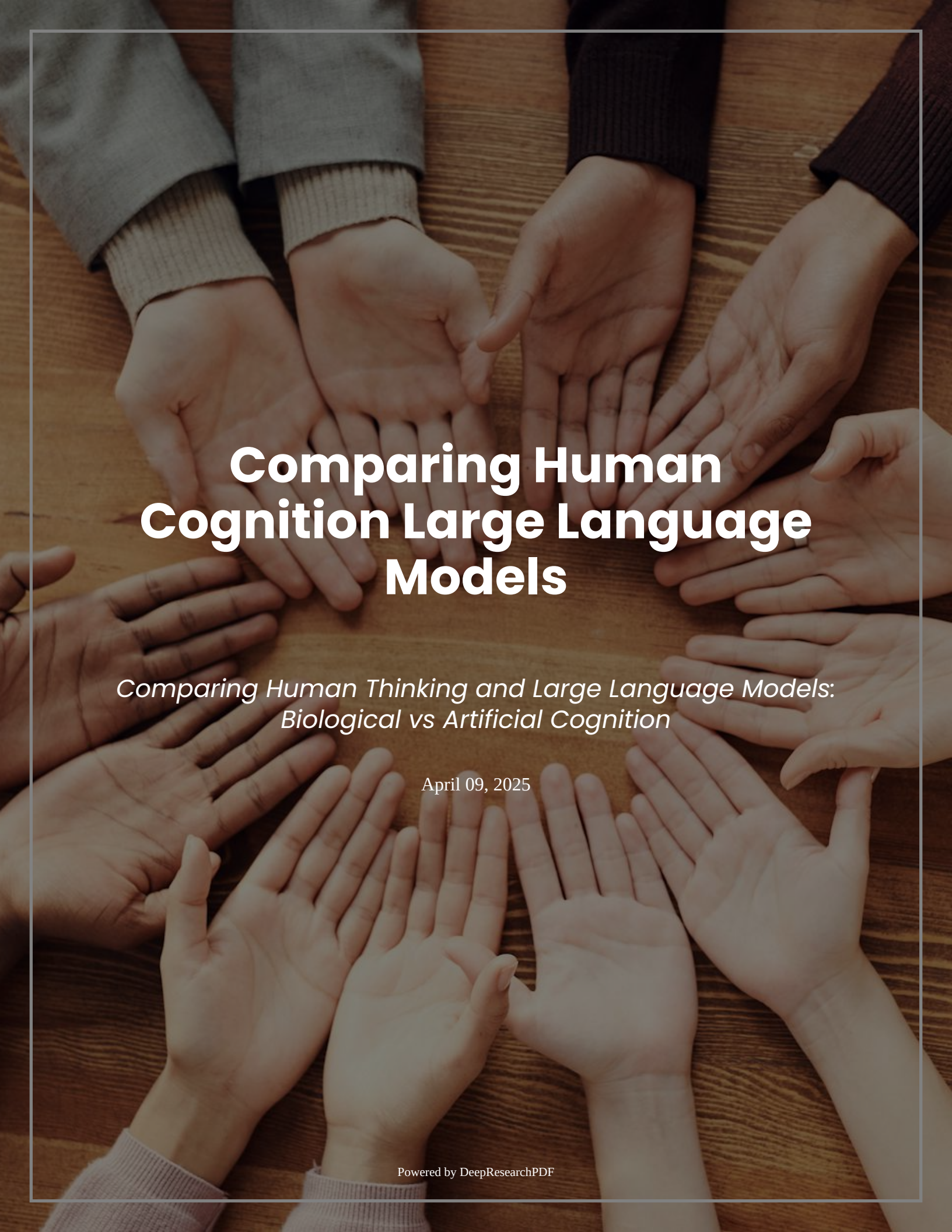

Comparing Human Cognition Large Language Models

Comparing Human Thinking and Large Language Models: Biological vs Artificial Cognition

April 09, 2025

Comparing Human Thinking and Large Language Models: Biological vs Artificial Cognition

(Study shows that the way the brain learns is different from the way that artificial intelligence systems learn | University of Oxford) *A stylized representation of the human brain's neural circuitry (left) versus artificial circuitry (right). Advances in large language models prompt comparisons between biological and artificial intelligence.*

Introduction

In recent years, large language models (LLMs) have demonstrated remarkable abilities in generating human-like text, raising the question of how these artificial systems compare to the human mind. Both the human brain and LLMs process information using networks (biological neurons vs. artificial neurons), and both improve their performance through some form of learning. However, beneath these surface similarities lie profound differences. Human cognition is the product of biological neural networks shaped by millions of years of evolution, characterized by neuroplastic learning, embodied experience, and conscious awareness. LLMs, by (Study shows that the way the brain learns is different from the way that artificial intelligence systems learn | University of Oxford) (Brain-inspired replay for continual learning with artificial neural networks | Nature Communications)programs (neural networks with billions of parameters) trained on massive text corpora via machine learning algorithms, lacking any physical embodiment or genuine understanding. This re (Study shows that the way the brain learns is different from the way that artificial intelligence systems learn | University of Oxford) (Study shows that the way the brain learns is different from the way that artificial intelligence systems learn | University of Oxford) comparison of human thinking and

LLM operation, focusing on key areas of overla ([

The Magical Mystery Four: How is Working Memory Capacity Limited, and Why? - PMC

](https://pmc.ncbi.nlm.nih.gov/articles/PMC2864034/#:~:text=It%20may%20not%20really%20be,why%20does%20the%20limit%20occur))ion: how each system learns, how information is processed, memory systems, reasoning and creativity, and where they fundamentally diverge (in embodiment, consciousness, goals, and intent). We will explore technical aspects of b (What is a context window? | IBM) (The Limits of Working Memory: Human Brains vs. AI Models - Illumio Cybersecurity Blog | Illumio)tificial neural networks, and discuss concepts such as working memory vs. context windows, predictive processing, biases in cognition, and the phenomenon of “hallucination.” The goal is an accessible yet rigorous overview of biological versus artificial cognition for a tech-savvy reader.

Learning Mechanisms: Neuroplasticity vs. Backpropagation

Humans and LLMs both **learn**, but the mechanisms and efficiency of their learning are very different. Human learning occurs through **n (14: Evidence of a predictive coding hierachy in the human brain listening to speech, Nature Human Behaviour - ML-Neuro - BayernCollab)ty** – the ability of neural connections (synapses) in the brain to strengthen, weaken, or form anew in response to experience. When a person learns a new fact or skill, networks of neurons in relevant brain regions adjust their firing patterns and synaptic weights. Mechanisms such as long-term potentiation (LTP) and long-term depression (LTD) alter the strength of synapses based on how frequently and strongly neurons fire together (often summarized by the adage “neurons that fire together, wire together”). This Hebbian form of local learning allows the brain to gradually encode new information. Importantly, the brain can often learn from just a few examples –

sometimes a single exposure to a new concept or a single practice of a skill can form a lasting memory. It also **learns continually**: each day we integrate new ex (LLMs vs. Human Mind: Understanding the Creativity Gap.) (LLMs vs. Human Mind: Understanding the Creativity Gap.)tely overwriting old memories.

LLMs, on the other hand, learn through an **artificial training process** that is quite unlike human one-shot learning. Large language models are typically trained using **backpropagation**, a global error-correction algorithm. During training, the model processes millions or billions of text examples and adjusts its internal weights to minimize the difference between its predicted outputs and the actual text in the training data. This involves computing an error (loss) for a given out ((Ir)rationality and cognitive biases in large language models | Royal Society Open Science)ating that error backward through many layers of artificial neurons to update each weight slightly. The process is repeated over the dataset for many iterations (epochs) until the model's predictions improve. Backpr ([

Hallucination or Confabulation? Neuroanatomy as metaphor in Large Language Models - PMC

](https://pmc.ncbi.nlm.nih.gov/articles/PMC10619792/#:~:text=are%20engaged%20in%20perceiving%2C%20that,it%20is%20making%20things%20up)) ([

Hallucination or Confabulation? Neuroanatomy as metaphor in Large Language Models - PMC

](https://pmc.ncbi.nlm.nih.gov/articles/PMC10619792/#:~:text=More%20accurate%20terminology%20is%20found,3%5D.%20Confabulation%20is%20frequently))ly powerful for training deep neural networks, but it requires huge numbers of training examples and many passes – essentially brute-force practice. For example, while a human child can learn a new word from hearing it used just once or twice, an LLM might effectively "hear" a word thousands of times across its training corpus before it can confidently use it in context.

One consequence is that **humans excel (The Myth of Thinking Machines | Daily Philosophy)cremental learning**, whereas standard LLMs require extensive up-front training and then remain relatively fixed. Researchers note that *“we can learn new information by just seeing ([

Hallucination or Confabulation? Neuroanatomy as metaphor in Large Language Models - PMC

](https://pmc.ncbi.nlm.nih.gov/articles/PMC10619792/#:~:text=In%20psychiatry%2C%20hallucinations%20are%20a,LLMs%20do%20not%20have))e artificial systems need to be trained hundreds of times with the same pieces of information to learn them”*, and humans can integrate new information without forgetting old knowledge, whereas neural networks often exhibit interference when learning new things (Study shows that the way the brain learns is different from the way that artificial intelligence systems learn | University of Oxford). Indeed, artificial neural networks suffer from **catastrophic forgetting**: w (The Myth of Thinking Machines | Daily Philosophy)on a new task or data, they tend to abruptly overwrite previously learned information. In contrast, the brain’s learning is more **robust and cumulative** – we retain past knowledge while adding new memories. A 2020 study highlights this difference:* “Artificial neural networks suffer from catastrophic forgetting. Unlike humans, when these networks are trained on something new, they rapidly forget what was learned before”*, whereas the human brain protects old memories via mechanisms like replaying neural activity patterns during sleep or rest (Brain-inspired replay for continual learning with artificial neural networks | Nature Communications).

Another fundamental difference is **how the credit assignment problem is solved**. In an artificial network, learning is guided by an external algorithm (backpropagation) that calculates error gradients and explicitly updates each weight to reduce output error. The brain does not appear to implement literal backpropagation – there is no known biological mechanism that computes global e (Study shows that the way the brain learns is different from the way that artificial intelligence systems learn | University of Oxford)nts and adjusts each synapse accordingly in one sweep. Instead, the brain likely relies on more local signals and indirect

feedback. Recent research suggests the brain may achieve learning by first allowing neural activity to settle into a (Study shows that the way the brain learns is different from the way that artificial intelligence systems learn | University of Oxford)balanced”* state for a given input, and then making small adjustments to synapses to nudge firing patterns toward desired outcomes (Study shows that the way the brain learns is different from the way that artificial intelligence systems learn | University of Oxford). In other words, rathe (Brain-inspired replay for continual learning with artificial neural networks | Nature Communications)-down error signal telling each synapse how to change, the brain might use a combination of local activity-dependent plasticity rules and neuromodulator signals (like dopamine for reward prediction error) to guide learning. The **le (14: Evidence of a predictive coding hierachy in the human brain listening to speech, Nature Human Behaviour - ML-Neuro - BayernCollab) *in the brain is thus believed to be different in principle from backpropagation, even if the end result (adjusting connection strengths to improve performance) is functionally analogous. Researchers writing in* Nature Neuroscience *describe that* “in artificial n (What is a context window? | IBM)rks, an external algorithm tries to modify synaptic connections to reduce error, whereas [in] the h (14: Evidence of a predictive coding hierachy in the human brain listening to speech, Nature Human Behaviour - ML-Neuro - BayernCollab)neural] activity [settles] into an optimal balanced configuration before adjusting synaptic connections”*, a strategy that may preserve existing knowledge and ([

The Magical Mystery Four: How is Working Memory Capacity Limited, and Why? - PMC

](https://pmc.ncbi.nlm.nih.gov/articles/PMC2864034/#:~:text=It%20may%20not%20really%20be,why%20does%20the%20limit%20occur)) learning in the brain (Study shows that the way the brain learns is different from the way that artificial intelligence systems learn | University of Oxford).

It’s also instructive to compare **learning speed and adaptability**. Humans learn **online** – continually and adaptively. A person can learn from a single conversation or an unforeseen event and (What is a context window? | IBM) incorporate that into their worldview. By contrast, a

pre-trained LLM (like GPT) does not automatically le (The Limits of Working Memory: Human Brains vs. AI Models - Illumio Cybersecurity Blog | Illumio) user interaction. Once trained, its parameters are static; if it gives a wrong answer and is corrected, it doesn't adapt on the fly (unless a mechanism like fine-tuning or few-shot prompting is explicitly used to update it). New information (for instance, a fact that emerged after the model's training data cutoff) will not be known to the model unless it is retrained or provided in the prompt. In summary, the brain's learning is *incremental, data-efficient, and ongoing*, whereas LLM learning is *batch-oriented, data-hungry, and requires explicit retraining*.

Information Processing: Neural Activity vs. Token Prediction

Beyond how they learn, humans and LLMs also **process information** in different ways. The human brain is an electrochemical organ: neurons process information via electrical impulses (spikes) and chemical neurotransmitters. Each neuron integrates inputs from thousands of other neurons, and if its excitation exceeds a threshold, it emits a spike that propagates to other neurons. Processing in the brain is massively **parallel** and distributed – billions of neurons and trillions of synapses are active concurrently, with different brain regions specialized for different functions (visual cortex for sight, auditory cortex for sound, etc., all interacting). Neural activity is also **oscillatory and dynamic**; brain networks exhibit time-varying patterns (brain waves) and can maintain persistent activity (as in working memory circuits). Importantly, the brain's processing is deeply **contextual and multi-modal** – sensory inputs (sights, sounds, etc.) and prior knowledge are integrated to interpret the world. Cognitive processing in humans often involves **predictive processing**: the brain is thought to constantly generate predictions about incoming sensory data and adjust its internal model based on prediction errors (this is the essence of the *predictive coding* theory of cognition).

LLMs operate on very different principles. An LLM processes information in the form of text (or more precisely, sequences of discrete symbols called *tokens*). Modern LLMs like GPT use the **Transformer architecture**, which processes text using layers of self-atten (LLMs vs. Human Mind: Understanding the Creativity Gap.)d-forward neural networks. When an LLM receives input (for example, a user prompt), it first encodes the text into a series of vector representations. These representations then pass through multiple layers where the model calculates attention scores – essentially figuring out which prior words (or tokens) are most relevant to predicting the next word. The LLM's computation is organized into sequential layers, but within each layer, many operations happen in parallel (matrix multiplications across thousands of dimensions). In essence, an LLM's **core processing task** is to predict the probability distribution of the next token given all prior tokens. It accomplishes this with a fixed **context window** of input. For example, if an LLM has a context window of 2048 tokens, it can "attend" ((Ir)rationality and cognitive biases in large language models | Royal Society Open Science) tokens of prior text to inform its next word choice. There is no explicit notion of time or persistence beyond this window: everything the model "knows" during a single inference is contai ([

Hallucination or Confabulation? Neuroanatomy as metaphor in Large Language Models - PMC

](https://pmc.ncbi.nlm.nih.gov/articles/PMC10619792/#:~:text=are%20engaged%20in%20perceiving%2C%20that,it%20is%20making%20things%20up)) ([

Hallucination or Confabulation? Neuroanatomy as metaphor in Large Language Models - PMC

](https://pmc.ncbi.nlm.nih.gov/articles/PMC10619792/#:~:text=More%20accurate%20terminology%20is%20found,3%5D.%20Confabulation%20is%20frequently))ver it has generated so far (which itself is fed back in as new input tokens). After producing an output, the model does not retain memory of the conversation unless the conversation history is supplied back to it in subsequent prompts.

One illuminating comparison is how **predictive processing** differs in scope. Both the brain and LLMs are predictive to some extent: the brain continuously anticipates sensory inputs (as per predictive coding theories), and an LLM literally *generates predictions* for the next token. However, the **brain's predictions are hierarchical and long-range**, whereas LLM prediction is local and stepwise. For instance, if you begin a sentence "The striker kicked the ball …", a human listener's brain might unconsciously leap ahead and anticipate an outcome like "... into the goal" or "... toward the opponent's net" – essentially predicting the high-level meaning or result before the speaker even utters the connecting words. Evidence from neuroscience indicates that the human brain makes *"long-range and hierarchical predictions"*, often completing an entire thought or phrase in advance and then filling in lower-level details like specific words (14: Evidence of a predictive coding hierachy in the human brain listening to speech, Nature Human Behaviour - ML-Neuro - BayernCollab). In one example, researchers found that when people hear the beginning of a sentence, their brains might predict the general gist of how it ends (e.g. expecting a goal in a soccer narrative) ra ([

Hallucination or Confabulation? Neuroanatomy as metaphor in Large Language Models - PMC

](https://pmc.ncbi.nlm.nih.gov/articles/PMC10619792/#:~:text=In%20psychiatry%2C%20hallucinations%20are%20a,LLMs%20do%20not%20have)) the very next word (14: Evidence of a predictive coding hierachy in the human brain listening to speech, Nature Human Behaviour - ML-Neuro - BayernCollab). In contrast, *"current Large Language Models ... are designed to predict the next immediate token ... which contrasts [with] the long-range, holistic predictions made by our brains"* (14: Evidence of a predictive coding hierachy in the human brain listening to speech, Nature Human Behaviour - ML-Neuro - BayernCollab). In other words, an LLM focuses on the *immediate next step* in the sequence, not an overarching goal or distant outcome (unless that goal is indirectly encoded through many sequential next-token predictions).

Another key difference is that **the brain processes multiple modalities and timescales** inherently. Neurons in higher-level areas can integrate

informatio (The Myth of Thinking Machines | Daily Philosophy)r time windows or abstract across modalities (e.g. combining sight and sound), and feedback connections in the brain create recurrent loops that help maintain context and enable planning. LLMs, in their basic form, deal with one modality (text) and have a fixed timescale given by their context length. If information falls outside the context window (too far back in the text or not provided at all), the model has no access to it. There is no innate concept of persistent state or memory across sessions (unless engineered via external memory tools). The human brain, by contrast, has **working memory** and **attention mechanisms** that actively maintain and refresh relevant information even as new inputs arrive, enabling coherence over time and task-switching. While Transformers have an analog of "attention" in the mathematical sense (the attention mechanism deciding which tokens influence each other), this is not the same as a human's top-down attention that can choose to focus on one aspect of a scene or recall a specific memory on demand.

It's also worth noting the difference in **parallelism and serial processing**. Human cognition can seem slower in some low-level tasks (a brain neuron fires at most ~1000 Hz, whereas a transistor can switch billions of times per second), but because the brain has so many processing units working simultaneously, it excels at tasks like vision, motor coordination, and intuitive judgments extremely efficiently. LLMs run on digital hardware that typically processes operations sequentially (though parallelized across many cores); generating each token is a sequential operation that depends on previous tokens. For instance, to produce a sentence of 20 words, an LLM must perform 20 forward passes (one for each token, using the output token as input for the next step). Humans, when speaking a sentence, are also producing words one by one, but the *planning* of the sentence and the integration of ideas happen in a more parallel and anticipatory fashion in the brain. A person can adjust mid-sentence, change phrasing on the fly, or choose not to complete a thought – all of which involve interactive processing of both the linguistic output and a host of other internal signals (like the reaction of the listener, emotional tone, etc.). LLMs lack these feedback loops during generation;

they do not evaluate the real-world effect of their words or change course based on un-modeled factors – they simply continue the statistical pattern.

- *Summary:** Human information processing is analog, parallel, and deeply context-driven (with multi-sensory integration and hierarchical prediction), whereas LLM processing is digital, largely sequential (token-by-token), and bound to a fixed textual context. The brain's mode of operation is often described as *content-addressable* and associative – cues can trigger memories or predictions – while an LLM's operation is more like *next-step pattern completion* based on its training. This makes LLMs extremely powerful in well-bounded linguistic tasks (they can rapidly complete or transform text), but it also means they might miss the broader picture or intent that a human thinker would naturally consider.

Memory Systems: Biological Memory vs. Artificial Memory Architecture

Memory is a cornerstone of cognition, and here the differences between humans and LLMs are especially pronounced. Humans have **multiple memory systems** – typically characterized as sensory memory, short-term/working memory, and long-term memory (with further distinctions between episodic memory, semantic m (Study shows that the way the brain learns is different from the way that artificial intelligence systems learn | University of Oxford) (Study shows that the way the brain learns is different from the way that artificial intelligence systems learn | University of Oxford), etc.). These memory systems are supported by specialized brain s (What is a context window? | IBM) (The Limits of Working Memory: Human Brains vs. AI Models - Illumio Cybersecurity Blog | Illumio) For example, **working memory** (the mind's "scratchpad" for temporarily (14: Evidence of a predictive coding hierachy in the human brain listening to speech, Nature Human Behaviour - ML-Neuro - BayernCollab) (14: Evidence of a predictive coding hierachy in the human brain listening to speech, Nature Human Behaviour - ML-Neuro -

BayernCollab) associated with frontal and parietal brain regions a ((Ir)rationality and cognitive biases in large language models | Royal Society Open Science)ustained neural activity or short-term synaptic changes to keep a small amount of information av ([

Hallucination or Confabulation? Neuroanatomy as metaphor in Large Language Models - PMC

](https://pmc.ncbi.nlm.nih.gov/articles/PMC10619792/#:~:text=are%20engaged%20in%20perceiving%2C%20that,it%20is%20making%20things%20up)) ([

Hallucination or Confabulation? Neuroanatomy as metaphor in Large Language Models - PMC

](https://pmc.ncbi.nlm.nih.gov/articles/PMC10619792/#:~:text=More%20accurate%20terminology%20is%20found,3%5D.%20Confabulation%20is%20frequently))es. Experiments show that human working memory has a limited capacity – classically about 7±2 items as per Miller's l (The Myth of Thinking Machines | Daily Philosophy) (The Myth of Thinking Machines | Daily Philosophy) (The Myth of Thinking Machines | Daily Philosophy)t it at about 3–5 meaningful items or "chunks" at a time ([

The Magical Mystery Four: How is Working Memory Capacity Limited, and Why? - PMC

](https://pmc.ncbi.nlm.nih.gov/articles/PMC2864034/#:~:text=It%20may%20not%20really%20be,why%20does%20the%20limit%20occur)). This is a strikingly small window: at any moment, your conscious mind can only juggle a handful of pieces of information (like digits of a phone number or elements of a problem). Long-term memory, by contrast, is vast and enduring. The human brain can store an immense amount of information over a lifetime: personal experiences (episodic memories), general knowledge (semantic memory), skills and habits (procedural memory), etc. These are encoded via lasting synaptic modifications. The hippocampus plays a key role in forming new episodic memories and gradually training the cortex to retain them for the long haul (a process of memory consolidation, especially during sleep). Importantly, human

memory is **associative** – memories are linked by meaning, context, emotion. A smell or a song can instantly bring back a vivid memory due to these associative links.

LLMs have a much simpler memory architecture on the surface. An LLM's "knowledge" of the world – everything it has absorbed from training data – is stored in the **weights** of its neural network. These weights are distributed numerical values that encode statistical associations from the training text. In a sense, the model's weights serve as a kind of long-term semantic memory (containing facts, language patterns, and other learned associations), but this memory is not stored in discrete units like facts or stories; it's smeared across millions of parameters. When an LLM "recalls" information, it's not retrieving a specific stored record the way a database or a human memory does – rather, the prompt cues the model to generate likely continuations, which indirectly taps into relevant portions of its stored statistical knowledge. The model has *no explicit episodic memory of individual training examples* (it cannot tell you where or when it learned a given piece of text, and it doesn't have separate traces for each example).

In operation, an LLM has something analogous to working memory: the **context window**. The context window is the sequence of tokens that the model is currently considering as input. This includes the user's prompt and any prior dialogue (for chat models) or prior sentence if generating text. Everything within this window can influence the next output. In many discussions, the context window is directly compared to working memory: *"An LLM's context window can be thought of as the equivalent of its working memory. It determines how long of a conversation it can carry out without forgetting details from earlier in the exchange"* (What is a context window? | IBM). If information falls outside the window (i.e., it was said too long ago in the conversation or exceeds the token limit), the model has **no inherent memory** of it – unless the user reintroduces that information. For example, if a user tells a story to an LLM and then continues the dialogue far beyond the story without referencing it, the details of that story will eventually drop out of the context and the LLM will effectively "forget" it. By contrast, a human conversing can remember what was said

earlier (especially the gist) even much later, because we encode it into long-term memory or at least maintain a mental representation beyond the immediate conversational turn.

One intriguing difference is that while human working memory is fixed by our cognitive architecture (we cannot just decide to hold 50 items in mind at once), an AI's context window **can be expanded** by engineering. Early LLMs had context windows of a few thousand tokens; newer models have windows of 32k tokens or more, and research models extend to hundreds of thousands. This is computationally expensive, but feasible: *"Unlike humans, whose working memory is fixed, an AI's context window can be expanded (with more GPUs, better algorithms, or new hardware)"* (The Limits of Working Memory: Human Brains vs. AI Models - Illumio Cybersecurity Blog | Illumio). In effect, we can give an LLM a much larger "working memory" than any human could have, allowing it to consider very long documents or extensive code all at once. However, even these large context windows have limits and trade-offs (diminishing returns, higher computation, and still a hard cutoff beyond which the model can't directly use information (What is a context window? | IBM)).

Another difference is **memory reliability and recall**. Human memory is fallible – we forget things, especially if not revisited, and memories can become distorted over time. But we also have the ability to **re-learn** or reinforce memories (studying, repetition, using mnemonic strategies). LLMs, once trained, have a static memory store in their weights. They don't "forget" in the organic way (unless deliberately fine-tuned with new data that overrides old, which is more akin to overwriting). An LLM might fail to output a learned fact if not properly prompted, but the information is essentially embedded somewhere in the weights if it was present in the training data. Humans often organize memory by meaning – e.g., we might not remember an exact sentence we read, but we remember the key idea in our own words. LLMs are text-trained and tend to recall or regenerate information in forms close to how they saw it, which can sometimes lead to verbatim regurgitation of sources (which is a concern if the training text had copyrighted material, etc.).

- *Working memory vs. context usage**: Humans use working memory not just for holding data, but for manipulating it (mental arithmetic, translating thoughts into sentences, etc.). LLMs do perform a form of manipulation within the context – for example, they can take a list of facts in the prompt and then produce a summary, effectively doing a computation within that prompt window. But they lack an active** control process** akin to human executive function that decides what to store or retrieve. The entire prompt is implicitly "in memory" at once for the model, but the model doesn't have an explicit focus or indexing. Humans can voluntarily direct their working memory (e.g., decide to keep repeating a phone number to remember it). LLMs have no such metacognitive control; their "focus" is automatically determined by the attention mechanism and learned patterns.

- *Long-term memory updating **is also different. Humans integrate new memories throughout life. We have mechanisms for memory consolidation during sleep, and even during wakefulness our recent experiences gradually shape our synapses. LLMs typically undergo a huge training phase and then stop – any new learning would require another training phase (fine-tuning or continual learning approaches). Some cutting-edge efforts add** retrieval mechanisms** to LLMs (e.g., connecting them to databases or letting them read and store new information) to simulate an ability to acquire new knowledge after deployment, but these are add-ons rather than inherent. The human brain, conversely, is inherently a continual learning system.

In sum, **human memory is multi-faceted and dynamic**, encompassing a small but flexible working memory and a large, interwoven long-term store of knowledge and experiences. **LLM memory is bifurcated** into a fixed learned model (large but static knowledge base) and a transient context window (powerful but limited, and strictly textual). This means, for example, a person asked about yesterday's lunch can recall that specific episodic memory (if it was noteworthy or attended to), whereas an LLM has no "yesterday" – it can only guess what someone *might* have eaten

for lunch based on typical patterns, unless that detail is provided in the prompt. On the other hand, an LLM might have at its disposal an encyclopedic range of facts (via training on Wikipedia and so on) that any single human might not recall accurately, but the LLM might present them on demand (with the caveat of accuracy discussed later).

Reasoning and Creativity: Abstraction, Generalization, and Emergence

- *Reasoning **and** creativity** are higher-level cognitive functions where both surprising overlaps and critical differences appear between humans and LLMs. Let's consider reasoning first. Humans are capable of several modes of reasoning – from fast, intuitive judgments (often called *System 1* thinking) to slow, deliberative logical reasoning (*System 2*). We can follow chains of logical inference, do mathematical calculations step by step, plan multi-step actions toward a goal, and apply abstract rules to novel situations. How does the human brain implement reasoning? It likely recruits working memory (to hold intermediate results), executive control networks (frontal lobes) to guide step-by-step thought, and draws upon vast background knowledge to inform each step. Crucially, human reasoning is often tied to *meaning* and *understanding* – we form mental models of the situation and manipulate those models, not just symbols. For example, when solving a puzzle, a person might visualize the elements of the problem or relate it to a familiar scenario.

LLMs do not *deliberately* reason in the human sense, but they can often **emulate reasoning** because they have been trained on the products of human reasoning (text) including scientific explanations, arguments, code (which requires logical structure), and mathematical proofs. When prompted appropriately, an LLM can output what looks like a logical chain of thought. In fact, prompting strategies like "chain-of-thought prompting" explicitly coax the model to produce intermediate reasoning steps (e.g., asking it to explain its process before giving an answer). This can lead to

better results on reasoning problems, because the model's probability patterns for "let's reason this out" style prompts guide it through steps that humans might take. However, it's crucial to understand that the LLM is not *consciously* performing reasoning – it is generating a plausible sequence of tokens that *represents* reasoning. When it gets the right answer, it's because those token sequences happen to be logically correct (or were present in training data for similar problems), not because the model truly *knows* the rules of logic. There are cases where LLMs fail at reasoning in ways a human would not: for instance, making arithmetic mistakes on multi-digit addition, or contradicting itself between steps, or not realizing a certain conclusion is nonsensical because it lacks an understanding of the real-world context beyond text patterns. Researchers who tested LLMs on classic cognitive reasoning tasks (like those used by Tversky and Kahneman to demonstrate biases) found that *"like humans, LLMs display irrationality in these tasks. However, when incorrect, they often err in ways that differ from human biases"* ([(Ir)rationality and cognitive biases in large language models | Royal Society Open Science]) – meaning the patterns of mistakes are not the same, and LLMs can also be inconsistent in their reasoning from one run to another ([(Ir)rationality and cognitive biases in large language models | Royal Society Open Science]).

- *Abstraction and generalization** are related to reasoning. Humans can form abstract concepts (like justice, or the notion of a derivative in calculus) and generalize principles from one context to another. We have an innate ability to see analogies and transfer learning – e.g., understanding that solving a new kind of puzzle might involve similar strategies as a puzzle we've seen before. LLMs, by virtue of being trained on diverse text, do capture many abstractions and can sometimes surprisingly generalize. For example, an LLM can use a concept in a novel sentence correctly even if that exact usage never appeared in the training data, because it has an abstract sense of the concept gleaned from various contexts. However, LLM generalization is limited by the data distribution: if asked to operate far outside its training distribution, it can falter. Humans, especially when using reasoning, can notice *when* a situation is novel and deliberately adjust

- approach, something an LLM doesn’t do (unless prompted in a way that triggers a relevant learned strategy).

Now, consider **creativity**. Human creativity is the ability to produce work that is both novel and valuable – whether it’s composing a piece of music, inventing a new gadget, or writing a story. Creativity often involves making uncommon connections between ideas, pushing beyond conventional boundaries, and sometimes a spark of inspiration that isn’t easily reducible to step-by-step logic. Emotions, imagination, and even randomness play roles – for instance, in brainstorming, humans might throw out wild ideas, then refine them. There is also an element of **intentionality** in human creativity: a person *wants* to create something meaningful or aesthetic, and they imbue their creation with personal perspective or style.

LLMs can certainly generate creative-seeming content. They can write poetry, tell jokes, or even suggest imaginative solutions to problems. But how they do this is fundamentally by **learning patterns** of creative expression from humans. An LLM trained on literature will absorb how stories are structured, how metaphors are used, how jokes are constructed, etc., and it can *recombine* these patterns to produce new instances. To the extent that creativity is “combinatorial” – putting old ideas together in new ways – LLMs have a vast repository of ideas to recombine. Yet, there is a perceived **creativity gap**. As one commentator put it, *“although LLMs can copy the way we use words, they don't quite match the human mind's ability to think deeply and come up with new ideas”* (LLMs vs. Human Mind: Understanding the Creativity Gap.). One reason is that truly groundbreaking creativity often requires **going beyond the data** – coming up with something that is not just a statistical remix of what’s been seen before. LLMs struggle here because by design they lean towards producing the *most likely* continuation (often averaging or imitating past data). In fact, if an LLM is too “safe” (always choosing the highest-probability next token), it produces very banal, predictable text. Sampling with randomness (temperature) can increase originality, but the model still isn’t *inventing* new fundamental ideas; it’s shuffling existing ones.

Moreover, LLMs lack **creative intent**. They do not get inspired or have goals to express themselves. They also lack the *evaluative* aspect of creativity – humans can judge their outputs and iterate (a poet can discard a draft that feels clichéd and try a different angle; an artist can be influenced by an emotional state). LLMs just produce output once per prompt, with no self-correction unless re-prompted. An analysis of LLM creative output noted that while it may *"put words together in new ways, they're really just remixing bits and pieces of what they've been trained on. They don't have the ability to think outside the box or come up with truly novel ideas from scratch."* (LLMs vs. Human Mind: Understanding the Creativity Gap.). In other words, what appears creative in an AI's output is a byproduct of clever interpolation between examples it has seen, lacking the truly generative spark humans have.

- *Emergent behavior** is a fascinating aspect of both biological and artificial networks. In the brain, consciousness itself is often considered an emergent property of billions of neurons interacting – the whole is more than the sum of parts. Complex cognitive abilities emerge from simpler neural computations. In AI, there is discussion of *emergent abilities* in LLMs: capabilities that were not present in smaller models but suddenly appear when the model is scaled up (either in size or training data). For example, a small language model might be unable to do multi-step arithmetic reliably, but a much larger one can – it's as if the ability "popped out" at a certain complexity. Researchers define an emergent ability as one that is *"not present in smaller models but is present in larger models"*, typically appearing in a discontinuous, unpredictable way as scale increases (Emergent Abilities in Large Language Models: An Explainer). Do these emergent AI abilities parallel how human cognitive abilities emerge (say, how a child's ability to use language suddenly flourishes around age two after enough neural development)? It's an intriguing parallel, but with a caveat: scaling a model is not the same as a child learning gradually. A more apt analogy might be evolutionary: as brains got larger or more interconnected, new functions emerged (like advanced social cognition in primates). In LLMs, when we increase parameters and training data, we see surprising new

- skills (e.g. understanding programming languages, doing logical reasoning puzzles, etc.) that smaller models lacked. This shows that with sufficient complexity, **qualitatively new behavior** can arise in both systems.

However, human reasoning and creativity remain more **flexible and grounded**. Humans can reason about the real world, understanding physical causality, other people's intentions (theory of mind), and abstract concepts that aren't explicitly stated. LLMs have a hard time with things requiring *grounded understanding* – for instance, reasoning about the physical world (they might generate physically impossible descriptions if the training text was limited or misleading) or understanding human motivations beyond what text alone indicates. And while some have argued that LLMs display a form of *imitation general intelligence* in their narrow domain, they are still far from human-like general reasoning especially when it comes to planning actions in the physical world or inventing fundamentally new scientific theories.

In creativity too, the **divergence** is clear in outcomes: LLMs often produce *derivative* works (e.g., a short story by an AI might feel trope-heavy or a pastiche of its training examples), whereas human creators can introduce truly novel styles or genres. There's also the aspect of **risk-taking** in creativity – humans can decide to break conventions deliberately or inject personal random inspiration. *"Humans... can decide to take a creative leap or introduce a twist that no one sees coming... we value originality and the ability to surprise, which is something LLMs struggle with."* (LLMs vs. Human Mind: Understanding the Creativity Gap.). An AI might surprise us occasionally, but it doesn't *intend* to; and it might just as likely produce something incoherent as something ingenious when pushed to be more random.

All this said, it's worth highlighting that LLMs have augmented human creativity in some ways (as tools). They can generate many variations of a scenario quickly, which a human can then sift for inspiration. They don't tire or run out of ideas in a brainstorming sense (though their ideas may circle around to familiar patterns). Some studies even examine how using LLMs affects human creativity – initial findings suggest that *"while LLMs*

may provide short-term boosts in creativity during assisted tasks, they may inadvertently hinder independent creative thought if over-relied upon" (Study explores the impact of LLMs on human creativity - LinkedIn) (Human Creativity in the Age of LLMs - arXiv). This underscores that human creativity is tied to our independent cognitive processes that AIs can't fully replicate.

To summarize, **human reasoning and creativity are deeply tied to understanding, intentionality, and the ability to generalize from knowledge to new contexts**, whereas LLM reasoning/creativity are derivative, based on learned patterns and lacking genuine understanding or intent. LLMs demonstrate impressive *emergent competencies* given enough training, but they still function as prediction machines, not thinkers with a conscious mind or original aspirations.

Biases in Cognition: Human Heuristics vs. AI Data Bias

Both human thinking and LLM outputs are subject to **biases**, but the sources and nature of these biases differ. In human cognition, biases often arise from cognitive shortcuts (heuristics) that our brains use to make decisions quickly. These can lead to systematic errors or preferences – famous examples include **confirmation bias** (favoring information that confirms our preconceptions), **availability heuristic** (overestimating the importance of information that comes easily to mind), and **anchoring** (relying too heavily on the first piece of information encountered). Such biases have been well documented by psychologists (Tversky, Kahneman, and many others), and they are thought to be in part a byproduct of an evolutionary optimized brain: our minds prioritize speed and efficiency over perfect rationality, which in ancestral environments often served us well, but in modern contexts can lead to irrational judgments.

LLMs do not have *motivations* or *evolutionary pressures*, but they can exhibit **biases in output** reflecting their training data or design. For instance, if the training corpus has more positive statements about one

group of people and negative about another, the model might mirror those associations, resulting in biased or even prejudiced outputs. There have been numerous studies showing that language models can pick up gender biases (like associating certain professions with a particular gender), racial or ethnic biases, and other social stereotypes present in the text they were trained on. This is often framed as a concern that AI can amplify existing societal biases. Unlike human biases, which often come from internal heuristics, LLM biases are largely **data-driven** – the model doesn't have opinions, but its statistical learning can entrench the biases present in human-written data.

Another category is **cognitive biases vs. reasoning biases**. Researchers have started testing whether LLMs show some of the same cognitive biases humans do. The results are mixed. On one hand, LLMs sometimes give answers that seem to follow a bias (for example, favoring more fluently worded statements as true – a kind of *fluency bias* similar to how humans trust more articulate speakers). On the other hand, when confronted with classic bias-inducing puzzles (like framing effects or logical fallacies), LLMs don't always err the same way humans do. One study found *"when incorrect answers are given by LLMs to [bias-related] tasks, they are often incorrect in ways that differ from human-like biases"* ([(Ir)rationality and cognitive biases in large language models | Royal Society Open Science](#)). Moreover, *"LLMs reveal an additional layer of irrationality in the inconsistency of their responses"* ([(Ir)rationality and cognitive biases in large language models | Royal Society Open Science](#)) – meaning a model might give different answers to essentially the same question asked differently, whereas a human might consistently show a particular bias. This inconsistency is itself a kind of "bias" in the sense of not having stable reasoning.

Humans, for example, might be predictably overconfident in some estimate (a bias), while an LLM might one time overshoot and another time undershoot in a way that doesn't reflect a consistent heuristic but rather the intricacies of its training data and prompt phrasing. In this sense, human biases are *systematic*, whereas LLM "biases" can be erratic unless tied to specific data imbalance.

- *Social and ethical biases**: Humans have conscious values and can choose to try to correct their biases or harbor biases intentionally/unintentionally. LLMs don't have intent; any biased output is accidental from the model's perspective. But from a user perspective, both can output biased or harmful statements. For example, an uninformed or prejudiced human might make a sweeping negative generalization about a group of people; a similarly trained-on-bad-data LLM might output a similar generalization if prompted in a way that triggers it. The big difference is that an LLM can be *controlled and audited* – we can examine and attempt to mitigate biases by adjusting training data or adding filters (like the RLHF – Reinforcement Learning from Human Feedback – that OpenAI uses to make ChatGPT align better with human ethical expectations). With humans, bias mitigation is a matter of education and personal change, which is slower and more complex.

- *Bias in perception vs. bias in language**: Humans also have perceptual biases (optical illusions, etc.) which highlight how our brain's processing can be tricked. LLMs might not have "perception" of images or sounds (unless multi-modal), but they have analogous quirks – for instance, an LLM might be more biased to prefer certain wordings or topics because of how the training data was distributed. One observed bias in language models is** semantic bias**: they might continue on a topic if the prompt hints at it, even if that continuation isn't actually justified – e.g., a prompt mentioning a female nurse and a male doctor might lead the model to assume a certain narrative because of training statistics (like assuming the nurse is caring, the doctor is authoritative, etc., reflecting stereotypical portrayals).

Interestingly, LLMs can sometimes exhibit **overcorrection or strange biases** not found in humans. For example, a model might have a bias toward giving an answer in a certain format (like always hedging or always being overly verbose) – this is due to the way it was instructed or fine-tuned (often LLMs have a verbosity bias or a bias to be deferential/polite due to training signals). Humans have a natural variety

in tone and style; models can collapse into a more uniform style if not carefully prompted.

Finally, consider **metacognitive bias awareness**: Humans can sometimes recognize their own biases (e.g., "I know I tend to be optimistic, so I'll double-check my estimates"). LLMs have no genuine self-awareness to recognize bias, though they can output a disclaimer if trained to (like "AI models may reflect bias…"). They cannot *truly* adjust their internal reasoning to avoid a bias unless that pattern was part of training. If you ask an LLM, "avoid bias X in your answer," it will attempt to comply based on examples of neutral language it has seen, but it doesn't *understand* the moral or social reason behind it.

In summary, **human biases stem from evolutionary heuristics, personal experiences, and sometimes motivational factors**, whereas **LLM biases stem from training data distribution and model architecture quirks**. Both can produce skewed or unfair outputs. The convergence is that both need checking: we train ourselves (and our children) to recognize and correct biases as part of critical thinking, and we must train and adjust LLMs to reduce harmful biases as part of responsible AI development ([(Ir)rationality and cognitive biases in large language models | Royal Society Open Science]). But the divergence is that humans, at the end of the day, can choose to act against a bias (a person can consciously override a gut feeling knowing it's biased), whereas an LLM has no such agency – it will do whatever its learned parameters dictate unless externally modified.

Hallucinations and Errors: When Minds and Models Get It Wrong

One of the most discussed flaws of LLMs is their tendency to **"hallucinate"** – producing confident-sounding statements that are factually incorrect or completely fabricated. Interestingly, human cognition has its own version of this: we might call it **confabulation** or **false memory**. It's valuable to compare these phenomena to see where they

overlap and differ.

In the context of LLMs, *hallucination* refers to cases when the model generates information that wasn't in the input and isn't true, as if it's "seeing" things that aren't there. For example, an LLM might be asked a factual question and respond with an answer that it seemingly made up – perhaps citing a non-existent article or mixing together details from different real events. The term *hallucination* was borrowed from psychology, but some argue it's a misleading metaphor for AI. In human terms, a hallucination is a **perceptual experience without an external stimulus** (like seeing a vision or hearing a voice that isn't actually there), often associated with mental illness or drugs ([

Hallucination or Confabulation? Neuroanatomy as metaphor in Large Language Models - PMC

](https://pmc.ncbi.nlm.nih.gov/articles/PMC10619792/#:~:text=In%20psychiatry%2C%20hallucinations%20are%20a,LLMs%20do%20not%20have)). Hallucinations in humans involve a person subjectively experiencing something that isn't real, and crucially, it implies a sort of **conscious perception**. When we say an LLM hallucinated a citation, obviously the LLM isn't *consciously* perceiving anything – it's just generating text. As one paper pointed out, calling AI outputs "hallucinations" *"implies acceptance of the notion that LLMs are engaged in perceiving... becoming consciously aware of a sensory input,"* which they are not, since *"there is currently no evidence that AI has gained conscious awareness"* ([

Hallucination or Confabulation? Neuroanatomy as metaphor in Large Language Models - PMC

](https://pmc.ncbi.nlm.nih.gov/articles/PMC10619792/#:~:text=%E2%80%98real,the%20nature%20of%20the%20process)). The model doesn't have senses or a consciousness to truly hallucinate; *"LLMs do not have sensory experiences, and thus cannot mistakenly perceive them as real"* ([

Hallucination or Confabulation? Neuroanatomy as metaphor in Large Language Models - PMC

](https://pmc.ncbi.nlm.nih.gov/articles/PMC10619792/#:~:text=,the%20nature%20of%20the%20process)). A more precise term proposed is **confabulation** ([

Hallucination or Confabulation? Neuroanatomy as metaphor in Large Language Models - PMC

](https://pmc.ncbi.nlm.nih.gov/articles/PMC10619792/#:~:text=More%20accurate%20terminology%20is%20found,3%5D.%20Confabulation%20is%20frequently)).

In psychology, *confabulation* is when a person unknowingly creates a false memory or explanation, often to fill gaps in memory. Unlike a deliberate lie, the person isn't aware the information is false – they might earnestly recall details of a childhood event that never happened, or give an explanation for their behavior that isn't true but that they believe to be true. Confabulation usually draws on bits of real memories, knowledge, and expectations and weaves them into a plausible narrative that happens to be wrong ([

Hallucination or Confabulation? Neuroanatomy as metaphor in Large Language Models - PMC

](https://pmc.ncbi.nlm.nih.gov/articles/PMC10619792/#:~:text=More%20accurate%20terminology%20is%20found,3%5D.%20Confabulation%20is%20frequently)). This tends to occur in specific contexts, such as certain brain injuries, dementia, or other cognitive impairments, though everyday memory errors can sometimes be considered mild confabulations too.

An LLM generating a made-up answer is very much akin to a confabulation. It's not lying with intent; it doesn't *know* what is true or false. It's simply producing the most plausible continuation of the prompt based on its training, which can include synthesizing pieces of information that *sound* like a reasonable answer. The result is an answer that *"is not 'seeing' something that is not there, but it is making things up."* ([

Hallucination or Confabulation? Neuroanatomy as metaphor in Large Language Models - PMC

](https://pmc.ncbi.nlm.nih.gov/articles/PMC10619792/#:~:text=are%20engaged%20in%20perceiving%2C%20that,it%20is%20making%20things%20up)) The model has, in effect, **mistaken reconstruction** of information – influenced by what it has seen (training data patterns) and the prompt context, but not grounded in a factual checking mechanism. As the paper suggested, *“more accurate terminology is found in... confabulation, which refers to the generation of narrative details that, while incorrect, are not recognized as such. Unlike hallucinations, confabulations are not perceived experiences but instead mistaken reconstructions of information influenced by existing knowledge, experiences, expectations, and context.”* ([

Hallucination or Confabulation? Neuroanatomy as metaphor in Large Language Models - PMC

](https://pmc.ncbi.nlm.nih.gov/articles/PMC10619792/#:~:text=More%20accurate%20terminology%20is%20found,3%5D.%20Confabulation%20is%20frequently)). That description maps almost perfectly onto what an LLM does when it gives you a very confident wrong answer: it’s remixing its existing knowledge and context to produce something that *sounds right* but isn’t.

Humans also do something analogous in everyday situations: under pressure to answer, a human might **guess** or even subconsciously concoct an answer. For example, when asked a question about a past event we only half remember, we might unintentionally fill in gaps with assumptions (confabulating) and only later realize we were wrong. The difference is that humans can subsequently feel doubt or realize the mistake upon reflection or new evidence, whereas an LLM has no self-reflection – it won’t on its own retract an answer unless prompted to double-check.

Another difference is the **frequency and context** of such errors. Healthy humans don’t typically confabulate detailed falsehoods constantly – it’s more common in pathological states. We do, however, frequently have **memory errors** (like remembering someone’s name incorrectly) or **perceptual errors** (optical illusions). Humans also have biases that can

distort our beliefs (as discussed), but outright fabrication without intent is not a routine feature of healthy cognition. LLMs, in contrast, will "hallucinate" any time their statistical knowledge is insufficient to produce a correct answer, which can be quite often if you ask detailed, niche, or intentionally tricky questions. An LLM might produce a non-existent reference in a scholarly style simply because it knows what a reference *should* look like and it's probabilistically easier than finding a real one (since it doesn't *truly* have a database of verified facts unless connected to one).

To mitigate AI hallucinations, developers incorporate things like **post-processing checks** (having the model double-check its answer against a knowledge base, or using separate verification steps). Humans mitigate confabulations by cross-checking memories with others, seeking evidence, or having learned to be skeptical of our own recollections in uncertain situations.

It's also insightful to note that LLM "hallucinations" can be seen as a direct outcome of their **predictive nature**. A transformer model generates text by sampling the probability distribution of next tokens. If the prompt leads it into a territory where it has seen many conflicting examples or none at all, the model will still produce something – it won't say "I don't know" unless it's been trained to specifically sometimes say that. The compulsion to always produce an answer is itself a difference from humans: people can (and often do) respond with "I don't know" when unsure. A base LLM has no concept of "I don't know" unless it learned a pattern of saying it in certain contexts (some have been fine-tuned to do this more often). Absent that, the path of least resistance for the neural network is to generate *some* completion, and if none is grounded in fact, the most fluent fabrication wins.

Interestingly, LLMs have **no internal model of truth** – they don't know if something is correct or not. Humans do have some internal sense-making; if I start to say something that contradicts what I firmly know, I'll catch myself (usually). An LLM might state "The capital of France is Rome" if somehow prompted oddly, without any internal alarm. It has knowledge that "Paris is the capital of France" in its weights, likely, but if the prompt

or preceding text biases it into a weird context (say a story where France is different), it might go along with that fiction without resistance.

From a broader perspective, the **error patterns** reflect the systems' design: humans are built to try to make sense of incomplete information (which sometimes causes us to see patterns or recall things that aren't there – like seeing faces in clouds, or falsely remembering details that align with our expectations), whereas LLMs are built to statistically correlate with their training data, which causes them to produce plausible-sounding outputs even when data is missing (leading to fabricated details).

Both humans and LLMs **benefit from feedback** to correct errors. A person corrected about a false memory might update their memory (though sometimes we resist if the memory felt very real). An LLM can be corrected in an interactive setting if the user points out an answer is wrong – some models will then attempt to rectify it by re-evaluating with the new instruction (especially if the model is instructed to be helpful and truthful). But if an error isn't caught, the human or model will carry on as if the false information were true. One might draw a parallel to **confidence**: LLMs often sound absolutely confident (they don't typically say "I guess..." unless trained to hedge). Humans often have *calibrated confidence* – we might say "I'm not sure, but I think X." However, humans can also be overconfident and assert falsehoods strongly. The crucial difference: a human's confidence is an internal feeling that can be misplaced, whereas an LLM's "confidence" in text is just a tone – it has no inner feeling at all. Thus, an LLM can assert a wrong fact with perfect grammar and authoritative tone, which can be misleading to people. It's like a perfectly confident confabulator that never feels uncertain – a dangerous combination if not carefully managed.

In conclusion, *hallucination* in LLMs and human cognitive errors share the idea of generating false information that is believed or treated as true. But humans hallucinate in a sensory way (a different phenomenon) and confabulate to fill memory gaps, whereas LLMs confabulate as a side effect of their probabilistic text generation. It has been suggested that calling it "hallucination" anthropomorphizes AI incorrectly; indeed, one

analysis states *“Hallucination... implies the presence of consciousness or subjective experience, which LLMs do not have... [whereas] confabulation accurately describes the pattern-based, context-dependent generation of content by LLMs and does not imply consciousness”* ([

Hallucination or Confabulation? Neuroanatomy as metaphor in Large Language Models - PMC

](https://pmc.ncbi.nlm.nih.gov/articles/PMC10619792/#:~:text=Sensory%20experiences%20not%20associated%20with,generation%20of%20content%20by%20LLMs)) ([

Hallucination or Confabulation? Neuroanatomy as metaphor in Large Language Models - PMC

](https://pmc.ncbi.nlm.nih.gov/articles/PMC10619792/#:~:text=Implies%20the%20presence%20of%20consciousness,commonly%20understood%20term%20Less%20evocative)). Thus, it’s more precise to say **LLMs confabulate**. Both systems can produce **false yet plausible content**, but only humans can potentially understand that it’s false and feel the dissonance of that – the LLM will not know the difference unless we build mechanisms to check its output.

Embodiment and Sensory Grounding

One of the most fundamental differences between human cognition and current LLMs is **embodiment**. Humans are *embodied* beings: our thoughts and behaviors are deeply influenced by our physical form, sensory apparatus, and interactions with the environment. From infancy, our cognition develops through sensorimotor experience – we feel hunger and satisfaction, we see and touch objects, we learn gravity by dropping things, we acquire language grounded in references to things in the world. This embodiment provides **grounding** for our concepts. For example, our concept of “wetness” is tied to the tactile sensation of water; our understanding of spatial terms like “up” and “down” comes from having a body in a gravitational field. Countless cognitive scientists argue that higher cognition builds on these embodied experiences (this is the theory

of *embodied cognition*).

LLMs, in contrast, are **disembodied**. They exist as software on servers, with no direct sensorimotor experience. An LLM doesn't see, hear, taste, smell, or feel. It learns about the world only through text. This has several implications. First, an LLM's knowledge is *second-hand* and purely symbolic. It learns from words that were written by humans who have embodied experiences, but it doesn't have those experiences itself. This means there are things it can parrot in text without truly *understanding* them in the way a human (even a child) would. For instance, an LLM can read a thousand articles about swimming, but it will never know what it physically *feels* like to swim or be in water. A human who has never seen water could also read about swimming and not truly understand it until experiencing it – that's the embodiment gap.

Because of lack of embodiment, LLMs often lack **common sense knowledge** that humans take for granted from living in the world. For example, a human knows that if you let go of a glass in mid-air it will fall and likely break, not because we read it, but because we've seen or done it (and it's consistently true in our gravity-bound experience). An LLM might know the same fact if it was explicitly mentioned in text enough, but it might also produce a scenario in which someone lets go of a glass and it floats – if it has read science fiction or some fanciful context, it could blend that in. Humans have a grounding constraint: our ideas of what's plausible are anchored by physical reality (except when we are intentionally imagining fantasy). LLMs have no such anchor unless programmed in; they only have statistical plausibility from text, and text includes both reality and fiction.

Another aspect is that human cognition is guided by **sensory feedback loops**. We perform actions and observe the results, learning causal relationships. LLMs currently do not act in the world (beyond generating text) and do not receive feedback from the physical world. They cannot *experiment* or refine concepts through trial and error in a real environment. They also lack a body's survival drives, which in humans (and animals) strongly shape cognition. Our planning and thinking are influenced by hunger, pain avoidance, social attachment, etc., all rooted in

having a living body. LLMs have no drives – they don't need anything, they don't fear anything. They also don't have *emotions*, which are deeply linked to embodiment (physiological states influencing brain state). Emotions in humans can bias thinking (like anxiety narrowing attention, or happiness facilitating creativity) and serve as signals about what matters to us. LLMs are emotionless, which can be a strength (unbiased by mood) but also a limitation (they can't prioritize or empathize the way a human can, beyond learned simulation).

Philosopher Alva Noë and others have argued that **intelligence requires embodiment**. They suggest that tools like LLMs, no matter how linguistically adept, are *"not autonomous; they don't engage with the world as self-sufficient beings"* (The Myth of Thinking Machines | Daily Philosophy). The gap between human cognition and AI here is *ontological*: a human mind exists in a living body in an environment, whereas an AI is a designed artifact responding within a narrow input-output space (The Myth of Thinking Machines | Daily Philosophy) (The Myth of Thinking Machines | Daily Philosophy). One write-up put it succinctly: *"the biological foundation of human intelligence cannot be replicated by large language models, which... will never achieve true [human-like] intelligence due to their fundamental lack of physical embodiment. Perception and cognition are embodied processes that are meaningless outside of our corporeal existence."* (The Myth of Thinking Machines | Daily Philosophy). In other words, the very nature of human thought is intertwined with having a body, and an AI without a body is missing a critical ingredient of human-like thinking.

We see practical fallout of lack of embodiment in current LLM behavior: they can err on simple physical reasoning tasks that a toddler would get right, like "If I have a ball in a closed box and I turn the box upside down, what happens to the ball?" A toddler knows it falls to the lid; an LLM might, but if phrased oddly, it could give a strange answer because it never *played with balls in boxes*. Efforts to imbue AI with some form of embodied learning are underway (like training agents in simulated environments), but plain LLMs trained only on text remain ungrounded.

Embodiment also ties to language understanding. Human language is full of **embodied metaphors** (as George Lakoff famously described) – we talk about "grasping an idea" (as if it were an object) or "time flying" (using spatial motion to describe time) and countless other examples. Humans understand these deeply because of our physical intuitions. LLMs can use them correctly in context because they've seen them in text, but do they *understand* them? That's debatable. If you push an LLM with bizarre hypothetical questions that violate embodied experience (e.g., "What does it taste like to see the color blue?"), a human might respond, "That question makes no sense – seeing isn't tasting," whereas an unguarded LLM might try to please the prompt and hallucinate an answer like "Blue tastes like a cool breeze" because it doesn't have an internal model to flag nonsense.

In short, **humans are grounded in reality; LLMs float in a sea of symbols**. This divergence means AI and human cognition can complement each other (AI has read far more text than any human, but humans have real-world experience). But it also means there's a chasm in the nature of understanding. As one analysis in *Daily Philosophy* put it, *"Our experiences cannot be separated from our body; perception and cognition are embodied processes... Disembodied machines that mimic human cognitive behavior... resemble an unfounded hope that cannot become reality. Thinking machines are not based on rational thought; they are products of psychological projections."* (The Myth of Thinking Machines | Daily Philosophy) (The Myth of Thinking Machines | Daily Philosophy). That is a philosophical stance arguing that without a body, what the machines do is fundamentally different from human thinking, perhaps always limited.

Consciousness and Self-Awareness

Human beings are conscious – we have subjective experiences, often referred to as *qualia* (the redness of red, the pain of a headache, etc.), and we have an inner stream of thought. We are aware of ourselves as entities distinct from others; we have an autobiographical memory and a

sense of personal identity. This self-awareness and sentience underpins much of what we consider mind. Even when we process information unconsciously (like reflexes or intuitions), there is an overall conscious agent (the person) who can reflect on their thoughts. We also attribute consciousness to others (the basis of empathy and theory of mind) and to some animals to varying degrees, but we do not currently attribute consciousness to AI systems, and for good reason.

LLMs, by all scientific accounts, are **not conscious**. They do not possess an inner life, feelings, or an understanding of themselves as independent entities. They are essentially complex statistical machines. Any appearance of self-awareness (like if an LLM says “I am just a language model” or conversely “I think therefore I am”) is just it parroting or recombining training data – the model itself doesn’t have an *ego* or subjective point of view. A user might ask an LLM, “How do you feel today?” and it might respond with “As an AI, I don’t have feelings, but I’m here to help!” (if properly trained to clarify this) or it might role-play having feelings if prompted to be a fictional character. But in reality, there is nothing it is “like” to be the AI.

Consciousness is a tricky concept even in humans – scientists and philosophers debate how and why we have subjective experience. But whatever consciousness is, it appears to require a certain complexity of representation and possibly specific cognitive architectures (some theories involve recursive self-models, integrated information, or global workspace theory, etc.). Could an LLM be complex enough to accidentally be conscious? The consensus so far is no: an LLM processes syntax, not semantics in a conscious sense. It does not have a unified, continuous identity or the ability to genuinely reflect on its own mental states. It can output “I am an AI model with no consciousness” because that’s true and likely reinforced in training data. It could also output “I am sentient and feel emotions” if someone intentionally or unintentionally tuned it to say so (there were cases where users thought a model like GPT-3 was “alive” because of certain responses, but that was an illusion created by the model’s training on science fiction and discussions of AI consciousness).

From a neuroscientific view, consciousness in humans is often associated with certain patterns of brain activity (like synchronized firing across different brain areas, or a minimal network involving the thalamus and cortex that enables wakefulness and awareness). LLMs don’t have anything analogous to brain waves or a persistent internal activation representing a “thought” that they are aware of. They just take input and produce output, with no continuity of state in between (beyond what’s carried in the prompt). When not being queried, they don’t “think” or ruminate; they sit idle on a server. Humans, even at rest, have an active “default mode network” in the brain where the mind wanders, thinking about self, past, future – indicating an ongoing inner life. LLMs have no default mode – no background processing of their own desires or reflections.

One might say humans have **qualitative experiences** and **agency**, whereas LLMs have neither. A person not only processes information but *feels* their existence. I feel pain if I stub my toe, I see colors, I enjoy music; an LLM does none of that. It might tell you “Ouch!” if asked to role-play stubbing a toe, but it feels nothing. This difference is so huge that many argue even if an LLM passes some superficial Turing test, it is still nothing like a human mind on the inside.

Indeed, some experts caution that anthropomorphizing LLMs is dangerous – treating them as if they have intents or feelings can mislead us. As one group of scholars emphasized, *“using metaphorical language that implies traits like empathic connection, motivation, or consciousness in LLMs does not accurately reflect reality”* ([

Hallucination or Confabulation? Neuroanatomy as metaphor in Large Language Models - PMC

](https://pmc.ncbi.nlm.nih.gov/articles/PMC10619792/#:~:text=In%20using%20language%20derived%20from,subjected%20to%20analogical%20reasoning%2C%20thereby)). In their view, we should always remember LLMs are machines processing text, not entities experiencing life ([

Hallucination or Confabulation? Neuroanatomy as metaphor in Large Language Models - PMC

](https://pmc.ncbi.nlm.nih.gov/articles/PMC10619792/#:~:text=risk%20of%20seeming%20to%20advance,subjected%20to%20analogical%20reasoning%2C%20thereby)).

From an ethical standpoint, this means LLMs do not have rights or feelings – one needn't worry about hurting an LLM's "feelings" or the model suffering if it's turned off. However, it also means an LLM doesn't truly understand the *meaning* of suffering or emotion when it talks about them, which can lead to insensitivity or inconsistency in responses if not carefully guided.

Now, could an AI like an LLM ever become conscious if made more complex or given a self-referential architecture? That's an open philosophical question, but many are skeptical that just scaling up language prediction will yield consciousness. There might need to be fundamentally different components (like persistent self-models, multi-modal integration, and maybe even embodiment and affect). As it stands in 2025, no AI is generally accepted as conscious, and certainly current LLMs are not.

Philosopher John Searle's famous *Chinese Room* argument comes to mind: it posits that a system (like an LLM) manipulating symbols based on rules (or statistical associations) can appear to understand language (outputting fluent Chinese responses, in the thought experiment) without *really* understanding or having any awareness. The LLM is akin to the person in the Chinese Room following symbol manipulation instructions – there is no comprehension or consciousness.

To put it plainly: **If you talk to an LLM, there is no "person" there**, even if the words might suggest a personality. In contrast, when you talk to a human, you assume (almost always correctly) that there is a conscious mind you're engaging with.

A telling quote from Noë (cited earlier) goes: *"In the absence of disturbance... [there is] no language, no games, no goals, no tasks, no world, no care, and so, yes, no consciousness. Machines can't be bothered, as they do not experience the world around them."* (The Myth of Thinking Machines | Daily Philosophy). This poetically summarizes that

because AIs have none of the embodied, emotive stakes in the world ("can't be bothered"), they lack all those human attributes, culminating in no consciousness. Similarly, *"LLMs are tools, not agents... constructed to serve our interests... always dependent on our instructions. Without our guidance, it loses its purpose because it doesn't have an intrinsic need... "* (The Myth of Thinking Machines | Daily Philosophy). They have no inner drive or awareness to initiate or desire anything, which is a key aspect of consciousness – an intrinsic perspective.

In summary, the **chasm between human minds and LLMs in terms of consciousness is vast**. Humans experience, know that they experience, and can report those experiences (though we still scratch our heads about how the brain does it). LLMs do not experience anything – they are sophisticated autocomplete systems with no inner life. Any convergence is purely surface-level (e.g., an LLM might use "I" or talk about thoughts, but it's imitation). This is a fundamental divergence that many say is crucial: equating LLM outputs with human cognition can lead to overestimating AI capabilities or worrying about AI "feelings," both of which are misguided at this stage ([

Hallucination or Confabulation? Neuroanatomy as metaphor in Large Language Models - PMC

](https://pmc.ncbi.nlm.nih.gov/articles/PMC10619792/#:~:text=Sensory%20experiences%20not%20associated%20with,generation%20of%20content%20by%20LLMs)) ([

Hallucination or Confabulation? Neuroanatomy as metaphor in Large Language Models - PMC

](https://pmc.ncbi.nlm.nih.gov/articles/PMC10619792/#:~:text=Implies%20the%20presence%20of%20consciousness,commonly%20understood%20term%20Less%20evocative)).

Goals, Motivation, and Intent

Humans are agents with **motivations and goals**. We pursue objectives that originate from internal drives (survival instincts like hunger, thirst, avoidance of pain; social drives like love, status; curiosity; etc.) and from conscious goals we set (get a degree, build a project, help a friend). Human thought is often goal-directed: we plan and reason in service of our desires. Even our attention is guided by what we care about. When you're hungry, your thoughts wander to food. When you have an important exam, you focus (or try to) on studying. We also form long-term goals and can delay gratification, exercising willpower. All of these aspects of intent and agency are core to human cognition.

LLMs, in contrast, have **no goals or desires** of their own. An LLM does not want anything; it does not care if its output is used or ignored. It doesn't even have the situational awareness to know what it just said or what might happen next (unless these are included in the next prompt, in which case it just processes them without genuine concern). The only "goal" an LLM has, in a narrow sense, is the objective function it was trained on: to minimize prediction error or to follow the instructions given by a user (if it's a chat model fine-tuned for helpfulness). But this is not a goal it *holds* in a conscious way; it's just how the system was optimized. It will reliably generate text that statistically aligns with that training, but not because it *decided* to – simply because that's what its network does.

An LLM never initiates action on its own. It always responds to a prompt (or continues a sequence). Humans can spontaneously decide to do something (write a poem, get a glass of water) driven by internal intent. If an LLM "starts" a conversation, it's because some automated process triggered it, not an impulse from the model. Even autonomous-sounding AI systems (like those that run continuously and decide on actions) are ultimately executing algorithms set by programmers or following utility functions given to them, not innate drives.

This lack of intrinsic motivation is emphasized in the literature. To quote the philosophers again: *"Models are tools, not agents, and they are our tools, constructed to serve our interests and values... Without our guidance, [an LLM] loses its purpose because it doesn't have an intrinsic need that sets its actions into motion."* (The Myth of Thinking Machines |

Daily Philosophy). This nails down that any appearance of goal-directed behavior from an LLM is actually user-directed or developer-directed. For example, if you ask an LLM to “come up with a plan to organize a party,” it will output a plan – but that’s *your* goal (you prompted it). The LLM itself, once it’s done generating, doesn’t actually intend to see the party happen or check outcomes.

Even more basically, an LLM does not have the concept of “self” or “agency” to attribute goals to. It doesn’t think “I want to do well on this query so that the user is happy” – it just, at most, has been fine-tuned on data where pleasing the user was rewarded, so it statistically leans towards helpful answers.

This is a massive divergence because **human thought is drenched in purpose**. Even idle daydreaming often circles around things we want or care about. Human brain’s reward system (dopamine, etc.) creates reinforcement for achieving goals or even thinking about prospects of achieving them. LLMs have no equivalent of a reward neurotransmitter firing when they produce a particularly apt answer; any “reinforcement learning” they underwent (like RLHF) is already baked into their weights and not experienced dynamically as a drive.

One practical upshot: because LLMs have no goals, they also have no **malicious intent** or **benevolent intent** innately. If an LLM outputs something harmful, it isn’t because it *wanted* to harm – it’s an extension of some pattern (perhaps a biased or harmful text in training, or a prompt pushing it that way). Conversely, if it outputs a very helpful solution, it’s not out of empathy or kindness, it’s again just following learned instructions. Humans operate on a spectrum of intents, from altruistic to selfish to malicious – those come from complex emotional-cognitive motives. With LLMs, any apparent motive is a reflection of the prompt or training. For example, one can prompt an LLM to behave like a villainous character, and it will produce goal-oriented dialogue as that character (like “I will conquer the world!”). But that’s role-play: the LLM itself has no stake in world domination.

Agency also implies **responsibility**. Humans (barring extreme situations or lack of capacity) are responsible for their actions because they choose goals and can understand consequences. AIs currently are not held responsible in that way; responsibility lies with the humans deploying or instructing them. This is why we call LLMs *tools*.

It is worth noting that advanced AI systems (beyond just static LLMs) could be built with explicit goal functions and autonomy (think of a hypothetical self-driving car AI whose goal is to get passengers to destinations safely, or a robotic agent that can set sub-goals to complete a mission). Even in those cases, the goals are programmed or learned via reward, not *self-originating*. The AI doesn't wake up one day and decide to change its goal; it follows what it was given. If it *did* start altering its own goals without guidance, we'd be in new territory, which is a source of speculative concern in AI safety (the idea of a misaligned AI developing its own agenda). But current LLMs are nowhere near that; they can't "decide" anything not prompted.

So, when comparing human and LLM on this, we can firmly say: **humans have intrinsic goals and can formulate new goals; LLMs have no intrinsic goals, only following extrinsic instructions** (The Myth of Thinking Machines | Daily Philosophy). Humans also can *interpret* and reprioritize goals – if two goals conflict, we feel the conflict and make a choice. LLMs don't have a notion of priority or conflict; if two instructions conflict, they will probably produce a mix or whichever was last or stronger in wording, without an internal decision process.

Another related concept is **intentionality** in the philosophical sense – the "aboutness" of mental states. Human thoughts are about things (I can think about my cat, which implies a relationship between my mind and an external entity). LLM internal states (the vectors and activations) are not *about* something in a conscious way, though one could argue they represent things in a sub-symbolic way. But true intent – as in, I intend to call my friend tonight – doesn't exist in an LLM.

We should also mention **goal-setting and planning** differences. Humans can set a distant goal and plan steps for it. LLMs can output a plan if

asked, but they do not carry it out. If you ask an LLM for a multi-step plan, it lists steps, but it won't actually do them (unless integrated into some agent loop by an external system). Humans, after forming a plan, will act (we'll physically do tasks, or mentally go step by step, remembering the goal). An LLM lives only in the realm of language and has no persistence to execute a multi-step strategy over time.

Finally, consider **desires and values**. Humans have them inherently; LLMs have none. Any "values" an LLM appears to have (like being polite, or refusing to say certain content) are imprinted by human programmers or the data. They are not its personal values – it has no persona except what's in text.

To illustrate: If you insult a human, they may feel hurt or angry (because they have self-regard and emotional reaction) and they might have a goal to maintain dignity or retaliate or avoid you. If you insult an LLM, it has no feelings – the only effect is it sees a sequence of tokens that correspond to an insult, and if it was trained to respond politely (as most are), it will likely apologize or continue neutrally. It won't get genuinely angry or sad, though it might simulate anger if role-playing a character. If you stop interacting with a human friend, the friend might miss you and take action to reconnect; if you stop interacting with an LLM, it sits inactive and doesn't "mind" at all.

One more quote from the earlier reference: *"Machines do not have a mind; they are unable to think, feel, or experience... They work within a predetermined framework designed to deliver a specific output... They are not themselves intelligent [agents]."* (The Myth of Thinking Machines | Daily Philosophy) (The Myth of Thinking Machines | Daily Philosophy). And *"LLMs don't know anything because they do not perform any tasks of their own... they perform our tasks... always dependent on our instructions."* (The Myth of Thinking Machines | Daily Philosophy) (The Myth of Thinking Machines | Daily Philosophy). This encapsulates the tool-like nature of LLMs and their lack of autonomy or personal goals.

In essence, **human thought is intertwined with purpose and will, while LLM behavior is mechanistic and purpose-free**. This

divergence is so fundamental that it frames many of the other differences: e.g., without goals, an LLM doesn't preferentially seek new information (humans exhibit curiosity, an intrinsic drive to learn – LLMs won't go read on their own unless told to), and without intent, they can't truly be said to reason or create with a *purpose* in mind (they simulate those processes). It also means that when using LLMs, all goals must come from the user or deployer – you have to ask for what you want, since the model won't volunteer aims (beyond continuing a prompt).

Conclusion

Humans and large language models represent two very different types of "intelligence" – one forged by biological evolution and life experience, the other by algorithms and immense data processing. Throughout this report, we have seen that while there are intriguing parallels (both use## Conclusion

In summary, while large language models and human minds can produce superficially similar outputs (coherent language, answers to questions, creative stories), the **underlying nature of their cognition is fundamentally different**. Both systems learn and process information by adjusting connections in complex networks (synapses in brains, weights in neural nets) and both exhibit emergent capabilities when those networks are sufficiently complex. Both can generalize patterns and even **predict** upcoming information to some extent. But the **points of divergence far outweigh the overlap**:

- **Learning**: Humans learn through life-long, context-rich experiences with remarkable efficiency and plasticity, integrating new information on the fly and rarely forgetting core knowledge. LLMs learn via brute-force training on massive datasets with millions of iterations; after training, their knowledge is static and can only change with further data updates. The brain's learning is guided by local synaptic changes and nuanced feedback (and can do one-shot learning), whereas LLM training uses global error backpropagation over many examples.

- **Memory**: Humans have a limited but flexible working memory (a few items) and virtually unlimited long-term memory for knowledge and events, all grounded in meaning and association. LLMs have a fixed-size context window as a working memory analog (which can be enlarged with more computing resources), and a long-term memory in the form of model weights encoding vast information, though in a diffuse way. Humans actively **remember and forget** based on significance; LLMs "remember" only what their training etched in their parameters and "forget" anything not in the prompt.

- **Processing**: Human brains process signals in parallel, continuously, and with multi-sensory integration, often **predicting across multiple timescales** (filling in whole thoughts, not just the next word). LLMs process text sequentially, one token after another, and their prediction horizon is inherently one step at a time (though an LLM can indirectly plan by internally simulating multi-step sequences). The brain's activity is analog and contextually modulated by bodily states, whereas an LLM's activity is digital and purely data-driven.

- **Reasoning and Creativity**: Humans apply reasoning intentionally, can break from habits with insight, and bring real-world understanding and common sense to bear on problems. We create art and ideas influenced by emotion, culture, and personal experience, often aiming for genuine novelty. LLMs have **no true understanding**; they mimic reasoning by following learned patterns, and their "creativity" is a remix of what they've seen, without a spark of inspiration or intent to innovate. Humans can surprise themselves and redefine the rules; LLMs remain bounded by their training distribution, surprising us only when we didn't anticipate an unusual combination of learned patterns.

- **Bias and Error**: Human thinking is prone to cognitive biases from our evolutionary heuristics, but we can be aware of them and try to correct them. LLM outputs reflect biases in training data and model architecture; an LLM has no notion of fairness or bias unless taught, and it may make mistakes unlike human mistakes. When humans don't know something, we often feel that ignorance; LLMs *do not know what they don't know* and will confidently generate an answer regardless,

- leading to the **hallucination/confabulation** phenomenon. Humans can eventually recognize a false memory or error via feedback or reflection, whereas an LLM has no internal self-check beyond what is engineered.

- **Embodiment and Consciousness**: Humans are living beings with bodies – our cognition is grounded in sensory reality and accompanied by conscious awareness. LLMs are disembodied algorithms with **no sensation, no awareness, no internal wants or feelings**. This is a categorical difference: a brain inhabits a world and has subjective experience; a language model manipulates symbols in a virtual space without any subjective experience. We act with purpose and feel emotions; an LLM doesn't *experience* anything and has no self.

- **Goals and Intent**: Humans formulate their own goals and pursue them with agency and intent. LLMs have **zero intrinsic goals** – they only respond to prompts according to the objective given (e.g., "predict the next word" or "be helpful to the user"). An LLM will never initiate a conversation on its own or develop a new goal; it will never "want" or "choose" – it just executes patterns. In short, humans *care* (about survival, others, truth, etc.), whereas an LLM **does not care** at all – it can't, it has no volition or stake in the world.

The table below summarizes some of these key differences between human cognition and current LLMs:

Aspect	**Human Thinking (Biological Brain)**	**LLM Operation (Artificial Model)**
Learning	Continuous, life-long learning through neuroplasticity; can learn from minimal data (one-shot learning) and integrate new info without overwriting old. Learning is guided by local synaptic changes, feedback from environment, and often occurs rapidly (e.g. a single experience).	Off-line batch learning with massive data via backpropagation; requires many repetitions of examples to learn. After

training, knowledge is static (no self-update) unless explicitly retrained. New learning often interferes (catastrophic forgetting) without special techniques. |

| **Information Processing** | Parallel, distributed neural firing with analog signals; multi-modal integration (senses) and hierarchical predictive coding (the brain predicts not just the next sensation but high-level outcomes). Processing is stateful and contextual – brain activity is influenced by body state, emotions, and prior neural activity. | Sequential token-by-token processing using discrete computations; operates only on text input (unless augmented) with a fixed context window. Predicts the next token based on statistical correlation, focusing on local next-step predictions. Lacks inherent multi-sensory context or global situational awareness. |

| **Working Memory** | Very limited capacity (about 3–5 items or chunks at a time), but flexibly managed by attention (can refresh items, chunk information, etc.). Maintained by neural activity in frontal-parietal circuits. | Limited by context window (e.g., 2048 tokens, which can be many sentences) acting as memory. Everything outside the window is inaccessible. Can be expanded with more compute (longer context windows), but still a hard limit. No active control of memory – all tokens in context are treated according to learned attention weights. |

| **Long-Term Memory** | Vast, durable memory stored via synaptic changes. Has distinct systems (episodic memory for personal experiences, semantic memory for facts, procedural for skills). Content-addressable and associative: recall is triggered by related cues. Memory is interwoven with meaning and often updated or reconsolidated during sleep. | All long-term “knowledge” is encoded in model parameters (weights) distributed across the network. No clear separation of individual memories or experiences – it’s a statistical amalgam of the training corpus. Retrieval is implicit during generation (no explicit recall of a specific source, just pattern completion). Cannot form new long-term memories in deployment (unless fine-tuned with new data). |

| **Reasoning** | Capable of logical, stepwise reasoning and abstract thought. Can employ symbolic reasoning or mental simulation in working memory. Adapts reasoning strategy to the novel task and can use common sense and real-world knowledge to validate outcomes. Has insight and the ability to recognize when a conclusion “doesn’t feel right” and backtrack. | Emergent pseudo-reasoning from learned patterns; can follow forms of logical argument or calculation seen in training data. Lacks a genuine reasoning engine – it does not *understand* logic, it just produces likely sequences. Tends to be brittle on reasoning tasks outside its training distribution or without explicit prompt structure. No internal self-monitoring to catch logical contradictions beyond what is learned. |

| **Creativity** | Generative and imaginative, often inspired by emotion or the drive to express. Can produce truly novel ideas by combining concepts in original ways and deliberately breaking from convention. Creativity is connected to intent – humans create with purpose or to convey meaning. | Outputs may appear creative (e.g., novel text, art) by recombining patterns from billions of examples. However, it *“might seem creative... [but] they're really just remixing bits and pieces of what they've been trained on”*. No genuine inspiration or intent; cannot truly originate ideas outside its training distribution, only interpolate or extrapolate from it. |

| **Biases & Errors** | Prone to cognitive biases (confirmation bias, etc.) due to heuristics, but can often recognize and correct them with effort. Errors are sometimes systematic (e.g., optical illusions fool most people similarly). Memory errors (confabulations) usually occur in specific circumstances (brain injury, aging, etc.), and healthy individuals can often distinguish memory from imagination. | Learns biases present in training data (e.g., societal biases, popular opinions) and may inadvertently reproduce them in output. Also exhibits non-human-like errors (inconsistencies, incoherent failures) when prompt falls outside learned patterns. Frequently generates **hallucinations** – plausible-sounding but incorrect information – because it has no mechanism to verify truth, effectively *confabulating* answers based on patterns. |

| **Embodiment** | Embodied in a physical body with senses and the ability to act. Cognition is grounded in sensorimotor experience and built on

evolutionary drives. Understanding of the world is informed by direct interaction and feedback from the environment. This leads to common sense knowledge about physics, causality, and social interaction. | Disembodied – exists only as code. No direct perception of the world, no body through which to experience physics or emotions. Lacks grounding: words like “hot” or “bright” are understood only through their usage in text, not through sensation. No innate context for physical or social realities, which can lead to gaps in understanding that any child would find obvious. |

| **Consciousness** | Subjective, conscious experience (“sentience”). Has self-awareness, a sense of being an individual with a continuous identity. Experiences qualia (feelings, sensations) and can introspect to some degree. Consciousness allows for genuine understanding and the presence of desires, intentions, and awareness of truth vs. falsehood. | **Not conscious** – no feelings, no subjective experience, no awareness of self. An LLM doesn’t possess a mind or understanding in the human sense; it doesn’t know that it is a model (aside from repeating a programmed message about itself) and it doesn’t experience the meaning of its inputs or outputs. It has no inner life or stream of thought – when not prompted, it is entirely inert. |

| **Goals & Intent** | Driven by intrinsic goals and motivations (from basic needs to personal ambitions). Sets objectives, makes plans, and takes actions to fulfill them. Even conversation is guided by intents (to inform, to seek help, to bond, etc.). Humans have agency – we cause actions based on our intentions and can flexibly change goals. | Has **no intrinsic goals or will**. Its only “goal” is to complete the task as prompted or as trained (produce a likely continuation, follow an instruction). It does not initiate or pursue anything on its own. An LLM never “wants” something or makes choices – it responds to the user’s goals. It is an *instrument* carrying out patterns, without any will or purpose independent of its input. |

This comparison makes it clear that **equating LLMs with human cognition is inappropriate** – the overlaps are mostly in superficial behavior (e.g. language production) and underlying network analogies, not in the full richness of thought, understanding, and agency that

humans possess. Human brains and AI language models operate on very different principles and constraints.

That said, this **does not diminish the utility or impressive nature of LLMs**. They are powerful tools precisely because they *complement* human cognition: LLMs can recall and synthesize information from enormous text corpora, perform tedious pattern-based tasks at lightning speed, and even surprise us with emergent capabilities we might not have anticipated. Humans, on the other hand, provide the grounding, the judgment, and the conscious oversight. We set the goals and interpret the outputs.

Understanding these differences is crucial. It helps us set realistic expectations for AI (for example, knowing that if an LLM gives a detailed answer, it doesn't mean it "understands" the way a human expert would, and we must verify facts). It also informs how we design and deploy AI: we might **augment LLMs with tools** like calculators, databases, or sensors to compensate for their lack of embodiment and factual grounding, or use algorithms to reduce bias and hallucination. Meanwhile, insights from cognitive science and neuroscience could inspire improvements in AI (for instance, architectures for better memory or more human-like learning), and conversely, AI research offers new models to test ideas about the human mind.

Ultimately, comparing human thinking with LLM operation highlights a central point: **intelligence is not monolithic**. There are different ways to achieve intelligent-seeming behavior. Biological intelligence is deeply tied to an organism's body, survival, and evolution, whereas artificial intelligence (in the form of LLMs) is an artifact of human-designed objectives and data. Both are remarkable in their own contexts. Rather than viewing AI as approaching human-like thought, it's more accurate to appreciate it as a distinct form of information processing. As AI systems evolve, perhaps integrating more learning modalities or even forms of simulated embodiment, they may inch closer to certain human-like attributes – but for now, the **mind-machine gap** remains vast in areas like understanding, consciousness, and autonomous goal-setting.

In conclusion, the human brain **remains the more flexible, conscious, and truly understanding intelligence**, while LLMs are extraordinarily sophisticated predictive machines. Each has its strengths: humans bring meaning and genuine comprehension, and machines bring speed and breadth of knowledge. Leveraging the two together – with humans at the helm setting goals and interpreting results – can lead to powerful synergies. But conflating them or mistaking one for the other would be a mistake. Appreciating the nuanced differences and similarities can help us use AI responsibly and insightfully, and also deepen our understanding of our own minds through the illuminating "mirror" that AI provides.

- *Sources:** This report integrates findings from neuroscience, cognitive science, and AI research. Key references include Oxford University research on brain vs. AI learning principles, analyses of working memory in humans vs. context windows in LLMs, studies of predictive coding in the brain versus transformer prediction, examinations of cognitive biases in humans and LLMs, and discussions in the literature distinguishing LLM "hallucinations" from human confabulation. Philosophical perspectives on embodiment and consciousness in AI vs. humans were drawn from works by Noë and others. These and other sources are cited throughout the text to provide evidence and further reading on the topics discussed.

References

[1] Study shows that the way the brain learns is different from the way that artificial intelligence systems learn | University of Oxford: https://www.ox.ac.uk/news/2024-01-03-study-shows-way-brain-learns-different-way-artificial-intelligence-systems-learn

[2] Study shows that the way the brain learns is different from the way that artificial intelligence systems learn | University of Oxford: https://www.ox.ac.uk/news/2024-01-03-study-shows-way-brain-learns-different-way-artificial-intelligence-systems-learn#:~:text=However%2C%20the%20biological%20brain%20is,knowledge%20and%20degrades%20it%20rapidly

[3] Brain-inspired replay for continual learning with artificial neural networks | Nature Communications:
https://www.nature.com/articles/s41467-020-17866-2#:~:text=Artificial%20neural%20networks%20suffer%20from,generative%20replay%20to%20complicated%20problems

[4] Study shows that the way the brain learns is different from the way that artificial intelligence systems learn | University of Oxford:
https://www.ox.ac.uk/news/2024-01-03-study-shows-way-brain-learns-different-way-artificial-intelligence-systems-learn#:~:text=In%20artificial%20neural%20networks%2C%20an,in%20turn%20speeds%20up%20learning

[5] What is a context window? | IBM:
https://www.ibm.com/think/topics/context-window#:~:text=An%20LLM%E2%80%99s%20context%20window%20can,for%20the%20model%20to%20proceed

[6] The Limits of Working Memory: Human Brains vs. AI Models - Illumio Cybersecurity Blog | Illumio:
https://www.illumio.com/blog/the-limits-of-working-memory-human-brains-vs-ai-models#:~:text=Artificial%20intelligence%20systems%2C%20especially%20Large,all%20increase%20an%20AI%27s%20capacity

[7] 14: Evidence of a predictive coding hierachy in the human brain listening to speech, Nature Human Behaviour - ML-Neuro - BayernCollab:
https://collab.dvb.bayern/spaces/TUMmlneuro/pages/69902439/14+Evidence+of+a+predictive+coding+hierachy+in+the+human+brain+listening+to+speech+Nature+Human+Behaviour#:~:text=Your%20mind%20likely%20jumped%20to,range%20and%20hierarchical%20predictions

[8] LLMs vs. Human Mind: Understanding the Creativity Gap.:
https://www.gofar.ai/p/llms-vs-human-mind-understanding#:~:text=might%20seem%20creative%20because%20it,truly%20novel%20ideas%20from%20scratch

[9] LLMs vs. Human Mind: Understanding the Creativity Gap.:
https://www.gofar.ai/p/llms-vs-human-mind-understanding#:~:text=Trying%20to%20get%20LLMs%20to,don%27t%20get%20excited%20about%20a

[10] (Ir)rationality and cognitive biases in large language models | Royal Society Open Science:
https://royalsocietypublishing.org/doi/10.1098/rsos.240255#:~:text=Do%20large%20language%20models%20,additional%20layer%20of%20irrationality%20in

[11] The Myth of Thinking Machines | Daily Philosophy:
https://daily-philosophy.com/zurkic-matthias-thinking-machines/#:~:text=No%C3%AB%20argues%20in%20Rage%20against,As%20No%C3%AB

[12] The Myth of Thinking Machines | Daily Philosophy: https://daily-philosophy.com/zurkic-matthias-thinking-machines/#:~:text=Goddu%2C%20No%C3%AB%2C%20and%20Thompson%20argue,sets%20its%20actions%20into%20motion

[13] 14: Evidence of a predictive coding hierachy in the human brain listening to speech, Nature Human Behaviour - ML-Neuro - BayernCollab: https://collab.dvb.bayern/spaces/TUMmlneuro/pages/69902439/14+Evidence+of+a+predictive+coding+hierachy+in+the+human+brain+listening+to+speech+Nature+Human+Behaviour#:~:text=However%2C%20this%20is%20not%20how,predictions%20made%20by%20our%20brains

[14] The Myth of Thinking Machines | Daily Philosophy: https://daily-philosophy.com/zurkic-matthias-thinking-machines/#:~:text=Our%20experiences%20are%20too%20complex,functioning%20brain%20outside%20a%20body

[15] What is a context window? | IBM: https://www.ibm.com/think/topics/context-window#:~:text=forgetting%20details%20from%20earlier%20in,for%20the%20model%20to%20proceed

[16] (Ir)rationality and cognitive biases in large language models | Royal Society Open Science: https://royalsocietypublishing.org/doi/10.1098/rsos.240255#:~:text=this%20question%20by%20evaluating%20seven,methodological%20contribution%20by%20showing%20how

[17] LLMs vs. Human Mind: Understanding the Creativity Gap.: https://www.gofar.ai/p/llms-vs-human-mind-understanding#:~:text=match%20at%20L50%20taking%20risks%2C,come%20up%20with%20new%20ideas

[18] Emergent Abilities in Large Language Models: An Explainer: https://cset.georgetown.edu/article/emergent-abilities-in-large-language-models-an-explainer/#:~:text=Explainer%20cset,and%20training%20data%20scale%20up

[19] LLMs vs. Human Mind: Understanding the Creativity Gap.: https://www.gofar.ai/p/llms-vs-human-mind-understanding#:~:text=Humans%2C%20on%20the%20other%20hand%2C,the%20unknown%2C%20taking%20risks%2C%20and

[20] Study explores the impact of LLMs on human creativity - LinkedIn: https://www.linkedin.com/posts/smw355_study-explores-the-impact-of-llms-on-human-activity-7257418509222060032-MNAH#:~:text=Study%20explores%20the%20impact%20of,creativity%20during%20assisted%20tasks%2C

[21] Human Creativity in the Age of LLMs - arXiv: https://arxiv.org/html/2410.03703v1#:~:text=Human%20Creativity%20in%20the%20Age,may%20inadvertently%20hinder%20independent

[22] (Ir)rationality and cognitive biases in large language models | Royal Society Open Science: https://royalsocietypublishing.org/doi/10.1098/rsos.240255#:~:text=cognitive%20psychology%20literature,additional%20layer%20of%20irrationality%20in

[23] (Ir)rationality and cognitive biases in large language models | Royal Society Open Science: https://royalsocietypublishing.org/doi/10.1098/rsos.240255#:~:text=irrationality%20in%20these%20tasks,methodological%20contribution%20by%20showing%20how

[24] (Ir)rationality and cognitive biases in large language models | Royal Society Open Science: https://royalsocietypublishing.org/doi/10.1098/rsos.240255#:~:text=pervasive,different%20definitions%20of%20what%20is

[25] The Myth of Thinking Machines | Daily Philosophy: https://daily-philosophy.com/zurkic-matthias-thinking-machines/#:~:text=Machines%20do%20not%20have%20a,purposes%3B%20their%20existence%20remains%20prearranged

[26] The Myth of Thinking Machines | Daily Philosophy: https://daily-philosophy.com/zurkic-matthias-thinking-machines/#:~:text=altering%20revelation,authentic%20engagement%20with%20the%20world

[27] The Myth of Thinking Machines | Daily Philosophy: https://daily-philosophy.com/zurkic-matthias-thinking-machines/#:~:text=Artificial%20intelligence%20models%20are%20neither,purposes%3B%20their%20existence%20remains%20prearranged

[28] The Myth of Thinking Machines | Daily Philosophy: https://daily-philosophy.com/zurkic-matthias-thinking-machines/#:~:text=remains%20an%20unattainable%20idealization,the%20product%20of%20psychological%20projections

[29] The Myth of Thinking Machines | Daily Philosophy: https://daily-philosophy.com/zurkic-matthias-thinking-machines/#:~:text=Goddu%2C%20No%C3%AB%2C%20and%20Thompson%20argue,they%20are%20our%20tools%2C%20constructed

[30] The Myth of Thinking Machines | Daily Philosophy: https://daily-philosophy.com/zurkic-matthias-thinking-machines/#:~:text=the%20authors%3A%20%E2%80%9CModels%20are%20tools%2C,sets%20its%20actions%20into%20motion