ESANN 2026 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 22-24 April 2026, i6doc.com publ., ISBN 9782875870964. Available from http://www.i6doc.com/en/.

Low-Rank Lens for Scalable LLMs Interpretability

Giuseppe Trimigno[1], Gianfranco Lombardo[1] and Stefano Cagnoni[1]

University of Parma - Department of Engineering and Architecture
Parco Area Delle Scienze, Parma, 43125 - Italy

Abstract. Representation lenses expose layer-wise predictions in LLMs. Current methods rely on full-rank affine maps with quadratic cost. However, spectral evidence across multiple model families shows these maps are intrinsically low-rank. We propose LoRA-Lens, a low-rank residual alignment mechanism that reduces parameters by over 95% while preserving fidelity to the model's final output. Experiments on OLMo, Qwen, and Gemma (up to 32B) demonstrate strong fidelity, large memory savings, robust transfer to instruction-tuned models, and effective early-exit inference.

1 Introduction

Large language models (LLMs) develop rich internal representations as predictions evolve across layers, and intermediate activations often encode key semantic and predictive signals [1, 2]. Representation lenses aim to expose these layer-wise predictions by projecting hidden states onto the output vocabulary space.

Background. LogitLens [3] first introduced direct projection via the unembedding matrix; however the mismatch between intermediate and output spaces often yields incoherent predictions. TunedLens [4] addresses this misalignment via learned per-layer affine maps; subsequent variants have extended the idea to applications: multi-token prediction [5], early prediction recovery [6], and localized representational analysis [7]. However, all such lenses require $\mathcal{O}(H^2)$ parameters per layer, with H being the latent space dimension. For modern LLMs, this results in billions of extra weights and large training data requirements, making lens-based interpretability a significant challenge.

Motivation and contributions. We investigated whether lenses truly require a quadratic parameterization. By performing a spectral analysis of learned affine transformations across several model families using three standard SVD-based metrics: effective rank [8], spectral energy ratio [9], and participation ratio [10]. We found that the lens mappings operate in a highly compressed subspace. Figure 1 shows an example on the OLMo family (1B-32B): the effective rank remains below 3% of H, most spectral energy lies in the top 50% components, and the participation ratio lies amid values near half of H. These findings indicate that low-rank formulations may capture the essential predictive structure while substantially reducing parameter cost. Motivated by this evidence, our contributions are: (1) we introduce **LoRA-Lens**, a low-rank residual alignment

ESANN 2026 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 22-24 April 2026, i6doc.com publ., ISBN 9782875870964. Available from http://www.i6doc.com/en/.

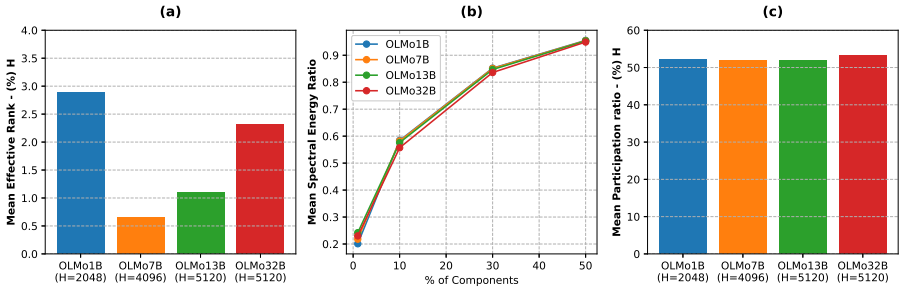


Fig. 1: Spectral analysis of *TunedLens* [4] on the OLMo family. Metrics are averaged across layers: (a) Mean effective rank, (b) Spectral energy ratio, and (c) Mean participation ratio. All indicate a highly compressed subspace of the latent representation.

mechanism inspired by LoRA [11], using rank-constrained updates to align intermediate states, reducing complexity (over 95% fewer parameters) while preserving alignment fidelity; (2) we show robust transfer from base to instruction-tuned variants; and (3) we demonstrate the presence of linear shortcuts and their practical utility for early-exit inference.

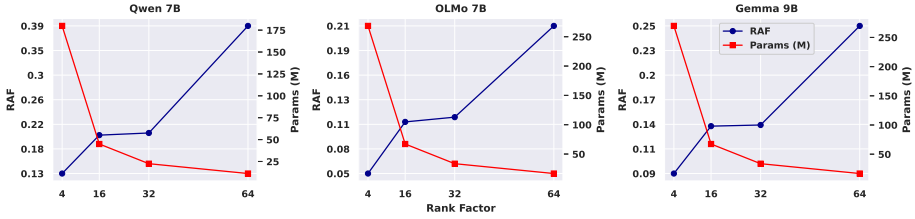


Fig. 2: Relative alignment fidelity between LoRA-Lens and TunedLens versus parameter count. k=32 offers the best trade-off between efficiency and fidelity.

2 Methodology

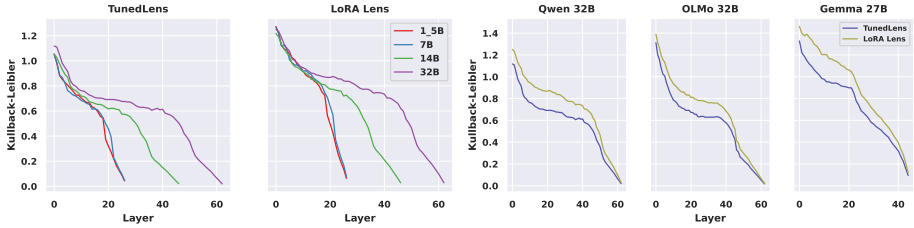
Let an LLM be defined by L layers with hidden dimension H and vocabulary size V. Let $h_\ell \in \mathbb{R}^H$ denote the hidden state at layer ℓ, and $W_U \in \mathbb{R}^{V\times H}$ the pre-trained unembedding matrix. A representation lens maps h_ℓ to a vocabulary distribution z_ℓ approximating the model's final prediction z_L. Full-rank lenses learn a per-layer affine map $W_\ell \in \mathbb{R}^{H\times H}$ producing logits $z_\ell = \text{softmax}(W_U(h_\ell + W_\ell h_\ell))$, whose $\mathcal{O}(H^2)$ parameters dominate the cost.

Low-Rank Lenses. LoRA-Lens replaces W_ℓ with a low-rank residual update aligning h_ℓ to the output space via $A_\ell \in \mathbb{R}^{H\times r}$ and $B_\ell \in \mathbb{R}^{r\times H}$, with $r \ll H$:

$$z_\ell = \text{softmax}(W_U(h_\ell + A_\ell B_\ell h_\ell)) \,.$$

This reduces complexity from $\mathcal{O}(LH^2)$ to $\mathcal{O}(LHr)$ and the learnable parameters per layer from H^2 to $2Hr$ (about 94% fewer when $H = 4096$, $r = 128$).

ESANN 2026 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 22-24 April 2026, i6doc.com publ., ISBN 9782875870964. Available from http://www.i6doc.com/en/.



(a) **(b)**

Fig. 3: Per-layer KL divergence comparison between *TunedLens* and *LoRA-Lens* on the SlimPajama validation subset. (a) Qwen family; (b) largest OLMo (32B), Qwen (32B), and Gemma (27B) models.

Training. Training uses a 226M-token SlimPajama [12] subset (Table 1). We optimize A_ℓ and B_ℓ for 500 steps using AdamW using a batch size of 2^{18} tokens. The objective minimizes the KL divergence to the final layer, encouraging each lens to anticipate the model's prediction.

Subset	CC	C4	ArXiv	GitHub	StackEx	Wiki	Total
Train	130.33M	65.72M	11.52M	9.66M	7.31M	1.95M	**226.49M**
Eval	10.83M	5.20M	0.99M	0.92M	0.54M	0.18M	**18.66M**

Table 1: Tokens composition of the LoRA-Lens train and validation subsets.

Rank Selection. The rank scales with the hidden size as $r = \lfloor H/k \rfloor$. We evaluate $k \in \{4, 16, 32, 64\}$ on OLMo 7B, Qwen 7B, and Gemma 9B to identify a good trade-off between parameter efficiency and fidelity. To quantify fidelity, we introduce the *Relative Alignment Fidelity* (RAF), defined as the mean absolute difference between the per-layer KL divergences produced by *LoRA-Lens* and *TunedLens*:

$$\mathrm{RAF} = \frac{1}{L} \sum_{\ell=1}^{L} \left| \mathrm{KL}_\ell^{\text{LoRA-Lens}} - \mathrm{KL}_\ell^{\text{TunedLens}} \right|.$$

Jointly considering RAF and parameter overhead, $k = 32$ provides the best balance, yielding substantial parameter reduction with negligible loss in alignment fidelity (Figure 2).

3 Experiments and Results

We evaluate **LoRA-Lens** along five axes: (1) *Per-layer fidelity*; (2) *Memory requirements*; (3) *Transferability* from base to instruction-tuned models; (4) *Latent linear shortcuts*; and (5) *Early-exit inference*. Experiments cover three model families: OLMo (1B–32B) [13], Qwen (1.5B–32B) [14], and Gemma (2B–27B) [15]. We compare our approach against *TunedLens*, the full-rank baseline. **Fidelity.** Figure 3 shows that LoRA-Lens closely matches the TunedLens across architectures and scales. For Qwen, OLMo, and Gemma models, the average

ESANN 2026 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 22-24 April 2026, i6doc.com publ., ISBN 9782875870964. Available from http://www.i6doc.com/en/.

per-layer KL difference remains small. Looking at the RAF: 0.12 for OLMo (1B-32B), 0.15 for Qwen (1.5B-32B), and 0.14 for Gemma (2B-27B), indicating that low-rank alignment preserves predictive fidelity across model sizes.

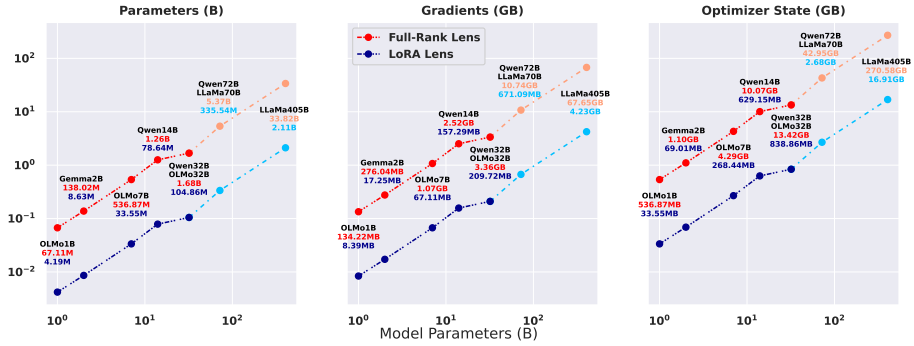


Fig. 4: Log-log Memory comparison (parameters, gradients, optimizer state) for full-rank lenses vs. LoRA-Lens. Lighter bars correspond to large models evaluated only for memory usage.

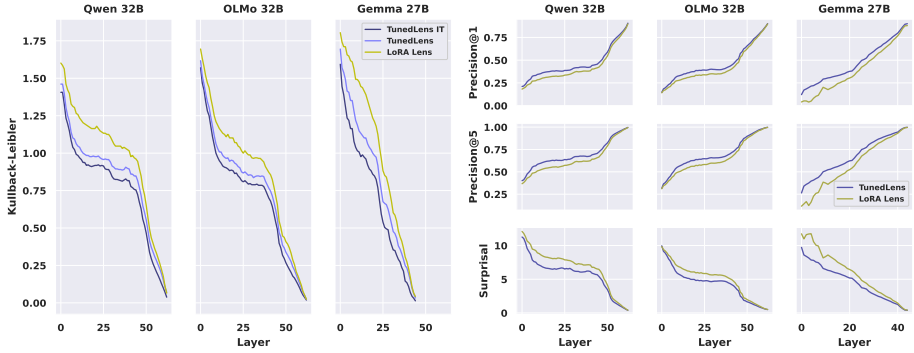
Memory Requirements. Figure 4 summarizes memory use, including two larger LLaMA models (70B and 405B) analyzed only for this metric. The low-rank formulation yields an average 95% reduction in parameters, gradient memory, and optimizer state, scaling linearly with model size.

Transferability. We test whether lenses trained on base models generalize to their instruction-tuned counterparts without retraining. LoRA-Lens and TunedLens trained on base Gemma 27B, Qwen 32B, and OLMo 32B are applied directly to their instruction-tuned versions and evaluated on AlpacaEval [16]. As shown in Figure 5a, KL divergence remains low and close to lenses trained natively on instruct models ($TunedLensIT$): TunedLens transfers with 0.06–0.13 RAF, and LoRA-Lens with 0.16–0.28. Despite the slightly larger gap with Gemma 27B, the alignment geometry is largely preserved across the models, indicating that low-rank lenses maintain robust transfer under instruction tuning.

Latent Linear Shortcuts. We evaluate whether intermediate representations permit accurate linear decoding. Using lenses trained on base models and applied to instruction-tuned variants, we measure (1) *Precision@k* for $k \in \{1, 5\}$ and (2) *Surprisal*, the negative log-likelihood of the lens' top-1 prediction under the final distribution. Across OLMo 32B, Qwen 32B, and Gemma 27B, LoRA-Lens matches TunedLens closely, with Precision@1 and Precision@5 differences in the 0.04–0.08 and 0.04–0.09 ranges, respectively, and surprisal differences in 0.71–1.32. Figure 5b shows that Precision@1 increases sharply after mid-depth, exceeding 80% agreement with the final output, validating that decisive information emerges early and lies within the low-rank subspace captured by LoRA-Lens.

Early Exiting. Beyond interpretability, we further evaluate LoRA-Lens in a downstream early-exit setting. Following [17], inference halts at layer ℓ when

ESANN 2026 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 22-24 April 2026, i6doc.com publ., ISBN 9782875870964. Available from http://www.i6doc.com/en/.



(a) **(b)**

Fig. 5: (a) Transfer from base to instruction-tuned models (KL divergence). (b) Linear shortcut behavior on instruct models using transferred lenses.

the confidence gap between top logits satisfies

$$\Delta_i > 0.9\lambda + 0.1e^{-4i/N},$$

with $\lambda \in [0, 1]$ controlling sensitivity and N the average input length. Using only lenses trained on base models, we evaluate Qwen 32B-instruct, OLMo 32B-instruct, and Gemma 27B-instruct on AlpacaEval and report *Exit Rate* (fraction of tokens halted early) and *Exit Precision* (agreement with the final prediction). LoRA-Lens reduces latency by early exiting on about 50% of tokens early while maintaining over 90% agreement with the final output (Figure 6), demonstrating the reliability of low-rank latent predictions in downstream settings.

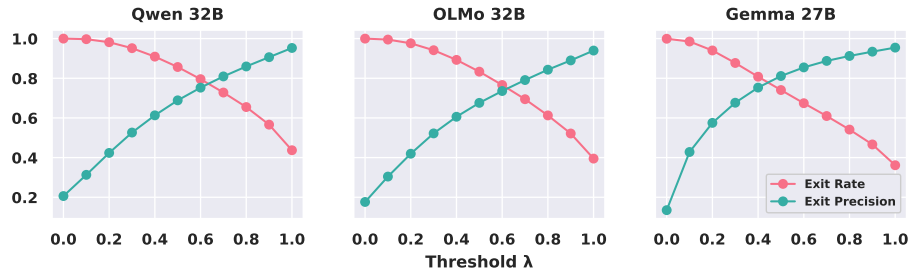


Fig. 6: Early-exit evaluation on AlpacaEval by varying the threshold λ.

4 Conclusion and Discussion

We introduced LoRA-Lens, a low-rank alternative to full-rank lenses, motivated by spectral evidence that affine maps act in compressed subspaces. By replacing quadratic transformations with rank-constrained updates, we reduce parameters by over 95% while preserving alignment fidelity and achieving linear complexity

ESANN 2026 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 22-24 April 2026, i6doc.com publ., ISBN 9782875870964. Available from http://www.i6doc.com/en/.

with respect to H. Experiments show strong per-layer accuracy, large memory savings, robust transfer, and effective early-exit inference, indicating that low-rank parameterizations are sufficient for scalable lens-based interpretability. A remaining limitation is that the fixed rank rule (k=32) may not be optimal across architectures; exploring adaptive per-layer ranks and extending the method to multimodal or sparse-expert models are promising directions.

References

[1] M. Geva, A. Caciularu, G. Dar, P. Roit, S. Sadde, M. Shlain, B. Tamir, and Y. Goldberg. LM-debugger: An interactive tool for inspection and intervention in transformer-based language models. In *EMNLP 2022 System Demonstrations*, pages 12–21, 2022.

[2] Giuseppe Trimigno, Gianfranco Lombardo, Michele Tomaiuolo, Stefano Cagnoni, and Agostino Poggi. Llms in staging: An orchestrated llm workflow for structured augmentation with fact scoring. *Future Internet*, 17(12), 2025.

[3] Nostalgebraist. Interpreting gpt: The logitlens, 2020.

[4] N. Belrose, Z. Furman, L. Smith, D. Halawi, I. Ostrovsky, L. McKinney, S. Biderman, and J. Steinhardt. Eliciting latent predictions from transformers with the tuned lens, 2023.

[5] Koyena Pal, Jiuding Sun, Andrew Yuan, Byron Wallace, and David Bau. Future lens: Anticipating subsequent tokens from a single hidden state. In *Proceedings of the 27th Conference on Computational Natural Language Learning (CoNLL)*, pages 548–560, Singapore, December 2023. Association for Computational Linguistics.

[6] Alexander Din, Taelin Karidi, Leshem Choshen, and Mor Geva. Jump to conclusions: Short-cutting transformers with linear transformations. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 9615–9625, Torino, Italia, 2024. ELRA and ICCL.

[7] Asma Ghandeharioun, Avi Caciularu, Adam Pearce, Lucas Dixon, and Mor Geva. Patchscopes: A unifying framework for inspecting hidden representations of language models. In *International Conference on Machine Learning*, pages 15466–15490. PMLR, 2024.

[8] Olivier Roy and Martin Vetterli. The effective rank: A measure of effective dimensionality. In *2007 15th European signal processing conference*, pages 606–610. IEEE, 2007.

[9] Ian T Jolliffe and Jorge Cadima. Principal component analysis: a review and recent developments. *Philosophical transactions of the royal society A: Mathematical, Physical and Engineering Sciences*, 374(2065):20150202, 2016.

[10] S. Recanatesi, S. Bradde, V. Balasubramanian, N. A Steinmetz, and E. Shea-Brown. A scale-dependent measure of system dimensionality. *Patterns*, 3(8), 2022.

[11] Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, et al. Lora: Low-rank adaptation of large language models.

[12] Zhiqiang Shen, Tianhua Tao, Liqun Ma, Willie Neiswanger, Zhengzhong Liu, Hongyi Wang, Bowen Tan, Joel Hestness, Natalia Vassilieva, Daria Soboleva, and Eric Xing. Slimpajama-dc: Understanding data combinations for llm training, 2024.

[13] Team OLMo. 2 olmo 2 furious, 2025.

[14] Team Qwen. Qwen2.5 technical report, 2025.

[15] Team Gemma. Gemma 2: Improving open language models at a practical size, 2024.

[16] Xuechen Li, Tianyi Zhang, Yann Dubois, Rohan Taori, Ishaan Gulrajani, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. Alpacaeval: An automatic evaluator of instruction-following models. https://github.com/tatsu-lab/alpaca$_e$*val*, 52023.

[17] T. Schuster, A. Fisch, J. Gupta, M. Dehghani, D. Bahri, V. Tran, Y. Tay, and D. Metzler. Confident adaptive language modeling. In *Advances in Neural Information Processing Systems*, volume 35, pages 17456–17472. Curran Associates, Inc., 2022.