Lexical Reference Frames

Interpreting LM Residual Streams in Native Semantics

Jian Gu Hongyu Zhang Chunyang Chen Aldeida Aleti

arXiv:2610.22169v2 | 10 November 2026

A model-native, context-aware reading of every residual-stream state

Publication: unverified as of 2026-11-10 arXiv v2 Official code

The middle of a language model has no fixed vocabulary frame

Fixed lens	Point cloud
One direction per token and layer. The same map reads every context.	Describes which states are near or far. Does not tell how later layers use them.

- Logit lens reuses the output frame.
- Jacobian lens averages over prompts.

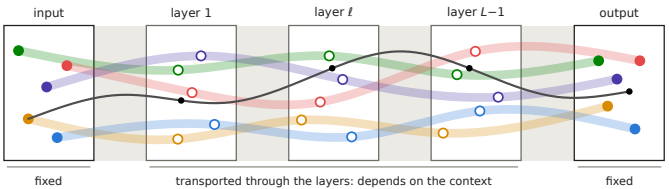
- Distances need not predict intervention effects.
- A hidden state needs a model-defined reading.

Question

What vocabulary directions do the remaining layers use to read this state in this context?

Paper, Sec. 1, PDF pp. 1–2.

A state's native frame is pulled back through the remaining layers



- Input and output frames are fixed by the model's vocabulary.
- In between, later layers transport the state and attend to its prefix.
- Therefore the relevant token directions depend on layer and context.

Paper, Fig. 2, PDF p. 4. Native transport keeps the prefix fixed while carrying the target state forward.

A frame has a reading side and a writing side

$$b_\ell(v) = \frac{\partial c_v}{\partial x_\ell(t)} = J_\ell^\top (w_v - \bar{w}) \qquad s_\ell(v) = J_\ell^+ s_v$$

$J_\ell = \partial N(T_\ell)/\partial x_\ell(t)$; c_v is the centred output logit.

Lexical readout	Lexical basis
Local gradient of token v's centred logit. Which small state change raises v fastest?	Smallest local move whose transported effect matches output basis s_v. How can we write v selectively?

$z_\ell(v) = \langle w_v, N(T_\ell(x_\ell(t))) \rangle$: transported coefficient used to read the state.

Paper, Sec. 3.2, Eqs. (1)–(2), PDF p. 4; basis construction in Appendix C.1, p. 17.

Context variation gives every fixed lens a direction ceiling

$$\frac{1}{m}\sum_i r^\top \hat{b}_i = r^\top \bar{b} \le \|\bar{b}\| \longrightarrow \sqrt{\kappa_\ell(v)}$$

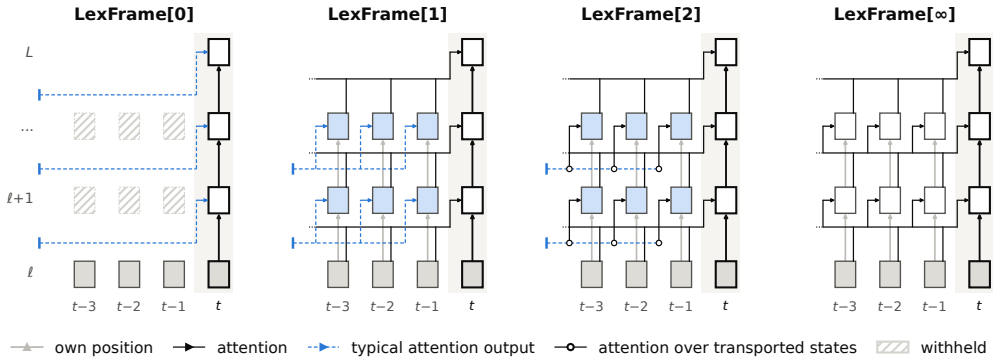
- $\hat{b}_i$: unit native readouts of one token across contexts.
- $\kappa_\ell(v)$: mean cosine of those readouts across sequences.
- A state-independent direction cannot follow context-dependent rotation.

Mid-depth result

Fixed-direction bound: 0.30 (SmolLM3) and 0.35 (OLMo3).
LexFrame[1] versus native readouts: cosine 0.97 in both models.

Paper, Sec. 3.2 and Sec. 4.1, PDF pp. 4, 6; proof in Appendix C.2, p. 17. The bound is for direction similarity, not reading KL.

LexFrame order controls how much prefix attention is restored



- Order 0: only the target stream; later attention is replaced by a typical output.
- Order 1: target attends to order-0 prefix states.
- Order 2: prefix states exchange one more round; LexFrame[∞] is the full model.

Paper, Fig. 3, PDF p. 5. Higher order means one more round of information exchange between positions.

Order 0 keeps the model's layers and estimates withheld attention

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

- G_k is a per-layer ridge regression from attention input to attention output.
- Fit once on unlabelled calibration text; no target logits are used.
- LexFrame[1] restores the target's real attention to the prefix, while prefix states remain order 0.
- The next order measures the contribution of one additional prefix exchange.

Reading the design

Every frame runs the remaining layers. The chosen order controls only how far context information travels between positions.

Paper, Sec. 3.3, Eq. (3), PDF p. 5; Appendix A.1, p. 13.

Higher order approaches the model, with a measurable compute cost

- LexFrame[n] is exact when at most n layers remain.
- LexFrame[∞] returns the model's output at every layer.
- Reading all layers at order n costs about $(n+1)(L+1)/2$ forward-pass units.

Calibration	one forward pass over unlabelled text
Reading	$n+1$ streams through the remaining layers

Trade-off

More attention rounds improve fidelity; one fixed matrix multiply is cheaper at read time, but a Jacobian lens also has a costly fit.

Paper, Appendix C.2–C.3, PDF p. 17. Costs are counted in forward passes, independent of hardware.

Evaluation uses two open models and one controlled text source

- SmolLM3-3B: 36 layers, tied input/output embeddings.
- OLMo3-7B: 32 layers, untied embeddings.
- WikiText-103; 256 calibration sequences and 64 test sequences, each 128 tokens.
- Jacobian lens fitted on 100 prompts; linear probe is a supervised reference.

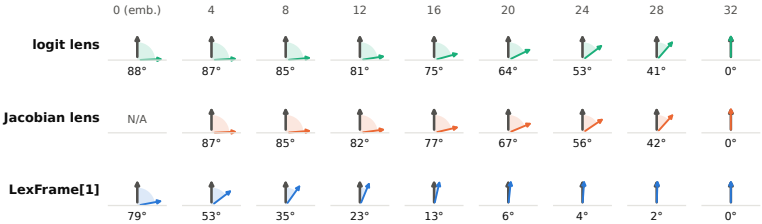
Metrics

KL: model distribution to reading, nats.
Top-1: agreement with model prediction.
ECE: calibration against the true next token.

- Layer averages exclude embedding and output.
- One deterministic run; intervals bootstrap sequences.

Paper, Sec. 4 and Appendix A, PDF pp. 5–6, 13–14.

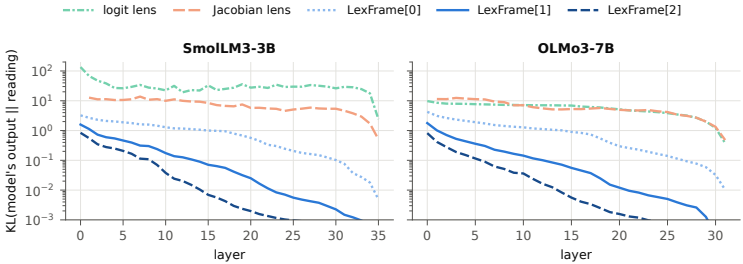
One attention round passes the fixed-lens ceiling



- OLMo3 layer 16: logit lens is 75° from native; Jacobian lens 77°; LexFrame[1] 13°.
- Four layers before output, lenses remain at least 41° away.
- At mid-depth, LexFrame[1] reaches cosine 0.97 versus the fixed-direction bound 0.30/0.35.

Paper, Fig. 1 and Sec. 4.1, PDF pp. 2, 6. Direction angles in the figure are OLMo3-7B.

LexFrame[1] readings nearly reproduce the model distribution



Layer-averaged reading fidelity

KL, SmolLM3/OLMo3: logit lens 29.06/5.66; Jacobian lens 7.51/6.19; LexFrame[0] 0.87/0.93; LexFrame[1] 0.16/0.15; LexFrame[2] 0.06/0.05.
Top-token agreement: LexFrame[1] 0.86/0.86; LexFrame[2] 0.92/0.93.

Paper, Fig. 4 and Table 1, PDF pp. 6–7. Lower KL is closer to the model.

A worked example reveals a choice before the output layer

Context and model output

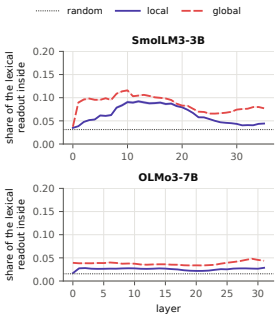
"A review of Boulter's performance in The"
Model: Guardian 0.29 runner-up: Times 0.22

Layer	Logit lens	Jacobian lens	LexFrame[1]
0	"oretical"	–	"New"
12	"comparison"	"..."	"Times"
16	"orem"	"Illustrated"	"Guardian"
20	"odor"	"Guardian"	"Guardian"
24	"Guardian"	"Guardian"	"Guardian"

- At layer 12, LexFrame exposes the runner-up; at layer 16, it reads the final choice.
- The two lenses reach "Guardian" only at layers 20 and 24.

Paper, Appendix B, Table 9, PDF p. 16. A single-position example; aggregate results are in Tables 1 and 7.

State-space distance poorly predicts what substitution changes



- Replace one position's state with another real state from the same layer.
- No point-cloud distance correlates above 0.28 with the downstream output change.
- The divergence between LexFrame[1] readings correlates at 0.94/0.95.

Why

The model's lexical readouts lie mostly outside the leading directions in which the state cloud varies.

Paper, Sec. 4.2, Table 2 and Fig. 5, PDF p. 7; 7,961 substitution pairs per layer.

The direction that measures a token is not the one that writes it

- Move each state by 10% of its norm along a token direction.
- Count how often that token gains the most logit.

Direction	SmolLM3	OLMo3
LexFrame[1], readout	0.54	0.73
LexFrame[1], basis	0.84	0.97
Native frame, readout	0.56	0.79
Native frame, basis	0.98	1.00

Interpretation

The readout is a gradient. The basis is a local move whose output effect best matches the token's lexical direction.

Paper, Sec. 4.3, Table 3, PDF p. 8. Values are shares of moves where the target token gains the most.

Intermediate-layer coefficients need the transport

State used for output coordinates	Smol KL	Smol Top-1	OLMo KL	OLMo Top-1
Layer state	8.36	0.04	5.64	0.15
Fixed map	5.19	0.13	5.46	0.13
LexFrame[1] transport	0.16	0.87	0.15	0.84
Native transport	0.00	1.00	0.00	1.00

- Coefficients are output unembedding scores after transport.
- A local gradient dotted with the layer state is not the amount of a token carried by that state.

Paper, Sec. 4.3, Table 4 and Appendix C.1, PDF pp. 8, 17.

Lexical bases steer predictions with fewer side effects

Model	Direction	Reach	Other-logit change	Clean
SmolLM3	LexFrame[1] readout	0.92	0.23	0.61
SmolLM3	LexFrame[1] basis	0.89	0.12	0.71
SmolLM3	Native basis	0.90	0.11	0.73
OLMo3	LexFrame[1] readout	0.94	0.20	0.64
OLMo3	LexFrame[1] basis	0.95	0.07	0.82
OLMo3	Native basis	0.97	0.06	0.85

- Reach: share of runner-up tokens made top-1.
- Other-logit change: mean change to other tokens at the first flip, in nats.
- Clean: flips with other-token change below 0.1 nats.

Paper, Sec. 4.4, Table 5, PDF p. 9.

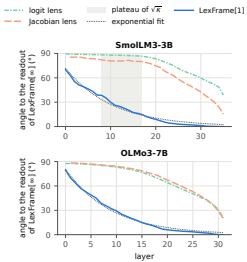
Coefficient differences capture nonlinear downstream effects

Model	Attribution	Layer	Attention	MLP
SmolLM3	LexFrame[1], difference	0.88	0.79	0.88
SmolLM3	Native, linear	0.62	0.75	0.63
OLMo3	LexFrame[1], difference	0.90	0.85	0.90
OLMo3	Native, linear	0.68	0.84	0.71

- Reference effect: remove a component's write and run the remaining model.
- Difference of transported coefficients needs no linearization.
- The largest gap is for MLP writes, where downstream response is less linear.

Paper, Sec. 4.4, Table 6, PDF p. 9. Entries are layer-averaged correlations with measured logit effects.

LexFrame closes the angle to native readouts gradually



- Mean angle shrinks by 9.2% (SmolLM3) and 10.5% (OLMo3) of what remains per layer.
- Exponential fit $R^2 = 0.98/0.99$.
- The authors relate this pattern to exponential relaxation in an overdamped system.

Paper, Appendix B, Fig. 8, PDF pp. 14–15. This is an empirical fit, not a derived Transformer-dynamics law.

Readings suggest early exits, but LexFrame itself does not save compute

Model	Reading	Settled-layer share	Exit skip share
SmolLM3	Logit lens	0.03	0.00
SmolLM3	Jacobian lens	0.09	0.00
SmolLM3	LexFrame[1]	0.82	0.45
OLMo3	Logit lens	0.11	0.04
OLMo3	Jacobian lens	0.09	0.03
OLMo3	LexFrame[1]	0.81	0.42

- Threshold chosen for 95% model agreement on half the sequences; evaluated on the other half.
- LexFrame transports the state through later layers to read it, so these figures measure potential depth, not achieved speedup.

Paper, Appendix B, Table 7, PDF pp. 15–16.

The key fidelity checks survive longer context and fit ablations

128 to 1,024 tokens

- LexFrame[1] KL: 0.16/0.15 to 0.25/0.18.
- LexFrame[2] KL: 0.06/0.05 to 0.15/0.08.
- Ranking of methods is unchanged.

Remove the fitted regression

- Mean-only KL: 1.26 (SmolLM3), 0.52 (OLMo3).
- Full regression KL: 0.16/0.15.
- Readout cosine and attribution also decline.

Paper, Appendix B, Fig. 9 and Table 8, PDF pp. 15–16.

Takeaways and evidence boundaries

1. A hidden state is best read in vocabulary directions induced by its remaining layers and context.
2. One round of target-to-prefix attention largely closes the middle-layer gap to the native frame in these two models.
3. Readout gradients and lexical basis moves answer different questions; transport is essential for intermediate coefficients.

What the experiments establish

Strong direction, reading, substitution, steering and attribution results on two open models and WikiText-103.

- One deterministic run per setting; no independent retraining replication.
- Early-exit savings are potential skips, since LexFrame itself computes through later layers.

Paper, Secs. 4–5 and Appendix A, PDF pp. 5–9, 13–14.