Lexical Reference Frames

Interpreting LM Residual Streams in Native Semantics

Paper: Jian Gu, Hongyu Zhang, Chunyang Chen, Aldeida Aleti

arXiv:2610.22169v2 | Research discussion, 12 November 2026

Read a hidden state through the layers that will use it.

Publication not found as of 2026-11-12 | Official public code

Paper v2, revised 18 October 2026; presentation is an independent reading.

A fixed output direction can misread an intermediate state

What should "this state means token v" mean?

How the remaining model uses the state, given its prefix.

- Logit lens: reuse the output frame.
- Jacobian lens: one average map per layer.
- Point-cloud geometry: describe where states vary.

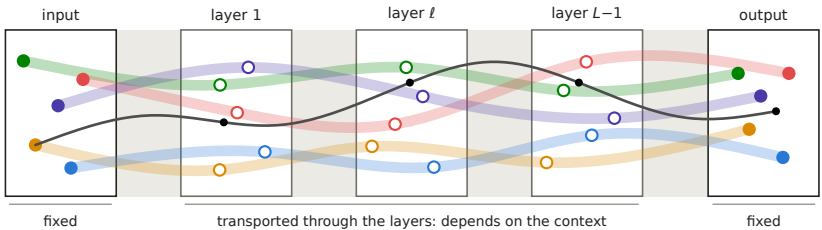
OLMo3, layer 16 of 32:

Direction	Angle to native readout
Logit lens	75°
Jacobian lens	77°
LexFrame[1]	13°

The same token has different local readout directions in different contexts.

Paper, Sec. 1, Fig. 1, PDF p. 2. Angles averaged over states and labels.

The vocabulary stays fixed; its local frame moves



- Input embedding and output unembedding fix the two interfaces.
- Between them, the lexical frame depends on the remaining layers and context.
- A frame supplies coefficients to read, and basis vectors to write.

Paper, Fig. 2, PDF p. 4. Author schematic: coloured lexical bases, black hidden state.

Native transport defines a state-specific lexical reading

$$x_\ell(t) \xrightarrow[\text{prefix at layer } \ell \text{ given}]{T_\ell\text{: remaining model}} x_L(t) \xrightarrow{N,\ W} z_\ell$$

$$z_\ell(v) = w_v^\top N(T_\ell(x_\ell(t))), \qquad c_v = z_v - \frac{1}{|V|}\sum_u z_u$$

- $x_\ell(t) \in \mathbb{R}^d$: residual state; N: final normalization.
- $W \in \mathbb{R}^{|V|\times d}$: unembedding, with row $w_v^\top$.
- Native transport evolves the prefix normally after the cut.
- The complete native reading equals the model output at every cut.

Paper, Secs. 3.1–3.2, PDF pp. 3–4, Eq. (2).

Readout, basis and coefficient answer different questions

$$J_\ell = \frac{\partial N(T_\ell)}{\partial x_\ell(t)}, \qquad s_v = W^+ e_v, \qquad \bar{w} = \frac{1}{|V|}\sum_v w_v$$

Readout:	$b_\ell(v) = J_\ell^\top (w_v - \bar{w})$	local sensitivity
Basis:	$s_\ell(v) = J_\ell^+ s_v$	selective output move
Coefficient:	$z_\ell(v) = w_v^\top N(T_\ell(x_\ell(t)))$	transported value

Why gradient and basis differ

The steepest increase can also raise correlated tokens. A pseudoinverse instead tries to realize a specified output change with a small state move.

Paper, Sec. 3.2, PDF p. 4, Eqs. (1)–(2); App. C.1, PDF pp. 16–17.

Context variation bounds every fixed readout direction

For one label, let $\hat{b}_i$ be native unit readouts across m states.

$$\frac{1}{m}\sum_i r^\top \hat{b}_i = r^\top \bar{b} \le \|\bar{b}\|, \qquad \|\bar{b}\|^2 = \frac{1}{m} + \left(1 - \frac{1}{m}\right)\hat{\kappa}.$$

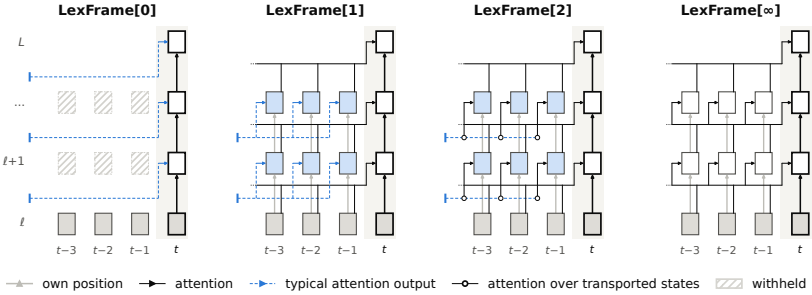
- r is any fixed unit direction; $\hat{\kappa}$ is mean pairwise cosine.
- With many states, the bound approaches $\sqrt{\kappa_\ell(v)}$.
- At mid-depth, the label-averaged bound is 0.30 / 0.35.

Scope of the result

This bounds mean directional cosine, not output KL. A lens with state-dependent final normalization needs that extra qualification; the paper reports it below the bound empirically.

Paper, Sec. 3.2, PDF p. 4; App. C.2, Eq. (5), PDF p. 17. Model order: SmolLM3 / OLMo3.

Order controls cross-position information paths



- Order 0: state-only transport, using typical attention writes.
- Order 1: target attends to a prefix evolved at order 0.
- Order 2: prefix evolves at order 1; one more interaction is retained.

Order counts nested cross-position exchanges, not attention layers.

Paper, Sec. 3.3, Fig. 3, PDF p. 5. Order ∞ recovers native transport.

Typical attention retains input-dependent average behavior

$$\hat{a}_k(y) = \bar{a}_k + G_k(y - \bar{y}_k)$$

- y: attention input at the current position.
- $\bar{y}_k$, $\bar{a}_k$: calibration means of input and attention write.
- G_k: layerwise ridge regression, fitted in closed form on unlabelled text.
- The actual residual blocks, normalizations and MLPs remain in the transport.

What is fitted?

Attention outputs from their inputs. Neither output logits nor answer labels are training targets. The model weights stay fixed.

Paper, Eq. (3), PDF p. 5; App. A.1, PDF p. 13. Ridge selected by two-fold cross-validation.

Finite orders become exact near the output

$$h_k^n(t) = F_k\big(h_{k-1}^n(t), \mathrm{Att}_k\big(h_{k-1}^n(t); \{h_{k-1}^{n-1}(s)\}_{s<t}\big)\big)$$

- Every stream starts from $h_\ell^n = x_\ell$; order 0 is the base case.
- Induction: order n agrees with the model for its first n remaining layers.
- Thus $L - \ell \leq n$ implies exact reading, readouts and basis.

Where approximation is informative

The complete native reading repeats the final output. Finite orders reveal what survives when selected future context interactions are withheld.

Paper, App. C.2, Eq. (6), PDF p. 17. Reading order n uses $n + 1$ streams.

Evaluation separates distributions, directions and interventions

Models and data

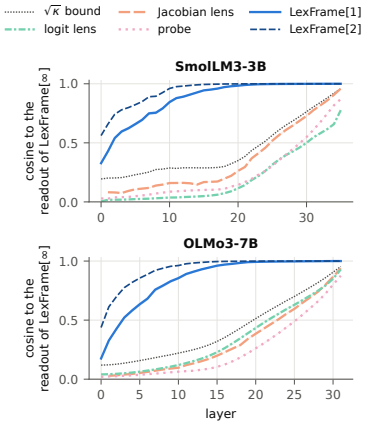
- SmolLM3-3B: 36 layers, tied embeddings.
- OLMo3-7B: 32 layers, untied embeddings.
- WikiText-103; 128-token sequences.
- Calibration: 256 training sequences.
- Reading test: 64 test sequences.

Comparisons and metrics

- Logit lens; Jacobian lens fitted on 100 prompts; direction-only ridge probe.
- KL(model $\|$ reading), in nats.
- Top-1 agreement with model.
- ECE against actual next tokens.
- Native gradients and ablations as intervention references.

Paper, Sec. 4, PDF pp. 5–7; App. A.1, pp. 12–13. Means over layers 1 to $L-1$. Tuned lens omitted.

One interaction recovers native directions from mid-depth



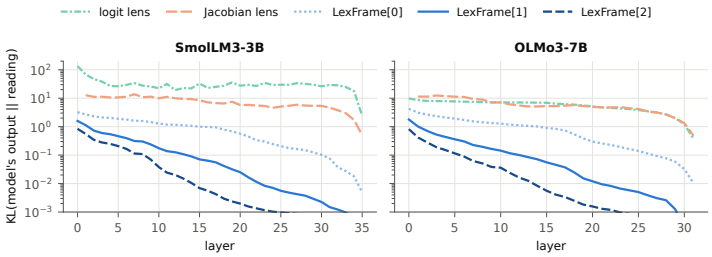
At layers 18/36 and 16/32

Cosine	Smol	OLMo
Fixed-direction bound	0.30	0.35
LexFrame[1]	0.97	0.97

- Blue solid: order 1; dashed: order 2.
- Grey dotted: coherence bound.
- Lowest layers retain a larger approximation gap.

Paper, Fig. 7, PDF p. 14; Sec. 4.1, p. 6. Direction comparison over 64 frequency-ranked labels.

Most distribution error disappears at order 1



Layer-mean KL ↓	Logit	Jacobian	Order 0	Order 1	Order 2
SmolLM3	29.06	7.51	0.87	0.16	0.06
OLMo3	5.66	6.19	0.93	0.15	0.05

Paper, Fig. 4, PDF p. 6; Table 1, p. 7. Vertical axis is logarithmic; KL in nats.

Faithful readings also have useful probabilities

	Top-1 agreement ↑		ECE ↓	
Method	Smol	OLMo	Smol	OLMo
Logit lens	0.04	0.15	0.632	0.107
Jacobian lens	0.14	0.13	0.376	0.191
LexFrame[0]	0.64	0.64	0.018	0.045
LexFrame[1]	0.86	0.86	0.027	0.018
LexFrame[2]	0.92	0.93	0.020	0.020
Full model	—	—	0.018	0.023

- Top-1 compares to the model; ECE compares to actual next tokens.
- At the embedding, order-1 agreement is only 0.42 / 0.39.
- Closer model agreement need not monotonically improve ECE.

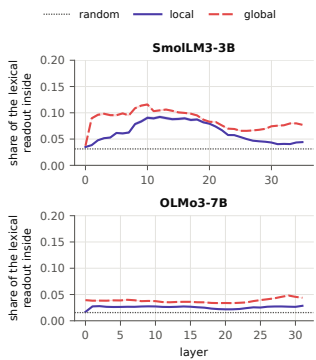
Paper, Table 1 and Sec. 4.1, PDF pp. 6–7. ECE uses 15 bins.

A lexical distance predicts substitution effects better

Replace one real hidden state with another; finish the model.

Layer-mean rank correlation with the output effect:

Distance	Smol	OLMo
Cosine	0.27	0.28
Euclidean	0.08	0.12
Geodesic	0.09	0.07
Jacobian-lens reading	0.28	0.24
LexFrame[1] reading KL	0.94	0.95



Paper, Table 2 and Fig. 5, PDF p. 7. Plot: squared readout share in 64 local/global principal directions; grey: random. 7,961 replacement pairs/layer.

The basis writes a token more selectively than its gradient

Direction	Smol	OLMo
Fixed output basis	0.45	0.80
Order-1 readout	0.54	0.73
Order-1 basis	0.84	0.97
Native readout	0.56	0.79
Native basis	0.98	1.00

- Move by $0.1\|x\|$; ask if the intended token gains *most*.
- A readout raises correlated tokens; an inverse transport targets a lexical output move.
- Coefficients still require nonlinear transport: $b_\ell(v)^\top x_\ell$ is a slope-related pairing, not the token's logit.

Paper, Table 3, Sec. 4.3, PDF p. 8. 64 states and 64 labels/layer; explicit FP32 Jacobian.

Basis steering reduces changes to other tokens

Task: move the runner-up token above the current top token.

$$\delta_b \propto b_\ell(u) - b_\ell(v), \qquad \delta_s \propto s_\ell(u) - s_\ell(v)$$

Model	Direction	Reach ↑	Side ↓	Clean ↑
Smol	Order-1 readout	0.92	0.23	0.61
	Order-1 basis	0.89	0.12	0.71
	Native basis	0.90	0.11	0.73
OLMo	Order-1 readout	0.94	0.20	0.64
	Order-1 basis	0.95	0.07	0.82
	Native basis	0.97	0.06	0.85

Paper, Table 5, PDF p. 9. Steps: 0.01–1 $\times \|x\|$. Side: other-token change at first flip, nats. Clean: fraction of all predictions flipped with Side < 0.1. Effects evaluated in the complete model.

A finite difference can outperform the exact local gradient

Remove a component write q at one position; keep its prefix.

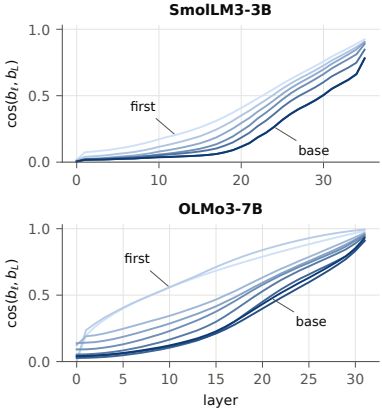
$$A^{(n)} = c_v^{(n)}(x) - c_v^{(n)}(x - q), \qquad A^{(n)}_{\text{linear}} = b^{(n)}(v)^\top q$$

Model	Attribution	Layer	Attention	MLP
Smol	Order-1 difference	0.88	0.79	0.88
	Native linear (AtP)	0.62	0.75	0.63
OLMo	Order-1 difference	0.90	0.85	0.90
	Native linear (AtP)	0.68	0.84	0.71

Approximate transport retains nonlinear finite effects, especially for MLP writes.

Paper, Table 6 and Sec. 4.4, PDF p. 9. Correlation with actual centred-logit ablation effects; 64 sequences. Equations restate the intervention in compact notation.

The output frame becomes less adequate during training



Mean KL: first trained → final

	Logit lens	Order 1
Smol	4.82 → 29.06	0.15 → 0.16
OLMo	2.32 → 5.66	0.02 → 0.15

- Darker curves: later checkpoints.
- The plot compares native readouts at each layer to the final layer.
- Every checkpoint is calibrated anew.

Paper, Fig. 6 and App. B, PDF pp. 13–14. Checkpoint-wise calibration; Smol's earliest checkpoint is already trained.

Predictions settle early, but this reader does not save compute

	Settled layer share		Threshold-skipped share	
Method	Smol	OLMo	Smol	OLMo
Logit lens	0.03	0.11	0.00	0.04
Jacobian lens	0.09	0.09	0.00	0.03
Order 0	0.58	0.58	0.23	0.20
Order 1	0.82	0.81	0.45	0.42
Order 2	0.89	0.90	0.53	0.54

- Settled: uninterrupted agreement back from the last layer (retrospective).
- Exit threshold selected on half the sequences for 0.95 model agreement.

LexFrame still runs the remaining layers: skipped depth is not speedup.

Paper, Table 7, PDF p. 16; explicit computational qualification in App. B, p. 15.

Faithfulness has a concrete computational cost

- Calibration: one forward over calibration text plus one ridge solve per layer.
- Reading all layers of a sequence at order n costs approximately

$$\frac{(n+1)(L+1)}{2} \quad \text{full-forward equivalents.}$$

- Jacobian-lens fitting costs about Pd backward passes; its later reading is cheap.
- Basis construction needs regularized inverse-Jacobian solves; steering uses 30 LSMR steps.

Practical role

A faithful offline reference and analysis tool. A cheaper adaptive reader would be needed to turn early readable predictions into efficient inference.

Paper, App. A.1, PDF p. 13; App. C.3, p. 17. Measured Jacobian fit cost: 18 / 44 times calibration plus test readings at orders 0–2.

Three results change how hidden states should be interpreted

1. **Meaning is downstream and contextual.**
   A native frame explains why one direction per layer misses intermediate readouts.

2. **Values, gradients and inverse directions differ.**
   Use transported coefficients to read, differences to attribute, and basis vectors to steer selectively.

3. **Order makes future context interaction controllable.**
   One interaction is already accurate from mid-depth in these models; it remains expensive.

Paper, Sec. 5, PDF p. 9; Apps. C–D, pp. 16–18.

Evidence scope and sources

- Tested: two model families on WikiText-103; context extension to 1,024 tokens; runner-up flips and single-position write ablations.
- Tuned lens is omitted, so no direct distribution comparison is established.
- Hallucination detection, knowledge editing and distillation are proposed uses, without downstream evaluations here.

Primary sources

- Gu, Zhang, Chen, Aleti (2026). *Lexical Reference Frames: Interpreting LM Residual Streams in Native Semantics*.
  arXiv:2610.22169v2 Official implementation
- Publication not found as of 2026-11-12; arXiv and author citation list a preprint.

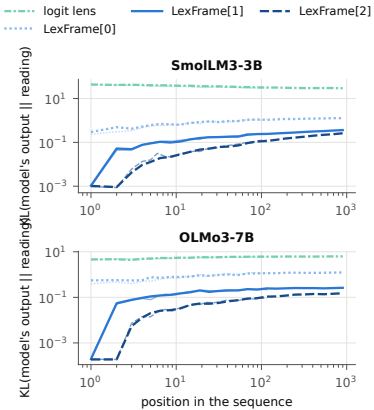
Backup: regression matters more than a constant write

Model	Typical attention	KL ↓	Cosine ↑	Attribution ↑
Smol	Regression	0.16	0.89	0.88
	Mean only	1.26	0.73	0.67
	Zero	0.76	0.71	0.68
OLMo	Regression	0.15	0.87	0.90
	Mean only	0.52	0.74	0.77
	Zero	0.47	0.74	0.77

- All rows evaluate order 1; the substitution changes its order-0 prefix stream.
- Three disjoint calibration texts: KL 0.160–0.163 / 0.146–0.151.
- Cosine here is layer-averaged, unlike the mid-depth value of 0.97.

Paper, Table 8, PDF p. 16; App. B. Point estimates in this ablation.

Backup: longer contexts degrade the finite orders gradually



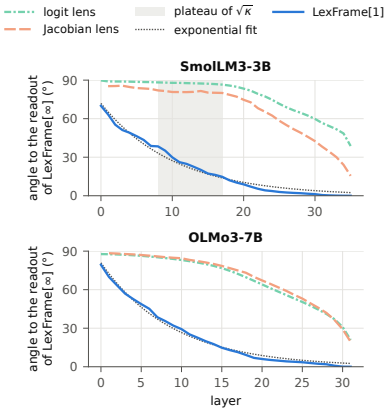
Paper, Fig. 9 and App. B, PDF p. 15.

128 → 1,024 tokens: mean KL

	Order 1	Order 2
Smol	0.16 → 0.25	0.06 → 0.15
OLMo	0.15 → 0.18	0.05 → 0.08

- Thick curves: 1,024; thin: 128 tokens.
- Both plot axes are logarithmic.
- 64 test sequences; calibration token budget held fixed.
- Order 1 still matches the model's top token at most states.

Backup: the direction gap fits an exponential across depth



Paper, Fig. 8 and App. B, PDF pp. 14–15.

- Mean angle shrinks by 9.2% / 10.5% of the remaining angle per layer.
- Exponential fit: $R^2 = 0.98$ / 0.99.
- Smol coherence plateau: layers 8–17, $\sqrt{\kappa} \approx 0.29$.
- Plateau absent at 2.5T training tokens, present from 8.1T; absent in all 9 OLMo checkpoints.

Author interpretation: overdamped relaxation. The empirical fit does not establish a physical dynamics model.

Backup: redundant coordinates remain local after transport

$$s_v = W^+ e_v, \qquad \sum_v s_v w_v^\top = W^+ W = I \quad \text{if } W \text{ has full column rank.}$$

$$\sum_v s_\ell(v)(J_\ell^\top w_v)^\top = J_\ell^+ J_\ell$$

- The vocabulary is larger than state dimension: a redundant frame.
- The right side is a projection onto output-effective state directions (ignoring ridge).
- A readout paired with the state measures a rescaling response:

$$b_\ell(v)^\top x_\ell = \frac{d}{d\alpha} c_v(\alpha x_\ell)\Big|_{\alpha=1}.$$

- Normalization breaks degree-one homogeneity: coefficient $\neq$ gradient pairing.

Paper, App. C.1, Eq. (4), PDF p. 17. Dual relation uses uncentred rows; centring and ridge add qualifications.

Backup: both candidates appear before the final choice

OLMo3 prefix: "A review of Boulter 's performance in The"

Layer	Logit lens	Jacobian lens	Order 1
8	nto	urga	Bill
12	comparison	newline fragments	Times
16	orem	Illustrated	Guardian
20	odor	Guardian	Guardian
24	Guardian	Guardian	Guardian
32	Guardian	Guardian	Guardian

- Final model: Guardian 0.29; Times 0.22.
- First among listed layers: order 1 at 16, Jacobian at 20, logit at 24.

Paper, Table 9 and worked example, PDF p. 16. Selected rows transcribed; newline token displayed descriptively.

Backup: numerical and statistical protocols

- Means exclude embedding and final layer: layers $1, \ldots, L-1$ equally weighted.
- Fixed seeds and sampled positions; one run, deterministic up to rounding.
- 95% percentile bootstrap over sequences, at least 300 resamples.
- Reported maximum table CI half-widths: 12% of KL, 0.03 nats for steering Side, 0.04 otherwise; fitted-term ablation uses point estimates.
- Released half precision for most gradients; FP32 explicit Jacobian in the composition test.

$$\min_{\delta} \|J_\ell \delta - s_v\|^2 + \lambda \|\delta\|^2, \qquad \lambda = 10^{-3} \frac{\|J_\ell^\top s_v\|^2}{\|s_v\|^2}.$$

Paper, App. A.1, PDF pp. 12–13. Basis solved in closed form for composition, with 30 LSMR iterations for steering.