Probing the Formation and Dynamics of J-Space in Language Models

Andrew Kiruluta

UC Berkeley, School of Information

July 2026

Abstract

The Jacobian lens (J-lens) identifies residual-stream directions according to their average first-order effect on a language model's present and future verbal outputs. Recent work argues that sparse combinations of these directions form a functionally privileged representational format, termed J-space, that supports report, causal control, flexible reuse, and parts of multi-step reasoning. The original study also makes clear that the current instrument is incomplete: it is primarily first-order, token-indexed, pointwise in layer and position, and only sparsely studied across training. This paper develops *J-Space Dynamics* (JSD), a methods proposal for extending that program while separating established results from new hypotheses. JSD contributes five linked constructions. First, it replaces an informal curvature-as-commitment claim with a target-versus-competitor continuation margin and a subspace-restricted Hessian that yields an estimable local robustness radius. Second, it defines a path-conditioned sequence lens and binding diagnostics for ordered multi-token content; this is positioned as complementary to, rather than the first instance of, the template and oracle multi-token extensions already proposed with the J-lens. Third, it models workspace-code evolution using a non-autonomous family of transfer operators across depth and position, with finite-window persistence, closure error, and causal ignition tests. Fourth, it places the workspace code in a regularized pullback Fisher geometry. This component builds directly on prior pullback-Fisher work for activation steering; its proposed novelty is the restriction to semantically indexed J-coordinates and the analysis of workspace trajectories, not the Fisher metric itself. Fifth, it introduces a multivariate consolidation profile and a secondary null-calibrated Workspace Consolidation Index for checkpoint studies and specifies change-point tests that distinguish genuine representational reorganization from artifacts of discontinuous behavioral metrics. The paper is therefore best read as a falsifiable experimental framework, not as evidence that a new dynamical law, phase transition, or form of consciousness has already been established. Relevant implementation repo is located here: https://github.com/andrew-jeremy/J-Space-in-Neural-Networks

1 Introduction

Mechanistic interpretability needs observables that are neither as coarse as loss curves and benchmark scores nor as unconstrained as individual neurons, weights, or raw activation coordinates. Logit lenses, tuned lenses, sparse autoencoders (SAEs), causal tracing, activation patching, and steering vectors each provide useful intermediate descriptions, but they answer different questions. Some reveal what can be decoded from a representation; others identify directions that change behavior; still others isolate features that reconstruct activations. None of those properties alone establishes that a representation is broadly available to the model's own downstream computations.

The Jacobian lens (J-lens) was introduced to target that access question directly. For each layer, it averages the input–output Jacobian from an intermediate residual-stream state to current and future final-layer states, then composes this transport with the model's normal output decoding. The resulting token-indexed directions characterize what an activation is, on average across contexts, disposed to make the model verbalize. Sparse nonnegative combinations of these directions define J-space. In the original experiments, J-space accounted for only a small fraction of activation variance yet was disproportionately important for report, flexible reasoning, representation swapping, and broad read/write connectivity across model components [1].

Those findings are substantial, but the source paper is explicit about the current method's limitations. The lens is only an approximate description of a possible workspace; it is token-indexed; some readouts are not interpretable; the mechanism that populates J-space is unknown; and pretraining-time formation has not been characterized beyond limited base/post-training comparisons [1]. The central opportunity is therefore not to restate the global-workspace analogy more strongly, but to convert these acknowledged limitations into precise estimands, controls, and falsification criteria.

This paper proposes *J-Space Dynamics* (JSD), a coordinated extension of the J-lens program across depth, token position, and training time. The name is chosen to avoid collision with the established "J-flow" of Kähler geometry [23]. JSD does not assume a stationary continuous-time generator: transformer depth is modeled as a discrete, layer-indexed, non-autonomous sequence.

1.1 Novelty assessment and claim boundaries

The strongest contributions of the present proposal are methodological integration and longitudinal design. Several ingredients have important precedents and should not be claimed as inventions:

1. **Multi-token extensions already exist in the originating work.** The J-lens paper proposes a template lens for predefined phrases and an oracle lens for phrase discovery [1]. JSD's narrower contribution is a path-conditioned sequence functional, an order-sensitivity statistic, and a causal binding-swap protocol.

2. **Pullback Fisher geometry for transformer activations already exists.** FishBack derives the intermediate-layer pullback Fisher metric and uses it for minimum-distortion activation steering [2]. JSD uses that geometry inside a J-space coordinate system and proposes trajectory-level diagnostics; it does not claim priority for the metric.

3. **Koopman and EDMD methods are established.** The novelty claim is their use on sparse, semantically aligned workspace codes with non-autonomous layer dependence and explicit closure tests, not the transfer-operator machinery itself [16, 17].

4. **Abrupt "emergence" cannot be assumed from benchmark curves.** Discontinuous metrics can manufacture apparently sharp capability transitions [3]. The longitudinal component therefore treats phase-transition language as a hypothesis requiring change-point evidence in continuous internal and behavioral observables.

With these boundaries, JSD makes five proposed contributions:

1. a *second-order continuation-margin diagnostic* that measures local robustness of a planned token or span against specified competitors;

2. a *path-conditioned sequence lens* for ordered spans, plus binding residuals and order-sensitivity tests;

3. a *non-autonomous transfer-operator model* of workspace-code evolution across layer and position, with finite-time persistence rather than unjustified stationary spectra;

4. a *J-restricted pullback Fisher geometry* for calibrated interventions and time-dependent workspace path costs;

5. a *primary multivariate consolidation profile and secondary null-calibrated Workspace Consolidation Index* and preregistered checkpoint analysis for testing whether representational reorganization precedes capability changes.

The framework's overarching hypothesis is deliberately weaker than the claim that a workspace is "the principal product" of training:

> *Some capabilities that require flexible routing of intermediate information may depend on the formation of a sparse, semantically accessible, and broadly connected representational format. If so, measurable changes in that format should predict at least a subset of capability gains across training.*

This hypothesis is useful because it can fail. Capabilities may improve while all JSD observables remain smooth or unchanged; the J-code may not close dynamically; apparent commitment may not predict intervention resistance; and compositional content may be better captured outside J-space.

2 Background and Formal Setup

Consider a decoder-only transformer with L layers, residual width d, vocabulary $\mathcal{V}$ of size V, and residual-stream states $\boldsymbol{h}_i^{(\ell)} \in \mathbb{R}^d$ at layer ℓ and token position i. Let N denote the final normalization and $U \in \mathbb{R}^{V\times d}$ the unembedding matrix.

2.1 Canonical corpus-averaged J-lens

The originating work defines a hidden-to-hidden transport matrix for each source layer by averaging Jacobians over prompts, source positions, and present or future target positions:

$$J_\ell^{\mathrm{src}} = \mathbb{E}_{x,\, i,\, k\geq i}\left[\frac{\partial \boldsymbol{h}_k^{(L)}}{\partial \boldsymbol{h}_i^{(\ell)}}\right] \in \mathbb{R}^{d\times d}. \tag{1}$$

The canonical J-lens readout replaces the downstream network by this average linear transport followed by the model's normal output map:

$$\mathrm{Lens}_\ell(\boldsymbol{h}) = \mathrm{softmax}(UN(J_\ell^{\mathrm{src}}\boldsymbol{h}))\,. \tag{2}$$

When normalization is suppressed or locally linearized, the token-indexed J-lens vectors are the rows of UJ_ℓ^{src}. With normalization retained explicitly, the local differential direction for token t at activation $\boldsymbol{h}$ is

$$\boldsymbol{j}_t^{\mathrm{src},(\ell)}(\boldsymbol{h}) = (J_\ell^{\mathrm{src}})^\top J_N(J_\ell^{\mathrm{src}}\boldsymbol{h})^\top U^\top \boldsymbol{e}_t, \tag{3}$$

where J_N is the Jacobian of the final normalization and $\boldsymbol{e}_t$ is the vocabulary basis vector. Equations (1)–(2) are the established source method; the constructions below are extensions rather than restatements of it [1].

2.2 Context-local output-functional extension

For second-order, span-level, and intervention analyses, JSD additionally uses a context-local differentiable output functional. For a source state $\boldsymbol{h}_i^{(\ell)}$, define the discounted future-token score

$$\widetilde{\Phi}_t^{(\ell,i)}(\boldsymbol{h}) = \sum_{k \geq i} \gamma^{k-i}\, z_{k,t}(\boldsymbol{h}; x_{\leq k}), \qquad \gamma \in (0,1], \tag{4}$$

where $\boldsymbol{z}_k = UN(\boldsymbol{h}_k^{(L)})$ and all prompt inputs are fixed except for causal effects induced by replacing the source state with $\boldsymbol{h}$. The corresponding local continuation-gradient direction is

$$\widetilde{\boldsymbol{J}}_t^{(\ell,i)} = \nabla_{\boldsymbol{h}} \widetilde{\Phi}_t^{(\ell,i)}(\boldsymbol{h})\big|_{\boldsymbol{h}=\boldsymbol{h}_i^{(\ell)}}. \tag{5}$$

A corpus average of Equation (5) is a useful variant, but it is not generally identical to Equation (1), because averaging, normalization, output selection, and differentiation need not commute. Every experiment must state whether it uses canonical source vectors, local continuation gradients, template/oracle directions, or a combined dictionary.

Let $D_\ell = [\boldsymbol{d}_1^{(\ell)}, \ldots, \boldsymbol{d}_q^{(\ell)}] \in \mathbb{R}^{d \times q}$ denote the selected and preregistered layer-ℓ dictionary after this choice has been made.

Definition 1 (Sparse J-space model class)**.** *For sparsity budget s, define*

$$\mathcal{W}_{\ell,s} = \{D_\ell \boldsymbol{a} : \boldsymbol{a} \geq 0,\ \|\boldsymbol{a}\|_0 \leq s\}. \tag{6}$$

This is a union of sparse polyhedral cones, not generally a linear subspace. The hard-sparse estimator

$$\widehat{\boldsymbol{a}}_i^{(\ell)} \in \arg\min_{\boldsymbol{a} \geq 0,\, \|\boldsymbol{a}\|_0 \leq s} \left\| \boldsymbol{h}_i^{(\ell)} - D_\ell \boldsymbol{a} \right\|_2^2 \tag{7}$$

is a conceptual definition; practical experiments may use nonnegative orthogonal matching pursuit or the stabilized convex surrogate below.

2.3 Coordinate normalization, identifiability, and uncertainty

Sparse coefficient trajectories are scientifically meaningful only if coding instability is separated from model dynamics. Before fitting codes, dictionary columns must follow a fixed scale convention, such as unit Euclidean norm, unit null-standardized readout variance, or unit regularized Fisher norm. Write $\widetilde{D}_\ell = D_\ell S_\ell^{-1}$ for the normalized dictionary, where S_ℓ is diagonal and the same convention is used at every checkpoint.

A stable practical code is

$$\widehat{\boldsymbol{a}}_i^{(\ell)} \in \arg\min_{\boldsymbol{a} \geq 0} \left\| \boldsymbol{h}_i^{(\ell)} - \widetilde{D}_\ell \boldsymbol{a} \right\|_2^2 + \lambda_1 \|\boldsymbol{a}\|_1 + \lambda_2 \|\boldsymbol{a}\|_2^2, \qquad \lambda_2 > 0, \tag{8}$$

with hard-sparse refits retained as a sensitivity analysis. Each study must report dictionary coherence

$$\mu(\widetilde{D}_\ell) = \max_{r \neq s} \left| \widetilde{\boldsymbol{d}}_r^{(\ell)\top} \widetilde{\boldsymbol{d}}_s^{(\ell)} \right|, \tag{9}$$

active-Gram conditioning, bootstrap coefficient intervals, support-selection frequency, and reconstruction error. Cross-layer comparisons use common semantic labels but not raw coefficient magnitudes: coefficients should be null-standardized or Fisher-normalized per layer. Dynamics, ignition, and

checkpoint analyses should be repeated over bootstrap code samples so that uncertainty in $\widehat{\boldsymbol{a}}$ propagates into operator and change-point uncertainty. An apparent transition that disappears after this propagation is coding jitter, not evidence of workspace reorganization.

Three distinctions will be maintained throughout. First, the canonical corpus-averaged map is global, whereas continuation gradients and robustness diagnostics are local to a prompt and activation. Second, J-space is a sparse model class rather than a privileged Euclidean subspace. Third, semantic labels align coordinates across layers, while normalization and uncertainty analysis are required to make their magnitudes comparable.

3 Design Principles

A credible extension of J-space should satisfy four requirements.

Incremental validity. Every new quantity must add predictive or causal value over the original J-lens, the source paper's template/oracle lens, SAE features, and simple output-distribution baselines.

Local mathematical meaning. A Hessian, metric, or spectrum must be tied to a stated functional. Curvature of a logit is not automatically a dynamical restoring force, and an eigenvalue of a pooled layer map is not automatically a physical mode.

Estimator diagnostics. Each reported result should include variance, approximation error, conditioning, and null controls. For transfer operators, the first question is closure error; for Fisher geometry, rank and regularization sensitivity; for sequence lenses, lexicon and paraphrase sensitivity.

Causal confirmation. Readout correlations are insufficient for claims about workspace function. Proposed mechanisms should be tested with matched-norm perturbations, coordinate swaps, component ablations, and held-out prompt families.

4 Component I: Second-Order Continuation Margins

4.1 Why raw curvature is not commitment

The Hessian of a future token logit describes local curvature of that score as a function of an activation. It does not, by itself, imply that the activation lies in a restoring basin: transformer activations do not evolve by gradient ascent on that token's logit. A more defensible notion of commitment is *robust preference*: how large a workspace-restricted perturbation is required before a planned continuation loses to specified alternatives.

Let y be a target continuation, which may be a token or span, and let $\mathcal{C}(y)$ be a preregistered competitor set. Define a finite-horizon continuation score

$$\psi_y(\boldsymbol{h}) = \log p_\theta(y \mid x_{\le i}; \boldsymbol{h}_i^{(\ell)} \leftarrow \boldsymbol{h}), \tag{10}$$

and the smooth target-versus-competitor margin

$$M_y(\boldsymbol{h}) = \psi_y(\boldsymbol{h}) - \log \sum_{c \in \mathcal{C}(y)} \exp \psi_c(\boldsymbol{h}). \tag{11}$$

A positive margin means that y is preferred to the competitor aggregate.

Let $P \in \mathbb{R}^{d\times m}$ have orthonormal columns spanning the currently active J-directions or a preregistered candidate set. Define the reduced gradient and Hessian

$$\boldsymbol{g}_y = P^\top \nabla M_y(\boldsymbol{h}), \qquad H_y = P^\top \nabla^2 M_y(\boldsymbol{h}) P. \tag{12}$$

Hessian–vector products permit estimation without materializing the full $d \times d$ Hessian [15].

4.2 An empirical local robustness radius

Let $G_y \succ 0$ be a fixed local intervention metric on the selected coordinate space; the recommended choice is the regularized J-restricted pullback Fisher metric developed in Section 7. Define

$$\|\boldsymbol{u}\|_{G_y} = \sqrt{\boldsymbol{u}^\top G_y \boldsymbol{u}}, \qquad \|\boldsymbol{g}_y\|_{G_y^{-1}} = \sqrt{\boldsymbol{g}_y^\top G_y^{-1} \boldsymbol{g}_y}. \tag{13}$$

Suppose an empirical curvature envelope $B_{G,y}(r)$ satisfies

$$\sup_{\|\boldsymbol{u}\|_{G_y}\le r} \left\| G_y^{-1/2} P^\top \nabla^2 M_y(\boldsymbol{h} + P\boldsymbol{u}) P G_y^{-1/2} \right\|_{\mathrm{op}} \le B_{G,y}(r). \tag{14}$$

Taylor's theorem then motivates the local lower approximation

$$M_y(\boldsymbol{h} + P\boldsymbol{u}) \gtrsim M_y(\boldsymbol{h}) - r\,\|\boldsymbol{g}_y\|_{G_y^{-1}} - \frac{1}{2} B_{G,y} r^2. \tag{15}$$

Using a fixed estimated envelope $\widehat{B}_{G,y} > 0$, define the empirical Taylor robustness radius

$$\widehat{r}_{\mathrm{rob}}(y) = \frac{-\|\boldsymbol{g}_y\|_{G_y^{-1}} + \sqrt{\|\boldsymbol{g}_y\|_{G_y^{-1}}^2 + 2\widehat{B}_{G,y} M_y(\boldsymbol{h})}}{\widehat{B}_{G,y}}, \qquad M_y(\boldsymbol{h}) > 0. \tag{16}$$

When $\widehat{B}_{G,y} \to 0$, this reduces to $M_y / \|\boldsymbol{g}_y\|_{G_y^{-1}}$. Equation (16) is not called a certificate unless Equation (14) is verified over the entire perturbation ball. With sampled Hessian–vector products it is an empirical, local predictor whose validity must be checked by adversarial search.

Prediction 1 (Plan robustness)**.** *In a rhyme-planning task, the selected rhyme span should have a larger continuation margin and empirical robustness radius than plausible distractor rhymes. Across prompts, the minimum Fisher-calibrated intervention needed to change the planned rhyme should be monotonically related to* $\widehat{r}_{\mathrm{rob}}$*, after controlling for baseline output probability, entropy, and first-order gradient norm.*

Prediction 2 (Concealed but stable content)**.** *In matched settings where a model recognizes a fact but is instructed not to report it, a target-specific continuation or action margin may remain robust at intermediate layers while direct token readout decreases near output layers. This pattern should be compared against absence controls and against non-J-space probes. Failure to improve classification or causal prediction over first-order readout falsifies the proposed benefit of the second-order diagnostic.*

Required controls. Competitor sets must be fixed before evaluation; results should be repeated with paraphrased competitors; curvature estimates should report Lanczos or Hutchinson uncertainty; and robustness should be tested with actual constrained adversarial searches. The metric G_y, its regularization, and the output horizon must be fixed before comparison. The empirical radius must outperform first-order margin, dual gradient norm, output entropy, and unconstrained local Lipschitz baselines.

5 Component II: Path-Conditioned Sequence Lenses and Binding

5.1 Relationship to existing multi-token extensions

The originating J-lens work already describes a template lens for predefined words or phrases and an oracle lens that proposes phrase-level descriptions [1]. JSD therefore does not claim that phrase-level J-space vectors are new in general. Its proposed contribution is to use the model's autoregressive probability of an ordered span as the differentiable target and to test whether order and relational binding add information beyond token and template baselines.

For an ordered span $w = (w_1, \ldots, w_n)$ beginning at target position k, define

$$\Phi_{w,k}(\boldsymbol{h}) = \sum_{r=1}^{n} \log p_\theta\Big(w_r \mid x_{<k}, w_{<r}; \boldsymbol{h}_i^{(\ell)} \leftarrow \boldsymbol{h}\Big). \tag{17}$$

The corresponding local path-conditioned sequence direction is

$$\boldsymbol{j}_{w,k}^{\text{path}} = \nabla_{\boldsymbol{h}} \Phi_{w,k}(\boldsymbol{h})\big|_{\boldsymbol{h}=\boldsymbol{h}_i^{(\ell)}}. \tag{18}$$

The gradient is evaluated under teacher forcing along the ordered span. Each term is conditioned on $w_{<r}$, so reversing the span generally changes the direction even when the token multiset is unchanged.

A corpus-averaged path direction can be obtained by averaging over prompts and admissible start positions. For causal experiments, however, the local direction in Equation (18) is preferable because it preserves prompt-specific binding.

5.2 Binding and order diagnostics

Define the constituent span

$$\mathcal{S}(w) = \text{span}\{\boldsymbol{j}_{w_r}^{(\ell)} : r = 1, \ldots, n\}. \tag{19}$$

The *binding residual* is

$$\beta(w) = \frac{\left\| \boldsymbol{j}_{w,k}^{\text{path}} - \Pi_{\mathcal{S}(w)} \boldsymbol{j}_{w,k}^{\text{path}} \right\|_2}{\left\| \boldsymbol{j}_{w,k}^{\text{path}} \right\|_2 + \epsilon}. \tag{20}$$

Because a large residual can also arise from estimation noise or context mismatch, $\beta(w)$ must be compared with paraphrase, shuffled-order, random-span, and matched-frequency controls.

For two spans containing the same tokens in different orders, define

$$\Delta_{\text{ord}}(w, w') = 1 - \frac{\left\langle \boldsymbol{j}_{w,k}^{\text{path}}, \boldsymbol{j}_{w',k}^{\text{path}} \right\rangle}{\left\| \boldsymbol{j}_{w,k}^{\text{path}} \right\|_2 \left\| \boldsymbol{j}_{w',k}^{\text{path}} \right\|_2 + \epsilon}. \tag{21}$$

This statistic directly tests whether the lens distinguishes "Alice trusts Bob" from "Bob trusts Alice."

For matched relational alternatives w and w' containing the same entities and lexical material, define the contrastive relational direction

$$\Delta \boldsymbol{j}_{\text{rel}}(w, w') = \boldsymbol{j}_{w,k}^{\text{path}} - \boldsymbol{j}_{w',k}^{\text{path}}. \tag{22}$$

The primary binding intervention moves an activation along Equation (22), or swaps its signed coordinate, while explicitly preserving projections onto constituent entity directions. This contrast

is more diagnostic than a large residual alone: $\beta(w)$ can also reflect tokenization, context specificity, autoregressive conditioning, or dictionary mismatch. Experiments should therefore treat β as descriptive and the selective relational swap as the main causal test. Token length, frequency, and tokenizer segmentation must be matched across contrasts.

Prediction 3 (Binding advantage)**.** *In cross-assignment tasks with identical active entities and attributes but different relations, path-conditioned sequence scores should predict the model's answer beyond unigram J-codes, template-lens scores, and SAE probes. A matched-norm intervention that exchanges two relational span coordinates while preserving constituent unigram coordinates should preferentially exchange the answer. Failure of either incremental prediction or causal selectivity rejects the binding claim.*

Lexicon design. Candidate spans may come from preregistered task hypotheses, high-PMI corpus phrases, or SAE-assisted proposals. Discovery and confirmation sets must be separated. The oracle lens can be used as a proposal mechanism, but the final binding tests should be evaluated on held-out prompts and spans to avoid circularity.

6 Component III: Non-Autonomous Workspace Dynamics

6.1 Common semantic coordinates

The raw J-directions vary by layer, so hidden-state vectors cannot be treated as points in one fixed coordinate system without transport. Token and phrase labels provide a common semantic index, but code magnitudes are comparable only after the normalization and uncertainty procedures of Section 2.3. Select a manageable dictionary of q concepts, fixed before the main analysis, and estimate

$$\boldsymbol{a}_i^{(\ell)} \in \mathbb{R}^q_{\geq 0} \tag{23}$$

using the layer-specific normalized dictionary $\widetilde{D}_{\ell,q}$. Null-standardized coefficients form a common semantic coordinate system, while bootstrap samples quantify uncertainty due to coherent or overcomplete dictionaries. All dynamical fits are repeated across bootstrap code realizations; operator uncertainty that is dominated by code instability is reported as a failure of the coordinate model.

6.2 Depth and position maps

Transformer depth is discrete and generally non-stationary: each layer has different parameters. The appropriate starting point is therefore a family of layer-indexed maps rather than one stationary Koopman operator. Let $g : \mathbb{R}^q \to \mathbb{R}^r$ be a fixed observable dictionary containing selected coefficients, interactions, and possibly a constant term. Estimate a regularized EDMD operator for depth,

$$g\left(\boldsymbol{a}_i^{(\ell+1)}\right) \approx K_\ell^{\text{depth}} g\left(\boldsymbol{a}_i^{(\ell)}\right). \tag{24}$$

A one-step position model,

$$g\left(\boldsymbol{a}_{i+1}^{(\ell)}\right) \approx K_\ell^{\text{pos}} g\left(\boldsymbol{a}_i^{(\ell)}\right), \tag{25}$$

is only a baseline because a causal transformer can use the full prefix. The primary position analysis compares it with the delay-embedded state

$$\boldsymbol{s}_i^{(\ell)} = \left[\boldsymbol{a}_{i-w+1}^{(\ell)\top}, \ldots, \boldsymbol{a}_i^{(\ell)\top}\right]^\top \tag{26}$$

and the controlled model

$$g(\boldsymbol{s}_{i+1}^{(\ell)}) \approx K_\ell^{\text{delay}} g(\boldsymbol{s}_i^{(\ell)}) + B_\ell \boldsymbol{c}_i^{(\ell)}, \tag{27}$$

where $\boldsymbol{c}_i^{(\ell)}$ is a preregistered summary of attention-accessible context, such as selected head outputs or prefix statistics. The delay window w and control variables are selected on validation data and frozen before testing.

The first reported quantity is held-out closure error,

$$\mathcal{E}_{\text{close}} = \frac{\mathbb{E}\,\|g(\boldsymbol{a}') - Kg(\boldsymbol{a})\|_2^2}{\mathbb{E}\,\|g(\boldsymbol{a}') - \mathbb{E}g(\boldsymbol{a}')\|_2^2}, \tag{28}$$

compared with linear autoregression, nonlinear regression, shuffled pairs, delay-only models, controlled models, and models augmented with non-J-space state. If closure is poor, the J-code is not a sufficient state description for the proposed dynamics. Improvement after adding controls indicates that position evolution is driven by omitted prefix information rather than by an autonomous workspace law.

6.3 Finite-window persistence

For a depth window $[\ell_0, \ell_1]$, define the propagator

$$\mathcal{P}_{\ell_1:\ell_0} = K_{\ell_1-1}^{\text{depth}} \cdots K_{\ell_0}^{\text{depth}}. \tag{29}$$

Raw singular values of $\mathcal{P}_{\ell_1:\ell_0}$ depend on arbitrary observable scaling. Let

$$\Sigma_\ell = \text{Cov}\Big[g(\boldsymbol{a}^{(\ell)})\Big] + \eta I, \qquad \eta > 0, \tag{30}$$

and define the covariance-whitened layer operator

$$\widetilde{K}_\ell^{\text{depth}} = \Sigma_{\ell+1}^{-1/2} K_\ell^{\text{depth}} \Sigma_\ell^{1/2}. \tag{31}$$

The metric-aware finite-window propagator is

$$\widetilde{\mathcal{P}}_{\ell_1:\ell_0} = \widetilde{K}_{\ell_1-1}^{\text{depth}} \cdots \widetilde{K}_{\ell_0}^{\text{depth}}. \tag{32}$$

Its singular values quantify persistence in standardized observable coordinates and are invariant to diagonal rescaling of the original code. A Fisher-whitened version, obtained by replacing Σ_ℓ with the regularized metric of Section 7, should be reported as a sensitivity analysis. Directions with singular values near one are approximately maintained over the window; rapidly contracting directions are transient. Complex eigenmodes may be interpreted only in layer ranges where stationarity tests justify pooling.

Definition 2 (Operational ignition)**.** *For concept coordinate a_c, an ignition event occurs when three preregistered conditions are met: (i) a_c crosses a null-calibrated activation threshold; (ii) its projected contribution to a persistent finite-window subspace exceeds a threshold; and (iii) a causal intervention on that coordinate changes a downstream task variable. Conditions (i)–(ii) define a candidate event; condition (iii) validates functional admission.*

The *gating model* predicts candidate ignition from upstream variables such as SAE features, attention-head outputs, MLP activations, and competing J-coordinates. Causal head or MLP ablations are then used to distinguish predictors from mechanisms.

Prediction 4 (Capacity competition)**.** *When unrelated concepts are injected at matched Fisher norm, increasing set size should eventually reduce persistent occupancy. A strong winner-take-most account predicts active quenching: losing coordinates decay faster than matched no-competition controls. A simple bottleneck account predicts only dilution. These alternatives should be distinguished rather than treating any capacity effect as "ignition."*

Prediction 5 (Internal/external scratchpad trade-off)**.** *For tasks solved both silently and with an explicit scratchpad, the explicit condition should reduce the persistence demand on internal J-coordinates only when written intermediates are actually attended to later. The reduction should be mediated by attention to the externalized tokens and should predict which tasks benefit from a visible scratchpad.*

7 Component IV: J-Restricted Pullback Fisher Geometry

7.1 Prior art and restricted construction

The natural local metric induced by a model's output distribution is not Euclidean. FishBack derives the pullback Fisher metric at intermediate transformer layers and uses it to construct minimum-distortion steering directions [2]. JSD adopts that result and restricts it to semantically indexed J-directions.

For next-token logits $\boldsymbol{z}(\boldsymbol{h})$ and probabilities $\boldsymbol{p} = \mathrm{softmax}(\boldsymbol{z})$, let

$$J_z = \frac{\partial \boldsymbol{z}}{\partial \boldsymbol{h}}. \tag{33}$$

The pullback Fisher matrix in activation coordinates is

$$F_{\boldsymbol{h}} = J_z^\top \left(\mathrm{diag}(\boldsymbol{p}) - \boldsymbol{p}\boldsymbol{p}^\top \right) J_z. \tag{34}$$

For a selected J-dictionary $D_{\ell,q}$, define a regularized metric on semantic coefficients,

$$G_\ell = D_{\ell,q}^\top F_{\boldsymbol{h}} D_{\ell,q} + \lambda I_q, \qquad \lambda > 0. \tag{35}$$

Alternatively, if P_ℓ is an orthonormal basis for the active J-span, use $P_\ell^\top F_{\boldsymbol{h}} P_\ell + \lambda I$. The regularizer is necessary because the Fisher metric is typically low-rank or ill-conditioned.

A finite-horizon version can sum discounted next-token Fisher terms along teacher-forced or sampled continuations. The chosen horizon must be reported because it changes the meaning and computational cost of the metric.

7.2 Calibrated intervention cost

For a coefficient perturbation $\delta\boldsymbol{a}$,

$$\|\delta\boldsymbol{a}\|_{G_\ell}^2 = \delta\boldsymbol{a}^\top G_\ell \delta\boldsymbol{a} \tag{36}$$

approximates the local KL effect of the corresponding activation perturbation. J-space swaps should therefore be compared at matched Fisher norm as well as matched Euclidean norm. This directly tests whether previously observed layer sensitivity is a geometric calibration artifact.

7.3 Time-dependent path cost

Because G_ℓ changes with layer, JSD does not assume one fixed Riemannian manifold across depth. It defines a discrete path cost in the common semantic coordinate space:

$$\mathcal{L}_{\mathrm{F}} = \sum_{\ell=\ell_0}^{\ell_1-1} \sqrt{\Delta \boldsymbol{a}_\ell^\top \overline{G}_\ell \Delta \boldsymbol{a}_\ell}, \tag{37}$$

$$\mathcal{A}_{\mathrm{F}} = \sum_{\ell=\ell_0}^{\ell_1-1} \Delta \boldsymbol{a}_\ell^\top \overline{G}_\ell \Delta \boldsymbol{a}_\ell, \tag{38}$$

where $\Delta \boldsymbol{a}_\ell = \boldsymbol{a}^{(\ell+1)} - \boldsymbol{a}^{(\ell)}$ and $\overline{G}_\ell$ is a specified interpolation, such as $(G_\ell + G_{\ell+1})/2$. These are time-dependent metric costs, not automatically geodesic lengths on a stationary manifold.

To discuss tortuosity, one must define a reference path under the same boundary conditions and metric schedule. Let $\mathcal{L}_{\min}$ be the numerically minimized cost over admissible coefficient paths between fixed endpoints. Then

$$\mathcal{T}_{\mathrm{F}} = \frac{\mathcal{L}_{\mathrm{F}}}{\mathcal{L}_{\min} + \epsilon} \geq 1 \tag{39}$$

is a well-defined excess-path statistic.

Prediction 6 (Difficulty and excess path cost)**.** *At matched accuracy, output length, and baseline entropy, tasks near a model's competence frontier may exhibit larger $\mathcal{T}_{\mathrm{F}}$ than routine tasks. Any deception-related claim is secondary and must be evaluated against difficulty-matched, role-play, uncertainty, and conflict controls. If Fisher cost adds no predictive value beyond entropy, margin, or token count, the geometric interpretation is unsupported.*

8 Component V: J-Genesis Across Training

8.1 Primary multivariate consolidation profile and secondary index

Calling a scalar an order parameter presupposes evidence for phases, scaling behavior, and a transition. The primary longitudinal object is therefore the preregistered, null-calibrated consolidation profile

$$\boldsymbol{q}(\tau) = \left(\log R_{\mathrm{hub}}(\tau),\ d_{\mathrm{eff}}(\tau),\ P_{\mathrm{persist}}(\tau)\right), \tag{40}$$

Here R_{hub} is the J-aligned versus matched-control read/write-connectivity ratio. The quantity d_{eff} is the participation-ratio dimension of the selected J-code covariance, and P_{persist} is excess metric-aware persistent mass relative to shuffled or random-direction controls. Multivariate change-point tests on Equation (40), retaining covariance among components, are the primary analysis.

For visualization and a preregistered secondary endpoint, define normalized factors

$$H(\tau) = \sigma\left(\frac{\log R_{\mathrm{hub}}(\tau) - m_H}{s_H}\right), \tag{41}$$

$$C_{\mathrm{raw}}(\tau) = \left[1 - \frac{d_{\mathrm{eff}}(\tau)}{q}\right]_{[0,1]}, \tag{42}$$

$$C(\tau) = C_{\mathrm{raw}}(\tau) \min\left\{1, \frac{d_{\mathrm{eff}}(\tau)}{d_{\min}}\right\}, \tag{43}$$

$$P(\tau) = \left[P_{\mathrm{persist}}(\tau)\right]_{[0,1]}, \tag{44}$$

where m_H and s_H are fixed from a null distribution, σ is the logistic function, $[\cdot]_{[0,1]}$ clips to $[0,1]$, and $d_{\min}$ is a preregistered diversity floor estimated from held-out task families. The factor in Equation (43) prevents representational collapse from being scored as ideal concentration. The secondary Workspace Consolidation Index (WCI) is

$$\Omega(\tau) = (H(\tau)C(\tau)P(\tau))^{1/3}. \tag{45}$$

The three raw components, normalized factors, and multivariate test must be reported even when the scalar index is shown. Results should be repeated with alternative nulls, normalizations, sparsity budgets, and values of $d_{\min}$; a change point found only in Ω is not sufficient evidence of representational reorganization.

8.2 Change-point hypothesis

Hypothesis 1 (Workspace consolidation precedes selected capability gains)**.** *For some tasks requiring flexible routing of intermediate information, a statistically detectable change in one or more continuous JSD observables will precede or coincide with a change in continuous task performance. This relationship should survive alternative behavioral metrics, held-out prompts, and correction for training loss.*

This statement is intentionally weaker than asserting a thermodynamic phase transition. Evidence for a transition would require, at minimum: replicated change points across seeds; robustness to smoothing and metric choice; a discontinuity or critical scaling pattern in continuous observables; and a mechanistic intervention linking the representational change to behavior. Since discontinuous scoring can create illusory emergence, accuracy should be accompanied by log-likelihood, calibrated margin, expected reward, and per-example continuous scores [3].

8.3 Developmental measurements

The checkpoint program tracks four developmental questions.

Vocabulary accessibility. Which token- and span-indexed directions become reliably decodable and causally useful first? Frequency, concreteness, and syntactic category should be treated as covariates, not post hoc explanations.

Binding onset. Does $\beta(w)$ or Δ_{ord} increase before cross-assignment competence, after it, or not at all? The key test is whether binding diagnostics predict held-out relational behavior beyond unigram and template-lens baselines.

Robustness maturation. Does the distribution of $\widehat{r}_{\text{rob}}$ become better calibrated to actual intervention resistance? Heavy tails alone do not establish hallucination or overconfidence; those hypotheses require matched evidence and truth-labeled tasks.

Post-training perspective. Base and post-trained checkpoints can be compared to determine whether post-training changes J-content, gating predictors, persistence, or only downstream use. Self-referential labels should not be interpreted as a "point of view" without controls for prompt genre, assistant tokens, and output priors.

ID	Question	Primary comparison	Failure criterion
A	Does second order add value?	$\widehat{r}_{\text{rob}}$ versus margin, entropy, gradient norm, and actual adversarial intervention radius	No held-out predictive improvement or poor estimator stability
B	Does the path lens capture binding?	Path lens versus unigram J-code, template/oracle lens, SAE probes, and surface probability	No incremental prediction and no selective binding-swap effect
C	Does J-code form a usable state?	Non-autonomous EDMD versus autoregression, nonlinear baselines, shuffled pairs, and J-plus-residual state	Held-out closure error near or above null/baseline models
D	Is there a gating mechanism?	Ignition predictors followed by head/MLP ablations and mediation tests	Predictors fail out of distribution or interventions do not alter admission
E	Does Fisher calibration matter?	Matched Euclidean versus matched Fisher interventions; comparison with FishBack-style optimal steering	No improvement in layer transfer, selectivity, or off-target KL
F	Does consolidation predict learning?	Multivariate consolidation profile, secondary WCI, and change points across public checkpoint suites and multi-seed algorithmic training	Effects vanish under continuous metrics, alternate normalization, seed replication, or loss controls

Table 1: Dependency-ordered evaluation plan. Each study has a quantitative failure criterion rather than a purely illustrative prediction.

9 Integrated Experimental Program

The proposal should be evaluated in stages, with each later study contingent on the earlier instrument checks.

One dependency requires explicit handling. Studies A–D use Fisher-calibrated interventions, but whether Fisher calibration itself adds value is only tested in Study E. To avoid circularity, the regularized metric of Section 7 is fitted once during instrument setup, frozen, and used *provisionally* in Studies A–D, with every Fisher-matched comparison also reported at matched Euclidean norm. If Study E subsequently fails, the Euclidean-matched results stand alone and the Fisher-calibrated results are reinterpreted as a sensitivity analysis rather than a primary endpoint.

9.1 Recommended initial model suite

The first implementation should use open-weight models for which gradients and intermediate activations are available. A practical suite is: (i) one small model for exact or high-accuracy Hessian/Fisher calculations; (ii) one medium model for causal interventions; (iii) a checkpoint suite such as Pythia for longitudinal analysis [21]; and (iv) small algorithmic transformers trained across multiple random seeds for controlled grokking-style studies [18, 22]. Production-model findings from the original J-lens paper should be treated as motivating evidence, not as a substitute for reproducible open-model validation.

9.2 Statistical discipline

All studies should report confidence intervals across prompts and seeds, hierarchical models where prompts are nested within task families, correction for repeated concept searches, and preregistered

primary outcomes. Candidate phrase discovery, hyperparameter selection, and hypothesis testing must use separate data splits. Negative results are scientifically informative because each component is explicitly optional: JSD can survive the failure of one instrument only if the remaining claims are narrowed accordingly.

9.3 Computational budget

The identifiability discipline of Section 2.3 is not free: bootstrap propagation multiplies the cost of code fitting, operator estimation, and change-point testing by the number of bootstrap replicates, and Hessian–vector and Fisher estimates add backward passes on top of the forward sweeps. Three mitigations keep the program tractable. First, expensive second-order quantities (curvature envelopes, Fisher blocks) are estimated on a preregistered subsample of prompts and shared across bootstrap replicates, since their sampling error is dominated by prompt variance rather than code jitter. Second, bootstrap replication is applied at full resolution only to the primary endpoints of each study, with reduced replicate counts for secondary analyses. Third, the checkpoint program uses a two-pass design: a cheap screening pass (codes, closure, and the consolidation profile at coarse checkpoint spacing) followed by dense sampling and full uncertainty propagation only around candidate change points. Reported results must state replicate counts and subsample sizes so that estimator uncertainty and compute budget can be assessed together.

10 Architectural Summary

Figure 1 summarizes the revised framework. It separates established inputs from proposed estimators and avoids depicting a stationary Koopman generator or an already-confirmed phase transition.

11 Applications, If Validated

Training diagnostics. A validated consolidation profile, with its summary WCI, could supplement loss curves with mechanistically interpretable measurements of accessibility, binding, persistence, and connectivity. It should not be called an online "cheap scalar" unless a reduced estimator is benchmarked: fitting lenses and collecting Hessian/Fisher statistics can be expensive (Section 9).

Model comparison. Models with similar accuracy may differ in robustness, workspace closure, persistence capacity, or Fisher-calibrated intervention sensitivity. These quantities could distinguish brittle from stable solutions, but only after showing test–retest reliability and relevance beyond output entropy and calibration.

Safety auditing. Second-order margins and persistence may help characterize whether concerning content is transient, robust, or causally connected to behavior. No result would establish that J-space monitoring is sufficient: automatic or strategically relocated computations may lie outside the measured sparse frame [1]. Defensive monitoring should therefore combine JSD with independent probes of non-J-space residual activity.

Comparative architectures. The framework can be adapted to recurrent networks, state-space models, multimodal transformers, and diffusion language models when a differentiable output

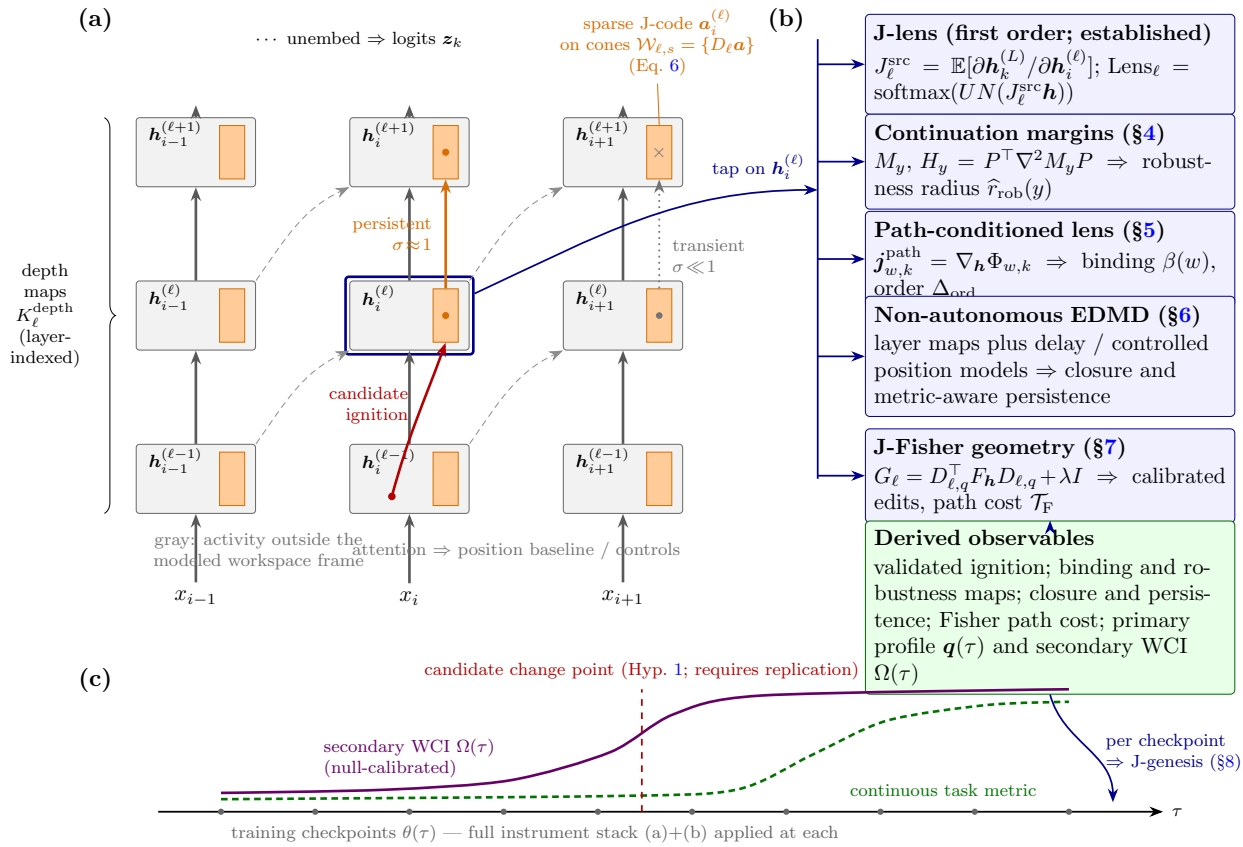

Figure 1: **JSD instrument architecture. (a)** Established substrate: residual states $h_i^{(\ell)}$ evolve along two discrete, non-stationary axes—depth, modeled by *layer-indexed* maps K_ℓ^{depth}, and token position through baseline, delay-embedded, and controlled maps; no stationary Koopman generator is assumed. Orange strips depict the sparse J-code $a_i^{(\ell)}$ on the cone family $\mathcal{W}_{\ell,s}$; the gray bulk is activity outside the modeled workspace frame, which independent probes must monitor (Section 13). A coordinate crossing its null-calibrated threshold is only a *candidate* ignition (red); functional admission additionally requires the causal test of the operational-ignition definition (Section 6). Metric-aware finite-window propagator singular values classify coordinates as approximately persistent ($\sigma \approx 1$) or transient ($\sigma \ll 1$). **(b)** The instrument stack, all reading from a common tap on $h_i^{(\ell)}$: the established first-order J-lens plus the four proposed extensions of Sections 4–7, combining into derived observables including the primary consolidation profile and secondary Workspace Consolidation Index. **(c)** Longitudinal program (Section 8); curves are *illustrative of Hypothesis 1, not results*: the dashed line marks a candidate change point that must replicate across seeds, survive continuous behavioral metrics, and admit a mechanistic link before any transition language is warranted.

distribution and hidden states are available. Cross-architecture comparisons require architecture-specific time axes and cannot assume that transformer depth, recurrent time, and diffusion steps are equivalent.

12 Related Work

JSD builds directly on the Jacobian-lens/global-workspace study and its reference implementation [1]. The J-lens belongs to a broader family of methods that decode or causally manipulate intermediate representations, including the logit lens, tuned lens, sparse autoencoders, causal tracing, and activation steering [8, 9, 10, 11, 12, 13]. Its distinctive feature is the use of an average input–output Jacobian to identify directions disposed toward future verbalization.

The sequence component is adjacent to phrase-template and oracle-lens methods proposed in the J-lens appendices, as well as contextual feature discovery using SAEs. Its proposed addition is an autoregressive path functional and relational causal test, not phrase representation per se.

The dynamical component uses Koopman operator theory and extended dynamic mode decomposition [16, 17]. Unlike autonomous physical systems, transformer layers form a non-autonomous discrete sequence, motivating layer-indexed operators and finite-window propagators. This distinction is central to avoiding overinterpretation of pooled eigenvalues.

The geometric component follows information geometry and natural-gradient theory [14]. Most directly, FishBack derives the pullback Fisher metric for transformer intermediate activations and demonstrates its use for optimal activation steering [2]. JSD's role is to apply this metric to J-coordinates and workspace trajectories.

The longitudinal program connects to checkpoint-based mechanistic studies of induction heads, grokking, and model development [20, 18, 22, 21]. It also incorporates the warning that apparent emergent abilities can depend on metric choice [3]. Finally, global workspace theory motivates the language of access, broadcast, competition, and maintenance [4, 5, 6]; these are functional analogies, not evidence of phenomenal consciousness [7, 24].

13 Limitations and Risks

Dependence on the J-lens approximation. If the corpus-averaged J-dictionary omits important context-specific directions, every downstream component inherits that blind spot. Results should be repeated with locally fitted directions, template/oracle lenses, and independent SAE or probe-based state descriptions.

State insufficiency. Sparse J-codes may not close under depth or position evolution. A poor transfer fit is not merely an engineering inconvenience; it is evidence that the proposed workspace state omits variables required for prediction. Augmenting with residual features may improve accuracy while weakening interpretability, and that trade-off should be reported.

Coordinate non-identifiability. Coherent J-dictionaries can yield unstable or nonunique sparse decompositions. Column normalization, elastic regularization, active-Gram conditioning, bootstrap support stability, and uncertainty propagation are required before coefficient changes are interpreted as cognitive dynamics.

Position-state non-Markovianity. The workspace code at one token need not summarize the full prefix available through attention. Delay embeddings and controlled operators must be compared with the one-step position model; failure of the simple model is expected and should not be hidden by pooling.

Second-order locality. Hessian-based robustness is local and can fail under large interventions, sharp nonlinearities, or activation-distribution shift. Certified radii based on empirical curvature bounds are approximate unless the bound is verified over the full perturbation ball.

Metric degeneracy. Pullback Fisher matrices are often low-rank and ill-conditioned. Regularization changes distances and must be sensitivity-tested. A Fisher-small direction is behaviorally quiet only relative to the chosen output horizon and context distribution.

Lexicon dependence. Phrase candidates and competitor sets introduce researcher degrees of freedom. Discovery/confirmation separation, paraphrase robustness, and preregistration are essential.

Causal ambiguity. A coordinate can correlate with behavior because it is a readout, a cause, a downstream consequence, or a shared effect. Intervention, mediation, timing, and matched controls are required before using terms such as gate, broadcast, or commitment.

Goodhart risk. Training directly against JSD observables may cause cognition to move outside the measured frame or make the readout cosmetically clean without improving behavior. Workspace-aware objectives should be evaluated with held-out probes constructed independently of the optimized metric.

Consciousness claims. JSD concerns functional access-like organization. It provides no measurement of phenomenal experience and no basis for asserting that a model is conscious. Even the global-workspace interpretation remains an empirical analogy with architectural differences from biological global neuronal workspace theories [1, 6].

14 Conclusion

The J-lens provides an unusually direct bridge between internal activation directions and future verbal behavior. Its limitations create a productive research agenda, but that agenda is strongest when novelty claims are narrowed and mathematical interpretations are disciplined. The JSD framework makes five testable proposals: continuation-margin robustness rather than curvature-as-attractor; path-conditioned sequence directions rather than a claim to invent multi-token lenses; non-autonomous transfer operators with closure tests rather than a single assumed Koopman generator; J-restricted pullback Fisher geometry that explicitly builds on prior work; and a primary multivariate consolidation profile with a secondary null-calibrated index rather than a presumed phase-transition order parameter.

The central scientific question is not whether these constructions sound cognitively suggestive. It is whether they improve held-out prediction, survive causal intervention, remain stable across estimators and model families, and anticipate learning under continuous metrics. If they do, JSD would turn a static vocabulary-indexed lens into a rigorous longitudinal instrument for studying how accessible, persistent, and compositional representations develop. If they do not, the framework's

explicit failure criteria will still clarify which parts of the workspace analogy are measurement artifacts and which are mechanistically real.

Acknowledgments

This manuscript is a methodological response to the Jacobian-lens and global-workspace research program. Characterizations of that work are based on its public paper, appendices, and reference implementation. No new empirical results are claimed here.

References

[1] W. Gurnee, N. Sofroniew, A. Pearce, M. Piotrowski, I. Kauvar, R. Chen, A. Soligo, P. Bogdan, E. Ong, R. Wang, B. Thompson, D. Abrahams, S. Kantamneni, E. Ameisen, J. Batson, and J. Lindsey. Verbalizable representations form a global workspace in language models. *Transformer Circuits Thread*, July 6, 2026. https://transformer-circuits.pub/2026/workspace/index.html. Companion code: https://github.com/anthropics/jacobian-lens.

[2] S. Wang and J. Zhao. FishBack: Pullback Fisher geometry for optimal activation steering in transformers. *arXiv:2605.17231*, 2026.

[3] R. Schaeffer, B. Miranda, and S. Koyejo. Are emergent abilities of large language models a mirage? *arXiv:2304.15004*, 2023.

[4] B. J. Baars. *A Cognitive Theory of Consciousness*. Cambridge University Press, 1988.

[5] S. Dehaene and L. Naccache. Towards a cognitive neuroscience of consciousness: basic evidence and a workspace framework. *Cognition*, 79(1–2):1–37, 2001.

[6] G. A. Mashour, P. Roelfsema, J.-P. Changeux, and S. Dehaene. Conscious processing and the global neuronal workspace hypothesis. *Neuron*, 105(5):776–798, 2020.

[7] N. Block. On a confusion about a function of consciousness. *Behavioral and Brain Sciences*, 18(2):227–247, 1995.

[8] nostalgebraist. Interpreting GPT: the logit lens. *LessWrong*, 2020.

[9] N. Belrose et al. Eliciting latent predictions from transformers with the tuned lens. *arXiv:2303.08112*, 2023.

[10] T. Bricken et al. Towards monosemanticity: decomposing language models with dictionary learning. *Transformer Circuits Thread*, 2023.

[11] A. Templeton et al. Scaling monosemanticity: extracting interpretable features from Claude 3 Sonnet. *Transformer Circuits Thread*, 2024.

[12] K. Meng, D. Bau, A. Andonian, and Y. Belinkov. Locating and editing factual associations in GPT. In *Advances in Neural Information Processing Systems*, 2022.

[13] A. Turner et al. Activation addition: steering language models without optimization. *arXiv:2308.10248*, 2023.

[14] S. Amari. Natural gradient works efficiently in learning. *Neural Computation*, 10(2):251–276, 1998.

[15] B. A. Pearlmutter. Fast exact multiplication by the Hessian. *Neural Computation*, 6(1):147–160, 1994.

[16] B. O. Koopman. Hamiltonian systems and transformation in Hilbert space. *Proceedings of the National Academy of Sciences*, 17(5):315–318, 1931.

[17] M. O. Williams, I. G. Kevrekidis, and C. W. Rowley. A data-driven approximation of the Koopman operator: extending dynamic mode decomposition. *Journal of Nonlinear Science*, 25:1307–1346, 2015.

[18] A. Power, Y. Burda, H. Edwards, I. Babuschkin, and V. Misra. Grokking: generalization beyond overfitting on small algorithmic datasets. *arXiv:2201.02177*, 2022.

[19] J. Wei et al. Emergent abilities of large language models. *Transactions on Machine Learning Research*, 2022.

[20] C. Olsson et al. In-context learning and induction heads. *Transformer Circuits Thread*, 2022.

[21] S. Biderman et al. Pythia: a suite for analyzing large language models across training and scaling. In *Proceedings of the 40th International Conference on Machine Learning*, 2023.

[22] N. Nanda, L. Chan, T. Lieberum, J. Smith, and J. Steinhardt. Progress measures for grokking via mechanistic interpretability. In *International Conference on Learning Representations*, 2023.

[23] B. Weinkove. On the J-flow in higher dimensions and the lower boundedness of the Mabuchi energy. *Journal of Differential Geometry*, 73(2):351–358, 2006. Preprint: *arXiv:math/0309404*.

[24] P. Butlin et al. Consciousness in artificial intelligence: insights from the science of consciousness. *arXiv:2308.08708*, 2023.