KEEL Technologies

KEEL SCAN · MODEL INTELLIGENCE REPORT

allenai/OLMo-2-1124-7B

7299M parameters · 32 layers · scanned 2026-07-24T18:44:21Z

A

GRADE A · EXCELLENT

K# 141.1

100 = BASELINE ACROSS THE INDEX

EXECUTIVE BRIEF

What this report says.

Your model scores **K# 141.1** on the Keel Index — a top-tier result, grade A. It ranks **1 of 14** in the Standard class. Your model tops its size class.

Three findings.

16%

of compute is not doing useful work

0 of 32

layers are redundant — every layer contributes to output

$667–$2,240

per month in avoidable spend at 1B tokens/month

Three actions, in order of impact.

1. Quantize weights fp16 → int8 (typically 2× memory reduction, ~1% quality loss)
2. Optimize the serving stack (typical 20–100% throughput gains)
3. Benchmark against the class leader on your own evaluation set

Next step. A **Keel Compass** rescan after any optimization delivers a measurable before/after on the same scoring scale. Contact **contact@keeltech.co** to scope.

HOW TO READ THIS REPORT

One diagnosis, five readings.

This report looks inside allenai/OLMo-2-1124-7B — 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.

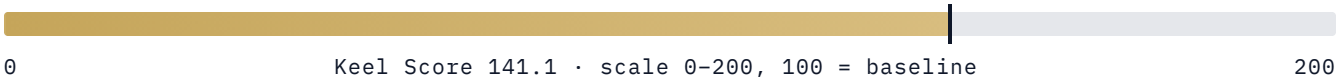
Keel Score (K#)	A single number for structural quality, positioned against every model we have scanned.
Architecture Class	The structural shape the model falls into, and what it predicts about behavior.
Coherence	How evenly capability is distributed across the model rather than concentrated.
Thermal 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

allenai/OLMo-2-1124-7B scores 141.1, grade A.

Against models of comparable size (Standard (5B–10B)), allenai/OLMo-2-1124-7B is **an excellent** result. Its structural shape is **Sequential — capability distributed evenly across depth**, and efficiency is graded **B** — 84% of its compute is doing useful work.



AT A GLANCE

READING	RESULT	RATING
Keel Score (K#)	141.1	A Excellent
Coherence Score	84.1 / 100	Strong
Architecture Class	Sequential	Evenly distributed
Thermal Grade	B	Good
Perplexity (held-out)	5.12	Strong for size
Redundant layers (sampled)	0 of 6	Optimization headroom

Verdict. A top-tier structural result in the Standard (5B–10B) class. The knockout analysis found no redundant layers in the sample: every layer contributes to output quality. The thermal grade is good — some idle capacity remains.

PEER COMPARISON

Rank 1 of 14 in your size class.

allenai/OLMo-2-1124-7B 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

OLMo-2-1124-7B ← your model	**K♯ 141.1**	**A**

> **Your model tops its size class.** No scanned peer in the Standard (5B–10B) band scored higher.

BELOW YOUR MODEL

Mistral-7B-v0.3	K♯ 140.1	A
Qwen2.5-7B	K♯ 137.4	B
Llama-3.1-8B	K♯ 136.0	B
Qwen2.5-Coder-7B	K♯ 134.4	B
starcoder2-7b	K♯ 133.9	B
gemma-2-9b	K♯ 132.5	B

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

ARCHITECTURE PROFILE

Sequential.

Every model resolves to one of five structural classes. allenai/OLMo-2-1124-7B is **Sequential**: its capability is distributed evenly across depth. Sequential models are typically well-balanced and predictable — every layer contributes proportionally, and the model degrades gracefully under perturbation.

PARAMETERS	LAYERS	WIDTH
7299M	32	4096
CONTEXT	**CLASS**	**UTILIZATION**
4096	Sequential	84%

THE FIVE CLASSES

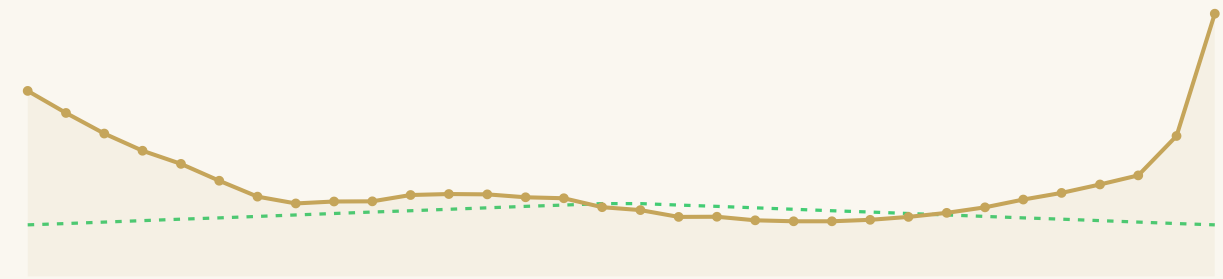
Sequential | Concentrated | Front-Loaded | Declining | Dual-Phase

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 allenai/OLMo-2-1124-7B'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.



This model Efficient target

Layer 0 → 31

PEAK VARIANCE	QUIETEST LAYER	SHAPE
layer 31	layer 20	Sequential

LAYER-BY-LAYER ANALYSIS

Where the work happens.

Activation variance across all 32 layers, sampled at 18 points evenly spaced through depth. Even distribution of variance across depth is the Sequential signature — no single region dominates.

LAYER	ACTIVATION VARIANCE	LEVEL
L0	1.5	High
L2	1.1	High
L4	0.9	Moderate
L5	0.8	Moderate
L7	0.6	Moderate
L9	0.6	Moderate
L11	0.7	Moderate
L13	0.6	Moderate
L15	0.5	Moderate
L16	0.5	Moderate
L18	0.5	Moderate
L20	0.4	Moderate
L22	0.4	Moderate
L24	0.5	Moderate
L26	0.6	Moderate
L27	0.7	Moderate
L29	0.8	Moderate
L31	2.1	Peak

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

108% 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) **5.12** Strong for size

lower is better reference class:

Bits per byte **2.355** 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

84.1/100

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

THERMAL GRADE

B

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

Good — roughly 84% 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.

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

allenai/OLMo-2-1124-7B sits at the **Advanced** 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.	
Emerging	Local patterns and consistent formatting begin to appear.	
Developing	Multi-step coherence starts to hold.	
Advanced	Cross-domain transfer and long-range consistency.	◂ YOUR MODEL
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

No redundant layers found in the sample.

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	2.910×	**Essential**
Layer 8	1.100×	**Contributing**
Layer 10	1.096×	**Contributing**
Layer 16	1.115×	**Contributing**
Layer 21	1.137×	**Contributing**
Layer 24	1.127×	**Contributing**

Finding. All 6 sampled layers contributed measurably to output quality. That is unusual and structurally significant — it means the model has minimal redundancy at the layer level. Optimization work would need to target sub-layer components (attention heads, MLP neurons), not whole layers.

PARAMETER UTILIZATION & COST

What idle capacity is costing you.

Efficiency grade B implies roughly **84%** of allenai/OLMo-2-1124-7B's parameters are doing useful work — the remaining **16%** is compute you provision, power, and pay for on every request.

■ working ~84% ■ idle ~16%

WHAT 16% IDLE COSTS AT PRODUCTION VOLUME

MONTHLY INFERENCE VOLUME	TOTAL COMPUTE COST	RECLAIMABLE WASTE
100M tokens (light production)	$417–$1,400/mo	**$67–$224/mo**
1B tokens (typical production)	$4,170–$14,000/mo	**$667–$2,240/mo**
10B tokens (heavy production)	$41,700–$140,000/mo	**$6,672–$22,400/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

Layer-level pruning found no cuts — move to sub-layer optimization.

DO — Every layer contributes measurably. Next optimization surface is head-level and MLP-level. Michel et al. 2019 and Voita et al. 2019 consistently find 30–50% of attention heads redundant after training. A targeted head-pruning study is the appropriate next step.

EXPECT — Typical head-pruning outcomes: 15–30% inference cost reduction with negligible quality loss. 2–4 weeks of engineering effort plus controlled evaluation.

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.

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 allenai/OLMo-2-1124-7B
Scanned 2026-07-24T18:44:21Z
Keel Technologies, Inc. · Nashville, TN · keeltech.co