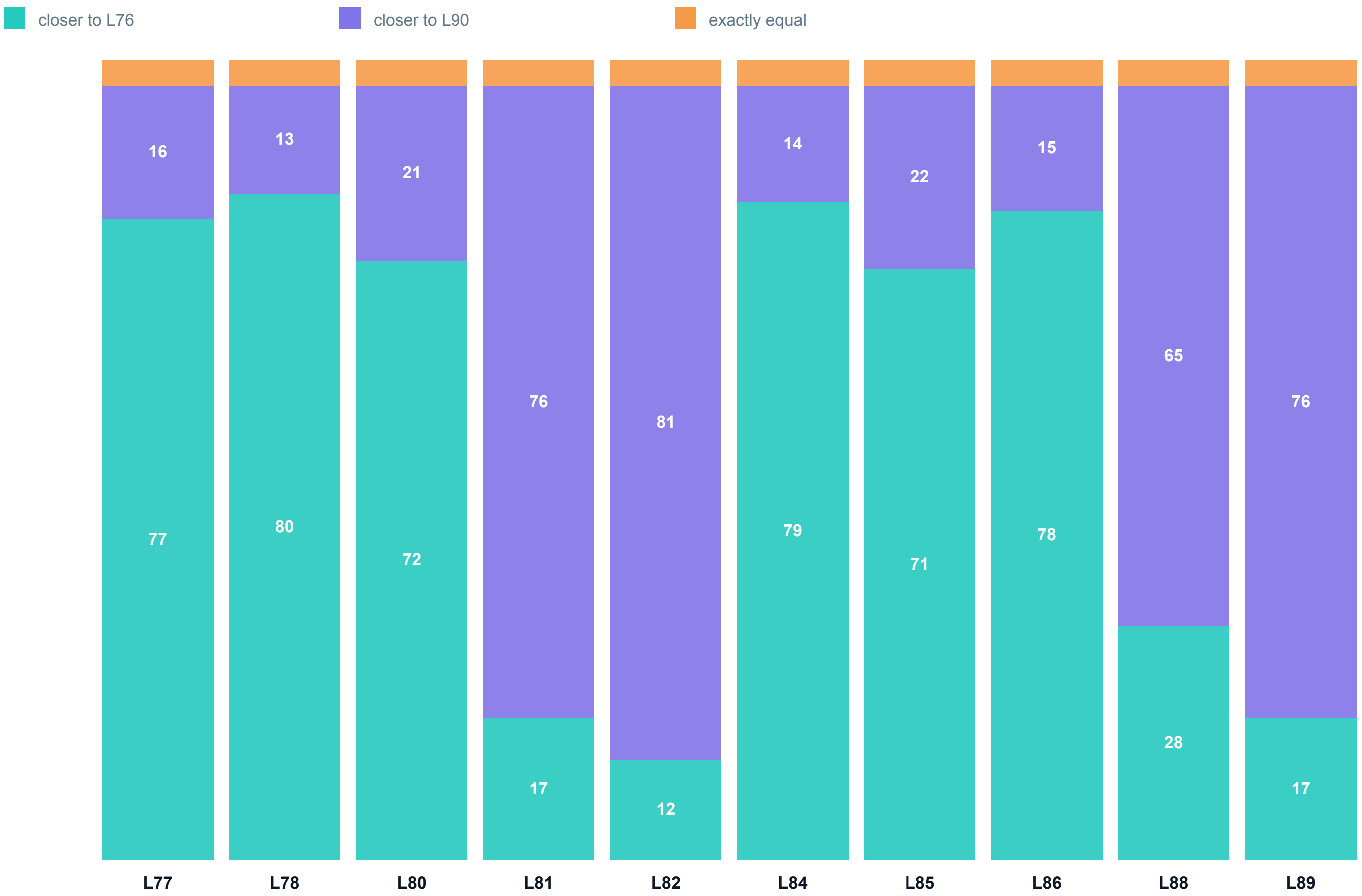
WHAT IT MEANS

A coarse graph can make two layers appear similar while finer geometries remain different. This is not model weakness; it is an observed loss of information caused by the representation.

V2.8 · 960 DESCRIPTIVE ROWS

Between two anchors: closer to 76 or 90?

For ten late intermediate layers, 96 family, representation, and metric combinations were described against anchors 76 and 90. Stacked bars show only which anchor has the smaller registered distance. Rows are not independent samples and do not establish biographies of individual experts.



531
closer to L76

399
closer to L90

30
exactly equal

V2.7 + V2.9

Replication wall: not similar, identical

Nine preregistered KDA layers were fully repeated. Each ring represents 5,376 of 5,376 byte-exactly reproduced expert tensors. In addition, 300 of 300 bound decision cells replicated. The observed depth structure is therefore not a one-off read artifact of the measurement path.

All nine A/B layers byte-exact

V2.7: 120/120 · V2.9: 180/180 decision cells identical

TECHNICAL CLAIM

Byte equality demonstrates determinism of the bound scanner, device, and data path. It proves neither mathematical correctness nor functional meaning of the model.

SINGLE-MODEL DOSSIER

Llama 3.2 1B Instruct 4-bit

All nine registered features are rendered within this checkpoint. Each micro-glyph summarizes the directly sealed tensor roles of one pilot layer as minimum, interquartile range, median, and maximum; the barcode below shows q50/RMS by role. Scales are deliberately model-local.

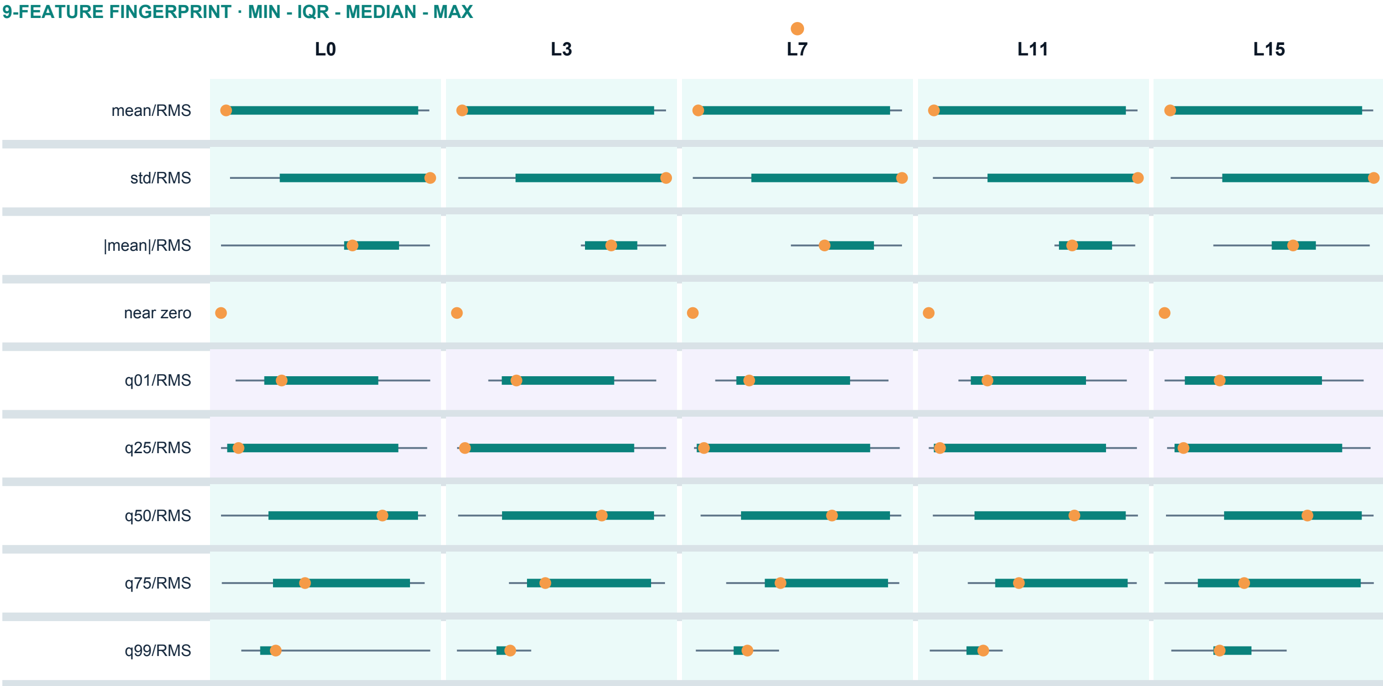
16
LAYERS

138
TENSOR ROWS

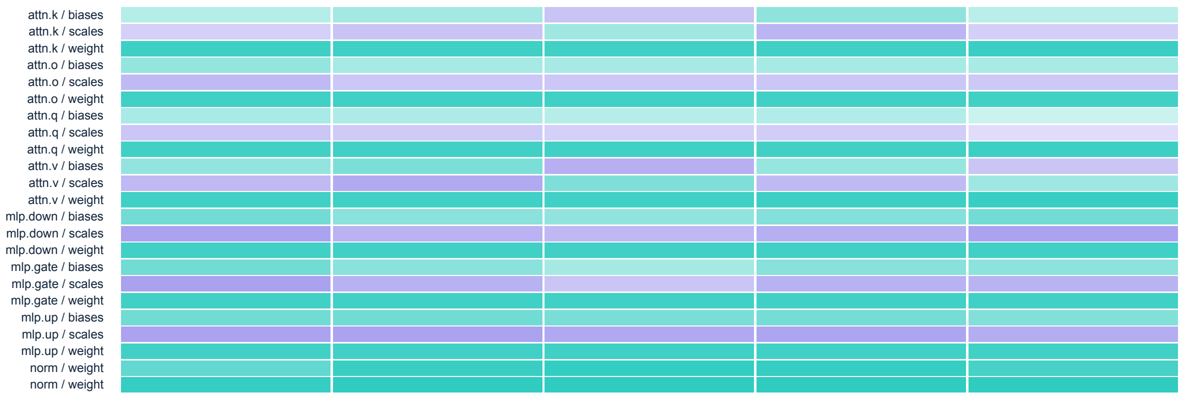
195.8 MiB
MEASUREMENT PAYLOAD

23
TENSOR ROLES

9-FEATURE FINGERPRINT · MIN - IQR - MEDIAN - MAX



DIRECT TENSOR-ROLE BARCODE · Q50/RMS



Scales remain local to each model and feature; no model ranking.

A/B L7: 23/23 byte-exact

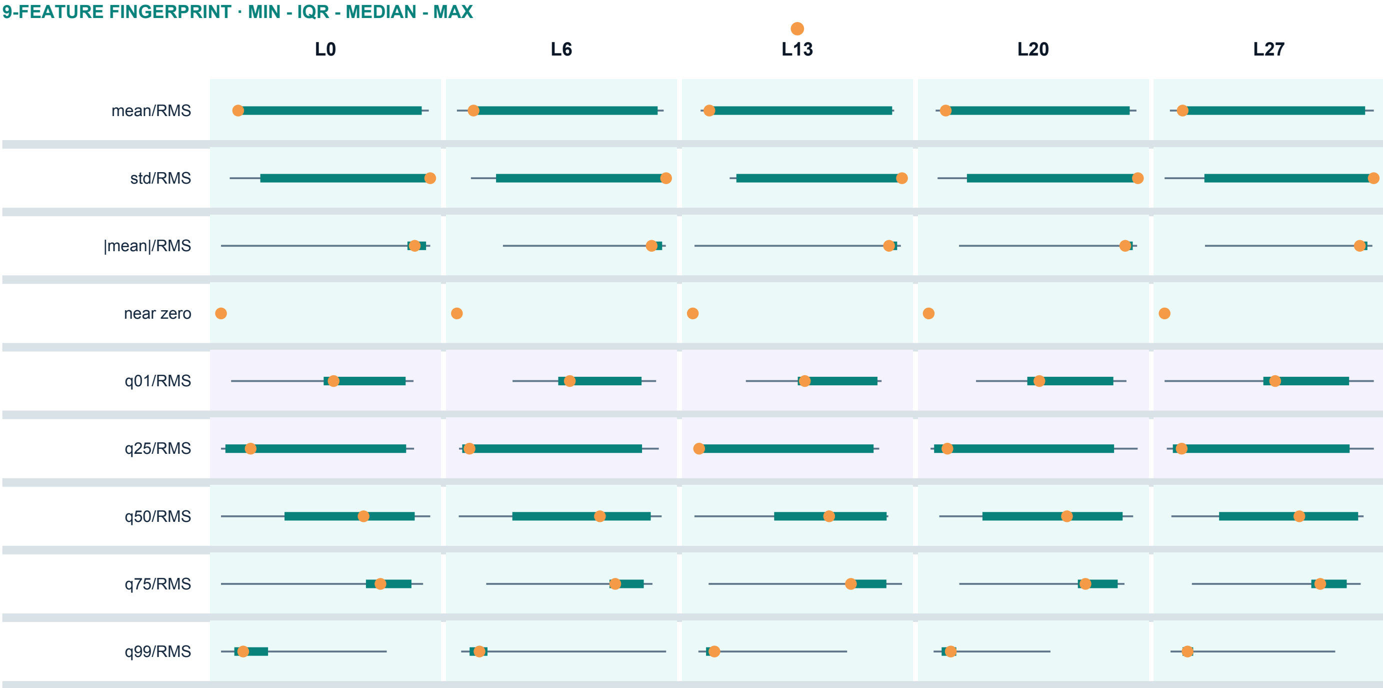
SINGLE-MODEL DOSSIER

DeepSeek R1 Distill Qwen 1.5B 4-bit

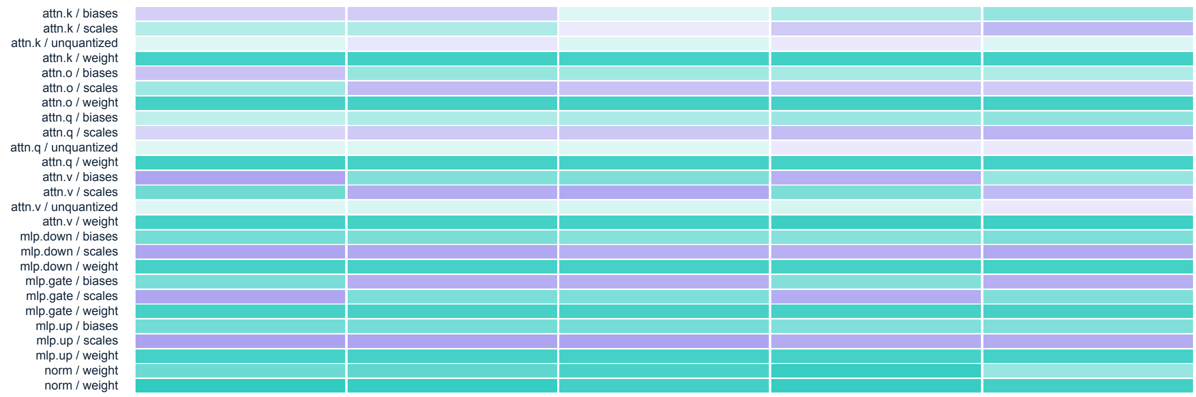
All nine registered features are rendered within this checkpoint. Each micro-glyph summarizes the directly sealed tensor roles of one pilot layer as minimum, interquartile range, median, and maximum; the barcode below shows q50/RMS by role. Scales are deliberately model-local.

28	156	150.7 MiB	26
LAYERS	TENSOR ROWS	MEASUREMENT PAYLOAD	TENSOR ROLES

9-FEATURE FINGERPRINT · MIN - IQR - MEDIAN - MAX



DIRECT TENSOR-ROLE BARCODE · Q50/RMS



Scales remain local to each model and feature; no model ranking.

A/B L13: 26/26 byte-exact

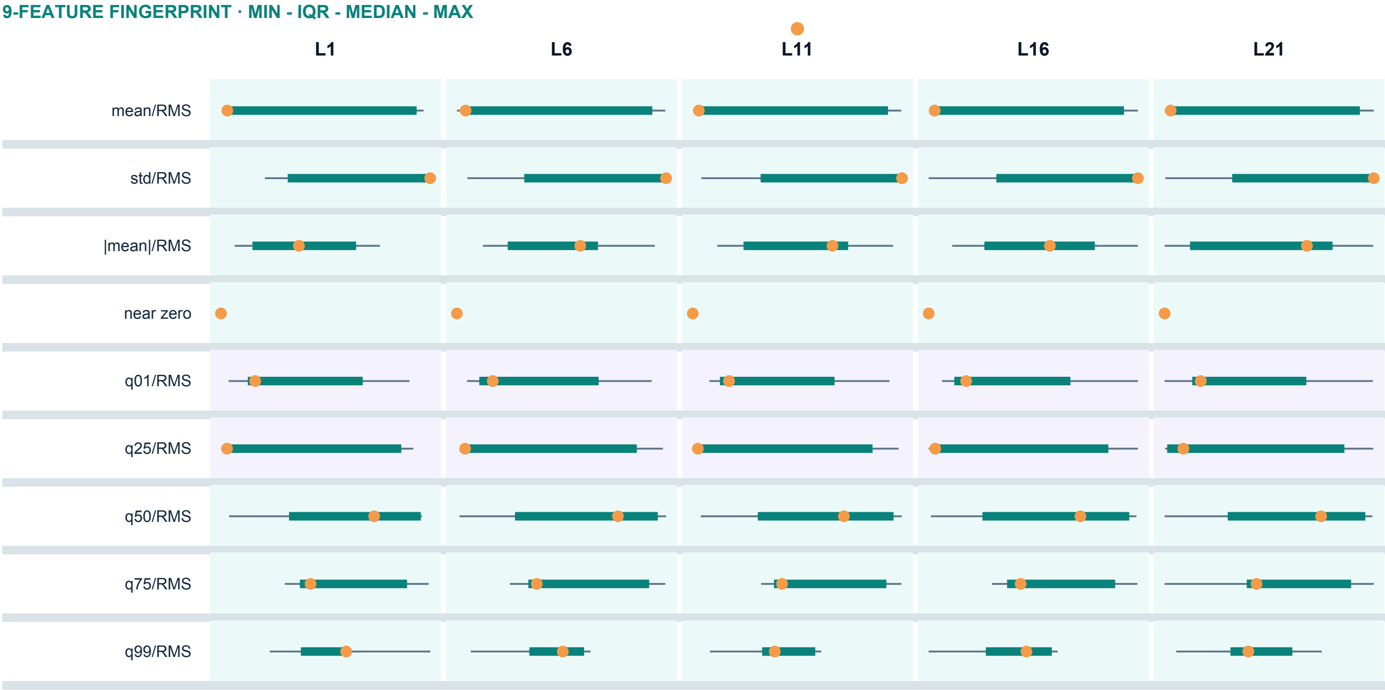
SINGLE-MODEL DOSSIER

ModernBERT Embed Base 8-bit

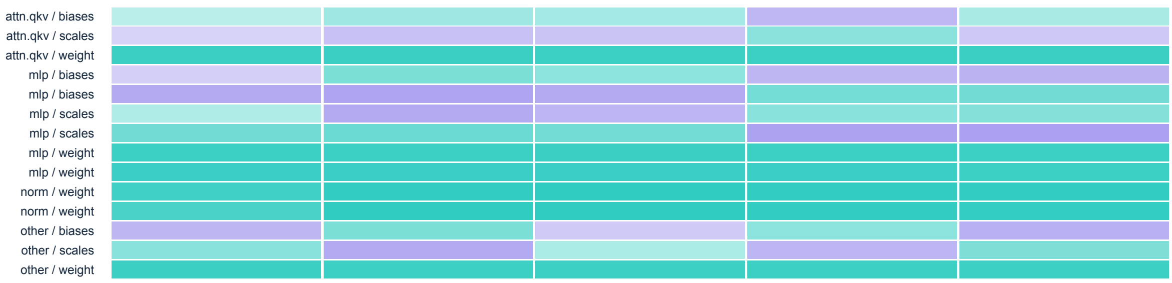
All nine registered features are rendered within this checkpoint. Each micro-glyph summarizes the directly sealed tensor roles of one pilot layer as minimum, interquartile range, median, and maximum; the barcode below shows q50/RMS by role. Scales are deliberately model-local.

22	**84**	**30.5 MiB**	**14**
LAYERS	TENSOR ROWS	MEASUREMENT PAYLOAD	TENSOR ROLES

9-FEATURE FINGERPRINT · MIN - IQR - MEDIAN - MAX



DIRECT TENSOR-ROLE BARCODE · Q50/RMS



Scales remain local to each model and feature; no model ranking.

A/B L11: 14/14 byte-exact

SINGLE-MODEL DOSSIER

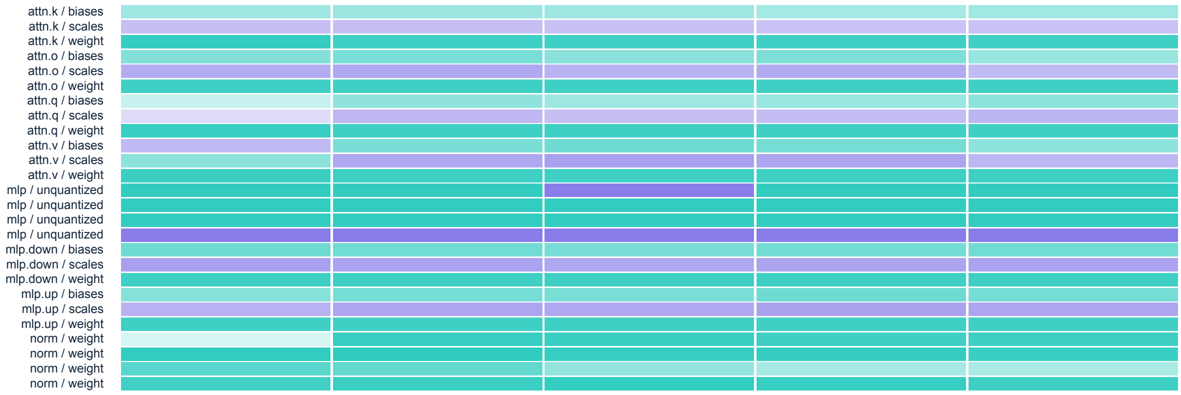
Apertus 8B 4-bit

All nine registered features are rendered within this checkpoint. Each micro-glyph summarizes the directly sealed tensor roles of one pilot layer as minimum, interquartile range, median, and maximum; the barcode below shows q50/RMS by role. Scales are deliberately model-local.

32 LAYERS	156 TENSOR ROWS	702.1 MiB MEASUREMENT PAYLOAD	26 TENSOR ROLES

9-FEATURE FINGERPRINT · MIN - IQR - MEDIAN - MAX

DIRECT TENSOR-ROLE BARCODE · Q50/RMS



Scales remain local to each model and feature; no model ranking.

A/B L15: 26/26 byte-exact

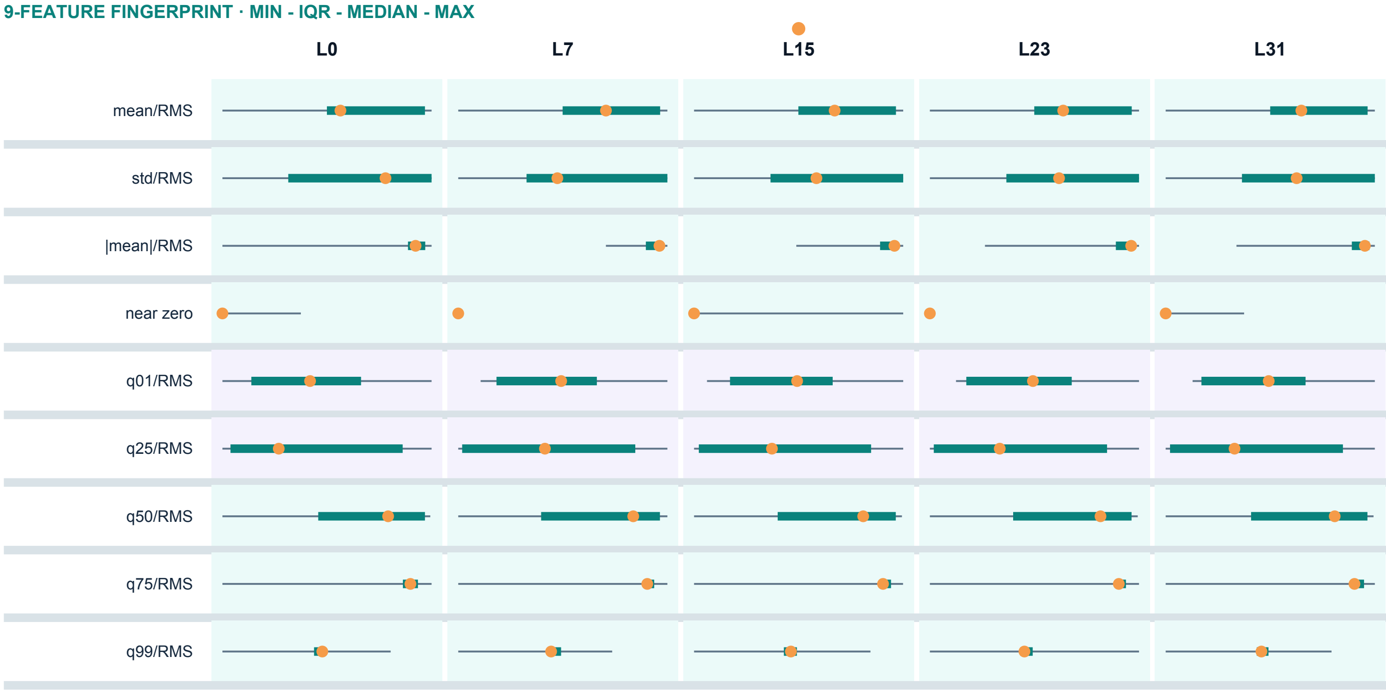
SINGLE-MODEL DOSSIER

Apertus 8B Instruct 8-bit

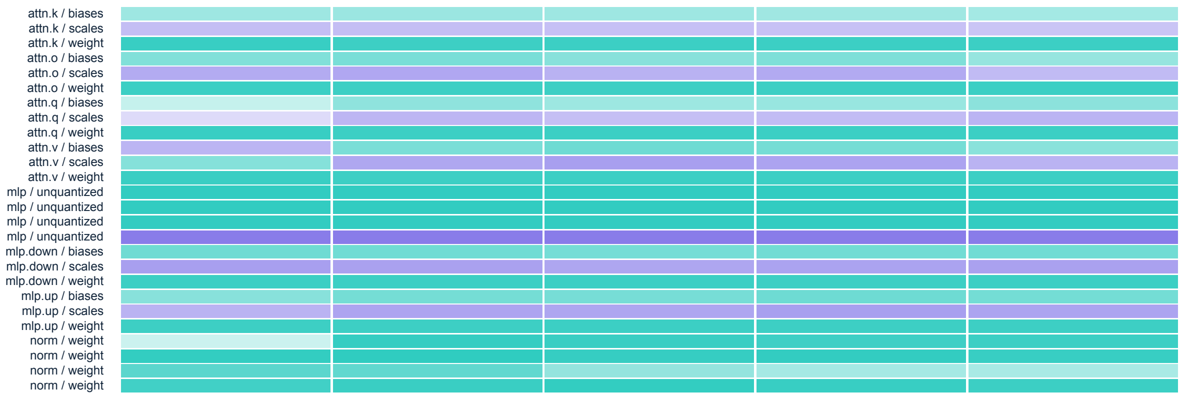
All nine registered features are rendered within this checkpoint. Each micro-glyph summarizes the directly sealed tensor roles of one pilot layer as minimum, interquartile range, median, and maximum; the barcode below shows q50/RMS by role. Scales are deliberately model-local.

32	156	1.30 GiB	26
LAYERS	TENSOR ROWS	MEASUREMENT PAYLOAD	TENSOR ROLES

9-FEATURE FINGERPRINT · MIN - IQR - MEDIAN - MAX



DIRECT TENSOR-ROLE BARCODE · Q50/RMS



Scales remain local to each model and feature; no model ranking.

A/B L15: 26/26 byte-exact

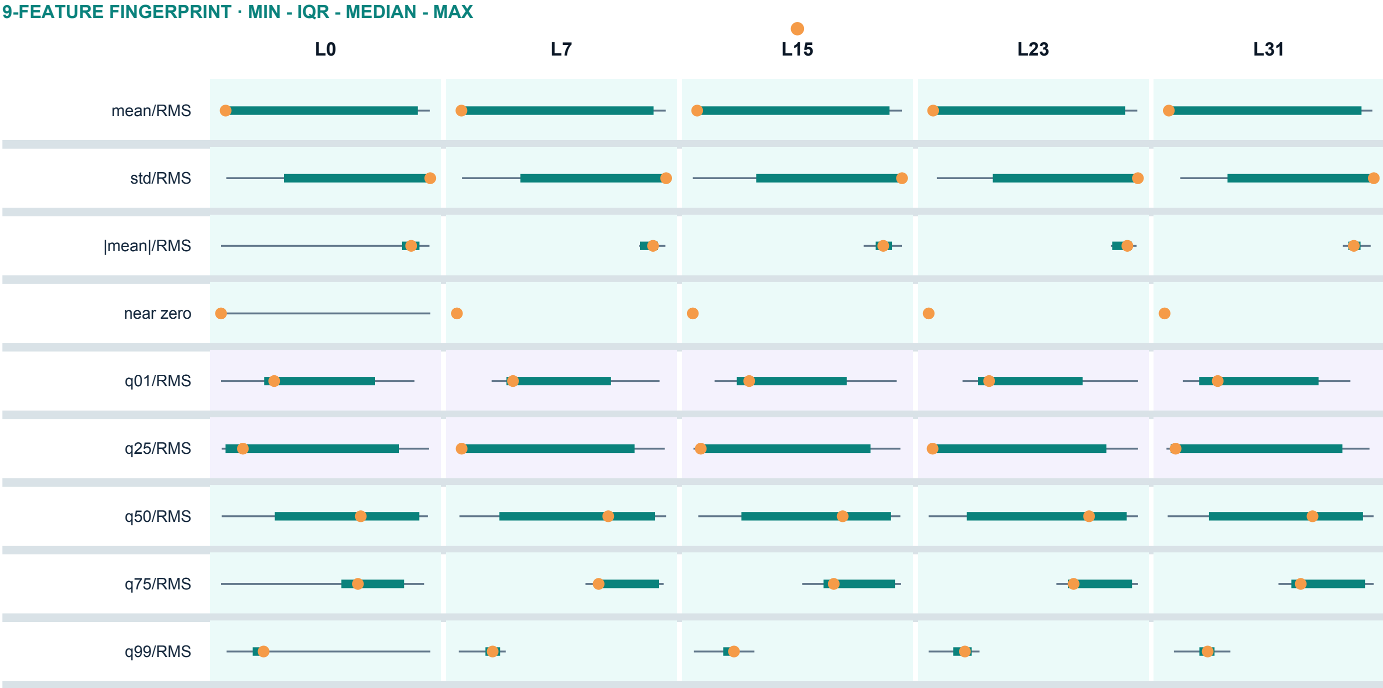
SINGLE-MODEL DOSSIER

Llama 3.1 8B Instruct 4-bit

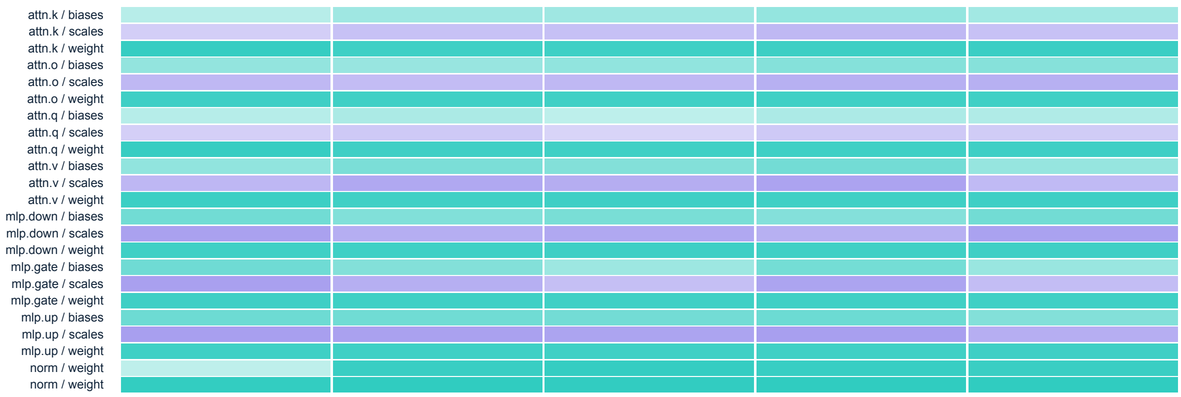
All nine registered features are rendered within this checkpoint. Each micro-glyph summarizes the directly sealed tensor roles of one pilot layer as minimum, interquartile range, median, and maximum; the barcode below shows q50/RMS by role. Scales are deliberately model-local.

32	138	702.1 MiB	23
LAYERS	TENSOR ROWS	MEASUREMENT PAYLOAD	TENSOR ROLES

9-FEATURE FINGERPRINT · MIN - IQR - MEDIAN - MAX



DIRECT TENSOR-ROLE BARCODE · Q50/RMS



Scales remain local to each model and feature; no model ranking.

A/B L15: 23/23 byte-exact

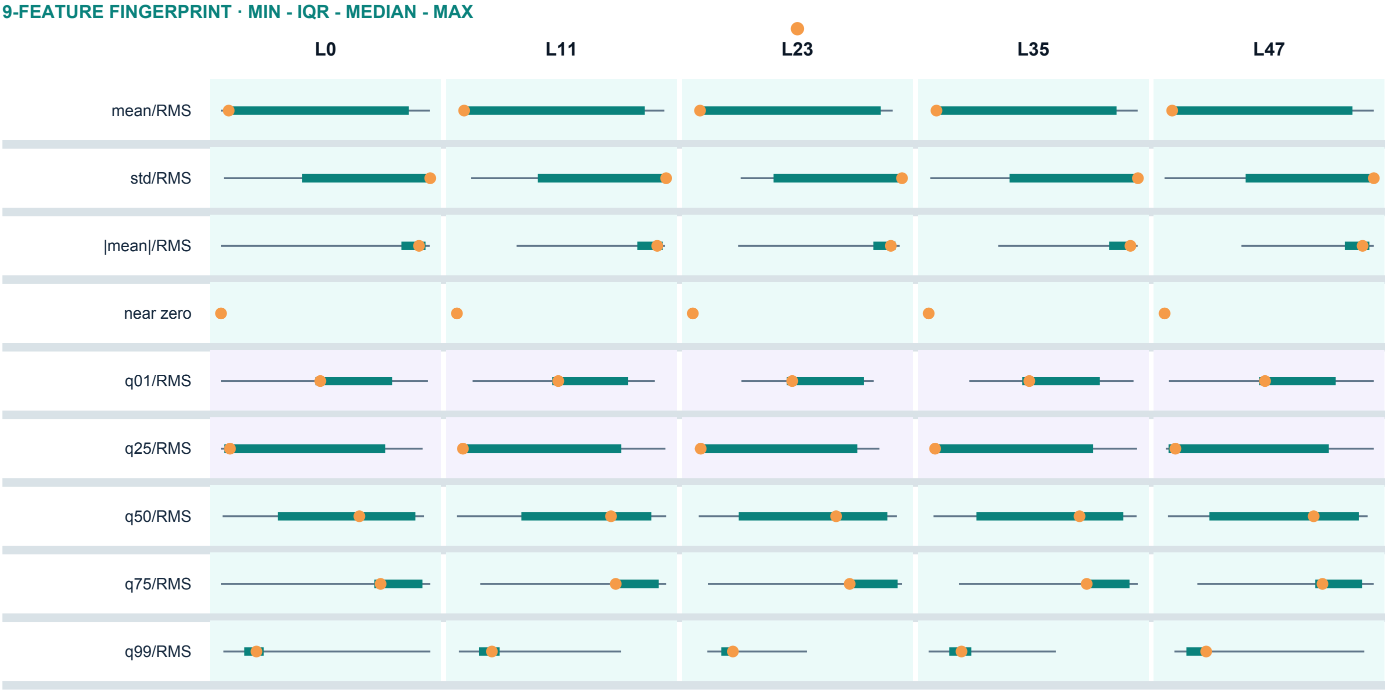
SINGLE-MODEL DOSSIER

DeepSeek R1 Distill Qwen 14B 6-bit

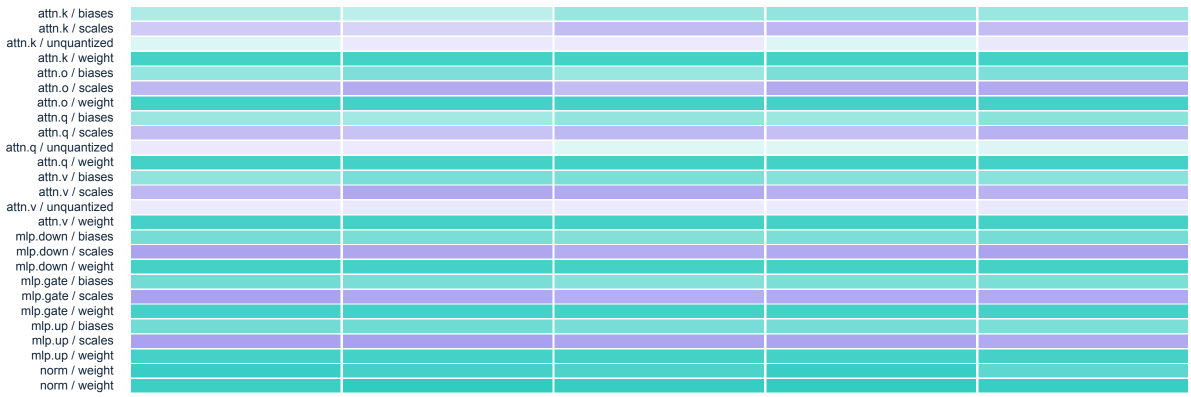
All nine registered features are rendered within this checkpoint. Each micro-glyph summarizes the directly sealed tensor roles of one pilot layer as minimum, interquartile range, median, and maximum; the barcode below shows q50/RMS by role. Scales are deliberately model-local.

48	156	1.25 GiB	26
LAYERS	TENSOR ROWS	MEASUREMENT PAYLOAD	TENSOR ROLES

9-FEATURE FINGERPRINT · MIN - IQR - MEDIAN - MAX



DIRECT TENSOR-ROLE BARCODE · Q50/RMS



Scales remain local to each model and feature; no model ranking.

A/B L23: 26/26 byte-exact

ARCHITECTURE ONLY

One wind tunnel, eight architecture contracts

Kimi remains the flagship depth case; seven further sealed checkpoints establish technical portability. Only architecture, layer count, tensor contract, and measurement coverage are shown jointly. Feature values, quality, and capability are not ranked across models.

Model	Architecture	L	Contract	Measured positions	Role
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

JOINTLY OPEN

layers · Dense/MoE · attention organization · tensor roles · quantization · shards · measurement coverage

CLOSED: FEATURE DISTANCES · SIMILARITY · QUALITY · WINNER/LOSER

DISCUSSION

What the evidence supports - and what it does not

SUPPORTED

- Kimi exhibits a heterogeneous registered decision profile across measured KDA intervals.
- The striking late boundaries and all nine A/B layers replicate exactly.
- Graph representations can demonstrably mask finer differences.
- The lifecycle supports seven further checkpoint-specific tensor contracts.

NOT SUPPORTED

- No activation, router, or token claim.
- No expert biographies across layers.
- No cause of model power or quality.
- No scientific model ranking.
- No exact localization inside excluded MLA gaps.

STORED STRUCTURE

measurable

REPEATABILITY

demonstrated

FUNCTIONAL MEANING

live study remains open

CONCLUSION

Paper I showed that stored weights can be reproducibly measured on constrained hardware and ended at the two-layer boundary. Paper II turns that boundary into a depth object: not a smooth continuum, but an evidence-bound sequence of stably reproducible, representation-dependent transitions. The seven catalog models show that the wind tunnel is not a Kimi-specific tool. Its scientific strength lies in the boundary: it precisely describes stored structure without inventing activation, semantics, or quality.

SOURCES / HASH-BOUND EVIDENCE

References and sealed evidence

[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.

SEALED INTERNAL EVIDENCE

- Paper I V1-V2.4 publication bundle
- Kimi V2.5-V2.10 result databases and result seals
- Seven centrally verified catalog result databases
- Architecture atlases, freeze manifests, and post-seal verifications
- RC4 visualization dataset with SHA-256 inventory

```
cross_model_scientific_comparison_performed = false

new_weight_measurement_performed = false
```