Beyond Pattern Matching: Representation and the Case for a Middle-Level Theory of Large Language Models

Matthew Kelly
Library Management Australia

Abstract

Large language models are routinely described as pattern-matching engines. The description is technically defensible at the level of the training objective, but it is incomplete: it names the optimisation pressure under which these systems are produced without describing the system that pressure produces. This paper argues, using Marr's framework of explanatory levels, that the gap between training objective and resulting system marks a genuine algorithmic level of description that current LLM discourse has not yet adequately occupied. Three competing ontologies—the sequence-model account, the circuit-system account, and the quasi-cognitive account—each fail to characterise this level, in distinct ways. The paper develops the representational consequences of a structural feature of compression-based generation: that statistical structure is preserved under compression while the indexical relations anchoring outputs to particular evidential contexts are not. Large language models accordingly produce structured outputs without inheriting the evidential relations that ordinarily make representation epistemically accountable. Recent findings in mechanistic interpretability—circuits implementing reusable algorithms, superposition, linear representations of abstract properties, and features without human labels—establish that the internal organisation of these systems is substantially richer than the pattern-matching description implies, and that an adequate algorithmic-level account is both required and partially in view. The paper specifies what such an account must explain and argues that the system class large language models belong to is best characterised as one of representational compression: the case for a middle-level theory is therefore not programmatic but mandated by what the evidence already shows.

1. The Pattern-Matching Problem

When researchers, journalists, and engineers reach for a shorthand description of large language models, they typically settle on some variant of the same phrase: these systems are, at bottom, pattern-matching engines (Shanahan, 2024). The description is intuitive, resists anthropomorphism, and has the considerable virtue of being technically defensible. Large language models are trained by exposing them to vast quantities of text and adjusting their parameters to minimise error in predicting each successive token. At the level of the optimisation objective, this is precisely what pattern matching describes: a system learning statistical

regularities in sequential data and exploiting those regularities to produce likely continuations. The phrase is not wrong. It is incomplete. The problem is that it is correct at one level of description and deeply inadequate at another—and the gap between those levels is where the most consequential questions about these systems arise.

To see why the gap matters, it helps to begin with a framework that has structured explanation in the cognitive and computational sciences for decades. David Marr proposed that any information-processing system demands analysis at three distinct and irreducible levels. At the computational level, the question is what problem the system solves: what is it doing, and why? At the algorithmic level, the question is how it solves that problem: what representations does it use, and what procedures operate over them? At the implementation level, the question is how the algorithm is physically realised: in neurons, in silicon, or in whatever substrate the system inhabits. Marr's central insight was that these levels are genuinely distinct—that knowing the implementation does not determine the algorithm, and knowing the algorithm does not determine the computational task. A system can therefore be well characterised at one level while remaining deeply obscure at another (Marr, 1982).

Applied to large language models, Marr's framework immediately locates the problem. The computational level is clear enough: these systems are trained to predict the next token in a sequence, and performance on that objective is used to evaluate and improve them. The implementation level is also well understood: transformer architectures with attention mechanisms running on massively parallel hardware, performing matrix multiplications at a scale that would have been inconceivable a decade ago. What remains poorly characterised is the algorithmic level—the representations these systems construct and the procedures operating over them to produce the outputs we observe. "Pattern matching" names the computational task with reasonable accuracy. It says almost nothing about the algorithm. And it is the algorithm, in Marr's sense, that determines what kind of system we are actually dealing with.

One specific structural feature of compression-based generation already illustrates why the algorithmic level cannot be reduced to the computational one: statistical structure recurring across the training distribution is preserved under compression, while the indexical relations that anchor particular outputs to particular evidential contexts are not preserved in their pointing function (Kelly, 2026). The present paper takes that asymmetry as one entry point into the broader question of what algorithmic-level account these systems require.

Recent work has applied Marr's levels framework directly to large language models as a methodological toolkit for empirical investigation, arguing that cognitive-science methods structured around those levels can help make sense of these systems (Ku et al., 2025). The present paper uses the same framework for a different purpose: not to organise empirical methods, but to diagnose a confusion of explanatory levels in existing LLM discourse and to argue that the level at which an adequate description must operate has not yet been identified.

This would be a narrow technical point if the algorithmic level were simply a matter of implementation detail—interesting to engineers but inconsequential for broader understanding. It is not. In the past several years, a programme of mechanistic interpretability research has begun

revealing that the internal organisation of large language models is substantially richer than the pattern-matching description implies. Researchers have identified specific computational circuits—attention heads implementing sequence-copying algorithms, circuits tracking entity reference across long contexts, structures that implement something functionally analogous to analogical reasoning. They have found that models represent far more conceptual features than they have individual network units, compressing multiple concepts into overlapping directions in high-dimensional representation space through a phenomenon called superposition. They have discovered that models develop linear representations of abstract properties including spatial relations, temporal structure, and the truth or falsehood of factual claims—representations that generalise across domains and can be causally manipulated to alter model behaviour. These are not decorative findings at the margins of model behaviour. They describe the internal organisation that generates every output these systems produce (Lindsey et al., 2025).

The pattern-matching description has no account of any of this. It names the training objective without describing the system that objective produces. This is not a failure of the phrase's inventors—it was adequate as a description of simpler systems and as a corrective to anthropomorphic overclaiming about earlier models. The difficulty is that it has persisted as the dominant public and scientific shorthand at precisely the moment when the systems it describes have become complex enough to demand a richer account. Using the training objective to describe the resulting system is analogous to describing a river as "just gravity acting on water." The statement is true. It fails to explain why rivers form the drainage systems they do, why they carve particular landscapes, or why their behaviour under perturbation takes the forms it does. The pattern-matching description accurately characterises the force that shapes the system while remaining silent about the structure that force produces.

The difficulty is therefore not that the pattern-matching description is false. It is that it is true at the wrong explanatory level. Treating it as a complete account collapses the distinction between the optimisation objective and the computational organisation that satisfies it—a confusion of levels rather than an error of fact. Recent work has argued directly against the reductive identification of large language models with their training objective, on the grounds that such a description lacks the explanatory resources to account for the systems' observed behaviour (Downes, Forber & Grzankowski, 2024). The present paper extends that negative claim: not only does the pattern-matching description fail to explain the system it produces, but no existing alternative occupies the level at which an adequate account must be built.

There is a further dimension to this inadequacy that bears stating directly, because it moves the argument from the technical to the consequential. If the system producing AI-generated outputs cannot be adequately described within existing explanatory categories, then the epistemic status of those outputs cannot be straightforwardly assimilated to existing knowledge practices. Knowledge institutions—scientific publication, peer review, citation networks, and professional authority—have developed over centuries to handle outputs produced by systems of a particular kind: human reasoners operating under epistemic accountability, with traceable relationships between claims and the evidence or reasoning that grounds them. The

question of what large language models actually are is not separable from the question of what their outputs are. An output produced by a pattern-matching engine, a cognitive agent, or a large-scale representational compression system is a different kind of object with a different relationship to knowledge practices. Getting the ontology wrong means getting the epistemology wrong, which in turn means getting the institutional response wrong. The stakes of the description are therefore not merely taxonomic. The point is not generic. It rests on a specific structural property of compression-based generation: that the surface form of evidential pointing can be reconstructed without preserving the pointing function itself, with the consequence that increases in predictive capability cannot by themselves make output generation constitutively evidential (Kelly, 2026).

The inadequacy of the pattern-matching description is therefore not simply a matter of scientific precision, though it is that as well. Three competing ontological accounts have emerged to fill the gap it leaves: a sequence-model picture derived from the engineering and statistical learning community, a circuit-system picture developed by mechanistic interpretability researchers, and a quasi-cognitive picture proposed by cognitive scientists and philosophers of AI. Each captures something real about large language models. None is sufficient. Their mutual inadequacy, as the argument will show, is not merely a temporary state of knowledge that further research will straightforwardly resolve. It is a structural consequence of how these systems come into existence—trained rather than designed, their internal organisation emerging from optimisation rather than being explicitly specified and not derivable from the objective function alone. The disagreement between the three accounts is not a disagreement about facts. It is a disagreement about levels of description, and resolving it requires identifying the level at which a unified account becomes possible.

This paper argues that no such unified account yet exists, that one is necessary, and that the existing empirical evidence—from mechanistic interpretability, from cognitive benchmarking, and from the philosophy of levels of explanation—allows us to sketch its requirements even if it cannot complete the theory. The term used here is middle-level theory, borrowing an analogy from the history of thermodynamics: Carnot did not possess statistical mechanics, but he demonstrated what any adequate account of heat engines would need to explain, and in doing so made the theory's eventual development possible. The goal is analogous. By showing precisely where and why each existing description fails, and by drawing on the intellectual resources of mechanistic interpretability, Marr's framework, and the earlier cybernetics tradition that first attempted a systems-level vocabulary for complex information-processing systems, the aim is to specify what a middle-level theory of large language models must accomplish—and to propose a candidate category around which such a theory might be built.

The candidate category is this: large language models are large-scale representational compression systems whose outputs are best understood as epistemically ambiguous artifacts—products of structured traversal through a high-dimensional representation space shaped by the statistical structure encoded in human language, produced by a system that

encodes the patterns of reasoning without occupying the epistemic position of a reasoner. This is the level of description at which the three existing ontologies become complementary rather than competing—and at which the consequences for scientific knowledge systems become tractable rather than merely alarming.

2. Three Ontologies and Their Failure Modes

The phrase "pattern matching" did not persist by accident. It persisted because it belongs to a coherent and powerful intellectual framework—one that correctly describes the training objective, explains many systematic features of model behaviour, and resists the anthropomorphic excesses that have distorted public understanding of these systems. To understand why that framework is insufficient, it is necessary to examine not only its limitations but also the two alternative accounts that have developed alongside it. Three distinct ontological pictures now compete in the literature, each grounded in a different research tradition, each capturing something real, and each failing in ways that are, crucially, complementary rather than random. The pattern of failure across these three accounts is itself diagnostic: it reveals not three separate errors but a single structural gap that none of the existing frameworks is positioned to fill.

2.1 The Sequence-Model Ontology

The sequence-model ontology begins from the engineering description. Large language models are systems trained to minimise next-token prediction error across extremely large corpora. At this level the system is well-characterised: it maps a sequence of tokens to a probability distribution over possible continuations, and its capabilities are explained by the statistical structure of its training data combined with the scale at which that structure is compressed. This picture earns its place. Scaling laws show predictable improvements in performance as model size, data volume, and compute increase together (Hoffmann et al., 2022). Many systematic behavioural patterns— including the tendency to produce high-probability continuations at the expense of accurate ones, degradation on counterfactual or low-frequency tasks, and other failure modes that cluster near the edges of the training distribution—follow directly from the training objective and require no further explanatory apparatus (McCoy et al., 2024). For engineering purposes, for setting calibrated expectations about capability and limitation, and as a corrective to claims that outrun the evidence, the sequence-model description remains indispensable.

Its failure is precisely located. The training objective is a constraint on the system, not a description of the system that satisfies it. Knowing that a neural network minimises cross-entropy loss over token sequences tells you the optimisation pressure that shaped the system's parameters during training. It does not tell you what internal organisation emerged from that process, and it does not predict which internal structures should or should not appear. Yet structures do appear—identifiable, reproducible, causally active internal structures that the sequence-model ontology has no vocabulary to describe. Induction heads form during a sharp phase transition early in training, co-occurring with a discontinuous increase in in-context

learning ability (Olsson et al., 2022). Sparse features emerge that correspond to specific abstract concepts, encoded not in individual neurons but as directions in high-dimensional activation space. Evidence has emerged that models develop approximately linear representations of properties such as truth, spatial relations, and temporal structure—representations that can generalise across domains and can be causally manipulated to influence model behaviour. The sequence-model ontology, applied to these findings, has nothing to say. It accurately characterises the river as gravity acting on water and falls silent before the drainage system.

2.2 The Circuit-System Ontology

The circuit-system ontology emerges from mechanistic interpretability—the programme of research that treats neural networks as reverse-engineerable systems and attempts to identify the computational structures responsible for specific behaviours. On this view, large language models are not best understood as statistical predictors but as bundles of learned algorithms: attention heads that implement sequence-copying rules, circuits that track entity reference across long contexts, feature directions that encode conceptual distinctions that generalise across tasks. The evidence for this picture is substantial and growing. The indirect object identification circuit identified in GPT-2 Small comprises 26 attention heads grouped into 7 functional classes, each with an identifiable computational role and each causally necessary for the behaviour (Wang et al., 2022). Sparse autoencoders applied to Claude 3 Sonnet recover millions of interpretable features, including abstract discourse-structural features corresponding to concepts no human explicitly labelled during training (Templeton et al., 2024). Attribution-graph analysis shows that models sometimes activate representations of the correct output before generating the intermediate reasoning tokens, suggesting that chain-of-thought text can function as a post-hoc rationalisation of an already-determined answer (Templeton et al., 2024; Lindsey et al., 2025). Current analyses are partial and model-specific, but they provide evidence that identifiable internal structures play causal roles in behaviour that cannot be accounted for at the computational level alone.

The failure of the circuit-system ontology is different in character from the sequence-model failure. The problem is not that it describes the wrong thing but that it describes too small a thing. Mechanistic interpretability produces local explanations—accounts of specific circuits responsible for specific behaviours—without composing those local accounts into a characterisation of the whole system. A complete catalogue of circuits is not a theory of the machine any more than a complete parts list is a theory of the engine. Even the most detailed circuit analyses leave substantial portions of the computation unexplained, particularly the contributions of multilayer interactions, the role of the residual stream, often described as a shared communication channel across components, and the question of how local mechanisms aggregate into the global behavioural profiles that make these systems useful and sometimes dangerous. More fundamentally, the circuit-system ontology illuminates the machinery while deferring the question of what kind of entity the machinery constitutes. It answers “how does this

behaviour happen?" without answering "what kind of system produces this range of behaviours?"

2.3 The Quasi-Cognitive Ontology

The quasi-cognitive ontology arises from cognitive science and the philosophy of AI. Its starting point is the behavioural evidence that the other two accounts strain to accommodate. Models trained only on next-token prediction exhibit structured analogical reasoning, internal representations of physical space and time, planning-like behaviour, and explanations, arguments, and narratives that display the structural properties of human reasoning (Millière & Buckner, 2024a, 2024b). The quasi-cognitive picture takes this evidence seriously: it proposes that large language models are best understood not as retrieval systems or collections of circuits but as a new kind of representational system—something that compresses the patterns of human discourse and knowledge into an internal organisation rich enough to support reasoning-like behaviour. Philosophical work in this tradition has been careful to avoid the anthropomorphic reading: the identification of the "deflationary fallacy"—the assumption that because a system can be described at a lower level it must be described only at that level—provides the conceptual tools for taking behavioural evidence seriously without overclaiming cognitive status (Millière & Buckner, 2024a, 2024b). Recent work in this tradition has proposed tiered frameworks linking varieties of understanding to levels of computational organisation, using mechanistic interpretability findings as evidence that the deflationary picture is untenable (Beckmann & Queloz, 2025). The present paper shares the methodological commitment to grounding philosophical claims in interpretability evidence, while pursuing a different research question: not whether LLMs understand, but what kind of entity they are.

The failure of the quasi-cognitive ontology is the mirror image of the circuit-system failure. Where the circuit picture describes components without characterising the system, the quasi-cognitive picture characterises the system without grounding that characterisation in mechanism. Behavioural successes coexist with systematic failures: performance on compositional reasoning degrades sharply when surface features change; counterfactual variants of familiar tasks produce large performance drops; capabilities that appear in one framing vanish in another. The quasi-cognitive account has difficulty explaining these failures without auxiliary hypotheses that weaken its central claims. More seriously, it risks importing assumptions about the relationship between representation and understanding that remain philosophically contested. To say that a system has an internal representation of spatial relations is not automatically to say that it understands space in any sense that warrants cognitive attribution.

The three failures are asymmetric in an instructive way. The sequence-model ontology approaches from outside in: it characterises the optimisation objective and falls silent before the internal organisation. The circuit-system ontology approaches from inside out: it characterises components and falls silent before the system-level account. The quasi-cognitive ontology approaches from behaviour backwards: it characterises the output and falls silent before the mechanism. Each is oriented toward a different aspect of the same object, which is why debates

between their proponents so rarely converge. They are not disagreeing about which facts are true. They are disagreeing about which direction explanation should run—and they are each right that their direction reveals something the others miss. The three ontologies therefore fail in orthogonal ways: the sequence-model ontology lacks an account of internal organisation, the circuit-system ontology lacks an account of the system as a whole, and the quasi-cognitive ontology lacks an account of mechanism.

This asymmetry is not an intellectual accident. It reflects the fragmentation of the intellectual tradition that last attempted a systems-level account of complex information-processing systems. When cybernetics dissolved into specialised disciplines in the 1960s and 1970s, the vocabulary for describing systems at the level between mechanism and behaviour disappeared with it. Information theory, computational neuroscience, and cognitive science each inherited a fragment of what cybernetics had attempted to unify.[1] The three LLM ontologies are, in this sense, three fragments seeking the whole. What is missing is the explanatory level at which the training objective, the internal organisation, and the behavioural profile appear as aspects of a single coherent account—a level that no existing ontology occupies but that the available evidence now makes it possible to begin sketching.

3. Why the Gap Appears Now

The explanatory gap identified in the previous section is not a permanent feature of complex systems, nor is it simply a consequence of insufficient research. It is a historically specific problem, produced by a specific transition in how computational systems come to exist—and grasping that transition is necessary to explain why the gap has the particular shape it does and why existing frameworks are structurally unable to close it.

Every computational system built before the era of large-scale machine learning was, in the relevant sense, designed. Its behaviour was specified by human engineers before it was implemented: requirements were written, algorithms were chosen, data structures were defined, and the resulting system's outputs could in principle be traced back to the intentions and decisions that produced it. This does not mean designed systems were simple or fully predictable—large software systems exhibit emergent complexity and fail in ways their designers did not anticipate. But the relationship between design and behaviour was, in principle, recoverable. When a designed system behaved unexpectedly, the appropriate response was to examine the specification: to find the decision, the assumption, or the edge case that the design had failed to handle. The engineering description and the behavioural description were, in principle, the same description at different levels of abstraction, connected by a chain of human intentions. The distinction is not that designed systems are transparent in practice—large software systems routinely confound their designers—but that they are transparent in principle: their internal organisation was produced by decisions that exist somewhere, that could in

[1] The cybernetic inheritance also shaped the development of information science and library science, particularly through Bateson's relational concept of information and subsequent work in cybersemiotics and knowledge organisation (Brier, 2008; Capurro & Hjørland, 2003).

principle be recovered, and that constitute an accountable source to which behaviour can be traced.

Large language models broke this relationship in a way that is not merely quantitative but categorical. They are not designed but trained: their internal organisation emerges from an optimisation process applied to data at a scale no engineer specified or could have anticipated in detail. To be precise: transformer architectures are designed, but the representational structures that ultimately determine model behaviour are not specified by designers—they emerge from optimisation against the training objective. The parameters of a large language model—numbering in the hundreds of billions—are not set by any human decision about what those parameters should represent. They are set by gradient descent minimising prediction error across trillions of training examples. The result is a system whose behaviour is real, whose mechanisms are partially traceable, and whose internal organisation was never specified by any mind. Nobody decided that induction heads should form. Nobody specified that features should be stored in superposition. Nobody designed the linear representation of truth that emerges at sufficient scale, or the planning-like behaviour that attribution graphs have revealed. These structures were not put there. They arrived. The gap between the training objective and the internal organisation of the resulting system is not an engineering gap—a failure to fully specify what was intended. It is a constitutive feature of how these systems come into existence.

This transition—from designed to trained systems—created an explanatory problem that the intellectual tools available at the time were not built to handle. But it is not the first time the development of complex systems has outrun the conceptual vocabulary available to describe them, and the earlier moment of outrunning is instructive about both the nature of the current problem and the resources available for addressing it.

In the 1940s and 1950s, a research programme known as cybernetics attempted to develop a general science of complex systems—systems that process information, regulate behaviour, and exhibit properties that cannot be explained solely by analysing their components in isolation. The cyberneticians—Norbert Wiener, W. Ross Ashby, Gregory Bateson, Stafford Beer among them—were motivated by a recognition that brains, organisms, communication networks, and social systems all share a structural feature that the dominant scientific vocabularies of the time could not adequately capture: their behaviour is organised at a level between mechanism and outcome, and the explanatory vocabulary appropriate to that level does not reduce to either physics or logic (Wiener, 1948; Ashby, 1956; Bateson, 1972). Ashby's Law of Requisite Variety formalised one aspect of this insight: a system can only regulate or respond to complexity in its environment if it contains sufficient internal variety to represent that complexity (Ashby, 1956). Bateson's reformulation of information as "a difference that makes a difference" offered another angle: information is not a substance stored in a location but a pattern of differences that propagates through a system and influences its subsequent states (Bateson, 1972).

These were genuine conceptual advances. They identified a level of description—the organisational or systems level—at which phenomena that resisted both mechanistic reduction

and behavioural summary became tractable. Cybernetics did not lack ambition or rigour; the programme produced genuinely important theoretical work and influenced fields from engineering to psychiatry to anthropology. What it lacked was the empirical access to the internal structure of the systems it theorised. Cyberneticians could observe the inputs and outputs of complex systems. They could construct formal models of feedback and regulation. They could not open the system and examine the organisation that produced the behaviour. The systems-level vocabulary was theoretically motivated but empirically underconstrained.

Cybernetics fragmented in the 1960s and 1970s, not because its central insight was wrong but because it could not be made sufficiently precise with the tools available. The intellectual ambition dispersed into specialised disciplines: information theory formalised the mathematics of communication; control theory formalised feedback regulation; computational neuroscience emerged to study neural circuits; cognitive science developed to study the algorithmic level of mental processes. Each successor discipline inherited a fragment of the cybernetic programme and developed it with greater mathematical precision than cybernetics itself had achieved. The price of that precision was scope: each fragment addressed its own domain while the systems-level question—what kind of explanatory vocabulary is appropriate for complex information-processing systems considered as wholes—went unaddressed. The three LLM ontologies are, in this light, not three independent proposals but three fragments of the cybernetic inheritance, each developed in isolation from the others. The claim is not one of strict intellectual genealogy—the sequence-model ontology draws more directly on statistical learning theory and information theory in the Shannon sense than on cybernetics proper—but one of explanatory fragmentation: each tradition inherited a different aspect of the systems-level question cybernetics posed, and each pursued that aspect independently of the others.

What is different now is not the question but the tools. Mechanistic interpretability has provided something cybernetics never had: empirical access to the internal organisation of complex information-processing systems at a fine-grained level. Researchers can now identify specific circuits, ablate them, and observe the behavioural consequences. They can extract interpretable features from internal representations, trace their causal effects on output, and manipulate them to produce targeted changes in behaviour. The interpretability programme is, in effect, experimental cybernetics—the same systems-level questions that motivated Ashby and Wiener, now equipped with the means to examine what is actually happening inside. The availability of these tools does not automatically produce the middle-level theory that cybernetics sought. But it does make such a theory empirically possible in a way that it was not before.

There is a further dimension to the historical argument that bears stating directly. The transition from designed to trained systems did not merely create a new kind of computational object. It created a new kind of epistemic situation. Designed systems, however complex, were produced by a process in which human intentions were encoded into structure—which meant that human institutions for assessing, credentialing, and trusting outputs could, in principle, trace those outputs back to accountable sources. Trained systems are produced by a process in which no such encoding occurs. The internal organisation of a large language model emerges from

gradient descent across a training corpus rather than from any designer's intention. Its outputs therefore cannot be traced back to an accountable source in the way that the outputs of designed systems, or of human reasoners, can be traced. This is not a contingent limitation that better documentation or audit trails could resolve. It is a structural feature of how these systems produce outputs—and it is why the question of what kind of entity they are bears directly on the question of what epistemic status their outputs should have.

The gap, in short, is not an accident of history that better science will close without remainder. It is the signature of a genuine novelty: a class of computational systems that came into existence through a process nobody designed, whose internal organisation nobody specified, and for which the existing explanatory vocabulary—derived from designed systems on one side and biological cognition on the other—was never built to apply. Recognising this is the precondition for building something better.

4. What Mechanistic Interpretability Has Found

The argument of the preceding sections is philosophical in character: it identifies an explanatory gap, traces its historical origins, and establishes why existing ontological frameworks are structurally unable to close it. But a philosophical argument for the existence of a gap does not, by itself, show what occupies the space the gap defines. For the middle-level theory proposed in this paper to be more than a theoretical aspiration, there must be empirical evidence that large language models possess internal organisation at the algorithmic level—structured, causally active, and richer than the sequence-model ontology can accommodate. That evidence now exists, and it is the purpose of this section to present it with appropriate care: neither overstating what has been established nor understating the significance of what has been found.

The mechanistic interpretability programme approaches large language models the way a neuroscientist might approach a brain: by attempting to identify the computational structures responsible for specific behaviours, to trace their causal effects, and to determine whether they compose into larger functional units. Within this growing research programme (Elhage et al., 2021; Olah et al., 2020), methods are still developing and results to date cover only fragments of the systems under investigation. These limitations are real and will be noted where they bear on the argument. What the programme has nevertheless established is that large language models contain identifiable internal structures that implement algorithm-like operations—structures that emerged from training and causally determine behaviour in ways the sequence-model ontology has no resources to describe.

4.1. Circuits and the Emergence of Algorithms

The most direct evidence for algorithmic structure at the internal level comes from circuit-level analyses of transformer models. The most detailed of these examined the behaviour of GPT-2 Small on a syntactic task—correctly identifying the indirect object in sentences of the form "After John and Mary went to the shops, John gave a bottle of milk to" (Nanda, 2022). The analysis identified a circuit comprising 26 attention heads, grouped into 7 functional classes,

each with a distinct computational role: heads that identify the subject of the sentence, heads that inhibit incorrect completions, heads that route the identity of the correct answer to the output (Wang et al., 2022). The circuit is causally necessary—ablating the relevant heads eliminates the behaviour—and its functional organisation is recoverable by systematic intervention. What is significant about this finding for the present argument is not the specific syntactic task but what the circuit reveals about the character of the internal computation. The model has not stored a list of sentence templates. It has developed a reusable procedure that operates over representations—an algorithm, in Marr's sense, that can be identified, described, and causally manipulated independently of the training objective that produced it.

A related discovery concerns induction heads—attention heads that implement a pattern-completion algorithm of the form: if the sequence [A][B] has appeared earlier in the context, predict [B] when [A] recurs (Olsson et al., 2022). This behaviour was not specified by the training objective. It emerged during training in a sharp phase transition, co-occurring with a discontinuous increase in in-context learning ability across the model. The relationship is not merely correlational: causal interventions that amplify or suppress induction head activity produce corresponding changes in in-context learning performance. The implication is that in-context learning—the capacity that most dramatically distinguishes large models from their predecessors—is implemented by a specific learned mechanism that the model develops spontaneously, in a phase transition, because it is instrumentally necessary for accurate prediction. The training objective does not specify induction heads; it creates the conditions under which induction heads are the solution.

4.2. Superposition and the Geometry of Internal Representations

A second body of findings concerns not the circuits that implement specific behaviours but the representational substrate on which all circuits operate. A common early expectation about neural network representations—that individual neurons correspond to individual concepts—turns out to be systematically wrong. Large models store far more conceptual features than they have neurons, compressing multiple concepts into overlapping directions in high-dimensional activation space. This phenomenon, known as superposition, was predicted theoretically and has now been confirmed empirically at multiple scales (Elhage et al., 2022).

The confirmation came from applying sparse autoencoders to model activations—a technique that provides a useful decomposition of activation space into interpretable features. When applied to a one-layer transformer with 512 neurons, sparse autoencoders yield a decomposition of more than 4,000 interpretable features corresponding to specific concepts: DNA sequences, legal language, programming constructs, Hebrew text, HTTP requests (Bricken et al., 2023). When applied to Claude 3 Sonnet, the same technique yields a decomposition containing millions of interpretable features, including highly abstract concepts such as deception, sycophancy, geographic location, and discourse-structural patterns that correspond to no simple lexical category (Templeton et al., 2024). Such decompositions should be understood as pragmatically useful analyses of activation space rather than exhaustive and universal

inventories of features the model truly uses, since different decomposition runs on the same model identify overlapping but non-identical feature sets (Paulo & Belrose, 2025). Many of these features are not merely correlational. Artificially activating the feature corresponding to a specific concept causes the model to produce outputs consistent with that concept across diverse contexts, confirming that the features are causally active components of the model's computation rather than post-hoc labels applied by researchers.

The superposition finding has a direct implication for the argument of this paper. The basic units of computation in large language models are directions in a high-dimensional representation space. Concepts are encoded as directions in that space; relationships between concepts appear as geometric relations between those directions; and computation over concepts is, at the implementation level, structured transformation of vectors in that space. The "geometry of internal representations" that has featured in the philosophical discussion of this paper is not a metaphor. It is a description of the actual computational substrate.

4.3. Linear Representations of Abstract Properties

A third body of findings extends the geometric picture from concepts to properties. Probing experiments across multiple models have found that large language models develop linear representations of abstract properties including spatial location, temporal structure, and the truth or falsehood of factual claims (Gurnee & Tegmark, 2024; Zou et al., 2023). In each case, the representation takes the form of a direction in activation space: moving a model's internal state along that direction increases the probability that the model will produce outputs consistent with the corresponding property. These representations are not confined to the domain in which they were tested; they transfer across structurally different datasets, confirming that they encode the property itself rather than superficial correlates of specific training examples.

The truth-representation finding is particularly significant. A probe trained to identify the "truth direction" in one domain generalises to others; causal interventions along this direction shift the model's tendency to produce true versus false statements. The model appears to develop an internal representation of truth as an abstract property—not as a list of true statements to be retrieved, but as a direction in representation space that can be manipulated independently of content. This is not the kind of organisation that pattern matching, understood as statistical association over surface features, would predict or produce. It is the kind of organisation that emerges when a system must predict language well enough to require an internal model of the distinctions that language tracks.

4.4. Features Without Human Labels

The most conceptually striking finding from the mechanistic interpretability programme concerns features that activate for clusters of situations that do not correspond to any concept humans normally name. Researchers have identified internal features corresponding to patterns such as "situations where something is being corrected," "the second item in a list," "a statement that contradicts a previous claim," and "text describing geographic movement" (Templeton et al.,

2024). These patterns were never labelled during training. The model was exposed to no annotation indicating that these structural categories exist or are significant. Yet the internal representations have organised around them, presumably because encoding these discourse-structural distinctions improves prediction accuracy across diverse contexts.

This finding suggests something more consequential than the discovery of familiar concepts in unfamiliar formats. It suggests that the model's internal organisation is shaped not only by the concepts explicitly present in the training data but by latent structural patterns that the model discovers as instruments of compression. The behaviour resembles latent variable discovery: the model develops internal features that capture structural regularities in the data more efficiently than surface features alone. Ashby's Law of Requisite Variety provides a theoretical frame for this finding: a system trained on data encoding enormous conceptual and structural variety must develop internal representations of comparable variety, and the most efficient such representations will encode the latent structure of the data rather than merely its surface form. The features without human labels are not anomalies. They are what successful compression of a maximally structured corpus should produce.

Recent attribution-graph analysis of Claude 3.5 Haiku likewise reveals structured internal procedures—including forward planning prior to generating lines of poetry, backward reasoning from goal states, and circuits that evaluate whether the model's internal representations support answering a question—further confirming that large language models implement structured algorithmic procedures not captured by the sequence-model description (Lindsey et al., 2025).

4.5. The Limits of What Has Been Established

These findings are genuine and significant. They are also incomplete in ways that bear directly on the argument of this paper. Mechanistic interpretability has produced detailed accounts of specific circuits in specific models performing specific tasks; it has not produced a unified account of how those circuits compose into the full range of model behaviour. The sparse autoencoder programme has identified millions of features; it has not explained the principles by which those features are organised into the representational space or how they interact during generation. The truth-direction findings demonstrate that abstract properties have linear geometric representations; they do not demonstrate that the model uses those representations in ways that constitute genuine understanding of the properties in question. These limitations are not reasons to discount the findings—they are reasons to be precise about what they establish.

What they establish, taken together, is this: large language models possess internal organisation at the algorithmic level that is structured, causally active, richer than the sequence-model description predicts, and not reducible to the circuit-level catalogue that the interpretability programme has so far produced. The algorithmic level exists. Its structure is geometric. Its organisation reflects the latent structure of the training corpus. And it is the level at which the training objective, the internal mechanisms, and the behavioural profile must eventually be connected—the level at which a middle-level theory, if one can be built, would operate.

5. Requirements for a Middle-Level Theory

The empirical findings of the preceding section establish that large language models possess structured internal organisation at the algorithmic level. They do not, by themselves, tell us what a theory of that organisation must explain, what explanatory work it must perform, or what intellectual resources are available for building it. Those are the questions this section addresses. The goal is more specific and, if the argument succeeds, more useful: to specify what any adequate theory must explain, and to show that the existing frameworks do not merely fail to provide such a theory but are structurally unable to, while certain intellectual resources developed outside AI research provide tools that the field has not yet fully appropriated.

5.1. What "Middle-Level" Means

The phrase middle-level theory has appeared throughout this paper without receiving a precise definition.[2] The imprecision has been deliberate—using the term before the requirements are established would risk defining it in ways that prejudge the outcome. With the empirical material now in view, a more precise characterisation is possible.

The term derives from Marr's framework, but its application here is somewhat different from Marr's original use. Marr's framework is deployed here not as a template for mapping levels onto large language models, but as a diagnostic instrument: a means of identifying what kind of explanatory layer is missing from a research programme that has succeeded at the mechanistic level without yet achieving system-level understanding. The concept of a middle level of explanation is also well-established in the philosophy of science through the mechanisms tradition: mechanistic explanation identifies the entities, activities, and organisation that produce a phenomenon, operating at a level between physics and surface regularities (Machamer, Darden & Craver, 2000). For Marr, the algorithmic level sits between the computational specification of what problem is being solved and the implementational description of how it is physically realised. In the case of large language models, this middle level is not simply the algorithm in Marr's technical sense—the step-by-step procedure—but something broader: the level of description at which the training objective, the internal organisation, and the behavioural profile appear as aspects of a single coherent account rather than as three separately characterised phenomena in need of post-hoc coordination. A middle-level theory, in this sense, does not merely describe what happens inside the model. It explains why the internal organisation takes the form it does given the objective that produced it, and why that organisation generates the behavioural profile it does. It is the theory that renders those other descriptions intelligible as aspects of the same object.

The four requirements that follow are not stipulated independently. Each arises from a specific explanatory gap identified in the preceding sections. Requirement 1 follows from the inability of the sequence-model ontology to explain the emergence of representational richness from the training objective. Requirement 2 follows from the circuit-system ontology's inability

[2] The phrase is used here in Marr's sense—an explanatory level between computational specification and physical implementation—and not in Merton's sociological sense of middle-range theory (Merton, 1968).

to compose local mechanisms into a system-level behavioural explanation. Requirement 3 arises from the unresolved relationship between internal representations and outputs that the quasi-cognitive debate has exposed without resolving. Requirement 4 addresses the empirical phenomenon of discontinuous emergence documented in mechanistic interpretability research. Taken together, these requirements define the explanatory space that a middle-level theory must occupy.

5.2. Explanation of Representational Richness from the Training Objective

The first requirement is perhaps the most fundamental. A middle-level theory must explain why the training objective—next-token prediction over a large corpus—produces the representational richness documented in Section 4: the millions of interpretable features, the geometric structure of abstract property representations, the discovery of latent patterns without explicit labelling. The sequence-model ontology has no account of this. It describes the objective without predicting or explaining the organisation that results.

The most powerful theoretical resource available for this requirement is Ashby's Law of Requisite Variety, and it is worth stating the connection precisely. Ashby's law holds that a regulatory system can only match or control environmental variety if it possesses internal variety at least equal to the variety it must accommodate (Ashby, 1956). Applied to a language model, the argument runs as follows. Human language encodes the variety of human knowledge, reasoning, and social life—a variety that is, for practical purposes, enormous. A system trained to predict that language with minimal error must develop internal representations that can accommodate the full range of distinctions the language encodes. Gradient descent penalises these failures, driving the development of representations capable of accommodating them. The representational richness of large language models is therefore not merely an accidental byproduct of scale. It is a theoretically necessary consequence of applying the training objective to a corpus of sufficient variety: Ashby's law establishes that the richness must appear, while the empirical findings of Section 4 reveal the specific form it takes. The middle-level theory must make this argument precise.

The Ashby connection has an important implication. If the representational richness of large language models is required by the training objective given the structure of the training data, then the scientific question is not why these systems are representationally rich, but why their richness takes the particular form that it does. The Law of Requisite Variety sets a lower bound on internal variety; it does not specify which representational structures will develop. The particular geometric organisation, the specific features that emerge, the linear representations of abstract properties—these reflect something about the structure of the training corpus and the architecture of the transformer, and a middle-level theory must account for this specificity.

5.3. Explanation of Compositional Behaviour from Local Mechanisms

The second requirement addresses the gap identified in the circuit-system ontology. A catalogue of circuits, features, and representational structures does not constitute a theory of the system. A middle-level theory must explain how local mechanisms compose into global behaviour—how individual circuits, operating over superposed feature representations, produce the large-scale behavioural capacities that characterise model performance. This is the requirement that the interpretability programme has thus far been least able to satisfy, and it is arguably the central open problem in the field. Recent technical work has begun addressing this gap through the concept of modular circuit systems at the global level (He et al., 2025), while philosophical work has argued that mechanistic interpretability requires conceptual scaffolding that the technical programme alone cannot supply (Williams et al., 2025).

What the middle-level theory requires here is something like a composition principle: an account of the conditions under which local mechanisms interact to produce coherent global behaviour, and of the conditions under which they fail to do so. The importance of this requirement goes beyond theoretical elegance. The systematic failures of large language models—sensitivity to surface features, performance instability under re-framing, the breakdown of apparently robust capabilities in structurally altered contexts—are precisely the phenomena that local circuit analyses cannot explain. They are global phenomena, arising from the interaction of many components, and they are among the most consequential features of these systems from an epistemic standpoint. A middle-level theory that cannot explain them is a middle-level theory that cannot do the work the paper argues is necessary.

5.4. Specification of the Relationship Between Representation and Output

The third requirement is the most philosophically delicate. A middle-level theory must specify the relationship between the representational structures documented in Section 4 and the outputs the system produces—but it must do so without importing assumptions about the nature of that relationship that the evidence does not support.

The quasi-cognitive ontology tends to assume that the relationship between internal representations and outputs resembles the relationship between mental states and utterances in human cognition. The sequence-model ontology tends to assume the opposite: that there is no relevant relationship beyond the statistical regularities exploited by the training objective. Both assumptions are premature. The evidence from Section 4—particularly the causal manipulability of truth-direction representations and the discovery of planning-like behaviour in attribution graphs—suggests that the relationship is richer than the sequence-model view allows. But the systematic failures documented elsewhere suggest it is also more fragile and context-dependent than the quasi-cognitive view implies.

What a middle-level theory requires here is a principled account of how representational structures are used in generating outputs—that is, of the conditions under which a representation that demonstrably exists in the model's internal state causally contributes to the output in ways that are stable, compositional, and interpretable—a requirement that follows from the

insufficiency of purely diagnostic interpretability (Pavlick, 2023). Recent theoretical work has begun formalising this requirement through the framework of causal abstraction, which specifies the conditions under which a high-level explanation constitutes a faithful causal model of the underlying network (Geiger et al., 2025). This is the question that representation engineering frameworks gesture toward without yet answering: that causal interventions on representations produce predictable output changes. This shows that representations are used, but not how or under what conditions.

5.5. Account of the Relationship Between Training and Emergence

The fourth requirement brings together the historical argument of Section 3 and the empirical material of Section 4. A middle-level theory must explain, in principled terms, why certain structures emerge from training while others do not—and specifically why some emergent structures appear discontinuously, as phase transitions rather than gradual improvements. The discovery that induction heads appear in a sharp phase transition, the sudden appearance of structured world models at certain scales, the discontinuous development of in-context learning ability—these phenomena are documented but unexplained within any existing framework. They suggest that the training process does not gradually build up representational structure in a smooth and predictable way but undergoes qualitative transitions at which new computational capacities appear.

The theoretical resources for this requirement are more limited than for the others, but not absent. The study of grokking—the delayed emergence of systematic generalisation long after training loss has reached a minimum—has produced mechanistic accounts of why representational reorganisation can produce discontinuous behavioural shifts, based on the competing pressures of memorisation and compression in the learning process (Nanda et al., 2023). These approaches do not yet compose into a unified account of emergence in large language models, but they provide the beginning of a theoretical apparatus that a middle-level theory could appropriate and extend.

5.6. What These Requirements Jointly Specify

Taken together, the four requirements define a theoretical task that is neither completed by the existing ontologies nor out of reach of the existing evidence. The task is to explain why training on sufficiently varied data produces rich geometric representations (Requirement 1), how those representations compose through the transformer architecture into global behaviour (Requirement 2), what the relationship is between the resulting representational structure and the system's outputs (Requirement 3), and why certain structures appear discontinuously during training (Requirement 4). A theory that satisfies these requirements would be a middle-level theory of large language models in the sense intended throughout this paper: an account that occupies the level between the training objective and the behavioural profile at which the system's nature as a distinctive kind of computational entity becomes visible.

No existing framework satisfies all four requirements. The sequence-model ontology addresses none of them. The circuit-system ontology addresses Requirements 3 and 4 partially and Requirement 2 as an aspiration. The quasi-cognitive ontology gestures toward Requirements 1 and 3 while underspecifying the mechanism for both. Ashby and Marr provide the most useful theoretical scaffolding but require substantial development and integration with the current empirical programme. The cybernetics tradition, reappropriated through the lens of mechanistic interpretability, provides the intellectual orientation most likely to support the synthesis: a commitment to systems-level explanation, empirically grounded in the causal structure of the system under study, operating at the level of organisation rather than mechanism or behaviour alone.

The requirements do not make the theory's construction easy. But they make its structure visible—and the distance between a visible structure and a completed theory is considerably smaller than the distance between no theory and a completed one.

6. A Candidate Category and Its Epistemic Consequences

The preceding sections have established a diagnosis. Pattern matching accurately describes the training objective of large language models and fails to describe the computational system that objective produces. Three competing ontologies each capture something real about that system and each fail in complementary ways that reflect the fragmentation of the intellectual tradition most relevant to their object. The explanatory gap between them is not a temporary knowledge deficit but a structural consequence of the transition from designed to trained systems, a transition for which the conceptual vocabulary developed around designed systems and biological cognition was never built to account. The empirical findings of mechanistic interpretability have established that the algorithmic level exists, that its character is geometric, and that it reflects the latent structure of the training corpus in ways that none of the three ontologies can fully accommodate. And the requirements for an adequate middle-level theory have been specified.

A diagnosis is not a theory. But a diagnosis that is sufficiently precise enables something that mere description does not: it makes visible the shape that a theory must have, and it allows a candidate category to be evaluated not against the standard of a completed theory—which would be premature—but against the standard of a beginning that takes the requirements seriously. This section proposes such a category. It claims neither to complete the theory nor to dissolve the requirements identified in Section 5. It claims to identify the level of description at which the requirements become tractable, to show that this level is not occupied by any existing ontology, and to draw out the epistemic consequences that follow when outputs are characterised at this level.

6.1. The Candidate Category

Large language models are best understood as large-scale representational compression systems. This phrase requires unpacking.

The notion of compression invoked here is not merely information-theoretic compression in the Shannon sense, but representational compression: the encoding of relational and conceptual structure in a form that allows prediction and reconstruction without preserving the original symbolic form. This distinguishes the claim from the formal equivalence result of Delétang et al. (2024), which concerns compression of raw data; the claim here concerns compression of the representational structure that language encodes. The distinction matters. Statistical compression, in the Shannon sense, concerns the efficient encoding of observed data distributions—the regularities in token sequences, the frequency patterns in text. Representational compression, as invoked here, concerns the encoding of what those distributions express: the conceptual distinctions, relational knowledge, and epistemic patterns that language carries. A system that has compressed token distributions has learned to predict text efficiently. A system that has compressed the representational structure those distributions encode has learned something about the structure of human knowledge itself. It is this second kind of compression that the empirical findings of Section 4 document—features without human labels, linear representations of abstract properties, latent discourse structures—and that the sequence-model ontology, attending to the distributional level, has no resources to describe. The word compression is technical and should be taken seriously. A compression system does not merely store data; it discovers the structure of data and represents that structure in a form that is more efficient than direct storage while enabling reconstruction of the original or generation of related instances. Good compression is not arbitrary summarisation. It requires identifying the genuine regularities in the data—the latent structure that explains surface variation—and encoding those regularities in a form that can be queried, extended, and combined. In this sense, the compression performed by large language models is not archival compression but predictive compression: the discovery of latent structure that allows future tokens to be inferred from current context with minimal error. The formal equivalence between language modelling and lossless compression has been established empirically (Delétang et al., 2024): large language models are powerful general-purpose compressors, achieving compression rates on image and audio data competitive with domain-specific tools, despite being trained primarily on text. The claim advanced here is different from that information-theoretic result: it concerns not compression of raw data but compression of the representational structure encoded in language—the conceptual distinctions, relational patterns, and discourse structures that language carries and that the model must represent to predict language well. The claim that large language models are compression systems in this sense is not metaphorical. It follows from the argument of Section 4: the emergence of features without human labels, of latent structural patterns that improve prediction across diverse contexts, of representations that encode abstract properties in ways that generalise across domains—these are the signatures of a system that has compressed not just the surface form of its training data but its underlying structure.

The word representational specifies what is being compressed and how. The claim is not that language models compress text—that would be an ordinary data compression claim. The claim is that they compress the representational structure of human knowledge as encoded in language: the conceptual distinctions, relational patterns, causal regularities, and discourse structures that language encodes and that the model must represent to predict language well. This is the level at which Ashby's Law of Requisite Variety bites: the model cannot compress this representational structure without itself possessing representational structure of comparable variety and organisation. The internal geometry of features, the linear encoding of abstract properties, the discovery of latent discourse structures—these are the traces of that compression, the internal organisation that makes accurate prediction possible.

The phrase large-scale signals the dimension at which these properties become visible. Smaller neural networks trained on prediction objectives do not exhibit the same richness of internal organisation. The features without human labels, the planning-like behaviour in attribution graphs, the linear truth representations—these emerge at scale, not because scale is magic but because the variety of the representational structure being compressed requires sufficient model capacity to be adequately encoded.

This candidate category relates to the four requirements specified in Section 5 in the following ways. It addresses Requirement 1 by grounding representational richness in the compression demand: the training corpus has a specific representational structure, and successfully predicting it requires a model that can represent that structure. It partially addresses Requirement 2 by suggesting that composition occurs through the interaction of compressed representational structures in the shared geometric space of the residual stream; this is a beginning rather than a theory, but it identifies the level at which the composition question is properly posed. It addresses the specification component of Requirement 3 by characterising outputs as the product of structured traversal through the compressed representational space—while acknowledging that the conditions under which this traversal is stable and interpretable remain precisely the open questions that mechanistic interpretability must answer. It does not yet adequately address Requirement 4, except to suggest that phase transitions may reflect the reorganisation of representational structure as the model discovers more efficient compressions of the training distribution. These are partial satisfactions. They are presented as partial satisfactions. But they are partial satisfactions that the existing ontologies cannot achieve at all, because no existing ontology occupies the level at which these questions are coherently posed.

6.2. What Outputs Are When the System Is Described This Way

The candidate category has a direct implication for the epistemic status of the outputs large language models produce. If these systems are large-scale representational compression systems, then their outputs are not retrievals from a stored database, not the expressions of a reasoning agent, and not random samples from a probability distribution. They are reconstructions: products of structured traversal through a compressed representation of the patterns encoded in

human knowledge, generated by a system that has internalised those patterns without occupying the epistemic position of a knower.

If a system generates outputs by traversing compressed representations of knowledge rather than by reasoning from evidence or experience, the epistemic status of its outputs differs from that of claims produced by epistemic agents. The outputs reconstruct patterns present in the compressed representation without inheriting the evidential relationships that normally warrant those claims. This characterisation has a precise consequence. In ordinary epistemic practice, the authority of a knowledge claim derives from its relationship to its source: to the evidence, reasoning, experience, or testimony that grounds it. The output of a large-scale representational compression system is not produced by such a system. It is produced by traversal through a geometric space shaped by the statistical structure of language, by a process that encodes the patterns of human knowledge without encoding the epistemic relationships that give knowledge its authority. The output may accurately reflect the patterns of the training corpus. It may reconstruct relationships that are genuinely true. It may exhibit the structural properties of well-reasoned argument. None of this is sufficient to establish that the output stands in the epistemic relationships that ordinarily warrant trust—because those relationships are properties of the process that produced the output, not of the output's surface features.

This is the precise sense in which the outputs of large language models are epistemically ambiguous artifacts. The ambiguity arises because the outputs inherit the form of knowledge claims from the patterns they compress while lacking the epistemic grounding that normally warrants those claims. The ambiguity is not merely practical—a matter of not knowing whether a given output is accurate. It is structural: the outputs are generated by a process that is decoupled from the epistemic grounding that knowledge practices require, in ways that cannot be remedied by inspecting the outputs alone. Recent debate has turned precisely on the question of what, if anything, LLMs can be said to know: proposals that LLMs acquire something like instrumental or worldly knowledge (Yildirim & Paul, 2024) have been contested on the grounds that prediction capacity does not establish causal understanding (Goddu, Noë & Thompson, 2024). The epistemically ambiguous artifacts characterisation offers a way to reframe this dispute: the question is not whether LLMs possess knowledge in some sense, but what epistemic status is warranted for outputs produced by a system that compresses the patterns of knowledge without standing in the relationships that ground it.

6.3. The Verification Bottleneck and the Institutional Consequences

The characterisation of large language model outputs as epistemically ambiguous artifacts connects the ontological argument of this paper to an institutional problem that knowledge systems are already encountering. As AI-generated content enters scientific literature, professional practice, and public discourse, existing verification institutions are confronted with a new kind of object: outputs that exhibit the surface features of credible knowledge claims while being produced by a process that does not ground those features in the epistemic relationships they ordinarily signal. The challenge is not merely one of accuracy. It is one of category: existing

verification practices were designed for outputs produced by accountable epistemic agents, and they apply imperfectly to outputs produced by large-scale representational compression systems.

This institutional mismatch is a reason to develop new verification practices—practices calibrated to the specific epistemic properties of the outputs, rather than practices imported from the evaluation of human-produced knowledge claims. Where Floridi (2023) characterises large language models as exhibiting agency without intelligence—a behavioural description—the present paper offers a complementary ontological account: the kind of system that produces such behaviour, and why that system's outputs carry a distinctive epistemic status that existing verification practices are not designed to handle. What those practices require, and what the candidate category makes possible, is a clear account of what kind of system produced the outputs in question. Verification practices designed for epistemically ambiguous artifacts will differ structurally from verification practices designed for knowledge claims: they will attend not only to the surface accuracy of the output but to the distribution of the training corpus, the domain coverage of the underlying representations, and the structural features of the query that determine which regions of the representational space are traversed. This is the agenda that a middle-level theory of large language models would underwrite—and that the absence of such a theory has so far prevented from being systematically articulated.

7. Conclusion

This paper began with a phrase—pattern matching—and the observation that its persistence as the dominant description of large language models is symptomatic of a deeper problem. The phrase is technically defensible. It accurately characterises the optimisation objective by which these systems are trained. But accuracy at one level of description is not accuracy at another, and the level at which pattern matching is accurate—the level of the training objective—is precisely the level that tells us least about what kind of computational system the objective produces.

The three ontologies that have developed to fill the gap the pattern-matching description leaves are not independent proposals. They are fragments of a fragmented intellectual tradition—each oriented toward a different level of the system, each succeeding within its orientation, each silent about what its orientation cannot see. The sequence-model ontology approaches from the training objective inward and stops at the boundary of internal organisation. The circuit-system ontology approaches from the components outward and stops short of a systems-level account. The quasi-cognitive ontology approaches from behaviour backward and stops short of the mechanism. Their failure modes are complementary because they are looking at the same object from positions that the existing intellectual landscape has made available, without the framework that would allow those perspectives to compose.

That framework does not yet exist. The paper has not provided it. What the paper has provided is a specification of what it must explain—representational richness from the training objective, compositional behaviour from local mechanisms, the relationship between internal structure and output, and the discontinuous emergence of new capacities during training—and a candidate category that identifies the level at which those explanatory tasks are coherently posed.

Large-scale representational compression systems are not a relabelling of an old phenomenon. They are an entity type that the existing ontologies, by virtue of their orientation, cannot quite see: a system that compresses not text as text but the representational structure of human knowledge as encoded in language, and that traverses that compressed structure to produce outputs that reconstruct the patterns of knowledge without occupying the epistemic position of a knower.

The value of the candidate category is precisely that it explains what the three rival ontologies each capture and why each is incomplete. The sequence-model description is correct because the training objective is what produced the compression. The circuit-system description is correct because the internal organisation of the compression—its geometry, its features, its circuits—is causally responsible for the outputs. The quasi-cognitive description is correct because the outputs reconstruct patterns from a representational space shaped by human knowledge, and those patterns include the patterns of reasoning. All three are correct, and none is sufficient, because none describes the level at which they connect.

At that level, the outputs of large language models appear as epistemically ambiguous artifacts: products of structured traversal through compressed representations of human knowledge, produced by a system that has internalised the patterns of knowledge without standing in the epistemic relationships those patterns normally presuppose. This characterisation has consequences for how such outputs should be received, verified, and integrated into knowledge practices —consequences that the absence of an adequate ontology has so far prevented from being systematically drawn. Drawing them out fully is the work of a subsequent inquiry. What this paper has attempted to establish is the prior condition: that the inquiry cannot be conducted adequately without first answering the question these pages have addressed. What kind of entity is a large language model? Not a pattern-matching engine, not a cognitive agent, not a circuit catalogue awaiting completion. Something that required a new kind of training to produce, that has arrived at an internal organisation no designer specified, and that generates a kind of output for which existing epistemic practices were not designed. A theory adequate to that entity is still to be built. The level at which it must operate is now, at least, visible.

References

Ashby, W. R. (1956). *Introduction to cybernetics.* Wiley.

Bateson, G. (1972). *Steps to an ecology of mind.* Chandler Publishing Co.

Beckmann, P., & Queloz, M. (2025). Mechanistic indicators of understanding in large language models. *arXiv*. https://arxiv.org/abs/2507.08017

Bricken, T., Templeton, A., Batson, J., Chen, B., Jermyn, A., Conerly, T., Turner, N. L., Anil, C., Denison, C., Askell, A., Lasenby, R., Wu, Y., Kravec, S., Schiefer, N., Maxwell, T., Joseph, N., Tamkin, A., Nguyen, K., McLean, B., ... Olah, C. (2023). Towards monosemanticity: Decomposing language models with dictionary learning. *Transformer Circuits Thread.* Anthropic. https://transformer-circuits.pub/2023/monosemantic-features

Delétang, G., Ruoss, A., Duquenne, P.-A., Catt, E., Genewein, T., Mattern, C., Grau-Moya, J., Wenliang, L. K., Aitchison, L., Orseau, L., Hutter, M., & Veness, J. (2024). Language modeling is compression. *In Proceedings of the International Conference on Learning Representations (ICLR 2024).* https://openreview.net/forum?id=jznbgiynus

Downes, S., Forber, P., & Grzankowski, A. (2024). LLMs are not just next token predictors. *arXiv*. https://arxiv.org/abs/2408.04666

Elhage, N., Nanda, N., Olsson, C., Henighan, T., Joseph, N., Mann, B., Askell, A., Bai, Y., Chen, A., Conerly, T., DasSarma, N., Drain, D., Ganguli, D., Hatfield-Dodds, Z., Hernandez, D., Jones, A., Kernion, J., Lovitt, L., Ndousse, K., Amodei, D., ... Olah, C. (2021). A mathematical framework for transformer circuits. *Transformer Circuits Thread.* Anthropic. https://transformer-circuits.pub/2021/framework/index.html

Elhage, N., Hume, T., Olsson, C., Schiefer, N., Henighan, T., Kravec, S., Hatfield-Dodds, Z., Lasenby, R., Drain, D., Chen, C., Grosse, R., McCandlish, S., Kaplan, J., Amodei, D., Wattenberg, M., & Olah, C. (2022). Toy models of superposition. *Transformer Circuits Thread.* Anthropic. https://transformer-circuits.pub/2022/toy_model/index.html

Brier, S. (2008). *Cybersemiotics: Why information is not enough!* University of Toronto Press.

Capurro, R., & Hjørland, B. (2003). The concept of information. *Annual Review of Information Science & Technology, 37*(1), 343–411. https://doi.org/10.1002/aris.1440370109

Floridi, L. (2023). AI as agency without intelligence: On ChatGPT, large language models, and other generative models. *Philosophy & Technology, 36*, 15. https://doi.org/10.1007/s13347-023-00621-y

Goddu, M. K., Noë, A., & Thompson, E. (2024). LLMs don't know anything: Reply to Yildirim and Paul. *Trends in Cognitive Sciences, 28(11)*, 963–964. https://doi.org/10.1016/j.tics.2024.06.008

Gurnee, W., & Tegmark, M. (2024). Language models represent space and time. *In Proceedings of the International Conference on Learning Representations (ICLR 2024).* https://openreview.net/forum?id=jE8xbmvFin

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*, 1–63.

He, Y., Zheng, W., Dong, Y., Zhu, Y., Chen, C., & Li, J. (2025). Towards global-level mechanistic interpretability: A perspective of modular circuits of large language models. *In Proceedings of the 42nd International Conference on Machine Learning.* PMLR 267. https://proceedings.mlr.press/v267/he25x.html

Hoffmann, J., Borgeaud, S., Mensch, A., Buchatskaya, E., Cai, T., Rutherford, E., Casas, D. d. L., Hendricks, L. A., Welbl, J., Clark, A., Hennigan, T., Noland, E., Millican, K., van den Driessche, G., Damoc, B., Guy, A., Osindero, S., Simonyan, K., Elsen, E., ... Sifre, L. (2022). Training compute-optimal large language models. *In Advances in Neural Information Processing Systems, 35*, 30016–30030. https://arxiv.org/abs/2203.15556

Kelly, M. (2026). Statistical structure and the failure of pointing: A system-class law for compression-based generative systems. *PhilSci Archive* preprint 29332. https://philsci-archive.pitt.edu/29332/

Ku, A., Campbell, D., Bai, X., Geng, J., Liu, R., Marjieh, R., McCoy, R. T., Nam, A., Sucholutsky, I., Veselovsky, V., Zhang, L., Zhu, J.-Q., & Griffiths, T. L. (2025). Levels of analysis for large language models. *arXiv*. https://arxiv.org/abs/2503.13401

Lindsey, J., Gurnee, W., Ameisen, E., Chen, B., Pearce, A., Turner, N. L., Citro, C., Abrahams, D., Carter, S., Hosmer, B., Marcus, J., Sklar, M., Templeton, A., Bricken, T., McDougall, C., Cunningham, H., Henighan, T., Jermyn, A., Jones, A., Persic, A., Qi, Z., Thompson, T. B., Zimmerman, S., Rivoire, K., Conerly, T., Olah, C., & Batson, J. (2025). On the biology of a large language model. *Transformer Circuits Thread.* Anthropic. https://transformer-circuits.pub/2025/attribution-graphs/biology.html

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.

McCoy, R. T., Yao, S., Friedman, D., Hardy, M., & Griffiths, T. L. (2024). Embers of autoregression show how large language models are shaped by the problem they are trained to solve. *Proceedings of the National Academy of Sciences, 121(41)*, e2322420121. https://doi.org/10.1073/pnas.2322420121

Merton, R. K. (1968). *Social theory and social structure.* Free Press.

Millière, R., & Buckner, C. (2024a). A philosophical introduction to language models—Part I: Continuity with classic debates. *arXiv*. https://arxiv.org/abs/2401.03910

Millière, R., & Buckner, C. (2024b). A philosophical introduction to language models—Part II: The way forward. *arXiv*. https://arxiv.org/abs/2405.03207

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

Nanda, N. (2022). *A walkthrough of interpretability in the wild.* https://www.neelnanda.io/mechanistic-interpretability/walkthrough-ioi

Olah, C., Cammarata, N., Schubert, L., Goh, G., Petrov, M., & Carter, S. (2020). Zoom in: An introduction to circuits. *Distill.* https://doi.org/10.23915/distill.00024.001

Olsson, C., Elhage, N., Nanda, N., Joseph, N., DasSarma, N., Henighan, T., Zeugmann, Z., Askell, A., Bai, Y., Chen, A., Conerly, T., Drain, D., Ganguli, D., Hatfield-Dodds, Z., Kernion, J., Lovitt, L., Ndousse, K., Amodei, D., & Olah, C. (2022). In-context learning and induction heads. *Transformer Circuits Thread.* Anthropic. https://transformer-circuits.pub/2022/in-context-learning-and-induction-heads/index.html

Pavlick, E. (2023). Symbols and grounding in large language models. *Philosophical Transactions of the Royal Society A, 381(2251)*, 20220041. https://doi.org/10.1098/rsta.2022.0041

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

Shanahan, M. (2024). Talking about large language models. *Communications of the ACM, 67(2)*, 68–79. https://doi.org/10.1145/3624724

Templeton, A., Conerly, T., Marcus, J., Lindsey, J., Bricken, T., Chen, B., Pearce, A., Citro, C., Ameisen, E., Jones, A., Cunningham, H., Turner, N. L., McDougall, C., MacDiarmid, M., Freeman, C. D., Sumers, T. R., Rees, E., Batson, J., Jermyn, A., Carter, S., Olah, C., & Henighan, T. (2024). Scaling monosemanticity: Extracting interpretable features from Claude 3 Sonnet. *Transformer Circuits Thread.* Anthropic. https://transformer-circuits.pub/2024/scaling-monosemanticity/index.html

Wang, K., Variengien, A., Conmy, A., Shlegeris, B., & Steinhardt, J. (2022). Interpretability in the wild: A circuit for indirect object identification in GPT-2 small. *arXiv*. https://arxiv.org/abs/2211.00593

Wiener, N. (1948). *Cybernetics: Or control and communication in the animal and the machine.* MIT Press.

Williams, I., Oldenburg, N., Dhar, R., Hatherley, J., Fierro, C., Rajcic, N., Schiller, S. R., Stamatiou, F., & Søgaard, A. (2025). Mechanistic interpretability needs philosophy. *arXiv*. https://arxiv.org/abs/2506.18852

Yildirim, I., & Paul, L. A. (2024). From task structures to world models: What do LLMs know? *Trends in Cognitive Sciences, 28(5)*, 404–415. https://doi.org/10.1016/j.tics.2024.02.008

Zou, A., Phan, L., Chen, S., Campbell, J., Guo, P., Ren, R., Pan, A., Yin, X., Mazeika, M., Li, A.-K., Li, Z., Gromov, A., Yin, N., Hendrycks, D., Fredrikson, M., & Zou, J. (2023). Representation engineering: A top-down approach to AI transparency. *arXiv*. https://arxiv.org/abs/2310.01405