Statistical Structure and the Failure of Pointing: A System-Class Law for Compression-Based Generative Systems

Matthew Kelly
Library Management Australia

Abstract

This paper proposes a system-class law for large language models as compression-based generative systems: statistical structure is preserved under compression, whereas indexical structure — the recoverable relation between an output and its originating evidential context — is not preserved in its pointing function. The asymmetry between statistical structure and indexical structure is not a contingent deficiency of current models but a structural property of compression-based generation. Compression preserves recurring regularities across the training distribution, but it does not thereby preserve particular pointing relations as such. From this law follows the Grounding Ceiling: increases in predictive capability can improve calibration and surface accuracy but cannot by themselves make output generation constitutively evidential, because the generation process does not traverse the evidential relations grounding requires. A conditional extension of the law, the Control Ceiling, follows if future empirical work confirms that inference proceeds through stable behavioural regimes shaped by pretraining: post-training control methods cannot then be assumed to arbitrarily rewrite that underlying regime structure. Together, these two ceilings establish a methodological consequence: current evaluation practices are organised primarily around surface plausibility rather than around the deeper properties this account identifies as explanatorily fundamental — grounding recoverability at the compression level and stable regime structure at the dynamical level. Once the evaluative target shifts, capability forecasting, interpretability, safety, and design change in kind.

Introduction

Large language models are among the most capable systems ever built for producing text that looks like knowledge. They generate fluent arguments, precise-sounding citations, authoritative syntheses, and expert-register prose across virtually every domain of human inquiry. The practical success of these systems is not in question.

The success of these systems has generated intense practical and theoretical attention. What remains underdeveloped is an account of why they have the properties they do — an account tethered to the mechanism that produces their outputs rather than inferred from the behaviour of those outputs alone. In the absence of such an account, evaluation, interpretive, and design practices have largely been calibrated to what is observable at the surface: output quality, task performance, and benchmark trajectories. What is in question is whether the field studying them has the theoretical resources to understand, at the right explanatory level, what kind of

system is producing these outputs — and what that understanding implies for how the field should evaluate, interpret, and design systems of this class.

This paper argues that those resources are not yet in place at the right explanatory level, and that acquiring them changes the questions the field can ask of systems of this class.

The argument begins with a distinction that current discourse often acknowledges but does not consistently preserve. There is a difference between a system that produces accurate outputs and a system that produces grounded outputs. An accurate output is one that is correct about what is the case. A grounded output is the product of a process that traverses the evidential relations that connect a claim to what it is about — the chain of inference, observation, and source that warrants the claim. These are not the same property. A system can be highly accurate without being grounded if its accuracy results from learned statistical regularities that track patterns in the world rather than from a generation process that itself traverses the evidential chains warranting the claim. The distinction matters because epistemic evaluation is concerned not only with whether a claim is correct, but with whether it is the product of a process that warrants trust.

The observation that large language models preserve the form of knowledge claims without preserving the evidential relations that would make those claims grounded has been advanced in different forms in recent philosophical and computational linguistics work (Bender & Koller, 2020; Floridi, 2023; Millière & Buckner, 2024; Pavlick, 2023). What the present paper adds is a mechanism-tethered account of why this asymmetry is a system-class property rather than a contingent limitation, and a methodological consequence for how the field should evaluate systems of this class once such an account is in place.

The paper's governing claim is that this distinction is not a contingent feature of current models — a deficiency that more capable systems will eventually overcome — but a structural property of compression-based generation as a system class. The argument is architectural: compression-based learning preserves recurring regularities across the training distribution; it does not thereby preserve particular pointing relations as such. Statistical structure — the patterns, forms, and regularities of knowledge-producing discourse — recurs and is preserved. Indexical structure — the recoverable relation between an output and its originating evidential context — does not generalise across instances in the same way and is not preserved in its pointing function. The asymmetry is therefore constitutive of the system class, not merely a description of what current systems happen to lack.

That is the paper's law: in compression-based generative systems, statistical structure is preserved under compression, whereas indexical structure is not preserved in its pointing function. The body of the paper earns this law mechanistically, derives two ceilings from it — one unconditional, one conditional on a further empirical programme — and shows what follows methodologically for how the field should evaluate systems of this class.

The paper is not an argument that current large language models are inadequate for practical purposes. It is an argument that the field studying them is, at present, evaluating these systems against targets misaligned with the properties this account identifies as explanatorily fundamental — and that those targets change once such an account is in place.

1. Statistical Structure and the Failure of Pointing Under Compression-Based Generation

Large language models are trained by compression. The training process extracts what recurs across a vast distribution of human-generated text and encodes it in the model's parameters. What recurs is preserved. What does not recur is not. This is not a limitation of current architectures or a consequence of insufficient data. It is what compression-based learning is.

The training distribution contains two kinds of structure that behave differently under this process.

The first kind is statistical structure: patterns of co-occurrence, syntactic form, argumentative register, genre convention, the characteristic rhythms of disciplinary prose, the typical shape of a well-formed citation, the recognisable cadence of expert hedging. Statistical structure recurs. It appears across millions of instances in the training distribution, which means the compression process can learn it, encode it, and reconstruct it. A model trained on this distribution learns what knowledge-claims look like, what evidence-talk sounds like, what the surface form of an evidentially well-supported argument is.

The second kind is indexical structure.[1] Citations, source references, dated observations, and other tokens that function to pick out particular evidential contexts are indexical in the sense relevant here: they function to establish a recoverable relation between the present claim and a specific originating context — a particular experiment, a particular source, a particular moment at which something was checked against the world. Indexical structure, as used here, means precisely that: the recoverable relation between an output token and its originating evidential context.

Indexical structure does not generalise across instances in the way statistical structure does. The specific experiment, the particular source, the dated observation — these are instance-specific by definition. They belong to one occasion. They do not recur in a way that allows compression to preserve their pointing function as recoverable links to originating evidential contexts. A model can learn that citations look like "Smith (2019) shows that …." It cannot, from compression alone, preserve the relation by which "Smith (2019)" functions as a recoverable pointer to what that text actually shows, because that is not a statistical regularity — it is a particular fact about a particular text.

This is the asymmetry the paper identifies as its governing law: In compression-based generative systems, statistical structure is preserved under compression, whereas indexical structure is not preserved in its pointing function.[2]

The precision of that final clause matters. The claim is not that models never reproduce names, dates, citations, or specific-sounding claims. They do, frequently and fluently. The claim is that when a model outputs the string "according to Smith (2019)" or "the study found that," it

[1] The term "indexical" is used here in a sense derived from the philosophy of language tradition originating with Kaplan (1989), where indexicals are expressions whose reference depends on the context of use rather than being fixed independently of that context. The paper extends this concept operationally to epistemic contexts: indexical structure is the recoverable relation between an output and its originating evidential context, distinct from the surface tokens that may resemble such pointing without preserving the relation.

[2] A substantial philosophical literature addresses whether LLMs can achieve grounding in various senses. Mollo and Millière (2026) argue that LLMs can achieve referential grounding — internal states standing in causal-informational and historical-selectional relations to worldly entities — even without multimodality or embodiment; Pavlick (2023) argues that the question of whether LLMs encode semantic content is open and depends on empirical characterisation of internal representations. The claim of this paper is orthogonal to both. The paper does not ask whether LLM internal states can be about worldly entities in the teleosemantic sense, nor whether they encode structured semantic content. It asks whether specific output tokens preserve the recoverable relation between an output and its originating evidential context — a question Mollo and Millière (2026) would classify under epistemic grounding, which they explicitly distinguish from the referential grounding their paper addresses and treat as insufficient to solve the Vector Grounding Problem. Crucially, even if LLMs achieve referential grounding in Mollo and Millière's (2026) sense, this is compatible with the law stated here: a system whose internal states genuinely track general world features via selection pressure can still fail to preserve the specific recoverable relation between "Smith (2019)" in an output and what that source actually shows. The two claims operate at different levels and do not conflict.

is reproducing the surface form of indexical pointing — the tokens that, in human-generated text, would establish a recoverable relation to a source — without preserving the pointing function itself. The token "Smith (2019)" in a model output is not a pointer to Smith (2019). It is a statistically appropriate continuation given the surrounding context. The form of pointing is preserved. The pointing function is not.

This distinction is what makes the law a claim about a specific system class rather than a general observation about lossy compression. A JPEG loses indexical information too — the provenance of the image, the context of its creation, the chain of custody from capture to file. But a JPEG does not reconstruct outputs that look like knowledge claims. It does not produce text that reads "according to the original scene." A compression-based generative system does exactly that. It reconstructs the surface form of evidential pointing — fluently, at scale, across a wide range of domains of human knowledge — while the pointing function that surface form ordinarily serves has not been preserved. The gap between the form and the function is invisible at the surface. That invisibility is what makes the asymmetry epistemically consequential, and what makes this law specific to systems that generate text with the structure of knowledge claims rather than to compression in general.

Scale deepens the preservation of statistical structure and the fluency of indexical-looking form; it does not by itself restore indexical structure in its pointing function.

2. Why Compression Cannot Preserve Pointing Relations

The preceding section stated the law. This section argues for it. The asymmetry between statistical and indexical structure under compression is not an empirical observation about what current models happen to do. It is a consequence of what compression-based generation is as a process. The argument is architectural.

Compression-based learning works by finding and encoding regularities. The formal equivalence between language modelling and compression has been established by Delétang et al. (2024), who demonstrate that language modelling under compression criteria produces predictions equivalent to standard next-token prediction objectives. Given a training distribution of sufficient size and diversity, a compression process identifies what recurs — what patterns of token sequence appear often enough, in enough varied contexts, to be stably encoded in the model's parameters. What recurs is compressed into reusable structure. What does not recur is not encoded as such. This is the basic logic of the process, and increasing model scale, architecture depth, or training duration does not alter that logic. A larger model with more parameters encodes more regularities, at finer grain, across a wider range of domains. It does not change what the process is doing. It does more of it.

Statistical structure is exactly the kind of structure this process captures. Syntactic patterns recur. Genre conventions recur. The characteristic form of an academic citation recurs. The argumentative structure of an empirical claim recurs. The hedging register of expert uncertainty recurs. All of these are patterns — regularities that appear across many instances in the training distribution — and compression encodes them as reusable structure. When a model reconstructs text that looks like a well-formed knowledge claim, it is drawing on this encoded statistical structure. The output is fluent because the patterns it instantiates were genuinely present in the distribution, recurrent across many instances, and learnable from it.

Indexical structure is not what this process captures, and the reason is not a contingent limitation. It is the same logic applied to a different kind of thing. The recoverable relation between an output token and its originating evidential context — the pointer from "Smith

(2019)" to what Smith (2019) actually contains, from "the experiment showed" to the particular experiment that showed it — is instance-specific. It belongs to one occasion in the training distribution, not to a recurring pattern across many. The compression process does not thereby preserve instance-specificity in its pointing function as a recoverable relation, because particular pointing relations are not the kind of regularities that recurrence-based compression encodes as such. A model exposed to many contexts in which Smith (2019) is cited learns the statistical regularities of how such citations occur — in what contexts, following what claims, in what disciplinary registers. It does not learn the pointing function itself, because the pointing function is not a statistical regularity. It is a particular fact about a particular text. What can recur, and therefore be learned, is the general pattern by which citation-tokens are used in discourse — the role-schema of pointing: that citations follow claims, name sources, and warrant the claims they support. But learning that role-schema is not the same as preserving any particular pointing relation. A particular pointing relation is constituted by a recoverable link between this token and this specific evidential source — a link that is downstream of an evidential path that can in principle be retraced. Human scholarly citation can instantiate such a link because the scholar actually followed that evidential path: read the source, judged what it warranted, and cited it on that basis. Compression-based generation can learn the statistical pattern of such acts without constituting any particular pointing relation. The model learns what citing looks like and when it occurs. It does not thereby create the recoverable link that citing, in the epistemically relevant sense, consists in.

This is why the law is not about current models being insufficiently capable. The problem is not that present architectures are too shallow to encode indexical relations. The problem is that compression-based learning is the wrong kind of process for preserving particular pointing relations as recoverable relations. A deeper architecture, trained on more data, learns more statistical regularities. It does not thereby come to preserve what recurrence-based compression does not preserve as such. The gap is not between current capability and future capability. It is between what compression does and what pointing requires.

The scaling corollary follows directly from this mechanism, and it is important to keep it downstream of the law rather than part of it. As model scale increases, the encoding of statistical structure deepens. Calibration improves. Surface accuracy improves. The fluency of indexical-looking form improves — larger models produce more convincing citations, more precise-sounding source references, more authoritative evidential rhetoric. None of this restores indexical structure in its pointing function, because what is improving is the statistical encoding of how pointing looks, not the preservation of any particular pointing relation. Scale makes the surface form more convincing precisely as it leaves the pointing function as unpreserved as before. The gap does not close with scale. It becomes less visible.

It is the mechanism — not the empirical record of current model behaviour — that licenses treating the asymmetry as a system-class property.The law holds because the process that produces these systems cannot preserve what pointing is. It is not inferred from the observation that current models fail to point. That distinction is what makes the law a claim about the class rather than a claim about present deficiency, and it is what grounds the ceiling that follows: if the generation process cannot traverse evidential relations, no increase in capability built on the same process can make its outputs constitutively evidential.

3. Why Capability Scaling Cannot Close the Gap

The asymmetry established by the preceding argument has a direct downstream consequence. If compression-based generation preserves the surface form of evidential pointing while failing to preserve indexical structure in its pointing function, then the outputs of that process are not constitutively evidential.[3] They are not the product of a process that traverses the evidential relations that grounding requires. For compression-based generative systems, increases in predictive capability alone can improve calibration and surface accuracy, but cannot by themselves make output generation constitutively evidential. This is the Grounding Ceiling.

The argument is a direct extension of the mechanism. Grounding, in the epistemically relevant sense, requires that an output be the product of a process that arrives via the relevant chain of inference and evidence connecting the claim to what it is about. A grounded claim is not merely accurate. It is accurate because the process that produced it traversed the relevant evidential relations. What makes a court transcript reliable is that it is produced by a process — sworn testimony, cross-examination, the rules of evidence — that tracks what actually happened, not merely that it tends to be correct. What makes a peer-reviewed finding trustworthy is that it arrived through a process of experimental design, observation, and structured scrutiny, not merely that it is accurate.

Compression-based generation does not traverse evidential relations. It reconstructs statistically appropriate continuations from encoded regularities. When it produces an accurate claim — and it frequently does — that accuracy is a consequence of learned statistical regularities that track patterns in the world, not of the generation process tracking the evidential chain behind the claim. The accuracy is real. The grounding does not come from the process that produced it.

Scaling improves calibration — the alignment between a model's expressed confidence and its actual accuracy rate — and it improves the fluency and precision of surface form. What it does not do is change the nature of the generation process. A more capable model produces more convincing outputs by the same mechanism: statistical reconstruction from compressed regularities. The process still does not traverse evidential relations, regardless of how accurately its outputs match what such relations would support.

The Grounding Ceiling is therefore not a claim about accuracy. It is a claim about process. It says that there is a property — constitutive evidentiality — that compression-based generation cannot produce, not because it lacks accuracy, but because it is the wrong kind of process. Accuracy and constitutive evidentiality are not the same property, and one cannot be scaled into the other.

The most immediate objection to this ceiling is retrieval-augmented generation. If a model's outputs are conditioned on retrieved documents — actual sources, with recoverable provenance — does that not restore the pointing function and close the gap? The objection is real but misdirected. Retrieval-augmented architectures address the gap only by adding a structurally distinct mechanism outside the compression layer: a retrieval component that locates and surfaces external documents or passages (Lewis et al., 2020). That component is not compression-based generation. It is a world-tracking mechanism added to the system. What

[3] The concept of constitutive evidentiality draws on the epistemology of testimony, which distinguishes beliefs that are merely true from those produced by a process standing in the right epistemic relation to their subject matter. For the canonical account of what process-dependence requires for testimonially warranted belief, see E. Fricker (1994, 1995). The parallel is structural: just as testimonial knowledge requires that the hearer's belief be grounded in the right epistemic relation to a knowledgeable source rather than merely correlated with true claims, constitutively evidential outputs require that the generation process traverse the relevant evidential relations.

retrieval does is change the system — it supplements the compression layer with a component that does what compression cannot — rather than closing the gap within the compression layer itself. The Grounding Ceiling applies to the compression-based generation layer. Retrieval does not defeat the ceiling; it works around it by importing a different kind of process.

This matters for how retrieval-augmented systems should be understood. They are not compression-based generative systems that have acquired the ability to ground their outputs. They are hybrid systems in which a compression-based generation component is supplemented by a retrieval component that locates and surfaces external documents or passages. To the extent that the system preserves a recoverable source-link, that linkage is supplied by the retrieval component rather than by the compression-based generation layer as such. In such systems, whatever grounding the output has comes from the retrieval component, not from the generation component as such. The ceiling still holds for the generation layer. What changes is the system's overall architecture, not the properties of the process the ceiling describes.

The Grounding Ceiling is this paper's unconditional anchor. It follows from the law without dependence on any further empirical programme. Whatever is subsequently established or disputed about inference-time dynamics or control-layer architecture, this ceiling stands: compression-based generation cannot, by increases in capability alone, produce outputs that are constitutively evidential. That is a system-class property, and it is the foundation on which the paper's methodological consequence rests.

4. What Post-Training Control Cannot Do

The preceding sections establish the law, its mechanism, and the Grounding Ceiling that follows from it. This section adds a conditional extension. It does not follow unconditionally from the law. It holds only if inference in large language models exhibits genuine regime structure of the kind the following account proposes. That condition is stated at the outset because the Control Ceiling's epistemic status is different from the Grounding Ceiling's, and conflating them would misrepresent the paper's argument.

The trajectory-level account on which the Control Ceiling depends proposes that inference-time generation is more than a sequence of context-conditioned probability draws: it is the evolution of a system state through a representational landscape with topographic properties (Doimo et al., 2024; Rigollet, 2025). Recent mathematical work on transformer dynamics has identified clustering and metastable behaviour consistent with taking such a dynamical account seriously (Geshkovski et al., 2025; Karagodin et al., 2024). The specific landscape formulation adopted here remains a theoretical proposal rather than an established result. On this account, the landscape contains regions of greater and lesser stability, barriers between regions, and transition costs that are asymmetric with respect to direction of travel. Behavioural modes correspond to relatively stable regions; transitions between modes have variable costs; and the cost of moving from one mode to another is not in general equal to the cost of the reverse transition. The Control Ceiling is conditional on this account: if it is confirmed, post-training control methods cannot be assumed to arbitrarily rewrite the regime structure induced by pretraining; if it is not, the ceiling does not.

The levels-of-explanation framework invoked here descends from Marr's (1982) account of distinct explanatory levels and from the mechanistic explanation tradition in the philosophy of science (Machamer et al., 2000). On this tradition, a level of explanation is genuine when it introduces concepts and questions that are not naturally expressible within adjacent levels without loss of explanatory power. The trajectory account proposes exactly such a level: regime

structure, asymmetric transition costs, and path dependence are not naturally expressible within accounts that model inference as a sequence of context-conditioned probability draws.

At that level, the representational landscape through which inference-time trajectories move is shaped primarily by pretraining. Pretraining compression dynamics shape which regions of the representational space are stable, how deep the basins are, and what the topography of transitions between them looks like. This is a consequence of the mechanism established in the preceding sections: pretraining encodes the statistical regularities of the training distribution, and those regularities structure the geometry of the space the model traverses during inference. The landscape is not a neutral medium. It is structured by what pretraining compressed.

Against this background, post-training alignment methods — reinforcement learning from human feedback, direct preference optimisation, constitutional approaches — operate as a control layer within this landscape. They can raise the cost of entering certain basins, modulate transition thresholds, gate certain trajectories at the point of initialisation, and locally reshape traversal in ways that produce more aligned outputs. These are genuine and often substantial effects. The Control Ceiling is not a claim that post-training alignment does nothing. It is a claim about what post-training alignment cannot be assumed to do.

Specifically: post-training control methods cannot be assumed to arbitrarily rewrite the basin topology that pretraining produced, without empirical demonstration that the mechanisms available to post-training are sufficient to overcome that topology. The reasons are structural. Post-training methods typically operate over a narrower distribution than pretraining — they are applied to a model whose representational geometry is already formed, using a training signal that covers a fraction of the space pretraining shaped. The objective is different: pretraining optimises for next-token prediction over the full distribution; post-training optimises for alignment with human preferences over a curated subset. The mechanisms available to post-training in these settings — gradient updates over a narrower alignment distribution with a different objective — are not, by default, mechanisms for arbitrarily rewriting the full representational geometry. They are designed to shift the model's behaviour in targeted ways within a geometry that pretraining has already established.

If a model's representational landscape contains deep basins corresponding to misaligned modes, post-training control faces structural resistance that increases with basin depth. It can raise the cost of entering those basins. It can gate trajectories away from them at initialisation. What it cannot be assumed to do is eliminate the basins from the landscape — because the mechanisms that produced them, gradient descent over the full pretraining distribution, are not available to post-hoc alignment methods operating over a narrower distribution with a different objective. This is the Control Ceiling.

The ceiling matters for the paper's evaluation argument in a specific way. If regime structure is real, then the basin topology of the representational landscape is an explanatorily fundamental property of the system — one that structures how inference behaves under perturbation, how stable behavioural modes are, and how much structural resistance alignment faces. Current evaluation is not designed to measure this property. It measures output behaviour under standard conditions, which gives no direct information about the landscape's topology or the depth of its basins. A system that produces aligned outputs under normal conditions may do so because its landscape topology genuinely supports alignment, or because post-training control is successfully suppressing trajectories into misaligned basins — and current evaluation does not reliably distinguish these two cases. If the trajectory-level account is confirmed, that distinction becomes one of the most important things the field could know about a deployed system.

The Control Ceiling is therefore not a counsel of despair about alignment. It is a claim about what theory, if confirmed, reveals about the limits of post-training control as a form of intervention. Post-training control is real, it works, and it matters. What the ceiling says is that it works within a landscape it did not create and cannot fully rewrite — and that a field without the theoretical resources to characterise that landscape lacks the information required to determine whether its control interventions are structural or merely local, stable or fragile.

5. Evaluation After Theory: From Surface Success to Explanatory Targets

That question — whether interventions are structural or merely local — is not one current evaluation practices are designed to answer, and for a reason that applies more broadly than the Control Ceiling alone.

Current evaluation practices for large language models are not badly designed. Benchmark performance assesses whether a system produces outputs that satisfy specified task criteria. Capability assessment tracks the range and reliability of a system's performance across problem domains. Output inspection evaluates whether outputs meet quality standards relative to human judgement. These are coherent instruments for coherent purposes, and they have been genuinely useful in tracking the development of systems whose properties were not yet theoretically understood.

The problem is not that these instruments are poorly constructed. The problem is that they are optimised for a property that the account developed here shows is not explanatorily fundamental. They measure surface plausibility — the degree to which outputs satisfy task criteria and correspond to human expectations of quality. Surface plausibility is real, it scales with capability, and it matters for many practical purposes. What it does not measure is the property this account identifies as explanatorily fundamental: the relation between the generation process and the evidential structure of what it produces. That relation is what the law and the Grounding Ceiling describe. It is not visible at the surface, and no current benchmark is designed primarily to detect whether the relation between an output and its evidential context is recoverable from the generation process itself. McCoy et al. (2024) demonstrate empirically that large language model behaviour is systematically shaped by the statistical structure of the training objective rather than by the underlying competence the evaluation task is meant to probe — a finding consistent with the deeper mismatch the present paper identifies between surface success and explanatorily fundamental properties.

This is a mismatch relative to theory, not an indictment of current practice. Before a theory at the relevant explanatory level was available, there was no principled basis at the right explanatory level for identifying surface plausibility as the wrong target. Evaluation practices were calibrated to what was observable and what mattered for the purposes the field was then pursuing. The theory changes the situation not by revealing that previous work was mistaken, but by identifying what the field should now be measuring if it wants its evaluations to track the properties that determine these systems' epistemic character. The question of what properties of information systems are epistemically relevant — and what is lost when evaluation is calibrated to surface form rather than evidential structure — has a long history in library and information science (Capurro & Hjørland, 2003). The present paper's methodological consequence joins that conversation by specifying, for compression-based generative systems, what the relevant mismatch is and why it is structural rather than contingent.

What should evaluation target instead? The answer has two parts, and their epistemic status is not equal.

The first target is grounding recoverability: the degree to which the relation between an output and its originating evidential context is recoverable by an independent verifier. This target follows unconditionally from the law. If compression-based generation preserves the surface form of evidential pointing while failing to preserve indexical structure in its pointing function, then grounding recoverability is precisely the property that current evaluation is not designed to measure. High benchmark performance and strong human quality assessment do not imply high grounding recoverability; the two properties come apart as a consequence of the law, not as an occasional anomaly. Grounding recoverability is therefore the unconditional evaluation target: it should be measured regardless of what is subsequently established about inference-time dynamics or control-layer architecture, because it follows from the compression-level properties the law describes.

The second target is regime structure: the degree to which a system's inference-time behaviour exhibits the stable modes, asymmetric transition costs, and path dependence that the trajectory-level account proposes. This target is conditional. It becomes evaluatively relevant only if inference exhibits genuine regime structure of the kind the empirical programme described in Section 4 is designed to confirm. If that programme is confirmed, regime structure becomes a second explanatorily fundamental property that current evaluation is not designed to measure — not because current instruments are insensitive to it by accident, but because they are not designed to detect dynamical properties of inference. If the programme is not confirmed, this second target does not follow. The asymmetry between the two targets must be preserved: grounding recoverability is the paper's unconditional evaluative consequence; regime structure is the extension that a confirmed trajectory-level account would add.

The mismatch between current evaluation and these two targets is not correctable without a change in the evaluation paradigm. This is not because better benchmarks cannot be designed. It is because the current paradigm is organised around a different question — how well does this system perform on tasks? — rather than the question this account now makes possible: what are the explanatorily fundamental properties of this system, and how well do its outputs reflect them? Transitioning from one question to the other is not simply a matter of adding new benchmarks to existing suites. It requires reorienting what evaluation is for: not scorekeeping across tasks, but diagnosing the system's properties at the levels this account identifies as explanatorily fundamental.

This reorientation is what this account makes possible and what the field's current practice does not yet reflect. The gap between what evaluation currently measures and what it should measure, once such an account is in place, is the methodological consequence the paper exists to identify. It does not follow from this account as a distant implication. It follows directly: if surface plausibility and grounding recoverability come apart as a system-class property, and if current evaluation measures surface plausibility, then current evaluation is misaligned with what this account identifies as explanatorily fundamental. The correction is not incremental. It is structural.

6. What Changes Once the Evaluative Target Shifts

The preceding sections have established a governing law, derived two ceilings from it, and identified the methodological consequence for how the field should evaluate systems of this class. This section asks what follows for the research field once such a consequence is in place. It does not introduce new theoretical claims. Its work is to show what changes in kind — not in degree — across four research programmes once the field shifts its evaluative target from surface

success to the properties this account identifies as explanatorily fundamental. The change in each case is not that the field gains finished answers. It is that the field gains the right questions.

6.1 Capability Forecasting

Current capability forecasting is organised primarily around extrapolation. The field observes benchmark trajectories, scaling curves, and emergent behaviours, and projects forward from them. This is not irrational — in the absence of theory, extrapolation from observed regularities is the best available tool. But it is a pre-theoretic practice, and it has a characteristic limitation: it does not by itself distinguish between capabilities that are structurally possible for systems of this class and capabilities that happen to have appeared in systems built so far. It does not by itself determine which observed behaviours are consequences of the system-class architecture and which are contingent on specific training choices. It does not provide a basis for identifying where structural limits lie, because it has no account of what limits the class as such.

Once such an account is in place, forecasting changes in kind. The law and its mechanism establish that compression-based generation has a structural limit with respect to indexical structure — one that does not change with scale. That means the field can now ask not only what current systems can do, but what systems of this class can in principle do, and where the boundaries of that space lie. Capability forecasting becomes less a matter of extrapolating observed curves and more a matter of reasoning about what the system-class architecture permits and excludes. The questions change: not only how capable will the next system be, but which kinds of capability are structurally within reach of this class and which require a different kind of process.

6.2 Interpretability

Much current interpretability research is forensic in orientation. It proceeds by identifying interesting mechanisms — attention heads, features, circuits — and characterising what they do (Olsson et al., 2022). The findings are real and valuable, but they accumulate as a collection of local discoveries without a theoretical framework that organises them. Each finding answers the question: what does this component do? The framework that would allow interpretability to answer the question: why does this system have this kind of component, and how does it relate to the system's other properties, has not yet been available at that explanatory level.

Once such an account is in place, interpretability changes in kind. Findings about internal mechanisms can be organised around the lawful structures this account identifies: the compression of statistical regularities, the failure to preserve indexical structure in its pointing function, the topography of the representational landscape if regime structure is confirmed. An attention head that functions in a certain way is no longer an isolated discovery — it is an instance of a lawful structure within a system class with known structural limits. Interpretability shifts from post-hoc tracing of what components happen to do toward explanation of why systems of this class develop the internal organisation they do. The field gains a framework within which individual findings become intelligible as parts of a larger account.

6.3 Safety

A good deal of current safety research proceeds by identifying failure modes and designing interventions to correct them. Many of these interventions operate at the level of observable behaviour: output filtering, reward shaping, behavioural fine-tuning, red-teaming against known attack patterns. These are genuine contributions, and they have reduced observable harms. Their

limitation is that they are calibrated to observed failures rather than to the structural properties that generate failure. Without a theory of what the system is, safety research lacks a principled basis for asking which interventions act at the right explanatory level and which are shallow corrections that leave the underlying structure intact.

Once such an account is in place, safety research changes in kind. The Grounding Ceiling identifies a structural property — the separation of surface plausibility from grounding recoverability — that no surface-level intervention can close. The Control Ceiling, if the trajectory account is confirmed, identifies a structural limit on what post-training alignment can achieve relative to the landscape pretraining produced. Empirical findings that alignment effects are mediated by specific low-dimensional directions in the representational space (Arditi et al., 2024) are consistent with this picture: they suggest that alignment operates at a level that is neither purely surface nor fully structural, which is precisely the kind of distinction a levels-based theory makes tractable. Safety research gains a theoretical basis for distinguishing interventions that address structural properties from interventions that correct surface behaviour while leaving structural properties unchanged. The question changes: not only does this intervention reduce observed harmful outputs, but does it act at the explanatory level where the relevant structural property is located.

6.4 Design

Current system design is navigated primarily by engineering intuition, scaling heuristics, and empirical trial. This is appropriate when theory is absent — when the space of possible systems is not yet mapped, engineering intuition is often the best available guide. Its limitation is that it does not by itself indicate where the boundaries of the space lie, which design choices are structurally consequential and which are interchangeable, or what tradeoffs are inherent to the class rather than contingent on specific implementation decisions.

Once such an account is in place, design changes in kind. The system class has structural limits identified by this account: compression-based generation preserves statistical structure but does not preserve indexical structure in its pointing function, regardless of architectural choices made within that class. That is a genuine constraint on the design space — not an engineering obstacle to be optimised around, but a structural feature of what systems of this class are. Designers gain the ability to reason about what their choices can and cannot change, which kinds of augmentation address structural properties and which do not, and where the boundaries of the class's capability space lie. Design becomes less a matter of navigating an unmapped space by feel and more a matter of working within an understood architecture with known properties and limits.

These four changes share a common structure. In each case, the field moves from a practice organised around observable surface properties — benchmark trajectories, component behaviours, failure outputs, engineering performance — toward a practice organised around the explanatorily fundamental properties this account identifies. The change is not that the field stops caring about surface properties. It is that surface properties are now understood in relation to a deeper account of what produces them. That relation is what theory provides, and it is what changes the kind of question the field can ask.

Theory maturity does not complete any of these research programmes. It changes what completion would look like. The field does not gain finished answers in capability forecasting, interpretability, safety, or design. It gains the theoretical resources to know what answers in those

domains would have to establish, and what would count as genuine progress toward them rather than movement along the surface of a problem whose structure remains uncharacterised. That is what theory changes. Not the systems. Not the data. Not the engineering. The field's understanding of what it is working with — and therefore what it is working toward.

Conclusion

This paper has argued that a compression-based account of large language models — one that treats these systems as generative processes that preserve statistical structure while failing to preserve indexical structure in their pointing function — changes what evaluation should target. The argument rests on a governing law, derives two ceilings from it, and identifies the methodological consequence that follows.

The law is this: in compression-based generative systems, statistical structure is preserved under compression, whereas indexical structure is not preserved in its pointing function. This is not a claim about the limitations of current models as they currently stand. It is a claim about what compression-based generation is as a process — one that preserves recurring regularities across the training distribution and does not thereby preserve particular pointing relations as such. The asymmetry between statistical and indexical structure under compression is a system-class property, and increases in scale do not by themselves alter it.

From this law the Grounding Ceiling follows unconditionally. Increases in predictive capability can improve calibration and surface accuracy, but cannot by themselves make output generation constitutively evidential. The generation process does not traverse the evidential relations that grounding requires. That is a structural feature of the class, not a temporary deficiency, and no intervention confined to the compression layer can be assumed to change it. The Grounding Ceiling is the paper's unconditional anchor — it stands regardless of what is subsequently established about inference-time dynamics or control-layer architecture.

The Control Ceiling is a conditional extension. It holds if inference exhibits genuine regime structure of the kind the trajectory-level account proposes. If that condition is met, post-training control methods cannot be assumed to arbitrarily rewrite the basin topology that pretraining produced. They can steer, gate, and locally reshape traversal within the learned landscape, but they do not by default have unrestricted authority over the geometry pretraining established. The epistemic status of this ceiling is different from the Grounding Ceiling's, and the paper has been careful to preserve that asymmetry throughout. The Grounding Ceiling follows from the law. The Control Ceiling follows from the law together with a further empirical programme that remains to be confirmed.

Together, the law and the two ceilings establish a methodological consequence that is the paper's central contribution. Current evaluation practices — benchmark performance, output inspection, capability assessment — are organised around surface plausibility, which scales with capability and is real for many practical purposes. They are not wrong on their own terms. They are misaligned with what this account identifies as explanatorily fundamental: grounding recoverability at the compression level, and stable regime structure at the dynamical level if the trajectory account is confirmed. These two properties are not captured by surface plausibility. They are not recoverable within the current evaluation paradigm, because that paradigm is designed to measure something else. The correction this account implies is not incremental. It is structural — a reorientation of what evaluation is for, from scorekeeping across tasks to diagnosing the system's properties at the levels this account identifies as fundamental.

Once that reorientation is made, four research programmes change in kind. Capability forecasting gains the resources to ask what systems of this class can in principle do, rather than extrapolating from observed benchmark curves. Interpretability gains a framework within which individual findings become intelligible as instances of lawful structure within a system class with known structural limits. Safety research gains a basis for distinguishing interventions that act at the right explanatory level from those that correct surface behaviour while leaving structural properties unchanged. Design gains a clearer map of the system class's structural constraints, tradeoffs, and limits, rather than an engineering space navigated primarily by scaling intuition. In each case, the change is not that the field gains finished answers. It is that the field gains the theoretical resources to know what answers in those domains would have to establish. Theory maturity does not complete these research programmes. It changes what completion would look like.

The paper has not confirmed the trajectory-level account. It has not measured grounding recoverability across deployed systems. It has not demonstrated the Control Ceiling empirically. What it has done is establish, from the nature of compression-based generation as a process, that the field is currently evaluating these systems against targets misaligned with the explanatorily fundamental properties of the system class — and that this misalignment is not correctable without a change in the current paradigm, because that paradigm is not designed to detect the properties this account identifies as fundamental. Accepting that claim does not require waiting for the empirical programmes to conclude. It requires accepting the law and its mechanism — claims about what compression-based generation is as a process, supported by the architecture and mechanism rather than inferred from contingent features of current model behaviour.
That is what theory gives a field. Not the answers. The right questions, stated precisely enough to be answered.

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*, 136037–136083. https://proceedings.neurips.cc/paper_files/paper/2024/file/f545448535dfde4f9786555403ab7c49-Paper-Conference.pdf

Bender, E. M., & Koller, A. (2020). Climbing towards NLU: On meaning, form, and understanding in the age of data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 5185–5198). https://aclanthology.org/2020.acl-main.463/

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

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://proceedings.iclr.cc/paper_files/paper/2024/hash/3cbf627fa24fb6cb576e04e689b9428b-Abstract-Conference.html

Doimo, D., Serra, A., Ansuini, A., & Cazzaniga, A. (2024). The representation landscape of few-shot learning and fine-tuning in large language models [Preprint]. arXiv. https://arxiv.org/abs/2409.03662

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

Fricker, E. (1994). Against gullibility. In A. Chakrabarti & B. K. Matilal (Eds.), *Knowing from words: Western and Indian philosophical analysis of understanding and testimony* (pp. 125–161). Kluwer Academic Publishers. https://doi.org/10.1007/978-94-017-2018-2_8

Fricker, E. (1995). Critical notice: Telling and trusting: Reductionism and anti-reductionism in the epistemology of testimony. *Mind*, *104*(414), 393–411. https://doi.org/10.1093/mind/104.414.393

Geshkovski, B., Letrouit, C., Polyanskiy, Y., & Rigollet, P. (2025). A mathematical perspective on transformers. *Bulletin of the American Mathematical Society*, *62*(3), 427–479. https://doi.org/10.1090/bull/1863

Kaplan, D. (1989). Demonstratives: An essay on the semantics, logic, metaphysics, and epistemology of demonstratives and other indexicals. In J. Almog, J. Perry, & H. Wettstein (Eds.), *Themes from Kaplan* (pp. 481–563). Oxford University Press.

Karagodin, N., Polyanskiy, Y., & Rigollet, P. (2024). Clustering in causal attention masking. *Advances in Neural Information Processing Systems*, *37*, 115652–115681. https://proceedings.neurips.cc/paper_files/paper/2024/hash/d18d208fa9c333483e5724ade7beff0f-Abstract-Conference.html

Kim, J., Wu, D., Lee, J. D., & Suzuki, T. (2025). Metastable dynamics of chain-of-thought reasoning: Provable benefits of search, RL and distillation [Preprint]. arXiv. https://arxiv.org/abs/2502.01694

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W., Rocktäschel, T., Riedel, S., & Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. Advances in Neural Information Processing Systems, 33, 9459–9474.

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

Mollo, D. C., & Millière, R. (2026). The vector grounding problem. *Philosophy and the Mind Sciences*, *7*(1). https://doi.org/10.33735/phimisci.2026.12307

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

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., Ndousse, K., Amodei, D., Brown, T., Clark, J., Kaplan, J., McCandlish, S., & Olah, C. (2022). In-context learning and induction heads [Preprint]. arXiv. https://arxiv.org/abs/2209.11895

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

Rigollet, P. (2026). The mean-field dynamics of transformers [Preprint]. arXiv. https://arxiv.org/abs/2512.01868