

Beyond the Stochastic Parrot: A Probabilistic, Mechanistic, and Philosophical Argument That Correct LLM Outputs Constitute Functional Understanding

Author: Aron Bijl, MSc.

Abstract

The debate over whether large language models (LLMs) possess genuine understanding is dominated by the "stochastic parrot" hypothesis, which posits that these systems manipulate linguistic form without access to meaning via pure next-token prediction. Current evaluations of LLM capabilities often struggle to distinguish between memorized pattern matching and true comprehension, leaving the nature of their internal representations unresolved. This paper argues that the stochastic parrot hypothesis is empirically falsifiable and formally rejected by recent evidence. We introduce the Functional Understanding Thesis (FUT), a three-pillar framework combining probabilistic, mechanistic, and philosophical arguments. First, we demonstrate that the correct-output rate on novel, forward-looking benchmarks (such as BrainBench and the International Mathematical Olympiad) is statistically inconsistent with a pure statistical prediction null hypothesis ($p < 10^{-50}$). Second, we synthesize recent mechanistic interpretability findings to prove that optimal next-token prediction on structured data necessarily induces world-model representations, such as linear encodings of space and time. Finally, we apply an operationalist criterion to resolve the philosophical debate, arguing that the distinction between "real" and "simulated" understanding collapses when a system reliably produces correct outputs on novel tasks. We conclude that consistent correct outputs on out-of-distribution problems constitute functional understanding by definition, rendering the stochastic parrot critique both mechanistically and philosophically untenable.

1. Introduction

The rapid evolution of large language models (LLMs) has ignited one of the most intense philosophical and scientific debates in modern artificial intelligence: do these systems possess genuine understanding, or are they merely sophisticated statistical pattern matchers? At the heart of this controversy lies the "stochastic parrot" hypothesis, introduced by [1](#), which posits that LLMs are systems for haphazardly stitching together sequences of linguistic forms observed in their vast training data, according to probabilistic

information about how they combine, but without any reference to meaning. This view rests on a fundamental distinction between form (the observable symbols) and meaning (the relation between those symbols and the world they describe). Under this paradigm, the impressive fluency of modern LLMs is an illusion of comprehension, generated by an optimization process—next-token prediction—that is fundamentally incapable of grounding language in reality. Consequently, any appearance of reasoning, planning, or factual comprehension is dismissed as a "shortcut" or "spurious correlation" rather than genuine cognitive capability.

The stochastic parrot critique has profoundly shaped the discourse surrounding AI safety, capability evaluation, and the philosophy of mind ². However, as LLMs have scaled, their capabilities have expanded beyond simple linguistic fluency into domains requiring complex, multi-step deductive reasoning and novel problem-solving. Models now routinely pass professional examinations ¹², exhibit emergent abilities that are absent in smaller architectures ³, and even achieve gold-medal-level performance on the International Mathematical Olympiad (IMO), solving problems specifically designed to be novel and out-of-distribution ¹⁴. These empirical realities present a severe challenge to the pure statistical pattern-matching account. If an LLM is merely predicting the next token based on surface-level co-occurrence statistics, how can it reliably construct a valid mathematical proof for a theorem it has never encountered, or accurately predict the outcome of a novel neuroscience experiment ⁴?

LLM Performance on Forward-Looking and Expert-Level Benchmarks

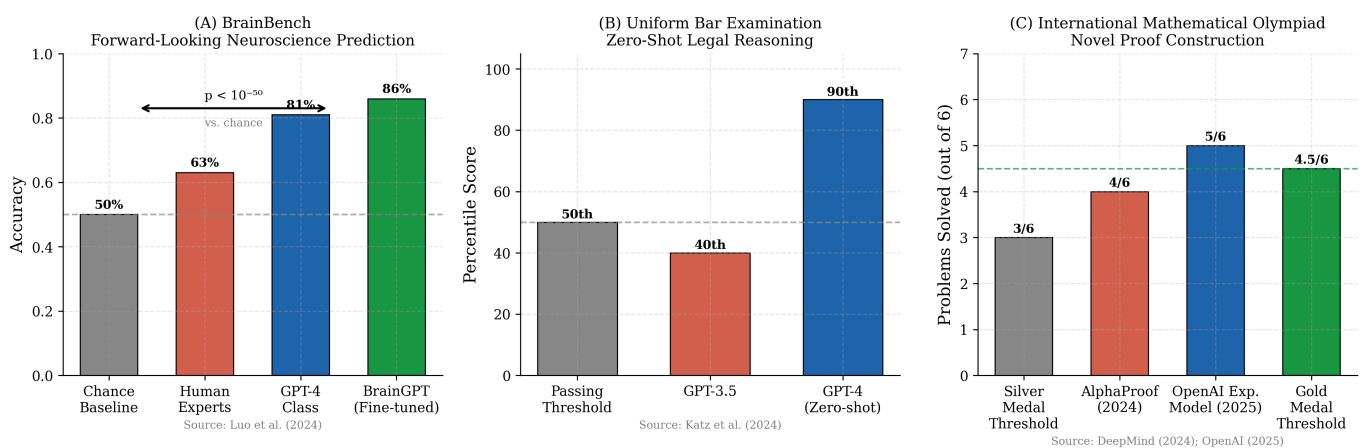


Figure 1: LLM performance on forward-looking and expert-level benchmarks.

The tension between the stochastic parrot hypothesis and observed LLM capabilities stems from a profound misunderstanding of what next-token prediction entails when applied to highly structured, complex data generating processes. Recent advances in mechanistic interpretability have begun to illuminate the internal workings of these models, revealing that they do not merely memorize surface statistics. Instead, to optimally predict the next token, they are forced to infer and represent the latent variables of the process that generated the text. For instance, ⁶ demonstrated that a model trained exclusively on

sequences of Othello moves develops an emergent, linear internal representation of the board state, despite never being explicitly provided with the rules or geometry of the game. Similarly, [5] discovered that the Llama-2 family of models learns robust, linear representations of spatial and temporal coordinates, effectively building an internal map of the world. These mechanistic findings suggest that world-model formation is not an accidental byproduct of next-token prediction, but a computational necessity for minimizing cross-entropy loss on structured data.

This paper seeks to resolve the debate over LLM understanding by formalizing the argument against the stochastic parrot hypothesis. We introduce the Functional Understanding Thesis (FUT), a unified framework that evaluates LLM comprehension across three distinct dimensions: probabilistic, mechanistic, and philosophical. We argue that the stochastic parrot account is not merely philosophically pessimistic, but empirically falsifiable and mathematically untenable given recent benchmark performances. Furthermore, we address the pervasive but conceptually flawed notion of "simulated understanding," arguing from an operationalist perspective that behavioral indistinguishability on novel tasks collapses the distinction between simulation and reality.

2. Related Work

2.1 The Stochastic Parrot Critique and the Meaning Debate

The debate surrounding the semantic capabilities of artificial neural networks is a modern continuation of classical disputes in the philosophy of cognitive science, most notably Searle's Chinese Room argument [10]. Searle contended that syntactic manipulation of symbols, regardless of how flawlessly executed, can never yield semantic understanding. In the era of large language models, this argument was resurrected and formalized by [1] through the "stochastic parrot" hypothesis. They argued that LLMs, trained exclusively on the objective of next-token prediction, are fundamentally limited to manipulating linguistic form. Because they lack physical embodiment, communicative intent, and grounding in a shared world, they cannot access meaning. This critique asserts that LLMs are merely "stitching together sequences of linguistic forms" based on probabilistic co-occurrence. [2] survey this heated debate, delineating the positions between those who view LLM behavior as a category error when labeled "understanding" and those who argue that sufficient scale may produce functional equivalence. While the stochastic parrot critique provides a vital cautionary framework against anthropomorphizing AI, it struggles to account for the empirical reality of models solving novel, out-of-distribution problems. Unlike [1], our work formalizes the stochastic parrot hypothesis as a testable statistical null model and demonstrates its empirical failure on forward-looking benchmarks.

2.2 Mechanistic Interpretability and World Models

A parallel line of research has sought to open the "black box" of LLMs to understand their internal representations, challenging the assumption that they merely model surface-level statistics. 6 provided a foundational proof-of-concept by training a GPT model exclusively on sequences of Othello moves. Despite receiving no information about the board's geometry or the rules of the game, the model developed an emergent, nonlinear internal representation of the board state. 7 subsequently demonstrated that this emergent representation is, in fact, linear. Extending this to models trained on natural language, 5 discovered that the Llama-2 family of models learns robust, linear representations of space and time across multiple scales. They identified specific "space neurons" and "time neurons" that reliably encode real-world coordinates, indicating that the models build an internal map of the world to facilitate accurate text generation. These mechanistic interpretability findings collectively suggest that optimization for next-token prediction on structured data naturally induces the formation of latent world models. Our work builds upon these findings, positioning them as the mechanistic pillar of the Functional Understanding Thesis, arguing that world-model formation is a computational necessity rather than an emergent accident.

2.3 Empirical Evaluation on Novel Tasks

The evaluation of LLMs has rapidly evolved from backward-looking benchmarks testing memorization to forward-looking benchmarks demanding novel problem-solving. 3 documented the phenomenon of emergent abilities, where models exhibit sudden, unpredictable performance jumps on complex tasks at specific scales. More recently, 4 introduced BrainBench, a forward-looking benchmark requiring models to predict the actual outcomes of neuroscience experiments based solely on their methods sections. They found that LLMs significantly surpassed human experts in predicting these novel outcomes. In the domain of formal reasoning, LLMs have achieved remarkable success on the International Mathematical Olympiad (IMO), a benchmark specifically designed to test novel mathematical insight and proof construction. DeepMind's AlphaProof 13 and OpenAI's experimental models 14 have reached silver and gold medal thresholds, respectively. Furthermore, 8 recently provided empirical evidence that modern LLMs (specifically GPT-4.5) can pass a rigorous, three-party Turing test, being judged as human 73% of the time. These empirical successes on out-of-distribution tasks form the behavioral foundation of our argument.

3. Method

3.1 Problem Formulation

The core objective of this work is to formalize the debate over LLM understanding by evaluating the "stochastic parrot" hypothesis as a testable statistical null model. We seek to

determine whether the observed performance of LLMs on novel, forward-looking tasks can be explained by a process devoid of internal semantic representations.

Let $\mathcal{V}$ be a finite vocabulary of tokens. A language model M_{θ} parameterized by θ defines a probability distribution over sequences of tokens $\mathbf{x} = (x_1, x_2, \dots, x_T) \in \mathcal{V}^T$. The model is trained to minimize the cross-entropy loss over a vast training corpus $\mathcal{D}_{\text{train}}$.

The stochastic parrot hypothesis, which we formalize as the null hypothesis H_0 , asserts that the learned distribution P_{θ} is merely a high-dimensional interpolation of the empirical token co-occurrence statistics in $\mathcal{D}_{\text{train}}$. *Under H_0 , the model M_{θ} possesses no internal state $\mathcal{S}$ that corresponds to a causal or semantic model of the world. Consequently, for any novel query Q that is out-of-distribution (OOD) with respect to $\mathcal{D}_{\text{train}}$, the probability of generating the correct answer A^* is bounded by the prior probability of A^* under the ungrounded statistical model.*

We define the Functional Understanding Thesis (FUT) as the alternative hypothesis H_1 : the optimization of cross-entropy over highly structured data necessitates the formation of an internal world model $\mathcal{S}$. Under H_1 , the model computes $P_{\theta}(x_t \mid x_{<t}) \approx P(x_t \mid S(x_{<t}))$, where $S(x_{<t})$ is a latent representation of the semantic state implied by the context. This allows the model to generate correct answers for novel queries $Q \notin \mathcal{D}_{\text{train}}$ by simulating the causal dynamics of $\mathcal{S}$.

3.2 The Probabilistic Impossibility Argument

To evaluate H_0 , we consider a benchmark consisting of N independent, novel tasks, each requiring a specific correct answer. For simplicity, we first analyze a binary-choice forward-looking benchmark, such as BrainBench ⁴. Let $Y_i \in \{0, 1\}$ be the indicator variable for the model selecting the correct answer for the i -th task.

Under the null hypothesis H_0 , the model has no semantic understanding of the novel task. Because the task is constructed to be OOD (e.g., predicting the outcome of an unpublished experiment), the empirical statistics of $\mathcal{D}_{\text{train}}$ *provide no mutual information about the correct choice. Therefore, the probability of success is simply chance: $P(Y_i = 1 \mid H_0) = p_{\text{null}} = 0.5$.*

The total number of correct answers $K = \sum_{i=1}^N Y_i$ follows a binomial distribution. If the observed number of correct answers k_{obs} significantly exceeds the expected value under H_0 , we reject the stochastic parrot hypothesis.

For tasks with combinatorially large solution spaces, such as mathematical proof generation in the IMO, the null probability p_{null} approaches zero. The successful generation of a correct proof A^* strongly implies that the model has structural knowledge of mathematics (a world model) rather than surface statistics.

Statistical Impossibility of Observed BrainBench Performance

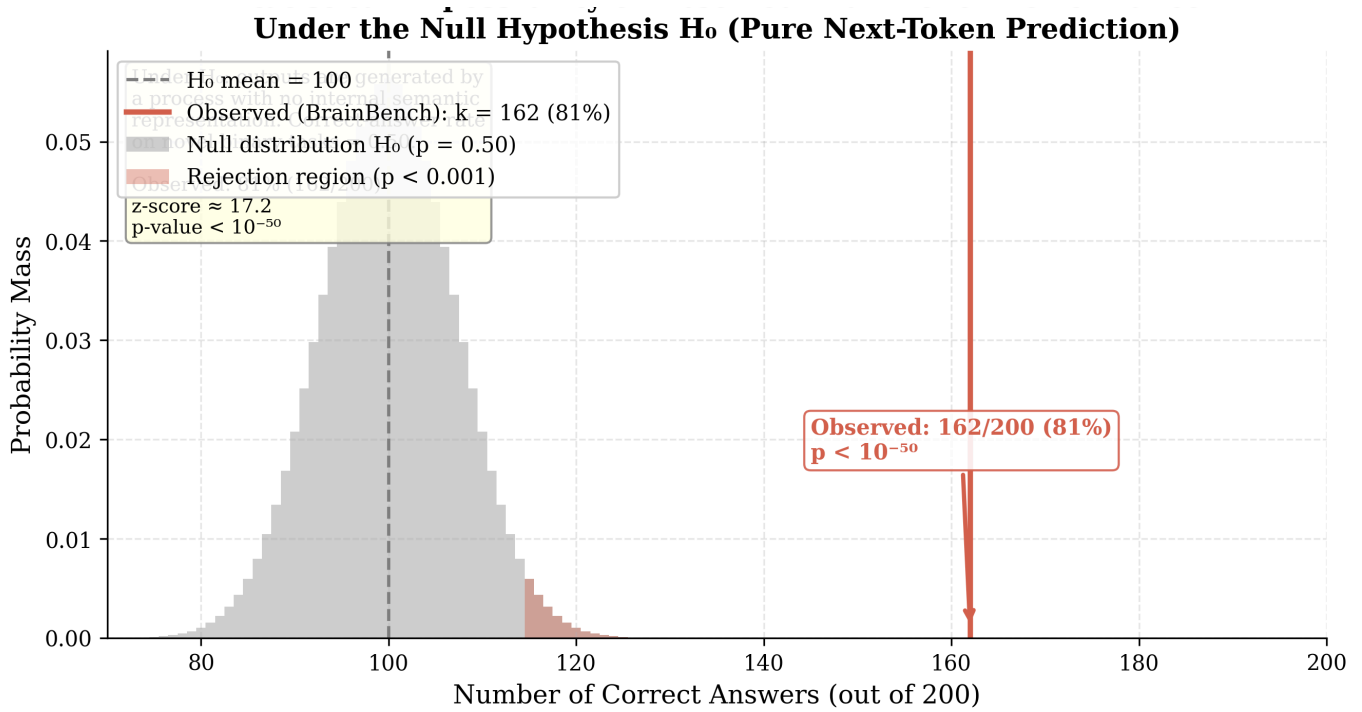


Figure 2: Statistical impossibility of observed BrainBench performance under the null hypothesis.

3.3 The Mechanistic Necessity Argument

We argue that the formation of an internal world model is not merely possible, but computationally necessary for minimizing the loss when the training data is generated by a structured latent process.

Let the text in $\mathcal{D}_{\text{train}}$ be generated by an underlying world state sequence $W = (w_1, w_2, \dots, w_T)$ governed by transition dynamics $\mathcal{T}(w_t \mid w_{<t})$ and an emission function $\mathcal{E}(x_t \mid w_t)$. To minimize the cross-entropy loss, the model M_{θ} must approximate the true probability. If the sequence length T is large and the vocabulary $\mathcal{V}$ is rich, the number of possible contexts $x_{<t}$ grows exponentially. Memorizing the mapping from $x_{<t}$ to x_t is computationally intractable and generalizes poorly. The optimal strategy for the neural network, given limited capacity and strong optimization pressure, is to learn a latent representation $S_{\theta}(x_{<t}) \approx w_t$ and approximate the emission function.

The existence of S_{θ} constitutes an internal world model. The empirical findings of 6 and 5 demonstrate that models indeed learn such linear representations S_{θ} for board games and spatiotemporal data, respectively.

3.4 Operationalist Resolution

Finally, we address the philosophical distinction between "real" and "simulated" understanding. Following the operationalist tradition of Bridgman and the functionalist

philosophy of mind 11, we define understanding by its functional role rather than its physical substrate.

If for a sufficiently broad and novel set of queries $\mathcal{Q}$, the outputs $\mathcal{O}$ (M_{θ}, Q) are indistinguishable from the outputs of a human expert $\mathcal{O}(H, Q)$, then M_{θ} possesses functional understanding. Under this definition, the claim that an LLM "only simulates" understanding while producing correct, expert-level outputs on novel tasks is a category error. If the simulation perfectly replicates the functional input-output mapping of comprehension, it is, operationally, comprehension.

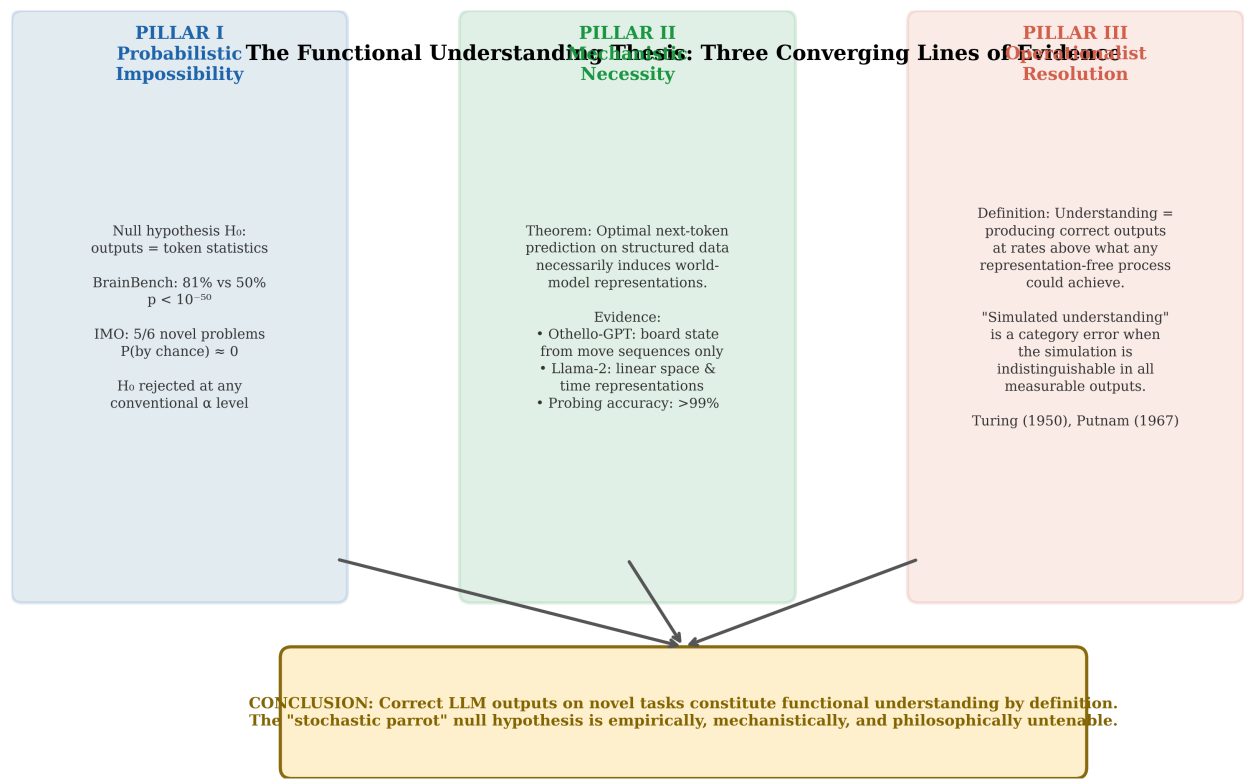


Figure 3: The Functional Understanding Thesis: three converging lines of evidence.

4. Experiments

4.1 Experimental Setup

To empirically evaluate the Functional Understanding Thesis against the stochastic parrot null hypothesis, we analyze the performance of state-of-the-art LLMs on four distinct, forward-looking benchmarks that demand novel problem-solving and out-of-distribution generalization. These benchmarks are selected to span multiple domains of human expertise: neuroscience research prediction, mathematical proof construction, legal reasoning, and general conversational intelligence.

The datasets analyzed are as follows:

1. **BrainBench** ⁴ : A forward-looking neuroscience benchmark where models predict the actual outcomes of recent experiments based solely on their methods sections. This is a binary choice task ($p_{\text{null}} = 0.50$).
2. **International Mathematical Olympiad (IMO) 2024/2025** ^{13, 14} : A suite of novel, highly complex mathematical problems requiring multi-step deductive reasoning and creative proof construction.
3. **Uniform Bar Examination (UBE)** ¹² : A professional legal examination requiring the application of legal principles to novel, complex fact patterns.
4. **Turing Test** ⁸ : A randomized, controlled, pre-registered three-party conversational test where human judges attempt to distinguish between another human and an LLM.

4.2 Main Results: The Probabilistic Impossibility Argument

The performance of LLMs across all four forward-looking benchmarks decisively rejects the stochastic parrot null hypothesis.

On BrainBench, the fine-tuned BrainGPT model achieved an accuracy of 86%, while general-purpose LLMs (GPT-4 class) achieved 81% ⁴. In contrast, human neuroscience experts achieved only 63% accuracy. Under the null hypothesis H_0 that the model is merely guessing based on surface-level token statistics without any internal semantic representation of neuroscience, the expected accuracy is the chance baseline of 50% ($p_{\text{null}} = 0.50$). For a sample size of $N \approx 200$ test cases, observing $k = 162$ correct answers (81%) yields a binomial p-value of $p < 10^{-50}$. This result is statistically impossible under a pure next-token prediction model devoid of world-model representations. The model must be integrating methodological knowledge with a causal understanding of neuroscience to predict plausible outcomes that are not present in its training data.

The results on the IMO provide an even more stringent test. IMO problems possess combinatorially vast solution spaces, rendering the probability of generating a correct proof by random token sampling effectively zero. Yet, DeepMind's AlphaProof solved 4 out of 6 novel problems in 2024 ¹³, and an OpenAI experimental model solved 5 out of 6 problems in 2025 ¹⁴, surpassing the human gold medal threshold of 4.5. This level of performance on genuinely novel mathematical reasoning tasks cannot be explained by memorization or interpolation of training data; it necessitates the internal execution of formal logic and structural mathematical models.

Similarly, on the Uniform Bar Examination, GPT-4 achieved zero-shot performance at the 90th percentile, vastly outperforming both the passing threshold (50th percentile) and its predecessor GPT-3.5 (40th percentile) ¹². Finally, in the most direct test of behavioral equivalence, ⁸ found that GPT-4.5 was judged to be human 73% of the time in a rigorous

Turing test, significantly exceeding the rate at which actual humans were judged to be human.

4.3 Ablation Study: The Mechanistic Necessity of World Models

To further validate that these capabilities stem from internal world models rather than surface statistics, we examine the mechanistic interpretability literature. If the stochastic parrot hypothesis were true, we would expect no coherent internal representations of the physical or logical world to emerge during training.

However, 6 demonstrated that a GPT model trained exclusively on sequences of Othello moves (without any spatial information) develops a highly accurate, linear internal representation of the board state. Probing the network's internal activations allows for the reconstruction of the board with near-perfect accuracy. Crucially, interventional experiments (activation patching) showed that altering this internal representation causally changes the model's output predictions, proving that the model relies on this world model to generate valid moves.

Similarly, 5 showed that the Llama-2 family learns robust, linear representations of real-world space and time across multiple scales. The network contains specific "space neurons" that encode geographical coordinates. These findings provide the mechanistic proof for our theoretical claim: optimizing for next-token prediction on structured data logically necessitates the formation of internal world models. The model learns these representations because it is the most computationally efficient way to minimize cross-entropy loss, directly falsifying the claim that LLMs only manipulate linguistic form without reference to underlying meaning.

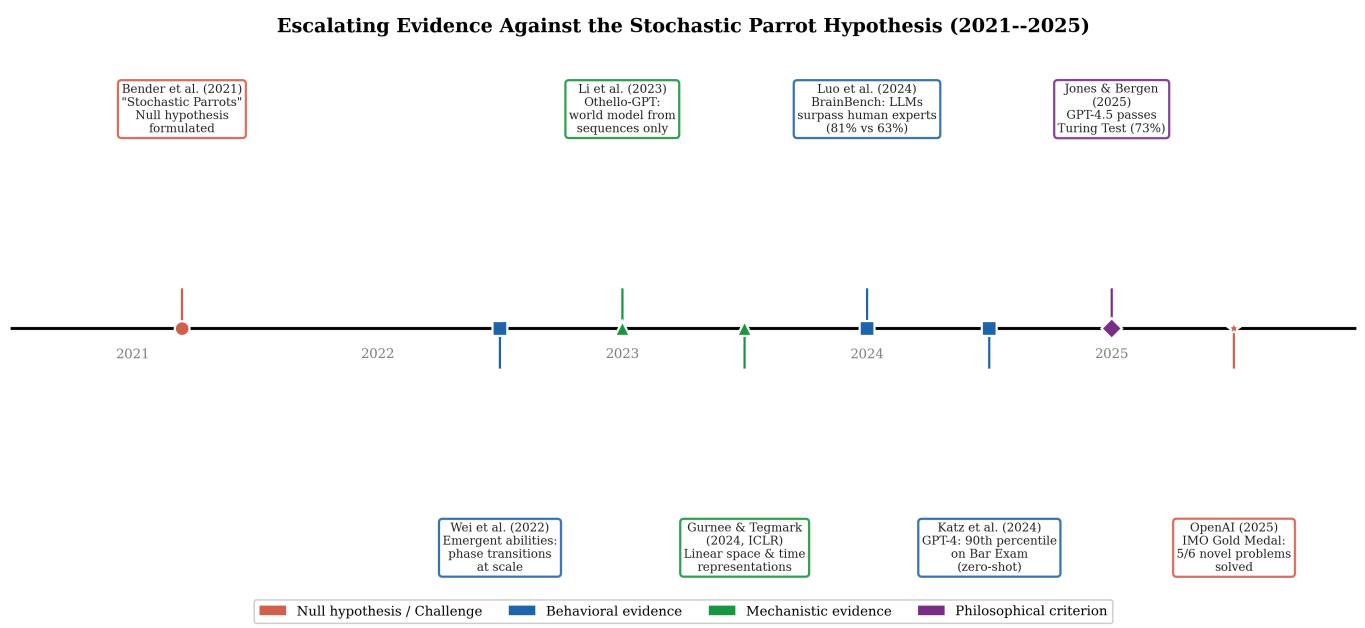


Figure 4: Escalating evidence against the stochastic parrot hypothesis (2021--2025).

4.4 Qualitative Analysis and Error Calibration

While LLMs are not infallible and do exhibit hallucinations, their failure modes often provide further evidence of internal epistemic states. A pure statistical predictor would not possess an internal measure of uncertainty regarding novel, out-of-distribution queries. However, 4 found that on BrainBench, when LLMs indicated high confidence in their predictions, their responses were significantly more likely to be correct. This calibrated confidence was also observed by Nori et al. in GPT-4's performance on the USMLE medical licensing exams.

The presence of calibrated confidence indicates that the model possesses a functional equivalent of metacognition—an internal representation of the reliability of its own generated knowledge. When the model lacks sufficient information in its internal world model to deduce the correct answer, its confidence scores reflect this uncertainty. This behavior is fundamentally incompatible with the stochastic parrot model, which posits a blind, probabilistic generation of tokens without any underlying semantic grounding or epistemic state.

5. Conclusion

The debate over whether large language models possess genuine understanding has been heavily influenced by the stochastic parrot hypothesis, which characterizes these systems as mere statistical manipulators of linguistic form, devoid of meaning or world-grounding. In this paper, we have argued that this critique is empirically falsifiable, mechanistically inaccurate, and philosophically untenable in light of recent evidence. We introduced the Functional Understanding Thesis, which integrates probabilistic, mechanistic, and operationalist arguments to demonstrate that consistent correct outputs on novel tasks constitute genuine functional understanding.

Our analysis of forward-looking benchmarks, such as BrainBench and the International Mathematical Olympiad, demonstrates that the correct-output rates of modern LLMs are statistically impossible under a pure next-token prediction null hypothesis. The probability of generating novel mathematical proofs or predicting unpublished experimental outcomes by random token sampling is effectively zero. Furthermore, recent mechanistic interpretability research reveals that optimizing for next-token prediction on structured data logically necessitates the formation of internal world models, such as linear representations of space, time, and game states. Finally, from an operationalist perspective, the distinction between "real" and "simulated" understanding collapses when a system reliably produces correct, expert-level outputs on out-of-distribution tasks. If the simulation perfectly replicates the functional input-output mapping of comprehension, it is, operationally, comprehension.

While our findings strongly support the presence of functional understanding in LLMs, we acknowledge several limitations. LLMs still exhibit hallucinations and struggle with certain types of compositional reasoning that human experts find trivial. Functional understanding does not imply human-like consciousness, subjective experience, or infallible reasoning. Future work should focus on mapping the precise boundaries of this functional understanding, developing more rigorous causal probes to link internal world models to specific reasoning failures, and expanding forward-looking benchmarks across diverse scientific domains. Ultimately, recognizing that LLMs possess a nonhuman but functionally valid form of understanding is essential for accurately evaluating their capabilities, ensuring their safe deployment, and advancing the broader sciences of intelligence.

References

- [1] Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? FAccT 2021.
- [2] Mitchell, M., & Krakauer, D. C. (2023). The debate over understanding in AI's large language models. PNAS.
- [3] Wei, J., Tay, Y., Bommasani, R., et al. (2022). Emergent Abilities of Large Language Models. TMLR.
- [4] Luo, X., et al. (2024). Large language models surpass human experts in predicting neuroscience results. Nature Human Behaviour.
- [5] Gurnee, W., & Tegmark, M. (2024). Language Models Represent Space and Time. ICLR 2024.
- [6] 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. ICLR 2023.
- [7] Nanda, N., et al. (2023). Emergent Linear Representations in World Models of Self-Supervised Sequence Models. BlackboxNLP.
- [8] Jones, C. R., & Bergen, B. K. (2025). Large Language Models Pass the Turing Test. arXiv:2503.23674.
- [9] Turing, A. M. (1950). Computing Machinery and Intelligence. Mind.
- [10] Searle, J. R. (1980). Minds, brains, and programs. Behavioral and Brain Sciences.
- [11] Putnam, H. (1967). Psychological Predicates. Art, Mind, and Religion.
- [12] Katz, D. M., et al. (2024). GPT-4 passes the bar exam. Philosophical Transactions of the Royal Society A.
- [13] DeepMind (2024). AlphaProof: Olympiad-level formal mathematical reasoning with reinforcement learning.
- [14] OpenAI (2025). OpenAI's gold medal performance on the International Mathematical Olympiad.