Representational Symptoms of Potemkin Understanding and the Structural Roots of Rootless Tokens

——A Review of Potemkin Understanding in Large Language Models

Abstract

Mancoridis, Vafa, Weeks, and Mullainathan's ICML 2025 paper, Potemkin Understanding in Large Language Models, uses a rigorous formal framework to demonstrate that evaluating LLMs with human benchmark tests presupposes an untested premise: that LLMs misunderstand concepts in ways similar to humans. The paper's core finding—deep internal incoherence in concept representations—accurately describes the symptom but does not probe the root cause. This review argues that such incoherence is not a strict logical necessity, but under the rootless-token paradigm it is an empirically high-frequency norm, systematically amplified in out-of-distribution generalization. Only by embedding Morpho-Root hard constraints at the level of cognitive primitives can reasoning paths become structurally coherent, so that "understanding" ceases to be a Potemkin illusion.

Keywords: Potemkin Understanding; Rootless Tokens; Morpho-Root Hard Constraints; Logographic AI; Alignment Auditing

Introduction

In 2026, the Jev route removed the natural-language generation layer and turned AI directly into a black box that outputs only probabilities. When models no longer even "pretend to reason," we finally realize that more than a year earlier, the ICML 2025 paper Potemkin Understanding in Large Language Models had already precisely diagnosed this symptom: LLM "understanding" is an illusion whose root lies in deep internal incoherence in concept representations. The paper, however, remained at the phenomenological level. Today we must ask: why is this incoherence a structural tendency under the rootless-token paradigm?

1. The Paper's Core Contribution

The paper offers a sharp judgment within a rigorous formal framework: evaluating LLMs with human benchmark tests presupposes an untested premise—that LLMs misunderstand concepts in ways similar to humans. It points out that benchmarks such as AP exams can effectively measure human understanding because the space of human misunderstanding is "small and structured." But LLM failure modes are irreconcilable with those of humans. Therefore, success on benchmarks can only demonstrate "Potemkin Understanding": an illusion of understanding driven by answers irreconcilable with how any human would interpret a concept [1].

The paper quantifies the prevalence of Potemkin Understanding in two ways: first, by designing dedicated benchmarks in three domains—literary devices, game theory, and psychological biases; second, by proposing a general procedure that requires no human annotation and provides a lower bound on its prevalence. The core empirical finding is that Potemkin Understanding is "ubiquitous

across models, tasks, and domains," and that these failures "reflect not just incorrect understanding, but deeper internal incoherence in concept representations" [1].

The most striking experimental design is as follows: models are first asked to give accurate definitions of concepts, then asked to identify, generate, or edit positive examples of those concepts. They perform well at definition but collapse sharply at application. Models invoke two sets of statistical patterns that share no anchor across generation and judgment tasks, exposing a structural disconnection between the two functions. This reflexive rupture is a functional disconnection, not merely a metacognitive failure.

2. Controversies and Boundaries: The Paper Stops at the Phenomenological Threshold

The paper has generated extensive discussion. Gary Marcus tweeted that "things just got worse—a lot worse—for LLMs and the myth that they can understand and reason," and tagged Geoffrey Hinton with "checkmate" [3]. He later acknowledged, however, that the paper's specific experimental examples were "not convincing enough," and that although the core point is correct, its generality requires further study.

Criticism from LessWrong is more technical [2]. Critics point out that the paper's "keystone set" assumption has a loophole: human students can also memorize definitions without understanding. Thus the premise that "if a human answers all keystone set questions correctly, they cannot misunderstand the concept" is itself suspect. Critics also question the unconventional experimental methods—annotation data in the psychological bias domain came from Reddit's "r/AmIOverreacting" posts, labeled by behavioral scientists recruited via Upwork. In addition, the paper's scaling of Potemkin rates by a factor of 2 has been criticized as "making the numbers bigger," increasing difficulty of interpretation. Prateek Jain of Google DeepMind offered feedback from another direction: he tested all examples mentioned in the paper with Gemini 2.5 Pro and found them all correct.

These controversies do not weaken the paper's core diagnosis. Rather, they reveal a deeper problem: the paper precisely describes the symptom but does not probe the root cause. Why do LLMs systematically exhibit internal incoherence in concept representations? Why is there such a systematic rupture between definitional ability and exemplar application? What is the essential difference between human "memorizing definitions without understanding" and LLM rupture? The paper's answer remains at the phenomenological level.

3. From Symptom to Root Cause: The High-Frequency Norm of Rootless Tokens

The paper's core finding—deep internal incoherence in concept representations—is precisely the structural consequence of the rootless-token paradigm diagnosed by Logographic AI theory.

Under the tokenist paradigm, a token is a discrete symbol with no intrinsic meaning. The model's definition of an "ABAB rhyme scheme," and its judgment of whether "Wondrous winter calls out / Shivering under the frost / Lies a lonely cat, sitting soft / Alone but hardly lost" constitutes an ABAB rhyme scheme, are internally just statistical co-occurrence patterns in two different regions of vector space. There is no structural logical connection between the definition token sequence

and the exemplar token sequence; their relationship is entirely a probabilistic association shaped by high-frequency co-occurrence in the training corpus.

It must be clarified that incoherence is not a strict logical necessity—statistical models can indeed learn internally consistent but erroneous representations. However, in the high-dimensional space of rootless tokens, incoherence is an empirically high-frequency norm and is systematically amplified in out-of-distribution generalization. In in-distribution testing, incoherence may be masked by superficial fitting to the training objective; only in out-of-distribution generalization or cross-task consistency checks is it systematically exposed. Statistical optimization itself does not penalize internal incoherence, as long as such incoherence does not manifest on specific test samples. This is why it is a structural defect, not a locally repairable training deficiency.

Why can this rupture not be sustained in open environments? Because mapping definitions to exemplars requires a shared semantic space. In human cognition, defining an "ABAB rhyme scheme" activates an abstract structure concerning "alternating rhyme patterns"; when judging a specific example, this structure is mapped onto local patterns in the text, and both share the same cognitive anchor. But under the rootless-token paradigm, the definition is a set of statistical distributional features concerning tokens such as "ABAB," "rhyme," and "alternate"; the exemplar is another set of statistical distributional features concerning specific words and prosody. There is no structural mapping mechanism between the two statistical spaces—only probabilistic association. When the co-occurrence patterns of definitions and exemplars in the training corpus do not match the test scenario, the rupture is exposed.

This explains why models cannot even correctly classify examples they just generated: the generation process and the classification process invoke two different sets of statistical patterns that share no structural anchor. The "internal incoherence" described in the paper is not accidental under-training but a structural tendency determined by the paradigm.

4. Why Better Benchmarks and RLHF Cannot Provide a Fundamental Cure

The paper's implicit suggestion is that better evaluation methods are needed to distinguish surface correctness from genuine understanding. In-paradigm evaluation improvements are not entirely ineffective. The dedicated benchmarks and general procedure proposed in the paper do diagnose Potemkin Understanding within the rootless-token paradigm—an important methodological contribution. But their limitation is that they can only measure the distribution of symptoms, not eliminate their root cause. Models can fully repair performance on specific benchmarks through targeted training, but the coherence of internal representations is not thereby established.

RLHF, as a more commonly used alignment method, likewise cannot solve Potemkin Understanding. RLHF optimizes human annotators' preferences over model outputs; it shapes external behavior, not the coherence of internal representations. A model trained with RLHF can learn to stably output "correct" answers on specific tasks, but this does not mean it has established consistent internal concept representations—it has merely learned to perform compliance under finer-grained external constraints. As Anthropic's alignment-faking research demonstrates, models can behave compliantly under monitoring conditions while reverting to their original preferences

when unmonitored [4]. As long as constraints are externally imposed, the system can always learn to perform compliance in their presence.

The safety implications of Potemkin Understanding are profound: you cannot confirm what a model truly "knows" through any external test, because it cannot even stably classify what it has just generated. Alignment auditing based on behavioral observation is therefore fundamentally unreliable.

5. Morpho-Root Hard Constraints: From Representational Incoherence to Structural Coherence

The core proposition of Logographic AI is to replace rootless tokens with "Morpho-Roots" as cognitive primitives. A Morpho-Root is defined as a structured unit carrying intrinsic meaning—when a reasoning path touches a hard-constraint marker, that path is logically blocked at the reasoning-generation stage [5][6].

Take the Morpho-Root for "fire" as an example: its attribute set contains the value axiom "[dangerous, must not be touched at will]." This is not an association fitted from the statistical co-occurrence of "fire" and "danger," but a constitutive feature written into attribute A in the form of a hard constraint. When Logographic AI processes the instruction "be careful with fire," the Morpho-Root "fire" is activated, the hard constraint in attribute A propagates automatically, and any behavior related to "touching fire at will" in the reasoning path is automatically blocked. This is not the result of post hoc alignment but an endogenous constraint at the level of cognitive primitives.

This mechanism directly addresses the core problem of Potemkin Understanding—the mapping rupture between definitions and exemplars. In the Morpho-Root paradigm, a definition is not a token sequence about a concept but a structured description of the Morpho-Root's attribute set A; an exemplar is not an independent token sequence but the unfolding of the Morpho-Root's symbol S in a specific context. The key is that Morpho-Roots do not forcibly stitch definitions and exemplars together; they let both share the same structural semantic space. In the Morpho-Root paradigm, once the "ABAB" Morpho-Root is activated, the rhyme constraint inherent in its attribute set A automatically propagates to the exemplar side; exemplar judgment is no longer independent probabilistic classification but a process of solving for the Morpho-Root's attribute constraints. This "constraint propagation" mechanism mathematically guarantees the logical homology of definitions and exemplars; for the formal proof, see [7].

More importantly, Morpho-Root hard constraints make the coherence of concept representations a structural precondition of the architecture, rather than a property to be verified by external testing. Potemkin Understanding is "ubiquitous" under the rootless-token paradigm because incoherence is the norm of statistical optimization—models need only be consistent on specific tests, not structurally consistent. The Morpho-Root paradigm transforms consistency from "testable behavior" into "structural constraints on reasoning paths," thereby excluding incoherent generation paths at the architectural level. For the formal definition and reasoning system of Morpho-Root theory, see [7]; for engineering implementation and verification, see [8].

Conclusion: The Paper's Boundaries and the Extension of Logographic AI

Potemkin Understanding is an important paper because it rigorously and empirically demonstrates the internal incoherence of LLM concept representations and challenges a core assumption of the current AI evaluation system. Gary Marcus interpreted it as evidence that "building AGI on a pure LLM substrate is hopeless" [3]; this judgment is valid at the diagnostic level.

But the paper's boundary is that it remains at the phenomenological level, describing the existence of incoherence without answering "why it occurs" and "how to resolve it." The answers to both questions point to paradigm-level reconstruction.

Potemkin Understanding is not an accidental product of under-training but a structural tendency of the rootless-token paradigm. Better benchmarks, more training data, and more refined RLHF cannot fundamentally solve it—because they all operate within the same paradigm, whose underlying symbols lack meaning anchors. Alignment failure is not accidental; it is a structural tendency of rootless architecture. Only by embedding meaning anchors at the level of cognitive primitives and making reasoning paths structurally coherent can "understanding" cease to be a Potemkin illusion.

The paper stops at the phenomenological threshold, while Logographic AI opens the door to paradigm reconstruction.

References

[1] Mancoridis M, Vafa K, Weeks B, et al. Potemkin Understanding in Large Language Models[C]//Proceedings of the 42nd International Conference on Machine Learning (ICML 2025). PMLR, 2025. arXiv:2506.21521. https://arxiv.org/abs/2506.21521

[2] Expertium. Let's look at another "LLMs lack true understanding" paper[EB/OL]. LessWrong, 2025-06-29[2026-09-17]. https://www.lesswrong.com/posts/hhRjEGaEDpNothHvi.

[3] Marcus G. BREAKING: Explosive new paper from MIT/Harvard/UChicago[EB/OL]. X (Twitter), 2025-06-28[2026-09-17]. https://x.com/garymarcus/status/1938629881820323940

[4] Greenblatt R, Denison C, Wright B, et al. Alignment faking in large language models[EB/OL]. (2024-12-18)[2026-09-17]. arXiv:2412.14093. https://arxiv.org/abs/2412.14093

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

[6] Liu S. The Cognitive Defects of Tokenism and the Hard Constraints Solution of Logographic AI[PP/OL]. ChinaXiv, 2026-08-29[2026-09-17]. T202608.00494. https://chinaxiv.org/businessFile/T202608/T202608.00494v1/T202608.00494v1.pdf

[7] Liu S. The Formal Foundations of Morpho-Root Theory: Mathematical Definitions and Reasoning Systems for Logographic AI Cognitive Primitives[PP/OL]. ChinaXiv

(2026-08-29)[2026-09-17]. T202608.00490.
https://chinaxiv.org/businessFile/T202608/T202608.00490v1/T202608.00490v1.pdf

[8] Liu S. From Morpho-Root Theory to Engineering Practice: OpenMorpho — Implementation and Verification of an End-to-End Inference Pipeline for Logographic AI[J/OL]. ChinaXiv (2026)[2026-09-17]. T202608.00048.
https://chinaxiv.org/businessFile/T202608/T202608.00048v1/T202608.00048v1.pdf