J-SPACE AS A CANDIDATE ACCESS-CLOSURE REGIME

Verbalizable Representations, Closure, and the Measurement of Remainder in Language Models

CF Dietz

Public Review Draft 1.1 | July 2026

Revised conceptual analysis and research proposal. No new empirical data are reported.

Abstract

Recent mechanistic interpretability research reports that large language models maintain a sparse, functionally privileged family of verbalizable internal representations. Termed J-space, these representations are causally implicated in verbal report, directed modulation, intermediate reasoning, flexible reuse across tasks, and selective impairment under intervention, while much routine computation proceeds outside them. This paper explains J-space for non-specialist readers and asks what, if anything, it contributes to the operational claims of Consciousness, Closure, and the Cosmos (CCC).

We argue that J-space is best treated as a candidate local access-closure regime. Its contents appear to stabilize distinctions, preserve functional identity across context-dependent transformations, compete for limited access, and become available to several downstream operations. The paper further distinguishes access closure, self-model closure, response closure, and phenomenal closure. In particular, J-space can register conflict that does not govern the final output, showing that what is accessible to a model need not be what its response system commits to.

A second contribution is a concrete proposal for testing CCC remainder in language models. Existing J-space residuals, sparse-autoencoder reconstruction errors, and attribution-graph error nodes are not themselves CCC remainder. CCC defines remainder as best-achievable mismatch within a closure-constrained model family. To make that definition applicable to transformers, we separate a candidate system closure, comprising a sparse frame, interaction graph, and internal access structure, from the output grammar through which the model acts and the observation grammar through which investigators measure it.

The resulting program compares J-space with rotated, logit-lens, sparse-autoencoder, phrase-lens, circuit-tracing, natural-language, and full-activation alternatives under fixed description-length costs. It proposes tests of observation expansion, ambiguity gaps, transition-model discrimination, forced factorization, perturbation recovery, access-response dissociation, and training-induced grammar change. The program could support a limited claim that reportable and flexibly usable cognition depends on selectively stabilized representational regimes. It would not establish phenomenal consciousness, irreducible presence, or CCC's cosmological extension.

Keywords

J-space; Jacobian lens; global workspace; closure; remainder; response closure; observation grammar; mechanistic interpretability; access consciousness; large language models; model selection

Plain-Language Summary

A language model does not usually perform its internal work in sentences. At every layer it carries a large numerical pattern containing many kinds of information. Some of that information helps with grammar, local word structure, pattern recognition, or familiar routines. Only a small part appears in a form the model can readily name, hold active, and pass to whatever operation the current task requires.

The Jacobian lens is a method for locating this verbalizable part of a model's internal activity. It asks, in effect: if this intermediate activation were nudged slightly, which words would the model become more disposed to say now or later? The researchers who developed the method found that a sparse, changing set of these word-linked directions is unusually important. The model can report them, use them as intermediate steps, and apply several different operations to the same representation. They call this family J-space.

A helpful analogy is a shared whiteboard. Most computation occurs away from the whiteboard. Information placed on it can be named, held active, revised, and used by several internal routines. The analogy has strict limits. J-space is not a room, a central observer, or evidence that anyone is experiencing what is written there.

One experiment shows why the whiteboard should not be confused with the model's final answer. When researchers force the model to choose an option it ordinarily disfavors, an internal conflict signal, especially the token BUT, appears in J-space. The model nevertheless usually continues by defending the forced choice. What is available in the workspace and what the response system finally commits to can therefore come apart.

CCC uses the word Closure for the stabilization of distinctions, identities, and allowed transformations. J-space is interesting from that perspective because its active contents appear to perform those functional jobs. A model can stabilize France rather than China, retain the identity of the selected country, and then use it in different operations involving capitals, languages, currencies, or continents.

A second distinction concerns the measuring instrument. The J-lens is not a transparent window onto everything the model represents. The model's tokenizer and vocabulary, together with the lens's averaged Jacobian, sparsity rule, and chosen layer range, determine what it can reveal. A phrase lens or a context-conditioned lens may expose content the ordinary J-lens misses. When a richer instrument succeeds, the improvement initially counts as reduced estimation or approximation error, not as a change in the model itself.

The paper does not argue that J-space is consciousness. It does not show that a model has subjective experience, and it offers no evidence for CCC's cosmological proposal. Its value is methodological. J-space may let researchers turn CCC's abstract concepts of closure, integration, stabilization, and remainder into a controlled model-comparison program.

Remainder in CCC is not simply whatever activation lies outside J-space. Much useful automatic processing occurs there. Remainder is the best mismatch that remains after the strongest model permitted by a specified closure has been fit to the system. The proposed experiments therefore compare several candidate representations, penalize their complexity, and ask which one best predicts both ordinary behavior and the effects of causal interventions.

1. Introduction

Large language models present a distinctive scientific problem. Their behavior is expressed in ordinary language, but the computation producing that behavior occurs in high-dimensional numerical states that are not directly legible. A model may answer correctly, conceal a relevant consideration, plan several words ahead, or use an intermediate concept without stating any of this in its output. Surface conversation therefore gives only partial access to the process that generated the conversation.

Gurnee and colleagues report a major advance on this problem. Using a technique called the Jacobian lens, they identify a sparse family of internal representations that are poised for later verbalization and are causally implicated in report, deliberate modulation, intermediate reasoning, and flexible task-dependent use. They call the family J-space and interpret it as a functional analogue of a global workspace [2]. The result is important even if one rejects every philosophical claim about machine consciousness. It suggests that a learned transformer can develop a limited, shared representational format that stands apart from a larger volume of automatic processing.

Consciousness, Closure, and the Cosmos develops a much broader framework [1]. CCC begins by separating the occurrence of presence from the organization of content. It defines Closure as stabilization into a coherent regime of distinctions, identity criteria, and admissible transformations. It also defines remainder as the best-achievable mismatch left when a closure-constrained model family is fitted to an empirical structure. The minimal portion of CCC is intended as an operational research program. Its maximal extension treats cosmogenesis as the onset of a stable closure grammar. Those two levels are explicitly separable.

The appearance of J-space raises an obvious but dangerous question. Does J-space support CCC, or does CCC support J-space? A loose answer would invite overstatement. J-space was developed to test properties associated with access consciousness and global workspace theory, not Closure. CCC did not predict the Jacobian lens, the token-indexed geometry of J-space, its middle-layer organization, or its broadcast circuitry. Conversely, an internal workspace in a transformer does not establish irreducible presence or a consciousness-first cosmology. Mere compatibility is not confirmation.

The defensible relation is narrower. J-space provides a concrete and unusually inspectable model system in which the operational components of Closure can be reformulated and tested. Its active representations appear to stabilize distinctions, retain identity while entering different downstream operations, and compete for limited access to a shared causal format. That makes J-space a candidate local access-closure regime. The word candidate is essential. The proposed identification is an interpretation of existing evidence, not a result already established by either paper.

This manuscript has six aims. First, it explains J-space in language accessible to readers who do not work in machine learning. Second, it states exactly which CCC claims the J-space results support, which they merely motivate, and which they do not bear on. Third, it distinguishes access closure, self-model closure, response closure, and phenomenal closure. Fourth, it separates the model's candidate internal closure from the output grammar through which it acts and the observation grammar imposed by the investigator's lens and feature vocabulary. Fifth, it develops a frame-and-graph extension of CCC suitable for superposed transformer representations. Sixth, it proposes a program for measuring remainder through held-out prediction, causal interventions, observation expansion, transition-model comparison, and complexity-controlled testing.

The analysis follows a four-part evidential discipline.

1. An empirical result is a finding reported by the J-space authors or another cited study.
2. A Closure interpretation is a description of that finding in CCC's vocabulary.
3. A prospective hypothesis is a new relation that must be tested.
4. An excluded inference is a claim the present evidence does not support.

This discipline matters because the philosophical interest of J-space can easily exceed its evidential reach. The central conclusion is deliberately constrained: J-space gives moderate empirical support to a minimal claim that flexible, reportable cognition is organized through a selectively stabilized access regime. CCC contributes a testable language for asking whether that regime is explanatory rather than merely redescribed, provided the model's organization is kept distinct from the grammar through which investigators observe it. This provides no evidence of phenomenal presence or cosmic closure; it makes closure and remainder experimentally consequential.

2. J-Space in Plain and Technical Terms

2.1 The residual stream

A transformer processes text as a sequence of tokens. At each token position it maintains a high-dimensional vector called the residual stream. Successive layers read from and write to this vector. Early layers primarily reflect the local input. Intermediate layers progressively assemble context, abstractions, and task-relevant information. Final layers prepare the probability distribution for the next token.

The residual stream is not a sentence hidden inside the model. A single activation can simultaneously carry grammatical information, entity information, task state, formatting details, latent plans, and many other features. These features may be superposed in the same dimensions [16]. Inspecting individual coordinates usually reveals little.

2.2 The Jacobian lens

The Jacobian lens asks how an intermediate activation is disposed to affect later verbal output. For a residual-stream state h at layer l, the method computes an average linearized map from small changes in h to changes in the model's final residual states at the present and later token positions. That map is then composed with the model's vocabulary readout. In simplified notation:

$$J_l = E[\partial h_{final,t'} / \partial h_{l,t}], \quad t' >= t$$

and

$$lens(h_l) = softmax(W_U \, norm(J_l \, h_l))$$

Here W_U is the model's ordinary unembedding matrix, which converts final activations into token scores. The expectation is taken over many prompts and token positions. Averaging is crucial. A Jacobian calculated on one prompt would capture how a representation happens to be used in that context. The averaged Jacobian aims to recover a more general disposition to make a concept verbalizable across contexts [2].

Each vocabulary token therefore receives a direction in residual-stream space. A high score for "spider," for example, means that the current activation is aligned with a direction that, across many contexts, tends to make the

model more likely to say "spider" now or later. The method is causal in construction, but it remains a first-order approximation. Its readout is limited by the model's tokenizer, the averaging distribution, linearization, and the assumption that a nameable concept can be associated with a token direction.

2.3 J-space is not an ordinary subspace

The term J-space can be misleading. The token directions are overcomplete. There are more vocabulary directions than residual-stream dimensions, and they can span the full residual stream. There is no unique decomposition of an activation into those directions.

Gurnee and colleagues therefore add a sparsity and positivity constraint. Let v_i be the J-lens vector for token i. For an allowed sparsity k, define:

$$F_k = \text{Union}_{|S| \le k} \{ \text{Sum}_{i \in S} a_i v_i : a_i \ge 0 \}.$$

Geometrically, F_k is a union of low-dimensional cones, not one linear subspace. The J-space component of an activation is the nearest sparse nonnegative combination of J-lens vectors under the chosen approximation procedure. The non-J-space component is the vector left after that approximation.

The paper usually sets k at no more than roughly 25, based on an empirical occupancy estimate. That number should not be read as 25 complete thoughts. Several token directions can jointly express one concept, and related concepts can activate as a semantic neighborhood. In a single layer, the simultaneously prominent content can be much smaller.

2.4 What the experiments show

The evidence for J-space is not based on readout alone. The researchers intervene on the inferred directions and examine what changes.

Verbal report. When a model is asked to think of an item and then name it, the item appears in the J-lens before it is spoken. Swapping the active direction for another item changes the report. Injecting a concept into earlier user-turn activations can cause the model to name that concept later when explicitly asked what thought it detects.

Directed modulation. A model instructed to hold a concept in mind while copying unrelated text places concept-related directions in J-space even though the copied output remains unchanged. It can also carry an unspoken arithmetic computation while writing a different sentence. Task instructions can increase or suppress the relevant directions, although suppression is imperfect.

Internal reasoning. In two-hop questions, J-space can reveal an unspoken intermediate. For "the number of legs on the animal that spins webs," the model represents spider before producing eight. Swapping spider for ant changes the answer toward six. Similar interventions redirect rhyme planning, multilingual antonym reasoning, and strategy selection.

Flexible generalization. The same country representation can be moved from one context to another and then processed by different operations. A France-to-China swap can change Paris to Beijing, French to Chinese, Europe to Asia, and euro to yuan. This is not merely an association with one output. The representation functions as a reusable argument.

Selectivity. Many routine operations continue without J-space. A model can preserve the language of a passage, detect a language anomaly, maintain grammar, or perform some classifications while a relevant J-space coordinate is swapped or suppressed. The same underlying information becomes J-space-dependent when the model must name it or apply an arbitrary context-specified operation to it.

The causal concentration is striking. For certain concept vectors, the J-space component carries only about 6 to 7 percent of the vector's variance, yet it accounts for most of the concept's effect on report. The much larger non-J-space component has little direct effect on report once the relevant J-space coordinates are clamped. Heavy J-space ablation reduces performance on controlled multi-hop reasoning from near ceiling to near zero while leaving a substantial majority of ordinary next-token predictions unchanged [2].

2.5 Structural evidence

J-space appears workspace-like only through an intermediate range of layers. In the models studied, early readouts are noisy or low-rank, middle layers contain persistent abstract content, and final layers become increasingly tied to the imminent output. The exact boundaries are model-specific. They should not be treated as a universal one-third/two-thirds law.

The active J-space content is also limited. Sparse decomposition explains less than 10 percent of activation variance at the measured occupancy. Related concepts can coexist more readily than unrelated items. New task categories can displace old ones. More difficult covert computations compete more strongly for the same token-level workspace.

Finally, J-space is embedded in privileged circuitry. MLP components amplify J-aligned directions, and a subset of attention heads preferentially transmits them between token positions. Ablating those heads reduces the stability of J-lens contents and sharply reduces later report of injected concepts, while matched random-head ablations have much smaller effects [2].

2.6 What the evidence does not show

J-space is not a complete inventory of model cognition. The lens is limited to single-token directions unless extended. It represents content as a sparse bag of concepts and does not by itself reveal relational binding. Early-layer information may be inaccessible to the lens rather than absent from the model. Automatic processing can remain highly structured outside J-space. A well-practiced harmful policy might therefore evade J-space monitoring.

J-space is also not evidence of phenomenal consciousness. The researchers explicitly investigate a functional notion of access. The fact that J-space ablation changes experiential language shows that J-space contributes to the representation and production of phenomenological discourse. The same effect occurs when the model describes another person's experience, which supports a representational interpretation rather than a direct inference to subjective feeling [2].

The proper conclusion is functional: J-space is a sparse, causally reusable format for selected verbalizable content.

3. The Minimal Operational Core of CCC

CCC begins with a distinction between presence and content [1]. C denotes irreducible presence, the bare first-person datum that experience occurs. c denotes localized consciousness-with-content, including perception, memory, attention, affect, agency, self-modeling, and reportable cognition. The distinction does not posit two substances. It separates the occurrence of experience from the organization of what is experienced.

The operational work of CCC concerns organization. Closure is defined as stabilization into a coherent regime of distinctions, identity criteria, and admissible transformations. A stable content regime determines what counts as the same, what counts as different, and what changes are allowed while identity is preserved. CCC characterizes this process through a triad.

Capacity concerns whether distinctions, salience, and identities are stable enough to support determinate content.

Variational stability concerns selection among alternative parsings under constraints. A closure is favored when it occupies a minimum or metastable basin in an objective trading fit against structural cost.

Consistency concerns coordination across perspectives. A shared world requires sufficiently compatible distinctions and identities among agents, although perfect agreement is unnecessary.

CCC then distinguishes a closure state from a closure grammar. A state is one stabilized configuration at closure time tau. A grammar is the rule family that determines which states are admissible. A system can change state without changing grammar, or enter a regime in which the old grammar no longer supplies adequate states.

The paper's formal proxy represents a candidate closure as K = (Pi, E), where Pi is a partition of degrees of freedom and E is a sparse interaction graph. For empirical structure p and a model family Q(K) constrained by K, state-level remainder is:

$$R(K \mid p) = \min_{q \in Q(K)} D_{KL}(p \| q). \quad (1)$$

The closure score adds structural costs:

$$S(K \mid p) = R(K \mid p) + \lambda C_E(E) + \gamma C_{\Pi}(\Pi). \quad (2)$$

The selected closure minimizes S. Grammar-level remainder minimizes state-level remainder over all states admitted by a fixed grammar Omega:

$$R_{\Omega}(p) = \min_{K \in K(\Omega)} R(K \mid p). \quad (3)$$

CCC emphasizes that observed remainder can have at least three sources. R_e is estimation remainder from finite data and noise. R_a is approximation remainder from an inadequate model family or representation. R_c is closure remainder attributable to the nontrivial distinctions and constraints imposed by the closure under finite access and resources. Separating them is an empirical obligation, not something philosophy can stipulate.

The formalism also defines three bridge-ready quantities. Delta S is the score gap between the best and second-best closure. Φ_{cl} is the increase in remainder caused by forcing an integrated system to factorize. Kappa is the rate at which remainder declines after perturbation. CCC proposes that clarity should increase as selected remainder falls and the score gap widens, that unity should increase with integration, and that novelty should produce a transient rise in remainder followed by restabilization [1].

Two boundaries are essential. First, these quantities operate at the level of organized content. They do not explain why presence exists. Second, the cosmological extension is not required by the operational core. J-space can therefore be relevant to CCC even if one remains agnostic about C and rejects every cosmological proposal.

4. J-Space as a Candidate Access-Closure Regime

4.1 Stabilized distinctions

The clearest correspondence concerns ambiguous input. In the J-space experiments, a token embedding can be interpolated between two concepts, such as Germany and France. Early activations follow the graded mixture. Near the onset of the workspace band, individual cases increasingly settle toward one interpretation or the other. At maximal ambiguity, the distribution becomes bimodal across prompts rather than remaining centered between the alternatives. The J-space component sharpens somewhat earlier than the full activation [2].

This is a concrete example of selection among alternative distinctions. The model does not merely preserve the input's continuous ambiguity. Context and learned constraints push it toward a determinate identity. In CCC language, the event resembles closure-state selection under variational stability.

The resemblance is not yet a formal test. The J-space study did not define competing closure candidates, calculate S, or measure a score gap. Ordinary nonlinear classification, attractor-like dynamics, and global-workspace ignition can also describe the effect. The result therefore supports the generic stabilization claim but does not uniquely confirm CCC.

The result is compatible with a transition-like or bifurcation-like interpretation, but the current evidence does not identify a saddle-node or any other specific dynamical class. Sharp selection and prompt-level bimodality do not by themselves reveal stable and unstable fixed points, their collision as a control parameter varies, hysteresis, critical slowing, or a common dynamical map iterated across layers [25]. Because successive transformer layers have different weights, treating depth as time in one autonomous dynamical system requires an explicit effective construction.

The proper prospective claim is therefore comparative. Smooth nonlinear classification, piecewise-linear thresholding, stochastic winner-take-all selection, cusp-like bistability, and saddle-node-like effective dynamics should be fit as rival descriptions. A bifurcation account earns explanatory status only if it predicts perturbation thresholds, recovery profiles, susceptibility near the boundary, and held-out transitions better than simpler alternatives. Section 9 converts that requirement into a test.

4.2 Identity criteria and admissible transformations

The flexible-generalization experiments provide a deeper correspondence. A concept direction is not tied to one response. The model treats the swapped-in country as the same argument across several operations. The identity "China" is maintained while the admissible transformation changes from capital to language, continent, or currency.

This maps closely onto CCC's definition of Closure. The relevant regime contains:

1. a stabilized distinction, China rather than France;
2. an identity criterion, the concept remains functionally the same across contexts;
3. a family of admissible transformations, each producing a context-appropriate relation or attribute.

Global workspace theory predicts flexible broadcast, so the evidence supports that theory more directly than it supports CCC. Closure nevertheless adds a useful level of description. Broadcast alone says that information becomes widely available. Closure asks what identity is being preserved and what transformations remain valid within the selected regime.

4.3 Capacity and selective access

J-space is capacity-limited in a functional sense. Only selected content becomes strongly verbalizable and causally available for flexible reasoning. The majority of activation variance remains elsewhere. Unrelated contents displace one another more readily than related contents, and difficult covert computations interfere with simultaneous content.

This fits the capacity component of Closure. Determinate, reportable content requires a selective stabilization of salience and identity. It also supplies a possible operational reading of CCC's local horizon m. Information can be present and causally useful in the model while remaining outside the format used for direct report and arbitrary downstream control. Local unknowability need not mean literal absence. It can mean restricted access to a representation encoded in another format.

4.4 Broadcast and integration

J-space content is preferentially read, amplified, and transmitted by model components. The J-space study identifies MLP and attention pathways that behave like broadcast machinery. A single concept direction can be used by several downstream operations, and ablating broadcast heads selectively disrupts J-space continuity and report [2].

This is a plausible mechanistic substrate for an integrated access closure. A representation is not only selected. It is inserted into a causal regime in which multiple transformations can act on it. CCC's integration quantity Phi_cl could be operationalized by measuring how much best-achievable predictive mismatch increases when those pathways are forced to factorize. That step remains prospective.

4.5 Closure state and closure grammar

A tentative mapping can now be stated.

The active sparse J-space configuration at one layer and token position is best treated as a local coordinate of a candidate closure state. A complete access state may extend across several layer-position cells and the read/write relations that connect them.

The model's learned weights, available feature directions, layer organization, and read/write circuitry jointly constrain the family of states that can be stably instantiated. They are therefore candidates for a closure grammar.

A prompt, task instruction, or intermediate inference changes the selected state within that grammar.

Training can alter the probability of existing states, the geometry of the available frame, the interaction graph, or all three. Only the latter two warrant the stronger claim that the grammar itself changed.

This mapping is useful but not automatic. A closure grammar in CCC is a family of admissibility constraints, not simply the total parameter set of a neural network. A defensible identification requires showing that the proposed weights and circuits delimit a stable family of states and transformations across contexts.

Because the active coordinates are recovered through a particular observation grammar, the mapping remains a hypothesis about the system rather than a direct transcription of it. A richer or differently structured instrument may describe the same internal regime in different coordinates.

4.6 Four closures that must be separated

The J-space findings sharpen CCC by motivating four distinct levels of organization.

Access closure occurs when selected content becomes available for report, deliberate manipulation, and flexible inference.

Self-model closure occurs when content is organized around a relatively stable agent or Assistant perspective.

Response closure occurs when late-stage processing commits the system to an output token, action, or continuation. It concerns what becomes response-dominant, not everything that remains accessible upstream.

Phenomenal closure, in CCC's stronger vocabulary, would be the organization of actual presence into a coherent experiential point of view.

Two dissociations are visible in the present evidence. First, the J-space study reports workspace-like organization in a pretrained base model before post-training gives the Assistant's perspective a privileged role. Post-training then causes safety assessments, emotional reactions, preference conflicts, and self-monitoring signals to occupy the workspace more systematically [2]. Access architecture and self-model organization therefore appear dissociable.

Second, access closure can diverge from response closure. When the model is prefilled to select an option it ordinarily disfavors, the all-caps token BUT and related conflict directions become prominent in J-space immediately after the commitment. In most trials, however, the model continues by defending the forced choice rather than voicing the objection [2]. The conflict remains accessible while the response system commits elsewhere.

That internal conflict is not itself CCC remainder. It is a represented state omitted from the final response. It becomes relevant to remainder only if an output-only closure model fails to predict later revision, termination, or susceptibility to counter-steering, while a model that includes the access state succeeds.

The resulting nonidentity is strict: access closure is not self-model closure; access closure is not response closure; and none of these establishes phenomenal presence.

The distinction is philosophically important. A model can contain a functionally global workspace without a stable self. It can contain a self-model without phenomenal consciousness. It can register an objection without allowing that objection to govern its answer. Evidence at one level does not transfer automatically to the next.

Table 1. Evidential status of proposed J-space/CCC claims.

Claim	Status	Reason
J-space is privileged for verbal report and flexible reasoning.	Empirical finding	Causal swaps, clamping, ablation, and flexible-generalization experiments support this claim in the models tested.
J-space is a candidate local access-closure regime.	Interpretive claim	The mapping is structurally motivated by stabilization, preserved identity, selective access, and reusable transformations, but has not yet been tested with the CCC objective.
Access closure and response closure can diverge.	Empirical finding with interpretive consequence	Preference-violation trials show conflict content in J-space while the model usually continues the imposed response. The closure vocabulary describes the dissociation; it does not make the latent token remainder.
An active J-space configuration is a local coordinate of a closure state.	Prospective hypothesis	A full state may extend across several layer-position cells; fixed candidate families and predictive state scores are still required.

Claim	Status	Reason
The learned frame and read/write circuitry form a system closure grammar.	Prospective hypothesis	Weights and circuits constrain admissible states, but grammar-level adequacy and changes across training remain to be measured.
The J-lens transparently reveals the model's intrinsic workspace.	Rejected	The lens is an observation grammar constrained by vocabulary, context averaging, sparsity, and available intervention channels.
The non-J-space residual is CCC remainder.	Rejected	Most non-J-space activity is competent automatic computation. CCC remainder is best-achievable model mismatch under specified constraints.
J-space establishes phenomenal consciousness or irreducible presence C.	Rejected	The experiments concern functional access, report, and representational organization, not the occurrence of subjective presence.
J-space supports cosmic closure.	Rejected	A local computational workspace in an engineered transformer provides no direct evidence about cosmogenesis or the onset of physical law.

5. Evidential Asymmetry and the Limits of Mutual Support

The relation between the papers is evidentially asymmetric.

J-space provides empirical material that can modestly increase confidence in a minimal cognitive-closure hypothesis. If a cognitive system must flexibly report and operate on content, one would expect some selective regime that stabilizes identities and makes them broadly reusable. J-space appears to be such a regime. The evidence is not decisive because several competing theories predict similar organization, including global workspace theory, attention-schema theory, recurrent-processing accounts, integrated-information approaches, predictive processing, and generic learned-bottleneck explanations [4-6,21-24].

CCC, by contrast, supplies no independent empirical evidence that J-space exists. It does not derive the Jacobian lens, predict its sparse token-indexed frame, locate it in middle layers, or identify the attention heads that broadcast its content. CCC's support for J-space is conceptual and methodological. It explains why a system of stabilized distinctions, identities, and transformations would be theoretically important and proposes quantities through which competing interpretations could be tested.

The strongest defensible statement is therefore:

J-space is an empirical candidate that instantiates part of CCC's minimal operational vocabulary, while CCC supplies a prospective model-selection framework for testing whether that interpretation earns explanatory preference.

Several apparent correspondences should not be promoted beyond that statement.

Consistency. CCC's third closure component concerns coordination among perspectives. Most J-space experiments concern one model and one forward pass. They do not test intersubjective alignment among agents. Post-training comparisons show a privileged Assistant standpoint, not shared-world consistency.

Phenomenological time. Layer progression and token progression provide computational orderings. They do not establish that phenomenological succession is generated by closure updates or that zero remainder would arrest experience.

C and c. If c means actual localized presence with content, J-space does not establish c. It establishes an organization that could structure reportable content if presence were independently present. The cautious expression is c-structuring mechanism, not artificial consciousness.

Experiential language. J-space ablation makes descriptions of experience more mechanical and less embodied. The effect generalizes from self-description to descriptions of other people. The result therefore shows that J-space carries representations used to construct phenomenological discourse. It does not identify a site of feeling.

Cosmology. A transformer is an engineered system operating within an already stable physical and linguistic world. Its local access architecture has no direct evidential bearing on the onset of spacetime, physical law, or a first closure grammar.

The value of the comparison lies precisely in maintaining these limits. A framework gains credibility when it prevents overreach as well as when it generates new interpretations.

6. Three Different Things That Can Be Called a Residual

The most consequential terminological risk concerns remainder. At least three different quantities appear in the relevant literature, and they must not be conflated.

The preference-violation result provides a useful warning. An internal BUT that never appears in the response is not "remainder made visible." It is content represented in access closure but excluded from response closure. CCC remainder is not an omitted word or a latent objection; it is a comparative mismatch between an empirical target and the best model allowed by a specified closure. The conflict signal becomes evidence about remainder only when including it measurably improves held-out prediction of later behavior or intervention effects.

6.1 The non-J-space component

For activation h and its nearest sparse J-space approximation h_J, the J-space paper defines a leftover:

r_J = h - h_J.

This is an algebraic residual. It is large. Most activation variance lies in it. It contains grammar, formatting, token bookkeeping, automatic classifications, numerical mechanisms, and other causally useful representations. In some tasks it carries the information needed for correct behavior even when that information is absent from J-space. It is therefore neither error nor unstructured openness.

The appendix of the J-space paper sometimes calls nodes carrying this content "remainder terms" in a J-lens attribution graph. In the present manuscript they should be called residual terms. The name remainder is reserved for CCC's model-relative quantity.

6.2 Reconstruction error and error nodes

Sparse autoencoders, cross-layer transcoders, template lenses, oracle lenses, and natural language autoencoders all leave reconstruction error. Anthropic's circuit-tracing work represents the portion of an MLP output not reconstructed by a cross-layer transcoder as an error node. Error nodes make visible that an interpretability dictionary is incomplete. They can dominate out-of-distribution prompts and obscure causal paths [8].

These errors are methodologically important, but they are still not automatically CCC remainder. They can arise because the dictionary is too small, the training data are unrepresentative, the model family is misspecified, the activation is nonlinear or relational, or the human label is too coarse. Those are primarily candidates for estimation and approximation remainder.

Circuit tracing makes another crucial point. A replacement model can match ordinary activations or outputs while using a different mechanism. Even exact local reconstruction does not guarantee mechanistic faithfulness. Anthropic therefore validates attribution graphs with interventions and reports that perturbation discrepancies can compound across layers [8]. This shows why output fit alone is insufficient for a Closure test.

6.3 CCC remainder

CCC remainder is:

$$R(K \mid p) = \min_{q \in Q(K)} D_{KL}(p \| q).$$

It is defined only after specifying:

1. the empirical target p;
2. the candidate closure K;
3. the grammar of admissible candidates;
4. the model family Q(K);
5. system-access, observation-access, and resource limits;
6. the complexity scheme used to prevent a permissive model from winning trivially.

The quantity is not the norm of a leftover vector. It is the best distributional mismatch achievable by a model that respects the candidate closure.

This distinction changes the research question. The question is not, "How much activation variance is outside J-space?" The answer is already known to be most of it. The question is, "How well can the strongest complexity-controlled model constrained by J-space explain the observables that an access closure is supposed to explain, especially under causal intervention?"

A small J-space variance fraction can coexist with a small functional remainder if those few dimensions carry the information relevant to report and flexible action. Conversely, a visually interpretable J-space can have high remainder if it fails to predict intervention effects or generalize outside selected demonstrations.

7. A Frame-and-Graph Closure Proxy with an Explicit Observation Grammar

CCC's original proxy represents a closure by a partition and sparse graph. J-space exposes two limitations of that choice. First, transformer representations are superposed and overcomplete: J-space is a sparse frame or union of cones, not a partition into disjoint blocks. Second, the interpretability instrument cannot be folded into the system it is meant to measure. What the model can internally access and what an investigator can observe are different constraints.

We therefore represent a candidate system closure as:

$$K_{sys} = (F_k, E, A_{sys}). \quad (4)$$

F_k is a sparse frame or union of cones defining admissible content states. E is a sparse directed graph specifying permitted read, write, and broadcast relations among features, layers, and token positions. A_{sys} specifies the internal channels through which model components can access and transform those states.

The investigator's observation grammar is represented separately:

$$O = (D_{obs}, M_{obs}, A_{obs}). \quad (5)$$

D_{obs} is the vocabulary or feature dictionary available to the instrument. M_{obs} is the measurement map, such as an averaged Jacobian, sparse-autoencoder encoder, template decoder, or natural-language bottleneck. A_{obs} specifies the layers, positions, logits, weights, and intervention channels available to the investigator.

This separation is consequential. The same model and the same candidate system closure can be studied under several observation grammars. Moving from a single-token J-lens to a phrase lens changes O, not K_{sys}. If measured remainder falls, the first interpretation is that observer-side estimation or approximation improved. The model's internal closure has not changed merely because the instrument became more expressive.

7.1 Three grammars

Three kinds of grammar must be kept distinct.

System grammar, Omega_sys, constrains which internal states, identity relations, and transformations the trained model can stably implement. Its candidates include the learned feature geometry, read/write circuitry, and task-sensitive access policies.

Output grammar, Omega_out, constrains what the model can directly emit or do. For a text-only language model, the tokenizer, vocabulary, autoregressive interface, and available tool calls delimit the action space. This output grammar may help explain why a verbalizable format is computationally privileged, but it is not identical to the system's full representational organization.

Observation grammar O constrains what researchers can detect, name, and manipulate. The J-lens, a sparse autoencoder, an attribution graph, and a natural-language autoencoder impose different observational distinctions. Agreement among them is evidence; no single instrument is ontologically privileged by definition.

The three grammars can align, but they need not. A token-indexed J-space may reflect genuine organization shaped by the model's output grammar while still being incompletely described by a token-indexed observation grammar. The distinction prevents the vocabulary of the instrument from being mistaken for the final vocabulary of the system.

7.2 Closure states under an observation grammar

Let V_l^O contain the feature directions supplied by observation grammar O at layer l. For activation h_l,t, define sparse observed coordinates:

$$z_{l,t}^{O} = \arg\min_{z \geq 0, \|z\|_0 \leq k} \|h_{l,t} - V_l^O z\|^2. \qquad (6)$$

For the ordinary J-lens, V_l^O is the token-indexed J-frame. The selected active set and coefficients are an observed coordinate of a candidate access-closure state. They should not automatically be identified with the whole state. A complete state may extend across several layers and token positions, include relations among concepts, and depend on automatic subclosures not represented in the sparse readout.

7.3 The empirical target

The cleanest behavioral target is the original model's conditional output distribution:

$$p_M(y \mid x)$$

where x is a prompt and y is the next token or a later response segment. For an open-weight model, the complete next-token distribution can be read directly, avoiding sampling error at the output level.

Because J-space is claimed to mediate report and flexible reasoning rather than all model computation, the primary prompt distribution should contain tasks for which that claim applies. Automatic continuation and classification tasks should serve as negative controls, not as the sole target. A full model of the transformer would require additional closures for automatic processing and late response commitment.

7.4 Behavioral remainder

Let q_K,O be a resource-bounded surrogate that receives only the state and connections allowed by K_sys as measured through O. Define:

$$R_{out}(K_{sys}; O) = E_{x \sim D} D_{KL}[p_M(. \mid x) \,||\, q_{K,O}(. \mid z_{K,O}(x))]. \qquad (7)$$

The minimization over q_K,O is implicit: the surrogate is trained to its best held-out performance within a preregistered family. D is divided into training, validation, and untouched test distributions. This quantity asks whether the selected system closure, as accessed through a specified instrument, preserves what matters for behavior. It does not require reconstructing the entire residual stream.

7.5 Intervention remainder

A surrogate can fit ordinary outputs through shortcuts. A stronger test includes causal interventions. Let I be drawn from a specified ensemble, such as concept swaps, direction ablations, head ablations, or small perturbations. Let p_M^I be the original model's output distribution after intervention and q_K,O^I the surrogate's prediction. Define:

$$R_{int}(K_{sys}; O) = E_{x,I}\, D_{KL}[p_M^I(. \mid x) \,||\, q_{K,O}^I(. \mid z_{K,O}(x))]. \qquad (8)$$

An optional internal term can compare predicted and observed downstream feature changes. The primary principle is that a closure model must predict how the system responds when its proposed distinctions and connections are altered. This incorporates the mechanistic-faithfulness lesson from circuit tracing [8,9].

7.6 Complexity grounding

Without a structural penalty, a model with full activations and unrestricted edges can fit best by construction. CCC's tradeoff terms must therefore be grounded rather than tuned until a preferred result appears.

Minimum description length offers a natural implementation [14,15]. Express output KL in nats per observation and add the code length needed to specify both the candidate closure and, when instruments are being compared, the observation grammar:

$$S_{MDL}(K_{sys}; O) = R_{out}(K_{sys}; O) + [L_{code}(K_{sys}, q_{K,O}) + L_{code}(O)] / N. \qquad (9)$$

The code can include the cost of identifying active atoms, quantized coefficients, graph edges, layer and token-position specifications, decoder parameters, task-specific side information, and the dictionary or measurement map used by O. If the same observation grammar is fixed across candidates, L_code(O) is a constant and drops out. If observation grammars differ, its cost must be counted rather than treated as free access.

A fixed coding scheme allows fit and structural complexity to be compared in common units. Sensitivity ranges should be preregistered. If the preferred closure changes only when code costs are adjusted post hoc, the result is not explanatory.

7.7 Estimation, approximation, and closure remainder

The three sources of observed remainder can now be assigned more precisely.

Estimation remainder R_e arises from finite prompts, noisy logits, imperfect Jacobian estimates, random initialization, and unstable fitting. It should shrink with repeated estimates, bootstrap control, larger samples, and more reliable measurement.

Approximation remainder R_a arises when O or the surrogate family cannot express the relevant structure. It is probed by enriching the observation grammar: single-token J-lens, context-conditioned J-lens, phrase or template lens, oracle or natural-language lens, sparse autoencoders, cross-layer transcoders, nonlinear decoders, and full-activation controls. If remainder falls when D_obs or M_obs becomes more expressive, the original mismatch was at least partly observer-side approximation error.

Closure remainder R_c is the excess mismatch caused by the nontrivial system constraints after data, observation access, and decoder capacity are controlled. Here R denotes a preregistered target, such as R_out, R_int, or a fixed combination of them. Let U be a resource-matched surrogate with the same observation grammar, data, and parameter budget but without the sparse-frame or graph restriction. Then:

$$R_{c\_hat}(K_{sys}; O) = R(K_{sys}; O) - R(U; O). \qquad (10)$$

This is an operational contrast, not an exact metaphysical decomposition. It is defensible only when U is genuinely matched and the candidate families are nested or otherwise comparable. Richer observation can reduce R_e and R_a; it does not by itself eliminate the possibility that a nontrivial system closure still imposes R_c.

7.8 Grammar-level remainder

A system grammar Omega_sys defines a family of frames, capacities, internal access policies, and edge structures. Under a fixed observation grammar O, its remainder is:

$$R_{\Omega_{sys}}(p_M; O) = \min_{K_{sys} \in K(\Omega_{sys})} R(K_{sys}; O \mid p_M). \quad (11)$$

Persistently high grammar-level remainder indicates that adjusting active states within the current system grammar is not enough. The grammar may need multi-token features, relational binding, nonlinear manifolds, additional modalities, or explicit automatic-processing subclosures. Persistently high remainder that disappears only when O is enriched instead indicates an inadequate observation grammar. The distinction converts a broad philosophical contrast into a model-comparison problem.

8. Anthropic's Interpretability Methods as an Observation and Access Ladder

The relevant Anthropic research does not provide one definitive view of model internals. It provides several instruments with different vocabularies, access channels, costs, and failure modes. Each instrument defines an observation grammar. Some form approximately nested expansions of access; others are non-nested alternatives that must be compared by held-out fit, causal faithfulness, and description length.

8.1 Output, chain of thought, and two-axis computation

The weakest observation condition sees only what the model says. This is inexpensive but incomplete. Anthropic's chain-of-thought research finds that factors which causally alter an answer are often omitted from the written reasoning, including in reward-hacking settings [12]. Output-only descriptions therefore provide a useful high-remainder baseline. They should not be treated as transparent records of internal computation.

Transformer computation unfolds along two axes. Within one forward pass, information is transformed across depth. Across autoregressive steps, information written at one token position becomes context that later positions can attend to and transform. The J-space study reports that explicit chain-of-thought mathematics is substantially more robust to J-space ablation than the same problems answered directly, consistent with the model externalizing intermediate results into the token sequence [2].

Once an intermediate is written, it is not lost merely because it is now in the past. It remains available in the context until it is truncated or otherwise inaccessible. In Closure terms, computational closure time tau can therefore be ordered over a layer-position graph rather than identified with layer depth alone. Chain of thought is both a record and an intervention: by writing an intermediate state, the model changes the state available to subsequent computation. It is not a passive window onto a computation that would otherwise have remained unchanged.

8.2 The J-lens as an observation grammar

The J-lens is cheap after its averaged Jacobian has been computed. It produces token-indexed, causally grounded directions and supports direct read/write interventions. Those advantages make it an unusually useful observation grammar for reportable and flexibly reusable content [2].

It is not a transparent transcript of an intrinsic workspace. Its observation grammar includes the tokenizer vocabulary, the corpus over which the Jacobian is averaged, the choice to include present and future output effects, the sparse nonnegative decomposition, the selected layer band, and the ranking rule used to display tokens. Each choice affects which distinctions become visible.

Context averaging requires particular care. The averaged Jacobian seeks a direction with a general disposition to influence verbalization across contexts, but it is still applied to a context-specific activation. Averaging therefore does not simply remove contextual content. It may, however, suppress context-specific causal maps that fail to survive the average. A concept absent from the J-lens may be absent from access closure, represented outside the token frame, or hidden by the observation map.

A direct control is to compare a global lens J_l with context-conditioned or context-clustered lenses $J_{l,c}$ on held-out prompts. If conditioned lenses reduce output and intervention remainder at matched complexity, the global lens incurred approximation remainder. If they merely overfit the contexts used to construct them, the apparent gain should disappear out of sample.

8.3 Template and oracle extensions

The J-space paper introduces template directions for multi-token concepts and an oracle-style lens that generates free-form phrases whose reconstructed directions explain activation structure [2]. These methods enlarge verbal access. If they reduce remainder on multi-token and relational tasks while preserving intervention accuracy, the difference should be assigned to approximation or access limits of the original J-lens, not to closure remainder.

The extensions also introduce new risks. Template methods can skip toward the answer, and free-form language decoders can generate plausible but ungrounded descriptions. Their role is comparison, not adjudication by authority.

8.4 Sparse autoencoders

Sparse autoencoders learn large dictionaries of feature directions. Anthropic reports millions of interpretable features in a production-scale model, including abstract, multilingual, and safety-relevant features [7]. SAEs supply a broader candidate frame than the token-indexed J-lens and can reveal automatic or nonverbalizable structure.

They remain incomplete. A feature dictionary can miss mechanisms, split one concept across several features, merge distinct concepts, or reconstruct activations without recovering the true causal organization. SAE reconstruction error is therefore a measure of instrument adequacy, not CCC remainder.

8.5 Cross-layer transcoders and attribution graphs

Circuit tracing replaces MLP computation with sparse cross-layer features and constructs prompt-specific attribution graphs linking embeddings, features, error terms, and output logits [8,9]. This work is especially relevant to CCC because it supplies a candidate interaction graph E and explicit measures of graph completeness.

It also demonstrates the central difficulty. Error nodes represent activity the replacement model does not explain. A local error-corrected replacement can reproduce the original forward pass, yet its response to interventions may diverge from the underlying model. Mechanistic faithfulness must therefore be measured, not assumed. In CCC terms, graph reconstruction error and graph incompleteness are candidates for R_e and R_a until they survive observation expansion and perturbation controls.

8.6 Natural language autoencoders

Natural language autoencoders train one model to verbalize an activation and another to reconstruct the activation from that text. Their objective makes natural-language explanation an information bottleneck rather than an after-the-fact label [11]. They can surface rich relational descriptions and have been used in model auditing.

Their text can still be noisy or confabulatory. A verbalizer and reconstructor can coordinate on a code that sounds meaningful without perfectly matching the original model's causal concepts. NLA reconstruction quality and intervention prediction should therefore be evaluated separately. NLAs are a richer observation grammar, not a privileged truth oracle.

8.7 Causal introspection

Anthropic's introspection experiments inject known concept representations and ask models to identify or control them. Some models can report injected concepts, distinguish prior internal states from text inputs, and modulate representations on instruction, but the abilities are unreliable and context-dependent [10]. These experiments help validate access relations between internal state and report. They do not establish human-like introspection or phenomenal awareness.

8.8 The observation-expansion prediction

CCC states that increased access should generally reduce best-achievable remainder without arbitrary retuning. With the system/observation distinction in place, this becomes an observation-expansion prediction.

For genuinely nested observation grammars, enlarging D_obs or A_obs should weakly reduce unpenalized held-out remainder because the richer instrument can reproduce the poorer one. After description-length penalties are

included, the added access must earn its cost. Non-nested methods, such as J-lenses, sparse autoencoders, and natural-language bottlenecks, require matched MDL comparisons rather than a simple monotonic ordering.

A practical sequence is:

- output only;
- global single-token J-lens;
- context-conditioned and multi-token J-lens variants;
- phrase, template, or oracle observation grammars;
- sparse-autoencoder or cross-layer feature dictionaries;
- attribution graphs and richer intervention channels;
- full residual-stream and weight access as a high-capacity control.

A systematic failure of richer, properly nested observation to reduce predictive and intervention remainder would weaken the interpretation of the mismatch as observer-side limitation. It would suggest that the instruments are not ordered as assumed, that the model families are incomparable, or that the remainder framework is not tracking the claimed structure. Full access is not automatically the best explanation: it must still pay for complexity and demonstrate causal generalization.

Table 2 summarizes the main instruments and their failure modes.

Table 2. Anthropic interpretability methods as an observation and access ladder.

Observation method	Additional access supplied	Principal risk or limitation
Output and chain of thought	Observable answers and written reasoning; written intermediates also become new computational context.	Causal factors may be omitted, and chain of thought changes the computation rather than passively reporting it.
Global J-lens and J-space	Cheap token-indexed readout with direct concept-level interventions.	Single-token observation vocabulary, local linearization, context averaging, bag-of-concepts structure, and incomplete coverage of automatic processing.
Context-conditioned, template, and oracle lenses	Context-specific causal maps, multi-token words, and free-form phrase descriptions of activation content.	Conditioned maps can overfit; template methods may skip toward outputs; phrase decoders may be noisy or confabulatory.
Sparse autoencoders	Large learned dictionaries that can expose abstract and nonverbalizable features.	Feature splitting, merging, dead or missing features, and reconstruction without mechanistic faithfulness.
Cross-layer transcoders and attribution graphs	Prompt-specific feature interactions, candidate causal edges, and explicit unexplained error nodes.	Replacement-model errors, graph pruning, missed mechanisms, and intervention effects that may diverge from graph predictions.
Natural language autoencoders	Expressive natural-language bottlenecks trained to reconstruct activations.	Verbalizer-reconstructor coordination can produce plausible text that is not fully causally faithful.
Full activations and weights	Maximum observation access available to the investigator, without an interpretive bottleneck.	High dimensionality and model capacity make explanation and complexity control difficult; unconstrained fit can be trivial.

9. A Concrete Experimental Program

The comparison becomes scientifically substantive only when it produces experiments that could fail. The following program is designed for an open-weight transformer with access to residual streams, logits, J-lens code, and at least one broader feature dictionary. Exact replication on Claude would require Anthropic's internal access, but the central tests can be implemented on supported open models.

9.1 General design

Prompt distributions should be divided into four families.

Access tasks require explicit report or arbitrary context-specified operations. Examples include two-hop reasoning, concept report, flexible entity transformations, covert arithmetic followed by a novel operation, and introspective report of injected concepts.

Automatic tasks use the same underlying information without requiring its explicit naming or flexible rerouting. Examples include language continuation, anomaly detection, line wrapping, local grammatical classification, and routine next-token prediction.

Ambiguity tasks interpolate between two candidate concepts under varied carrier contexts.

Out-of-distribution tasks include multi-token concepts, multilingual prompts, obfuscated inputs, long contexts, and concept combinations not used to estimate the lens.

Every analysis should use untouched held-out prompts. Candidate closures should be specified before outcome inspection. Random rotations, matched random directions, alternative lenses, and capacity-matched decoders should be included from the beginning.

9.2 Test 1: Does J-space minimize access-task remainder?

Evaluate candidate pairs $H_i = (K_{sys}^i, O_i)$, not unqualified representations. Each pair asserts that an observed frame tracks a causally relevant coordinate system in the model. The same resource-bounded surrogate architecture and coding convention should be used throughout.

The comparison set should include:

- H_J: a k-sparse nonnegative J-frame observed through the global J-lens;
- H_{rot}: a fixed orthogonal rotation of the J directions with broad geometry and capacity preserved;
- H_{logit}: logit-lens directions under the same layer and decoder budget;
- H_{SAE}: capacity-matched sparse-autoencoder features;
- H_{phrase}: expanded phrase or template directions;
- H_{full}: a compressed full-residual control with the same parameter and description-length budget.

The rotated and logit controls ask whether the J-frame contributes more than generic compression or vocabulary projection. The SAE, phrase, and full-residual candidates test whether the J-lens observation grammar omits causally relevant structure. None is treated as the system ontology in advance.

The primary outcome is held-out output KL on access tasks. The secondary outcome is intervention KL under concept swaps and ablations. Each representation must be paired with an explicit observation grammar; when those grammars differ, their dictionary, fitting, and decoder costs enter the same description-length convention.

Prediction. H_J should beat H_{rot} and H_{logit} at matched complexity on report and flexible reasoning. H_{phrase} should improve on H_J for multi-token concepts. H_{SAE} may outperform J-space on tasks whose relevant computation is automatic or poorly aligned with verbal output. H_{full} may fit best before penalties but should pay a larger description-length cost.

Disconfirmation. J-space does not outperform rotated or generic sparse controls; its advantage disappears on untouched tasks or interventions; or its fit depends on choosing sparsity, layer bands, observation map, or penalties after seeing the results.

This test would supply the first direct measurement of J-space remainder in the CCC sense. Existing variance-explained figures do not, and an instrument earns system-level interpretation only by predictive and intervention success.

9.3 Test 2: Does observation expansion reduce remainder?

Use two complementary forms of observation expansion.

Vocabulary expansion holds the causal map approximately fixed while enlarging what can be named:

- single-token J-frame;
- single- and multi-token template frame;
- phrase frame;
- a converged oracle-generated frame.

Observation-map expansion tests whether a globally averaged lens conceals context-sensitive structure:

- global averaged Jacobian J_l;
- context-clustered or task-conditioned Jacobians J_l,c;
- language-conditioned frames and, in cross-model comparisons, tokenizer-conditioned frames on semantically matched multilingual tasks.

Use the one-sided frame distance proposed in the J-space paper to verify approximate containment where possible. At every step, refit the same decoder, evaluate untouched prompts, and retain the same code-length convention. Context-conditioned instruments must be charged for the information used to select the context class.

Prediction. Output and intervention remainder should decrease on concepts and relations newly expressible at the richer level. Vocabulary expansion should help most on multi-token and relational tasks. Context-conditioned maps should help only where the global average suppressed a stable context-specific causal relation. Semantically matched single-token controls should change less.

Interpretation. The reduction estimates estimation or approximation remainder in the original observation grammar. The remainder that persists after observation expansion and matched decoding is a better candidate for mismatch imposed by the system closure.

Disconfirmation. Richer instruments do not improve held-out fit, improvements occur only on construction prompts, context-conditioned lenses overfit, or the apparent gain disappears when code length and decoder capacity are matched.

9.4 Test 3: Does Delta S track commitment under ambiguity?

For each ambiguous prompt, define two candidate closure states K_A and K_B corresponding to the two concept interpretations. At each layer, fit the best state-specific surrogate and compute:

$$\Delta S_l = S(K_{second} \mid p_l) - S(K_{best} \mid p_l). \qquad (12)$$

The score can combine sparse reconstruction, output predictive KL, and fixed state complexity. Candidate definitions must be set before examining the transition.

Prediction. Early layers should show a small score gap. Near the J-space onset, the gap should widen as one interpretation becomes causally dominant. At maximally ambiguous inputs, individual prompts should display large gaps in opposite directions, producing bimodality across contexts. The sign and magnitude of Delta S should predict which concept survives a weak counter-steer and how much intervention is required to force a switch.

The transition type must then be tested rather than named. Fit rival effective descriptions to held-out trajectories: smooth nonlinear classification, piecewise-linear thresholding, stochastic winner-take-all selection, cusp-like bistability, and saddle-node-like dynamics [25]. A specific bifurcation claim requires an explicit state variable and effective update rule, not the visual impression of a sharp boundary.

A saddle-node-like account would receive distinctive support only if the inferred dynamics exhibit the relevant fixed-point structure and predict additional signatures, such as increased susceptibility near the boundary, reduced

recovery rate, perturbation-dependent basin crossing, or hysteresis where the construction permits it. Simpler models should be preferred when they predict equally well at lower description length.

Disconfirmation. Delta S does not widen near behavioral commitment, rotated frames show the same effect, the gap fails to predict perturbation robustness, or a simpler classifier explains held-out transitions as well as the proposed dynamical model.

9.5 Test 4: Does forced factorization produce Phi_cl?

Use the identified J-space broadcast heads and MLP read/write paths to define an integrated graph E. Construct factorized controls by:

- removing the top J-broadcast heads;
- blocking cross-token propagation of selected J coordinates;
- separating early J-state formation from downstream readers;
- randomly removing an equal number of layer-matched heads or edges.

For each condition, refit the best surrogate permitted by the factorized graph. Define:

$$\text{Phi\_cl} = \text{R\_factorized} - \text{R\_integrated}. \quad (13)$$

Prediction. Phi_cl should be positive and substantially larger for the targeted factorization than for matched random factorization on report and flexible reasoning. It should be smaller on automatic tasks that bypass J-space.

The existing broadcast-head ablations already show selective behavioral effects. The proposed test adds the missing model-comparison step and determines whether the interaction graph earns its complexity cost.

Disconfirmation. Targeted factorization raises remainder no more than random damage, or the effect reflects general loss of model quality rather than the proposed access functions.

9.6 Test 5: Can a stabilization rate kappa be measured?

At a selected layer and token position, apply controlled perturbations of increasing strength toward a competing concept. Track the selected state, output distribution, and predictive remainder through later layers and subsequent token positions.

Define kappa over a preregistered interval as the negative slope of excess remainder after perturbation:

$$\kappa = -d[R_{perturbed}(\tau) - R_{clean}(\tau)] / d\tau. \quad (14)$$

Prediction. Weak perturbations should often be corrected or absorbed, producing positive kappa toward the original state. Stronger perturbations should cross a threshold and settle into the alternative state. The threshold should correlate with Delta S. Disrupting broadcast or self-correction circuitry should reduce kappa.

The ambiguity boundary supplies a stronger transition test. If the commitment reflects critical or bifurcation-like dynamics, susceptibility should increase and recovery should slow as the input approaches the boundary. In that case kappa, or another preregistered recovery statistic, should decline near the transition. A generic high-gain classifier need not show the same recovery profile.

A feedforward transformer does not recover through literal recurrence within one layer. Here tau is an ordering over a layer-position graph: later layers in the current pass and later token positions in subsequent passes. The interpretation remains computational, not phenomenological.

Disconfirmation. Recovery slopes are unstable across prompts, depend entirely on arbitrary distance metrics, show no relation to state confidence and causal circuitry, or fail to distinguish the proposed transition account from simpler nonlinear alternatives.

9.7 Test 6: State change or grammar change under training?

Compare a base model, a post-trained Assistant model, and a counterfactual-reflection fine-tune. Fit the same candidate grammar to all checkpoints.

Three outcomes are possible.

State redistribution. The same frame and graph explain all checkpoints, but the probabilities of particular states change.

Graph change. The feature frame remains comparable, but new read/write relations are needed.

Frame change. New directions or concept combinations are required to achieve comparable remainder.

Use the J-space paper's frame distance, cross-check it with CKA or aligned feature matching, and compare grammar-level remainder under a fixed versus refitted grammar.

Prediction. Ordinary post-training may largely reweight Assistant-perspective states, while reflection training may either strengthen existing ethical directions or introduce a more substantial grammar modification. The current causal ablation results show that implanted ethical directions mediate behavior, but they do not by themselves decide whether the grammar changed [2].

Disconfirmation. Apparent grammar change disappears after alignment of feature frames, or fixed-grammar remainder rises only because of poor cross-checkpoint estimation.

9.8 Test 7: Report and response remainder

Use prompt hints, hidden variables, or forced commitments known to change the model's answer. Compare three closure candidates:

- K_surface, based only on written chain of thought and final output;
- K_access, based on J-space states as well as output;
- K_rich, based on J-space plus NLA, SAE, or attribution-graph access.

The target is the distribution of answer changes under controlled prompt and activation interventions. Preference-violation trials provide a particularly clean access-response dissociation. Before the conflict readout is treated as explanatory, its candidate directions should be tested by ablation, steering, clamping, and mediation analysis against matched controls. The model can then be given controlled opportunities to terminate, revise, explain, or respond to a weak counter-steer.

Prediction. K_surface should have higher intervention remainder when a causally relevant factor is not verbalized. K_access should improve prediction when conflict directions such as BUT forecast revision, termination, or sensitivity to counter-steering. K_rich should improve further when the relevant factor is represented outside the ordinary J-frame.

The latent conflict token is not itself remainder. Remainder is the difference in best-achievable predictive mismatch between the response-only and access-informed closure models. This test links Anthropic's chain-of-thought findings to CCC's local-access idea without assuming that every hidden causal factor appears in J-space.

Disconfirmation. Internal methods do not predict interventions better than surface report, conflict signals have no out-of-sample relation to later behavior, or apparent success occurs only because the hidden factor is available in prompt text to the surrogate.

9.9 Test 8: Grammar inadequacy and automatic cognition

J-space is not expected to explain every model behavior. To test grammar-level inadequacy, evaluate tasks known to proceed outside J-space. If K_J has high remainder but an SAE or circuit-based grammar has low remainder, the result supports nested closures rather than J-space universality.

A complete model might require:

- an access closure for flexible report;
- automatic syntactic and semantic subclosures;
- a self-model closure installed by post-training;
- a response or motor closure in final layers.

The nested-closure hypothesis predicts that different tasks will select different but interacting regimes. It is disconfirmed if a single unconstrained representation wins only by absorbing every function without stable structure, or if no reproducible task-dependent organization appears.

Table 3 summarizes the core tests and their failure criteria.

Table 3. Proposed tests, predictions, and disconfirming outcomes.

Test	Predicted result	Disconfirming result
Access-task remainder	At matched description length, J-space outperforms rotated and generic sparse controls on report and flexible reasoning, especially under causal intervention.	No held-out advantage over matched controls, or an advantage that depends on post hoc choices of sparsity, layer range, decoder, or penalty.
Observation expansion	Multi-token or context-conditioned observation grammars reduce output and intervention remainder only where the global single-token lens lacks relevant access.	Richer observation does not improve held-out fit, improves only construction prompts, overfits context classes, or loses its advantage after complexity matching.
Ambiguity gap and transition model	The best-versus-second-best score gap widens near commitment, predicts counter-steering resistance, and any specific dynamical model outperforms simpler rivals on held-out perturbations.	The gap fails to track commitment, rotated controls perform similarly, or ordinary nonlinear classification predicts as well as the proposed bifurcation account.
Integration (Phi_cl)	Targeted factorization of J-space broadcast paths raises remainder more than layer-matched random factorization on access tasks.	Targeted factorization is no worse than random damage, or the result is explained by nonspecific model degradation.
Stabilization rate (kappa)	Weak perturbations are corrected downstream; recovery and susceptibility covary with state confidence, distance from the ambiguity boundary, and relevant circuitry.	Recovery is metric-dependent, unstable across prompts, unrelated to state confidence, or fails to distinguish rival transition models.
Training and grammar	Checkpoint comparisons distinguish state redistribution from changes in the feature frame, internal access policy, or interaction graph.	Apparent grammar change vanishes after principled frame alignment or reflects only estimation failure.
Report and response remainder	Access-informed models predict revision, termination, and answer changes under hidden causal factors better than written chain of thought or final output alone.	J-space and richer internal methods fail to outperform surface report on untouched intervention tests.
Grammar inadequacy and automatic cognition	J-space has high remainder on automatic tasks while broader feature or circuit grammars recover the missing structure, supporting nested closures.	No reproducible task-dependent organization appears, or one unconstrained representation wins only by absorbing every function without stable structure.

10. Controls, Alternative Explanations, and Failure Modes

A credible Closure interpretation must survive explanations that do not invoke Closure terminology.

10.1 Next-token leakage

The J-lens is constructed from effects on future output. It could merely reveal an early next-token predictor. The J-space paper addresses this by showing intermediate concepts that differ from the eventual answer, by comparing with tuned and logit lenses, and by demonstrating that intermediate swaps take effect earlier than answer swaps [2]. New remainder tests must preserve that separation. Motor-layer signals should be analyzed separately from workspace-layer states.

10.2 Correlation without use

A decodable feature can be present without being used by the model. Causal swaps, clamping, and mediation analyses are therefore indispensable. The surrogate must predict intervention effects, not merely correlate with clean outputs.

10.3 Generic low-dimensional bottleneck

Any sparse representation may appear efficient. Rotated J directions, random sparse frames, PCA, task-specific probes, and capacity-matched SAEs are required controls. A J-space advantage must be semantic, causal, and cross-task, not merely a consequence of compression.

10.4 Instrument-induced categories and grammar conflation

The system grammar, output grammar, and observation grammar can be conflated. A model may organize a distinction internally because it is useful for computation; it may privilege a verbal format because its action space is token-based; and the investigator may then recover that distinction because the J-lens uses the same vocabulary. Those three facts are related but not identical.

J-lens averaging, tokenizer boundaries, template generation, NLA language, sparse-feature dictionaries, and human labels can impose the categories researchers later "discover." Context averaging may suppress task-specific maps; tokenization may fragment multi-token concepts; free-text decoders may force relational states into familiar prose. Such effects belong first to the observation grammar.

Controls should therefore compare global and context-conditioned lenses, alternative tokenizers or languages where feasible, multi-token frames, nonverbal feature dictionaries, and causal interventions. Cross-method convergence reduces the risk of instrument-induced categories. A concept supported by J-lens, SAE features, attribution paths, and intervention predictions is more credible than one supported by one verbalizer. When richer observation lowers remainder without changing the model, the reduction should be assigned to R_e or R_a rather than to a newly improved system closure.

10.5 Surrogate shortcuts

A high-capacity decoder can infer answers from prompt features while ignoring the proposed closure state. Input channels must be restricted and audited. Intervention tests should be out of distribution relative to surrogate training. Mechanistic predictions should include intermediate feature changes, not output alone.

10.6 Ad hoc complexity weights

If k, layer range, graph density, or MDL penalties are adjusted separately for every task, the framework loses force. The coding scheme and sensitivity range must be set in advance. A useful closure should remain preferred across a defensible range.

10.7 Data quality and distribution shift

Remainder can rise because of insufficient prompts, poor Jacobian estimation, low signal, tokenization, or out-of-distribution inputs. Sample-size curves, bootstrap intervals, matched prompt families, and explicit OOD tests are required. If the quantity tracks data quality more strongly than structure, CCC's intended interpretation is weakened.

10.8 Model specificity

The present evidence concerns particular Claude checkpoints and selected open models. A workspace may differ with architecture, scale, modality, tokenizer, or training. Cross-model replication is part of the claim, not an optional afterthought.

10.9 No privileged closure

CCC's calibration requires ambiguity when the data do not privilege a parsing. If several frames achieve statistically and description-length equivalent scores, the correct conclusion is underdetermination. The method must not force J-space to win.

10.10 Scope error

J-space is an access candidate. High remainder on automatic tasks does not by itself falsify that local claim. Conversely, success on access tasks does not justify a claim about the whole model. The target domain of each closure must be specified before testing.

11. Philosophical Implications

11.1 A model system for organized access

The philosophical importance of J-space is not that it settles whether machines are conscious. It is that it turns a traditionally elusive functional distinction into an intervention-ready object. Researchers can read selected contents, alter them, disrupt their broadcast, compare them across training, and observe which behaviors depend on them.

This tractability makes language models useful model systems for theories of access. Biological consciousness research often relies on report, lesion, imaging, and temporal correlation. In a transformer, the candidate representation can sometimes be directly manipulated. That does not make the model more conscious than a brain. It makes one aspect of its computation easier to investigate.

11.2 What support for minimal CCC would amount to

The minimal CCC claim would gain meaningful support if a frame-and-graph closure:

1. separates a candidate system closure from the observation grammar used to measure it;
2. achieves a superior fit-complexity tradeoff on the functions it purports to explain;
3. predicts causal intervention effects;
4. displays state-selection gaps under ambiguity;
5. discriminates among rival transition models rather than naming a sharp boundary after the fact;
6. shows a measurable integration cost under forced factorization;
7. exhibits reproducible stabilization after perturbation;
8. becomes inadequate in principled ways when system access or observation grammar is restricted.

This would support Closure as a useful structural vocabulary and model-selection program. It would not show that reality itself is fundamentally made of closure, and it would not establish the ontological status of remainder.

11.3 What failure would mean

If J-space does not beat matched alternatives, if richer observation fails to reduce approximation error where it should, if score gaps do not predict commitment, if transition claims collapse into ordinary classification, or if the entire result depends on ad hoc penalties, the proposed bridge fails. CCC would then need a different operationalization or should be abandoned for this domain.

Failure at the J-space level would not refute the first-person datum C. It would refute a proposed scientific model of organized content. Maintaining that separation is one of CCC's strengths.

11.4 Consciousness remains unresolved

Access consciousness and phenomenal consciousness were distinguished to prevent functional availability from being equated with experience [3]. J-space offers evidence for a functional access architecture. Anthropic's introspection results offer evidence that some models can sometimes report and control internal states [10]. Neither result tells us whether there is something it is like to be the model [17,18].

CCC's C/c distinction reinforces the same point. No third-person pattern can by itself demonstrate the occurrence of irreducible presence under the framework's own definitions. J-space can at most illuminate how content would be organized and made reportable.

11.5 The cosmological extension remains separate

The maximal CCC proposal treats cosmogenesis as the onset of a stable closure grammar. Nothing in the J-space data tests that proposal. The analogy between a transformer settling on a concept and a cosmos acquiring stable laws is too broad to carry evidence. The present manuscript therefore treats cosmology as outside its empirical scope.

11.6 A constructive-empiricist reading

The most defensible posture is constructive rather than triumphal. J-space may be called an access closure if that description organizes findings, predicts new interventions, and survives comparison with alternatives. One can accept the model without claiming that its vocabulary reveals the final ontology of cognition. This is consistent with CCC's stated philosophy-of-science posture [1,19].

The relation between the papers can now be stated precisely.

J-space supports the plausibility of local closure-like organization.

CCC supplies a way to ask whether that organization is explanatory rather than merely redescribed.

Only the proposed remainder experiments can convert that relation from conceptual correspondence into stronger empirical support.

12. Conclusion

J-space is a sparse, causally privileged family of verbalizable representations in the language models studied by Gurnee and colleagues. Its contents can be reported, deliberately modulated, used as intermediate steps, and passed to several context-dependent operations. Much automatic computation proceeds outside it. These findings justify treating J-space as a candidate local access-closure regime.

The qualification is decisive. Access closure is not self-model closure, response closure, or phenomenal closure. A model can carry an internal objection while its response system commits elsewhere. That objection is not CCC remainder. The non-J-space component, reconstruction errors, and attribution-graph error nodes are not CCC remainder either. They become relevant only after the empirical target, candidate system closure, observation grammar, surrogate family, and complexity costs have been specified.

The revised frame-and-graph proxy supplies that specification. It represents the candidate system through a sparse feature frame, causal interaction graph, and internal access structure, while representing the investigator's instrument separately through a dictionary, measurement map, and observation channels. Remainder is the best held-out mismatch achieved under those paired constraints. Context-conditioned lenses, vocabulary expansion, MDL penalties, rotated controls, alternative dictionaries, ambiguity experiments, transition-model comparisons, forced factorization, perturbation recovery, access-response trials, and training comparisons separate observer-side failure from a possible closure-imposed mismatch.

The mutual support between J-space and CCC is therefore real but limited. J-space gives CCC a tractable empirical case. CCC gives J-space a broader vocabulary of state, grammar, integration, stabilization, response commitment, and remainder. Neither paper confirms the other's strongest claims. A successful experimental program would support the minimal proposition that reportable and flexibly usable cognition depends on selectively stabilized

representational regimes. It would leave the existence of presence and the structure of the cosmos exactly where the present evidence leaves them: open.

References

1. Dietz CF. Consciousness, Closure, and the Cosmos: A consciousness-first research posture via nested closure regimes. Version 3.3. Unpublished manuscript; 2026.
2. Gurnee W, Sofroniew N, Pearce A, Piotrowski M, Kauvar I, Chen R, et al. Verbalizable Representations Form a Global Workspace in Language Models. Transformer Circuits Thread; 2026.
3. Block N. On a confusion about a function of consciousness. Behavioral and Brain Sciences. 1995;18(2):227-247.
4. Baars BJ. A Cognitive Theory of Consciousness. Cambridge University Press; 1988.
5. Dehaene S, Changeux JP. Experimental and theoretical approaches to conscious processing. Neuron. 2011;70(2):200-227.
6. Butlin P, Long R, Elmoznino E, Bengio Y, Birch J, Constant A, et al. Consciousness in Artificial Intelligence: Insights from the Science of Consciousness. arXiv:2308.08708; 2023.
7. Templeton A, Conerly T, Marcus J, Lindsey J, Bricken T, Chen B, et al. Scaling Monosemanticity: Extracting Interpretable Features from Claude 3 Sonnet. Transformer Circuits Thread; 2024.
8. Ameisen E, Lindsey J, Pearce A, Gurnee W, Turner NL, Chen B, et al. Circuit Tracing: Revealing Computational Graphs in Language Models. Transformer Circuits Thread; 2025.
9. Lindsey J, Gurnee W, Ameisen E, Chen B, Pearce A, Turner NL, et al. On the Biology of a Large Language Model. Transformer Circuits Thread; 2025.
10. Lindsey J. Emergent Introspective Awareness in Large Language Models. Transformer Circuits Thread; 2025.
11. Fraser-Taliente K, Kantamneni S, Ong E, Mossing D, Lu C, Bogdan PC, et al. Natural Language Autoencoders Produce Unsupervised Explanations of LLM Activations. Transformer Circuits Thread; 2026.
12. Chen Y, Benton J, Radhakrishnan A, Uesato J, Denison C, Schulman J, et al. Reasoning Models Don't Always Say What They Think. Anthropic; 2025.
13. Kullback S, Leibler RA. On information and sufficiency. Annals of Mathematical Statistics. 1951;22(1):79-86.
14. Rissanen J. Modeling by shortest data description. Automatica. 1978;14(5):465-471.
15. Grünwald PD. The Minimum Description Length Principle. MIT Press; 2007.
16. Elhage N, Hume T, Olsson C, Schiefer N, Henighan T, Kravec S, et al. Toy Models of Superposition. Transformer Circuits Thread; 2022.
17. Chalmers DJ. Facing up to the problem of consciousness. Journal of Consciousness Studies. 1995;2(3):200-219.
18. Nagel T. What is it like to be a bat? Philosophical Review. 1974;83(4):435-450.
19. van Fraassen BC. The Scientific Image. Oxford University Press; 1980.
20. Lawson H. Closure: A Story of Everything. Routledge; 2001.
21. Graziano MSA, Webb TW. The attention schema theory: a mechanistic account of subjective awareness. Frontiers in Psychology. 2015;6:500.
22. Lamme VAF. Towards a true neural stance on consciousness. Trends in Cognitive Sciences. 2006;10(11):494-501.
23. Tononi G. An information integration theory of consciousness. BMC Neuroscience. 2004;5:42.
24. Friston K. The free-energy principle: a unified brain theory? Nature Reviews Neuroscience. 2010;11:127-138.
25. Strogatz SH. Nonlinear Dynamics and Chaos: With Applications to Physics, Biology, Chemistry, and Engineering. 2nd ed. Westview Press; 2015.