ARTICLE

From Tabula Rasa to Inductive Bias: Reframing Locke's Problem in the Age of Generative AI

Xufeng Zhang[a] and Han Li[b]

[a]Newlane University, Lehi, USA; [b]Lanzhou University, Lanzhou, China

ARTICLE HISTORY
Compiled December 16, 2025

ABSTRACT
Large language models (LLMs) often appear to vindicate a radical empiricist picture: train on vast corpora of experience-like text, and capacities emerge without explicit symbolic rules. Yet contemporary machine learning research repeatedly emphasizes that what is learned, how quickly it is learned, and how well it generalizes depend crucially on prior constraints: architectural structure, training objectives, optimization dynamics, and representational bottlenecks. These constraints constitute inductive biases in a precise, technical sense. This paper develops a philosophical argument that uses LLMs as a case study to reassess the classical tabula rasa thesis associated with Locke and its descendants. I defend two claims. First, even the most "data-driven" generative models are saturated with structural priors that make learning possible at scale; thus, their success cannot be straightforwardly read as a triumph of unstructured empiricism. Second, once this is appreciated, the rhetorical appeal of a "blank slate" conception of the infant mind weakens further: if artificial systems trained on orders of magnitude more linguistic input than any child still require rich inductive biases, it is implausible that human cognition begins wholly unstructured. I then show how contemporary debates about whether LLMs "understand" language recapitulate the older dispute about whether experience alone can generate semantics and conceptual structure. The upshot is a more balanced, non-caricatured empiricism: experience matters, but explanation must explicitly account for the learning system's prior structure.

KEYWORDS
inductive bias; empiricism; John Locke; tabula rasa; large language models; semantic grounding

1. Introduction

Generative AI has re-opened old philosophical questions in a new register. Large language models (LLMs) trained on massive datasets produce fluent text, solve exam-style problems, write executable code, and adapt to novel tasks via prompting. This empirical fact has been rhetorically mobilized in two opposite directions. One line of commentary treats LLMs as an existence proof for radical empiricism: give a sufficiently flexible learner enough experience, and sophisticated cognition will "emerge" without substantive innate structure. A second line instead sees LLMs as sophisticated engines of statistical mimicry that cannot yield genuine understanding, thereby reaffirming the necessity of grounding, embodiment, or inborn conceptual resources.

CONTACT Xufeng Zhang. Email: xufeng@nau.edu

This paper aims to shift the terms of this debate by focusing on a technical point that tends to be obscured when LLMs are treated as philosophical symbols. In machine learning, no learner is truly "blank." All learning requires inductive bias, understood as a restriction on the hypothesis space or a preference ordering over possible generalizations (Mitchell 1980). Indeed, modern learning theory contains formal results—including, but not limited to, the No Free Lunch theorems—that make some form of bias unavoidable if generalization is to be possible (Wolpert and Macready 1997). Importantly, the inductive bias of contemporary deep learning systems is not confined to explicit regularizers. It is distributed across architecture, objective function, optimization procedure, tokenization, data curation, and interface-level fine-tuning procedures such as reinforcement learning from human feedback (RLHF) (Ouyang et al. 2022). The philosophical moral is not that LLMs are "innate" in any human sense, but that their empirical success cannot be attributed to experience alone, conceived as undifferentiated input.

A further reason the "LLMs vindicate empiricism" line of thought is attractive is that contemporary pretraining is often described as "self-supervised" and "task-agnostic." The model is not given explicit instructions about grammar, semantics, or reasoning; it is trained to predict continuations of text, and capabilities appear to scale with data and compute (Kaplan et al. 2020; Hoffmann et al. 2022). It is therefore easy to slide from "no explicit supervision" to "no substantive prior structure." But this slide is precisely what the present argument resists. "Task-agnostic" does not mean "bias-free." Even an apparently generic objective (next-token prediction) embodies a determinate notion of success, and the machinery built to optimize it encodes strong assumptions about what kinds of regularities are worth representing (Vaswani et al. 2017; Lavie, Gur-Ari, and Ringel 2024).

This paper also insists on a second clarification that is often neglected in philosophical discussion: the "experience" of LLMs is not raw worldly contact but an engineered, curated, and filtered textual artifact. Many modern training corpora are derived from large-scale web scrapes; their composition depends on filtering choices, deduplication procedures, and blocklists, and these choices are known to have systematic downstream effects (Dodge et al. 2021; Bender et al. 2021). In other words, even before we consider architectural priors, what counts as the model's "input" is already shaped by a pipeline that embeds normative and methodological commitments. This will matter when we later ask whether LLMs support an experience-only picture of learning. If the "experience" itself is structured and selected, it cannot straightforwardly serve as an analogue of an unstructured sensory stream.

The core thesis of the paper can be stated in the following conditional:

> If even the most "data-driven" generative models cannot learn or generalize without strong structural inductive biases, then portraying the human infant mind as an entirely unstructured blank slate is more a rhetorical victory than an explanatory one.

This thesis is not intended as a knockdown refutation of Locke. Historically sensitive scholarship already emphasizes that Locke's anti-nativism is more nuanced than a cartoonish "mind as empty storage" picture (Uzgalis 2001). The target is rather a strong *tabula rasa* ideology that treats learning as the mere accumulation of experiential content, while downplaying, or effectively ignoring, the prior organization of the learning system. My strategy is to use LLMs as a vivid contemporary case study to clarify why such a view is explanatorily unstable.

Finally, it is worth specifying what sort of philosophical leverage LLMs can provide. The paper does *not* assume that LLMs are good psychological models of human

learners; nor does it claim that any specific architectural feature in transformers corresponds to a biologically innate module. Instead, the paper uses LLMs as a "stress test" for a family of empiricist slogans. If the most celebrated contemporary example of large-scale, data-intensive learning nevertheless requires substantial prior structure, then the bare appeal to "experience" cannot, by itself, discharge the explanatory burdens involved in accounting for cognition.

I proceed as follows. Section 2 reconstructs the tabula rasa theme and articulates the "Locke problem" as a family of explanatory pressures: how can learners generalize far beyond the data available, acquire structured concepts, and do so robustly across environments? Section 3 then introduces inductive bias as a methodological constraint in both philosophy of mind and machine learning, arguing that any serious empiricism must specify the learner's prior structure. Section 4 offers a detailed account of where inductive bias lives in LLMs: transformer architecture, next-token prediction objectives, RLHF fine-tuning, tokenization, data curation, and implicit regularization induced by optimization. Section 5 draws the central philosophical lesson: LLMs do not instantiate radical empiricism; they instead illustrate how experience and structure cooperate. Section 6 connects this lesson to the "do LLMs understand?" controversy, treating it as a contemporary form of the older question whether semantics can be generated from form alone (Bender and Koller 2020; Harnad 1990). Section 7 considers several objections and replies. Section 8 concludes with a constructive proposal: rather than choose between "blank slate" and "fully innate," we should adopt an explicitly *bias-aware empiricism* that treats prior structure as part of the explanandum.

2. The Tabula Rasa Thesis and the Locke Problem

Locke famously rejects innate speculative principles and innate ideas. The metaphor of the mind as "white paper" is offered to motivate a methodological demand: if we want to understand the origin and scope of knowledge, we should trace the contents of thought to experience and reflect on the operations by which the mind composes complex ideas from simpler ones (Locke 1975). This is often summarized—sometimes misleadingly—as the doctrine that the human mind is a blank slate at birth. Two clarifications are crucial.

First, Locke's rejection of innateness is directed primarily at *innate propositional contents* or *innate items of knowledge*. He is not committed to denying that the mind has powers, capacities, or structural features that make experience usable. Indeed, even a minimal account of how experience yields ideas must presuppose faculties of perception, attention, memory, abstraction, and comparison. Locke can plausibly be read as relocating explanatory work away from innate *ideas* and toward the mind's *operations* on experiential inputs (Uzgalis 2001). The extreme "blank slate" interpretation becomes most plausible only when Locke's metaphor is detached from his broader theory of mental operations.

Second, and more important for the present argument, the philosophical debate that Locke helped shape concerns not merely whether some contents are innate, but whether *experience alone* can explain the emergence of rich cognitive structure. This is the pressure point sometimes called "the Locke problem" in cognitive science and philosophy: how can finite, noisy, contingent experience support the acquisition of abstract concepts, compositional structures, and far-reaching generalizations? The problem is not unique to language acquisition, though language provides a paradigmatic case. Learners, including infants, appear to possess expectations about objects, agency,

number, and causality that guide learning in ways that are difficult to derive from raw sensory input alone (Spelke 2000). More broadly, human cognition displays what looks like *structure-sensitive* learning: people generalize in ways that respect hierarchical organization, causal dependencies, and relational patterns, rather than merely correlational surface regularities (Tenenbaum et al. 2011; Lake et al. 2017).

A historical bridge is instructive here. In mid-twentieth-century linguistics, Skinner-style behaviorism proposed an experience-driven account of language as learned verbal behavior. Chomsky's influential review argued that such an account fails to explain key properties of linguistic competence—in particular, the systematicity, productivity, and rule-governed character of language (Chomsky 1959). Regardless of how one evaluates the details, the review framed the dispute as an explanatory challenge for an empiricist program: merely pointing to reinforcement histories and stimulus control is insufficient to account for the structured nature of language. For present purposes, the important point is not to import Chomsky's theoretical commitments into the LLM debate, but to note a recurrent dialectical pattern: in domains where cognition exhibits strong structure and far-reaching generalization, appeals to "experience" without an accompanying account of the learner's internal organization are explanatorily fragile.

It is helpful to distinguish three versions of the Locke problem, each more demanding than the last:

The Generalization Problem. How can a learner move from observed instances to unobserved cases reliably? Even if experience provides a stock of examples, there are infinitely many possible hypotheses consistent with any finite dataset. Without some preference ordering—some bias—the learner cannot justify any particular "inductive leap" (Mitchell 1980). In philosophical terms, this is a descendant of Humean underdetermination: data alone does not determine theory.

The Structure Problem. How can learners acquire structured representations (e.g., compositional concepts, hierarchical syntax, causal models) rather than merely flat associations? A representational system must support operations such as variable binding, substitution, and systematic recombination if it is to explain the productivity of thought and language. Here the explanatory pressure concerns not just *which* generalization is chosen, but what the system can even represent and manipulate.

The Semantics Problem. How can symbols, words, or internal representations come to be *about* anything? Even if statistical learning yields patterns of use, does it thereby yield meaning, or does meaning require grounding in perception, action, and shared forms of life (Harnad 1990; Bisk et al. 2020)? This is the point at which empiricism intersects with debates about reference, intentionality, and the nature of conceptual content.

These three problems are historically entangled. Classical empiricists hoped to solve them by appealing to association, abstraction from experience, and the gradual construction of complex ideas. Nativists, in contrast, argued that experience underdetermines the learned structure, motivating inborn constraints. Contemporary cognitive science often rejects the stark dichotomy, emphasizing that learning is shaped by both input and prior structure (Samet 2008; Tenenbaum et al. 2011). Crucially, the contemporary landscape encourages a reframing: the central question is not "innate vs. learned" in a binary sense, but *what kinds of priors* are required for learning to be tractable and robust in a given environment.

The contribution of generative AI to these debates is double-edged. On one hand, LLMs show that large-scale statistical learning over linguistic data can produce impressive behavioral competence (Brown et al. 2020; OpenAI et al. 2023). On the other hand, close inspection reveals that this competence depends on engineered structural constraints at every level. If so, then LLMs are not a straightforward model of radical empiricism. Rather, they may function as an instructive analogue: they dramatize how far "experience" can take a learner, *given* powerful inductive biases.

3. Inductive Bias as a Methodological Requirement

In machine learning, "inductive bias" is not a pejorative label for arbitrary prejudice; it is a necessary condition for learning. Mitchell's classic formulation defines inductive bias as "any basis for choosing one generalization over another, other than strict consistency with the training instances" (Mitchell 1980). This conception generalizes readily to philosophy of mind. Any theory of learning that claims to explain cognitive competence by appealing to experience must specify what allows experience to be *interpretable* and *generalizable* for the learner.

The No Free Lunch theorems sharpen this into a formal constraint: averaged over all possible objective functions, no optimization algorithm outperforms any other; gains on one class of problems must be paid for by losses on another (Wolpert and Macready 1997). Although the theorems are sometimes overextended, the underlying message is robust: performance requires assumptions about the structure of the environment. Philosophically, this is a structural analogue of Hume's problem of induction: if we refuse to specify any bias, we cannot explain why inductive inference works as well as it does in the actual world.

To make the methodological point more precise, it helps to distinguish (at least) three ways inductive bias enters a learning explanation.

3.1. *Representational Bias, Search Bias, and Data-Selection Bias*

First, there is *representational bias*: the set of hypotheses the learner can express. In classical learning theory this is modeled as a hypothesis space $\mathcal{H}$. If the true regularities of the environment are not well-approximated within $\mathcal{H}$, no amount of data will rescue performance. Architectural choices in deep learning instantiate representational bias by defining the class of functions that are easy or hard to represent.

Second, there is *search or procedural bias*: even if multiple hypotheses in $\mathcal{H}$ fit the training data equally well, the learning procedure must select among them. This selection can be explicit (regularization terms, early stopping) or implicit (optimization dynamics that prefer certain parameter configurations) (Neyshabur et al. 2017; Soudry et al. 2018). Search bias matters because many modern learners are heavily overparameterized: there can be many distinct solutions with near-zero training error, and yet generalization differs widely among them (Zhang et al. 2017).

Third, there is *data-selection bias*: the distribution of experiences the learner receives is itself structured. In human development this includes ecological regularities, social scaffolding, and pedagogical cues. In LLM training it includes corpus construction, filtering, and deduplication pipelines (Dodge et al. 2021). Data-selection bias is often misdescribed as "just the environment," but explanatorily it plays the same role as a prior: it changes what the learner treats as typical or salient, and hence changes the generalizations it is likely to form.

The present paper emphasizes that all three dimensions are active in LLMs and, plausibly, in human learning. This is why bare appeals to "experience" are inadequate: experience is always received through a representational and procedural filter, and it is always drawn from some distribution.

3.2. *Bias as Constraint and Bias as Prior*

Inductive bias can be modeled either as a constraint or as a probabilistic prior. In a constraint-based picture, the learner rules out large regions of hypothesis space. In a Bayesian picture, the learner assigns a prior probability $P(h)$ over hypotheses $h \in \mathcal{H}$ and updates by Bayes' rule given data D, producing a posterior $P(h \mid D)$. The Bayesian perspective has been influential in cognitive science precisely because it makes explicit how learning depends jointly on priors and evidence (Tenenbaum et al. 2011).

For the purposes of the Locke problem, what matters is that both pictures reject a "no prior structure" view. A purely uniform prior over an enormous hypothesis space is practically equivalent to having no guidance at all; learning becomes sample-inefficient and unstable. Conversely, a structured prior can make learning tractable by privileging hypotheses with certain forms (e.g., causal graphs, compositional structures, smooth functions). The theoretical interest of LLMs is that, while they are not explicitly Bayesian learners in the usual sense, many of their observed generalization behaviors can be understood as the consequence of implicit priors induced by architecture and training objectives (Xie et al. 2021).

3.3. *Why "Bias" is not an Ad Hoc Escape Hatch*

A natural worry is that appeals to inductive bias could be used to immunize any theory from criticism: whenever learning is hard, simply posit a stronger prior. The appropriate methodological response is to treat inductive bias as part of what must be *explained*. In both cognitive science and machine learning, a good bias is not an arbitrary stipulation; it is one that fits the structure of the target environment and yields robust generalization under plausible data conditions (Goyal and Bengio 2022; Lake et al. 2017). Inductive bias therefore functions as an explanatory *commitment*: once one specifies the bias, one is committed to a certain pattern of successes and failures across tasks and environments.

This point will matter when we later compare LLMs and human learners. The claim is not simply "humans must have some bias" (which is trivial), but that the kinds of rapid and robust generalization humans display plausibly require rich structural priors. LLMs will be used as an argument by analogy, not to identify the content of those priors, but to undermine the credibility of the idea that cognition begins wholly unstructured.

4. Where Inductive Bias Lives in Large Language Models

This section provides a detailed account of the principal sources of inductive bias in contemporary LLMs. The goal is not an engineering tutorial. Rather, it is to make explicit the layers of structure that must be presupposed before "data-driven learning" can occur. For philosophical purposes, the important point is that these biases are not

optional decorations; they are conditions of learnability and generalization.

4.1. *Architectural Bias: Transformers as Structured Hypothesis Spaces*

The dominant architecture for LLMs is the transformer (Vaswani et al. 2017). At a high level, a transformer is a parameterized function that maps sequences of discrete tokens to probability distributions over the next token. But this description hides architectural commitments.

First, the transformer treats language as a sequence and builds representations via self-attention. Self-attention implements a content-addressable mechanism: each token can weight information from other tokens, subject to learned queries and keys. This provides a bias toward representing *relations* among tokens rather than only local features, and it enables long-range dependencies to be integrated in parallel. In effect, the architecture assumes that what matters for prediction is the pattern of interactions among elements in a context window, and it provides a computational primitive tailored to that assumption.

Second, transformers are permutation-sensitive only via positional encodings. Without positional information, self-attention is permutation invariant. The need to add positional encodings means that the architecture embodies an explicit decision about what kind of sequence structure is relevant. Recent theoretical work argues that, even in certain limits, transformers can exhibit biases toward permutation-symmetric functions in sequence space, and that learnability depends systematically on context length and symmetry properties (Lavie, Gur-Ari, and Ringel 2024). The philosophical point is not the specific asymptotic regime, but that "architecture" is a substantive constraint on what functions are representable and easily learnable.

Third, transformers are deep compositional systems. Layering attention and feed-forward modules yields a hierarchy of representations. This depth introduces a bias toward *multi-stage feature extraction*: earlier layers can capture local statistical patterns, while later layers can aggregate and abstract. Such hierarchies are not entailed by experience; they are a design choice that creates the possibility of representing abstract regularities as iterated compositions of simpler computations.

Finally, transformer design inherits and amplifies a broader family of "relational" inductive biases. Self-attention can be interpreted as implementing a learned, soft form of message passing among token representations. This connects it to a wider literature on relational inductive biases in deep learning (Battaglia et al. 2018; Bronstein et al. 2021). The relevance to the Locke problem is that relational inductive bias provides a principled way to represent structure that is not reducible to local co-occurrence. Even if the model is trained on raw text, its architecture makes certain relational abstractions easier to discover and exploit than they would be in a generic function approximator.

4.2. *Objective Bias: Next-Token Prediction and Its Consequences*

LLMs are typically trained with an autoregressive objective: maximize the likelihood of the next token given preceding tokens. GPT-style models exemplify this approach (Brown et al. 2020). Such training has often been treated as the epitome of data-driven learning: no task-specific labels, only a generic prediction objective. Yet the objective encodes a substantive bias: it treats linguistic competence as whatever is useful for predicting continuations in text.

This matters in at least three ways.

First, next-token prediction privileges *distributional form*. To the extent that semantic regularities manifest in patterns of co-occurrence, the objective can internalize them. But the objective is indifferent to truth, grounding, and reference except insofar as these affect token statistics. As a result, the model can become an excellent predictor of text without thereby becoming a reliable guide to the world. This underwrites concerns that LLMs may generate plausible but false outputs, a point that motivates alignment procedures such as RLHF (Ouyang et al. 2022).

Second, the objective encourages *compressive abstraction*. Predicting text well requires capturing long-range dependencies and latent topics. This creates pressure to infer hidden variables that explain document-level coherence, which can in turn support in-context learning behaviors. Theoretical work makes this connection precise in stylized settings by showing how in-context learning can emerge as a kind of implicit Bayesian inference over latent document concepts (Xie et al. 2021). On a philosophical reading, the model is not simply memorizing strings; it is pressured to construct internal states that act like "hypotheses" about what is going on in the text.

Third, next-token prediction induces a distinctive notion of generalization. Generalization is evaluated by perplexity or log-likelihood on held-out text, not by grounded performance in sensorimotor environments. This constrains what success means. Philosophically, it implies that any inference from LLM success to human cognition must be careful: the model optimizes for linguistic continuation, whereas humans use language for communication embedded in action and shared practices. One can therefore regard LLM training as a highly specialized form of empiricism: empiricism about textual regularities under a particular success criterion.

4.3. *Alignment and Fine-Tuning Bias: RLHF as Normative Structure*

A further layer of bias enters through fine-tuning and alignment procedures, especially RLHF (Ouyang et al. 2022). RLHF begins with supervised fine-tuning on demonstrations, then trains a reward model from human preference rankings, and finally performs reinforcement learning to optimize for higher reward.

Two philosophical observations follow.

First, RLHF makes explicit that "learning from data" is not value-neutral. It injects human normative judgments about helpfulness, harmlessness, and instruction-following into the model's objective landscape. Thus, the model is not merely fitting linguistic statistics; it is being shaped to satisfy human evaluative standards. This undermines simplistic empiricist narratives that treat LLMs as passive mirrors of textual experience. Even if the pretraining phase is framed as passive absorption, RLHF explicitly reorients the system toward interactive norms of inquiry and conversation.

Second, RLHF illustrates how biases can be layered: the base model's inductive bias is already substantial, but alignment adds another bias toward conversational cooperation, deference to user intent, and socially acceptable discourse. These biases are not reducible to the raw pretraining corpus. They are engineered constraints that change what the model is likely to say, and, crucially, what kinds of generalizations it will manifest in interactive settings. In many practical deployments, users primarily encounter RLHF-shaped behavior; therefore, philosophical claims about "what LLMs are like" should be cautious about attributing properties to pretraining alone.

4.4. *Optimization Bias: Implicit Regularization and the Geometry of Generalization*

Even holding architecture and objective fixed, training dynamics matter. Deep learning models generalize well despite being massively overparameterized, a fact that has motivated extensive investigation into implicit regularization: optimization procedures can bias solutions toward certain function classes even without explicit penalties (Neyshabur et al. 2017; Zhang et al. 2017). In simpler settings, gradient descent can be shown to converge to max-margin solutions, demonstrating that the training algorithm selects among many zero-training-error solutions according to implicit preferences (Soudry et al. 2018).

For our purposes, the lesson is straightforward: the trained model is not merely the result of "data plus architecture." It is also shaped by a training process that systematically prefers some representations over others. Thus, the empiricist picture of learning as imprinting from experience is incomplete unless it includes the dynamics by which experience is integrated. In philosophy-of-science terms, the optimization process is part of the inferential "method" of the learner; it functions as a constraint on what hypotheses are effectively reachable.

4.5. *Representational Bottlenecks: Tokenization as a Prior over Linguistic Structure*

Tokenization is often treated as a preprocessing detail, yet it imposes a representational bias by determining the atomic units the model can manipulate. Subword tokenization methods, such as byte pair encoding approaches (Sennrich, Haddow, and Birch 2016) and related tokenizers (Kudo and Richardson 2018), reflect an assumption about linguistic structure: words can be decomposed into reusable subunits that capture morphology, frequent character sequences, or other regularities. This design increases learnability by enabling the model to share parameters across related word forms and handle open-vocabulary phenomena.

Philosophically, tokenization is a clear example of "structure before experience." The learner does not encounter raw text as an unstructured stream; it is given a segmentation scheme that defines what counts as a symbol-like unit. This should caution against taking LLM performance as evidence that purely unstructured exposure can yield high-level competence. Indeed, one might say that tokenization plays a role analogous to perceptual segmentation in human cognition: it determines what the system treats as the "parts" over which it can learn systematic relations.

4.6. *Data and Corpus Construction: The Model's "Experience" is Curated*

So far, we have treated "data" as if it were simply given. But in practice, LLM pretraining depends on large-scale corpora whose construction embodies substantial methodological and normative choices. For example, the widely used C4 dataset was created by filtering Common Crawl, and later analysis documented that this filtering changes corpus composition in non-trivial ways (Dodge et al. 2021). The Pile was explicitly constructed as a diverse mixture of subcorpora, motivated by the hypothesis that diversity improves generalization (Gao et al. 2020). Such corpus design decisions are not merely practical conveniences; they materially influence what generalizations the model can form, and they therefore function as a form of inductive bias.

Two points deserve emphasis. First, corpus construction affects what the model treats as typical linguistic practice. If certain registers, genres, or communities are overrepresented or filtered out, then the model's learned distribution will inherit these imbalances. Second, curation decisions can encode normative commitments. Dodge et al. report that blocklist filtering in C4 disproportionately removes text from and about minority identities (Dodge et al. 2021). This is a concrete mechanism by which "the data" embeds social and political structure. Philosophically, this reinforces the idea that the LLM's "experience" is not an unstructured given but a filtered artifact, and it complicates any attempt to use LLMs as exemplars of raw empiricism.

There is also a methodological lesson for debates about tabula rasa. If someone claims that LLMs show how a system can become competent purely by exposure, we should ask: exposure to what, under what curation regime, and with what implicit structure in the distribution of examples? Once these questions are taken seriously, the romantic picture of an agent passively absorbing undifferentiated experience becomes harder to sustain.

4.7. *Mechanistic Bias: In-Context Learning and Algorithmic Circuits*

Perhaps the most philosophically suggestive LLM capability is in-context learning: the ability to adapt behavior based on examples in the prompt without parameter updates. Mechanistic interpretability research suggests that certain attention heads implement algorithm-like behaviors, such as "induction heads" that complete repeated token patterns (Olsson et al. 2022). Such results are preliminary and model-dependent, but they support a general point: the transformer architecture makes certain computations *easy to implement*. In-context learning is not a miracle of data alone; it is enabled by structural affordances of attention and by training conditions that reward exploiting those affordances (Xie et al. 2021).

Recent theoretical and empirical work goes further, suggesting that in-context learning can implement recognizable learning algorithms in the model's forward pass. Garg et al. show that transformers can be trained to perform in-context learning for simple function classes, including linear functions, with performance comparable to task-specific estimators (Garg et al. 2022). Akyürek et al. provide evidence that, at least in linear settings, in-context learners can approximate gradient descent, ridge regression, or least-squares predictors, with transitions between these behaviors depending on model depth and data conditions (Akyürek et al. 2022). Von Oswald et al. strengthen the mechanistic connection by constructing equivalences between self-attention transformations and gradient-descent-like updates, and by experimentally identifying similarities between transformer behavior and optimization in stylized regression tasks (Oswald et al. 2023).

These results matter for the Locke problem because they show that "learning from examples" can be partially internalized as computation within the model—but only because the architecture supports such internalization. In-context learning is therefore a paradigmatic case of bias-driven learnability: the model can behave as if it were performing a form of inference over the prompt because the computational substrate (attention over sequences) makes that form of inference readily expressible and trainable.

In sum, LLMs display a layered hierarchy of inductive biases. Architecture defines representational and computational primitives; objectives define what counts as success; optimization dynamics select among solutions; tokenization constrains symbolic

units; data curation shapes the distribution of "experience"; alignment procedures introduce normative preferences; and mechanistic circuits exploit architectural affordances. The existence of these biases does not diminish the remarkable role of scale and data (Kaplan et al. 2020; Hoffmann et al. 2022). But it does imply that LLM success cannot be interpreted as a vindication of an experience-only theory of learning.

5. Reassessing Empiricism: Lessons and Limits from LLMs

With the foregoing in view, we can articulate the central philosophical moral. LLMs do not instantiate radical empiricism. Rather, they show that large-scale learning requires a cooperation between (i) massive exposure to structured input and (ii) prior constraints that render that input learnable and generalizable.

5.1. *Why "Scale Alone" Does Not Eliminate Bias*

A tempting reply to the emphasis on inductive bias is to concede that biases exist but insist that, at sufficiently large scale, the biases become negligible. On this view, the role of bias is merely to get learning started; the real explanatory work is done by data and compute.

This reply misconstrues how inductive bias functions. Bias is not an additive booster that can be dialed down as data grows. It is a constitutive feature of the learning system, shaping what it can represent and what it will converge to, at all scales. Scaling laws show that performance improves predictably with model size, data, and compute (Kaplan et al. 2020), and subsequent work shows that data-model tradeoffs matter for compute-optimal training (Hoffmann et al. 2022). But these results do not imply that architecture and optimization are irrelevant; rather, they hold under fixed modeling choices. Scaling thus describes how performance changes *within* a biased hypothesis space, not how learning becomes unbiased.

Indeed, the very existence of stable scaling laws suggests that the learning problem has exploitable structure and that the model class is well-matched to it. If the bias were negligible, we would not expect consistent, systematic improvements under scaling within a particular architecture family. In philosophy-of-science terms, the stability of scaling is evidence that the method is well-calibrated to the domain; methods are never domain-neutral in a way that would support strong tabula rasa rhetoric.

5.2. *The Bitter Lesson and the Misreading of Generality*

A related but distinct line of thought is encapsulated in Sutton's "bitter lesson": across the history of AI, general methods that leverage computation tend to outperform approaches that bake in extensive domain-specific structure (Sutton 2019). This lesson is sometimes invoked to support a strong empiricist stance: if general methods win, then perhaps cognitive competence arises from generic learning plus lots of data, not from rich priors.

The present paper endorses part of Sutton's methodological attitude while rejecting the empiricist overextension. The bitter lesson is most plausibly read as a warning against *narrow, brittle, human-coded domain knowledge* that fails to scale. But it does not follow that scalable methods have no inductive bias. On the contrary, what makes a method "general" in practice is often that it embodies *broad* structural pri-

ors (e.g., compositionality, relational processing, smoothness, invariances) that apply across many tasks (Goyal and Bengio 2022; Battaglia et al. 2018). In that sense, the bitter lesson is compatible with the thesis of this paper: successful learning requires bias, but the most valuable biases are those that are powerful, reusable, and aligned with deep regularities of the environment.

This distinction matters for Locke. A strong blank-slate ideology treats any mention of prior structure as a retreat from empiricism. But the machine-learning notion of inductive bias shows that the relevant contrast is not bias vs. no bias; it is *which* biases, at *what level of generality*, and with *what empirical consequences*.

5.3. *From LLMs to Human Cognition: An Inference to the Best Explanation*

How does this bear on the tabula rasa thesis? The argument is not that human minds are "transformers" or that architectural priors in machines correspond one-to-one with innate structures in humans. The argument is instead an inference to the best explanation under comparative constraints.

Consider the following asymmetry. Human children acquire language and broad world knowledge with far less linguistic input than LLMs, and they do so in ways tightly integrated with perception, action, and social interaction. LLMs, by contrast, are trained on enormous corpora and still require substantial engineered inductive biases to generalize well and to behave coherently in interactive contexts. If a system with dramatically more "experience" still depends on rich prior structure, it is implausible that a system with dramatically less experience could succeed with less structure, unless we posit compensating forms of inductive bias in the human case.

This inference should be handled carefully. One might object that human "experience" is multimodal and temporally dense, and that the total information available to a developing child is not directly comparable to token counts in a text corpus. This is correct, and it is one reason the paper does not attempt any crude numerical comparison between "tokens seen" and "words heard." However, the point remains that human learning is not merely data-rich; it is *structure-rich*. Caregivers scaffold attention, provide feedback, and embed language in joint action. Such scaffolding itself constitutes a form of inductive bias at the level of the learning environment. Once again, we see that learning explanations must include prior structure somewhere: either in the learner, in the environment, or (most plausibly) in their interaction.

In cognitive science, the classic point is not that concepts are innate as explicit propositions, but that learning is guided by expectations about what kinds of hypotheses are plausible. Core knowledge proposals, for instance, treat infants as having domain-specific expectations about objects and agents that guide learning (Spelke 2000). Bayesian models of cognition similarly emphasize that learners bring priors to data, and that those priors may reflect evolved structure or learned biases acquired early in development (Tenenbaum et al. 2011). The LLM case supports this general methodological stance: explanation must include prior structure. Once this is accepted, the strong blank-slate rhetoric loses explanatory force. It can still be true that experience is indispensable and that many specific beliefs are acquired. But it is not explanatory to say that experience *alone* yields cognitive structure, unless one specifies how the learner's architecture makes such extraction possible.

5.4. *A More Plausible Empiricism: Bias-Aware, Not Bias-Denying*

If the target is a strong tabula rasa ideology, what replaces it? I propose what might be called *bias-aware empiricism*. It has three commitments:

(1) Experience is necessary for most specific knowledge and for calibrating cognition to local environments.
(2) Learning requires prior constraints that make some generalizations easier than others; these constraints can be innate, developmental, or culturally scaffolded.
(3) The explanatory task is to identify the interaction between input and constraints, rather than to celebrate one factor while erasing the other.

This view is compatible with a nuanced reading of Locke. Locke's emphasis on experience as the source of ideas need not imply that the mind is computationally unstructured. But it is incompatible with the stronger claim that humans begin cognitively unstructured and that all structure is written in by experience. The LLM case, precisely because it is sometimes treated as the purest instance of data-driven learning, helps to expose that stronger claim as a rhetorical simplification.

6. Does the Model "Understand"? LLM Controversies as Contemporary Locke Problems

The debate over whether LLMs understand language is often treated as novel. In fact, it can be seen as a contemporary expression of the semantics problem within the Locke problem family: can meaning and conceptual structure arise from experience with linguistic form alone?

6.1. *Meaning from Form? The Bender–Koller Challenge*

A prominent critique argues that training on form alone cannot yield meaning. Bender and Koller distinguish linguistic form from communicative intent and argue that systems trained only on form cannot, by that fact alone, acquire the relation between expressions and speakers' intended meanings (Bender and Koller 2020). The critique does not deny that LLMs can be useful or impressive; it challenges an inference from behavioral fluency to human-like understanding.

This critique aligns with older concerns about symbol grounding. A purely formal system may manipulate symbols according to syntactic rules without thereby achieving intrinsic semantics (Harnad 1990). Searle's Chinese Room thought experiment similarly aims to show that syntactic manipulation, even if it produces correct outputs, need not constitute understanding (Searle 1980). While Searle's argument is controversial, the underlying question is relevant: what is required for linguistic competence to count as semantic competence?

6.2. *Grounding and Shared Experience*

Bisk and colleagues argue that language understanding research is hampered by a failure to relate language to the physical and social world. They emphasize that successful communication relies on shared experience of the world, and that text-only training may be insufficient for deeper forms of understanding (Bisk et al. 2020). This connects the LLM debate to a broader empiricist-nativist discussion: experience matters, but

the relevant experience may not be reducible to exposure to text. Human language acquisition is embedded in perceptual and social contexts, suggesting that semantics may require forms of interaction that LLMs lack.

The present paper adds a complementary observation: even if we bracket grounding worries, LLM competence is not a simple product of exposure to text. It depends on the biases identified in Section 4. Thus, the contemporary semantics debate should not be framed as "text-only experience yields everything" versus "text-only experience yields nothing." A more illuminating framing is: given a particular architecture and objective, what kinds of semantic-like structure can be learned from textual distributional information, and what kinds plausibly require broader forms of coupling to the world? This framing is continuous with the Locke problem: it treats meaning as a target for explanation, not as a label to be bestowed or withheld.

6.3. *Anthropomorphism, "As-If" Psychology, and the Lure of Understanding Talk*

A further complication is the tendency to anthropomorphize. As Shanahan argues, the more fluent and conversational LLMs become, the easier it is to slip into talk of the system as knowing, believing, or understanding (Shanahan 2022). This tendency is not merely rhetorical; it can shape the explanatory standards we implicitly apply. If we treat conversational smoothness as evidence of genuine mentality, we may overlook the extent to which that smoothness is a product of alignment and interface-level fine-tuning (Ouyang et al. 2022). Conversely, if we insist that only grounded, embodied agents can understand, we may underappreciate the sophistication of the inferential structures that large-scale training can produce within language.

Bias-aware empiricism suggests a cautious middle path. One can acknowledge that LLMs exhibit impressive *competences*—including forms of abstraction, analogy, and in-context adaptation (Garg et al. 2022; Akyürek et al. 2022)—while remaining non-committal about whether these competences amount to full-blown understanding. Philosophically, this resembles an "as-if" stance: LLMs can behave as if they had beliefs and reasons in constrained interactional contexts, because their learned internal structures support patterns of inference that track many human expectations. But to move from "as-if" to "is" would require a more substantive account of grounding, reference, and participation in social practices (Harnad 1990; Bisk et al. 2020).

6.4. *The Counterpressure: Emergent Competence and the Temptation of Empiricism*

On the other side, researchers have documented striking emergent capabilities in large models. Work on GPT-3 emphasizes few-shot learning from prompts (Brown et al. 2020), and the GPT-4 report describes broad performance across benchmarks (OpenAI et al. 2023). Bubeck and collaborators argue that early versions of GPT-4 display a constellation of capabilities that invite renewed discussion of general intelligence (Bubeck et al. 2023). Such results encourage the thought that semantics, concepts, and reasoning may emerge from large-scale pattern learning.

How should we adjudicate this? The present paper does not attempt to settle the metaphysics of understanding. Instead, it offers a diagnostic: the debate mirrors the tension between (i) the thought that experience with linguistic patterns might suffice to generate cognitive structure, and (ii) the thought that something beyond such ex-

perience is required for meaning. Importantly, the inductive bias analysis complicates both sides.

For the skeptic, the existence of strong inductive biases in LLMs supports the claim that model competence is not merely an accumulation of text experience. The model's behavior is produced by a structured architecture optimized for prediction. Thus, even if the model lacks grounding, its outputs may reflect sophisticated internal structure, not mere parroting. At the same time, for the enthusiast, the presence of strong biases undermines a simplistic empiricist reading of emergence: capabilities arise not from data alone, but from the interaction of data with architectural and objective constraints.

6.5. *A Middle Position: Semantic Competence as an Achievement of Structured Learning in Context*

A plausible middle position is that semantic competence is not an all-or-nothing property but a cluster of abilities that can be partially instantiated. LLMs may exhibit robust *inferential* and *pragmatic* competencies within linguistic contexts while lacking full *grounded* reference. This resembles how some empiricist accounts treat concepts as patterns of use and inference rather than as intrinsically referential mental atoms. Yet even on this middle position, inductive bias remains central: the model's inferential abilities depend on prior structure, and human understanding likely depends on even richer structure, given humans' integration of language with perception and action.

In this sense, the LLM understanding debate functions as a contemporary Locke problem: can experience with linguistic form yield semantics and concepts, and if so, under what structural constraints? The lesson is not a return to crude nativism, but a rejection of the idea that experience alone, conceived as undifferentiated input, can do the explanatory work.

7. Objections and Replies

7.1. *Objection 1: Inductive Bias in LLMs Is Trivial Compared to Data*

One might argue that the biases identified above are minimal and generic. Transformers, next-token prediction, and gradient descent are not domain-specific, so the model is still fundamentally an empiricist learner: the real content comes from data.

7.1.0.1. Reply.. The objection conflates *domain-specificity* with *explanatory significance*. A bias can be highly general and still be crucial. Mitchell's point is precisely that any nontrivial generalization requires bias, whether or not that bias is tailored to a specific domain (Mitchell 1980). Moreover, the transformer architecture is not content-free: it embodies assumptions about sequence processing, relational integration via attention, and hierarchical composition (Vaswani et al. 2017; Battaglia et al. 2018). Tokenization embodies assumptions about subword structure (Sennrich, Haddow, and Birch 2016; Kudo and Richardson 2018). Optimization embodies preferences among solutions (Soudry et al. 2018). These are not trivial in the sense relevant to explanation; they shape what is learnable and how.

Finally, appeals to scaling laws do not support the claim that bias is negligible. Scaling laws describe how performance improves under increasing resources *within* a fixed class of biased learners (Kaplan et al. 2020; Hoffmann et al. 2022). They do not

show that the learner becomes unbiased. If anything, the stability of scaling suggests that the biases are well-matched to the structure of language data.

7.2. *Objection 2: The Human Case Is Not Analogous; Biology Is Different*

A critic may grant that LLMs have inductive biases while insisting that this has little relevance to the human mind. Human learning occurs in embodied agents with evolved neural circuitry, motivations, and social interactions. Therefore, the need for engineered biases in artificial systems does not imply anything about innateness or structure in humans.

7.2.0.1. Reply.. The argument does not depend on a tight analogy between transformers and brains. It depends on a comparative constraint: if learning is possible at all, some form of inductive bias is necessary. The human case must therefore involve bias, whether evolved, developmental, or socially scaffolded. The LLM case is relevant because it is often claimed to be a paradigm of "experience-only" learning; showing that even this paradigm depends on rich priors undercuts the plausibility of experience-only explanations more generally.

Moreover, cognitive science provides independent evidence that humans bring structured expectations to learning (Spelke 2000; Tenenbaum et al. 2011; Lake et al. 2017). The LLM analysis complements this evidence by providing a mechanistic demonstration of how far experience can go *given* architectural constraints. If anything, the embodied and social richness of human development suggests the presence of additional constraints that LLMs lack, strengthening the conclusion that a strong blank-slate picture is implausible.

7.3. *Objection 3: Locke Was Not a "Strong Blank Slate" Thinker*

A historically informed objection is that the paper attacks a strawman. Locke's antinativism is nuanced, and he does not deny mental powers and operations. Therefore, the argument does not refute Locke but only a vulgarized ideology.

7.3.0.1. Reply.. This objection is largely correct as a point of intellectual history. The target is indeed a strong tabula rasa ideology rather than Locke's most sophisticated position. However, the point is not merely exegetical. Locke's metaphors have played a major role in shaping popular and sometimes academic rhetoric about human malleability. The philosophical task is to assess the explanatory adequacy of that rhetoric. LLMs provide a timely case study showing that even in engineered systems, successful learning requires substantial prior structure. This supports a more accurate and fruitful reading of empiricism: anti-nativism about propositional content can be combined with serious attention to prior structure. In that sense, the argument can be read as rehabilitating a nuanced empiricism rather than rejecting it.

7.4. *Objection 4: Inductive Bias Is Compatible with Tabula Rasa*

One might argue that the tabula rasa thesis never meant "no structure whatsoever." It meant only "no innate ideas." If so, then the existence of inductive biases in LLMs

is irrelevant: biases are not innate ideas.

7.4.0.1. Reply.. Two points. First, if tabula rasa is interpreted so weakly that it allows substantial innate structural constraints, then it becomes compatible with bias-aware empiricism and ceases to support the strong "blank slate" rhetoric. My argument presses precisely on this: once bias is acknowledged as essential, the slogan "blank slate" loses its explanatory bite and must be replaced by an explicit account of what the learner contributes.

Second, the semantics problem shows why structure matters. Even if we deny innate propositional content, we still need to explain how concepts, reference, and understanding arise. The LLM debate makes this vivid: text-only learning can yield impressive competence, yet questions remain about grounding and meaning (Bender and Koller 2020; Bisk et al. 2020; Harnad 1990). To address these questions, we need theories that specify not only input but also the system's structural capacities to connect input to the world. A purely negative thesis about "no innate ideas" is insufficient for explanation.

7.5. *Objection 5: "The Bitter Lesson" Undermines the Paper's Moral*

A final objection appeals directly to Sutton. If the bitter lesson is that scalable general methods dominate, then emphasizing inductive bias risks encouraging the wrong research program: building more structure into systems rather than leveraging computation and data (Sutton 2019). The present paper, on this view, confuses "bias" with brittle hand-engineering and therefore draws the wrong lesson.

7.5.0.1. Reply.. The objection is helpful because it highlights an ambiguity in how "structure" is discussed. The paper does not recommend adding narrow, domain-specific rules to LLMs. Instead, it argues that successful learning already depends on structural priors—including broad, reusable biases such as relational processing and compositional representation (Battaglia et al. 2018; Bronstein et al. 2021). A bias-aware empiricism is compatible with the bitter lesson precisely because it distinguishes between (i) ad hoc domain knowledge that fails to scale and (ii) general inductive biases that make scaling effective.

Indeed, the historical moral Sutton draws is not "no structure," but "structure that scales." From this perspective, acknowledging inductive bias is not a deviation from generality; it is a necessary step in explaining why general methods work in the first place. The paper's claim is therefore not in tension with the bitter lesson but supplies a philosophical interpretation of it: general methods succeed because their inductive biases align with deep regularities in the environment, and those biases remain indispensable even as data increases.

8. Conclusion

LLMs have often been treated as icons in a renewed empiricism-versus-nativism dispute. Their apparent "learning from text alone" is taken by some to vindicate radical empiricism, and by others to expose the hollowness of ungrounded pattern matching. This paper has argued that both readings risk missing the most instructive lesson. Contemporary machine learning makes clear that no successful learner is a

blank slate: inductive bias is necessary for generalization (Mitchell 1980; Wolpert and Macready 1997). LLMs, despite their data-driven appearance, incorporate strong biases distributed across architecture (Vaswani et al. 2017; Lavie, Gur-Ari, and Ringel 2024), objective functions (Brown et al. 2020), alignment procedures (Ouyang et al. 2022), optimization dynamics (Neyshabur et al. 2017; Soudry et al. 2018), representational bottlenecks such as tokenization (Sennrich, Haddow, and Birch 2016; Kudo and Richardson 2018), and even corpus construction choices that shape what counts as "experience" (Dodge et al. 2021; Gao et al. 2020). These biases are not incidental; they shape learnability and generalization at every scale (Kaplan et al. 2020; Hoffmann et al. 2022).

Once this is appreciated, the rhetorical force of a strong tabula rasa picture diminishes. If even the most "data-driven" generative models require substantial structural priors to learn from enormous corpora, then treating human infants as wholly unstructured learners becomes less plausible still. This does not entail a crude nativism about innate propositional content. Rather, it motivates a more methodologically responsible empiricism: experience is indispensable, but explanation must explicitly theorize the learner's prior structure and its interaction with input. Seen in this light, the current "LLMs understand or not" controversy is not an isolated quarrel; it is a new iteration of the old question whether semantic and conceptual structure can be generated from experience with form alone (Bender and Koller 2020; Harnad 1990; Bisk et al. 2020). The correct response is neither to declare victory for empiricism nor to retreat to slogans about innateness, but to adopt bias-aware explanations that treat structure as an indispensable part of the cognitive story.

Disclosure statement

No potential conflict of interest was reported by the author(s).

References

Akyürek, Ekin, Dale Schuurmans, Jacob Andreas, Tengyu Ma, and Denny Zhou. 2022. "What Learning Algorithm is In-Context Learning? Investigations with Linear Models." *arXiv preprint* https://arxiv.org/abs/2211.15661.

Battaglia, Peter W., Jessica B. Hamrick, Victor Bapst, Alvaro Sanchez-Gonzalez, Vinicius Zambaldi, Mateusz Malinowski, Andrea Tacchetti, et al. 2018. "Relational Inductive Biases, Deep Learning, and Graph Networks." *arXiv preprint* https://arxiv.org/abs/1806.01261.

Bender, Emily M., Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 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*, 610–623. https://dl.acm.org/doi/10.1145/3442188.3445922.

Bender, Emily M., and Alexander Koller. 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*, 5185–5198. https://aclanthology.org/2020.acl-main.463/.

Bisk, Yonatan, Ari Holtzman, Jesse Thomason, Jacob Andreas, Yoshua Bengio, Joyce Chai, Mirella Lapata, et al. 2020. "Experience Grounds Language." In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, 8718–8735. https://aclanthology.org/2020.emnlp-main.703/.

Bronstein, Michael M., Joan Bruna, Taco Cohen, and Petar Veličković. 2021. "Geomet-

ric Deep Learning: Grids, Groups, Graphs, Geodesics, and Gauges." *arXiv preprint* https://arxiv.org/abs/2104.13478.

Brown, Tom B., Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, et al. 2020. "Language Models are Few-Shot Learners." *arXiv preprint* https://arxiv.org/abs/2005.14165.

Bubeck, Sébastien, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, et al. 2023. "Sparks of Artificial General Intelligence: Early Experiments with GPT-4." *arXiv preprint* https://arxiv.org/abs/2303.12712.

Chomsky, Noam. 1959. "A Review of B. F. Skinner's *Verbal Behavior*." *Language* 35 (1): 26–58. https://www.jstor.org/stable/i217081.

Dodge, Jesse, Maarten Sap, Ana Marasović, William Agnew, Gabriel Ilharco, Dirk Groeneveld, Margaret Mitchell, and Matt Gardner. 2021. "Documenting Large Webtext Corpora: A Case Study on the Colossal Clean Crawled Corpus." In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 1286–1305. Association for Computational Linguistics. https://aclanthology.org/2021.emnlp-main.98/.

Gao, Leo, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, et al. 2020. "The Pile: An 800GB Dataset of Diverse Text for Language Modeling." *arXiv preprint* https://arxiv.org/abs/2101.00027.

Garg, Shivam, Dimitris Tsipras, Percy Liang, and Gregory Valiant. 2022. "What Can Transformers Learn In-Context? A Case Study of Simple Function Classes." *arXiv preprint* https://arxiv.org/abs/2208.01066.

Goyal, Anirudh, and Yoshua Bengio. 2022. "Inductive Biases for Deep Learning of Higher-Level Cognition." *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences* 478 (2266): 20210068. https://royalsocietypublishing.org/rspa/article/478/2266/20210068/56687/Inductive-biases-for-deep-learning-of-higher-level.

Harnad, Stevan. 1990. "The Symbol Grounding Problem." *Physica D: Nonlinear Phenomena* 42 (1–3): 335–346. https://doi.org/10.1016/0167-2789(90)90087-6.

Hoffmann, Jordan, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, et al. 2022. "Training Compute-Optimal Large Language Models." *arXiv preprint* https://arxiv.org/abs/2203.15556.

Kaplan, Jared, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. "Scaling Laws for Neural Language Models." *arXiv preprint* https://arxiv.org/abs/2001.08361.

Kudo, Taku, and John Richardson. 2018. "SentencePiece: A Simple and Language Independent Subword Tokenizer and Detokenizer for Neural Text Processing." In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 66–71. https://aclanthology.org/D18-2012/.

Lake, Brenden M., Tomer D. Ullman, Joshua B. Tenenbaum, and Samuel J. Gershman. 2017. "Building Machines That Learn and Think Like People." *Behavioral and Brain Sciences* 40: e253. https://doi.org/10.1017/S0140525X16001837.

Lavie, Itay, Guy Gur-Ari, and Zohar Ringel. 2024. "Towards Understanding Inductive Bias in Transformers: A View From Infinity." *arXiv preprint* https://arxiv.org/abs/2402.05173.

Locke, John. 1975. *An Essay Concerning Human Understanding*. Oxford: Clarendon Press. Clarendon Edition of the Works of John Locke.

Mitchell, Tom M. 1980. "The Need for Biases in Learning Generalizations." *Technical Report* Rutgers University, https://www.cs.cmu.edu/ tom/pubs/NeedForBias1980.pdf.

Neyshabur, Behnam, Ryota Tomioka, Ruslan Salakhutdinov, and Nathan Srebro. 2017. "In Search of the Real Inductive Bias: On the Role of Implicit Regularization in Deep Learning." *arXiv preprint* https://arxiv.org/abs/1705.03071.

Olsson, Catherine, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, et al. 2022. "In-context Learning and Induction Heads." *arXiv preprint* https://arxiv.org/abs/2209.11895.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya,

Florencia Leoni Aleman, et al. 2023. "GPT-4 Technical Report." *arXiv preprint* https://arxiv.org/abs/2303.08774.

Oswald, Johannes Von, Eyvind Niklasson, Ettore Randazzo, Joao Sacramento, Alexander Mordvintsev, Andrey Zhmoginov, and Max Vladymyrov. 2023. "Transformers Learn In-Context by Gradient Descent." In *Proceedings of the 40th International Conference on Machine Learning*, Vol. 202 of *Proceedings of Machine Learning Research*, 35151–35174. PMLR. https://proceedings.mlr.press/v202/von-oswald23a.html.

Ouyang, Long, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, et al. 2022. "Training Language Models to Follow Instructions with Human Feedback." In *Advances in Neural Information Processing Systems*, NeurIPS 2022, https://arxiv.org/abs/2203.02155.

Samet, Jerry. 2008. "The Historical Controversies Surrounding Innateness." https://plato.stanford.edu/entries/innateness-history/.

Searle, John R. 1980. "Minds, Brains, and Programs." *Behavioral and Brain Sciences* 3 (3): 417–457. https://www.cambridge.org/core/journals/behavioral-and-brain-sciences/article/minds-brains-and-programs/DC644B47A4299C637C89772FACC2706A.

Sennrich, Rico, Barry Haddow, and Alexandra Birch. 2016. "Neural Machine Translation of Rare Words with Subword Units." *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics* 1715–1725. https://doi.org/10.18653/v1/P16-1162, https://aclanthology.org/P16-1162/.

Shanahan, Murray. 2022. "Talking About Large Language Models." *arXiv preprint* https://arxiv.org/abs/2212.03551.

Soudry, Daniel, Elad Hoffer, Mor Shpigel Nacson, Suriya Gunasekar, and Nathan Srebro. 2018. "The Implicit Bias of Gradient Descent on Separable Data." *Journal of Machine Learning Research* 19 (70): 1–57. https://www.jmlr.org/papers/volume19/18-188/18-188.pdf.

Spelke, Elizabeth S. 2000. "Core Knowledge." *American Psychologist* 55 (11): 1233–1243. https://doi.org/10.1037/0003-066X.55.11.1233.

Sutton, Richard S. 2019. "The Bitter Lesson." Online essay. March 13, 2019, https://www.incompleteideas.net/IncIdeas/BitterLesson.html.

Tenenbaum, Joshua B., Charles Kemp, Thomas L. Griffiths, and Noah D. Goodman. 2011. "How to Grow a Mind: Statistics, Structure, and Abstraction." *Science* 331 (6022): 1279–1285. https://doi.org/10.1126/science.1192788.

Uzgalis, William. 2001. "John Locke." https://plato.stanford.edu/entries/locke/.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. "Attention Is All You Need." In *Advances in Neural Information Processing Systems*, NeurIPS 2017, https://arxiv.org/abs/1706.03762.

Wolpert, David H., and William G. Macready. 1997. "No Free Lunch Theorems for Optimization." *IEEE Transactions on Evolutionary Computation* 1 (1): 67–82. https://www.cs.ubc.ca/ hutter/earg/papers07/00585893.pdf.

Xie, Sang Michael, Aditi Raghunathan, Percy Liang, and Tengyu Ma. 2021. "An Explanation of In-context Learning as Implicit Bayesian Inference." *arXiv preprint* https://arxiv.org/abs/2111.02080.

Zhang, Chiyuan, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. 2017. "Understanding Deep Learning Requires Rethinking Generalization." *arXiv preprint* https://arxiv.org/abs/1611.03530.