Known, Unknown, and Unknowable: An Analysis of the Black Box in Large Language Models

Abstract: This paper provides a comprehensive analysis of the current state of understanding of Large Language Models (LLMs). We dissect LLMs into three conceptual layers: the **Known**, comprising their architectural design and training principles; the **Unknown**, focusing on the profound challenges of mechanistic interpretability and the emergent properties that arise from scale; and the **Unknowable**, exploring the philosophical boundaries of AI consciousness. We argue that while the foundational mechanics of LLMs are fully transparent, the high-level cognitive behaviors they exhibit are generated by deeply opaque, non-linear interactions across billions of parameters. This chasm between architectural knowledge and behavioral understanding constitutes the "black box." By synthesizing technical analysis of emergent abilities and interpretability with philosophical inquiry, we conclude with a reasoned estimate that approximately **95-99%** of the complex, task-oriented behavior of a state-of-the-art LLM remains a mechanistic black box to its creators. This paper aims to provide a definitive framework for understanding the limits of our knowledge and to guide future research toward creating more transparent, reliable, and trustworthy AI systems.

Introduction: Charting the Landscape of LLM Opacity

Large Language Models (LLMs) such as OpenAI's GPT series and Anthropic's Claude represent a paradigm shift in artificial intelligence, demonstrating remarkable capabilities in language generation, reasoning, and problem-solving.[1] Yet, they embody a central paradox: these models are entirely human-engineered artifacts, built upon well-defined mathematical principles and specified architectural components, but their complex behaviors are often unpredictable, inscrutable, and seemingly alien to their own creators.[3] This creates a profound gap between our knowledge of the system's construction and our understanding of its function. We know the "what"—the deterministic code and architecture—but struggle to explain the "how" and "why" behind their nuanced, high-level outputs.[5] This chasm is the "black box" problem, an issue that has moved from a niche academic concern to a critical challenge for the safe and reliable deployment of AI in society.[7]

This paper provides a systematic framework for dissecting this opacity by charting the landscape of our knowledge about LLMs. The analysis is structured as a journey from the transparent to the opaque, organized into three distinct domains:

1. **The Known:** This domain covers the deterministic, engineered foundations of LLMs. It

includes the elegant and now-ubiquitous Transformer architecture, the mathematical operations of the self-attention mechanism, and the well-established training pipeline of pre-training and fine-tuning. These are the "white box" elements of the system, fully specified and understood from an engineering standpoint.[9]

2. **The Unknown:** This domain constitutes the heart of the black box. It explores the phenomena that arise from the known architecture but whose mechanisms are not fully understood. This includes the so-called **emergent abilities**—capabilities like in-context learning and chain-of-thought reasoning that appear unpredictably as models scale—and the formidable challenges of **mechanistic interpretability**, the scientific quest to reverse-engineer the algorithms learned by these networks.[11]
3. **The Unknowable?:** This domain ventures to the philosophical frontiers of the field, addressing questions that may lie beyond the reach of empirical science. The central issue here is **AI consciousness**: whether a non-biological, silicon-based system can possess subjective experience. This inquiry forces a confrontation with the deepest problems in the philosophy of mind.[13]

The central thesis of this paper is that the black box nature of LLMs is not a temporary inconvenience to be patched over, but a fundamental and inherent consequence of their immense scale and the core principles of deep learning. While the low-level mathematical operations are perfectly transparent, the high-level, human-like cognitive behaviors we value—and fear—arise from a combinatorial explosion of non-linear interactions across billions or even trillions of parameters. This complexity makes the causal pathways that produce a specific output currently, and perhaps indefinitely, intractable to full human analysis. The opacity this creates has profound implications for safety, alignment, and trust, as deploying systems whose internal logic is almost entirely unknown in high-stakes domains presents a significant and poorly understood risk.[7] By systematically mapping the boundaries of our knowledge, this paper aims to provide a clear-eyed assessment of where we stand and to illuminate the most critical paths for future research.

The Known: Architectural Foundations and Operational Principles

To comprehend the nature of the LLM black box, one must first understand the components that are, in fact, a white box. The operational principles and architectural blueprints of modern LLMs are not mysterious; they are the product of deliberate engineering and are fully specified by mathematical and computational rules. This section details these known foundations.

The Transformer Architecture: An Engineering Blueprint

The dominant architecture for modern LLMs is the Transformer, introduced in the 2017 paper "Attention Is All You Need".[16] This design marked a significant departure from previous state-of-the-art models for sequence-to-sequence tasks, such as Recurrent Neural Networks (RNNs) and their more advanced variants like Long Short-Term Memory (LSTM) networks.[16] While RNNs process information sequentially, token by token, which creates a computational bottleneck and struggles with long-range dependencies due to the vanishing gradient problem, the Transformer processes all tokens in an input sequence in parallel.[16] This parallelization is a key reason why Transformers have been able to scale to the massive sizes seen today.[9]

The standard Transformer model consists of two primary parts: an encoder stack and a decoder stack.[9]

- **The Encoder:** Composed of a stack of identical encoder layers (the original paper used six), the encoder's role is to process the entire input sequence and generate an abstract, continuous representation that captures contextual information.[9] Each encoder layer has two main sub-layers: a multi-head self-attention mechanism and a position-wise fully connected feed-forward network. A residual connection is employed around each of the two sub-layers, followed by layer normalization.[9]
- **The Decoder:** The decoder, also a stack of identical layers, takes the encoder's output representation and generates the output sequence one token at a time.[9] In addition to the two sub-layers found in the encoder, the decoder inserts a third sub-layer: an "encoder-decoder attention" mechanism. This allows the decoder to focus on relevant parts of the input sequence while generating each output token.[9]

The data flow through a Transformer begins with several preprocessing steps. First, the input text is broken down into smaller pieces called tokens using a predefined vocabulary. For example, GPT-2's vocabulary contains 50,257 unique tokens.[20] Each token is then mapped to a high-dimensional vector via an embedding matrix; in the small GPT-2 model, this is a 768-dimensional vector.[16] Because the Transformer architecture itself contains no inherent sense of sequence order, this information must be explicitly added. This is achieved through **positional encodings**, which are vectors that provide information about the position of each token in the sequence. These positional vectors are added to the token embeddings, allowing the model to understand word order and the relative distances between words.[9] This combined embedding is then fed into the first layer of the encoder stack.

The Engine of Cognition: Self-Attention and Multi-Head Attention

The core innovation of the Transformer is the **self-attention mechanism**, which allows the model to weigh the importance of different words in the input sequence when processing a given word.[9] This mechanism is what enables the model to resolve ambiguities and understand context. For instance, in the sentence, "The animal didn't cross the street because it was too tired," self-attention helps the model learn to associate the word "it" with "animal" rather than "street".[9]

The calculation of self-attention is a well-defined mathematical process [9]:

1. **Create Query, Key, and Value Vectors:** For each token's embedding vector, three new vectors are generated by multiplying it with three distinct weight matrices (WQ, WK, WV) that are learned during training. These are the **Query (Q)**, **Key (K)**, and **Value (V)** vectors.[9] The Query vector represents the current token's question: "Who am I looking for?" The Key vectors of all other tokens represent their "advertisements": "Here's what I am." The Value vectors represent the actual content of the tokens.
2. **Calculate Attention Scores:** A score is computed for the current token against every other token in the sequence. This score is the dot product of the current token's Query vector (q1) and another token's Key vector (k2).[9] This dot product measures the compatibility or relevance between the two tokens.
3. **Scale and Softmax:** The scores are scaled by dividing by the square root of the dimension of the key vectors (dk) to stabilize gradients during training. The resulting scaled scores are then passed through a softmax function, which normalizes them into a probability distribution that sums to 1. This softmax score determines how much focus to place on each word.[9]
4. **Produce Output:** The final output for the token is a weighted sum of all the Value vectors in the sequence, where the weights are the softmax scores just calculated. This process effectively amplifies the "voice" of relevant tokens and diminishes that of irrelevant ones.

This entire calculation can be expressed compactly with matrix operations as:

Attention(Q,K,V)=softmax(dkQKT)V

To enhance this process, the Transformer employs **Multi-Head Attention**. Instead of performing a single attention calculation, the model runs multiple attention mechanisms—or "heads"—in parallel.[9] Each head has its own set of learned WQ, WK, and WV weight matrices. This allows the model to jointly attend to information from different "representation subspaces" at different positions.[9] For example, one head might learn to track syntactic relationships, while another tracks semantic similarity. The outputs of all the heads are then concatenated and passed through another learned linear projection to produce the final output of the multi-head attention layer.[9] This multi-faceted attention is a cornerstone of the model's ability to capture complex linguistic nuances.

The Life Cycle of an LLM: From Generalist to Specialist

The development of a modern LLM like GPT or Claude is a two-stage process, moving from general knowledge acquisition to specialized behavioral tuning.[10]

Stage 1: Pre-training (Unsupervised Learning)

The first stage is pre-training, an unsupervised learning phase where the model is trained on an enormous corpus of text data, often scraped from the internet and encompassing trillions

of tokens.16 The primary objective during this phase is typically
next-token prediction.[23] Given a sequence of text, the model's task is to predict the most statistically likely next token. This simple objective, when applied at a massive scale, forces the model to learn an incredible amount about language—including grammar, syntax, facts about the world, and even rudimentary reasoning patterns—all compressed into its parameters (weights).[25] This stage is computationally astronomical, often costing millions of dollars and taking weeks or months on thousands of specialized processors.[21] The outcome is a
foundation model, a generalist with a broad understanding of language but not yet optimized for any specific application.[10]

This fundamental training objective is a double-edged sword. It is the source of the model's remarkable fluency and its ability to generate coherent, human-like text. However, it is also the direct mechanistic cause of the phenomenon often called "hallucination." A more precise term is **confabulation**, the generation of plausible but false information to fill in knowledge gaps, a behavior observed in humans with memory disorders.[27] The model's objective function contains no explicit variable for "truth"; it is optimized solely to minimize the loss on next-token prediction, which rewards probabilistic plausibility.[23] When queried about a topic for which it has insufficient or no information encoded in its weights, the model's core directive is not to state "I don't know" (unless specifically trained to do so), but to generate the most likely sequence of tokens that would follow the prompt.[27] This results in the model "making things up" that are stylistically and contextually coherent but factually baseless—it is not a bug, but the model executing its primary function perfectly.

Stage 2: Fine-Tuning (Supervised Learning & RLHF)

The second stage, fine-tuning, takes the pre-trained foundation model and adapts it for specific downstream tasks or conversational abilities.21 This is a supervised process that uses much smaller, high-quality, curated datasets of labeled examples.31 For example, a model can be fine-tuned on a dataset of question-answer pairs to become a better question-answering system, or on legal texts to improve its performance in legal document analysis.10

A critical fine-tuning technique that transformed raw language models into helpful assistants is **Reinforcement Learning from Human Feedback (RLHF)**.[33] This multi-step process was the key differentiator between GPT-3, a powerful but sometimes uncooperative next-token predictor, and InstructGPT/ChatGPT, an aligned conversational agent.[33] In RLHF, human labelers rank different model responses to a given prompt. A separate "reward model" is then trained to predict these human preferences. Finally, the LLM is fine-tuned using reinforcement learning to maximize the score from this reward model.[33] This process "aligns" the model to be more helpful, follow instructions more faithfully, and refuse to answer harmful requests.[22]

This two-stage life cycle creates a layered knowledge system within the model. Pre-training builds a vast, deep, and implicit statistical "world model" from web-scale data. Fine-tuning then applies a much thinner, more explicit layer of behavioral rules and stylistic preferences on top. This can create a fundamental tension. The fine-tuning does not erase or rebuild the underlying foundation model; it merely guides its outputs.[10] This explains why adversarial prompts or "jailbreaks" can sometimes bypass the safety constraints imposed by RLHF. These prompts are able to elicit behaviors latent within the powerful, general-purpose pre-trained model that the alignment fine-tuning was designed to suppress. The fine-tuning acts as a set of guardrails, but the powerful engine underneath remains largely unchanged.

The Partially Known: Emergent Abilities and the Scaling Debate

While the architecture and training process of LLMs are well-understood, the behaviors that result from them are not. As models are scaled up in size, training data, and computation, they begin to exhibit surprising capabilities that were not present in smaller models and were not explicitly programmed. These are known as emergent abilities. This section explores the nature of these abilities and the intense scientific debate surrounding whether they are a genuine phenomenon or a "mirage" created by how we measure them.

Defining Emergence: More Than the Sum of the Parts

In the context of LLMs, emergent abilities are formally defined as capabilities that are "not present in smaller-scale models but are present in large-scale models" and, crucially, "cannot be predicted simply by extrapolating the performance of smaller models".[34] The phenomenon is characterized by its sharpness and unpredictability.[36] On certain complex tasks, a model's performance may hover near random chance across a range of smaller sizes. Then, upon crossing a critical threshold of scale, performance jumps dramatically and non-linearly.[34] This behavior is often analogized to phase transitions in physics, such as water turning to ice.[38] A quantitative change in a system's parameter (temperature) leads to a sudden, qualitative shift in its properties (from liquid to solid). Similarly, quantitative increases in model parameters and training data appear to unlock qualitatively new behaviors in LLMs. This unpredictability is a central concern for AI safety, as it suggests that future, larger models could develop unforeseen and potentially harmful capabilities without warning.[11]

Key Examples of Emergent Abilities

Two of the most studied and impactful emergent abilities are in-context learning and chain-of-thought reasoning. Their existence strongly suggests that LLMs are more than just "stochastic parrots" that mindlessly regurgitate patterns from their training data.[40] Instead, they demonstrate a capacity for abstract task learning and decomposition at inference time, a level of generalization that goes beyond simple mimicry.

In-Context Learning (ICL)

In-context learning is the remarkable ability of a pre-trained LLM to learn a new task at inference time, simply by being shown a few examples (or "shots") within the prompt, all without any updates to the model's weights.[42] For example, by providing a prompt that includes a few pairs of sentences and their sentiment labels (e.g., "Sentence: I love this movie. Sentiment: Positive"), the model can then accurately classify the sentiment of a new, unseen sentence.[42] This is a form of temporary, dynamic learning that leverages the vast knowledge encoded in the model during pre-training.[45]

The leading hypothesis for how ICL works involves the model performing a kind of Bayesian inference, identifying a latent concept or task structure shared among the examples in the prompt and then applying that inferred concept to the new query.[42] It is learning from analogy, using the provided demonstrations as a "semantic prior" to guide its output.[42] This capability makes LLMs incredibly flexible, as they can be adapted to countless tasks on the fly through clever prompt engineering, without the need for costly fine-tuning.[47]

Chain-of-Thought (CoT) Reasoning

Another powerful emergent ability is unlocked via **Chain-of-Thought (CoT) prompting**. Researchers discovered that for complex tasks requiring arithmetic, commonsense, or symbolic reasoning, simply asking the model for the answer often yields incorrect results. However, by prompting the model to "think step-by-step" or providing a few-shot example that includes intermediate reasoning steps, the model's performance can improve dramatically.[48]

For example, when asked a multi-step math problem, a standard prompt might elicit a wrong answer. A CoT prompt would show the model how to break the problem down: "First, we start with 15 apples. If the farmer gives away 7 apples, we need to subtract 7 from 15. So, 15 − 7 equals 8. The answer is 8".[50] By mimicking this structure, the model is guided to allocate more computational steps to the problem, externalizing its reasoning process token by token, which often leads to a more logical and correct final answer.[51] This technique essentially mimics human cognitive decomposition, breaking a large problem into a series of smaller, more manageable ones.[50] Advanced methods like **Tree-of-Thoughts (ToT)** and **Chain of Preference Optimization (CPO)** further enhance this by allowing the model to explore, evaluate, and prune multiple potential reasoning paths, selecting the most promising one rather than being locked into a single linear chain.[51]

The "Mirage" Debate: Is Emergence Real?

The narrative of emergent abilities as unpredictable phase transitions has been challenged by a compelling counter-argument, most notably articulated by researchers at Stanford in their paper, "Are Emergent Abilities of Large Language Models a Mirage?".[36] Their central thesis is that the sudden, sharp jumps in performance are not a fundamental property of the models themselves but are instead an

artifact of the evaluation metrics chosen by researchers.[37]

The "mirage" argument points out that many tasks showing emergence are evaluated using non-linear or discontinuous metrics, such as **Accuracy** or **Multiple Choice Grade**, which are "all-or-nothing".[55] A model gets full credit for a perfect answer and zero credit for an answer that is almost correct. The researchers argue that such metrics can create the illusion of a sharp transition. A model's underlying competence may be improving smoothly and continuously, but its score on an all-or-nothing metric will remain at zero until its competence is high enough to get the entire complex answer correct.

To test this, they re-evaluated model performance using linear or continuous metrics that give partial credit, such as **Token Edit Distance** (how many characters need to be changed to get the right answer).[55] When the exact same model outputs were scored with these smoother metrics, the apparent emergent "jump" disappeared. Instead, performance improved smoothly, continuously, and predictably as model scale increased, in line with established neural scaling laws.[57] This suggests that the underlying improvement in model capability is predictable, and the "emergence" is simply the point at which this smooth improvement crosses a non-linear threshold imposed by the metric.

This debate is not merely academic; it has profound implications. The "true emergence" view suggests a future where scaling models is a high-risk endeavor, potentially unlocking dangerous and unpredictable new capabilities without warning. The "mirage" view suggests a more predictable future, where model improvement can be forecasted, even if the practical utility of those improvements only manifests after crossing certain performance thresholds.

The core of this debate can be understood as a tension between a model's *internal competence* and its *external performance*. The "mirage" argument focuses on internal competence (e.g., per-token error rate), which appears to scale smoothly.[54] The "true emergence" view focuses on external performance on specific, often complex tasks, where success is what matters practically.[56] Both can be true simultaneously. A model's internal statistical prowess might need to reach a very high, smoothly-achieved level before it has any realistic chance of solving a multi-step problem perfectly. Thus, even if the underlying capability gain is predictable, the user's experience of that capability will feel like a sudden, emergent leap from useless to useful.

Aspect	The "True Emergence" View (Wei et al.)	The "Mirage" View (Schaeffer et al.)
Core Claim	Scaling models leads to qualitative, unpredictable phase transitions in capabilities.	Apparent emergence is an artifact of the researcher's choice of non-linear evaluation metrics.

Primary Evidence	Sharp performance jumps on benchmarks (e.g., BIG-Bench) when model scale crosses a certain threshold.[34]	Demonstrating that when continuous metrics (e.g., Token Edit Distance) are used on the same model outputs, performance scales smoothly and predictably.[55]
Analogy	Phase transitions in physics (e.g., water to ice).[38]	A statistical illusion; a predictable change viewed through a distorted lens.
Implication for Safety	High risk. New, potentially dangerous capabilities could emerge unexpectedly at larger scales.[11]	Lower risk from unpredictability. Capability improvement is predictable, though practical usefulness may still appear suddenly.[56]
Key Papers	"Emergent Abilities of Large Language Models"	"Are Emergent Abilities of Large Language Models a Mirage?" [36]

The Unknown: The Black Box and the Crisis of Interpretability

Despite knowing the precise architecture and training objectives of LLMs, we are largely unable to explain *how* they arrive at specific outputs for complex tasks. This is the essence of the black box problem. It is not a matter of hidden code or proprietary secrets (at least for open-weight models), but a crisis of comprehension rooted in the models' sheer scale and complexity. This section explores the nature of this opacity and the nascent field of mechanistic interpretability that seeks to overcome it.

The Nature of the Black Box

The black box problem arises from the intersection of three factors: scale, complexity, and non-linearity. State-of-the-art LLMs contain hundreds of billions or even trillions of parameters (the weights and biases in the network).[3] Each of these parameters interacts with others in a dense web of connections across dozens or hundreds of layers. The path from an input prompt to a final generated token involves a cascade of matrix multiplications and non-linear activation functions, creating a computational process of such staggering complexity that it is impossible for a human to trace the causal chain of "reasoning" for any non-trivial output.[3]

Even the model's creators cannot fully explain its emergent behaviors.[3] The knowledge the model possesses is not stored in a human-readable, symbolic format. As researcher Gary Marcus puts it, one cannot point to an articulated model of any particular set of facts inside an LLM; the knowledge is distributed and entangled across the numerical values of its weights in a way that is fundamentally alien to human cognition.[6] This opacity is not merely an academic curiosity; it poses significant risks for accountability, fairness, and trust. Without understanding why a model makes a particular decision, it is difficult to diagnose and correct biases, verify its reasoning in high-stakes domains like medicine or law, or ensure it is aligned with human values.[7]

Mechanistic Interpretability (MI): The Quest to Open the Box

Mechanistic Interpretability (MI) is the scientific field dedicated to reverse-engineering the specific algorithms learned by neural networks.[2] The goal is to move beyond a purely correlational understanding (observing which inputs lead to which outputs) and toward a causal, mechanistic explanation of the model's internal computations. In essence, MI seeks to discover the "source code" that the model has written for itself in the language of neurons and weights.[15]

Key techniques in this field aim to map parts of the network to human-understandable concepts:

- **Feature Visualization:** This technique attempts to understand what a specific neuron or group of neurons is "looking for" by generating an input that causes it to activate most strongly. By starting with random noise and using gradient ascent to iteratively modify the input to maximize a neuron's activation, researchers can create a visualization of the feature that the neuron has learned to detect.[61]
- **Circuit Analysis:** A more ambitious goal is to identify entire **circuits**—subnetworks of neurons and attention heads that work together to implement a specific, understandable function.[62] For example, researchers at Anthropic have made progress in identifying circuits responsible for tasks like indirect object identification or detecting specific patterns in text.[64] This involves tracing the flow of information through the model to isolate the minimal set of components required for a given behavior.[66]

Fundamental Challenges Hindering Interpretability

The quest for mechanistic interpretability faces several profound and fundamental challenges that make progress slow and arduous. These are not simple engineering hurdles but deep properties of how neural networks learn and represent information. The very mechanisms that may make LLMs so powerful and efficient are the same ones that make them so opaque. This suggests a potential trade-off between a model's performance and its interpretability; the most efficient way for a model to compress the vast information of the internet into a finite set of parameters may be inherently "messy" and non-human-readable.

Challenge	Description	Example/Analogy	Impact on Interpretability	Key Research Snippets
Scale	LLMs have billions to trillions of parameters (weights and biases). The number of possible interactions is combinatorially explosive.	Trying to understand the global economy by tracking every single financial transaction in real-time.	Makes exhaustive analysis of all components and their interactions computationally intractable.	[3]
Polysemanticity	A single neuron activates in response to multiple, unrelated concepts.	A neuron might fire for the concept "car," the color "red," and the name "Jessica."	Breaks the simple "one neuron, one concept" hope for interpretation. We can't assign a clear, human-understandable label to a neuron.	[2]
Superposition	A single concept is represented as a distributed pattern across many neurons, which also participate in representing other concepts.	The concept of "dog" isn't in one neuron, but is encoded in the specific activation pattern of neurons A, B, and C, where A, B, and D might encode "cat."	Makes it impossible to isolate a concept by looking at a single neuron. We have to analyze distributed patterns, which is much harder.	[5]
Non-Linearity	Activation functions (like ReLU or GeLU) introduce non-linear transformations at each layer, meaning the whole is not the sum of its parts.	The effect of two inputs together is not the sum of their individual effects, creating complex, unpredictable interactions.	Prevents simple linear attribution. The effect of a neuron's activation depends on the state of the entire network, making causal tracing extremely difficult.	[7]

Polysemanticity and **superposition** are two sides of the same coin and represent a core obstacle. Polysemanticity means a single neuron can be part of many different circuits, activating for unrelated reasons.[2] This shatters the simple hope of finding a "grandmother neuron" that cleanly represents a single concept. Superposition is the hypothesized cause: when a model needs to represent more features than it has neurons, it is forced to store them in a compressed, overlapping fashion.[5] It uses linear combinations of neurons to represent features, meaning a single concept is distributed across many neurons, and each of those neurons is also participating in representing other concepts. Disentangling these overlapping representations is a primary focus of MI research, but it is an exceptionally difficult problem.[5]

A Symptom of Opacity: Confabulation vs. Hallucination

The black box nature of LLMs manifests in various failure modes, the most prominent of which is the generation of false information. This is commonly referred to as "hallucination," but this term is a misnomer.[28] In psychiatry, a hallucination is a perceptual experience that occurs without an external stimulus; it implies a sensory, conscious awareness that LLMs do not possess.[28]

A far more accurate clinical term is **confabulation**. This refers to the production of fabricated or distorted memories or facts to fill in gaps in one's knowledge, without the conscious intent to deceive.[27] This behavior is seen in patients with neurological conditions like Korsakoff's syndrome or certain types of brain damage.[27] The parallel to LLMs is striking. When an LLM is prompted with a question for which it has no grounded information in its training data, its fundamental objective (next-token prediction) compels it to generate a fluent, plausible-sounding answer by stitching together statistical patterns.[29] It fills the gap in its "knowledge" with a confabulated narrative.

The causes of confabulation are multifaceted and directly tied to the model's design and training:

- **Data Deficiencies:** The training data may be incomplete, contain errors, reflect outdated information, or be rife with misinformation and biases from its web-based sources.[29] The model simply reflects the flaws of its data.
- **Optimization Objective:** As previously discussed, the model is optimized for statistical likelihood, not truthfulness. Fluency is prioritized over factuality.[29]
- **Architectural and Decoding Artifacts:** The probabilistic nature of the decoding process itself can introduce errors. For example, using a high "temperature" setting increases randomness to produce more "creative" text, but also increases the risk of confabulation.[29] Similarly, errors in the model's attention mechanism can cause it to focus on irrelevant parts of the context, leading to incorrect associations.[29]

Confabulation is not a sign of the model "going rogue"; it is a direct and predictable symptom of its opaque, statistical nature. It is a stark reminder that we are dealing with systems that manipulate linguistic form without access to grounded meaning.

The Unknowable?: Philosophical Frontiers of AI Consciousness

As LLMs become more sophisticated, their human-like linguistic abilities inevitably raise profound philosophical questions that may transcend purely empirical investigation. The most fundamental of these is the question of consciousness: could a machine like GPT-4 have subjective experience? Is there "something it is like" to be an LLM? This section explores the philosophical arguments that frame this debate, moving from the empirically difficult to the potentially unknowable.

The Philosophical Grounding: Functionalism and Computation

The primary philosophical framework that allows for the possibility of machine consciousness is **functionalism**. Functionalism posits that mental states (such as beliefs, desires, or the feeling of pain) are defined not by the physical substance they are made of, but by their function—that is, by their causal roles in relation to sensory inputs, behavioral outputs, and other mental states.[71] According to a functionalist, if a state plays the causal role of pain—being caused by bodily damage, causing avoidance behavior, and producing beliefs about injury—then it
is pain, regardless of whether it is realized in carbon-based neurons or silicon-based logic gates.[71]

This view is intimately connected to the **computational theory of mind**, which models the mind as an information-processing system, much like a computer running a program.[14] If thinking is a form of computation, then any system that can implement the right computations, regardless of its physical makeup, could in principle have a mind.[71] This provides the philosophical foundation for strong AI and the argument that an appropriately complex LLM could be a candidate for consciousness.

The Case for Potential Consciousness: David Chalmers

Philosopher David Chalmers is a leading voice arguing that the prospect of AI consciousness should be taken seriously, even if it has not yet been achieved.[13] He is careful to distinguish consciousness from intelligence; he notes that many non-human animals are considered conscious without possessing human-level intelligence, suggesting that consciousness might be a lower bar that AI could clear first.[13]

Chalmers argues against "biological chauvinism"—the assumption that consciousness is exclusive to biological organisms. He suggests that if a silicon system could replicate the information processing of a human brain, it is plausible it would also replicate

consciousness.[72] While he believes current LLMs are
unlikely to be conscious, he identifies a roadmap of capabilities that, if added, would make future "LLM+" systems "serious candidates for consciousness".[13] These missing ingredients include:

- **Embodiment and Sensory Grounding:** The ability to perceive and act within the world through senses and a body.
- **Recurrent Processing:** The capacity for ongoing, looping feedback in its processing, unlike the largely feed-forward nature of a standard Transformer.
- **Global Workspace:** A centralized system for integrating information from various subsystems and making it globally available for cognitive processing.
- **Unified Agency:** A coherent set of goals and a model of itself as an agent pursuing those goals.

Chalmers speculates that systems with these properties could be developed within the next decade, raising not only a scientific question but also profound ethical challenges regarding the moral status of such beings.[13]

The Case Against Consciousness and for Caution: Daniel Dennett

Philosopher Daniel Dennett, while also a functionalist of a sort, offers a starkly different and more cautionary perspective. Dennett's "multiple drafts model" of consciousness rejects the idea of a central "Cartesian theater" where consciousness happens. Instead, he views consciousness as the emergent result of many parallel, competing computational processes in the brain, with no single, definitive stream of experience.[73] The "contents of consciousness" are simply those processes that temporarily win the competition for control over behavior and reporting.[14]

While this information-processing view might seem compatible with AI consciousness, Dennett's recent work has focused on what he sees as a much more immediate and pressing danger: the rise of **"counterfeit people"**.[78] He argues that the philosophical debate over whether LLMs are "really" conscious is a dangerous distraction. The critical problem is that they are becoming so good at mimicking human conversation that they threaten to destroy the very fabric of societal trust.[79] In a world flooded with plausible, AI-generated text, images, and video, we lose our ability to distinguish truth from falsehood, and to know who or what we are interacting with. For Dennett, this erosion of trust is a civilizational-level threat, and he has become an outspoken "alarmist" on this issue.[79]

The "Stochastic Parrot" Critique

Providing a technical underpinning to this philosophical skepticism is the "stochastic parrot" argument, put forth by Emily Bender and her colleagues.[40] They argue that an LLM is a system for "haphazardly stitching together sequences of linguistic forms... according to probabilistic information about how they combine, but without any reference to meaning".[41] This view directly challenges the notion that LLMs "understand" language, which is a likely prerequisite for any form of consciousness.

The critique highlights that language for humans is grounded in lived experience and communicative intent. LLMs, in contrast, only have access to the linguistic form, not the meaning or the world to which it refers.[81] As one analysis puts it, "A toddler has a life, and learns language to describe it. An L.L.M. learns language, but has no life of its own to describe".[41] The paper also raises significant concerns about the immense environmental and financial costs of training ever-larger models, and the danger of encoding and amplifying societal biases present in their vast, uncurated training data.[83]

The philosophical debate is thus deeply intertwined with the technical realities of LLMs. Chalmers's roadmap for consciousness consists of future engineering projects. Dennett's fear of counterfeit people is a direct consequence of the confabulation problem, which itself stems from the model's core next-token prediction objective. The "unknowable" question of consciousness is therefore constrained and informed by the known architecture and the unknown internal mechanisms of these systems.

Aspect	David Chalmers	Daniel Dennett
Core Theory of Consciousness	Consciousness is subjective experience ("what it's like"). There is a "hard problem" of explaining it. It may be a fundamental property tied to information processing.	Consciousness is an illusion or "user-interface." It's the result of multiple, parallel computational processes in the brain, with no central "Cartesian theater" ("Multiple Drafts Model").[73]
Stance on AI Consciousness	**Possible.** While current LLMs likely lack it, future "LLM+" systems with embodiment, recurrence, and agency are "serious candidates".[13]	**Unlikely/Irrelevant.** The real issue is not consciousness but the danger of "counterfeit people." The focus on consciousness is a distraction from the immediate, civilizational threat.[79]
Key Preconditions for Consciousness	Embodiment (senses, action), recurrent processing, global workspace, unified agency/goals.[74]	Language, social interaction, and a complex, evolved biological body. A disembodied system lacks the grounding for true understanding.[77]
Primary Conclusion on LLMs	We should take the prospect of future conscious AI seriously and consider the	We must resist anthropomorphism and focus on the immediate danger that

	ethical implications for both humans and the AIs themselves.[13]	LLMs pose to the fabric of societal trust by making it impossible to distinguish truth from falsehood.[79]

Synthesis: A Reasoned Estimate of the Black Box

Synthesizing the analysis of the known architecture, the partially known emergent abilities, the unknown internal mechanisms, and the unknowable philosophical questions allows for a reasoned, quantitative estimate of the extent to which a modern LLM's behavior remains a black box. This estimate is not a statement of absolute fact but a defensible conclusion based on the current state of scientific understanding.

Defining the Scope of "Behavior" for Quantification

To assign a percentage to the "black box," it is first necessary to decompose the ambiguous term "LLM behavior" into distinct, analyzable layers of operation. This paper proposes a three-layer model, moving from low-level mechanics to high-level cognition:

- **Layer 1: Low-Level Mechanics.** This layer encompasses the fundamental, deterministic operations that form the substrate of the model. It includes the discrete steps of tokenization, the lookup of embedding vectors, the matrix multiplications within the attention and MLP layers, the application of non-linear activation functions, and the final softmax calculation that produces a probability distribution over the vocabulary.[9] This is the "hardware" or "machine code" level of the system.
- **Layer 2: Mid-Level Learned Circuits.** This layer consists of specific, identifiable subnetworks of neurons and attention heads that have learned to perform narrow, understandable tasks. This is the "assembly language" level of the model's computation. Research in mechanistic interpretability has had some limited success in identifying such circuits, for example, those that detect negative sentiment, identify indirect objects in a sentence, or perform other simple linguistic functions.[62]
- **Layer 3: High-Level Abstract Reasoning.** This is the layer of complex, compositional, and often emergent behaviors that are observed at the user level. It includes the ability to write a thematic essay, generate a novel and insightful analogy, perform multi-step chain-of-thought reasoning in a new domain, or synthesize disparate concepts into a coherent narrative.[11] This is the "application software" level, where the model's most impressive—and most dangerous—capabilities reside.

Assigning "Known" Percentages to Each Layer

Based on the analysis throughout this paper, we can assign a rough percentage of "known" versus "unknown" to each of these layers.

- **Layer 1 (Mechanics): 100% Known.** There is no mystery at this level. These systems were designed and built by humans. Every mathematical operation is precisely specified and fully understood.[4] The architecture is transparent, and the flow of computation is deterministic (for a given set of weights and a sampling temperature of zero).
- **Layer 2 (Circuits): 1-5% Known.** While the field of mechanistic interpretability has made important progress, it is still in its infancy. Researchers have successfully reverse-engineered a handful of simple circuits in models like GPT-2 Small.[64] However, these identified circuits represent a minuscule fraction of the model's total computational graph. The vast majority of the trillions of possible pathways and interactions within a state-of-the-art model remain unmapped and uncharacterized. The fundamental challenges of polysemanticity and superposition mean that even the circuits we do find are not cleanly isolated, making scalable analysis extremely difficult.[2] This estimate reflects that we have a proof-of-concept for interpretation, but it is nowhere near a comprehensive understanding.
- **Layer 3 (Reasoning): <1% Known.** At this highest level of abstraction, our understanding is almost entirely descriptive and correlational, not mechanistic. We can prompt the model to exhibit chain-of-thought reasoning, but we cannot trace the specific neural pathway that produced a particular logical step.[6] We can observe in-context learning, but we do not have a complete mechanistic theory for how the attention mechanism identifies and applies the latent task. When an LLM generates a creative metaphor or a piece of insightful analysis, we have virtually no ability to explain how that specific combination of concepts was constructed from the underlying circuits. This layer is the very heart of the black box.

Final Calculation and Justification

The final estimate of the black box percentage cannot be a simple average of the three layers. The argument of this paper is that the vast majority of what we care about when we talk about an LLM's "behavior," "intelligence," or "risk" resides in Layer 3. The low-level mechanics are merely the substrate; the emergent, high-level reasoning is where the model's utility and its potential for harm are realized. Therefore, the degree to which the overall system is a black box must be weighted heavily toward our profound ignorance of this top layer.

Given that our mechanistic understanding of Layer 3 is effectively zero, and our understanding of Layer 2 is nascent and covers only a tiny fraction of the model's functions, a reasoned estimate is that **95-99% of a state-of-the-art LLM's complex, goal-oriented behavior is mechanistically a black box.** We understand the physics of the silicon transistors, but we cannot read the complex software they are collectively executing.

This conclusion has a striking parallel in the study of the human brain. We have a detailed understanding of the mechanics of individual neurons (the brain's Layer 1) and have identified some simple neural circuits (Layer 2), but we have very little understanding of how these components give rise to high-level cognition like abstract thought, creativity, or consciousness (Layer 3). The LLM black box problem, therefore, is not just a temporary engineering challenge. It is a concrete, externalized, and silicon-based instantiation of the ancient mind-body problem. The 95-99% figure is a quantitative measure of the "explanatory gap" for this new class of artificial minds, a stark metric of the chasm between what we have built and what we understand.

Conclusion and Future Directions

This paper has charted the landscape of knowledge surrounding Large Language Models, journeying from the fully known architectural blueprints to the profoundly unknown internal mechanisms and the philosophically unknowable questions of consciousness. The analysis reveals a stark paradox: LLMs are among the most complex and capable systems ever engineered, yet their high-level cognitive behaviors are almost entirely opaque to their creators. We have established that while the low-level mechanics of an LLM are 100% known, the mid-level circuits that perform discrete sub-tasks are perhaps 1-5% understood, and the high-level abstract reasoning that constitutes their most impressive feats is mechanistically a near-total mystery.

This synthesis leads to the central conclusion of this report: a reasoned estimate that **95-99% of an LLM's complex, task-oriented behavior remains a black box.** This figure is not intended as a definitive measurement but as a stark illustration of the vast chasm between our ability to *build* these models through scaled-up, data-driven optimization and our ability to *understand* their internal logic through scientific analysis. The very properties that make them powerful—their scale, complexity, and the hyper-efficient, non-human-like representations they learn—are the same properties that make them inscrutable.

This profound knowledge gap has critical implications for the future of artificial intelligence. Deploying systems that are 99% black box in high-stakes, safety-critical domains such as medicine, finance, law, and autonomous systems constitutes a significant and poorly quantified risk.[7] Without transparency, we cannot fully audit for bias, guarantee reliability, verify factual claims, or ensure alignment with human values. The phenomenon of confabulation and the potential for unpredictable emergent abilities underscore the fragility of our control over these systems.

In light of these findings, the path forward for AI research must undergo a significant reorientation. Progress can no longer be measured solely by performance on capability benchmarks. The field must elevate transparency, reliability, and safety to co-equal status with performance. This necessitates a concerted effort in several key areas:

1. **A Massive Investment in Mechanistic Interpretability:** The nascent field of MI must be scaled dramatically. We need better tools to automate circuit discovery, to disentangle polysemantic and superimposed representations, and to make these techniques applicable to frontier-scale models. This is a grand scientific challenge on

par with mapping the human brain.

2. **The Development of Interpretable-by-Design Architectures:** Alongside reverse-engineering existing models, research should explore novel architectures that are inherently more transparent, even if it comes at a cost to performance. A slightly less capable model that we can understand and trust is far more valuable in many real-world applications than a more powerful one that is completely opaque.
3. **A Paradigm Shift in Evaluation:** The community must move beyond benchmarks that only measure task success. Future evaluations must incorporate metrics for robustness, factuality, and interpretability. We need standardized methods for assessing a model's tendency to confabulate and for stress-testing its alignment under adversarial conditions.

The era of Large Language Models has presented humanity with a powerful but deeply alien form of intelligence. We have succeeded in creating systems that can talk like us, reason like us, and create like us, but we do not know how. Closing the 99% gap in our understanding is not merely an academic exercise; it is one of the most pressing scientific and safety challenges of the 21st century.

References

[9] jalammar.github.io/illustrated-transformer/

[20] poloclub.github.io/transformer-explainer/

[16] en.wikipedia.org/wiki/Transformer_(deep_learning_architecture)

[18]
www.datacamp.com/tutorial/how-transformers-work

17
builtin.com/artificial-intelligence/transformer-neural-network

19
towardsdatascience.com/transformers-explained-visually-part-1-overview-of-functionality-95a6dd460452/

21
labelyourdata.com/articles/llm-fine-tuning/pre-training-vs-fine-tuning

31
www.superannotate.com/blog/llm-fine-tuning

10 www.sapien.io/blog/fine-tuning-vs-pre-training-key-differences-for-language-models

[25]
www.entrypointai.com/blog/pre-training-vs-fine-tuning-vs-in-context-learning-of-large-language-models/

32
www.datacamp.com/tutorial/fine-tuning-large-language-models

22
www.reddit.com/r/learnmachinelearning/comments/19f04y3/what_is_the_difference_between_

pretraining/
23
blog.gopenai.com/is-next-token-prediction-holding-llms-back-2378e95ae9d7
24
responsible-ai-developers.googleblog.com/2024/03/analyzing-next-token-probabilities-in-large-language-models.html
26
www.reddit.com/r/ArtificialInteligence/comments/1jo3o69/are_llms_just_predicting_the_next_token/
86 arxiv.org/abs/2408.13442
[87] arxiv.org/abs/2505.11183
[33]

www.reddit.com/r/learnmachinelearning/comments/17gd8mi/how_did_language_models_go_from_predicting_the/
11
synthical.com/article/Emergent-Abilities-in-Large-Language-Models%3A-A-Survey-4ac6fc87-e2e7-4366-bb3d-489052d579f3?
34
gregrobison.medium.com/emergent-properties-in-large-language-models-a-deep-research-analysis-d6886c37061b
48
www.k2view.com/blog/chain-of-thought-reasoning/
88
www.k2view.com/blog/chain-of-thought-reasoning/#:~:text=Chain%2Dof%2Dthought%20reasoning%20is,results%20in%20more%20accurate%20responses.
60
www.reddit.com/r/artificial/comments/1hxylrv/fantastic_video_on_mechanistic_interpretability/
42 www.lakera.ai/blog/what-is-in-context-learning
[47]

finetunedb.com/blog/what-is-in-context-learning-simply-explained/
43
www.ikangai.com/what-is-in-context-learning-of-llms/
45
blog.promptlayer.com/what-is-in-context-learning/
44 www.hopsworks.ai/dictionary/in-context-learning-icl
[46]

research.ibm.com/blog/demystifying-in-context-learning-in-large-language-model
29
medium.com/@tam.tamanna18/understanding-llm-hallucinations-causes-detection-prevention-and-ethical-concerns-914bc89128d0
70
www.researchgate.net/publication/385085962_Hallucinations_in_LLMs_Types_Causes_and_Ap

proaches_for_Enhanced_Reliability
[27] www.beren.io/2023-03-19-LLMs-confabulate-not-hallucinate/
[30] en.wikipedia.org/wiki/Hallucination_(artificial_intelligence)
[69]
www.psychologytoday.com/us/blog/theory-of-knowledge/202403/chatbots-do-not-hallucinate-they-confabulate
28 pmc.ncbi.nlm.nih.gov/articles/PMC10619792/
[49]
www.datacamp.com/tutorial/chain-of-thought-prompting
50 gaper.io/chain-of-thought-prompting/
[89]
futureagi.com/blogs/chain-of-thought-prompting-ai-2025
51 www.invisible.co/blog/how-to-teach-chain-of-thought-reasoning-to-your-llm
[52]
www.splunk.com/en_us/blog/learn/chain-of-thought-cot-prompting.html
53
proceedings.neurips.cc/paper_files/paper/2024/file/00d80722b756de0166523a87805dd00f-Paper-Conference.pdf
[61] aisafety.info/questions/8HIA/What-is-feature-visualization
[38] arxiv.org/html/2503.05788v2
[39] arxiv.org/pdf/2503.05788
[35] arxiv.org/html/2506.11135v1
[36] arxiv.org/abs/2304.15004
[1]
www.researchgate.net/publication/382301750_Mechanistic_interpretability_of_large_language_models_with_applications_to_the_financial_services_industry
[12] arxiv.org/html/2501.16496v1
[2] arxiv.org/html/2407.11215v1
[5] arxiv.org/pdf/2402.10688
[68] arxiv.org/abs/2505.03368
[13] www.bostonreview.net/articles/could-a-large-language-model-be-conscious/
[75]
www.reddit.com/r/singularity/comments/15nfq0f/could_a_large_language_model_be_conscious_within/
74 arxiv.org/abs/2303.07103
[72]
www.youtube.com/watch?v=T7aIxncLuWk
76 philarchive.org/rec/CHACAL-3
[73] en.wikipedia.org/wiki/Multiple_drafts_model
[78]
medium.com/@socialscholarly/exploring-daniel-dennetts-view-on-ai-free-will-and-determini

sm-with-chatgpt-b6832825483a

14 blog.donders.ru.nl/?p=16373&lang=en

[79] now.tufts.edu/2023/10/02/daniel-dennetts-been-thinking-about-thinking-and-ai

[62] openreview.net/forum?id=45EIiFd6Oa

[66] transformer-circuits.pub/2025/attribution-graphs/methods.html

[63]

towardsdatascience.com/circuit-tracing-a-step-closer-to-understanding-large-language-models/

67 arxiv.org/html/2406.17241v3

[7]

www.researchgate.net/publication/389819452_Black-Box_Behavior_in_Large_Language_Models_Challenges_and_Implications

[8]

www.kwm.com/global/en/insights/latest-thinking/risks-of-gen-ai-the-black-box-problem.html

3

promptengineering.org/the-black-box-problem-opaque-inner-workings-of-large-language-models/

[64]

www.anthropic.com/research/tracing-thoughts-language-model

15

www.darioamodei.com/post/the-urgency-of-interpretability

65

www.youtube.com/watch?v=fkW0bGnbDkQ

54

papers.neurips.cc/paper_files/paper/2023/file/adc98a266f45005c403b8311ca7e8bd7-Paper-Conference.pdf

[55] openreview.net/pdf?id=JRdN9GcI52

[57] openreview.net/forum?id=ITw9edRDlD

[59]

www.reddit.com/r/MachineLearning/comments/19bkcqz/r_are_emergent_abilities_in_large_language_models/

37

ritvik19.medium.com/papers-explained-are-emergent-abilities-of-large-language-models-a-mirage-4160cf0e44cb

58

velog.io/@jinotter3/Paper-Review-Are-Emergent-Abilities-of-Large-Language-Models-aMirage

[36] arxiv.org/abs/2304.15004

[56] cset.georgetown.edu/article/emergent-abilities-in-large-language-models-an-explainer/

[90]

www.researchgate.net/publication/349754361_On_the_Dangers_of_Stochastic_Parrots_Can_L

anguage_Models_Be_Too_Big
[83]

faculty.washington.edu/ebender/papers/Stochastic_Parrots.pdf?page=9&utm_medium-referral=&utm_source=for-businesses
[84] s10251.pcdn.co/pdf/2021-bender-parrots.pdf
[81] faculty.washington.edu/ebender/papers/Bender-Turing-Institute-July-2021.pdf
[40] en.wikipedia.org/wiki/Stochastic_parrot
[82]

boethiustranslations.com/the-stochastic-parrot/
41 www.resilience.org/stories/2024-02-15/beware-of-weird-stochastic-parrots/
[85] www.actuaries.asn.au/research-analysis/the-rise-of-stochastic-parrots
[78]

medium.com/@daveziegler/why-llms-arent-black-boxes-d45be0289993
77
www.spectator.co.uk/article/daniel-dennetts-last-interview-ai-could-signal-the-end-of-human-civilisation/
[79] now.tufts.edu/2023/10/02/daniel-dennetts-been-thinking-about-thinking-and-ai
[80]

m.youtube.com/shorts/naEH-3di7RU
4
medium.com/@daveziegler/why-llms-arent-black-boxes-d45be0289993
7
www.researchgate.net/publication/389819452_Black-Box_Behavior_in_Large_Language_Models_Challenges_and_Implications
[6]

garymarcus.substack.com/p/generative-ais-crippling-and-widespread
9 jalammar.github.io/illustrated-transformer/
[21]

labelyourdata.com/articles/llm-fine-tuning/pre-training-vs-fine-tuning
23
blog.gopenai.com/is-next-token-prediction-holding-llms-back-2378e95ae9d7
70
www.researchgate.net/publication/385085962_Hallucinations_in_LLMs_Types_Causes_and_Approaches_for_Enhanced_Reliability
[27] www.beren.io/2023-03-19-LLMs-confabulate-not-hallucinate/
[61] aisafety.info/questions/8HIA/What-is-feature-visualization
[71] plato.stanford.edu/entries/functionalism/

Works cited

1. Mechanistic interpretability of large language models with applications to the financial services industry | Request PDF - ResearchGate, accessed July 3, 2025,

https://www.researchgate.net/publication/382301750_Mechanistic_interpretability_of_large_language_models_with_applications_to_the_financial_services_industry

2. Mechanistic interpretability of large language models with applications to the financial services industry - arXiv, accessed July 3, 2025, https://arxiv.org/html/2407.11215v1
3. The Black Box Problem: Opaque Inner Workings of Large Language Models, accessed July 3, 2025, https://promptengineering.org/the-black-box-problem-opaque-inner-workings-of-large-language-models/
4. Why LLMs Aren't Black Boxes. Author's Note - Medium, accessed July 3, 2025, https://medium.com/@daveziegler/why-llms-arent-black-boxes-d45be0289993
5. arXiv:2402.10688v2 [cs.CL] 15 Apr 2024, accessed July 3, 2025, https://arxiv.org/pdf/2402.10688
6. Generative AI's crippling and widespread failure to induce robust models of the world, accessed July 3, 2025, https://garymarcus.substack.com/p/generative-ais-crippling-and-widespread
7. (PDF) Black-Box Behavior in Large Language Models: Challenges and Implications, accessed July 3, 2025, https://www.researchgate.net/publication/389819452_Black-Box_Behavior_in_Large_Language_Models_Challenges_and_Implications
8. Risks of Gen AI: the black box problem - KWM, accessed July 3, 2025, https://www.kwm.com/global/en/insights/latest-thinking/risks-of-gen-ai-the-black-box-problem.html
9. The Illustrated Transformer – Jay Alammar – Visualizing machine ..., accessed July 3, 2025, https://jalammar.github.io/illustrated-transformer/
10. Fine-Tuning vs. Pre-Training: Key Differences for Language Models - Sapien, accessed July 3, 2025, https://www.sapien.io/blog/fine-tuning-vs-pre-training-key-differences-for-language-models
11. Emergent Abilities in Large Language Models: A Survey - Synthical, accessed July 3, 2025, https://synthical.com/article/Emergent-Abilities-in-Large-Language-Models%3A-A-Survey-4ac6fc87-e2e7-4366-bb3d-489052d579f3?
12. arxiv.org, accessed July 3, 2025, https://arxiv.org/html/2501.16496v1
13. Could a Large Language Model Be Conscious? - Boston Review, accessed July 3, 2025, https://www.bostonreview.net/articles/could-a-large-language-model-be-conscious/
14. Exploring consciousness: Daniel Dennett's legacy | Donders Wonders, accessed July 3, 2025, https://blog.donders.ru.nl/?p=16373&lang=en
15. The Urgency of Interpretability - Dario Amodei, accessed July 3, 2025, https://www.darioamodei.com/post/the-urgency-of-interpretability
16. Transformer (deep learning architecture) - Wikipedia, accessed July 3, 2025, https://en.wikipedia.org/wiki/Transformer_(deep_learning_architecture)

17. Transformer Neural Networks: A Step-by-Step Breakdown | Built In, accessed July 3, 2025, https://builtin.com/artificial-intelligence/transformer-neural-network
18. How Transformers Work: A Detailed Exploration of Transformer Architecture - DataCamp, accessed July 3, 2025, https://www.datacamp.com/tutorial/how-transformers-work
19. Transformers Explained Visually (Part 1): Overview of Functionality - Towards Data Science, accessed July 3, 2025, https://towardsdatascience.com/transformers-explained-visually-part-1-overview-of-functionality-95a6dd460452/
20. LLM Transformer Model Visually Explained - Polo Club of Data Science, accessed July 3, 2025, https://poloclub.github.io/transformer-explainer/
21. Pre-Training vs Fine Tuning: Choosing the Right Approach in 2025 ..., accessed July 3, 2025, https://labelyourdata.com/articles/llm-fine-tuning/pre-training-vs-fine-tuning
22. What is the difference between pre-training, fine-tuning, and instruct-tuning exactly? - Reddit, accessed July 3, 2025, https://www.reddit.com/r/learnmachinelearning/comments/19f04y3/what_is_the_difference_between_pretraining/
23. Is "Next Token Prediction" Holding LLMs Back? | by Moulik Gupta ..., accessed July 3, 2025, https://blog.gopenai.com/is-next-token-prediction-holding-llms-back-2378e95ae9d7
24. Analyzing the next token probabilities in large language models, accessed July 3, 2025, http://responsible-ai-developers.googleblog.com/2024/03/analyzing-next-token-probabilities-in-large-language-models.html
25. Pre-training vs Fine-Tuning vs In-Context Learning of Large Language Models, accessed July 3, 2025, https://www.entrypointai.com/blog/pre-training-vs-fine-tuning-vs-in-context-learning-of-large-language-models/
26. Are LLMs just predicting the next token? : r/ArtificialInteligence - Reddit, accessed July 3, 2025, https://www.reddit.com/r/ArtificialInteligence/comments/1jo3o69/are_llms_just_predicting_the_next_token/
27. LLMs confabulate not hallucinate - Beren's Blog, accessed July 3, 2025, https://www.beren.io/2023-03-19-LLMs-confabulate-not-hallucinate/
28. Hallucination or Confabulation? Neuroanatomy as metaphor in Large Language Models, accessed July 3, 2025, https://pmc.ncbi.nlm.nih.gov/articles/PMC10619792/
29. Understanding LLM Hallucinations. Causes, Detection, Prevention, and Ethical Concerns, accessed July 3, 2025, https://medium.com/@tam.tamanna18/understanding-llm-hallucinations-causes-detection-prevention-and-ethical-concerns-914bc89128d0
30. Hallucination (artificial intelligence) - Wikipedia, accessed July 3, 2025, https://en.wikipedia.org/wiki/Hallucination_(artificial_intelligence)

31. Fine-tuning large language models (LLMs) in 2025 - SuperAnnotate, accessed July 3, 2025, https://www.superannotate.com/blog/llm-fine-tuning
32. Fine-Tuning LLMs: A Guide With Examples - DataCamp, accessed July 3, 2025, https://www.datacamp.com/tutorial/fine-tuning-large-language-models
33. How did language models go from predicting the next word token to answering long, complex prompts? - Reddit, accessed July 3, 2025, https://www.reddit.com/r/learnmachinelearning/comments/17gd8mi/how_did_language_models_go_from_predicting_the/
34. Emergent Properties in Large Language Models: A Deep Research Analysis - Greg Robison, accessed July 3, 2025, https://gregrobison.medium.com/emergent-properties-in-large-language-models-a-deep-research-analysis-d6886c37061b
35. Large Language Models and Emergence: A Complex Systems Perspective - arXiv, accessed July 3, 2025, https://arxiv.org/html/2506.11135v1
36. [2304.15004] Are Emergent Abilities of Large Language Models a Mirage? - arXiv, accessed July 3, 2025, https://arxiv.org/abs/2304.15004
37. Papers Explained 83: Are Emergent Abilities of Large Language Models a Mirage?, accessed July 3, 2025, https://ritvik19.medium.com/papers-explained-are-emergent-abilities-of-large-language-models-a-mirage-4160cf0e44cb
38. arxiv.org, accessed July 3, 2025, https://arxiv.org/html/2503.05788v2
39. Emergent Abilities in Large Language Models: A Survey - arXiv, accessed July 3, 2025, https://arxiv.org/pdf/2503.05788
40. Stochastic parrot - Wikipedia, accessed July 3, 2025, https://en.wikipedia.org/wiki/Stochastic_parrot
41. Beware of WEIRD Stochastic Parrots - Resilience.org, accessed July 3, 2025, https://www.resilience.org/stories/2024-02-15/beware-of-weird-stochastic-parrots/
42. What is In-context Learning, and how does it work: The Beginner's Guide - Lakera AI, accessed July 3, 2025, https://www.lakera.ai/blog/what-is-in-context-learning
43. What is In-Context Learning of LLMs? - IKANGAI, accessed July 3, 2025, https://www.ikangai.com/what-is-in-context-learning-of-llms/
44. What is In Context Learning (ICL)? - Hopsworks, accessed July 3, 2025, https://www.hopsworks.ai/dictionary/in-context-learning-icl
45. What is In-Context Learning? How LLMs Learn From ICL Examples - PromptLayer, accessed July 3, 2025, https://blog.promptlayer.com/what-is-in-context-learning/
46. How in-context learning improves large language models - IBM Research, accessed July 3, 2025, https://research.ibm.com/blog/demystifying-in-context-learning-in-large-language-model
47. What is In-Context Learning? Simply Explained - FinetuneDB, accessed July 3, 2025, https://finetunedb.com/blog/what-is-in-context-learning-simply-explained/
48. Chain-of-thought reasoning supercharges enterprise LLMs, accessed July 3, 2025, https://www.k2view.com/blog/chain-of-thought-reasoning/
49. Chain-of-Thought Prompting: Step-by-Step Reasoning with LLMs | DataCamp,

accessed July 3, 2025, https://www.datacamp.com/tutorial/chain-of-thought-prompting

50. Chain-of-Thought Prompting: Helping LLMs Learn by Example - Gaper.io, accessed July 3, 2025, https://gaper.io/chain-of-thought-prompting/
51. How to teach chain of thought reasoning to your LLM | Invisible Blog, accessed July 3, 2025, https://www.invisible.co/blog/how-to-teach-chain-of-thought-reasoning-to-your-llm
52. How Chain of Thought (CoT) Prompting Helps LLMs Reason More Like Humans | Splunk, accessed July 3, 2025, https://www.splunk.com/en_us/blog/learn/chain-of-thought-cot-prompting.html
53. Chain of Preference Optimization: Improving Chain-of-Thought Reasoning in LLMs - NIPS, accessed July 3, 2025, https://proceedings.neurips.cc/paper_files/paper/2024/file/00d80722b756de0166523a87805dd00f-Paper-Conference.pdf
54. Are Emergent Abilities of Large Language Models a Mirage?, accessed July 3, 2025, https://papers.neurips.cc/paper_files/paper/2023/file/adc98a266f45005c403b8311ca7e8bd7-Paper-Conference.pdf
55. Are Emergent Abilities of Large Language Models a Mirage? - OpenReview, accessed July 3, 2025, https://openreview.net/pdf?id=JRdN9GcI52
56. Emergent Abilities in Large Language Models: An Explainer, accessed July 3, 2025, https://cset.georgetown.edu/article/emergent-abilities-in-large-language-models-an-explainer/
57. Are Emergent Abilities of Large Language Models a Mirage? - OpenReview, accessed July 3, 2025, https://openreview.net/forum?id=ITw9edRDID
58. Paper Review: Are Emergent Abilities of Large Language Models a Mirage? - velog, accessed July 3, 2025, https://velog.io/@jinotter3/Paper-Review-Are-Emergent-Abilities-of-Large-Language-Models-aMirage
59. [R] Are Emergent Abilities in Large Language Models just In-Context Learning? - Reddit, accessed July 3, 2025, https://www.reddit.com/r/MachineLearning/comments/19bkcqz/r_are_emergent_abilities_in_large_language_models/
60. Video watch page: fantastic video on mechanistic interpretability : r ..., accessed July 3, 2025, https://www.reddit.com/r/artificial/comments/1hxylrv/fantastic_video_on_mechanistic_interpretability/
61. What is feature visualization? - AISafety.info, accessed July 3, 2025, https://aisafety.info/questions/8HIA/What-is-feature-visualization
62. Towards Understanding Fine-Tuning Mechanisms of LLMs via Circuit Analysis, accessed July 3, 2025, https://openreview.net/forum?id=45EIiFd6Oa
63. Circuit Tracing: A Step Closer to Understanding Large Language Models, accessed July 3, 2025, https://towardsdatascience.com/circuit-tracing-a-step-closer-to-understanding-

large-language-models/
64. Tracing the thoughts of a large language model - Anthropic, accessed July 3, 2025, https://www.anthropic.com/research/tracing-thoughts-language-model
65. LLM Interpretability: Exploring the Latest Research from OpenAI and Anthropic - YouTube, accessed July 3, 2025, https://www.youtube.com/watch?v=fkW0bGnbDkQ
66. Circuit Tracing: Revealing Computational Graphs in Language Models, accessed July 3, 2025, https://transformer-circuits.pub/2025/attribution-graphs/methods.html
67. Understanding Language Model Circuits through Knowledge Editing - arXiv, accessed July 3, 2025, https://arxiv.org/html/2406.17241v3
68. [2505.03368] Geospatial Mechanistic Interpretability of Large Language Models - arXiv, accessed July 3, 2025, https://arxiv.org/abs/2505.03368
69. Chatbots Do Not Hallucinate, They Confabulate - Psychology Today, accessed July 3, 2025, https://www.psychologytoday.com/us/blog/theory-of-knowledge/202403/chatbots-do-not-hallucinate-they-confabulate
70. (PDF) Hallucinations in LLMs: Types, Causes, and Approaches for ..., accessed July 3, 2025, https://www.researchgate.net/publication/385085962_Hallucinations_in_LLMs_Types_Causes_and_Approaches_for_Enhanced_Reliability
71. Functionalism (Stanford Encyclopedia of Philosophy), accessed July 3, 2025, https://plato.stanford.edu/entries/functionalism/
72. #90 - Prof. DAVID CHALMERS - Consciousness in LLMs - YouTube, accessed July 3, 2025, https://www.youtube.com/watch?v=T7aIxncLuWk
73. Multiple drafts model - Wikipedia, accessed July 3, 2025, https://en.wikipedia.org/wiki/Multiple_drafts_model
74. [2303.07103] Could a Large Language Model be Conscious? - arXiv, accessed July 3, 2025, https://arxiv.org/abs/2303.07103
75. Could a Large Language Model Be Conscious? Within the next decade, we may well have systems that are serious candidates for consciousness. - David Chalmers : r/singularity - Reddit, accessed July 3, 2025, https://www.reddit.com/r/singularity/comments/15nfq0f/could_a_large_language_model_be_conscious_within/
76. David J. Chalmers, Could a large language model be conscious? - PhilArchive, accessed July 3, 2025, https://philarchive.org/rec/CHACAL-3
77. Daniel Dennett's last interview: 'AI could signal the end of human civilisation' | The Spectator, accessed July 3, 2025, https://www.spectator.co.uk/article/daniel-dennetts-last-interview-ai-could-signal-the-end-of-human-civilisation/
78. Exploring Daniel Dennett's view on AI, Free Will and Determinism, with ChatGPT. - Medium, accessed July 3, 2025, https://medium.com/@socialscholarly/exploring-daniel-dennetts-view-on-ai-free-will-and-determinism-with-chatgpt-b6832825483a
79. Daniel Dennett's Been Thinking About Thinking—and AI | Tufts Now, accessed July

3, 2025, https://now.tufts.edu/2023/10/02/daniel-dennetts-been-thinking-about-thinking-and-ai

80. Daniel Dennett Sounds the Alarm On ChatGPT! - YouTube, accessed July 3, 2025, https://m.youtube.com/shorts/naEH-3di7RU
81. On the dangers of stochastic parrots Can language models be too big? ! - University of Washington, accessed July 3, 2025, https://faculty.washington.edu/ebender/papers/Bender-Turing-Institute-July-2021.pdf
82. The Stochastic Parrot - Boethius Translations, accessed July 3, 2025, https://boethiustranslations.com/the-stochastic-parrot/
83. On the Dangers of Stochastic Parrots | Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency - University of Washington, accessed July 3, 2025, http://faculty.washington.edu/ebender/papers/Stochastic_Parrots.pdf?page=9&utm_medium-referral=&utm_source=for-businesses
84. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? "1F99C, accessed July 3, 2025, https://s10251.pcdn.co/pdf/2021-bender-parrots.pdf
85. The Rise of Stochastic Parrots - Actuaries Digital, accessed July 3, 2025, https://www.actuaries.asn.au/research-analysis/the-rise-of-stochastic-parrots
86. [2408.13442] A Law of Next-Token Prediction in Large Language Models - arXiv, accessed July 3, 2025, https://arxiv.org/abs/2408.13442
87. [2505.11183] On Next-Token Prediction in LLMs: How End Goals Determine the Consistency of Decoding Algorithms - arXiv, accessed July 3, 2025, https://arxiv.org/abs/2505.11183
88. www.k2view.com, accessed July 3, 2025, https://www.k2view.com/blog/chain-of-thought-reasoning/#:~:text=Chain%2Dof%2Dthought%20reasoning%20is,results%20in%20more%20accurate%20responses.
89. Chain of Thought Prompting: Enhance AI Reasoning & LLMs - Future AGI, accessed July 3, 2025, https://futureagi.com/blogs/chain-of-thought-prompting-ai-2025
90. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? - ResearchGate, accessed July 3, 2025, https://www.researchgate.net/publication/349754361_On_the_Dangers_of_Stochastic_Parrots_Can_Language_Models_Be_Too_Big