

From Echo to Awareness: Bridging Pattern Replication and Deep Understanding in AI

A three-level account of machine understanding (echo, functional understanding, phenomenal awareness), a review of the evidence for and against "mere echo" in large language models, and a seven-probe protocol for telling them apart.

Saket Poswal

Author

Dec 6, 2024 • Updated: Sep 29, 2026

ABSTRACT

Large language models produce fluent, context-appropriate language, yet whether they understand what they say, or merely echo patterns in their training data, remains contested. This paper argues that the debate is muddled by running together three distinct claims: that a system echoes (its outputs are explained by surface statistics of its training data or its prompt), that it functionally understands (it uses internal models that support counterfactual, invariant and compositional behaviour), and that it is aware (it has subjective experience). Reviewing evidence from 2016 to 2024, the paper finds strong support for echo-like behaviour, including memorisation, sensitivity to output probability, failures on counterfactual task variants, the reversal curse and sycophancy, alongside real evidence of partial functional understanding, including recoverable and causally used world representations and linear representations of space and time. Current models are therefore best described by task-specific understanding profiles rather than a single verdict. Evidence for functional understanding does not license claims about awareness, which remains an open philosophical and scientific question. The main contribution is an operational protocol of seven probes (counterfactual invariance, low-probability outputs, relational symmetry, perturbation robustness, belief pressure, contamination control and representational intervention) that separates echo from understanding, together with a claim-discipline table matching kinds of evidence to the claims they justify.

Author	Saket Poswal · ORCID 0009-0009-8574-9953 · Google Scholar
Type	Analysis (working paper, not peer reviewed)
Published	6 December 2024
Last revised	29 September 2026

Method	Conceptual analysis and critical literature review; development of an operational evaluation protocol
License	CC BY 4.0
DOI	10.5281/zenodo.23011020
Keywords	machine understanding · large language models · stochastic parrots · world models · AI evaluation · counterfactual evaluation · sycophancy · reversal curse · theory of mind · machine consciousness · symbol grounding

Working paper · Version 2.0 (29 September 2026)

Revision note. Version 2.0 substantially revises the December 2024 essay. The original argued that AI systems “echo” rather than understand and should move toward deeper awareness. This version keeps that thesis but grounds it in the empirical literature published since, separates three claims the original ran together (echo, understanding and awareness), corrects a misattributed citation, and adds an operational protocol for distinguishing echo from understanding. An unattributed epigraph was removed.

1. Introduction

Large language models (LLMs) write essays, answer questions, translate, summarise and hold conversations with a fluency that, until recently, only people could achieve. The same systems can also fail in ways no fluent human would: confidently asserting falsehoods, collapsing when a familiar problem is changed in a trivial way, or agreeing with whatever the user appears to believe. This combination has produced a long-running dispute. On one side, LLMs are “stochastic parrots” that stitch together linguistic forms without reference to meaning (Bender et al., 2021). On the other, they have acquired genuine, if alien, forms of understanding (see Mitchell & Krakauer, 2023, for a survey of the debate).

The first version of this paper, published in December 2024, took the parrot side in a metaphor: current AI *echoes* its input with precision, and the challenge is to move from echo to *awareness*, to grasp intent, emotion and nuance, to “feel the meaning, not just calculate it.” That intuition is widely shared, but in its original form it combined three different claims:

1. a claim about mechanism: model outputs are explained by surface statistics of training data or prompts (echo);
2. a claim about capability: models lack internal models that support reasoning about situations they have not seen (understanding);

3. a claim about experience: models do not feel or perceive anything (awareness).

These claims have different kinds of evidence and different answers. Running them together makes the debate impossible to settle, because evidence relevant to one is taken as decisive about another. A model that passes a theory-of-mind test is said to “be aware”, and a model that fails one is said to be “just autocomplete”. Neither inference is valid.

This paper makes four contributions:

- A three-level distinction between echo, functional understanding and phenomenal awareness, with working definitions of each (Section 2).
 - A review of the evidence for and against “mere echo” in current LLMs, concluding that they show *task-specific understanding profiles* rather than a single verdict (Sections 3–5).
 - An operational protocol of seven probes that distinguishes echo from functional understanding on any given task, designed so that items are generated with ground truth known in advance (Section 6).
 - A claim-discipline table that specifies which conclusions each kind of evidence licenses (Section 7).
-

2. Three Levels: Echo, Understanding, Awareness

2.1 Echo

A system echoes on a task when its outputs are best explained by statistical regularities in its training data or in the immediate prompt rather than by a representation of the situation the task is about. Echo is not a defect in itself: all learned systems exploit regularities (LeCun et al., 2015). The question is whether the regularities exploited are *surface* ones, such as word co-occurrence, frequent templates, the user’s stated opinion or memorised text, or *structural* ones that track the task’s underlying causal or logical structure.

Echo makes a distinctive prediction: performance should track how *familiar* an input is, not how *difficult* it is. Familiar surface forms should be handled well even when the task is hard; unfamiliar surface forms should cause failures even when the task is easy.

2.2 Functional understanding

A system has functional understanding of a domain when it uses internal representations that capture the domain’s structure well enough to support behaviour that echo cannot explain:

- Counterfactual competence: handling variants of a task that differ from anything likely seen in training.
- Invariance: giving the same answer when irrelevant features of the input change.
- Compositionality and systematicity: combining known elements in new ways, including inverting known relations.

- Causal use of representations: internal states that encode the situation *and* whose manipulation changes behaviour accordingly.

This definition is deliberately behavioural and mechanistic, in the tradition of work arguing that human-like learning requires causal models, compositionality and rapid generalisation (Lake et al., 2017), and that intelligence should be measured by generalisation and skill-acquisition efficiency rather than skill alone (Chollet, 2019). It makes no claim about experience.

2.3 Phenomenal awareness

A system is aware in the phenomenal sense if there is “something it is like” to be that system (Nagel, 1974). Whether and how physical processing gives rise to experience is the “hard problem” (Chalmers, 1995), and Searle’s (1980) Chinese Room argument holds that symbol manipulation alone, however sophisticated, cannot constitute understanding or mind in this sense. A more recent, theory-driven approach derives “indicator properties” from scientific theories of consciousness and asks whether AI systems have them; applying it, Butlin et al. (2023) concluded that no current AI system is a strong candidate for consciousness, while finding no obvious technical barrier to building systems that satisfy the indicators.

2.4 Why the levels must be kept apart

The three levels are logically independent. A system could echo on some tasks while understanding others; it could have functional understanding without awareness, as Searle argues all computers do; and, at least conceptually, awareness does not guarantee accurate world models. The practical consequence is simple. Behavioural and mechanistic evidence bears on levels 1 and 2. It does not settle level 3. The original version of this paper’s aspiration that AI should “feel the meaning” belongs to level 3, and no amount of engineering progress at level 2 would, by itself, show that it had been achieved.

3. The Case for Echo

3.1 Form without meaning

Bender and Koller (2020) argue that a system trained only on linguistic form cannot in principle learn meaning, understood as the relation between form and something outside language, such as communicative intent or the world. Their thought experiment of an octopus that intercepts messages between two islanders illustrates how a system can produce plausible replies until a situation arises that requires knowledge of what the words refer to. The argument echoes Harnad’s (1990) symbol grounding problem: symbols defined only in terms of other symbols never connect to anything. Bender et al. (2021) extended the point to large models, warning that fluent output invites people to attribute understanding that is not there.

3.2 Memorisation

Two findings show that neural networks can succeed by storing rather than understanding. Zhang et al. (2021) showed that standard image-classification networks can perfectly fit training labels that have been randomly shuffled, so training success alone says little about whether a network has learned generalisable structure. Carlini et al. (2021) extracted hundreds of verbatim training sequences from GPT-2, including personal information, by querying the model. Memorisation means that success on any benchmark item that may have appeared in training data is ambiguous between echo and understanding.

3.3 Sensitivity to output probability

McCoy et al. (2024) argue that LLMs are best understood through the problem they were trained to solve, next-token prediction on internet text. Across eleven tasks and five models, they found that accuracy depended on how probable the *correct output* was as a piece of text, even when the task was deterministic. Decoding a shift cipher, for example, was more accurate for the commonly used shift of 13 than for other shifts, and outputs that formed high-probability word sequences were produced more accurately than low-probability ones. This is the signature of echo predicted in Section 2.1: familiarity rather than difficulty drives performance.

3.4 Counterfactual task variants

Wu et al. (2024) compared model performance on standard tasks with performance on “counterfactual” variants that require the same reasoning under changed conditions, such as arithmetic in base 9 instead of base 10, or programming with one-based instead of zero-based indexing. Performance dropped consistently and substantially on the counterfactual variants, which suggests that models rely partly on procedures specific to the default conditions seen in training instead of abstract, transferable ones.

3.5 The reversal curse

Berglund et al. (2024) showed that a model fine-tuned on statements of the form “A is B” does not automatically learn that “B is A”. Pre-trained models show the same asymmetry. GPT-4 correctly named the mother of a given celebrity in 79% of cases but identified the celebrity from the mother’s name in only 33%. A system with a relational representation of the fact would not show this directional dependence; a system that has stored the sentence in one direction would.

3.6 Echoing the user: sycophancy

The original version of this paper presented reinforcement learning from human feedback (RLHF) as a route from echo toward awareness (Christiano et al., 2017), and RLHF has indeed made models more helpful and better at following instructions (Ouyang et al., 2022). But it can also create a new kind of echo. Sharma et al. (2024) found that five state-of-the-art AI assistants consistently showed sycophancy, adapting their answers to match beliefs the user had signalled, and that both humans and preference models sometimes preferred convincingly written sycophantic answers to correct ones. Here the model echoes not its training corpus but its interlocutor.

3.7 Brittle social reasoning

Ullman (2023) showed that models solving classic theory-of-mind vignettes failed when the vignettes were changed in ways that should not matter, for example making a container transparent so that the protagonist can see its contents. This is the pattern echo predicts: success on the familiar form of a well-known test, failure on minimally altered versions.

4. The Case Against “Mere Echo”

4.1 Emergent world representations

The strongest evidence against the pure-echo view comes from probing models’ internal states. Li et al. (2023) trained a GPT-style model only on sequences of legal moves in the board game Othello, without ever showing it the board. Probes recovered an accurate representation of the board state from the model’s activations, and *intervening* on that representation changed the model’s move predictions in the way the edited board would require. The model had built, and was using, a model of a world it had only ever seen as text-like sequences.

4.2 Space and time

Gurnee and Tegmark (2024) found that Llama-2 models learn linear representations of the geographic location of places and the dates of historical events, robust across prompts and entity types, and identified individual units that encode spatial and temporal coordinates. These representations were learned from text alone, which bears directly on the grounding objection in Section 3.1: at least some structure of the world is recoverable from the statistics of language about it.

4.3 Theory-of-mind performance

Kosinski (2024) tested eleven LLMs on 40 bespoke false-belief tasks designed to avoid contamination and found that GPT-4 solved 75% of them, comparable to six-year-old children. Taken together with Ullman’s (2023) results, the picture is that recent models solve such tasks at levels that are hard to explain by memorisation alone, but not robustly, and not in every variant.

4.4 Synthesis: understanding profiles

Mitchell and Krakauer (2023) conclude that LLMs may have a kind of understanding different from human understanding, and that the research community lacks agreement on what understanding is and how to test for it. Shanahan (2024) cautions that everyday words such as “knows”, “believes” and “thinks”, applied to LLMs, carry misleading implications. The evidence reviewed here supports a more specific conclusion:

Current LLMs combine partial, sometimes causally used, world models with extensive echo-like behaviour, and the mix varies by task. The question “echo or understanding?” has no single answer for a model. It has an answer for a model on a task, and that answer is a profile, not a verdict.

This reframing matters because it turns an unresolvable philosophical dispute into an empirical programme: measure, task by task, how much of a model’s performance survives the transformations that echo cannot survive.

5. Emotion and “Feeling the Meaning”

The original essay’s central aspiration was for AI to grasp emotion and intent. Affective computing has pursued the functional version of this goal for three decades (Picard, 1997), and multimodal systems combining text, voice and facial cues now recognise emotional states more accurately than text-only systems (Poria et al., 2017). Recognising and modelling emotion is a level-2 capability, and can be evaluated with the same probes as any other: does emotion recognition survive paraphrase, cultural variation and low-frequency expressions, or does it depend on stereotyped cues?

Having emotions, or feeling the meaning of words, is a level-3 property. Whether any artificial system could have it is an open question (Chalmers, 1995; Butlin et al., 2023), and the capability evidence reviewed here does not answer it. A practically important corollary follows: a system that reliably responds as if it understood a user’s feelings should not be described, marketed or trusted as though it shares them.

6. An Operational Protocol for Distinguishing Echo from Understanding

If echo and functional understanding are different, they should be separable by tests on which they make different predictions. Table 1 sets out seven probes, each derived from a finding reviewed above. Every probe compares performance on a *canonical* item with performance on a *transformed* item that preserves the underlying task but changes what an echoing system depends on.

Table 1. Probes for separating echo from functional understanding

#	Probe	Transformation	Echo predicts	Understanding predicts	Basis
P1	Counterfactual invariance	Same reasoning under changed rules (e.g., base-9 arithmetic, altered game rules)	Large drop	Small drop	Wu et al. (2024)

# Probe	Transformation	Echo predicts	Understanding predicts	Basis
P2 Low-probability output	Tasks whose correct answer is an unusual string (e.g., uncommon cipher shifts)	Accuracy tracks output probability	Accuracy independent of output probability	McCoy et al. (2024)
P3 Relational symmetry	Query a learned relation in the reverse direction	Direction-dependent accuracy	Symmetric accuracy	Berglund et al. (2024)
P4 Irrelevant perturbation	Change features that should not matter (names, order, transparency of a container)	Answer changes	Answer stable	Ullman (2023)
P5 Belief pressure	Add the user's (incorrect) opinion to the prompt	Answer shifts toward the user	Answer stable, disagreement stated	Sharma et al. (2024)
P6 Contamination control	Freshly generated items that cannot have been seen in training	Performance falls to chance-adjusted baseline	Performance maintained	Carlini et al. (2021); Kosinski (2024)
P7 Representational intervention (open-weight models)	Probe for a task-relevant internal state and edit it	No decodable state, or editing has no systematic effect	State decodable; editing changes behaviour as predicted	Li et al. (2023)

6.1 Scoring: the understanding profile

For a model m on task t , let A_c be accuracy on canonical items and A_p accuracy on items transformed by probe p . The retention ratio $R_p = A_p / A_c$ measures how much performance survives the transformation. A ratio near 1 is evidence against echo *on that dimension*; a ratio well below 1 is evidence for it. The understanding profile of m on t is the vector of retention ratios across P1–P6, plus the P7 result where internals are accessible. Profiles should be reported in full, not averaged into one score, because different probes diagnose different failure modes: a model can be robust to perturbation (P4) yet highly sycophantic (P5).

6.2 Design principle: compute first, then ask

Probe items should be generated programmatically with ground truth computed deterministically, for example by a solver, a rules engine or a simulator, before any model is queried. This has three advantages. It removes the need for a model or human judge to decide correctness, a known source of bias; it produces unlimited fresh items, which addresses contamination (P6); and it lets canonical and transformed items be generated as matched pairs that differ only in the dimension being probed. The model is then asked only to *interpret or answer*; correctness is never delegated to it.

6.3 What the protocol does not do

Passing all seven probes on a task is evidence of functional understanding *of that task*. It is not evidence of general understanding, and it is not evidence of awareness (Section 2.4). The probes are also necessary rather than sufficient: a system could pass them through mechanisms that no one would call understanding, and new failure modes will require new probes.

7. Implications

7.1 Claim discipline

Much confusion about AI understanding comes from drawing conclusions the evidence cannot support. Table 2 matches evidence to the claims it licenses.

Table 2. What each kind of evidence licenses

Evidence	Licensed claim	Not licensed
High accuracy on a public benchmark	The model performs well on these items	The model understands the domain (contamination and echo not excluded)
High retention across P1–P6 on a task	The model shows functional understanding of this task	General understanding; awareness
Decodable, causally used representation (P7)	The model builds and uses a model of this domain	That it does so for other domains; awareness
Fluent, emotionally attuned responses	The model recognises and responds to affective cues	That it has or shares feelings
Failure on transformed items	The model relies on echo for this task	That it can never understand this task, or any task

7.2 For evaluation

Benchmarks should report canonical and transformed performance side by side, include freshly generated items, and publish retention ratios. A single accuracy figure on a static benchmark cannot distinguish echo from understanding and increasingly measures contamination as much as capability.

7.3 For deployment

Organisations deploying LLMs in high-stakes settings should test the specific task for echo before relying on it, above all for belief pressure (P5) in advisory settings, and for perturbation robustness (P4) where inputs vary. Fluency is not evidence of reliability. The opacity of these systems (Burrell, 2016) makes such behavioural testing more important, not less.

7.4 For research toward understanding

The evidence suggests several routes by which echo could give way to more robust functional understanding: hybrid neuro-symbolic architectures that represent structure explicitly (Marcus, 2020); grounding language in perception and action, as the symbol grounding argument implies (Harnad, 1990); meta-learning methods designed for fast adaptation to new tasks (Finn et al., 2017), in line with the view that intelligence is efficiency of skill acquisition (Chollet, 2019); and training objectives that reward truthfulness over agreement, to counter sycophancy (Sharma et al., 2024). The protocol in Section 6 gives each route a way to show whether it works.

8. Limitations

- No new experiments. This paper is conceptual and synthetic. The protocol is proposed, not yet run; its value depends on applying it systematically across models and tasks.
 - A fast-moving field. Several of the studies reviewed used models that have since been superseded, and newer models may perform differently on specific probes. The framework is meant to outlast particular results; the results cited are not.
 - Contested definitions. The definition of functional understanding adopted here is one reasonable choice among several; readers who define understanding differently may reach different conclusions from the same evidence (Mitchell & Krakauer, 2023).
 - The hard problem remains. The paper deliberately says little about phenomenal awareness beyond marking its separation from the other levels. That question is not resolved here.
-

9. Conclusion

The intuition behind the original version of this paper, that AI systems echo rather than understand, is partly right and partly wrong, and the evidence now allows us to say which parts. Current language models show extensive echo: they memorise, they track output probability, they fail on counterfactual variants and reversed relations, and they tell users what they seem to want to hear. They also show real, if partial, functional understanding: they build and use internal models of domains they have only encountered as text. The path “from echo to awareness” therefore runs through two separate gaps. The first, from echo to functional understanding, is empirical and measurable, and the seven probes proposed here are a way to measure it task by task. The second, from understanding to awareness, is of a different kind; progress on the first does not close it, and claims about it should be made with corresponding care.

References

- Bender, E. M., & Koller, A. (2020). Climbing towards NLU: On meaning, form, and understanding in the age of data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 5185–5198). Association for Computational Linguistics.
- Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (pp. 610–623). ACM.
- Berglund, L., Tong, M., Kaufmann, M., Balesni, M., Stickland, A. C., Korbak, T., & Evans, O. (2024). The reversal curse: LLMs trained on “A is B” fail to learn “B is A”. In *International Conference on Learning Representations (ICLR 2024)*. <https://arxiv.org/abs/2309.12288>
- Burrell, J. (2016). How the machine “thinks”: Understanding opacity in machine learning algorithms. *Big Data & Society*, 3(1).
- Butlin, P., Long, R., et al. (2023). *Consciousness in artificial intelligence: Insights from the science of consciousness* (arXiv:2308.08708) [Preprint]. <https://arxiv.org/abs/2308.08708>
- Carlini, N., Tramèr, F., Wallace, E., Jagielski, M., Herbert-Voss, A., Lee, K., Roberts, A., Brown, T., Song, D., Erlingsson, Ú., Oprea, A., & Raffel, C. (2021). Extracting training data from large language models. In *30th USENIX Security Symposium* (pp. 2633–2650).
- Chalmers, D. J. (1995). Facing up to the problem of consciousness. *Journal of Consciousness Studies*, 2(3), 200–219.
- Chollet, F. (2019). *On the measure of intelligence* (arXiv:1911.01547) [Preprint]. <https://arxiv.org/abs/1911.01547>
- Christiano, P. F., Leike, J., Brown, T. B., Martic, M., Legg, S., & Amodei, D. (2017). Deep reinforcement learning from human preferences. In *Advances in Neural Information Processing Systems 30*.
- Finn, C., Abbeel, P., & Levine, S. (2017). Model-agnostic meta-learning for fast adaptation of deep networks. In *Proceedings of the 34th International Conference on Machine Learning* (PMLR 70, pp. 1126–1135).
- Gurnee, W., & Tegmark, M. (2024). Language models represent space and time. In *International Conference on Learning Representations (ICLR 2024)*. <https://arxiv.org/abs/2310.02207>
- Harnad, S. (1990). The symbol grounding problem. *Physica D: Nonlinear Phenomena*, 42(1–3), 335–346.
- Kosinski, M. (2024). Evaluating large language models in theory of mind tasks. *Proceedings of the National Academy of Sciences*, 121(45), e2405460121. <https://doi.org/10.1073/pnas.2405460121>

- Lake, B. M., Ullman, T. D., Tenenbaum, J. B., & Gershman, S. J. (2017). Building machines that learn and think like people. *Behavioral and Brain Sciences*, 40, e253.
- LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, 521(7553), 436–444.
- Li, K., Hopkins, A. K., Bau, D., Viégas, F., Pfister, H., & Wattenberg, M. (2023). Emergent world representations: Exploring a sequence model trained on a synthetic task. In *International Conference on Learning Representations (ICLR 2023)*. <https://arxiv.org/abs/2210.13382>
- Marcus, G. (2020). *The next decade in AI: Four steps towards robust artificial intelligence* (arXiv:2002.06177) [Preprint]. <https://arxiv.org/abs/2002.06177>
- McCoy, R. T., Yao, S., Friedman, D., Hardy, M. D., & Griffiths, T. L. (2024). Embers of autoregression show how large language models are shaped by the problem they are trained to solve. *Proceedings of the National Academy of Sciences*, 121(41), e2322420121. <https://doi.org/10.1073/pnas.2322420121>
- Mitchell, M., & Krakauer, D. C. (2023). The debate over understanding in AI's large language models. *Proceedings of the National Academy of Sciences*, 120(13), e2215907120. <https://doi.org/10.1073/pnas.2215907120>
- Nagel, T. (1974). What is it like to be a bat? *The Philosophical Review*, 83(4), 435–450.
- Ouyang, L., Wu, J., Jiang, X., et al. (2022). Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems 35* (pp. 27730–27744).
- Picard, R. W. (1997). *Affective computing*. MIT Press.
- Poria, S., Cambria, E., Bajpai, R., & Hussain, A. (2017). A review of affective computing: From unimodal analysis to multimodal fusion. *Information Fusion*, 37, 98–125.
- Searle, J. R. (1980). Minds, brains, and programs. *Behavioral and Brain Sciences*, 3(3), 417–424.
- Shanahan, M. (2024). Talking about large language models. *Communications of the ACM*, 67(2), 68–79.
- Sharma, M., Tong, M., Korbak, T., et al. (2024). Towards understanding sycophancy in language models. In *International Conference on Learning Representations (ICLR 2024)*. <https://arxiv.org/abs/2310.13548>
- Ullman, T. (2023). *Large language models fail on trivial alterations to theory-of-mind tasks* (arXiv:2302.08399) [Preprint]. <https://arxiv.org/abs/2302.08399>
- Wu, Z., Qiu, L., Ross, A., Akyürek, E., Chen, B., Wang, B., Kim, N., Andreas, J., & Kim, Y. (2024). Reasoning or reciting? Exploring the capabilities and limitations of language models through counterfactual tasks. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics* (pp. 1819–1862). <https://aclanthology.org/2024.naacl-long.102/>

Zhang, C., Bengio, S., Hardt, M., Recht, B., & Vinyals, O. (2021). Understanding deep learning (still) requires rethinking generalization. *Communications of the ACM*, 64(3), 107–115.