THE MISSING REFERENCE FRAME: CONSENSUS GEOMETRY AND OBSERVABLE DYNAMICS IN LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Intermediate states are increasingly used to read, compare, and manipulate the internal computation of language models, yet they are measured in native residual coordinates that carry no scientific reference. An invertible change of hidden basis leaves the model's function intact but changes every Euclidean length, angle, and shrinkage penalty computed from raw activations. We call this the missing reference frame. The model itself supplies the information needed to fix it: under a hidden reparameterization $h \mapsto hT$, the input embedding and the output decoder transform reciprocally, and balancing these two lexical interfaces has a unique positive-definite solution, the matrix geometric mean of their Gram operators, which is the one reference in which both interfaces induce the same geometry. Because coordinatewise models, shrinkage, and the direction of a lexical gradient are all defined relative to a reference, this model-induced reference is indispensable for results that describe the model rather than its parameterization. LEXICAL LENS builds on it in three modules. Consensus Geometry fixes this data-free, model-wide reference and reads every state through the decoder's lexical covectors as a vocabulary-indexed Semantic State. Observable Dynamics estimates, without any decoder signal, the part of the terminal state that a source layer predictably supports, which defines Semantic Transitions across depth. Information Projection passes the Riesz-lifted exact forward-KL gradient through the same frozen response operator and allocates three coefficients in a globally convex problem. Across ten models from four families, SmolLM2, MiniCPM5, Granite 4.2, and Ministral 3 (135M to 14B parameters, 322 layers), resetting the Euclidean identity after a chart change or using a lexical covector as a vector distorts the measured quantity by 14% to 92%. LEXICAL LENS lowers macro held-out forward KL from 2.984 for a full-affine Tuned Lens trained on the same data to 2.033, improves every layer and every held-out sequence, and also improves NLL and next-token accuracy on every model. The decoder-free Observable Dynamics alone accounts for most of this gain, and both the lens and its Semantic States remain stable as calibration data grow, while Semantic Transitions need the largest budgets.

1 INTRODUCTION

The hidden states of a language model are becoming an increasingly important interface for understanding what the model represents and how its computation develops. Probes and predictive lenses read information from intermediate layers, patching and editing methods intervene on them, and recent work studies semantic directions, sparse features, and verbalizable representations as objects that can be compared or manipulated (Belinkov, 2022; Belrose et al., 2023; Ghandeharioun et al., 2024; Huben et al., 2024; Fang et al., 2025; Gurnee et al., 2026). This use of internal representations rests on a measurement requirement. To say that a concept moves, strengthens, or changes direction across depth, the same semantic object has to be measured at every layer, with lengths, angles, and gradient directions that mean the same thing throughout.

Native residual coordinates do not meet that requirement. For an invertible $T \in GL(d)$, the reparameterization $h \mapsto hT$ with compensating changes of the adjacent weights leaves the computed function unchanged (Wang & Wang, 2025), yet it changes every Euclidean norm, every coordinatewise

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.

model, every ridge penalty, and the identification of a lexical gradient with a hidden direction. A method that treats the model's native axes as orthonormal therefore measures a property of the parameterization rather than of the model. We call this the *missing reference frame*. It is sharpest for flexible readouts: a layerwise translator such as the Tuned Lens (Belrose et al., 2023) can absorb the choice of coordinates, the evolution toward the endpoint, and the output correction at once, so predictive accuracy does not reveal which part of the map reflects the model.

Language models nevertheless contain a natural source of reference. Their input embedding and output decoder refer to the same vocabulary but serve different functional roles, and analyses of weight tying and of input and output embeddings show that these roles are related but not geometrically identical (Press & Wolf, 2017; Inan et al., 2017; Derby et al., 2020; Lopardo et al., 2026). Invariance of the logits forces the two interfaces to transform reciprocally under a hidden reparameterization, so they form a primal/dual pair rather than two copies of one Euclidean embedding. With the structured semantic geometry found in activations (Gurnee & Tegmark, 2024; Park et al., 2024; Jiang et al., 2024; Park et al., 2025; Valois et al., 2025), this suggests that semantic structure is intrinsic to the model without any residual basis being semantically privileged. We call semantic structure grounded in the frozen model's own lexical interfaces *model-inherent semantics*. It separates semantic content from the coordinates used to measure it.

Our central idea is to separate reference, dynamics, and readout instead of asking one layerwise translator to solve all three. We first derive a single reference from the frozen model: requiring the two lexical interfaces to agree as closely as possible yields a strictly convex problem whose unique solution is the matrix geometric mean of their Gram operators, so the reference is data-free and transforms correctly under every hidden reparameterization. We then ask, without any decoder signal, which terminal quantities are predictably available from each source depth. Reading States and their depthwise changes through the decoder's lexical covectors turns both answers into vocabulary-indexed Semantic States and Semantic Transitions, which can be compared across depths, models, and calibration samples. Only after those two pieces are fixed do we fit the terminal lexical distribution, with a small convex projection that cannot redefine the geometry used to compare states.

Our contributions are:

- We identify the missing reference frame in depthwise language-model analysis and derive the Consensus reference as the unique positive-definite balance of the two reciprocal lexical interfaces. On ten models, resetting the Euclidean identity after a chart change or treating a lexical covector as a hidden vector distorts the measured quantity by 14% to 92%, so structured estimators need the model-induced reference to describe the model rather than its parameterization.
- We introduce LEXICAL LENS, which separates reference, predictive support, and readout into Consensus Geometry, Observable Dynamics, and Information Projection. Beyond a closed-form backbone, each layer selects one response scale by decoder-free leave-sequence-out risk and three coefficients by a convex exact-KL projection that never increases the calibration objective.
- On ten models from four families, LEXICAL LENS lowers macro forward KL from 2.984 for a full-affine Tuned Lens trained on the same sequences to 2.033, improving all 322 layers and 640 held-out sequences as well as NLL and next-token accuracy. The decoder-free LEXICAL LENS* alone yields 22.1%, and the lens and its semantic readings stabilize with more calibration data.

2 RELATED WORK

Predictive Probes and Lenses. Probing asks what can be decoded from intermediate states, but probe accuracy conflates properties of the representation with the flexibility of the readout (Belinkov, 2022). For autoregressive models, the Logit Lens applies the output decoder to intermediate states, the Tuned Lens learns one affine translator per layer, the Future Lens anticipates later tokens, and cross-layer casting approximates deeper computation with learned linear maps (nostalgebraist, 2020; Belrose et al., 2023; Pal et al., 2023; Yom Din et al., 2024). Spectral analyses of the embedding and unembedding identify which directions make intermediate vocabulary projections informative (Cancedda, 2024), Patchscopes uses the model's own downstream computation as the readout (Ghandeharioun et al., 2024), and the Jacobian Lens isolates representations poised to affect verbalizable outputs (Gurnee et al., 2026). These methods differ in how much geometry and future computation they absorb into the readout. LEXICAL LENS fixes the geometry first and restricts what is learned

to a response scale and three coefficients. Our claims are predictive rather than causal, with causal abstraction providing the separate standard for functional necessity (Geiger et al., 2025).

Latent Semantics and Lexical References. Recent work increasingly treats semantics as structure in representation space. Formal work on the linear representation hypothesis links concept directions to probing and steering and explains how such directions can emerge from next-token prediction, while empirical studies recover spatial, temporal, categorical, hierarchical, and multi-token concept geometry (Park et al., 2024; Jiang et al., 2024; Gurnee & Tegmark, 2024; Park et al., 2025; Valois et al., 2025). Representation engineering and sparse autoencoders construct learned feature coordinates, whereas SemanticLens maps internal components into an external foundation-model semantic space (Zou et al., 2023; Huben et al., 2024; Dreyer et al., 2025). For language models, the lexical boundary gives a stronger model-internal reference: input and output embeddings share token identities but need not induce the same geometry. Recent analyses of weight tying make this distinction explicit, showing different semantic and contextual structure in the two interfaces and measurable bias toward the output space when they are tied (Bertolotti & Cazzola, 2024; Lopardo et al., 2026). Vocabulary-defined semantics, semantic-aware layer freezing, pseudo-inverse tying, and SemRF develop related vocabulary-grounded constructions (Gu et al., 2024; 2025; 2026a;b). Consensus Geometry differs by treating the two frozen lexical views as a primal/dual pair, reconciling them once through their reciprocal transformation law, and using the resulting model-wide reference at every depth.

Language Model Geometry and Dynamics. Depthwise analyses must distinguish change in represented content from change of coordinates. Transformer gauge-symmetry results characterize function-preserving reparameterizations of hidden coordinates (Wang & Wang, 2025), while recent spectral analyses find structured layerwise shifts and critical layers in pretrained language models (Liu et al., 2025). Complementary work reports that low-order linear surrogates capture depthwise sensitivity increasingly well as model scale increases (Nettasinghe & Joseph, 2026), and learning-dynamics analyses decompose how predictions move in logit space under a parameter update (Ren & Sutherland, 2025). Our dynamics problem is narrower: given a fixed reference and a source depth, identify which terminal quantities are predictably supported by that source. This view is closest to observable-based dynamical modeling such as Koopman operators (Otto & Rowley, 2021), but we do not assume a Markov-complete hidden state or a semigroup over layers. Response estimation uses covariance shrinkage (Chen et al., 2010; Ledoit & Wolf, 2021). The decoder-facing fit is an exact forward-KL information projection onto a three-field family (Cho et al., 2026).

3 Approach

A frozen language model with L blocks produces, at every position, normalized residual states $x_\ell \in \mathbb{R}^d$ after block ℓ and a terminal state $y = x_L$, and the frozen decoder maps a state to logits over the vocabulary $\mathcal{V}$. A lens predicts the terminal distribution p_y from x_ℓ, and its primary error is the forward KL $D_{\mathrm{KL}}(p_y \| p_{\hat{z}})$. Lexical Lens separates three questions that a flexible layerwise map would otherwise conflate: which reference makes hidden quantities meaningful, which terminal quantities a source state predictably supports, and how the frozen decoder should realize them. The first uses only the lexical interfaces, the second uses calibration states but never the decoder, and the third uses exact forward KL but changes only three coefficients. Appendix A gives the implementation, the complexity analysis, the experimental setup, and all proofs.

3.1 Consensus Geometry

A semantic analysis needs a reference before it fits a predictor. Let $E, O \in \mathbb{R}^{V_s \times d}$ be the input-embedding and output-decoder rows on the tokenizer-linked lexical support, with the decoder's common-logit direction removed exactly. A chart change $h \mapsto hT$ must preserve every lexical pairing, $(ET)(OT^{-T})^\top = EO^\top$, so the interfaces transform contragrediently,

$$E \mapsto ET, \qquad O \mapsto OT^{-T}. \tag{1}$$

A raw chart does not specify which Euclidean metric is scientific, and the native identity is one arbitrary choice. We instead seek the reference in which the two interfaces disagree least.

With $G_E = E^\top E$ and $G_O = O^\top O$, the paired disagreement $\|ET - OT^{-T}\|_F^2$ depends on T only through $G = TT^\top$ and equals $\mathrm{tr}(G_E G) + \mathrm{tr}(G_O G^{-1})$ up to a constant. This function is strictly

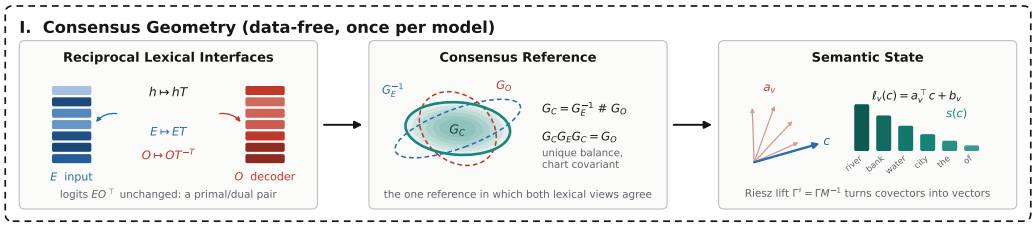

Figure 1: **Fix the reference before studying depth.** The input embedding and output decoder transform reciprocally under a hidden change of coordinates. Consensus Geometry balances them once per model into the reference G_C, which defines the Semantic State and the Riesz lift used later.

convex on the positive-definite cone, and its unique minimizer is the matrix geometric mean

$$G_C = G_E^{-1} \# G_O = G_E^{-1/2} \big(G_E^{1/2} G_O G_E^{1/2}\big)^{1/2} G_E^{-1/2}, \qquad G_C G_E G_C = G_O. \tag{2}$$

The second identity explains the name. In any chart where G_C is the identity, the input and output interfaces have the same Gram operator, $T^\top G_E T = T^{-1} G_O T^{-T}$ for $T = G_C^{1/2}$, so the two lexical views reach consensus on every length and angle. Under $h \mapsto hT$ the solution transforms as $G_C \mapsto T^{-1} G_C T^{-T}$, which is exactly the law of a State metric, so the reference is a property of the model rather than of its chart. Each interface alone already yields a covariant candidate, G_E^{-1} or G_O, but the two generally disagree; G_C is their geodesic midpoint, the unique choice that treats them symmetrically, and exchanging their roles inverts it. It depends only on frozen parameters and tokenizer identities, not on calibration states, layerwise targets, or KL. Its realization is factored into an orthogonal *execution gauge* U, in which States are measured as $c_\ell = x_\ell U$, and a positive reciprocal factor P that carries the State metric $M = PP^\top$ in these coordinates.

Semantic State. Compiling the frozen decoder into the gauge gives centered lexical covectors $\{a_v\}_{v \in \mathcal{V}}$, the *Semantic Basis*. For a State c, the *Semantic Logit* $\ell_v(c) = a_v^\top c + b_v$ is one lexical coordinate of its meaning, and the vocabulary-wide vector $s(c) = (\ell_v(c))_{v \in \mathcal{V}}$ is its *Semantic State*. The same basis defines the convex lexical potential $F(c) = \log \sum_v \exp(a_v^\top c + b_v)$ with mean map $m = \nabla F$, whose Bregman divergence is exactly forward KL, $D_F(z, y) = D_{\mathrm{KL}}(p_y \| p_z)$. Semantic States serve as the common readout of all later analysis: they are indexed by tokens rather than by hidden axes, the pairing $a_v^\top c$ is unchanged by any chart change, and they can therefore be compared across depths, models, and calibration samples. Module I thus owns the reference, the frozen readout, and the distinction between vectors and covectors. A State residual is a vector, whereas a lexical gradient is a covector and transforms with T^{-T} as in Equation (1). The Riesz map $\Gamma \mapsto \Gamma^\sharp = \Gamma M^{-1}$ induced by the reference converts the second into the first, and it is the only coordinate-consistent way to turn a lexical gradient into a hidden direction.

3.2 Observable Dynamics

Once the reference is fixed, the remaining question is statistical: which terminal quantities can be recovered from the source state at depth ℓ. Module II answers it without the decoder, the softmax, or KL.

Let $X, Y \in \mathbb{R}^{N \times d}$ collect calibration source and terminal States in the gauge, drawn from Q sequences of m positions each. Module II first fits the minimal coordinatewise persistence field, the affine-diagonal *backbone* $B = \mathbf{1} b^\top + XD$ with D diagonal, and forms the *State Residual* $E_{\mathrm{st}} = Y - B$. The empirical source States define one continuous family of response operators,

$$H_X(\rho) = X(X^\top X + \rho I)^{-1} X^\top, \qquad \rho = \rho_0 e^s, \tag{3}$$

with spectrum in $[0, 1]$ and no rank to select. The origin ρ_0 is a closed-form shrinkage level computed from the source Gram (Chen et al., 2010); it fixes $s = 0$ and is not fitted. A single scalar is chosen by exact leave-sequence-out State risk measured in the reference metric,

$$s^* = \arg\min_s \; \frac{1}{2d} \sum_i w_i \big\| \big(E_{\mathrm{st}}^- - H_X^-(\rho) E_{\mathrm{st}}^-\big)_i P \big\|_2^2, \tag{4}$$

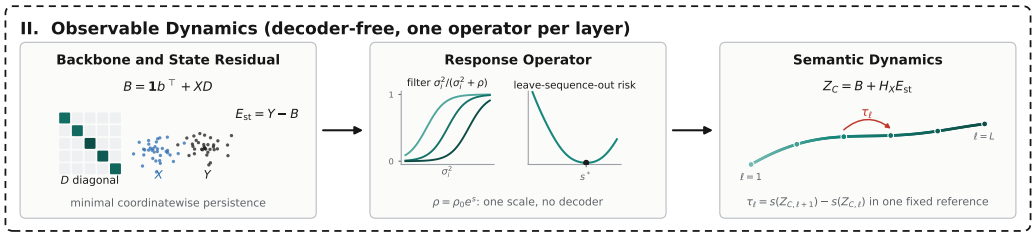

Figure 2: **Identify observable terminal motion inside the fixed reference.** An affine diagonal backbone and one ridge response operator, whose scale is selected by decoder-free leave-sequence-out State risk, predict the terminal State directly from each depth and define Semantic Dynamics.

where the superscript $^-$ means that the backbone and the response at every position are computed with that position's entire sequence removed, and $w_i = 1/(Qm)$ gives each sequence equal weight. Sequence exclusion matters because the positions of one context are repeated measurements rather than independent samples. The selected operator $H_X = H_X(\rho_0 e^{s^*})$ is frozen, and it defines the *State Response* $R_{\text{state}} = H_X E_{\text{st}}$ and the Module-II State $Z_C = B + R_{\text{state}}$. H_X is a *predictive realizability* operator: it restricts residual structure to what the source State cloud supports. The module is *observable* in the sense of observable-based dynamics (Otto & Rowley, 2021): it does not model a hidden layer-to-layer transition, but only those terminal quantities that are recoverable from the source States, and it certifies each by held-out State risk rather than assuming it. LEXICAL LENS* is the decoder-free lens that deploys Z_C, that is, Modules I and II without Information Projection.

Semantic Dynamics. Reading the Module-II trajectory through the Semantic Basis defines the dynamic semantic objects. The *State Transition* $\Delta Z_\ell = Z_{C,\ell+1} - Z_{C,\ell}$ becomes semantic only after the basis is applied: the *Semantic Transition* is $\tau_\ell = s(Z_{C,\ell+1}) - s(Z_{C,\ell})$, its label-wise coordinate is the *Semantic Delta* $\delta_{v,\ell} = a_v^\top \Delta Z_\ell$, and the depth sequence $\{\tau_\ell\}$ is *Semantic Dynamics*. They express, token by token, what the model predictably adds to its terminal reading between consecutive depths. These objects subtract States estimated at different depths, which is meaningful because every depth is estimated by one chart-covariant procedure in one fixed reference. Their stability across calibration samples, tested in Section 4.4, decides whether they describe the model or the sample.

3.3 INFORMATION PROJECTION

Modules I and II answer general questions about meaning and predictability. Information Projection is the lens-specific addition: it allocates the fixed fields to the terminal distribution and adds no dynamics.

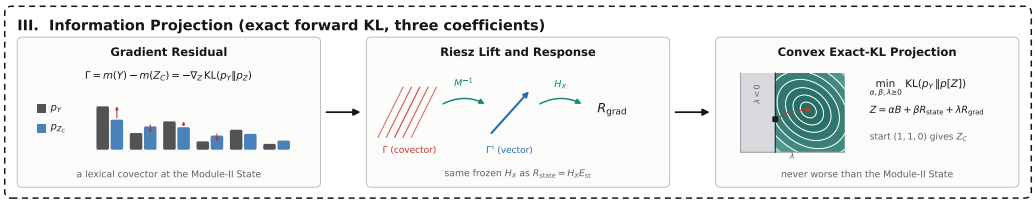

Figure 3: **Apply Information Projection after geometry and dynamics are fixed.** The exact forward-KL Gradient Residual is Riesz-lifted and passed through the same frozen operator, and a convex projection fixes three coefficients without redefining Consensus Geometry or Observable Dynamics.

At the Module-II anchor, exact forward KL defines the *Gradient Residual* $\Gamma = m(Y) - m(Z_C)$, the negative gradient of $D_{\text{KL}}(p_Y \| p_Z)$ at $Z = Z_C$. Unlike E_{st}, Γ is a covector. It is Riesz-lifted to the State vector $\Gamma^\sharp$, and only then does the same frozen operator act on it, giving the *Gradient Response* $R_{\text{grad}} = H_X \Gamma^\sharp$. The final prediction solves one exact-KL problem over three coefficients,

$$(\alpha^*, \beta^*, \lambda^*) = \arg \min_{\alpha,\beta\in\mathbb{R},\ \lambda\geq 0} D_{\text{KL}}\Big(p_Y \,\Big\|\, p\big[\alpha B + H_X(\beta E_{\text{st}} + \lambda \Gamma^\sharp)\big]\Big). \tag{5}$$

Because the logits are affine in the three coefficients and the log-partition function is convex, Equation (5) is convex for the frozen H_X, and its KKT conditions certify the optimum both in the interior and on the face $\lambda = 0$. The point $(1, 1, 0)$ reproduces Z_C, so the final State can never have higher calibration KL than the Module-II State. Equation (5) is an information projection of the terminal distribution onto the source-realizable family $\mathrm{span}\{B, R_{\mathrm{state}}\} + \mathbb{R}_+ R_{\mathrm{grad}}$ (Cho et al., 2026). The coefficients have fixed roles, with α scaling persistence, β the source-predictable State innovation, and $\lambda \geq 0$ the source-realizable lexical correction, and they compile into one deployable diagonal-plus-factorized map per layer.

4 EXPERIMENTS

The evaluation addresses three research questions. RQ1 (reference) examines whether native coordinates can stand in for the reciprocal reference, by measuring what the two common shortcuts do to a quantity under a hidden chart change. RQ2 (fidelity) measures how faithfully the complete lens recovers each model's own terminal distribution relative to existing readouts, and which module carries the gain. RQ3 (stability) examines whether performance and semantic readings stabilize as calibration evidence grows. Appendix B reports per-model fidelity, the fitted estimator structure, the measured fitting cost, semantic convergence by family, and a broader discussion.

Setups. We evaluate ten base models from four families: SmolLM2 135M, 360M, and 1.7B (Ben Allal et al., 2025), MiniCPM5 1B and 2B (OpenBMB, 2026), Granite 4.2 3B and 8B (IBM Granite Team, 2026), and Ministral 3 3B, 8B, and 14B (Liu et al., 2026). All are dense decoder-only Transformers, spanning widths 576 to 5120 and 322 intermediate layers; SmolLM2 ties its input and output embeddings, and the other families keep them separate. Calibration uses $Q = 64$ sequences of length 128 from the WikiText-103 training split (Merity et al., 2017) and all 127 causal next-token positions of each; evaluation uses a disjoint set of 64 test sequences that is never used for selection or stopping. Baselines are the Logit Lens and a full-affine Tuned Lens per layer (Belrose et al., 2023), identity-initialized and trained on the *same* calibration positions under forward KL. The comparison is thus at a matched and deliberately small budget of 8,128 positions per layer, the regime of per-model analysis, in which the translator has up to 2.6×10^7 parameters per layer. We report forward KL to the model's own terminal distribution as the primary metric, together with the negative log-likelihood of the observed next token (NLL) and next-token accuracy $\mathbf{1}[\arg\max_v p(v) = y_{\mathrm{next}}]$. Metrics are averaged over held-out positions and layers, and macro values average the per-model means; Δ is a paired sequence-level difference with a 95% interval from 2000 paired sequence bootstraps.

The frame experiment applies dense chart changes T built from a Householder reflection and a log-uniform spectrum with condition number $\kappa \in \{2, 4, 8\}$ to every layer and all calibration positions of each model. Calibration convergence refits the unchanged estimator on nested prefixes of the calibration stream with $Q \in \{8, 16, 32, 64\}$ sequences while the held-out stream stays fixed. Performance stability is measured by the State KL gap and the Final KL gap, the paired held-out KL differences of LEXICAL LENS* and LEXICAL LENS to their own $Q = 64$ fits. Semantic stability is measured on the Module-II trajectory by the Semantic State Agreement, the vocabulary-wide correlation between Semantic States estimated at adjacent budgets, and the Semantic Dynamics Agreement, the same correlation between Semantic Transitions. Depth is normalized as ℓ/L, layerwise gains are averaged within ten equal depth bins, and in every figure a line is a family mean and the shaded band of the same color spans the models of that family.

4.1 REFERENCE FRAME VERIFICATION

The frame experiment addresses RQ1 by measuring two shortcuts that analyses of raw activations take implicitly. *Identity reset* declares the new chart orthonormal and compares squared State norms before and after the change. *Covector as vector* treats the exact-KL Gradient Residual, a lexical covector, as if it were a hidden vector and compares its transformed array with the correctly transported direction.

Table 1 shows that neither shortcut is harmless. Identity reset changes squared State norms by 14% to 66% across layers and models, with means of 34%, 42%, and 56% at $\kappa = 2, 4, 8$. Treating the Gradient Residual as a vector misplaces it by 38% to 92%, with means of 46%, 59%, and 72%. Both distortions appear on all ten models and all 322 layers and grow monotonically with κ for every

Table 1: Relative distortion (%) of the two invalid shortcuts.

	Identity reset			Covector as vector		
Model	$\kappa = 2$	$\kappa = 4$	$\kappa = 8$	$\kappa = 2$	$\kappa = 4$	$\kappa = 8$
SmolLM2 135M	38.0	48.1	61.4	40.0	58.2	80.7
SmolLM2 360M	36.7	45.4	58.0	39.3	47.3	58.4
SmolLM2 1.7B	33.3	41.4	55.1	43.0	55.1	68.7
MiniCPM5 1B	34.5	44.3	58.8	44.0	57.5	71.7
MiniCPM5 2B	35.3	45.4	59.7	46.2	61.7	76.4
Granite 4.2 3B	27.6	33.9	48.3	39.7	48.7	61.2
Granite 4.2 8B	28.4	34.4	48.5	39.3	47.2	58.5
Ministral 3 3B	35.3	44.9	58.8	40.1	49.9	63.2
Ministral 3 8B	35.6	45.6	59.6	65.0	82.1	91.2
Ministral 3 14B	31.9	39.8	54.0	66.1	82.9	91.6
Mean	33.7	42.3	56.2	46.3	59.1	72.2

model, because they are properties of the chart rather than of the model; their size varies with the anisotropy of each model's State cloud and lexical gradient, but no model escapes either shortcut.

These controls make the value of the Consensus reference concrete, because every non-trivial operation of LEXICAL LENS depends on a reference. The backbone of Observable Dynamics is diagonal only relative to a chosen basis, the ridge penalty and the State risk in Equations (3) and (4) measure lengths, and the Gradient Response in Equation (5) needs a Riesz lift. In native coordinates the last two inherit exactly the two distortions of Table 1, and the first changes its model class, so the lens and its semantic readings would describe the parameterization. The Consensus reference is induced by the model's own interfaces, fixed without data, and covariant by construction, which is what makes Tables 2 and 3 statements about the frozen model. A full-affine translator escapes the distortion only by absorbing any chart into its weights, which is also why its fit cannot separate reference, dynamics, and readout. Native coordinates that carry the transported Consensus metric are the same reference in another chart; what is indispensable is the reference itself.

4.2 TERMINAL DISTRIBUTION PREDICTION

RQ2 is answered first at the macro level. The Logit Lens is far from the terminal distribution (macro KL 11.130, NLL 14.467, next-token accuracy 5.50%), and the Tuned Lens closes most of that gap (2.984, 6.165, 17.69%). LEXICAL LENS reaches macro KL 2.033, NLL 5.266, and accuracy 20.10%, a 31.9% reduction relative to the Tuned Lens (macro paired $\Delta = -0.950$ nats, 95% CI $[-0.961, -0.940]$), while the decoder-free LEXICAL LENS* already reaches 2.323, 5.561, and 18.07%. All lenses are evaluated on the same held-out positions, so the comparison isolates the estimator.

Table 2 resolves the result by model. The reduction ranges from 25.4% on Ministral 3 3B to 41.0% on MiniCPM5 1B and shows no systematic decline with scale: it grows with model size within SmolLM2 (26.6% to 32.9%) and Ministral 3 (25.4% to 36.0%) and changes by less than four points within Granite 4.2 and MiniCPM5. LEXICAL LENS improves forward KL on **322/322 layers** and on **640/640 held-out sequences**. It also has the lowest NLL on every model and on all 322 layers, by 0.51 to 1.13 nats, and the highest next-token accuracy on every model, by 0.1 to 3.7 points. Figure 4 places both lenses against the two baselines at every depth. The LEXICAL LENS curves fall steeply over the last third of the network, where the Tuned Lens plateaus; averaged over models, the gain over the Tuned Lens grows from 0.84 nats in the first third of the layers to 1.13 nats in the last third.

4.3 MODULE DECOMPOSITION

The decoder-free Observable Dynamics carries most of the gain. LEXICAL LENS* reduces macro KL from 2.984 to 2.323 (22.1%, macro paired CI $[-0.670, -0.650]$). It beats the Tuned Lens on all ten models, on all 640 held-out sequences, and on 314 of 322 layers, the exceptions being eight early and

Table 2: Per-model held-out forward KL. "Red." is the relative reduction of LEXICAL LENS against the Tuned Lens, and Δ is the paired difference between them.

Model	Layers	Tuned Lens	LEXICAL LENS*	LEXICAL LENS	Red.	Δ [95% CI]
SmolLM2 135M	29	2.372	1.897	**1.742**	26.6%	-0.630 [-0.665, -0.597]
SmolLM2 360M	31	2.573	2.010	**1.801**	30.0%	-0.772 [-0.799, -0.745]
SmolLM2 1.7B	23	2.954	2.277	**1.982**	32.9%	-0.973 [-1.011, -0.935]
MiniCPM5 1B	23	2.678	1.698	**1.579**	41.0%	-1.099 [-1.131, -1.066]
MiniCPM5 2B	41	3.117	2.143	**1.946**	37.6%	-1.171 [-1.199, -1.145]
Granite 4.2 3B	39	3.260	2.796	**2.322**	28.8%	-0.938 [-0.979, -0.899]
Granite 4.2 8B	39	3.761	3.211	**2.695**	28.4%	-1.066 [-1.108, -1.029]
Ministral 3 3B	25	2.708	2.328	**2.020**	25.4%	-0.688 [-0.715, -0.663]
Ministral 3 8B	33	3.146	2.479	**2.153**	31.6%	-0.993 [-1.025, -0.961]
Ministral 3 14B	39	3.267	2.395	**2.092**	36.0%	-1.175 [-1.209, -1.140]

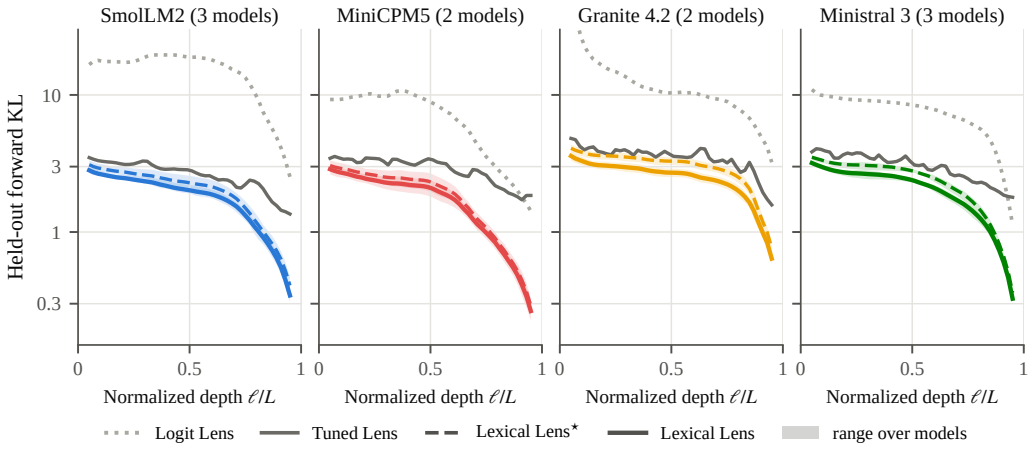

Figure 4: **Layerwise held-out forward KL by family** (log scale). LEXICAL LENS stays below both baselines at every layer of every model.

middle layers of Granite 4.2 and Ministral 3 3B. A diagonal backbone, one frozen ridge operator, and a single scale thus outperform a learned $d \times d$ affine translator on its own objective without any decoder evaluation during fitting. Information Projection then lowers held-out KL by a further 12.5% (macro paired $\Delta = -0.290$, CI $[-0.294, -0.286]$) and improves on all 322 layers. Figure 5 resolves both steps by depth and family. The reduction of both lenses is positive at almost every depth, rises in the second half of the network where the Tuned Lens plateaus, and the additional gain of Information Projection, the gap between the solid and dashed lines, is present at all depths. It is largest on Granite 4.2 and narrows toward the last layers on the other families. The lexical correction is used on every layer of every model ($\lambda^* > 0$ on 322/322 layers).

The same structure makes fitting cheap: one spectral factorization per layer serves every scale and sequence deletion in Equation (4), and the scale search and Equation (5) need about 21 and 22 evaluations, without gradients or optimizer moments.

4.4 CALIBRATION CONVERGENCE

RQ3 concerns stability along the nested calibration trajectory. In Table 3, both KL gaps shrink at every doubling for every model, from 0.588 and 0.636 nats at $Q = 8$ to 0.184 and 0.175 at $Q = 32$ (State and Final), with every per-model interval excluding zero; the two gaps stay close, so Modules II and III converge together. Semantic States agree to 0.972 already between $Q = 8$ and 16 and to

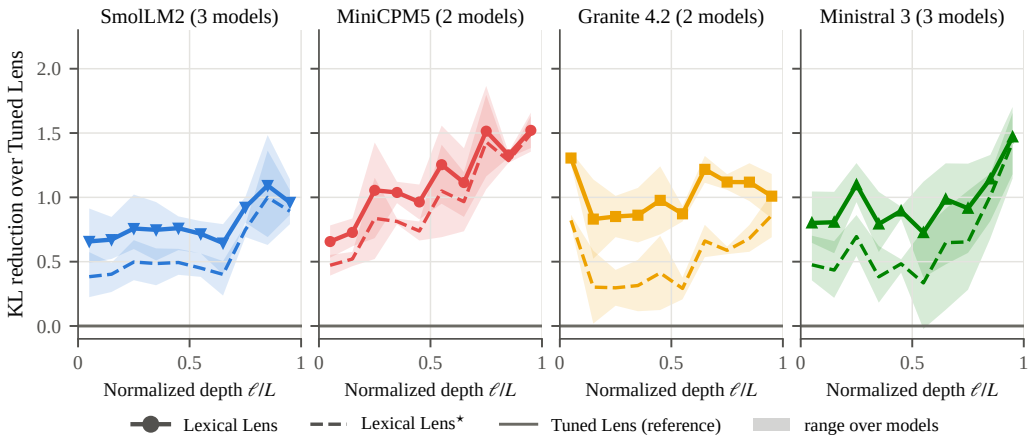

Figure 5: **Layerwise KL reduction over the Tuned Lens by family.** The gap between the solid and dashed lines is the contribution of Information Projection.

Table 3: Calibration convergence per model. KL gaps are in nats.

	State KL gap			Final KL gap			Semantic State Agreement			Semantic Dynamics Agreement		
Model	Q=8	16	32	Q=8	16	32	8→16	16→32	32→64	8→16	16→32	32→64
SmolLM2 135M	0.486	0.334	0.137	0.466	0.314	0.118	0.967	0.971	0.980	0.762	0.776	0.824
SmolLM2 360M	0.515	0.352	0.143	0.523	0.355	0.136	0.962	0.964	0.975	0.753	0.752	0.799
SmolLM2 1.7B	0.595	0.423	0.183	0.661	0.438	0.180	0.964	0.962	0.972	0.732	0.720	0.759
MiniCPM5 1B	0.492	0.330	0.134	0.523	0.340	0.136	0.961	0.962	0.976	0.770	0.771	0.818
MiniCPM5 2B	0.592	0.397	0.170	0.653	0.416	0.165	0.962	0.960	0.974	0.765	0.753	0.798
Granite 4.2 3B	0.653	0.423	0.217	0.684	0.426	0.195	0.984	0.985	0.988	0.808	0.795	0.822
Granite 4.2 8B	0.661	0.452	0.264	0.745	0.488	0.242	0.984	0.983	0.987	0.819	0.802	0.829
Ministral 3 3B	0.615	0.425	0.191	0.697	0.443	0.184	0.978	0.979	0.983	0.785	0.782	0.800
Ministral 3 8B	0.636	0.439	0.200	0.716	0.460	0.197	0.981	0.982	0.985	0.768	0.763	0.786
Ministral 3 14B	0.631	0.437	0.202	0.695	0.455	0.198	0.981	0.981	0.985	0.778	0.766	0.791
Mean	0.588	0.401	0.184	0.636	0.413	0.175	0.972	0.973	0.981	0.774	0.768	0.802

0.981 between 32 and 64, above 0.96 for every model. Semantic Transitions, which are differences of nearby States, are harder to estimate (0.77 to 0.80 on average) and agree most at the largest budgets. The Semantic State read at each depth is thus a property of the frozen model already at small budgets, whereas the finer change between layers benefits most from more data.

5 Conclusion

Depthwise semantic analysis should fix its reference before fitting a flexible predictor, and the language model itself provides one. The reciprocal transformation of its input and output interfaces determines this reference uniquely and without data, whereas the shortcuts of raw-coordinate analysis distort measured quantities under function-preserving reparameterizations. Built on this reference, Lexical Lens is more faithful than a learned affine translator on ten models from four families in forward KL, NLL, and next-token accuracy, with Observable Dynamics carrying most of the gain. Its Semantic States are stable already at small calibration budgets, and its Semantic Transitions converge more slowly as the finer object. Semantic geometry, source-supported dynamics, and task-specific readout are best treated as distinct objects whenever internal representations are compared, edited, or read across depth.

AI USE STATEMENT

We used generative AI tools to aid or polish writing and for proving mathematical claims, namely to draft and check the proofs of the uniqueness, covariance, and convexity results; no other disclosure category applies to this work. We have reviewed all AI-assisted work, verified every proof step by hand, and checked every reported number against the experiment outputs.

REPRODUCIBILITY STATEMENT

The complete code package is provided as supplementary material. It implements all three modules and both baselines and contains the experiment commands and the scripts that rebuild every table and figure from the raw outputs. The Experiments section and the technical appendix specify the models, data, protocol, and estimation details, and the appendix contains complete proofs.

REFERENCES

Yonatan Belinkov. Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, 48(1):207–219, 2022. doi: 10.1162/coli_a_00422. URL https://aclanthology.org/2022.cl-1.7/.

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, 2023. URL https://arxiv.org/abs/2303.08112.

Loubna Ben Allal, Anton Lozhkov, Elie Bakouch, Gabriel Martín Blázquez, Guilherme Penedo, Lewis Tunstall, Andrés Marafioti, Hynek Kydlíček, Agustín Piñeres Lajarín, Vaibhav Srivastav, et al. SmolLM2: When smol goes big – data-centric training of a small language model, 2025. URL https://arxiv.org/abs/2502.02737.

Francesco Bertolotti and Walter Cazzola. By tying embeddings you are assuming the distributional hypothesis. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 3584–3610. PMLR, 2024. URL https://proceedings.mlr.press/v235/bertolotti24a.html.

Rajendra Bhatia. *Positive Definite Matrices*. Princeton Series in Applied Mathematics. Princeton University Press, 2007.

Nicola Cancedda. Spectral filters, dark signals, and attention sinks. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 4792–4808. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.acl-long.263. URL https://aclanthology.org/2024.acl-long.263/.

Yilun Chen, Ami Wiesel, Yonina C. Eldar, and Alfred O. Hero. Shrinkage algorithms for MMSE covariance estimation. *IEEE Transactions on Signal Processing*, 58(10):5016–5029, 2010. doi: 10.1109/TSP.2010.2053029.

Gunhee Cho, Jae Kwang Kim, and Yumou Qiu. A tutorial on Bregman projection in statistics, 2026. URL https://arxiv.org/abs/2606.21714.

Steven Derby, Paul Miller, and Barry Devereux. Analysing word representation from the input and output embeddings in neural network language models. In *Proceedings of the 24th Conference on Computational Natural Language Learning*, pp. 442–454. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.conll-1.36. URL https://aclanthology.org/2020.conll-1.36/.

Maximilian Dreyer, Jim Berend, Tobias Labarta, Johanna Vielhaben, Thomas Wiegand, Sebastian Lapuschkin, and Wojciech Samek. Mechanistic understanding and validation of large AI models with SemanticLens. *Nature Machine Intelligence*, 7:1572–1585, 2025. doi: 10.1038/s42256-025-01084-w. URL https://www.nature.com/articles/s42256-025-01084-w.

Junfeng Fang, Houcheng Jiang, Kun Wang, Yunshan Ma, Shi Jie, Xiang Wang, Xiangnan He, and Tat-Seng Chua. AlphaEdit: Null-space constrained knowledge editing for language models. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=HvSytvg3Jh.

Atticus Geiger, Duligur Ibeling, Amir Zur, Maheep Chaudhary, Sonakshi Chauhan, Jing Huang, Aryaman Arora, Zhengxuan Wu, Noah Goodman, Christopher Potts, and Thomas Icard. Causal abstraction: A theoretical foundation for mechanistic interpretability. *Journal of Machine Learning Research*, 26(83):1–64, 2025. URL https://www.jmlr.org/papers/v26/23-0058.html.

Asma Ghandeharioun, Avi Caciularu, Adam Pearce, Lucas Dixon, and Mor Geva. Patchscopes: A unifying framework for inspecting hidden representations of language models. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 15466–15490. PMLR, 2024. URL https://proceedings.mlr.press/v235/ghandeharioun24a.html.

Jian Gu, Aldeida Aleti, Chunyang Chen, and Hongyu Zhang. Vocabulary-defined semantics: Latent space clustering for improving in-context learning, 2024. URL https://arxiv.org/abs/2401.16184.

Jian Gu, Aldeida Aleti, Chunyang Chen, and Hongyu Zhang. A semantic-aware layer-freezing approach to computation-efficient fine-tuning of language models. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 8019–8033. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.findings-acl.420. URL https://aclanthology.org/2025.findings-acl.420/.

Jian Gu, Aldeida Aleti, Chunyang Chen, and Hongyu Zhang. Rethinking weight tying: Pseudo-inverse tying for stable lm training and updates, 2026a. URL https://arxiv.org/abs/2602.04556.

Jian Gu, Aldeida Aleti, Chunyang Chen, and Hongyu Zhang. SemRF: A semantic reference frame for residual-stream dynamics in language models, 2026b. URL https://arxiv.org/abs/2606.32022.

Wes Gurnee and Max Tegmark. Language models represent space and time. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=jE8xbmvFin.

Wes Gurnee, Nicholas Sofroniew, Adam Pearce, Mateusz Piotrowski, Isaac Kauvar, Runjin Chen, Anna Soligo, Paul Bogdan, Euan Ong, Rowan Wang, T. Ben Thompson, David Abrahams, Subhash Kantamneni, Emmanuel Ameisen, Joshua Batson, and Jack Lindsey. Verbalizable representations form a global workspace in language models. *Transformer Circuits Thread*, 2026. URL https://transformer-circuits.pub/2026/workspace/index.html.

Robert Huben, Hoagy Cunningham, Logan Riggs Smith, Aidan Ewart, and Lee Sharkey. Sparse autoencoders find highly interpretable features in language models. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=F76bwRSLeK.

IBM Granite Team. Granite 4.2 LLMs: How they're built. Hugging Face blog, 2026. URL https://huggingface.co/blog/ibm-granite/granite-4-2.

Hakan Inan, Khushayr Khosravi, and Richard Socher. Tying word vectors and word classifiers: A loss framework for language modeling. In *International Conference on Learning Representations*, 2017. URL https://openreview.net/forum?id=r1aPbsFle.

Yibo Jiang, Goutham Rajendran, Pradeep Kumar Ravikumar, Bryon Aragam, and Victor Veitch. On the origins of linear representations in large language models. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 21879–21911. PMLR, 2024. URL https://proceedings.mlr.press/v235/jiang24d.html.

Olivier Ledoit and Michael Wolf. Shrinkage estimation of large covariance matrices: Keep it simple, statistician? *Journal of Multivariate Analysis*, 186:104796, 2021. doi: 10.1016/j.jmva.2021.104796.

Alexander H. Liu, Kartik Khandelwal, Sandeep Subramanian, Victor Jouault, Abhinav Rastogi, Adrien Sadé, Alan Jeffares, Albert Jiang, Alexandre Cahill, Alexandre Gavaudan, et al. Ministral 3, 2026. URL https://arxiv.org/abs/2601.08584.

Xuyuan Liu, Lei Hsiung, Yaoqing Yang, and Yujun Yan. Spectral insights into data-oblivious critical layers in large language models. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 4860–4877. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.findings-acl.251. URL https://aclanthology.org/2025.findings-acl.251/.

Antonio Lopardo, Avyukth Harish, Catherine Arnett, and Akshat Gupta. Weight tying biases token embeddings towards the output space. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 40795–40810. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.findings-acl.2027. URL https://aclanthology.org/2026.findings-acl.2027/.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. Pointer sentinel mixture models. In *International Conference on Learning Representations*, 2017. URL https://openreview.net/forum?id=Byj72udxe.

Buddhika Nettasinghe and Geethu Joseph. As language models scale, low-order linear depth dynamics emerge, 2026. URL https://arxiv.org/abs/2603.12541.

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

OpenBMB. MiniCPM5: On-device language models. Model release, 2026. URL https://github.com/OpenBMB/MiniCPM.

Samuel E. Otto and Clarence W. Rowley. Koopman operators for estimation and control of dynamical systems. *Annual Review of Control, Robotics, and Autonomous Systems*, 4:59–87, 2021. doi: 10.1146/annurev-control-071020-010108.

Koyena Pal, Jiuding Sun, Andrew Yuan, Byron C. 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)*, pp. 548–560. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.conll-1.37. URL https://aclanthology.org/2023.conll-1.37/.

Kiho Park, Yo Joong Choe, and Victor Veitch. The linear representation hypothesis and the geometry of large language models. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 39643–39666. PMLR, 2024. URL https://proceedings.mlr.press/v235/park24c.html.

Kiho Park, Yo Joong Choe, Yibo Jiang, and Victor Veitch. The geometry of categorical and hierarchical concepts in large language models. In *International Conference on Learning Representations*, 2025.

Ofir Press and Lior Wolf. Using the output embedding to improve language models. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pp. 157–163. Association for Computational Linguistics, 2017. URL https://aclanthology.org/E17-2025/.

Yi Ren and Danica J. Sutherland. Learning dynamics of LLM finetuning. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=tPNHOoZFl9.

Pedro H. V. Valois, Lincon S. Souza, Erica K. Shimomoto, and Kazuhiro Fukui. Frame representation hypothesis: Multi-token llm interpretability and concept-guided text generation. *Transactions of the Association for Computational Linguistics*, 13:1436–1458, 2025. doi: 10.1162/tacl.a.48.

Hong Wang and Kelly Wang. Complete characterization of gauge symmetries in transformer architectures. In *NeurIPS 2025 Workshop on Symmetry and Geometry in Neural Representations*, 2025. URL https://openreview.net/forum?id=KrkbYbK0cH.

Alexander Yom 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)*, pp. 9615–9625. ELRA and ICCL, 2024. URL https://aclanthology.org/2024.lrec-main.840/.

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, Shashwat Goel, Nathaniel Li, Michael J. Byun, Zifan Wang, Alex Mallen, Steven Basart, Sanmi Koyejo, Dawn Song, Matt Fredrikson, J. Zico Kolter, and Dan Hendrycks. Representation engineering: A top-down approach to ai transparency, 2023. URL https://arxiv.org/abs/2310.01405.

A Technical Details

A.1 Consensus Geometry Construction

States are the final normalization applied to each residual stream without its learned gain, and the gain and any output scale are absorbed into the decoder rows. The lexical interfaces E and O are restricted to the tokenizer-linked lexical support, and the decoder's common-logit direction, which shifts every logit equally and leaves the softmax unchanged, is quotiented out exactly. For a model with tied embeddings the two interfaces share one table but still differ through the absorbed gain and the quotient.

Both Gram operators are normalized to unit determinant, so a global rescaling of either interface changes nothing. G_C is realized through the SVD of the Cholesky cross-factor $L_O^\top L_E$, which avoids squaring the spectral condition number, and the resulting analysis map is factored by a right polar decomposition into the orthogonal gauge U and the positive factor P. When the centered decoder Gram is singular, the construction is restricted to its machine-precision-identifiable quotient and completed orthogonally. No ridge, jitter, whitening, or calibration data enters Module I.

A.2 Observable Dynamics Estimation

One factorization of the source design serves every scale: a thin SVD $X = U_X \Sigma V^\top$ when $N < d$, and otherwise an eigendecomposition of $X^\top X$ with $U_X = XV\Sigma^{-1}$. Both give the same nonzero singular system, so $H_X(\rho) = U_X \operatorname{diag}(\sigma_i^2/(\sigma_i^2+\rho))U_X^\top$ without any rank truncation. The path is parameterized internally by the precision $1/\rho$, which includes the zero-response limit exactly. The only nontrivial identity is the sequence-excluded response. For a sequence block g of m positions and any right-hand side W,

$$[H^- W]_g = W_g - (I - H_{gg})^{-1}\big(W_g - [HW]_g\big),$$

where H_{gg} is the $m \times m$ diagonal block of H_X. One batched Cholesky factor of $I - H_{gg}$ per scale serves all d output coordinates, and the leave-sequence-out backbone subtracts each block's per-coordinate sufficient statistics, so no model is refit per sequence. The deployed backbone and operator use all Q sequences.

The origin is $\rho_0 = \hat{a}\,\|X\|_F^2/d$, where $\hat{a} \in [0,1]$ is the OAS shrinkage intensity computed from the trace invariants of the centered source Gram, evaluated from raw moments so that no State is recentered. Equation (4) is minimized by bounded Brent search over $s \in [-16, 16]$, and the selected risk is asserted not to exceed the risk at the origin $s = 0$. No selected scale reached the bound.

A.3 Information Projection Optimization

Module III dualizes in the execution gauge fixed by Module I, so the production Riesz lift has the identity array representation while remaining a type conversion from covectors to vectors. The covariance of the reciprocal lift ΓM^{-1} under $GL(d)$ is verified separately in the frame experiment.

Equation (5) is solved by projected trust-region BFGS in three dimensions from $(1, 1, 0)$, with the calibration KL asserted not to exceed that of Z_C. Every KL value and directional derivative is accumulated over vocabulary blocks. Model execution is bf16, statistics accumulate in FP32, and spectral factorizations and small solves use FP64.

A.4 Computational Complexity

One fixed reference, one factorization per layer, and a low-dimensional convex readout make Lexical Lens economical in time and space. Table 4 lists the dominant costs, with N calibration positions, N_e held-out positions, width d, vocabulary V, lexical support V_s, and factor width $q = \min(N, d)$.

Time Complexity. The reference is computed once per model at cost $\Theta(V_s d^2 + d^3)$. One factorization per layer, at cost $\Theta(Ndq)$, serves every scale and all Q sequence deletions, so exact leave-sequence-out cross-fitting costs one batched $m \times m$ Cholesky solve per scale instead of Q refits, replacing more than a thousand refactorizations per layer at $Q = 64$. Module II never evaluates the decoder, and Module III is a three-dimensional convex problem that has no learning rate, epoch budget, or stopping rule to tune. The Tuned Lens instead trains $d^2 + d$ parameters per layer by stochastic optimization.

Space Complexity. Fitting stores no gradients or optimizer moments, whereas a Tuned Lens trained with Adam holds four copies of its $d^2 + d$ parameters. Decoder evaluations stream the vocabulary in blocks of size V_b, so no $N \times V$ logit tensor is formed and peak decoder memory is $\Theta(Nd + V_b d)$. The deployed map $\hat{c} = \alpha^*(b + c \odot \operatorname{diag} D) + (cW)V^*$ shares one source carrier W between the two responses, $V^* = \beta^* V_{\text{state}} + \lambda^* V_{\text{grad}}$, and stores $2d + 2dq$ numbers, a saving over an affine translator when $N < d$ and of the same d^2 order otherwise. Semantic agreements need no vocabulary scan, because the vocabulary-wide covariance of two Semantic States is a quadratic form in data-free moments of the Semantic Basis, so they cost $\Theta(N_e d^2)$ instead of $\Theta(N_e V d)$.

Table 4: Principal objects and dominant costs (symbolic and hardware independent).

Stage	Principal object	Shape	Dominant cost
Consensus Geometry	gauge U, factor P	$d \times d$	$\Theta(V_s d^2 + d^3)$ once per model
Source factorization	$U_X \Sigma V^\top$	$N \times q$	$\Theta(Ndq)$ once per layer
State-risk evaluation	$H_X^- E_{\text{st}}$ at one scale	$N \times d$	$\Theta(Nqd + Nmd)$, no refactorization and no decoder
Gradient Residual	$\Gamma = m(Y) - m(Z_C)$	$N \times d$	one streamed decoder scan
Information Projection	(α, β, λ)	3	streamed decoder scans, three directional derivatives each
Held-out metrics	KL, NLL, accuracy	N_e	one streamed decoder scan per layer
Semantic agreement	Semantic State and Dynamics	N_e	$\Theta(N_e d^2)$ from data-free lexical moments
Deployment	one State map	$2d + 2dq$	$\Theta(dq)$ per State

A.5 Experimental Setups

Models. Table 5 lists the ten base models. All are dense decoder-only Transformers, SmolLM2 ties its input and output embeddings, and the Ministral 3 runs use the released base weights.

Data. WikiText-103 (raw, fixed dataset revision), seed 42, sequence length 128, and all 127 eligible causal positions per sequence. Calibration uses the first 64 sequences of a deterministic training-split stream (8,128 positions), and the held-out set is 64 sequences of the test split (8,128 positions), used only for reporting. Each sequence carries total weight $1/Q$ in every calibration objective. The SmolLM2 results come from release v9.91 and the other models from releases v9.81.6 to v9.88.0, whose modules are numerically identical to v9.91.

Metrics. For a predicted State z at a held-out position with terminal State y and observed next token y_{next}, forward KL is $D_{\text{KL}}(p_y \| p_z)$, NLL is $-\log p_z(y_{\text{next}})$, and next-token accuracy is

Table 5: Model suite. "Layers" counts evaluated intermediate layers.

Model	Parameters	Width d	Layers	Vocabulary
SmolLM2 135M	0.135B	576	29	49,152
SmolLM2 360M	0.36B	960	31	49,152
SmolLM2 1.7B	1.7B	2048	23	49,152
MiniCPM5 1B	1.08B	1536	23	130,560
MiniCPM5 2B	2.52B	2048	41	130,560
Granite 4.2 3B	3B	2560	39	100,352
Granite 4.2 8B	8B	4096	39	100,352
Ministral 3 3B	3B	3072	25	131,072
Ministral 3 8B	8B	4096	33	131,072
Ministral 3 14B	14B	5120	39	131,072

$\mathbf{1}[\arg\max_v p_z(v) = y_{\text{next}}]$. All three come from one streamed decoder scan of z. The agreements are Pearson correlations over the vocabulary between $s(Z_{C,\ell})$, or between τ_ℓ, estimated at adjacent budgets, averaged over held-out positions and layers.

Reference Frame. Chart changes are $T = \mathcal{H}\ \mathrm{diag}(\sigma)$ with a fixed Householder reflection $\mathcal{H}$ and a log-uniform spectrum of condition number κ, rescaled to a fixed volume. Identity-reset distortion is the relative change of squared State norms over all calibration rows. Covector-as-vector distortion compares the raw transformed Gradient Residual, computed at the mean source State of each layer, with its correctly transported direction.

Baseline. The Logit Lens applies the frozen final normalization and decoder to each intermediate state. The Tuned Lens uses one identity-initialized full-affine translator per layer followed by the frozen final normalization and decoder, trained under terminal-distribution forward KL with Adam (learning rate 10^{-3}, 128-row batches, 4 to 32 epochs, patience 3, relative tolerance 10^{-4}) on exactly the calibration positions used by LEXICAL LENS. Stopping uses a training-stream plateau only.

A.6 PROOFS

Throughout, States are row vectors, a chart change acts as $h \mapsto hT$ with $T \in GL(d)$, and a State metric M transforms as $M \mapsto T^{-1}MT^{-T}$ so that $hMh^\top$ is invariant.

Consensus Reference Uniqueness. Write $A = G_E \succ 0$, $B = G_O \succ 0$, and $G = TT^\top$. Expanding the square gives $\|ET\|_F^2 = \mathrm{tr}(AG)$, $\|OT^{-T}\|_F^2 = \mathrm{tr}(BG^{-1})$, and a cross term $\mathrm{tr}(EO^\top)$ that does not depend on T, so the paired disagreement equals $f(G) = \mathrm{tr}(AG) + \mathrm{tr}(BG^{-1})$ up to a constant. The first term is linear. For a symmetric direction $H \neq 0$, the second derivative of $t \mapsto \mathrm{tr}(B(G+tH)^{-1})$ at $t = 0$ is $2\|B^{1/2}G^{-1}HG^{-1/2}\|_F^2 > 0$, so f is strictly convex and has at most one minimizer. Because $\mathrm{tr}(AG) \geq \lambda_{\min}(A)\|G\|$ and $\mathrm{tr}(BG^{-1}) \geq \lambda_{\min}(B)\|G^{-1}\|$, f diverges both as G grows and as G approaches the boundary of the cone, so a minimizer exists. Its stationarity condition $A - G^{-1}BG^{-1} = 0$ is the Riccati equation $GAG = B$. With $K = A^{1/2}GA^{1/2}$ it becomes $K^2 = A^{1/2}BA^{1/2}$, whose unique positive-definite root gives $G = A^{-1/2}(A^{1/2}BA^{1/2})^{1/2}A^{-1/2} = A^{-1}\#B$ (Bhatia, 2007).

Consensus Reference Covariance. By Equation (1), $A \mapsto T^\top AT$ and $B \mapsto T^{-1}BT^{-T}$. The matrix $G' = T^{-1}GT^{-T}$ satisfies $G'(T^\top AT)G' = T^{-1}GAGT^{-T} = T^{-1}BT^{-T}$, so by uniqueness it is the transformed solution, which is the transformation law of a State metric. Exchanging A and B turns $GAG = B$ into $G^{-1}BG^{-1} = A$, so the dual problem is solved by G^{-1}.

Riesz Lift Covariance. A lexical covector transforms as $\Gamma \mapsto \Gamma T^{-T}$. Its lift then transforms as $\Gamma T^{-T}(T^{-1}MT^{-T})^{-1} = \Gamma M^{-1}T$, which is the law of a State vector. The raw array transforms with T^{-T} instead, which coincides with T only when T is orthogonal. This is why the covector-as-vector distortion vanishes only for orthogonal chart changes and grows with κ.

Information Projection Convexity. For position i, let $G_i \in \mathbb{R}^{|\mathcal{V}|\times 3}$ collect the logit directions induced by B_i, $R_{\text{state}i}$, and $R_{\text{grad}i}$, and let $\theta = (\alpha, \beta, \lambda)$. The logits are affine in θ and the log-partition function is convex, so $\mathcal{L}(\theta) = \sum_i w_i D_{\text{KL}}(p_{Y,i} \| p_i(\theta))$ is convex with Hessian $\sum_i w_i G_i^\top [\text{diag}(p_i) - p_i p_i^\top] G_i \succeq 0$. The feasible set $\mathbb{R}^2 \times \mathbb{R}_+$ is convex and satisfies Slater's condition, so the KKT conditions are necessary and sufficient: all three gradient components vanish at an interior optimum, and on the face $\lambda = 0$ the first two vanish and $\partial_\lambda \mathcal{L} \geq 0$. The point $(1, 1, 0)$ is feasible and gives $B + H_X E_{\text{st}} = Z_C$, hence $\mathcal{L}(\theta^*) \leq \mathcal{L}(1, 1, 0)$.

B Additional Results and Discussion

B.1 Per-Model Fidelity

Figure 6 places the three fitted lenses of Table 2 side by side. On every model Lexical Lens* already lies well below the Tuned Lens, and Information Projection moves every model further in the same direction. The two steps are of comparable size on Granite 4.2, whereas on every other model most of the reduction is already achieved by Observable Dynamics, most clearly on MiniCPM5 1B with nearly 90% of the gain. Table 6 reports the two secondary metrics. Lexical Lens has the lowest NLL on every model and the highest next-token accuracy on every model. Lexical Lens* improves NLL on every model but not always accuracy, which is the metric most sensitive to the decoder-facing correction added by Information Projection.

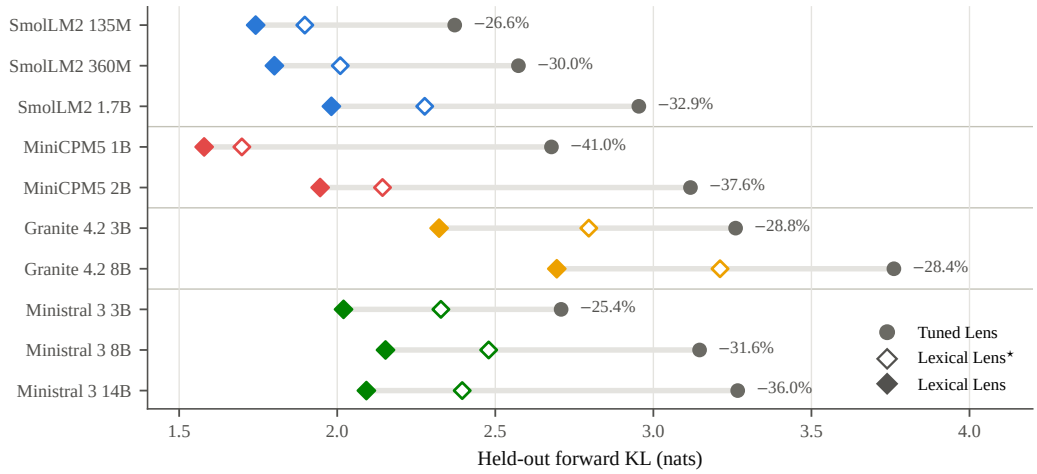

Figure 6: **Per-model held-out forward KL.** Percentages are the reduction of Lexical Lens relative to the Tuned Lens.

B.2 Estimator Structure

Figure 7 reports the four fitted quantities against depth. The selected scale is positive on all but the last layer of SmolLM2 135M and SmolLM2 360M, so the State-risk criterion almost always prefers stronger shrinkage than the closed-form origin, and it decreases with depth as the source becomes more informative about the terminal State; it never reaches the bound of the search interval. The backbone coefficient stays close to one, between 0.95 and 1.04 on average, so the affine-diagonal persistence field is already well calibrated, and the State Response is mildly amplified, between 1.05 and 1.25 on average, with a peak in the middle layers. The Gradient-Response coefficient is strictly positive on all 322 layers and varies by more than an order of magnitude across families, from 0.15 on SmolLM2 135M to 7.8 on Ministral 3 3B on average. Its profile is nearly flat in depth within each family, which indicates that the size of the lexical correction is mainly a property of the decoder geometry of a family rather than of a particular layer.

Table 6: Per-model held-out NLL and next-token accuracy.

	NLL↓				Next-token accuracy↑			
Model	Logit Lens	Tuned Lens	LEXICAL LENS*	LEXICAL LENS	Logit Lens	Tuned Lens	LEXICAL LENS*	LEXICAL LENS
SmolLM2 135M	19.04	6.040	5.711	**5.532**	6.0%	16.8%	15.8%	**16.9%**
SmolLM2 360M	14.47	6.034	5.544	**5.307**	5.2%	17.1%	17.1%	**18.7%**
SmolLM2 1.7B	23.26	6.064	5.470	**5.156**	3.7%	18.4%	18.3%	**20.3%**
MiniCPM5 1B	10.87	6.644	5.748	**5.596**	6.8%	14.5%	15.7%	**17.9%**
MiniCPM5 2B	10.77	6.388	5.510	**5.323**	8.6%	17.3%	19.3%	**21.0%**
Granite 4.2 3B	16.29	6.445	5.965	**5.488**	3.6%	16.7%	16.5%	**19.0%**
Granite 4.2 8B	18.65	6.571	6.012	**5.528**	3.2%	16.5%	17.3%	**19.4%**
Ministral 3 3B	10.63	5.647	5.294	**4.981**	4.8%	19.3%	19.0%	**21.7%**
Ministral 3 8B	10.39	5.886	5.268	**4.946**	5.7%	19.4%	20.0%	**22.4%**
Ministral 3 14B	10.33	5.927	5.091	**4.802**	7.5%	20.8%	21.8%	**23.8%**

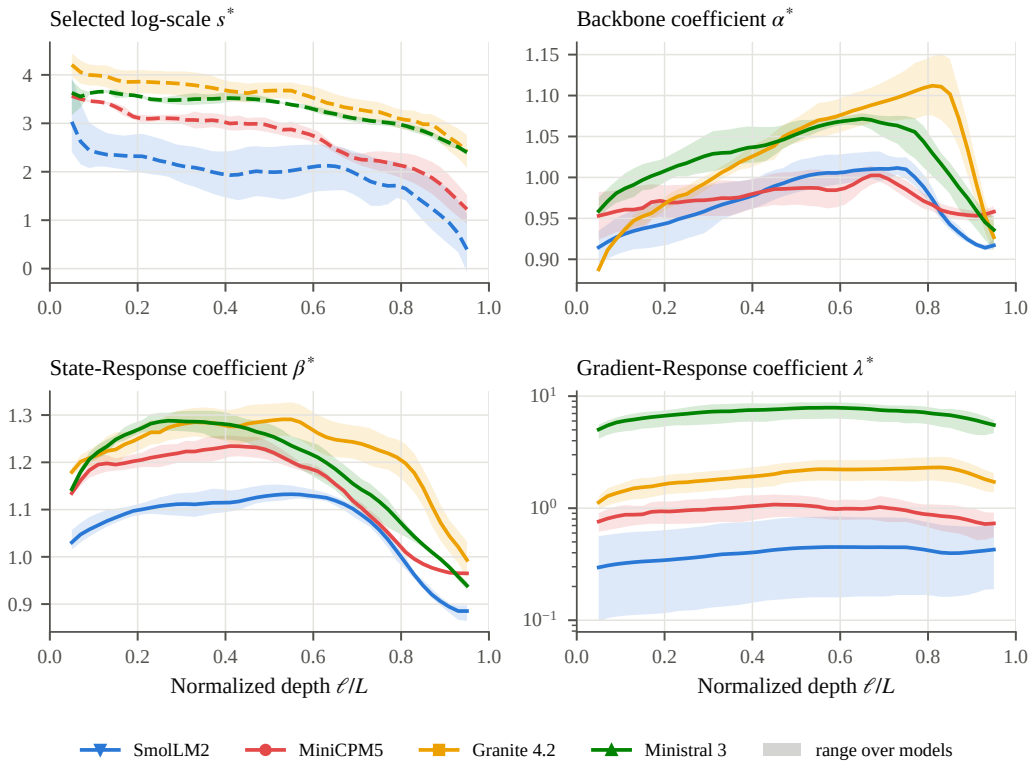

Figure 7: **Fitted estimator structure against depth.** The scale s^* belongs to Observable Dynamics (dashed) and the three coefficients to Information Projection (solid, λ^* on a log scale).

B.3 FITTING COST

Table 7 reports the measured fitting effort per layer. The scale search of Observable Dynamics needs about 21 decoder-free State-risk evaluations and Information Projection about 22 exact-KL evaluations, with no learning rate, epoch budget, or stopping rule to tune. The Tuned Lens averages 16.3 epochs and 1043 optimizer steps per layer, and 20 of 322 layers reach the 32-epoch cap without a training plateau, so its fitting cost is both larger and less predictable.

Table 7: Measured fitting effort per layer. State risk and Projection count evaluations of Equations (4) and (5); Plateau is the fraction of Tuned Lens layers that reach the training plateau within 32 epochs.

	Lexical Lens		Tuned Lens		
Model	State risk	Projection	Epochs	Steps	Plateau
SmolLM2 135M	23.2	17.2	23.2	1485	72%
SmolLM2 360M	21.8	16.9	18.0	1152	81%
SmolLM2 1.7B	21.3	16.6	20.9	1338	91%
MiniCPM5 1B	20.1	20.9	12.5	801	100%
MiniCPM5 2B	19.6	20.1	12.8	821	100%
Granite 4.2 3B	19.8	18.7	14.9	957	97%
Granite 4.2 8B	20.7	22.8	14.8	947	97%
Ministral 3 3B	19.8	28.2	16.4	1050	96%
Ministral 3 8B	20.2	27.8	15.6	1001	97%
Ministral 3 14B	20.1	25.9	13.7	875	100%
Mean	20.7	21.5	16.3	1043	93%

B.4 Semantic Convergence

Figure 8 resolves the two agreement measures by family. Semantic State Agreement is highest on Granite 4.2 and Ministral 3 and rises most at the largest budget on SmolLM2 and MiniCPM5. Semantic Dynamics Agreement is lowest on SmolLM2, where it varies most across models, and highest on Granite 4.2; on seven of ten models it dips slightly from $8 \to 16$ to $16 \to 32$ before rising at $32 \to 64$, so transitions benefit most from the largest budgets.

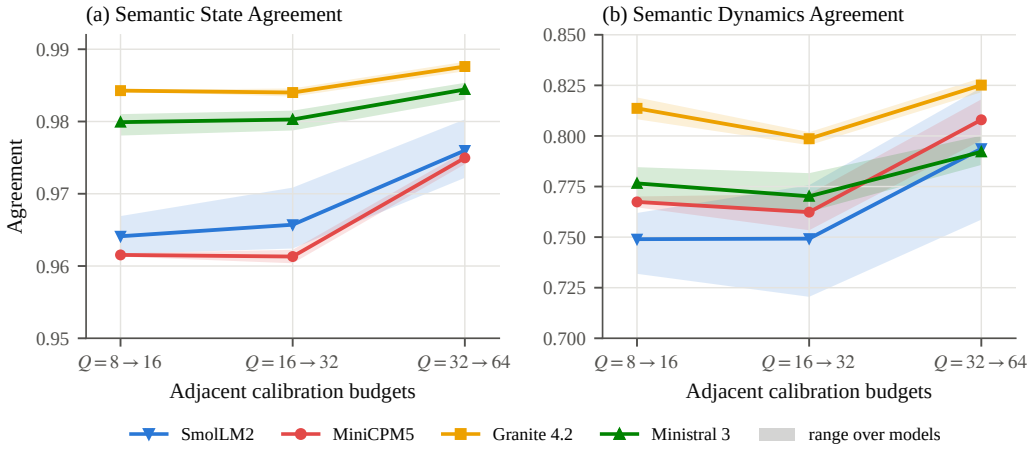

Figure 8: **Semantic convergence by family.** (a) Semantic State Agreement and (b) Semantic Dynamics Agreement between adjacent calibration budgets.

B.5 Discussion

A fixed semantic reference is particularly useful for large language models. Increasing width and depth adds internal coordinates, but it does not make those axes semantically canonical, and the frame experiment shows that silently treating them as orthonormal changes norms and gradient directions by amounts that grow with the non-orthogonality of a function-preserving reparameterization. Consensus Geometry instead derives one model-wide semantic reference from the frozen lexical interfaces and keeps it fixed across depth. It is not a better choice of coordinates but the object that gives lengths,

angles, and gradient directions a meaning; native coordinates remain a valid chart when the transported metric is carried with them.

Within this reference, model-inherent semantics takes a concrete operational form. Centered lexical covectors define token observables directly in the fixed frame and support comparisons, signed lexical contrasts, Semantic Logits, and depthwise Semantic Transitions without training a separate semantic encoder, while the reference types every quantity so that State residuals and lexical gradients meet only through the Riesz lift. Observable Dynamics asks the narrower question of which part of the terminal State a source layer predictably supports and answers it without the decoder, and this answer accounts for most of the predictive gain (Section 4.3). Information Projection smooths the remaining decoder-facing residual with the same operator and fits only three amplitudes, so each gain can be attributed either to source-supported dynamics in a fixed reference or to a calibration of fields fixed beforehand.

Using the same reference across layers also has practical value. The same semantic object can be inspected at different depths without redefining its coordinates, the closed-form dynamics estimator and the decoder-facing projection remain compatible by construction, and the compiled map keeps a diagonal-plus-factorized form instead of a dense translator learned by stochastic optimization. The scope is deliberate. Model-inherent semantics is defined by the lexical interfaces of one frozen model rather than a universal ontology, Observable Dynamics identifies predictable terminal quantities rather than causally necessary components, and the evidence covers dense decoder-only models on English text with 128-token contexts. This distinction keeps the semantic reference, predictive dynamics, and decoder-facing objective separate.