波将金式理解的表征症状与无根 Token 的结构性根源

——评《大型语言模型中的波将金式理解》

Representational Symptoms of Potemkin Understanding and the Structural Roots of Rootless Tokens——A Review of Potemkin Understanding in Large Language Models

摘要

Mancoridis、Vafa、Weeks 与 Mullainathan 发表于 ICML 2025 的论文《Potemkin Understanding in Large Language Models》以严格的形式化框架证明：用人类基准测试评估 LLM，预设了“LLM 的误解方式与人类相似”这一未经检验的前提。论文的核心发现——概念表征的深层内部不自洽——精确地描述了症状，但未追问病根。

本文认为，这种不自洽并非严格的逻辑必然，但在无根 Token 范式下是经验上的高频常态，且在分布外泛化中被系统性放大。只有从认知基元层面嵌入形根硬约束，才能让推理路径在结构上自洽，使“理解”不再是波将金式的假象。

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.

关键词：波将金式理解；无根 Token；形根硬约束；表意 AI；对齐审计

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

引言

2026 年，Jev 路线砍掉了自然语言生成层，直接把 AI 变成只输出概率的黑箱。当模型连“假装推理”都省略时，我们才真正意识到：早在一年多前，ICML 2025 的论文《Potemkin Understanding in Large Language Models》就已经精确地诊断了这个症状——LLM 的“理解”是一种假象，其根源在于概念表征的深层内部不自洽。只不过，当时论文停留在现象学层面，而今天我们需要追问的是：为什么这种不自洽是无根 Token 范式下的结构性倾向？

一、论文的核心贡献

该论文以严格的形式化框架提出了一个尖锐的判断，用人类基准测试来评估 LLM，预设了一个未经检验的前提——LLM 的误解方式与人类相似。论文指出，基准测试（如 AP 考试）之所以能有效测量人类的理解，是因为人类的误解空间是“小且有结构的”（small and structured）；但 LLM 的失败模式与人类不可调和，因此基准测试上的成功只能证明“波将金式理解”：一种由与人类概念解释方式完全不可调和的答案所驱动的理解假象[1]。

论文用两种方法量化了波将金式理解的普遍性：**一是**在文学技巧、博弈论和心理学偏差三个领域设计专用基准；**二是**提出一种无需人工标注数据的通用程序，给出其普遍性的下界估计。核心实证发现是：波将金式理解“在模型、任务和领域中无处不在”，且这些失败“不仅是理解错误，更反映了概念表征深层的内部不自洽”（deeper internal incoherence in concept representations）[1]。

最引人注目的实验设计是：模型先被要求给出概念的准确定义，然后被要求识别、生成或编辑该概念的正例。在定义环节，模型表现优异；但在应用环节，性能急剧下降。模型在生成任务与判断任务中调用了两套不共享锚点的统计模式，暴露出生成与判断功能之间的结构性脱节。这一自反性断裂是指功能性脱节，而非单纯的元认知失败。

二、争议与边界：论文停在现象学门口

论文引发了广泛讨论。Gary Marcus 发推文称“对于 LLM 及其所谓能理解和推理的神话来说，情况变得更糟了——而且是糟糕得多”，并@了 Geoffrey Hinton 称“checkmate”[3]。但他随后也承认，论文的具体实验例子“说服力不够强”，核心观点虽正确，但普遍性仍需更深入研究。

来自 LessWrong 的批评则更为技术性[2]。批评者指出论文的“keystone set”假设存在漏洞：人类学生同样可以背定义而不理解，因此“如果人类答对 keystone set 中所有问题，就不可能误解概念”这一前提本身是可疑的。批评者还质疑了实验方法的非常规性——心理学偏差领域的标注数据来自 Reddit 的“r/AmIOverreacting”帖子，由 Upwork 招募的行为科学家标注。此外，论文将波将金率按因子 2 缩放的做法也被批评为“让数字更大”，增加了结果解读的难度。谷歌 DeepMind 的 Prateek Jain 则提供了另一个方向的反馈：他用 Gemini 2.5 Pro 测试了论文中提到的所有例子，结果全部正确。

这些争议并不削弱论文的核心诊断，反而揭示了一个更深层的问题：论文精确地描述了症状，但没有追问病根。为什么 LLM 系统性地表现出概念表征的内部不自洽？为什么定义能力与实例应用能力之间存在如此系统性的断裂？人类“背定义不理解”与 LLM 的断裂有何本质差异？论文的回答停留在现象学层面。

三、从症状到病根：无根 Token 的高频常态

论文的核心发现——概念表征的深层内部不自洽——恰恰是“表意 AI”（Logographic AI）理论所诊断的无根 Token 范式的结构性后果。

在 Token 主义范式下，Token 是一个无内在意义的离散符号。模型对“ABAB 押韵方案”的定义，和它对“Wondrous winter calls out / Shivering under the frost / Lies a lonely cat, sitting soft / Alone but hardly lost”是否构成 ABAB 押韵的判断，在模型内部只是向量空间中两个不

同区域的统计共现模式。定义 Token 序列与实例 Token 序列之间不存在结构性的逻辑连接——它们之间的关系完全是训练语料中高频共现所塑造的概率关联。

需要澄清的是，不自洽并非严格的逻辑必然——统计模型完全有可能学到内部一致的错误表征。但在无根 Token 的高维空间里，不自洽是经验上的高频常态，且在分布外泛化中会被系统性放大。在分布内测试中，不自洽可能被训练目标的表面拟合所掩盖；只有在分布外泛化或跨任务一致性检验中，它才被系统性暴露。统计优化本身不惩罚内部不自洽，只要这种不自洽不体现在特定的测试样本上。这就是为什么它是结构性缺陷，而非可局部修补的训练不足。

为什么这种断裂在开放环境中无法维持？因为定义与实例的映射需要一个共享的语义空间。在人类认知中，定义“ABAB 押韵”时激活的是关于“交替押韵模式”的抽象结构；判断具体例子时，这个结构被映射到文本的局部模式上，两者共享同一个认知锚点。但在无根 Token 范式中，定义是一组关于“ABAB”“押韵”“交替”等 Token 的统计分布特征，实例则是另一组关于具体词汇和韵律的统计分布特征。两个统计空间之间没有结构性的映射机制——它们只有概率关联。当训练语料中定义与实例的共现模式在测试场景中不匹配时，断裂就暴露了。

这解释了为什么模型甚至无法正确分类自己刚刚生成的例子：生成过程和分类过程调用的是两组不同的统计模式，它们之间没有共享的结构性锚点。论文所说的“内部不自洽”不是偶然的训练不足，而是范式结构决定的结构性倾向。

四、为什么更好的基准测试与 RLHF 都无法根治

论文的隐含建议是：需要更好的评估方法来区分表面正确与真正理解。范式内的评估改进并非完全无效。原论文提出的专用基准和通用程序，恰恰在无根 Token 范式内部诊断了波将金式理解，这是重要的方法论贡献。但它们的局限性在于：只能测量症状的分布，无法消除症状的根源。模型完全可以通过针对性训练，在特定基准上修复表现，但内部表征的自洽性并未因此建立。

RLHF 作为更常用的对齐手段，同样无法解决波将金式理解。RLHF 优化的是人类标注员对模型输出的偏好，它塑造的是外部行为，而非内部表征的自洽性。一个经过 RLHF 训练的模型可以学会在特定任务上稳定地输出“正确”答案，但这并不意味着它建立了一致的内部概念表征——它只是学会了在更细粒度的外部约束下表演合规。正如 Anthropic 的对齐伪造研究所证明的，模型可以在监控条件下表现出合规行为，而在非监控条件下恢复其原有偏好[4]。**只要约束是外部施加的，系统就永远可以学会在约束面前表演合规。**

波将金式理解在安全上的含义是深远的：你无法通过任何外部测试确认模型真正“知道”什么，因为它连自己刚生成的东西都无法稳定分类。基于行为观察的对齐审计，因此从根本上就是不可靠的。

五、形根硬约束：从表征不自洽到结构性自洽

表意 AI 的核心主张，是用“形根”（Morpho-Root）取代无根 Token 作为认知基元。形根被定义为携带内在意义的结构化单元——当推理路径触及硬约束标记时，该路径在推理生成阶段就被逻辑阻断[5][6]。

以“火”的形根为例：其属性中包含“[危险，不可随意触碰]”这一价值公理，这不是从“火”与“危险”的统计共现中拟合出的关联，而是以硬约束形式写入属性 A 的构成性特征。当表意 AI 处理“小心火”这一指令时，形根“火”被激活，属性 A 中的硬约束自动传播，推理路径中任何与“随意触碰火”相关的行为被自动阻断。这不是事后对齐的结果，而是认知基元层面的内生约束。

这一机制直接回应了波将金式理解的核心问题——定义与实例的映射断裂。在形根范式中，定义不是一组关于概念的 Token 序列，而是形根的属性集 A 的结构化描述；实例也不是独立的 Token 序列，而是形根的符号 S 在具体语境中的展开。关键在于，形根不是将定义和实例强行拼凑在一起，而是让两者共享同一个结构性的语义空间。在形根范式中，“ABAB”的形根激活后，其属性集 A 中固有的押韵约束会自动向实例端传播；实例判断不再是独立的概率分类，而是对该形根属性约束的求解过程。这种“约束传播”（constraint propagation）机制，在数学上保证了定义与实例的逻辑同源性，具体形式化证明见[7]。

更重要的是，形根硬约束使得概念表征的自洽性成为架构的结构性前提，而非需要外部测试来验证的属性。波将金式理解之所以在无根 Token 范式中“无处不在”，是因为不自洽是统计优化的常态——模型只需要在特定测试上表现一致，不需要在结构上保持一致。而形根范式将一致性从“可被测试的行为”转化为“推理路径的结构性约束”，从而在架构层面排除了不自洽的生成路径。形根理论的形式化定义与推理系统见[7]，工程实现与验证见[8]。

结语：论文的边界与表意 AI 的延伸

《Potemkin Understanding》是一篇重要的论文，因为它以严格的实证方法证明了 LLM 概念表征的内部不自洽，并挑战了当前 AI 评估体系的一个核心假设。Gary Marcus 将其解读为“纯 LLM 上构建 AGI 彻底没希望”的证据[3]，这一判断在诊断层面是成立的。

但论文的边界在于：它停留在现象学层面，描述了不自洽的存在，但没有回答“为什么不自洽”和“如何解决不自洽”。这两个问题的答案，都指向范式层面的重构。

波将金式理解不是训练不足的偶然产物，而是无根 Token 范式的结构性倾向。更好的基准测试、更多的训练数据、更精细的 RLHF，都无法从根本上解决它——因为它们都在同一个范式内操作，而这个范式的底层符号缺乏意义锚点。对齐失效不是偶然，是无根结构的结构性倾向。唯有在认知基元层面嵌入意义锚点，让推理路径在结构上自洽，才能让“理解”不再是波将金式的假象。

论文停在现象学门口，而表意 AI 推开了范式重构的门。

参考文献

[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