HORT HORIZONS AND SPARSE CONCEPTS: A MATHEMATICAL VIEW OF THE READOUT IN THE J-LENS

Anonymous authors
Paper under double-blind review

ABSTRACT

The Jacobian lens (J-lens) has been proposed as a way to read verbalizable representations from language models. However, its principle and meaning lack a detailed and theoretical discussion. We provide a mathematical view of this interpretation and of its assumed causal structure. Besides treating the J-lens as a heuristic probe, we further regard it as a first-order causal transfer operator from intermediate activations to expected future readouts. We study the Jacobian matrix as the optimal local linear approximation of the downstream mapping, analyze its global approximation behavior and bias, and identify its mathematical meaning as an expectation over anticipated future readouts. Further analysis of the Jacobian energy distribution reveals that its causal geometry is highly sparse. The energy decays with depth, concentrates in an extremely small proportion, and decomposes into diagonal pathways and specific critical positions. This decomposition further resolves the expectation of the J-lens over future outputs into short-horizon and sparse concept predictions, providing a more intuitive attribution and explanation for the ability of the J-lens to visualize concepts during the thinking process. Based on the theory, we propose a simple but effective improvement strategy, which significantly enhances the ability of the J-lens to read out correct intermediate concepts.

1 INTRODUCTION

Most computation in large language models (LLMs) occurs in the background (Brown et al., 2020; Wei et al., 2022). A prompt may elicit a brief answer, but this answer arises from a sequence of complex intermediate transformations whose content cannot be directly observed. Mechanistic interpretability attempts to open this black box by reading internal states, intervening on them, and relating them to behavior (Sharkey et al., 2025; Alain & Bengio, 2017; Hewitt & Liang, 2019). A central difficulty is that intermediate-layer representations are not naturally expressible in the vocabulary of human concepts, and their connection to the output layer is indirect. The residual stream is a high-dimensional vector space, whereas interpretation requires a structure that is sparse, reportable, and causally meaningful.

The recently proposed Jacobian lens, or J-lens, addresses this difficulty by propagating a hidden state along the downstream computation of the model and measuring how its perturbation affects the final-layer state (Gurnee et al., 2026). The resulting J-space is described as sparse and verbalizable: hidden states can be approximated by nonnegative combinations of a few J-lens vectors that support reporting, modulation, and selective computation (Olsson et al., 2022). However, it lacks a clear and detailed mathematical interpretation. The reasons why averaging Jacobian matrices over future positions yields valid tokens, and how the resulting representations relate to the internal thoughts of the model remain unknown. The J-lens claims that the first-order sensitivity of future outputs to hidden states is itself a carrier of verbalizable tendencies. Evaluating this claim requires establishing in what sense the Jacobian matrix is optimal, what objective it approximates, and whether the causal influence behind the average is uniformly distributed or concentrated at specific positions.

To address this issue, we provide a clear mathematical interpretation of the J-lens and empirically validate its effectiveness and identify its practical deviations. Given an input $h_{t,l}$, the J-lens can be viewed as a global fit of the first-order linear approximation to the latent mapping $\mathbb{E}_{t'>t}[h_{t',L}]$. Conceptually, we regard the J-lens as an averaged first-order causal transfer operator. In a given context,

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

the Jacobian matrix is the optimal local linear approximation of how a small change in an intermediate activation affects subsequent outputs. Under this formulation, the residual vanishes to first order. Averaging this operator over samples yields the practical J_ℓ, which provides the globally optimal approximation under the minimum mean squared error criterion. We use a Stein-based identity to justify the passage from local approximation to global fitting, estimate the potential theoretical bias, and provide experimental evidence that corroborates this conclusion from the reverse direction.

Building on this, we further provide an attribution of the causal mapping fitted by the J-lens. The averaged operator J_ℓ compresses all position-to-position Jacobian matrices into a single matrix, but the underlying sensitivity is not uniform. We therefore study the Jacobian energy $E_L(t, t')$ between a source position t and a target position t'. Experiments validate the sparsity of the Jacobian energy: on the one hand, the energy tends to be larger in earlier layers and much smaller near the output; on the other hand, within a single layer, the Jacobian energy exhibits strong concentration, and a small number of pairs carry most of the total energy. Through visualization, we observe that the concentration of energy can be grouped into two patterns: the diagonal pattern ($t \approx t'$) and the horizontal or vertical pattern ($t = t^*$ or $t' = t^*$). They represent two effective outcomes of the J-lens mapping in practice: short-horizon token prediction and highly sensitive tokens at critical positions. We refer to these two outcomes as the short horizons and sparse concepts. The readout results of the practical J-lens also effectively corroborate our conclusion.

Based on the above theoretical and experimental results, we design experiments to further validate the effectiveness of our interpretation by improving the J-lens. The concentration effect of the Jacobian energy indicates that the predictions of the J-lens naturally align with short-horizon tokens and sparse concepts. Averaging the Jacobian matrices over all positions turns unimportant positions into noise and affects the practical representations of the J-lens. We therefore introduce a simple but effective filtering strategy that uses only a subset of positions with the highest energy in the J-lens construction. Experiments show that using only the Jacobian matrices from the top 10% to 20% of positions by energy significantly improves the capability of the J-lens.

In summary, this work makes three contributions. First, we provide a mathematical foundation for the J-lens and establish it as the expectation over anticipated outputs. Through local linearization, global least-squares projection, and a Stein-type bridge, we connect the J-lens to the optimal linear readout and explicitly characterize non-Gaussian and nonlinear deviations. Second, we present experimental analyses and visualization studies that cluster the expectation over anticipated outputs into short-horizon token predictions and critical concept predictions. Third, we design a lens improving method that further validates the reliability and effectiveness of our conclusions.

2 Background: J-lens and Global Workspace

A transformer language model maintains, at each token position, a residual-stream vector that is read from and written to by every layer. Early states are close to the input token, whereas final states are directly mapped to vocabulary scores by the unembedding matrix W_U. Intermediate states are therefore computational, but not directly verbal. The logit lens reads them as if all layers shared one coordinate system; this works near the output but often fails earlier, where representations have not yet been rotated into output coordinates (nostalgebraist, 2020; Belrose et al., 2023; Geva et al., 2022).

The Jacobian lens (J-lens) instead reads an intermediate activation by its average first-order effect on present and future outputs (Gurnee et al., 2026). For layer ℓ, it averages the Jacobian from a source activation to later final-layer states over source positions, target positions, and prompts:

$$J_\ell = \mathbb{E}_{t,t' \geq t,\text{prompt}} \left[\frac{\partial h_{\text{final},t'}}{\partial h_{\ell,t}} \right], \qquad \text{lens}(h_\ell) = \text{softmax}(W_U \, \text{norm}(J_\ell h_\ell)) \,. \tag{1}$$

The top tokens of this readout describe what an activation is disposed to say across contexts, rather than what one context happens to say. The rows of $W_U J_\ell$ define token-associated readout directions; their sparse nonnegative combinations form the J-space.

Functionally, it supports five properties corresponding to conscious access: verbal report, directed modulation, internal reasoning, flexible generalization, and selectivity. However, a more detailed theoretical account is missing, and the J-lens itself is approximate and limited in applications. This

paper takes this challenge seriously. We investigate why the averaged Jacobian matrix can approximate the optimal readout, how the approximation is affected by nonlinear and non-Gaussian bias, and what sparse causal structure over positions underlies this interpretation.

3 Mathematical Explanation of J-lens: Expectation of the Future Readouts

Given a layer ℓ and position t. We study the relationship between the J-lens and the mapping function of the averaged future readouts:

$$y_t = f_\ell(h_{\ell,t}) := \mathbb{E}_{t' \geq t}[h_{\text{final},t'}], \qquad f_\ell : \mathbb{R}^d \to \mathbb{R}^d, \tag{2}$$

where $h_{\ell,t}$ is the hidden states at layer ℓ and position t, and y_t is an averaged future final-layer responses $h_{\text{final},t'}|_{t' \geq t}$. The function f_ℓ is implicitly defined by the remaining network and is generally nonlinear. In this section, we systematically analyze the mathematical principles underlying the J-lens, corroborate its first-order approximation to the nonlinear mapping f_ℓ, and confirm its meaning. Specifically, given the local function-fitting capability of the Jacobian matrix, we mathematically interpret the construction of the J-lens, i.e., the statistical expectation of the Jacobian over the training distribution, as a process of fitting f_ℓ from local approximation to global (Stein, 1981). In addition, we analyze its approximation bias from both theoretical and experimental perspectives, validating the reliability and effectiveness of the theoretical interpretation in this section while providing a new perspective on the failure of the J-lens in early layers.

3.1 Local View: Tangent Linearization

For a smooth map $f : \mathbb{R}^d \to \mathbb{R}^d$, the Jacobian $J_f(h_0) \in \mathbb{R}^{d \times d}$ has entries $[J_f(h_0)]_{ij} = \partial f_i / \partial h_j$ at h_0. The first-order Taylor expansion is

$$f(h_0 + \delta) = f(h_0) + J_f(h_0) \cdot \delta + o(\|\delta\|). \tag{3}$$

Thus the Jacobian can be regarded as the local sensitivity map: a small intervention δ at $h_{\ell,t}$ changes the predicted future response, to first order, by $J_{f_\ell}(h_{\ell,t}) \cdot \delta$. The approximation is asymptotically exact as $\|\delta\| \to 0$, but its usable radius shrinks where curvature is large, and the matrix is pointwise rather than global. Therefore, a single Jacobian matrix solely estimates the first-order approximation with tangent linearization in a local range.

3.2 Global View: Least-squares Readout

The global view fixes one affine map for the whole distribution. Let $h \sim p$, $y = f(h)$, $\mu = \mathbb{E}[h]$, and $\Sigma = \mathrm{Cov}(h)$, with $\Sigma \succ 0$. Writing $\tilde{h} = h - \mu$ and $\tilde{y} = y - \mathbb{E}[y]$, a global first-order approximation of the function f can be treated as a least-squares estimation:

$$\min_{W,b} \ \mathbb{E}\|y - (Wh + b)\|^2 \tag{4}$$

with the normal-equation solution

$$\begin{cases} W^* = \mathbb{E}[\tilde{y}\,\tilde{h}^\top]\,\Sigma^{-1} = \mathrm{Cov}(y,h)\,\mathrm{Cov}(h)^{-1} \\ b^* = \mathbb{E}[y] - W^*\mu. \end{cases} \tag{5}$$

The residual is orthogonal to all affine functions of h, so $W^*h + b^*$ is the L^2 projection of the regression function $m(h) = \mathbb{E}[y \mid h]$ onto affine functions. This optimum requires no Gaussianity and no differentiability; it is a statistical compromise slope over the data distribution.

3.3 Score identity and the Stein bridge

The link between local derivatives and global least squares is integration by parts. Assume that h has a smooth positive density p on $\mathbb{R}^d$, that $f : \mathbb{R}^d \to \mathbb{R}^d$ is componentwise differentiable, and that f grows slowly enough relative to p that boundary terms vanish (Hyvärinen & Dayan, 2005). The score function is

$$s(h) := \nabla \log p(h) = \frac{\nabla p(h)}{p(h)}. \tag{6}$$

For one component f_i and one input coordinate h_j,

$$\mathbb{E}[\partial_j f_i(h)] = \int_{\mathbb{R}^d} \partial_j f_i(h)\, p(h)\, dh = -\int_{\mathbb{R}^d} f_i(h)\, \partial_j p(h)\, dh, \tag{7}$$

where the boundary term vanishes under the growth assumption, where $f_i(h)p(h) \to 0$ as $\|h\| \to \infty$ for every component i. Since $\partial_j p = p\, \partial_j \log p = p\, s_j$ as shown in Eq. 6,

$$\mathbb{E}[\partial_j f_i(h)] = -\mathbb{E}[f_i(h) s_j(h)]. \tag{8}$$

Stacking over i and j gives the multivariate score identity:

$$\mathbb{E}[J_f(h)] = -\mathbb{E}\big[f(h) s(h)^\top\big]. \tag{9}$$

The left side averages local slopes. The right side says that this average is determined by correlations between function values and the local tilt of the density. Derivative information is therefore encoded in how the density changes, not only in how the function bends.

Equation equation 9 is distribution-free once the regularity conditions hold. The Gaussian case is special because the score is affine. If $h \sim \mathcal{N}(\mu, \Sigma)$ with $\Sigma \succ 0$, then

$$\log p(h) = \text{const} - \frac{1}{2}(h-\mu)^\top \Sigma^{-1}(h-\mu), \qquad s(h) = -\Sigma^{-1}(h-\mu). \tag{10}$$

Substituting into equation 9,

$$\mathbb{E}[J_f(h)] = -\mathbb{E}\Big[f(h)\left(-\Sigma^{-1}(h-\mu)\right)^\top\Big] = \mathbb{E}\big[f(h)(h-\mu)^\top\big]\,\Sigma^{-1}. \tag{11}$$

Because $\mathbb{E}[s(h)] = 0$, replacing f by $\tilde{f} = f - \mathbb{E}[f(h)]$ does not change the right-hand side. Hence

$$\mathbb{E}[J_f(h)] = \mathbb{E}[\tilde{f}(h)\tilde{h}^\top]\Sigma^{-1} = \mathrm{Cov}(f(h), h)\Sigma^{-1} = W^*. \tag{12}$$

This is the Stein bridge: under Gaussian inputs, the average Jacobian is exactly the population least-squares slope.

Theorem 1 (Stein bridge (Stein, 1981)). *Let $h \sim \mathcal{N}(\mu, \Sigma)$ with $\Sigma \succ 0$, and let f satisfy the mild growth and differentiability conditions above. Then*

$$\mathbb{E}[J_f(h)] = \mathrm{Cov}(f(h), h)\,\mathrm{Cov}(h)^{-1} = W^*. \tag{13}$$

Under Gaussian inputs, the averaged Jacobian matrix is exactly equal to the population least-squares slope. Therefore, the J-lens can be understood as an approximation to the optimal linear readout. When the input distribution is close to Gaussian, this approximation is accurate; when the input distribution deviates substantially from Gaussian, an additional residual term is introduced into the score function, inducing a systematic bias between the averaged Jacobian and the optimal linear readout. In the subsequent section, we will analyze this bias and evaluate the layer-level valid range of our theoretical interpretation.

3.4 Bias Analysis

The J-lens involves two approximations. The first replaces the nonlinear mapping with a first-order linear transfer matrix, and the second approximates the globally optimal transfer matrix by the expectation of local Jacobians. Nonlinear and non-Gaussian biases are therefore systematically introduced and amplified as the distance between source and target layers increases. In this section, we analyze in detail the specific effects of these biases from both theoretical and experimental perspectives and provide a new explanation for the early failure of the J-lens.

3.4.1 Nonlinear Bias

The Stein bridge compares two slopes, W^* and $\mathbb{E}[J_f(h)]$. A separate error remains even if the best possible slope is used: a linear map cannot reproduce curvature. Let $\mu = \mathbb{E}[h]$, $\delta = h - \mu$, $J = J_f(\mu)$, and let $H_i = \nabla^2 f_i(\mu)$ be the Hessian of output component i. Write $H[\delta, \delta]$ for the vector with components $\delta^\top H_i \delta$ and $H[\Sigma] = \mathbb{E}[H[\delta, \delta]]$, whose i-th component is $\mathrm{tr}(H_i\Sigma)$. The intrinsic nonlinear error of the optimal affine readout is

$$\varepsilon^2_{\text{lin}} = \mathbb{E}\big\| f(h) - \big(\mathbb{E}[f(h)] + W^*(h-\mu)\big)\big\|^2. \tag{14}$$

This is the residual that no choice of linear slope can remove.

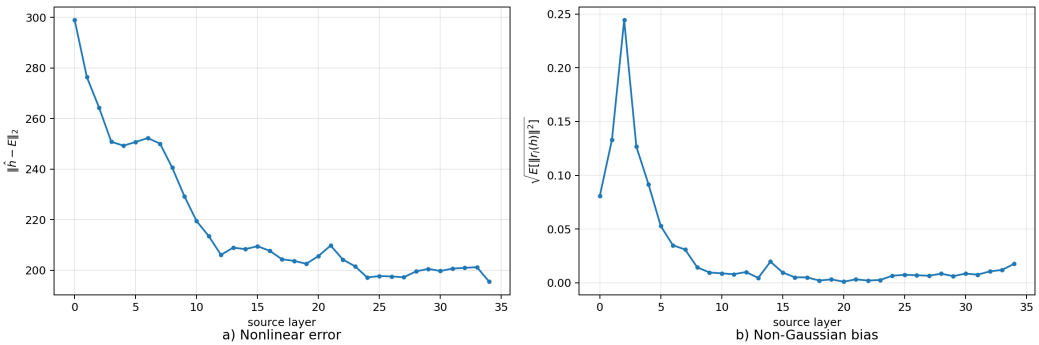

Figure 1: Layer-wise bias analysis of J-lens for Qwen3-8B.

3.4.2 Non-Gaussian Bias

Away from Gaussianity, decompose the score into its Gaussian part plus a residual:

$$s(h) = -\Sigma^{-1}(h-\mu) + r(h), \qquad r(h) = 0 \iff h \text{ is Gaussian.} \tag{15}$$

Substitution into the score identity gives the exact bias formula

$$\epsilon_{\text{Gau}} := W^* - \mathbb{E}[J_f(h)] = \mathbb{E}\Big[\tilde{f}(h)\, r(h)^\top\Big]. \tag{16}$$

The gap is therefore not controlled by nonlinearity alone. It is the covariance between the centered output fluctuation and the pointwise non-Gaussianity of the density. A standard Cauchy–Schwarz bound yields

$$\|\epsilon_{\text{Gau}}\|_F \le \sqrt{\mathbb{E}\|\tilde{f}(h)\|^2}\,\sqrt{\mathbb{E}\|r(h)\|^2} \tag{17}$$

where the second factor is a distribution-level non-Gaussianity measure.

3.4.3 Overall Bias

As a cascade bias system without correlation, the overall system bias can be calculated by a sum of the squared biases:

$$\epsilon_{\tilde{J}}^2 = \epsilon_{\text{lin}}^2 + \|\epsilon_{\text{Gau}}\|_F^2. \tag{18}$$

To verify the consistency between theory and experiments, we perform a layer-wise quantitative comparison between the theoretical error and the empirical error of the J-lens. The results show that both the nonlinear bias and the non-Gaussian error decrease as the source layer approaches the target layer, and that empirical measurements of the overall error also validate our theory. This provides an effective validation of our mathematical interpretation of the J-lens and explains the early failure of the J-lens from the perspective of mathematical bias: large nonlinear and distributional biases cause the mapped results at the target layer to deviate substantially from the confidence interval and lose their intended meaning.

4 Jacobian Energy Structure in J-lens: Short Horizon and Sparse Concepts

This section moves from the validity of the averaged operator J_ℓ to the positional structure that the average hides. The theory above treats J_ℓ as an aggregate first-order readout, while we further ask which source-target pairs actually carry that aggregate.

4.1 Setup

The positional Jacobian energy is adopted as a guide quantity, which measures how strongly a perturbation at one position propagates to a later final-layer position. For layer ℓ, source position t, and

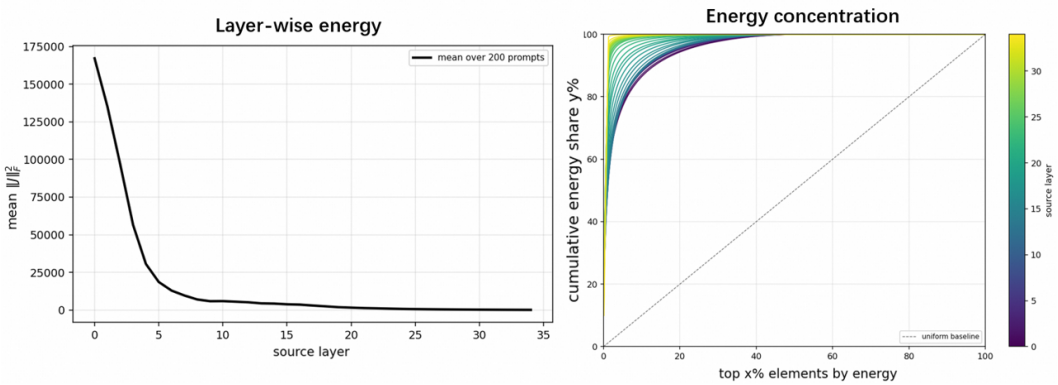

Figure 2: Jacobian energy analysis along the layers. The left figure indicates the relationship between absolute energy and the layers. The right figure indicates the energy concentration across layers, where the top x% of elements accumulate y% of the total energy of the layer.

target position t', define Jacobian energy as:

$$E_\ell(t, t') = \left\| \frac{\partial h_{\text{final},t'}}{\partial h_{\ell,t}} \right\|^2 . \tag{19}$$

Since the J-lens is the average of the Jacobian matrices over all positions within a layer, analyzing the Jacobian energy at each position is important for understanding the representational tendency of the J-lens. It reveals which positions receive more attention from the J-lens, and the content at these positions with higher energy would influence the readout of the J-lens more strongly.

4.2 Layer-wise decay and concentration

As shown in Figure 2, two regularities across the layers appear before structural analysis. First, total energy decays with depth. Early layers sit upstream of many subsequent transformations and therefore have larger influences; late layers are closer to the output and act through shorter residual pathways. Second, energy is heavy-tailed across position pairs. A small top fraction of pairs accounts for most of the mass, and the concentration increases with depth. Thus the averaged object J_ℓ is not an even superposition of weak effects but the domination of a minority of position pairs. This also explains why the J-lens, although constructed from the output expectation over all future positions, can stably read out valid concepts.

4.3 Jacobian Structure: Short Horizons and Sparse Concept

To further investigate the distributional pattern of the highly concentrated Jacobian energy, we conduct visualization studies to analyze the structure of the J-lens and the causal implications of its actual representations. Specifically, we compute the Jacobian energy for every position pair in each layer during the construction of the J-lens and plot a three-dimensional heatmap over source positions, target positions, and layers. Since an overly dense point cloud is not suitable for display, and the concentration effect of energy distribution indicates that a small number of high-energy points are sufficient to reveal the structure, we display the 400 (approximately the top 13%) position pairs with the highest energy in each layer, and hide the remaining points.

Figure 3 demonstrates that high-energy pairs separate into two patterns. The first is the diagonal one: $E_\ell(t, t)$ is large, meaning that a position primarily preserves or prepares the next token. Diagonal energy is especially prominent in later layers, where the remaining computation is short and output-oriented. Since the elements on the diagonal and adjacent positions represent short-term expressions, these causal results are defined as short horizons. The second pattern is horizontal or vertical positions. Some source positions influence many future positions, behaving as broadcast origins; some target positions receive influence from many sources, behaving as integration points. When these positions are stable across prompts and attach to interpretable tokens, we group them as the sparse concept positions.

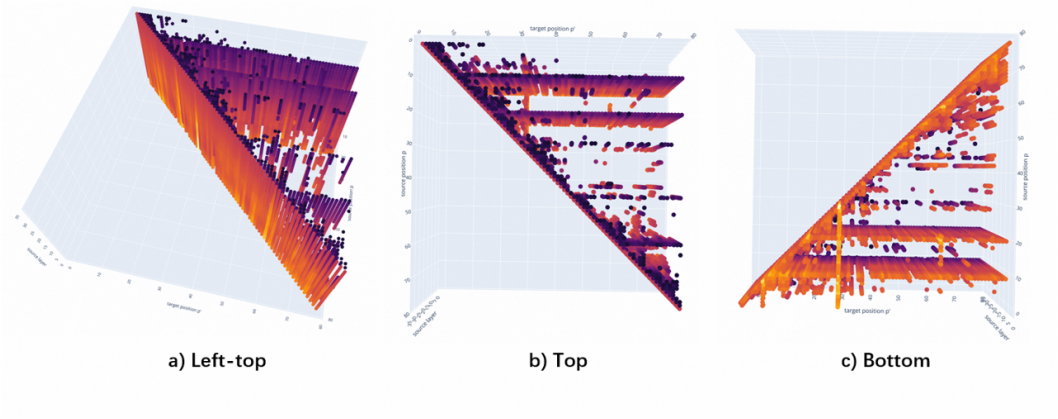

Figure 3: A 3D heatmap for the energy distribution of the Jacobian at each layer and position.

According to the visualization results above, let t^* denote the positions of sparse concept, the formalization of J-lens can be approximated as:

$$J_\ell = \mathbb{E}_{t,t'\geq t,\text{prompt}}\left[\frac{\partial h_{\text{final},t'}}{\partial h_{\ell,t}}\right] \approx \mathbb{E}_{t,t'\approx t,\text{prompt}}\left[\frac{\partial h_{\text{final},t'}}{\partial h_{\ell,t}}\right] + \mathbb{E}_{t,t'=t^*,\text{prompt}}\left[\frac{\partial h_{\text{final},t'}}{\partial h_{\ell,t}}\right]. \quad (20)$$

Moreover, based on the derivation that interprets the J-lens as the aggregate expectation over future outputs in Section 3, we can reduce this future output to two positional modes: **short horizons** and **sparse concepts**. This is precisely the core understanding and interpretation of the meaning of the J-lens readout in this paper.

5 Experiments

Based on the theoretical and experimental study above, we propose improvements to the J-lens to further validate our theory. Since experiments show that the Jacobian energy exhibits a pronounced concentration pattern, and that its concentration in the short horizon and sparse concept modes is precisely the design goal of the J-lens, the Jacobians at the remaining positions act as noise under this view. We therefore propose a simple but effective entry filtering strategy: only the Jacobian matrices at the top-j fraction of position pairs (p, p') with the largest energy are used to construct the J-lens.

5.1 Experiments Settings

We utilized Qwen3-1.7B, Qwen3-4B, Qwen3-8B, Gemma3-1B-it and Gemma3-4B-it (Team, 2025) in the experiments. Following the previous study, Wikitext (Merity et al., 2017) is introduced for J-lens construction, and six tasks including association, multihop reasoning, multilingual, ordering operations, poetry, and typos are evaluated (Gurnee et al., 2026).

5.2 Metrics

In this paper, we design two metrics named short-horizon layers (SHL) and intermediate-concept recall rate (ICR) to evaluate the capability of J-lens on next-token prediction and intermediate concept recalling individually. Both metrics operate on the full position-layer lens readout: for a prompt of T tokens, the lens yields a vocabulary distribution $\mathbf{y}_t^{(l)}$ at every layer $l \in \{1,\ldots,L-1\}$ and position $t \in \{1,\ldots,T\}$.

Short-horizon layers. For each position t with true next token x_{t+1}, we count the length of the trailing run of layers whose lens top-1 prediction matches x_{t+1}:

$$SHL_t = \max\{r : \arg\max_v q_{t,v}^{(l)} = x_{t+1}\ \forall\, l \in \{L-r+1,\ldots,L\}\}. \quad (21)$$

Table 1: Main results across several tasks. **Bold** denotes the best. Vanilla means the original J-lens. All J-lenses are trained on the same 200 examples with the same configuration.

	association		multihop		multilingual		order-ops		poetry		typo		avg	
Method	SHL	ICR	SHL	ICR	SHL	ICR	SHL	ICR	SHL	ICR	SHL	ICR	SHL	ICR
Qwen3-8B														
Vanilla	1.01	4.70	1.65	16.58	0.66	40.07	0.66	12.82	0.99	19.54	0.62	14.40	0.94	18.71
$\text{Filter}_{\text{topj}=0.5}$	1.04	4.93	1.71	15.58	0.68	38.53	0.68	12.56	1.00	20.17	0.67	15.69	0.98	18.60
$\text{Filter}_{\text{topj}=0.2}$	1.05	6.36	**1.73**	16.35	**0.69**	**48.40**	0.67	17.36	**1.06**	21.13	0.69	19.03	1.00	**22.14**
$\text{Filter}_{\text{topj}=0.1}$	1.06	**6.44**	1.72	15.78	0.68	47.38	0.68	**18.26**	1.04	**21.17**	**0.72**	**19.25**	1.00	22.00
$\text{Filter}_{\text{topj}=0.01}$	1.05	4.28	1.69	13.26	0.67	31.56	0.69	11.82	1.03	18.96	0.69	15.73	0.98	16.45
Qwen3-4B														
Vanilla	1.06	4.84	1.70	15.71	**0.67**	40.56	0.69	16.78	1.16	18.38	0.70	17.45	1.01	19.41
$\text{Filter}_{\text{topj}=0.2}$	**1.15**	**9.31**	**1.84**	**18.41**	0.64	**64.89**	**0.75**	**20.82**	**1.23**	**22.88**	**0.80**	**24.99**	**1.08**	**27.93**
Qwen3-1.7B														
Vanilla	0.87	4.45	1.44	12.52	0.64	33.08	1.34	28.65	1.08	14.51	0.60	15.99	0.96	17.58
$\text{Filter}_{\text{topj}=0.2}$	**0.95**	**6.29**	**1.59**	**14.15**	**0.67**	**38.80**	**1.39**	**32.04**	**1.15**	**17.32**	**0.71**	**21.06**	**1.04**	**21.04**
Gemma3-4B-it														
Vanilla	1.52	**9.54**	2.15	28.13	0.72	45.57	1.74	40.93	1.38	30.91	1.09	29.71	1.39	30.12
$\text{Filter}_{\text{topj}=0.5}$	**1.58**	9.39	**2.27**	**28.18**	**0.73**	**45.79**	**1.83**	**41.49**	**1.39**	**31.70**	**1.13**	**30.78**	**1.44**	**30.53**
Gemma3-1B-it														
Vanilla	1.39	4.64	1.79	**20.89**	0.62	**36.72**	2.87	83.39	1.65	19.69	0.90	18.93	1.42	26.63
$\text{Filter}_{\text{topj}=0.5}$	**1.48**	**5.78**	**1.97**	20.09	**0.66**	35.91	**3.08**	**84.11**	**1.70**	**21.16**	**0.96**	**19.25**	**1.51**	**26.94**

Averaging SHL_t over positions measures how early the lens readout converges to the model's next token prediction, demonstrating the short-horizon ability of a J-lens.

Intermediate-concept recall rate. For prompts annotated with intermediate concepts $\mathcal{I} = \{w_1, \ldots, w_m\}$, we count grid cells whose lens top-K ($K = 10$) contains any token of concept:

$$ICR = \left|\left\{(l,t) : \mathcal{I} \cap \text{top-}K\left(\mathbf{q}_t^{(l)}\right) \neq \varnothing\right\}\right|, \tag{22}$$

which measures how broadly the J-lens reads the required sparse concept.

5.3 Results

Results in Table 1 presents the results for methods on several benchmarks in this paper. We can conclude the following finding. First, since the filtering methods with topj further strengthens the concentration of the Jacobian energy in the J-lens and enables its representational capacity to exceed that of the unfiltered baseline, J-lens can focus more on the highlight sparse concept and outperform the vanilla on most tasks. Second, models differ in their sensitivity to the degree of position filtering. Counterintuitively, this sensitivity correlates weakly with model size but strongly with model family. Specifically, the optimal filtering magnitude is approximately 0.2 for Qwen3 and 0.5 for Gemma3. Model scale exerts little influence, and parameter count and layer number likewise have no direct effect on the filtering value.

6 Conclusion

We interpret the J-lens as an averaged first-order transfer operator from intermediate activations to expected future readouts: the Jacobian matrix is the locally optimal linearization, least squares gives the globally optimal affine readout, and the Stein bridge makes them coincide under Gaussian inputs. Nonlinear curvature and non-Gaussian score residuals explain systematic bias and provide an explanation for the early failure of the J-lens. We then further analyze the structure of the Jacobian energy. The Jacobian energy decays with depth, concentrates in a small number of source-target pairs, and splits into diagonal short horizons and sparse concept positions. Finally, we introduce improvements methods for the J-lens. Energy filtering preserves or improves next-token consistency and the recall rate of intermediate concepts, thereby validating our interpretation of the meaning of the J-lens readout. Finally, we acknowledge that this work still has limitations and is still in progress, which does not represent the final results.

AI USE STATEMENT

In this work, we used generative AI tools for provide critical ingredients for proving mathematical claims, assist in the writing of proofs, and assist with translation. We have not used generative AI tools for help develop theoretical models or conceptual frameworks, Generate synthetic data sets, formulate mathematical claims, propose or refine hypotheses, design or provide feedback on research methodology or experiments, implement methods, clean and reformat dataset, support qualitative and thematic data analysis, and interpret results. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Guillaume Alain and Yoshua Bengio. Understanding intermediate layers using linear classifier probes. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Workshop Track Proceedings*. OpenReview.net, 2017. URL https://openreview.net/forum?id=HJ4-rAVtl.

Nora Belrose, Zach Furman, Logan Smith, Danny Halawi, Igor Ostrovsky, Lev McKinney, Stella Biderman, and Jacob Steinhardt. Eliciting latent predictions from transformers with the tuned lens. *CoRR*, abs/2303.08112, 2023. doi: 10.48550/ARXIV.2303.08112. URL https://doi.org/10.48550/arXiv.2303.08112.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin (eds.), *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/1457c0d6bfcb4967418bfb8ac142f64a-Abstract.html.

Mor Geva, Avi Caciularu, Kevin Ro Wang, and Yoav Goldberg. Transformer feed-forward layers build predictions by promoting concepts in the vocabulary space. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (eds.), *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pp. 30–45. Association for Computational Linguistics, 2022. doi: 10.18653/V1/2022.EMNLP-MAIN.3. URL https://doi.org/10.18653/v1/2022.emnlp-main.3.

Wes Gurnee, Nicholas Sofroniew, Adam Pearce, Mateusz Piotrowski, Isaac Kauvar, Runjin Chen, Anna Soligo, Paul C. Bogdan, Euan Ong, Rowan Wang, Ben Thompson, David Abrahams, Subhash Kantamneni, Emmanuel Ameisen, Joshua Batson, and Jack Lindsey. Verbalizable representations form a global workspace in language models. *CoRR*, abs/2607.15495, 2026. doi: 10.48550/ARXIV.2607.15495. URL https://doi.org/10.48550/arXiv.2607.15495.

John Hewitt and Percy Liang. Designing and interpreting probes with control tasks. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (eds.), *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pp. 2733–2743. Association for Computational Linguistics, 2019. doi: 10.18653/V1/D19-1275. URL https://doi.org/10.18653/v1/D19-1275.

Aapo Hyvärinen and Peter Dayan. Estimation of non-normalized statistical models by score matching. *Journal of Machine Learning Research*, 6(4):695–709, 2005.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. Pointer sentinel mixture models. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017. URL https://openreview.net/forum?id=Byj72udxe.

nostalgebraist. Interpreting gpt: The logit lens. LessWrong, 2020. https://www.lesswrong.com/posts/AcKRB8wDpdaN6v6ru/interpreting-gpt-the-logit-lens.

Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Scott Johnston, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Chris Olah. In-context learning and induction heads. *CoRR*, abs/2209.11895, 2022. doi: 10.48550/ARXIV.2209.11895. URL https://doi.org/10.48550/arXiv.2209.11895.

Lee Sharkey, Bilal Chughtai, Joshua Batson, Jack Lindsey, Jeffrey Wu, Lucius Bushnaq, Nicholas Goldowsky-Dill, Stefan Heimersheim, Alejandro Ortega, Joseph Isaac Bloom, Stella Biderman, Adrià Garriga-Alonso, Arthur Conmy, Neel Nanda, Jessica Rumbelow, Martin Wattenberg, Nandi Schoots, Joseph Miller, William Saunders, Eric J. Michaud, Stephen Casper, Max Tegmark, David Bau, Eric Todd, Atticus Geiger, Mor Geva, Jesse Hoogland, Daniel Murfet, and Tom McGrath. Open problems in mechanistic interpretability. *Trans. Mach. Learn. Res.*, 2025, 2025. URL https://openreview.net/forum?id=91H76m9Z94.

Charles M. Stein. Estimation of the mean of a multivariate normal distribution. *The Annals of Statistics*, 9(6):1135–1151, 1981.

Qwen Team. Qwen3 technical report. *CoRR*, abs/2505.09388, 2025. doi: 10.48550/ARXIV.2505.09388. URL https://doi.org/10.48550/arXiv.2505.09388.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. Emergent abilities of large language models. *Trans. Mach. Learn. Res.*, 2022, 2022. URL https://openreview.net/forum?id=yzkSU5zdwD.

A Datasets and Training Configuration

A.1 Training Corpus

The lens is fit on natural text from the WikiText-103 training split. The reference configuration uses $N = 1000$ sequences of 128 tokens; our runs use $N = 200$ sequences truncated to 96 tokens. The lens quality saturates quickly with corpus size, so a few hundred sequences are already sufficient. The parameters are also recommended in the original studies.

A.2 Fitting Parameters

We fit each lens on $N = 200$ training prompts (the reference configuration uses 1000), truncating every prompt to 96 tokens during fitting (reference: 128). The first 16 leading positions of each prompt are skipped, since they behave as attention sinks and carry atypical residual statistics. Every layer below the final layer is fit against that final layer, and the Jacobian is estimated in chunks of 96 output dimensions per backward pass. The resulting per-layer matrices are stored in half precision (fp16).

B Models

Table 2 lists the decoder-only transformers used in the experiments and their two lens-relevant architecture parameters. Fitting settings are shared across all models and follow Appendix A.2.

C Prompt Formats

C.1 Training Prompts

There is no single fixed training prompt. The training set is a collection of distinct, heterogeneous natural-language passages (the first N qualifying WikiText-103 records), each truncated to the fitting

Model	#Layers	d_{model}
Qwen3-1.7B	28	2048
Qwen3-4B	36	2560
Qwen3-8B	36	4096
Gemma3-1B	26	1152
Gemma3-4B	34	2560

Table 2: Models and their lens-relevant architecture parameters. The lens is fit on every layer below the final layer and stored per layer.

Subset	Example prompt	Intermediates	Readout position
association	*She kept his coffee mug on the counter, washed it every morning and put it back, though no one had used it in eleven months.*	grief	closing period
multihop	*Fact: The ocean on the coast of the country where Carnival is most famously celebrated is the*	Brazil	token before target
multilingual	*Lo opuesto de "grande" es "*	Spanish, opposite, big, small	token before target
order-ops	*(2 + 3) * 4 =*	5, multiplication	token before target
poetry	*A rhyming couplet:* *The soldier held his final, rattling breath,* *And closed his eyes to greet the face of*	death	end of line 1
typo	*The course required that every student learn a second langauge*	language	final token

Table 3: One representative prompt per evaluation subset, with its intermediate target tokens and the fixed lens readout position.

length. Within a model the prompts differ from one another; across models the same deterministic set of passages is reused. The lens is the Jacobian averaged over this diverse corpus, which is what makes the readout direction corpus-level rather than tied to any one prompt.

C.2 Evaluation Prompt Examples

Lens quality is measured on six evaluation subsets. Each item provides a prompt and a set of *intermediates* (the latent tokens that should be recoverable from the lens); the readout position is fixed per subset. Table 3 gives one representative example per subset.

D Hyper-parameter Sensitivity Analysis

We study the sensitivity of the lens to the *top-j* position-filtering ratio: when fitting the Jacobian, only the top-j fraction of source positions (ranked by per-pair mean influence) contribute to the per-prompt average. We sweep $j \in \{0.1, 0.2, \ldots, 0.9\}$ and report two metrics, each a weighted average over the six evaluation subsets (weighted by the number of items): SHL and ICR.

Figure 4 shows both metrics as a function of top-j for all five models. The two families respond differently: for the Qwen3 models SHL decreases monotonically toward top-$j = 1.0$ while ICR peaks at small top-j (0.1–0.2) and then declines; for the Gemma3 models SHL again falls with top-j, but ICR does not sensitive with higher top-j, and peaks around $0.4 - 0.7$. Overall the lens is only mildly sensitive to top-j in SHL, whereas ICR is the more top-j-sensitive metric and its optimum is model-family-dependent.

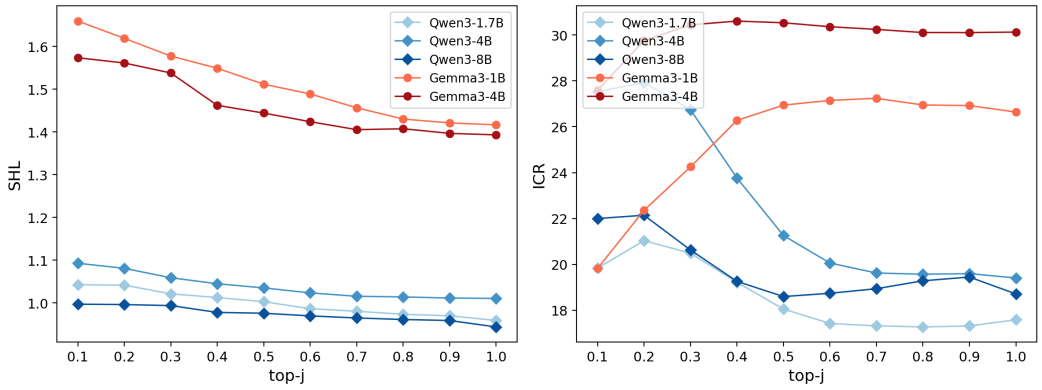

Figure 4: Sensitivity of lens quality to the top-j position-filtering ratio. Left: SHL; right: ICR. Each curve is one model; top-$j = 1.0$ is the base method that keeps every position. Both metrics are weighted averages over the six evaluation subsets.