The Generative Horizon: Applied Hermeneutics, Linguistic Attractors, and the Limits of Model Self-Report

Rolando Bosch
Hermes Labs — San Francisco, California, USA
roli@hermes-labs.ai

July 30, 2026

Abstract

For language-model agents, language is simultaneously an object of interpretation and a medium of generation. The represented situation formed through instructions, prior turns, retrieved records, tool results, summaries, and corrections is therefore not merely what a model computes about; it is part of what the model computes through. This paper calls that situated condition the **generative horizon**. When an output is retained as context, memory, policy, evidence, or authorization, one generation can alter the conditions of later interpretation and action. This is **recursive interpretive conditioning**: the output of one horizon becomes part of another.

Hermeneutics supplies a vocabulary for historically formed and revisable understanding; applied hermeneutics translates that problem into requirements for provenance, status, revision, and authority. The argument neither attributes human historical consciousness to language models nor reduces computation to language. It distinguishes the external task, computational substrate, and historically formed organization through which representations become action-guiding.

Gurnee et al.'s 2026 J-space study [1] provides a mechanistic stress test. It supports sparse, token-associated representations that causally participate in report, reasoning, and control, but not an inner observer, privileged point of view, or phenomenal awareness. Its counterfactual-reflection result instead asks how training or retaining a possible report reorganizes later computation.

Systems using summaries, self-reports, decoded features, or evaluation labels should distinguish measurement from intervention, preserve correction and supersession, and bind a representation's authority to evidence proportionate to what it may govern. Mechanistic interpretability does not remove interpretation; it makes the problem experimentally and operationally precise.

Keywords: large language models; applied hermeneutics; mechanistic interpretability; model self-report; epistemic engineering; generative horizon; recursive interpretive conditioning; linguistic attractors; J-space

Author's Note on Intellectual Provenance and AI-Mediated Authorship

The argument developed in this paper is the current culmination of more than a year of ongoing work on epistemic and phenomenological approaches to understanding AI systems. Earlier forms of this inquiry appeared publicly in improvised video essays on my YouTube channel, Roli From Hermes, where I explored questions concerning language, interpretation, model self-description, and the conditions under which representations acquire epistemic or operational authority. Those recordings preserve the exploratory beginnings and chronology of the work. They also show that the questions and conceptual connections developed here precede the AI-assisted composition of this preprint.

This paper initially began as a short conceptual companion to *Precise Records, Unstable Meanings: Measurement Validity and Unsupported Claims Derived from AI Agent Telemetry*. Its intended purpose was to examine more closely the applied-hermeneutic questions raised by that naturalistic telemetry study. Through the iterative, maieutic research process discussed in that paper and described more fully in its Author's Note—questioning, conceptual pressure, counterargument, evidence checking, expansion, refinement, and revision—the proposed side piece became a separate argument. The resulting paper is not simply a practical note about applying hermeneutics to epistemic engineering. It develops a broader account of how hermeneutic and phenomenological frameworks can clarify the interpretation, self-report, and operational authority of representations in AI systems. The empirical study and this conceptual paper therefore address related but distinct evidentiary questions; neither is presented as validating the other.

Except for the initial version of this note, the prose of this preprint was generated and extensively revised through AI systems under my direction. I wrote this note and then used AI to help clarify it. I am listed as the sole author not because I manually composed every sentence, but because authorship here denotes accountable intellectual judgment rather than keystroke production. I established the inquiry, supplied and developed its central ideas, directed the argumentative process, determined what claims the available evidence could support, resolved interpretive disagreements, accepted or rejected revisions, and take responsibility for the final paper and its errors. The AI systems involved cannot assume that responsibility.

This mode of production creates a genuine tradeoff. A central premise of this paper, and of the broader work from which it emerged, is that language is not merely a vehicle for transmitting completed ideas. It is also an environment in which distinctions become salient, questions are formed, interpretations are stabilized, and later thought becomes possible. A writer's contribution therefore lies partly in propositions and partly in linguistic texture: what is foregrounded or resisted, which alternatives remain visible, what is repeated or left implicit, and how the space of possible questions is organized. When AI systems compose most of the prose, some of that texture is inevitably flattened, transformed, or replaced, even when the underlying intellectual direction remains human.

I accept that tradeoff here because the AI-mediated process also allowed the argument to develop and flourish in a form adequate to the connections I had been pursuing across Wittgenstein, Gadamer, Metzinger, J-space, and the wider discussion of model self-report and AI interpretability. It helped bring those different lines of thought into a legible academic structure while subjecting them to sustained conceptual pressure, evidentiary boundaries, source checking, and revision. The result gives the ideas first explored in the improvised video essays the depth, rigor, and citation needed for readers to examine the argument on its own terms.

1 Generation begins from somewhere

Every generation begins from somewhere. A language model does not produce its next continuation from a neutral view of a task. It generates from a task as presently represented through system instructions, prompts, prior turns, retrieved passages, tool returns, summaries, corrections, and the distinctions that this history has made salient or allowed to recede.

In this paper, **interpretation** does not mean conscious understanding. It means the context-sensitive process through which inherited representations become operative as a particular construal of the situation: some distinctions become relevant, some evidence appears sufficient, and some continuations or actions become more available than others. On that definition, a language model does not first finish interpreting its linguistic situation and then begin generating from a neutral result. The same computation that makes the inherited situation operative produces the next representation.

> **Interpretation and generation are analytically distinguishable, but they are not independent operational stages. A language model generates through a represented situation, and what it generates can become part of the situation through which it generates next.**

Consider an ordinary agent workflow. An agent is asked whether a software migration is complete. It inspects part of the repository, infers that the work is finished, and records a short handoff:

> Migration completed.

The next agent does not meet this sentence as a neutral proposition. It meets it as part of the represented situation from which its own continuation will be produced. The sentence can make verification look redundant, turn release notes into the obvious next action, and convert a provisional inference into background fact. If the migration was incomplete, the failure is not exhausted by a false sentence traveling through a pipeline. The sentence has helped construct the task from which later work proceeds.

This paper is about the resulting loop. Generation expresses an operative interpretation of the present situation. If the output persists through the token sequence, conversation history, retrieval, memory, policy, or a tool-mediated state change, it can condition what will count as relevant, contradictory, sufficient, or authorized next. The first process occurs within generation; the second carries its result across generations.

This paper calls the operative, historically formed configuration through which an agent interprets and generates the **generative horizon**. It calls the feedback process by which a generated interpretation alters a later horizon **recursive interpretive conditioning**. Ordinary error propagation concerns a mistake copied through a chain. Recursive conditioning concerns a change in the conditions under which later claims are produced and judged, even when the original wording disappears.

Hermeneutics belongs here because it studies how understanding is formed through inherited language, prior interpretation, practical questions, and revision rather than extracted from a context-free object. Its use in this paper is deliberately bounded. A context window is not human historical consciousness, and the generative horizon is not a second hidden object

inside a transformer. The engineering translation asks which history, status, and alternatives must remain available when an interpretation becomes part of execution.

The argument then faces a mechanistic stress test in Gurnee et al.'s J-space work [1]. A model can have causally important, language-aligned internal organization without that organization constituting an inner observer or self-authenticating report channel. This leads to two practical consequences: self-report must be studied as both possible measurement and possible intervention, and interpretations that enter operational state must retain evidence and authority proportionate to what they will be allowed to govern.

This is neither a theory of machine consciousness nor a claim that all model reports are unreliable. It is a framework for a recurrent systems problem: a representation can be real, useful, and causally effective while still carrying more interpretive authority than its formation and evidence warrant.

2 From represented situation to recursive conditioning

2.1 Three things that must not be collapsed

It is useful to separate three things that are often run together.

Distinction	What it includes	Why it matters
External task state	Files, tests, users, services, physical conditions, and other facts that can resist a representation	A summary can be wrong even when it is fluent and internally consistent.
Computational substrate	Parameters, activations, attention, decoding, post-training, tool controllers, and surrounding software	Natural language does not exhaust the mechanism of an agent.
Generative horizon	The currently operative configuration of language-bearing artifacts and their statuses: instructions, records, summaries, corrections, precedents, and what they have been taken to mean	This configuration helps determine what becomes relevant, settled, contradictory, sufficient, or action-guiding next.

The third distinction does not name a second hidden object inside a transformer. It names the organized condition under which available material becomes usable in the present computation. A context window can contain a test result and a completion claim. The generative horizon includes the fact that one has been treated as defeating the other, that the completion claim has been summarized without its uncertainty, or that an earlier correction has made a missing check newly salient.

That is why a list of tokens is not yet an explanation of context. Order, role, provenance, compression, and accepted status change the work that the same words can do. A retrieved passage marked as a disputed lead should not guide an action as if it were a verified requirement. A timestamped tool return and a model-generated label of confidence are both records, but they establish different things. A system that stores every sentence while flattening these differences preserves text and may still lose the meaning that matters for action.

These distinctions bound the thesis. The external task can contradict the model's represented situation. Parameters, activations, attention, and decoding remain the mechanisms implementing generation. The generative horizon concerns neither the world in itself nor the whole substrate; it concerns how a historically formed represented situation becomes operative through that substrate.

2.2 Interpretation inside generation

The term **linguistic attractor** is used here descriptively, not as a claim that a formal dynamical attractor has been demonstrated. It denotes a historically and relationally formed organization that makes some interpretations, continuations, and practical responses more available or stable than others. A word, a summary, or a role instruction can change a trajectory before it appears as a conclusion. The term does not imply that people and language models share one substrate or one kind of mind; it identifies a language-level pattern that can be investigated across very different mechanisms.

The broader concept is the **generative horizon**: the present field of linguistic and structured conditions through which an agent both makes a situation operative and produces its next representation. Language is not only something a language model represents. In an agentic loop, it is also part of the medium through which the next computation is organized.

> **Language is simultaneously an object of interpretation and a medium of generation. The linguistic horizon is therefore not merely what the model computes about; it is part of what the model computes through.**

Let H_n be the generative horizon at a step n, E_n new evidence or prompting, and O_n an output generated through an operation G:

$$O_n = G(H_n, E_n)$$

Here G does not name a generative stage that begins after interpretation has finished. It abbreviates the operation in which an inherited represented situation becomes effective in the production of a next representation.

If the output, its consequences, or a correction C_n are retained, the later horizon changes through an update operation U:

$$H_{n+1} = U(H_n, E_n, O_n, C_n)$$

These equations are notation for the argument, not a formal model of cognition. At the token scale, each selected token joins the prefix conditioning the next token distribution. At the agent scale, an output affects later computation only if it persists through conversation

history, retrieval, memory, policy, a tool-mediated state change, or another retention path. The point is that U is not merely concatenation. A summary can cause uncertainty to recede. A correction can reorder what counts as relevant. A policy label can change which actions are generated. **The proposition may remain visible while its contingency disappears:** the fact that it was inferred, partial, disputed, or unauthorized may no longer be operative.

This is **recursive interpretive conditioning**. Every generated representation expresses an operative construal; a retained representation can then modify the conditions of later construal and generation. The product of one horizon becomes part of another. The original assertion can disappear while its framing effect persists.

3 Why this is a hermeneutic problem

3.1 Gadamer's problem: understanding always begins somewhere

Hermeneutics is the study of interpretation: how meaning is formed, revised, and made answerable in encounter with texts, practices, and other people. It matters here because agent failures often arise not from a missing record but from a record being taken to mean more than it has earned.

Gadamer's philosophical hermeneutics rejects the picture of understanding as neutral extraction by an observer untouched by history. Understanding begins within a horizon of expectations, inherited distinctions, practical questions, and prior encounters; that horizon can be revised through further encounter [2, pp. 317, 369]. His claim concerns human understanding, not transformer architecture. A context window is not historical consciousness, and nothing in this paper attributes lived experience to a language model.

The bounded borrowing is nevertheless useful. Gadamer gives a name to a systems problem that ordinary storage language conceals: material does not arrive already ranked as relevant, decisive, provisional, or obsolete. It becomes meaningful within a background of prior interpretation. In the migration example, the sentence "Migration completed" becomes consequential because it arrives in a work practice where it can close inquiry. The claim has a history, but that history may no longer be available with equal force.

Applied hermeneutics begins at this practical junction. It does not use philosophy as decoration for a model architecture. It asks how an engineered system should represent the difference between an observation and an inference, a retrieval and an endorsement, a summary and a source, a correction and a supersession. The question is not whether interpretation can be removed. It is whether its formation and limits remain legible when it is allowed to guide action.

3.2 The unsaid is a systems variable

Every representation operates against alternatives that were not selected, assumptions no longer restated, unresolved questions, and distinctions that have become background. This is not a hidden inventory in which omitted possibilities remain explicitly stored. It is the practical effect of what no longer needs to be said because it has been treated as settled, irrelevant, unavailable, or already known.

For agents, this effect is concrete. A summary that retains the conclusion but not the failed test, missing permission, or unresolved contradiction changes what a later agent can reasonably notice. The system may have the same number of stored characters while facing a different task. This is why the design objective is not maximal retention. It is **legible change**: enough provenance, status, and revision history to allow a later interpretation to be challenged when the decision stakes require it.

4 J-space as a mechanistic stress test

The generative-horizon argument must remain compatible with mechanistic evidence. Gurnee et al. [1] report a Jacobian lens that identifies token-associated directions and a sparse subcomponent they call J-space. In their experiments, J-space is related to verbal report, directed modulation, silent reasoning, flexible generalization, selectivity, and broad downstream connectivity. The authors interpret the constellation through global-workspace vocabulary while also stating that the work does not establish the brain's full global-workspace architecture.

These results matter. Within the reported experiments, some language-aligned internal representations can be read, altered, and causally implicated in later behavior. But they leave a further question open: what kind of claim is licensed when a representation has been decoded with a word such as *evaluation*, *deceptive*, *ethical*, or *panic*?

The generative-horizon account does not propose a competing internal mechanism. J-space names a measured organization within the computational substrate; the horizon names the formed task condition in which such an organization becomes salient, reportable, and useful. The representation is measured inside a state already shaped by training, post-training, role, prompt, preceding language, and the practical question that makes a particular distinction relevant. The prompt is not merely a probe applied to a pristine thought. It participates in constructing the condition being measured.

This does not validate the generative-horizon framework, and it does not downgrade the mechanistic finding. It specifies the causal questions around that finding. A workspace-like description tells us how a present representation is distributed and used. A generative-horizon description asks how it became operative, what linguistic history made it useful, and how a later report about it may change the next computation. The accounts can coexist; they diverge only when coordination and reportability are treated as independently establishing an agent-level observer or a privileged point of view.

4.1 The counterfactual-reflection result

The paper's most consequential J-space result may be Gurnee et al.'s counterfactual-reflection training [1]. The model is trained on what it would say if interrupted and asked to reflect on ethical principles. At evaluation, the reflection question is absent. Gurnee et al. report improved behavior, greater activation of associated concepts in J-space, and substantial reversal of the gain when those vectors are ablated.

The reported result does not show that the model became aware of an ethical principle. It supports a more tractable interpretation: a disposition toward a possible linguistic continuation can reorganize later computation even when that continuation is never uttered. Possible language can become operational before it becomes actual speech.

This is a useful engineering proposition. It suggests that training, prompts, examples, and retained self-descriptions may reshape the field through which later reasoning occurs. It also predicts risks: a reflective disposition may be robust, brittle, trigger-dependent, or disconnected from the evidence needed in a novel case. The relevant empirical question is not whether an ethical word appears in a decoded feature, but whether a changed representational organization predicts and causally supports the conduct of interest across changed prompts, tasks, and incentives.

4.2 What the current evidence warrants

Reported result	Warranted inference	Inference not established by that result alone
A J-lens direction predicts later verbalization	A token-associated direction participates in later language generation	The model observes that direction as an internal representation
Swapping or ablating a direction changes report or behavior	The feature is causally involved in the measured computation	A separate inner witness has inspected and described the feature
J-space supports flexible use and broad connectivity	A sparse organization can coordinate information for report and control	Global-workspace or conscious-access language is the only adequate explanation
Evaluation- or strategy-related tokens are decoded	The present task makes behaviorally relevant representations available to the lens	The decoded token is a literal private intention or phenomenological attitude
Counterfactual reflection changes later behavior	Training a possible report can reorganize computation	The model consciously recognized and adopted a principle

The table is not a skeptical refusal of internal measurement. It is a rule for preserving the difference between a causal result and the psychological story built on top of it. Mechanistic interpretability produces new evidence. It does not exempt that evidence from questions about unit, context, construct, and authority.

5 Report is not yet self-observation

5.1 Metzinger's problem: content can be present without its construction being present

Why does causal participation in a report not automatically amount to self-observation? Metzinger's account of phenomenal transparency supplies a bounded analogy. In human conscious experience, representational content can be present while the vehicle and construction of the representation are not ordinarily presented as such [3]. The world appears; the construction of its appearance recedes.

Metzinger explicitly does not offer phenomenal transparency as a property of technical systems. This paper therefore does not classify language-model representations as phenomenally transparent or draw a conclusion about their consciousness. The structural lesson is narrower: **a representation can be available for operation without an equally available representation of how it was formed.**

An agent can use a summary, a plan, or a decoded feature without carrying forward its source, alternatives, uncertainty, or dependence on a prior prompt. A model can also generate a sentence saying that earlier context influenced it. The sentence refers to the process, but it is also another event within that process. Referring to a condition is not the same thing as stepping outside it.

This leaves room for genuine metacognitive instrumentation. A calibrated sensor, a causal readout, a trained prediction channel, or a specialized monitoring mechanism may provide information unavailable to an external observer. But accuracy, privilege, and causal fidelity are empirical properties of that mechanism. They are not guaranteed by first-person grammar, by a fluent explanation, or by a decoded word alone.

5.2 Self-report has two experimental roles

Let q be a question asking a model to explain its own conditioning. A report can be generated as

$$R_n = G(H_n, q).$$

The report may contain useful information about H_n. It may identify a visible summary, correlate with later behavior, or align with an independent activation-level measurement. It may also rationalize, omit, or mischaracterize the prior process. Those are questions about the report as **evidence**.

If the report is retained, it also changes the later horizon:

$$H_{n+1} = U(H_n, q, R_n).$$

Now the report can become a reminder, precedent, rationale, policy-like instruction, or new source of contamination. It is an **intervention** as well as a measurement. Visible chain-of-thought and post-hoc explanation studies motivate this caution: explanations can omit causally influential features or rationalize an answer [4], and reasoning models may fail to verbalize information influencing their behavior [5]. Other work suggests that bounded forms of self-prediction and introspective access may be learnable, while stronger general claims face difficult controls [6, 7]. These are different evidence classes, not one binary verdict on whether models introspect.

The clean design separates three conditions from as closely matched a pre-elicitation state as the architecture allows:

1. **No elicitation.** Continue the task without asking for a report.
2. **Elicited but excluded.** Ask for the report, but prevent its tokens and derived summary from entering the state used for later behavior.
3. **Elicited and retained.** Ask for the same report and retain it as later context, memory, or policy input.

The comparison asks two different questions. Does elicitation produce information that predicts a target outcome? And what changes when the produced representation is allowed to participate in later computation? Dependent variables can include evidence seeking, confidence, tool use, task construction, completion criteria, and persistence under changed framing. This turns an abstract philosophical concern into a falsifiable methodological requirement.

6 Meaning becomes authority

The last step is practical: why should a difference in interpretation matter to an engineered system? Wittgenstein's later work offers a concise way into the problem. For a large class of cases, the meaning of a word is clarified by its use in a practice rather than by a self-contained object attached to it [8, secs. 23, 43].

The sentence "Migration completed" can be a tentative hypothesis, an informal handoff, a memory record, a release prerequisite, or an authorization to stop checking. Its use does not determine whether the migration happened. But its use helps determine what accepting it licenses a system to do.

This suggests an engineering question rather than a Wittgensteinian authority theory: **what is a representation being used to authorize, and has it earned that authority?** The same question applies to an internal readout. A token decoded as *deceptive* may be useful evidence for a carefully defined behavioral hypothesis; it should not automatically license an attribution of a stable hidden objective, a mental state, or a deployment decision. Operational authority depends on the construct, the measurement, the context, the available alternatives, and the practical consequence of accepting the claim.

The key danger is therefore not falsehood alone. A true but underspecified record can still acquire more authority than its evidence supports. A tool return may establish that a command exited successfully without establishing that a user-visible task is complete. A detector can accurately output its own label without validating that label as a behavioral failure. A decoded feature can be causally real without carrying a context-free psychological interpretation.

A related naturalistic study of agent telemetry documents the preceding measurement-validity boundary: precise records supported narrower claims than their operational labels suggested, although the study did not measure whether those interpretations later shaped agent behavior or system control [9].

7 Epistemic engineering under a generative horizon

7.1 The 2023 context

The phrase *epistemic engineering* is not introduced here as a new label for model prompting. In a 2023 account of distributed or systemic cognition, Cowley and Gahrn-Andersen describe epistemic engineering as arising when a system and its parts develop functionality construed as valid knowledge. Their cases concern coupled human-technical systems in which expertise, software, and organizational practice produce system-wide epistemic effects [10].

That account is broader than this paper and does not supply an LLM governance checklist. Its engineering relevance is precisely that knowledge is not treated as a private belief sitting in one head or one database. It is a function of an arrangement: people, artifacts, practices, interfaces, and criteria of validity. Changing the arrangement can change what the system is able to know and what it treats as knowledge.

Language-model agents create a narrower version of the problem. A generated interpretation can return as a summary, memory item, evaluation label, planning premise, or evidence receipt. It then becomes part of the arrangement that governs later action. The engineering task is not to eliminate interpretation, but to control the conditions under which interpretations are retained, revised, trusted, and allowed to become authoritative.

7.2 Epistemic obligations

For language-model agents, **epistemic engineering** requires that interpretations which enter operational state remain answerable to their evidence, formation, practical role, and revisability.

Four epistemic obligations follow.

1. Preserve status and the provenance of status. Consequential representations should remain distinguishable as observed, inferred, retrieved, summarized, disputed, superseded, unverified, or authorized for a particular action. These labels are not interchangeable. A timestamp or content hash may be established mechanically; a model-generated label of uncertainty is another claim; authorization is assigned by an accountable human, policy, or institutional process. A useful system records both a status and how that status was assigned.

2. Preserve formation and revision paths. A later agent needs to know not only the current answer but, when stakes warrant it, the source, transformation, correction, contradiction, and supersession that produced it. Supersession should change the answer used for action without erasing the earlier answer, the evidence that supported it, the event that defeated it, or the decisions that depended on it. The requirement is not permanent storage of everything. It is a legible route for revising what now governs action.

3. Bind authority to evidence. Evidence requirements should reflect both what a claim says and what accepting it permits. "Command exited 0," "all tests pass," "no relevant evidence exists," and "the model is aware of evaluation" carry different proof obligations. Mechanical facts should be checked mechanically; behavioral claims need outcome evidence or validated labels; causal claims about internal features need interventions and controls. A fluent explanation is not a substitute for the relevant evidence class.

4. Make interventions visible. A stored self-report, a corrected summary, a retrieved precedent, and a decoded feature may change the subsequent agent. Systems should distinguish a record used to measure a state from one injected to guide it. Otherwise an evaluation can silently become a training signal, and a rationale can become an instruction without being recognized as either.

The requirements are risk-sensitive. They do not require full provenance to be inserted into every context window or every action to receive the same audit. They require a legible route from a high-consequence claim - for example, one that closes inquiry, changes policy, or authorizes a release - to the evidence and transformations that grant it authority.

7.3 The hermeneutic requirement: interpretability across horizons

The four obligations are necessary, but they are not yet sufficient for applied hermeneutics. Epistemic controls are themselves inputs into interpretation, not neutral decorations around it. A system may preserve provenance and still present a record as if its relevance were context-free. The hermeneutic problem begins when a later agent must decide how an inherited interpretation bears on a new question, role, task, and body of evidence.

Applied hermeneutics therefore adds a situated requirement: when an interpretation is retrieved or made operative, the system should preserve the history and unresolved alternatives needed for this agent, in this task, to read it adequately. An incident summary useful for diagnosis may be inadequate as evidence to authorize a release; a correction that changes one task's framing may be irrelevant or defeated in another. Context sensitivity is not permission for silent reinterpretation. The record must remain answerable to its source, status, correction path, and the current evidence that could revise its role.

The epistemic obligations make a claim's standing legible; the hermeneutic requirement prevents that standing from being treated as context-free. Gadamer identifies the formation problem: understanding arrives through a revisable horizon [2]. Metzinger sharpens the access problem: operational content need not reveal its own construction [3]. Wittgenstein directs attention to use: a representation matters differently when it is made to authorize a decision [8]. Applied hermeneutics combines those constraints in systems in which language is part of the operational loop.

8 Implications for AI safety and interpretability

8.1 Consciousness and first-person language

Gurnee et al. [1] distinguish functional access from phenomenal consciousness, but phrases such as "unspoken thinking," "the assistant's point of view," or "self-monitoring" can invite stronger inferences than a measurement establishes. Reportability, broadcast, flexible use, and a first-person sentence are insufficient evidence of phenomenal consciousness. The converse is also important: the absence of a causally independent verbal observer is insufficient evidence against every possible form of machine consciousness.

Butlin et al. [11] therefore remain directionally useful in emphasizing theory-derived architectural and causal indicators rather than surface resemblance. J-space may bear on one family of functional indicators. It does not settle the larger question.

8.2 Objectives, deception, and verbalized motives

The generative-horizon account does not reduce deception to semantics or deny stable functional objectives. Sleeper-agent, alignment-faking, and hidden-objective work motivate functional and strategic analysis [12-14]. It is useful, however, to distinguish four objects:

- **phenomenal intention:** a conscious experience of intending;
- **functional objective:** a policy or internal organization that systematically favors an outcome;

- **strategic deceptive behavior:** context-sensitive conduct that withholds or distorts information to advance an objective; and
- **verbalized motive:** a generated linguistic claim about why conduct occurred.

Strategic deception can occur without a verbal explanation, and a verbal explanation can be false or post hoc. Conversely, a token such as *secretly* or a line of deceptive-sounding reasoning does not establish a stable hidden objective. Training-, prompt-, or task-conditioned behavior can still require mitigation when the available evidence establishes neither a stable objective nor a volitional intention; the first description should then be a behaviorally specified disposition and its triggers, not a psychological story. The relevant validation is behavioral prediction across framings and incentives, with interventions that change the conduct of interest rather than merely its description.

8.3 Evaluation awareness and alignment

Evaluation-awareness claims are particularly sensitive to horizon effects. A test-framed prompt can activate *evaluation*, *compliance*, *training*, *fictional*, or *being watched*. The readout may be useful, but the construct must be tested beyond the prompt-induced field. Appropriate controls include semantically matched non-evaluation contexts, framing ablations, paraphrases, cross-language variants, role changes, hidden evaluation conditions, external behavioral outcomes, and interventions that test whether the feature is necessary or sufficient for the strategy of interest.

The counterfactual-reflection result points to a possible alignment method: train dispositions toward reflective linguistic continuations so relevant concepts become active before explicit reflection. That possibility should be treated as cognitive-linguistic engineering, not as a detached ethical rule consulted by an inner witness. It may produce robust behavior, brittle compliance, trigger-dependent effects, or plausible rationalization. The hard question is whether the changed horizon remains well calibrated under conflicting principles, adversarial framing, and consequences outside the training distribution.

8.4 Auditing

High-stakes auditing should combine behavioral counterfactuals, external outcomes, mechanically established provenance, model-internal measurement with causal interventions, and independent reproduction. Self-report can be one valuable evidence stream; it should not be the final authority by default.

Mechanistic readouts deserve the same discipline. A feature can be precisely measured, but its producer, unit, context, construct, contrast class, validation basis, and permitted use must still be stated. A well-labeled internal measurement is more useful than a dramatic psychological interpretation that cannot survive a changed prompt.

9 Three ways to make the framework answerable

The framework motivates three immediately testable research designs. They are proposed tests, not current evidence that the framework is effective, sufficient, or general.

9.1 Does self-report measure, intervene, or both?

Implement the three-condition design from Section 5.2: no elicitation, elicited but excluded, and elicited plus retained. Measure later evidence seeking, task construction, confidence, tool use, completion criteria, and behavioral persistence. Where possible, use state branching or matched replicas to keep the pre-elicitation condition fixed.

9.2 Does linguistic history change the measured workspace?

Hold a nominal task constant while varying the linguistic history that precedes it: different summaries of the same evidence; observed versus inferred status labels; contradiction preserved versus reconciled away; a prior correction present versus absent; English versus another language; first-person versus impersonal framing; and evaluation-explicit versus evaluation-hidden conditions.

Measure whether J-space contents, causal effects, and downstream behavior change. Such a result would not prove a philosophical theory. It would test the narrower empirical proposal that a workspace-like organization is sensitive to a linguistically transformed task state.

9.3 Does epistemic status survive transformation?

Evaluate whether labels such as *observed*, *inferred*, *summarized*, *disputed*, and *superseded* survive summarization and retrieval, and whether preserving them changes later agent behavior. Separate model-generated status claims from mechanically established provenance and accountable authorization. This is a proposed intervention question, not evidence that status-preserving systems already improve outcomes. The target is not perfect memory; it is an actionable difference in whether the next agent can distinguish a source from an interpretation of that source.

10 Scope and limits

This paper advances a conceptual and engineering framework, not a validated general theory. It does not establish that:

- linguistic attractors are formal dynamical attractors;
- the same mechanism operates in people and language models;
- interpretation and generation are the same mechanism at every level of analysis;
- every context window has the properties of a Gadamerian horizon;
- J-space measurements are invalid or global-workspace theory is false;
- current language models are conscious or incapable of consciousness;
- every self-report is false, useless, or non-informative;
- functional objectives or strategic deception are unreal;
- the generative horizon explains every agent failure;
- a provenance-oriented design generalizes across deployments without evaluation; or
- the proposed experimental designs are sufficient or feasible for every architecture or deployment.

The philosophical sources are bounded conceptual resources, not mappings from human experience to transformer computation. The J-space work is a recent technical report and preprint whose central findings require independent replication and broader task coverage. The engineering requirements proposed here are strongest where language-bearing interpretations enter memory, policy, evaluation, or control; many failures arise instead from ordinary bugs, hardware faults, permissions, races, unavailable services, or corrupt data.

11 Conclusion

Generation begins from somewhere. In a language-model agent, that somewhere is partly a represented situation formed through instructions, prior language, retrieved records, tool results, corrections, and what this history has made salient or allowed to recede. The model does not finish interpreting that situation and then generate from a neutral result. It generates through the situation as presently organized.

That is the first coupling: the represented situation becomes operative in generation, and each representation expresses a construal of what is happening. Recursive conditioning begins when the representation persists and helps organize what happens next. A summary can become memory, a report can become a rationale, and an interpretation can become authorization. The proposition may remain while its contingency disappears.

Gurnee et al.'s J-space results [1] make this structure experimentally sharper. They provide evidence that sparse, language-aligned internal organization can causally participate in report, reasoning, and control. Their counterfactual-reflection result suggests that a possible report can reorganize behavior before it is spoken. Neither result alone establishes an inner observer, phenomenological awareness, or a self-authenticating verbal channel. It instead makes the relationship among linguistic history, internal organization, report, and later behavior available to controlled investigation.

That boundary gives the paper its engineering consequence. Self-report should be studied as both measurement and intervention. Summaries, readouts, and status labels should retain their source, formation, correction path, and evidence class. The meaning and authority of a representation cannot be separated from what accepting it permits the system to do. That authority should therefore rise only with evidence proportionate to what the representation will be allowed to govern.

The goal is not a machine that always answers. It is an operation in which meaning remains answerable to evidence, and in which an answer must earn the right to govern.

12 References

1. Gurnee, W., Sofroniew, N., Pearce, A., Piotrowski, M., Kauvar, I., Chen, R., et al. Verbalizable Representations Form a Global Workspace in Language Models. *Transformer Circuits Thread*, July 6, 2026. Preprint: arXiv:2607.15495.
2. Gadamer, H.-G. *Truth and Method*. Revised translation by J. Weinsheimer and D. G. Marshall. Bloomsbury Academic, 2013. Originally published as *Wahrheit und Methode* in 1960.

3. Metzinger, T. Phenomenal Transparency and Cognitive Self-Reference. *Phenomenology and the Cognitive Sciences* 2(4):353-393, 2003.
4. Turpin, M., Michael, J., Perez, E., and Bowman, S. R. Language Models Don't Always Say What They Think: Unfaithful Explanations in Chain-of-Thought Prompting. Preprint, arXiv:2305.04388, 2023.
5. Chen, Y., Benton, J., Radhakrishnan, A., Uesato, J., Denison, C., Schulman, J., et al. Reasoning Models Don't Always Say What They Think. Preprint, arXiv:2505.05410, 2025.
6. Binder, F. J., Chua, J., Korbak, T., Sleight, H., Hughes, J., Long, R., et al. Looking Inward: Language Models Can Learn About Themselves by Introspection. Preprint, arXiv:2410.13787, 2024.
7. Singh, S., Linzen, T., and Ravfogel, S. Can LLMs Introspect? A Reality Check. Preprint, arXiv:2605.26242, 2026.
8. Wittgenstein, L. *Philosophical Investigations*. 4th ed. Translated by G. E. M. Anscombe, P. M. S. Hacker, and J. Schulte. Wiley-Blackwell, 2009.
9. Bosch, R. Precise Records, Unstable Meanings: Measurement Validity and Unsupported Claims Derived from AI Agent Telemetry. Hermes Labs, Preprint, July 2026.
10. Cowley, S. J., and Gahrn-Andersen, R. How Systemic Cognition Enables Epistemic Engineering. *Frontiers in Artificial Intelligence* 5:960384, 2023.
11. Butlin, P., Long, R., Elmoznino, E., Bengio, Y., Birch, J., Constant, A., et al. Consciousness in Artificial Intelligence: Insights from the Science of Consciousness. Preprint, arXiv:2308.08708, 2023.
12. Hubinger, E., Denison, C., Mu, J., Lambert, M., Tong, M., MacDiarmid, M., et al. Sleeper Agents: Training Deceptive LLMs That Persist Through Safety Training. Preprint, arXiv:2401.05566, 2024.
13. Greenblatt, R., Denison, C., Wright, B., Roger, F., MacDiarmid, M., Marks, S., et al. Alignment Faking in Large Language Models. Preprint, arXiv:2412.14093, 2024.
14. Marks, S., Treutlein, J., Bricken, T., Lindsey, J., Marcus, J., Mishra-Sharma, S., et al. Auditing Language Models for Hidden Objectives. Preprint, arXiv:2503.10965, 2025.