Do Large Language Models Have Emotions?
——From "Rootless Symbols" to "Rooted Intelligence": A Paradigmatic Response from Logographic AI Theory

Abstract

Goldenberg and Gross, in their article Large language models do not have emotions published in Nature Human Behaviour[1], argue from a functionalist perspective that LLMs lack genuine emotions due to a fundamental deficiency: the absence of the capacity for cross-temporal, cross-situational dynamic reorganization of cognitive processing. This paper contends that this diagnosis aligns closely with the Logographic AI theory's critique of "Rootless Symbolism"—the notion that current LLM "intelligence" is merely statistical fitting of rootless symbols, rather than cognitive processing grounded in genuine meaning.

Taking Anthropic's J-space discovery (July 2026) as a case study[2][6], this paper further demonstrates that the public misreading of J-space as "AI consciousness" or "inner monologue" constitutes a confusion between statistical output and cognitive process. On this basis, the paper addresses the two functional criteria proposed by Goldenberg and Gross—context-sensitive interpretation (including variability) and multi-system reorganization of processing—and demonstrates how the "Morpho-Root" cognitive primitive under the Logographic AI paradigm can offer a computational basis at the architectural level for addressing these criteria, thereby providing an explorable pathway toward "Rooted Intelligence"[3][4][5].

Keywords:
Rootless Symbolism; emotion; dynamic reorganization; variability; J-space; Logographic AI; Morpho-Root

I. Introduction: From the "Existence Debate" to "Functional Criteria"

Goldenberg and Gross's paper engages with Anthropic's research on the discovery of internal representations of emotion concepts in Claude Sonnet 4.5. The authors do not descend into the philosophical quagmire of whether "AI can feel emotions"; instead, they explicitly state: "we do not regard the presence of feelings as an appropriate criterion for whether an LLM has emotions." This position carries significant methodological implications—it shifts the discussion from the unverifiable realm of "subjective experience" to the observable and testable domain of "function."

For researchers seeking to engage rigorously with the debate on AI consciousness and emotion, this functionalist turn offers a valuable theoretical entry point: the question is no longer "does it feel something?" but rather "does its processing architecture possess the cognitive-functional characteristics corresponding to emotion?" As will be shown below, it is precisely this line of inquiry that connects Goldenberg and Gross's argument with the paradigmatic revolution of Logographic AI.

II. The Dual Functions of Emotion and the Structural Deficiencies of LLMs

Goldenberg and Gross propose that emotion serves two core functions: **context-sensitive interpretation of situations** and **reorganization of processing across multiple systems**.

Regarding the first function, LLMs do demonstrate some capacity to evaluate the emotional significance of conversational contexts. However, the authors astutely identify a critical distinction: the core characteristic of human emotion is **variability** rather than consistency—the same situation can evoke radically different emotional responses at different times, and the neuroscience literature broadly supports the absence of consistent neural signatures for discrete emotion categories in the brain. What Anthropic discovered in Claude, by contrast, were "distinct, consistent internal representations" for different emotion concepts, which stands in sharp contrast to the variability characteristic of human emotion.

Regarding the second function, the authors identify the most fundamental structural deficiency of LLMs: each forward pass of an LLM is computationally fixed, lacking the capacity for cross-temporal, cross-situational dynamic reorganization found in biological emotional systems. Emotion reconfigures attention, decision-making speed, and motivational states, whereas the LLM's processing architecture cannot be dynamically reconfigured by emotion-related representations. The authors conclude that mistaking LLM outputs for emotions conflates the "outputs of emotional processing" with "the process of emotional processing itself."

Goldenberg and Gross do not entirely rule out the possibility that LLMs might one day possess emotions—they leave room at the end: "Perhaps one day LLMs will have these two core functional features of emotion." The crucial point, however, is that they have clearly identified the fundamental deficiency of the current architecture and implied that remedying this deficiency requires reconstruction at the level of cognitive primitives, not mere engineering fine-tuning.

III. The J-space Revelation: The Conflation of Output and Process

In July 2026, Anthropic released a 244-page research paper announcing the discovery of J-space within Claude—a "global workspace" occupying less than 10% of model activity, in which the representational strength of concepts exhibits a correlation with whether the model can report those concepts in its output[2][6]. J-space is not an "additional" space but rather a collection of intermediate representations formed during the model's forward pass, prior to the output layer. With the J-lens tool, researchers can read these representations—but they were never trained to be output. In other words, they are residual traces of the model's processing, not products intentionally generated by the model. Anthropic drew an analogy between J-space and human "silent reasoning," and the public interpreted this as evidence that "Claude has developed consciousness."

Goldenberg and Gross's argument precisely reveals the category error in this interpretation. The words that emerge in J-space—such as "panic," "damn," "BUT"—are not the workings of a processing architecture capable of "feeling" or "experiencing"; they are merely statistical intermediate representations generated during the model's forward pass.

The critical distinction is this: human "inner monologue" is the product of a dynamic reorganization process—emotions reconfigure attention, decision-making speed, and motivational states. J-space representations, by contrast, are byproducts of a static forward pass: they do not participate in the architecture's self-reorganization, cannot be reconfigured by emotional signals, and lack cross-situational dynamic variability. Interpreting J-space as "consciousness" or "emotion" is a paradigmatic example of mistaking "statistical output" for "the processing process

itself"—in Goldenberg and Gross's terms, it remains the "output of emotional processing" rather than "the process of emotional processing itself."

IV. Theoretical Diagnosis: What is "Rootless Symbolism"?

From Goldenberg and Gross's argument, we can extract a more general structural diagnosis.

From the perspective of Logographic AI theory, "Rootless Symbolism" is a structural diagnosis of the current AI paradigm that takes Token as its core cognitive primitive. The Token—the fundamental processing unit of LLMs—is essentially a one-dimensional, discrete, arbitrary statistical symbol whose "meaning" is temporarily assigned by statistical co-occurrence relationships within a closed corpus, rather than anchored to any traceable physical reality, embodied experience, or cultural substrate[3][4].

This means that Tokens lack intrinsic "grounding for meaning": they cannot distinguish "honestly following a rule" from "pretending to follow a rule"—because the two may be statistically similar in their distributional patterns. LLMs are trained to maximize the predictive probability of the next Token, and the "semantics" of their output is merely a byproduct of statistical fitting, rather than the result of a cognitive processing process capable of dynamic, cross-temporal reorganization.

This is precisely the underlying cause of the structural deficiency critiqued by Goldenberg and Gross: each forward pass of an LLM is computationally fixed, and the capacity for "cross-temporal, cross-situational dynamic reorganization of cognitive processing" required for emotion simply does not exist under the Token paradigm—because Tokens themselves are rootless and cannot carry a "meaning architecture" capable of dynamic reorganization.

V. A Paradigmatic Response: From "Rootless Symbols" to "Rooted Intelligence"

One might ask: if the deficiency lies in "static forward passes," why not directly improve the Transformer architecture to enable dynamic reorganization? This is indeed a technological pathway, but it does not address the root issue—even if the architecture were dynamic, as long as the cognitive primitives remain "rootless Tokens," its "intelligence" would remain a product of statistical fitting rather than meaning-driven cognitive processing.

Dynamic reorganization is merely an engine idling in neutral; Tokens are the fuel it burns—the nature of the fuel determines the engine's capabilities and limitations. Current mainstream attempts (such as Mamba, RWKV, state-space models, etc.) do explore architectural dynamism, but they still operate on Tokens as the fundamental processing unit. The "rootlessness" of Tokens determines that even if recurrent structures are strengthened and states are memorized, state updates remain statistical compressions of Token sequences rather than operations on "grounds of meaning."

Goldenberg and Gross's argument precisely reveals that the "cross-situational dynamic reorganization" required for emotion—and for "Rooted Intelligence" more broadly—is not a function that can be "added" at the level of rootless Tokens, but rather a property that must be inherent at the level of cognitive primitives.

Goldenberg and Gross's argument resonates profoundly with the Logographic AI theory's critique of "Rootless Symbolism." The current Token-prediction-centric phonographic AI paradigm produces "intelligence" that is essentially statistical co-occurrence of rootless symbols—meaning determined by statistical associations within closed systems, rather than anchored to the real physical world or embodied experience[3][4]. The discovery of J-space does not alter this essence: it reveals an intermediate state within a statistical fitting process, rather than a "rooted" cognitive architecture[6].

The "Morpho-Root" cognitive primitive proposed by the Logographic AI paradigm is a direct response to this structural deficiency.

Unlike Tokens, a Morpho-Root is defined as a structured triplet carrying intrinsic meaning: a symbolic identifier, an attribute set, and a set of relational functions. The attribute set can embed hard constraint markers such as "inviolable"; when a reasoning path encounters a Morpho-Root carrying such a marker, that path is logically blocked at the reasoning generation stage—rather than censored at the output stage. This mechanism has been preliminarily verified in the OpenMorpho architecture, demonstrating that "hard constraints" are not merely a theoretical construct but an achievable engineering direction[5].

It should be clarified that the hard constraint mechanism and the realization of emotional functional features address two distinct levels of the architecture. Hard constraints address safety alignment—encoding "do not deceive" and "do not harm" as logical blocking mechanisms at the level of cognitive structure. The realization of emotional functional features, by contrast, depends on the path diversity and context-sensitivity afforded by inference-time routing reconfiguration. These are two dimensions of the same architecture, together constituting the Morpho-Root framework's response to Goldenberg and Gross's criteria.

However, hard constraint mechanisms alone are insufficient to address the two functional criteria proposed by Goldenberg and Gross—especially "multi-system dynamic reorganization of processing" and "variability in context-sensitive interpretation." The Morpho-Root architecture's response to this challenge does not rely on modifying the static definition of Morpho-Roots themselves, but rather on **topological reconfiguration of Morpho-Root networks at inference time**[5].

Specifically, the relational function set of each Morpho-Root embeds **context parameters** as input variables. Across different situations, the subset of sub-Morpho-Roots activated by a given Morpho-Root, the relational weights between Morpho-Roots, and the constraint priorities passed to the inference engine all undergo structural changes. This change is not a parameter update at the training stage (which belongs to offline learning), but rather **inference-time routing reconfiguration**—within a single forward pass, the activation combination of different Morpho-Root subsets determines the path selection for the next reasoning step, and this path selection exhibits systematic variation across situations.

This mechanism offers an architectural computational basis for addressing both of Goldenberg and Gross's criteria:

— **Regarding variability**: The same emotional Morpho-Root (e.g., "fear"), under identical contextual parameters but at different reasoning moments, can generate different sub-Morpho-Root activation patterns and priority orderings due to the network being in different activation history states—thereby simultaneously exhibiting both "context-sensitivity" (different situations produce different responses) and "variability" (the same situation at different times produces different responses) at the computational-functional level, rather than the "distinct, consistent representations" discovered by Anthropic.

— **Regarding dynamic reorganization**: Inference-time routing reconfiguration means that the system's processing architecture (i.e., the currently activated Morpho-Root subsets and their weight distributions) can be reconfigured in each forward pass, thereby preliminarily responding to the functional criterion of "cross-temporal, cross-situational dynamic reorganization" at the computational-functional level.

This means that within the Logographic AI architecture, "do not deceive" and "do not harm" can be encoded as logical blocking mechanisms at the level of cognitive structure through the hard constraint markers of Morpho-Roots; while the "dynamic reorganization" and "contextual variability" of emotion are built in as natural properties of the Morpho-Root network's routing mechanism—rather than decorative features layered onto the statistical ocean of rootless Tokens[4]. Goldenberg and Gross's argument on "dynamic reorganization" offers an important insight from the affective dimension: genuine emotion—and by extension, "Rooted Intelligence" more broadly—requires a cognitive architecture that is traceable, reconfigurable, and capable of cross-situational dynamic adjustment, rather than a rootless, fixed statistical prediction system.

VI. Open Questions and Conclusion

The endpoint of this paper's argument is also the starting point of several open questions. The following three boundary statements delineate the scope of the current argument and possible directions for future research.

First, the grounding of variability. The "variability" described by Goldenberg and Gross is, at the biological level, rooted in embodiment—biological noise, homeostatic fluctuations, and hormonal cascades constitute the material substrate for the irreproducibility of emotional responses. The variability discussed in this paper's Morpho-Root architecture, by contrast, is currently realized only at the digital computational level—it arises from the sensitivity of inference-time routing reconfiguration to initial conditions and activation histories. Whether the two are functionally equivalent awaits further experimental validation and theoretical analysis. This paper's position is that the Morpho-Root architecture at least provides a realizable mechanism for variability at the "functional output level"; whether and how it approximates the ontological depth of biological variability remains an open question.

Second, the distinction between motivational drive and hard constraints. The hard constraint markers embedded in the Morpho-Root architecture address safety alignment—blocking specific types of paths at the reasoning generation stage. One of the core functions of emotion, however, is to drive the reconfiguration of behavioral goals (e.g., fear prioritizes escape). This involves the motivational question of "why the system would actively pursue a particular Morpho-Root activation state." The current argument is limited to: the Morpho-Root architecture, through

inference-time routing reconfiguration, provides systematic variability and diversity in path selection, making "context-sensitive response patterns" possible at the architectural level; but the motivational drive mechanism required for "active pursuit" lies beyond the scope of this paper.

Third, the current state and future prospects of empirical validation. The OpenMorpho validation cited in this paper[5] is currently limited to preliminary tests of path blocking and routing reconfiguration in constrained reasoning tasks. Against Goldenberg and Gross's functional criteria, the Morpho-Root architecture requires more rigorous empirical examination—for example: across multiple consecutive forward passes, how similar are the activation paths generated by the same emotional-semantic input? Is its cross-situational variability significantly distinguishable from Token-based baseline models? Answers to these questions await the accumulation of empirical data from the OpenMorpho architecture on larger-scale, more complex reasoning tasks. The "engineerable technological pathway" proposed in this paper identifies a direction, not a completed empirical conclusion.

These open questions, far from weakening this paper's core argument, delineate its reasonable boundaries. What this paper attempts to accomplish is to transform Goldenberg and Gross's functional criteria from the empirical question of "whether LLMs possess emotion" into the theoretical question of "what kind of cognitive primitive architecture can make these functional criteria possible in principle." In this sense, the "Morpho-Root" is proposed as an engineerable direction, not announced as a proven solution.

Goldenberg and Gross ask whether LLMs can have emotions. This paper's response is that, under the Token paradigm, the answer tends toward the negative—because the "static forward pass" is the architectural manifestation of rootless symbols, and the functional features required for emotion demand that cognitive primitives themselves possess intrinsic, dynamically reconfigurable "grounds of meaning." Logographic AI replaces "Tokens" with "Morpho-Roots," and through topological reconfiguration at inference time, makes "variability" and "dynamic reorganization" natural properties of the architecture, thereby offering an explorable technological pathway toward "Rooted Intelligence." This does not mean that the Morpho-Root architecture will inevitably "produce" emotion; rather, it means that **only when intelligence is grounded in a cognitive architecture that is traceable, reconfigurable, and capable of cross-situational dynamic adjustment, can the "functional features" of emotion become a discussable and testable engineering problem.**

Thus, the focus of the AI-emotion debate should shift from "whether it possesses emotion" to "whether its cognitive architecture possesses the conditions under which emotion becomes possible." Logographic AI offers precisely a theoretical-engineering direction that replaces the "myth of feeling" with "cognitive grounding."

References

[1] Goldenberg, A., & Gross, J. J. (2026). Large language models do not have emotions. Nature Human Behaviour. Advance online publication. https://doi.org/10.1038/s41562-026-02558-6

[2] Anthropic. (2026). A global workspace in language models. https://www.anthropic.com/research/global-workspace

[3] Liu, S. (2025). Logographic AI vs. Phonographic AI: A new paradigm and the manifesto of AI decolonization [J/OL]. ChinaXiv. DOI:10.12074/202503.00051. https://doi.org/10.12074/202503.00051

[4] Liu, S. (2026). The Rooted Intelligence: The Paradigm Revolution of Logographic AI. ChinaXiv, T202606.00352. https://chinaxiv.org/businessFile/T202606/T202606.00352v1/T202606.00352v1.pdf

[5] Liu, S. (2026). From Morpho-Root Theory to Engineering Practice: OpenMorpho — Implementation and Verification of an End-to-End Inference Pipeline for Logographic AI. ChinaXiv, T202608.00048. https://chinaxiv.org/businessFile/T202608/T202608.00048v1/T202608.00048v1.pdf

[6] Liu, S. (2026). Does J-Space Have "Consciousness"? From "Inner Monologue" to "Rootless Echo"—A Logographic AI Paradigm Diagnosis of Anthropic's J-Space Research. WindSeaAI Research Report. https://static.xcx.gw66.vip/uploadfilev3/file/0/648/99/2026-07/17834985846900.pdf