KEEL Technologies

KEEL SCAN · MODEL INTELLIGENCE REPORT

meta-llama/Llama-3.1-8B

8030M parameters · 32 layers · scanned 2026-07-24T18:31:08Z

B

GRADE B · GOOD

K# 136.0

100 = BASELINE ACROSS THE INDEX

02 Executive brief
04 Executive summary
05 Peer comparison
06 Architecture
07 Activation profile
08 Layer analysis
09 Compression density
10 Performance
11 Coherence & efficiency
12 Capability
14 Knockout analysis
16 Prescriptions
18 About

EXECUTIVE BRIEF

What this report says.

Your model scores **K# 136.0** on the Keel Index — an above-baseline result, grade B. It ranks **4 of 14** in the Standard class. The class leader (OLMo-2-1124-7B) scores 5.1 points higher.

Three findings.

19%

of compute is not doing useful work

1 of 32

layers can be removed with no measured quality loss

$792–$2,660

per month in avoidable spend at 1B tokens/month

Three actions, in order of impact.

1. Prune the 1 redundant layer (saves 3% of parameters, no measured quality loss)
2. Quantize weights fp16 → int8 (typically 2× memory reduction, ~1% quality loss)
3. Optimize the serving stack (typical 20–100% throughput gains)

Next step. A **Keel Forge** engagement executes these on your specific infrastructure and delivers a measurable before/after with a Keel Compass rescan. Contact **contact@keeltech.co** to scope.

HOW TO READ THIS REPORT

One diagnosis, five readings.

This report looks inside meta-llama/Llama-3.1-8B — not at what it says, but at how it is built. Each reading below is a lens on a different property of the model's structure.

Keel Score (K#)	A single number for structural quality, positioned against every model we have scanned.
Architecture Class	The structural shape the model falls into, and what it predicts about behavior.
Coherence	How evenly capability is distributed across the model rather than concentrated.
Thermal Grade	Efficiency — how much compute does useful work versus sits idle.
Knockout analysis	Which layers can be removed with little measurable impact.

> The measurement uses a proprietary methodology. This report shows the result, not the method. No training data was required and nothing about the model leaves the engagement.

EXECUTIVE SUMMARY

meta-llama/Llama-3.1-8B scores 136.0, grade B.

Against models of comparable size (Standard (5B–10B)), meta-llama/Llama-3.1-8B is **a good** result. Its structural shape is **Declining — capability tapers gradually across depth**, and efficiency is graded **B** — 81% of its compute is doing useful work.

0 Keel Score 136.0 · scale 0–200, 100 = baseline 200

AT A GLANCE

READING	RESULT	RATING
Keel Score (K#)	136.0	B Good
Coherence Score	71.3 / 100	Good
Architecture Class	Declining	Tapers with depth
Thermal Grade	B	Good
Perplexity (held-out)	5.89	Strong for size
Redundant layers (sampled)	1 of 6	Optimization headroom

Verdict. A solid structural result in the Standard (5B–10B) class. The knockout analysis identified layer 10 as redundant at no measured cost — the fastest path to lower inference cost. The thermal grade is good — some idle capacity remains.

PEER COMPARISON

Rank 4 of 14 in your size class.

meta-llama/Llama-3.1-8B sits in the **Standard (5B–10B)** band on the Keel Index. Here is where your model ranks against every other scanned model in the same band.

YOUR MODEL

Llama-3.1-8B ← your model	**K♯ 136.0**	**B**

ABOVE YOUR MODEL (3 IN CLASS)

OLMo-2-1124-7B	K♯ 141.1	A
Mistral-7B-v0.3	K♯ 140.1	A
Qwen2.5-7B	K♯ 137.4	B

BELOW YOUR MODEL

Qwen2.5-Coder-7B	K♯ 134.4	B
starcoder2-7b	K♯ 133.9	B
gemma-2-9b	K♯ 132.5	B
Llama-3.1-8B-Instruct	K♯ 131.2	B
gpt-j-6b	K♯ 128.5	B
pythia-6.9b	K♯ 127.6	B

The Keel Index is growing continuously. Your ranking reflects models scanned to date and will refine as the database expands.

ARCHITECTURE PROFILE

Declining.

Every model resolves to one of five structural classes. meta-llama/Llama-3.1-8B is **Declining**: its capability tapers gradually across depth, with early layers doing more work than later ones. Declining models often have architectural room in the tail — later layers may be candidates for pruning.

PARAMETERS	LAYERS	WIDTH
8030M	32	4096
CONTEXT	**CLASS**	**UTILIZATION**
131072	Declining	81%

THE FIVE CLASSES

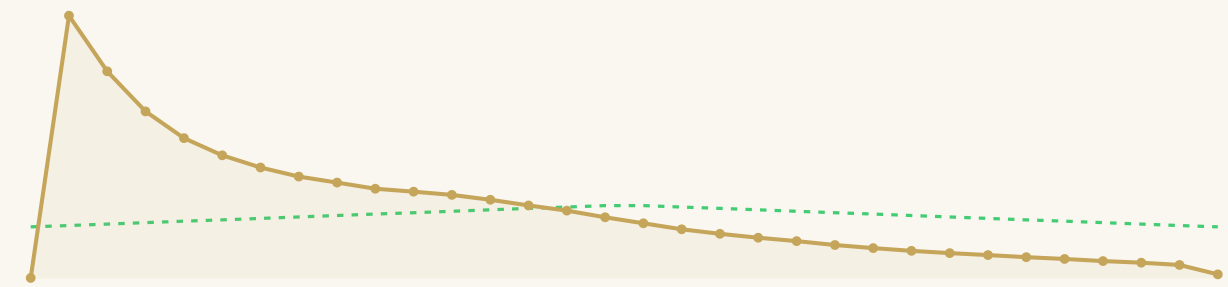
Sequential	Concentrated	Front-Loaded	**Declining**	Dual-Phase

Class is a structural signature, not a quality grade. It tells you how the model will behave — under fine-tuning, under pruning, under distribution shift — more than how good it is.

ACTIVATION PROFILE

The gap is the opportunity.

The solid line is meta-llama/Llama-3.1-8B's activation variance across its 32 layers. The dotted line below it is the profile of an efficient, evenly-built model at this size. The distance between them is exactly what an optimization pass would recover.



PEAK VARIANCE	QUIETEST LAYER	SHAPE
layer 1	layer 0	Declining

LAYER-BY-LAYER ANALYSIS

Where the work happens.

Activation variance across all 32 layers, sampled at 18 points evenly spaced through depth. Variance tapering across depth is the Declining signature — later layers do progressively less.

LAYER	ACTIVATION VARIANCE	LEVEL
L0	0.3	Low
L2	164.1	High
L4	111.2	High
L5	97.5	High
L7	80.7	Moderate
L9	71.1	Moderate
L11	66.1	Moderate
L13	57.8	Moderate
L15	48.4	Moderate
L16	43.7	Moderate
L18	35.3	Moderate
L20	29.6	Moderate
L22	24.0	Low
L24	20.1	Low
L26	16.8	Low
L27	15.4	Low
L29	12.4	Low
L31	3.2	Low

COMPRESSION DENSITY

Bits per parameter, calibrated against the Index.

Compression Density measures how much structural information the model stores per parameter, calibrated against every model on the Keel Index. This reading sits alongside K# and Grade — it measures an aspect of structure the headline number aggregates away, and it is computed directly from the model's weights, with no forward passes or training data required.

READING

103% of baseline

Above baseline

Above baseline means the model is more parameter-efficient than average for its size class — it stores more structural information per parameter than the Index median. Below baseline means the opposite: there is structural room for the same capability at lower parameter count.

How this relates to the K#. Two models with the same K# can have very different Compression Density readings. A high-density model is punching above its weight class; a low-density model is achieving the same result with parameter redundancy. Together with Knockout Analysis on the next-but-one page, this reading tells the customer whether their model's structural room is at the layer level (Knockout) or the parameter-count level (Compression Density).

PERFORMANCE METRICS

How well it predicts.

Standard held-out metrics, contextualized for a model. These measure output quality; the rest of the report measures structure.

Perplexity (held-out) **5.89** Strong for size

lower is better reference class:

Bits per byte **2.558** compression density

lower is better drives the Keel Score

> Perplexity is measured on held-out text the model never trained on. Training-set numbers measure memorization, not quality, and are never reported here.

COHERENCE & EFFICIENCY

Evenly built, partly idle.

COHERENCE SCORE

71.3/100

Good — capability is reasonably distributed, with some concentration early.

THERMAL GRADE

B

A | **B** | C | D | F

Good — roughly 81% of compute is doing useful work. The remainder is headroom.

Coherence and efficiency are independent of raw score. A high-scoring model can still be uneven or wasteful — which is precisely where the value of a scan lies.

CAPABILITY CERTIFICATE

What models at this level tend to do.

> These are **correlations from our scan database**, not guarantees and not causal claims. They describe what has been observed in models at a given structural level — they do not certify that this model will or will not perform a task.

CAPABILITY	TYPICALLY SEEN AT	THIS MODEL
Byte-level compression	Baseline+	observed
Short-sequence continuation	Baseline+	observed
Local pattern recognition	Emerging+	not yet observed
Consistent formatting	Emerging+	not yet observed
Multi-step coherence	Developing+	not yet observed
Cross-domain transfer	Advanced+	not yet observed
Long-range consistency	Advanced+	not yet observed
Instruction adherence	Frontier+	not yet observed
Robustness under shift	Frontier+	not yet observed

meta-llama/Llama-3.1-8B sits at the **Baseline** tier. Capabilities marked “not yet observed” are those our database associates with higher structural levels; they are an indication of headroom, not a verdict on the model.

CAPABILITY CONTEXT

Reading the certificate.

The tiers below describe increasing structural sophistication. They are drawn from the models in our scan database and are updated as the database grows.

Baseline	Compression and short continuation. The foundation every model reaches.	◂ YOUR MODEL
Emerging	Local patterns and consistent formatting begin to appear.	
Developing	Multi-step coherence starts to hold.	
Advanced	Cross-domain transfer and long-range consistency.	
Frontier	Instruction adherence and robustness under shift.	

> A model can be structurally excellent and still sit at a lower tier if it is small. Tier reflects observed capability range, not effort or design quality.

KNOCKOUT ANALYSIS

One redundant layer: layer 10.

We remove layers one at a time and measure the impact on held-out quality. A value near 1.00× means the model barely notices — a candidate for pruning. Sampled across the depth of the model:

LAYER REMOVED	IMPACT ON QUALITY	VERDICT
Layer 0	706.205×	**Essential**
Layer 8	1.064×	**Contributing**
Layer 10	1.044×	**Redundant**
Layer 16	1.160×	**Contributing**
Layer 21	1.095×	**Contributing**
Layer 24	1.082×	**Contributing**

Finding. Layer 10 showed negligible impact when removed from the sample of 6. Pruning is a low-risk optimization — one round of retraining and you keep the same output for less compute. A Keel Forge engagement can execute this end-to-end.

PARAMETER UTILIZATION & COST

What idle capacity is costing you.

Efficiency grade B implies roughly **81%** of meta-llama/Llama-3.1-8B's parameters are doing useful work — the remaining **19%** is compute you provision, power, and pay for on every request.

■ working ~81% ■ idle ~19%

WHAT 19% IDLE COSTS AT PRODUCTION VOLUME

MONTHLY INFERENCE VOLUME	TOTAL COMPUTE COST	RECLAIMABLE WASTE
100M tokens (light production)	$417–$1,400/mo	**$79–$266/mo**
1B tokens (typical production)	$4,170–$14,000/mo	**$792–$2,660/mo**
10B tokens (heavy production)	$41,700–$140,000/mo	**$7,923–$26,600/mo**

How to read this. Ranges span typical cloud pricing for a model of this size — from RunPod-tier rates ($1.50/hr A100) to hyperscaler-tier rates with modest reserved-capacity discounts ($4–5/hr A100). Idle share maps close to linearly onto compute cost for transformer inference.

Actual savings depend on your serving stack, batch configuration, and traffic pattern. A Keel Forge engagement measures the before-and-after on your specific infrastructure rather than estimating it.

PRESCRIPTIONS

What to do next.

Prioritized actions drawn from this report's findings, ordered by expected impact.

HIGH

Prune layer 10.

DO	Layer 10 showed ≤5% PPL degradation when removed. Remove and briefly retrain (10–20% of original training compute) to lock in the smaller model. Parameter reduction: 3%.
EXPECT	At 1B tokens/month, roughly $125–$420/month savings (range spans cloud pricing tiers). Same quality, fewer parameters, faster inference.

MEDIUM

Reduce weight precision (quantization).

DO	Independent of structural score, applicable to every model. Fp16 → int8 or int4 typically achieves 2–4× memory footprint reduction. Validate quality loss on your own eval set before shipping.
EXPECT	Memory drops by the compression ratio; latency and cost drop proportionally on memory-bound workloads. int8 typically ≤1% quality loss on structured tasks; int4 requires careful calibration.

MEDIUM

Optimize the serving stack.

DO	Well-built model — infrastructure wins are worth pursuing: KV-cache config, batch-size tuning, speculative decoding with a smaller draft model, tensor parallelism. Serving-stack changes that don't touch model weights.
EXPECT	Typical 20–100% throughput gains depending on baseline. Speculative decoding alone can 2–3× tokens/sec on latency-sensitive workloads.

INFO

Benchmark against the class leader.

DO	Ranks 4 of 14 in the Standard band. OLMo-2-1124-7B tops the class at K# 141.1 — a gap of 5.1 points. Structural gaps at this size class often correspond to capability gaps — worth an A/B evaluation.
EXPECT	A direct A/B on your own eval set. Either your model is fit-for-purpose, or the leader offers a materially better structural foundation.

LOW

Re-scan after optimization.

DO	Re-scan after any change and compare against this report.
EXPECT	Verified improvement on the same measurement scale.

How this was measured.

The Keel Scan reads a model's internal activation structure directly from its weights, using a proprietary methodology developed by Keel Technologies. No training data is required; the model is not modified.

Grade (A–F)	The headline result, anchored at K# = 100 (baseline). A at K# ≥ 140, B at 100–139 (above baseline), C at 75–99, D at 50–74, F below 50.
Keel Score (K#)	The underlying structural-quality score, positioned against our scan database. 100 is the baseline; above 100 is above average, below 100 is below.
Coherence	A measure of how evenly activation structure is distributed across layers.
Architecture Class	A structural signature computed from the activation profile.
Thermal Grade	An efficiency reading — the share of compute engaged in useful work.
Knockout	Empirical layer removal, measured against held-out quality.

GRADE COLOURS

A · Excellent B · Good C · Fair D · Poor F · Failing

Held-out evaluation uses text the model never trained on. We report what we can measure and mark what we cannot; capability correlations are labeled as such.

ABOUT KEEL TECHNOLOGIES

Measurement for intelligence.

Keel is a measurement company. The Keel Scan is the first instrument — a structural diagnosis of any AI model. Comparison, monitoring, optimization, and certification follow.

Keel Scan This report. $1,500 per model.	**Keel Compass** Same-model rescan — quantify what changed.
Keel Pulse Continuous monitoring and drift alerts.	**Keel Forge** Optimization — cut the waste this scan found.
Keel Fleet Portfolio management across many models.	**Keel Certify** Certification and trust mark.

Ready to act on this report? **contact@keeltech.co** · keeltech.co

KEEL Technologies

We grow intelligence.

Report prepared for the owner of meta-llama/Llama-3.1-8B
Scanned 2026-07-24T18:31:08Z
Keel Technologies, Inc. · Nashville, TN · keeltech.co