LLMs as boundary phenomena: A comment on Nefdt (2026)

Brett Reynolds[*]
Humber Polytechnic & University of Toronto

12th February 2026

Abstract

The debate over whether large language models "really think" reproduces a familiar pattern: a boundary case meets a binary classification. Nefdt (2026) organizes this debate through a 2×2 matrix crossing language with cognition, placing LLMs in a "missing quadrant" of language without cognition. But the binary frame undersells his own insights. His hedging on proxies, partial capacities, and qualified attributions points toward something the table can't express: that language and cognition are cluster concepts with graded membership. Homeostatic property cluster (HPC) theory, with its projection-purpose analysis, offers a better framework. Cognitive predicates serve purposes that converge for core cases but diverge at boundaries – the same phenomenon that makes a tomato a vegetable or a fruit depending on whether you're a greengrocer or a botanist. The debate persists because the purpose divergence is invisible in the core case, and resists resolution because disputants treat cluster kinds as if they had essences. But the categories themselves aren't fixed: LLMs are reshaping what "cognition" means, not just testing whether they have it.

1 Introduction

Everyone has an opinion about whether ChatGPT "really thinks". The debate reproduces a pattern familiar from philosophy of science: a boundary case meets a binary classification, and the deadlock gets mistaken for a factual disagreement.

Nefdt (2026) offers a sophisticated analysis. He frames the conceptual landscape through a 2×2 matrix (his Table 1) crossing language with cognition. Humans have both. Non-human animals have cognition without language. LLMs, Nefdt argues, fill the "missing quadrant": language without cognition (p. 4). They're "purely linguistic agents unplugged from integration with both larger cognitive structure and the world in which it evolved" (p. 12).

The table carves out genuine conceptual space and steers between the eliminativism of Bender and Koller (2020) and the maximalism of Cappelen and Dever (2025). But its binary structure undersells Nefdt's own insights. His discussion is full of hedging, proxies, and qualified answers – scalar claims that a binary frame can't capture.

[*] Contact: brett.reynolds@humber.ca

This note argues that HOMEOSTATIC PROPERTY CLUSTER (HPC) theory (Boyd, 1991, 1999), with its projection-purpose analysis, handles the LLM case better. HPC theory deals with kinds defined not by necessary and sufficient conditions but by clusters of co-occurring properties held together by causal mechanisms. Which properties matter depends on the projection purpose – what the categorization is for. This explains why the debate is so persistent, why its disagreements resist resolution, and why the categories at stake are themselves in flux. Nefdt's hedging is evidence for the framework, not a weakness of his analysis.

2 The binary and its discontents

Nefdt's Table 1 is a 2×2 matrix crossing language with cognition:

	+Cognition	–Cognition
+Language	Humans	???
–Language	Non-human animals	Rocks

Table 1: Conceptual possibilities, adapted from Nefdt (2026), p. 4.

The open question is whether anything occupies the top-right cell: language without cognition. Nefdt's answer is that LLMs fill it. They "occupy a hitherto vacant part of conceptual space" (p. 4). The table presupposes that LANGUAGE and COGNITION are necessary-and-sufficient-condition categories. Something either has language or it doesn't, either has cognition or it doesn't, and the philosophical work consists in deciding which cell a system occupies. One might object that the table is scaffolding, not a metaphysical commitment – a heuristic to organize the debate, not a claim that the categories are really binary. But even heuristic binaries constrain the conceptual space they organize: a table with two columns invites two-valued answers. The hedging that runs through Nefdt's discussion is his own analysis pressing against that constraint.

Consider the evidence. His NO BRAINER principle states that "LLMs only model one aspect of cognition, namely (statistical) linguistic processing" (p. 6). This already blurs the distinction between columns: if statistical linguistic processing is an "aspect of cognition", then LLMs don't simply lack cognition. They instantiate part of it.

COGNITION UNPLUGGED goes further, drawing on Casto et al.'s (2025) distinction between "linguistic understanding" and "deep understanding". Nefdt concludes that purely linguistic agents have "statistically-based proxies for more cognitively loaded states" (p. 8). But proxies aren't absences. This isn't just epistemic hedging (we're uncertain which cell LLMs belong in). It's evidence of ontological continuity: a proxy for reasoning is on the same continuum as reasoning, better explained by continuity than by binary uncertainty.

The hedging continues throughout section 4. On perspective: LLMs can be "trained to execute a particular point of view" (p. 9), but "the individual phenomenal level is missing" (p. 10). On time: they don't clearly have or lack temporal cognition; rather, "they could depending on the kinds of structures they employ" (p. 12). On cognitive agency generally: Nefdt titles his section 4.3 "Why I'm neither a realist nor an eliminativist" and locates himself in "a nonempty position in between" the two poles (p. 12) – a position his own table has no cell for.

These aren't binary verdicts. They're positions along a continuum. The properties Nefdt distributes between "language" and "cognition" (inferential reasoning, perspective-taking, temporal processing, phenomenal experience) don't sort cleanly into two groups. They cluster, with graded and contested membership at the margins. As Mahowald et al. (2024) show, language and thought "dissociate" in LLMs, but the dissociation is partial and uneven, not a clean binary divide. Partial and uneven dissociations are exactly what HPC theory was developed to handle.

3 HPC reframing

HPC theory was developed by Boyd (1991, 1999) to handle natural kinds that resist definition by necessary and sufficient conditions. The framework is contested (see, e.g., Magnus 2014 for objections), but the core insight – that kind membership can be graded and mechanism-dependent – is widely shared even among critics. Biological species are the paradigm case, and ring species illustrate it vividly. In the *Ensatina* salamander complex of western North America (Wake, 1997), adjacent populations share morphology, colouration, and reproductive compatibility, but these properties grade continuously around the ring: where the endpoints meet in southern California, the populations can't interbreed despite geographic overlap. There's no point where one species stops and another begins. The cluster holds, but membership is irreducibly graded.

The framework applies equally to language and cognition. In humans, properties like inferential reasoning, pragmatic implicature, analogical mapping, perspective-taking, planning, emotional response, and embodied experience cluster together. They do so because of homeostatic mechanisms: shared neural architecture, developmental trajectories, social interaction patterns, and embodied engagement with the world.

But the clustering is contingent, not definitional. The properties co-occur in humans because of the particular causal structure of human biology and development, and LLMs instantiate a novel combination from this cluster. They display some properties typically associated with cognition (inferential reasoning, analogical mapping, something like perspective-taking) while lacking others (embodied experience, persistent memory, emotional states, autonomous goal-formation). This doesn't make them a clean occupant of a "missing quadrant." It makes them a graded member of the cluster: exactly the kind of entity HPC theory was designed to handle and that binary categories systematically misclassify.

Powell (2020) distinguishes convergent from contingent properties, which helps clarify what to expect in a novel system. Convergent properties recur across unrelated systems because similar functional pressures produce similar solutions. Pattern extraction and inferential reasoning may be convergent: any system under sufficient pressure to predict and generate natural language is likely to develop something functionally similar. Contingent properties depend on specific implementation history. The absence of embodiment in LLMs is a contingent feature of their engineering, not a deficit. The absence of phenomenal consciousness may be contingent too, or it may reflect deeper architectural constraints – an empirical question, not a definitional one. Either way, HPC theory predicts what we observe: a system that shares some cluster properties and lacks others, with the specific combination shaped by causal mechanisms rather than by a definition.

The "homeostatic" in HPC does real work. It requires not just co-occurring properties but causal mechanisms that maintain the clustering through feedback. Boyd's framework targets natural kinds shaped by gene flow and developmental constraint (see Khalidi 2013 on etiological kinds more generally). In LLMs, the candidate mechanism is the training process itself: gradient descent against a loss landscape shaped by natural-language statistics. Representational convergence supports this: independently trained models across different architectures and modalities arrive at similar internal representations (Huh et al., 2024), suggesting the property profile is constrained by the task, not stipulated by engineers. The mechanistic story is incomplete, as it was when Boyd first applied HPC theory to biological species.

4 The tomato move

The cluster structure is only half the HPC story. The framework also asks: which properties matter, and for whom? Even granting that LLMs occupy some definite cell, we'd need to ask, definite relative to what?

To a greengrocer, a tomato is a vegetable: it's savoury, shelved with the peppers and onions. To a botanist, it's a fruit: it develops from the ovary of a flower and contains seeds. Neither classification is wrong. Each serves a different PROJECTION PURPOSE: the interest or analytical goal that determines which similarities count and therefore what falls inside the category. The disagreement dissolves once you specify which purpose you're serving. But perspectival doesn't mean inconsequential. The US Supreme Court ruled in *Nix v. Hedden* (1893) that tomatoes are vegetables for tariff purposes – an ontological question settled, with characteristic confidence, by a customs schedule. Whether LLMs "really think" has analogous consequences for regulation, liability, and intellectual-property law.

In Goodman's (1955) terms, the issue is about projectibility: not all predicates project equally well to new cases. *Green* projects from observed emeralds to unobserved ones; *grue* doesn't. A predicate is projectible when it lets you predict further properties of new instances. When cognitive scientists apply false-belief tasks to LLMs (Kosinski, 2024) – the same paradigm developed for chimpanzees (Premack & Woodruff, 1978) and children (Wimmer & Perner, 1983) – they're projecting *cognitive*, and it predicts some capacities well. When Kallini et al. (2024) find that LLMs struggle with impossible languages but handle natural ones readily, they're projecting *linguistic*, and it predicts others. For LLMs, the question isn't which predicate applies but which one projects usefully, and for whom. Each analytical perspective yields a different answer.

Under a neuroscience projection, we ask what mechanism produces the behaviour. LLMs lack the integration properties that hold the cognition cluster together in biological systems: sensorimotor loops, persistent memory, autonomous goal-formation. By this criterion, their capacities are linguistic: produced without the broader cognitive architecture, however sophisticated the computation.

Under a functional projection, we ask what the system does. LLMs draw inferences, construct analogies, adopt perspectives, and plan multi-step solutions. By this criterion, their capacities are cognitive: they exhibit the functional profile of cognition regardless of the underlying mechanism.

Under a phenomenological projection, we ask whether there is something it is like to be the system (Nagel, 1974). The answer for LLMs is not obviously yes – and may not be decidable from the outside. By this criterion, the question of cognition can't be resolved.

Same system, same capacities, three verdicts.

Nefdt's Table 1 implicitly places more weight on what produces the behaviour than on what the behaviour achieves. LLMs lack the neural substrates and embodied integration associated with cognition in humans, so they go in the "language without cognition" cell. But this is a consequence of the projection chosen, not a discovery about LLMs. Under the functional projection, the same systems would land in "language and cognition."

The persistence of the "do LLMs really think?" debate is predicted by projection-purpose analysis. It's a dispute about which predicate to project, disguised as an ontological one. Resolving it doesn't require discovering a hidden fact about LLMs. It requires specifying the purpose of the categorization.

5 Why this matters

Nefdt's analysis is better served by the HPC framework than by his table. His hedging, his "neither realist nor eliminativist" stance, his acknowledgment of proxies and partial capacities – these are exactly what HPC theory predicts for boundary cases. The framework doesn't force him to choose a cell. It lets him say what he already wants to say: that LLMs share some cluster properties with cognitive agents, lack others, and that the combination is genuinely novel. The binary table does have one virtue: it forces a commitment. HPC's flexibility is also its risk, since a framework that accommodates everything explains nothing. The claim here is narrower: for *this* boundary case, the cluster structure is more informative than the binary.

But the framework also diagnoses two specific patterns in the debate. The first explains why the debate persists; the second, why the disagreements within it resist resolution.

The first pattern is projection mismatch. Cognitive predicates like *believes* and *reasons* serve multiple purposes that normally converge. Under one, a predicate projects well when it accurately predicts the cluster of associated properties: calling someone a believer predicts contextual stability, responsiveness to reasons and integration with action. Under another, it projects well when it provides the right tools for interaction: treating someone as a believer lets you coordinate expectations and hold them accountable. For humans, both purposes agree, so people slide between them without noticing. At the LLM boundary, they come apart. Section 4 showed this at the analytical level, with neuroscience, functional, and phenomenological purposes each placing LLMs differently. The same divergence operates at the level of vocabulary choice. Shanahan (2024) foregrounds predictive accuracy: cognitive predicates carry implications that don't transfer to LLMs, so they risk anthropomorphism and mislead about what to expect. Cappelen and Dever (2025) foreground productive engagement: without cognitive predicates, we lack the tools to figure out LLMs' place in our social structures. The observations can largely be shared; what differs is the purpose the predicates serve. Once the purposes are named, the apparent contradiction dissolves into two defensible answers to two different questions.

The second pattern is essentialism about the categories themselves. Bender and Koller (2020) and Piantadosi and Hill (2022) both ask whether LLMs understand language – a shared question under a shared functional projection. But each treats a different property of the meaning cluster as criterial. Bender and Koller (2020, p. 5187) take meaning to require a relation between linguistic form and communicative intent, so world-directed grounding becomes the essential property: LLMs lack it, so they don't understand language. Piantadosi and Hill (2022) treat meaning as conceptual role constituted by relations among internal representational states, so relational coherence becomes the essential property: LLMs have it, so they do. Each account selects one property from the cluster and elevates it to a necessary condition – exactly the move HPC theory diagnoses as an error for cluster kinds. Cluster kinds channel substantive disagreements this way: when multiple properties are available, a prior view about what matters gets expressed as a claim about what the kind requires. If meaning is a homeostatic property cluster, no single property is definitional; the properties co-occur contingently, held together by mechanisms, and different systems can instantiate different subsets. The two camps disagree about everything except the assumption that guarantees the disagreement: that one property of the meaning cluster must be definitional. An essentialist frame makes a perspectival choice look like a factual mistake.

HPC kinds aren't static. They drift when their environment changes, as biological species do under ecological pressure. LLMs have changed the environment in which cognitive categories operate. Before LLMs, the properties in the cognition cluster co-occurred so reliably in humans that fine-grained distinctions among them – *thinks* vs. *processes language*, *understands* vs. *pattern-matches* – were of little practical consequence. Now the distinctions are urgent: regulators, journalists, and courts need them. The debate over LLM cognition isn't just about categorizing a novel entity within existing kinds. The entity is reshaping the kinds. The homeostatic mechanisms that held the cognition cluster together – the reliable co-occurrence of inference, understanding, grounding, and phenomenal experience in embodied agents – are under new selection pressure, and the cluster is reorganizing in real time.

The question isn't whether LLMs "really" have cognition. It's which properties cluster, under what mechanisms, for what analytical purpose. That question is itself in motion: LLMs are reshaping the conceptual ecology they've entered. HPC theory and projection-purpose analysis are designed for exactly this kind of moving target, and they have more interesting answers than yes or no.

Acknowledgements

This note was drafted with the assistance of Claude Code (Opus 4.6). I have reviewed and revised all content and take full responsibility for the final text.

References

Bender, E. M., & Koller, A. (2020). Climbing towards NLU: On meaning, form, and understanding in the age of data. In D. Jurafsky, J. Chai, N. Schluter & J. Tetreault (Eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 5185–5198). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.463

Boyd, R. (1991). Realism, anti-foundationalism and the enthusiasm for natural kinds. *Philosophical Studies*, *61*, 127–148. https://doi.org/10.1007/BF00385837

Boyd, R. (1999). Homeostasis, species, and higher taxa. In R. A. Wilson (Ed.), *Species: New interdisciplinary essays* (pp. 141–185). MIT Press. https://doi.org/10.7551/mitpress/6396.003.0012

Cappelen, H., & Dever, J. (2025). *Going whole hog: A philosophical defense of AI cognition* [arXiv preprint]. https://arxiv.org/abs/2504.13988

Casto, C., Ivanova, A., Fedorenko, E., & Kanwisher, N. (2025). *What does it mean to understand language?* [arXiv preprint]. https://arxiv.org/abs/2511.19757

Goodman, N. (1955). *Fact, fiction, and forecast*. Harvard University Press.

Huh, M., Cheung, B., Wang, T., & Isola, P. (2024). Position: The Platonic representation hypothesis [arXiv:2405.07987]. *Proceedings of the 41st International Conference on Machine Learning*, 20617–20642.

Kallini, J., Papadimitriou, I., Futrell, R., Mahowald, K., & Potts, C. (2024). Mission: Impossible language models. In L.-W. Ku, A. Martins & V. Srikumar (Eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (volume 1: Long papers)* (pp. 14691–14714). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.acl-long.787

Khalidi, M. A. (2013). *Natural categories and human kinds: Classification in the natural and social sciences*. Cambridge University Press. https://doi.org/10.1017/CBO9780511998553

Kosinski, M. (2024). Evaluating large language models in theory of mind tasks. *Proceedings of the National Academy of Sciences*, *121*, e2405460121. https://doi.org/10.1073/pnas.2405460121

Magnus, P. D. (2014). NK≠HPC. *The Philosophical Quarterly*, *64*(256), 471–477. https://doi.org/10.1093/pq/pqu010

Mahowald, K., Ivanova, A. A., Blank, I. A., Kanwisher, N., Tenenbaum, J. B., & Fedorenko, E. (2024). Dissociating language and thought in large language models. *Trends in Cognitive Sciences*, *28*(6), 517–540. https://doi.org/10.1016/j.tics.2024.01.011

Nagel, T. (1974). What is it like to be a bat? *Philosophical Review*, *83*(4), 435–450. https://doi.org/10.2307/2183914

Nefdt, R. M. (2026). *What it's like to be an LLM* [Manuscript, University of Cape Town / University of Bristol]. https://philpapers.org/rec/NEFWIL

Piantadosi, S. T., & Hill, F. (2022). *Meaning without reference in large language models* [arXiv preprint]. https://arxiv.org/abs/2208.02957

Powell, R. (2020). *Contingency and convergence: Toward a cosmic biology of body and mind*. MIT Press.

Premack, D., & Woodruff, G. (1978). Does the chimpanzee have a theory of mind? *Behavioral and Brain Sciences*, *1*(4), 515–526. https://doi.org/10.1017/S0140525X00076512

Shanahan, M. (2024). Talking about large language models. *Communications of the ACM*, *67*(2), 68–79. https://doi.org/10.1145/3624724

Wake, D. B. (1997). Incipient species formation in salamanders of the *Ensatina* complex. *Proceedings of the National Academy of Sciences*, *94*(15), 7761–7767. https://doi.org/10.1073/pnas.94.15.7761

Wimmer, H., & Perner, J. (1983). Beliefs about beliefs: Representation and constraining function of wrong beliefs in young children's understanding of deception. *Cognition*, *13*(1), 103–128. https://doi.org/10.1016/0010-0277(83)90004-5