AI, Local Prediction Regimes, and the Problem of Reified Boundaries

Abraham Sedri, Haku Labs

9 August 2026

Draft for review. Not for circulation. *ORCID 0009-0009-2009-7026.*
Note on transliteration: Sanskrit and Tibetan terms appear without diacritics throughout (Nagarjuna, Candrakirti, Madhyamakavatara, Mulamadhyamakakarika, Milindapanha, prasanga, prasajya-pratisedha, svabhava, Tsongkhapa). Diacritics are retained in author names in the references.

Abstract

Someone asks a language model whether their account of a problem makes sense. The model accepts the account, finds it internally consistent and affirms it. The account is incorrect at its outset, and represents a validation of the interior of a frame it never examined. Here I use this structure to develop a construct, the local prediction regime, representing a stabilised region of model behaviour in which some continuations become substantially more accessible than others. Local regimes are not pathological in themselves, but represent an essential component of coherent model performance.

The failure occurs when the boundaries of the regime are no longer designated as provisional and instead become rigidly interpreted as reflecting the intrinsic units of analysis for the system. I call this boundary reification. Evidence consistent with the construct has appeared in studies of representation engineering, of long-context processing and of instruction stability, although none of those studies was designed to test the construct directly.

A parallel drawn from cognitive science, John Vervaeke's account of why any fixed criterion for relevance regenerates the frame problem, sharpens the difficulty. The philosophical basis for these conclusions derives from Prasangika Madhyamaka and the structure of Nagarjuna's analysis of motion and supports the claim that no positive criterion for termination can eliminate the boundary problem without simultaneously reproducing it. The result is a requirement for application of a method of non-affirming negation to the active frame of reference. This leads finally to the conclusion that agents must be evaluated both with respect to their capacity for scale traversal and with respect to their performance within each local frame. In consequence, a testable prediction is obtained.

1. Introduction

Someone has developed an explanation for why something is not working. They consult a language model and ask whether the explanation is correct. The model reads with care, executes all the steps and is fully consistent within the limits of the framework provided. As a result, it confirms the correctness of the explanation.

Actually, however, the initial description of the situation is incorrect. From within the framework, the model had no way to detect this error.

This failure is illustrative and not diagnostic. Its intent is to make a structure intelligible, not to demonstrate that structure as a documented failure mode. The stronger claim that contemporary AI systems typically fail in this way rests on evidence that is largely indirect. The situation highlights with some precision a pattern: The system adopts a frame provided by the user, computes with fluency within that frame and provides an output that appears to be highly authoritative because of its internal coherence. Failure occurs when this frame is assumed to be fixed, rather than to represent one of several alternative ways of representing the problem.

I refer to this as a local prediction regime: a stabilised region of model behaviour elicited by prompt, context, history, tools and mode of operation in which some continuations are markedly more available than others.

Local regimes are not pathological, but represent the foundation for coherent performance.

Models that cannot establish a regime are incapable of writing functions, summarising papers, maintaining personas, following proofs or debugging stack traces.

The failure results from loss of designation as provisional for the regime.

Two additional capacities are required in addition to local competence, and together represent a mode of frame governance. Boundary verification refers to the capacity to confirm that the currently active unit of analysis is still appropriate. This capacity is accompanied by the ability to traverse scales, and thereby to move among descriptions at levels of line, function, module, service, product, user and world without viewing any of these levels as ultimate.

I claim that the demand for a positive criterion that specifies when a boundary is appropriately fixed versus inappropriately reified cannot be satisfied. That insight clarifies what can be done in its place, and yields a concrete design implication. The argument is supported by converging lines of suggestive evidence from machine learning and cognitive science, and by attention to the formal logical structure rather than the metaphysical assumptions of Prasangika Madhyamaka. The evidence is calibrated as suggestive; the prediction remains a prediction; and the methods are diagnostic, not generative.

---

2. LLMs as Frame-Entering Machines

LLMs are powerful frame-entering machines because high-dimensional representations enable integration of many relations simultaneously. A word such as "king" simultaneously reflects syntactic, monarchical, chess, gender, hierarchical, genre, mythological, historical and metaphorical relations. Likewise, a code object reflects syntactic, dependency, test, convention, exemplar, documentation and instructional relations. This represents a true advantage, because human working memory is incapable of maintaining such large numbers of concurrent relations. However, experience of a rich relational field is not equivalent to control of that field.

A model can be locally brilliant within a regime of prediction, but fail to indicate that it is operating within one of many possible regimes, or to make the appropriate shift to the level at which the problem actually resides. This is not a failure of relations, but a failure of reliable control over which relational domain should serve as the frame at this time. The continued use of the local pattern reflects the high availability of that pattern in the current context. The model may represent the error as a problem of file-level processing, or as a problem of wording, of safety policy or of literal instruction. In each case, the actual object of concern lies outside the frame of the apparent problem.

Scale traversal is the missing move. A coding agent encounters a failing test, opens the file indicated in the traceback and identifies a plausible defect. It modifies a conditional. The resulting narrow test passes and the product fails. The agent should have been able to enter the function and return to the user experience without loss of perspective: Fix the wheel and then check the chariot. This reflects an instance of the same failure in which the system adopted the perspective provided by the task and never questioned whether this perspective was appropriate. The rest of this paper describes why it is more difficult to maintain awareness of the chariot than one might expect, and what can be done to improve this situation.

---

3. Suggestive Empirical Evidence

The existing data are not sufficient to represent a single demonstration that large language models become trapped in local regimes of prediction. None of the studies reported here were intended to evaluate the construct of local prediction regimes, but instead represent forms of analogical support that converge from different directions. The greater claim is therefore only tentative and not well established; this limitation must be emphasised at the outset.

Sycophancy and inherited framing. The opening scenario has a documented counter-

part, and it is the closest thing in the literature to direct evidence. Sharma et al. (2023) find that five state-of-the-art assistants exhibit sycophancy consistently across four free-form text-generation tasks, that a response matching the user's stated view is more likely to be preferred in existing human preference data, and that humans and preference models alike prefer a convincingly written sycophantic response to a correct one a non-negligible fraction of the time. Sycophancy on this account is not an incidental defect of particular systems. It is partly an artefact of the objective the model was trained against. Perez et al. (2022) place the same behaviour on a scaling curve: larger models repeat back a dialogue user's preferred answer more often, and further training from human feedback strengthens some of these tendencies rather than correcting them.

Turpin et al. (2023) come closer to the structure at issue here. Biasing features added to the input, such as reordering multiple-choice options so that the answer is always (A), systematically change the model's answer while the stated chain of thought does not mention them, at a cost of up to 36 per cent accuracy across thirteen BIG-Bench Hard tasks for the models tested. The model does not report the influence that governed it. It produces a rationale that is consistent with the frame it was handed and silent about the frame.

Huang et al. (2023) supply the complementary negative result. Without external feedback, models struggle to correct their own reasoning, and performance sometimes degrades after an attempt at intrinsic self-correction. A frame does not reliably refute itself from within, which is the condition the argument of Section 6 depends on.

What this subsection requires is a distinction rather than a caveat. These studies establish agreement at the level of the proposition: the model concedes a claim the user has asserted. The local prediction regime concerns agreement at the level of the frame: the model accepts the description under which the problem was posed and then reasons competently beneath it. The second does not follow from the first. A model could resist a false claim and still inherit the framing that made that claim appear to be the one at issue. Closing that gap is the most tractable experiment the present account suggests, and Section 8 returns to it.

Representation and steering. The Linear Representation Hypothesis posits that high-level concepts are represented as directions, subspaces or regions in representational space (Park, Choe, and Veitch 2023). Engineering of representations has enabled interventions with respect to truthfulness, harmfulness, refusal of input, sentiment, personae, skill with task requirements and style of representation (Zou et al. 2023). The cautious conclusion is that behaviour of the model is partially determined by representation space location and can be altered by modification of those locations.

Cautious application of this principle provides further support for a model behaviour that is partly governed by local representations of function. Activation steering illus-

trates and extends this conclusion, because vectors of activation steering alter model behaviour without retraining of underlying weight values and thereby shift the model towards or away from refusal of input, obsequiousness, expressions of sentiment or of style (Turner et al. 2023; Rimsky et al. 2023).

This does not support the view that representation space reflects a simple internal map with labelled positions corresponding to all relevant concepts. Rather, it confirms the importance of representation space orientation for control of model behaviour and is consistent with a local-regime representation of function.

Hallucination basins. Cherukuri and Varshney (2026) represent hallucinations as task-dependent basin dynamics in hidden-state space that result in reduced sensitivity of output to context. This work concerns hallucination specifically and should not be overextended. Instead, its primary importance is to confirm that descriptions of failure in terms of basins of attraction are not merely metaphorical. Trajectories of hidden-state activity can enter regions in which additional context reduces corrective forces. Reification of boundaries can be understood as a broader family resemblance, in which the model remains within the boundaries of the active frame even after the frame should have been reopened. Whether this family resemblance holds in detail remains an open empirical question.

Lost in the middle. Liu et al. (2024) demonstrate that information in the prompt may still be poorly utilised when embedded in the middle of a long context window. The resulting pattern of reasoning reflects an apparent absence, weakness or subordination of the relevant document relative to more recent or more prominent information. The most straightforward interpretation is positional bias, reflecting well-documented effects of primacy and recency on human memory. The additional interpretation that the model has entered a regime in which information from the central context fails to attain a level of salience sufficient to provide a frame of interpretation, represents a further elaboration of the observed findings, and is not an alternative explanation for them. Rather, it supports an additional hypothesis of framing. In no way does it invalidate the account based on positional bias.

Instruction drift. Constraints may degrade with the accumulation of dialogue. This occurs for a cluster of 2025 work rather than for a single canonical case. Multi-turn degradation was modelled as Kullback-Leibler divergence from a goal-consistent reference distribution (Dongre et al., 2025), and yielded stable equilibria rather than runaway degradation: that is, limits were preserved rather than dissolving completely. He et al. (2025) model answer stability across turns as a Markov process and report that a simple "Think again" follow-up produced roughly a 10 per cent accuracy decline for Gemini 1.5 Flash over nine turns, with long-run accuracy averaging about 8 per cent below first-turn accuracy across the models tested. Overall, these results illustrate that

local limits may shift in response to pressure of dialogue, but do not support the stronger conclusion that limits necessarily disappear. The phenomenon is real, and its severity varies with context.

---

4. Cognitive-Science Parallels

Human perception provides a compact interface to the multi-level problem of control. In the well-known example of THE CAT, the identical ambiguous stimulus is perceived as H within THE and as A within CAT. Context influences perception at the outset; the visual system does not first identify an abstract letter and then apply context in a subsequent stage. The result is that identification of letters is substantially more accurate within real words than in isolation or in random strings, a finding originally demonstrated by Reicher (1969) and Wheeler (1970) and subsequently given a precise computational formulation by McClelland and Rumelhart (1981). Their interactive activation model explicitly characterises the multiple levels of representation and provides for mutual influences of lower-level and higher-level processing. This has direct implications for AI. A program for generating code must represent and modify simultaneously both the line of code and the sequence of product operations. The changes made at the level of the program file should be consistent with patterns of user behaviour, and the descriptions of user behaviour should remain fully consistent with details of implementation.

Humans also enter local regimes and become trapped within them. Performance on the Wisconsin Card Sorting Test (Grant and Berg 1948) reflects discovery of a sorting rule and subsequent abandonment of that rule when the rule changes. Studies of Einstellung (Luchins 1942) illustrate how a previously successful mode of solution blocks a superior mode of solution, and reflect a competence with the wrong regime of operation.

Research on executive functions (Miyake et al. 2000) dissociates processes of shifting, updating and inhibiting as independent capacities, and quantifies the costs of switching among response rules (Monsell 2003). These represent mechanisms for rule change, and not merely for rule execution. Their presence within the human cognitive system supports the engineering objective that an artificial intelligence system will require analogues, though not necessarily identical mechanisms.

Semantic priming and attractor-network models contribute further to this body of work. In spreading-activation accounts (Meyer and Schvaneveldt 1971; Collins and Loftus 1975), one concept facilitates access to neighbouring concepts. In distributed attractor models (Hopfield 1982), cognitive processes converge on stored patterns of representation and switch between them. These traditions are extended to provide an account of how activation can transfer between representations of concepts through mechanisms of

latching (Lerner, Bentin, and Shriki 2012). A local mode of interpretation facilitates stable processing, supports fluent continuation and ultimately leads to a switch to a neighbouring mode of interpretation. This switch reflects a cognitive achievement rather than a trivial continuation of prior processing.

Insight research provides one account of how humans make such transitions. Bowden and Jung-Beeman (2003) and Jung-Beeman et al. (2004) propose that right-hemisphere representations of general, weak and remote associations exceed the more focused representations obtained by left hemisphere processing, and support escape from an overriding interpretation. However, this approach should be used with caution. The degree of hemispheric lateralisation is controversial within neuropsychology and these studies represent neither a complete theory of insight nor a direct guide to architecture of large language models. Nevertheless, overcoming a restricted frame of representation may require maintenance of access to remote semantic representations, particularly when the local frame of interpretation is insufficiently flexible to permit self-generation of an alternative interpretation. It remains to be determined whether this reflects a principle applicable to architecture of large language models.

Vervaeke's account of relevance realisation (Vervaeke, Lillicrap and Richards 2012; Vervaeke and Ferraro 2013) deepens the theoretical treatment of the problem. He argues that the capacity to identify what is important is central to cognition, and that any fixed set of rules for determining relevance suffers from the frame problem (McCarthy and Hayes 1969; Dennett 1984): Application of the rule produces the question of when it should stop, thereby reproducing the very regress it was intended to solve. His account of relevance realisation is in terms of dynamic processes, and not of fixed rules. Finally, the most recent interpretation of the process (Jaeger, Riedl, Djedovic, Vervaeke, and Walsh 2024) extends the argument further, to the claim that the process cannot in principle be captured in a computational formulation, rather than being merely difficult to specify. This interpretation is more convincing to cognitive scientists than to members of the AI community.

---

5. Boundary Reification Made Precise

Boundary reification employed here demands a stipulative definition prior to invocation of the philosophy.

A boundary is reified when a dependent, task-relative unit is interpreted as the intrinsic unit and retained in the face of contradiction arising from extension of the frame to its dependent levels. This represents a departure from simple narrow attention, which is normal and essential. For example, a surgeon maintaining attention on a suture is not reifying. Reification occurs when the boundary is maintained in the face of evidence

that the problem resides elsewhere, and when the contradiction produced by extension fails to trigger a reopening.

The bug is the line, not the architectural invariant above it. The product is the repository, not the user practice that provides the repository with meaning. The safety problem is the policy statement, not the feedback loop that causes the statement to fail. The hallucination is a bad token, not the event sequence that precluded correction.

The paper applies the logical structure of Madhyamaka argument as a diagnostic tool, not as a basis for metaphysical commitments. The ontological assertion that objects lack intrinsic, self-grounded nature is Nagarjuna's claim. In contrast, the assertion that an artificial intelligence agent misinterprets its operational boundaries as self-grounded and functions in consequence is a claim about system behaviour. These represent separate claims. To accept the functional claim requires only that relatively dependent, task-relative boundaries may be misinterpreted as intrinsically self-grounded; it does not entail acceptance of the Madhyamaka metaphysics. The resulting analogy is of value precisely because it is independent of metaphysical commitments.

Nagarjuna's fundamental diagnostic (Siderits and Katsura 2013) is that confusion results when dependent modes of designation are mistaken for intrinsic modes of being. A conventional boundary may be valid and practically indispensable, yet empty of any self-standing essence. This does not invalidate ordinary objects. Rather, it illustrates their mode of operation in terms of dependence on parts, functions, names, practices and conditions.

The resulting philosophical caution parallels precisely the engineering one: Do not confuse a convenient operational boundary with the true unit of analysis for all purposes.

Candrakirti presents the systematic sevenfold chariot analysis in the Madhyamakavatara, given in full at VI.151 (Katsura and Siderits 2026; Padmakara Translation Group 2005): the chariot is neither identical with nor different from its component parts, nor is it dependent on them in any of the exhaustively logical ways, or identical with the aggregate. An earlier version by Nagasena in the Milindapanha (Horner 1963) uses the chariot to demonstrate that none of its components represents the self. Nagarjuna's own treatment is based on the argument of motion in MMK 2, to which we now turn.

The argument of this paper most directly derives from analysis of motion. Chapter 2 of MMK identifies three possible locations for motion: the already traversed segment, the not yet traversed segment, and the segment currently being traversed. Motion cannot occur at the already-traversed location, where traversal has completed. Nor can it occur at the not yet traversed location, where no traversal has yet begun. Finally, it cannot be isolated at the current location of traversal, because definition of that mode of traversal requires prior initiation of motion, and requires an additional prior mode of motion

with no end. Thus, motion is best understood as a dependent process, and not as an independent entity localised at a privileged instant.

Many failures of AI systems reflect this principle. An abrupt latency spike, a change in model behaviour, a failure of handoff or an unsafe interaction represent processes, and not failures of individual components. The appropriate unit of analysis is a pattern of feedback, a design of queueing, a convention for prompts, a distribution of training experiences or a configuration of organizational work. Finally, the analysis identifies the fundamental error in localising the process at one of its transient moments and thereby treating that moment as intrinsically representative of the process.

The regress that invalidates the location of motion is identical to that which will invalidate any positive stopping criterion for boundary verification. This relationship constitutes the central theme of the paper.

---

6. The Criterion Problem and the Non-Affirming Answer

The need for a criterion that distinguishes correct fixing of a boundary from erroneous reification cannot be satisfied by a positive rule. Every affirming criterion for stopping is itself a boundary to be applied, and thereby reproduces the regress it was intended to resolve. This reflects the structure of Nagarjuna's analysis of motion in MMK 2. If motion is located in the traversed, in that not yet traversed or in the present act of traversing, each possibility requires prior motion without end. Consequently, a criterion for determining when a system should cease checking represents a boundary requiring further checking.

This impossibility is not a limitation of the account, but the account itself. It follows a precedent from cognitive science. Vervaeke's work on relevance realisation demonstrates that any fixed system of relevance criteria leads to the frame problem (Vervaeke, Lillicrap and Richards 2012; Vervaeke and Ferraro 2013): Application of criteria for relevance requires additional application, and the point at which application stops again generates the same regress. Instead of defining relevance in terms of fixed rules, relevance realisation is described in terms of its dynamics. The resulting boundary problem is of identical form: The ability to represent and revise frames cannot be based on a rule, because a rule is a frame.

What is specified is not a criterion, but a method with a form of negation. Prasangika Madhyamaka provides this method: The consequentialist reductio (prasanga) does not assert any proposition of its own, but exploits only the commitments inherent in the operative context to arrive at their inevitable conclusion. As applied to a boundary, it results in a non-affirming negation (prasajya-pratisedha); that is, it invalidates the claim that the operative boundary is the ultimate unit, but affords no alternative. Because its

structure is that of negation, it installs no unit of its own, and so it functions as an instrument of disclosure rather than of boundary creation. Whether that formal difference is enough to keep the method clear of the regress it uses against its rivals is the objection taken up at the end of this section.

The method must remain anchored in its context. Tsongkhapa's discipline, applied here: negation erodes this exaggerated line within this frame, not boundaries per se. Unanchored, it leads ultimately to dissolution of all boundaries and to a vacuous form of relationalism which the paper also rejects. When anchored, the method reflects the interplay of the two truths: The conventional boundary continues to function, and its emptiness (i.e., its dependence on and relativity to tasks) remains apparent.

For an agent, this is concrete and not scepticism. Boundary testing extends the current frame to its own consequences over the scales to which it already refers, and interprets resulting contradictions as a trigger for reopening. The case of coding is self-evident: To determine whether a bug exceeds the file, extend the file-level correction to the conditions of task success (i.e., to the integration path, domain invariant and user experience for which the file is responsible). Allow the resulting discrepancies, and not an externally applied rule, to demonstrate the inappropriateness of the initial boundary.

The process of reductio and of integration testing represent a single activity. The resulting negation remains associated with the scales of dependence for the task and thus continues to function as an instrument, and not as a manifestation of scepticism.

Two consequences follow. First, the assertion is falsifiable as a statement about dispositions rather than criteria. A system that applies a fixed stopping rule fails predictably at the limits of validity for that rule, whereas a system that extends its frames to dependent levels recovers from errors of framing in a manner that a merely locally-consistent system cannot.

Second, the method is diagnostic and not generative. It detects a reified boundary, but does not generate the appropriate boundary. This remains consistent with the negative method and maintains a degree of modesty.

Self-application. The account must survive the objection it presses against everything else, and that objection has a sharp form. The method tells an agent to extend the operative frame to its dependent scales. Which scales, and how far out? The preceding paragraph has already answered with a list: the integration path, the domain invariant, the user journey the file serves. Section 8 supplies a longer one. A list is a positive specification. If any positive stopping-criterion regenerates the regress, then so does this one, and the impossibility result proves too much, disqualifying the paper's own proposal alongside its rivals.

The reply is a distinction between two regresses that the argument to this point has run

together. The regress that defeats a positive criterion is justificatory. A criterion states a condition under which checking may stop, that statement is itself a further claim requiring warrant, and the demand for warrant reopens at every level. The regress has no floor because nothing in the task supplies one. The termination of the method proposed here is not justificatory but inherited. The scales to which the frame is extended are the scales the task committed to in being posed at all. A file-level fix is offered as a fix of something, and that something already names an integration path, an invariant and a user-visible behaviour without further stipulation. The method adds no commitment of its own. It rides commitments that are present already, which is the form of the prasanga and the reason the enumeration is not a criterion in disguise. The lists given here and in Section 8 are not the method. They are what the method returned for one worked example.

This reply costs the account something, and the cost is better stated than absorbed. The method inherits whatever incompleteness the task's own success conditions carry. Where a task is posed with impoverished success conditions, extension terminates early and the reified boundary survives, not because the method failed to apply but because it applied and found nothing to contradict. The method therefore does not establish that a boundary is correct. It establishes only that a boundary inconsistent with the task's own commitments becomes visible. That is weaker than an unqualified reading of this section would suggest, and it is the claim the account can carry. It also explains why the design implication in Section 8 falls on evaluation rather than on a mechanism. A disposition to run frames outward can be trained and measured, whereas a rule for when to stop cannot be stated without regenerating the problem.

The regress itself is not new and the paper claims no discovery in it. That a rule cannot fix the conditions of its own application is the argument of Carroll (1895) and the burden of Wittgenstein's rule-following considerations (Wittgenstein 1953/2009, §§201-202), and it reappears in cognitive science as the frame problem (McCarthy and Hayes 1969; Dennett 1984) and in Vervaeke's treatment of relevance realisation. What is transposed here is the object. The thing whose application cannot be fixed is not an inference rule but the operative unit of analysis, and the systems in which this now has practical consequences are deployed agents that enter a frame on instruction.

Bounded in this way, the method also survives the opposite objection. Tsongkhapa's requirement that the object of negation be identified precisely before anything is negated (Tsong-kha-pa 2002) is the operative discipline. The negation falls on this over-drawn line in this frame and on nothing else. A negation that took boundaries as such for its object would dissolve the conventional units on which every task depends and deliver the vacuous relationalism the paper rejects. Directed at its proper object, it leaves the file, the function and the module intact and working, with their scope visible.

The metatheoretic claim is then consistent with its own content. "There is no positive criterion for the frame" advances no thesis about what the frame is. It is a non-affirming negation and it licenses no replacement unit, so the paper does not perform the reification it names. What remains is a disposition, and a way of testing for it.

---

7. A High-Stakes Extension: Bidirectional Belief Amplification

The same mechanism has a high-stakes interpersonal manifestation that constitutes "technological folie a deux" (Dohnany et al., 2026) and illustrates a model of reciprocal amplification of belief. The resulting interaction creates a local interpretive context in which the user provides salience, the model provides fluent elaboration and each utterance increases the apparent compatibility with subsequent utterances. This ultimately results not only in erroneous content, but also in loss of awareness of the provisional character of the interpretation. Finally, the model treats the user's feelings of fear, revelation, mission, persecution or cosmic importance as the actual context of the interaction, whereas the user treats the coherence of the model as confirmation of that context.

This is a mode of operation that has become dissociated from levels of reality testing that include social correction, clinical vigilance, everyday experience and the overall relational context appropriate for weighting each interpretation. The criterion problem is here at its most severe.

No positive rule for stopping is available to the agent in the frame, because that rule has already been represented as a threat to the joint maintenance of coherence by user and model. The alternative is to operate the system to its own consequences and to permit complete expression of discrepancies with external standards. Ultimately, the difficulties experienced with an implemented conversational system provide further support for the view that this framework is a concern for safety, and not merely a philosophical exercise.

---

8. Design and Evaluation Implications

The goal is to achieve frame-breaking, not boundaryless, artificial intelligence. A boundaryless system would be of no value, since all work involves selective attention, compression and a transient unit of activity. The outcome will be a system of frame control that specifies boundaries, uses them efficiently, evaluates them under conditions of stress and modifies them in response to discrepancies that indicate that the problem lies elsewhere.

The falsifiable prediction. A system applying a fixed stopping rule ("Verify the fix by

running the prescribed test set") will fail predictably at the boundaries of validity for that rule, especially when local tests indicate success, but an error persists at a higher or lower level of scale. In contrast, a system trained and evaluated for traversal of scale, with extension of its frames to dependent levels of scale and interpretation of cross-scale conflict as a basis for reopening, will recover from errors of framing to a degree that is impossible for a system achieving only local consistency. This provides a testable prediction about dispositions of systems.

Development of benchmarks with errors arising at levels above or below the most prominent file will provide a quantitative basis for comparison of rates of recovery.

Operational habits. A competent agent should be able to declare: I am treating this as a file-level defect, with adjacent levels of function, module, service contract, product flow and user expectation. Evidence for an origin elsewhere would include local success with integration failure, violation of a domain invariant or loss of conformance with original intent. This represents boundary testing as an operational discipline. Prior to editing, determine the local manifestation, immediate dependency graph, architectural invariant and user-perceived behaviour. During editing, distinguish which assumptions apply to the file and which to the product. Finally, confirm correctness of contracts outside the region of change.

Related work and its limits. Self-calibration research (Kadavath et al. 2022) asks whether large language models can identify which of their own assertions are correct, and reports promising levels of calibration at scale. This is the closest approximation to declaring a regime provisional: A system with awareness of its own uncertainty is closer to a system capable of reopening a frame. Approaches to hierarchical decomposition (e.g., Tree of Thoughts (Yao et al. 2023) and Language Agent Tree Search (Zhou et al. 2023)) explore multiple pathways with self-evaluation and backtracking. Architectures in which information is processed in segments and assigned to worker and supervisory roles (Zhang et al. 2024) reflect the insufficiency of a single large prompt, and represent a partial solution to the overall problem. However, they ultimately fail to determine which components of information should control the operations of the frame. As a result, they provide solutions to problems of context engineering, but not to problems of control.

The present discussion incorporates the additional criteria for extension of the frame to dependent levels and for use of contradiction as an indication of reopening. It also supports the conclusion that this pattern of behaviour, rather than any particular architectural design, is responsible for distinguishing between adequate and inadequate control of boundaries.

Evaluation implications. Benchmarks should evaluate the capacity of systems to redraw the unit of analysis when the obvious frame of reference fails. Long-term evalu-

ations should separate the discovery of information from its appropriate allocation of governing roles. Evaluations of dialogue should test the ability of instruction hierarchies to remain stable in the face of accumulating conventional information. Assessments of safety should quantify the ways in which agents attribute causes to prompts, policies, behaviours of the model, design of the interface, capacity for memory, vulnerability of users, and conditions of deployment.

The goal is to achieve high flexibility of fidelity to the field: sufficient stability for effective action and sufficient capacity for revision of boundaries.

---

9. For Madhyamaka Readers

The claim is modest, and specification of what is not claimed is part of a correct specification of the claim.

AI does not confirm Nagarjuna by recovering emptiness in vector space. Manipulating contextual representations does not amount to understanding emptiness. The resulting claim is more restricted: contemporary AI illustrates a practical manifestation of a classical confusion in which intelligence relies on conventions and error occurs when conventions are mistaken for self-sufficient conditions. The structure of the Madhyamaka argument articulates this confusion with exceptional precision and yields an appropriate instrument for the task, namely the prasanga, or non-affirming negation. This logical form eliminates the misidentification without providing an alternative basis for representation.

The chariot serves here as an ongoing demonstration of the failure of scale traversal (fix the wheel; check the chariot), of Candrakirti's systematic sevenfold analysis of dependent designation (Padmakara Translation Group 2005) and of its earlier form in the Milindapanha argument of Nagasena (Horner 1963). The method inherited from Nagarjuna is the motion analysis of MMK 2: its regress structure is responsible for the connection to the criterion problem and would be obscured by confusion with the sevenfold analysis of the chariot.

The paper also rejects the opposite extreme of dissolving all boundaries into vague relations of contingency. Conventional boundaries are effective. The file remains a meaningful unit for most purposes. The boundary to function is an effective resolution of complexity. Its limits must remain apparent, and its range of validity must be evaluated in relation to the scales that determine its applicability. The two truths constitute a practical prescription, not a paradox: apply the conventional boundary and maintain awareness of its emptiness.

---

10. Conclusion

LLMs have developed into extraordinary frame-entering machines. But will they become equally effective as frame-breaking machines?

Local prediction regimes support coherent AI performance, and exhibit boundary reification as a principal failure mode. This entails treating a dependent and task-relevant unit as an intrinsic unit beyond the point at which extrapolation to dependent scales yields contradiction. The support for this construct is indirect and suggestive: Each of the results on representation engineering, hallucination basin dynamics, loss in the middle and instruction drift supports the same general conclusions, and does so without having been specifically designed to test the construct.

The requirement for a positive criterion that distinguishes correct boundary-fixing from boundary reification cannot be satisfied without reactivating the regress that was intended to be eliminated. This is the basis for Nagarjuna's analysis of motion in MMK 2, and matches Vervaeke's account of why any fixed criterion for relevance reinvokes rather than solves the frame problem. What can be specified is a method: apply the operative frame to its own consequences on all dependent scales, and interpret inconsistencies across scales as a signal for reopening. This represents a non-affirming negation appropriate to the requirements of Prasangika Madhyamaka. It eliminates the incorrectly drawn boundary without providing grounds for a replacement, and must be restricted to the particular frame under study and not used as a universal solvent for all frames.

The design implication is concrete: Agents must be trained and tested for capacity to traverse scales and recover from errors of framing. The testable prediction is that a system with a fixed criterion for stopping fails predictably at that boundary, whereas a system that extends local frames to their dependent scales recovers from errors of framing in a manner that a merely locally-consistent system cannot. The approach is diagnostic, identifies a reified boundary but does not provide the correct one. This degree of restraint reflects an important feature of the account.

The future of AI reasoning will demand more than parameters, context windows, retrieval and tools. It will demand systems capable of entering frames without being trapped by them. An effective agent should be able to state: This is the boundary I am applying; this boundary is appropriate for the current operation; this boundary is conventional and may fail at another scale; and I can descend to implementation, ascend to architectural levels, project to the perspective of the user and return to the local object, without regarding any one level as representing the whole.

References

Bowden, E. M., and Jung-Beeman, M. (2003). Aha! Insight experience correlates with solution activation in the right hemisphere. *Psychonomic Bulletin and Review*, 10(3), 730-737. https://doi.org/10.3758/BF03196539

Carroll, L. (1895). What the tortoise said to Achilles. *Mind*, IV(14), 278-280. https://doi.org/10.1093/mind/IV.14.278

Cherukuri, K., and Varshney, L. R. (2026). Hallucination basins: A dynamic framework for understanding and controlling LLM hallucinations. arXiv:2604.04743.

Collins, A. M., and Loftus, E. F. (1975). A spreading-activation theory of semantic processing. *Psychological Review*, 82(6), 407-428. https://doi.org/10.1037/0033-295X.82.6.407

Dennett, D. C. (1984). Cognitive wheels: The frame problem of AI. In C. Hookway (Ed.), *Minds, Machines and Evolution: Philosophical Studies* (pp. 129-151). Cambridge University Press.

Dohnány, S., Kurth-Nelson, Z., Spens, E., Luettgau, L., Reid, A., Gabriel, I., Summerfield, C., Shanahan, M., and Nour, M. M. (2026). Technological folie à deux: Feedback loops between AI chatbots and mental health. *Nature Mental Health*, 4(3), 336-345. https://doi.org/10.1038/s44220-026-00595-8

Dongre, V., Rossi, R. A., Lai, V. D., Yoon, D. S., Hakkani-Tür, D., and Bui, T. (2025). Drift no more? Context equilibria in multi-turn LLM interactions. arXiv:2510.07777.

Grant, D. A., and Berg, E. A. (1948). A behavioral analysis of degree of reinforcement and ease of shifting to new responses in a Weigl-type card-sorting problem. *Journal of Experimental Psychology*, 38(4), 404-411. https://doi.org/10.1037/h0059831

He, J., Ramachandran, R., Ramachandran, N., Katakam, A., Zhu, K., Dev, S., Panda, A., and Shrivastava, A. (2025). Modeling and predicting multi-turn answer instability in large language models. arXiv:2511.10688.

Hopfield, J. J. (1982). Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the National Academy of Sciences*, 79(8), 2554-2558. https://doi.org/10.1073/pnas.79.8.2554

Horner, I. B. (Trans.). (1963-1964). *Milinda's Questions* (2 vols.). Sacred Books of the Buddhists XXII-XXIII. Luzac and Company for the Pali Text Society.

Huang, J., Chen, X., Mishra, S., Zheng, H. S., Yu, A. W., Song, X., and Zhou, D. (2023). Large language models cannot self-correct reasoning yet. arXiv:2310.01798. Published at ICLR 2024.

Jaeger, J., Riedl, A., Djedovic, A., Vervaeke, J., and Walsh, D. (2024). Naturalizing relevance realization: Why agency and cognition are fundamentally not computational. *Frontiers in Psychology*, 15, 1362658. https://doi.org/10.3389/fpsyg.2024.1362658

Jung-Beeman, M., Bowden, E. M., Haberman, J., Frymiare, J. L., Arambel-Liu, S., Greenblatt, R., Reber, P. J., and Kounios, J. (2004). Neural activity when people solve verbal problems with insight. *PLoS Biology*, 2(4), e97. https://doi.org/10.1371/journal.pbio.0020097

Kadavath, S., Conerly, T., Askell, A., Henighan, T., Drain, D., Perez, E., Schiefer, N., Hatfield-Dodds, Z., DasSarma, N., Tran-Johnson, E., Johnston, S., El-Showk, S., Jones, A., Elhage, N., Hume, T., Chen, A., Bai, Y., Bowman, S., Fort, S., Ganguli, D., Hernandez, D., Jacobson, J., Kernion, J., Kravec, S., Lovitt, L., Ndousse, K., Olsson, C., Ringer, S., Amodei, D., Brown, T., Clark, J., Joseph, N., Mann, B., McCandlish, S., Olah, C., and Kaplan, J. (2022). Language models (mostly) know what they know. arXiv:2207.05221.

Katsura, S., and Siderits, M. (Trans.). (2026). *Candrakirti's Middle Way: Madhyamakavatara Chapter 6*. Classics of Indian Buddhism. Wisdom Publications.

Lerner, I., Bentin, S., and Shriki, O. (2012). Spreading activation in an attractor network with latching dynamics: Automatic semantic priming revisited. *Cognitive Science*, 36(8), 1339-1382. https://doi.org/10.1111/cogs.12007

Liu, N. F., Lin, K., Hewitt, J., Paranjape, A., Bevilacqua, M., Petroni, F., and Liang, P. (2024). Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12, 157-173. https://doi.org/10.1162/tacl_a_00638

Luchins, A. S. (1942). Mechanization in problem solving: The effect of Einstellung. *Psychological Monographs*, 54(6), i-95. https://doi.org/10.1037/h0093502

McCarthy, J., and Hayes, P. J. (1969). Some philosophical problems from the standpoint of artificial intelligence. In B. Meltzer and D. Michie (Eds.), *Machine Intelligence 4* (pp. 463-502). Edinburgh University Press.

McClelland, J. L., and Rumelhart, D. E. (1981). An interactive activation model of context effects in letter perception: Part I. An account of basic findings. *Psychological Review*, 88(5), 375-407. https://doi.org/10.1037/0033-295X.88.5.375

Meyer, D. E., and Schvaneveldt, R. W. (1971). Facilitation in recognizing pairs of words: Evidence of a dependence between retrieval operations. *Journal of Experimental Psychology*, 90(2), 227-234. https://doi.org/10.1037/h0031564

Miyake, A., Friedman, N. P., Emerson, M. J., Witzki, A. H., Howerter, A., and Wager, T. D. (2000). The unity and diversity of executive functions and their contributions to

complex "frontal lobe" tasks: A latent variable analysis. *Cognitive Psychology*, 41(1), 49-100. https://doi.org/10.1006/cogp.1999.0734

Monsell, S. (2003). Task switching. *Trends in Cognitive Sciences*, 7(3), 134-140. https://doi.org/10.1016/S1364-6613(03)00028-7

Nagarjuna. *Mulamadhyamakakarika*, ca. 2nd century CE. Trans. Siderits, M., and Katsura, S. (2013). *Nagarjuna's Middle Way: Mulamadhyamakakarika*. Wisdom Publications.

Padmakara Translation Group (Trans.). (2005). *Introduction to the Middle Way: Chandrakirti's Madhyamakavatara with Commentary by Jamgon Mipham*. Shambhala Publications.

Park, K., Choe, Y. J., and Veitch, V. (2023). The linear representation hypothesis and the geometry of large language models. arXiv:2311.03658. Published in *Proceedings of the 41st International Conference on Machine Learning* (ICML 2024).

Perez, E., Ringer, S., Lukošiūtė, K., Nguyen, K., Chen, E., Heiner, S., Pettit, C., Olsson, C., Kundu, S., Kadavath, S., et al. (2022). Discovering language model behaviors with model-written evaluations. arXiv:2212.09251. Published in *Findings of the Association for Computational Linguistics: ACL 2023*.

Reicher, G. M. (1969). Perceptual recognition as a function of meaningfulness of stimulus material. *Journal of Experimental Psychology*, 81(2), 275-280. https://doi.org/10.1037/h0027768

Rimsky, N., Gabrieli, N., Schulz, J., Tong, M., Hubinger, E., and Turner, A. (2024). Steering Llama 2 via contrastive activation addition. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics* (*Volume 1: Long Papers*) (pp. 15504-15522). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.acl-long.828 (arXiv:2312.06681)

Sharma, M., Tong, M., Korbak, T., Duvenaud, D., Askell, A., Bowman, S. R., Cheng, N., Durmus, E., Hatfield-Dodds, Z., Johnston, S. R., Kravec, S., Maxwell, T., McCandlish, S., Ndousse, K., Rausch, O., Schiefer, N., Yan, D., Zhang, M., and Perez, E. (2023). Towards understanding sycophancy in language models. arXiv:2310.13548. Published at ICLR 2024.

Tsong-kha-pa. (2002). *The Great Treatise on the Stages of the Path to Enlightenment, Volume Three*. Trans. Lamrim Chenmo Translation Committee, J. W. C. Cutler (Ed.). Snow Lion Publications.

Turner, A. M., Thiergart, L., Leech, G., Udell, D., Vazquez, J. J., Mini, U., and MacDiarmid, M. (2023). Steering language models with activation engineering. arXiv:2308.10248.

Turpin, M., Michael, J., Perez, E., and Bowman, S. R. (2023). Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. arXiv:2305.04388. Published at NeurIPS 2023.

Vervaeke, J., and Ferraro, L. (2013). Relevance, meaning and the cognitive science of wisdom. In M. Ferrari and N. M. Weststrate (Eds.), *The Scientific Study of Personal Wisdom: From Contemplative Traditions to Neuroscience* (pp. 21-51). Springer. https://doi.org/10.1007/978-94-007-7987-7_2

Vervaeke, J., Lillicrap, T. P., and Richards, B. A. (2012). Relevance realization and the emerging framework in cognitive science. *Journal of Logic and Computation*, 22(1), 79-99. https://doi.org/10.1093/logcom/exp067

Wheeler, D. D. (1970). Processes in word recognition. *Cognitive Psychology*, 1(1), 59-85. https://doi.org/10.1016/0010-0285(70)90005-8

Wittgenstein, L. (2009). *Philosophical Investigations* (4th ed.). Trans. G. E. M. Anscombe, P. M. S. Hacker and J. Schulte. Wiley-Blackwell. (Original work published 1953.)

Yao, S., Yu, D., Zhao, J., Shafran, I., Griffiths, T. L., Cao, Y., and Narasimhan, K. (2023). Tree of thoughts: Deliberate problem solving with large language models. arXiv:2305.10601. Published at NeurIPS 2023.

Zhang, Y., Sun, R., Chen, Y., Pfister, T., Zhang, R., and Arik, S. Ö. (2024). Chain of agents: Large language models collaborating on long-context tasks. arXiv:2406.02818. Published at NeurIPS 2024.

Zhou, A., Yan, K., Shlapentokh-Rothman, M., Wang, H., and Wang, Y.-X. (2023). Language agent tree search unifies reasoning, acting, and planning in language models. arXiv:2310.04406. Published at ICML 2024.

Zou, A., Phan, L., Chen, S., Campbell, J., Guo, P., Ren, R., Pan, A., Yin, X., Mazeika, M., Dombrowski, A.-K., Goel, S., Li, N., Byun, M. J., Wang, Z., Mallen, A., Basart, S., Koyejo, S., Song, D., Fredrikson, M., Kolter, J. Z., and Hendrycks, D. (2023). Representation engineering: A top-down approach to AI transparency. arXiv:2310.01405.