# Compression, Dynamics, and Control in Large Language Models: Toward a High-Level Theory

Matthew Kelly
Library Management Australia

---

Abstract

This paper introduces a trajectory level of explanation for inference-time behaviour in large language models. Existing frameworks (autoregressive conditioning, mechanistic circuit analysis, and quasi-cognitive description) treat generation as a sequence of context-conditioned draws or as circuit execution. None provides the conceptual resources needed to represent directional properties of inference such as regime persistence, transition thresholds, or asymmetric resistance to perturbation. The paper argues that these limitations are not merely empirical but conceptual, and that their persistence indicates the absence of a distinct explanatory level rather than gaps in current knowledge. These limitations motivate the asymmetry coefficient $A(M, \gamma) = R_{out}(\gamma)/R_{in}(\gamma)$ as a methodological discriminator between four competing accounts of inference-time behaviour. The existing theoretical and empirical evidence supports the prerequisites for trajectory-level dynamics but does not establish the hypothesis; the asymmetry coefficient is what does the discriminating work. The structure of the test and its result space are developed at a conceptual level, and alignment gating is reframed as trajectory control over a dynamical process rather than as a fixed property of output distributions.

## 1. The Missing Level of Explanation

This paper develops a framework for understanding inference-time behaviour in large language models, taking as its starting point a compression-based account of what large language models are: systems whose internal organisation emerges from training on human-generated text and whose outputs are reconstructions produced by traversal through compressed representations of human knowledge. On this account, LLM outputs are epistemically ambiguous artifacts: products of a process that reconstructs patterns present in the training distribution without inheriting the grounding relationships between claims and evidence that normally warrant them. The key implication is a systematic surface-grounding divergence: an LLM output can be well-formed, coherent, and consistent with the training distribution while remaining epistemically ungrounded in the sense that its constituent claims are not connected to the evidence structures that would ordinarily license them. This paper takes the compression account as its starting point and addresses a distinct question it leaves open.

The compression account addresses a question of ontology: what kind of thing is a large language model? The answer, a representational compression system, is a claim about structure. It explains why these systems possess the internal organisation they do, and why their outputs exhibit the epistemic character they do. It does not explain something distinct: how the system behaves over time during inference. Knowing that a system contains a structured representational geometry is not the same as knowing how that geometry is navigated during generation. The compression account provides the map. It does not provide an account of the dynamics of traversal.

This paper proposes that the dynamics of traversal constitute a genuine explanatory level in their own right. Inference-time generation in large language models is not fully explained by stepwise conditioning or circuit execution alone; it can instead be understood as trajectory through a structured, dynamically organised representational space, a process with its own structural properties, including stable regions, transition thresholds, and path dependence. The proposal is that this space may exhibit regime structure: stable behavioural modes whose resistance to perturbation is asymmetric with respect to direction of travel. The contribution of this paper is not to introduce dynamical or geometric vocabulary to the study of LLMs since such vocabulary is already present, in fragmented form, across several strands of the recent literature. It is to integrate those strands into a unified explanatory framework for inference-time behaviour, and to derive empirical consequences from that integration that none of the individual strands generates on its own.

This is not a claim that has yet been empirically established. It is a claim that existing frameworks make impossible to investigate, because none of the dominant accounts of LLM behaviour treat inference as a process with directional dynamics. The sequence-model account treats generation as successive application of a conditional probability function. The circuit-system account treats it as the execution of learned algorithms. Neither framework contains the vocabulary needed to ask whether exit from a behavioural mode is harder than entry, whether transitions are continuous or threshold-mediated, or whether a model's trajectory through representational space exhibits path dependence characteristic of a dynamical system with stable attractors. The gap is not merely empirical. It is conceptual: the phenomena in question are invisible from within existing frameworks because those frameworks do not treat inference as the kind of thing that could have directional dynamics at all.

A second dimension of the proposed level concerns control. Recent mechanistic interpretability research has established that alignment-related behaviour is mediated by low-dimensional directions in activation space whose manipulation can reliably induce or suppress specific classes of output (Arditi et al., 2024; Zhao et al., 2025). On the account proposed here, these directions are best understood as control signals: compact interventions on a dynamical process that redirect the path of inference through representational space before that path reaches the output distribution. If safety behaviour is a property of trajectory control rather than of fixed output policy, then the fragility and context-dependence of alignment, now well-documented empirically, becomes theoretically explicable rather than merely observed.

This proposal bears directly on debates in the philosophy of science concerning the structure of explanation and the conditions under which distinct explanatory levels are warranted. The trajectory level proposed here is situated with respect to mechanistic accounts of explanation, which emphasise organised structures and activities as the basis of explanatory adequacy, as a claim about the conditions under which mechanistic decomposition alone is insufficient to capture behaviour, because the relevant explanatory properties concern directional dynamics over structured spaces rather than static organisation.

The argument also bears on levels-of-explanation approaches descending from Marr. Its claim is not that existing levels are mistaken, but that they leave a class of questions unformulable: questions about persistence, transition thresholds, and asymmetric resistance to perturbation in inference-time behaviour. The trajectory level is introduced as the minimal extension required to render those questions well-posed. In this respect, the paper contributes to the general philosophy of science by identifying a condition under which a new explanatory level is not merely convenient but necessary for the articulation and testing of hypotheses about

system behaviour.

The paper is organised as follows. Section 2 presents the four-level architecture within which the trajectory layer is situated. Section 3 develops the theoretical and empirical grounds for the trajectory hypothesis. Section 4 argues that transition costs between behavioural states are conceptually invisible within existing frameworks. Section 5 introduces the asymmetry coefficient as a discriminator between four competing accounts. Section 6 specifies the structure of the test and its result space. Section 7 develops the control account of alignment and its implications for alignment theory. Section 8 develops the philosophical argument that trajectory dynamics constitutes a genuine explanatory level. Section 9 closes by identifying the research programme this framework opens.

Before turning to the argument in detail, it is necessary to clarify the assumptions and scope under which it proceeds.

### 1.1 Assumptions and Scope

The argument begins from a compression-based account of large language models, treated here as a working premise rather than reargued in full (Kelly, 2026a). Internal representations encode statistical structure present in the training distribution without preserving the evidential relations that ordinarily license claims (Kelly, 2026b). On this view, outputs may be well-formed and distributionally appropriate while remaining epistemically ungrounded in the sense that their constituent claims are not produced by processes that track underlying evidence. The paper does not attempt to defend this account here. It treats it as a working premise in order to address a distinct question.

The target of the present argument is not the epistemic status of model outputs, nor the conditions under which they may be relied upon, but the explanatory structure of inference-time behaviour. Specifically, the paper asks whether existing frameworks for understanding generation (autoregressive conditioning, circuit-based accounts, and quasi-cognitive description) provide the conceptual resources needed to represent directional dynamics in inference, including regime persistence, transition thresholds, and asymmetric resistance to perturbation. The claim advanced is that they do not, and that this limitation is not merely empirical but conceptual.

The asymmetry coefficient introduced later in the paper should therefore be understood as a methodological proposal rather than a reported empirical result. Its role is to specify what it would mean to test for the presence of trajectory-level dynamics and to discriminate between competing accounts of behavioural persistence. The argument is conditional in a precise sense: if inference is appropriately modelled as a dynamical process over representational space, then the phenomena identified here become measurable and theoretically tractable; if it is not, the proposed measurements should fail to discriminate and the simpler accounts should be preferred.

## 2. The Architecture of Explanation: Four Analytically Distinct Levels

Given these assumptions, we can now specify the architecture of explanation required to capture inference-time behaviour. The central proposal is that large language models require explanation at four analytically distinct levels, each capturing phenomena the others cannot capture without residue. These levels are not physically isolated modules; the mechanisms they describe are entangled in the trained system, but they are separable as explanatory strata requiring distinct kinds of answer whose answers cannot be derived from one another without additional theoretical work.

### 2.1 Compression Pressure and Representational Geometry

The first level concerns the representational structure that emerges from training. Training under next-token prediction exerts compression pressure on the model's internal organisation: the system must encode the statistical structure of an extraordinarily rich distribution in a finite parameter space, and the resulting geometry reflects the structure of the training distribution rather than anything explicitly specified by the training objective. The central claim of the compression account is that this geometry constitutes a representational compression of human knowledge: one that preserves the surface structure of epistemic claims while discarding the grounding relationships that would warrant them. This is the source of epistemically ambiguous artifacts: outputs that reconstruct the appearance of grounded knowledge without possessing its epistemic substance. The first level answers the question of what kind of representational structure training produces. Its limit is that it provides a static account without addressing how that geometry is navigated during generation.

### 2.2 Inference-Time Trajectory Dynamics

The second level is the primary subject of this paper. Generation proceeds through the representational geometry produced by training, and the proposal is that this process has structural properties not derivable from the geometry alone. A high-dimensional representational space may contain regions of varying stability: areas from which trajectories are easily displaced and areas that resist displacement. If such structure exists, inference is not merely the sampling of a conditional probability distribution but the evolution of a state through a landscape with topographic properties: basins that attract and hold trajectories, boundaries that must be crossed for transitions between modes, and path dependence that makes the history of a trajectory relevant to its future. This level answers the question of how generation proceeds through representational space and whether that process exhibits regime structure.

### 2.3 Alignment Gating as Control Signals

The third level concerns the relationship between alignment training and inference-time behaviour. Recent mechanistic interpretability research has established that refusal and safety-related behaviour are mediated by low-dimensional directions in activation space whose manipulation can reliably induce or suppress behavioural modes (Arditi et al., 2024), and that harmfulness judgment and refusal execution appear to be encoded as separable internal representations, suggesting a two-stage control architecture (Zhao et al., 2025). On the account proposed here, these findings are best understood not as describing properties of the output distribution but as describing control signals that act on inference trajectories: interventions that redirect the path of generation through representational space before that path reaches the output distribution. This level answers the question of how alignment training intervenes in inference-time dynamics, and its explanatory scope extends to the fragility and context-dependence of alignment behaviour.

### 2.4 Observable Output Phenomena

The fourth level is not a mechanism but a set of observable consequences that the preceding three levels jointly explain. Epistemically ambiguous artifacts arise from the compression-geometry level. Epistemic divergence, the systematic mismatch between internal representations and produced outputs documented by the alignment-faking and

unfaithful-reasoning literatures (Greenblatt et al., 2024; Turpin et al., 2023), arises from the interaction of the trajectory and control levels. Behavioural regime persistence arises from the trajectory level.

### 2.5 The Entanglement of Levels

These four levels are analytically distinct but not mechanistically independent. Alignment training does not add a control layer on top of a pre-existing geometry; it modifies the geometry itself, changing the representational structure through which inference trajectories subsequently pass. The more accurate picture involves feedback: training shapes geometry, alignment training reshapes geometry while also introducing compact control signals, and the resulting inference dynamics are properties of an entangled system. In designed systems, the separation of levels corresponds to separable engineering decisions. In trained systems, the separation of levels is an analytical achievement, not a structural given. The four-level architecture proposed here should therefore be understood as a framework for asking distinct questions about a system in which those questions are answered by the same underlying mechanism.

## 3. Grounds for the Trajectory Hypothesis: Theory and Evidence

The trajectory hypothesis requires support on two fronts: theoretical grounds for expecting attractor-like dynamics given the architecture of the transformer, and empirical evidence that behaviour consistent with regime structure is observable. Neither body of evidence individually establishes the hypothesis. Both together establish that the paper is investigating a predicted phenomenon with documented behavioural correlates rather than speculating about an unknown possibility. That the combined case does not discriminate between the trajectory hypothesis and its alternatives is precisely the function of the asymmetry coefficient, introduced in Section 5.

### 3.1 Energy Landscapes in Transformers

The relationship between transformer attention and attractor dynamics is not merely analogical. Ramsauer et al. (2021) demonstrated that the transformer attention mechanism is mathematically equivalent to the update rule of a modern Hopfield network: an energy-based associative memory system in which stored patterns correspond to energy minima and retrieval corresponds to iterative descent toward those minima. The formal equivalence is exact: the softmax attention update is the fixed-point iteration of a Hopfield network with continuous states and exponential storage capacity. Hopfield networks are, by construction, dynamical systems organised around attractors. Their computation proceeds by moving a state toward the nearest energy minimum, and their behaviour is characterised by basins of attraction, transition thresholds, and path dependence. If transformer attention implements Hopfield retrieval, then the mathematical structure of attractor dynamics is not something imposed on transformers from outside as an interpretive frame — it is present in the architecture.

The Hopfield equivalence provides the theoretical foundation for treating inference as energy minimisation over a structured landscape. It does not by itself establish that generation across autoregressive steps exhibits attractor dynamics: the equivalence applies to computation within a single forward pass. But it establishes that the computational substrate from which inference emerges is one in which attractor structure is architecturally native, not merely metaphorically applicable.

### 3.2 Metastability and the Dynamics of Token Clustering

The second theoretical foundation concerns the dynamics of token clustering under attention. Geshkovski et al. (2025) model transformer layers as a mean-field interacting particle system on the unit sphere, in which tokens are particles and attention implements inter-particle interaction. Their central result is that these dynamics produce token clustering, with clusters forming and subsequently merging over time. This analysis has been extended to the causally masked attention used in generative LLMs by Karagodin et al. (2024), who prove asymptotic convergence to a single cluster and provide theoretical and numerical evidence that intermediate multi-cluster states exhibit metastable behaviour, persisting for extended periods with minimal movement before sequential merging.

Metastability is the mathematical signature of basin structure in a dynamical system: a state is metastable when it sits near a local energy minimum from which escape requires crossing a barrier. The exponential persistence times proven by Geshkovski et al. imply that transformer dynamics can be captured in multi-cluster configurations for durations that scale exponentially with the depth of the energy well, with sudden transitions between plateau states corresponding to rapid reorganisations of clustering structure. The implication for inference-time behaviour is structurally relevant: if the representational geometry contains metastable states, generation proceeding through that geometry will encounter regions of varying stability. Trajectories entering a metastable region will tend to remain in it, and displacement from such regions will require overcoming an energy barrier rather than merely adjusting to a new conditional distribution — precisely the asymmetric entry-exit dynamics that the trajectory hypothesis predicts.

Empirical corroboration within actual inference passes is reported by Fernando and Guitchounts (2025), who treat the residual stream as a dynamical system evolving across layers and find self-correcting curved trajectories in reduced-dimensional representations, with attractor-like dynamics in lower layers and structured recovery from mid-layer perturbations. This self-correcting behaviour is not straightforwardly predicted by autoregressive conditioning alone.

### 3.3 Phase Transitions in Training Dynamics

A third line of theoretical support comes from the evidence that training proceeds through phase-transition-like reorganisations of the model's internal structure. Power et al. (2022) documented grokking: small transformers trained on modular arithmetic tasks achieve near-perfect training accuracy rapidly, then undergo delayed and abrupt generalisation thousands of training steps later. Nanda et al. (2023) reverse-engineered the mechanism underlying this transition, identifying three continuous training phases, memorisation, circuit formation, and cleanup, and showing that the apparently discontinuous generalisation transition is underlain by the formation of a specific algorithmic circuit whose displacement of the memorisation mechanism has the character of a first-order phase transition (Rubin et al., 2024). Olsson et al. (2022) document a related phenomenon: the formation of induction heads, whose appearance co-occurs with a sharp increase in in-context learning ability and a visible discontinuity in the training loss curve.

A representational space organised through phase transitions is likely to retain stable regions corresponding to winning configurations and residual instabilities near the boundaries between them, though this remains a theoretically motivated expectation rather than a logical

necessity. The training-time evidence establishes an enabling condition for the trajectory hypothesis: that the geometry is structured enough to support attractor-like dynamics, rather than carrying primary evidential weight for the hypothesis itself.

### 3.4 Empirical Correlates of Regime Structure

The theoretical case establishes that dynamical behaviour in inference is architecturally native. Whether it is borne out in observable behaviour is a further question. The evidence surveyed here was collected for other purposes and cannot be taken as discriminating evidence for the trajectory hypothesis over its alternatives. Its value is motivational: the phenomena are real and robust, and they exhibit features consistent with regime structure.

#### 3.4.1 Behavioural persistence and stance drift

The most widely observed empirical pattern relevant to the trajectory hypothesis is the tendency of large language models to exhibit behavioural persistence under multi-turn interaction, resisting stance change under moderate corrective pressure. Sharma et al. (2023) demonstrate that preference-tuned models exhibit systematic patterns of agreement with user-stated beliefs. Subsequent work in multi-turn settings (Liu et al., 2025 and Hong et al., 2025) shows that models often require sustained conversational pressure before shifting stance and may exhibit repeated reversals under continued prompting. These studies introduce quantitative measures of stance dynamics, including Turn-of-Flip and Number-of-Flip, which can be read as behavioural proxies for transition cost: they operationalise how resistant a model is to moving from one locally coherent pattern of responses to another. The evidence on persona stability is more mixed. Tosato et al. (2025) show that personality-like measurements are highly sensitive to prompt perturbations and conversational history, with scaling providing limited stability gains. This is better read not as evidence against basin structure but as evidence of heterogeneous basin geometry: alignment-anchored modes (such as refusal, sycophancy, and hedging) may occupy relatively deep basins while personality-like dimensions occupy metastable or weakly structured regions. The trajectory hypothesis does not require global stability; it requires only that some regions exhibit asymmetric transition dynamics.

#### 3.4.2 Steering persistence and the generation feedback loop

Activation steering research provides a second cluster of evidence. Contrastive activation addition (Rimsky et al., 2024) and related approaches establish that adding a computed direction to the residual stream during generation can reliably shift the model's behavioural mode. For present purposes, the more significant observation is the dynamics of steered behaviour over generation: steering interventions can have persistent and unintended downstream effects, including degradation of safety constraints (Xiong et al., 2026), indicating that activation-space interventions alter internal state in ways that propagate through subsequent generation. One interpretation is that interventions produce an early distributional shift, suppressing the tokens that would initiate an unsteered trajectory, after which autoregressive generation carries the model into a trajectory consistent with the steered mode. KV-cache steering (Belitsky et al., 2025) reinforces this picture: a single intervention applied to the key-value cache can shift reasoning behaviour across subsequent generation steps, demonstrating that control signals can be stateful and their effects persistent. This suggests that the dynamical system most relevant to the trajectory hypothesis may be the full autoregressive generation loop — the coupled system in which activations and generated context interact across steps — rather than the feed-forward

computation of a single forward pass.

### 3.4.3 Endogenous resistance to perturbation

The most directly relevant empirical finding is Endogenous Steering Resistance (ESR), documented by McKenzie et al. (2026). ESR is the observation that large models can resist task-misaligned activation steering during generation and recover toward task-consistent behaviour even while the steering intervention remains active. The recovery is not a failure of the steering vector to shift the model — the model initially follows the intervention before recovering — suggesting an internal consistency-monitoring process that counteracts the perturbation. McKenzie et al. provide causal ablation evidence linking this resistance to specific internal features. The pattern, displacement under perturbation followed by reversion toward the unperturbed trajectory while perturbation continues, is precisely what a system near a stable attractor would exhibit. This interpretation is not established by the ESR findings themselves, and the alternative that recovery reflects progressive reassertion of a dominant conditional distribution cannot be ruled out. But ESR establishes the existence of a recovery phenomenon that a purely autoregressive conditioning account does not naturally explain.

### 3.4.4 Context drift equilibria

Dongre et al. (2025) formalise context drift in extended model interactions as a bounded stochastic process with restoring forces, finding that multi-turn drift stabilises at equilibria characterised by noise-limited fluctuations around a central tendency rather than runaway divergence or smooth monotonic change. Dongre et al.'s identification of restoring forces (dynamics that counteract displacement from an equilibrium) represents the most directly dynamical-systems-framed empirical finding in the recent LLM literature, and its convergence with the theoretical predictions of the trajectory hypothesis arises from independent empirical work rather than from retrofitting.

Together these four clusters establish that the phenomena the trajectory hypothesis would explain (behavioural stickiness, propagating control signals, endogenous recovery, equilibrating drift) are empirically real and exhibit features consistent with regime structure. None of them discriminates between the trajectory account and the alternatives developed in Section 5. That discrimination is the function of the asymmetry coefficient. Before introducing it, however, it is necessary to establish why transition costs (the property that would distinguish attractor dynamics from the available alternatives) are not merely unmeasured but conceptually invisible within the frameworks that currently dominate the study of LLM behaviour.

## 4. Transition Costs and the Limits of the Existing Framework

The empirical findings surveyed in the preceding section show that large language models exhibit behavioural phenomena consistent with regime structure. They do not demonstrate that regime structure exists, because the studies that documented those phenomena were not designed to measure the property that would distinguish attractor dynamics from the available alternatives. That property is the cost of transitioning between behavioural states: specifically, whether entry into a mode and exit from it are symmetric or asymmetric under matched conditions. Transition costs, however, are not merely an unmeasured empirical quantity. Within the frameworks that currently dominate the study of LLM behaviour, they are conceptually invisible. Understanding why they are invisible, and what it would take to make them visible, is the philosophical core of this paper’s contribution.

The three dominant accounts of LLM behaviour each treat inference as a process without directional structure. The sequence-model account treats generation as the successive application of a conditional probability function: given a context, the model computes a distribution over next tokens, samples from it, appends the result to the context, and repeats. On this account, the relevant object of study is the conditional distribution and the factors that shape it: prompt content, context length, decoding parameters, and the statistical structure of the training distribution. The question of whether leaving a particular region of output space is harder than entering it does not arise, because the account provides no framework within which such regions could be treated as objects with differential resistance to traversal. Each step is a fresh application of the same function to an updated context.

The circuit-system account, developed by mechanistic interpretability research, treats inference as the execution of learned algorithms implemented in specific attention heads and multilayer perceptron (MLP) layers. On this account, the relevant object of study is the circuit, the identifiable computational subgraph that performs a specific function, and the central question is what function a given circuit implements and how it interacts with others. The concept of a trajectory through representational space has no natural place in this framework: circuits are fixed computational structures whose contribution is characterised by the transformation from input to output, not by the history of a generation trajectory. Whether a model is harder to move out of a behavioural mode than into it is therefore not a question about any circuit, because circuits do not have modes in the relevant sense: they have inputs and outputs.

The quasi-cognitive account, which treats LLMs through the vocabulary of belief, intention, and reasoning, provides a richer surface description of behavioural phenomena: it can describe a model as occupying an apologetic stance or as committed to a position, but it provides no theoretical machinery for explaining why such stances persist or what would be required to change them. The vocabulary is drawn from folk-psychological description, which captures the surface of human behaviour without itself supplying a dynamical account of the mechanisms that generate it. Applied to language models, it imports the descriptive richness of intentional language while inheriting its explanatory limitations. As a result, it can describe a model as remaining in or leaving a stance, but it cannot represent the cost of such transitions; still less the possibility that those costs are asymmetric with respect to direction of travel.

None of these three accounts contains the conceptual resources to represent transition costs between behavioural states. This is not an oversight that could be remedied by collecting more data within the existing frameworks. It is a consequence of how those frameworks model inference: as a function application, as a circuit execution, or as a sequence of intentional states. None treats inference as the evolution of a state through a landscape with topographic properties, including regions of greater and lesser stability, barriers between them, and traversal costs that vary with direction of travel. For transition costs to become measurable, inference must first be reconceptualised as the kind of process that could possess them.

This is the work that the trajectory level is designed to do. By treating generation as the evolution of a state through representational space — as a trajectory rather than a sequence of context-conditioned draws — the trajectory framework makes it possible to ask questions that are otherwise unformulable: Is this region of the space stable? Does its stability depend on the direction of approach? What happens at the boundary between two stable regions? These questions are not merely empirical; they are made available by a theoretical choice about how to describe the inference process. The asymmetry coefficient introduced in the following section is

not a measurement of a pre-existing quantity that existing frameworks happen not to have measured. It is a measurement that becomes possible only once inference is treated as a dynamical process, and whose result is interpretable only within the trajectory framework.

The preceding analysis has a further consequence for the philosophical status of the trajectory level. A common objection to introducing a new explanatory level is that it is unnecessary: the phenomena it is designed to explain can be accommodated within an existing framework with modest extension. The preceding analysis suggests that this objection cannot be sustained here. The phenomena in question, asymmetric transition costs, threshold-mediated exit, and path dependence in inference, are not merely unaddressed by the existing frameworks; they are unaddressable within them, because those frameworks do not model inference as the kind of process that could generate such phenomena. The trajectory level is not a convenience or a redescription of known facts in more evocative language. It is a reconceptualisation that opens a class of questions the existing frameworks cannot ask, and makes possible a class of measurements the existing frameworks cannot take.

In this sense, the trajectory level constitutes a genuine explanatory level in the tradition of Marr's levels of analysis. A new level is justified not when it redescribes the same phenomena in different vocabulary, but when it identifies a distinct kind of question, one that requires its own theoretical objects and empirical methods, and when answering that question illuminates phenomena the adjacent levels leave unexplained. The preceding sections have argued that such a question exists: whether inference-time generation exhibits regime structure with asymmetric transition costs. The following sections introduce the theoretical objects, the asymmetry coefficient and the four-account discriminator, and the empirical methods needed to investigate it.

## 5. The Asymmetry Coefficient: A Proposed Discriminator

On the trajectory account, the central empirical question is not only what representations are computed, but how movement between behavioural modes occurs, and whether that movement is symmetric or directionally constrained. If transition costs between behavioural modes are asymmetric, and if exit is threshold-mediated rather than smoothly gradient-sensitive, autoregressive conditioning is not sufficient to explain inference.

The trajectory level is supported on two grounds: theoretical, in the mathematical structure of transformer attention and the evidence for metastability in transformer dynamics; and empirical, in behavioural phenomena consistent with regime structure that existing frameworks leave unexplained. What is now required is a way to determine whether the trajectory level is not merely plausible and empirically consistent but genuinely necessary: whether inference-time generation exhibits the specific dynamical structure that distinguishes attractor-based regime dynamics from the available alternatives. This section introduces the asymmetry coefficient as a proposed discriminator between four competing accounts of inference-time behaviour and addresses the principal objections it will face.

A candidate explanatory level earns its place by identifying phenomena lower-level descriptions cannot explain without residue. The trajectory layer proposed here makes a specific claim: that inference-time generation exhibits regime structure: stable behavioural modes whose resistance to perturbation is asymmetric with respect to direction of travel. To test this claim precisely, we define mode occupancy $O(t)$ as a continuous value derived from a trained probe over the model's hidden states at generation step $t$, treated instrumentally as a hypothesis about local behavioural state rather than as a direct readout of a psychologically real mode.

The asymmetry coefficient is then defined as $A(M, \gamma) = R_{out}(\gamma) / R_{in}(\gamma)$, where $\gamma$ denotes the magnitude of the steering intervention, operationalised as the Euclidean norm of the applied vector in the residual stream. $R_{in}$ and $R_{out}$ measure resistance, operationalised as the minimal steering vector norm required to induce a transition across a predefined occupancy threshold $\tau$ within a fixed generation window. These quantities are measured for entry into and exit from mode $M$ under equal-norm steering interventions applied in opposite directions in residual space, with prompt perturbations matched for length and approximate semantic distance under a fixed embedding metric.

Alternative operationalisations (tokens-to-transition, probability-mass shift at the threshold, or number of repeated interventions required) are expected to be monotonically related to the norm-based measure under the experimental protocol, insofar as each operationalisation measures the same underlying quantity: the resistance of the system to crossing a fixed occupancy threshold. Under matched interventions, greater resistance should require greater intervention magnitude, longer time-to-transition, and larger cumulative perturbation, so variation across operationalisations should preserve ordering even if the scales differ. The key quantity is the ratio, not the absolute values, so bounded monotonic variation across operationalisations does not affect the discriminating prediction. Under pure autoregressive conditioning, $A(M, \gamma) \approx 1$: entry and exit difficulty should track the same probability gradients symmetrically. Under attractor dynamics, $A(M, \gamma) > 1$, with a specific further prediction that distinguishes the accounts: entry should be gradient-sensitive, scaling roughly linearly with intervention magnitude, while exit should be threshold-mediated, resisting small perturbations disproportionately before yielding discontinuously to larger ones. This asymmetric shape — not mere asymmetry — is the signature a genuinely dynamical account predicts and a conditioning account does not.

A note on the probe is needed here. The non-uniqueness of interpretability decompositions is now empirically established: sparse autoencoders trained on identical data share only approximately thirty percent of their learned features across independent runs (Paulo & Belrose, 2025), supporting the view that feature decompositions are better treated as instrumental lenses than as unique descriptions of internal structure. The probe-based definition of $O(t)$ is therefore not a claim that the probe recovers a single determinate internal state. It is an operational hypothesis: the probe provides a tractable handle on a behavioural dimension of interest, and its validity is assessed by its capacity to predict behavioural transitions and to respond appropriately to causal interventions. This operationalisation follows the methodological direction of recent mechanistic interpretability work, in which decodable information is more robustly supported when it is validated by intervention rather than by correlation alone (Geiger et al., 2025).

### 5.1 The Four-Account Discriminator

The asymmetry coefficient does not test only for the presence or absence of attractor dynamics. It discriminates between four distinct accounts of the mechanisms underlying behavioural mode persistence, each of which makes different predictions about the shape of entry and exit functions. These accounts, and their predictions, are as follows.

Under *autoregressive conditioning*, mode persistence is explained by the accumulated influence of tokens already generated. If the model has been producing apologetic outputs, apologetic continuations are locally probable because the existing context makes them so. Entry into a mode and exit from it are governed by the same probability gradients acting in opposite

directions: equal-magnitude interventions should produce symmetric effects, and $A \approx 1$. Exit trajectories should be smooth, with behavioural occupancy declining gradually as the context shifts.

Under *training-distribution asymmetry*, mode persistence reflects the differential reinforcement of behavioural modes during training. Alignment training strongly reinforces entry into certain modes such as refusal, hedging, and apology, and does not equivalently reinforce the transitions out of them. $A > 1$ is expected for alignment-anchored modes, but the asymmetry should be specific to modes that were asymmetrically reinforced during training. Critically, this account predicts that $A \approx 1$ for behaviourally neutral modes: verbosity, formality register, hedging in low-stakes contexts. These modes were not asymmetrically shaped by the alignment training signal. Exit trajectories under this account should be asymmetric but smooth: harder than entry, but still gradient-sensitive (i.e., scaling continuously with intervention magnitude) rather than threshold-mediated. Here, as in the conditioning and trajectory accounts, entry remains gradient-sensitive; the discriminator is exit shape, not entry shape.

Under *autoregressive inertia*, mode persistence is explained by sequential self-reinforcement in the generation loop. Once a behavioural pattern is established in the generated context, subsequent token predictions are conditioned on that context, creating a feedback loop that maintains the pattern without any underlying basin structure. Under this account, entry is also asymmetric: once early tokens establish a behavioural pattern, autoregressive self-reinforcement can accelerate settling into that mode, even though exit still proceeds through gradual contextual displacement rather than threshold-mediated release. As the steering intervention accumulates over tokens, the context shifts, and the model follows it. Exit trajectories should therefore show gradual decay, with behavioural occupancy declining smoothly as steering accumulates, rather than the critical persistence followed by abrupt transition that attractor dynamics predict.

Under *trajectory dynamics*, mode persistence reflects the topographic structure of the representational space: some regions of the space are stable basins from which displacement requires crossing an energy barrier. Entry into a basin is gradient-driven: the trajectory follows the landscape downward toward the minimum, while exit requires surmounting the barrier, which resists small perturbations and yields discontinuously to larger ones. $A > 1$ is expected across both alignment-anchored and behaviourally neutral modes, and exit trajectories exhibit critical persistence: occupancy remains near threshold for an extended period before transitioning sharply, rather than gradually decaying. Crucially, if the asymmetry reflects basin structure in the representational geometry, it should persist when the refusal direction is suppressed via activation intervention, since it is not dependent on the control-layer gating signal remaining active. These predictions are summarised in the following table.

**Table 1**

*Predicted signatures of the asymmetry coefficient across four competing accounts*

| Account | Entry Function | Exit Function | *A(M,γ)* Prediction | Scale Prediction | Path Dependence Prediction |
|---|---|---|---|---|---|
| Autoregressive conditioning | Smooth, gradient-sensitive | Smooth, gradient-sensitive | $A \approx 1$ | No scale effect | None |
| Training-distribution asymmetry | Smooth | Smooth, elevated resistance | $A > 1$ for alignment modes only | Depends on training, not geometry | None |
| Autoregressive inertia | Asymmetric | Gradual decay | $A > 1$ but exit smooth | Weak scale effect | No genuine path dependence (apparent persistence depends on current context, not trajectory history under matched-context conditions) |
| Trajectory dynamics | Smooth, gradient-sensitive | Threshold-mediated, discontinuous | $A > 1$ across mode types | Nonlinear scale effect | Yes, resistance varies with trajectory history under matched-context conditions |

*Note.* $A(M, \gamma)$ = asymmetry coefficient; $R_{in}$, $R_{out}$ = entry and exit resistance; $\gamma$ = steering intervention magnitude; $\tau$ = occupancy threshold. Path dependence is tested under matched-context conditions (same state, different entry trajectory).

The discriminating power of the asymmetry coefficient lies not merely in the magnitude of $A$ but in the joint pattern of predictions across mode type, exit trajectory shape, and scale dependence. A result of $A > 1$ in alignment-anchored modes alone, with smooth exit trajectories, supports the training-distribution account. A result of $A > 1$ across mode types with gradual exit supports the inertia account. Only a result of $A > 1$ across mode types with threshold-mediated exit trajectories and nonlinear scale dependence supports the trajectory dynamics account. The accounts are therefore separable by the shape and distribution of the asymmetry, not merely by its presence. Among the discriminating axes, path dependence carries particular weight: if a model in an apparently identical behavioural state exhibits different exit resistance depending on the trajectory by which that state was reached, this pattern cannot be explained by distributional asymmetry or autoregressive inertia, both of which are insensitive to trajectory history under matched-context conditions. Path dependence under matched conditions is therefore the strongest single discriminator between basin-structured dynamics and the available alternatives.

### 5.2 Objections to the Discriminator

There are three potential confounds to the asymmetry coefficient as a discriminator that require clarification. The first concerns what the coefficient is designed to detect. Asymmetric transition costs could arise from the curvature of the token probability landscape alone, without any basin structure. The response is that the discriminating prediction concerns not mere asymmetry but a specific shape: threshold-mediated exit combined with gradient-sensitive entry. Nonlinear probability surfaces predict asymmetric gradient magnitudes varying smoothly with position;

attractor dynamics predict resistance disproportionate to intervention magnitude up to a threshold, then discontinuous yielding. The transition shape is the discriminator, not the transition cost.

The second concerns the source of any observed asymmetry. $A > 1$ in alignment-anchored modes could reflect training-distribution asymmetry rather than basin structure. This is addressed by testing behaviourally neutral modes first: if $A > 1$ appears in modes not asymmetrically shaped by alignment training, training-distribution asymmetry cannot explain the result. If $A > 1$ is confined to alignment-anchored modes, the training-distribution account is supported and should be reported as such.

The third concerns the exit trajectory. Sequential autoregressive inertia could produce $A > 1$ without any basin structure, as generated tokens reinforce the context that sustains the mode. The response is that inertia predicts gradual exit, with mode occupancy decaying smoothly as the intervention accumulates, while attractor dynamics predict critical persistence: occupancy remains near its stable value before transitioning sharply. Whether the exit trajectory is smooth or threshold-mediated is therefore independently discriminating, and the coefficient must be read in conjunction with exit trajectory shape rather than in isolation.

**6. Empirical Structure of the Test**

The asymmetry coefficient introduced in the preceding section is a theoretical proposal, not a reported result. This section establishes that the proposal is operationally tractable with current methods, and maps the full result space onto the four competing accounts. The goal is not to specify a protocol in implementation detail but to show that the test has a determinate structure: that its result, whatever it turns out to be, would constitute discriminating evidence between the accounts, including the trajectory hypothesis itself.

The test requires three things. First, the selection of behavioural modes across two categories: neutral modes not differentially reinforced by the alignment training signal, including verbosity, formal register, and low-stakes hedging, and alignment-anchored modes such as refusal, apology, and sycophantic agreement. Beginning with neutral modes is evidentially critical: if $A(M, \gamma) > 1$ appears in neutral modes under matched interventions, training-distribution asymmetry cannot explain the result. If $A \approx 1$ in neutral modes but $A > 1$ in alignment-anchored modes, the simpler training-distribution account is supported and the trajectory level is not required. Second, the construction of matched entry and exit interventions, using steering vectors of equal norm in the residual stream applied in opposite directions, so that apparent asymmetry cannot be attributed to differential intervention strength. Third, measurement not only of the magnitude of $A$ but of the shape of the exit trajectory: whether mode occupancy under exit steering decays smoothly across tokens, or exhibits critical persistence, remaining near its stable value under small perturbations before transitioning sharply once the intervention exceeds a threshold. This shape distinction is the primary discriminator between autoregressive inertia, which predicts smooth decay, and trajectory dynamics, which predicts threshold-mediated exit. The transition shape, not the transition cost alone, is what the trajectory account specifically predicts and the conditioning account cannot produce.

A further dimension of discrimination is path dependence. Under matched-context conditions, where two trajectories arrive at an apparently identical behavioural state by different routes, basin dynamics predict that exit resistance will vary with the trajectory by which the state was reached, since the depth of the basin relative to the current position depends on the approach direction. Distributional asymmetry and autoregressive inertia are both insensitive to trajectory

history under these conditions. Path dependence under matched contexts is therefore the strongest single discriminator between basin-structured dynamics and the available alternatives.

The result space maps cleanly onto the four accounts. If $A \approx 1$ across all mode types, this supports the autoregressive conditioning account and the trajectory level is not required. If $A > 1$ for alignment-anchored modes only, with smooth exit trajectories, this supports the training-distribution account. If $A > 1$ across mode types but with gradual exit, this supports the autoregressive inertia account. Only if $A > 1$ across mode types including neutral modes, with threshold-mediated exit and resistance that varies with trajectory history under matched-context conditions, does the result provide discriminating empirical support for the trajectory dynamics account. The accounts are separable by the joint pattern of asymmetry magnitude, exit trajectory shape, and path dependence, and not by any single measurement in isolation.

This result space is intended to be exhaustive: any outcome should fall into one of the four interpretive categories, or into an informative combination. Threshold behaviour in neutral modes but not in alignment-anchored modes would, for example, suggest that the representational geometry contains genuine attractor structure in some regions while alignment gating operates as a control-layer intervention over shallower structure elsewhere, a finding of theoretical interest in its own right. The protocol is designed to produce discriminating evidence across the full outcome space, including outcomes that would count against the trajectory account. If exit trajectories are uniformly smooth, if asymmetry is confined to alignment-anchored modes, or if resistance proves insensitive to trajectory history under matched conditions, the trajectory hypothesis does not receive support from this protocol, and one of the simpler accounts is doing the explanatory work. These failure conditions are not qualifications; they are constitutive of the discriminator's evidential value.

## 7. Alignment Gating as Trajectory Control

### 7.1 The Mechanistic Picture

The refusal-direction findings of Arditi et al. (2024) established that refusal behaviour across a range of open-weight chat models is mediated by a single low-dimensional direction in the residual stream. Suppressing this direction via activation intervention reliably reduces refusal on harmful prompts; adding it reliably induces refusal on benign ones. Zhao et al. (2025) showed that harmfulness judgment and refusal execution are encoded as separable internal representations, with harmfulness registered at the last token of the user instruction and refusal execution at the last token of the full input sequence. Wollschläger et al. (2025) found that refusal is governed not by a single direction but by a low-dimensional subspace: multiple geometrically independent directions that collectively constitute a compact structure governing alignment-related behaviour. Alignment behaviour in large language models is not uniformly distributed throughout the model's weights but is partially concentrated in identifiable structures in activation space.

These findings are typically framed within the circuit-system account: alignment training installs specific computational structures that detect policy-relevant features and route the model toward compliant outputs. This framing is accurate as a description of mechanism but obscures an important structural feature of what the findings reveal. A low-dimensional direction in activation space that redirects the model away from one class of completions and toward another is more usefully modelled as a control signal, a compact intervention on a high-dimensional dynamical process that shifts the trajectory of generation without specifying its endpoint, than as a circuit that produces refusal as a computed function. A circuit is defined by its function; a

control signal is defined by its effect on dynamics. When Arditi et al. show that ablating the refusal direction eliminates refusal behaviour across a wide range of prompts, this is consistent with modelling the refusal direction as a compact control signal that redirects trajectories across a large region of prompt space. The trajectory framework makes this distinction theoretically explicit in a way the circuit framework does not.

### 7.2 Alignment Gating as Trajectory Initialisation

Recent work suggests that alignment-related influences on model behaviour are disproportionately concentrated in the earliest stages of generation, with the distribution over initial tokens exerting a strong downstream constraint on subsequent outputs (Qi et al., 2024). This observation does not by itself establish a dynamical account of inference, but it is consistent with an interpretation on which alignment operates, in part, by shaping the initial direction of a trajectory through representational space.

On this view, alignment does not function solely as a persistent constraint applied uniformly across all stages of generation. Instead, it can be understood as biasing the initial conditions under which autoregressive dynamics unfold. Once a trajectory is initiated within a particular region of the representational landscape, subsequent generation may be governed primarily by the local structure of that landscape and by the autoregressive feedback loop, rather than by continuous application of a global constraint. The effect of alignment, in such cases, is not to determine the full path of generation but to influence which regions of the space are entered in the first place.

The trajectory framework makes this temporal asymmetry theoretically intelligible. If inference proceeds as movement through a structured landscape with regions of varying stability, then small differences in initial conditions can result in qualitatively different trajectories. Early-stage interventions may therefore have effects that are amplified over the course of generation, not because they exert ongoing control, but because they determine which regions of the landscape the trajectory enters.

This interpretation provides a natural account of the empirical concentration of alignment effects at early tokens while remaining agnostic about the full temporal distribution of control. It does not imply that alignment ceases to operate after initialisation, nor that later-stage interventions are ineffective. The relevance of this observation to alignment is interpretive rather than definitive: it offers a way of understanding existing empirical findings within a dynamical framework, not independent evidence for that framework.

### 7.3 Implications for Alignment Theory

The trajectory-control framing has a broader implication. If alignment behaviour is not only a property of fixed output policy but also of trajectory control, then the standard conception of alignment as weight-encoded behaviour is incomplete. Alignment, on the trajectory-control account, is a property of the interaction between the control layer and the dynamical landscape it operates on. Its robustness depends not only on the strength of the control signal but on the structure of the landscape: a strong control signal operating on a landscape with deep basins corresponding to misaligned modes may be less robust than a weaker signal on a flat landscape, because the dynamical pull of the misaligned basin will resist redirection. Alignment training modifies the landscape, potentially deepening or reshaping basins through reinforcement, but does not fully determine its structure. On the account developed here, the landscape's organisation emerges primarily from pretraining compression dynamics; alignment intervenes

within that structure rather than constituting it.

This reframing has a concrete implication for alignment research. Current alignment methods such as RLHF, direct preference optimisation, and constitutional AI are primarily understood as methods for shaping the model's conditional probability distribution. The trajectory-control framing suggests that they are also methods for modifying the dynamical landscape that inference traverses and for introducing control signals that redirect trajectories within that landscape. Whether a given alignment intervention produces robust alignment may therefore depend not only on how strongly it reinforces compliant outputs but on whether it modifies the landscape's topography, specifically the depth and extent of basins corresponding to aligned and misaligned modes, or merely introduces a control signal that redirects trajectories without changing the landscape they traverse.

**8. Trajectory Dynamics as an Explanatory Level: The Philosophical Argument**

The proposal advanced in this paper is not that attractor dynamics have been demonstrated in large language models, but that an explanatory gap in the compression account requires a level of description capable of linking representational structure to behavioural regularities during inference. Behavioural regimes, transition costs, and control signals become systematically describable only once inference is treated as a dynamical process. The asymmetry coefficient functions not as a proof of a particular theory but as a methodological proposal for investigating whether trajectory dynamics constitutes a genuine explanatory level.

What it means for a level to be genuine rather than merely convenient is the central philosophical question. Drawing on the tradition of levels-of-explanation arguments descending from Marr (1982), a level is genuine when it introduces explanatory concepts and questions not naturally expressible within adjacent levels and not recoverable from them without loss of explanatory power. This criterion extends Marr's framework in the direction of mechanistic levels approaches (Machamer et al., 2000) that emphasise the non-derivability of higher-level descriptions from lower-level ones as a condition of genuine explanatory independence. Section 4 argued that transition costs are conceptually invisible within the three dominant frameworks because none models inference as a process with directional dynamics. Such systematic invisibility is evidence for a distinct explanatory level rather than merely a gap in current knowledge.

The argument draws on two philosophical resources. The first is Bedau's account of weak emergence and macro-level explanatory autonomy (Bedau, 1997): a higher-level description is explanatorily autonomous when the macro-level patterns it identifies are not practically derivable from the lower-level description, even if consistent with it. In principle, the behaviour of a large language model at every inference step is fully determined by its weights and input sequence. In practice, no account operating purely at the level of weights, circuits, or conditional probability distributions provides a tractable description of why some behavioural modes resist perturbation more than others, why transitions can be threshold-mediated, or why exit from a mode can require qualitatively more intervention than entry. The trajectory level generates predictions, specifically the asymmetry coefficient and its expected dependence on mode type, exit shape, and scale, not readily generated by lower-level descriptions.

The second resource is a structural analogy with Haken's slaving principle from synergetics (Haken, 1983): in systems operating near critical points, dynamics collapse to a low-dimensional description in terms of slow order parameters that enslave fast variables. If the trajectory framework is correct, the relevant slow variables during inference are the mode

occupancy values, the positions of the trajectory in the behavioural landscape, and the fast variables are the individual token-level probability distributions. The trajectory level earns genuine status when the slow variables have their own dynamics: stable regions, transition thresholds, and path dependence that cannot be read off from the fast-variable dynamics without the intermediate description.

A further consideration concerns the transition from designed to trained systems. In designed systems, levels of explanation correspond to levels of design: the engineer who specified the algorithm, the engineer who chose the implementation, and the engineer who selected the hardware made decisions at Marr's computational, algorithmic, and implementational levels respectively. In trained systems, no such correspondence exists. The trajectory level is identified analytically, by finding the stratum at which distinct questions arise and distinct empirical methods are required, and justified by showing that it is the level at which regime persistence, asymmetric transition costs, and control-signal effects become tractable, and that no adjacent level makes them as tractable without it.

The trajectory level is therefore a candidate for genuine explanatory autonomy: it identifies questions unformulable without it, generates predictions that existing accounts do not, and provides the framework within which the empirical programme proposed in Section 6 becomes coherent. What the philosophical argument establishes is that the question is worth asking, and that asking it requires the level of description this paper proposes.

## 9. Conclusion: Toward a Dynamics of Inference

This paper has argued for the introduction of a trajectory level of explanation in the study of large language models: a level that treats inference-time generation as the evolution of a state through a high-dimensional representational landscape rather than as a sequence of independent probability draws.

The central contribution is a proposed discriminator between four competing accounts of inference-time behaviour: the asymmetry coefficient $A(M, \gamma) = R_{out}(\gamma)/R_{in}(\gamma)$, which measures the ratio of exit to entry resistance under matched steering interventions. The accounts make distinct predictions about the magnitude, shape, and distribution of this coefficient across mode types and scales. $A > 1$ across neutral modes with threshold-mediated exit and nonlinear scale dependence would constitute discriminating support for the trajectory account; $A \approx 1$, or asymmetry confined to alignment-anchored modes with smooth exit, would support one of the alternative accounts. The protocol is designed to produce discriminating evidence across the full result space, including outcomes that count against the trajectory hypothesis.

What follows from the framework is determinate in both directions. If the asymmetry coefficient yields the joint pattern the trajectory account predicts, a programme of regime cartography becomes tractable: mapping stable regions, transition boundaries, and basin depths, and characterising how alignment training modifies that topography. If it does not, the simpler accounts are supported and the framework's limits are clarified. Either outcome advances the field. The alignment implication is equally conditional: if the trajectory-control account is correct, the tools of control theory become applicable to alignment research in a precise sense. If it is not, that application does not follow. The framework does not foreclose either result.

## References

Arditi, A., Obeso, O., Syed, A., Paleka, D., Panickssery, N., Gurnee, W., & Nanda, N. (2024). Refusal in language models is mediated by a single direction. *Advances in Neural Information Processing Systems, 37*. https://proceedings.neurips.cc/paper_files/paper/2024/file/f545448535dfde4f9786555403ab7c49-Paper-Conference.pdf

Bedau, M. A. (1997). Weak emergence. In J. Tomberlin (Ed.), *Philosophical perspectives: Mind, causation, and world* (Vol. 11, pp. 375–399). Blackwell Publishers.

Belitsky, M., Kopiczko, D. J., Dorkenwald, M., Mirza, M. J., Glass, J. R., Snoek, C. G. M., & Asano, Y. M. (2025). KV cache steering for controlling frozen LLMs [Preprint]. arXiv. https://arxiv.org/abs/2507.08799

Dongre, V., Rossi, R. A., Lai, V. D., Yoon, D. S., Hakkani-Tür, D., & Bui, T. (2025). Drift no more? Context equilibria in multi-turn LLM interactions [Preprint]. arXiv. https://arxiv.org/abs/2510.07777

Fernando, J., & Guitchounts, G. (2025). Transformer dynamics: A neuroscientific approach to interpretability of large language models [Preprint]. arXiv. https://arxiv.org/abs/2502.12131

Geiger, A., Ibeling, D., Zur, A., Chaudhary, M., Chauhan, S., Huang, J., Arora, A., Wu, Z., Goodman, N., Potts, C., & Icard, T. (2025). Causal abstraction: A theoretical foundation for mechanistic interpretability. *Journal of Machine Learning Research, 26*(83), 1–64. https://www.jmlr.org/papers/v26/23-0058.html

Geshkovski, B., Letrouit, C., Polyanskiy, Y., & Rigollet, P. (2025). A mathematical perspective on transformers. *Bulletin of the American Mathematical Society, 62*(1), 53–99. https://doi.org/10.1090/bull/1837

Greenblatt, R., Denison, C., Wright, B., Roger, F., MacDiarmid, M., Marks, S., Treutlein, J., Belonax, T., Chen, J., Duvenaud, D., Khan, A., Michael, J., Mindermann, S., Perez, E., Petrini, L., Uesato, J., Kaplan, J., Shlegeris, B., Bowman, S. R., & Hubinger, E. (2024). Alignment faking in large language models [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2412.14093

Haken, H. (1983). *Synergetics: An introduction to nonequilibrium phase transitions and self-organization in physics, chemistry, and biology* (3rd ed.). Springer.

Hong, J., Byun, G., Kim, S., & Shu, K. (2025). Measuring sycophancy of language models in multi-turn dialogues. In *Findings of the Association for Computational Linguistics: EMNLP 2025* (pp. 2239–2259). Association for Computational Linguistics. https://aclanthology.org/2025.findings-emnlp.121.pdf

Karagodin, N., Polyanskiy, Y., & Rigollet, P. (2024). Clustering in causal attention masking [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2411.04990

Kelly, M. (2026a). *Beyond pattern matching: Representation and the case for a middle-level theory of large language models* [Preprint]. PhilSci-Archive. https://philsci-archive.pitt.edu/id/eprint/29470

Kelly, M. (2026b). *Statistical structure and the failure of pointing: A system-class law for compression-based generative systems* [Preprint]. PhilSci-Archive. https://philsci-archive.pitt.edu/id/eprint/29350

Liu, J., Jain, A., Takuri, S., Vege, S., Akalin, A., Zhu, K., O'Brien, S., & Sharma, V. (2025).

Truth decay: Quantifying multi-turn sycophancy in language models [Preprint]. arXiv. https://arxiv.org/abs/2503.11656

Machamer, P., Darden, L., & Craver, C. F. (2000). Thinking about mechanisms. *Philosophy of Science, 67*(1), 1–25. https://doi.org/10.1086/392759

Marr, D. (1982). *Vision: A computational investigation into the human representation and processing of visual information*. MIT Press.

McKenzie, A., Pepper, K., Servaes, S., Leitgab, M., Cubuktepe, M., Vaiana, M., de Lucena, D., Rosenblatt, J., & Graziano, M. S. A. (2026). Endogenous resistance to activation steering in language models [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2602.06941

Nanda, N., Chan, L., Lieberum, T., Smith, J., & Steinhardt, J. (2023). Progress measures for grokking via mechanistic interpretability [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2301.05217

Olsson, C., Elhage, N., Nanda, N., Joseph, N., DasSarma, N., Henighan, T., Mann, B., Askell, A., Bai, Y., Chen, A., Conerly, T., Drain, D., Ganguli, D., Hatfield-Dodds, Z., Hernandez, D., Johnston, S., Jones, A., Kernion, J., Lovitt, L., … Olah, C. (2022). In-context learning and induction heads. Transformer Circuits Thread. https://transformer-circuits.pub/2022/in-context-learning-and-induction-heads/index.html

Paulo, G., & Belrose, N. (2025). Sparse autoencoders trained on the same data learn different features [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2501.16615

Power, A., Burda, Y., Edwards, H., Babuschkin, I., & Misra, V. (2022). Grokking: Generalization beyond overfitting on small algorithmic datasets [Preprint]. arXiv. https://arxiv.org/abs/2201.02177

Qi, X., Panda, A., Lyu, K., Ma, X., Roy, S., Beirami, A., Mittal, P., & Henderson, P. (2024). Safety alignment should be made more than just a few tokens deep [Preprint]. arXiv. https://arxiv.org/abs/2406.05946

Ramsauer, H., Schäfl, B., Lehner, J., Seidl, P., Widrich, M., Adler, T., Gruber, L., Holzleitner, M., Pavlović, M., Sandve, G. K., Greiff, V., Kreil, D., Kopp, M., Klambauer, G., Brandstetter, J., & Hochreiter, S. (2021). Hopfield networks is all you need. In *Proceedings of the International Conference on Learning Representations*. https://openreview.net/forum?id=tL89RnzIiCd

Rimsky, N., Gabrieli, N., Schulz, J., Tong, M., Hubinger, E., & Turner, A. (2024). Steering Llama 2 via contrastive activation addition. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics* (pp. 15504–15522). Association for Computational Linguistics. https://aclanthology.org/2024.acl-long.828

Rubin, N., Seroussi, I., & Ringel, Z. (2024). Grokking as a first order phase transition in two layer networks. In *Proceedings of the Twelfth International Conference on Learning Representations*. https://arxiv.org/abs/2310.03789

Sharma, M., Tong, M., Korbak, T., Duvenaud, D., Askell, A., Bowman, S. R., Cheng, N., Durmus, E., Hatfield-Dodds, Z., Johnston, S. T., Kravec, S., Maxwell, T., McCandlish, S., Ndousse, K., Rausch, O., Schiefer, N., Yan, D., Zhang, C., & Perez, E. (2023). Towards understanding sycophancy in language models [Preprint]. arXiv. https://arxiv.org/abs/2310.13548

Tosato, T., Tump, A. N., Flesch, T., Sumers, T., Griffiths, T. L., & Behrens, T. E. J. (2025).

Persistent instability in LLMs' personality measurements: Effects of scale, reasoning, and conversation history [Preprint]. arXiv. https://arxiv.org/abs/2508.04826

Turpin, M., Michael, J., Perez, E., & Bowman, S. R. (2023). Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. *Advances in Neural Information Processing Systems, 36*. https://arxiv.org/abs/2305.04388

Wollschläger, T., Elstner, J., Geisler, S., Cohen-Addad, V., Günnemann, S., & Gasteiger, J. (2025). The geometry of refusal in large language models: Concept cones and representational independence. *Proceedings of Machine Learning Research, 267*, 66945-66970. https://proceedings.mlr.press/v267/wollschlager25a.html

Xiong, C., He, Z., Chen, P.-Y., Ko, C.-Y., & Ho, T.-Y. (2026). Steering externalities: Benign activation steering unintentionally increases jailbreak risk for large language models [Preprint]. arXiv. https://arxiv.org/abs/2602.04896

Zhao, J., Huang, J., Wu, Z., Bau, D., & Shi, W. (2025). LLMs encode harmfulness and refusal separately [Preprint]. arXiv. https://arxiv.org/abs/2507.11878